AN ANTHOLOGY OF PHILIPS RESEARCH

AN ANTHOLOGY OF PHILIPS RESEARCH

EDITED BY

H. B. G. CASIMIR AND S. GRADSTEIN

PUBLISHED BY

N.V. PHILIPS' GLOEILAMPENFABRIEKEN

ON THE OCCASION OF THEIR 75TH ANNIVERSARY

EINDHOVEN, NETHERLANDS, 1966

© N.V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands, 1966

INTRODUCTION

The 75th anniversary of the Philips Company provides a welcome opportunity to publish this volume which illustrates the scope of the work of our laboratories and surveys a number of important results. It deals primarily with work of the last fifteen years. Much of the earlier work is now only of interest if viewed in the context of the development of science and technology and we did not want to devote too many pages to history. On the other hand the work of a research organization is often long term work and to gain a proper perspective one has to consider a period of some length. Preparing an anthology of papers previously published in the Philips Technical Review was the most expedient way to reach the objects we had in mind.

It is regrettable, but probably unavoidable, that original publications are nowadays usually only understandable to a limited group of readers with specialized knowledge. Popularized science on the other hand is in continuous danger of losing the essential points and may thus easily become objectionable to the expert while remaining incomprehensible to the layman. From its early beginnings in 1936 the Philips Technical Review has tried to steer a middle course between these two extremes and to establish a level of treatment that does not require expert knowledge and great mathematical skill on the readers' part, although it may require a good deal of concentration and real interest. It is our firm belief that it is possible to describe at this level the essential features of our work and our results, even though the work itself could never have been done without application of more sophisticated methods.

The core of this book (part B) is formed by a selection of 17 articles, one for each year from 1950 onwards. Although the choice of 1950 as a starting date is somewhat arbitrary and partly determined by what we considered to be an acceptable size for the book, 1950 may also be regarded as a year in which the material disturbances and upheavals following the second world war had come to an end. We have tried to choose only articles dealing with fields where our laboratories have made important new

contributions; we have excluded accounts of technical realisations based on well-known techniques. Neither were purely didactic articles included in this selection, although they are an important feature of the Philips Technical Review. As a matter of fact we felt sometimes tempted to break our own set of rules, for some of these articles, especially those on semi-conductor devices, have generally been regarded as excellent introductions to their subject. It was gratifying that a relevant article could be found for each year. Also, more by good luck than by good management, we were able to touch upon the majority of subjects in which we have been so fortunate to pioneer, although this had often to be done by a description of later developments. We could not include Penning's original work on the "pig" (Philips ionization gauge) which was carried out before the war, but our series opens with the description of a further improvement of this device. Ferrites are represented by a spectacular application (B 5) which also brought new properties to light and hexaferrites by a paper on ferroxdure II and III (B 9). In this way justice could be done to our work in magnetism even though the real breakthrough is described in earlier publications. Something similar is true for the Stirling cycle, which came to a new span of life through the work of our laboratories in the forties: B 10 and B 16 describe further innovations of the hot gas engine and the Stirling refrigerator. A long series of investigations on thermionic emission finds expression through the final article of the series, which describes the most recent — and very important — advance in dispenser cathodes.

Of course our somewhat arbitrary method has its disadvantages too. We were able to touch upon most of the important subjects but obviously the number of pages is no indication of their relative importance. Semi-conductors and acoustics are both represented by one article. The article on semi-conductors (B 14) describes the alloy-diffusion technique for making transistors, a process originating from the Research Laboratories which has been used by Philips and its licencees for the manufacture of tens of millions of transistors. Acoustics is represented by stereo reverberation (B 7). This work gave rise to rather striking demonstrations and found an adequate measure of practical application but of course in recent years we have devoted more efforts to the study of semi-

conductors and semi-conducting devices than to work in acoustics. Also, the personal taste of the editors of this book has influenced the choice and since both of them are primarily physicists our selection may show a certain leaning towards phenomena, single devices and pieces of apparatus, rather than towards complicated systems. In this connection I should, however, like to call attention to the paper on △-modulation (B 3), a modulation method that is now finding wide-spread application in many telecommunication systems.

This selection of articles is supplemented in two ways:

It is preceded by five articles of a more general nature (part A). Two of these articles (A 1 and A 2) provide the necessary historical background and summarize work up to 1950, whereas two others (A 4 and A 5) try to give a more synthetic picture of our basic work in physics and chemistry. The remaining article of this series (A 3) expresses some of my personal views on the relations between science and industry.

It is followed by a third section of the book (part C) containing a number of articles on miscellaneous subjects most of them dating from the last ten years. Many of these articles deal with somewhat isolated results rather than with a whole body of work. Exceptions are the papers on welding techniques (C 1) and on bearings (C 8). Although some of the shorter articles describe important new inventions — we might specifically mention the superconducting dynamo (C 5) — it is in keeping with the festive occasion for which this book is published that others were chosen mainly for their artistic appeal. This playful spirit is even more apparent in some of the full page illustrations that are inserted as "extra's" in each of the three sections.

As a leader of a research organization that has to play its part in a world of ever faster scientific and technical progress I have at certain moments felt alarmed and almost discouraged, but looking at the present collection I cannot help feeling that our staff has contributed its share. These articles recall to us many of the highlights of our work. I hope that they will also convey at least some of this enthusiasm to other readers.

Eindhoven, May 1966. H. B. G. CASIMIR

VOL. 1 No. 1 JANUARY 1936

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, HOLLAND

INTRODUCTION

With the appearance of the first issue of this new periodical, the policy and aims should be outlined which will be pursued in these pages. The Philips Research Laboratories are continually receiving an everincreasing number of enquiries and requests from many quarters for more detailed data and particulars of the extensive range of Philips products, and especially for information as to their specific characteristics and practical applications. A large proportion of these enquiries comes from the engineering world and it is hoped by means of this periodical to establish permanent contact with these circles.

This journal will deal with the following items:

a) Technical descriptions of new Philips products, and reports on investigations relating thereto,

b) Information concerning the applications of these products, and

c) Technical articles of a more general nature on various questions relating to the above, which may be of interest to the reader.

An endeavour will be made to present the subject matter of the articles as simply as possible. Mathematical treatment will be resorted to neither more nor less than is necessary. Facts which can be concisely stated by mathematical formulae will in addition be explained in simple language, which will make it possible for those who do not wish to follow the mathematics to gather the gist of the article from the text alone. Additional clarity will further be strived for by the free use of diagrams and graphs. In this way every unnecessary show of learnedness will be avoided, but on the other hand care will be taken that this simplicity of expression does not lead to superficiality and indefiniteness when dealing with really complicated subjects.

Each contribution to this periodical will be, as far as possible, complete in itself. Owing to the wide range of products manufactured by Philips every article cannot be expected to command the same measure of interest from every reader. Nevertheless, it may be expected that many readers wil find matter of interest also in those contributions which do not concern their own immediate activities.

This periodical therefore will be not merely a journal embracing the activities of the Philips organisation nor a technical journal on popular lines, but a source of information of value and interest to the whole engineering profession. Our hope is that this publication will prove of practical interest and use to many, and in consequence merit a wide circle of readers.

G. HOLST.

NOTE FOR THE READER

The three parts of this book — easily distinguishable on perusing its pages — are each preceded by its own list of contents. Volume and page of the Philips Technical Review where the articles were originally published are indicated in the lists. The texts are interspersed with a number of one-page contributions for practical reasons (and, incidentally, in order to make browsing more attractive). These contributions, which also have been reproduced or abstracted from publications in Philips Technical Review, are listed separately at the end of the book.

For readers who are not acquainted with the Philips publications it may be mentioned that Philips Technical Review was founded in 1935, mainly on the initiative of Professor Dr. G. Holst, then the Director of Philips Research Laboratories, Eindhoven. The policy of this journal, laid down in the Introduction to its first issue (see opposite page), has remained substantially unaltered. Another journal, Philips Research Reports, which is of a strictly scientific nature, has been published by the Eindhoven laboratories since 1946.

It should, however, be pointed out that even the combined volumes of these two journals do not give a complete picture of our research activities and results. The bulk of the work is published as Conference papers or through the normal outlets — scientific and professional journals. These papers contributed by Philips Research Laboratories, Eindhoven, as well as by Mullard Research Laboratories, Redhill (England), by the Aachen and Hamburg laboratories of Philips Zentrallaboratorium GmbH, by the Laboratoires d'Electronique et de Physique appliquée, Limeil-Brévannes (France) and by MBLE Laboratoire de Recherche, Brussels, are regularly listed on the last pages of the issues of the Philips Technical Review and reprints of most of them can be obtained from the Eindhoven laboratories.

Specialized journals are in addition published by several Philips Divisions, for instance Philips Telecommunication Review, Electronic Applications, Medicamundi, and by affiliated Companies and Laboratories, for instance Mullard Technical Communications, Acta Electronica, etc. Books on many topics have been written by Philips authors, and many of these appear as volumes of the Philips Technical Library or of other series under the auspices of Centrex Publ. Co. at Eindhoven.

LIST OF CONTENTS

page

INTRODUCTION 5

NOTE FOR THE READER 9

ANTHOLOGY

A. HISTORY AND BACKGROUND 13

B. REPRESENTATIVE PAPERS 115

C. MISCELLANEOUS CONTRIBUTIONS 327

LIST OF SHORT ITEMS 469

A. HISTORY AND BACKGROUND

		page
1	**Scientific research of Philips' Industries from 1891 to 1951** W.de Groot (Vol.13, p.3)	15
2	**Fifty years of the gas-filled lamp** J.C.Lokker (Vol.25, p.2)	60
3	**Science and industry** H.B.G.Casimir (Vol.20, p.85)	75
4	**Some main lines of 50 years of Philips research in physics** H.B.G.Casimir (Vol.24, p.341)	81
5	**Contributions of the Philips Research Laboratories to solid-state chemistry** H.J.Vink (Vol.24, p.364)	94

SCIENTIFIC RESEARCH OF PHILIPS' INDUSTRIES FROM 1891 TO 1951

INTRODUCTION

With the invention, at the beginning of the nineteenth century, of the voltaic pile and the galvanic battery, providing a source of current with an almost unlimited voltage and a relatively small internal resistance, in principle also the possibility of electric light was created. In 1808 *Davy* succeeded in passing current through air between two carbon rods of a few millimetres diameter. The tips of the carbons and the gas discharge radiated a bright light. Upon the carbon tips being spaced a couple of centimetres apart, owing to convection the discharge assumed the shape of an arc (an inverted U), which name came to be used to describe all gas discharges with a high current density and with a thermionic cathode.

In a certain sense the carbon arc is the precursor of modern discharge lamps and, in so far as the light originates from the glowing carbon tips, also of the incandescent lamp. A second forerunner of the incandescent lamp is the platinum wire brought to incandescence by an electric current, an experiment which was also carried out by *Davy* (1802). Carbon-arc lamps were known about 1850 and, as was subsequently disclosed in the famous law-suit between *Edison* and *Goebel* (1893), also electric incandescent lamps had already been made at that time. These sources of light were fed from galvanic batteries. It was not, however, until the invention of the dynamo (*Siemens* 1866, *Gramme* 1869) that electric light came to be introduced on any large scale. In 1879, on the one hand the differential-arc lamp of *Von Hefner-Alteneck* made its appearance and, on the other hand, *Edison's* carbon-filament lamp, which was demonstrated at the Paris exhibition of 1881, together with the machines for generating the current and the means of distributing it. That exhibition aroused interest everywhere, and it was shortly after this that *E. Rathenau* founded the A.E.G. in Germany.

All this greatly interested the young Dutchman Gerard Leonard Frederik Philips, born 9th October 1858 at Zaltbommel, who was studying at the Polytechnical School (now the Technical University) at Delft, where he graduated as an engineer in 1883. Shortly after that he was given the opportunity to install electric lighting in some ships at Glasgow, and there he made the acquaintance of *William Thomson (Lord Kelvin)*, who was likewise greatly interested in electric light. Later on he conducted negotations for *Rathenau* in Berlin between the A.E.G. and the municipality of Amsterdam about an electric-lighting project, but these negotiations broke down on account of the high price per kWh (fl. 0.60) asked for by the A.E.G.

In 1890, when 6800 arc lamps and 118,000 incandescent lamps were already burning in Paris, four electric power stations in Berlin were supplying current to some 3000 arc lamps and 70.000 incandescent lamps, and in London a generating station was being built for feeding 600.000 lamps of 10 candle power, whilst in the U.S.A. 23,500 arc lamps and 2,800,000 incandescent lamps had already been installed, Ir. Philips conceived the idea of starting a new incandescent-lamp factory in the Netherlands, where others had already been established, among which were the *Pope* and the *De Kothinski* works. This plan materialised in the month of May 1891 with the opening of the factory at Eindhoven. In the course of time G.L.F. Philips and his younger brother Anton Frederik Philips, who joined the firm in 1895, turned this small factory, with a starting capital of 75,000 guilders, into a world-wide concern, the parallel of which is but scarcely found.

It is not the intention to enter here into the history of this concern, for this will be found in a book which is to be published shortly. Instead of that, a review will be given of the development of the scientific research, both in the pure and in the applied sciences, of the Philips Industries and in particular that which has been carried out in the Physical Research Laboratory founded in 1914.

The first research work connected with the Philips business was carried out by Gerard Philips himself. For some time before the factory was started he was studying the preparation of the carbon filament and various other processes in a primitive workshop at home. In November 1890, when writing to one of the many people with whom he was negotiating for the establishing of the new factory, Ir. Philips wrote: "I am able to produce perfectly homogeneous cellulose filaments on a business scale". Considering, however, what the management of a rapidly growing young industry involved, and bearing in mind that up to 1895 both the commer-



Ir. G. L. F. Philips engaged with his first experiments for the manufacture of carbon filaments from cellulose (1890).

cial side of the business and the management of the works were in the hands of one man, obviously this research work was at first limited to what was absolutely essential. Technical and scientific help soon became necessary, especially when from 1903 onwards new materials, such as osmium, tantalum and tungsten, the last of which was destined to be the filament material of the future, came to be used in the manufacture of incandescent lamps. About 1907, at the time that incandescent lamps began to be made with squirted tungsten wire, P. N. L. Staal, A. de Broekert and H. Gooskens*) joined the firm as Gerard Philips's assistants, these being followed, respectively in 1908 and 1909, by the chemical engineers J.C. Lokker and A. de Graaff, and shortly afterwards by the mechanical engineers H. de Jong, H. Reufel and W. Koning.

In December 1911 the brittle squirted wire was successfully replaced by the more rigid drawn wire.

Thus a technical staff of some size was formed, comprising people trained in mechanics and in chemistry, but it was not long before also the need of physicists began to be felt when, in 1913, a new development was announced in the lamp-making world, in the form of the gas-filled lamp.

In the General Electric Company's laboratory at Schenectady (U.S.A.), where *Edison*'s work was carried on and physicists were available, in 1909 *Coolidge* had found an entirely new method for drawing strong, thin wires of tungsten. There soon followed, in 1913, in the same laboratory, *I. Langmuir's* invention whereby a coiled filament could be brought to incandescence in an inert-gas atmosphere, thereby counteracting vaporization of the tungsten and thus making it possible to heat the filament to a higher temperature, so that the luminous output of the lamps could be greatly increased. Within a very short space of time a number of other inventions followed from the same laboratory, such as X-ray tubes with heated cathode and gas-filled rectifying tubes.

In November 1913 also Philips brought gas-filled tungsten lamps on the market, under the name of "half-watt lamps". In the manufacture of these lamps and also, for instance, in their photometry so many problems arose which lay more in the domain of physics than in that of chemistry or mechanics that Ir. Philips decided to engage a physicist. Thus on 2nd January 1914 G. Holst joined the firm in that capacity, and this was the beginning of the



Prof. Dr. G. Holst, after a painting by *S. Schröder*.

*) In this review the names of people connected with the Philips Concern are spaced out, while those of others are printed in italics.

foundation of a physical laboratory. Some months later E. Oosterhuis joined the staff of this laboratory as second physicist.

It will be shown how the work of the laboratory established in 1914, starting with the problems of the manufacture of incandescent lamps, has expanded into such a multifarious programme, from which such a wide variety of other products have been born. In this growth three main periods are to be distinguished:

1) 1914-1923, the period which saw the first world war and during which the laboratory was located in a part of the lamp factory;

2) 1923-1940, beginning with the opening of a new laboratory specially equipped for the purpose (in 1929 it was considerably enlarged) and ending with the occupation of the Netherlands by enemy forces in the second world war;

3) 1940 up to the present day, covering the years of occupation and the revival of activity after the country's liberation.



Dr. E. Oosterhuis, after a painting by *S. Schröder.*

THE LABORATORY OF 1914 - 1923

After studying at the Technical University at Zürich — where he took his doctor's degree in 1914 — Dr. Holst worked for a number of years in the laboratory of *H. Kamerlingh Onnes* at Leyden, where, inter alia, he took an active part in the discovery of superconductivity. Thus, in addition to his bent for purely scientific work, Dr. Holst was also technically interested in taking up his new appointment with the Philips' laboratory. Although it was considered as his main task to study the physical questions arising in the manufacture of the tungsten lamp, it was felt at the same time that it should not be left at that, but that it was necessary to penetrate to the very roots of the phenomena to be studied. From the tackling of subjects on such a wider basis physical science was primarily benefited, but then this in turn bore fruit for technical science, often in a surprising manner, in the form of improvements of existing products and the opening up of new fields of activity.

The study of the tungsten lamp was therefore the starting point: first the behaviour of tungsten wire in the processing, and further its behaviour upon being heated by an electric current to a high temperature in a glass bulb either in vacuo or in the atmosphere of an inert gas; the current had to be passed into the bulb via vacuum-tight and fused-in wires.

Owing to the extensive programme of work the staff soon had to be enlarged. Dr. Holst received successively the assistance of P. G. Cath, S. Weber, C. Bol, H. C. Burger, G. Hertz, Balth. van der Pol, A. Bouwers, A. E. van Arkel, W. Geiss, J. H. de Boer, P. Clausing, R. Vermeulen, W. de Groot, C. Zwikker and others, whilst contributions towards this research were likewise furnished by L. Hamburger, D. Lely and J. A. M. van Liempt, who belonged to the chemical staff of the works.

As regards the physical problems directly related to the incandescent lamp, first of all there was the photometry to be studied, particularly of the gas-filled lamps, since their luminous flux is distributed in space in a less surveyable manner than that of vacuum lamps with a linear filament. In a series of "Communications of the Philips' Laboratory" (1918-1919), which in a sense may be regarded as the forerunners of this journal,

conceptions in the field of illumination were explained and the basic elements of utility lighting, fittings, projectors, car lamps, sunlight lamps, etc. were dealt with.

Of a more fundamental nature was the investigation of radiation, such as the spectral energy distribution of incandescent lamps. At that time *Planck's* formula was still regarded with some scepticism and there was no thought of an "international relative luminosity curve" or an "international temperature scale". Much doubt existed as to the constant c_2 in *Planck*'s formula, which it was important to know in connection with optical pyrometry.

Owing to the gradual decline in the luminous intensity of the incandescent lamps, and particularly of the vacuum lamps, due to precipitation of tungsten on the inside of the bulb, a matter which will be reverted to in the next section of this review, the study of such thin layers of metal and of the chemical processes taking place in residual gases was taken in hand (Hamburger).

In connection with the filament problems, the electrical conductivity and the thermal conductivity of graphite and tungsten were investigated, as also, in connection with the gas filling, the thermal conductivity of gases and the vaporization of tungsten in a gaseous atmosphere (Weber). Further, following upon the investigations which Holst had himself carried out and had seen conducted by others in the cryogenic laboratory at Leyden, extensive tests were made with the argon-nitrogen system, the mixture of gas used for filling the half-watt lamps.

As far as the metallurgy of tungsten is concerned, with the aid of the then new methods of X-ray spectroscopy and X-ray diffraction a start was made with the study of the structure of tungsten and other metals (Burger, Van Arkel), of the physical (e.g. elastic) properties of tungsten (Geiss) and of the chemistry of tungsten compounds (Van Liempt and others).

Of importance, for its future consequences, was the investigation of electrical conductivity in gases and their emission of light when an electric current is passed through them. This came about as a result of the method applied in the factory for testing the degree of evacuation in vacuum lamps and the purity of the gas in gas-filled lamps with the aid of a *Tesla* discharge (spark test). Another, no less important, reason for this investigation being started was the occurrence of electrical breakdown between the extremities of the filament in gas-filled lamps, both when the lamps were burning and when the filament was broken.

To combat this breakdown it was necessary to make a thorough study of the passage of electricity through gases. The investigation of gas-discharge phenomena, already studied by *Faraday* and more intensively by *Plücker, Hittorf, Crookes, H. Hertz, Lenard, J. J. Thomson, Townsend* and others, reached a new stage when it appeared that these phenomena could be related to *Bohr*'s new theory of the atom (1913). This theory, just as was the case with *Planck*'s quantum theory, was at first received with a certain reserve. Moreover, the outbreak of the 1914-1918 war held back more or less the propagation of *Bohr*'s ideas. Of importance for the laboratory in this connection were the lectures given in Eindhoven by *P. Ehrenfest,* since 1912 professor in theoretical physics at the Leyden University and a keen advocate for the ideas held by *Bohr,* whilst in 1920 the laboratory staff was supplemented by G. Hertz, who in 1914, together with *J. Franck,* had established the relationship between *Bohr*'s theory and the phenomena of electric conductance in gases, for which, in 1926, they were both awarded the Nobel prize. Hertz's investigations will be discussed in the next section.

The most important result of the investigation of gas discharges in the period 1914-1923 was the deeper insight thereby gained into the phenomena of electrical breakdown, in particular at low pressure, in rare gases between plane, parallel, cold metallic electrodes, and the deeper knowledge gained of the transition from the non-self-sustained to the self-sustained discharge (Holst, Oosterhuis). One of the facts thereby established was that the production of electrons in the glow is due to the action of the positive ions upon the cathode and not, as *Townsend* imagined it to be, as a result of direct ionization of the gas by the positive ions.

The first practical outcome of the investigation into gas discharges was the appearance in this period of the neon glow lamp (1917) and the tungsten arc lamp with neon-filling (1920).

The investigation of rare gases was facilitated by the fact that since 1916 the Philips' works had been making (under the guidance of H. Filippo) their own liquid oxygen and nitrogen, from which the rare gases could be distilled.

Since during the war the importation of X-ray tubes was stopped medical practitioners in the Netherlands sent their tubes to the Philips' works for repair. Thus interest also came to be taken in these tubes, not only in respect to the gas-filled

tubes that were then being used but also in the new Coolidge vacuum X-ray tube with a heated tungsten filament as electron source.

The study of the manner in which the metal leads could be passed through the glass of the lamp bulb led to the production of ferrochromium alloys which have about the same coefficient of thermal expansion as that of glass and which can be fused onto glass. Prior to this, the leads were of platinum wire, later of iron wire coated with copper; for other glass-metal joints pure copper was used, which differs in expansion from glass but owing to its softness changes shape while cooling. With the new alloys (Bol, Bouwers, B. Jonas) there was no need to make allowance for any deformation during cooling, so that the leads and connections could be made much stronger and fusing-in was no longer confined to thin-walled tubes. This led to the construction of metal transmitting tubes and metal X-ray tubes. In connection therewith it was of importance that in 1916 Philips had started a glass works of their own (P. J. Schoonenberg).

In 1883 *Edison* had discovered that an electric current flowed through the vacuum between the poles of the filament in his incandescent lamps, a fact which we now know to be caused by electron emission. Following upon *Richardson*'s investigation of electron emission, in 1904 *Fleming* had invented the diode and in 1907 *Lee de Forest* had made the first triodes by adding a grid. Simple receiving triodes were produced by Philips in 1917, these being followed by transmitting triodes and diodes for rectifying alternating currents. Soon a systematic study of the phenomena in the radio valve was started. In 1922 Balth. van der Pol (a former pupil of *H. A. Lorentz*) was charged with radio research*).

In connection with the manufacture of radio valves, whereby use was made of the electron-emitting properties of tungsten, further investigations were carried out with a view to finding materials which could replace tungsten and give a greater emission for a smaller filament power. In 1921, as a result of new ideas about the theory of atoms, *Coster* and *Hevesy* discovered the element hafnium, related to zirconium. Great expectations were held about the thermionic emission of this element, and the preparation and study of this new material was energetically taken in hand in the Philips' laboratory. Though this did not culminate in the important results that were expected of it, the chemical and metallurgical experience thereby gained was useful in many respects.

The ever-increasing amount of research work to be carried out demanded more space, so that in 1922 it was decided to build a new laboratory, which was completed and taken into use in 1923.



Memorial tablet in the hall of the Research Laboratory presented to Ir. G. L. F. Philips on the occasion of his 50th anniversary as Engineer in 1933.

On 1st April 1922 Dr. Ir. G. Philips, who during the first world war had had the honorary degree of doctor in the technical sciences conferred upon him at Delft, resigned his co-directorship at the age of 63. He died at The Hague on 26th January 1942.

THE NEW LABORATORY 1923 - 1940

The new laboratory, situated in a part of Eindhoven which in 1923 was still on the outskirts of the town, was taken into occupation in November that year by a staff of 15 graduated physicists, chemists and engineers with about 20 assistants, instrument-makers, glass-blowers, etc. By 1939 there were 106 scientists and 360 assistants. The number of publications issued by the laboratory in the period 1914-1923 was about 150 and in the period 1923 - 1940 about 1500. A review of research done in this latter period must therefore be limited to the main lines and some outstanding points.

The work begun in the old laboratory was continued in the new under more favourable circumstances. With the larger space available and the increase of staff there was more opportunity for deeper and more far-reaching investigations. It should be mentioned that Dr. Holst always left his scientific staff a high degree of freedom.

*) Since 1949 Prof. Dr. Balth. van der Pol has been Director of the Comité Consultatif International des Radiocommunications at Geneva.



It is impossible, in this short survey, to give a proper account of the various ways in which Dr. Holst stimulated his many co-workers. The few places where Holst is explicitly named here, certainly give an imperfect impression in this respect. The same applies to Dr. Oosterhuis, who (until 1946) held the position of Vice-Director and supervised a team of workers engaged in practical radio research.

Of course research work was still bound to a certain extent to the factories and their production. The laboratory was expected not only to evolve new ideas leading to new products or to the improvement of existing ones — ideas which were suitable for patenting — but these ideas had also to be brought to fruition for manufacture. Further, a certain amount of service was called for, such as the calibrating of measuring instruments, examination and testing of materials, etc. Thus, side by side with the purely scientific work, more and more work of a different nature had to be done. This is the reason why the number of publications issued year by year did not grow in proportion to the growth of the staff.

For a better comprehension of the vastness of the laboratory work this has been divided into five main groups, viz:

I. Light and the production of light, including gas discharges.

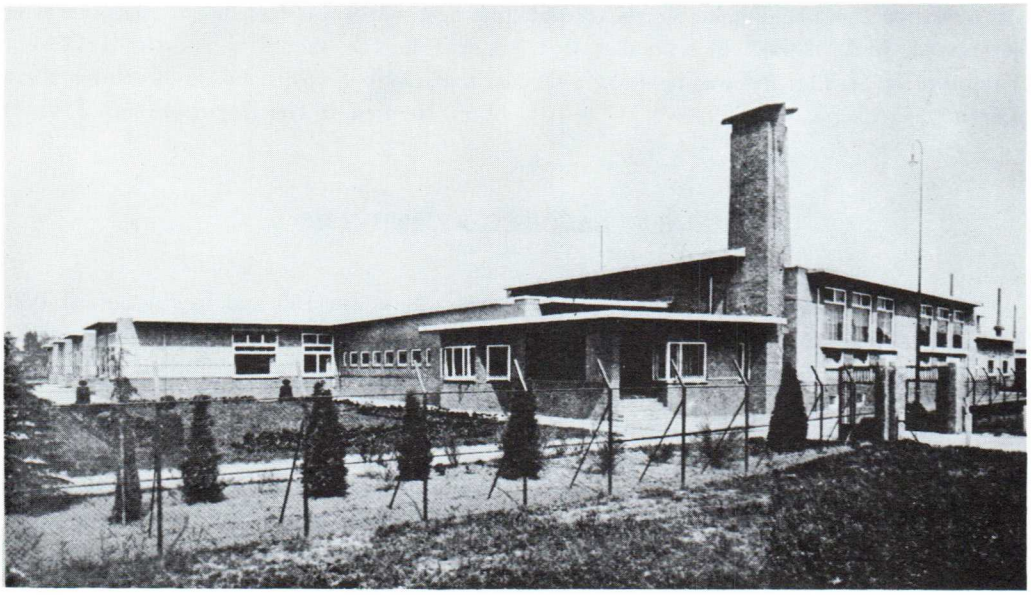
II. Electrotechnics and radio, including acoustics.

III. Chemistry, including metallurgy.

IV. X-ray investigations.

V. Mathematics and mathematical physics.

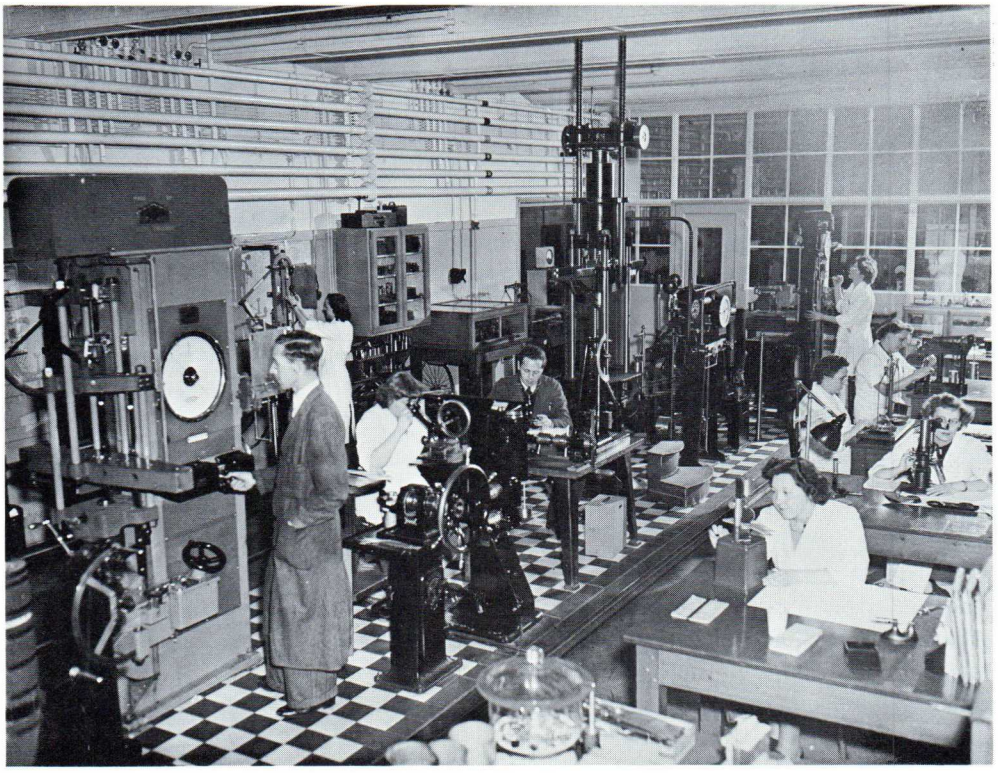
These main divisions have to be taken broadly, and furthermore there are important connections between them. The X-ray examination of crystals, for instance, forms a link between the groups III and IV, the examination of magnetic materials links up groups II and III, whilst the problems coming under the heading V mostly arise from other groups, especially from group II.



The Physical Research Laboratory in 1923.

I. Light and gas discharges

Investigations in connection with incandescent lamps, as far as the new laboratory is concerned, were more or less terminated about 1925 by Zwikker's thesis on the physical properties of tungsten as functions of temperature, which work can be placed side by side with similar work in the U.S.A. In the lamp factory, where a physico-chemical laboratory continued to be maintained, for many years much important work was done by Van Liempt, Geiss and others in the field of the chemistry of tungsten compounds and the metallurgy of tungsten.



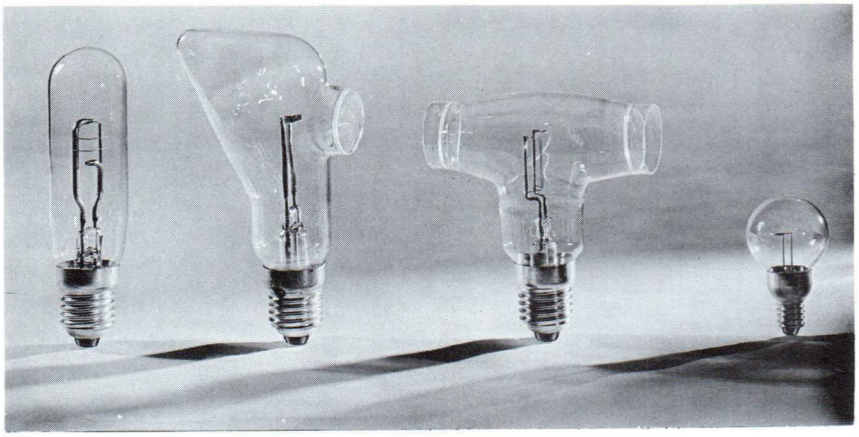
Department for testing materials used in the Philips' Works.

The investigation of gas discharges, which had been begun by Holst, Oosterhuis and Hertz in the old laboratory, was continued in the new one on a wider scale. Hertz had worked out methods for accurately measuring excitation and ionization potentials, mainly of rare gases, and established a relationship between these quantities and the spectrum (term diagram). The resonance lines of all rare gases were photographed with the vacuum spectograph and their wavelengths accurately determined. Furthermore, following up the investigations of Holst and Oosterhuis, the low-tension arc was investigated, when it appeared that an arc discharge can take place in a rare gas when there is a potential difference between anode and cathode lower than the lowest excitation potential. Other investigations concerned the negative glow discharge, especially the heating of the anode when electrons enter it, and the thermal effects arising at the electrodes of the tungsten arc lamp. Important work was by done by F. M. Penning in regard to electrical oscillations in a D.C. discharge in mercury vapour of low pressure; in deviation from *Langmuir's* findings (1925) it was proved that abnormal velocities of electrons in such a discharge are due to the said oscillations.

For Hertz's experiments an equipotential cathode was used which had been coated with barium oxide, obtained by oxidation of barium applied to the cathode in the form of an azide and then thermally decomposed. The experience gained with these oxide-coated cathodes was of direct use for the manufacture of radio valves. Tungsten as electron-emitting substance was very soon replaced by materials (dull-emitters) which already give adequate thermionic emission at a lower temperatue.

Oxide-coated cathodes soon came to be used also for rectifying tubes filled with argon and mercury and for gas-discharge lamps (mercury lamps, neon tubes).

The study of the spectra of rare gases had revealed that the atoms of these gases, namely of argon and neon, may be in an excited state which cannot be transformed into the ground state by radiation, such a state being called a metastable state. After the departure of Hertz in 1925, H. B. Dorgelo, who in collaboration with *Burger* and *Ornstein* had been studying in Utrecht the intensity rules for multiplets in spectra, which study was continued in Eindhoven, took up the study of metastable atoms.



Tungsten-ribbon lamps made in the laboratory to serve as light sources for various optical methods of measuring. The second lamp from the left has been so shaped that the light reflected by the bulb does not pass through the (plane-parallel ground) window and thus does not interfere with the measurement. The third lamp from the left has two flat windows placed somewhat obliquely, this being useful for optical pyrometry.

These atoms are capable of absorbing and re-emitting certain spectral lines which the normal gas allows to pass through unhindered. With the aid of this absorption the lifetime of metastable atoms was investigated with respect to the influence of temperature and gas pressure.

Owing to their energy of excitation, in a gas mixture metastable atoms of one atomic kind (say neon) may ionize other atoms (say of argon), such being possible under the condition that the ionization potential of the second gas is lower than the potential corresponding to the metastable levels of the first gas. In such gas mixtures various factors, such as the breakdown voltage, are greatly influenced by small amounts of the readily ionizable component.

This was further investigated by Penning. The fact that the lowering of the breakdown voltage is due to the action of metastable atoms was proved by demonstrating that the breakdown voltage of neon, lowered by the addition of argon, rose again when the neon was irradiated with light which is absorbed by the mestastable atoms, as a consequence of which these atoms, while emitting resonance radiation, show a transition to a non-metastable state and then to the ground state. It was also found possible by this means to raise the voltage drop of a positive column in a mixture of neon and argon by irradiation with neon light, and even to quench the discharge if the voltage applied is low enough.

The effect of the ionization through metastable atoms is also apparent when investigating the breakdown voltage V_s between plane parallel plates as a function of the product p_0d (p_0 = pressure reduced to 0 °C, d = distance between electrodes), thus determining the *Paschen* curve. Instead of just one minimum (optimum ratio of ionization and excitation) this curve then shows two minima, owing to an increase of metastable atoms and thus the formation of "secondary" ions accompanying increased excitation.

Another anomaly of the *Paschen* curve occurs at values of $p_0d < (p_0d)_{min}$. Normally V_s increases monotonically with decreasing p_0d, or, what amounts to the same thing, p_0d decreases monotonically with increasing V_s. In the case of helium it is remarkable that p_0d as a function of V_s shows a minimum and a maximum, so that in a certain range of p_0d values three critical values of the potential are found, V_1, V_2 and V_3. Breakdown only takes place if the applied voltage V_a answers to $V_1 < V_a < V_2$ or $V_a < V_3$. The same anomaly is found when in the case of stronger currents the voltage is investigated as a function of the current.

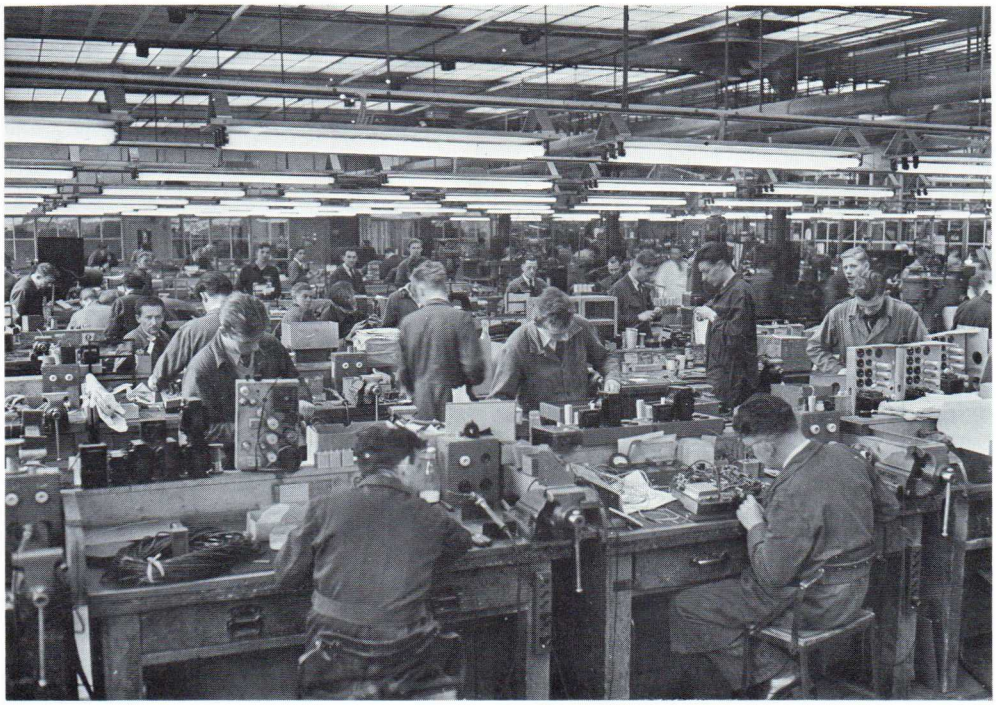
The study of the relation between the current i and the voltage V in discharges, as described above, led also to interesting data being collected in regard to the stability of gas discharges, a subject which

was further investigated later in the Laboratory for Technical Physics at Delft University by Prof. *H. B. Dorgelo*, *Chr. van Geel* and *C. Verhagen*.

Finally there were interesting investigations into the influence of magnetic fields upon discharges. Apart from the fact that the discharge as such forms a "current conductor" which in the magnetic field is subject to a force perpendicular to the current and to the magnetic lines of force, there is the influence of the magnetic field upon the paths of the individual electrons. The latter is particularly manifested at low pressures, where the electrons have a long free path. Following upon Penning's investigations into this field a vacuum meter was constructed in which the current in a gas placed in a magnetic field serves as a measure for the pressure. With this meter direct readings can be taken of pressures between 10^{-5} and 10^{-3} mm Hg and, for example, the improvement of the vacuum in pumping installations (such as in the case of the electron microscope and the cyclotron) can be followed from one minute to the next.

We must now consider the development of gas-discharge lamps. In 1923, following *Claude's* example (1910), the manufacture of neon tubes for advertising purposes was begun. These tubes had (cold) iron electrodes. Instead of neon, which gives a typical red light, other rare gases, such as helium, were used and also mixtures of a rare gas and mercury vapour. The colours could be given more variation by using coloured glass for the tubes. Later on, also fluorescent glass was used and fluorescent powders were applied to the inside of the tubes so as to produce new colour effects. With the introduction of the oxide-coated cathode, already applied for rectifying valves, it was possible to make neon tubes for stronger currents and a lower voltage, which came to be used as beacon lights for airfields and as light sources for special purposes (irradiation of plants).

Scientific research kept pace with this development. As is known, in the case of a discharge in an elongated tube *Faraday* had already distinguished a cathodic part (glow discharge) and an anodic part (positive column), the two being separated by „*Faraday*'s dark space". It is the positive column that produces the light in neon tubes, the glow discharge at the cathode being of no importance in this respect. The positive column was thoroughly investigated, theoretically as well as empirically, by M. J. Druyvesteyn, after *Schottky* had yielded an important contribution to the theory of this discharge. An important concept is that of the electron temperature, which characterizes the velocity distribution of the electrons in the column and can be measured with the well-known probe method of *Langmuir*. Druyvesteyn was able to deduce that in the absence of cumulative processes the electron temperature is, to a first approximation, proportional to the ionization potential of the gas.



Central workshop of the Physical Research Laboratory.

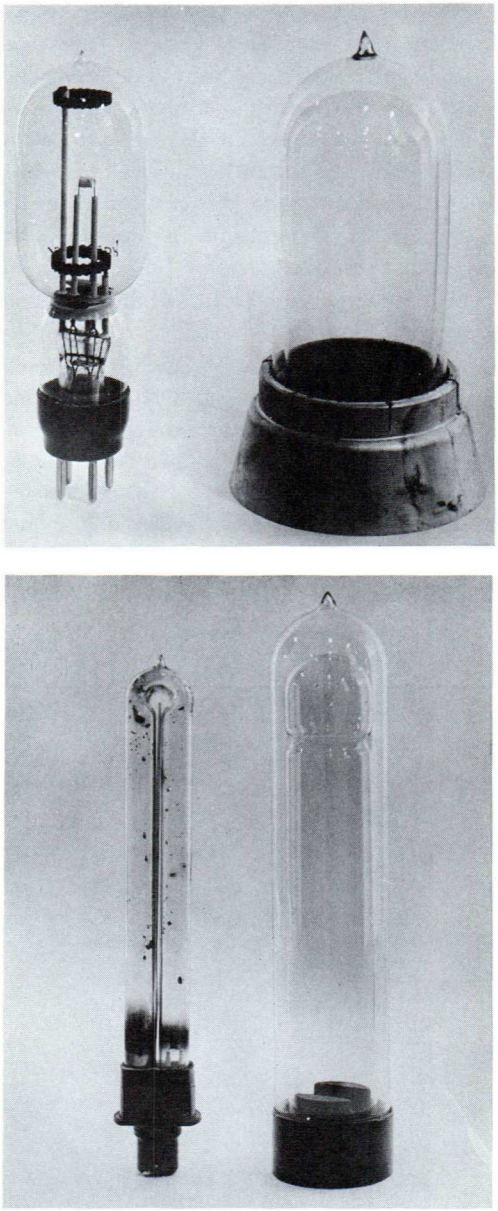
Among the discharge lamps with a filling of rare gas and metallic vapour the sodium lamps occupy a special place. It is well worth while outlining the development of the sodium lamp in the Philips' laboratory. Discharges in sodium vapour had already been investigated and used. In 1919, for instance, such a discharge was described by *R. J. Strutt (Lord Rayleigh*'s son). In 1923 *A. H. Compton* and *Van Voorhis* (Westinghouse) investigated a large number of gases and vapours for their ability of producing light and thereby found also the high light-yielding capacity of sodium vapour. They even patented a special kind of glass resistant to sodium, but this did not lead to any useful lamp being developed.

In 1925 *G. Hertz* demonstrated in the laboratory a low-tension arc in sodium vapour in a bulb of about 7 cm diameter. The bulb was fitted with an oxide-coated cathode and an anode, and the sodium was introduced into the bulb by electrolysis through the glass wall, with the cathode acting as negative pole and a bath of molten $NaNO_3$ as positive pole (*Warburg*'s method, 1890). This lamp was placed in a furnace with a temperature of about 250 °C in order to maintain a sufficiently high vapour pressure of the sodium.

Attention was again drawn to sodium as a source of light when G. Zecher set out to improve, in the Philips' laboratory, the yield and the colour of the light of neon tubes by introducing a small amount of lithium into the tube. This was an obvious means, since from the works of *Kirchhoff* and *Bunsen* it had become known that lithium gives a flame a bright red colour. However, something quite unexpected happened: the lithium attacked the wall of the tube and thereby released sodium, so that a sodium lamp was formed unintentionally. This lamp and a number of similar tubes, in which sodium was purposely added to a rare gas, have been used as polarimeter lamps in the laboratory for a number of years.

Meanwhile investigations conducted by *Pirani* in the Osram works had shown that under favourable conditions more than 95% of the energy absorbed by the column could be converted into sodium light, and the Osram works also brought a polarimeter lamp on the market.

At that time the investigations carried out in Philips' laboratory had turned in another direction. The use of the mercury lamp as a source of light for medical ray treatment had led to a closer study being made of the connection between vitamin D and rachitis. It appeared that vitamin D could be produced by irradiating ergosterine with light of a wavelength between 2800 and 2900 Å, which could be produced by a mercury lamp or a magnesium spark. This complexity of phenomena was investigated by E. H. Reerink and A. van Wijk. The demand arose for a special source of light for the production of vitamin D and for anti-rachitic ray treatment. Magnesium proved to be the most suitable element for this. This led to the construction of a low-tension arc lamp with a mixture of rare gas and magnesium vapour; the magnesium



An old sodium lamp (for direct current) and a lamp for connection (via a choke) to 220 V alternating current, each with its vacuum envelope. The D.C. lamps, in groups of 30 to 40 connected in series and fed from a rectifier, were used for the first installations for road lighting with sodium lamps (1932).

was contained in the anode and evaporated into the discharge.

This opened the way to the construction of other metallic-vapour lamps by the same means. In connection therewith, for instance, sodium was investigated anew and this led to the production of a low-tension arc lamp fed with direct current and filled with a mixture of neon and sodium vapour, which has a favourable luminous efficiency (50-60 lumens per watt) (De Groot and E. G. Dorgelo). In order to give the sodium sufficient vapour pressure the bulb was thermally insulated by placing it inside another, evacuated, bulb. Experimental lighting systems installed in a corridor in the laboratory and later on in one of the roadways on the factory site soon demonstrated the excellent qualities of sodium light for road-lighting purposes. After the construction of the lamp had been improved and it had been given the form of a single bulb contained in a *Dewar* flask, these experiments were continued on a more extensive scale, inter alia in cooperation with Prof. Dr. *H. C. J. H. Gelissen*, on stretches of road forming part of the Dutch network of highways (1932). Meanwhile Druyvesteyn and W. Uyterhoeven had been further investigating the positive column in mixtures of sodium and rare gas, whereby it was found that this discharge, in a tube placed in a *Dewar* flask, is less susceptible to changes in the ambient temperature than the low-tension arc lamp. It is to the merit of Bol, who had also given the low-tension arc lamp a shape suitable for manufacture, that the A.C. column lamp with vacuum envelope was developed into a practical unit which for the greater part can be manufactured mechanically, also as far as the glass envelope is concerned. The discharge tube proper has the shape of an elongated U. By 1940 more than 100,000 of these lamps had been installed. Abroad too, where similar lamps had likewise been developed, numerous lighting projects were carried out with sodium lamps, some of which by Philips.

Lighting with mercury lamps has also to be discussed. Compared with sodium lighting, at first Philips took little interest in mercury lighting. Since the investigations of *Küch* and *Retschinsky* (1907), which resulted in the appearance of the quartz-mercury lamp with mercury-pool electrodes (artificial sun), it had been found that with mercury vapour under high pressure (1-3 atm) a source of light of high efficiency could be obtained. The spectral composition of this light shows a striking lack of red and as a consequence colour rendering is inadequate. But with sodium there is still less colour rendering and when this drawback came to be generally accepted for the sake of the other, favourable, qualities of sodium light (efficiency, visual acuity, contrast) there was every inducement to try out also the mercury lamps. Great Britain set the example by installing experimental lighting with high-pressure mercury lamps, with oxide-coated cathodes, on some highways. Eindhoven, too, very soon started producing these lamps. But meanwhile further developments were taking place. The experimental and theoretical investigations carried out by W. Elenbaas very soon made it possible to survey the whole field of mercury discharges as functions of the various parameters (dimensions of the tube, mercury pressure, current and voltage). For instance, a principle of similarity could be worked out, whereby the number of essential parameters could be reduced and it could easily be predicted what the behaviour would be of a lamp with certain dimensions and containing a certain amount of mercury.

It was again Bol who succeeded in drawing practical conclusions from this study and arrived at a very compact construction of mercury lamps, which, however, only became possible after H. J. Lemmens had worked out a method of fusing tungsten leads to quartz, using only one intermediate glass. These lamps, only a few centimetres in length and with an internal diameter of 2 mm, were so designed that part of the mercury remained in the liquid state. Owing to the considerable heating of the wall of the tube these lamps were cooled with water. The internal pressure amounted to 100 atm and more. These lamps are being used for cinema projectors, in television studios and in searchlights, in general wherever there is no serious objection against the installation of a water-cooling system.

Another type, with a determined quantity of mercury which entirely evaporated during the operation of the lamp, had no need of any forced cooling. The equilibrium pressure was 5 to 10 atm. This type of lamp found extensive use for road lighting. In order to limit the preheating time the lamp proper, which has an internal diameter of 10 mm and a length of 35 mm, was placed in a normal incandescent lamp bulb filled with an inert gas to prevent oxidation of the leads. Notwithstanding the fact that at this high pressure the mercury spectrum shows, in addition to the much widened mercury lines, a continuous background extending into the red, colour rendering is inadequate, just as is the case with the lamp with a pressure of 1 atm. This drawback can be remedied by admixing incandescent light, but then the efficiency of the

whole is reduced. Later it was found possible to improve the colour rendering by coating the inside of the outer bulb with substances which fluoresce under the influence of the ultra-violet rays of the mercury light.

Even when corrected with incandescent light or by means of a (red) fluorescing substance, however, for various reasons these lamps are not suitable for indoor lighting. In this direction gas-discharge lamps appeared in a different form. Starting from the low-pressure mercury discharge, whereby the emission of the visible mercury lines is only small and mainly the ultra-violet resonance lines of 2537 Å and 1849 Å are produced, these ultra-violet rays were converted into visible light by means of fluorescence. Philips contributed much towards the development of these "TL" tubes. By coating the inner wall of the discharge tube with a suitable mixture of fluorescent substances ($MgWO_4$, $(Zn,Mn)_2SiO_4$, $Cd_2B_2O_5$-Mn, $(Zn,Be,Mn)_2SiO_4$, and later also other substances, such as halophosphates) it was possible to produce a white light with a spectral energy distribution sufficiently approximating that of daylight or of incandescent light.

Such lamps as these, with a gross output of 40 to 50 lm/W, are being used more and more for all forms of utility lighting, such as in the home, workshops, offices, shops, etc., and for road lighting.

J. Voogd and others have been occupying themselves with the photometry of gas-discharge lamps.

The development of gas-discharge lamps with their spectral energy distribution differing so greatly from incandescent light, their extensive applications for road lighting and the problems of colour rendering all called for a profound study of the properties of the human eye and the faculty of seeing under different levels of brightness. In the Philips' laboratory these problems were energetically tackled by P. J. Bouma and A. A. Kruithof. Bouma revived interest in the more or less forgotten work of *A. König* (1891) on the subject of seeing at low brightnesses and he strougly advocated the ideas of *Schrödinger* (1920) in regard to the use of the colour space for colorimetric problems. His division of the visible spectrum into eight sections as a means of judging colour rendering has come to be of almost universal use, at least in Europe. For the practical application of the eight-section analysis a special photometer was constructed by P. M. van Alphen.

II. Electrotechnics, radio and acoustics

From the manufacture of incandescent lamps a number of important electrotechnical products have emerged which are related to the property of the filament to emit electrons, an effect which, as already remarked, was discovered in principle by *Edison* and subsequently investigated quantitatively by *Richardson*.



Library of the Physical Research Laboratory.

As a result of this investigation there appeared, as we have seen, first rectifying valves and radio valves with tungsten filament in vacuo, whilst Philips followed the G.E.C. in producing gas-filled rectifying valves with coiled tungsten filament. An important improvement was the replacement of the helicoidal tungsten filament in these valves by an oxide-coated cathode (J. Bruynes).

With the introduction of the oxide-coated cathode many more uses were found for rectifying valves, because on the one hand larger emission currents were obtained, thereby extending the field of application to heavy currents, whilst on the other hand the lifetime of these tubes was so extended as to be dependent only upon the chance of accidental damage (breakage or breakdowns) and not upon exhaustion of the thermionic properties of the filament (J. G. W. Mulder).

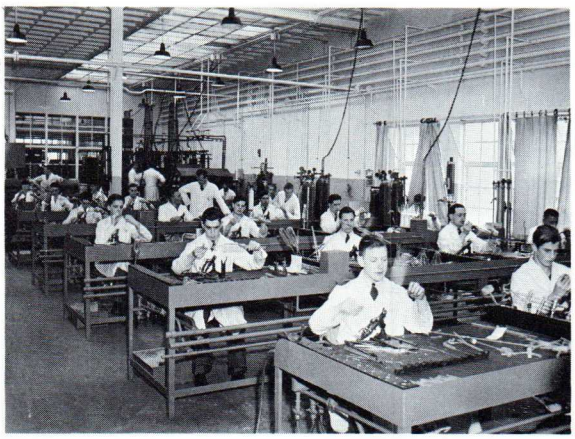
The field of rectifiers is an interesting example of a case, often seen, where a phenomenon first found as a small effect, so small that its existence might be doubted, ultimately finds application on a scale precluding any doubt as to its reality. The currents studied by *Richardson* could at first only be measured with a sensitive galvanometer. Now a rectifying valve with an emission current of 100 A is nothing rare.

Rectifying valves are now being made for high as well as low voltages. Valves for low voltages (some tens of volts) are used, for instance, for charging batteries and feeding cinema arc lamps (D. M. Duinker). In X-ray practice, on the other hand, rectifying valves are employed which can withstand voltages exceeding 100 kV, thanks to their being given a suitable shape. Another important product was the welding rectifier, characterized by a very heavy current (H. A. W. Klinkhamer).

Within the scope of rectifier research mention is to be made also of the work done in developing blocking-layer rectifiers, and in connection therewith the investigations with selenium (W. Ch. van Geel, N. W. H. Addink).

With the advent of the triode as receiving and transmitting valve round about 1914, radio entered upon a new era, a development to which Philips laboratory contributed in no small degree.

It is remarkable how often in the field of radio practice has been far in advance of the theory. The triode proved to answer its purpose well and was already being applied on a large scale ever before the relative problems, such as the calculation of the field between the electrodes and the behaviour of the electrons in that field, had been really mastered; the propagation of the radio waves over long distances had been found practicable long before scientists had got to the bottom of the actual reason for it and before the electromagnetic equations governing the propagation along the earth's surface had been satisfactorily solved. In the long run, however, it is essential to gain the fullest possible insight into the theory of the phenomena, and for that reason Philips have always devoted much attention to their theoretical investigation. Many other problems remained to be



A corner of the glass-blowing shop of the Physical Research Laboratory.

solved in connection with the practical application of radio valves, and so in the new laboratory radio investigations were divided among a number of groups of research workers. B. van der Pol was charged mainly with the conducting of theoretical radio investigations, whilst the more practical investigations were carried out by E. Oosterhuis, P. R. Dijksterhuis, Y. B. F. J. Groeneveld, H. Rinia, B. D. H. Tellegen and many others.

Van der Pol had begun in 1922 with the study of the triode. It had been found that the flow of electrons to the grid and the anode, respectively i_g and i_a, is a function of the voltages V_a and V_g respectively at the anode and at the grid with respect to the cathode. A three-dimensional plaster model was constructed with which this relationship could be visualized. Then attention was paid to the paths followed by the electrons under the influence of the field and to the secondary emission caused by the electrons impinging on the anode, the effect of which can be demonstrated with the plaster model. Further problems were the distribution of the electron stream between grid and anode and the effect that space charge has upon the field; for the study of these problems a foundation had been laid by the theoretical investigations of *Langmuir* and *Epstein*.

The problem of the motion of the electrons, which has always demanded attention in connection with the construction of radio valves, was subsequently investigated by P. H. J. A. Kleynen and J. L. H. Jonker with the aid of a model employing small steel balls made to roll over a sheet of rubber. This device has proved to be of great value in cases where the mathematical approach to the problem is too complex. Another means of circumventing mathematical difficulties was the measuring of fields of complex electrode systems with the aid of the electrolytic tank.

In connection with the phenomena arising in the radio valve mention is also to be made of the extensive investigation of fluctuation phenomena (noise) carried out by C. J. Bakker and M. Ziegler, which later on was extended to higher frequencies by M. J. O. Strutt, A. van der Ziel and others.

Further, there was the investigation made by H. Bruining into the secondary electron emission of solids, not only in connection with the occurrence of this emission in ordinary radio valves but also with a view to the construction of special valves in which secondary emission is brought about purposely in order to produce special effects.

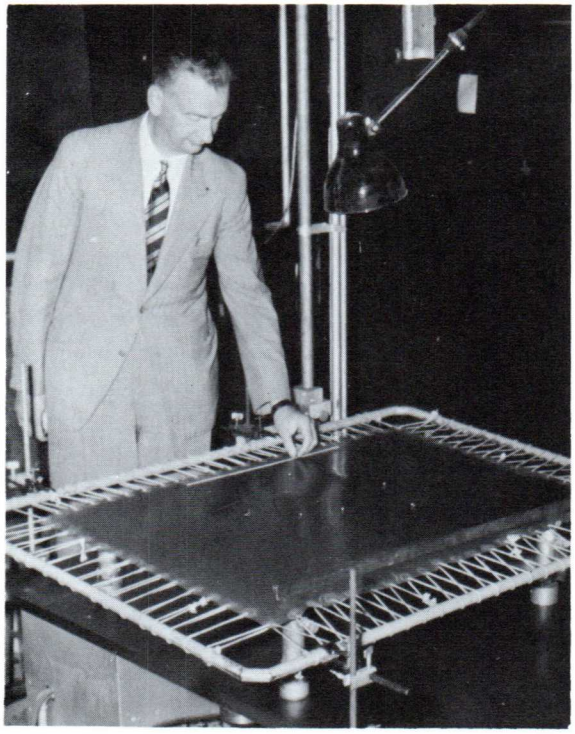
Following upon the study of the radio valve as such, its behaviour was investigated when employed as an amplifier or as an oscillator in a network comprising capacitances, resistances, self-inductances and mutual inductances.

An electric network is a system which as a rule is governed by a number of linear differential equations. With the introduction of radio valves in the network not only are negative resistances introduced, which make it possible for oscillations to be generated, but also non-linear terms then enter into the equations and make the problem more complicated.

Van der Pol succeeded in working out a non-linear differential equation in a simple form (the Van der Pol equation) which incorporates all essential data involved in the case of oscillations and the limiting of oscillations in networks. This equation,

$$\ddot{v} - \varepsilon(1 - v^2)\,\dot{v} + v = 0\,,$$

contains as parameter the factor ε, which to a considerable degree determines the behaviour of the solution. It appeared that within a certain range of values ($\varepsilon \gg 1$) the oscillation bears a character differing greatly from the known behaviour. Van der Pol named these relaxation oscillations. As opposed to "ordinary" oscillations as may occur when the system comprises mainly C's and L's, and which have a sharply defined cycle, while the amplitude depends upon various secondary conditions, relaxation oscillations may arise in systems which are governed mainly by C's and R's. In the latter case the amplitude of the oscillations is determined by the system, while their frequency is highly sensitive to external disturbances. Thus a system showing relaxation oscillations, such as a glow lamp shunted by a capacitor charged by a voltage source via a resistor, can easily be synchronized with a periodical signal, a principle widely used nowadays in television.



Small steel balls made to roll over a sheet of rubber give a picture of the paths followed by electrons in, for instance, electronic valves.

With the aid of relaxation oscillations it is also easy to demultiply frequencies (frequency dividing) and to produce the sub-harmonics, in contrast to the formation of higher harmonics, which can be obtained by connecting non-linear impedances to a normal oscillatory circuit.

As Van der Pol pointed out, the concept of relaxation oscillation is also of biological importance, where the rôle of "resistance" is performed by some diffusion phenomenon governing the biological process. Examples are the reaction of plant foliage to the alternation of day and night and the functioning of the human heart.

Van der Pol and J. van der Mark succeeded in constructing an electrical model of the heart, built up from a number of glow discharge lamps, capacitors and resistors, with which not only the functioning of the normal heart but also certain pathological aberrations could be imitated. To our minds the medical world was too sceptical about the value of that model at the time. Fortunately, however, medical scientists of the present day are showing more and more interest in the results of electro-physical and electrotechnical work, as for instance in neurophysiology. There are many indications that in the near future closer cooperation between the medical practitioner, the physicist and the electrical engineer, with mutual appreciation for the experiences and views of each, will prove to be of great advantage to all concerned, including the patients.

An important part of radio research is that concerning the propagation of electric waves and the interaction between these waves and matter. First of all mention is to be made of the investigations relating to skin effect and to the penetration of electromagnetic alternating fields in conductors. The practical side of this subject was made manifest in the manufacture of radio valves, whereby the valve was placed in a high-frequency magnetic field for degassing various parts mounted in it. It was of importance to be able to predict the degree of heating of the parts (plate, cylinder, grid) as a function of the field strength and of the orientation in the field; this problem was thoroughly investigated in the laboratory by M. J. O. Strutt and others. Of particular interest are the phenomena taking place in bodies of a magnetic material, for there it may happen that above the *Curie* point ($\mu \approx \mu_0$) the product $\mu\delta$ ($\mu = \mu_0\mu_r$ = permeability, δ = depth of penetration) is small with respect to the dimensions and below the *Curie* point ($\mu \gg \mu_0$) large. As demonstrated experimentally by J. L. Snoek, this has remarkable consequences when small objects of magnetic material are subjected to high-frequency heating. Upon the field strength being reduced the temperature remains high, owing to sufficient power being absorbed even when the field is weak. When, however, the temperature drops below the *Curie* point the absorption of power rapidly decreases and the field has to be made much stronger to heat the body to a high temperature again.

The radiation from aerials was investigated both empirically and theoretically. As far as the experimental work is concerned, extensive measurements were taken, for instance, of the field strengths in areas covered by various broadcasting stations all over the Netherlands (R. Veldhuyzen).

Theoretically the problem of wave propagation for a free radiating dipole had already been solved by *H. Hertz*. When the dipole is placed over an infinite, perfectly conducting, "flat earth" the field of the dipole and that of its reflected image can simply be added together. Considering, however, that the conductivity of the earth is finite and that it has a finite relative dielectric constant, the problem is much more complicated. For this case *Sommerfeld* arrived at an exact formula as far back as 1909, but this formula is not suitable for calculations. Therefore at the same time an approximative formula was given for calculating the field strength at the earth's surface. In 1919 the same problem was tackled once more, but in a different way, by *Weyl*, and in 1926 *Sommerfeld* showed that *Weyl's* result agreed with his own. But he then put his approximative formula in a somewhat different form. Much has been written on the question which of the two formulae, that of 1909 or that of 1926, was the "correct" one and what was to be decided by experiment. Important contributions on this subject were given by Van der Pol and his coworker K. F. Niessen (one of *Sommerfeld*'s pupils), who arrived at a strict solution in a new form, and further by the Americans *Norton* and *Burrows*; the exact experimental determinations carried out by the latter proved to be in agreement with the formulae given by *Weyl* and by Van der Pol and Niessen.

Another question of practical importance is what part of the energy radiated by a dipole is dissipated in the earth, since from that the efficiency of a transmitter can be calculated. This mathematically complicated problem was solved by Niessen (1940) by employing *Sommerfeld*'s exact formula.

The problem of the propagation of the waves over a spherical surface has been the subject of an intensive investigation by Van der Pol and H. Bremmer following upon *Watson*'s work. Since in the formulae the dimensions of the spherical surface and the wavelength are not bound to certain values, these apply also in other realms of physics, such as for the interaction between light waves and droplets of water (the rainbow). A remarkable theoretical result, for instance, is that the radiation of the rainbow shows polarisation. This can now easily be verified with the aid of "Polaroid" spectacles; it seems that meteorologists had never noticed it.

In the investigations referred to, the atmosphere was regarded as a homogeneous medium. Later,

Bremmer contributed important information on propagation in an inhomogeneous atmosphere (the ionosphere, "ducts"). In Great Britain the investigation of the ionosphere had been taken up experimentally on a wide basis by *Appleton*, who gave a lecture on this at Eindhoven in 1930. The root of the problem, for which no solution had then been found, lay in the relation between the virtual height (velocity of light in vacuo × half the reflection time) as a function of the frequency and the concentration of free electrons as a function of the height above the earth. From the discussions held at Eindhoven it appeared that, when ignoring the magnetic field, this relationship is given by an integral equation of *Abel*, so that the result can be written explicitly.

At certain frequencies the reflection time is infinite. Van der Pol saw in this a connection with the mysterious "delayed echos" which at that time were puzzling many radio amateurs. It is to be noted, however, that *Störmer* had quite a different explanation for this phenomenon and attributed it to charged particles coming from the sun at a great distance from the earth (10^6 km and more).

A remarkable effect of the ionosphere, discovered at Eindhoven, is the interaction of radio waves of different frequencies. In the empty space, owing to the linearity of the *Maxwell* equations, the principle of superposition applies exactly, so that two wave systems may penetrate each other without any mutual interference. In the ionosphere, however, the interaction is partly of a non-linear nature, with the result that waves propagated through the ionosphere over a powerful transmitter become modulated with the frequencies of that transmitter. Tellegen observed this at Eindhoven in the case of signals from the Beromünster station, which were subject to interference from the powerful Luxemburg station.

In addition to this, for the greater part, purely scientific research a considerable amount of work was directed towards the practical side of radio, in connection with both radio valves and further radio equipment. Above all, as was only natural, radio valves underwent repeated changes. There was a universal rational dimensioning of the various electrodes. In many cases the filament was replaced by an indirectly heated cathode. A second grid was introduced, first as space-charge grid and later as screen grid (*Hull*), so as to render the i_a-V_g characteristic less sensitive to anode voltage fluctuations and, furthermore, with the object of reducing the capacitance between the control grid and the anode, which was found desirable on account of the ever higher frequencies at which transmitting stations were working.

With the tetrode thus formed much trouble was experienced from the secondary emission of the anode. Tellegen therefore introduced between screen grid and anode a third grid (suppressor grid) to hold back the secondary electrons. Thus arose the five-electrode valve or pentode, which at first was used as output valve and subsequently came to be applied also in other circuits. On first sight this may not seem to be a very drastic change, but it has proved to be an exceptionally important improvement, since now, with a very few exceptions, all radio valves are pentodes.

Much work has been done in investigating the behaviour of these and other types of valves in various functions, such as for high-frequency and audio-frequency amplification (A. J. Heins van der Ven, J. van Slooten, H. van Suchtelen, and others).

Considerable attention has also been paid to the construction of transmitting valves, partly in connection with the employment of short waves (15-50 m) for long-distance radiotelephony. An experimental transmitter was built (J. J. Numans), provisionally equipped with a "dummy aerial", in which the energy normally radiated could be dissipated. A milestone was reached in the history of the laboratory in March 1927, when this transmitter was connected to an aerial and communication was established with what at that time was the Netherlands East Indies. This soon gained world fame, especially when, on 1st June 1927, *H. M. Queen Wilhelmina* used the transmitter (station PCJJ) to broadcast an address to the overseas territories.

In connection with the ever higher frequencies that were being used, there was also the development of the magnetron as transmitting tube. Philips Research Laboratory contributed much towards a proper understanding of the working of this tube. The treatises by K. Posthumus on the functioning of the magnetron with split anode still form the basis for all theoretical expositions in this field. Experiments in range finding by means of radio waves of 1 m and smaller generated by a magnetron transmitter were carried out in the laboratory before 1940; in the period 1940-'45 this principle was applied on a large scale elsewhere in the form of "radar".

The designing and manufacture of radio valves calls for great care and special methods. As higher



H.M. Queen Wilhelmina, accompanied by H. R. H. Princess Juliana, addressed Her subjects in overseas territories via the Philips' transmitting station PCJJ on 1st June 1927.

frequencies came to be used the connections had to be shorter and so different methods had to be found for constructing the electrode lead-ins. The solution was found by applying pressed glass or sintered glass (Lemmens) and special metals for the leads. This development came for a large part from the factory laboratories.

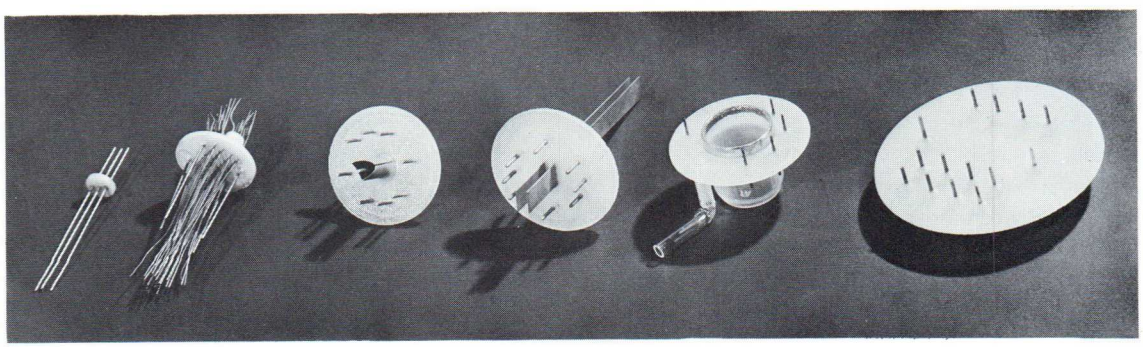
At first radio valves and some radio parts, such as high-tension units, audio-frequency transformers and resistance-capacitance couplings, used by amateurs and set makers, were the only radio products manufactured in the works. However, with the increasing popularity of radio broadcasting there was such a big demand for radio receivers that it became worth while to start manufacturing complete receiving sets. Looking back now, it is hardly imaginable that 25 years ago this was regarded as being something out of the ordinary.

Almost at once the need was felt for an apparatus that could be worked without batteries by feeding it entirely from the mains. This was made possible by employing the indirectly heated cathode which had meanwhile been developed at Eindhoven.

In the construction of radio sets, as already remarked, a great many problems were involved which had to be solved in the laboratory, in the beginning even in all sorts of constructional details (Bol, C. J. van Loon, J. M. Unk). On the one hand there was a need of certain circuits which had to be as efficient as possible while at the same time being easy to make and to repair, and on the other hand accurate methods of measuring were needed for testing the functioning of experimental circuits.

Much attention was devoted to the coils. By making these as loss-free as possible (according to a principle evolved in this laboratory by H. Rinia) greater selectivity was obtained and a "straight set" could be built with four tuned circuits, which for a long time answered the purpose very well. As the "ether" became more and more crowded with the increasing number of stations working within the allotted frequency bands, so that short waves came to be used for broadcasting, receivers had to be built on the superheterodyne principle.

Of great importance was the invention, by



Metal leads and supporting rods can be fused into bases of sintered glass in an almost unlimited number and in any order.

Posthumus, of negative feedback, for which patents were obtained already in 1928. This principle, which reduced distortion due to curvature of the valve characteristics, was very soon introduced in the Philips receivers. It proved, however, to have a much wider scope than this, so that at the present day it is being employed in practically all amplifiers and in all kinds of regulating devices, etc.

For those cases where a highly constant voltage (low internal resistance) is required for feeding radio apparatus a self-regulating high-tension supply unit was developed (Rinia, H. J. Lindenhovius). In connection with smoothing systems for the feeding of radio apparatus mention is to be made of the considerable amount of work done in the field of electrolytic capacitors (Van Geel, A. Claassen).

A difficulty encountered in the manufacture of radio sets lay in the lack of uniformity of electrical networks, especially since there are both direct and alternating current mains. J. W. Alexander constructed vibrator-converters capable of converting direct voltages of 100 to 200 V into an alternating voltage. Such vibrators are now commonly used with car radio sets for converting the battery voltage of 6 or 12 V into an alternating voltage of 220 V.

Another problem lies in the feeding of receivers in places where no mains are available. In the place of accumulators and dry-cell batteries the thermoelectric generation of current was thought of, but this is too uneconomical. This subsequently led to the development of the air engine, which before long will be able to take the place of the petrol engine now used for this purpose.

In the development of radiotelephony there is also the problem of conversion of the radio signal into audible vibrations of the air, thus into speech or music. As a consequence acoustics, and especially electro-acoustics, have become inseparably connected with radio. In 1925, when under the guidance of R. Vermeulen Philips started making loudspeakers, these were built on the electromagnetic principle, whereby the movement of an armature is transmitted to a diaphragm. Then there appeared the moving-coil loudspeaker, in which the cone-shaped diaphragm is connected to a small cylindrical coil through which the varying signal current flows and which is placed in a radial magnetic field. This magnetic field was at first produced by means of a soft-iron circuit excited with direct current. Philips very soon replaced this circuit by a permanent magnet. This development involved intensive research in connection with magnet steel, a subject which will be referred to again elsewhere. The new magnetic materials also made it necessary to give the magnets a different shape, so that attention had to be paid to the designing of magnetic circuits (A. Th. van Urk).

In addition to the work involved in the development of loudspeakers, much work has also been put into the development of power amplifiers, which in turn led to great activity in the field of line telephony (W. Six, H. G. Beljers, J. te Winkel).

In connection therewith attention may be drawn to the various measuring instruments developed in the laboratory, such as a measuring bridge for measuring losses in coils and capacitors, with which phase angles can be measured with an accuracy of 10^{-6} in a frequency range of 10^3 - 10^5 c/s (J. W. Köhler, C. G. Koops).

Further research work led to the development of various types of microphones, gramophone pick-ups and gramophone motors, and finally sound reproduction for the sound film and for broadcasting studios. An important part has been played in this by the Philips-Miller process, whereby a width-modulated track is cut in a celluloid tape coated with a lacquer, after which it is scanned by the known optical means. An advantage of this system was that the recording could be played back at once and thus corrected where necessary.

In this connection mention is to be made of a system, developed by K. de Boer, for stereophonic sound reproduction, both direct and by means of gramophone records or a Miller tape.

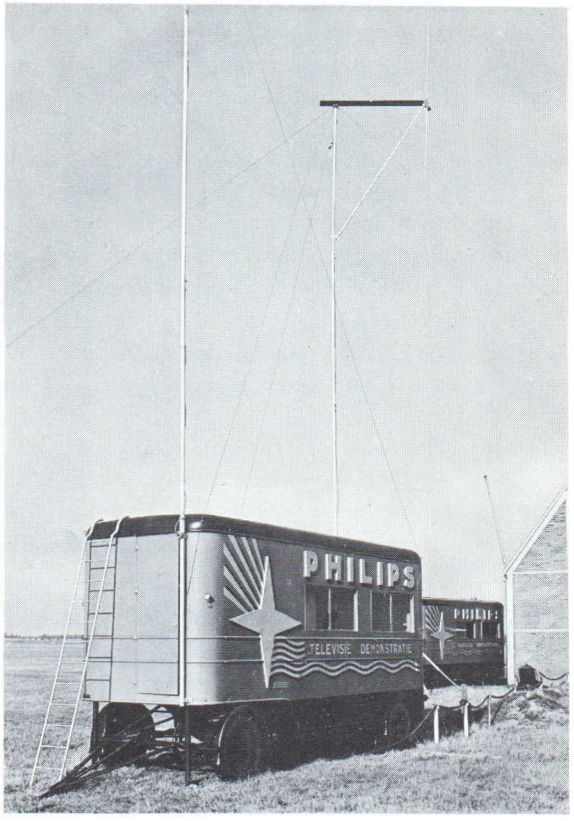
Important work has also been done in the field of the physiology of hearing. Particular mention is to be made of J. F. Schouten's investigations into the validity of the so-called *Ohm's* acoustic law, from which it appeared that in a mixture of frequencies which are a multiple of a certain fundamental frequency the latter can sometimes be heard even if it is not itself present in the mixture (the "residue" theory).

Further, L. Blok designed various signal generators, which have been applied, inter alia, for audiometric investigations.

Other developments in the technique of sound reproduction will be dealt with in the last section of this review.

The development of television began already in the thirties. In its earliest stages a *Nipkow* disc with 48 lines was used both at the transmitting

and at the receiving end. Subsequently a cathode-ray tube was used in the receiver, whilst the *Nipkow* disc, still employed for the transmission, was perfected for televising films (Rinia). Later on the *Zworykin* iconoscope, meanwhile developed in the U.S.A., came to be used for TV transmission, first for the 180-line system and later for larger numbers of lines (Van der Mark, G. Hepp, A. Venis). Experience was gathered in the construction of amplifiers (J. Haantjes and others) and the building of transmitters (W. Albricht). Experiments



In 1937 already a transportable installation for television transmission and reception was completed and taken on tour for demonstrations over a large part of Europe. It could be worked on the system of 405 lines or on that of 567 lines.

were also carried out with projection television by means of lenses, employing a cathode-ray tube with concave screen (M. Wolf). Complete mobile television apparatus was built and installed in two motor-vans, with which demonstrations were given in various places in Western Europe (Utrecht, Brussels, Antwerp, Copenhagen, Stockholm) and also in Eastern Europe (Budapest, Bucarest, Zagreb, Belgrade, Warsaw). Owing to the outbreak of the war these tours came to an abrupt end in 1939.

III. Chemistry

It would be saying too much to maintain that all the chemical problems tackled in the Philips laboratory in the course of time likewise emanated from the incandescent lamp, but for a very large part such is indeed the case, either directly or indirectly.

Disregarding the metallurgy of the tungsten and the making of the glass, in the manufacture of the incandescent lamp chemical problems arise from the action of the gases released from the wall of the bulb and the metal leads upon the glowing tungsten filament, whereby traces of water vapour have a particularly disastrous effect. In the dissociation of the water vapour by the filament oxygen is combined with the tungsten and forms a volatile tungsten oxide which is precipitated on the glass wall of the bulb. The remaining hydrogen in turn reduces this oxide to metallic tungsten. Thus, owing to this "*Langmuir*'s cycle", tungsten is continuously being transported to the wall of the bulb, so that gradually this is blackened. It is therefore essential to remove water vapour and other gases, and this is done by introducing into the bulb a substance, such as phosphorus, which combines with the residual gases into a non-volatile and harmless product. Such a substance is called a "getter". In radio valves, where phosphorus cannot be used on account of its high vapour pressure and residual gases are moreover obnoxious because in the ionized state they affect the oxide-coated cathode and give rise to grid currents, and in X-ray tubes, where residual gases increase the risk of breakdown, metals like barium and zirconium are used as getters.

Even though the bulb of an incandescent lamp may not contain any gases attacking the filament, still evaporation of the tungsten takes place. In course of time this evaporation likewise turns the bulb black and thus the yield of light is reduced. Furthermore, the filament itself is reduced in thickness and as a result eventually collapses. The same applies in the case of the gas-filled lamp. The prevention of the evaporation by the gas-filling as such is offset by the filament being heated to a temperature so much higher that the lamp has about the same lifetime, so that the only advantage gained is the higher efficiency.

The manner in which this evaporation of the tungsten takes place differs according to whether the lamp is filled with gas or evacuated. In a gas-filled lamp the tungsten atoms come into collision with the gas molecules time after time, so that they have an opportunity to combine into aggregates which move about in the gas in the form of sub-

microscopical flocculations and eventually settle upon the wall of the bulb. Owing to convection this shows a preference for that part of the lamp which in the burning position is uppermost, thus not the part through which most of the light passes. With the vacuum lamp the situation is different: the tungsten atoms travel in a straight line from the filament to the inner wall of the bulb. It can be imagined that upon striking the glass wall the tungsten atom is repelled like a minute metal ball, but we know that ultimately it adheres to the glass, so every time the atom strikes against the wall it must lose some of its kinetic energy. *Langmuir*'s conception was that upon collision with the wall the atom temporarily adheres to it and is then, as it were, again evaporated, thus implying a certain "adhesion time". By accurate experimentation this has been confirmed by Clausing. It is true that in the case of cadmium atoms only an upper limit (10^{-6} sec) could be found for this adhesion time, but in the case of argon atoms on glass at temperatures of 80 to 90 °K adhesion times of 10^{-4} to 10^{-5} sec were found.

This matter of the evaporation of tungsten has been dealt with at some length because the investigations carried out in connection therewith formed an introduction to further investigations into the adsorption of tungsten atoms and the resultant absorption of light. So long as the tungsten atoms remain isolated on the wall of the bulb there is no appreciable absorption of light, but the position is different when they form a continuous layer of metal. In order to minimize the absorption of light, therefore, before the mount is fused into the bulb the filament is sprayed not only with the getter but also with a little salt, say CaF_2. As soon as the filament is heated to a high temperature this salt evaporates and is precipitated on the wall of the bulb as an invisible thin layer. The tungsten atoms coming from the filament shoot into this layer of salt, the particles of which keep the atoms separated, so that very much less light is absorbed.

About 1920 little was known with certainty about the effect of such layers of salt. Investigations into their action carried out in Philips laboratory, mainly by J. H. de Boer, extended over a period of more than 15 years. It has thereby been found that the layer of salt is not to be regarded as a homogeneous mass but as an agglomeration of minute crystal lamellae about 10^{-6} cm thick lying criss-cross one on top of the other and thus forming a very large active surface, tens of times greater than the surface of the glass covered by them. The tungsten atoms are adsorbed on the surface of the crystals. The intensive study of the adsorption of atoms and molecules (e.g. caesium and iodine) in such layers of salt (De Boer, C. J. Dippel, C. F. Veenemans) has yielded very important results.

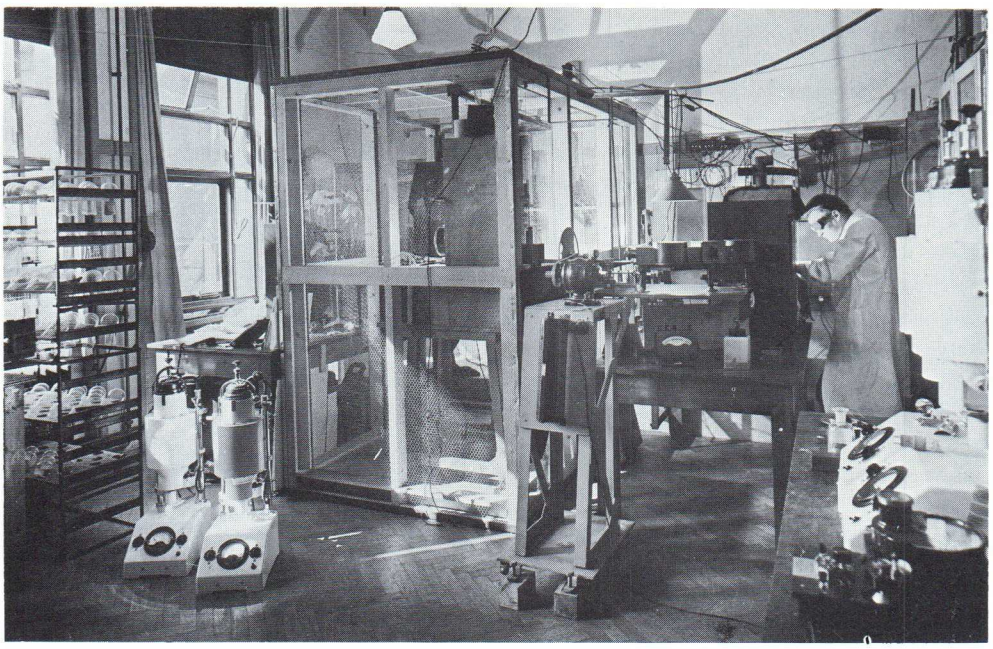
Metals like caesium may be adsorbed also on the surfaces of metals, in which case they have the property of reducing the work function of the metal and thus increasing the electron emission, both the thermionic emission and that brought about by irradiation with light — photo-electric effect. By oxidizing the adsorbed layer and again precipitating caesium onto the layer of oxide it is possible to produce complex layers with an exceptionally strong photo-electric effect (M. C. Teves). The study of this emission was of importance for the construction of photocells, for which there was a need in connection with the sound film. The deeper insight gained into the nature of these layers opened up new possibilities, such as the employment of the "light transformer" for converting infra-red rays into visible light (Holst, De Boer, Teves), the construction of the electron-multiplying valve and, later, the television pick-up tubes.

The investigations into the vaporization of CaF_2 also led to other compounds, such as H_3BO_3, K_2BF_4, being tried out for the same purpose, as a result of which the volatility of various compounds was studied. Why, it may be asked, is NaCl for instance a substance having a high melting point and a negligible vapour pressure at room temperature, whereas WCl_6 is a substance that readily melts and is easily evaporated? From the point of view of the theory of heteropolar chemical compounds the answer is simple. The molecule of NaCl is built up from a positive sodium ion and a negative chlorine ion, which together form an electric dipole, so that two NaCl molecules exercise a very strong attractive force upon each other. In the case of WCl_6, on the other hand, the central metallic ion is surrounded by six chlorine ions screening off the charge of the central ion, so that there is no great attraction between neighbouring molecules. The theory of the heteropolar chemical bond was formulated by *Kossel* in 1920, following upon *Bohr*'s work. It is mainly due to the work of Van Arkel in this direction that a more or less coherent explanation was established for the most important facts in anorganic chemistry, an explanation that is of great value as a basis for chemical thought and for education in chemistry. It is fortunate that this was more or less completed before

the conception of the homeopolar bond came into more prominence as a result of the development of quantum mechanics. Thus the one-sided-heteropolar aspect could gradually be introduced into the new system without losing its value.

In connection with these investigations the energy of formation of a number of molecules and crystal lattices was calculated and the relative stability of various molecule models investigated.

known that WCl_6 and hydrogen on the surface of a heated tungsten filament may react one upon the other and form tungsten, which is deposited on the wire. When a mono-crystalline wire is taken (*Pintsch* wire) also the growing metal becomes a mono-crystal. In the Philips laboratory it was found that dissociation of WCl_6 at elevated temperature takes place also without hydrogen; attempts were made to produce other metals in the pure



Measuring the spectral transmission of various materials in the ultra-violet range. On the right the source of ultra-violet radiation and the monochromator. In the centre, in a screening cage, the measuring apparatus with photoelectric cell.

The above conceptions were also applied to the relation between physical properties of homologous organic compounds, as for instance the relation between the boiling points of CH_4 and of the compounds obtained when replacing in CH_4 one or more hydrogen atoms by atoms of F, Cl, Br or I. Subsequently, after Van Arkel's appointment as professor at Leyden, this investigation was extended by him and his students to a large number of other compounds. Also worthy of mention is the study of the dielectric behaviour of organic dipole molecules in solution, which led, inter alia, to the "Van Arkel and Snoek formula", which was later investigated theoretically by *Onsager* and by *Böttcher*.

The study of the volatility of metal compounds had also important technical consequences. It was

state in this way. Perhaps the most striking result was the preparation of titanium, zirconium, hafnium and thorium from their iodides by precipitating the latter on a thin tungsten wire as core (De Boer and J. D. Fast). In this way titanium and zirconium, known as being greyish brittle substances of a doubtful metallic character, were obtained in the form of fine lustrous metallic products. The rods, consisting of a few large crystals, proved to be highly ductile, so that they could be drawn into wire and rolled into foil. This is particularly of importance in the case of titanium, considering the interest taken in this metal in recent years. Ductile zirconium has found important applications in vacuum technics. Applied to the anode of a transmitting valve it proves to be an excellent getter, on account of the almost unlimited capacity

of this metal to adsorb oxygen and other gases. The oxygen taken up in the metal has a negative charge, as appears from the fact that in an oxygen-charged rod of zirconium through which a direct current is passed the oxygen migrates towards the positive pole.

The knowledge of metals and their interaction with gases proved to be of great value to the laboratory staff when it was decided to manufacture welding rods (J. Sack, P. C. van der Willigen), in addition to the welding rectifiers and rectifying valves already being produced.

The study of tungsten wire and the behaviour of tungsten in processing led to extensive research in the domain of recrystallization. A substance particularly suitable for this is aluminium, which is easily deformed and recrystallized at comparatively low temperatures (< 600 °C), while the process of this recrystallization can be followed by etching in aqua regia after removal of the superficial oxide with caustic soda or hydrofluoric acid. This simple technique, supplemented by crystallographic study (W. G. Burgers, J. F. H. Custers), led to a deeper insight into the essence of the formation of crystal nuclei. This study is still being continued at the present day by Prof. *W. G. Burgers* in the inorganic-chemical laboratory of the Technical University at Delft.

Finally, in connection with applications in the field of electrotechnics and acoustics (transformers, loudspeakers), extensive research has been carried out in regard to the magnetic properties of metals, particularly of iron and iron alloys (G. J. Sizoo, W. F. Brandsma, Elenbaas, Jonas, Snoek, Six, G. W. Rathenau, J. J. Went, H. J. Meerkamp van Emden).

Special products, the fruit of years of study in this domain, were the rolled nickel-iron ("Fernicube") with strong anisotropic properties, for loading coils, and magnet steels with exceptionally high coercivity and high value of the (HB_{max}) product, which are widely used not only in loudspeaker magnets but also in pocket and bicycle dynamos.

Of importance for the investigations both of magnetic and non-magnetic metals was the work done by Snoek in studying the magnetic after-effects of iron, whereby it was found that these effects are related to the presence of traces of carbon and nitrogen, which influence not only the magnetic but also the elastic after-effects. Both these are governed by a sort of diffusion of C and N atoms present in interstitial places to neighbouring interstitial places.

Other materials studied at the time in the laboratory are the ferrites, which since 1934 have been further investigated by Van Arkel, Snoek and E. J. W. Verwey.



Preparation of magnet steel. Emptying the furnace in which the alloy has been melted by induced high-frequency currents.

This led, inter alia, to the view (De Boer and Verwey) that a substance like Fe_3O_4 derives its conductivity from the fact that ions of one and the same element but of different valency are present in crystallographically identical places and thus make it possible for an electron to pass over from one ion to another.

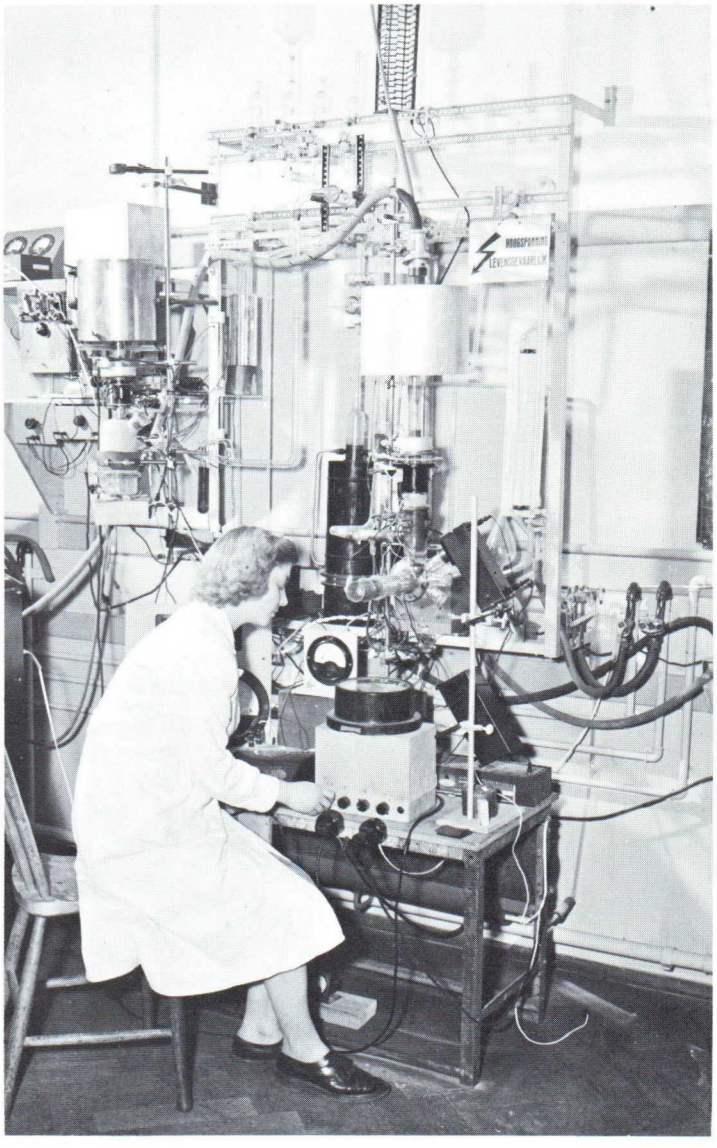
In the course of the resultant investigations our research workers became familiar with various problems of the solid substance, in particular with oxidic systems, of which the spinels form an interesting sub-group. The latter were further studied both in respect to their electrical properties (Verwey and others) and with regard to their magnetic properties (Snoek). The study of the electrical properties led to the development of semi-conducting materials, while that of the magnetic properties yielded the important magnetic material "Ferroxcube", which will be dealt with later.

The previously mentioned work by De Boer and others on the adsorption of atoms on crystals was extended to cases where atoms are built into a crystal lattice, a subject to which much attention was devoted in Germany by *Pohl* and his pupils. What is particularly due to De Boer is the disclosure of the fact that the place of a lacking negative ion in an ion lattice may be occupied by an electron and that, at the cost of only a little energy, this electron may be brought into a state in which it acts as a conduction electron.

It was at that time that the programme of work on solids was given shape, which later on was to occupy the minds of such a large part of the scientific staff of the laboratory and comprised, among others, the investigation of luminescent substances.

The luminescent substances (luminophores or phosphors) used for converting the ultra-violet light of gas-discharge lamps into visible light are prepared according to methods evolved by *Lenard* round about 1890. These investigations in the field of luminescence are to be regarded as classical and have not yet lost any of their value. They revealed in particular the fact that the luminescence of substances like zinc sulphide is due to the presence of extremely small admixtures of certain metals, such as copper and silver (activators).

When it became evident that large quantities of luminescent materials would be needed for lighting purposes Philips began to take up the manufacture of these substances and, at the same time, the study of the phenomena of luminescence. This study covered both the physical aspects — such as the spectral composition and the intensity of the luminescent light as a function of that of the incident rays, the decay of luminescence as a function of time in the case of discontinuous irradiation, the relation between luminescence and temperature, luminescence under the influence of cathode rays and X-rays — as well as the chemical aspects, such as the influence of the composition and of the conditions during preparation upon the luminescence, and the influence of admixtures (quenchers, sensitizers) upon the intensity of fluorescence. Often the chemical and the physical problems are so closely interwoven as to be inseparable, so that close cooperation between physicists and chemists or the combination of physicist and chemist in one person is essential for these investigations.



Phosphors are exposed to an electron beam in a vacuum tube and tested for their fluorescing properties.

Important physical results lay in the deeper insight thereby obtained into the mechanism of fluorescence and phosphorescence, the transfer of excitation energy in phosphors and the relationship between persistence and quenching as a function of temperature (F. A. Kröger, H. A. Klasens).

Chemical results lay in the deeper insight gained into the properties of zinc sulphide and related compounds and of substances such as Zn_2SiO_4-Mn_2SiO_4 and other manganese phosphors (Kröger).

In the foregoing some cases have already been mentioned where organic compounds were studied. Mostly these concerned simple organic molecules which are very closely allied to the inorganic compounds.

But work has also been done in the field of organic chemistry proper, namely in that of compounds with large molecules, such as gelatin, cellulose, proteins and synthetic resins. Gelatin was thoroughly studied because this substance serves as base for the Philips-Miller tape (Dippel, Vermeulen) used for sound reproduction; this tape consists of a celluloid carrier with a transparent coating of modified gelatin, covered by a non-transparent layer of HgS-sol in which the cutter of the recording apparatus traces a "sound track".

Artificial resins were originally studied with a view to the possibility of using these materials for making the bases of radio valves and loudspeaker baffles, and later for making various parts and the cabinets of radio sets from these materials. The phenols hardened with formaldehyde, which had been named bakelite after the Belgian chemist *Baekeland*, were placed on the market by Philips in a large number of varieties under the trade name "Philite".

A third group of organic compounds, the diazo compounds, was intensively investigated in the Philips laboratory when it was contemplated to produce dye-line paper, in addition to the mercury lamps destined for the dye-line process. With the aid of these compounds in various carriers, such as cellophane, paper, cellulose esters, etc., materials for photographic reproduction were produced which, via conversion of the products of dissociation through light into a developable latent metallic image, ultimately yield silver images with a resolving power of 1200 lines per mm, which appear to present interesting possibilities of application (Dippel, R. J. H. Alink, K. J. Keuning).

Attention was also given to the study of colloid-chemical problems. One result of these investigations, which was of importance for the manufacture of radio valves, was a new method of coating cathodes with a layer of oxide, covering them by means of electrophoresis with finely distributed carbonates of barium and strontium, which are afterwards turned into oxides. Also the insulating layer around the filament of indirectly-heated cathodes is applied in this way (De Boer, Verwey, H. C. Hamaker).

This concludes the review of chemical research in the Philips laboratory prior to 1940. A number of investigations carried out with metals and with non-magnetic and magnetic ceramic materials, glass and semi-conductors will be dealt with in the last section.

IV. X-rays

From the radio valve it is but one step to the X-ray tube. Any diode is in principle a source of X-rays, and it does in fact become so when the electrons strike the anode with sufficient energy. With the X-ray tube problems arise which are similar to those encountered with the radio valve, such as the shaping of the electrodes with respect to the nature of the electric field, the paths followed by primary and secondary electrons, and the focusing. To these are added the typical problems connected with high tensions, as for instance the efficient distribution of potential differences. Since only a small fraction of the electron energy is converted into X-rays and the rest is absorbed in the anticathode in the form of heat, it is a great problem how to carry off that heat; by providing for sufficient dissipation of this heat and giving the anticathode a suitable construction it has to be ensured that in the focus the material struck by the electrons does not melt.

Of physical importance is the measuring of the intensity of the X-rays and the strength of the dose. Much attention has therefore been devoted to this problem by A. Bouwers, who took up X-ray research in the Philips laboratory in 1920.

In the construction of X-ray tubes one is confronted with such problems as the safeguarding of the users of X-ray apparatus against scattered rays and high tension, the raising of the specific load, the requirement of easy handling, and the applications in connection with a proper formulation of the optical requirements which the apparatus has to answer; these applications may be of a medical (diagnostics and therapy), a physical (examination of crystals with X-rays) or a technical nature (examination of materials).

An important discovery was the possibility of fusing glass to metal, for which the chrome-iron already mentioned was used. This made it possible for Bouwers to build rugged X-ray tubes, which, since the part where the rays are generated is entirely enveloped in metal, afforded the maximum of safety against undesired radiation and, moreover, allowed of a simple safeguarding of the user against high tension by earthing the metal shield.

Tubes came to be developed for still higher voltages and for larger powers, both for diagnostic purposes and particularly for therapy.

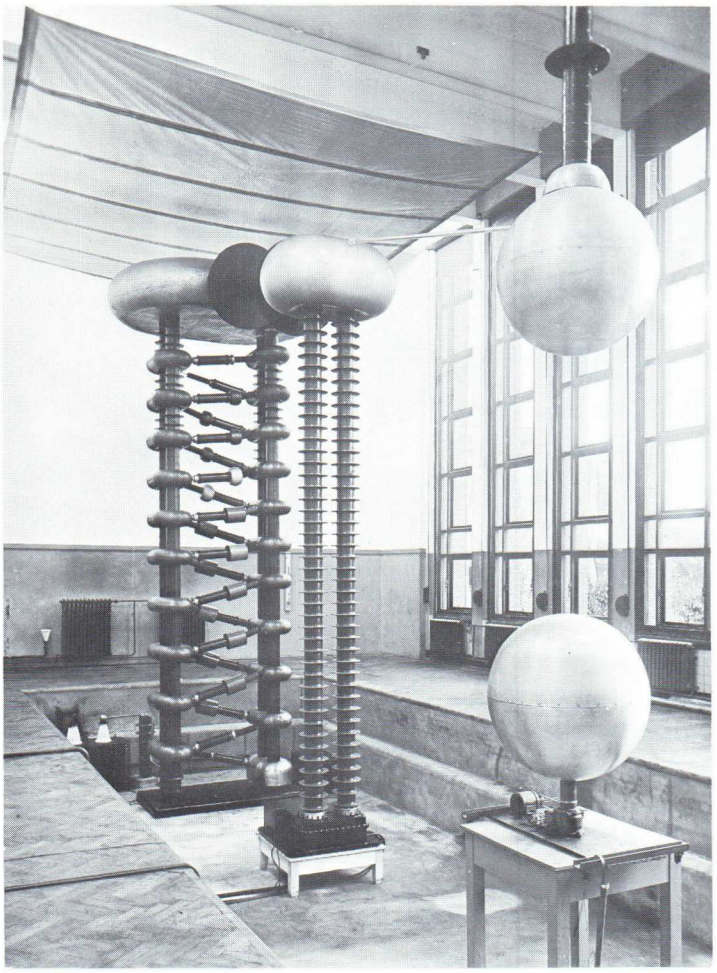
A special form of X-ray tubes which should be

mentioned is the rotating anode tube, which was first developed by Bouwers in cooperation with J. H. van der Tuuk. The problem of the dissipation of heat in these and other anodes was thoroughly investigated by W. J. Oosterkamp.

Further, special tubes were designed for crystallographic research with the aid of X-rays, in which the anticathode was made of special materials, such as molybdenum, copper, iron, etc., because these are required to yield approximately monochromatic rays of a known wavelength.

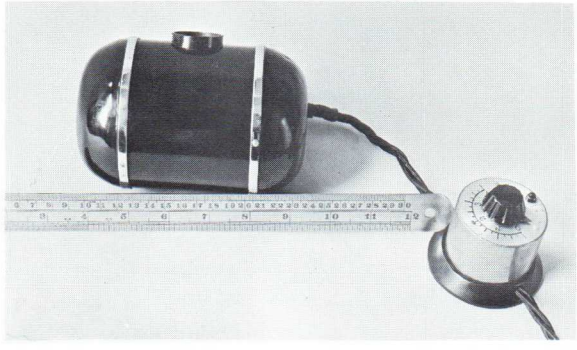
In the beginning all X-ray tubes were made in the laboratory, but later a separate factory was opened. The main features in the manufacture of X-ray tubes are well studied and well applied technology (getters!) and the exercising of extreme cleanliness in handling the materials.

Apart from the development of the X-ray tubes themselves, that of the apparatus for supplying the high tensions required for the working of the tubes, as also that of the control desks, forms a considerable part of Philips' activity in the field of X-rays. Often alternating voltage has to be transformed into direct voltage, for which special rectifying valves are needed, which formerly had a tungsten cathode but now have either an oxide-coated cathode or one of thoriated tungsten.



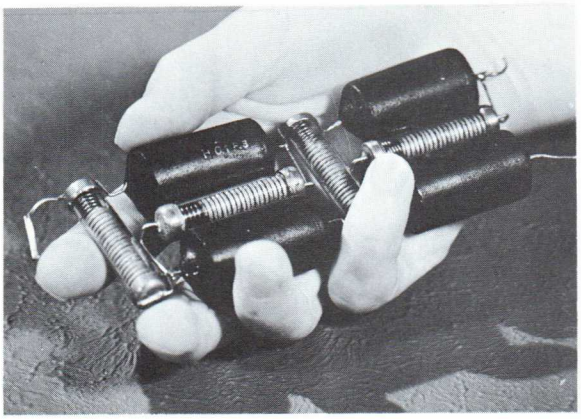
One of the first practical executions (1937) of a cascade generator, for 1.7 MV direct voltage.

Side by side with the aim towards higher tensions and greater powers, provision was also made for cases where a relatively low voltage and a small power suffice. Small X-ray apparatus was therefore developed in which the tube and the transformer are incorporated in one single unit. Such units serve as portable apparatus for diagnostic work and, for example, as X-ray apparatus in dentistry. It has even been possible to produce a complete X-ray apparatus so small that it can be carried in the pocket, and yet it yields a satisfactory beam of X-rays.



One of the first experimental X-ray units of very small dimensions (1933), a forerunner of the "Centralix" and "Oralix" apparatus.

From investigating current sources for high voltages Bouwers arrived at the principle of voltage multiplying by means of a cascade circuit. Subsequently it appeared that this principle had already been recorded by *Greinacher* and that it had also been applied by *Cockcroft* and *Walton* for nuclear-physical research in the Cavendish laboratory. As a special feature of Philips' development in this field is to be mentioned the elegant solution of the heating of the filaments in the valves by high-frequency current (A. Kuntke). A number of high tension generators of this type, for tensions

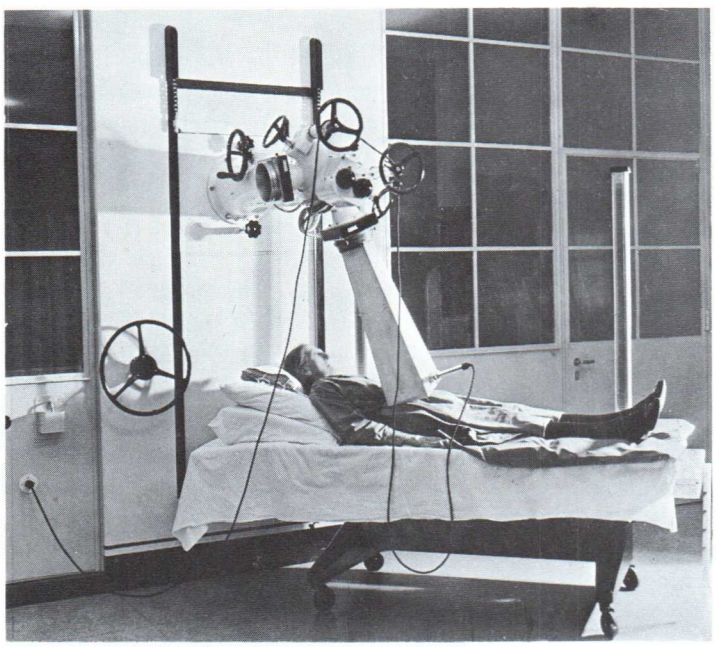


Small rectifier in cascade connection, consisting of selenium valves and capacitors. Connected to 220 V alternating voltage it yields 1200 V direct voltage, with which, for instance, a radiation counter tube can be fed.

up to 2 MV, have been built at Eindhoven for institutions in several countries.

Once it became possible to generate high tensions it was only a logical sequence to take up also research in nuclear physics. Bouwers and F. A. Heyn designed an ion-accelerating tube with which the ions of deuterium, formed in a separate ion source, could be accelerated to an energy of 1.2 MeV and focused upon a target of beryllium or lithium. Thus a powerful neutron source is formed by means of which materials placed in its vicinity can be made radioactive. In order to gain experience in the working of this apparatus, scientific nuclear research work was carried out with it. In the course of that work Heyn discovered the $(n, 2n)$ reaction with the elements Cu and Zn, by which reaction the nucleus is struck by one neutron and yields two neutrons, this being accompanied by the formation of a radioactive isotope of the same atomic number but lighter than the basic isotope. In 1939 A. H. W. Aten, Bakker and Heyn studied the transmutation of uranium and thorium by neutrons.

In this connection mention is also to be made of a neutron tube without a separate ion source constructed by Penning, whereby deuterium ions are accelerated in a gas discharge and the target is placed in the discharge tube itself. In order that sufficiently high voltages can be applied and the free path of the ions made large enough, the tube is filled with deuterium under a very low pressure and the discharge space is placed in a magnetic field so as to constrain the electrons to follow long paths, just as in the case of the manometer described earlier, thus making it possible for the discharge to be maintained under the low pressure.



Installation for X-ray therapy, working with a voltage of 400 kV, supplied in 1941 to the Academic Hospital at Groningen.

V. Mathematics and theoretical physics

With the growth of the laboratory more and more interest was taken in theoretical-physical and mathematical research. Whereas on the one hand, besides development work for practical applications, experimental physical research is

necessary to obtain a deeper insight into the problems, so on the other hand fruitful experimentation is only possible when at the same time the necessary attention is devoted to theoretical physics. This implies that the necessary mathematical apparatus has to be mastered and where possible further developed. This applies more or less to each of the groups already dealt with. Both the study of gas discharges and radio research, as well as the physico-chemical studies and the work on X-rays, had each in turn their own mathematical problems.

In regard to gas discharges there was the mathematical treatment by G. Hertz of the diffusion of electrons in an electric field. The method followed by Hertz has served as an example for many theoretical calculations in this field (Druyvesteyn, Penning, De Groot). This brought us a step nearer to the ideal: to explain the various forms of discharges with the aid of a small number of data concerning elementary processes (the probability of excitation and ionization in the case of collision of electrons and atoms, etc.), in much the same way as the kinetic gas theory relates the measurable quantities, such as pressure and temperature, to the elastic collisions between gas molecules.

Further it has already been seen, for instance, that with the aid of considerations of similarity the phenomena in discharges in mercury vapour of high pressure can be reduced to one single aspect. Once he had become familiar with this method of calculation, and following upon a publication by *Nusselt* (1916), Elenbaas was able to apply similar considerations to the phenomenon of thermal emissivity through natural convection. The need for this arose from the manufacture of blocking-layer rectifiers in connection with the dimensioning of cooling fins, but also other problems of convection can be considered in this light, such as the heat dissipation of horizontal cylinders, which brings us back to the gas-filled incandescent lamp designed by *Langmuir*.

The part of the work programme that lent most stimulation to the practising of mathematics was radio. Mention has already been made of important mathematical problems which arose in connection with the propagation of waves. Van der Pol took up the study of the wave equation and the potential equation, not only in three dimensions but also in the cases of fewer or more than three dimensions.

As is known, the propagation of waves can be treated in two ways, either as a whole, by the solution of a "wave equation", or by an approximative solution of the behaviour of a narrow beam or "ray". The latter method is well known from the theory of light; the image produced by a lens is not usually studied by starting from the representation of a wave but by investigating how the rays of light are refracted by the lens. In the case of radio waves Bremmer has investigated in how far results can be reached with this geometrical-optical approximation.

In addition to radio, particularly the theory of atoms has formed grounds for excursions into the field of mathematics. It is remarkable how closely the mathematics of these problems are associated with the mathematics encountered in the field of radio and acoustics. In 1925 *Schrödinger* showed, for instance, that the motion of an electron under the influence of an electromagnetic field has to be described by a wave equation which shows a certain resemblance to the equation representing the propagation of light or sound in an inhomogeneous medium, and that the classical consideration of an electron as a "particle" describing a "path" is to be compared to that conception as geometrical optics compare to wave optics. The quantity used by *Schrödinger* to play the part of "field strength" or "deformation" of the medium, and which he denotes by Ψ, has the property that $|\Psi^2|$ is a measure for the probability of finding a particle at a certain place. Owing to the similarity between these problems and those of the propagation of waves encountered in radio and acoustics, theorists in the respective fields soon understand each other and find interest in the results of each other's work.

In dealing with the problems connected with oscillations in networks or the propagation of radio waves one mostly has to do with more or less complicated differential equations. *Heaviside* showed that in many cases the differential symbol, $\mathrm{d}/\mathrm{d}t$, can be regarded as an algebraical quantity (usually represented by D), with which ordinary arithmetical operations can be carried out. In this way he was able to derive deep-lying results by simple means.

This operational or symbolic calculus was at first received with much scepticism, but later *Carson* (1926), for instance, found it to be justified. What *Carson*'s formula amounts to is the furnishing of a function $\mathrm{f}(p)$ corresponding to a function $\mathrm{h}(x)$ by the *Laplaec* transformation:

$$\mathrm{f}(p) = p \int_0^\infty \mathrm{e}^{-px}\, \mathrm{h}(x)\mathrm{d}x.$$

The shape of $\mathrm{f}(p)$ depends, of course, upon $\mathrm{h}(x)$. This function $\mathrm{f}(p)$ is said to the be "image" or the

"translation" of the function h(x) and it is denoted by the symbol:

$$f(p) \doteqdot h(x).$$

The functions x^n, e^{-x}, sin x and cos x, for instance, have as their images

$$\frac{n!}{p^n}, \quad \frac{p}{p+1}, \quad \frac{p}{p^2+1} \text{ and } \frac{p^2}{p^2+1},$$

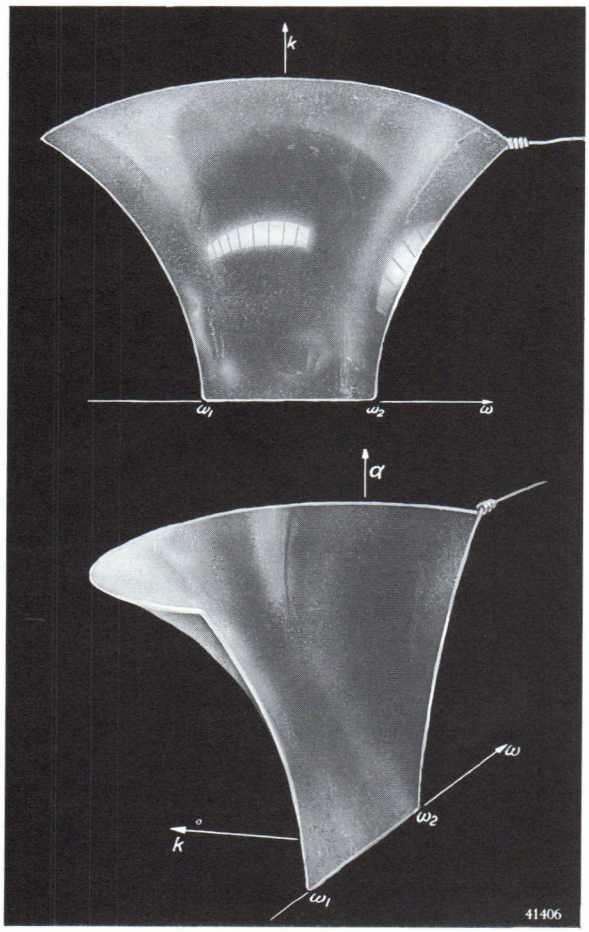
from which it is to be noted that often the image is a simpler function than the original. Thus relations between different originals can be deduced via simpler relations between their respective images.

Van der Pol and Niessen have worked out a great many images and from them derived new relationships between functions. Mention has already been made of the problem of the aerial over the flat earth, and for that, too, the symbolic calculus was employed, with the further help of the methods followed in the theory of the functions of a complex variable.

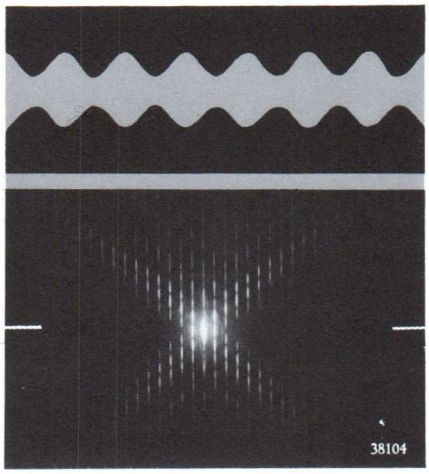
These results were obtained with the aid of the one-sided *Laplace* integral, with the integration extended from 0 to ∞. Later, Van der Pol and Bremmer went deeply into a symbolic calculus based upon the two-sided *Laplace* integral (integration from $-\infty$ to $+\infty$).

The space available does not permit us to enter into details regarding other mathematical work in connection with radio and acoustics. We can only refer to the numerous publications by Strutt on acoustic and antenna problems, Niessen's calculations on aerials and cavity resonators, C. J. Bouwkamp's work on radiation properties of antennae and acoustic and electromagnetic diffraction problems, that of F. H. L. M. Stumpers and Th. J. Weyers on frequency modulation, Tellegen's studies of network synthesis, from which the "gyrator" subsequently appeared as a new network element, and finally Kleynen's investigations of electric fields in radio valves, etc.

As regards mathematical and theoretical-physical contributions outside the realm of radio, reference is to be made to the investigations of J. Haringx into problems of applied mechanics, the publications by Bouma and G. Heller on the geometry of colour space, and those by Niessen on diamagnetism and by Verwey and J. Th. G. Overbeek on the theory of lyophobic colloids.



In the course of development of carrier telephony it was necessary to investigate how the behaviour of electric filters is affected by losses. This behaviour is governed by the *Laplace* differential equation. A graphical solution of this equation, with given boundary conditions, is obtained by stretching a film of soap between three-dimensional curves corresponding to the boundary conditions. The case illustrated here relates to a bandpass filter (frequency limits ω_1 and ω_2); k is a measure for the losses, α a measure for the damping.



Philips-Miller tape (upper picture) in which a sound track (10 times enlarged) has been cut. When light passes through the tape a diffraction spectum is obtained (lower picture) from which the *Fourier* analysis of the recorded sound can be read. In this case (sinusoidal signal) the spectrum consists only of components of the zero and first orders.

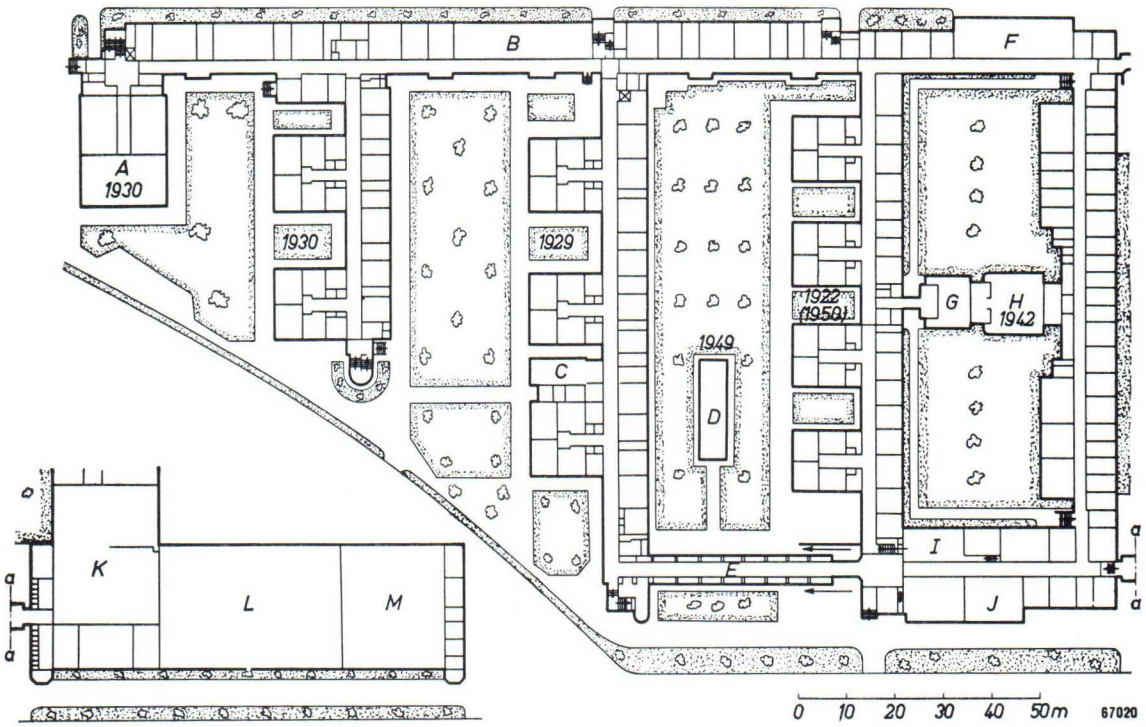
THE PERIOD 1940-1951

In the foregoing section an attempt has been made to give a review of the scientific research work undertaken in the period 1923-1940, and incidentally mention has been made of the technical products that emanated from that work.

The development of this research suffered a check when on May 10th 1940 enemy forces invaded the Netherlands and shortly afterwards Eindhoven and the Philips' works came under military occupation.

It is not the place here to enlarge upon the course of affairs during the period of occupation, which, as far as Eindhoven was concerned, lasted until September 1944. As everyone will realize, the state of tension arising from war conditions and the frequent acts of injustice, coupled with more direct causes such as scarcity of foodstuffs, clothing and means of transportation (bicycle tyres), and further a number of air attacks on the works, repeated air-raid warnings, etc., all tended to create an atmosphere adversely affecting work. It would be wrong to suppose, however, that scientific research was thereby brought more or less to a standstill. On the contrary, many investigations not directly concerned with practical applications were widened in scope and in some fields important results were reached, though care was taken to keep them secret from the occupying forces. The long working hours imposed by the enemy administrators were further turned to use for the exchange of experiences in all sorts of domains by organizing lectures, courses of instruction, etc.

After the liberation of Eindhoven in 1944 it took some time before a return to normal conditions could be established. In the first half of 1945 the northern half of the country was still in enemy occupation and even after its liberation there was no regular contact with the rest of the country owing



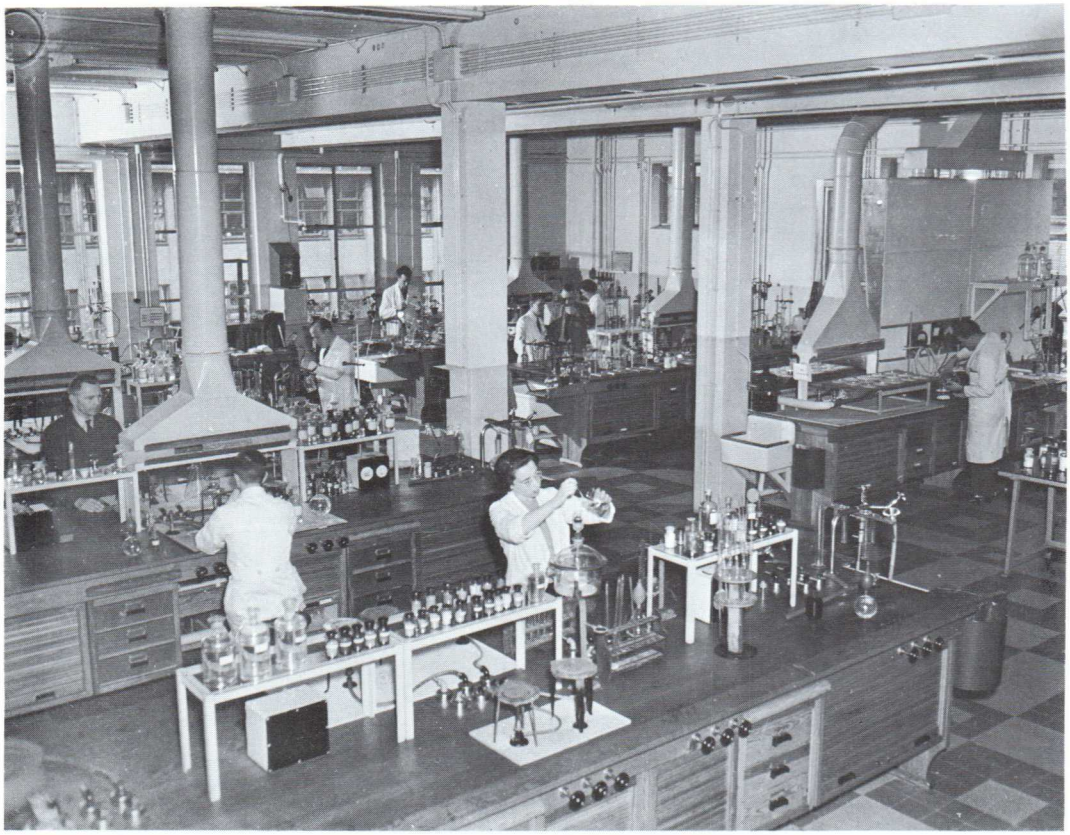
Ground plan of the Physical Research Laboratory as it has been since 1942. *A* high-tension room, *B* material-testing department, *C* television studio, *D* horticultural glass-house, *E* small greenhouses with artificial climate, *F* one of the chemical departments, *G* one of the battery rooms, *H* library, *I* installation for carrier telephony, *J* engine room, *K* glass-blowing shop, *L* central workshop, *M* room for testing transmitting valves. The wing *K-L-M* adjoins the main diagram at *a-a* on the right. Most of the buildings have either one or two upper storeys.

to the lack of railway communications. In order to help university students, with permission of the government in February 1945 a temporary academy was set up at Eindhoven where a large number of Philips' scientists assumed the rôle of professor, and this continued up to November 1945.

In this connection it may be well to point out that relations between the Philips laboratory and the higher educational authorities have never been confined to that special occasion. Long before that there had developed in the course of time a more and more intimate relationship between the laboratory and the universities, including the Technical University at Delft, partly on account of the fact that quite a number of scientists have in course of time left our laboratory to take up a professorship.

During the occupation one began to feel more and more the lack of literature from the outside world, so that as far as scientific work was concerned one had the feeling of living as it were in a vacuum. The possibility of publishing was also greatly restricted. The publication of Philips Technical Review, begun in 1936, had to be stopped in 1942, and publications in the Dutch journal "Physica" could only be made in the native language.

It was not until 1946 that journals and books began to come into the country again from abroad. On 1st January 1946 Philips Technical Review appeared again. Meanwhile, in October 1945, the first number had been issued of a new publication under the name of Philips Research Reports, containing scientific articles which bear a decidedly Philips character or which owing to their volume cannot easily be published elsewhere. Furthermore, part of the material suitable for publication which had been collected during the years 1940 to 1945 was published in book form.



Department for analytical chemistry in the new part of the Physical Research Laboratory (1950).

Meanwhile the need of more space was being felt in the laboratory, which in 1929 had already undergone a considerable expansion increasing ten-fold the amount of floor space originally available in 1923. A new wing had already been built in 1942, for the administrative staff and the library, but

until 1945 the latter had been housed elsewhere owing to the risk of fire due to air attacks. In 1950 a second floor was built onto the oldest single-storey part of the laboratory first erected in 1922, and that new floor was destined mainly for chemical work.

In 1946 a change was made in the management of the laboratory, Prof. Holst, whose 25th year of directorship had been celebrated in the laboratory in 1939, retired from that function. As adviser to the concern, he has still at the present day a word to say, be it indirectly, in the course of affairs, whilst at the same time he is devoting himself to the interests of the Technical University at Delft, where he holds of the Presidency of the Board of Governors. The direct management of the Philips laboratory was placed in the hands of a triumvirate formed by the physicist Prof. Dr. H. B. G. Casimir, the electrotechnician Ir. H. Rinia and the chemist Dr. E. J. W. Verwey.

For reasons which will be understood, in this section it will not be possible to do full justice to all the research work which has been undertaken in the Philips laboratory in the course of the last decennium and which, in part, is still in progress. Neither will it be possible to mention the names of all those engaged in the various investigations in the laboratory.

More so than in the preceding sections, here attention will be paid to materials and products and their applications.

Materials

It has been seen that in the period 1923-1940 scientific interest was directed for a large part towards gases, particularly as carriers of electric discharges. In addition, in various ways more and more interest came to be taken in solids. Mention has already been made of the investigation of thin crystal layers adsorbing foreign atoms. Activity in the field of radio demanded a closer study of the dielectric properties of glass and other insulators. For illumination engineering, too, glass had become an important material, as a result of the demand, for instance, for ultraviolet-transmitting glass and for intermediate glasses for combining glass with quartz glass. For the fusing of glass to metal suitable kinds of glass and alloys were required. In the technique of lighting fluorescent substances came to be applied more and more, so that the phenomenon of luminescence had likewise to be studied in the laboratory. Further, attention has also been drawn to magnetic materials, such as magnet steels and ferrites.

Generally speaking it may be said that round about 1940 interest was centred upon solids and the problems relating to the solid state, both from the point of view of purely scientific investigations and in respect of practical applications. In the following a number of materials will be briefly dealt with, but since it will be mainly the results that will be brought forward it is well to point out once more that in many cases these results are due for a large part to purely scientific research. As an example may be mentioned the investigations into "induced valency", which led to improved methods of preparing semi-conductors and luminescent materials, and the insight gained into the structure of "spinels", which resulted in a greater variety of "Ferroxcube" products.

Luminescent substances

The investigation of luminescent substances begun about 1935 led not only to the results already mentioned but also to a wider knowledge of luminescent tungstates and molybdates and of some activators such as Mn, Ti and U in different states of ionization.

Of particular practical importance were the silicates activated with Ti and magnesium arsenate activated with tetravelent manganese. With the latter substance, which shows a strong red fluorescence, it is possible, for instance, to improve considerably the colour of the light from the small high-pressure quartz mercury lamps.

The investigation of sulphides gave a better insight into the part played by "fluxes", such as NaCl, used in the preparation of luminescent substances. It appeared that it is the Cl^--ions that make it possible, for instance, for monovalent Cu^+-ions to be "built into" the lattice. The same can be reached by introducing into the lattice, which consists normally of bivalent ions Zn^{++} and S^{--}, trivalent cations such as Al^{3+}.

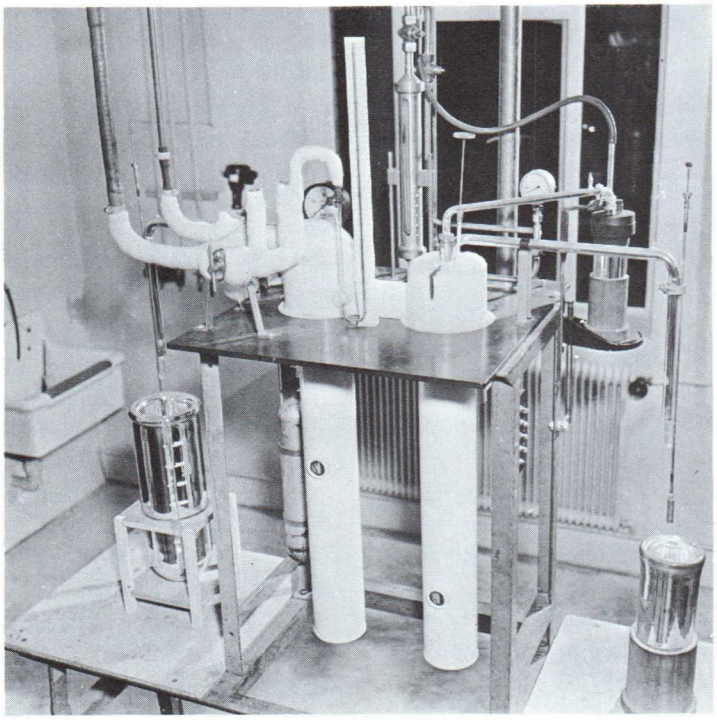
Results were also achieved in respect to luminescence brought about by bombardment with cathode rays; a deeper insight was gained into the phenomenon of saturation with increasing current.

For some purposes, such as radar, it is not the fluorescence but rather the phosphorescence (persistent after-glow) that is of importance. It was found possible to obtain a fluorescent screen with long persistence by making it in two layers, the first of which gives a blue fluorescence under the influence of the cathode rays, this blue fluorescence then being suitable for producing a green phosphorescence in the second layer. For other purposes, such as the televising of films,

substances with a short persistence are desired: a "field" produced on the fluorescent screen by an unmodulated electron beam is imaged onto the film, whilst behind the film is a photocell. This converts the light passed through, which is modulated by the variations in density of the film, into a signal which is passed to the transmitter.

ceramic materials (ferrites), because these have proved to be of great importance for high-frequency technique (radio and telecommunications).

Of theoretical importance was the interpretation of the losses which the magnetic ferrites show at very high frequencies (10^5 to 10^7 c/s): these losses were ascribed to the gyromagnetic effect



A simple hydrogen liquefactor has been constructed for the examination of solids at low temperature (20 °K). The hydrogen gas is fed in under high pressure and precooled to the temperature of liquid nitrogen, after which, through expansion, by employing heat exchangers according to the *Linde* method, it becomes partly liquid (*Joule-Kelvin* effect). This installation has an output of about 2 litres liquid hydrogen per hour.

Many other important applications of luminescent substances which have been investigated in the laboratory have to be passed over here.

Ceramic materials and glass

Ceramic materials are obtained by sintering together small particles which in themselves have a crystalline structure. We do not refer here to the more common materials like porcelain, steatite, etc., about which little research work has been done, but to special materials, as for instance those with a high dielectric constant, such as rutile (TiO_2), and titanates such as $BaTiO_3$, with which important work has been done in recent years.

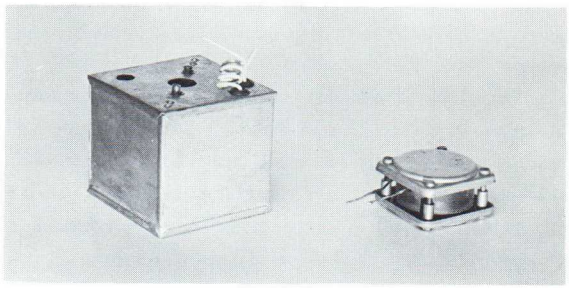
Particular mention is to be made of the magnetic

prophesied by *Landau*, and this conception was confirmed experimentally.

Of practical importance was the resultant knowledge gained of the fact that the less the initial permeability of the substance in the low-frequency range, and thus the greater its crystal anisotropy, the higher is the frequency at which the gyromagnetic losses become perceptible.

The ferritic materials, which combine great hardness and relatively light weight with a very small electric conductivity and favourable magnetic properties, such as low losses in a wide frequency range, are now being manufactured by Philips under the name of "Ferroxcube" in various compositions.

In addition to the ferrites, of equal importance are the ceramic semi-conductors, which will be dealt with below in a separate section devoted to semi-conductors.



Improved quality and saving of space are important advantages of the magnetic material "Ferroxcube" as compared with metallic magnetic materials. *Left:* can containing a bandpass-filter coil for carrier telephony, the core of which consists of nickel-iron wires and the jacket of dust-core material. The can is necessary on account of the small effective permeability of the dust-core material. Quality factor $Q = 220$ at 60 kc/s, volume 210 cm³. *Right:* coil with core and jacket of "Ferroxcube", $Q = 600$ at 60 kc/s, volume 44 cm³.

Glass is a material of fundamental importance for the manufacture of incandescent lamps and radio valves. Apart from the obvious physical properties, such as the softening point, coefficient of expansion and specific gravity, there are many other properties to be considered, such as the dielectric constant, dielectric losses in high-frequency electric fields, electric conductivity, spectral transmission, etc. Considering the large number of ingredients from which glass is mostly made, and consequently the numerous possible variations in composition, it would seem to be an almost impossible task to gain a clear insight into the effect of the composition of a glass upon its physical properties. Yet in recent times considerable success has been attained in this direction, particularly due to the better theoretical knowledge acquired as a result, i.a., of the work done by *Zachariasen*. It is now possible to form an idea of the internal structure of glasses and from that to predict their properties. Even though the reality proves to be more complicated than the theoretical model, the latter anyhow points the way to approximating the correct relations of the phenomena and for investigating those relations which are most promising for gaining the insight desired.

In cooperation with the glass works the following practical results have been obtained: soft glasses with small dielectric losses, glasses which have small dielectric losses independent of frequency, glasses which under the influence of X-rays do not discolour or which, on the other hand, very readily change colour, and glasses which are transparent to ultra-violet radiation.

Metals

Metals form an important and extensive group of materials. The investigation of tungsten and molybdenum and of metals like titanium, zirconium, hafnium and thorium has already been mentioned, as also that of alloys such as chrome-iron.

The investigation of magnet steels has likewise already been referred to. The very important "Ticonal" (an alloy of iron, titanium, cobalt, nickel and aluminium) with a high value of the product $(BH)_{max}$ and great coercive strength, was further improved by making it anisotropic by cooling in a magnetic field or by some other means.

The nickel-iron with anisotropic structure obtained by rolling, for use in loading coils, has also been mentioned elsewhere. Extensive investigations have been carried out as to the manner in which the texture of this material could be influenced by a suitable thermal and mechanical treatment.

Another group of products is formed by the welding rods, the coating of which has been the subject of particular study. Special mention is to be made of the new method of contact welding, for which purpose the rods are given a special coating with a high iron content.

It is partly in connection with this that extensive investigations have been carried out into the diffusion of gases in metals. This subject had become of real importance as soon as the manufacture of water-cooled metal transmitting valves and metal X-ray tubes was begun, and it proved to be of particular interest in connection with welding problems, especially as regards the penetration of hydrogen into the metal of the welding bead, causing porosity, cracks and fractures. In this connection mention is also to be made of more fundamental research, as for instance that concerning ageing phenomena such as occur, inter alia, in welding.

The absorption of gases by metals has particularly also been studied in the case of metals like zirconium and titanium, which are used as getters in X-ray tubes and transmitting valves. Incidentally reference is to be made here to the coating of anodes of transmitting valves with zirconium to give them a black surface with good heat-radiating properties and also to reduce secondary-electron emission. Investigations of a more technical nature in the field of metals led to new methods of shaping and improved methods for

determining the properties of samples of metal, an example of the former being the so called "lost wax" casting of metallic parts of instruments and apparatus which would be too costly if the ordinary methods of processing were followed, whilst an example of the latter is to be found in the micro-hardness meter.

This section on metals will be concluded by a short reference to investigations concerning oxidation and corrosion. Interesting aspects are presented by the method of hardening alloys by internal oxidation. A study of corrosion at high temperatures revealed that small quantities of molybdenum oxide may have a harmful effect. Investigations into the non-oxidizing property at high temperatures and the solderability of non-oxidizing metals were particularly of importance in connection with the development of the hot-air engine.

Semi-conductors

Between the insulators of an inorganic and an organic nature, such as glass, porcelain, amber, polystyrene, on the one hand and the metals as good conductors for electric current on the other hand, there is an important group of substances which as compared with metals show little conductivity, e.g. iron oxides, copper oxide, selenium, germanium, etc. These are of interest because they possess typical properties lacking in the other two groups.

Their principal property is a large, negative, temperature coefficient of their resistance, a property found in practically all representatives of this group. Some of them also show photo-conductivity, in that when exposed to light their resistance is reduced. Others, when provided in a certain way with electrodes, appear to have rectifying properties, which means to say that when the current flows in one direction the resistance is many times, say 1000 or more times less than in the other direction. There are also certain materials whose resistance appears to be dependent upon the electric potential gradient.

Although these properties have been known for a long time it is only during the last 10 to 15 years that it has been found possible to utilize them. Obviously a high temperature coefficient of the resistance offers many possibilities. Mention may be made, for instance, of the measuring of temperature and of radiation; furthermore, owing to the negative sign of the coefficient, there is the possibility of suppression of current peaks and application in delay devices, regulating apparatus, etc.

A peculiar feature of these materials is that often their specific resistivity is not a property inherent in the material, as is the case, for instance, with a metal, but that it depends to a large extent upon the antecedents, such as the temperature and the gas atmosphere applied in the preparation, the presence or not of impurities and many other factors.

In the field of oxidic semi-conductors some extensive crystal-chemical investigations have led to the production of semi-conducting materials answering high requirements of reproducibility and stability. Such a material is being marketed in various forms as "N.T.C. resistors" (negative temperature coefficient resistors).

Much attention was paid also to selenium. The resistance of selenium can be varied by the addition of foreign substances. This investigation has led to the production of rectifiers suitable for low voltages, e.g. 18 V, and, by connecting them in parallel, for high currents. On the other hand very small rectifiers are being made for currents of a few milliamperes, as used in the amplifying technique, for example, for modulator cells, which play an important part in carrier-telephony, and in series-connected small cells for rectifying high voltages. By a certain treatment of the surface of the selenium and choice of the electrode material it is possible, to a certain extent, to determine the voltages which are to be applied per rectifying unit.

During the years 1940-1945 there has been renewed interest in the crystal rectifier, a component used in the early days of radio. It appears that in the range of ultra-short waves, which has become of such importance, this element in a more perfected form offers advantages over rectifying valves. Extensive investigations are proceeding in the laboratory also in this field and have already led to a rapidly increasing manufacture of various types of crystal rectifiers.

There are indications that the possibilities of applications in the field of semi-conductors are not by any means exhausted, and that, by the very reason of their abnormal properties, we are only at the beginning of a development of great importance for electrotechnical engineering.

Products

This review will be concluded by referring to some products which have particularly come into existence in the last ten to fifteen years as a result of laboratory work. The reader is not to expect a catalogue with detailed descriptions. Many of the

products which will be mentioned have already been described more or less fully in recent volumes of this journal, whilst a description of some others will shortly be appearing in our columns. Suffice it, therefore, to mention in a few words only the most important products.

Sources of light

In the field of lighting we have already referred to the "TL" lamps, which as such have already long passed the laboratory stage of development. Further investigations in connection with these lamps are being carried out in regard to their colour rendering, for instance with new phosphors used in their manufacture. Particular mention is to be made of the application of "TL" lamps for irradiating plants. Since 1949 the Philips laboratory has had a large horticultural glasshouse at its disposal, in which various plants are being cultivated, and also a dozen smaller glasshouses illuminated exclusively with artificial light. With the large variety of phosphors available it is possible to investigate how the growth and flowering of plants are affected by the spectral energy distribution and the intensity and duration of radiation.

Further, there is the low-pressure mercury lamp without any fluorescent material but made with a glass which is transparent to ultra-violet radiation (2537 Å), viz. the so-called germicidal lamp. Other special sources of light are the flashlamp for the illumination of *Wilson* chambers and for other photographic instantaneous exposures, and a point source of light in the form of a gas-discharge lamp the light of which can be modulated up to high frequencies.



Glasshouse for biological investigations.

Radio, television, etc.

In the field of radio conditions had already reached such a stable state before 1940 that the normal development of valves and apparatus no longer belonged to the work of the laboratory, so that attention could be directed towards new developments. In the first place mention is to be made of the new valves for ultra-short waves (decimetric and centimetric waves). Development of these valves had already reached an advanced stage abroad during the war years, so that in 1945 Philips had a great deal to catch up with. Extensive investigations had already been carried out with radio valves in the metric wave range. Special tubes were developed based upon the velocity modulation of



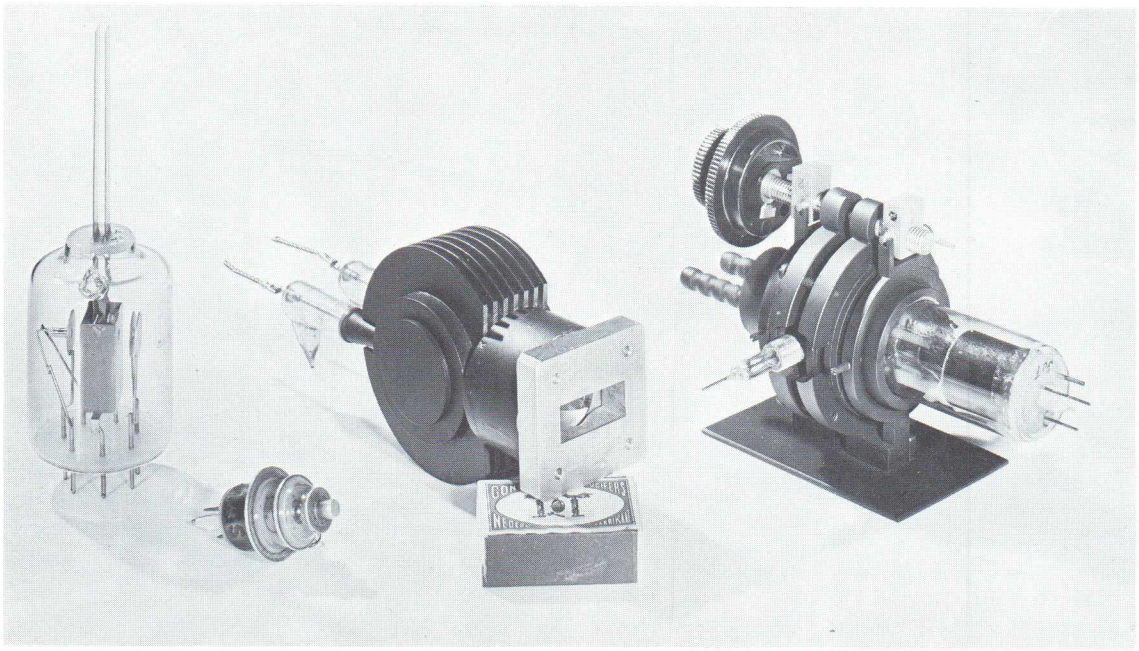
Cucumbers cultivated entirely under artificial light. The light is supplied by "TL" lamps. The temperature inside is prevented from rising too high by causing tap water to flow down the pane of glass in front of the lamps. Half-way down between the "TL" lamps is a TUV lamp, the ultra-violet radiation from which prevents the growth of algae on the glass.

electrons (the velocity-modulation valve, the klystron, the multireflection tube), employing a number of Philips' own inventions.

An important improvement of heated cathodes is the L-cathode, which contains a reservoir filled with barium oxide and enclosed by a porous wall of tungsten acting as the source of electron emission. With the aid of this cathode a disc-seal triode can be built for wavelengths down to 8 cm with a clearance of only 40 μ between cathode and grid. This new type of cathode offers important possibilities for many other applications as well.

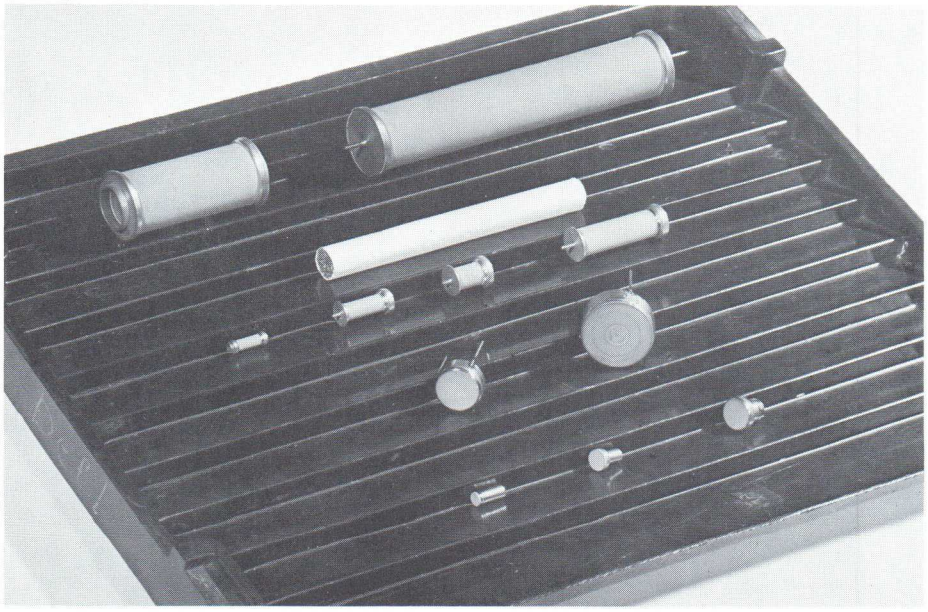
An intensive study is being made of the special circuits for generating ultra-short waves, by employing cavity resonators, and of modern means for conducting high-frequency energy with the aid of wave guides.

After 1945 the development of television had to be taken in hand again in the laboratory on account of the greater interest being shown in it everywhere.

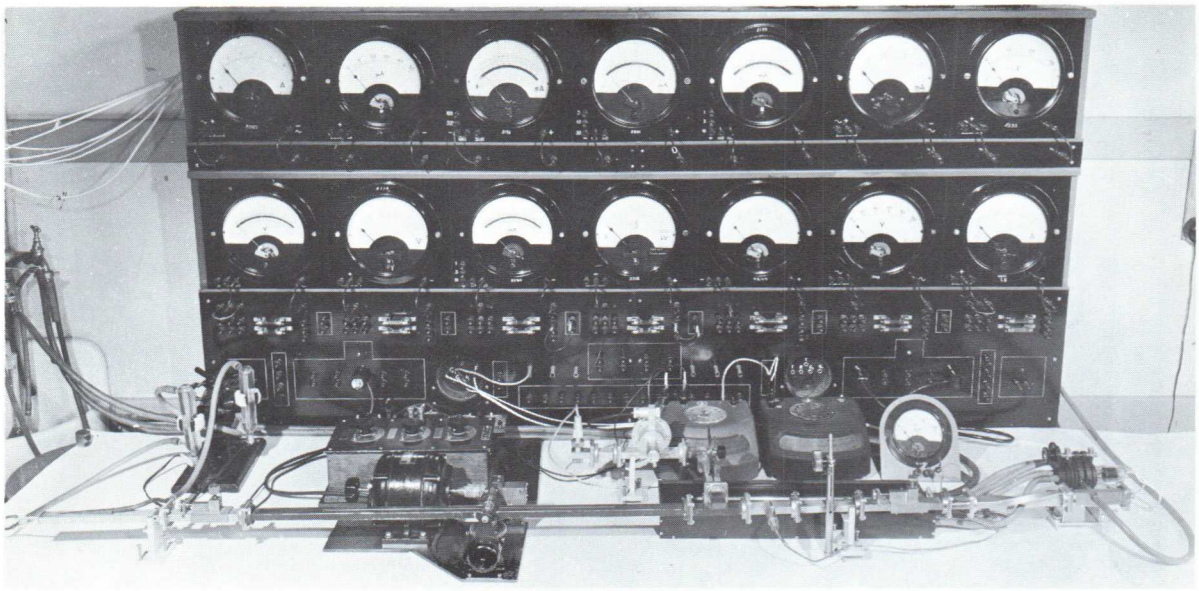


Tubes for centimetric waves. From left to right: a multireflection tube with a continuous output of 20 W on a wavelength of 12 cm; a receiving tube of the "lighthouse" shape for amplification of wide frequency bands (gain 10 at a bandwidth of 50 Mc/s, noise figure 10 db at a wavelength of 10 cm); a magnetron yielding pulses of 1000 kW at a wavelength of 3 cm; a velocity-modulation valve with a continuous output of 100 W at a wavelength of 3 cm.

At first a system of 567 lines was used, with 25 frames per second. In the audio channel frequency modulation was applied. As a result of many discussions on the question of the number of lines the transmitting apparatus was designed so that it could quickly be changed over to different numbers of lines. This proved to be of great value for judging the relative merits of various systems. When, later on, the Netherlands Television Committee advised the adoption of a system with 625 lines



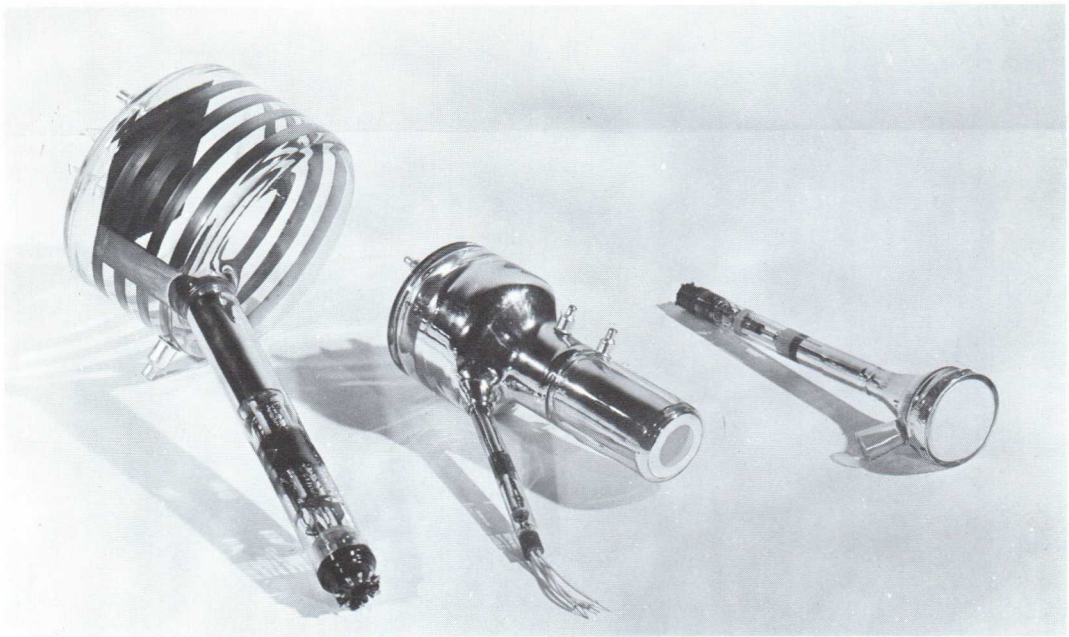
Some L cathodes (with a cigarette for comparison of size). The L cathode has a porous, metal, emitting surface and can withstand heavy loads, say 100 A per cm^2.



Set-up for taking measurements of a velocity-modulation valve (on the extreme right) at a wavelength of 3 cm. Here wave guides are used for conducting the energy.

and 25 frames per second the installation was made suitable for that system.

Further, attention was devoted both to the problems encountered at the receiving end and to those arising in the transmission. As far as reception is concerned, based upon the results of experiments already mentioned a system of projection television was worked out with the aid of a *Schmidt* optical system, using very small projection tubes with the screen coated with phosphors suitable for the purpose. Special precautions had to be taken against discoloration of the glass of these tubes under the influence of cathode rays and X-rays. In the optical system proper, use was made of a correction plate made of gelatin, for which a separate method of manufacture had to be developed.



Left: two pick-up tubes for television (an iconoscope and an image iconoscope). Right: a picture tube for projection television.

Of course attention had also to be paid to direct-view reception. One of the most important results of the work done in this connection was a considerable improvement in regard to flicker of the image by employing phosphors having a long persistence.

As regards TV transmission the most important advance made was in the construction of new pick-up tubes based on the principle of the image iconoscope. Further improvements lay in the television cameras and in the studio lighting.

Except for some short interruptions there have been regular experimental television transmissions from Eindhoven ever since 1946, for which purpose a studio, a control room and dressing rooms for artists were fitted out in the laboratory building. These experimental transmissions have helped much towards arousing the interest now being shown in this youngest branch of technical development both in the Netherlands and in adjoining countries.

Next in order to radio and television is the work done in the field of telecommunications. Ever since 1935 there has existed a department in the laboratory (organically not belonging to it but as part of the A.F. telephony department of the works) where apparatus for carrier telephony, i.a. a 17-channel system, and for A.F. telegraphy have been developed. About 1940, in cooperation with the laboratory, a modern carrier-telephony system for 48 channels was developed, in which the new core material for coils, "Ferroxcube", plays an important part. This system is already being used on an extensive scale in the Netherlands and in Switzerland, whilst it is to be installed also in Denmark in the near future. In this connection a word of grateful recognition is due to the Netherlands P.T.T. officials for their close cooperation. In 1945 the carrier-wave department was transferred from Eindhoven to the Philips' Telecommunication Industry at Hilversum, whilst a research group for telecommunications was established in the laboratory. Development work in Hilversum is proceeding further in the direction of the improvement and reduction in size of electrical and mechanical parts, with the aid of results reached in the laboratory.

Special carrier systems have been developed which are suitable also for short distances. Further, by developing such things as repeaters, a super-group through-filter and other filters, the laboratory has already contributed much towards a system now in course of development at Hilversum for incorporating some hundreds of channels in a coaxial cable.

In this laboratory work is continuing on new methods of modulation; for instance, a simplified form of pulse-code modulation has been studied, the quantum modulation, which makes it possible for conversations to be carried satisfactorily over long distances in spite of a high noise level. In addition, the possibilities of applying new materials and components for telecommunication apparatus are being studied.

In this connection mention is to be made of the new switching tubes with ribbon-shaped electron beam, which promise interesting applications in the field of telecommunications and electric computers.

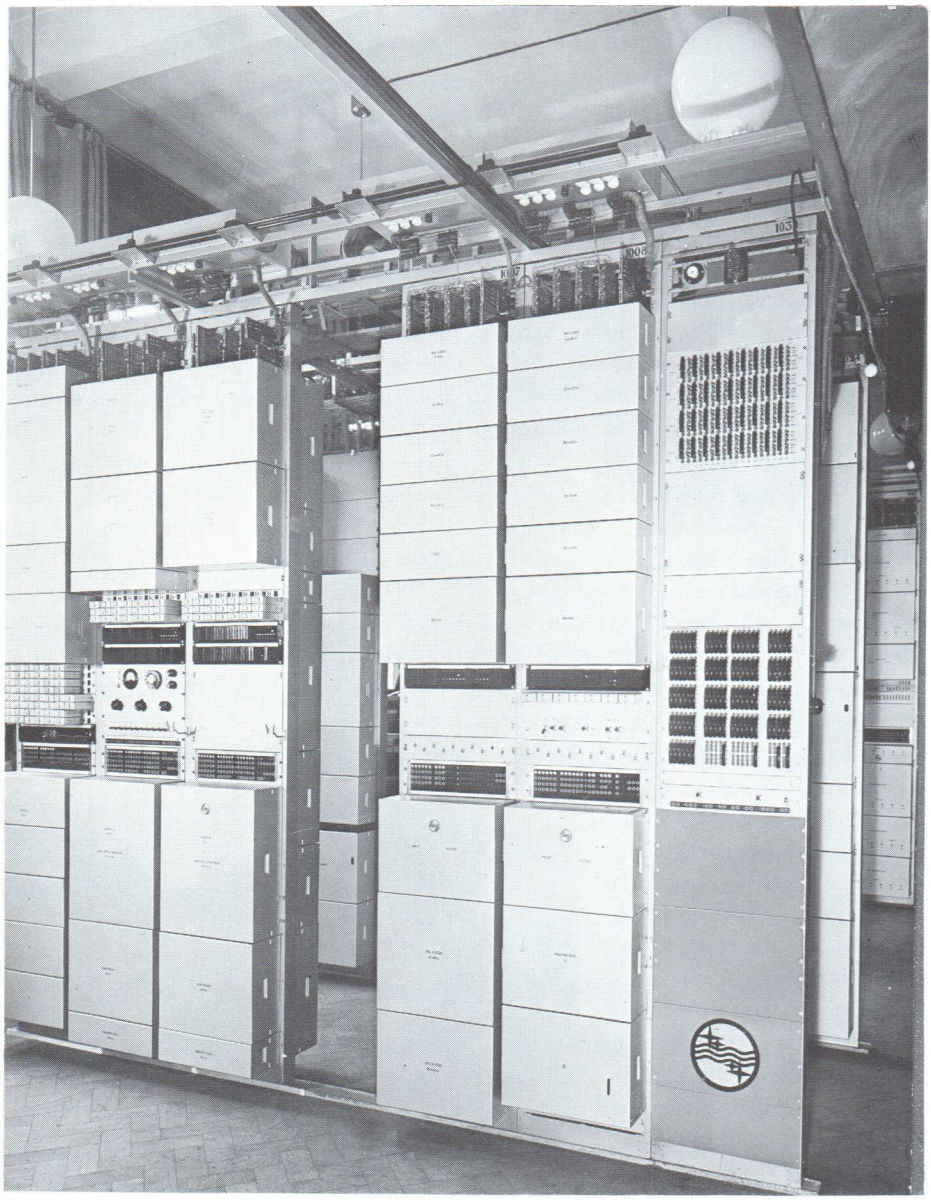
Special mention is to be made of the system of facsimile transmission worked out in the laboratory, by means of which documents can be transmitted and photographically recorded at the rate of 5000 cm^2/minute.

Of importance for meteorology is a newly developed radio sonde, in which extremely small radio valves with low current consumption are employed and with which experiments are now being carried out in conjunction with the Royal Netherlands Meteorological Institute at De Bilt.



Dr. A.F. Philips in front of the camera of the Eindhoven television transmitter.

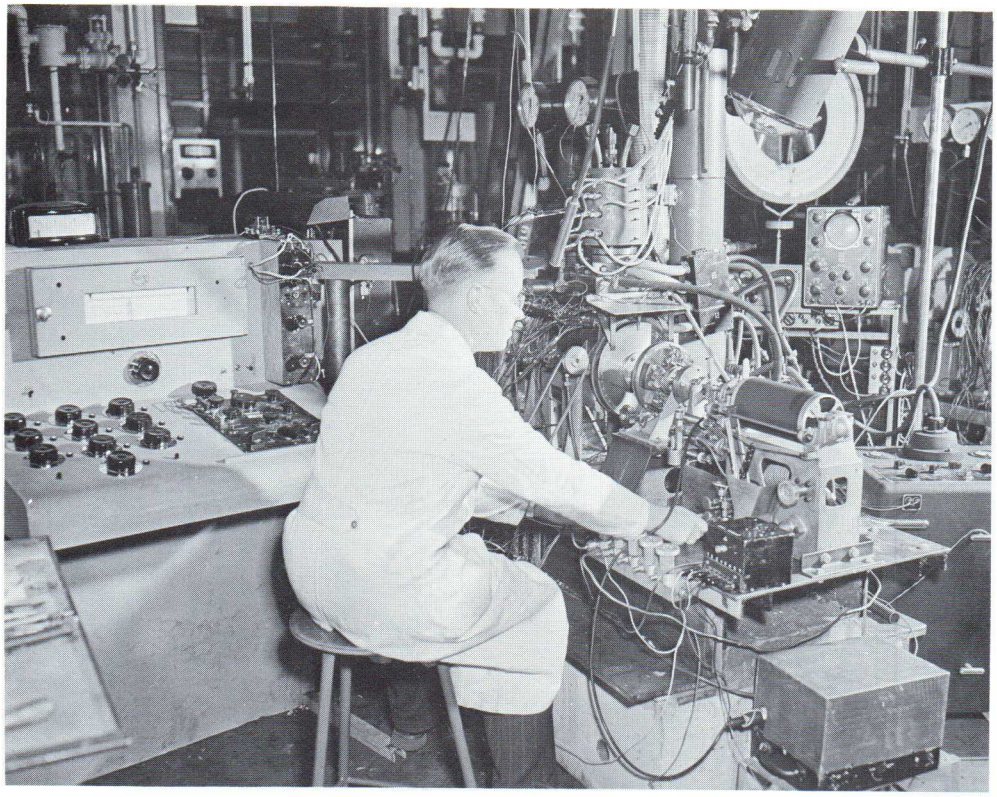
A product which certainly has some connection with radio, considering the need of a supply source in places where no electricity is available, but which represents the outcome of a development of its the piston of one cylinder is caused to act as the transfer piston for another, is of importance in connection with future possibilities the consequences of which cannot yet be fully predicted.



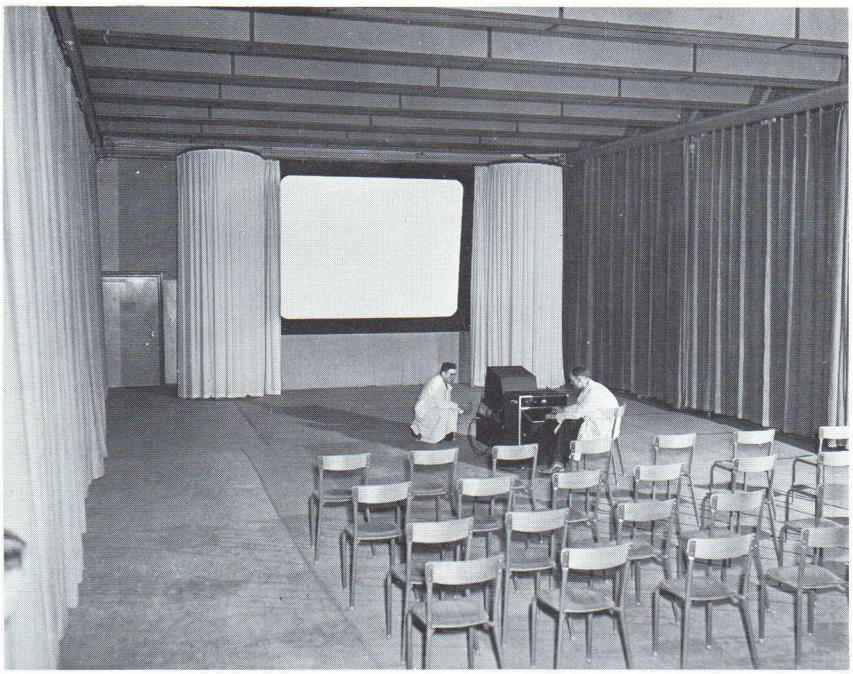
Experimental installation for carrier telephony.

own, is the air engine already referred to in this review. From a close study of the heat exchange between flowing gases and metallic walls and of materials suitable therefor, and further of the thermo-dynamics of the cycle, it has become possible to build a light high-speed motor with good thermal efficiency. A new principle, of application for systems with more than one cylinder, according to which

In the field of acoustics there are quite a number of new products to be mentioned. First of all the improved sound reproduction by means of stereophony, which has led also to the construction of small microphones. The ideal microphone, with small dimensions compared with the wavelength, also for high audio-frequencies, appeared to be most closely approximated by the condenser microphone,



Taking measurements in the testing of an experimental air engine. Energy input by electric heating, energy output via a loaded dynamo. Connected to the desk on the left are thermo couples measuring the temperature at various points (inside and outside). A *Farnboro* pressure indicator records the pressure in the cylinder as a function of time; the phase of this diagram is checked with the aid of a capacitive pressure indicator (connected to an oscillograph) and a stroboscope.



Hall for acoustical investigations (i.a. stereophony) and large-screen television (almost 3 m × 4 m).

which can now be made in a size of about 20 mm, and for special purposes even as small as 7 mm.

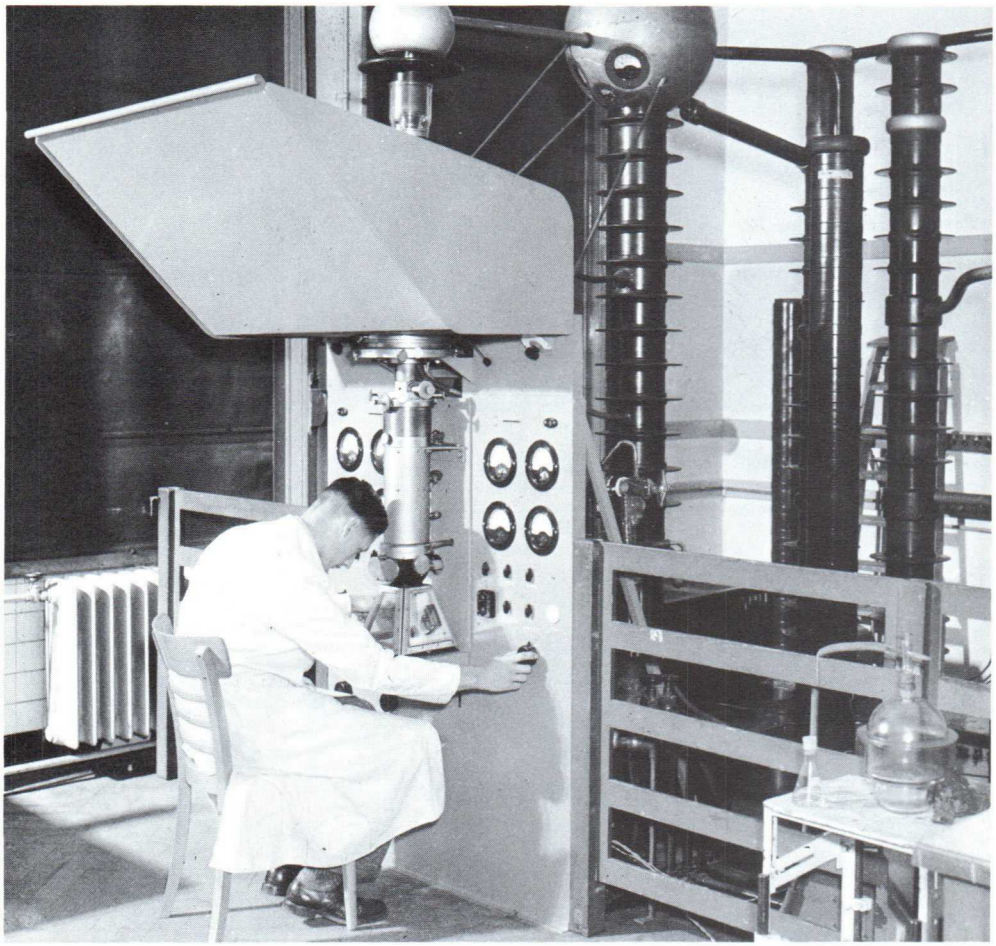
In the Philips laboratory the necessary attention has also been paid to the development, started in Denmark, of the magnetophone for recording and reproducing sound. This development obliged the manufacturers of gramophone records at last to relinquish the demand that it should be possible to play their records both electrically and mechanically, as a result of which considerable improvements became possible in the field of electrical gramophone reproduction by means of microgroove records. Thus the gramophone record may retain its position as an instrument of reproduction.

Physiological study and the resultant connection with the medical world led to the construction of hearing aids.

Experience gained in the amplification of signals of very low frequency, close to the limit of audibility, led further to the designing of a vector cardiograph, some specimens of which have been given to medical experts for testing.

Miscellaneous

Among the X-ray tubes there is a new one for contact therapy and, further, new high-powered tubes for X-ray diffraction with accessory apparatus.



Experimental electron microscope with an accelerating voltage of 400,000 V.

With the experience gained in the field of television it has been possible to construct an X-ray picture intensifier, by means of which the intensity of the screen picture can be increased to such an extent as to allow of screening patients in a room not blacked out. A number of these intensifiers are now being tried out in Philips' Medical Department.

As an offshoot, so to speak, of the development of cathode-ray tubes, aided by the experience acquired with high tensions and in collaboration



Microgram of influenza virus (specimen not entirely purified) magnified 21,000 times by the electron microscope. This virus occurs in the shape of small pellets as well as in a thread-like form.

with the Technical-Physical Laboratory of the Technical University at Delft, an electron microscope with a magnetic electron-optical system was constructed. An apparatus for 100 kV has been marketed, and in the laboratory an experimental apparatus for 400 kV has also been built. Recently, following upon previous experiments by Burgers, an emission-electron microscope has been developed, with which phenomena at the surface of metals (recrystallization) can be observed with a magnification of 1000.

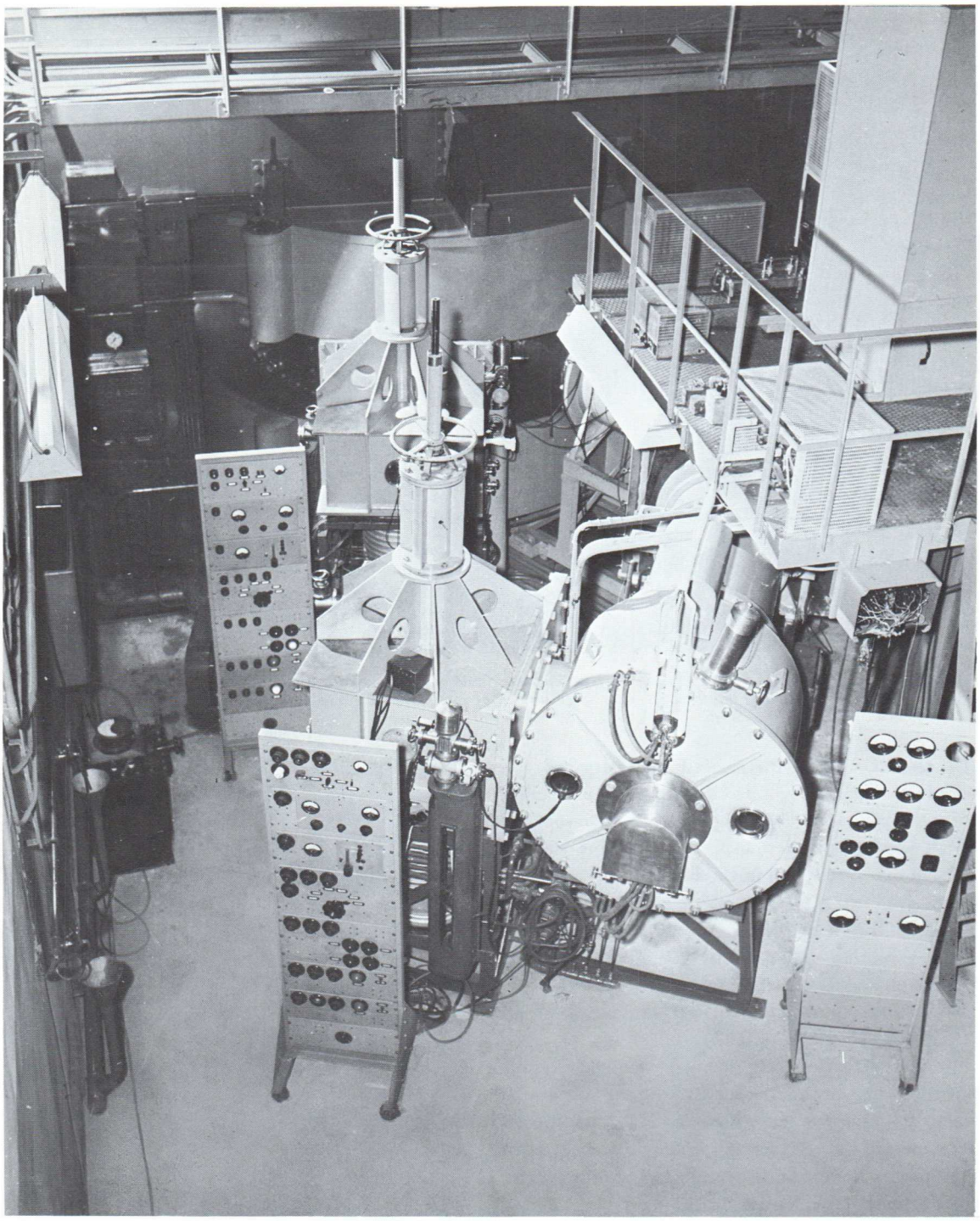
For the detection and measurement of X-rays and other kinds of rays, radiation counter tubes with accessory apparatus have been made, and a special electrometer with vibrating capacitor has been developed, which can be connected, for instance, to a dosimeter for X-rays.

Finally, nuclear-physical research has led to the construction of neutron generators with ion-accelerating tube (maximum voltage 1.2 MV) and also a synchrocyclotron, with which, for instance, deuterons can be accelerated to an energy of 28 MeV; this synchrocyclotron has been installed at Amsterdam.

Further, an experimental betatron with air coils has been developed, with which electrons can be accelerated to an energy of 9 MeV.



Portable radiation counter for detecting beta and gamma rays, fed entirely from built-in batteries. The counter reacts perceptibility to the rays from the luminescent paint on the hands of a watch.



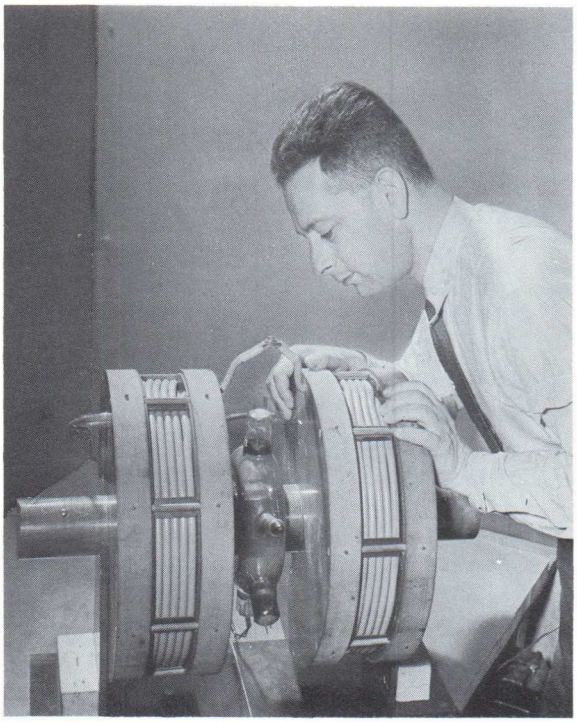
The synchrocyclotron installed in the Institute for Nuclear Research at Amsterdam. This cyclotron produces continuously deuterons with an energy of 28 million electronvolts. In the background the electromagnet, between the poles of which, where there is a maximum flux density of 1.38 Wb/m^2, are the two Dees (not visible in the photograph) inside which the deuterons describe spiral orbits. Between the Dees is an alternating voltage with a peak value of 15 kV and a frequency of about 10.7 Mc/s. The modulator (in the foreground on the right) varies the frequency periodically by an amount of 4%, so as to keep at least some of the deuterons to the desired orbits.

Conclusion

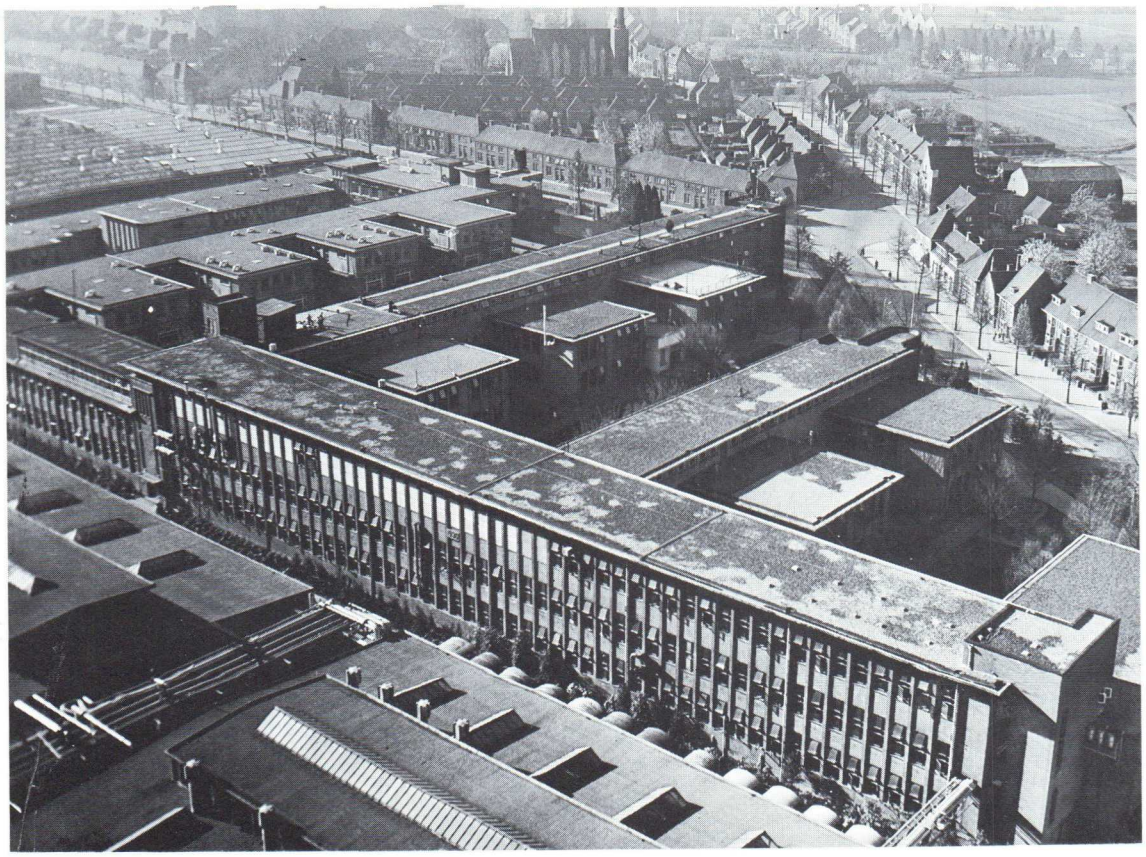
In the foregoing an attempt has been made to give a review of the activities of the Physical Research Laboratory of Philips' Industries in the course of the nearly forty years during which it has formed part of the Concern now celebrating the sixtieth anniversary of its foundation. It goes without saying that these activities could not have been achieved without the build-up of an extensive organization for the equipment of the laboratory and the workshops. In many cases these workshops have made substantial contributions to the realization of various projects.

The writer is fully aware that many sides of the development of the research work have been inadequately treated in this review. Due to lack of space, several subjects have received but scanty attention or have not been mentioned at all. However, the reader will be able to supplement this summary by referring to the Abstracts of Scientific Publications regularly appearing in Philips Technical Review and to the subject index given with the last number of each volume.

W. de Groot.



A betatron, with which electrons can be given an energy of 9 million electronvolts. The electrons travel round in the annular glass tube seen in the middle. The particular feature about the rest of the construction is that no iron yoke is used for the magnetic field.



FIFTY YEARS OF THE GAS-FILLED LAMP

by J. C. LOKKER *).

621.326.72



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

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

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

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

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

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

This lamp was a very considerable improvement

*) Formerly with Philips, now in retirement.

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

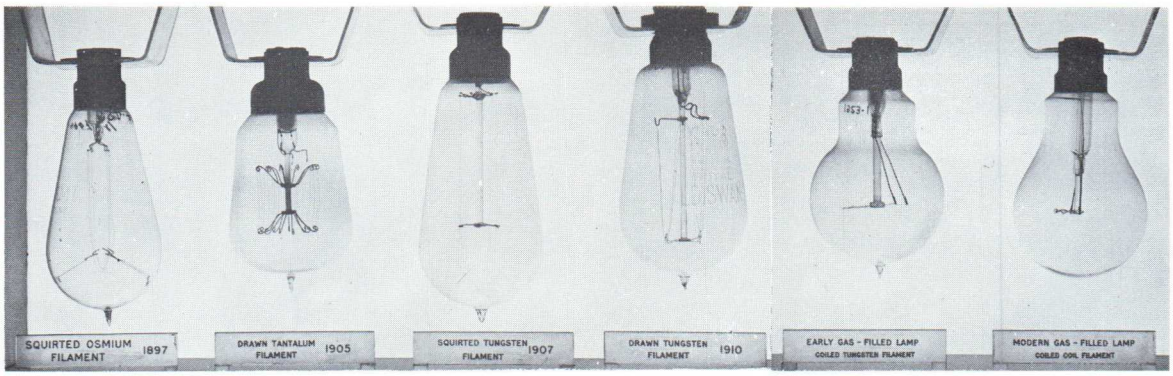


Photo Science Museum, London

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

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

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

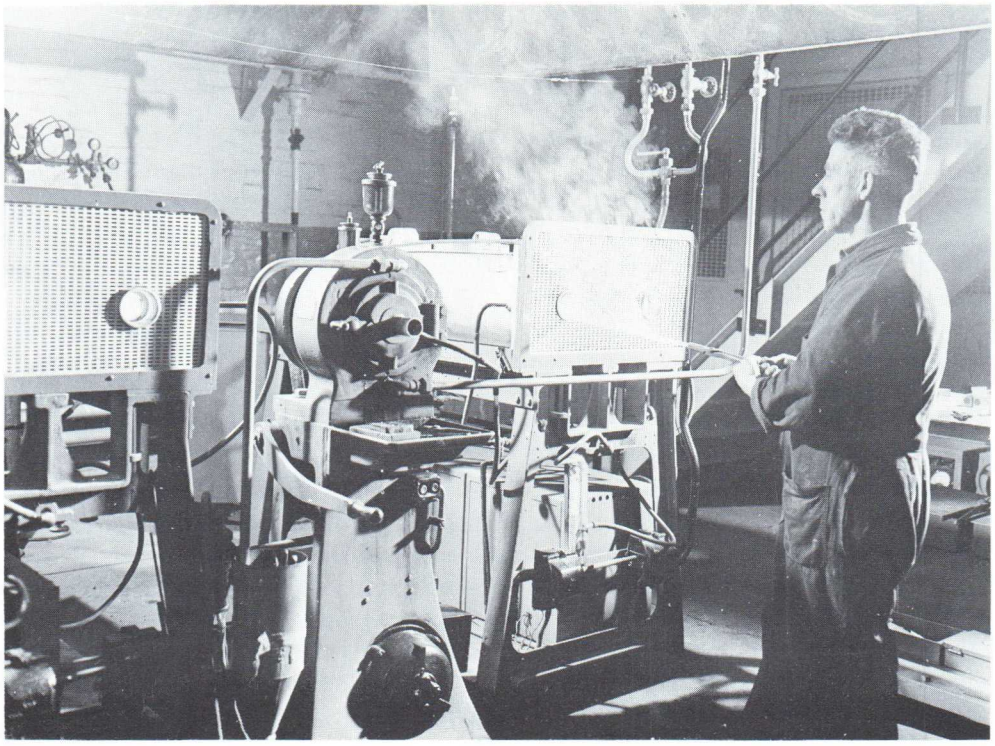
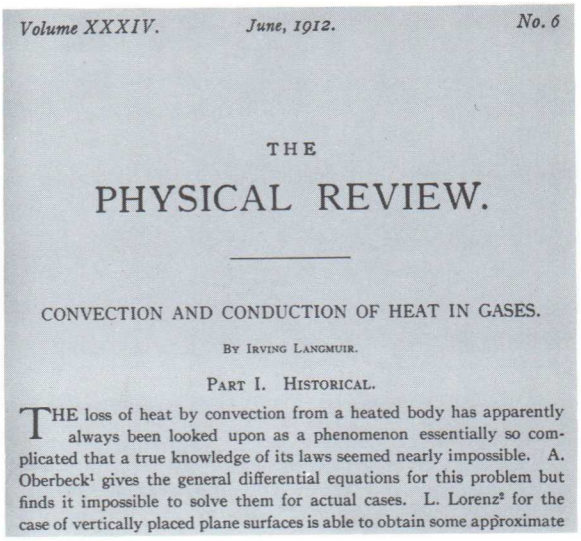


Fig. 2. Ductile tungsten wire is made by machine-hammering (swaging) sintered tungsten bars to increase their density after which they are drawn to the required thickness through hard-metal or diamond dies. The photograph shows a swaging machine, with the tungsten bar being introduced manually after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





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

THE

PHYSICAL REVIEW.

CONVECTION AND CONDUCTION OF HEAT IN GASES.

By Irving Langmuir.

Part I. Historical.

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

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

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

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

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

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

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

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

Langmuir's invention

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

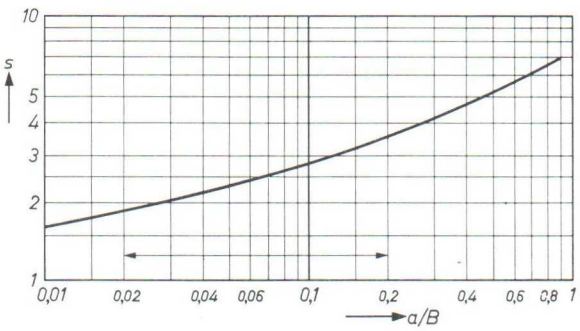


Fig. 4. Graphic representation of equation (4). In practice a/B varies only within the limits indicated along the abscissa.

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

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

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

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

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

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:

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

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	α 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]):

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

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

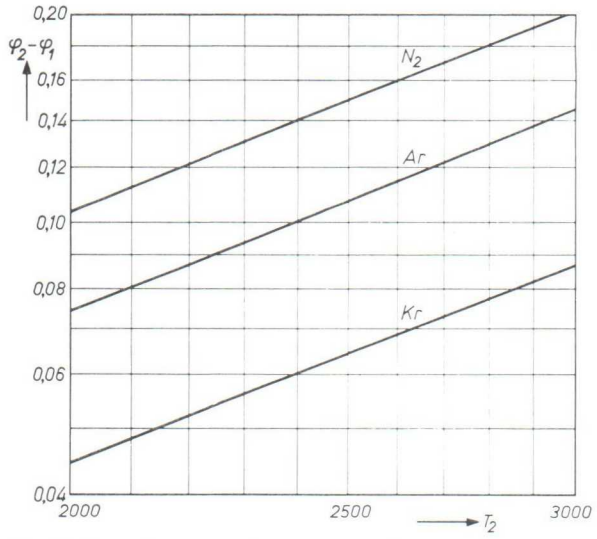


Fig. 5. Plot of the quantity $\varphi_2-\varphi_1$ in Langmuir's heat-loss equation (eq. 1) versus the filament temperature T_2 in °K, representing the theoretical values for different gases (in the case of N_2 no acccount is taken of dissociation).

[5]) Langmuir really gives the relation:

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

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

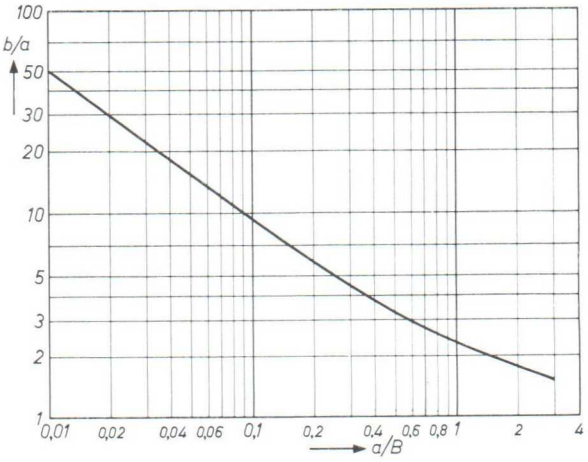


Fig. 6. Ratio of the diameter b of the stationary gas layer to the diameter a of the (cylindrical) incandescent filament, as a function of a/B, calculated using Langmuir's equation (8).

Realization of the invention

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

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

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

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

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

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

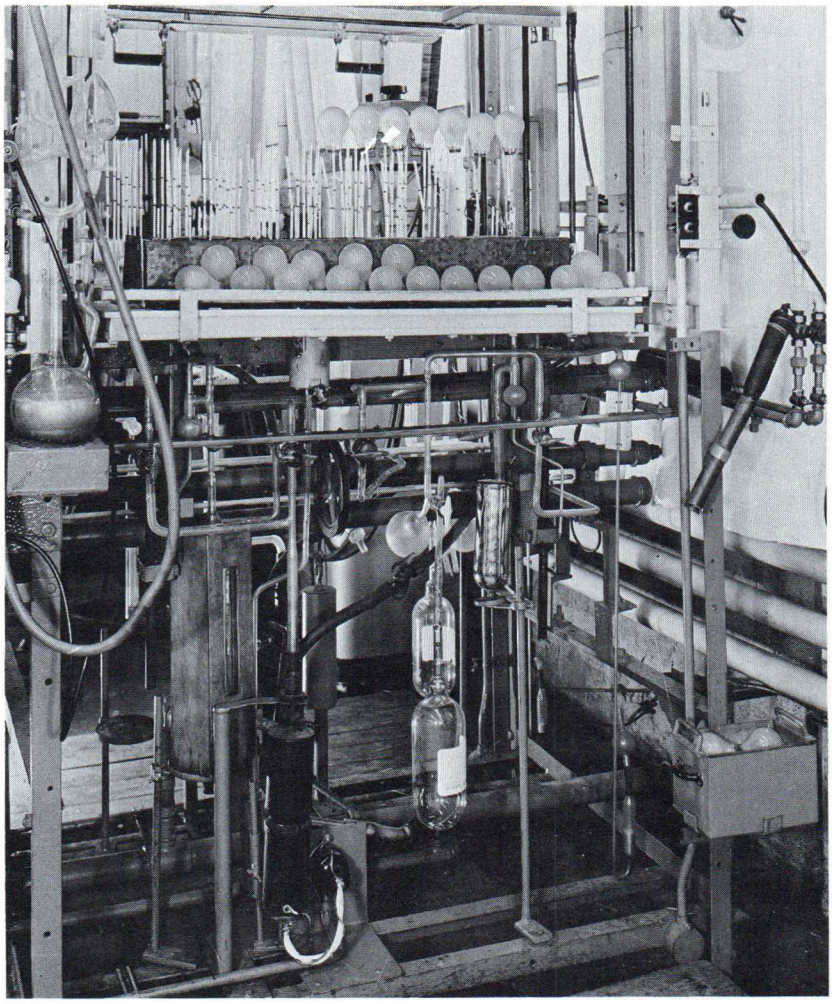
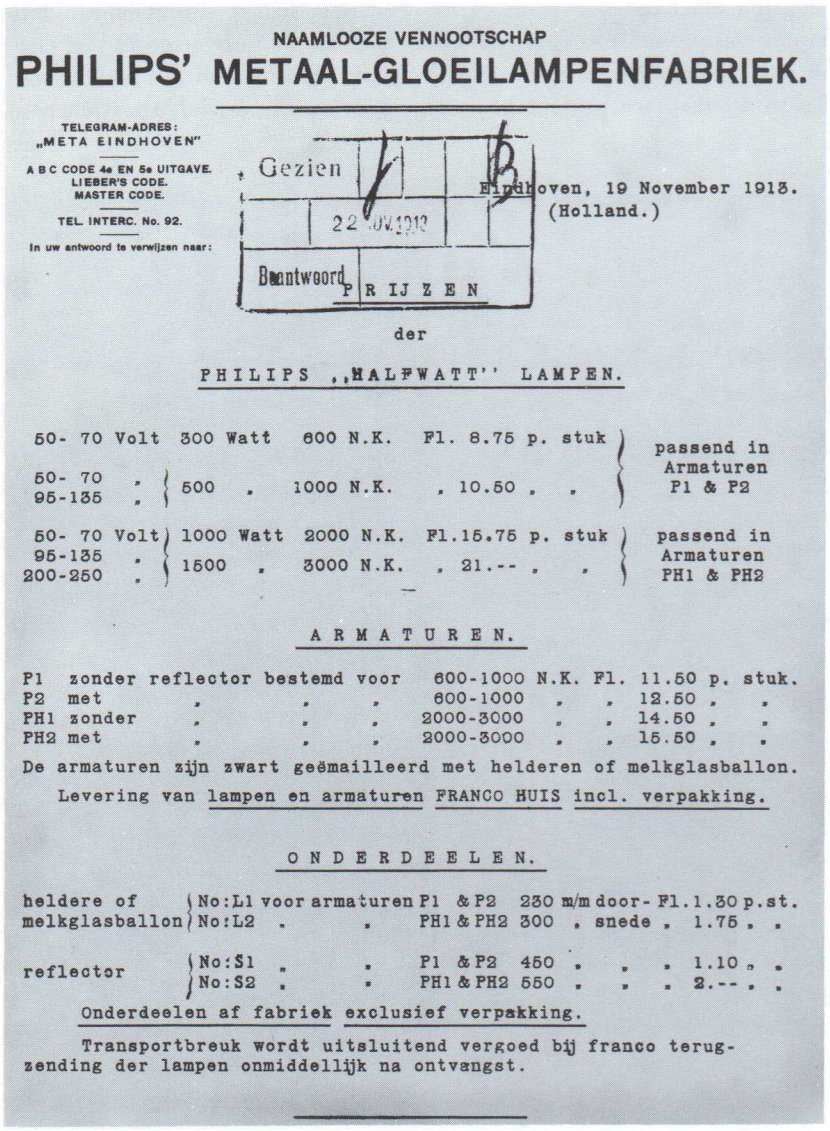


Fig. 7. When evacuating the lamp bulbs on the pumps the last traces of water vapour are removed by a liquid-air trap in the pump line (see the Dewar flask roughly in the middle of the photograph).

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

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



NAAMLOOZE VENNOOTSCHAP

PHILIPS' METAAL-GLOEILAMPENFABRIEK.

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

Gezien 22 NOV. 1913
Beantwoord

Eindhoven, 19 November 1913. (Holland.)

PRIJZEN

der

PHILIPS „HALFWATT" LAMPEN.

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

ARMATUREN.

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

De armaturen zijn zwart geëmailleerd met helderen of melkglasballon.

Levering van lampen en armaturen FRANCO HUIS incl. verpakking.

ONDERDEELEN.

heldere of melkglasballon	No:L1 voor armaturen	P1 & P2	230 m/m door-snede	Fl. 1.30 p.st.
	No:L2 " "	PH1 & PH2	300 " "	" 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.

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]).

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

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

[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.

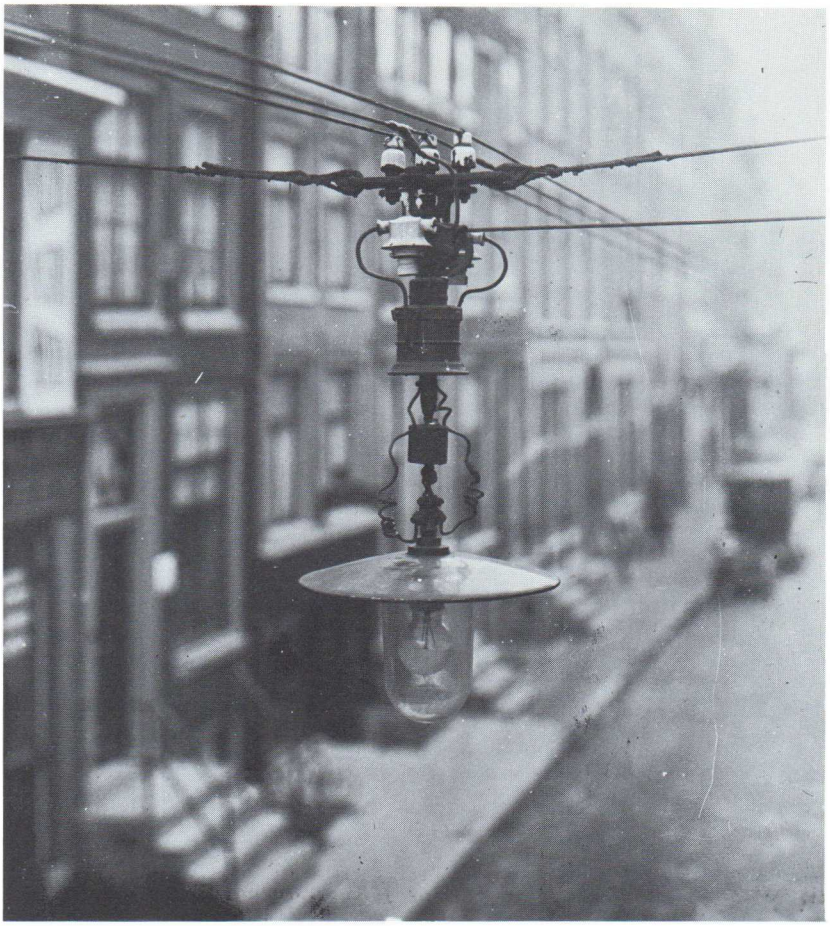
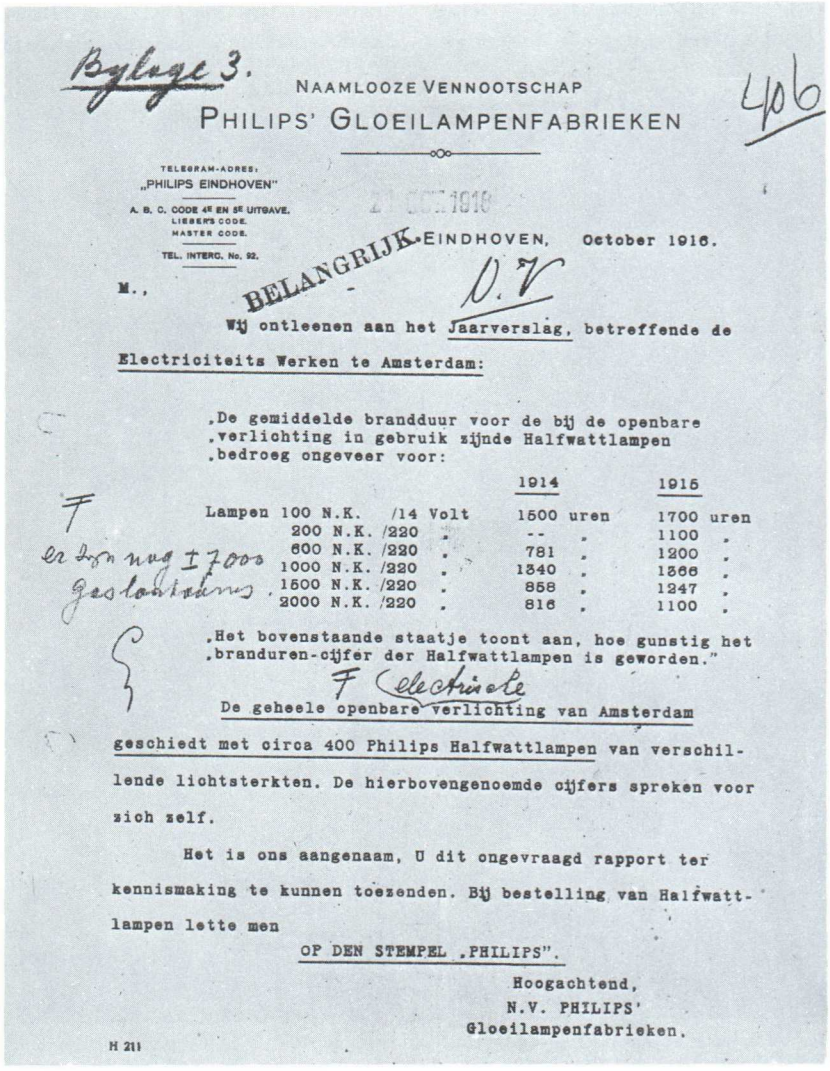


Fig. 9. Fitting containing a 100-candle gas-filled lamp, part of the lighting installation in an area of the Kerkstraat in Amsterdam in 1914. Each lamp required its own step-down transformer for 220/14 V. (Photograph by courtesy of the editor of "De Koppeling"; see also that journal, vol. **8**, 149, 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 difference between monatomic argon gas and diatomic nitrogen gas does not seem very considerable (respective molecular weights 40 and 28), but an undesirable effect of nitrogen is that it dissociates at high temperature, so that in fact nitrogen compares



Bylage 3.

NAAMLOOZE VENNOOTSCHAP
PHILIPS' GLOEILAMPENFABRIEKEN

406

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

EINDHOVEN, October 1916.

M.,

BELANGRIJK. N.V.

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

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

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

F
er zijn nog 17000 Gaslampen in dienst

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

F Electriciteit

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

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

OP DEN STEMPEL „PHILIPS".

Hoogachtend,
N.V. PHILIPS'
Gloeilampenfabrieken.

H 211

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

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

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

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

EINDHOVEN, 1 November 1914.

Belangrijke Prijsverlaging.

Philips' „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:		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 × 2 × 50 × 1 1 . . .	= 110 K.W.U.	1000 × 100 × 0,6 . . .	= 60 K.W.U.
ad Fl. 0,20	= Fl. 22,—	ad Fl. 0,20	= Fl. 12,—
Lampenverwisseling na gemiddeld 1000 uren: 2 × Fl 0,55 . .	= „ 1,10	Lampenverwisseling na gemiddeld 800 uren: ¾ × Fl. 2.70	= „ 3.37
Totale verlichtingskosten . .	Fl. 23,10	Totale verlichtingskosten	Fl. 15,37

EENE TOTALE BESPARING DUS VAN Fl. 7,73.

Spaart dus stroom en geld en vervangt Uw 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

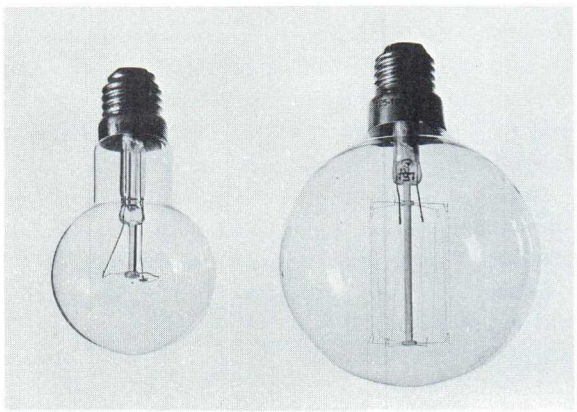


Fig. 12. The compact form of the filament and the favourable distribution of the gradually forming black tungsten deposits made it possible to produce gas-filled lamps (left) with a much smaller bulb than vacuum lamps of the same power.

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 li e is governed by the increasing blackening; on the other hand the useful life of the gas-filled lamp is limited by the occurence of thin or weak spots in the filament after a certain amount of tungsten has evaporated, causing the tungsten wire to break. Consequently, the choice of bulb diameter of a gas-filled lamp does not depend on blackening and useful life, but can be made as small as the temperature of the bulb wall permits (the bulb of course becoming hotter as its diameter decreases).

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

Further developments, notably of the filament

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

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

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

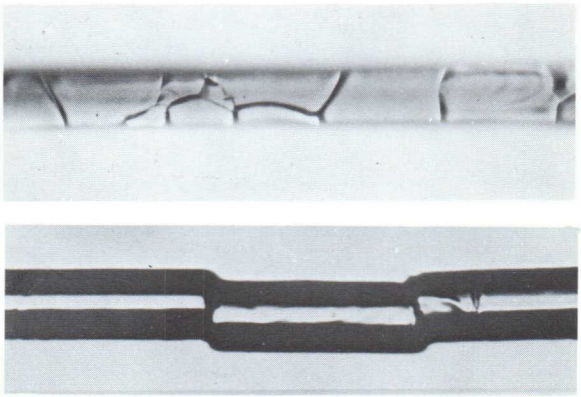


Fig. 13. *a*) Structure of a wire of pure tungsten after recrystallization.
b) "Offsetting" in a tungsten filament, occurring when a lamp having a filament with the structure shown in (*a*) had burnt for some time.

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

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

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

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

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

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

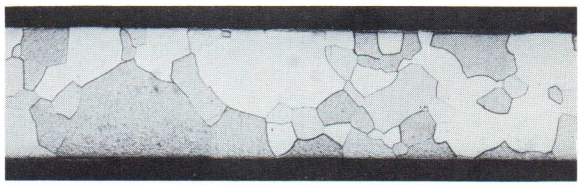


Fig. 14. Recrystallization texture of a tungsten wire containing impurities resulting from the Battersea process.

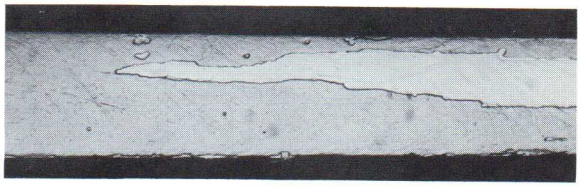


Fig. 15. Recrystallization texture of "doped tungsten", i.e. tungsten containing controlled amounts of additives. The elongated, wedge-shaped crystals give the wire considerable strength and prevent sagging when the wire is coiled.

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

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

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

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

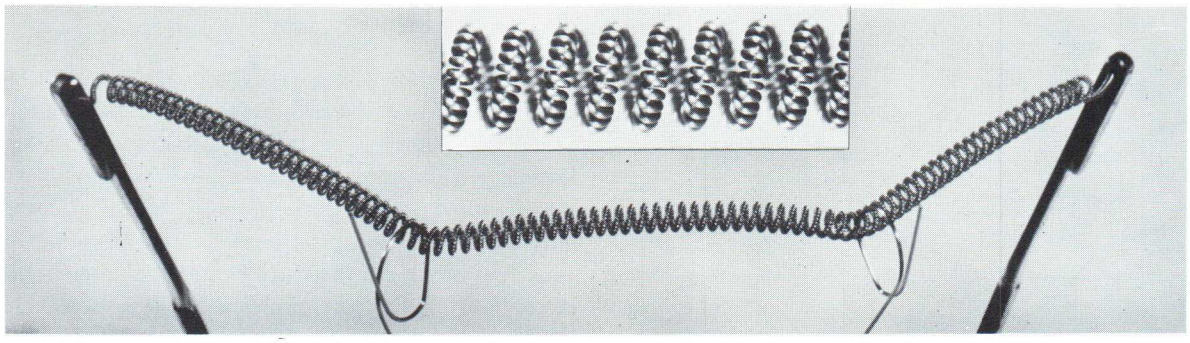


Fig. 16. Coiled-coil filament for a 100 W lamp. The inset shows a piece of the coiled-coil at higher magnification.

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

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

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

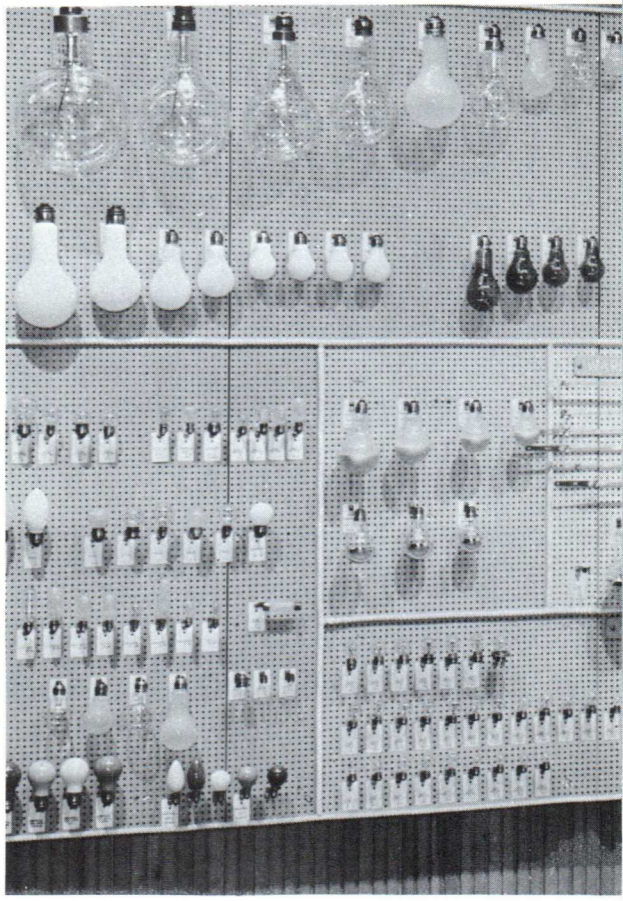
improvement of luminous efficiency particularly at lower wattages [13]). The coiled-coil lamp for low watt-

[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.



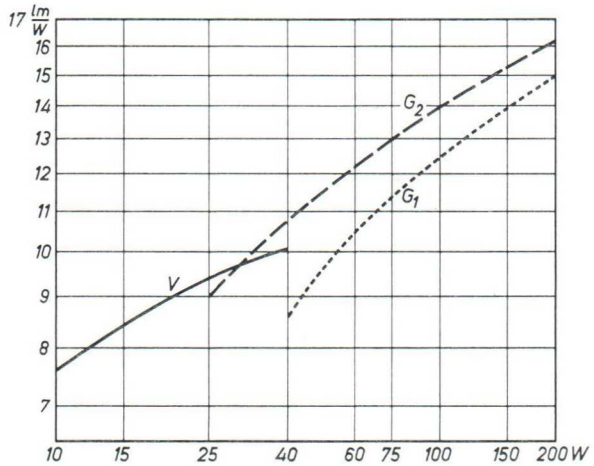


Fig. 17. Luminous efficiency of present-day vacuum lamps (V), of gas-filled single-coil lamps (G_1) and gas-filled coiled-coil lamps (G_2), as a function of wattage.

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.

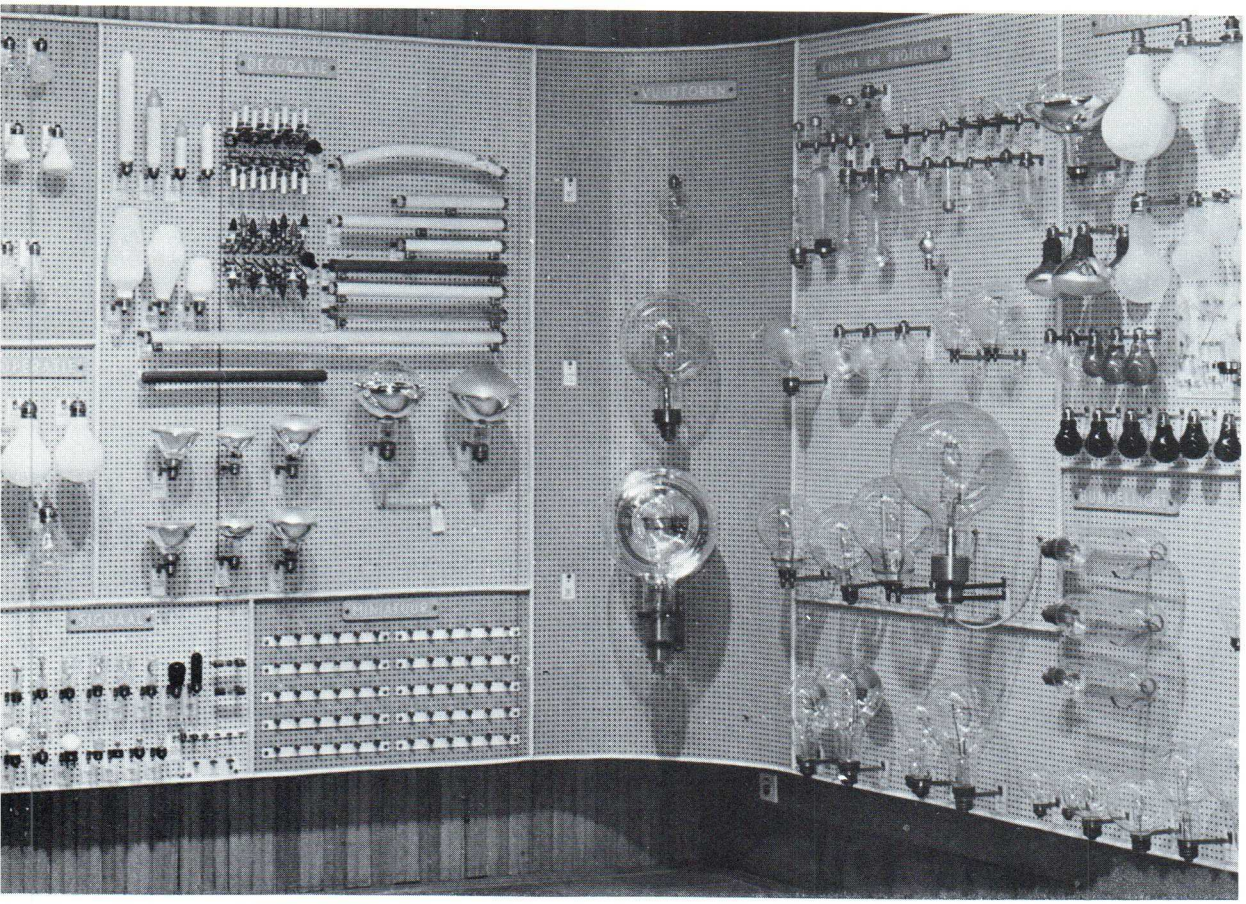
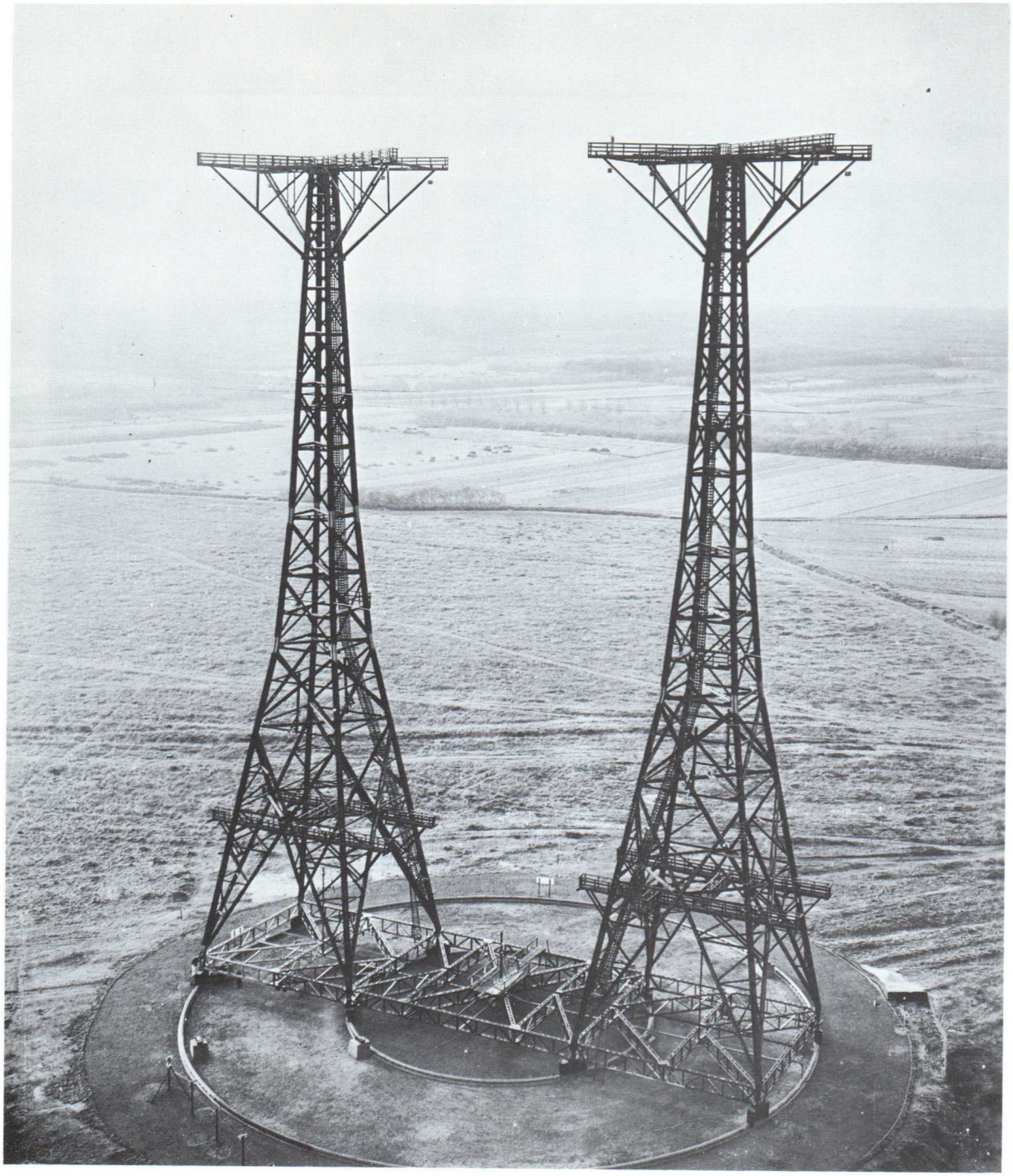


Fig. 18. A small selection from the tens of thousands of types of incandescent lamps nowadays produced by Philips. Most types come into the many categories of "special lamps" (projection lamps, car bulbs, window-display lamps, airfield, sports-field and lighthouse lamps, studio lamps, infrared-heating 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.



Picturesque early transmitting aerial

The PCJ transmitter on 31.3 m, built in 1926 in the Eindhoven laboratories and used for the first broadcast transmissions to the East-Indies (see this book, page 30) was transferred in 1937 to the PHOHI (Philips Omroep Holland-Indië) establishment at Huizen, where it was redesigned for 75 kW carrier-power and provided with a rotatable directional aerial, in order to transmit experimental programmes to particular regions of various continents.

The aerial consisted of eight vertical wires arranged in two parallel planes. Two *wooden* towers were used, so that the wires could be suspended from the towers themselves and the towers (200 feet high and weighing 18 tons each) could be placed at a relatively small spacing (exactly one wavelength). Even so the turntable frame, which occupied a rectangle 150 × 43 feet and was carried by 8 pairs of wheels running on circular railway tracks, had a weight of 40 tons. Wind forces, however, constituted the major part of the load: in the worst case a single pair of wheels would have to withstand about 13 tons in weight + 38 tons in wind forces.

The aerial was dismantled in 1958.

(See: P. J. H. A. Nordlohne, A rotating directional aerial for short wave broadcasting, Philips tech. Rev. **3**, 59-64, 1938.)

SCIENCE AND INDUSTRY

by H. B. G. CASIMIR.

001.658:37

During the Golden Anniversary of the State of Oklahoma, on June 17, 1957, an international symposium was organized at Oklahoma City under the general heading: science, industry and education. A number of the foremost leaders of research institutions from different countries were invited to give their views on problems related to this general theme. The following article is a slightly abridged version of the address delivered by Prof. H. B. G. Casimir, director of the Philips Laboratories and member of the board of directors of the Philips Industries.

The subject of the relations of science, industry and education is an extensive one about which a lot can be said and has been said, and it would be useless if I tried once more to survey the whole field. I shall therefore concentrate on three statements of which I trust that they are sufficiently controversial to form a contribution to discussions on this subject.

The first one relates to laboratory administration and runs as follows: in an industrial research laboratory it is impossible to relate research products budgetwise to industrial and commercial projects. Any attempt at operating a laboratory on a budget system will either be detrimental to research itself or it will lead to a mild form of cheating which makes the whole system illusory.

My second statement is that although industry is for its future dependent on fundamental research, very few fundamental results have come, will come or should come out of an industrial research laboratory.

My third statement: it is wrong to regard education and research as two distinct and separate tasks of a University.

Before proceeding any further I should make it clear what I understand by the words fundamental research, research, development and so on.

I shall describe what one might call the spectrum of research. At one end of the spectrum there is Fundamental Research with very heavy capital *F* and capital *R*. I should like to define this as research dealing with entirely new phenomena connected with basically and philosophically new principles in the description of nature. It is easy to give examples: the work of Faraday and Maxwell on the electromagnetic field, the work of Boltzmann and Gibbs on the kinetic explanation of thermodynamic phenomena, Einstein's theory of relativity, Bohr's quantum theory of the atom and the new quantum mechanics of Heisenberg, Schrödinger and Dirac belong clearly to this category. Next comes fundamental research with a more modest *f* and *r*. The aim is still the understanding of nature, but it is dealing with fields in which the basic principles are well understood. Work on molecular spectra, solid-state physics, electromagnetic diffraction, hydrodynamics would come in this group.

My next category is target-research. The method of working, the scientific approach is still the same, but now one has a definite technical target in mind: one is for instance no longer working on fluorescence in general, but is trying to find a substance which under electron bombardment gives a good output of red light with a more or less specified delay-time. In the engineering field a similar type of work is what I like to call project-research, but what is also frequently covered by the words advanced development. By this I understand work aimed at realizing a concrete technical object but with the qualification that it is still uncertain to what extent the thing we have in mind is possible. The first people who started to build electron microscopes knew almost nothing about the potential possibilities of this instrument; for those who came later in the field it was evident that an electron microscope could be

built, but it was impossible to tell beforehand what resolving power would be obtained with a reasonable amount of work. At present, unless we go to the very furthest limits of resolving power, it is almost possible to lay down specifications beforehand and to make a reasonable estimate of the cost of this development in time, manpower and facilities. This I consider characteristic of development, the possibility to estimate more or less accurately what is required to make a prototype of more or less specified characteristics. I have twice inserted more or less; this is not very elegant phrasing, but anyone who has had to do with the administration of research and development will agree that it is a necessary precaution.

Next in my list appear factory-engineering and trouble-shooting, activities that are related to development, but that often also give rise to problems of target-research. In our own organization the end of the spectrum is application-research, the work on specific applications of our own products. This would comprise such things as illumination-engineering, the art of adapting public address systems to existing buildings or the adaptation of a high-frequency furnace to a special industrial process. As I said before, this application-research has in our firm a somewhat separate position because of the nature of its relations with industrial and commercial activities, but from a more scientific and technical point of view it is a combination of the former types.

Perhaps the word spectrum is not too badly chosen for there is in reality a continuous transition and just as it is difficult in the solar spectrum to tell whether a certain colour is a bluish green or a greenish blue, one may differ in opinion about whether a certain type of work is target-research or development, and so on.

In our organization each industrial division is responsible for its own development, factory engineering and application work, whereas fundamental research, target-research and project-research or advanced development are carried out in our central research laboratory. I believe that this division has been found useful by all organizations that are dealing with a wide and diversified range of products. The basic principles underlying a simple radio receiver and a modern radar installation are not too different, but there is such a tremendous difference in style of technical realization that it is hard to imagine a central research and development organization able to cope competently with the development work for both the one and the other.

If you remember my definitions, you will notice that the work of our factory laboratories begins where results are reasonably certain and where costs can be predicted. The work in the research laboratories may be equally practical or applied in character, but there is always this distinction that there is a much higher degree of uncertainty and this leads directly to my first statement. In the case of factory development it is reasonable to ask whether the cost of developing a product can be earned back by sales of that product. Development costs of a specified product are part of the cost price of that product. It is impossible to make a similar breakdown of the expenses of a research laboratory: some of the work may not lead anywhere — much of it will later turn out to be useful in fields vastly different from those envisaged when the work was started. We originally studied ferromagnetic ferrites for a special application in carrier telephony: many years later we found that the bulk of our production goes into television receivers. From a long-range point of view our work on ferromagnetism has been succesful and profitable but if our nascent carrier telephony department had had to bear the initial expenses it would have been killed right at the start.

But, you will ask, why should the cheating come in? The reason is that the total budget of a research laboratory is determined fairly accurately by the number of scientific workers: if in a given field of research and a given country you want to have a laboratory where a number of research workers can work efficiently and are adequately provided with technical assistance, workshops and so on, then the cost per research worker will always come out the same. Further, a research institution should aim at continuity of staff, for it is an exception that a man, however competent he may be, does valuable work during his first half year or so in new surroundings, and some of the best research workers become productive only after several years. A slow and continuous flow of people is all right, but one should try to avoid discontinuous expansions or reductions.

The art of running a laboratory on a budget system consists therefore in splitting a sum total that is known beforehand into a number of separate items in such a way that those in charge of appropriations will fall for it. If I had to do this in my own institution I would start arguing in the following way: we have a fairly large group working on fluorescence. It is useful to both our lampworks and our cathode-ray tube factory. On the other hand some fairly abstract mathematical work on electron optics is being done, for which it is hard to find a suitable place on the budget.

Should the latter fact induce us to discontinue this work, which we think in the long run may prove quite useful? Better place it under the general heading of cathode-ray tubes, which sounds reasonable, and in order not too overload the budget for this division, put the whole fluorescence research on the budget for the lighting division, which is less crammed. And so on. It is this kind of artificial splitting and combining which I had in mind when I spoke of a mild form of cheating. Fortunately in our company we have done away with this system altogether. The research laboratories as a whole are each year allotted a lump sum calculated in such a way that it provides reasonable working conditions for the number of people employed and leaves room for slow expansion. Only in the case of large capital investments, new buildings, etc., do we apply for an additional budget. Of course we have to keep an eye on expenses, but no reference is ever made to a budget for a specific project. The notion of such a budget simply does not exist.

In my dictionary of English slang I find "Dutch reckoning" defined as a lump account without particulars. I am proud of being a Dutchman, so you can hardly be surprised that I consider this type of reckoning ideally suited to the needs of a research laboratory. So much about my first statement.

My second statement was that industry depends on the results of fundamental research — and here I mean to say Fundamental Research with capitals — but that such results do not come out of industrial laboratories. History gives ample evidence for this. Electrical engineering could not even be imagined without the pioneering work of Faraday and of Maxwell whose work in those days was entirely outside the scope of existing industries. There would be no radio without the work of Maxwell and Hertz, but it did not occur to the electrical industries of those days to do the work by which Hertz justly became famous. Speaking and working about electrons was almost anathema to the electrical engineer at the turn of the century, but today an evergrowing industry is based on the very notion of electrons. No company interested in power plants could dream that the work of the Curies, Rutherford and others would eventually lead to a new source of power. An industry may take a long range point of view, it may encourage fundamental research, it may not require results of immediate practical value, but it cannot be expected to work on basic principles and phenomena for which even our most vivid imagination cannot foresee whether, or in what field, they will find applications. Today at the very front of science, physicists are working on elementary particles, on concepts like parity, on the relations between mesons and nuclear forces. No one can foresee to what this will lead. No industry laboratory could be expected to do cosmic-ray work on a large scale or to be the first to build at its own cost a cosmotron, and yet if we deny the possibility of future applications we are probably just as shortsighted as the power engineer early in the century who derided electrons. The places for Fundamental Research are the University laboratories.

Here someone might interrupt me and object: What about the transistor? Now first of all let me say that the transistor is not only an extremely important and useful device that will have a far reaching influence on the whole field of electronics, but that the research that led to the transistor is a very beautiful piece of work indeed, combining well planned and yet imaginative experiments with a penetrating and elegant theoretical analysis. So it is not a question of grapes being sour: on the contrary they are very sweet; and though they may have been a little bit too high for us to pick, all the electronic industries got a generous helping after the picking. And yet I think the transistor is a good example for my thesis. So allow me to go into technical details for a moment and let me review the basic thoughts that play a role in the transistor.

First of all the idea of electrons, a notion connected especially with the theoretical work of my countryman Lorentz and the experiments of J. J. Thomson. Next the statistical theory of matter and its relation to thermodynamics (Maxwell, Boltzmann, Gibbs). The application of such methods to the motion of electrons led, in the hands of Lorentz and of Niels Bohr in his little known thesis, to important results. But there remained a number of very puzzling questions: How it is possible that the conduction electrons do not contribute appreciably to the specific heat; how can we imagine that an electron can move freely through a metal; and how is it possible that the sign of the so-called Hall effect is sometimes as if the current through a metal were carried by positive carriers, although we know that electrons are negatively charged? The solution of these riddles required entirely new notions.

The first was solved by the Fermi-Dirac statistics, a modification of classical statistical theory, brought about firstly by a recognition of the existence of quantum states and secondly by the Pauli exclusion principle. The application of Fermi-Dirac statistics to the motion of the electrons in a metal was worked out by Sommerfeld and turned out to be a great improvement. The solution of the second problem is due to Bloch and Peierls after some preliminary

and more qualitative work of Houston. It is based on new quantum mechanics where electrons are described by waves, a theory connected with the names of de Broglie and Schrödinger. It is found that these electron waves are free to move in a perfectly periodic field of force and the analysis of such waves by Brillouin and others led to the notion of electron bands. The third riddle was solved by Peierls; he was the first to introduce the notion of a hole in a filled electron band and to show that it behaves as it it were a positive electron. From then on the problems became less fundamental, but the essential ideas about semiconductors were still mainly the result of University work. Now the one new principle that was added by the Bell people to existing notions was the idea that electrons can persist for some time in p-type germanium, that is, germanium containing such impurities that the conduction is due to the motion of holes; similarly injected holes can persist for some time in n-type germanium. This is both an important experimental fact and an interesting theoretical idea, but from a philosophical point of view it is certainly not on a par with the great new ideas of quantum mechanics. The work on transistors is essentially a brilliant and novel application of known principles, that would be a beautiful piece of physical research anyway, but happens to be of great technical consequence too.

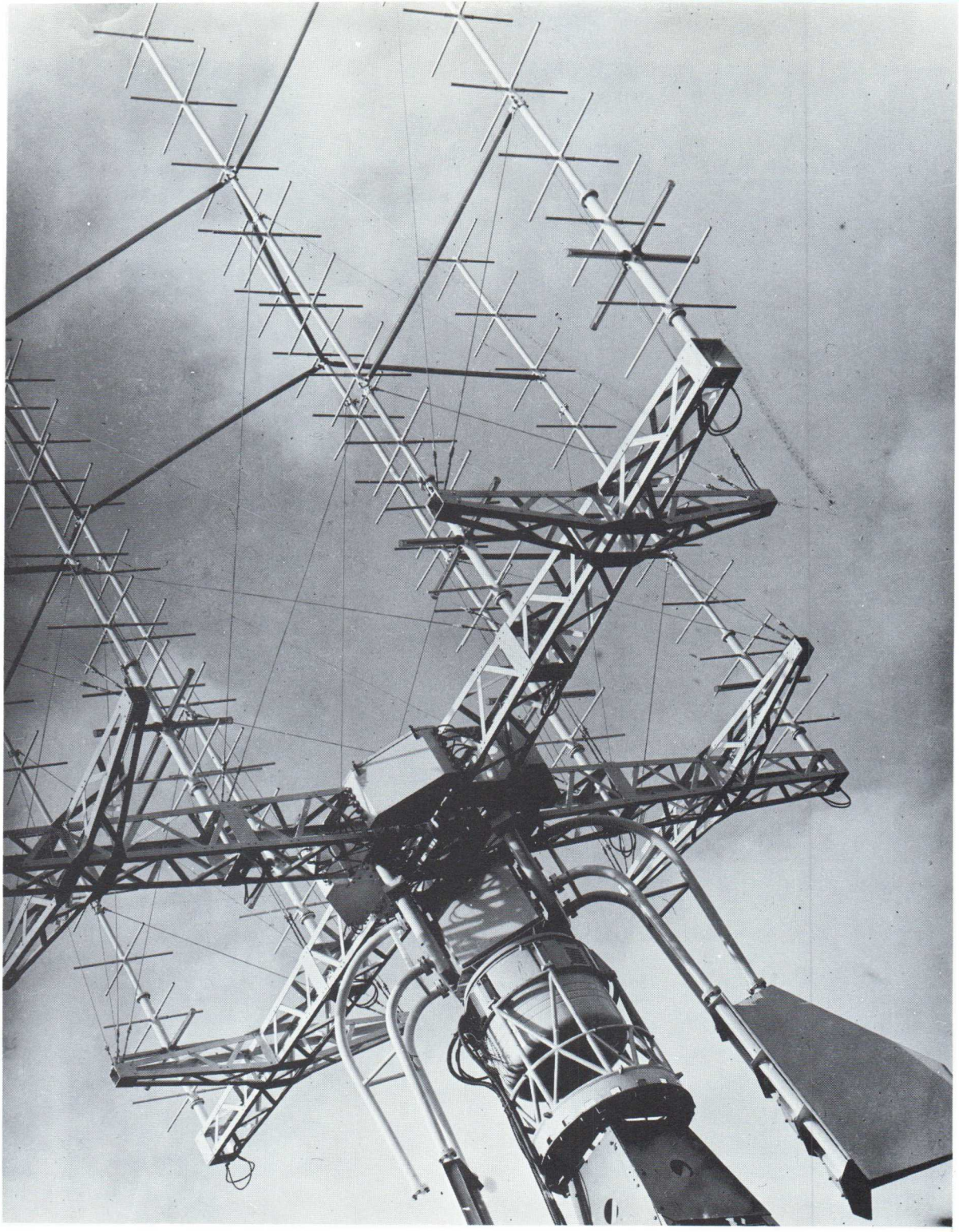
Once more, the fundamental notions that are required by industry are coming from the universities.

I would add the following. The rise of an industry depends on leadership that recognizes among other things the necessity of research. But when it comes to fundamental research we are again faced with the same situation on a different level. For the future of a single industry it is essential that it be broad-minded with respect to its own research; for the community as a whole and for the future of all its industries, it is essential to support also the most abstract type of fundamental work. Industrial laboratories are part of a far-seeing policy of industrialists. Wise statesmanship of the leaders of a nation will result in sponsorship of academic research.

Why then my third statement? Why is it wrong to say that Universities have not only an educational task but should *also* carry out research? Because this way of looking at things shows a basic misunderstanding of the nature of a University. It is the task of a University to live a scientific life and to initiate a younger generation in this way of thinking. This initiation requires instruction in certain techniques, but these are only means to a goal. There is only one form of real higher education and that does not consist in listening to lectures, in performing classroom experiments or in solving textbook problems: it consists in struggling on the borderline of knowledge, with the unknown and the unexpected, under the guidance of an inspired and inspiring teacher.

I have discussed, in the foregoing, one or two aspects of the relations between science and industry. I have emphasized the importance of industrial research for industry and of academic research for industrial research. But that is not the whole story. Although our technical civilization is the result of a joining of forces of philosophical inquisitiveness and industrial zeal, our desire to understand nature is not only justified by its practical importance. It corresponds to a noble urge of the human race that is also an aim in itself. Even the most die-hard materialist feels dimly that the mathematician dealing with an impressive edifice of abstractions and the astronomer probing the structure of the universe are doing valuable things and are in some way contributing to the richness of human experience.

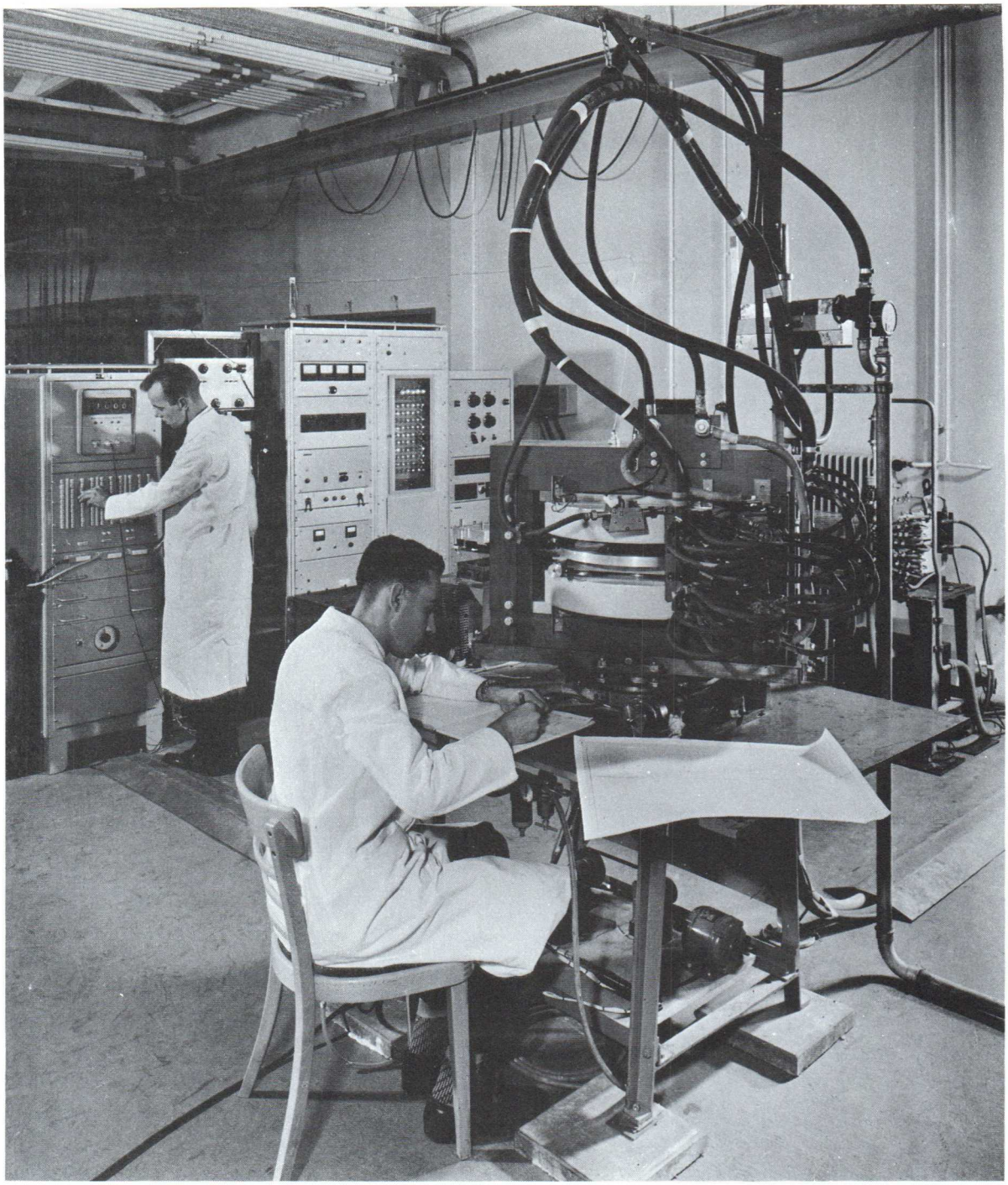
There enters some of this even in a simple piece of applied research. To feel that one's work is not only of practical use but has also inherent value and beauty, is one of the things that make life worth living. We research people should count ourselves fortunate beings.



Aerial for telemetering satellite-launcher

When the European Launcher Development Organization (ELDO) fires its fourth rocket from the Woomera range, the operation of the third rocket stage will be telemetered by means of equipment developed and built by Philips' Telecommunicatie Industrie at Huizen (Netherlands). The 60 feet high receiving aerial shown in the photograph forms part of this equipment. It contains 8 vertical and horizontal dipole arrays, from which the signals are fed separately to the polarization diversity receivers; this will reduce fading effects caused by the hot gases of the rocket exhaust. Phase differences between the voltages set up in the dipole arrays give rise to difference voltages which are a measure for the aiming error of the aerial in azimuth or elevation. These voltages are used for actuating a servo-system which will lock the aerial on to the rocket.

(From: Philips tech. Rev. **27**, No. 5, 1966.)



Photograph Adolf Morath

Development of an isochronous cyclotron

The shape of the "clover-leaf" poles required for ensuring adequate stabilization of particle orbits in an isochronous cyclotron cannot be derived analytically. The usual procedure is to build a model magnet and map very accurately the field obtained; the resultant particle orbits can be calculated by means of a computer and the changes required in the pole shape can then be found by trial and error. The photograph shows the field measurements being carried out on the test magnet, scaled down 5 times, of the first isochronous cyclotron developed at Philips Research Laboratories in Eindhoven (Geldrop section). The field is measured with a Hall probe allowing very accurate positioning, and is calibrated by means of a proton-resonance arrangement, while the results are recorded on punched tape for direct input to the computer, with suitable programs.

(See: N. F. Verster and H. L. Hagedoorn, Investigation of the magnetic field of an isochronous cyclotron, Philips tech. Rev. **24**, 106-120, 1962/63.)

SOME MAIN LINES OF 50 YEARS OF PHILIPS RESEARCH IN PHYSICS

by H. B. G. CASIMIR.

001.891.53

The field of gas discharges had already a honourable past when it was tackled by Holst and his associates as one of their first major topics [1])*). It had led to the discovery — or as we Dutchmen prefer to say, to convincing experimental proof — of the electron by J. J. Thomson and to the discovery of X-rays by W. C. Röntgen. *Fig. 1* shows one of the earliest

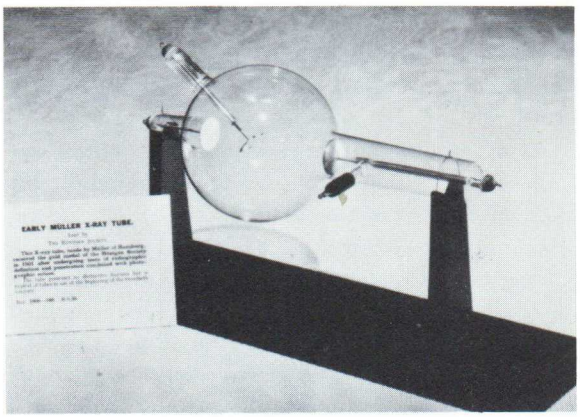
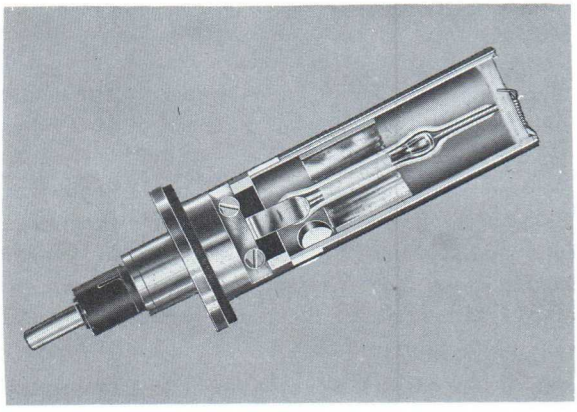


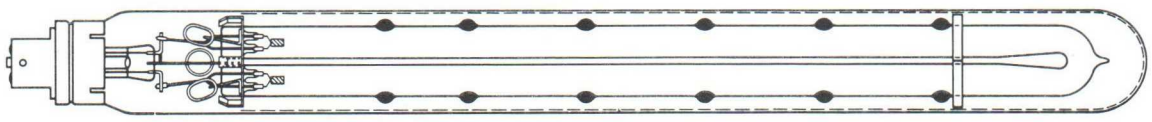
Fig. 1. One of the earliest commercial X-ray tubes, made by Messrs. C. H. F. Müller in 1901.

commercial X-ray tubes, made by Messrs. C. H. F. Müller, a firm that has since become a valuable member of the Philips group. Bohr's theory provided better understanding and hence also better possibilities for technical applications. I have no time to dwell on scientific results of this early work in the Philips Laboratories but should like to mention sodium lamps [2]) and super-high-pressure mercury lamps [3]) as spectacular technical results (*fig. 2a* and *b*). We have recently been able to improve the energy balance of a sodium lamp by a coating of suitably processed tin oxide which is transparant to visible light but was found to have a high reflectivity in the infrared [4]). In the late thirties interest in the basic aspects of gas discharges was somewhat on the decline but the rapidly growing field of experimental nuclear physics made good use of ion sources and Geiger-Müller tubes. Even today G.M. tubes have by no means been ousted by photomultipliers-cum-scintillators and solid state devices. As a matter of fact satisfactory stability and reproducibility were only obtained fairly recently [5]). I show part of the range that is at present manufactured by our factories (*fig. 3*). It was mainly developed by a small branch of our laboratories in the cyclotron institute at Amsterdam. Good fun, building cyclotrons; our latest design of isochronous cyclotron, of which we are rather proud, is one of the items of our Symposium programme [6]). I also would like to mention scalers, mainly because of the display tube used in many of them, a deflection tube with ten stable positions [7]) which gives me the opportunity to say a few words to commemorate Dr. J. L. H. Jonker, in whose group this tube was developed, whom many of you have known and to whom the Laboratories are indebted for creating a fine team and a fine tradition. He left us to join the new Technical University here at Eindhoven and was the inspiring leader of its department of electrical engineering

*) Literature references are listed at the end of this article.



b



a

Fig. 2. *a*) A modern sodium lamp of high efficiency, obtained i.a. by an infrared reflecting coating of tin oxide (indicated by broken line).
b) A 1000-W water-cooled super-high-pressure mercury lamp (SP 1000 W). This lamp is fitted in a metal reflector enlarging its effective width.

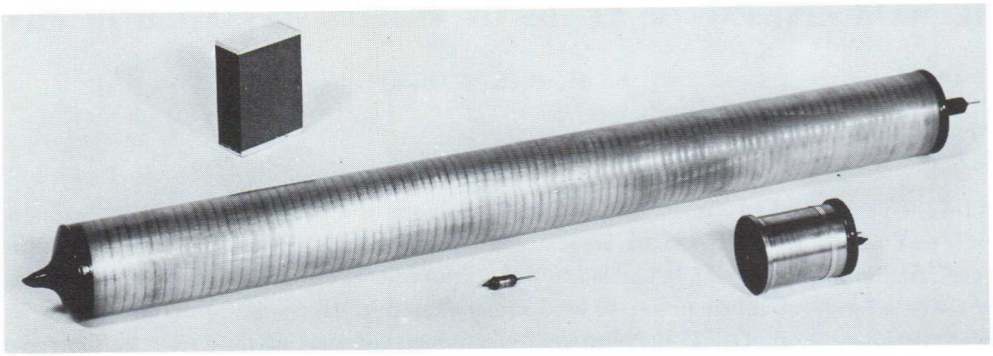


Fig. 3. Three Geiger-Müller counters of different sizes.

until his untimely death this summer. The E1T tube [7]) is the only surviving member of a family of tubes designed for switching and logical circuits which were made obsolete by the transistor even before they were thoroughly developed.

Back to gas discharges. Radar brought T.R. switches and hydrogen thyratrons and today there is a terrific activity in connection with gaseous lasers, possibilities of controlled fusion, direct conversion of heat into electrical energy, microwave devices and so on, while the older topics are by no means dead. Also, astronomers who in their thinking are so often ahead of us earthbound mortals have started an intensive study of magnetohydrodynamics. Today it is no longer difficult to convince a young man that gas discharges is an exciting subject. Of course you have to call it plasma physics, but as a father of four daughters I have to make worse concessions to fashion. We do not work at present on very high current discharges but we shall show work on several phenomena and devices, including sealed-off neutron generators [8]), beam plasma interactions [9]) and optical lasers [10]).

One interesting feature of laser research is that it offers much scope for ingenious optical arrangements; some examples will be shown. Laser modes always remind me of cascade generators. Why? Because they are pseudo-stationary modes in the sense that the wave motion is not only dying out because of energy dissipation but also because waves are leaking away into free space. Such situations are always rather tricky to treat. The most famous example is Gamow's theory of alpha radio-activity (*fig. 4*). An alpha particle is inside a potential well but can leak out through the barrier. Even in this simple case one had some difficulty to arrive at a satisfactory mathematical treatment. You will remember that the notion of penetration through a barrier encouraged Cockcroft and Walton to proceed with their experiments. Soon afterwards we began to produce high voltage generators industrially. *Fig. 5* shows the first generator we installed at Cambridge and which I think had a satisfactory record [11]).

One of our outstanding men in gas discharges was the late Dr. Penning, and one of his contributions made in 1937 was the Penning — or Philips — ionization gauge known to the initiated as „PIG" [12]). I show you an early model (*fig. 6*). Later the geometry was modified, but the essential idea remained the same, viz, the use of crossed electric and magnetic fields such that electrons cannot reach the anode from the cathode unless they suffer collisions. Penning emphasized that the device may be used as a manometer — as such it soon found widespread application in our own factories —, as an ion source — he made the first sealed-off neutron tube [13]) —, and as a pump. I remember clearly the talk he gave in 1936 to the Dutch Physical Society in Amsterdam. For his demonstrations only a rather poor pump was available, but he closed a valve and the manometer did the pumping. I was then a young theoretician reared on quantum mechanics and so the notion that a measurement influences the situation was familiar, but that it should *improve* the situation came rather as a surprise. Today „PIG's" are widely used. One of our competitors has kindly advertized its advant-

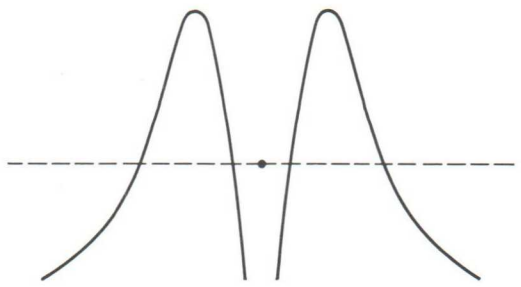


Fig. 4. Symbolic representation of Gamow tunnelling through a potential barrier.

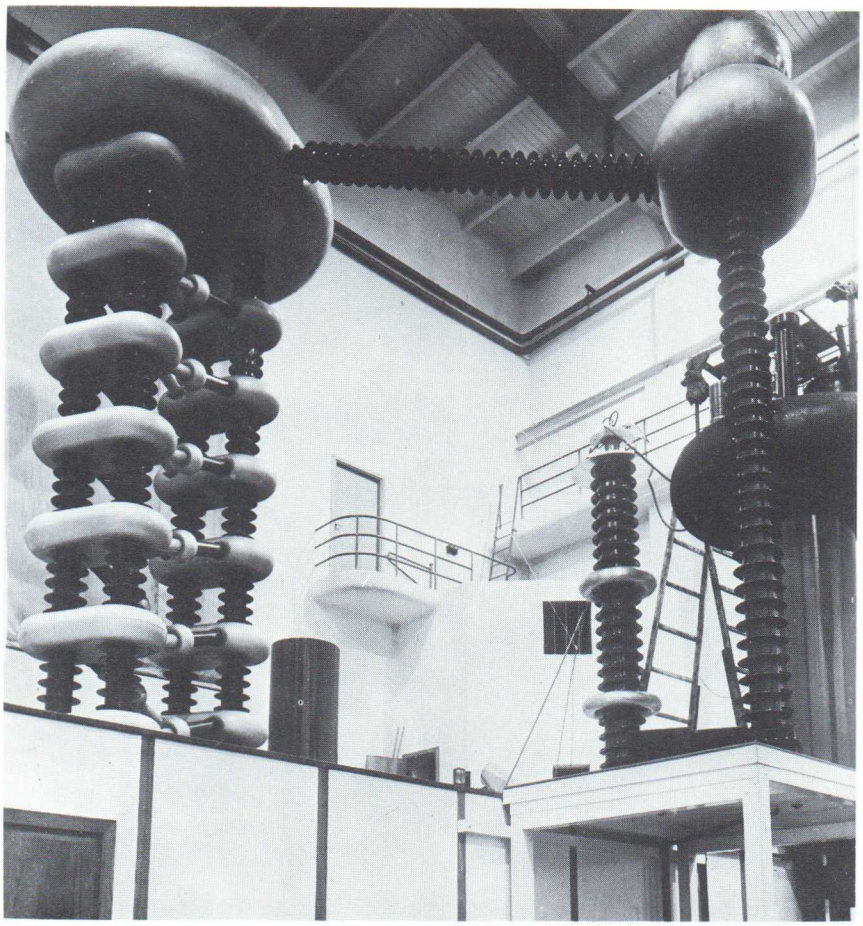


Fig. 5. First Philips cascade generator (1 MeV) installed at Cambridge in 1937.

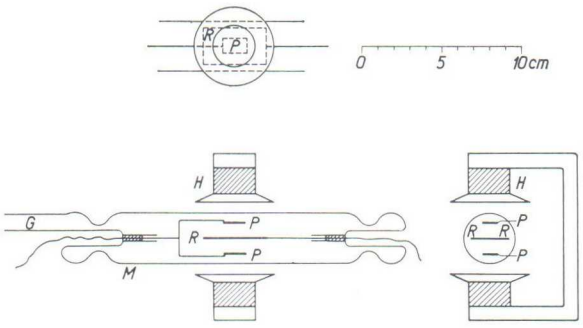


Fig. 6. Geometry of an early model of the Penning ionization gauge; taken from Penning's first publication on this subject.

ages. In our more mercenary moods we might even regret that the patents have run out.

In the thirties our laboratories began to study ferromagnetism, and this has remained one of our major subjects of research ever since. I shall not be able to do justice to the work of Snoek [14]) and many others in a quarter of an hour or so but perhaps this does not really matter because many of the results have become staple knowledge. The following diagram (*Table I*) may help to survey the situation. We can distinguish two classes of materials: metals and alloys on the one hand, semi-insulating oxides on the

Table I. Magnetic materials.

Main properties and use	Materials	
	metals	oxides
permanent	"Ticonal"	hexagonal ferrites (ferroxdure)
semi-permanent memories magnetic recording	thin films thin wire	square-loop ferrites iron-oxide powder (tape)
soft linear transformers, inductances, etc.	silicon iron permalloy etc.	cubic ferrites
soft non-linear magnetic amplifiers voltage regulators	silicon iron permalloy etc.	cubic ferrites
microwave applications		cubic ferrites hexagonal ferrites (ferroxdure, ferroxplana) garnets
magnetostriction	nickel iron	special ferrites

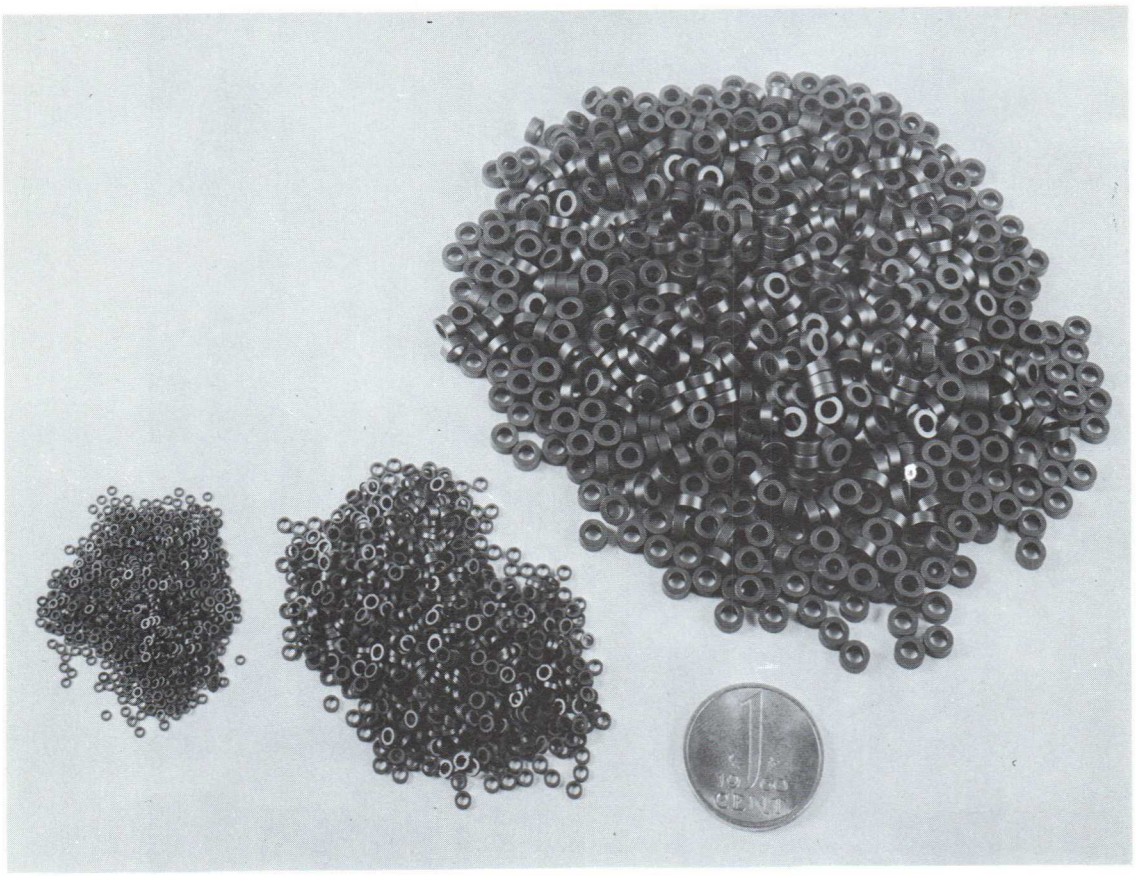


Fig. 7. Rings of square-loop ferrite for magnetic stores. Each pile contains 1000 rings.

other hand. This distinction usually coincides with a more fundamental distinction between ferromagnetic and ferrimagnetic materials. I shall come back to this question a little later. From the point of view of the electrical engineer we can distinguish several fields of application: permanent magnets are used for producing magnetic fields, magnetic materials used for storing information in terms of magnetization I have called semi-permanent, soft linear materials find application in transformers, filter coils and so on, whereas soft non-linear materials (they may be the same materials used at higher fields) are applied in magnetic amplifiers, voltage regulators, modulators and in oscillators with varying frequency. In the microwave field isolators or unilines are of special importance, and we should also remember

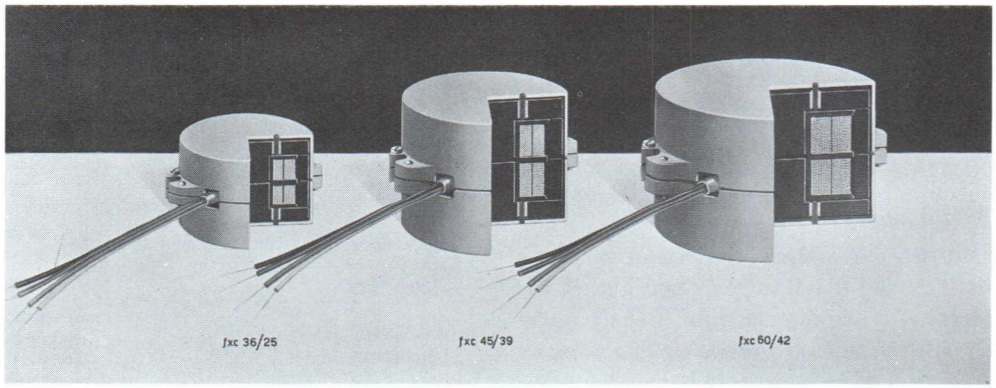


Fig. 8. Three Pupin coils, part of which has been cut away in order to show the windings and the ferroxcube core (black).



Fig. 9. Ferrite rings for the 12 accelerator stations of the 33-BeV proton synchrotron at Brookhaven. (Photograph by courtesy of Brookhaven National Laboratory.)

the Suhl type of parametric amplifier. The mentioning of the following examples may illustrate our activities in these various fields. "Ticonal" 5 [15]): this material was announced in 1937 and its figure of merit $BH_{max} = 5 \times 10^6$ was obtained by cooling in a magnetic field. "Ticonal" 11 [16]): this is a similar material but it is now a single crystal and was cooled in a magnetic field parallel to one of the cubic axes. Ferroxdure [17]): a substance that is chemically a hexaferrite, $BaO.6Fe_2O_3$, has a hexagonal crystal structure and a very high crystal anisotropy with a preferred direction of magnetization along the hexagonal axis. Its figure of merit is only 3.5×10^6 but a figure of merit does not always tell the whole story. These materials and the physico-chemical basis of their properties will be discussed in more detail by Dr. Vink. Each of these three little piles (*fig. 7*) contains 1000 tiny rings of a square-loop ferrite (to be precise it is a Mn-Mg-Zn-ferrite for the largest and a Cu-Mn-ferrite for the two smaller sizes) and therefore the material for storing 1000 bits of information. Ferrite cores are used in many shapes and sizes and one of the earliest examples of their application is found in Pupin coils (*fig. 8*) which have largely replaced the older coils using highly anisotropic nickel iron ribbon that was magnetically soft in one direction. A rather spectacular application of ferrites is made in large accelerators like the Brookhaven and CERN machines, and I show you one of the rings made for that purpose (cf. *fig. 9*). Finally my box of exhibits contains a couple of isolators for millimetric waves (*fig. 10*) and an ultrasonic oscillator of a type that is finding increasing application for ultrasonic cleaning (*fig. 11*).

Let me now turn to some of the more basic phys-

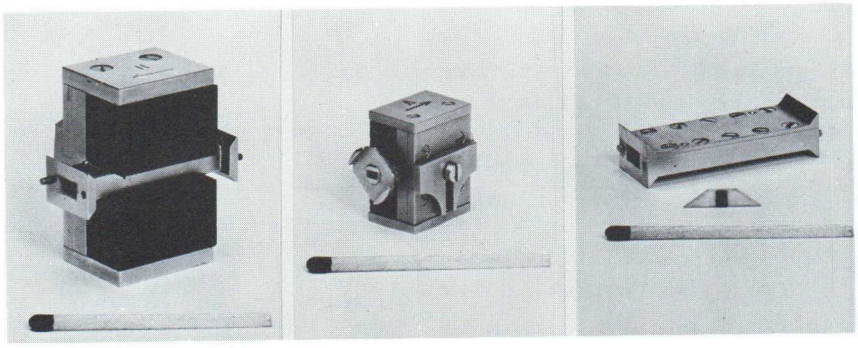


Fig. 10. Isolators for 8-mm waves (left and right) and 4-mm waves (centre).

ical aspects. I mentioned the difference between ferromagnetism and ferrimagnetism. Let me remind you of the basic ideas of ferrimagnetism, ideas that are due to Neel of Grenoble.

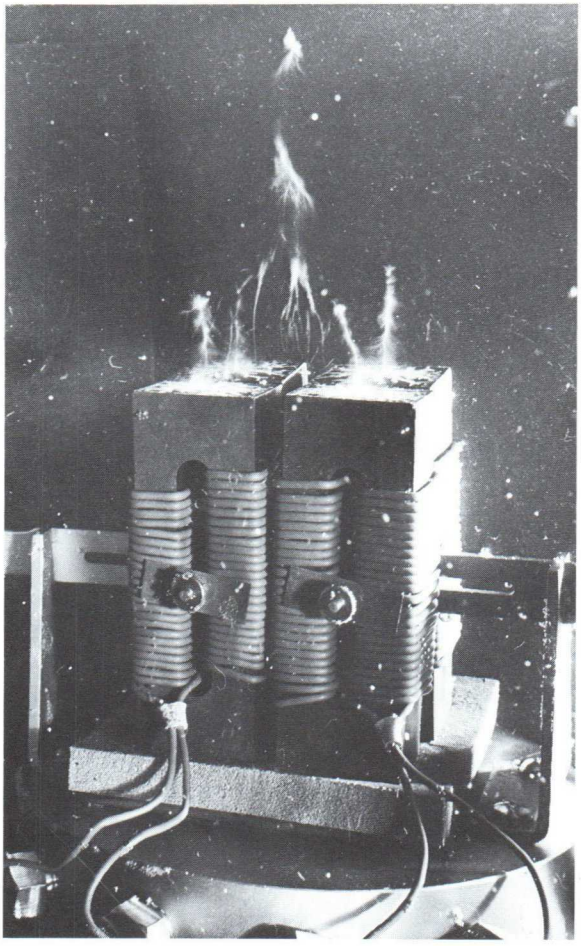


Fig. 11. Ferrite transducer for ultrasonic cleaning, photographed while operating in water.

In ferromagnetism the exchange force between neighbouring spins is such that it tends to orient these spins parallel to one another, in ferrimagnetism the interaction has opposite sign. But if there are various types of sites for the magnetic ions this interaction may yet result in a non-vanishing magnetic moment. In the simplest case there are two types of sites; let us call them A and B sites. Now it may well be that the interaction between A and B spins is larger than either the interaction between B and B and between A and A. Then the state of lowest energy is one in which all A spins are antiparallel to the B spins. For the total magnetic moment we can then write $M = M_A - M_B$. While the relation of these hypothetical A and B sites to actual crystal sites and the distribution of the various magnetic ions over these sites will be discussed in Dr. Vinks paper, a rather spectacular physical consequence of this theory is illustrated in the following figures (*fig. 12a* and *b*). Gorter and Schulkes [18]) prepared a lithium-chromiumferrite — the exact composition is $Li_{0.5}Fe_{1.25}Cr_{1.25}O_4$ — for which the temperature variation of M_A and M_B is such that at a certain temperature the resulting moment is zero. Van Wieringen [19]) has measured the g-factor by means of ferromagnetic resonance. Now the magnetic moments are not exactly spin-only values but contain a certain orbital contribution that is different for different ions. Therefore magnetic moment and mechanical moment do not vanish at exactly the same temperature and hence the curious behaviour of the g-factor.

Far more detailed data on the partial moments can be obtained by means of the Mössbauer effect. This method was recently applied by Van Loef [20])

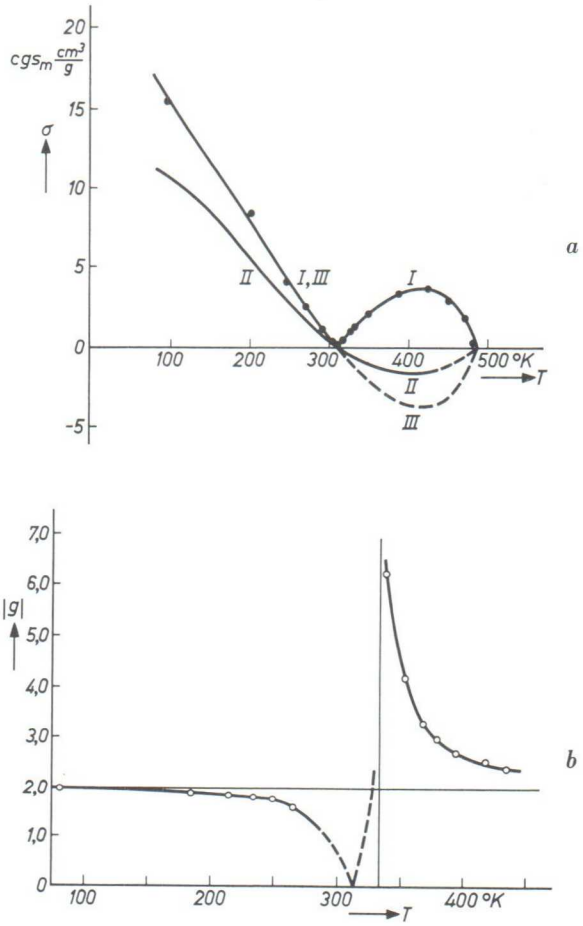
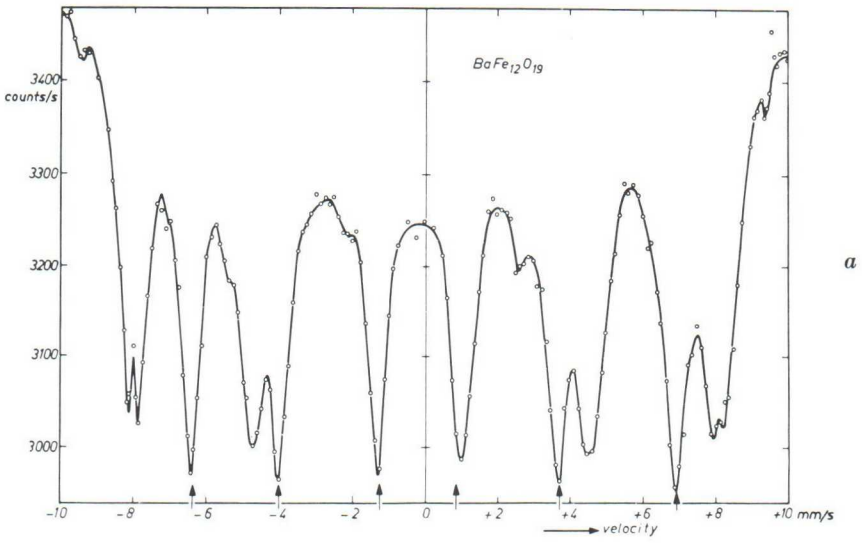


Fig. 12. *a*) Temperature variation of the saturation magnetization of $Li_{0.5}Fe_{1.25}Cr_{1.25}O_4$. Note that at a certain temperature the magnetization changes sign.
b) Curious behaviour of the g-factor as a function of temperature, in the same substance.



to ferroxdure (*fig. 13a*). This illustration shows his results, obtained with a cobalt source and a sample of $BaO.6Fe_2O_3$ containing a high percentage of Fe-57. We see not only the normal hyperfine structure but we can distinguish for one hyperfine component four different lines corresponding to four different crystal sites. From crystallographic analysis we know that there are five kinds of sites for the iron ions but one line may well be weak. From measurements at different temperatures Van Loef has calculated the temperature dependence of the partial moments; their resultant moment is in good agreement with experiments (fig. 13*b*, *c*). We are convinced that continuation of this work will yield further important information on the partial moments in ferrimagnetic substances.

Before leaving the field of ferromagnetism I should like briefly to recall the ferroxdure-ferroxplana-ferroxcone story. Let us consider a hexagonal crystal and assume that the azimuthal anisotropy of magnetization is slight so that we can write for the contribution of crystalline anisotropy to free energy:

$$F = K_1 \sin^2 \Theta + K_2 \sin^4 \Theta . \quad . \; . \; . \quad (1)$$

Then the direction of preferred orientation depends on the values of K_1 and K_2. The result is given in the following figure (*fig. 14*). You will notice that there are regions where $\Theta = 0$. That is the ferroxdure case. There are also regions with $\Theta = \pi/2$. This is the case of ferroxplana, of which we still hope that it will be developed into a high-frequency material of outstanding properties. But as was pointed out by Smit [21]) there is also a region where the preferred orientations lie on a cone. Instances of ferroxcone behaviour have really been found. This is very gratifying but so far we have not been able to think

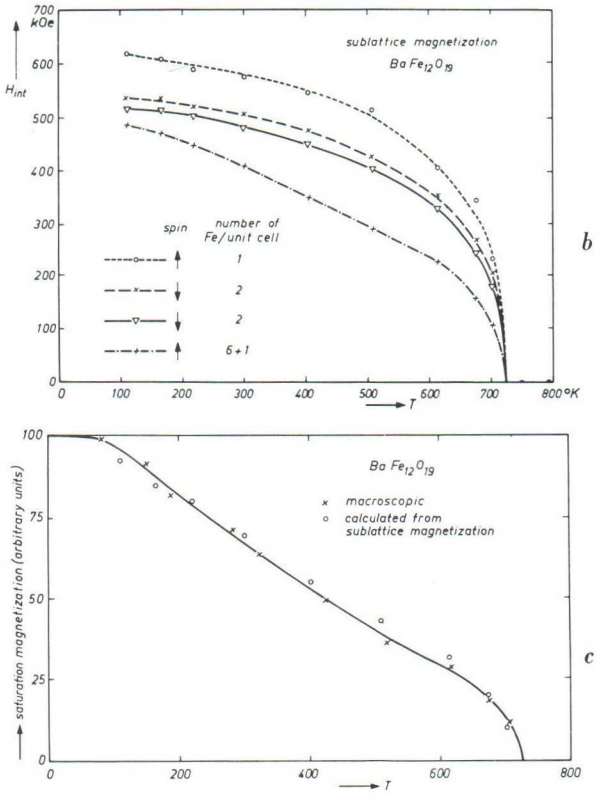


Fig. 13. *a*) Mössbauer absorption spectrum of Fe-57 in $BaO.6Fe_2O_3$.
b) Partial magnetic moments of Fe on four sublattices in $BaO.6Fe_2O_3$ as obtained from the Mössbauer spectrum.
c) Comparison between sum of partial moments and measured total saturation magnetization.

of any useful applications. Further material on preferred orientations will be presented by Dr. Enz in the course of this Symposium.

I mentioned optical one-way systems and microwave gyrators at the beginning of this lecture and

showed in passing one or two devices. I have not yet mentioned that the gyrator concept formulated by Tellegen [22]) was an important intermediary between Rayleigh's idea and Hogan's realization. Tellegen

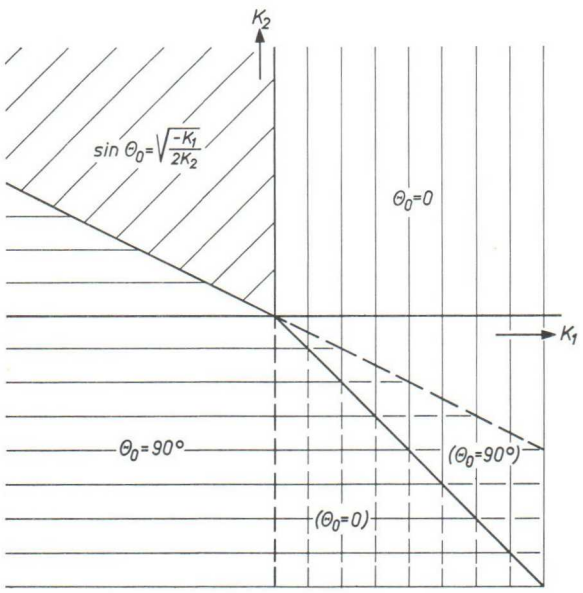


Fig. 14. In different areas of the K_1-K_2-plane different types of preferred magnetic orientation are found: preferred direction ($\Theta_0 = 0$); preferred plane ($\Theta_0 = 90°$); preferred cone ($\sin\Theta_0 = \sqrt{-K_1/K_2}$).

studied from a formal point of view four-pole networks described by:

$$\left.\begin{aligned} V_1 &= A_{11}i_1 + A_{12}i_2\,, \\ V_2 &= A_{21}i_1 + A_{22}i_2\,, \end{aligned}\right\} \quad . \; . \; . \; . \; . \quad (2)$$

for which the symmetry relation is violated, and especially idealized networks for which $A_{12} = -A_{21}$. He also pointed out that for systems in an external magnetic field possibilities for realizing such gyrators might well exist. I was able to show that the symmetry relations for passive four-pole networks are a special case of Onsager's relations for irreversible processes, which in turn are based on the symmetry between past and future of the equations of motion of the elementary particles and of the equations for the electromagnetic field [23]). Therefore the following general formulation holds: if a passive four-pole network incorporates magnetic moments M_1, M_2, ... and is subjected to external fields H_1, H_2, ... then

$$A_{12}(H_1, H_2, \ldots, M_1, M_2) = \\ = A_{21}(-H_1, -H_2, \ldots, -M_1, -M_2, \ldots) \quad . \; . \quad (3)$$

Of course, ferromagnetism makes one also think of radio astronomy. It is well known that the magnetic behaviour of matter must be ascribed to the existence of atomic currents. The meaning — and the only meaning — of the magnetization M is that the current density is given by:

$$i = \text{curl}\, M\,. \quad . \; . \; . \; . \; . \; . \quad (4)$$

Perhaps, when the notion of a spinning electron was introduced one may temporarily have been inclined to think in terms of dipoles, but Dirac's theory showed that this is wrong. In the fundamental state of a hydrogen atom for instance there is a current circulation around the nucleus. This current can to a high degree of accuracy be described by the analogon of the usual macroscopic equation

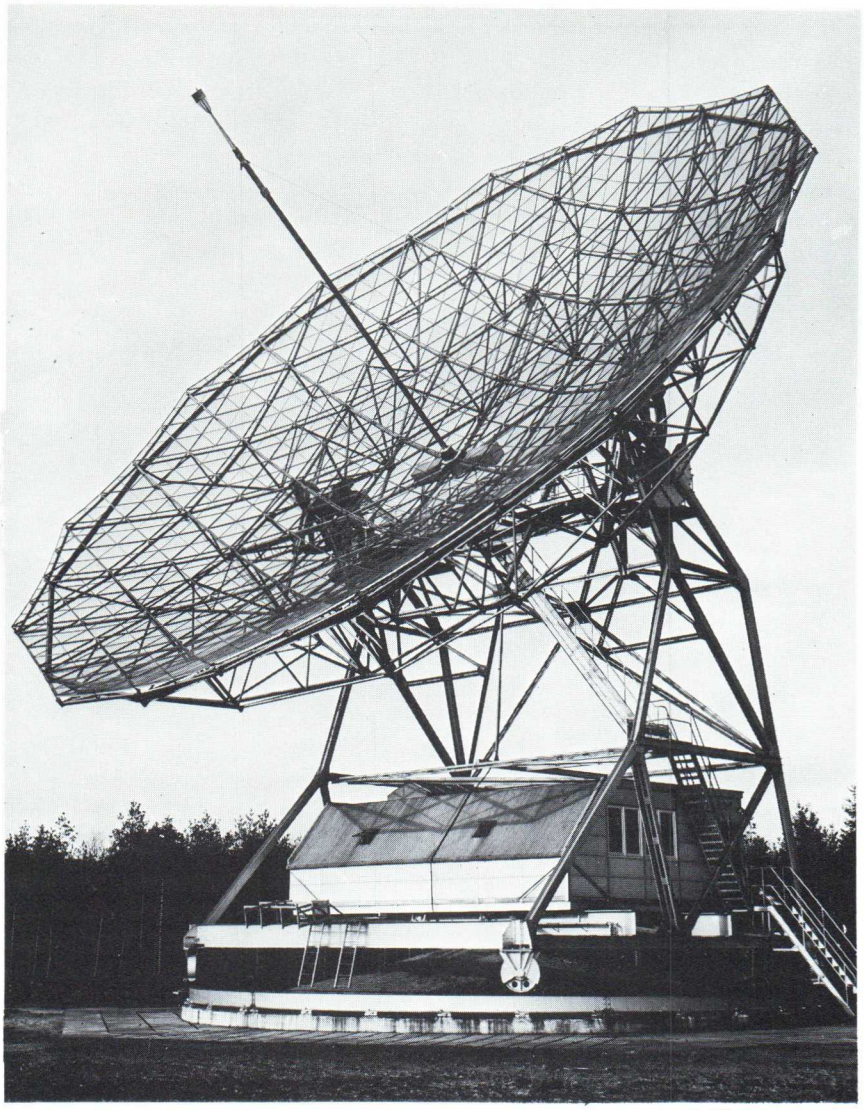
$$i = \text{curl}\, \psi^* \,\mathbf{s}\, \psi\,, \quad . \; . \; . \; . \; . \; . \quad (5)$$

where **s** is the operator for the spin magnetic moment. This current distribution produces at the nucleus a magnetic field that can be calculated by simple electrodynamics formulae. The magnetic moment of the nucleus orients itself in this field which leads to a hyperfine splitting corresponding to a wavelength of 21 cm. It was the Dutch astronomer Van der Hulst who predicted that hydrogen atoms in empty space should emit this line. It was observed for the first time by Purcell but the Dutch astronomers, who were somewhat delayed because of a fire that destroyed part of their apparatus, were among the first to obtain valuable information about the structure of the galaxy from such observations. *Fig. 15* shows the Dutch radiotelescope at Dwingeloo and I am glad to report that our laboratories have been able to assist in the design of several important pieces of equipment [24]).

If we regard ferromagnetism as one way of obtaining permanent circulating currents then a comparison with superconductivity suggests itself. There are very fundamental differences. Since the current in magnets is a spin current of the type curl $\psi^*(\sum \mathbf{s})\psi$ it can never circulate around a hole. That is why permanent magnets are not particularly suitable for producing homogeneous fields along the axis of a long cylinder. A current in a superconductor on the other hand is of the type $\psi^*\mathbf{A}\psi$, where **A** is the vector potential, and can circulate around holes. The difference appears very clearly in the following simple experiment that was carried out by Van Geuns. A thick-walled cylinder of NbSn is cooled below its transition point while the bore is filled by a closely fitting rod of "Ticonal". Below the transition temperature the "Ticonal" is pulled out and a persisting current is started. The following figure gives a schematic drawing of the currents in both situations (*fig. 16*). During our Symposium we shall demonstrate several experiments on superconductivity,

especially the very elegant unipolar generator conceived by Volger [25]). The Stirling-cycle engines for producing low temperatures will be demonstra - ed presently.

Sometimes it seems to me that our successes with ferromagnetic ceramics have in the past temporarily somewhat detracted our attention from the study of single crystals — although Van Arkel [26]) invented the beautiful method for growing single crystals of refractory metals by vapour phase transport reactions. If there is any truth in this remark we are in any case trying hard to mend our ways. The work of Penning, Polder and Okkerse [27]) on anomalous transmission of X-rays bears witness of our preoccupation with perfect, dislocation-free crystals. I should also like to mention the work of Hornstra [28]) who carefully analysed the various possible types of dislocations in complicated crystal structures.



Photograph H. Kleibrink, Leyden

Fig. 15. Antenna of the Dutch radiotelescope at Dwingeloo, at its completion in 1956 the largest movable reflector in the world. Several important pieces of equipment were developed in our Eindhoven laboratory.

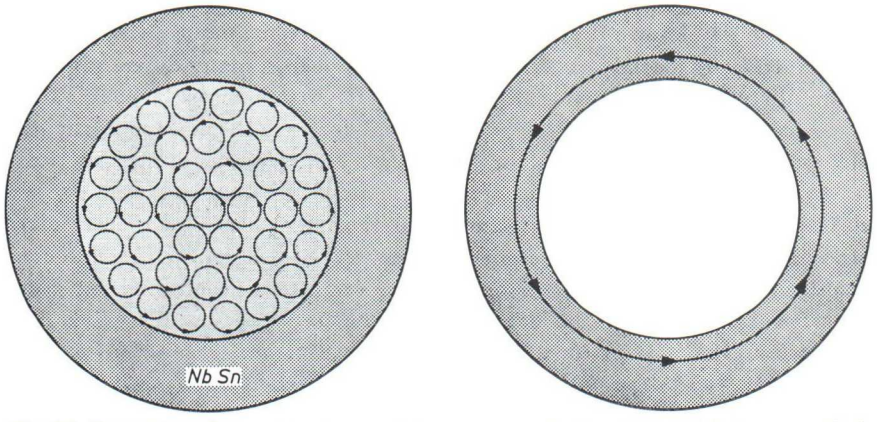


Fig. 16. Generation of a persistent current in a superconducting ring of NbSn by removal of a tightly fitting rod of "Ticonal".

Of course ferromagnetism is just one branch of solid state physics but I will not have time to touch upon our work on semiconducting materials and devices, on luminescence and photoconductivity, on metals and on surface phenomena. I trust these subjects will adequately be dealt with by others.

However, I should like to call your attention to

the importance of solid state work for the design of electron tubes. Our X-ray intensifier [29]) has become a useful tool in modern radiography (*fig. 17*). Its electron-optical system is simple and efficient but its main distinguishing feature is its combination of a fluorescent screen and a photo-electric layer which required a rather special technique. Our latest television camera-tube, the "Plumbicon", which will be extensively demonstrated, is as far as its principles of operation are concerned similar to the well-known vidicon. Yet it is an entirely different tube; its special performance is the result of a careful study of lead-oxide layers. And, although our centimetric and millimetric tubes require fine workmanship and careful design, they could never have been made without our dispenser cathodes [30]). We have done quite a bit of work on oxide cathodes, both empirical and theoretical. You may for instance remember the work of Vink and Loosjes [31]), who explained the conduction in porous layers of barium oxide as a conduction in the electron gas that fills the pores. But I should like to emphasize that the first idea of the L-cathode was only indirectly the result of fundamental research; in fact it was invented by the head of our tube shop, Mr. Lemmens. He was killed in a motoring accident a few months ago and I should like to pay a tribute to his memory. In our work we do not only depend on the achievements of the great scientific pioneers but also on the skill and intuition and perseverance of outstanding craftsmen.



Fig. 17. 9″ X-ray image intensifiers during fabrication.

The interactions we are dealing with in solid state physics are always electromagnetic in origin but electric forces may appear in many different guises. As simple Coulomb forces in heteropolar compounds, as screened Coulomb forces in the more sophisticated versions of the theory of electrons in metals, as exchange forces in homopolar compounds and in ferromagnetism. And they give also rise to attractive forces between neutral atoms and molecules even when these are so far apart that their wave functions do not overlap. Such forces are usually called Van der Waals forces because the forces introduced by Van der Waals to explain the behaviour of non-ideal gases are at least partly of this type. The theory has been given by F. London, who derived the following

formula for the interaction energy:

$$U = -\frac{3}{2}\frac{hc}{R^6}\sum_{n,r}\frac{\alpha_n \alpha_r}{\lambda_n + \lambda_r}. \quad \ldots \quad (6)$$

Here R is the mutual distance, the summation extends over all excited states of both atoms, λ_n and λ_r are the wavelengths corresponding to the excitation energies and α_n, α_r are partial polarizibilities. These London-Van der Waals forces give in some cases an appreciable contribution to the binding energy in solids and they play an important role in adsorption phenomena.

Hamaker [32]) was the first to point out that Van der Waals attraction between colloidal particles may be significant and later Verwey and Overbeek [33]) developed in detail the theory of stability of lyophobic colloids in which Van der Waals attraction is contending with the repulsion of electric double layers. A further step was made by Overbeek when he found that a suspension of quartz particles was more stable than was to be expected from his formulae. He concluded that at large distances the interaction must decrease more rapidly than the inverse sixth power and he suggested that this might be due to retardation effects. Polder and I [34]) were able to treat this problem by means of quantum electrodynamics: we found that London's formula has to be multiplied by a monotonically decreasing correction factor. In the limit $R \gg \lambda_n$ one finds:

$$U = -\frac{23}{8\pi^2}\times\frac{hc}{R^7}\alpha_1(0)\,\alpha_2(0), \quad \ldots \quad (7)$$

where $\alpha_1(0)$, $\alpha_2(0)$ are the static polarizibilities. This limiting case can be treated very simply by considering the change of zero point energy of empty space. Applying similar ideas to resonant cavities with movable walls I derived a formula for the attraction between two conducting plates. The force per cm^2 is given by:

$$K = \frac{\pi hc}{480}\frac{1}{d^4}, \quad \ldots \ldots \quad (8)$$

when d is the distance between plates. Sparnaay [35]) was able to measure this force and found good agreement between theory and experiment. More qualitative evidence for this type of forces was found by another group of our laboratories when they had to approach a metal surface with very thin wires ending in a little sphere.

Personally I think that this connection between colloid chemistry, Van der Waals forces, quantum field theory and zero point energy of empty space and universal attraction between metals is rather amusing. Also I should like to point out that there exists now a tendency to make smaller and smaller objects. If we ever learn to manipulate dimensions and distances of a few tenths of a micron then forces between metals might have a dominating influence.

In this rambling talk I have already mentioned many subjects and indicated many lines of research. I have spoken from the point of view of a physicist. This I felt I could safely do because the points of view of the chemist and the engineer will be well represented during the rest of this meeting. But I am afraid that one aspect of our work will receive inadequate attention and that is the mathematical aspect which does not lend itself so easily to short lectures or snappy demonstrations. Yet we have ever since the days of Van der Pol a pretty good tradition in applied mathematics. Now I have just time for explaining one little piece of mathematics and my choice is Van der Pauw's theorem [36]) because of its outstanding elegance and simplicity. If we have to determine a specific resistance we usually take a prismatic or cylindrical rod (*fig. 18*). Van der Pauw's theorem states that we can take a plate of known and constant thickness but otherwise of arbitrary shape with four small contacts at arbitrary points (*1*, *2*, *3*, *4*) on the circumference. Then, if we define R_{12} as the ratio of the voltage between *1* and *2* and current between *3* and *4*, and similarly R_{23}, we have:

$$e^{-\pi d R_{12}/\varrho} + e^{-\pi d R_{23}/\varrho} = 1. \quad \ldots \quad (9)$$

Let me indicate the proof. First take an infinite half plane. Then all potentials are logarithmic and the proof is elementary. Next map conformly to obtain the desired contour and remember that resistance is a conformal invariant. That is all. Very simple once you come to think of it which apparently no one did before Van der Pauw. And very useful: one measures two resistance values and reads ϱ/d from

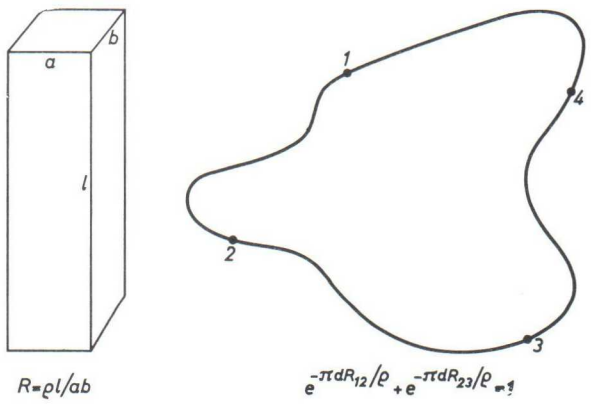


Fig. 18. Illustrating Van der Pauw's theorem.

a set of curves. We do not only apply this in the research laboratory but also in our factories to test specific resistance.

And this is the end of my talk.

BIBLIOGRAPHY

In general, only the earliest publication on each subject has been quoted.

1) G. Holst and L. Hamburger, Proc. Kon. Ned. Akad. Wet., Amsterdam, **18**, 872, 1915.
2) W. Uyterhoeven and C. Verburg, Bull. techn. Ass. Ing. Ec. polytechn. Brux. **29**, 138, 1933.
3) W. Elenbaas, Z. techn. Phys. **17**, 61, 1936.
4) M. H. A. van de Weijer, Philips tech. Rev. **23**, 246, 1961/62.
5) K. van Duuren, A. J. M. Jaspers and J. Hermsen, Nucleonics **17**, No. 6, 86, 1959.
6) N. F. Verster, H. L. Hagedoorn, J. Zwanenburg, A. J. J. Franken and J. Geel, Nucl. Instr. Meth. **18/19**, 88, 1962.
7) J. L. H. Jonker, A. J. W. M. van Overbeek and P. H. de Beurs, Philips Res. Repts **7**, 81, 1952.
8) O. Reifenschweiler, Nucleonics **18**, No. 12, 69, 1960.
9) M. T. Vlaardingerbroek, K. R. U. Weimer and H. J. C. A. Nunnink, Philips Res. Repts **17**, 344, 1962.
10) H. G. van Bueren, J. Haisma and H. de Lang, Physics Letters **2**, 340, 1962.
11) A. Bouwers and A. Kuntke, Z. techn. Phys. **18**, 209, 1937.
12) F. M. Penning, Physica **4**, 71, 1937.
13) F. M. Penning and J. H. A. Moubis, Physica **4**, 1190, 1937.
14) J. L. Snoek, Physica **3**, 463, 1936; New developments in ferromagnetic materials, Elsevier, Amsterdam 1949.
15) B. Jonas and H. J. Meerkamp van Embden, Philips tech. Rev. **6**, 8, 1941.
16) A. I. Luteijn and K. J. de Vos, Philips Res. Repts **11**, 489, 1956.
17) J. J. Went, G. W. Rathenau, E. W. Gorter and G. W. van Oosterhout, Philips tech. Rev. **13**, 194, 1951/52.
18) E. W. Gorter and J. A. Schulkes, Phys. Rev. **90**, 487, 1953.
19) J. S. van Wieringen, Phys. Rev. **90**, 488, 1953.
20) J. J. van Loef, to be published.
21) J. Smit, J. Phys. Radium **20**, 362, 1959.
22) B. D. H. Tellegen, Philips Res. Repts **3**, 81, 1948.
23) H. B. G. Casimir, Rev. mod. Phys. **17**, 343, 1945.
24) C. L. Seeger, F. L. H. M. Stumpers and N. van Hurck, Philips tech. Rev. **21**, 317, 1959/60.
25) J. Volger and P. S. Admiraal, Physics Letters **2**, 257, 1962.
26) A. E. van Arkel, Physica **3**, 76, 1923.
27) P. Penning and D. Polder, Philips Res. Repts **16**, 419, 1961.
28) J. Hornstra, J. Phys. Chem. Solids **5**, 129, 1958.
29) M. C. Teves and T. Tol, Philips tech. Rev. **14**, 33, 1952/53.
30) H. J. Lemmens, M. J. Jansen and R. Loosjes, Philips tech. Rev. **11**, 341, 1949/50.
31) R. Loosjes and H. J. Vink, Philips Res. Repts **4**, 449, 1949.
32) H. C. Hamaker, Physica **4**, 1058, 1937.
33) E. J. W. Verwey and J. Th. G. Overbeek, Theory of the stability of lyophobic colloids, Elsevier, Amsterdam 1948.
34) H. B. G. Casimir and D. Polder, Nature **158**, 787, 1946.
35) M. J. Sparnaaij, Nature **180**, 334, 1957.
36) L. J. van der Pauw, Philips Res. Repts **13**, 1, 1958.

Summary. A lecture in which numerous subjects of research at the Philips Laboratories are briefly touched upon. Gas discharges, ferromagnetism, solid-state physics, network theory are among the main items. The investigations in these variegated fields have resulted on the one hand in industrial products, such as lamps, counter tubes, cyclotrons, ferromagnetic materials and devices of various types, cathodes etc., many of which are illustrated in this lecture, on the other hand they have materialized in contributions to basic physical theory, some of which are discussed.



Manufacture of integrated circuits

In the planar technique several hundred integrated circuits are manufactured from one silicon wafer measuring a few cm. The successive operations in the manufacturing process require the use of a number of masks. In each mask the pattern required for one operation is repeated several hundred times. Each pattern is first drawn on a very large scale (200 : 1) by means of a drawing machine. The mask is then produced by repetition photography of the drawing. The drawings for the successive patterns must register very accurately; this is being checked here by stacking a number of drawings and viewing them against the light.

(From: Philips tech. Rev. **27**, No. 7, 1966.)

CONTRIBUTIONS OF THE PHILIPS RESEARCH LABORATORIES TO SOLID-STATE CHEMISTRY

by H. J. VINK *).

54-16

Introduction

In the past four decades scientists in the Philips Research Laboratories have been engaged in various areas of research in chemistry such as colloidal chemistry, surface phenomena, electrochemistry, organic chemistry, photo-chemistry, analytical chemistry, biochemistry, glass technology and solid-state chemistry. Especially the field of solid-state chemistry has been extensively studied. At this Symposium it seemed therefore worthwhile to take a retrospective view of the work in this field; to contemplate the problems that have been posed, the methods that have been used to solve them and the results that have been achieved.

The choice of the various topics of research is naturally determined by the fact that this laboratory serves the electronics industry. Chemistry has always played an important role in the development of electrical engineering and electronics, but it is especially in the last 30 years that the importance of chemistry to electrical engineering and even more to electronics has been made abundantly clear. The control and preparation of materials has always been of prime importance but the way chemistry nowadays is able to develop materials with new combinations of known physical properties or materials possessing physical properties that were hitherto unknown or did not occur in such a high degree makes it indispensable for the electronics industry. It is not going too far to state that in chemistry lie the foundations of the rapid expansion and increasing importance of the electronics industry.

However the technical importance of the various studies undertaken in this laboratory will only lightly be touched upon in the course of this review. A more general point of view will be taken here, viz, the interrelation between physical and chemical properties and exact chemical composition.

The physical and chemical properties of a substance depend on its chemical composition. On the other hand the electrical, optical, magnetic and chemical properties can often be conveniently used to determine the chemical composition. Therefore an interplay of physical and chemical methods can often lead to an understanding of the interrelation between physical and chemical properties and exact chemical composition. This interplay requires not only a close cooperation between chemists and physicists, but also, for both of them, a certain ability to understand each other's "language", methods and problems.

Research in this field is largely governed by two questions:

1) What kind of sites in what kind of crystal structures must be occupied by what kind of atoms (or groups of atoms) in what state of valency in order to provoke a certain (desired) combination of physical and chemical properties?
2) Is it possible to develop certain groups of compounds and ways of preparing them such that desired concentrations of specific kind(s) of atoms each in a proper state of valency do indeed occupy the right kind of site, in order to control reproducibly certain desired physical and chemical properties?

In this paper the approach towards these questions will have a chemical bias and will be divided into four main areas, viz, a) crystal chemistry, b) internal charge compensation, c) gases and metals, and d) thermodynamics.

Crystal chemistry

Cation distribution in spinels; experimental determination

An investigation that proved to be the starting point of a whole array of studies was begun by Verwey [1) 2) 3) 4) 5)] in the middle thirties **). The then known oxides of iron and other elements of the transition series showed quite a variety of interesting electrical and magnetic properties. Especially the fact that some of the magnetic oxides were also insulators made it appear worthwhile to start an investigation of these related oxidic compounds (Snoek [6)]). This investigation soon centred around the spinels, a large proportion of which can be described by the general formula XY_2O_4. The spinel structure is cubic and can be characterized by a cubic, almost close-packed arrangement of oxygen (O^{2-}) ions in which the metal ions lie on certain sites surrounded tetrahedrally by four O^{2-} ions and on

*) Philips Research Laboratories, Eindhoven.

**) Literature references are listed at the end of this article.

certain octahedral sites surrounded by six O^{2-} ions whose centres form the apices of a slightly distorted octahedron (*fig. 1*). The packing of the O^{2-} ions can be described by a parameter u. For $u = 0.375$ (3/8) the oxygen ions form an ideally cubic close-packed arrangement. Usually the value of u is somewhat larger than 0.375.

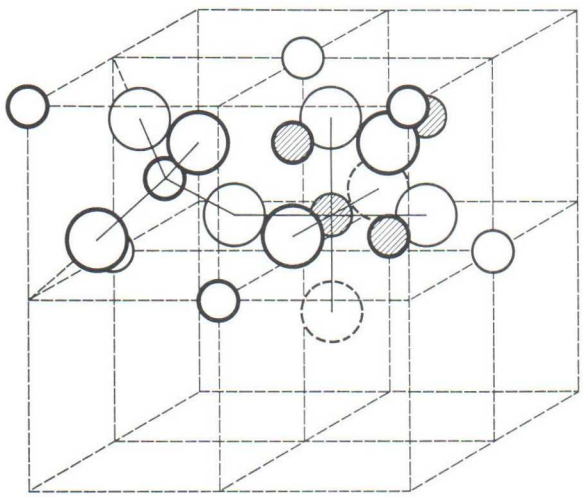


Fig. 1. Unit cell of spinel structure. The position of the ions in only two octants is shown. The dashed circles belong to other octants. The drawn lines indicate the fourfold and sixfold coordination of the respective metal-ion positions. Large circles: oxygen ions; small hatched circles: metal ions on octahedral sites; small unhatched circles: metal ions on tetrahedral sites. The figure is drawn for $u = 0.375$.

The distribution of the cations among the sites available may be:

a) "Normal" with X in the tetrahedral positions and 2Y in the octahedral positions. This may be formulated by $X[Y_2]O_4$, using brackets to indicate cations in the octahedral positions.

b) "Inverse" $Y[XY]O_4$.

c) "Mixed" $X_x\,Y_{1-x}[X_{1-x}Y_{1+x}]O_4$, with $0 < x < 1$.

The metal ions that together with oxygen can form spinel structures must have radii between 0.5 and 1.0 Å, but as there seems to be no restriction as regards valency, a great number of spinels exist.

It soon appeared that the magnetic and electrical properties and their combination depend strongly on the distribution of cations over the two types of sites. It is therefore of interest to know whether there is a relation between the kind (and valency) of an ion and a possible preference for one of the two types of site. The programme, therefore, was first to find out experimentally, for a number of spinels, how the ions were distributed and secondly to give an explanation for these experimental facts so that predictions could be made of the cation distribution and hence the physical properties of other spinels. With regard to the first part of this programme not only binary spinels but also mixed solutions between them, ternary or quaternary spinels had to be investigated in order to determine the relative preference of the various ions for a specific site.

X-ray diffraction immediately suggested itself for this kind of study. The distribution of the cations in a number of spinels has indeed been determined in this way, especially by comparing ratios of intensities of pairs of reflections that are more sensitive to a change in distribution parameter than to a change in u. However in view of the small difference in scattering power between the metal ions of the first transition series or between different ions (Fe^{2+}, Fe^{3+}) of the same atom, this method can only be used for a limited number of cases. For other spinels the study of the electrical and/or magnetic properties can very often be used to determine the cation distribution. This is an example of the fact, mentioned in the introduction, that physical properties that are to be controlled by the exact chemical composition are helpful in their turn in the determination of the chemical composition. The electrical conductivity for instance was used to establish the distribution of the Fe^{2+} and Fe^{3+} ions in Fe_3O_4 [7) 8)]. This spinel has a very high electronic conductivity of the order of $\sigma = 10^2\ \Omega^{-1}cm^{-1}$ at room temperature. Now it was known that substances like Cu_2O which, when pure, show a high resistivity, could have their electronic conductivity considerably increased by a slight deviation from stoichiometry, for instance an oxidation. This had been explained by Wagner and co-workers by assuming that in oxidized Cu_2O not only Cu^+ ions are present but also Cu^{2+} ions. As a consequence of this presence of homonymous ions in a different valency state, electronic conductivity was possible because electronic transport could now occur by exchange of valency electrons:

$$Cu^+Cu^{2+}Cu^+Cu^+Cu^+ \rightarrow Cu^+Cu^+Cu^{2+}Cu^+Cu^+.$$

Verwey and De Boer [7) 8)] explained the high electronic conductivity of Fe_3O_4 by assuming it to have the inverse structure $Fe^{3+}[Fe^{2+}Fe^{3+}]O_4^{2-}$ rather than the normal structure $Fe^{2+}[Fe^{3+}Fe^{3+}]O_4^{2-}$. As a matter of fact, the Fe ions of different valency occupying one type of crystallographic site (octahedral) will be more similar to the situation in oxidized Cu_2O than the case of the normal spinel structure where Fe^{2+} and Fe^{3+} ions occupy different crystallographic sites, viz, the tetrahedral and octahedral sites respectively.

Another physical property that can often be used is the saturation magnetization at 0 °K. Néel's theory of ferromagnetism postulated a strong

negative exchange interaction (TO) between ions on tetrahedral (T) and those on octahedral sites (O) and weaker negative TT and OO interactions. In simple cases, i.e. for which the TO interaction is dominant, it is possible to find the saturation magnetization (σ) at 0 °K, expressed in Bohr magnetons (μ_B) per formula unit by simply subtracting the sum of the ionic magnetic moments occurring on the T sites from those on the O sites or vice versa. Let us take as examples the compounds $MgFe_2O_4$ and $NiFe_2O_4$ both of which, because of their X-ray diffraction patterns, were thought to be inverse spinels $Fe^{3+}[Mg^{2+}Fe^{3+}]O_4^{2-}$ and $Fe^{3+}[Ni^{2+}Fe^{3+}]O_4^{2-}$. However the accuracy of these determinations was not very high. Therefore the saturation magnetization was determined for these compounds in order to substantiate the rather vague results of the X-ray diffraction investigations. According to these results and applying Néel's rules, the saturation magnetization values one expects to find are $\sigma = 0$ for $Fe^{3+}[Mg^{2+}Fe^{3+}]O_4^{2-}$ and $\sigma = \mu_{Ni^{2+}} = 2\mu_B$ for $Fe^{3+}[Ni^{2+}Fe^{3+}]O_4^{2-}$. However the values found were 1.1 μ_B and 2.3 μ_B respectively.

Now the value $\sigma = 1.1\ \mu_B$ for $MgFe_2O_4$ could be changed by different heat treatments. For instance quenching from 1250 °C increased its value to $\sigma = 1.4\ \mu_B$. Gorter [9]) explained these results by assuming that Mg-ferrite actually shows a "mixed" distribution, $Fe^{3+}_{1-x}Mg^{2+}_x[Mg^{2+}_{1-x}Fe^{3+}_{1+x}]O_4^{2-}$ with $x = 0.11$ and 0.14 respectively. The same explanation however does not apply to Ni-ferrite. Its saturation magnetization does not change with different heat treatments. Gorter therefore had to assume that $x = 0$ for Ni-ferrite, and further that the magnetic moment of Ni^{2+} also had a contribution from the orbit leading to a total spin moment of 2.3 μ_B. This value of the total spin moment of Ni^{2+} has been amply verified in other experiments. These simple and well known cases are given here to demonstrate how careful one must be in interpreting physical phenomena into chemical formulae. The facts that 1) ions can occur in more than one (unknown) state of valency, 2) the distribution in ternary systems has to be described by two parameters x and y and 3) the dilution by non-magnetic ions interferes in an unknown way with the interaction between the magnetic ions, all underline the care necessary in interpretation.

Cation distribution in spinels; theoretical explanation

Romeijn [10]) [11]), Gorter [12]) [13]) [14]) and Blasse [15]) prepared a great number of binary, ternary and quaternary spinels and studied the cation distribution using the methods mentioned above. Their combined results proved to be of an enormous complexity. This is not very surprising however as there are a great number of different kinds of energies to be taken into account. In the course of the years the following contributions to the lattice energy have been considered: a) electrostatic Coulomb-forces, b) Born repulsion, c) ordering, d) individual site preference of some cations having a tendency to be surrounded by a certain configuration of oxygen ions, e) crystal-field effects, f) ligand-field effects and g) polarization.

Of the corresponding energies only the Coulomb energy [16]) [17]) and the ordering energy [18]), by making certain assumptions, could be calculated with any accuracy (Verwey, Heilmann, De Boer and Van Santen). It appeared however that this Coulomb energy depends on the values of the oxygen parameter u and the length of the unit cell edge a (both themselves depending on the actual distribution) to such an extent that it proved to be virtually impossible to predict any cation distribution from this effect alone. Geometrical arguments replacing calculations of Born repulsion energies [10]) also did not lead to any conclusion. Further, the energies calculated for ordering phenomena proved to be a substantial fraction of the Coulomb energy, for cases where one or both of the sublattices are occupied by ions differing in charge. This means that there must exist a tendency to the formation of superstructures in these spinels. Such a superstructure will generally be accompanied by lattice distortions. Superstructures have indeed been found in a few cases: $Fe^{3+}[Fe^{2+}Fe^{3+}]O_4^{2-}$ [19]), $Fe^{3+}[Li^{+}_{0.5}Fe^{3+}_{1.5}]O_4^{2-}$ [20]), $Fe^{3+}[V_{1/3}Fe^{3+}_{5/3}]O_4^{2-}$ [21]), $Li^{+}_{0.5}Fe^{3+}_{0.5}[Cr^{3+}_2]O_4^{2-}$ [14]) and $In^{3+}_{2/3}V_{1/3}[In^{3+}_2]S_4^{2-}$ [22]), V being the symbol first for a vacancy on an octahedral and then on a tetrahedral site. However the transition temperatures of these ordered phases to a disordered one are too low compared with the high ordering energies involved. It has been argued by Van Santen [23]) that in ionic crystals the energy difference between long-range and short-range order is small, due to the fact that the Coulomb interactions are long-range interactions. This implies that the transition temperature will be low and also that above this temperature a fair degree of short-range disorder will still exist.

Already Verwey and Heilmann pointed out that apart from these electrostatic effects, individual site preferences of some cations must also be taken into consideration, such as the tendency of Zn^{2+}, Cd^{2+}, Ga^{3+}, In^{3+} (and Ge^{4+}) to be surrounded by four anions (sp^3 hybridization).

On the other hand it was suggested by Van Santen and Van Wieringen [24]) and by Romeijn [11])

that the behaviour of ions of the transition series with partially filled shells such as Ti^{3+}, V^{3+}, Cr^{3+}, Mn^{4+}, Mn^{3+}, Fe^{3+}, Mn^{2+}, Co^{3+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+} would be also governed by energy gained by orbital splitting due to the crystalline field.

Other workers made quantitative calculations about this effect. For the ions of the transition series they calculated energies of octahedral site preferences. From these values it could be deduced that ions with $3d^5$ and $3d^{10}$ configuration (Mn^{2+}, Zn^{2+}) should show no site preference at all, whereas the ions with $3d^3$ and $3d^8$ configuration (Cr^{3+} and Ni^{2+}) would show a preference for octahedral sites, in good agreement with experimental data. It is very unsatisfactory however that the cation distribution should be controlled by energies which at their highest (Cr^{3+}: 40 kcal/mol) have values less than 1% of the electrostatic Coulomb energies.

Blasse [15]), using the method of molecular orbitals, combined the effects of specific tendency to covalent binding and of crystal field stabilization. This approximation — the ligand field theory — leads to the following result, differing somewhat from that of the authors just mentioned: ions with $3d^3$ configuration prefer octahedral sites, ions with $3d^5$, $3d^6$, $3d^7$, $3d^9$ or $3d^{10}$ configuration prefer tetrahedral sites, whereas ions with $3d^0$, $3d^1$, $3d^2$, $3d^4$ and $3d^8$ do not show any pronounced preference. These results are of a qualitative nature only. It is therefore impossible to compare the energies due to this kind of arguments with those of the Coulomb interactions. It is reasonable to assume that they will be larger than the energies calculated for the crystal field stabilization but they will again be small compared to electrostatic energies.

Blasse also brought anion polarization into the picture; and he was able to show that polarization of the oxygen ions as a consequence of these being asymmetrically surrounded by differently charged cations can be decisive for the actual cation distribution.

For instance in $Fe^{3+}Ni^{2+}Rh^{3+}O_4^{2-}$, from every other point of view one should expect the configuration $Fe^{3+}[Ni^{2+}Rh^{3+}]O_4^{2-}$. Actually however, part of the Ni^{2+} ions occupy tetrahedral sites, in this way increasing the asymmetry with respect to charge of the immediate surroundings of the oxygen ions and thus increasing the energy gained by polarization.

Cation distribution in spinels; molecular engineering

Although therefore from a theoretical point of view the situation around the cation distribution in spinels as a whole leaves something to be desired, yet the chemist now has enough experimental data and theoretical considerations at his disposal to be able to perform "molecular engineering" with this structure. In this way a great many products could be developed differing in initial permeability, saturation magnetization, electrical resistivity, electrical losses, magnetic losses etc. It appeared that for all spinels the product of the static initial permeability and the dispersion frequency is substantially constant (Snoek's relation [25])). We will not go into all this but give just one very simple example of molecular engineering in this field of magnetic materials.

When $ZnFe_2O_4$ ($Zn^{2+}[Fe^{3+}Fe^{3+}]O_4^{2-}$), having a total magnetic moment of zero, is introduced into $NiFe_2O_4$ ($Fe^{3+}[Ni^{2+}Fe^{3+}]O_4^{2-}$), it gives rise to mixed crystals with a *higher* magnetic moment than $NiFe_2O_4$ [26]). The distribution of the different ions in the mixed crystals, according to the rules mentioned above will be

$$Zn_x^{2+}Fe_{1-x}^{3+}[Ni_{1-x}^{2+}Fe_{1+x}^{3+}]O_4^{2-},$$

leading to a saturation magnetization per molecule of

$$(1-x)\mu_{Ni^{2+}} + 2x\mu_{Fe^{3+}}.$$

This is also the case when $ZnFe_2O_4$ is introduced into other spinels (*fig. 2*).

For higher concentrations this effect disappears,

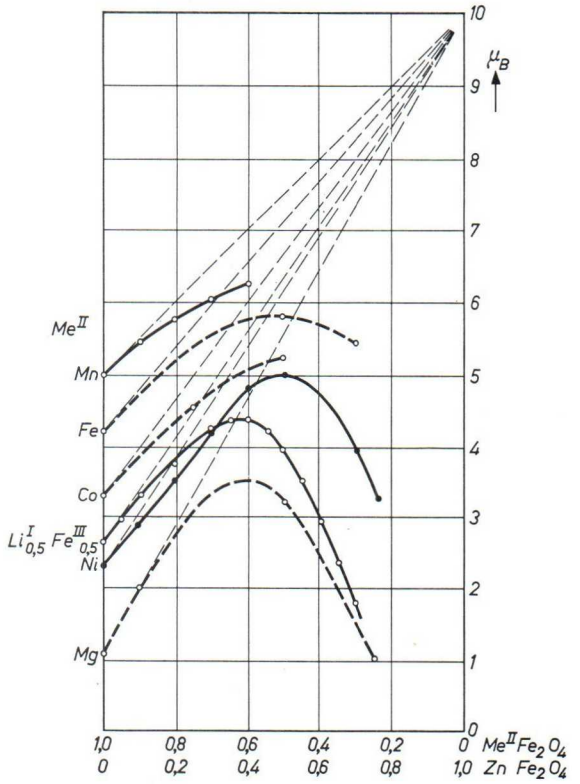


Fig. 2. The saturation magnetization in Bohr magnetons per "molecule" for mixed crystals of Zn ferrite with other ferrites.

the dilution with non-magnetic ions on the tetrahedral sites making the TO exchange small compared with the OO and TT interactions.

Hexagonal and trigonal ferrites; determination of their structure and composition

In extending the study of spinels Braun, Jonker and Wijn found an interesting group of new compounds in the system BaO - MeO - Fe_2O_3 (*fig. 3*) [27] [28] [29] [30] [31] [32] [33]), where Me stands for a divalent metal like Fe, Mn, Co, Zn and Mg.

These ferrites all crystallize with hexagonal crystal structures. Their compositions are given by the points **M**, **W**, **X**, **Y**, **Z** and **U** in fig. 3 and can be represented by the formulae:

M	$BaFe^{III}_{12}O_{19}$
W	$BaMe^{II}_2Fe^{III}_{16}O_{27}$
X	$Ba_2Me^{II}_2Fe^{III}_{28}O_{46}$
Y	$Ba_2Me^{II}_2Fe^{III}_{12}O_{22}$
Z	$Ba_3Me^{II}_2Fe^{III}_{24}O_{41}$
U	$Ba_4Me^{II}_2Fe^{III}_{36}O_{60}$

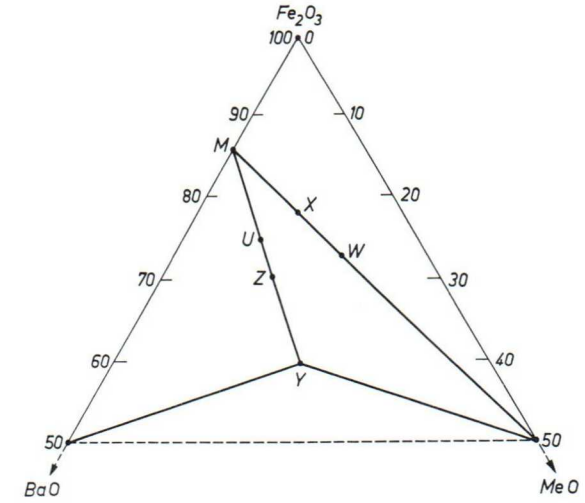


Fig. 3. Part of the phase diagram BaO-MeO-Fe_2O_3, where Me stands for a divalent metal such as Mn, Fe, Co, Ni, Zn or Mg.

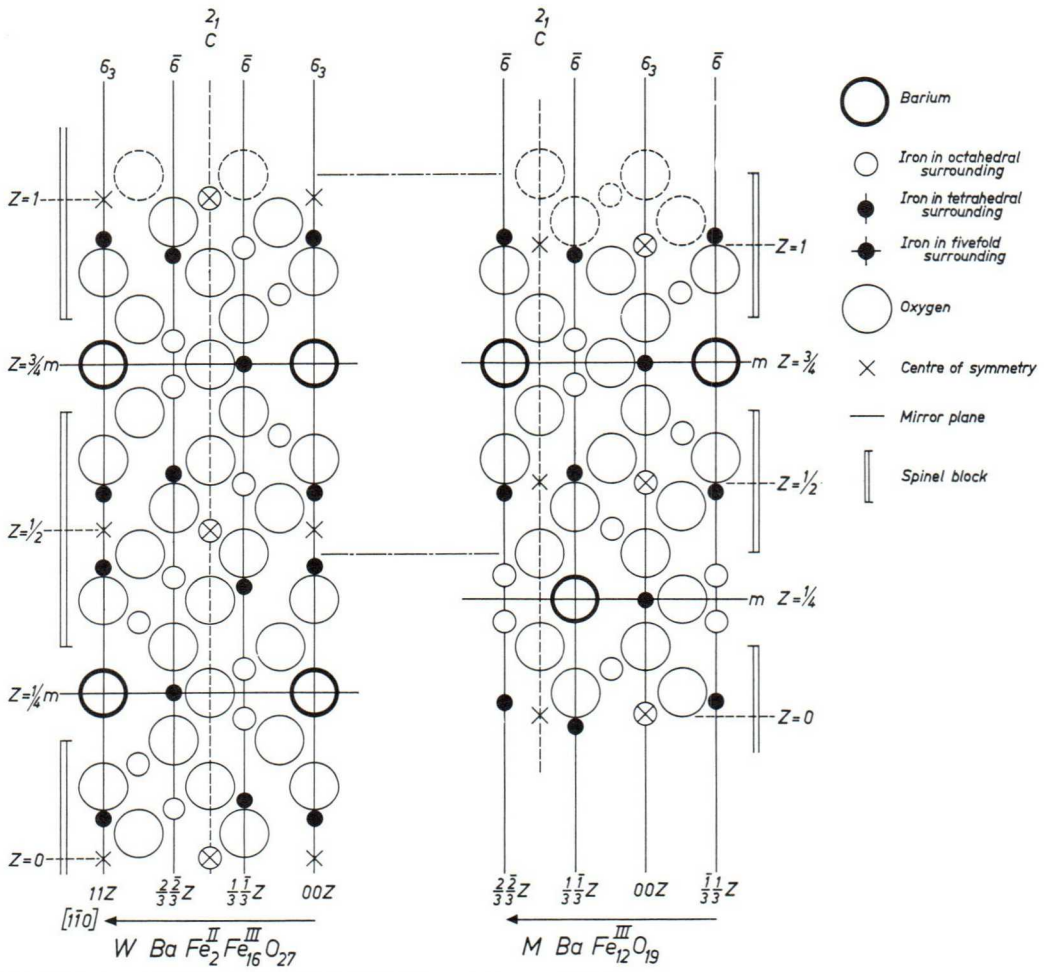


Fig. 4. Unit cell of **W** compared with that of **M**. Only atoms and symmetry elements in a mirror plane containing the *c*-axis are shown. The relative positions of all atoms in nine layers and those of the interstitial atoms between are identical in the two structures. The spinel blocks include 4 oxygen layers in **M** and 6 oxygen layers in **W**.

The preparation of those crystals often leads to a mixture of a number of compounds. Therefore it would have been impossible to determine the very existence and the precise composition of these compounds by the usual chemical methods. This study however was greatly facilitated by the fact that contrary to the magnetic spinels these hexagonal compounds show a strong magnetic anisotropy. By making use of this effect the finely powdered reaction products could be oriented in such a way that some of the compounds would have their hexagonal axis parallel and others perpendicular to an applied magnetic field. An X-ray diffraction pattern of such an oriented powder is greatly simplified. The different reaction products could now be "labelled" so to speak and in this way it was possible to improve the reaction procedure such that single crystals of the various compounds could be obtained. Braun [30] [31] was then able to determine the crystal structure of these compounds and from this their chemical composition. The complicated chemical formulae of these compounds have thus been determined not directly which would have been impossible, but by means of X-ray diffraction. This again exemplifies how the chemist can make use of physical properties to determine not only the composition but the very existence of complicated compounds.

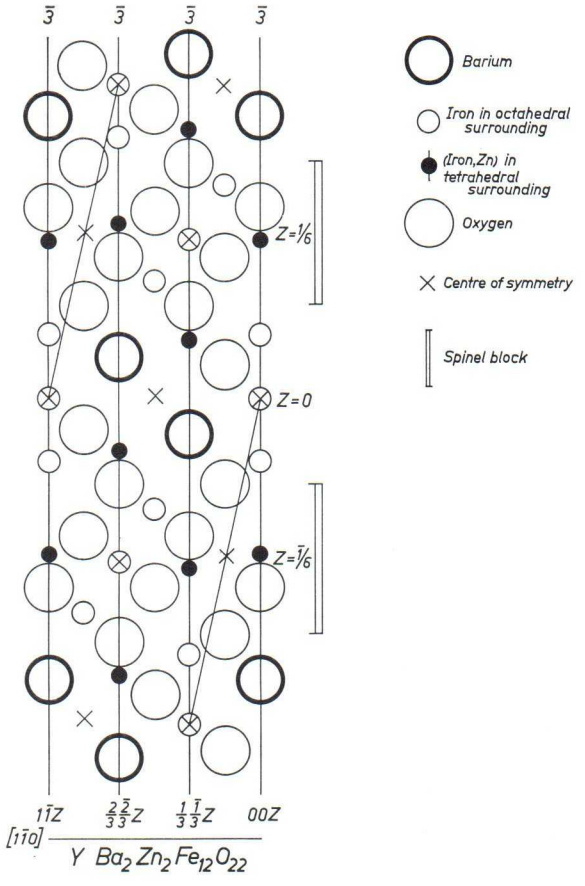


Fig. 5. Two thirds of a unit cell of **Y**. Only the atoms and symmetry elements in a mirror plane containing the *c*-axis are shown. The spinel blocks include 4 oxygen layers. The rhombohedral repeat distance is indicated.

In the determination of these structures use was made of the fact that a strong relationship appeared to exist between the crystal structure of these hexagonal ferrites and the well known spinel structure. These structures can be described in different ways [34]. In this paper the notation of Braun [31] will be followed. In the cubic close- packing of the O^{2-} ions in spinels, trigonal O^{2-} layers perpendicular to a [111] direction have a mutual distance of 2.33 Å. In the unit cell (32 O^{2-} ions) there are six of these layers.

Now the crystal structures of **M**, **W**, and **X** also contain such close-packed layers. However in some of these layers (the "Ba-layers") one quarter of the O^{2-} ions is substituted in an ordered way by Ba^{2+} ions that are about as large as the O^{2-} ions. The compounds can generally be described by an alternation of blocks with pure spinel structure and consisting of 4 or 6 layers followed by blocks of 1 or 2 "Ba-layers". The "Ba-layers" are connected with the spinel layers in a "hexagonal" way. For instance in the structure of **M** one can distinguish blocks of 4 layer with cubic close-packing of O^{2-} ions and a spinel-like distribution of Fe^{3+} ions over the tetrahedral and octahedral spinel sites between the O^{2-} ions (S_4) followed by one Ba-layer (B_1) also containing Fe ions. The unit cell consists of two equivalent parts each containing 5 layers with, in total, 19 O^{2-}, 1 Ba^{2+} and 12 Fe^{3+} ions, and with a thickness of $5 \times 2.33 = 11.6$ Å. Similarly the compound **W** contains blocks of 6 spinel layers (S_6) alternating with a "Ba-layer" (B_1). Here the spinel layers, as well as Fe^{3+} ions, also contain 2 Me^{2+} ions (see *fig. 4*). **X** consists of a regular alternation of layers with **M** and **W** structures. Therefore one has:

$\mathbf{M} = M_5 = B_1S_4B_1S_4B_1 \ldots$ thickness $11.6 = 5 \times 2.33$ Å
$\mathbf{W} = W_7 = B_1S_6B_1S_6B_1 \ldots$ thickness $16.4 = 7 \times 2.34$ Å
$\mathbf{X} = X_{12} = B_1S_4B_1S_6B_1 \ldots$ thickness $28.0 = 12 \times 2.33$ Å.

In the compounds **Y**, **Z**, **U** one encounters as a new building block, that of the double Ba-layer (*fig. 5*):

$\mathbf{Y} = Y_6 = B_2S_4B_2 \ldots\ldots$ thickness $14.7 = 6 \times 2.42$ Å
$\mathbf{Z} = Z_{11} = B_2S_4B_1S_4B_2 \ldots\ldots$ thickness $26.2 = 11 \times 2.38$ Å
$\mathbf{U} = U_{16} = B_2S_4B_1S_4B_1S_4B_2 \ldots$ thickness $37.8 = 16 \times 2.36$ Å.

Hexagonal and trigonal ferrites; magnetic properties

An elegant confirmation of the structure of these compounds is the fact that Gorter [35]) was able, by straightforward reasoning, to derive from these structures the ferrimagnetic character of the **W**, **Y** and **Z** compounds and to calculate the values of their saturation magnetization.

It has been said already that the new magnetic compounds show a magnetic anisotropy. There are two extremes of this anisotropy. For some of the compounds the ion-spins are all directed parallel to the *c*-axis, whereas in others they are perpendicular to this axis. In the first case the components are magnetically hard which gives us a valuable new class of materials in addition to that of the magnetically soft spinels. In the second case, components with a preferred plane of magnetization, the magnetization is virtually free to assume any direction in that plane, but energy is required for rotation out of the plane. Such a preferred plane occurs with the **Y** group and also with those compounds of the **Z** and **W** groups in which Co-ions are completely or partly substituted for Me. *Table I* illustrates this phenomenon more clearly.

Table I. Compounds with a preferred direction (↑) or a preferred plane (⊥).

Group	Divalent ion substituted for Me					
	Mn	Fe	Co	Ni	Zn	Mg
W	↑	↑	⊥	↑	↑	↑
Z	↑	↑	⊥	↑	↑	↑
Y	⊥	⊥	⊥	⊥	⊥	⊥

An important feature of the compounds showing a preferred plane is the fact that here the relation of Snoek mentioned earlier is no longer valid and that the compounds can therefore be used at much higher frequencies [34]).

It is very intriguing how complicated structures like those mentioned above can be formed out of a melt or by sintering of the oxides and that an irregular alternation of the different but very similar blocks of layers never seems to occur. No crystal was found with an averaged structure, so mistakes in growing occur much less often than one would expect from the close relationship of the layers.

Polytypes in SiC; growth phenomena and stability

This problem of how to explain the "memory" of a growing crystal for such a complicated ordered sequence of layers is encountered in a perhaps more pronounced degree in the well-known polytypes of SiC, although in these structures a slight disorder can often be found. SiC has been a subject of study in this laboratory for a number of years. In the course of these studies Lely [36]) developed a method of growing large pure SiC crystals in the laboratory, by recrystallization of polycrystalline material under its equilibrium Si pressure. Knippenberg [37]) studied especially the growth phenomena of SiC polytype crystals, using an improved Lely technique and also various other methods of preparation. One of his results was that independent of the way of preparation a definite statistical relationship seems to exist between the polytypes formed and the temperature of preparation. A qualitative picture of this relation is sketched in *fig. 6*. As can be seen, cubic SiC could

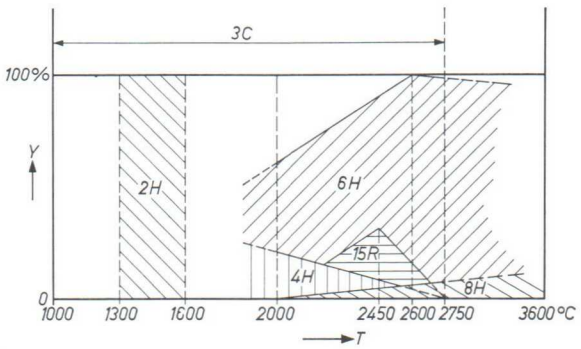


Fig. 6. Relation between structure of SiC and temperature of preparation. The relative amounts *y* of the different non-cubic structures (H and R) formed at the temperature *T* of preparation are shown qualitatively. Cubic (3C) SiC (unstable) is formed at all temperatures. The highest temperature of preparation was 2750 °C.

be formed at all temperatures but is always unstable. The cubic modification grows only under conditions of supersaturation and then in a "skeleton" formed by multiple twinning of anisotropic lamellae.

Polytypes in SiC; variation of the band gap

It is a remarkable fact that for these very closely related structures the variation of the band gap with the polytype is surprisingly large. *Table II* gives a comparison of the band gaps for the different polytypes. The second column in this table shows the way in which for every polytype the hexagonal (h) and cubic (c) layers follow each other along the *c*-axis.

One sees that the band gap seems to vary in a

Table II. Variation of the band gap in SiC with the polytype.

Polytype	Sequence of layers	Band gap (eV)	α
2 H	hh	> 3	0
4 H	hchc	3.1	0.50
15 R	hchcc	2.9	0.60
6 H	hcchcc	2.86	0.67
24 R	hcchcccc	2.72	0.75
3 C	cccc . . .	2.2	1

systematic way. This can be seen more clearly if one plots the band gap against the ratio a (Table II) of the number of cubic layers to the total number of layers per unit cell (*fig.* 7).

It is known that in mutually comparable structures there exists a relationship between band gap, lattice energy and single-bond energy, the band gap increasing with increasing lattice energy or increasing bond energy. Applying this to the series of polytypes of SiC one may expect an increase in single-bond energy and lattice energy for the respective polytypes going from the bottom to the top in Table II. The fact that cubic SiC is unstable seems to be in line with this correlation.

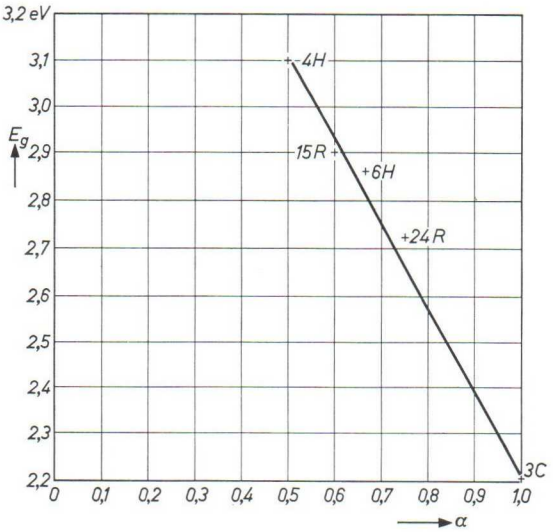


Fig. 7. Plot of the band gap E_g of various polytypes of SiC as a function of the ratio a of the number of cubic layers to the total number of layers per unit cell.

Topotactical reactions

We will now return once again to the system BaO-MeO-Fe_2O_3 (with Me = Zn, Co, etc.). Under proper conditions these compounds can be converted into each other by solid-solid reactions. For instance, when a pellet consisting of 1 equivalent $BaFe_{12}O_{19}$, 2 equivalents CoO and 2 equivalents Fe_2O_3 is fired for some hours at 1250 °C, the following solid-solid reaction takes place:

$$BaFe_{12}O_{19}\uparrow + 2CoO + 2Fe_2O_3 \rightarrow BaCo_2Fe_{16}O_{27}\perp,$$

the signs $\uparrow$ and $\perp$ indicating a preferred direction and a preferred plane of magnetization respectively. In this way a sintered aggregate of hexagonal crystals is obtained. A remarkable discovery, made by Lotgering [38]), was the fact that when the small grains of $BaFe_{12}O_{19}$ (**M**) present in the original pellet had been previously oriented by means of a static magnetic field, the c-axes of the new crystals (**W**) show the same orientation as the c-axes of the original **M** crystals, which disappear during the reaction. In the initial mixture only the **M** crystals are oriented. It is evident therefore that the preferential orientation of the **W** crystals is caused by chemical reactions that take place on the surface of the **M** crystals.

Such reactions are called topotactical reactions. A topotactical reaction starts as an epitaxial process; there is however, an essential difference from epitaxy. The final product of an epitaxial process is an intergrowth of the new phase and the substrate, whereas during topotactical reactions (like the one just mentioned) the substrates disappear completely during the reaction. In the example given, the amount of preferentially oriented material increases considerably during the reaction. It was found in further study that not all reactions of this kind lead to well oriented products, diffusion of reactants and intermediate products playing an important role. Of course as the reactions all take place at temperatures (1000-1300 °C) that are far above the Curie temperatures (300-500 °C), the magnetic fields of the ferrimagnetic grains originally present have disappeared and cannot play any part in the orientation.

Internal charge compensation

Controlled valency

Verwey and De Boer [7]), combining Wagner's conception of the occurrence of differently charged ions in non-stoichiometric compounds (giving electronic conductivity) and their own explanation of the conductivity of Fe_3O_4, generalized these ideas by stating that electronic conductivity could be expected for any compound in which equivalent crystallographic sites were occupied by ions of the same element occurring in two states of charge differing by not more than one electron unit.

Starting from these notions Verwey, Haaijman, Romeijn and Van Oosterhout [39]) [40]) developed the conception of controlled valency. According to this principle the introduction on Ni-sites of a monovalent ion like, say, Li^+ into the lattice of a non-conducting compound such as NiO, containing divalent Ni^{2+} ions, should bring about the formation of an equal amount of Ni^{3+} ions, in order to maintain electro-neutrality. This being the case, the lattice would then contain, on equivalent crystallographic points, Ni ions differing in charge by only one electronic unit. Therefore this Li-containing NiO should show electronic conductivity. This was in fact found. The conductivity of, for instance, pure NiO could be increased from 10^{-14} $\Omega^{-1}cm^{-1}$ to 10^2 $\Omega^{-1}cm^{-1}$ by the introduction of 10% Li.

Of course one condition has to be fulfilled, viz, the foreign ion (here Li^+) must show a smaller tendency to change its charge than the ions (Ni^{2+}) of the base crystal. Further, in order to get appreciable effects, the ionic radius of the foreign ion must not differ too much from that of the ion to be substituted, thus facilitating the incorporation.

It will be clear that in this way both *N*- and *P*-type semiconducting compounds can be obtained. The incorporation of Li^+ into NiO leading to the formation of Ni^{3+} ions among the Ni^{2+} ions gives *P*-type conduction, whereas incorporation of Ti^{4+} ions into Fe_2O_3 leading to the formation of Fe^{2+} ions among the Fe^{3+} ions will give rise to *N*-type conductivity.

Table III shows a number of mainly oxidic systems where this principle has been applied. Column 1 contains the formula of the poorly conducting mother substance, column 2 the substance containing the ion which replaces the bold-type ion of column 1. Columns 3 and 4 specify respectively the normal ions and the ions of deviating valency that are presumably present in the final system.

Table III. Systems in which ions of deviating valency have been introduced.

1	2	3	4	Crystal structure
NiO	Li_2O	Ni^{II}	Ni^{III}	
CoO	Li_2O	Co^{II}	Co^{III}	rocksalt
MnS	Li_2S	Mn^{II}	Mn^{III}	
CaTiO_3	La_2O_3	Ti^{IV}	Ti^{III}	
SrTiO_3	La_2O_3	Ti^{IV}	Ti^{III}	
BaTiO_3	La_2O_3	Ti^{IV}	Ti^{III}	
Ca$Mn^{IV}O_3$	La_2O_3	Mn^{IV}	Mn^{III}	perovskite
La$Mn^{III}O_3$	CaO	Mn^{III}	Mn^{IV}	
La$Mn^{III}O_3$	SrO	Mn^{III}	Mn^{IV}	
La$Fe^{III}O_3$	SrO	Fe^{III}	Fe^{IV}	
Zn**Fe**$_2O_4$	TiO_2	Fe^{III}	Fe^{II}	
Mg**Fe**$_2O_4$	TiO_2	Fe^{III}	Fe^{II}	
Ni**Fe**$_2O_4$	TiO_2	Fe^{III}	Fe^{II}	spinel
Co**Fe**$_2O_4$	TiO_2	Fe^{III}	Fe^{II}	
Fe$_2O_3$	TiO_2	Fe^{III}	Fe^{II}	
Fe$_2O_3$	SnO_2	Fe^{III}	Fe^{II}	hematite
Fe$_2O_3$	WO_3	Fe^{III}	Fe^{II}	
SnO_2	Sb_2O_5	Sn^{IV}	Sn^{II}	rutile
TiO_2	Ta_2O_5	Ti^{IV}	Ti^{III}	
MgWO_4	Cr_2O_3	W^{VI}	W^{V}	wolframite

However, the systems represented in Table III are by no means the only ones for which this principle of controlled valency works. Further research made it clear that this principle is quite general and is also valid for instance in compounds like CdS [41]), PbS [42]), CdTe [43]) [44]), which on introduction on cation sites of ions like Ag^+ may become *P*-type and of ions like Ga^{3+}, In^{3+}, Sb^{3+} may become *N*-type. Not only can foreign cations change the valency of the cations of the base crystal, but also foreign anions can bring this about. For instance, the introduction of Cl^- ions on S^{2-} sites in CdS may lead to the formation of an equal amount of Cd^+ ions between the normal Cd^{2+} ions and consequently to an *N*-type semiconductor [41]).

Experimental determination of electron-energy levels of impurities in Fe_2O_3

As has been said the foreign ion must show a smaller tendency to change its charge than the ions of the base lattice. Jonker [45]) has investigated this requirement more quantitatively for Fe_2O_3. Introduction of Ti gives rise to the formation of

$$(Ti_x^{4+}Fe_x^{2+}Fe_{1-2x}^{3+})_2O_3^{2-}$$

and *N*-type conduction. However, at first sight a second formula $(Ti_x^{3+}Fe_{1-x}^{3+})_2O_3^{2-}$ not involving any change of charge of the Fe ions, cannot be excluded. To compare the two possibilities one has to consider the energy balance of the reaction

$$Fe^{3+} + Ti^{3+} \rightarrow Fe^{2+} + Ti^{4+} + E.$$

The energies involved are the following.

1) The difference of the ionization potentials: $I_3(Fe) - I_4(Ti)$.
2) The polarization energies $\sum E_p$ caused by the electric field around the ions Fe^{2+} and Ti^{4+} having a charge that differs from that of the Fe^{3+} ions occurring in Fe_2O_3. Rough calculations show that this polarization energy for a pair of Fe^{2+} and Ti^{4+} ions is about 16 eV.
3) Crystal-field stabilization energies $\sum E_{cf}$ being of the order of 1 eV.

This leads to

$$E = I_3(Fe) - I_4(Ti) + \sum E_p + \sum E_{cf}.$$

Samples of Fe_2O_3 to which different foreign atoms where added in a concentration of 2 at. % were all fired in air at 1300 °C, and their resistivities measured at room temperature. The results are given in *Table IV*.

It can be seen from Table IV that two groups of elements can be distinguished, a first group leading to low values of ϱ and a second group having no influence on the resistivity. Obviously for the second group of elements the value of E of the reaction is negative.

Neglecting the relatively small values of $\sum E_{cf}$, the value for $I_3(Fe) - I_4(f.a.)$ must be compensated by $\sum E_p$, the polarization energies, in order to have controlled valency. From Table IV it follows that for Fe_2O_3, $\sum E_p$ is of the order of 17 eV, in good agree-

Table IV. Influence of 2 at.% substitution of foreign ions on the resistivity of Fe_2O_3. For each kind of ion the fourth ionization potential I_4 and the difference between this and the third ionization potential of iron I_3(Fe) is given.

Foreign atom (f.a)	ϱ Ω cm	I_4(f.a.) eV	I_3(Fe) − I_4(f.a.) eV
Ta	1.3	33.3	− 2.7
Zr	0.7	34.0	− 3.4
Nb	7.9	38.3	− 7.7
Ti	0.6	43.2	− 12.6
Sn	0.4	46.4	− 15.8
V	1.5×10^6	48.0	− 17.4
Cr	1.0×10^3	50.6	− 20.0
Mn	7.0×10^7	53.7	− 23.1
Fe	1.7×10^4	56.2	− 25.6

ment with rough calculations. Attention must be drawn to the fact that the polarization energy which is difficult to calculate, is in this way determined experimentally by measuring resistivities. The results of this can further be used to construct an electronic energy-level diagram for Fe_2O_3 with several types of impurities. As will be discussed further on, it is generally assumed that for transition-metal oxides this picture consists not of bands but of localized levels. The distance E_g between the occupied Fe^{3+} levels and the empty Fe^{2+} levels is found from

$$-E_g = I_3(\mathrm{Fe}) - I_4(\mathrm{Fe}) + \Sigma E_p + \Sigma E_{cf}.$$

Here ΣE_{cf} cannot be neglected and is about 2 eV. Taking $\Sigma E_p = 17$ eV and $I_4 - I_3 = -25.6$ eV one gets $E_g \approx 6.5$ eV. *Fig. 8* gives the corresponding electronic energy-level diagram of Fe_2O_3.

It can be remarked that the results obtained in this experiment and also the results of quite a number of other experiments in solid-state chemistry can often be qualitatively predicted by the chemist's "chemical feeling". Now this chemical feeling amounts to an almost unconscious knowledge about the relations between ions dissolved in water. That this feeling nevertheless can account for the results obtained can be understood quite easily when one realizes that the hydratation energy of the ions (6-11 eV) is about equal to the polarization energy (6-9 eV) in these oxidic and other ionic compounds.

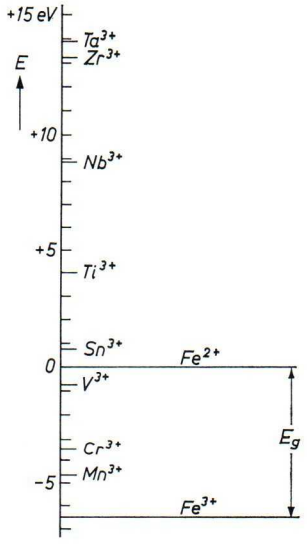


Fig. 8. Electronic energy-level diagram of Fe_2O_3 with the levels of substituted trivalent foreign ions.

Calculation of electronic energy levels in NiO

Van Houten [46]) was able to calculate the electron-energy scheme of NiO. He assumed 1) the crystal to be build up from Ni^{2+} and O^{2-} ions and 2) localization of the electrons. By taking into account a) the ionization energies or electron affinities of the ions involved, b) the Madelung energies, c) the polarization energies and d) the crystal field stabilization of the transition metal ions, a distance of 5.4 eV was found between the highest filled level (that of the Ni^{2+} ions) and the lowest empty level (that of the Ni^{+} ions). This large distance gives, just as did that of 6.5 eV for Fe_2O_3, a quantitative explanation of the fact that pure oxides of this type exhibit a very high resistivity.

Mobility of electronic charge carriers in transition-metal oxides

De Boer and Verwey [8]) drew attention to the fact that pure stoichiometric oxides of the transition metals are nearly all insulators. These oxides, however, contain metal ions with incompletely filled 3d or 4f shells. Especially for those ions having incomplete 3d shells the collective electron treatment of wave functions of Bloch-Wilson would lead one to expect metallic conduction for these oxides. It became clear that instead of the usual band picture of electron energies, these oxides must show sharp localized levels.

As has been said already the conductivity of these oxides could be enhanced substantially by the incorporation of an excess of one of their components or of ions with a deviating valency. Now the semiconductors made in this way mostly show a high positive temperature coefficient of conductivity:

$$\sigma = \sigma_0 e^{-Q/kT}.$$

The activation energy Q can have values lying between 0.1 and 0.5 eV and decreasing with increasing deviation from stoichiometry or concentraton of foreign ions [46]). It has long remained obscure whether this activation energy was caused by the energy

needed to liberate the charge carriers from their centres or by the fact that in contradistinction to semiconductors like Ge, the mobility involves an activation energy.

To investigate this question Jonker [47]) measured the conductivity of mixed crystals of $CoFe_2O_4$ with Fe_3O_4 or Co_3O_4, as a function of temperature and concentration. In the mixed crystals $Co_{3-x}Fe_xO_4$ in the vicinity of $x = 2$, one has for $x > 2$ an Fe^{2+} content of $x - 2$ and for $x < 2$ a Co^{3+} content of $2 - x$. Now applying the rules of occupation of tetrahedral and octahedral sites mentioned in the second section it can be concluded that for $x > 2$ $(Fe^{3+}[Co^{2+}_{3-x}Fe^{3+}Fe^{2+}_{x-2}]O^{2-}_4)$ one has electron conduction among the Fe ions and that for $x < 2$ $(Fe^{3+}[Co^{2+}Co^{3+}_{2-x}Fe^{3+}_{x-1}]O^{2-}_4)$ one has hole conduction among the Co ions.

With this series of mixed crystals, therefore, it is possible to vary the conductivity and even its nature without introducing centres, the concentration and kind of charge carriers being given directly by the ratio of the Co and Fe contents (*fig. 9*). By also

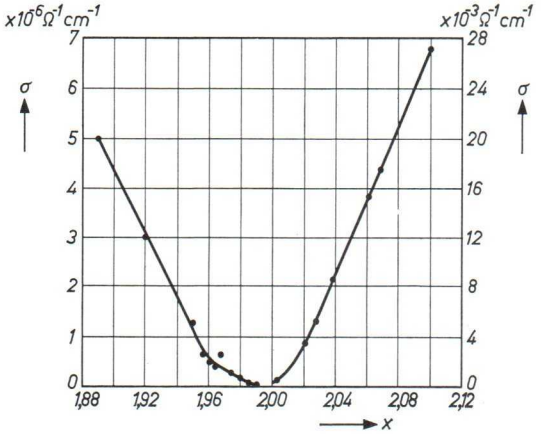


Fig. 9. Conductivity σ at room temperature of a series of mixed crystals of Co_3O_4 and Fe_3O_4 ($Co_{3-x}Fe_xO_4$) in the neighbourhood of the composition $CoFe_2O_4$ ($x=2$). For hole conductivity (excess cobalt: $x < 2$) the left ordinate has to be used; for electron conductivity (excess iron: $x > 2$) the right ordinate.

measuring the conductivity as a function of temperature it was found that for this compound the mobility showed an exponential dependence on temperature. At room temperature the mobilities turned out to be very small, e.g. $\approx 10^{-4}$ cm²/Vs for electrons and $\approx 10^{-8}$ cm²/Vs for holes, their activation energies being about 0.2 and 0.5 eV respectively. By measuring the Seebeck coefficient as a function of charge-carrier concentration (*fig. 10*) the number of available sites for the charge carriers proved to be about 10^{22} per cm³, this being roughly equal to the number of metal ions per cm³. This number is quite different from that to be expected from the

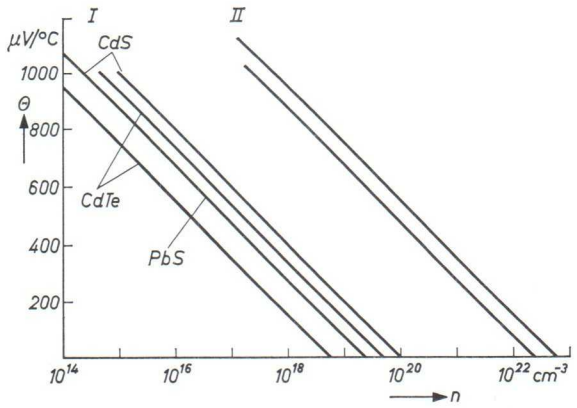


Fig. 10. Schematic survey of the Seebeck effect Θ of a number of compounds as a function of electron or hole concentration n, extrapolated to the intersection points with the abscissa. Group *I* consists of the Ge-type semiconductors CdS [41]), PbS [42]) and CdTe [43]). The two lines of group *II* enclose the measurements on NiO and $CoFe_2O_4$ [46]) [47]).

usual band theory of semiconductors ($\approx 10^{19}$) and in good agreement with the assumption that the charge carriers jump from one metal ion to the next, as should be the case for the localized electron-energy levels. The activation energy Q therefore refers in this case to the mobility and must be considered as being the energy q required to overcome the Landau trapping of the electron, i.e. the polarization of the lattice around the slowly moving charge carrier (see *fig. 11*).

In oxides that have been made conductive by means of a slight deviation from stoichiometry or by introduction of foreign atoms (centres), at low temperature the charge carriers will be trapped near

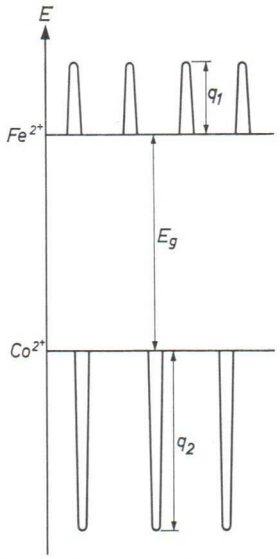


Fig. 11. Energy-level scheme for $Co_{3-x}Fe_xO_4$ ($x \approx 2$). For pure $CoFe_2O_4$ at low temperature the Fe^{2+} level is empty and only Fe^{3+} is present, whereas the Co^{2+} level is full and only Co^{2+} ions are present. q_1 and q_2 are the activation energies needed for jumps of electrons or holes.

these centres and there will be no conduction. However, in contradistinction to the usual band type semiconductor, here the trapped electron because of its localization, together with the deviating charge of the centre may form a permanent dipole, which can reorient itself by local jumps of the trapped carrier between equivalent positions around the centre. Dielectric loss measurements (Volger) yield the relaxation rate $1/\tau$, which is given by:

$$\frac{1}{\tau} = \frac{1}{\tau_0}\, e^{-\mu/kT}.$$

The value of the activation energy, μ, can be found by carrying out loss measurements as a function of temperature. It need not be equal to the activation energy q governing the motion of the dissociated charge carrier. From low temperature measurements on α-Fe_2O_3 with a slight excess of Fe, a value for μ of 0.005 eV was found [48]).

Van Houten [46]), measuring Seebeck coefficients in NiO heavily doped with Li as a function of temperature, was able to calculate from these measurements the energy needed to liberate the holes from their (Li^+Ni^{2+}) centres. It was found to be low: $E_a \approx 0.035$ eV. This seems to be the case for high concentrations in a large number of compounds of this kind. *)

Activators and co-activators in phosphors

We have seen that charge compensation plays an important role in the concept of controlled valency. This same idea of charge compensation was used by Kröger [49]) [50]) [51]) [52]) in elucidating the complicated state of affairs in the ZnS-type phosphors. He was able to demonstrate that the so-called activator atoms Ag, Au and Cu actually were effectively negatively-charged monovalent cations (Ag^+) replacing divalent Zn^{2+} ions (Ag'_{Zn}), their effective negative charge being compensated by Cl^- ions (coactivators) on S^{2-} sites ($Cl^{\cdot}_S$) brought into the crystal from the NaCl flux used as mineralizator. In the symbols dots denote effectively positive centres and the dashes effectively negative centres. If the activators Ag^+ and Cu^+ really occupy Zn sites and not for instance interstitial sites, then the trivalent ions Al^{3+}, Ga^{3+}, In^{3+}, Sc^{3+} substituting Zn^{2+} ions should also be able to act as coactivators, which indeed proved to be true. Kröger, Hoogenstraaten [53]) and Klasens [54]) and later Van Gool [55]) using these notions were able to explain for this kind of phosphor a surprising number of physical effects related to luminescence, including those of decay, trapping, killing, quenching and thermal glow.

General defect chemistry

So far we have encountered two ways of compensating the deviating charge of a foreign ion incorporated in a crystal structure. The first (controlled valency) takes place by a compensating change of the charge of the ions of the host crystal, the second by the simultaneous incorporation of other foreign ions having a valency such that the deviating charges of the foreign ions compensate each other. There is still another possibility. A well-known example is that of the incorporation of divalent Ca^{2+} ions into NaCl crystals (see for instance Haven [56])), where the effective positive charge of the Ca^{2+} ion replacing an Na^+ ion ($Ca^{\cdot}_{Na}$) is compensated by the effective negative charge of a missing Na^+ ion, that is by an effectively negatively charged Na^+-ion vacancy (V'_{Na}). Now Kröger and Vink [41]) [57]) [58]) [59]) [60]) asked themselves how it comes about that sometimes the deviating charge of the foreign ion is compensated by a corresponding change of valency of the ions of the base lattice, whereas in other cases the host lattice creates effectively charged atomic defects (like vacancies or interstitials) to arrive at the same end.

As a consequence of thermodynamical requirements the pure compound contains all kinds of imperfections: free electrons, holes and effectively charged and uncharged atomic imperfections. Starting from ideas developed by Schottky and Wagner, Kröger and Vink were able to work out a general theory showing how the concentrations of the different electronic and atomic imperfections depend on the concentrations of foreign ions in the lattice and on the composition of the gas phase that is in equilibrium with the crystals at the temperature of preparation. In this theory the effectively charged imperfections are of prime importance. In a simple case, for instance pure PbS, one has four of those effectively charged imperfections: free electrons, holes, effectively negatively charged lead vacancies V'_{Pb} and effectively positively charged sulphur vacancies $V^{\cdot}_S$. The concentrations depend on each other via relations similar to those of the usual law of mass action, e.g.

$$n\,p = K_i$$

for the electronic imperfections, with n representing the concentration of the electrons and p that of the holes, and

$$[V'_{Pb}][V^{\cdot}_S] = K'_S$$

for the atomic imperfections. The brackets denote concentrations. The influence of the composition of the gas phase can be expressed by

105

*) *Addition on reprinting (May 1966):*
Recent investigations on NiO semiconductors with low Li-concentrations, however, point to a correspondence between their behaviour and that of narrow-band semiconductors: all the activation energy Q is needed to liberate the charge carriers from the centres, whereas the mobility of the charge is nearly independent of T. The Hall effect, on the other hand, shows a remarkable sign reversal at the Néel point. (A. J. Bosman and C. Crevecoeur, Phys. Rev. 144, 763, 1966.)

$$p_{Pb} = K_r[V_S^{\bullet}]\,n \quad (\text{or } p_{S_2}^{\frac{1}{2}} = K_{ox}[V'_{Pb}]p),$$

with p_{Pb} the partial pressure of the lead atoms in the gas phase that is in equilibrium with the pure PbS and p_{S_2} that of the S_2 molecules.

Electro-neutrality must always be maintained:

$$n + [V'_{Pb}] = p + [V_S^{\bullet}].$$

This nomenclature and also a more thorough investigation of the deeper-lying aspects of this theory have been discussed in a paper by Kröger, Stieltjes and Vink [61]).

The actual calculation of equilibria involves simple but somewhat tedious and complicated mathematics. Brouwer [62]) published a method by means of which the calculations can be substantially simplified. It appears that for large ranges of the partial pressure of the components only two imperfections in the above equation for electro-neutrality have concentrations that need to be taken into account. So for relatively high lead pressures one has

$$n = [V_S^{\bullet}],$$

i.e. Pb excess in the solid and *N*-type conductivity. For relatively high sulphur pressures (low lead pressures) one has however

$$p = [V'_M],$$

which means S excess and *P*-type conductivity. In between there is a region where $n = p$. Now for PbS containing some Bi on Pb sites one has for the condition of electro-neutrality

$$n + [V'_{Pb}] = p + [V_S^{\bullet}] + [Bi_{Pb}^{\bullet}]$$

(trivalent Bi^{3+} ions on Pb^{2+} sites giving rise to effectively positive $Bi_{Pb}^{\bullet}$ centres).

It can now easily be shown that for sufficiently high concentrations of Bi two new simplified forms arise for the condition of electro-neutrality together with their corresponding ranges of lead pressures. For high lead pressures one has

$$n = [Bi_{Pb}^{\bullet}],$$

i.e. the effective charge of the foreign ion is compensated by an electronic defect (change in valency). In other words, for high lead pressure Bi is incorporated according to the principle of controlled valency, with avoidance of lead vacancies. On the other hand, for low lead pressures one has

$$[V'_{Pb}] = [Bi_{Pb}^{\bullet}],$$

i.e. the effective charge of the Bi ion is compensated by an effectively charged atomic imperfection.

Application of defect chemistry in CdS, PbS and SnS

By means of this theory it was possible to explain in a qualitative and quantitative way how in the systems CdS, ZnS, PbS, CdTe and SnS the properties of the compounds depend on the kind and concentration of the impurities introduced into the crystal and on the partial pressures of the components during the preparation. For instance, it was possible to explain why CdS containing Ga on Cd sites and prepared under sulphurizing conditions (low Cd pressures) has a bright red color and is an insulator, whereas with the same amount of Ga but prepared under reducing conditions (high Cd pressures) the yellow colour of pure CdS does not change with the incorporation of Ga, and the substance shows *N*-type conductivity, the charge carriers at room temperature having the same concentration as the Ga atoms [41]).

The explanation amounts to the fact that under "sulphurizing" conditions the Ga^{3+} ions are incorporated into CdS in the same way as the Cd^{2+} ions are incorporated into NaCl, i.e. by the formation of effectively charged atomic imperfections, viz, Cd vacancies. It is those vacancies that are responsible for the red colour. Under reducing conditions, however, the Ga^{3+} ions are incorporated according to the principle of controlled valency, viz, by the formation of an equal amount of Cd^{+} ions, which amounts to free electrons in this energy-band type of semiconductor.

Fig. 12 gives the concentration of electrons (n) and

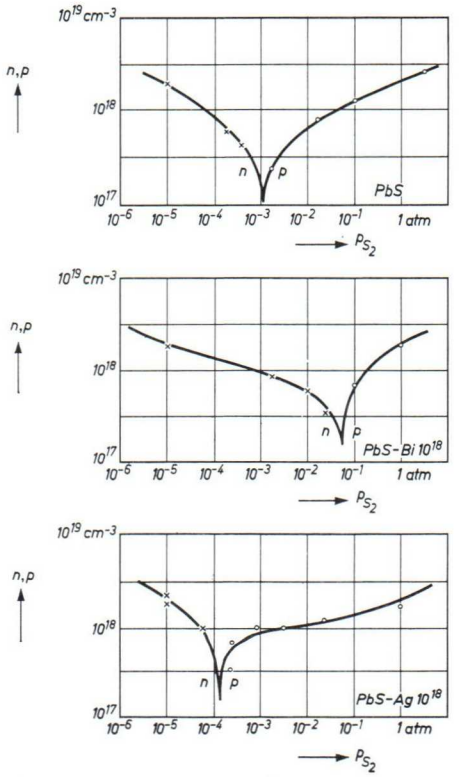


Fig. 12. Concentrations of electrons (n) and holes (p) in undoped PbS crystals (upper graph), in PbS crystals doped with 10^{18} Bi atoms per cm^3 (middle graph) and with 10^{18} Ag atoms per cm^3 (lower graph) as a function of the sulphur pressure p_{S_2} during heating at 1000 °K, as measured after subsequent quenching. Logarithmic scales along both axes.

holes (p) in doped and undoped crystals of PbS, equilibrated under various sulphur pressures at 1000 °K, as measured after quenching [42]). From the figure for pure PbS it can be seen that the plot is nearly symmetrical, showing a sharp minimum in the carrier concentration at a certain sulphur pressure; at higher and at lower sulphur pressures the concentrations increase rapidly, whereas still farther from the minimum the curves flatten and reach a constant slope, corresponding to an electron or hole concentration proportional to $(p_{S_2})^{\mp 1/4}$. The type of carriers is different on both sides of the minimum, the material being N-type in the region of low sulphur pressures and P-type in the region of high pressures. When bismuth ions are present (Bi_{Pb}) it is seen that there is a certain range of pressures where the crystals contain an electron concentration which is about equal to that of the added foreign ions. Similarly, there is a range in which crystals doped with silver (Ag'_{Pb}) contain a concentration of holes about equal to that of the added silver ions. In these ranges the foreign ions are incorporated according to the principle of controlled valency. Bloem [42]) was able to explain all these features and others not mentioned here by the application of the theories mentioned above.

For the pure substance these ideas permit the determination of the limits of the usually very narrow region of stability of solid semiconductors with respect to other phases as a function of the composition of the vapour phase and of the temperature along the three-phase equilibrium: solid semiconductor — vapour phase — other liquid or solid phase. *Fig. 13* gives as an example the limits of stability of solid SnS as a function of temperature, the pressure of the coexisting vapour phase also being given [63]).

Refined molecular engineering in phosphors and related compounds

Van Gool [55]) pointed out that one of the advantages of the considerations of this theory is the fact that the experimenter becomes aware of the large number of possibilities of explaining an often *limited amount* of experimental data, and is thus prevented from jumping to conclusions. For instance, in CdS containing chlorine on sulphur sites the requirement of electro-neutrality may read

$$n + [V'_{Cd}] + 2[V''_{Cd}] + [(Cl_S V_{Cd})'] = p + [V^{\cdot}_S] + 2[V^{\cdot\cdot}_S] + [Cl^{\cdot}_S],$$

leading to quite a number of possible simplified electro-neutrality conditions with their corresponding ranges of pressures. The knowledge of the simplified conditions that occur in reality as a function of the partial pressures of Cl_2 and S_2 gives a great deal of information about the system under consideration.

In this connection Van Gool [64]) gave examples of how to calculate the partial pressures of the relevant components from thermodynamical data; in particular he calculated how, in the system ZnS-H_2S/H_2/Cl_2, as a function of temperature, the partial pressures p_{Cl_2}, p_{S_2} and p_{Zn} depend on the initial pressures p_{H_2S}, p_{H_2} and p_{HCl} or p_{Cl_2} of the gas stream to be equilibrated with ZnS.

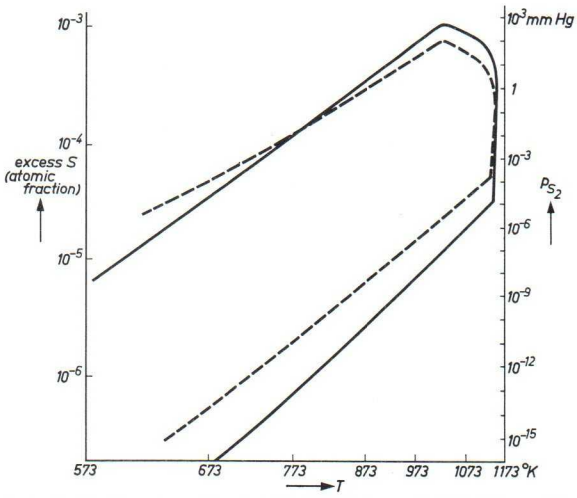


Fig. 13. Limits of stability of solid SnS. The solid curve (left vertical scale) gives the composition of solid SnS at the three-phase equilibrium: solid SnS — vapour phase — other liquid or solid phase, as a function of the temperature T. The composition is given in atom fractions of excess S atoms. The broken curve (right vertical scale) refers to the partial pressure p_{S_2} of S_2 molecules in the coexistent vapour phase. T is plotted reciprocally ($1/T$ scale), the quantities on the vertical scales are plotted logarithmically.

It will be clear from the foregoing that by controlling in this way the composition of the gas phase in equilibrium with the compound at its temperature of preparation, together with a thorough purification and doping of the compound, it is possible, in principle, to perform a very much refined sort of molecular engineering. However, as has been said, the number of possible defects is considerable. It therefore appears likely that although the chemist is nowadays able to prepare all kind of phosphors, photoconductors and other devices of almost any desired property, the explanation of what he is really doing is still lacking in many cases. Only by a much more careful study of the correlation between a *large variety* of physical properties (such as absorption, luminescence and its dependence on temperature and intensity of stimulation, glow curves, photoconductivity, quenching, electron-spin resonance) and the exact chemical composition obtained by this refined molecular engineering, a real insight

into the nature of centres responsible for the different effects might eventually be reached.

Gases and metals

Internal friction; Snoek effect

In 1939 Snoek [65]) discovered one of the causes of internal friction in solids. The solid solution of nitrogen in iron will be taken as an example. At zero stress the nitrogen atoms will be distributed in a completely disordered way over the interstitial sites. On the application of a unidirectional stress some of them will jump into sites in the immediate neighbourhood, if their energy will be lower in these sites. The relaxation time τ for this redistribution is determined by the jump frequency of the nitrogen atoms and, as a consequence, τ obeys the equation

$$\tau = \tau_0 e^{U/RT},$$

where U is the activation energy for the diffusion of nitrogen atoms. Because of this relaxation time, the strain lags behind the stress when a periodically varying stress is applied to the metal. The logarithmic decrement of the mechanical oscillations divided by π is called the loss angle $(1/Q)$. The maximum loss angle will be found for $\omega\tau = 1$, ω being the angular frequency of the imposed mechanical oscillation.

The phenomenon of internal friction has been applied in a wide range of investigations [66]), mainly because of two facts. Firstly, as only jumps over atomic distances are needed to give the inelastic effects, information related with movements over small distances can easily be obtained. For instance, Fast and Verrijp [67]) were able to measure the diffusion constant of nitrogen in iron at room temperature down to — 35 °C, the values lying between 10^{-16} and 10^{-20} cm²/s. Secondly, as the effect is very sensitive and proportional to the concentration of the jumping atoms, very small concentrations can be detected. For instance, concentrations of 0.001% by weight of N in Fe can be detected with an accuracy of 10% [68]).

Application of internal friction in the study of the system Fe-Mn-N

In this laboratory Fast and Verrijp have been studying the systems Fe-C and Fe-N extensively, mainly by means of this effect. Only one example will be given, viz, the interaction of interstitial N with substitutional Mn [69]). *Fig. 14* gives the loss angle as a function of temperature for Fe-N and for Fe-Mn-N with 0.5 and 2% Mn respectively, the N-concentration for the three systems amounting to about 0.03% by weight. One sees that on introduction of Mn new maxima come into existence lying to the left and to the right of the maximum of the pure Fe-N system. The new maximum at the higher temperature is interpreted as being caused by a nitrogen atom jumping around a Mn atom. The free enthalpy of formation of $Mn_{2.5}N$ being larger than that of Fe_4N, the N atom is more tightly bound around a Mn atom, thus causing the corresponding maximum

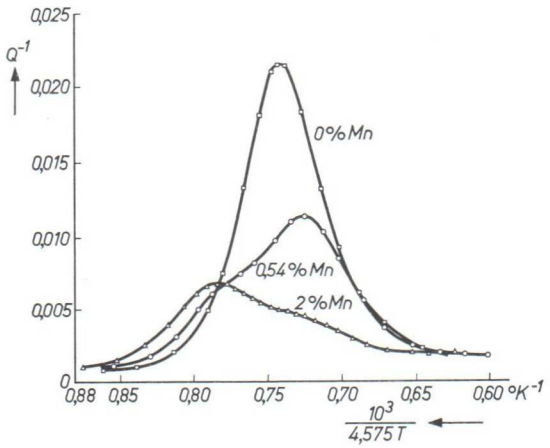


Fig. 14. Internal friction of iron containing 0.03% N with 0%, 0.54% and 2% Mn. The loss angle Q^{-1} is given as a function of the temperature T. The maximum loss angles lie at 22°, 29° and 6° C. Oscillation period 1.2 s.

loss angle to lie at a higher temperature. The new maximum to the left of the original one, occurring at higher Mn contents, is thought to be caused by an N atom jumping around a pair of Mn atoms lying next to each other. It can be made plausible that jumps around a pair of Mn atoms require less activation energy than those around Fe atoms [66]). The bonding of an N atom to such a pair of Mn atoms however is still stronger than that to a single Mn atom. This can be seen in *fig. 15*, where the loss angle is given as a function of temperature after different annealing times at 100 °C. During this annealing the N precipitates out in the form of Fe_4N. One sees that the free N and the N jumping around a single Mn atom disappear more quickly than the N attached to a Mn pair.

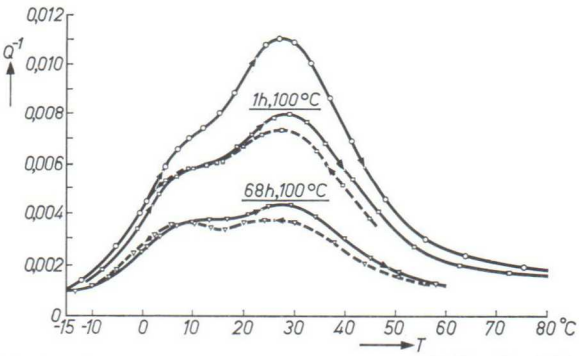


Fig. 15. Internal friction of iron containing 0.03% N and 0.54% Mn. Upper curve: before tempering; other curves: after 1 hour and 68 hours of tempering at 100 °C in order to partially precipitate the nitrogen in the form of Fe_4N.

Hardening by precipitation and internal oxidation

The precipitation of N into very small particles of Fe_4N causes a considerable increase of the hardness of iron. This precipitation hardening, as it is called, not only occurs in the system Fe-N (or Fe-C), but also in a number of other cases. For instance, Ag containing some Mg (1.5%) can be hardened considerably (the Vickers hardness increasing from 40 to 170) by internal oxidation of the Mg. The experimental and theoretical aspect of this internal oxidation has been given a great deal of consideration by Meijering and Druyvesteyn [70] [71] [72]. The conditions for internal oxidation are evident:

1) The oxygen must diffuse into the metal quicker than the foreign metal diffuses out. This means that one must have $C_O D_O > C_{Me} D_{Me}$, C_O being the surface concentration of oxygen atoms in equilibrium with the O_2 pressure or with an external oxide layer (e.g. MeO), C_{Me} the bulk concentration of the foreign metal and D_O and D_{Me} the respective diffusion constants.
2) The affinity of the dissolved metal for oxygen must be greater than that of the base metal.

With regard to the first point it can be said that in most metals D_O of the interstitial oxygen atoms will be larger than D_{Me} of the substitutional metal atoms but that in a limited number of metals only, e.g. Ag, Cu, Ni, Fe, Ti and Zr, the solubility of oxygen is large enough to obtain $C_O D_O > C_{Me} D_{Me}$. The second point, however, cannot be fulfilled by Ti and Zr. Therefore only Ag, Cu, Ni and Fe can be subjected to internal oxidation. This process was subjected to a rigorous mathematical treatment by Meijering and Druyvesteyn, one of their results being that to a good approximation the concentration of the oxide molecules for a dissolved divalent metal Me is given by

$$C_{ox} = C_{Me}(1 + C_{Me} D_{Me} / C_O D_O).$$

The concentration of oxide molecules is therefore somewhat larger than that of the original foreign metal. In a metal that has been subjected to a complete internal oxidation this cannot be the case throughout the volume. It is to be expected therefore that in a strip of metal being oxidized ("internally") from both sides, somewhere in the middle there must exist a region showing almost no oxidation. This region can be made visible by etching and can be seen in *fig. 16*. This "internal reaction" can be used as an extremely sensitive tool to see whether atoms like O and F dissolve and diffuse at all in certain metals. Use is thereby made of the fact that by etching, the transition between the "oxidized" part of the volume and that part of the metal that has not yet been reached by the diffusant can be made visible, as we have seen already.

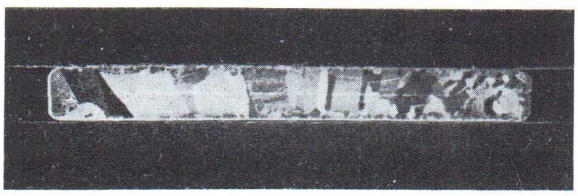


Fig. 16. Ag with 1.5 weight % Mn, 48 h on 800 °C in air; etched $NH_4OH + H_2O$. Magnification 10×. Although oxidation is completed, a "middle-line", poor in oxide, persists.

This was used by Meijering [73] to show that fluorine can be dissolved in and diffuses through silver. By this method it could also be demonstrated that oxygen dissolves in iron [74], although its concentration is so small that even by means of the method of internal friction its presence could not be detected.

Controlled precipitation in ferromagnetic alloys

Precipitation also plays a predominant role in making ferromagnetic alloys with high $(BH)_{max}$ values, because by this method, in principle, particles can be prepared small enough to contain only one single domain separated by non-magnetic material. In ferromagnetic alloys of a special composition the particles may be precipitated in the form of needles, thus minimizing the demagnetization. By carrying out this precipitation in a single crystal under the application of a static magnetic field, all needles and their magnetic moments can be oriented. Thus a material with a very high value of $(BH)_{max}$ can be obtained (Jonas, Meerkamp van Embden, Luteijn, De Vos, Fast and De Jong [75] [76] [77]). With alloys consisting of 35% Fe, 34% Co, 15% Ni, 7% Al, 4% Cu and 5% Ti, from which needles rich in Fe and Co are precipitated in a matrix rich in Ni and Al, values of $(BH)_{max} = 12 \times 10^6$ gauss oersted have been obtained. The orientation of these precipitates can be clearly seen by using the electron microscope (De Jong, Smeets and Haanstra [78]). *Fig. 17* shows two such electron-microscope photographs.

Preparation of pure metals by means of decomposition of gaseous compounds

We will now enter the field of interaction between gases and metals. This is another area of research that has been investigated extensively by the Philips Laboratories. The research started in 1923 when Van Arkel [79] [80] published a method to prepare pure and single-crystal tungsten. In this method the tungsten is deposited on a hot wire at 1600 °C by dissociation of WCl_6. The Cl_2

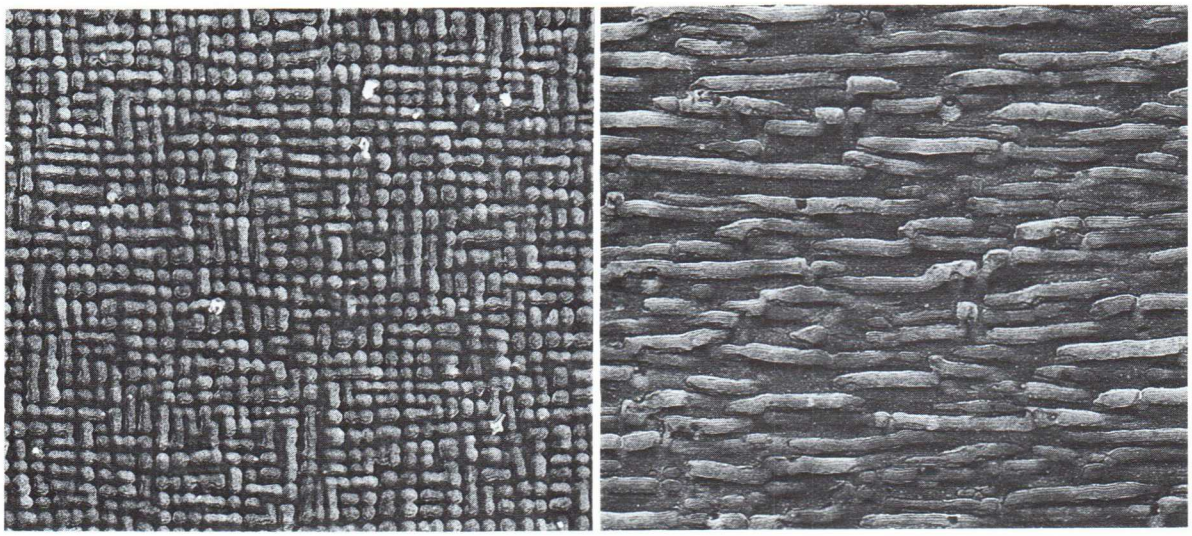


Fig. 17. *a*) Electron-microscope photograph of "Ticonal" with $(BH)_{max} = 11 \times 10^6$ gauss. oersted. The regular arrangement of the needles of the precipitates is shown in a plane perpendicular to their length. Magnification 70 000×.
b) The same as (*a*), but now the arrangement of the needles is shown in a plane parallel to their length.

liberated by this dissociation then reacts with W powder in another part of the reaction tube kept at 400 °C. The reaction product WCl_6 again dissociates at the hot wire and so it is understandable that by means of a relatively small amount of chlorine a large quantity of W powder can be converted into a thick wire of single crystals of W of extremely high purity.

In the last few years this method under the name of vapour-phase transport reactions has drawn new attention, as it often proves to be an excellent way of preparing pure single crystals or epitaxial layers of compounds.

Shortly after Van Arkel's experiments De Boer and Fast showed that this method could also be applied to the metals Ti, Zr and Hf [81] [82] [83] [84] [85] by using the tetra-iodides instead of the chlorides. Up to then these metals had been considered to be very brittle by nature. It appeared, however, that when prepared by the method just mentioned these metals were all very ductile indeed, so much so that it was first thought that perhaps they were other modifications. Very soon however it appeared that the brittleness of these metals is caused by the presence of small concentrations of oxygen, nitrogen or hydrogen.

In Zr powder for instance every grain of metal is covered by a thin layer of oxide. On pressing and sintering the powder this oxide dissolves in the metal causing it to become brittle. Ti, Zr and Hf can dissolve very high concentrations of these gases. Zr, for instance, may contain up to 30 at.% of oxygen [86] and 20 at.% of nitrogen [87], while Ti and Hf also can contain high concentrations of oxygen and nitrogen.

Permeation of gases through metals

It was further found that the dissolved atoms of these gases are present at the interstitial sites of the metal crystal and have a relatively high mobility. It is, however, impossible, once they have been dissolved in the metals (Ti, Zr or Hf), to drive these gases out by mere heating. Therefore although their solubility is high and their absorption is rapid, the permeation of oxygen and nitrogen through Ti, Zr or Hf is impossible because for that process also a third step, that of extraction, must be easy. This permeation of gases through metals is another aspect of the interaction between gases and metals that was carefully studied in these laboratories [88].

For instance, the case of hydrogen and iron [89] is in some way opposite to that of oxygen and zirconium. As De Boer and Fast showed, hydrogen atoms move very easily interstitially through iron and it is also easy for them to leave the metal and recombine to H_2 molecules. On the other hand it is only possible to introduce H atoms into the iron at room temperature when by some other means the H_2 molecules are first dissociated into atoms. This can be done either by dissociation at a hot W wire or in an aqueous solution with a *p*H value low enough for H atoms to be formed at the Fe interface.

Owing to these and other effects the permeation of gases through metals proved to be a complicated

process consisting of at least five successive stages, viz, absorption, adsorption, diffusion, de-absorption and de-adsorption, the theoretical aspects of which were carefully analysed.

Investigation of getters; the system Th-Al-Ce

As has already been indicated, Zr, that at high temperatures can absorb huge quantities of gases, is protected at low temperatures by a very thin layer of oxide. This property makes it useful as a getter. At low temperatures it can be mounted in the tube to be gettered and later on at any moment it can be heated to absorb gases. It goes without saying that "gettering" in general has been a subject of investigation. For instance, a great deal of research work has been carried out in order to understand the getter action of the Ceto getter. This getter contains mainly Th, Al and Ce. To understand its action Van Vucht made a study of the binary systems Th-Al, Al-Ce and Ce-Th and the ternary system Th-Al-Ce, using X-ray analysis and thermoanalytical and metallographical methods. It was found that pure Ceto has the structure of the intermetallic compound Th_2Al in which about one quarter of the Th atoms is replaced by Ce [90]).

Absorption of hydrogen in Th_2Al

To examine the action of this type of getter in more detail, the absorption of H_2 by the well-defined compound Th_2Al was carefully analysed by means of the determination of hydrogen-absorption isotherms at different temperatures, X-ray diffraction of Th_2Al containing hydrogen, neutron diffraction of deuterium solutions in Th_2Al and proton magnetic resonance [91]). It was found that the unit cell of Th_8Al_4 apart from other holes, contains 16 equivalent holes tetrahedrally surrounded by Th atoms arranged in 8 pairs of twin holes, having one tetrahedral plane in common, namely the plane parallel to the base of the cell (*fig. 18*). This lattice can be filled with H atoms up to the composition of $Th_8Al_4H_{15.4}$. It could be shown that the hydrogen atoms are all situated in the holes mentioned above.

Measurements of the hydrogen equilibrium pressures at various temperatures yielded isotherms which (around 225 °C) showed three points of inflexion (*fig. 19*). Two of these near the compositions $Th_8Al_4H_4$ and $Th_8Al_4H_{12}$ were considered to be remnants of plateaux, indicating two-phase regions at lower temperatures. This would imply that $Th_8Al_4H_8$ had to be considered as an intermediate hydride. X-ray diffraction showed that at room temperature a phase separation did in fact occur, the compositions of the two phases being about Th_8Al_4 and $Th_8Al_4H_5$. However, even at 83 °K a similar region was not found between $Th_8Al_4H_8$ and $Th_8Al_4H_{15.4}$. The changes in the cell dimensions with increasing hydrogen content seemed to indicate that at $Th_8Al_4H_8$ the hydrogen atoms had a strong tendency to order.

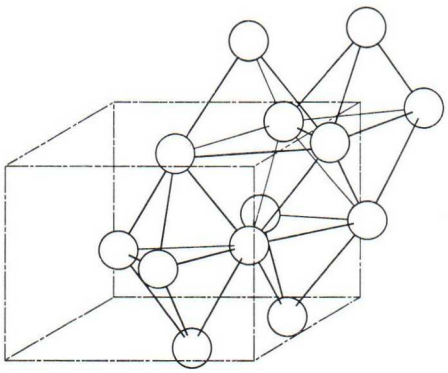


Fig. 18. Structure of Th_2Al (only Th atoms are shown). Arrangement of tetrahedral holes as "double holes".

Thermodynamics

The results obtained so far by Van Vucht could be further corroborated by a more theoretical analysis of the hydrogen-absorption isotherms [92]). A general expression for the Gibbs free energy could be set up in which the enthalpy contained a term corresponding to the energy of a single hydrogen atom in a "double hole" and a term corresponding to the energy of a pair of hydrogen atoms in a double hole. If ϑ is the fraction of holes occupied by a proton and q the fraction of holes that are occupied and

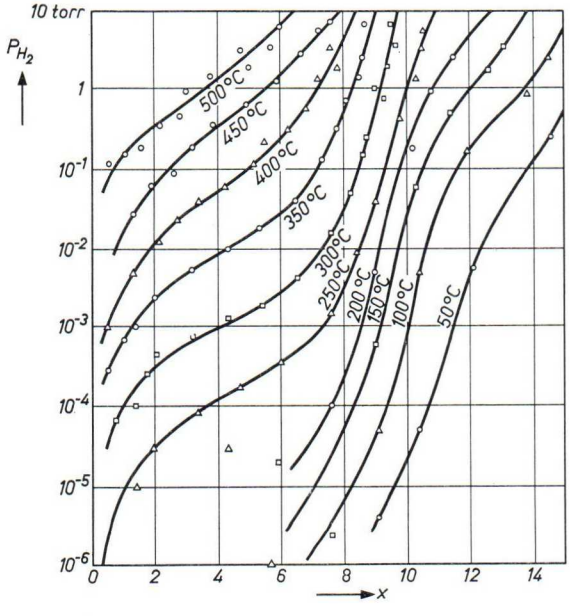


Fig. 19. Isotherms of the hydrogen equilibrium pressure p_{H_2} above $Th_8Al_4H_x$. This pressure is plotted logarithmically against x.

associated with a vacant twin, then at equilibrium

$$\frac{\partial G}{\partial q} = 0 \quad \text{and} \quad \frac{dG}{d\vartheta} = \tfrac{1}{2}\mu_{H_2} = \tfrac{1}{2}\mu^0_{H_2} + \tfrac{1}{2}RT \ln p_{H_2}.$$

From these conditions a general expression for the hydrogen-absorption isotherms as a function of hydrogen concentration could be obtained. Comparing these with the experimental isotherms it could be safely concluded that a strong tendency exists for the hydrogen atoms to fill the "double holes" with one atom only, which is in agreement with the indications obtained from the changes of the cell dimensions as a function of hydrogen content.

Retrograde solidus curves

The power of this method of pure thermodynamical reasoning has been often demonstrated by Meijering [93]). It was possible to derive an expression enabling one to predict (e.g. from eutectic data) whether in a T,x projection of a solidus-liquidus equilibrium of a binary system, the solidus would show a retrograde character or not [93]) [94]). The principle is very simple indeed. In a binary system two phases (solid and liquid, with compositions x_S and x_L atom fractions) are in equilibrium when at a certain temperature

$$\frac{dG_S}{dx_S} = (G_S - G_L)/(x_S - x_L) = \frac{dG_L}{dx_L}.$$

Now for coexistence of the two phases this must be valid also on changing T. Remembering that $\partial G/\partial T = -S$ and $G = H - TS$, the equation of Van der Waals for coexistent phases is obtained:

$$\frac{dx_S}{dT} = \frac{dS_S/dx_S - (S_S - S_L)/(x_S - x_L)}{d^2G_S/dx_S^2}.$$

The denominator being always positive, retrograde solidus lines $dx_S/dT > 0$ will be found for positive numerators. Taking for the entropy only Gibbs's entropy of mixing one gets for S_S and S_L the curves given in *fig. 20*. It was thus possible with this to explain many retrograde solidus curves and to predict others. For instance, from fig. 19 it could be predicted that the number of retrograde solidus curves would increase enormously with the application of refined techniques of determination of small solid solubilities, a prediction that in the course of the years for the cases of Ge and Si has been amply verified.

Regular solutions; the system Ni-Cu-Cr

A powerful tool used in those thermodynamical considerations [95]) [96]) [97]) [98]) was the concept of regular solutions. For binary regular solutions the Gibbs free energy can be written:

$$G = ax(1-x) + RT\ \{x \ln x + (1-x)\ln(1-x)\}.$$

i.e. adding a simple extra enthalpy term to the entropy of mixing. For a ternary regular system one gets:

$$G = axy + bxz + cyz + RT(x \ln x + y \ln y + z \ln z),$$

with x, y and z the atom fractions of the three components. The use of this relatively simple function was demonstrated for instance for the system Cu-Ni-Cr. From the known binary systems between body-centred cubic Cr and face-centred cubic Ni and Cu the ternary diagram of *fig. 21a* could be expected, i.e. a small body-centred region in the Cr corner, a face-centred region along the Cu-Ni side and a two-phase region in between. However, metallographical and X-ray investigations [99]) [100]) revealed the existence of an extended region of a three-

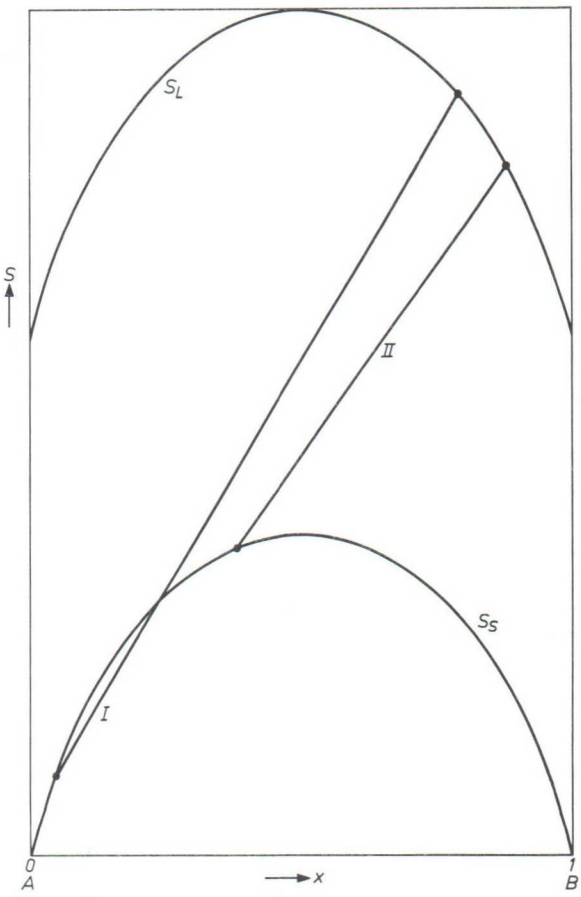


Fig. 20. The lower and upper curves give the entropies S_S and S_L of the solid and the liquid phase respectively as a function of the atom fraction x of B in the binary system A-B. Gibbs's expression for the entropy of mixing is used in both cases, and $1.1R$ is taken for the entropy of melting per mole. On applying Van der Waals's equation for coexistent phases it is seen that in example I one has $dx/dT > 0$, i.e. a retrograde solidus curve, whereas in example II this not the case.

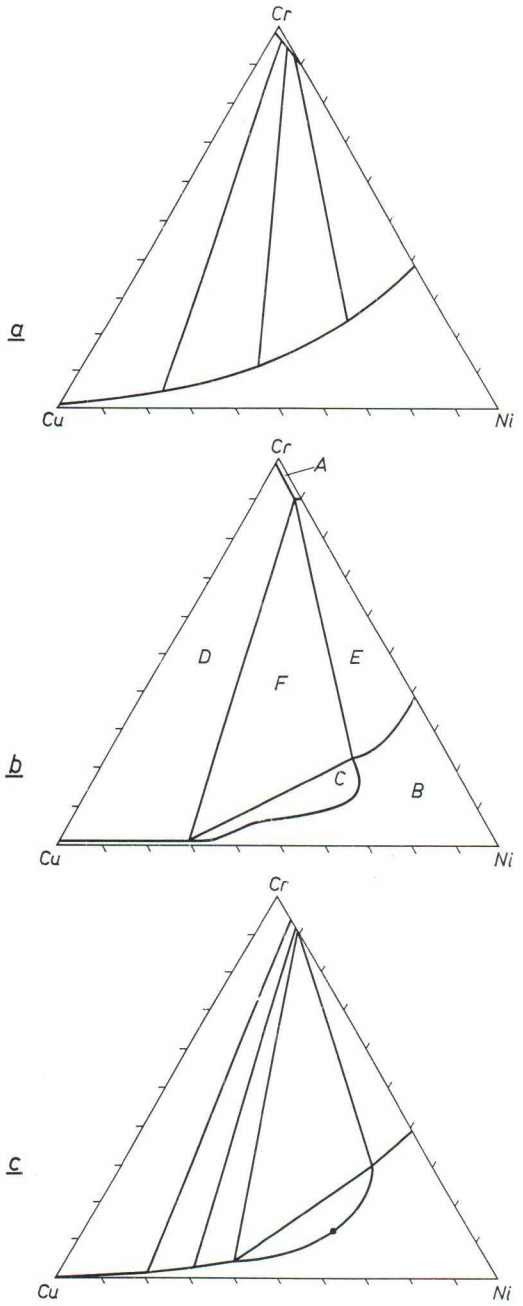


Fig. 21. *a*) Phase diagram of Cu-Ni-Cr at 930 °C obtained by graphical interpolation of the solubilities in the binary border systems.
b) Real phase diagram of the system Cu-Ni-Cr at 930 °C as determined experimentally.
A: one phase; body-centred.
B: one phase; face-centred.
C: two phases; face-centred.
D: two phases; body-centred and Cu-rich face-centred.
E: two phases; body-centred and Ni-rich face-centred.
F: three phases.
c) Phase diagram of the system Cu-Ni-Cr as calculated for 1200 °K. Between the Cr-rich and Cu-rich phases two connodes have been drawn. The dot denotes the critical point where the Cu-rich and the Ni-rich phases become identical.

phase equilibrium: a Cr-rich body-centred phase and *two* face-centred phases (fig. 21*b*). It now proved possible to calculate the values of a, b and c from the binary systems Ni-Cr, Cu-Ni and Cu-Cr respectively. Taking into account also the difference in free enthalpy between body-centred and face-centred Cr, it was possible, using the above expression for G, to calculate a phase diagram (1200 °K) resembling in many aspects the experimental one (fig. 21*c*).

Only a few items of the application of thermodynamics have been given in this relatively short section. It will be clear that in all areas of chemistry reviewed in this paper the use of thermodynamics is indispensable. For instance the work of Kröger and Vink and that of these authors together with Stieltjes mentioned in the section on internal charge compensation is in essence the application of thermodynamics to the defect chemistry of homogeneous and heterogeneous phase equilibria. The same can be said of the calculations by Van Gool also dealt with in that section.

Conclusion

In this paper a limited number of topics of research related to solid-state chemistry have been briefly dealt with. It is hoped that an impression has been given of the chemical climate in this laboratory. One important point, however, has not yet been stressed and certainly must not be overlooked. Many workers in this laboratory have been mentioned in this paper. It must be emphasized that they could not have done what they did without the continuous help, in the form of papers, lectures and discussions, from scientists in other laboratories. Only a very few of them have been mentioned. Every research worker knows, however, that progress is possible only by means of close cooperation between scientists all over the world.

BIBLIOGRAPHY

1) E. J. W. Verwey, Z. Kristallogr. **91,** 65, 1935.
2) E. J. W. Verwey, J. chem. Phys. **3,** 592, 1935.
3) E. J. W. Verwey and M. G. van Bruggen, Z. Kristallogr. **92,** 136, 1935.
4) A. E. van Arkel, E. J. W. Verwey and M. G. van Bruggen, Rec. Trav. chim. Pays-Bas **55,** 331, 1936.
5) E. J. W. Verwey, A. E. van Arkel and M. G. van Bruggen, Rec. Trav. chim. Pays-Bas **55,** 340, 1936.
6) J. L. Snoek, New developments in ferromagnetic materials, Elsevier, Amsterdam 1947.
7) E. J. W. Verwey and J. H. de Boer, Rec. Trav. chim. Pays-Bas **55,** 531, 1936.
8) J. H. de Boer and E. J. W. Verwey, Proc. Phys. Soc. **49,** extra part, 59, 1937.
9) E. W. Gorter, Nature **165,** 798, 1950.
10) F. C. Romeijn, Philips Res. Repts **8,** 304, 1953.
11) F. C. Romeijn, Philips Res. Repts **8,** 321, 1953.
12) E. W. Gorter, Philips Res. Repts **9,** 295, 1954.
13) E. W. Gorter, Philips Res. Repts **9,** 321, 1954.

[14]) E. W. Gorter, Philips Res. Repts **9,** 403, 1954.
[15]) G. Blasse, Philips Res. Repts, to be published.
[16]) E. J. W. Verwey and E. L. Heilmann, J. chem. Phys. **15,** 174, 1947.
[17]) E. J. W. Verwey, F. de Boer and J. H. van Santen, J. chem. Phys. **16,** 1091, 1948.
[18]) F. de Boer, J. H. van Santen and E. J. W. Verwey, J. chem. Phys. **18,** 1032, 1950.
[19]) E. J. W. Verwey and P. W. Haaijman, Physica **8,** 979, 1941.
[20]) P. B. Braun, Nature **170,** 1123, 1952.
[21]) G. W. van Oosterhout and C. J. M. Rooijmans, Nature **181,** 44, 1958.
[22]) C. J. M. Rooijmans, J. inorg. nucl. Chem. **11,** 78, 1959.
[23]) J. H. van Santen, Philips Res. Repts **5,** 282, 1950.
[24]) J. H. van Santen and J. S. van Wieringen, Rec. Trav. chim. Pays-Bas **71,** 420, 1952.
[25]) J. L. Snoek, Physica **8,** 426, 1941.
[26]) E. J. W. Verwey, P. B. Braun, E. W. Gorter, F. C. Romeijn and J. H. van Santen, Z. phys. Chem. **198,** 6, 1951.
[27]) J. J. Went, G. W. Rathenau, E. W. Gorter and G. W. van Oosterhout, Philips tech. Rev. **13,** 194, 1951/52.
[28]) H. P. J. Wijn, Nature **170,** 707, 1952.
[29]) P. B. Braun, Nature **170,** 708, 1952.
[30]) G. H. Jonker, H. P. J. Wijn and P. B. Braun, Philips tech. Rev. **18,** 145, 1956/57.
[31]) P. B. Braun, Philips Res. Repts **12,** 491, 1957.
[32]) G. H. Jonker, H. P. J. Wijn and P. B. Braun, Proc. Instn. Electr. Engrs **104 B,** Suppl. No. 5, 249, 1957.
[33]) G. H. Jonker, XVIe Congrès int. Chim. pure et appl., Paris 1957, section de chimie minérale, p. 117, Birkhäuser, Basel 1958.
[34]) J. Smit and H. P. J. Wijn, Ferrites, Philips' Technical Library, Eindhoven 1959.
[35]) E. W. Gorter, Proc. Instn. Electr. Engrs **104 B,** Suppl. No. 5, 255, 1957.
[36]) J. A. Lely, Ber. Dtsch. Keram. Ges. **32,** 229, 1955.
[37]) W. F. Knippenberg, Philips Res. Repts, to be published.
[38]) F. K. Lotgering, J. inorg. nucl. Chem. **9,** 113, 1959; **16,** 100, 1960/61.
[39]) E. J. W. Verwey, P. W. Haaijman, F. C. Romeijn and G. W. van Oosterhout, Philips Res. Repts **5,** 173, 1950.
[40]) E. J. W. Verwey, in: Semi-conducting materials, (Proc. Conf. Univ. Reading), p. 151, Butterworths, London 1951.
[41]) F. A. Kröger, H. J. Vink and J. van den Boomgaard, Z. phys. Chem. **203,** 1, 1954.
[42]) J. Bloem, Philips Res. Repts **11,** 273, 1956.
[43]) D. de Nobel, Philips Res. Repts **14,** 361, 1959.
[44]) D. de Nobel, Philips Res. Repts **14,** 430, 1959.
[45]) G. H. Jonker, Proc. int. Conf. on semiconductor physics, Prague 1960, p. 864, Academic Press, New York 1961.
[46]) G. H. Jonker and S. van Houten, Halbleiterprobleme **6,** 118, 1961 (Vieweg, Braunschweig).
[47]) G. H. Jonker, Phys. Chem. Solids **9,** 165, 1959.
[48]) J. Volger, Disc. Faraday Soc. No. 23, 63, 1957.
[49]) F. A. Kröger and J. E. Hellingman, J. Electrochem. Soc. **93,** 156, 1948.
[50]) F. A. Kröger, J. Opt. Soc. Amer. **39,** 670, 1949.
[51]) F. A. Kröger and J. E. Hellingman, J. Electrochem. Soc. **95,** 68, 1949.
[52]) F. A. Kröger and J. Dikhoff, Physica **16,** 297, 1950.
[53]) W. Hoogenstraaten, J. Electrochem. Soc. **100,** 356, 1953.
[54]) H. A. Klasens, J. Electrochem. Soc. **100,** 72, 1953.
[55]) W. van Gool, Philips Res. Repts Suppl. 1961, No. 3.
[56]) Y. Haven, Rep. Conf. Defects in crystalline solids, Bristol 1954, p. 261, Phys. Soc. London 1955.
[57]) F. A. Kröger and H. J. Vink, Physica **20,** 950, 1954.
[58]) F. A. Kröger and H. J. Vink, in: Halbleiter und Phosphore (Semiconductors and phosphors), Proc. int. Colloq. Garmisch-Partenkirchen 1956, p. 17, Vieweg, Braunschweig 1958.
[59]) F. A. Kröger and H. J. Vink, Solid State Physics **3,** 307, 1956.
[60]) F. A. Kröger and H. J. Vink, Phys. Chem. Solids **5,** 208, 1958.
[61]) F. A. Kröger, F. H. Stieltjes and H. J. Vink, Philips Res. Repts **14,** 557, 1959.
[62]) G. Brouwer, Philips Res. Repts **9,** 366, 1954.
[63]) W. Albers and H. J. Vink, Chem. Weekblad, to be published.
[64]) W. van Gool, to be published.
[65]) J. L. Snoek, Physica **6,** 591, 1939.
[66]) J. D. Fast, Métaux, Corr., Industr. **36,** 383, 431, 1961.
[67]) J. D. Fast and M. B. Verrijp, J. Iron Steel Inst. **176,** 24, 1954.
[68]) J. D. Fast and M. B. Verrijp, J. Iron Steel Inst. **180,** 337, 1955.
[69]) J. D. Fast, J. L. Meijering and M. B. Verrijp, Métaux, Corr., Industr. **36,** 112, 1961.
[70]) J. L. Meijering and M. J. Druyvesteyn, Philips Res. Repts **2,** 81, 1947.
[71]) J. L. Meijering and M. J. Druyvesteyn, Philips Res. Repts **2,** 260, 1947.
[72]) J. L. Meijering, Trans. Metallurg. Soc. AIME **218,** 968, 1960.
[73]) J. L. Meijering, Rev. Métall. **54,** 520, 1957.
[74]) J. L. Meijering, Acta metallurgica **3,** 157, 1955.
[75]) B. Jonas and H. J. Meerkamp van Embden, Philips tech. Rev. **6,** 8, 1941.
[76]) A. I. Luteijn and K. J. de Vos, Philips Res. Repts **11,** 489, 1956.
[77]) J. D. Fast and J. J. de Jong, J. Phys. Radium **20,** 371, 1959.
[78]) J. J. de Jong, J. M. G. Smeets and H. B. Haanstra, J. appl. Phys. **29,** 297, 1958.
[79]) A. E. van Arkel, Physica **3,** 76, 1923.
[80]) A. E. van Arkel and J. H. de Boer, Z. anorg. allg. Chem. **148,** 345, 1925.
[81]) J. H. de Boer and J. D. Fast, Z. anorg. allg. Chem. **153,** 1, 1926.
[82]) J. H. de Boer and J. D. Fast, Z. anorg. allg. Chem. **187,** 177, 1930.
[83]) J. H. de Boer and J. D. Fast, Z. anorg. allg. Chem. **187,** 193, 1930.
[84]) J. D. Fast, Z. anorg. allg. Chem. **239,** 145, 1938.
[85]) J. D. Fast, Z. anorg. allg. Chem. **241,** 42, 1939.
[86]) J. H. de Boer and J. D. Fast, Rec. Trav. chim. Pays-Bas **55,** 459, 1936.
[87]) J. D. Fast, Metallwirtschaft **17,** 641, 1938.
[88]) J. D. Fast, Chem. Weekblad **38,** 2, 19, 1941.
[89]) J. H. de Boer and J. D. Fast, Rec. Trav. chim. Pays-Bas **58,** 984, 1939.
[90]) J. H. N. van Vucht, Philips Res. Repts **16,** 1, 1961.
[91]) J. H. N. van Vucht, Philips Res. Repts **18,** 1, 21, 35, 1963 (No. 1).
[92]) J. H. N. van Vucht, Philips Res. Repts **18,** 53, 1963 (No. 1).
[93]) J. L. Meijering, The physical chemistry of metallic solutions and intermetallic compounds, Proc. Symp. Nat. Phys. Lab. 1958, Vol. 2, paper No. 5A.
[94]) J. L. Meijering, Philips Res. Repts **3,** 281, 1948.
[95]) J. L. Meijering, Philips Res. Repts **5,** 333, 1950.
[96]) J. L. Meijering, Philips Res. Repts **6,** 183, 1951.
[97]) J. L. Meijering and H. K. Hardy, Acta metallurgica **4,** 249, 1956.
[98]) J. L. Meijering, Acta metallurgica **5,** 257, 1957.
[99]) J. L. Meijering, G. W. Rathenau, M. G. van der Steeg and P. B. Braun, J. Inst. Metals **84,** 118, 1955/56.
[100]) J. L. Meijering, Chem. Weekblad **51,** 438, 1955.

Summary. A review is given of contributions of the Philips Research Laboratories to solid-state chemistry in the past four decades. Four areas are briefly reviewed, viz, crystal chemistry, internal charge compensation, gases and metals, and thermodynamics. Stress is laid on the fact that whereas the physical and chemical properties of a substance depend on its chemical composition, the exact chemical composition can often be determined only by making use of these physical and chemical properties. This interplay of physical and chemical methods, leading eventually to an understanding of the interrelation between physical and chemical properties and exact chemical composition, is illustrated by means of a number of diverse examples.

B. REPRESENTATIVE PAPERS

		page
1	**Construction and applications of a new design of the Philips vacuum gauge** F.M.Penning and K.Nienhuis (Vol.**11**, p.116)	117
2	**The manufacture of virus vaccine against influenza** A.J.Klein and E.Hertzberger (Vol.**12**, p.273)	125
3	**Delta modulation, a new modulation system for telecommunication** J.F.Schouten, F.de Jager and J.A.Greefkes (Vol.**13**, p.237)	135
4	**Electronic intensification of fluoroscopic images** M.C.Teves and T.Tol (Vol.**14**, p.33)	145
5	**The ferroxcube transformer core used in the Brookhaven Cosmotron** F.G.Brockman and M.W.Louwerse (Vol.**15**, p.73)	157
6	**Ferroxdure II and III, anisotropic permanent magnet materials** A.L.Stuijts, G.W.Rathenau and G.H.Weber (Vol.**16**, p.141)	169
7	**Stereo reverberation** R.Vermeulen (Vol.**17**, p.258)	176
8	**The gyrator, an electric network element** B.D.H.Tellegen (Vol.**18**, p.120)	186

9 **The projection of colour-television pictures** 192
T.Poorter and F.W. de Vrijer
(Vol.**19**, p.338)

10 **The Philips hot-gas engine with rhombic drive mechanism** 211
R.J.Meijer
(Vol.**20**, p.245)

11 **Motion-picture projection with a pulsed light source** 229
P.Hoekstra and C.Meyer
(Vol.**21**, p.73)

12 **Waveguide equipment for 2 mm microwaves** 239
C.W.van Es, M.Gevers and F.C.de Ronde
(Vol.**22**, p.113 and p.181)

13 **The orthocyclic method of coil winding** 261
W.L.L.Lenders
(Vol.**23**, p.365)

14 **The alloy-diffusion technique for manufacturing high-frequency transistors** 277
P.J.W.Jochems
(Vol.**24**, p.231)

15 **The "Plumbicon", a new television camera tube** 287
E.F.de Haan, A.van der Drift and P.P.M.Schampers
(Vol.**25**, p.133)

16 **A gas refrigerating machine for temperatures down to 20 °K and lower** 307
G.Prast
(Vol.**26**, p.1)

17 **Osmium dispenser cathodes** 319
P.Zalm and A.J.A.van Stratum
(Vol.**27**, p.69)

CONSTRUCTION AND APPLICATIONS OF A NEW DESIGN OF THE PHILIPS VACUUM GAUGE

by F. M. PENNING and K. NIENHUIS

531.788.732

The Philips vacuum gauge, with which the pressure of a gas is measured by means of the intensity of a gas discharge in a magnetic field, was introduced 12 years ago and has now found application in many fields and proved to be a most useful instrument. Meanwhile the need has also been felt for such a simple vacuum gauge for measuring still lower gas pressures. This need has been met by changing the construction of the present gauge. In its new design this instrument is also quite useful as leak detector.

A description of the Philips vacuum gauge was given in this journal 12 years ago [1]). The tube used in this gauge contains two parallel plates of zirconium as cathode and a ring as anode (*fig. 1a*) and is placed between the poles of a permanent magnet with its field perpendicular to the cathode. Owing to the peculiar course of the electrical and magnetic lines of force inside the tube the electrons travel to and fro an immense number of times before reaching the anode, thus making many more collisions with gas molecules than would be the case if they travelled directly from the cathode to the anode. This large number of collisions can be interpreted as being derived from an apparent elevation of the gas pressure.

Consequently it is possible for a gas discharge to take place between cold electrodes at a pressure about 1000 times lower than is possible in a tube of a similar construction without magnetic field.

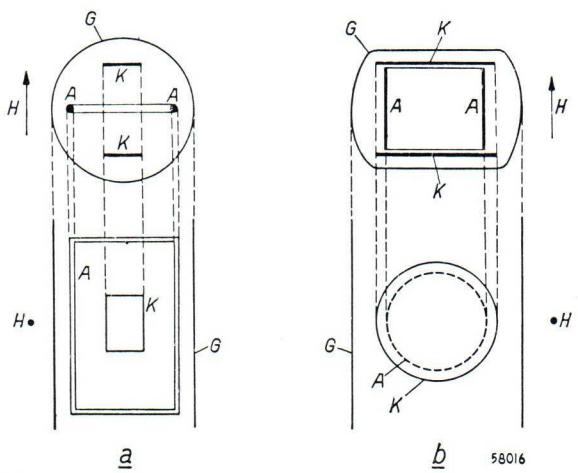


Fig. 1. Diagrammatic representation of the position of the cathode K and the anode A of the Philips vacuum gauge. G represents the glass envelope and H the direction of the magnetic field. *a*) The older design, *b*) the new design.

[1]) F. M. Penning, High-vacuum gauges, Philips Techn. Rev. **2**, 201-208, 1937.

It has been found experimentally that the magnitude of the discharge current is a good measure for the pressure.

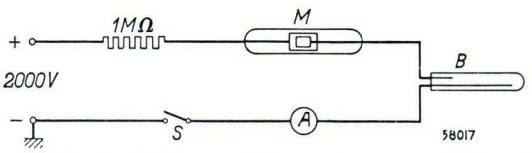


Fig. 2. The circuit diagram of the vacuum gauge M in series with the glow lamp B, acting as current indicator, and a micro-ammeter A. This circuit is applied both in the old design and in the new one.

This vacuum gauge, the circuit diagram of which is given in *fig. 2*, has in recent years been extensively applied in the manufacture of electronic valves and in apparatus in which the vacuum has to be maintained by means of a pump, as is the case with two kinds of apparatus previously described in this journal: the neutron generator [2]) and the electron microscope [3]). It can also be used, as described in literature on the subject, in combination with a "Pirani" manometer [4]) or with a thermocouple manometer [5]), in which case a larger pressure range is covered. Further, this instrument is very suitable for starting a certain signal [6]) or activating a relay [4]) [7]) at the moment that the pressure exceeds a certain limit.

[2]) F. A. Heyn and A. Bouwers, An apparatus for the transmutation of atomic nuclei, Philips Techn. Rev. **6**, 46-53, 1941.

[3]) A. C. Dorsten, W. J. Oosterkamp and J. B. le Poole, An experimental electron microscope for 400 kilovolts, Philips Techn. Rev. **9**, 193-201, 1947.

[4]) N. C. Picard, P. C. Smith and S. M. Zollers, A reliable high vacuum gauge and control system, Rev. sci. Instr. **17**, 125-129, 1946.

[5]) R. I. Garrod and K. A. Gross, A combined thermocouple and cold-cathode vacuum gauge, J. sci. Instr. **25**, 378-383, 1948.

[6]) R. S. Mackey, Non-linear indicator for vacuum gauge, Electronics, Febr. 1948, p. 140.

[7]) H. A. Thomas, T. W. Williams and J. A. Hipple, A mass spectrometer type of leak detector, Rev. sci. Instr. **17**, 368-372, 1946; Detecting vacuum leaks electronically, Westinghouse Engineer, July 1946, p. 108.

For such applications the pressure range of 2×10^{-3} to 10^{-5} mm Hg covered by this vacuum gauge is generally sufficient, but in the manufacture of valves which are exceptionally sensitive to gas still lower pressures have to be measured.

The new vacuum gauge

The new vacuum gauge specially designed for lower pressures has a different electrode system (fig. 1*b*). The cathode plates are round discs parallel

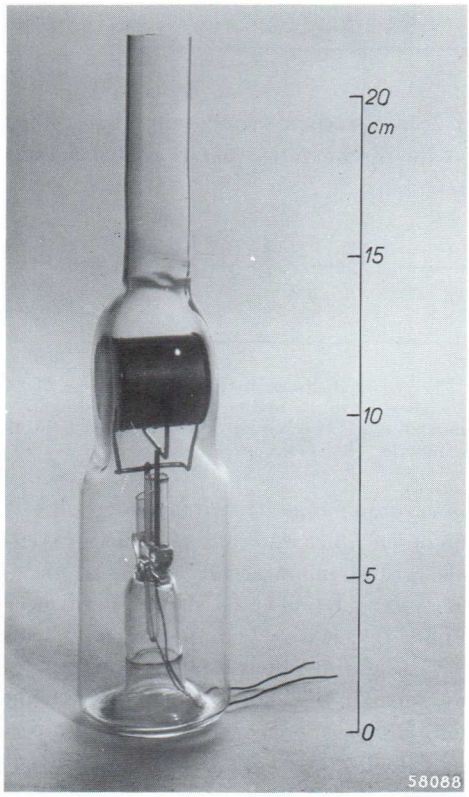


Fig. 3. Photograph of the new vacuum gauge.

to the axis of the glass tube, whilst the anode is in the form of a cylindrical jacket perpendicular thereto. Where the cathode plates are mounted the glass tube is pinched. Thanks to this construction the same permanent magnet can be used with the new vacuum gauge as with the old one, whilst the electrical circuit (fig. 2) also remains exactly the same. *Fig. 3* is a photograph of the new gauge. Compared with the old one it has the advantage that the whole of the discharge space is enveloped by the electrodes and the discharge can no longer be affected by the glass wall and the thin layer of metal deposited upon it by sputtering of the cathode.

In *fig. 4* the current flowing through the vacuum gauge has been plotted as a function of the gas pressure for the voltage (2000 V) and series resistance (1 megohm) usually applying also with the old type. This curve applies for air [8]), its trend depending more or less upon the nature of the gas. A comparison with the analogous curve (dotted line) for the old gauge (fig. 1*a*) shows that the sensitivity of the new vacuum gauge is about ten times greater, due, inter alia, to the larger effective surface area of the cathode. The curve for the new gauge has been drawn as a continuous line down to a pressure of about 10^{-6} mm Hg, but actually the discharge continues to take place at still lower pressures, in which range the sensitivity is represented approximately by the broken line; at such low pressures calibration is very difficult.

In the first description of the Philips vacuum gauge [9]) it was observed that when the pressure is reduced there is sometimes a sudden jump in the intensity of the current. With several specimens of the new design such a jump has been observed round about 10^{-4} mm Hg, thus at the upper limit of the pressure range. Below that pressure there were only small jumps.

Just as with the old construction, for currents between 10 and 1000 mA a small glowlamp (e.g. Philips 4662) can be used as indicator (see fig. 2),

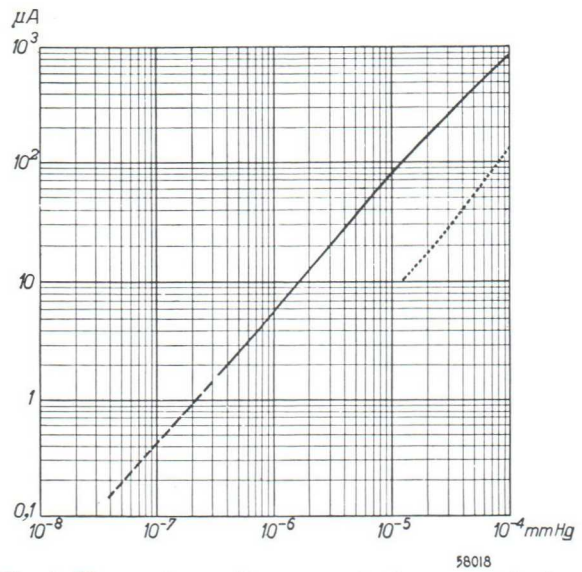


Fig. 4. The continuous line represents the current flowing through the new vacuum gauge plotted as a function of the gas pressure for the usual values of voltage and series resistance (2000 V, 1 megohm). The probable extension of the curve towards lower pressures is represented by the broken line; the dotted line relates to the old vacuum gauge.

[8]) The curve plotted is the average for three different tubes; the greatest deviation was 30 %.

[9]) F. M. Penning, Ein neues Manometer für niedrige Gasdrücke, insbesondere zwischen 10^{-3} und 10^{-5} mm, Physica, The Hague, **4**, 71-75, 1937.

but for currents below 10 mA a more sensitive device is needed. A light-spot galvanometer with say 100 scale divisions for 1 μA is highly suitable for this purpose, one scale division corresponding to a difference in pressure of the order of 10^{-8} to 10^{-9} mm Hg. In order to get results at all reproducible in the range below 10^{-5} mm Hg it is of course necessary to degas carefully all glass walls and electrodes, including the vacuum gauge tube itself.

Leak detectors

In any pumping installation the vacuum gauge can also be used for detecting and tracing leaks. To find the origin of a leak in the object to be evacuated (for instance a transmitting valve or the pumping installation itself) one usually goes over the whole apparatus part for part, replacing the atmospheric air by some other substance. When the leak is reached this substance flows in and the gauge usually registers a change in pressure. As substitute one often uses a liquid, for instance water, and if the apparatus is fitted with a liquid air trap, in which the penetrating water vapour immediately condenses, the pressure in the apparatus will drop to a very low level.

The use of a liquid as substitute for air has its objections. It is not always possible to carry the liquid to the suspected places, and moreover the method becomes very cumbersome when some atmospheric air remains between the liquid and the leak and first has to be pumped away. It is therefore better to use instead of a liquid a gas which can be sprayed upon the suspected places. In recent years this principle has in fact been applied in various ways. For this reason we shall confine our further considerations to gases as substitutes for air. The usual method, whereby a leak is detected through a difference in the pressure registered by the vacuum gauge when the suspected place is sprayed, will be reverted to farther on, after we have first dealt with some other methods.

One can use as indication, for instance, the change in colour of a gas discharge. Hitherto this discharge has usually been brought about by means of a Tesla coil, but this requires a rather high pressure (at least 10^{-3} mm Hg). At lower pressures the change in colour of the discharge can be observed in the Philips vacuum gauge (old design) itself if the tube is provided with a window far enough removed from the discharge to be kept free from any sputtered material (see *fig. 5*). When a leak is sprayed for instance with hydrogen one can clearly see the discharge inside the vacuum gauge change colour.

For smaller leaks other methods have been devised [10]), which were of great importance in nuclear research for testing the extensive vacuum apparatus required for separating U^{235} [11]). For this purpose the apparatus was fitted with a mass spectrograph [7]) sensitive especially to helium and

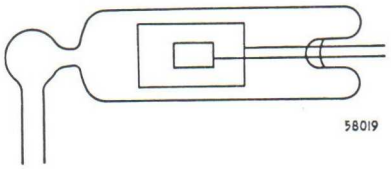


Fig. 5. A vacuum gauge (old design), used as leak detector, provided with a window through which the colour of the discharge can be seen to change when the leak is sprayed with some gas other than air.

the vacuum was tested by spraying with this gas. This method has the great advantage that helium can never occur as an impurity in the apparatus itself, such in contrast, for instance, to hydrogen. In most countries, however, helium is too expensive to be used for this purpose, whilst on the other hand the mass spectrograph is a rather complicated apparatus and the currents have to be amplified before being measured.

Simpler in many respects is a method where advantage is taken of the fact that a heated wall of palladium is only permeable for hydrogen and not for air. With this method hydrogen has therefore to be used as substitute for air. The principle of the method is illustrated in *fig. 6*:

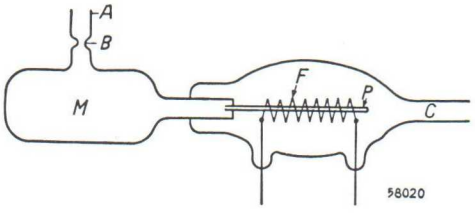


Fig. 6. Principle of a leak detector with palladium wall. M is the vacuum gauge which is evacuated via the tube A, sealed by fusing at B and then connected at C to the pump. P is the palladium tube which can be heated by means of the filament F.

the previously evacuated vacuum gauge M (e.g. an ionization vacuum gauge) is separated from the apparatus being tested by the palladium tube P which can be heated by means of the filament F;

[10]) For a more extensive review see the recent publication of S. Dushman, Scientific foundation of high vacuum technique, John Wiley and Sons, New York 1949.

[11]) R. B. Jacobs and H. F. Zuhr, New developments in vacuum engineering, J. appl. Phys. **18**, 34-48, 1947.

as soon as the jet of hydrogen strikes the leak hydrogen passes through P into M and raises the pressure.

The new vacuum gauge as leak detector

Owing to its high degree of sensitivity the new vacuum gauge described above is highly suitable for detecting small leaks. One can use, for instance, the arrangement shown in fig. 6, where M then represents the new gauge, which is first evacuated via the tube A and then sealed off at B, after which the whole device is connected at C to the pumping installation and the apparatus to be checked.

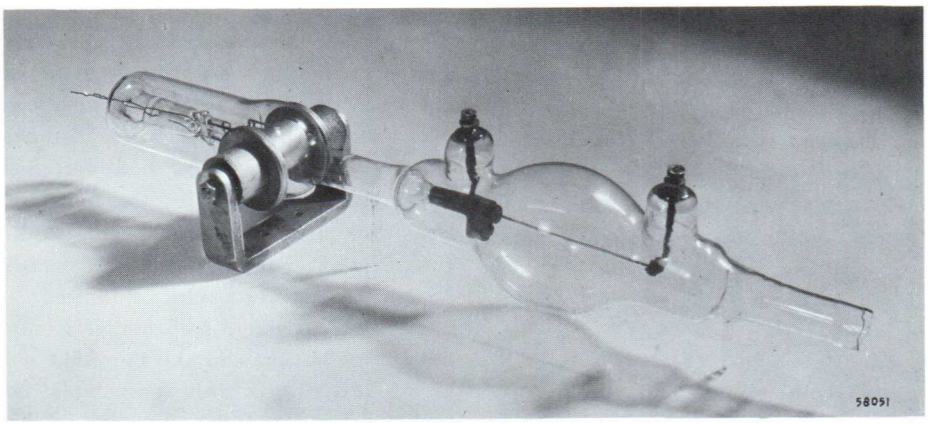


Fig. 7. The new vacuum gauge used as leak detector according to the principle of fig. 6. On the left the new gauge with magnet, on the right the envelope with the palladium tube heated by the direct flow of current via the contacts provided at the top of the envelope.

For tracing a leak the palladium tube is heated with the filament F and current is sent through the vacuum gauge. Since the discharge causes gas to disappear, after some time there will be a low equilibrium pressure, which is not zero because as a rule some hydrogen will still be passing through the heated palladium owing to hydrogen or hydrogen compounds being present in the pumping installation. To localize the leakage the pump is shut off by means of a cock and one goes over the outside of the apparatus with a jet of hydrogen in the manner described above. As soon as the current flowing through the vacuum gauge rises this gives an indication that the leak has been found. If there is any risk of an explosion a mixture of hydrogen with some other gas can be used, but then of course this reduces the sensitivity of the method.

Fig. 7 is an illustration of such a leak detector consisting of a vacuum gauge in the field of a permanent magnet and a bulb containing a palladium tube, the latter in this case being heated by the direct passage of current. The whole device is connected on the right-hand side to the apparatus being tested.

In *fig. 8* some examples are given of the variations in current as a function of time when a leak in an object (total volume about 2 litres) is exposed alternately to air and to hydrogen. The spraying with hydrogen begins where the arrow points upwards and ceases where the arrow points downwards. It is seen that the pressure continues to rise for some time after the spraying has been stopped. This is due to the fact that the leakage channel is still filled with hydrogen, the pressure ceasing to rise when all the hydrogen in this channel has been replaced by air and diffused through the pumping installation.

For these and subsequent measurements the leak was formed artificially by means of a very narrow glass capillary open to the outside air. Such a capillary is made by first reducing the diameter of the tube in a flame until the air channel inside is just visible, after which the tube is drawn out further and broken at its narrowest part. The leak can be enlarged by breaking off a piece of the capillary, and the tube can be sealed by holding the extremity in the flame for a moment. The passage is previously measured by allowing air to leak in for a considerable length of time. In this way it is possible to make capillaries passing through less than 10^{-7} $cm^3 \times$ mm Hg air per second.

The amount of hydrogen passing through the palladium depends upon the difference in pressure between the two spaces and thus upon the increase

of the pressure in the object. For a given leak the sensitivity of this method is therefore roughly inversely proportional to the volume of the object.

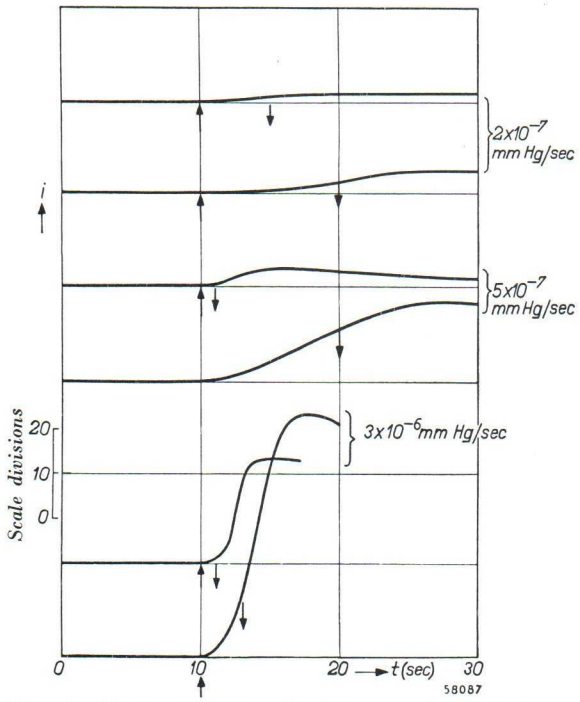


Fig. 8. Some examples of the current variation as function of time when a leak is alternately exposed to air and to hydrogen. Spraying with hydrogen begins at the arrow directed upwards and ceases at the arrow directed downwards. The initial current in air for $0 < t < 10$ sec is indicated at an arbitrary height above the t-axis. An idea of the sensitivity can be formed from the size of the scale divisions of the current meter given in the diagram. To the right of the curves the size of the leak in air is indicated in mm Hg per second.

As already stated, the principle of combining a heated palladium wall with an ionization vacuum gauge with hot cathode has been applied before [12]), the pressure being deduced from the ratio of the ion current to the electron current. The method described here is simpler because there is only one current to be measured and moreover this is greater than the non-amplified ion current in the other method. Another advantage is that when a leak is detected the hydrogen gas that has penetrated into the vacuum gauge can be more easily removed, so that one can pass over more quickly to the next test. In the method with a cold cathode the discharge itself will "burn away" the gas more quickly than when a hot cathode is used.

For the detection of still smaller leaks in valves which can be sealed off from the pumping device another method, without palladium tube, is to be preferred. A vacuum gauge is fused onto the valve to be tested and both of these are then evacuated as well as possible, after which they are sealed off from the pump (see for instance *fig. 9*). The current through the vacuum gauge is then switched on until it reaches a constant value. As already remarked, owing to the discharge a certain amount of gas disappears. In the state of equilibrium the amount burnt away per second by the discharge is obviously equal to the sum of that leaking in and that released from the walls, etc. If the amount of the gas released is negligible some idea of the size of the leak can already be formed from the current flowing through the gauge. On a former

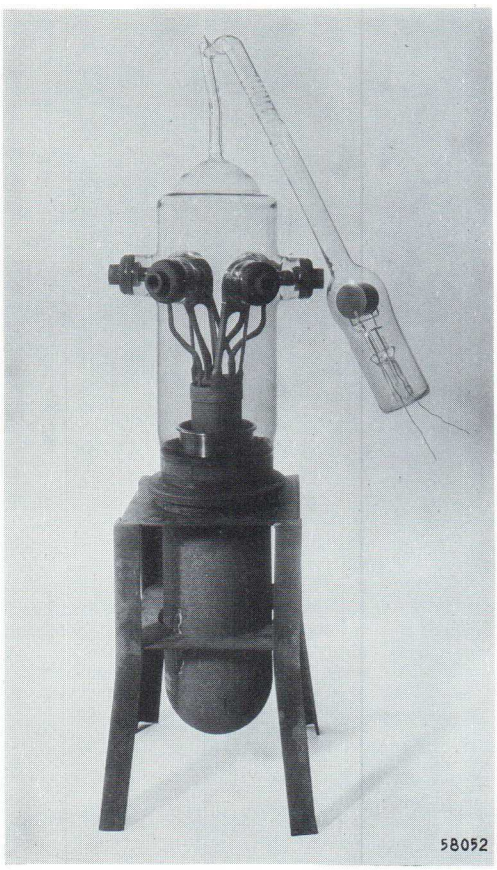


Fig. 9. Photograph of a transmitting valve with vacuum gauge fused on, by which means a leak was found in one of the metal caps for the current lead-ins.

occasion [9]) it had been observed that in a well degased tube about 20 i cm^3 $\times$ mm Hg air per second is burnt away, i representing the current in amperes. In the case just assumed this is in fact the amount leaking in per second.

In order to localize the leak one can now spray the valve to be tested with some gas other than air.

[12]) H. Nelson, The hydrogen gauge, an ultrasensitive device for location of air leaks in vacuum-device envelopes, Rev. sci. Instr. **16**, 273-275, 1945.

The current passing through the gauge will generally depend upon the nature of the gas used, because both the rate at which the gas leaks in and that at which the burning away takes place are related thereto, whilst the sensitivity of the instrument also depends upon the kind of gas. Good results

In this way it is possible to determine extremely small leaks of about 2×10^{-7} cm^3 $\times$ mm Hg/sec, corresponding to an increase in pressure in the volume of the manometer of 0.0025 mm per month. The sensitivity is several hundred times greater than that given for the mass spectrograph method.

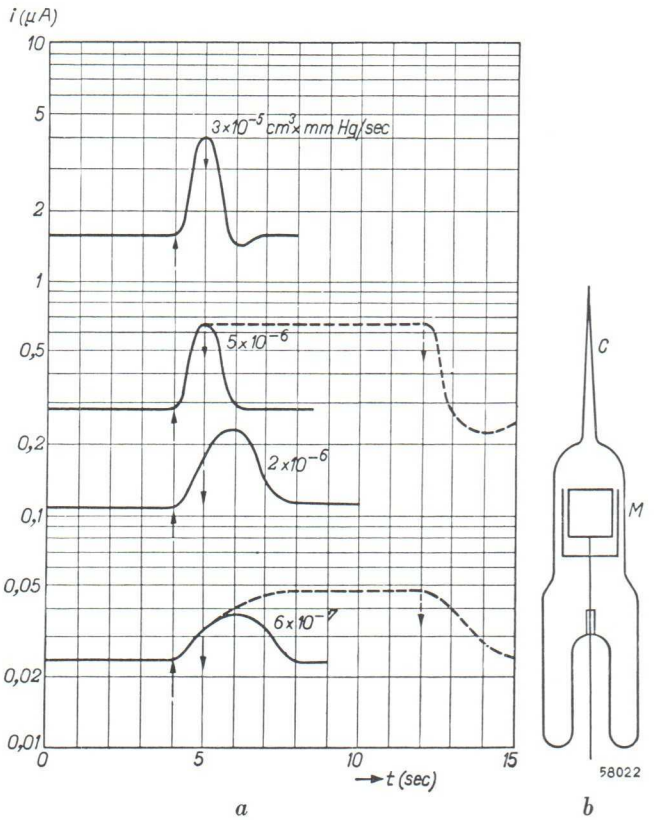


Fig. 10. *a*) Some examples of the detection of leaks in a valve evacuated with the vacuum gauge to which it was fused (the substitute gas was argon). The size of the leak is indicated at the right of the curves in cm^3 $\times$ mm Hg per second. When no more gas is released from the glass wall and the electrodes then the value at which the current through the gauge reaches an equilibrium, with air around the tube, ($t<4$ sec in the diagram) already gives an idea of the extent of the leakage. *b*) illustrates the manner in which the artificial leak in the form of a capillary *C* was connected to the vacuum gauge.

have been reached for instance with argon and hydrogen as substitutes. *Fig. 10a* gives some examples of the results obtained with argon. The leak was again in the form of a capillary shaped as described above, and this was directly fused onto the vacuum gauge as indicated in fig. 10*b*. Just as in fig. 8, the arrows indicate the beginning and the end of the spraying of the leak with the gas. When spraying for only one second the current returns to its initial value within a few seconds owing to the argon being burnt away. When spraying for 10 seconds (curves and arrows drawn in dotted lines) the current reaches within a few seconds a new state of equilibrium whereby just as much argon is burnt away as flows in.

An advantage of the latter method, however, is that the inflowing helium can also be measured when there is already a fairly high pressure of other gases in the apparatus. Moreover, to reach the maximum sensitivity with the method described here it is essential that both the object being tested and the manometer together must be well degased, so that the residual pressure is determined only by the leakage.

In most cases the valves to be tested will have a volume of the same order of size or greater than that of the tube of the vacuum gauge, in which case it will take longer for the air to be replaced by the gas injected. *Fig. 11a* gives an example where there is a volume of about 1 litre between the

leakage capillary and the vacuum gauge (see also fig. 11*b*). For the same size of leak both the initial value before spraying and the ultimate value after spraying will be the same as in the case of fig. 10. In both cases when the state of equilibrium is reached the amount of gas burnt away by the discharge equals the amount flowing in through the leak. When, however, a volume of 1 litre is placed in between the vacuum gauge and the capillary the transition from the initial value to the ultimate value takes place much more slowly. This may be seen when comparing fig. 11*a* with fig. 10*a*. In fig. 11*a* it has moreover been indicated by a dotted line what increase in current might be expected if the volume of 1 litre were omitted. With larger volumes it thus takes much longer to detect a leak, but apart from this the sensitivity is just as great as in the case of smaller volumes.

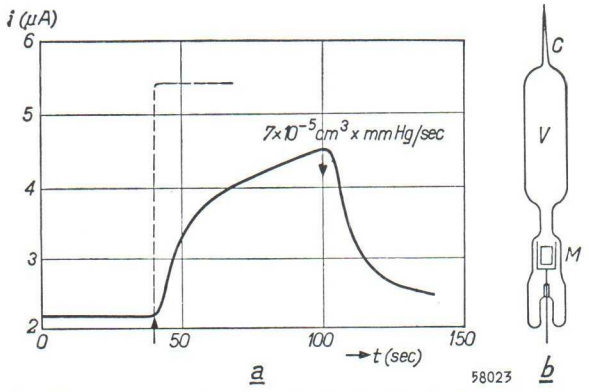
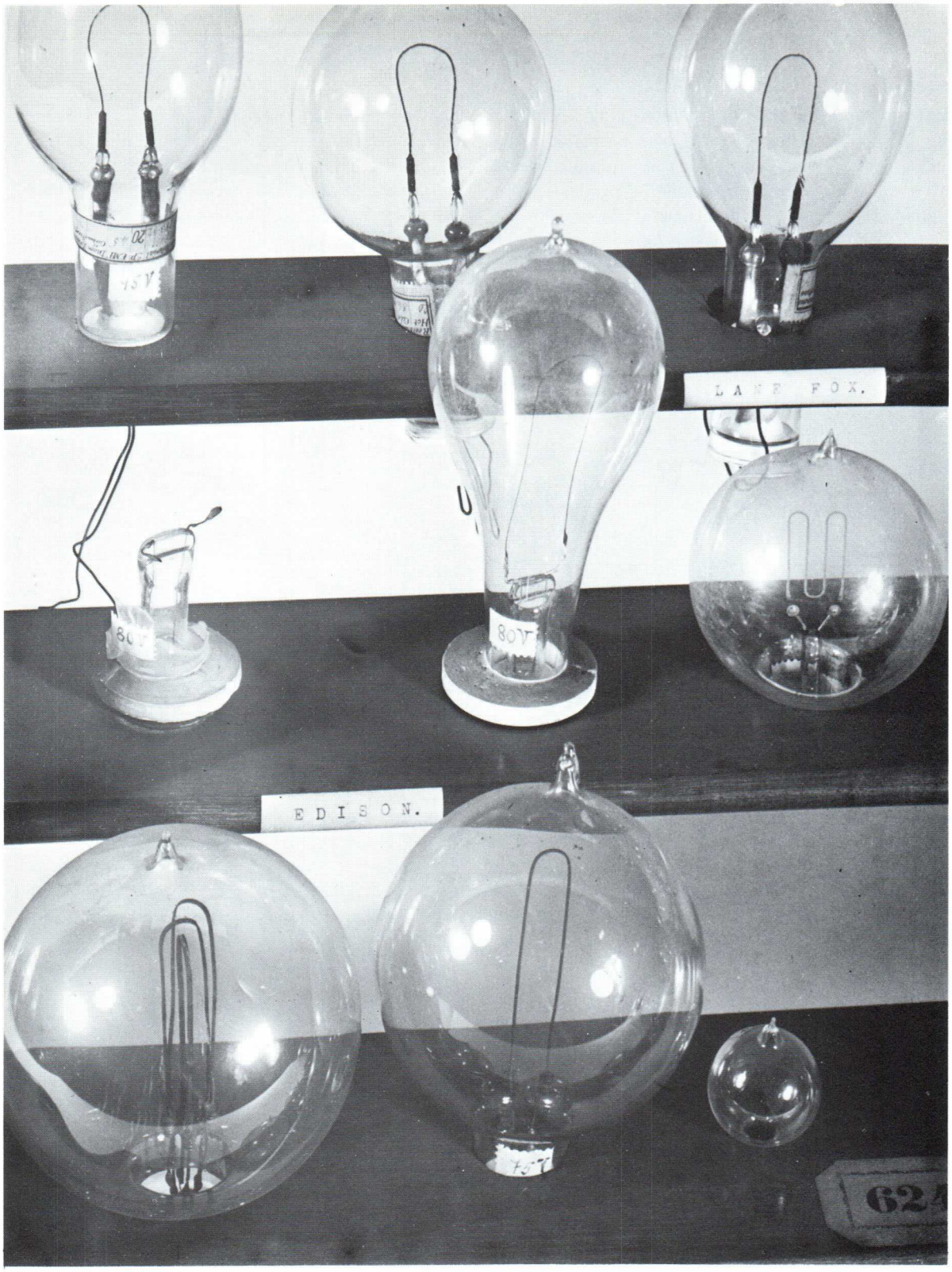


Fig. 11. *a*) An experiment similar to that in fig. 10 but with a volume V of 1 litre between the vacuum gauge and the capillary. The broken line shows the trend to be expected for this leak if the capillary were attached in the manner indicated in fig. 10*b*. *b*) The manner in which the artificial leak (capillary C) was affixed to the vacuum gauge.

Summary. The Philips vacuum gauge previously described in this journal is suitable for gas pressures of 2×10^{-3} to 10^{-5} mm Hg. In order to produce a vacuum gauge based upon the same principle but useful for lower pressures, the electrode system has been replaced by another consisting of two parallel round metal discs as cathode and an anode in the form of a cylindrical jacket. To keep the distance between the pole shoes of the permanent magnet enveloping the glass tube as short as possible this tube has been pinched at that point. These improvements have given the new instrument a sensitivity about ten times as great: it covers a range from 10^{-4} to 10^{-6} mm Hg and can probably be used for pressures even below 10^{-7} mm Hg. Vacuum gauges are also used as leak detectors and some methods followed for this purpose are briefly discussed. It is indicated how the new design of vacuum gauges can best be employed for such an application.



Ancient incandescent lamps

Some incandescent lamps from a collection in Teyler's Museum, Haarlem, Netherlands. These were bought by the Teyler Foundation in 1881 at the Paris exhibition of Electro-Technology. The whole collection comprises 20 lamps, made by Edison, Swan, Lane-Fox and Siemens.

(See: R. J. Forbes, Glass throughout the Ages, Philips tech. Rev. **22**, 282-299, 1960/61.)



THE MANUFACTURE OF VIRUS VACCINE AGAINST INFLUENZA

by A. J. KLEIN *) and E. HERTZBERGER *).

616.921.5-085.371

*For some time now a branch of the Philips Industries has been making Vitamin-D preparations **). During recent years this branch has developed into a many-sided industry turning out pharmaceutical products such as insulin, preparations of liver and hormones, plant-hormones, insecticides, weed-killers, etc. It is now engaged upon the production of virus vaccines, a peculiar form of mass production taking place partly in some large egg hatcheries and promising to be of great importance not only for the Netherlands but also far beyond its frontiers.*

Influenza as an epidemic

A number of infectious diseases occurring in the form of epidemics, such as cholera, plague, typhoid, smallpox, etc., have lost their grip on mankind thanks to counter-measures evolved by medical science. These counter-measures consist on the one hand in the prevention of infection by certain hygienic precautions and on the other hand

*) N.V. Philips-Roxane, Pharmaceutical-chemical Industry "Duphar", Weesp, Holland.

) See A. van Wijk, Lamp manufacture and vitamine research, Philips Techn. Rev. **3, 33-39, 1938.

in the immunization, by means of sera or vaccines, of individuals liable to infection.

During the last ten years it has been made possible to combat also influenza in this way. This disease often occurs in more or less harmless forms and as a result the public at large tends to underestimate its seriousness. The terrible epidemic of influenza that spread over the whole world in 1918 and 1919 and cost about 20 million people their lives is, however, evidence enough to show how important it is to have the means of combating this disease. Moreover, the seriousness of influenza is not to be judged solely from the number of deaths it causes. Influenza epidemics with low mortality figures, occurring, as they do, every other year or so, cause great losses to the national economy owing to the high percentage of workers laid up for weeks on end.

The possibility now offered of immunizing either certain individuals or whole groups of the people against influenza is therefore a great boon both from the medical as well as the social-economic point of view.

In this article something will be said about the principles underlying the preparation of the vaccine for immunization against influenza, the methods for the large-scale production of this vaccine and in particular those employed by Philips-Roxane at Weesp in Holland, and about the manner in which the vaccine is to be used in the event of an epidemic.

Immunization; serum and vaccine

What are the principles underlying immunization against an infectious disease?

The disease is caused by a morbific germ, in the form of a bacillum or a virus, getting into the body and very rapidly multiplying itself. The body reacts to this by producing specific antibodies rendering the morbific germs harmless. Roughly speaking the process of the disease can then be regarded as a contest between the morbific germs and the antibodies: if the former multiply themselves at a faster rate than the latter are produced then they win, whilst in the other case the patient wins and recovers. If a person is already carrying a quantity of the specific antibodies before the morbific germ gets into his body, then as soon as there is an infection that person has a good start, so to speak, and better chances of throwing off the attack — he is more or less "immune" to the disease.

One can become immune in three ways:

(a) Once a person has had an attack of the disease and fully recovered from it there is usually a surplus of antibodies left in the body and this surplus continues to be active for a time, so that one is naturally immune against another attack for some length of time and possibly, in the case of some diseases, for the rest of one's life.

(*b*) A person can be injected with a dose of antibodies taken from some other persons or from test animals who have passed through the process of the disease; a preparation of antibodies sufficiently concentrated for this purpose can usually be extracted from the blood of a test animal, and this supplies the preparations known as "sera" or anti-sera.

(*c*) The body of the person to be immunized can be incited to prepare antibodies without rendering that body ill. This is done by injecting a "vaccine", i.e. a preparation of inactivated germs. It is indeed surprising that this should be possible, considering that the morbific germ is "killed" and can no longer multiply itself, but it is nevertheless still capable of bringing about the production of the specific antibodies. In fact it is not so surprising when it is borne in mind that there are also "dead" substances (proteins) capable of causing the formation of corresponding "antibodies" in the human body. Anyhow this method is probably the most important weapon available to the physician in combating epidemics, more so than the serum method. Since it has not yet been found possible to extract antibodies in the pure state for any disease, and only for a few diseases is it possible to extract them in strong concentrations, usually rather large quantities of a serum (some tens of cm^3) have to be injected and owing to the sensitivity of the human body for foreign proteins this not infrequently gives rise to complications. For vaccination, on the other hand, usually only a few drops of the vaccine are sufficient because of the high concentration in which it can be produced, thanks to the multiplying power of the morbific germs (not yet inactivated!). Furthermore, antibodies taken from other people, and especially when they are taken from animals, being "generic-foreign material", are fairly soon expelled from the human system, so that the protection afforded by a serum (passive immunity) is of short duration — at most six weeks — whereas "active" immunity obtained by vaccination may last many months and even years.

The influenza virus

It is only possible to prepare a vaccine (and also a serum) after the germ causing the disease concerned has been identified, or rather — in the case

of the morbific germ being unknown — after it has been possible to isolate it. Such was first possible in the case of influenza in 1933, when Smith, Adrewes and Laidlaw [1]) succeeded in extracting in a more or less pure form a virus from the liquid used by influenza patients for gargling. Test animals infected with this virus "caught" influenza. While the patients were recovering these investigators also succeeded in detecting in their blood antibodies against the virus found; test animals injected with virus preparations previously mixed with the blood extract did not become ill.

Further investigations carried out by a number of research institutions in America, Great Britain, the Netherlands and other countries proved that in the case of influenza it is not a matter of one morbific germ but that there are several different "strains" of virus, distinguished, inter alia, by the fact that they cause the formation of different antibodies. According to their degree of relationship these strains of virus can be classified under a few groups which have been named influenza-A-virus, influenza-A′-virus and influenza-B-virus. A virus can be neutralized not only with its own antibody but also with antibodies of other viruses belonging to the same group. The practical importance of this fact will be explained later. Suffice it to say here that a number of laboratories now have preparations of the isolated strains of virus at their disposal, from which by cultivation they are able to produce the quantities of virus required for experiments and for the manufacture of vaccines and anti-sera.

Fundamentals of the preparation of virus vaccine

In essence the cultivation of a virus consists in causing it to multiply itself in a suitable medium. With the object of vaccination, this cultivation is necessary for two reasons: firstly, the extraction, purification and concentration of a virus from the excretions of influenza patients is a cumbersome process and yields only comparatively small quantities, whilst secondly it has been proved that for an effective immunization highly concentrated vaccines have to be used. Prior to 1941 such highly concentrated vaccines were not available and this is the reason why the first attempts to immunize groups of the population met with little success, although with mice and ferrets — the test animals mostly used for this work — a decided immunizing action of the inactivated virus preparations had been observed immediately.

The cultivation of a virus on any large scale was a problem never encountered before. Good methods were known for cultivating all sorts of bacteria, e.g. in bouillon or blood which may be made into a solid culture medium with agar-agar, at certain temperatures. One of the essential differences between a bacillum and a virus, however, is that the latter can only multiply on or in living tissues. The customary simple culture media for bacteria were therefore unsuitable and use had to be made of test animals or of tissues derived from test animals and kept alive in a suitable medium (mostly of rather complicated composition, i.e. the so-called tissue cultures). For a time the virus was cultivated, for instance, in the lungs of mice. This technique was far too cumbersome for the large-scale production of vaccine, and the purification (removal of undesirable proteins) and sterile preservation of the product (exclusion of foreign germs which would take part in the multiplication and make the vaccine useless) were difficult problems. Consequently the result could not be called satisfactory.

An excellent idea, which eliminated at once most of the difficulties, was to use embryonated chicken eggs as medium for the cultivation [2]). Here we have indeed a living tissue, easily obtainable at all times in large quantities, contained in a shell guaranteeing sterility — veritably the egg of Columbus. When this idea came to be further investigated it appeared that with a suitable choice of the spot inside the egg also an excellent opportunity was afforded for the purifying of the virus solution obtained. Since 1940 very effective methods have been developed upon this principle for the cultivation of influenza virus. It is such a method — worked out for the greater part empirically — that is being employed in the mass manufacture of influenza virus vaccine by Philips-Roxane at Weesp (Holland).

Before proceeding to describe the details of the method followed, it is to be pointed out that the cultivation of the virus is only one of the steps in the preparation of the virus vaccine. Other essential steps are: the purifying of the virus, its inactivation, possibly drying, and last but not least the assay and dilution. These processes, too, will be briefly described.

Incidentally, it may be mentioned that once the technique of manufacturing a virus vaccine has been mastered the possibility of immunization also against other diseases caused by a virus comes within the reach of the medical profession. A vaccine

[1]) W. Smith, C. H. Andrewes and P. P. Laidlaw, Lancet **225**, 66, 1933.

[2]) W. Smith, Brit. J. Exp. Path. **16**, 508, 1935.

can be made against parotitis (mumps) for instance, the methods for the preparation of which are entirely analogous to those followed for influenza vaccine.

Manufacture of the vaccine

Cultivation of the virus in embryonated chicken eggs

Fig. 1 is a cross section of a fertilized chicken egg after ten days' incubation (hatching normally takes 21 days). The most favourable part of the egg for the development and subsequent purification of the virus is the membrane of the allantoic cavity. This cavity, filled with the clear allantoic fluid, almost entirely envelops the embryo and takes up all the waste products while the embryo is developing, thus acting, as it were, as its bladder. On the tenth day of incubation the virus is injected into this cavity, where in two days' time it multiplies about 100,000 fold. In this way, from one or two eggs (according to the strain of the virus) an amount of virus is obtained sufficient for immunizing one person.

An enormous number of eggs are therefore required for manufacturing vaccine on a scale calculated for the immunization of a whole population. The incubation of the eggs for the required period of ten days and the cultivation of the virus during

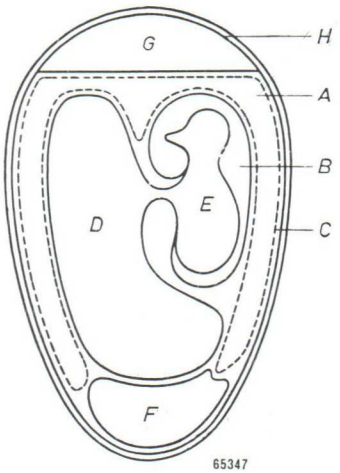


Fig. 1. Diagrammatic cross section of a fertilized chicken egg after 10 days' incubation. *E* embryo, *D* yolk sac, *B* amnion cavity, *A* allantoic cavity, *C* allantoic membrane, *F* albumin, *G* air space cell, *H* shell membrane.



Fig. 2. Machine for the simultaneous hatching of 16,000 eggs. The temperature inside the cabinet is maintained at a level of 37.5 °C. At the bottom are trays of water providing for the necessary humidity of the air, so that the eggs do not lose any of their liquid through the pores of the shell. The trays filled with eggs are placed in a large drum which can be rotated through an angle of 90 degrees: in one position of the drum the eggs are tilted 45° to one side and in the other position 45° to the other side. This turning of the eggs (in imitation of what the brooding hen regularly does) has to be done every day to prevent the heaviest parts of the egg from settling at the bottom of the shell. A wooden framework rotating around the drum provides for a uniform distribution of the heat and moisture content of the air inside the cabinet.

two days is therefore done in large incubators such as are used in the poultry-raising industry, each having a capacity of, say, 16,000 eggs (see *fig. 2*).

After three days' of incubation all eggs are examined in a darkened room, one by one, in a strong beam of light from a mercury lamp (see *fig. 3*). *Fig. 4* is a picture of a fertilized egg as examined in this way; the blood vessels formed during the first three days of incubation are clearly seen. At this stage it is thus possible to see whether an egg has been fertilized or not, and those that have not been fertilized are rejected. After ten days' incubation the eggs are examined again and marked at the spots where "inoculation" and "harvesting" is to be done. This is necessary because the eggs are inoculated by drilling a small hole in the shell and then injecting a few drops of a virus preparation: when this is being done care has to be taken not to puncture one of the blood vessels. The laboratory assistant examining the egg therefore marks it with a cross at a point suitable for the injection at the right height and at a safe distance from the blood vessels. Before injection takes place a second small hole has to be made in the air space of the egg (see fig. 1), so that as the virus is injected a corresponding amount of air can escape. For the subsequent harvesting the boundary of the air space is also marked out on the shell at this second examination of the egg.

The drilling of the holes is illustrated in *fig. 5* and the process of injection in *fig. 6*. Injecting is done in a glass-top case with two openings through which the assistant can pass her arms. After the inoculation the two holes in the egg are sealed with paraffin wax. Inside the case is a germicidal lamp (the TUV lamp), the ultra-violet rays from which keep the air and everything else inside the case sterile [3]); it is of the utmost importance that no living bacteria or active undesired viruses on the egg shell or on the point of the injection needle should get into the egg and there multiply together with the influenza virus (of course at this stage the influenza virus itself must not come under the germicidal action of the lamp rays). To keep the air inside the case sterile in spite of the arm holes, a certain pressure is maintained in the case by means of a ventilator; the stream of air entering at the back passes through a filter, where all dust is filtered out, past a second germicidal lamp and thus in a perfectly sterile condition into the work compartment, from which it flows out through the arm holes (see *fig. 7*).

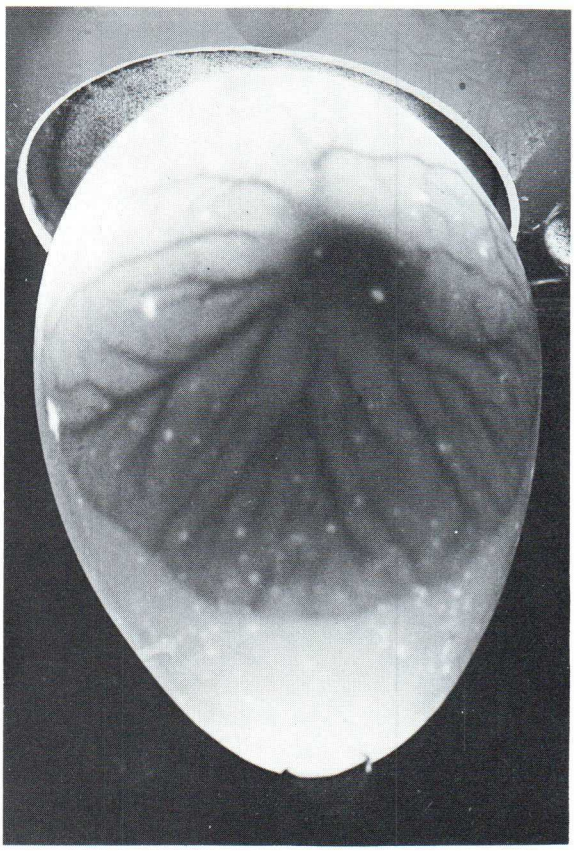


Fig. 4. An egg as seen when examining it after three days' incubation. The blood vessels are clearly seen. The boundary of the air space cell cannot be seen on the photograph owing to the halo effect.

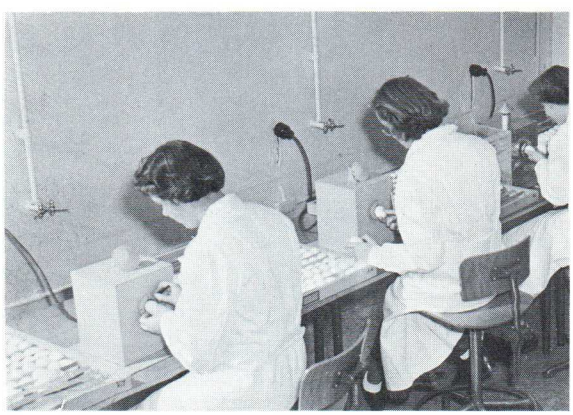


Fig. 3. Examination of the eggs in a strong beam of light to see whether they have indeed been fertilized; the same devices are also used for marking the eggs for subsequent inoculation and harvesting.

[3]) For a description of these lamps and their operation see J. Voogd and J. Daams, Inactivation of bacteria by ultraviolet radiation, Philips Techn. Rev. **12**, 111-119, 1950 (No. 4).



Fig. 5. The necessary holes are drilled in the eggs with a kind of dentist's drill.

There is not much to be said about the actual process of cultivation, which takes place in two days after the injection of the virus into the egg. It is rather peculiar that the best results are obtained when the incubating temperature during the cultivation period is kept at 36.5 °C, whilst the normal incubating temperature, maintained during the first 10 days of incubation, is 37.5 °C.

After the two days' cultivation the eggs are kept 24 hours in a refrigerator, so that the embryo dies off and the rest of the contents of the egg "sets" more or less, this being desired for the "harvesting". For the harvesting the part of the egg shell covering the air space is cut off without damaging the contents of the egg. Then with a pipette, connected via a flask to a water-jet pump, the liquid contents of the allantoic cavity are drawn off into the flask. This, too, is done in a sterilized case with arm holes, at which a number of laboratory assistants work together (see *fig. 8*). One flask can contain the culture from about 300 eggs.



Fig. 6. The case, kept sterile with TUV lamps, in which the incubated eggs are inoculated with the virus to be cultivated. The glass top does not allow the rays of the 2537 Å wavelength to pass through. The arm holes can be moved a little to the side to provide a comfortable position for the assistant's arms.

After sterilization by heating, the eggs from which the allantoic fluid has been extracted can be worked up into cattle food.

The virus preparations required for the inoculation of the eggs are kept in a refrigerator and by cultivation, likewise on incubated eggs, in a small thermostat (miniature incubator) can be constantly renewed.

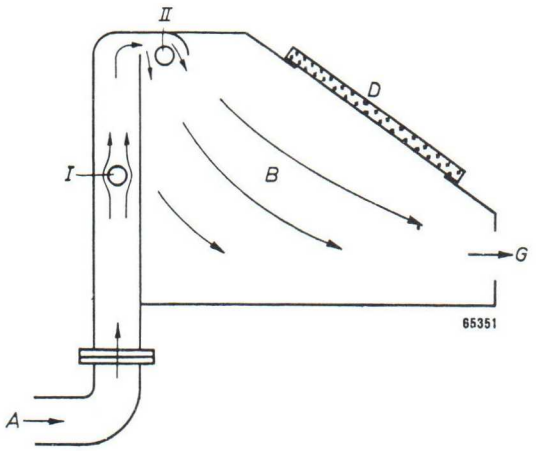


Fig. 7. Cross-sectional diagram of the inoculating case. The work compartment *B*, with glass top *D* and arm holes *G*, is kept sterile by the TUV lamp *II*. A certain pressure is maintained inside the case by the admission of filtered air at *A* with the aid of a ventilator. This air is sterilized as it passes the TUV lamps *I* and *II* and leaves the case through the arm holes.

Purifying and assaying the virus preparation

For purifying the virus, that is to say the removal of most of the undesired proteins present in the solution that has been collected, a process is applied which might be called "differential centrifugation". In a fairly large, normal, centrifuge with a speed of 3000 r.p.m. the flocculent precipitates from the virus solution are deposited on the wall of the machine. The remaining liquid containing the virus is run through a small centrifuge rotating at a very high speed (Sharples super-centrifuge, 50,000 r.p.m.); see *fig. 9*. This centrifuge is highly selective, all the molecules over a certain weight — and those only — being deposited on the wall. Thus the virus (which in this connection can be regarded as a heavy protein molecule) is deposited on the wall of the rotor in an extremely pure state, being subsequently shaken off with a suitable liquid. The speed of the centrifuge and the rate of flow have to be very precisely adjusted and continually checked,

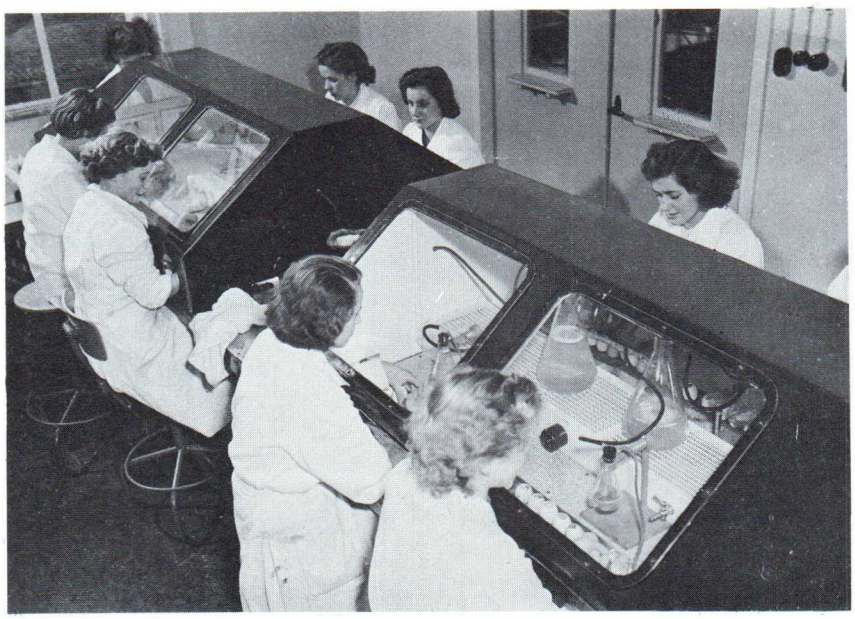


Fig. 8. Case, sterilized with TUV lamps, in which the cultivated virus is extracted from the eggs. Each of the assistants working at the case uses a pipette, connected to a suction pump, with which the liquid is drawn out of the allantoic cavity of the "decapitated" egg into the flask standing in the middle of the case. This flask can hold the culture taken from 300 eggs. In each suction line is a small intermediate flask which has to be inverted after every fourth egg for its contents to be transferred to the big flask. If one of the eggs happens to be a failure (for instance through a blood vessel bursting) the intermediate flask is replaced by a clean one, so that the contents of the big flask are not spoiled.

so as to ensure the right sedimentation. Since, during the flow of the liquid through the centrifuge, outside air is apt to be drawn in (see fig. 9*b*), also the centrifuge is set up inside a case sterilized with the aid of TUV lamps (*fig. 10*).

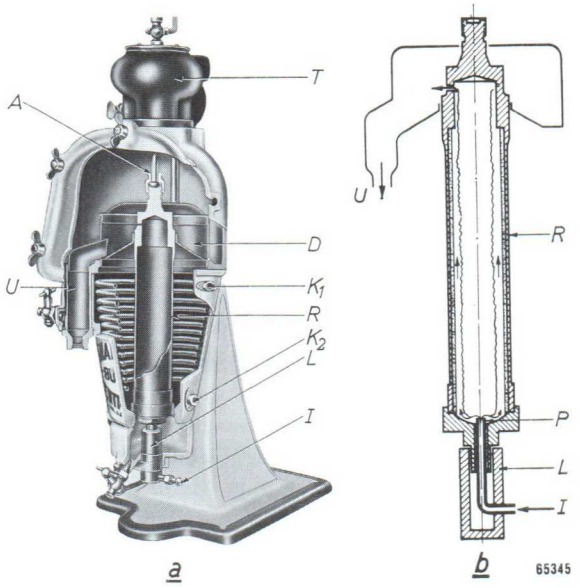


Fig. 9. *a*) The Sharples supercentrifuge. *T* driving steam or compressed-air turbine, *A* shaft rotating at 50,000 r.p.m., *R* rotor mounted on the shaft and through which the liquid to be centrifuged flows, *I* inlet, *U* outlet, *D* cover, *L* bearing leaving sufficient play for the rotor to find its own ideal axis of rotation, K_1 and K_2 water inlet and outlet for the cooling coil. In *b*) the sedimentation in the rotor (drawn very roughly) is made a little clearer. The liquid entering the rotor at *P* follows the rotation of the rotor almost immediately.

The highly concentrated virus preparation thus obtained is shaken with a small addition of formaldehyde, thereby completely inactivating the virus. After adding a preservative (sodium ethylmercuri-thiosalicylate) the virus solution can be kept in a refrigerator until it is needed.

Another method of preparing the virus for storage is by drying. Of course the water cannot be expelled from the solution by heating because then the virus is damaged. Boiling under reduced pressure is not practicable either, because then too much virus is lost (owing to excessive heating, concentration of salts or other causes). Very good results are obtained, however, by freezing the virus solution (in a vessel cooled with carbon dioxide) and then evaporating the ice by connecting the vessel to a vacuum pump. The dry residue containing the virus is then kept in a refrigerator at a temperature of —20 °C.

The last, and very important, step in the manufacture of the vaccine is the determination of the activity of the suspension (assaying) and the dilution to the desired strength.

There is a simple method of determining the concentration of an influenza-virus suspension, namely the Hirst test [4]). This is based upon the discovery that influenza-virus suspensions cause the red blood corpuscles of chickens to clump. This phenomenon allows of the quantitative measuring of the concentration, and the Hirst test has therefore become an indispensable means of continually checking the manufacture of the vaccine and of making the desired dilutions (see frontispiece). This cannot, however, be said to be an assaying in the strict sense of the word: virus suspensions of the same concentration but from different cultures may vary considerably as far as their activity as vaccine is concerned, since the virus may have a different degree of antigenic action, i.e. a different power (retained or originally possessed) to produce antibodies. It is therefore necessary to determine the activity of the virus suspension as vaccine by means of test animals.

This is done according to the method prescribed by the National Institute of Health in the U.S.A., which has also been made compulsory by the State Institute of Public Health in the Netherlands. A number of mice are inoculated with different doses of the vaccine and after 14 days samples of their blood are taken. This blood serum is mixed with active influenza virus and dripped into the nostrils of non-inoculated mice. The serum from the vaccinated mice should neutralize the active virus, in which case the mice to which the mixture has been administered will stay alive. The dose of vaccine just neutralizing the virus is taken as the standard strength on which further dilutions are to be based.

This assaying is done also in the laboratories at Weesp. To conclude the preparation of the vaccine the prescribed tests are carried out — partly on animals — for making sure that the vaccine does not contain any living virus, any living germs of some other genus, or any other harmful substances of whatever nature. Not until then can the vaccine be released for medical use.

This is not the place to go into the manner in which the vaccine is injected into human beings. All that need be said here is that the persons to be vaccinated must not be oversensitive to chicken protein. Fortunately this is only the case with about 0.5% of the population, and any over-

4) So long as a virus cannot be completely isolated it is of course not possible to determine its quantity by weighing, so that one has to have recourse to such means as biological effects, specific absorption of light or fluorescence, or suchlike.

sensitiveness (allergic reaction) can be determined by first injecting a very small dose into the skin (the actual vaccination is by subcutaneous injection).

Combating influenza with the virus vaccine

The fact that in the case of influenza it is not a matter of just one single virus but of various strains is of great influence in the combating of the disease. It cannot be known in advance what strain of the virus will be responsible for a future epidemic.

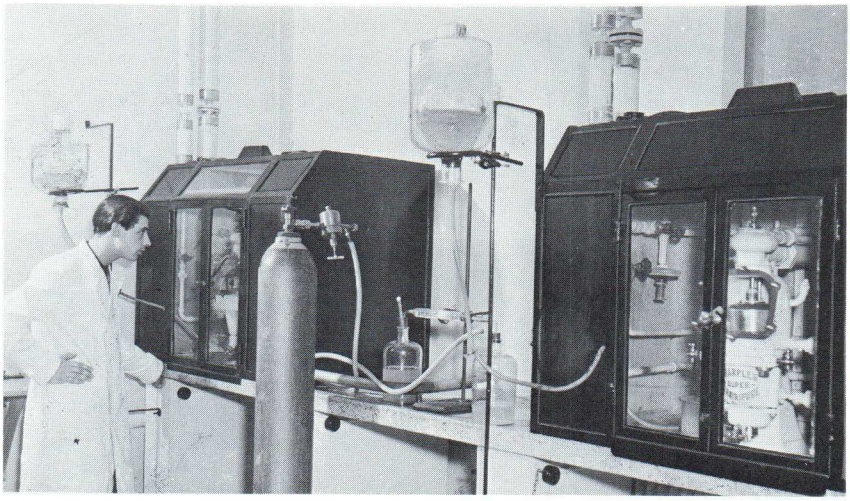


Fig. 10. Cases, kept sterile by TUV lamps, in which the Sharples supercentrifuges are mounted. The centrifuges have to be easily accessible because at regular intervals the speed has to be checked and if necessary corrected.

Therefore, to make an influenza vaccine which can be held in readiness to cope with the first attack of an approaching epidemic or, in normal times while there are still no signs of an epidemic, to safeguard against possible future infection people for whom influenza involves particular risks, this vaccine has to be "polyvalent", that is to say it has to contain various suitably chosen strains of virus which together have the widest possible antigenic action.

To hold in readiness this polyvalent vaccine is not, however, an ideal solution. It is true that in the event of an epidemic breaking out the majority of the people vaccinated will be safeguarded against the infection if they have been vaccinated early enough and if the strain of virus causing the epidemic is susceptible for one of the antibodies created. But if the polyvalent vaccine is made up, for instance, from a mixture of three strains then every person has to be inoculated with three times the amount of vaccine that would be needed in the case of a "monovalent" vaccine. This is to be deprecated on account of the undesired secondary reactions, which in any case have to be limited to a minimum, and the more so because with a given capacity of the factory producing the vaccine only one third of the number of people can be immunized which otherwise could be treated with a monovalent vaccine.

Furthermore, it may happen that the epidemic is caused by a strain of virus as yet entirely unknown, a mutant such as seems to appear fairly regularly every ten years and which may not react at all to the antibodies formed by the polyvalent vaccine. In that case everyone who has already been vaccinated has to be revaccinated.

With a view to this situation, in 1948 the World Health Organisation established a service for reporting to all countries any outbreak of influenza in any part of the world and for taking immediate steps to isolate and identify the virus. Thus, in all countries to which the epidemic may be expected to spread in the course of some time it is now possible to proceed at once to the production of a monovalent vaccine against the identified morbific germ of the disease.

This means that the industry making the influenza vaccine is obliged to equip itself for coping with very heavy "peak loads". In order to meet this requirement without involving the installation of large permanent plant that would be standing idle most of the time, Philips-Roxane have estab-

lished a remarkable organisation. The manufacture of the virus vaccine depends for a large part on the use of large incubators for hatching the thousands of eggs needed. Such incubators, however, are already available in the poultry industry, where as a rule they are idle for several months of the year, just in the summer and autumn when epidemics of influenza commonly break out. It has therefore been arranged with several hatcheries in the country (from whom the fertilized eggs for cultivating the vaccine are procured) for their machines to be used in the vaccine production process in case of need. By temporarily increasing the number of assistants making up the inoculating and harvesting teams (the inoculating teams work in the hatchery, whilst for harvesting the eggs are transported to the factory) the production capacity can now easily be raised to a multiple of the normal.

Even with such a greatly increased capacity it is still impossible, within the short time available before the reported epidemic spreads to the country, to produce sufficient vaccine for immunizing the whole population. But then, after all, this is not really necessary. Much harm can already be avoided by immunizing those groups of people most exposed to the infection and certain individuals whose incapacitation would help the epidemic to spread and thus aggravate the general dislocation of public services and industry. The vaccine available will therefore be reserved in the first place for the vaccination of doctors, nurses and suchlike, for personnel of electricity works, telephone exchanges and the foodstuffs industry, further for people living or working in communities, such as the military forces, seamen, pupils in boarding schools, workers in large factories, etc. For the same groups of people a preventive vaccination in normal times might also be considered. This would have to take place every year, because between 6 and 12 months after the injection immunization begins to decline.

Summary. As is the case with other infectious diseases, immunization is now also possible against influenza. The vaccine, a concentrated suspension of the inactivated influenza virus, is being manufactured on a large scale by cultivating the virus in the allantoic cavity of embryonated chickens' eggs. The injection of the virus into the egg, the cultivation and the harvesting of the greatly multiplied virus, as also the purification of the virus suspension and the assaying for correctly dosing as vaccine are described in this article. Since there are a number of strains of the influenza virus, in normal times a polyvalent vaccine is made containing a mixture of different strains and thus offering immunization against various strains of the virus. With the approach of an epidemic of influenza the specific strains of virus responsible for it is reported by an international organisation to the threatened countries, where steps can immediately be taken for producing a monovalent vaccine against that particular strain and vaccinating the population as far as necessary and possible.

DELTA MODULATION, A NEW MODULATION SYSTEM FOR TELECOMMUNICATION

by J. F. SCHOUTEN *), F. de JAGER and J. A. GREEFKES.

621.396.619.16

For the transmission of information such as speech or music the aim is to design a system possessing the optimum properties, having regard to the properties of the transmitting channel on the one hand and, on the other hand, the characteristics of the signal to be transmitted. Recently systems have been developed in which a quantization of the signal is applied both in amplitude and in time. Such a system is described in this article. Its essential feature is the employment of a special kind of negative feedback. Compared with other quantizing modulation systems the apparatus has been greatly simplified.

Signal and interference

A remarkable development is taking place in the technique of telecommunications, in that speech and music are being transmitted by means of a pattern of unit signals, such as already known in telegraphy.

The desire for such a drastic departure from the usual methods of transmission, such as amplitude modulation for instance, has arisen from the fact that it is becoming more and more difficult to ensure interference-free reception at very great distances. In modern carrier-telephony systems the signal transmitted by cable over a distance of 100 km is attenuated 240 db, i.e. a power ratio of 10^{-24} times. This attenuation, which is in the order of the ratio of the power emitted by the sun to that of a pocket-torch lamp, does not in itself constitute any difficulty: in transcontinental communications such attenuations may reach some tens of thousands of decibels, but they are compensated by means of a series of repeaters (relay stations) installed at intervals along the route. But, besides being attenuated, the signal is also distorted. The distortion may consist, for instance, in noise, crosstalk from neighbouring channels, and variations in the level of the signal. These effects, in themselves only small, act along the whole transmission circuit, so that the ultimate distortion of the signal received is a cumulation of very small effects.

When a table is covered with dust it can be cleaned by brushing the dust off, for we know what is the table and what is the dust. However, at the receiving end of the communication system referred to it is impossible to distinguish between signal and interference. This means that the "transmission dust" cannot be brushed off, and so with the usual systems of communication cumulation of interference is unavoidable.

There are, however, communication systems where the interference can indeed be distinguished from the signal, and with such systems, when intermediate repeater stations are used, a practically interference-free signal can be received even at very great distances. Before dealing with these systems we shall discuss some other modern methods of transmission which have, as it were, pointed the way to achieving the ultimate object of suppressing interference. These methods are illustrated in *fig. 1*.

How cumulation of interference is counteracted

Amplitude modulation on a carrier (fig. 1*b*) of the signal to be transmitted (fig. 1*a*) can be replaced by amplitude modulation on a series of equidistant pulses: pulse-amplitude modulation (fig. 1*c*). An audio or video signal is given as

*) Of Philips' Telecommunication Industries, Hilversum.

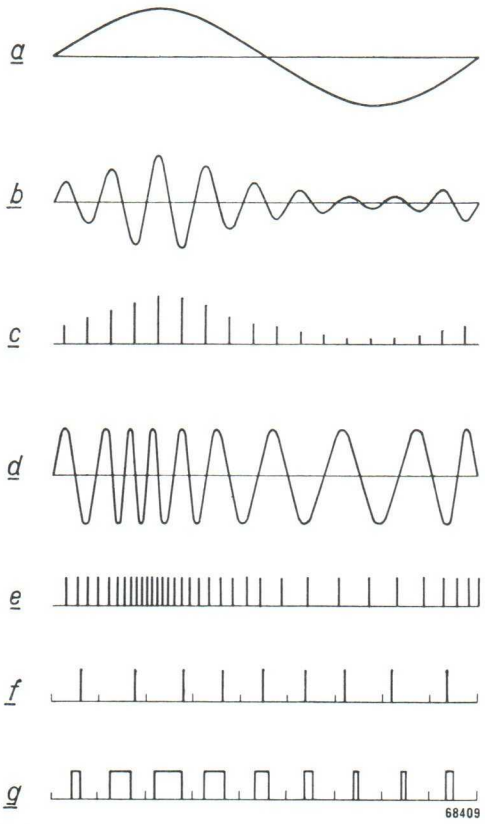


Fig. 1. Diagrammatic representation of various non-quantized modulation systems. *a*) The modulating signal. *b*) Carrier amplitude modulation. *c*) Pulse-amplitude modulation. *d*) Carrier frequency modulation. *e*) Pulse-frequency modulation. *f*) Pulse-position modulation (note the marked time intervals within which each pulse can move). *g*) Pulse-width modulation.

a continuous function; the time is the independent variable, while the dependent variable is the sound pressure, respectively a voltage at the pick-up. It appears that if the value of this voltage is known at equidistant instants the total signal, i.e. including all instantaneous values between those instants, can be wholly reconstructed, provided the frequency bandwidth of the signal is not greater than half the rate of sampling. Thus, the wider the range covered by the frequencies of the signal to be transmitted, the greater is the required sampling rate.

In pulse-amplitude modulation this principle is applied directly. In fact this method is analogous to carrier-amplitude modulation; only the continuous carrier is replaced by discrete pulses. For a frequency band of, say, 3400 c/s to be transmitted, a pulse-repetition frequency of about 8000 c/s is employed. As far as the separation of signal and interference is concerned this method does not, however, offer any advantage over carrier modulation.

Almost simultaneously with pulse-amplitude modulation, frequency modulation (fig. 1*d*) made its appearance in telecommunication technique. With this method it is not the amplitude but the frequency of the carrier that is modulated. Here, too, the continuous carrier can be replaced by a series of pulses, so that pulse-frequency modulation (fig. 1*e*) is obtained [1]). A special form of this is pulse-position modulation (fig. 1*f*), also called pulse-time modulation, whereby the signal is built up from pulses all having the same amplitude and width, called unit pulses. The information is contained in the time pattern of these pulses, such that each pulse occupies a certain place in the time interval allotted to it, according to the instantaneous value of the signal [2]). This offers an entirely new possibility of counteracting interference: if the effects of interference in the transmission channel are sufficiently limited so that they do not severely mutilate the unit pulses then the pattern is retained and can be restored by regenerating the partly mutilated pulses at the receiving end.

Still it is not possible to reach a complete separation between signal and interference, so that some interference remains and is cumulative. The interference in the transmission channel acts in the first instance on the amplitude of the signal transmitted. If the pulses were quite rectangular in shape complete separation of the signal from interference would be possible, provided the interference does not exceed a certain value. Since, however, the slope of the edges of the pulses always has a finite value, there is still some variation in the time at which a pulse is received, i.e. in the position of the pulse in its interval. The resultant interference cannot be separated from the signal and so there is again a cumulative effect.

There is much less interference with pulse-position modulation than with pulse-amplitude modulation, and there is much less interference in the case of carrier-frequency modulation than in that of carrier-amplitude modulation. Frequency modulation and pulse-position modulation have led to the fundamental conception that it is possible to reduce the effect of interference by increasing the frequency band of the high-frequency signal transmitted. This possibility, however, is not efficiently utilised

[1]) Pulse-frequency modulation is something that has existed in nature for millions of years, viz. in the transmission of signals via the nerves of animals. When a nerve is excited a series of electrical pulses of equal amplitude are propagated along the nerve; the number of pulses per second (up to some hundreds) is a measure for the strength of the stimulus.

[2]) Pulse-position modulation has been mentioned in this journal before; see C. J. H. A. Staal, An installation for multiplex-pulse modulation, Philips Techn. Rev. **11**, 133-144, 1949.

in pulse-position modulation, nor in its modified form of pulse-width modulation (based on the principle that the width of the pulses is varied according to the instantaneous value of the signal, fig. 1*g*).

Pulse-position modulation with the aid of rectangular pulses would, as already stated, give complete suppression of interference, but the frequency band required for undistorted transmission of these signals is infinitely large.

Quantization of the signal

For pulse-position modulation unit signals are employed. This implies a sort of quantization: only discrete values (in this case only the unit value) of the pulse amplitude are permitted. The information is determined, however, by a continuously variable pattern of these pulses, in the sense that any particular pulse may be at any point within the continuous time range allotted to it. One can now go a step further and transmit only discrete values — say, the multiples of a certain unit — of the continuous scale of instantaneous values of the signal. Something similar is regularly done in everday life when expressing length in kilometres, metres or millimetres, temperature in degrees or tenths of a degree, time in hours, minutes or seconds. The consideration behind this inaccurate quotation is the fact that in practice a limited accuracy, depending on the object in view, always appears to be sufficient.

The steps in which the instantaneous value of an audio signal is measured can likewise be chosen with a sufficient fineness to obtain a certain quality of reproduction. The loss in fidelity is offset by the advantage gained in being able to eliminate completely the active effects of interference, provided it is kept below a certain limit.

This can be illustrated with the following example, for which we shall revert to the method of pulse-position modulation. Complete quantization of this modulation means that in the time interval allotted to it each pulse can occur only at a discrete number of fixed places, as illustrated in *fig. 2*. If the interference brings about a displacement of the pulse positions which is less than half the difference in time between two of the discrete places then, in spite of this interference, at the receiving end one knows exactly what pulse position, thus what signal, is intended. The interference can therefore be entirely separated from the signal and completely eliminated, provided it does not exceed a certain threshold value.

Direct application of the quantization process to this or to one of the other modulating systems mentioned thus gives the desired suppression of interference. But, as will be shown later, from the point of view of efficiency of the transmission system it is uneconomical to proceed in this way.

Special methods of modulation have therefore been developed for transmission of quantized signals. One of these methods is that of delta modulation, which will now be discussed.

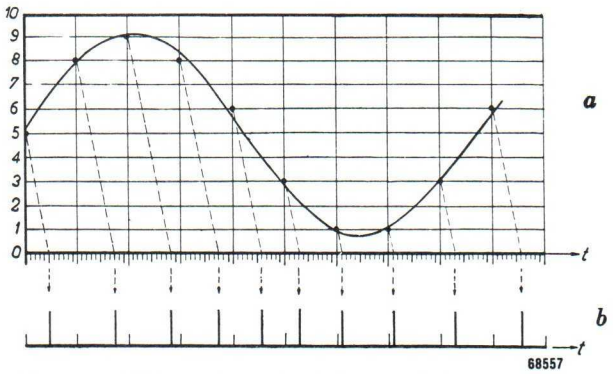


Fig. 2. Quantization of pulse-position modulation. *a*) The instantaneous signal value at the beginning of each interval is approximated by a quantized value. This is "translated" into the series of transmitted pulses (*b*) by a corresponding pulse position in the interval. In this diagram the quantization is carried out in 11 steps. Thus for each pulse there are 11 discrete positions available in its interval. This system is not applied in practice because it is not effective.

Delta modulation

Principle

Delta modulation is a pulse modulation system, pulses being produced at the sending end at equidistant instants. Not all pulses are, however, transmitted. The question whether a pulse is to be transmitted or not is decided in the following way. An "echelon curve" is formed from the pulses in a manner to be explained later. This echelon curve is compared with the modulating signal (*fig. 3b*). If at the moment when a new step has to be added the value of the modulating signal is greater than that of the step-shaped signal, then a positive step is added, but if the value of the modulating signal is smaller a negative step is formed. Thus the echelon curve oscillates and approximates to the modulating signal, being the quantized form thereof. The formation of a positive step is arranged to initiate the transmission of a pulse, whereas the formation of a negative step is not and results in a gap in transmission at that moment in time. What is then transmitted is the series of pulses shown in fig. 3*a*.

At the receiving end the step-shaped approximating signal is built up again from the series of pulses received and finally converted into a con-

tinuous signal giving the reproduction of the original modulating signal. It is clear that if the interference in the transmission channel does not mutilate or displace the pulses as such beyond recognition then this interference can be completely eliminated.

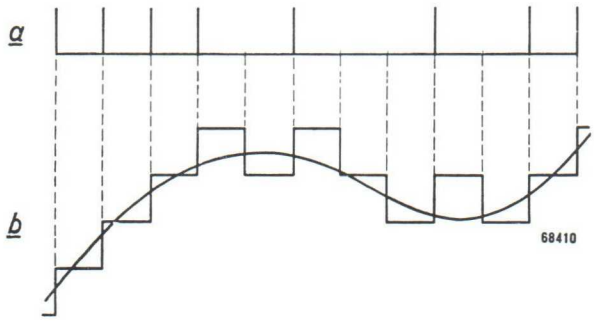


Fig. 3. Principle of delta modulation. The series of pulses emitted by the transmitter (*a*) is generated after comparison of the modulating signal (the continuous curve in *b*) with a step-like approximating signal with equidistant steps built up in the transmitter. When the approximating signal is smaller than the modulating signal a pulse is emitted, but if it is larger the pulse is suppressed. At the receiving end the approximating signal is formed from the series of pulses by integration and from that signal the modulating signal is reconstructed by smoothing out the steps.

This is only a very sketchy exposition and in practice the method is not so simple as that, but anyhow it gives the essence of the method, which is recapitulated in the following concise form.

With the system of delta modulation, from a series of equidistant pulses transmitted, which may have the value 1 or 0, at the receiving end a signal is built up which has to form the nearest possible approximation of the original modulating signal. This approximating signal is constructed not only at the receiving end but also at the transmitter. At the sending end the original signal can therefore always be compared with an approximation similar to that which would be obtained with the aid of the quantized pulses at the receiving end. If the approximating signal is smaller than the input signal then a pulse is sent out by the transmitter, but if the approximating signal is larger than the input signal the next pulse is suppressed.

The "auxiliary receiver" at the sending end, which supplies the approximating signal, is in its simplest form an integrating network. The output voltage from this network is compared with the original signal in a difference meter, a simple resistance network, followed by an amplifier. Thus a sort of inverse feedback is applied in the transmitting circuit. This feedback, however, differs from the usual inverse feedback in two respects. First, the feedback circuit is closed only at the moments when it has to be decided whether a pulse is to be transmitted or not. Second, the voltage acting on the feedback circuit is not proportional to the difference voltage but dependent only upon the polarity of the difference voltage. This inverse feedback is virtually quantized in position and amplitude.

In *fig. 4* a block diagram is given representing the circuit of the transmitter. We see here a pulse generator, *I.G.*, which yields a series of equal pulses with a repetition frequency of, say, 60 000 or 100 000 pulses per second. These pulses have to pass through a pulse modulator, *I.M.* If the voltage at the point *2* is negative then *I.M.* gives a pulse having the same sign as the incoming pulse; if, on other hand, the voltage at *2* is positive then a pulse with the opposite sign is sent to point *3*. The pulses coming from the pulse modulator *I.M.* are sent to an integrating network *I* (i.e. the "auxiliary receiver"). Here each positive pulse will raise the output voltage a certain amount and each negative pulse will lower it by the same amount. In this way an echelon curve or approximating signal *4* is obtained. In the difference meter *V.M.* this signal is compared with the modulating signal *5*, the output voltage always indicating the difference between the two signals and thus controlling the pulse modulator. The next pulse is therefore positive when the approximating voltage at that instant is lower than the input signal and negative when it is higher. The series of positive and negative pulses generated in this manner are

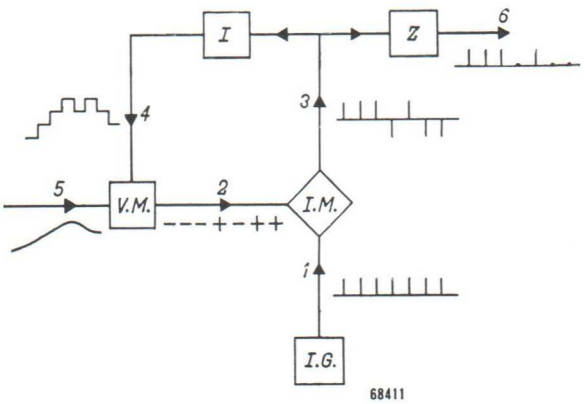


Fig. 4. Block diagram of the transmitter. The pulse generator *I.G.* supplies a series of identical equidistant pulses *I* which are passed on to the transmitter by the pulse modulator *I.M.* either unchanged or with opposite polarity. The action of the pulse modulator is governed by the polarity of the difference voltage *2* obtained in the difference meter *V.M.* by comparison of the modulating signal *5* and the step-like approximating signal *4* formed from the series of pulses *3* by integration in *I* ("auxiliary receiver"). The transmitter suppresses the negative pulses and passes the signal *6* to the aerial or the transmission line.

passed not only to the auxiliary receiver but also to a transmitter Z (for instance a radio relay system) which suppresses the negative pulses and transmits the positive ones.

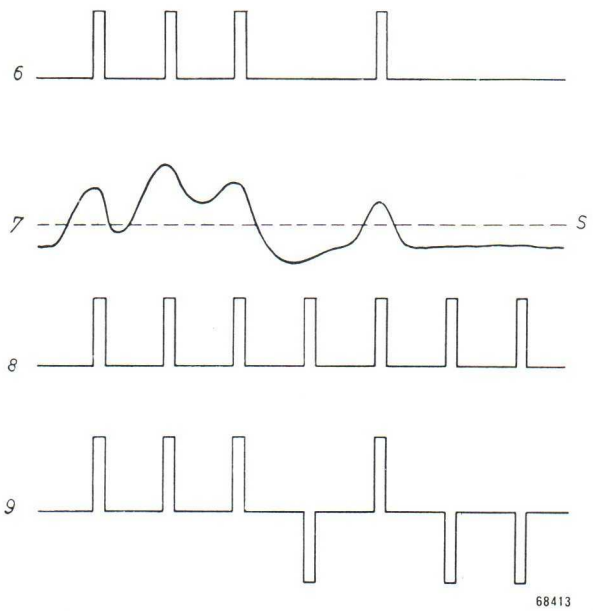


Fig. 5. The series of pulses *6* emitted by the transmitter of fig. 4 becomes distorted in the course of transmission, assuming, for instance, the shape of the signal 7. This signal influences a pulse modulator in the receiver in such a way that, provided the pulse amplitude exceeds a certain limit *s* (mostly chosen equal to half the original pulse amplitude), the pulses *8* produced by a pulse generator are passed on with the correct polarity. If the signal 7 lies below the level *s* the pulses produced by the generator are replaced in the pulse modulator by pulses of the opposite polarity. In this way the series of pulses *9* is obtained.

Fig. 5 shows to what changes the pulses on their way through the aether are subject. The pulse pattern as transmitted corresponds to *6*, while owing to the interference the pattern as received is more or less distorted, resembling, for example, that depicted by 7. Now in order to reconstruct from this pattern the step-shaped approximating signal, in the receiver (*fig. 6*) a pulse modulator is employed which in principle functions in the same way as the pulse modulator in the transmitter: the pulse (*8*) coming from a pulse generator *I.G.* is passed through unmodulated if the incoming pulse exceeds a certain value *s* (for which, as a rule, half the height of the undistorted pulse is taken), but if the incoming pulse is smaller than that value then the pulse modulator *I.M.* gives a pulse of the opposite sign. At the output of the pulse modulator a combination of pulses (*9*) is obtained which is identical to the signal *3* (see figs 4, 5 and 6). (The pulse generator has to be synchronized with the incoming pulses, the apparatus required for this having been omitted from the diagram in fig. 6 for the sake of simplicity.)

When this series of pulses is applied to an integrating network, *I*, the same step-shaped signal (*10*) is obtained as was generated at the transmitter (signal *4*). By passing this signal through a low-pass filter *F* the steps are smoothed out into a continuous signal *11* closely resembling the original input signal.

In this manner interfering signals of at most half the pulse amplitude can be permitted in the course of transmission without being perceptible in the signal.

Sound reproduction by the delta modulation system

Obviously the step-shaped approximating signal (and thus also the continuous output signal derived from it) will always show some small deviations with respect to the original signal. If the modulating signal is sinusoidal then at the receiver a sinusoidal signal is obtained against a background of noise, the so-called quantizing noise, due to the approximation of the instantaneous values of the speech signal by discrete values. The higher the rate of sampling, however, the less is this quantizing noise. In practice, with a pulse frequency of 40 000 pulses per second there is still a perceptible grating quality in the speech signal transmitted, but with a pulse frequency of 100 000 pulses per second there is practically no noticeable quantizing noise. Due to this high pulse frequency a much wider band is required for the transmission than is needed in the case of direct modulation of the signal on a carrier, since the bandwidth required is at least

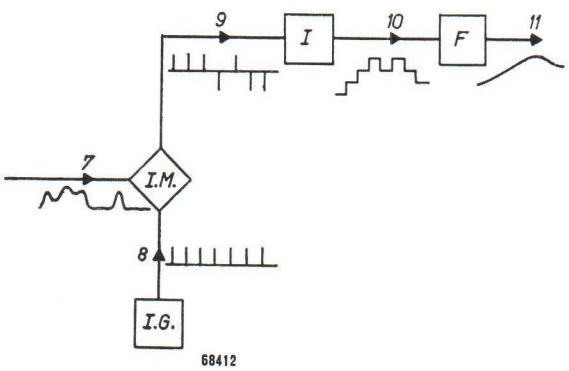
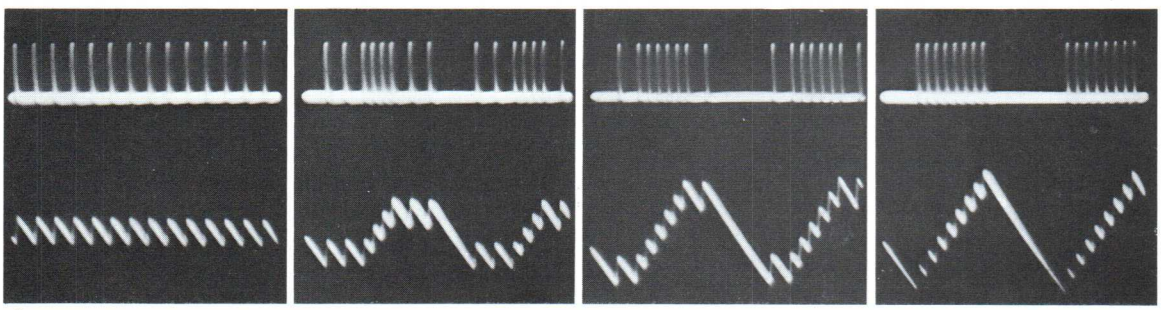


Fig. 6. Block diagram of the receiver. The distorted signal *7* influences the pulse modulator *I.M.* in the manner described under fig. 5. The pulses *8* produced by the pulse generator *I.G.* may be passed on with the correct or opposite polarity and the signal *9* thus produced is converted by an integrator *I* into the echelon curve *10*. From this the continuous-wave signal *11* is formed in the low-pass filter *F*, this signal being the reproduction of the original signal at the input of the transmitter.

equal to half the repetition frequency of the pulses. From the physical point of view this is to be regarded as the price to be paid for the suppression of the interference. Especially for radio relay systems, however, this does not constitute any serious objection, since in this frequency range it is easier to increase the bandwidth than to improve the signal-to-noise ratio.

In practice the pulse modulator is so arranged that only pulses of the values 1 and 0 appear at the output. The input signal is then more or less approximated by a sawtooth instead of an echelon.



a *b* *c* *d*

Fig. 7. The sawtooth-shaped approximating voltage and the corresponding series of pulses transmitted, for an A.F. sinusoidal input signal of 4000 c/s. The rate of sampling of the pulses supplied by the generator is 64 kc/s. *a*) The amplitude of the sine curve is zero; the pulses emitted are alternately 1 and 0. *b*) and *c*) The amplitude is gradually increased, as a result of which the pattern of the series of pulses is changed. *d*) The amplitude has by now become so large that the system reaches complete modulation, and even over-modulation. Even larger amplitudes will not produce any other pulse pattern.

This does not, however, alter the principle of the system.

In *fig. 7* some oscillograms of such an approximating signal are reproduced, together with their corresponding series of pulses. These approximating signals built up from unit pulses are smoothed out at the receiver with the aid of a low-pass filter, as already mentioned, and then give a reproduction of the (in this case sinusoidal) input signal. The frequency of the input signal in the case under consideration is 4000 c/s, the rate of sampling is 64 000 pulses per second. From the four oscillograms it is to be seen how the configuration of the pulses changes with increasing amplitude of the signal transmitted. In fig. 7*a* the input signal is zero. The pulses transmitted are then alternately 1 and 0, and since the sawtooth-shaped approximating signal then has the fundamental frequency of 32 000 c/s, thus containing no frequencies lying within the pass band of the filter, the value of the output signal behind the filter in the receiver will likewise be zero. In figs 7*b* and 7*c* the amplitude has been gradually increased, while in fig. 7*d* the maximum amplitude has been reached that can be observed at the receiver. Larger input amplitudes will not further modify the approximating signal, and thus at the receiver the same output signal will be obtained. It therefore appears that with delta modulation there is a limitation of the amplitudes to be transmitted. This limitation is reached when the signal voltage assumes a certain slope corresponding to the permanent emission of positive pulses. The higher the frequency of the input signal, the smaller is the maximum amplitude of the output signal, as will be shown below.

Other methods for transmitting quantized signals

From the foregoing it is seen that with the method of delta modulation two principles are applied: integration of pulses, and application of a special type of inverse feedback. Owing to the inverse feedback the essence of the method lies in the transmitter responding only to what has been changed with respect to the earlier situation. Each pulse transmitted is only a correction of the preceding signal.

Another system of modulation based on the use of quantized signals is that of pulse code modulation [3]). This system works as follows. As in the case with delta modulation, only the presence or absence of a pulse is of importance. To be able to distinguish more than two amplitude levels with the aid of this "all or nothing" principle, the instantaneous values of the message signal are denoted not by one pulse (present or not) but by a code group of 5 or 7 equidistant pulses each of

[3]) See, e.g., W. M. Goodall, Telephony by pulse code modulation, Bell System Technical Journal **26**, 395-409, 1947.

which may have the value 1 or 0. The position of a pulse in the code group determines its significance. This position principle is met with also in ordinary numerical systems (it was first employed systematically by Indian mathematicians about the year 600). In the decimal system all numbers can be denoted with the aid of only 10 symbols, the position of a symbol in a number determining its value (units, tens, hundreds, etc.). With the method of pulse code modulation, instead of the decimal scale, the simpler binary system is used. The position of a pulse in the code group determines whether it will have the value 1, 2, 4, 8, and so on. In this way, with code groups of five pulses, 32 different signal values can be denoted. This is illustrated in *fig. 8*.

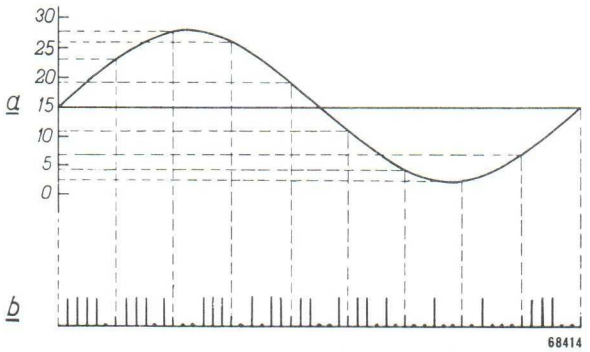


Fig. 8. Principle of pulse code modulation with code groups of 5 pulses. At the beginning of each time interval corresponding to a code group the value of the signal to be transmitted (*a*) is rounded off to one of the discrete values 0 to 31 and "translated" into a pulse pattern (*b*). This pattern is formed by the presence or absence of pulses at the five places marked by dots within the group. A binary system is applied, each place corresponding to a binary value and the measured amplitude thus being expressed according to the binary scale. From left to right the five points in a code group then represent 1, 2, 4, 8 and 16 amplitude "units".

Delta modulation may be regarded as a modified form of pulse code modulation in so far as both systems employ a coding of the signals to be transmitted. Pulse code modulation employs 5 or 7 coding units (sometimes even more), whilst delta modulation, at least in the simple form described here, employs a coding with the aid of one unit, viz. one step in the echelon curve. This is possible on account of the fact that with delta modulation only corrections of the signal are transmitted, whereas in the case of pulse code modulation the complete signal values are transmitted.

With the introduction of pulse code modulation a binary five-unit code came to be applied in telecommunications for the fourth time. In 1837 Cooke and Wheatstone built in London (between Euston and Camden Town) a transmission system of five copper wires connected at the receiving end to a telegraph instrument with five needles registering twenty different letters. In 1926 Rainey obtained a patent for a code system for picture telegraphy which may be regarded as the precursor of the present-day pulse-code modulation systems. The modern telex system, the principle of which was invented by Baudot in 1875, employs a code system usually of 5 units replacing the rather peculiar Morse code of telegraphy (the peculiarity about the Morse code is that the length of the signs is adapted to the frequency in which the signs occur, and therefore has a particular significance).

The systems can be compared only by means of the basic theory concerned with the transmission of information where the concept of information is given a well-defined meaning. This would lead us too far afield, so we shall mention only some of the results reached from the comparison and include in it also the (hypothetical) quantized pulse-position modulation as described on page 239. There are two kinds of interference to be distinguished, the interference which is inherent in the system itself and that which is introduced in the course of the transmission. In the case of non-quantized systems, in theory only a negligible interference (noise) arises in the transmitter and in the receiver, so that if the transmitter and the receiver were directly interconnected the reproduction would be quite free of interference. But, as already pointed out, along the path of transmission there is some interference which is cumulative. If this interference in the course of transmission remains below a certain threshold value then the transmission interference can be entirely eliminated by quantization. As we have seen, quantization means that an "interference" — the quantizing noise — is introduced in the transmitter itself, but this interference does not take part in the cumulation. Moreover, the quantizing noise can be made as small as desired by increasing the band covered by the high-frequency signal transmitted. The transmission interference, however, cannot be reduced.

Quantized pulse-position modulation is characterized by the fact that the threshold value of the transmission interference below which that interference is inactive is relatively low. If the pattern is not to be lost, the pulses may be displaced only slightly in the interval allotted to them. Further it can be shown mathematically that in this system the quantizing noise decreases inversely with increasing bandwidth f.

In coded systems, such as delta modulation and pulse code modulation, it is just the coding that allows of a much greater transmission interference. Furthermore in these two systems the quantizing noise appears to decrease far more with increasing bandwidth, according to a law of $f^{-5/2}$

and of approximately 2^{-f} respectively. This clearly shows the inefficiency, already mentioned, of quantized pulse-position modulation. The same inefficiency is found in all other quantized but not coded systems.

The reason why delta modulation and pulse code modulation are both more advantageous from the point of view of interference suppression than quantized, non-coded, normal systems lies in the use of inverse feedback in the system first mentioned and in the employment of more coding units in the second system. However, also in delta modulation a coding in more units can be applied, for instance by not only judging the difference measurement according to the sign but also dividing the difference into a number of steps and expressing it in the pulse picture transmitted. In this manner the signal-to-noise ratio can be improved for the same bandwidth. It can be shown that in order to obtain the same quality of reproduction fewer units are needed when inverse feedback is applied than when such is not employed.

Properties of delta modulation

Let us now return to the system of delta modulation described. From what has been said above it follows that from the point of view of quantizing noise delta modulation is less advantageous, for the same bandwidth, than pulse code modulation. It appears that in order to obtain a certain quality in the reproduction of speech in the case of delta modulation the band has to be about 50% wider than that required for the same quality with pulse code modulation. On the other hand, however, the apparatus is very much simpler. Not only is there no coding device required, but, since with delta modulation all the pulses are of equal significance, the requirements to be met for the circuit tolerances are less severe; in the difference meter only the polarity of the voltage is determined, so that the amplifiers for the difference voltage need not possess strictly linear characteristics. It is also permissible for a valve adjustment, or for the amplitude of the pulse, to change gradually with time, because in that case the inverse feedback ensures that the signal is not distorted. Finally, the construction of the feedback and integrating networks in the receiver is not very critical; deviations do not give rise to non-linear distortion (as is the case with the networks employed for pulse code modulation) but change only slightly the frequency characteristic of the whole system, which change can be corrected by a simple egalization. In the designing of the complicated quantization apparatus required for pulse code modulation great care has to be taken to ensure that it is linear, whereas for the quantization in the case of delta modulation a simple and essentially linear network is employed.

It would lead us too far to discuss the technical details of the circuit, but mention is to be made of the fact that the frequency characteristic of the network in the feedback circuit can be given an entirely different form for the frequencies above the audio range than for the lower frequencies, thereby reducing the quantizing noise.

Both with pulse code modulation and with delta modulation the possibilities are limited: the signals to be transmitted must not be too large, while at the receiver very small signals cannot be distinguished from the zero signal. In the case of pulse code modulation this limitation is governed by the size of the steps in the 5- or 7-unit code, the smallest signal to be transmitted corresponding to a code group with the "signal value" 1, and the largest corresponding to the transmission of a code group with the signal value 32 (5-unit code) or 128 (7-unit code). Thus the limitation is independent of frequency.

As already seen from fig. 7*d*, the limit in the case of delta modulation, for the large signal values, is reached when the signal voltage has a certain slope corresponding to the permanent transmission of positive pulses. If, however, the signal voltage is so small that notwithstanding the presence of the signal the approximating voltage retains the shape depicted in fig. 7*a*, and thus the pulses are alternately transmitted and suppressed, then nothing is to be perceived of the signal behind the low-pass filter in the receiver. By means of special circuits the smallest signal to be reproduced (threshold value) can be made somewhat smaller than what follows from this reasoning, but the minimum amplitude that can be observed at the receiver always remains finite.

As far as reproduction of the sound is concerned, the important consequence of the signal amplitude being limited owing to the existence of a maximum slope in the case of delta modulation is that the maximum amplitude that can be transmitted is inversely proportional to the frequency. Since in the case considered the threshold value of the signals received does not depend upon the frequency, this means that the dynamic range, i.e. the ratio of the maximum to the minimum amplitudes at the receiving end, becomes smaller as the frequency is raised.

This deviation from the ideal reproduction of the audio signal happens to coincide with a property inherent in the reproduction of speech and music and in the receptive qualities of the human ear, so that as far as its physiological effect is concerned it does not adversely influence the transmission: the maximum amplitude of the sound pressure in the audio spectrum decreases with increasing frequency, as is likewise the case with the dynamic range perception of the ear (especially at high frequencies).

A third quantity, one that does not occur in normal acoustics, is the gamma of the transmission system, by which is to be understood the number of distinguishable steps between the maximum and minimum amplitudes transmitted. With delta modulation the gamma becomes smaller as the frequency increases, and although physiological experiments have shown that in the range of speech frequencies this should lead to a deterioration in the quality of reproduction, in practice this is not evident.

For frequencies above 6000 c/s there is a reduction of the gamma with increasing frequency also in hearing. At these frequencies delta modulation therefore matches the properties of the human ear in two respects.

This brings us to the following highly important point. An optimum transmission system is characterized by the fact that it transmits what is to be transmitted and nothing more than that. When the signal power decreases as a function of the frequency then any possibilities offered by the system for the transmission of large powers at high frequencies are superfluous and undesired. When, as appears to be the case, the human ear perceives smaller dynamic ranges at the higher frequencies, in the optimum system it would be impossible for larger dynamic ranges than desirable to be transmitted. This explains why in the case of delta modulation a relatively small bandwidth suffices. The matching of this system to the peculiarities both of the speech signal and of the human ear partly neutralizes the mathematical deviations from the ideal reproduction.

The deeper significance of delta modulation

Being in essence based upon inverse feedback, delta modulation belongs to the family of regulating systems and servo-mechanisms, but with its quantization both in amplitude and in time it is a member that has been little studied.

As already stated, only corrections of the preceding signal are transmitted, this being brought about with the aid of the integrating network. Mathematical integration, however, would mean that even the distant past would contribute towards the present amplitude. Considering that there is a definite limit set to the time within which earlier signal values of a speech or music signal are related to the instantaneous value, those earlier values may gradually be forgotten. A resistance-shunted capacitor (as is present in our integrating *RC* network) thus supplies the first useful variation of the purely mathematical integration. A second variation consists in the fact that by adding more elements to the network an attempt is made to make the output signal correspond to the most probable continuation of the signal to be transmitted (that continuation where the first or more time derivatives of the signal voltage do not change with time). All the transmitter need do is then to transmit, via the difference meter, the corrections with respect to this most probable continuation.

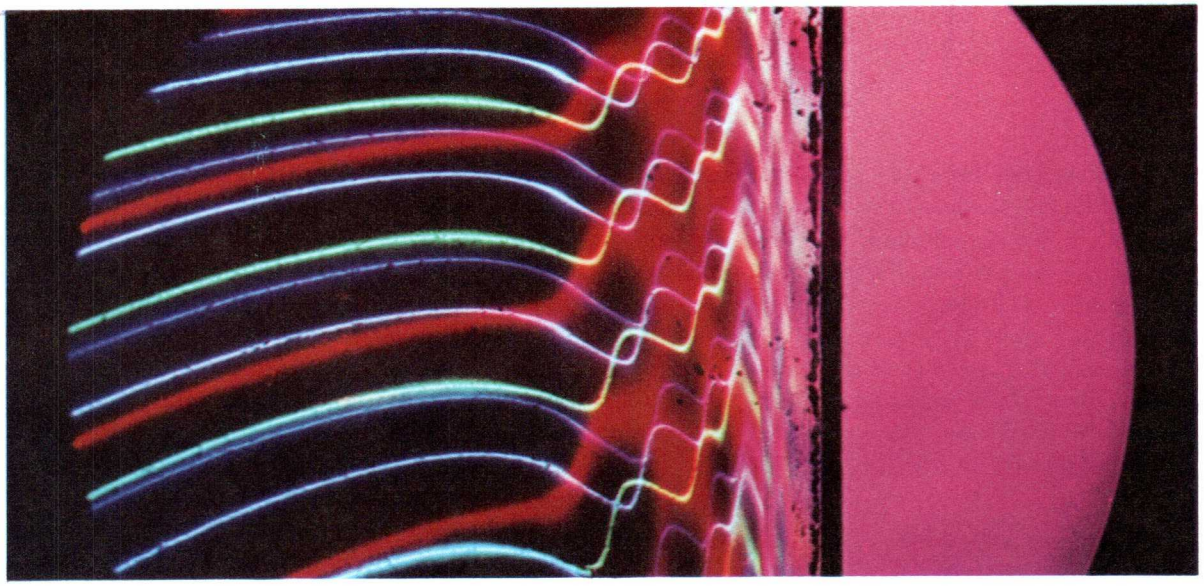
This can also be formulated in another way. Each audio or video signal — whether speech, music or picture — is characterized by certain correlation properties, i.e. by a certain relation existing between the successive signal values. Any transmission of the known part of the signal, thus of that part of the amplitude that follows from the correlation properties, is superfluous, for the very reason that it is already known. In any optimum transmission system, therefore, these correlation properties should be discounted at the transmitter and only what is to be regarded as new and individual should be transmitted.

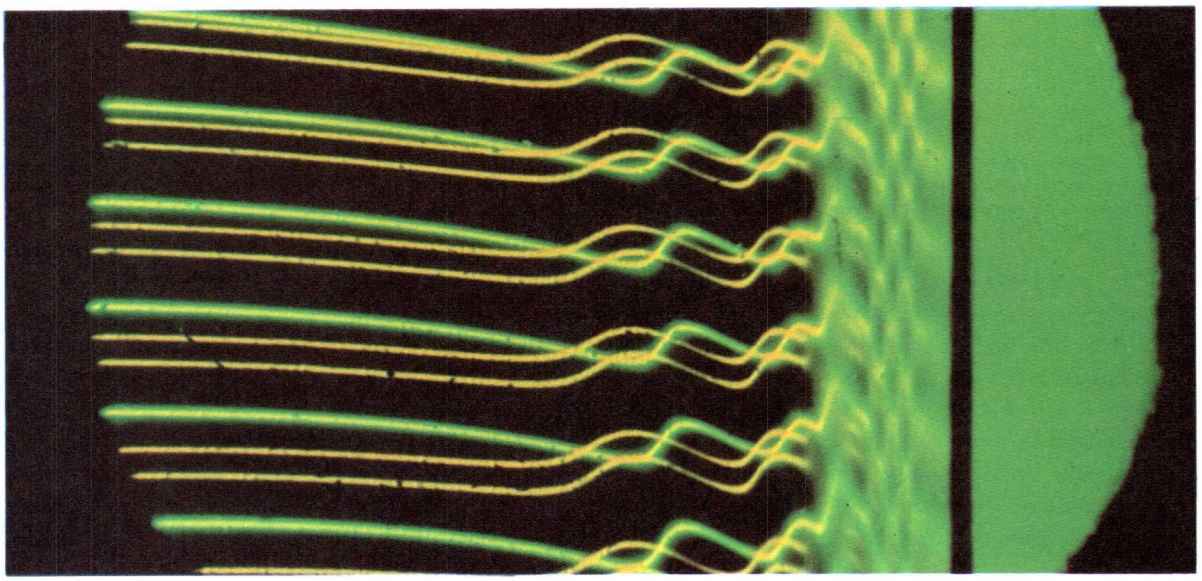
This implies that the signal transmitted in the case of an optimum transmission system possesses the statistical properties of noise, thus having no correlation [4]).

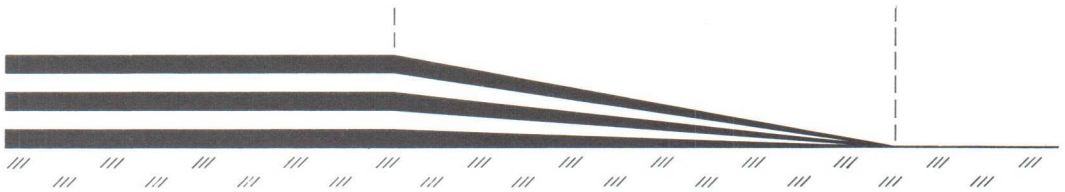
It is the function of the feedback circuit to take into account these correlation properties, which means, therefore, that the properties of the feedback network must be determined by the nature of the signal to be transmitted, on the one hand, and by that of the ear (or eye) on the other hand.

[4]) S. Goldman, Proc. Inst. Radio Engrs **36**, 584, 1948.

Summary. In this article the development of different modulation systems is described, a development which aims at achieving the transmission of audio or video signals as free as possible from interference. The interference occurring in carrier modulation systems and also in pulse-position modulation arises from very small effects of interference being cumulative in the course of transmission. This cumulation can be eliminated by quantizing the signals to be transmitted, both in time and in amplitude. The system of delta modulation, based on this principle, is described and compared with other systems employing quantization (pulse code modulation). It appears that for a good reproduction, using a reasonable bandwidth for the transmitted signal, the apparatus required for delta modulation is relatively very simple, because of coding being applied with the aid of only one unit. This is possible by reason of the fact that with delta modulation account is taken of the correlation properties of the signal, by employing a special kind of inverse feedback in the transmitter, and that the properties of the system are matched with those of speech and music and of the human ear.







Fizeau fringes observed with an interference mirror

When a monochromatic beam of light is directed upon an interference mirror (consisting of thin dielectric layers of alternate high and low refractive index) which is placed a short distance from a half-silvered mirror, a pattern of relatively sharp interference fringes (Fizeau fringes) can be seen in transmission. These fringes represent lines of constant (optical) thickness of the space between the mirrors, and can therefore in general never intersect, even when several wavelengths are used.

The photographs show the fringes for several different wavelengths at the same time (*above*: cadmium lines 468-480-509-644 nm; *below*: green mercury line 546 nm, and yellow mercury doublet 577-579 nm) in the neighbourhood of the edge of an interference mirror, where the layers gradually taper to zero thickness (see drawing). It can be seen that here the fringes for different wavelengths intersect. This remarkable effect can be explained in terms of the wavelength dependence of the phase with which a multilayer mirror reflects the light (phase dispersion).

(From: Philips tech. Rev. **23**, 380, 1961/62. Further reading: G. Bouwhuis, A dispersion phenomenon observable on dielectric multilayer mirrors, Philips Res. Repts. **17**, 130-132, 1962.)

ELECTRONIC INTENSIFICATION OF FLUOROSCOPIC IMAGES

by M. C. TEVES and T. TOL. 621.386.8:616-073.75:621.383.8

In essence, the generation of X-rays, which entails the production of a beam of electrons in a vacuum tube, falls within the technique of electronics. It is a noteworthy point that in X-ray diagnostics the subsequent stage, that is, the conversion to a visible picture of the X-rays which have passed through the object to be examined, is now also tending to become dependent on electronics for a satisfactory solution. In this context it is found that the number of X-ray quanta available virtually sets a limit to the amount of detail in the information that can be derived from fluoroscopy.

The problem of screen brightness in fluoroscopy

X-rays have the capacity to damage or destroy human tissue. In X-ray therapy, use is made of this fact by exposing morbid tissue to these rays for such periods and at such intensities that in total it will receive a certain, heavy, dose. In X-ray diagnostics, on the other hand, care must be taken that patients receive the smallest possible dose during examination so that they will not suffer any injurious consequences, even as a result of repeated examinations.

Using the conventional method of fluoroscopy, therefore, the radiologist must do the best he can with an X-ray image of very low brightness. In fact, for a given duration of an examination (a few minutes at the most), limitation of the total dosage means limitation of the dosage rate to which the patient may be subjected and, accordingly — dependent upon the absorption of the rays by the patient — limitation of the brightness that can be obtained from the screen. In gastric fluoroscopy, for example, with the X-ray tube operating on a voltage of, say, 90 kV and a current of 2 mA, the image obtained on a high-quality screen has an average brightness of about 0.003 cd/m^2. For comparative purposes it may be said that the average brightness of a landscape by the light of the full moon is 0.01 cd/m^2.

In order to be able to make a good visual assessment at such low brightness values, the radiologist must first remain in darkness for at least 15 minutes, so that his eyes may be adapted to the low brightness level. Even so, his visual acuity and contrast sensitivity [1]) are then fairly low. Hence, for the diagnosis of conditions involving slight abnormalities which often reveal only a low contrast, this is a very inconvenient limitation. It will therefore be easily appreciated that, since the inception of X-ray diagnostics, higher image brightnesses have been an ever-present need.

If more information is to be obtained from the fluoroscopic screen and at the same time a solution found for the problem of the necessity for dark adaptation of the eyes, an increase in image brightness to the extent of at least a factor of 100 to 1000 is needed, as pointed out by Chamberlain in 1942 [2]). It would not be possible to achieve this simply by increasing the efficacy of existing types of screen; these convert roughly 5% of the X-rays falling on them to visible radiation, and this radiation has approximately the most advantageous wavelength (i.e. in the yellow-green part of the spectrum, where, for a given strength of the radia-

[1]) Visual acuity is defined as $1/a$, where a is the angle subtended by the finest distinguishable detail. Contrast sensitivity is defined as $1/c$, where c is the lowest distinguishable contrast $\Delta B/B$; in this, B is the brightness of the background and $B \pm \Delta B$ the brightness of an object seen against that background.

[2]) W. E. Chaimberlain, Fluoroscopes and fluoroscopy, Radiology **38**, 383-413, 1942.

tion, the strongest impression of brightness is received by the eyes). Thus, the gain that is theoretically possible in this way would not exceed a factor of 20.

In recent years several methods, using electronic aids, have in fact been devised for intensifying the brightness of the fluoroscopic image to a considerable degree [3]). In the present article we shall describe an image intensifier designed in the Philips Laboratories at Eindhoven. This equipment is based on the well-known principle of the image converter, another development of this laboratory, a description of which was given as early as in 1934 [4]). The fluoroscopic images obtained with the aid of this instrument are roughly 1000 times brighter than those produced on the ordinary fluoroscopic screen.

Description of the intensifier

The more important details of the unit are illustrated in *fig. 1*. A fluoroscopic screen on a support of thin aluminium is mounted in an evacuated glass tube. A photocathode lies in contact with the screen. When X-rays pass through the glass wall of the tube to the screen, this fluoresces. The fluorescent light sets electrons free from the photocathode; at any point on the photocathode the number of photo-electrons released per second is proportional to the luminous intensity of the fluoroscopic screen at that point. The luminous image with all its variations in brightness is thus transformed into an "electronic image" with corresponding variations in current density. By means of an electric field of a certain configuration (electrostatic "electronic lens") this electronic image is reproduced on a second fluorescent screen, the viewing screen, reduced 9 times in size. The energy of the electrons falling on this screen is in part converted to fluorescent light, and a reproduction of the image on the fluoroscopic screen is thus seen on the viewing screen reduced 9 times. This reduced image is viewed through a simple microscope of about 9 × magnification, and finally the fluoroscopic image is thus seen erect and in its original size (both the electronic lens and the microscope produce inverted images).

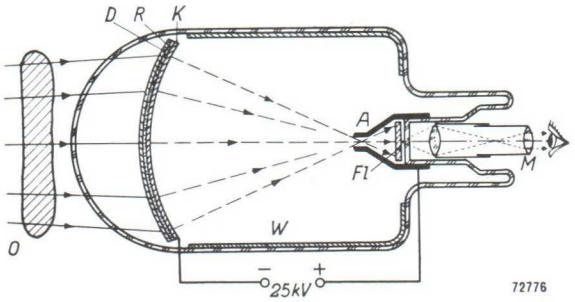


Fig. 1. Diagrammatic cross-section of the image intensifier. R = fluoroscopic screen on which the X-rays are received after passing through the object to be examined O, and the glass wall of the tube. D = support of fluoroscopic screen and photocathode K. The fluorescent radiations produced in R release electrons from the photocathode K. The "electronic image" is reproduced on a reduced scale by the electric field between K and the perforated anode A, on a viewing screen Fl; this image is observed through a simple microscope M. W is a conductive layer on the inside of the tube.

The image seen through the microscope is thus identical with that seen on an ordinary fluoroscopic screen; it is, however, about 1000 times brighter, assuming that the image intensifier and the fluoroscopic screen both receive X-rays of the same intensity.

How is the brightness intensification obtained?

Two factors (which are not in fact independent of each other) contribute towards the intensification of the image brightness.

The first and more readily understood factor is an increase in the total luminous flux. The electrons emitted by the photocathode are not only focused on the viewing screen by the electrical field, they are also accelerated by it; in the particular instrument under review a potential of 25 kV is applied between the viewing screen and the photocathode. The more energy possessed by an electron on arrival at the screen, the more fluorescent light it will produce. Notwithstanding the fact that only about 1 in 10 of the light quanta from the fluoroscopic screen liberates a photo-electron and that only some 10% of the electronic energy is converted to light by the viewing screen, the last mentioned screen — owing to the energy imparted to the electrons — is able to yield a total of 10 to 15 times as much luminous flux (in lumens) as an ordinary fluoroscopic screen, for the same object viewed.

The second, more important factor is the gain produced by the above-mentioned electron-optical

---

[3]) J. W. Coltman, Fluoroscopic image brightening by electronic means, Radiology **51**, 359-367, 1948.
R. J. Moon, Amplifying and intensifying fluoroscopic images by means of a scanning X-ray tube, Science **112**, 389-395, 1950.
R. H. Morgan and R. E. Sturm, The John Hopkins fluoroscopic screen intensifier, Radiology **57**, 556-560, 1951.
P. C. Hodges and L. Staggs, Electronic amplification of the röntgenoscopic image, Amer. J. Röntgenology Rad. Ther., **66**, 705-710, 1951.

[4]) G. Holst, J. H. de Boer, M. C. Teves and C. F. Veenemans, An apparatus for the transformation of light of long wavelength into light of short wavelength, Physica **1**, 297-305, 1934.

reduction of the size of the image. All the electrons coming from the photocathode contribute towards the final image, and the energy that each electron is capable of converting to light after passing through the electron-optical system is independent of whether the electrons are distributed over a large or over a small area. By employing a reduction of 9 times we concentrate the electrons in an area 9^2 times smaller, and the total luminous flux is thus emitted from an area that is about 80 times smaller than would be the case in reproduction on a scale of 1 : 1. This, by definition, represents an increase in brightness [5]) by a factor of 80.

The total gain in brightness is equal to the product of the "lumen gain" and the gain due to the reduction in size, i.e. in the tube in question, 10 to 15 times $9^2 = 800$ to 1200.

It is perhaps not superfluous to add that the gain brought about by the electron-optical reduction cannot be obtained by the ordinary methods of optical reduction. According to Abbe's law, the apparent brightness of an object does not vary as a result of the introduction of any kind of optical instrument between the object and the eye of the observer, that is, if such an instrument can be regarded as ideal, without the absorption and reflection losses that normally occur in lenses and mirrors (and assuming the exit pupil of the instrument to be at least as large as the pupil of the eye).

This will readily be appreciated on reference to *fig. 2a*. Let the brightness of the object be B, and the solid angle subtended by the lens when seen from the object ω. The lens then receives from a surface element df a luminous flux of $\Phi = B\omega df$. This luminous flux is concentrated in a smaller area df' in the image space. This area thus again emits the luminous flux Φ, this time, however, within the greater solid angle ω' (or, with diffusion, in an even larger angle). The brightness of df' is $B' = \Phi/\omega' df'$ and, since we know from the laws of optics that $\omega' df' = \omega df$ (the sine law), B' is equal to B.

The difference that is seen to exist in this case by no means implies that the analogy between electron-optics and light-optics is no longer valid. Abbe's law as stated in the above simplified form holds good only when the object and the image are located in media having the same refractive index, and in ordinary optical instruments this is usually the case. If the refractive index of the medium containing the image is higher than that of the object, the brightness of the image can certainly be greater than that of the object itself. This is exactly what happens in electron-optics. The potential field in which the electrons are moving may be regarded as a medium the refractive index of which is continuously variable; in fact, the index is at any point proportional to the velocity of the electrons. The refractive index at the site of the viewing screen is therefore about 100 times higher than that at the photocathode.

The analogy is thus not only restored formally, but may again be illustrated in a simple way. Owing to the continuous variability of the refractive index, the "rays" are now not straight, but curved. The path of the rays in the electrostatic electron-lens therefore takes the form shown in fig. 2*b*, in contrast with fig. 2*a*; it will be seen that the flux radiated from the surface element df within a solid angle almost equal to 2π is employed and is concentrated in a much smaller solid angle at the surface df'.

By reason of the influence of the velocity of the electrons, it is seen that the "lumen intensification" and the brightness gain due to the reduction in size virtually spring from the same cause, i.e. the presence of the accelerating field.

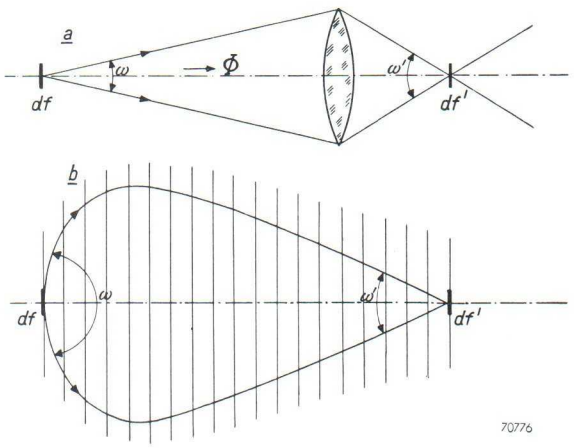


Fig. 2. *a*) Path of rays on optical reproduction of a surface df, at df' (illustrating Abbe's law).
b) The same, as applicable to electron-optical reproduction with an electrostatic lens, as in the image intensifier.

In the light of Abbe's law it will now be understood why in the last stage of the image intensifier, viz. the optical enlargement by the microscope of the image on the viewing screen, the gain in brightness arising from the electron-optical reduction is not neutralised (provided that the microscope meets the condition as regards the exit pupil to which reference has already been made).

It may possibly strike the reader that in the foregoing, particularly in our explanation of the lumen gain, the luminous flux of the viewing screen is each time compared with that of an ordinary fluoroscopic screen and not, as might appear more obvious, with the luminous flux of the fluoroscopic screen in the image intensifier itself. The reason is that the last-mentioned screen need not be observed and that the luminosity curve of the fluorescent light emitted by it may therefore be such as to be quite unsuitable from the point of view of direct observation; in fact the fluorescent

[5]) According to a recent standardisation decision on lighting technology terms, we should speak here of luminance.

radiation might even be entirely invisible, for instance ultra-violet or infra-red, provided only that the photocathode be sensitive to such radiation. Our image intensifier is equipped with a photocathode the quantum efficiency of which is highest (about 10%) when it is exposed to blue light. A phosphor giving blue fluorescence was accordingly chosen for the fluoroscopic screen. For the viewing screen, on the contrary, if a choice from different materials of the same luminous efficiency is available, that colour should be employed of which the fluorescent spectrum coincides as nearly as possible with the maximum in the spectral sensitivity curve of the eye. This would be a phosphor emitting yellow-green light. Clearly, if we compared the brightness of such a viewing screen with that of the blue emission of the fluoroscopic screen in the image intensifier, we would calculate a much greater brightness intensification value than 1000 times, but at the same time one that would have no practical significance.

For this reason it would be as well to dwell for a moment on the precise meaning of the name "image intensifier" or "brightness intensifier". In effect the instrument is an image converter, which converts the "image" of the invisible X-rays to a visible image, by analogy with the well-known instruments for transforming infra-red radiations into visible images (see the article referred to in note [4])) which are employed in numerous different forms. The fact that this conversion in our case involves as intermediate stage a visible fluorescent image (as in normal fluoroscopy) is attributable only to the impossibility of liberating photo-electrons direct from a photocathode with any degree of efficiency (and with not too high an initial velocity) by means of the X-rays themselves. Reasonable efficiency in this release of electrons is naturally of the highest importance in the fluoroscopic image intensifier, as otherwise the effect of the subsequent gain would be lost and we might then just as well employ only the "image conversion" as obtained from the ordinary fluoroscopic screen.

Details of construction

The blue fluorescence of the fluoroscopic screen as matched to the photocathode is obtained from zinc sulphide activated with silver, and the yellow-green fluorescing viewing screen is prepared from very finely divided zinc sulphide-zinc selenide. The effective diameter of the cathode face is 13.5 cm, this thus representing the maximum size of the object to be examined. The image on the viewing screen is only 1/9th of this size, viz. 1.5 cm across.

The electrical field producing the electron-optical image is generated basically by means of only two electrodes, one of which (the cathode) is formed by the concave spherical surface of the combined fluoroscopic screen and photocathode, as well as by a conductive layer on the inside wall of the glass tube. The other electrode (the anode) is also spherically curved and is concentric with the photocathode; it has in it an aperture to allow the electrons to pass to the viewing screen behind it [6]). Between the cathode and the anode a potential of 25 kV is applied.

This simple design, employing only two electrodes, has proved to be very effective. Definition in the image is ample and the field strength on the only slightly curved photocathode is very low, this being an asset for a reason which is given in another paragraph. When the tube was still in the experimental stage, the configuration offered sufficient latitude, viz, in the radii of curvature of cathode and anode, allowing for example a variation in the amount of reduction of the image. For practical purposes it was later found to be better for the inner lining of the tube to function as a separate electrode and to maintain it at a slightly positive potential with respect to the photocathode. Without this, only the electrons emitted from the central part of the photocathode pass through the opening in the anode, the peripheral parts thus not being reproduced. With the auxiliary potential applied the peripheral electrons also pass through the anode, the effective image area of the fluoroscopic screen being thus almost doubled. At the same time, the slight "pin-cushion" distortion of the image that occurs without the auxiliary voltage is in this way almost entirely eliminated. Another important advantage of the auxiliary potential — which is variable — is that by means of it the image on the viewing screen can be adjusted for maximum definition whilst the tube is in use. Consequently, the tolerances in the electrode spacing, during the manufacture of the glass tube, can be relatively wide.

The 25 kV applied between cathode and anode may be obtained for instance from the well-known small high-tension generator developed in this Laboratory for projection television [7]). This generator also supplies the necessary voltage for the inner tube coating. The cathode is earthed, so that only one of the three poles of the image intensifier need be suitable for carrying high voltage. Only a very small amount of power is required to operate the tube, the photo-current for normal fluoroscopy being in the order of 10^{-9} A. For this reason and also in view of the fact that in this case no stringent con-

[6]) For a more detailed description of the principle of this two-electrode lens the reader is referred to an article by P. Schagen, H. Bruining and J. C. Francken, Philips Research Reports **7**, 119-130, 1952 (No. 2).

[7]) G. J. Siezen and F. Kerkhof, Philips Techn. Rev. **10**, 125-134, 1948.

ditions need be imposed as regards voltage stability (a basic advantage in purely electrostatic electron-optical systems), an even simpler high-tension unit would suffice.

Fig. 3 depicts the complete image amplifier tube. It is cylindrical, the diameter being 17.5 cm; the over-all length, including the small microscope, is 45 cm. To protect the tube against jarring and to afford local protection from stray X-rays, the tube proper is contained in a lead-lined aluminium jacket which can be mounted quite simply on a stand; see *fig. 4*. Tube and jacket together weigh 7.5 kg approx.

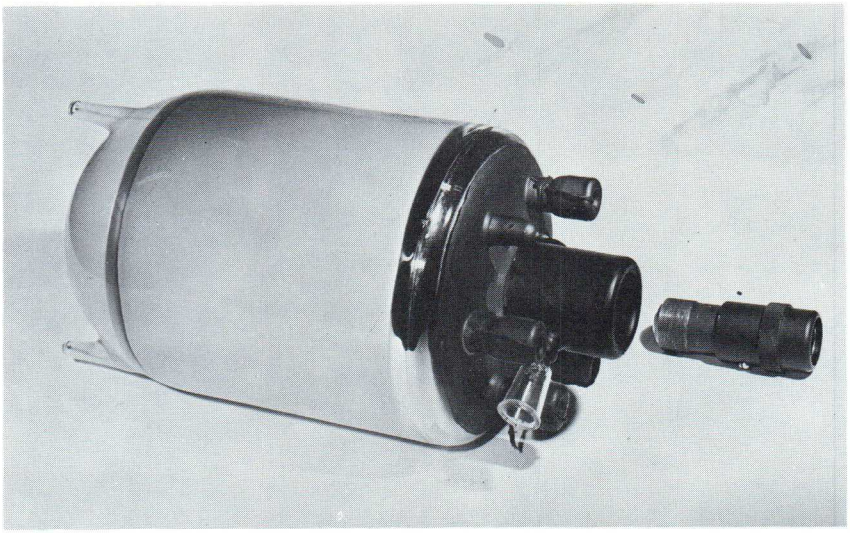


Fig. 3. The Philips X-ray image intensifier, producing an image that is approximately 1000 times brighter than that of the ordinary fluoroscopic screen. The tube is 45 cm in length, 17.5 cm in diameter. The effective area of the screen is a circle 13.5 cm in diameter. On the right will be seen the viewing end into which the microscope shown alongside is fitted.

Definition and contrast in the image

When the image intensifier is in use it is not necessary for the radiologist to allow time for his eyes to adapt themselves; examinations can be made in a room with almost normal illumination. This is in itself a great practical advantage that will almost certainly outweigh the drawback that the image cannot be surveyed with the naked eye, but only through an optical system. Moreover, it may be expected, as explained in the introduction, that the radiologist will be in a position to perceive smaller details than is possible with the ordinary fluoroscopic screen because of the higher visual acuity and contrast sensitivity of the eye at the increased brightness level. In order that this expectation — which we shall enlarge upon presently — may be realised, it is an essential condition that the blurring and loss of contrast introduced by the image intensifier itself, as by any other form of optical instrument, shall be kept within certain limits.

Definition

In normal fluoroscopy the definition of the image will be determined by the finite dimensions of the focus of the X-ray tube (half-shadow width or geometrical blurring) as well as by the thickness and grain size of the fluoroscopic screen (intrinsic or screen blurring). Screen blurring is usually about 0.7 mm, sometimes less, down to 0.4 mm; the geometrical blurring is generally slightly less. The resulting definition in the fluoroscopic image is thus somewhat less than 1 mm (the total blurring being equal, approximately, to the square root of the sum of the squares of the geometric blurring and the screen blurring [8]).

Screen blurring can be reduced considerably by using thinner screens (and finer grain). This is not done in the case of ordinary fluoroscopy, as such screens would transmit a larger amount of X-rays, which are thus ineffective, resulting in reduced screen efficiency [9]). Moreover, better definition would be

[8]) See H. A. Klasens, The blurring of X-ray images, Philips Techn. Rev. **9**, 364-369, 1948.

[9]) As far as efficiency is concerned there is an optimum thickness, for in screens that are too thick, apparently too much absorption of the emitted light will take place in the screen itself. (This conflict between useful absorption of the incident radiation and wasteful absorption of the rays emitted in the image intensifier again occurs in the photocathode and in the viewing screen.)

of little use, since, at the low brightness levels of normal fluoroscopic screens, visual acuity is in any case too low for the perception of finer detail.

In the image intensifier the position is very different. Here, too, there is a primary lack of definition in the image owing to the intrinsic blurring in the fluoroscopic screen employed. However, there is now no objection to the use of a relatively thin screen, thus sacrificing some of the gain in brightness to benefit the definition. The grain size of the viewing screen used is very fine and the screen itself very thin (the penetration depth of electrons at 25 kV is about 12 μ), so that the intrinsic blurring in this screen, referred to natural image size, is less than that of the fluoroscopic screen. Blurring by the electron-optical reproduction system is at least one order of size less (see formula (26) in the article referred to in note [6])) and can therefore be ignored, as can, also, the limitation in the resolving power of the microscope



Fig. 4. The complete equipment for fluoroscopy by means of the image intensifier. The tube is housed in a lead-lined aluminium jacket and the whole can be conveniently mounted on a stand. Left: the small high-tension unit for supplying the necessary 25 kV, D.C.

A real advantage is obtained by so doing, for the visual acuity at the brightness level of the image intensifier is so high that the eye can reap the benefit of the increased definition.

For this reason a thinner fluoroscopic screen is used in the image intensifier than is provided for ordinary fluoroscopy. On the other hand the reduction in the screen thickness cannot be carried too far, for a reason other than that connected with efficiency, as will be seen presently.

In the case of the image intensifier there are also other possible causes of blurring. These must of course not detract from the available gain in definition, and in fact, in our model, they do not. through which the screen is viewed. No appreciable blurring occurs in the photocathode either, since this is in direct contact with the fluoroscopic screen, and, in order to avoid too much absorption of the liberated photo-electrons, is extremely thin.

Owing to these favourable properties, the over-all blurring effect produced by the image intensifier is not more than 0.4 mm. In order that the advantage of this low "intrinsic" blurring may not be lost, the previously mentioned geometrical blurring must also be limited as much as possible. It is therefore desirable to use in conjunction with the intensifier an X-ray tube having a very small focus (fine focus tube).

Contrast

On passing through the subject to be examined, the X-rays are attenuated to an extent that varies from one point to another, and it is the resultant "radiation contrast" that supplies the required information regarding the subject, i.e. the image. The processes involved in the conversion of the X-rays to fluorescent light, then into photoelectrons and finally into light again, are wholly linear, owing to the very small electronic current occurring in the image intensifier. In principle, therefore, the contrasts in the image are unchanged (gamma = 1), in contradistinction to, for example, photographic reproduction (in which case gamma may be as much as 2.5 to 3).

For various reasons, however, the image intensifier does occasion a certain amount of fogging, i.e. a brightness contribution more or less uniformly distributed over the whole image, which results in a loss of contrast. The glass window of the tube through which the X-rays enter partly scatters the rays and thus delivers diffuse radiation to the fluoroscopic screen; similarly those X-rays which are not absorbed by the fluoroscopic screen but which fall on the wall of the tube are in part rediffused towards the screen. Again, a fraction of the blue fluorescent light which the very thin photocathode is not capable of absorbing is reflected from the inner wall of the tube, back to the photocathode, where it liberates electrons in the wrong places; up to a point this also applies to the persistence of the fluoroscopic screen. Furthermore, the photocathode exhibits a certain amount of thermionic emission which is independent of the intensity of illumination received by it; field-emission is also liable to occur at the photocathode. Both these effects contribute additional light uniformly spread over the whole of the viewing screen. A proportion of the total number of electrons arriving at the viewing screen is reflected there and tends to liberate positive ions from the walls of the anode space (as well as from possible residual gases), and these ions may be diffused in the direction of the anode aperture where they are accelerated by the field in the direction of the photocathode; here they may liberate more electrons and this will also produce additional diffuse light. Finally, at the viewing screen there is also some diffusion of the light generated by the screen itself, resulting in a kind of halo which is intensified by repeated reflection in the glass support of the screen, whose granules are in contact with it.

Very severe fogging would occur if a part of the light from the viewing screen were to return to the photocathode (a kind of "feedback"), but this is prevented by coating the rear of the viewing screen with a thin layer of aluminium which transmits practically all the high-velocity electrons, but completely intercepts light that would be reflected back [10]).

Suitable precautions have also been taken to reduce to a minimum all the other sources of fogging summarised above. As a result the loss in contrast is kept within reasonable limits. We will mention only one or two of these precautions. For example, the inner lining of the tube is of a kind that does not reflect blue light too well. The shape of the electrodes is such that high field strengths do not occur at the photocathode, and field emission is accordingly low. For the same purpose and also in order to limit thermionic emission, the threshold potential of the cathode should be fairly high.

Accurate data regarding the degree of fogging cannot be given at this time, but for orientation purposes we may assume that this represents 10% of the total luminous flux which the viewing screen would emit without the fogging. (This value is probably not very far from the actual figure.) From the aspect of contrast this means that where we would have had to observe contrast between a brightness of 105 and an average of 100 for example, we will now have a contrast of 115 : 110. The loss is thus of little importance. When contrast is very pronounced the depreciation is greater (10 as against 100 would become 20 to 110), but such contrasts are generally easily distinguished in any case.

Fundamental limitation of minimum perceptible subject detail

Let us now ascertain in how far the radiologist, using the image intensifier, really will be able to perceive more detail, that is, details of smaller dimensions and lower contrasts, than with the ordinary fluoroscopic screen. The ultimate answer to this question can be obtained only with the intensifier in actual use. While, however, insufficient medical experience has been gained as yet with the instrument to enable one to express in numerical values the advance which the introduction of the intensifier undoubtedly represents, it is possible to show theoretically that the minimum detail to be perceived cannot decrease below a certain limit. This is attributable to the circumstance that X-rays are in the form of quanta; attention was first drawn to this consequence by the scientists Sturm and Morgan [11]).

The energy quanta, or photons, of magnitude

[10]) See J. de Gier, Philips Techn. Rev. **10**, 97-104, 1948.

[11]) R. E. Sturm and R. H. Morgan, Screen intensification systems and their limitation, Amer. J. Röntgenology Rad. Ther. **62**, 617-634, 1949.

$h\nu$, of which all radiations consist (h = Planck's constant and ν = frequency of oscillation), are not emitted by the source perfectly regularly, nor are they absorbed regularly by a medium such as the fluoroscopic screen or the retina of the eye. It is possible to speak of an average number N of quanta per given interval of time, on a given surface, and this determines the intensity of the radiation emitted or absorbed. In the case of a number of such time intervals, however, the number of quanta within each interval will not be equal to $\overline{N}$; random fluctuations about this average occur, whose order of magnitude, computed on the basis of probability, is given by the so-called standard deviation $\sqrt{\overline{N}}$. The order of magnitude of the relative fluctuations in the radiation is defined by the coefficient of variation, which is equal to $\sqrt{\overline{N}}/\overline{N} = 1/\sqrt{\overline{N}}$.

These natural fluctuations, the "noise" in the radiation, set a fundamental limitation on the detail which the radiation is capable of conveying, or, in more concrete terms, on the minimum radiation contrasts which can be communicated by means of the quanta absorbed.

This argument is immediately applicable to the human eye. Let us take as starting point a surface element of a given object, with a size of 0.1 mm². Those quanta will contribute towards the perception of brightness which travel from this surface element to the pupil of the eye within the "accumulation time" T of the eye. Only a fraction of these is effectively absorbed in the retina, i.e. produce the stimulus conveyed to the brain by the optic nerve. The fluctuations in this number of quanta can be regarded as fluctuations, in effect, in the observed brightness of the surface in question, and it is found that the difference in brightness between this surface and its background definitely cannot be perceived if the contrast is not at least 3 to 5 times as great as the relative fluctuation already mentioned. The lower limit of perceptible contrast is thus expressed in terms of the "noise" in the light quanta effectively absorbed. It will now be seen at once that as the size of the surface observed is increased (i.e. more quanta are permitted to contribute to the sensation of brightness), so also does it become possible to appreciate lower contrasts; see for instance the well-known contrast-detail diagrams obtained in the fluoroscopic examination of phantoms (artificial test-objects), *fig. 5* [12]). It will also be clear that when the brightness level is raised, contrast sensitivity increases. However, this takes place more slowly than might be concluded from the decrease in the relative fluctuations, since, when the brightness increases, the percentage of light quanta effectively absorbed drops (owing amongst other things to a saturation effect in the optic nerve), as also does the accumulation time T.

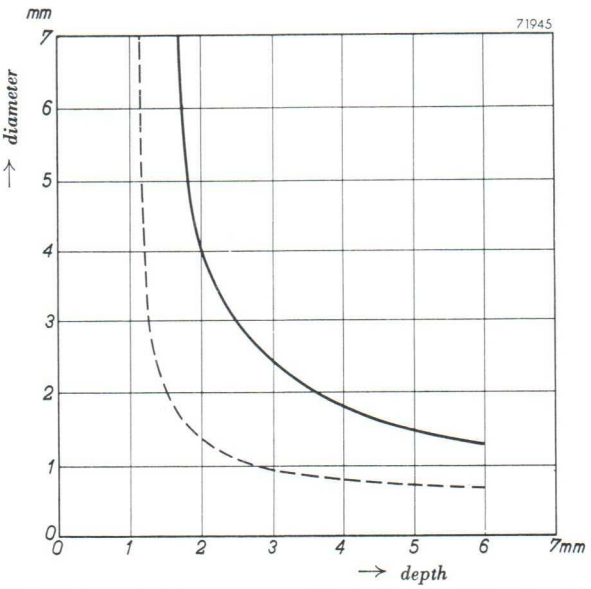


Fig. 5. Example of a contrast-detail diagram as obtained with the aid of an X-ray "phantom". The "phantom" consists of a number of plates of "Philite", one of which has in it a number of holes of different diameters and depths. The full line represents as a function of the hole diameter, the depth of a hole (contrast!) necessary to make it just visible in normal fluoroscopy. (The dotted line shows the results of provisional tests with the image intensifier under the same conditions.)

Let us now turn to X-ray diagnostics, again taking a surface of 0.1 sq. mm of the fluoroscopic screen. In chest examinations 100 000 X-ray quanta fall on such an area per second. Of this the fluoroscopic screen absorbs 65%. The accumulation time T of the eye in normal fluoroscopy is about 0.1 sec. The number of X-ray quanta involved in providing information about details of the above-mentioned size of the lungs under examination is therefore about 6500. One X-ray quantum of 70 kV absorbed liberates an average of 5000 light quanta in the screen. Thus we have more than 3×10^7 effective quanta; these occur in "packets" of 5000 quanta at a time, however. Consequently — and this is the cardinal point of the whole argument — the large number of quanta that would yield the low relative fluctuation of 1/5600 (if all the quanta occurred individually) cannot furnish any more information than the original 6500 X-ray quanta were capable of doing, with the much greater fluctuation of 1/80.

[12]) See G. C. E. Burger, Phantom tests with X-rays, Philips Techn. Rev. **11**, 296, 1950.

In normal fluoroscopy this is not really important. In fact, of the light quanta obtained, only a very small proportion falls on the pupil of the eye (see *fig. 6*): in the retina only 120 quanta are

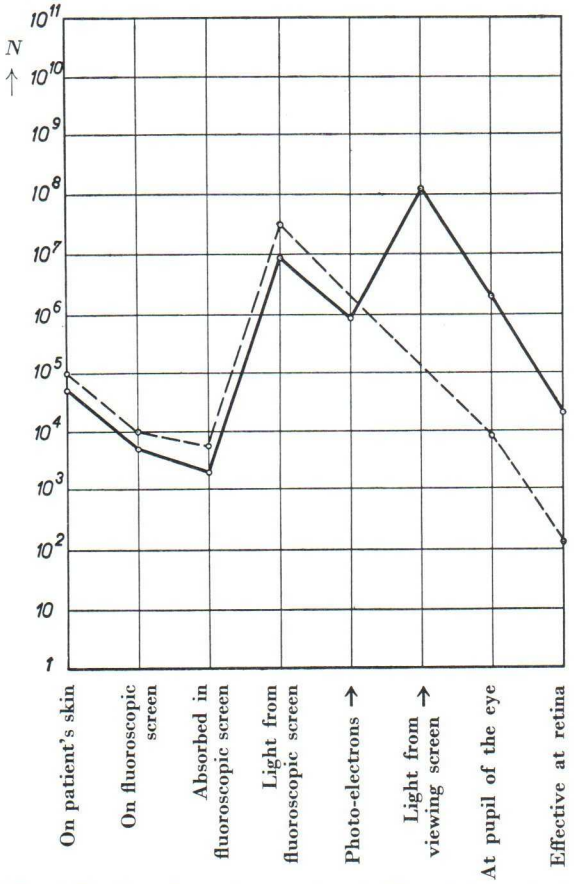


Fig. 6. Number of quanta occurring at different stages during a chest examination by means of X-rays, within a beam corresponding to an area of 0.1 sq.mm on the fluoroscopic screen. Voltage on X-ray tube 70 kV, current 2.5 mA, distance from X-ray focus to screen 70 cm.
Dotted line: ordinary fluoroscopy; accumulation time 0.1 sec.
Full line: with image intensifier; the accumulation time is now slightly shorter.

effectively absorbed, which is much less than the available number of X-ray quanta. In X-ray images of objects 0.1 sq. mm in size, therefore, the eye will not perceive contrasts of less than 3 to 5 times 1/11 and a large part of the information contained in the X-ray quanta is thus not used.

The image intensifier, however, has entirely altered this state of affairs. Initially the situation is rather less favourable, since, out of the above mentioned 100 000 X-ray quanta falling per second on 0.1 sq. mm of the thinner fluoroscopic screen, fewer are absorbed by the thinner screen and, moreover, the accumulation time T of the eye is shorter at the higher brightness level of the intensifier. We may take the number of effectively absorbed quanta to be about 1750. The light quanta produced by them in the fluoroscopic screen, however, are now put to better use (see fig. 6); it may be estimated that the area observed now yields roughly 20 000 actively absorbed light quanta, which is considerably more than the number of active X-ray quanta. This means that contrast sensitivity is then high enough for the eye to register the whole of the information contained in the X-ray quanta — but, obviously, not more.

This bears a direct analogy with optical enlargements. For example, in the image produced by a microscope objective, the fine details of the object are resolved only up to a certain limit; the "information" regarding the object is limited by the resolving power of the objective (which is in turn determined by the lens aberrations and ultimately by the wavelength of the light). The resolving power of the eye is much lower and the image is therefore observed through a high-magnification eye-piece. Now the greatest effective magnification of the eye-piece is that at which the details as separated by the objective can be seen as separate by the eye. If the eye-piece is any stronger, the details are certainly seen at larger angles, but the wealth of detail is not increased, the information contained in the image being by then already exhausted. It may be said that the extra enlargement is "ineffective". Similarly it may be said of the image intensifier that the part of the brightness intensification which corresponds to the factor by which the number of effective light quanta in the eye exceeds the number of effective X-ray quanta in the fluoroscopic screen, may be regarded as ineffective.

The conclusions to be drawn from the above argument are as follows:

1) Of the brightness intensification factor of 1000 which our instrument yields, a maximum factor of about 20 is "effective" [13]; the remaining factor of 50 plays no part in reducing the smallest perceptible contrast, but is of great practical significance as it makes vision easier (no adaptation necessary).
2) Should it be found desirable to make use of the properties of the image intensifier to reduce the X-ray dosage to the patient (for the same duration of observation), this will be unavoidably

[13]) This factor is greater than the ratio 1750/120 to be derived from the example in fig. 6 since, owing to the decrease in the accumulation time T, as also in the percentage of quanta producing stimuli in the optic nerve, a relatively higher brightness is required to ensure that the eye absorbs as many effective light quanta as there are effectively absorbed X-ray quanta.

accompanied by a loss of "information" which is, after all, determined by the number of X-ray quanta. If the same amount of information as that obtained from the ordinary fluoroscopic screen is considered adequate, the dosage for the purpose of a chest examination could, theoretically, be reduced at most to about 1/15th; this would still leave in hand a gain in brightness to a factor of about 60, which would represent a marked decrease in adaptation time. For thicker subjects the permissible reduction factor would be smaller.

The image intensifier and radiography

Similar considerations to those outlined above will make it clear that the use of the image intensifier can never provide the same high quality of image as that obtained in X-ray photography (radiography). The photograph owes its superiority to the fact that the accumulation time of the photographic emulsion is not limited, as is that of the human eye, added to which the emulsion retains what it has accumulated, the dosage time being thus no longer than the accumulation time. Unless a series of exposures is to be made, the dosage is therefore not usually dangerous and the intensity of the X-rays and/or the exposure time can be adjusted for each subject so as to yield the best average density. For a representative grade of film it will be found that to this end 20 000 X-ray quanta per 0.1 sq.mm (about 0.001 ergs of energy) must be effectively absorbed. The exposed film is examined with the aid of a viewing lantern; the illumination level can be raised as desired and matters so arranged that the light quanta received by the retina are in any case sufficient for perception of all the information the photograph has to offer (*fig.* 7). Now the difference in image quality between normal fluoroscopy, image intensifier and radiography is immediately expressed by the difference in number of useful quanta, i.e. for a chest examination 120 : 1750 : 20 000. With subjects that entail a higher degree of absorption (e.g. gastric fluoroscopy), the supremacy of the photograph is still more pronounced; it remains at 20 000 quanta, whereas in the two other cases the numbers of quanta are much lower.

Under these circumstances it will be obvious that the principle of the electronic image intensifier has no advantage over conventional X-ray photography. It can be employed to advantage, however, in fluorography. In this Review, on several occasions, the fact has been discussed that in this technique, whereby a reduced X-ray image is obtained with the aid of a camera, it is a difficult matter to obtain sufficient light to expose the film, and this is the reason why extremely fast lenses have been developed for this class of work [14]). The image intensifier, which furnishes a very bright, reduced image that is capable of being photographed quite easily on a 1 : 1 scale, provides a very convenient solution to the problem. Apart from this, the image intensifier may possibly prove to be the answer to the old problem of X-ray cinematography. As far as the mechanical aspect is concerned, X-ray fluorography has already brought a solution within sight, but the next step towards successful X-ray cinematography, although apparently only a small one, calls for some caution, since a very large number of photographs taken in succession means that the total X-ray dosage to the patient will again be fairly high. A reduction in the X-ray intensity with a view to limiting the total dosage

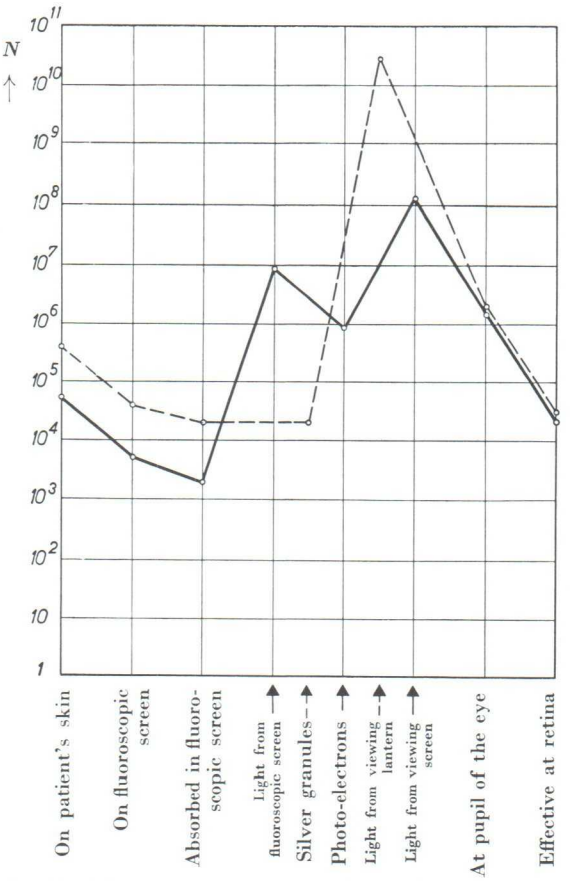


Fig. 7. Same as in fig. 6 but for normal radiography (dotted line). Full line; with image intensifier, as reproduced from fig. 6.

[14]) W. Hondius Boldingh, Fluorography with the aid of a mirror system, Philips Techn. Rev. **13**, 269-281, 1952 (No. 10).

during the taking of the film is possible only up to a certain point, seeing that even with very fast lenses too little light is otherwise obtained to ensure sufficient density of the film. This difficulty is overcome by the extra brightness provided by the image intensifier; even at appreciably lower X-ray intensities than those usually employed in fluorography, the image on the viewing screen will ensure a well-exposed film without necessitating particularly fast lenses. The fact that a reduction in the X-ray intensity detracts from the image quality (available information) has already been explained. It is a well-known phenomenon, however, that some slight sacrifice in image quality can safely be made in cinematography, since the summation of the impressions received from the successive pictures provides additional information.

Summary. In order to reduce to a minimum the risk of injury to the patient as a result of X-ray examination, the X-ray dosage that the patient receives must be limited as far as is practicable. Therefore, the attainable brightness of the image on a standard fluoroscopic screen is so low that long visual adaptation periods are required before such images can be studied; even then, visual acuity and the contrast sensitivity of the eye are both fairly low at such low brightness levels. Improvement can be brought by increasing the brightness of the fluoroscopic image ("intensifying"). The electronic tube by means of which this is now made possible functions on the principle of the image converter. The X-rays produce on a fluoroscopic screen an image whose radiation liberates electrons from a photocathode in contact with the screen. The electrons are focussed by an electrostatic electronic lens system working at a potential of 25 kV, to produce another image, reduced 9 times in size, on a second fluorescent screen, the viewing screen, which is observed through a simple microscope of 9 × magnification. The gain in brightness, to a factor of 800-1200 with respect to normal fluoroscopy, is produced partly by the accelerating field which imparts energy to the electrons, and partly by the electron-optical reduction of the fluoroscopic image. The resolving power of the image intensifier is higher than that of the best fluoroscopic screen. Owing to a number of causes the ultimate image is subject to a certain amount of fogging, which does reduce the contrast, although not to a serious degree. From a discussion of the lower limit of perceptible detail as based on spontaneous fluctuation in the number of X-ray and light quanta it is seen that the brightness can be intensified up to a factor of about 20 with a corresponding increase in the amount of "information" to be derived from the image. Beyond that point the information is restricted by the number of contributing X-ray quanta, but the much higher gain factor of the image intensifier offers the advantage that the image can be viewed without preliminary dark adaptation of the eyes, in a room with practically normal illumination. In conjunction with film cameras the intensifier paves the way to X-ray cinematography without the necessity for increasing the X-ray dosage to a hazardous extent.

VOL. 8 No. 12, p. 353-384 DECEMBER 1946

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, HOLLAND

NON-METALLIC MAGNETIC MATERIAL FOR HIGH FREQUENCIES

by J. L. SNOEK.

621.318.323.2:538.245

It is a well-known phenomenon that when an a.c. field is generated in metallic coils eddy currents are set up in the core. The resultant loss of energy can be minimized, in the case of low frequencies, by using a laminated core with insulation between the plates. The higher the frequency, however, the finer the distribution of the core laminations has to be, and as a matter of fact for rather high frequencies cores are being used that are made up of wire and even grains (so-called dust cores); in the latter kind the lines of force run in part through the non-magnetic insulation of the grains, so that there is always a fairly large minimum "air-gap".

Entirely new methods have become possible as the result of researchwork with ferrites, substances of the type of MFe_2O_4, with M denoting a bivalent metal. Some of these ferrites have a cubic structure and are capable of forming mixed crystals in any proportions. With a few exceptions these non-metallic substances are ferromagnetic. Their specific resistivity is 10^7-10^{12} times as high as that of iron, so that the problem of eddy currents is relegated entirely to the background, even in the range of radiofrequencies and without laminations. By suitable composition and heat treatment it is also possible to keep the hysteresis and residual losses relatively low, whilst the initial permeability may be of the order of 1000. By means of a pressure or spraying process this material, called "Ferroxcube", can be moulded into various shapes, and coils of improved quality or smaller volume have been made with it for different purposes. Thanks to its being homogeneous it makes a good magnetic screen.

Metallic and non-metallic ferromagnetics

Magnetic properties were most probably discovered in non-metallic substances much sooner than in metals. Anyhow magnetite (magnetic iron ore, Fe_3O_4) was known in ancient times as a non-metallic ferromagnetic mineral.

It was not until the nineteenth century that any real technical use was made of these magnetic properties, except for compasses. And when, following upon the epoch-making discoveries of Faraday and others, an electrotechnical industry sprang up and a demand arose for suitable magnetic materials, it was not these minerals that were sought but rather iron or alloys with iron as basic component; magnetic oxides were considered to be nothing more than a scientific curiosity.

When it came to using metallic cores in magnetic alternating fields one very soon came up against the phenomenon of eddy currents discovered by Foucault, and the losses due thereto. In order to reduce these losses one proceeded to laminate the cores; this offered the additional advantage that the plate-shaped material could be properly purified by annealing, thereby improving its magnetic properties. By simple deduction one can arrive at the following formula for the eddy current losses W_e of a laminated core (assuming the magnetic induction to be equal at all points of a core section):

$$W_e = \frac{4}{3}(af B_{max})^2 \cdot \frac{d^2}{\varrho} \cdot V \cdot 10^{-16} \text{ watt} \quad . \quad (1)$$

Here f is the frequency in c/sec., B_{max} the amplitude of the magnetic alternating induction in gauss, d the thickness in cm, s the specific resistance of the core material in ohm. cm, V the volume of iron in cm^3 and a the so-called form factor, which in the case of an induction sinusoidally variable with time amounts to $\pi/2\sqrt{2} = 1.11$.

This formula shows that there are only two ways of reducing eddy current losses for a given frequency and induction, viz:

1) by reducing the plate thickness (d);
2) by increasing the specific resistance (s).

First of a long series

The development of ferroxcube, a non-metallic magnetic material doing away with the problem of eddy current losses, has been one of the major contributions to the art of Electronics from Philips Research Laboratories. It opened the way to small high-Q coils and innumerable other devices for H.F. applications. Many articles on the characteristics of ferroxcube, its varieties and the concurrent theoretical investigations, its various applications, and on the development of other ceramic magnetic materials (ferroxdure, ferroxplana) have been published in Philips Technical Review. The first page of the first of this long series of articles is shown here.

THE FERROXCUBE TRANSFORMER CORE USED IN THE BROOKHAVEN COSMOTRON

by F. G. BROCKMAN *) and M. W. LOUWERSE **). 621.318.134: 621.384.61

For the last ten or twenty years fundamental research in nuclear physics has been associated with the building of ever larger machines for particle acceleration. This evolution which has tied up the physicist's work with that of the engineer has frequently been commented on. However, the evolution continues, and in recent years the personal assistance of the engineer has been superseded by the material cooperation of whole industrial organizations. This article will give an account of the cooperative studies that have been made in the development of ferrites for high energy particle accelerators by the Brookhaven National Laboratories, U.S.A., and the Philips Laboratories in the Netherlands and the U.S.A.

Introduction

In the efforts to produce particles of very high energies for nuclear experiments, a substantial advance was made in the past year when the proton accelerator of the Brookhaven National Laboratories, U.S.A. — the Cosmotron — was brought into successful operation. While the highest particle energy obtained with earlier accelerators did not surpass 1000 million electron volts, the Cosmotron has produced protons with 2300 MeV, and it is expected that the energy can be increased to 3000 MeV.

The completion of this machine, which was achieved by a group of workers including M. S. Livingston, J. P. Blewett, G. K. Green and L. J. Haworth, is a remarkable feat of enterprise and engineering. A few data that will be mentioned presently may give some idea of the vastness of the project. An account of the design was given by these Brookhaven workers in 1950 in the Review of Scientific Instruments, and a detailed description of the completed machine has just appeared in the same journal [1]).

The present article will focus attention on the fact that, in one of the chief elements of the Cosmotron, the "transformer" producing the radio-frequency electric field for the actual acceleration of the protons, a large ferrite core plays an important part. The major portion of this core was assembled from Ferroxcube material manufactured by Philips at Eindhoven. The specifications for this core and the problems encountered in selecting the material and in the manufacturing will be described. A few remarks will be made on the development of other ferrite materials for similar purposes.

Basic design of the Cosmotron

A short description of the design principles of the Cosmotron as published in the first article quoted in [1]), will first be given.

The protons travel in an annular evacuated chamber of cross-section $30'' \times 7''$. This chamber is composed of four quadrants of about 30 feet radius, which are connected together by four short straight sections (plan view, *fig. 1*). The protons are made to follow a path roughly on the axis of the chamber by magnetic fields perpendicular to the plane of the path, which are produced by four powerful electromagnets disposed along the path. In the photograph of the Cosmotron (*fig. 2*) the yokes of these magnets, which enclose the annular acceleration chamber, can be seen.

*) Philips Laboratories, Inc., Irvington-on-Hudson, N.Y., U.S.A.
**) Laboratory of the Philips Ceramics Factory, Eindhoven, Netherlands.

[1]) M. S. Livingston, J. P. Blewett, G. K. Green and L. J. Haworth, Design study for a three-Bev proton accelerator, Rev. sci. Instr. **21**, 7-22, 1950.
The detailed description is contained in a series of articles in Rev. sci. Instr. **24**, 723-870, 1953 (No. 9).

One of the straight field-free sections is used for the injection of the protons into the chamber. The protons are obtained from a hydrogen ion source and accelerated by a Van de Graaff high voltage generator to an initial energy of 3.5 MeV for injection. In another straight section the acceleration of the injected protons is effected: under the influence of a longitudinal electric field alternating at a suitable frequency, their energy is increased by a certain amount after every revolution. To keep the protons (mass M, charge e) on a circular path of radius R, a magnetic field of flux density

$$B = \frac{Mv}{eR} \quad \ldots\ldots\ldots \quad (1)$$

is required. With the momentum Mv corresponding to the *initial* energy of 3.5 MeV, the required value of B is 0.029 Wb/m². In order to keep the protons on the same orbit within the vacuum chamber

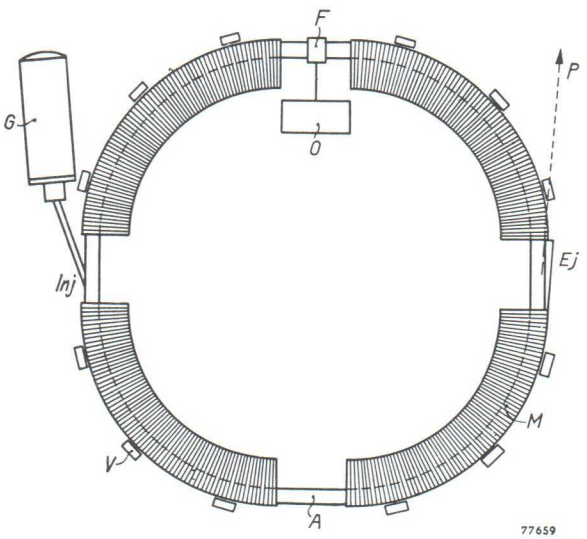


Fig. 1. Plan-view of the Cosmotron. The accelerated protons move on (or near) the axis of the annular evacuated chamber A, which has a diameter of about 60 feet. The protons are injected into the chamber at *Inj*, with an initial energy of 3.5 MeV, from a Van de Graaff high tension generator G. Four electromagnets M keep the protons on their track. The acceleration of the protons occurs at F, where an alternating electric field, generated by the oscillator O, gives the protons a "kick", once per revolution. When the protons have acquired their final energy, they emerge in a beam P from the annular chamber at the ejector Ej. V represent a number of high vacuum pumps for the evacuation of the acceleration chamber.

while their energy and momentum is increasing, the magnetic field is increased continuously, up to a maximum value of 1.4 Wb/m². This value will allow the protons to be kept on their track with a final energy of 3100 MeV. The beam of accelerated protons can be ejected from the annular chamber by a suitable deflector device accommodated in another straight section of the chamber.

At the injection energy of 3.5 MeV, the proton velocity is 2.6×10^7 m/sec. As the path along the axis of the annular chamber has a total length of about 70 meters, the time for one revolution is initially about $^1/_{350\,000}$ second. This means that the accelerator section proper must carry an alternating electric field of frequency about 350 000 cycles/second, in order to give the circulating protons a "kick" each revolution in the correct phase. At a final energy of 3000 MeV, the proton velocity (for the calculation of which the relativistic formula must now be applied [2]) is approximately 10 times the initial value, so that during the course of the acceleration of a burst of protons the frequency of the accelerating electric field must be gradually increased from 0.35 to about 4 Mc/s.

Although this brief description would suffice for our further discussion of the accelerator, a few further details of the design study published by the Brookhaven group [1]) may be mentioned in this section for the sake of completeness. One of the major problems of the project was the supply of the large magnetic energy required for keeping the protons on their track. The field of 1.4 Wb/m² in a space 70 metres long and roughly 36″×9.35″ in cross-section (gross dimensions, as against the above-mentioned net dimensions inside the chamber) involves a magnetic energy of 12 megajoules. This formidable magnetic energy [3]) must be built up during the acceleration time of one burst of protons and taken away relatively quickly after the ejection in order to make the apparatus ready for the next burst. As it would be very uneconomical to dissipate the magnetic energy after each burst, it is stored mechanically in the form of the kinetic energy of a large flywheel and recycled. In order that the build-up of the magnetic energy may be accomplished with a reasonable peak power, and that the acceleration may be achieved with not-too-large "kicks", the acceleration time for one bunch of protons must be relatively long. An acceleration time of 1.0 sec was chosen. (This means that the protons will make roughly three million revolutions during their acceleration.) The cooling of the magnet and the available power of the motor-

[2]) A convenient graph for the general velocity-energy relationship may be found in Philips tech. Rev. **11**, p. 68, 1949.

[3]) For a comparison, the reader is referred to the description of the Amsterdam synchrocyclotron given in this Review **12**, 354, 1950: it was there mentioned that the large electromagnet contains a magnetic energy of 0.65 megajoules, whereas the magnetic energy in the gap of a high power loudspeaker amounts to about 0.25 joules.

generator coupled to the flywheel are such that the acceleration of a bunch of protons may be repeated every 5 seconds.

Another important problem was the simultaneous variation of the magnetic field and of the frequency of the accelerating electric field. Because of the essentially non-linear behaviour of an electromagnet as a circuit element, it would be difficult to realize an exactly predetermined increase of the magnetic field with time. The designers of the Cosmotron avoided this difficulty by regarding the magnetic field as the independent variable. The expected time-variation of the magnetic field (*fig. 3*) was derived from the circuit constants and model measurements. From the rate of change dB/dt of the magnetic field it can be deduced by how much the momentum (and the energy) of the protons must increase after each revolution in order to keep them on the same orbit. This required energy gain per revolution is plotted against time in *fig. 4*; it is seen to decrease during the acceleration period from an initial value somewhat greater than 1000 eV to a final value of rather more than 600 eV per revolution. The velocity of the protons resulting at each instant of the acceleration time (and which depends only on the instantaneous value of B) determines the frequency that is required at that instant for the accelerating electric field; cf. curve f in fig. 3. The problem of controlling the frequency to obtain this time variation with the required precision, was solved at Brookhaven in an elegant manner: the derivative of the magnetic field is measured continuously by means of a pick-up coil, and the integrated signal of this

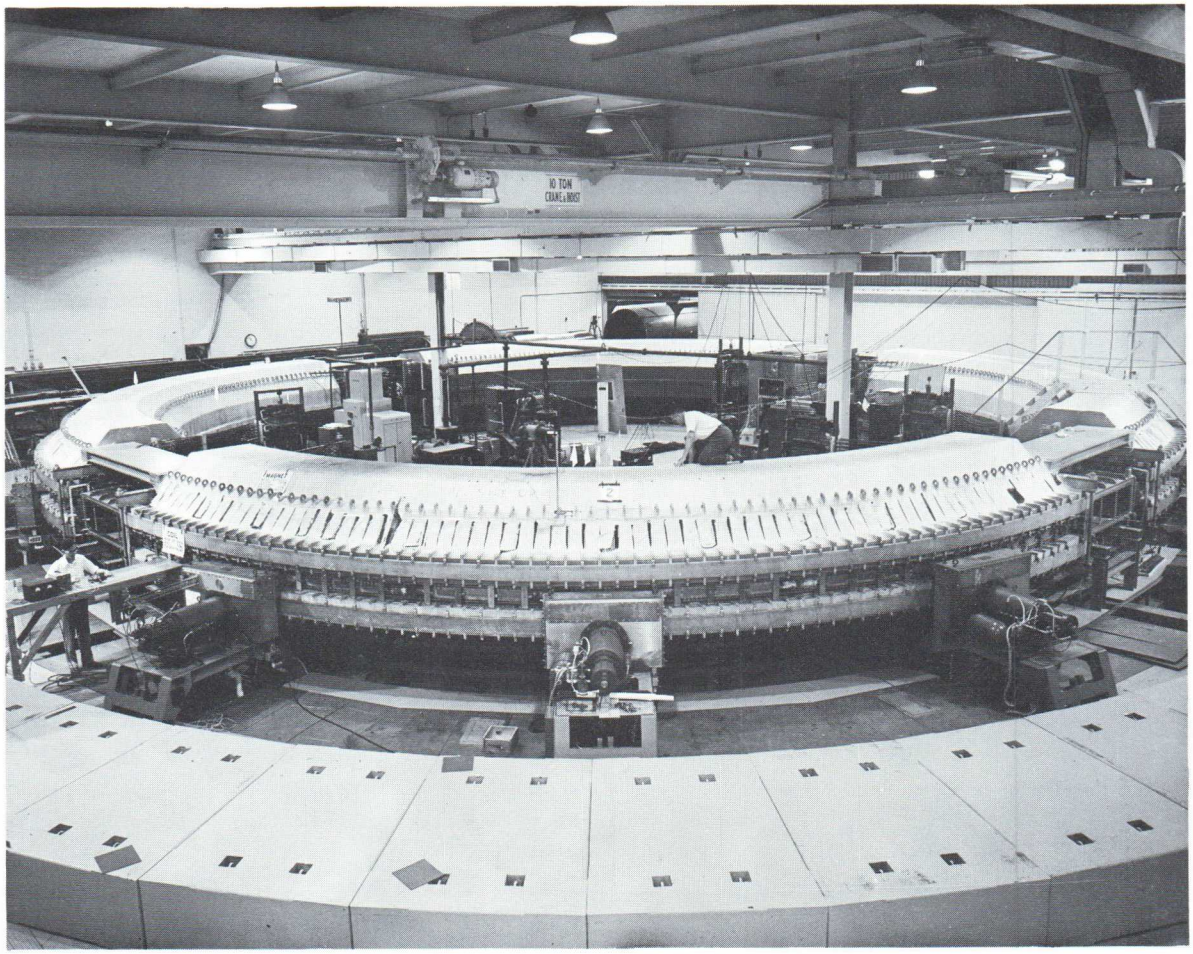


Fig. 2. General view of the Cosmotron, in the Brookhaven National Laboratory, U.S.A. The four large electromagnet quadrants, surrounding the four sections of the annular acceleration chamber, have a total weight of 2200 tons. In the background is the Van de Graaff generator placed in a horizontal tank. Farther to the right is the accelerator unit. The concrete wall seen in the foreground protects personnel from the radiation of the machine.

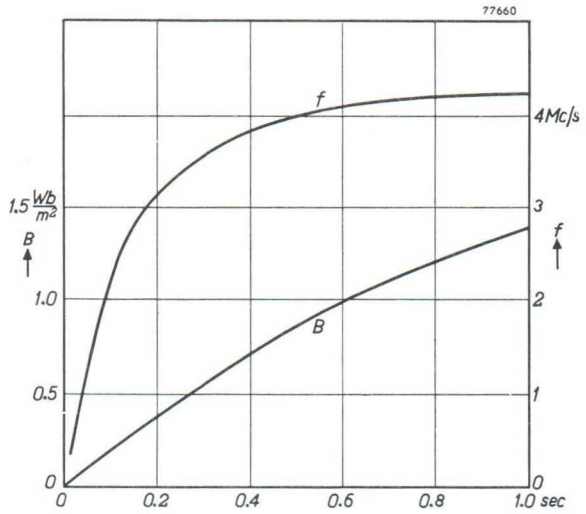


Fig. 3. Variation with time of the flux density B of the magnetic field during one accelerating period. This field, which forces the protons to move in their circular orbit, is regarded as the independent variable: the frequency of the accelerating electric field must then vary according to the curve f, in order to ensure that the protons circulating at an increasing rate will remain on their track inside the annular chamber.

coil, is used for controlling the frequency of a "permeability-tuned" LC-oscillator producing the accelerating radio frequency field [4]). The tuned circuit of this oscillator employs an inductance containing a Ferroxcube III core with an initial permeability of about 1000. The control signal provides an increasing bias magnetization of the core, causing its effective permeability to drop sharply, thus decreasing the inductance. In this way the required wide frequency sweep (fig. 3) is obtained, together with the required frequency stability. A frequency error as small as 0.3% would cause the proton orbit radius to exceed the latitude allowed by the annular chamber. With the oscillator described, when the temperature of the Ferroxcube core is held constant to 0.1 °C, the oscillator frequency is stable and reproducible at any control signal value to better than 0.1%. The oscillation amplitude is held constant at a value such that the peak accelerating field is about twice the highest instantaneous value needed (see above).

With a view to the phase stability of the revolving protons, these must pass through the alternating electric field at a certain phase angle Θ after its peak. If the protons are focussed to a bunch at a phase of about $\Theta=60°$, the peak value of the accelerating radio-frequency voltage must be about 2200 V. Matters of orbit stability, which of course were of prime importance in the design, need not be considered here, though in the last section they will have to be mentioned.

A Ferroxcube core is thus encountered in the tuning circuit of the oscillator, at the heart of the Cosmotron. This small core, however, did not offer any specific manufacturing problems and need not be discussed in this article. The Ferroxcube core forming the subject of this article is also part of the radio-frequency accelerating device, but serves the purpose of *transferring* the radio-frequency energy to the proton beam.

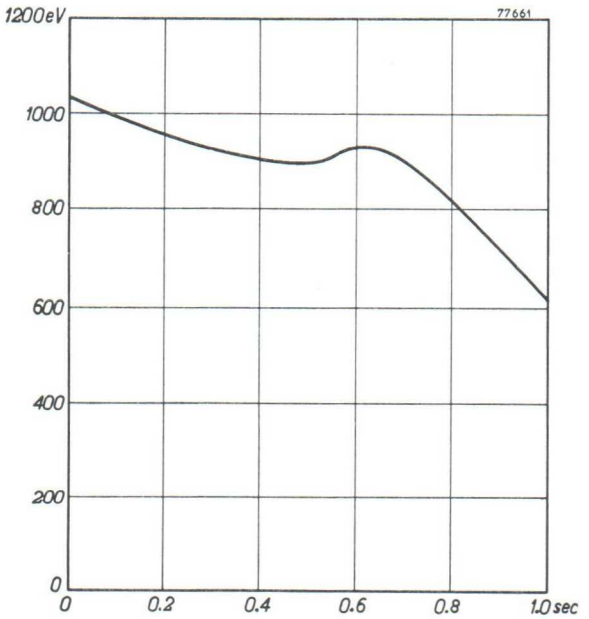


Fig. 4. Required energy gain per revolution of the protons as a function of time.

[4]) A. I. Pressman and J. P. Blewett, A 300 to 4000 kilocycle electrically tuned oscillator, Proc. Inst. Rad. Eng. **39**, 74-77, 1951.

The accelerating unit

The manner in which the longitudinal radio-frequency electric field is established in the acceleration path of the protons is one of the unique features of the Cosmotron as compared with other types of particle accelerators. The method [5]) may be roughly described by regarding the large circular path of the protons as the single turn secondary winding of a transformer, whose primary winding consists of a coil (having one or more turns) in the anode circuit of a power amplifier (*fig. 5*) driven by the radio-frequency oscillator. Both windings are effectively coupled by a ferromagnetic core (*fig. 6*).

The Brookhaven workers were able to consider the adoption of this novel system because, at the time that the design work on the Cosmotron had begun, the development at the Philips Laboratories

[5]) A proposal to accelerate particles by a method along these lines was already made by A. Bouwers, in his book: Elektrische Höchstspannungen, J. Springer, Berlin 1939, p. 83.

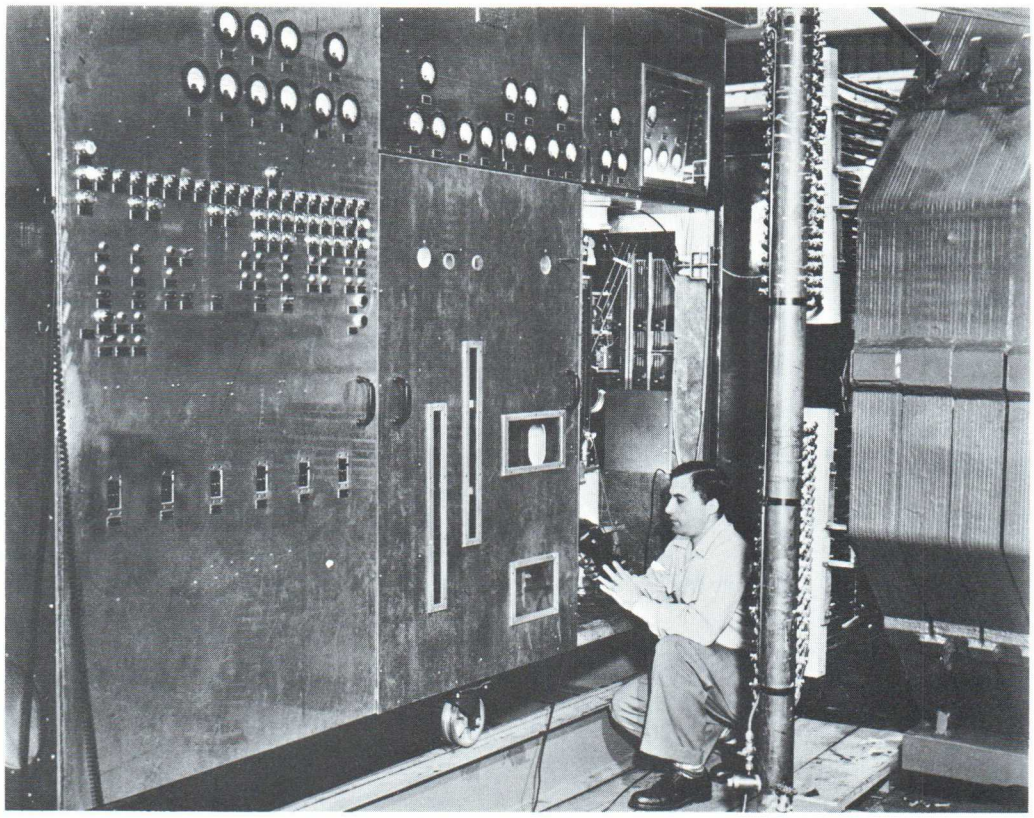


Fig. 5. Power amplifier supplying the radio-frequency energy to the primary of the transformer.

at Eindhoven of ferromagnetic materials suitable for very high frequencies had become known [6]). The usual transformer steel, even with extremely fine lamination, would not do in this case because of excessive eddy current losses. Powdered cores of iron or of "Permalloy" also offered no satisfactory prospects (see below). The new materials, however, owing to their very high resistivity, could be expected to be adequate even with sections several centimeters in thickness.

Quantitatively, the conditions imposed on the core material were derived from the formula for the input impedance Z of the above-mentioned transformer (cf. the first article quoted in [1]), eq. 17). This formula in simplified form says that Z is proportional to

$$f \mu_r \left(\frac{1}{Q} + j \right),$$

f being the frequency, μ_r the relative permeability of the core material and Q the usual figure of merit of the primary winding including the effects of core losses. In order to obtain the desired accelerating voltage (2200 V peak) with a reasonable anode current of the oscillator, a high impedance Z is desired, while for reducing power loss and heat dissipation in the transformer, the real component of Z should, of

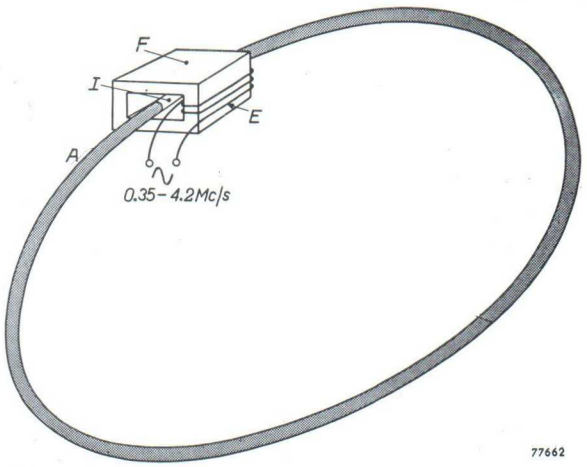


Fig. 6. Simplified representation of the acceleration system of the Cosmotron. The metallic annular acceleration chamber A constitutes the secondary winding of a transformer with primary winding E and ferromagnetic core F. I is an insulating section of the annular chamber, across which the accelerating voltage is produced. — In reality the core F is enclosed in a copper shield having the gap I in the centre. The shield forms an autotransformer with the windings. Thus the electric field is restricted to the gap and radio interference prevented.

[6]) J. L. Snoek, Non-metallic magnetic materials for high frequencies, Philips tech. Rev. **8**, 353-360, 1946; J. L. Snoek, New developments in ferromagnetic materials, Elsevier Publishing Co., Amsterdam 1947.

course, not be too large. A core material with a high permeability μ_r, preferably not less than 500, and low losses resulting in a reasonably high value of Q, say, not less than 0.5 even at the highest frequency used (4.2 Mc/s), was therefore desirable. Comparative measurements of μ_r and Q for ring core samples of several materials were performed by the Brookhaven group. Their results [1]), shown in *table I*, clearly pointed to the superiority of the ferrite material for this application, though they did not yet conclusively demonstrate that ferrites could do the job.

It should be emphasised that the system outlined in fig. 6 was not the only practicable solution for the problem of transferring energy to the proton beam from a radio frequency source. An alternative solution considered by the Brookhaven group involved the use of a *tuned* system, for example a resonant cavity, driven at a varying frequency and continuously tuned over the desired frequency sweep. Although the first system was chosen for the Cosmotron, the alternative system has gained new interest, as it will probably be used in a new type of accelerator which can yield even higher particle energies and which is now the subject of study by a number of groups including that at Brookhaven. We shall return to this further development at the end of this article.

Table I. Properties of magnetic core materials at 1 megacycle, as measured by the Brookhaven group [1]).

Material	μ_r	Q
4-79 Mo Permalloy (0.001″ laminations)	600	0.6
Fe powder (80 mesh grade B)	35	2.0
81 Permalloy powder (120 mesh)	75	1.6
2-81 Mo Permalloy powder (120 mesh)	125	3.0
Typical ferrite	1000	6.0

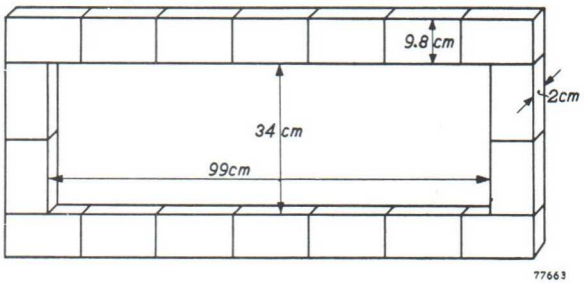


Fig. 7. Dimensions of one of the 24 "picture-frames", made of bricks of Ferroxcube IV, manufactured by Philips for the core of the Cosmotron.

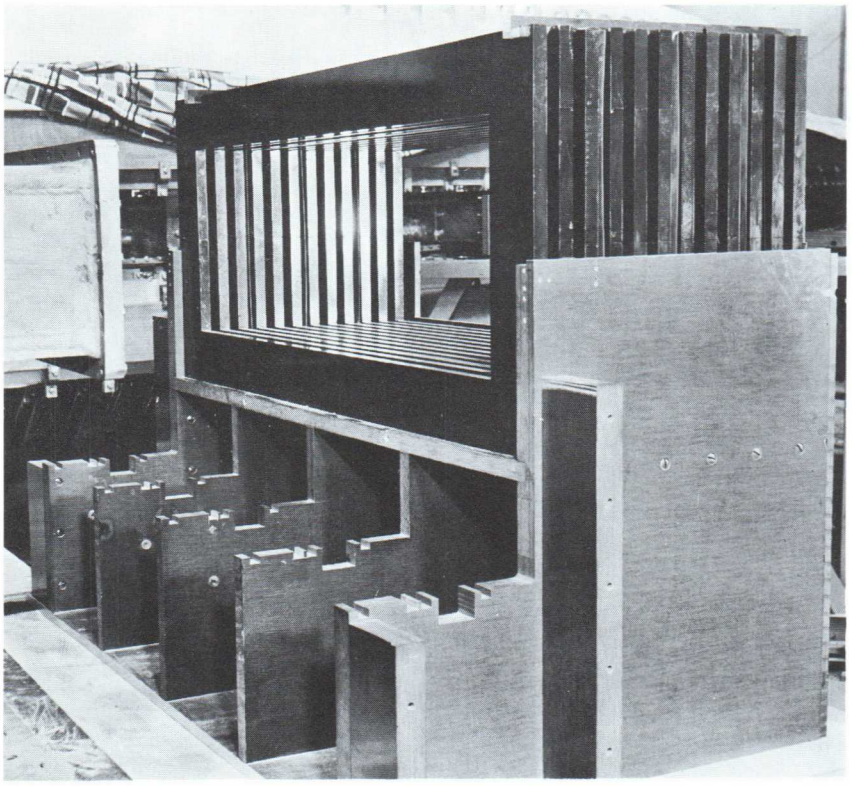


Fig. 8. Bank of 12 Ferroxcube "picture-frames", during the assembly of the core of the Cosmotron. On the left, part of the annular acceleration chamber, which runs through the "window" of the core, can be seen.

The adoption of the transformer system presented the makers of ferrites with two practical problems: the choice of the ferrite material best suited for the purpose, and the actual manufacturing of the large core structure. The "window" of the core must be about 100 cm wide and 34 cm high, to allow space for the annular vacuum chamber and the primary winding. In order to keep the magnetic flux density sufficiently low, a core cross-section of 10×50 cm was required. This would mean a total weight of the ferrite core of about 750 kg.

In 1948 the Brookhaven group consulted the Philips Laboratories, Irvington, regarding a suitable fer-

rite for the core structure. After the characteristics of different materials for the purpose were investigated, part of the final core was assigned to Philips in November 1949. This part consisted of 24 "picture frames" of Ferroxcube IV, of the size shown in *fig. 7* and composed of bricks with dimensions 170.0 × 98.0 × 20.0 mm. The delivery of the whole batch was completed in May 1950. The frames were installed in the Cosmotron in two banks of 12 each. The photograph in *fig. 8*, made during the course of the assembly of the Cosmotron, shows one of these banks.

A few details of the preliminary investigations and of the manufacturing method will now be given.

Choice of material and manufacturing of the Ferroxcube frames

The ferrite chiefly manufactured at the time was Ferroxcube III, a solid solution of manganese-, zinc- and ferrous ferrites. This type of ferrite has by far the lowest hysteresis losses, a property of great importance for telephone applications, and in fact, the use of ferrite in carrier telephony was then commercially the most important [7]). The different varieties of Ferroxcube III have initial permeabilities [8]) varying from 800 to 1700, resistivities of the order of 100 ohm cm (a factor of 10 million times greater than that of iron) and they yield Q-values of 10 at frequencies up to 0.3 to 0.8 Mc/s. From the latter property, it was estimated that core sections with a thickness of a few centimeters would not introduce excessive eddy-current losses. This conclusion appeared to be verified by tests performed on samples of the material of the small size usual in carrier telephony and other applications. When, however, bricks of this material having dimensions of several cm and suitable for the assembling of a test core for the Cosmotron were made at Eindhoven and tested at Brookhaven, it was discovered that their initial permeability μ_i decreased abnormally rapidly with increasing frequency resulting in a very low value at about 2 Mc/s (*fig. 9*). Closer investigation of this unexpected phenomenon at Irvington revealed the existence of a *dimensional effect*, which could be shown to be due to the material possessing a very high dielectric constant [9]).

The rapid decrease of μ_i, which could be interpreted, in general, as a resonance effect, was shown to shift to higher frequencies when air gaps were introduced into the core, an increasing air gap with a given frequency thus giving rise to a larger effective permeability (*fig. 10*). This result made it very probable that the permeability observed with no air gap was not a true constant of the material. Further evidence for this was obtained from experiments with laminated cores. Again, the apparent resonance effect was shifted to higher frequencies as the core was subdivided more and more in a direction perpendicular to the magnetic flux (*fig. 11*). Since the effect of ohmic eddy currents certainly could not

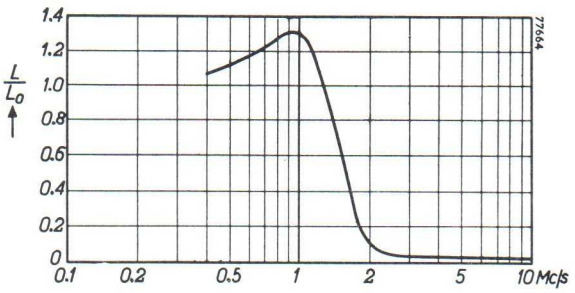


Fig. 9. The permeability of Ferroxcube III, measured on the sample bricks, for the Cosmotron core, plotted against frequency. The ratio of the self-inductance L of a ring-coil with a core of these bricks to that of the same coil at low frequencies L_0, is used as a measure of the permeability.

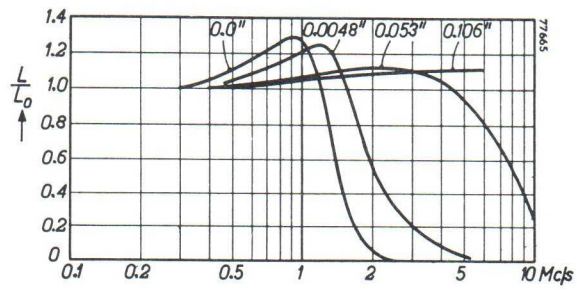


Fig. 10. Same as in fig. 9, with air gaps in the core of 0, 0.12, 1.35 and 2.7 mm respectively. The sharp drop in permeability occurs at higher frequencies as the air gaps get larger.

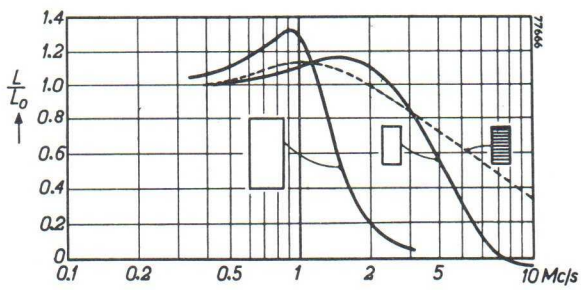


Fig. 11. Same as fig. 9, for cores of different cross-sections and subdivided into laminations as shown. The smaller the dimensions, the higher the frequencies at which the decrease occurs. (In one of the curves, negative values of L are seen to occur; this means that a dimensional resonance frequency has been passed and that the induction in the core then is oscillating in antiphase with the exciting alternating field.)

[7]) G. H. Bast, D. Goedhart and J. F. Schouten, A 48-channel carrier telephone system, Philips tech. Rev. **9**, 161-170, 1947 and **10**, 353-362, 1948/49.

[8]) Cf., for example, J. J. Went and E. W. Gorter, The magnetic and electrical properties of Ferroxcube materials, Philips tech. Rev. **13**, 181-193, January 1952.

[9]) F. G. Brockman, P. H. Dowling and W. G. Steneck, Dimensional effects resulting from a high dielectric constant found in a ferromagnetic ferrite, Phys. Rev. **77**, 85-93, 1950.

account for the results, one was led to consider the dielectric properties, and a relative dielectric constant ε_r of quite a surprising value was found, being as high as 50 000. The phenomena observed could then be explained by the superposition upon the usual eddy currents due to ohmic conduction of another type of eddy current which flows by dielectric displacement. Such currents will be out of phase with the ohmic eddy currents, and they can lead to a dimensional resonance in an inductor wound on a core of large enough cross-section. The wavelength of an electromagnetic wave in such a material is given approximately by:

$$\lambda \approx \frac{c}{f} \cdot \frac{1}{\sqrt{\mu_r \varepsilon_r}} \quad (c = \text{velocity of light in vacuo}).$$

Standing wave effects can be noted when the smaller dimension of a core, perpendicular to the magnetic flux, is $^1/_2$ wavelength.

Interesting though this discovery was from a scientific point of view, it showed that Ferroxcube III would not be a suitable material for the Cosmotron core. However, another type of ferrite, Ferroxcube IV (nickel zinc ferrite), was available at Eindhoven, and a variety of this type met the specifications in a satisfactory manner. This material has a sufficiently small dielectric constant and a much higher resistivity than Ferroxcube III (of the order of 10^6-10^7 ohm cm), so that no dimensional resonance effects could occur at the frequencies to be used, even with bricks of large dimensions. Consequently, the frames for the Cosmotron were made from Ferroxcube IV. The permeability versus frequency curve of one large lamination is shown in *fig. 12*.

The large size of the core structure, which in this way influenced the choice of the material, also created some difficult manufacturing problems. Ferrite materials are prepared by *ceramic* methods, i.e. the powdered materials, after the addition of suitable binders, are given the shape of the objects to be manufactured by pressing, extruding or the like, the shaped parts then being dried and fired at a high temperature. During the firing a shrinkage occurs, which in the case of ferrites may amount to 20-24% (on the linear dimensions). This large shrinkage, which is illustrated by *fig. 13*, imposes very strict requirements on the uniformity of temperature during drying and firing of large bricks such as were required for the Cosmotron core, in order to avoid cracks. The bricks were moulded in an 80 ton press and then dried very slowly, remaining for several days in a large drying chamber at about 40 °C. The firing was then performed in special ovens in an oxygen atmosphere, with a temperature cycle covering a period of several days. Such an oven is shown in *fig. 14*. The firing process, of course, must be conducted very carefully, as the reactions occurring during firing are essential for the development of the special magnetic properties of a ferrite.

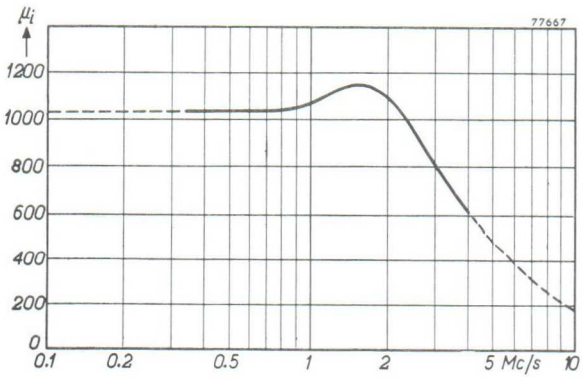


Fig. 12. Relative initial permeability μ_i of Ferroxcube IV, measured on one of the picture frames for the Cosmotron, plotted against frequency. A decrease is seen to occur at much higher frequencies than in the case of Ferroxcube III.

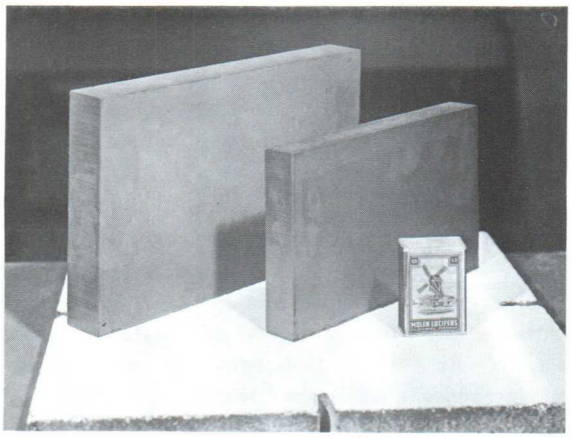


Fig. 13. A brick of Ferroxcube, before and after firing. During the firing at high temperature a shrinkage of 20-24% on linear dimensions occurs.

The bricks then had to be cemented together to form the "picture frames" as outlined in fig. 7. The unavoidable air gaps between adjacent bricks were reduced to a few microns by grinding the faces of the bricks with this precision; it will be appreciated that severe requirements were imposed on the rectangularity of the bricks, the permissible deviation of the 90° angle amounting to 1′ of arc in the largest surface of the brick and 1′30″ in the second largest.

It was established at Eindhoven that a plastic of the aethoxyline group, hardening at a temperature of about 180 °C, would be a suitable cement for ferrite materials. Experience with a sample picture

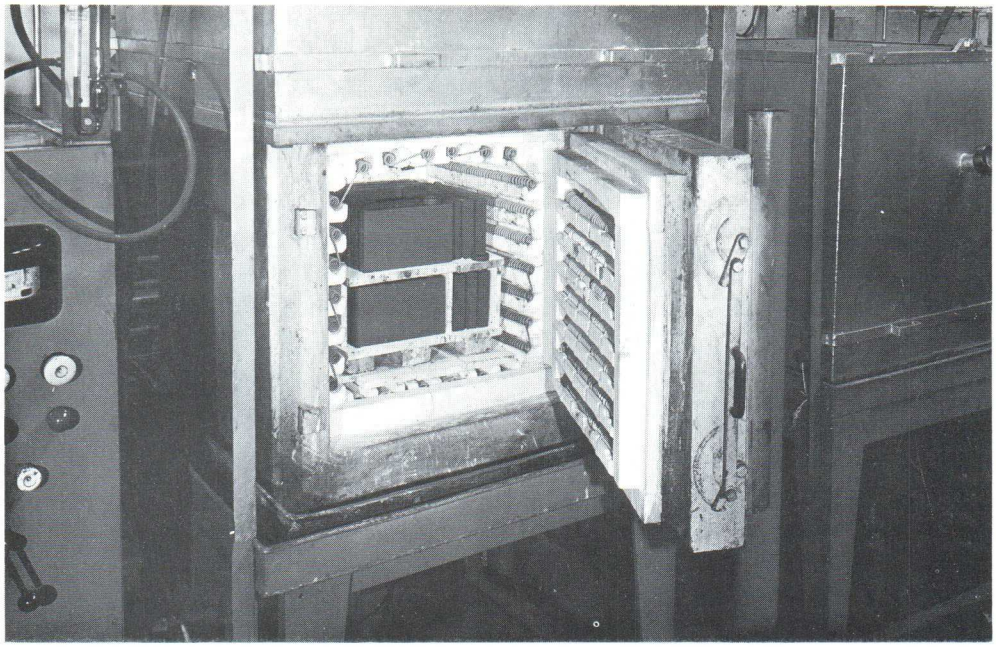


Fig. 14. One of the electric ovens in which the Ferroxcube IV bricks for the Cosmotron core were fired.

frame, used for preliminary measurements, had shown however, that the simple cementing and baking procedures applied to similar joints of other materials and also of Ferroxcube III were not giving water vapour, entering from the sides, access to the thin plastic films in the joint, causing deterioration in a few months.

A special procedure was therefore developed for

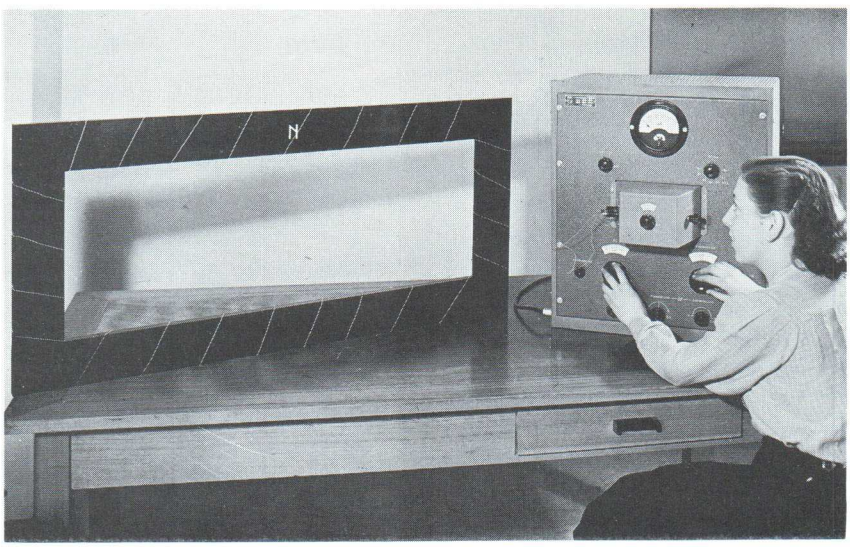


Fig. 15. One of the "picture-frames" for the Cosmotron core, during control measurements in the Ceramic Factory at Eindhoven.

satisfactory in the case of Ferroxcube IV. This could be ascribed to the *porosity* of the sintered Ferroxcube IV material, a property essential for giving the oxygen gas access to the innermost parts of the sample during firing, but unfortunately also the cementing of the Ferroxcube IV bricks. The main steps of this procedure involved a thorough cleansing of the surfaces to be cemented, a repeated impregnation in vacuo with the plastic cement and baking of the bricks. In the cementing

process, a special tool enabled the bricks of a frame to be pressed together by a constant force during baking, so that most of the plastic layer applied between adjacent bricks is driven out of the joints during that part of the polymerization period when the viscosity is lowest. In this way the gap lengths

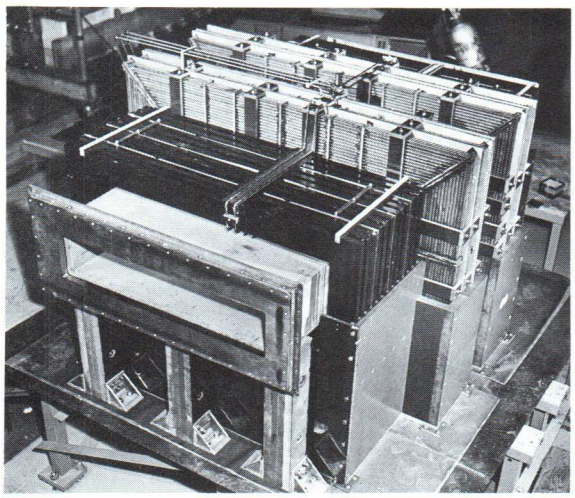


Fig. 16. The completed ferrite core of the Cosmotron, assembled in position. The Ferroxcube "picture frames" are seen in the centre and in the background of the photograph.

of non-ferromagnetic material were reduced to the small dimensions mentioned above. As a final precaution, the cemented frames were coated with a water repellent polyvinyl lacquer.

The breaking strength of the joints obtained is considerably larger than that of the Ferroxcube IV slabs themselves. This had been established by experiments, and it was proved again in shipping the frames to Brookhaven: when a box containing one of them was dropped from the hoisting crane, the frame was broken at two places, but not at any of the joints.

The photograph *fig. 15* shows one of the completed frames in the course of testing at Eindhoven. *Fig. 16* is a photograph of the complete core, assembled at Brookhaven and containing the two banks of picture frames.

Operation of the completed core under the load conditions imposed at Brookhaven has been reported to be completely satisfactory. The frames are separated by air spaces, and since the core is surrounded by a copper shield, forced air cooling can be applied. Even uncooled, however, the core functions without appreciable temperature rise. For further information about the performance of the oscillator and accelerator unit as a whole, the reader should refer to the recent publication by the Brookhaven group [1]).

The application of ferrites in "strong-focusing" accelerators

Since the Cosmotron was put into operation great interest has been aroused by the discovery by members and consultants of the Brookhaven group of the so-called "strong-focusing" principle [10]). Without entering into a detailed discussion, it may be said that a new method of reducing the amplitudes of radial and vertical oscillations of the revolving particles will allow the cross-section of the annular vacuum chamber to be reduced considerably, compared with the cross-section of the Cosmotron. Thus, with about the same magnetic energy — and therefore, very roughly, with a total cost of the same order — particles could be kept on a circular track with a diameter, say about 10 times that of the Cosmotron. In this way a particle energy of about 10 times that of the Cosmotron, i.e., of about 30 000 MeV, could be obtained (the momentum according to eq. (1) is proportional to the product BR, and at these high energies the energy is approximately proportional to the momentum; see also the graph of fig. 10 in the article quoted in [2]).

Plans for accelerators using this principle are actually being studied at several places. Such an accelerator would not make use of a single accelerating unit like that of the Cosmotron, but it would contain a large number of accelerating units, based on the continuously tuned resonator principle mentioned above as an alternative possibility. No

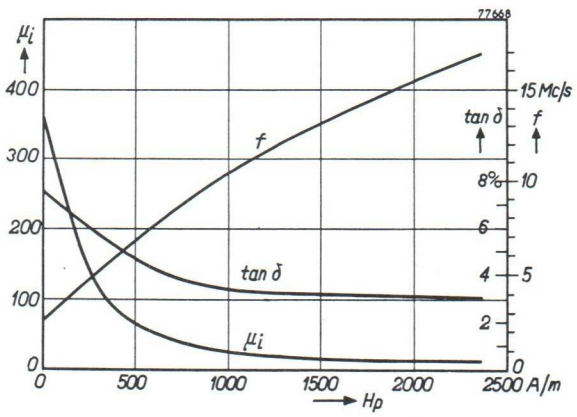
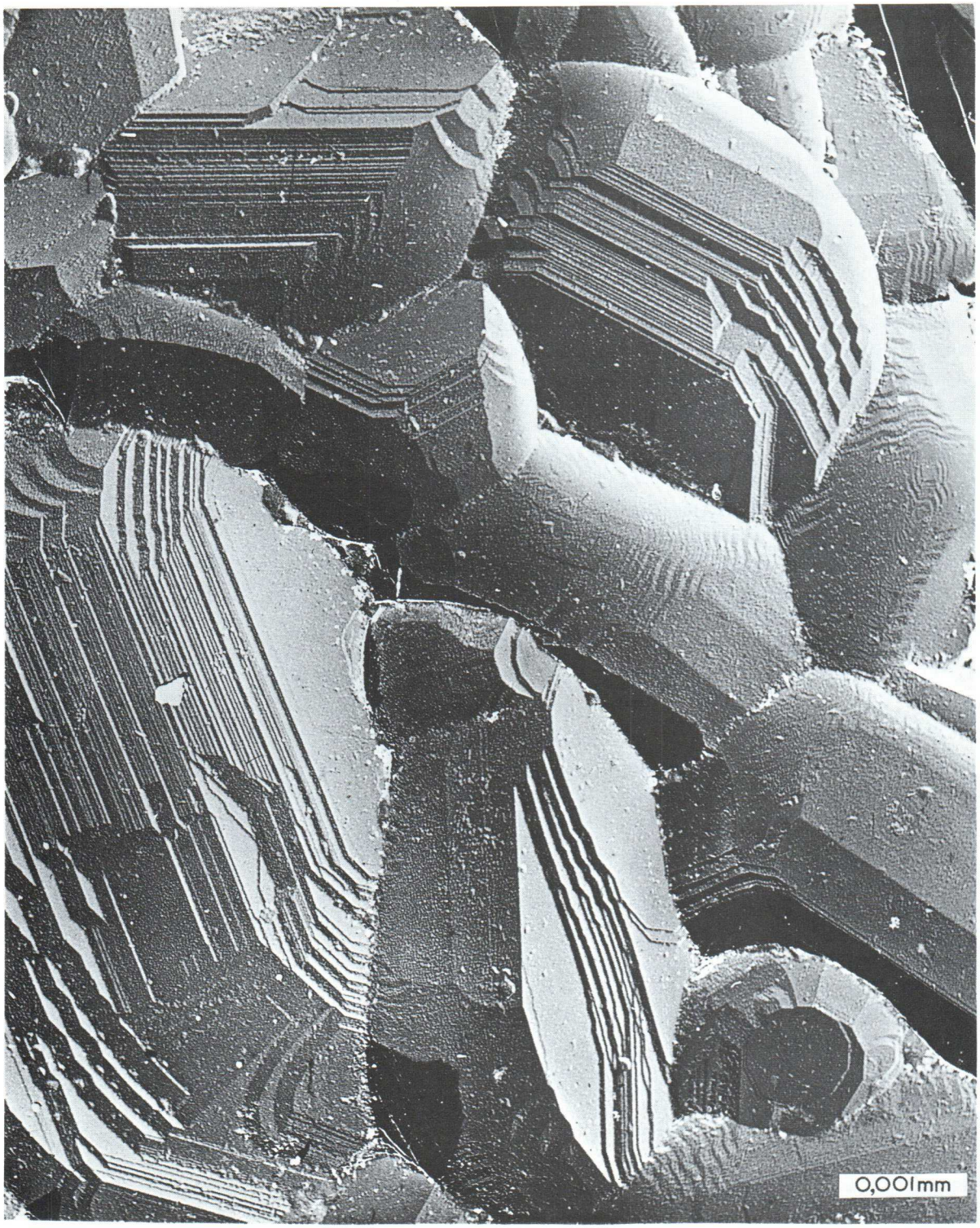


Fig. 17. A Ferroxcube material suitable for accelerators employing tuned resonant cavities has been developed at Eindhoven. The permeability μ_i of this material is varied by means of a polarizing field H_p and the curve f shows the frequency variation obtainable in this way (1 : 6). The figure of merit Q (= 1/tan δ) of a cavity filled with this material is greater than 10 even at the lowest frequencies.

[10]) E. D. Courant, M. S. Livingston and H. S. Snyder, The strong-focusing synchrotron — a new high energy accelerator, Phys. Rev. **88**, 1190-1196, 1 Dec. 1952.

transformer then would be needed, but oddly enough, ferrites again would have to be employed, on an even larger scale than in the Cosmotron: the resonators would be resonant cavities, containing massive cores of ferrite. The resonant frequency of such a cavity can be changed by a saturating magnetic bias field, much in the same way as in the master oscillator of the Cosmotron described above. As several tons of ferrite would go into each cavity, a total quantity of several hundred tons would be needed for a 30 000 MeV accelerator! The ferrite to be used will have to meet very exacting requirements. One of the tentative designs calls for a lowest resonant frequency of 3 Mc/s and an upper limit of 17 Mc/s. This frequency sweep (about 1 : 6) implies a corresponding permeability change of 1 : 36, to be obtained by applying a polarizing magnetic field. Another requirement is a Q at all frequencies of interest of 10 or more, at a maximum flux density at the lower frequency limit of 0.01 Wb/m^2. In co-operation with the Brookhaven group, the Philips Laboratories at Eindhoven have developed a ferrite which satisfies these requirements, as is illustrated by *fig. 17*. Samples of this ferrite are now under test at two laboratories associated with the high energy accelerator design.

Summary. The Cosmotron proton accelerator, designed and built by the Brookhaven National Laboratory, U.S.A., has already delivered a beam of protons with an energy of 2300 million electron volts and is expected to yield protons of 3000 MeV. The accelerator unit of this machine consists of a large "transformer", the ferromagnetic core of which is excited by an alternating current whose frequency during the acceleration period of one second varies from 0.35 Mc/s to 4.2 Mc/s. At these high frequencies only ferrite materials, which have very low eddy-current losses, could be used for the core. The major part of the ferrite core was manufactured by Philips at Eindhoven after a suitable variety of Ferroxcube meeting the permeability and figure of merit requirements had been selected by the Brookhaven Laboratory in co-operation with the Philips Laboratories at Irvington, U.S.A., and at Eindhoven. Some glimpses of the problems encountered in the manufacture of the Ferroxcube core are given in this article. Mention is also made of the development at Eindhoven of another variety of Ferroxcube which will be suitable for the accelerator units of machines based on the principle of "strong focussing", which are expected to yield particles of energies of 30 000 MeV or more.



Electron microscope photograph of ferroxdure

Surface of a block of ferroxdure ($BaFe_{12}O_{19}$) containing a slight excess of BaO, sintered at 1200 °C, sawn through and then reheated to a somewhat higher temperature. Photographs such as this can provide information on diffusion and grain growth in the process of recrystallization.

The photograph was made with a carbon replica (shadowed with palladium), i.e. a carbon skin evaporated on to a polystyrene impression prepared from the unpolished surface of the ferroxdure. The photograph was taken with an electron microscope EM 75 kV with an anastigmat lens.

(From: Philips tech. Rev. **18**, 237, 1956/57.)

FERROXDURE II AND III, ANISOTROPIC PERMANENT MAGNET MATERIALS

by A. L. STUIJTS, G. W. RATHENAU *) and G. H. WEBER. 621.318.124:538.246.2

*The $(BH)_{max}$ product, i.e. the maximum value of the product of the induction B and the magnetizing force H for points on the demagnetization curve, is a measure of the quality of any permanent magnet material. In Ferroxdure **), a ceramic permanent magnet material already discussed in this Review, values of $(BH)_{max}$ up to about 7200 J/m³ [0.9 × 10⁶ gauss oersted] are attained. This article describes the development of an anisotropic type of Ferroxdure, which has a $(BH)_{max}$ product comparable to that of modern magnet steels, viz. up to about 28 000 J/m³ [3.5 × 10⁶ gauss oersted].*

Characteristic properties of Ferroxdure

Ferroxdure **) is the name given to a new group of ceramic permanent magnet materials, some properties of which have been fully described in an earlier issue of this Review [1]). The magnetically essential phase of Ferroxdure I (which is now in quantity production) is the compound $BaFe_{12}O_{19}$, whose crystal structure is hexagonal, with close-packed oxygen ions. One of the oxygen sites in this structure is occupied by a Ba^{2+} ion; the Fe^{3+} ions occupy five non-equivalent lattice sites in the interstices between the oxygen ions.

The saturation magnetization of the compound $BaFe_{12}O_{19}$ is low compared with that of iron; extrapolation to absolute zero gives $J_0 = 0.66$ Wb/m² [$4\pi J_0 = 6600$ gauss] [2]) as against 2.18 Wb/m² [21 800 gauss] for iron. Again, at room temperature the value J_s for this compound is 0.475 Wb/m² [$4\pi J_s = 4750$ gauss], whereas the corresponding value for iron is 2.15 Wb/m² [21 500 gauss]. This relatively low saturation value and also the variation of the paramagnetic susceptibility as a function of temperature, suggest that the source of magnetization in the compound is what may be described as non-compensated antiferromagnetism, or ferrimagnetism [3]). Although the magnetic moment of each Fe^{3+} ion in this ionic compound is more than twice the moment per atom of metallic iron, only one third of the total moment of these ions is really effective, since the residual elementary moments are oriented antiparallel to each other and therefore cancel. Moreover, the magnetization is all the lower in relation to that of iron owing to the fact that the structure of the compound is greatly "diluted" with oxygen ions.

The strong coercive force exhibited by Ferroxdure is attributable to a pronounced magnetic anisotropy, in this case mainly crystal anisotropy, as will now be explained. In every Ferroxdure crystal there exists one direction of easy magnetization (preferred direction), parallel to the hexagonal axis, and only a strong magnetic field can turn the magnetization away from this preferred direction. It is to this distinct crystal anisotropy, combined with a relatively low saturation magnetization, that Ferroxdure owes its magnetic hardness. As a result of the low saturation magnetization, the couple

$$M = J_s H \sin \varepsilon, \quad \ldots\ldots \quad (1)$$

which a magnetic field at an angle ε to the preferred

*) Now professor at the University of Amsterdam; formerly with the Philips Laboratories at Eindhoven.

**) Known in U.S.A., U.K., and some other countries under the trade name "Magnadur".

[1]) J. J. Went, G. W. Rathenau, E. W. Gorter and G. W. van Oosterhout, Ferroxdure, a class of new permanent magnet materials, Philips tech. Rev. **13**, 194-208, 1951/52.

[2]) Rationalized Giorgi-units are employed throughout this article (see Philips tech. Rev. **10**, 55-60, 1948/49 and **13**, 194, 1951/52. For convenience, the formulae and values as expressed in unrationalized cgs-units (Gauss units) are also quoted, in square brackets.

[3]) See article referred to in note [1]).

direction can exert to rotate the magnetization away from this direction is relatively weak, whereas the couple

$$M = K \sin 2\Theta \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

tending to turn the magnetization back to the preferred direction is strong (K is the anisotropy constant and Θ the angle which the direction of of magnetization in a field H makes with the preferred direction).

Given this process of rotation alone, the coercive force, as computed for a material in which the preferred directions of the constituent crystals are distributed at random, is:

$$_JH_c = 0.96\, K/J_s\,.$$

For Ferroxdure at room temperature, $K = 3.1 \times 10^5$ J/m^3 [3.1×10^6 erg/cm^3] and $J_s = 0.475$ Wb/m^2 [$4\pi J_s = 4750$ gauss]: hence a value of $\mu_0 {}_JH_c = 0.79$ Wb/m^2 [$_JH_c = 7900$ oersted] is calculated. In practice, however, a value of about 0.3 Wb/m^2 [3000 oersted] is obtained in the case of Ferroxdure I, which suggests that Bloch wall movements, which occur far more readily, are also involved. Now, no stable Bloch walls can be formed in particles of diameter less than a certain critical value. In the case here considered, the critical diameter is about 1μ, which, again by reason of the distinct crystal anisotropy and relatively low saturation magnetization of the material, is considerably larger than the critical diameter for iron and cobalt crystals. Accordingly, an effective sintering process, producing a compact material whose constituent particles are smaller than the critical diameter, is the first essential. In practice, however, the sintered product will invariably contain a number of larger particles, which tend to diminish the coercive force by fostering the formation and movement of Bloch walls.

Permanent magnets are often evaluated in terms of the $(BH)_{max}$ product, that is, the maximum value of the product of the induction B and the field H for points on the demagnetization curve. The volume of material required to maintain a given magnetizing force in a particular air-gap, broadly speaking, is inversely proportional to the $(BH)_{max}$ product. Hence there is every reason to increase the value of this product as far as possible. The formula for an "ideally permanent" magnet ($_JH_c = \infty$, $\mu_0 {}_BH_c = B_r$) is (see *fig. 1*):

$$\left.\begin{aligned} (BH)_{max} &= \frac{1}{\mu_0}\left(\frac{B_r}{2}\right)^2 \\ \text{or,}\quad \left[(BH)_{max} &= \left(\frac{B_r}{2}\right)^2\right] \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (3)$$

In the case of Ferroxdure I, $\mu_0 {}_JH_c \gg B_r/2$ [$_JH_c \gg B_r/2$], which allows the above formula to hold reasonably well, as will be seen from the fact that the theoretical maximum is:

$$(BH)_{max} \approx \frac{1}{\mu_0}\left(\frac{0.21}{2}\right)^2 \approx 8800 \text{ J/m}^3$$

$$\text{or,}\quad \left[(BH)_{max} \approx \left(\frac{2100}{2}\right)^2 \approx 1.1 \times 10^6 \text{ gauss Oe}\right],$$

as compared with 7200 J/m^3 [0.9×10^6 gauss oersted], which is the value established by experiment.

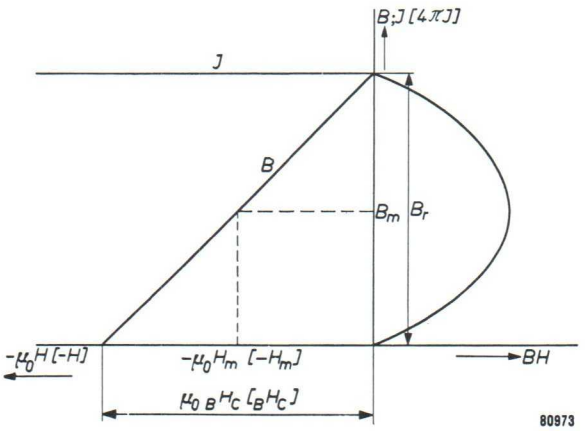


Fig. 1. Demagnetization curve for a material of which $_JH_c = \infty$. In the case here considered,

$$(BH)_{max} = \frac{1}{\mu_0}\left(\frac{B_r}{2}\right)^2, \quad \text{or} \quad \left[(BH)_{max} = \left(\frac{B_r}{2}\right)^2\right].$$

Whereas a low remanence limits the value of $(BH)_{max}$ in Ferroxdure I, the fact that $\mu_0 {}_JH_c \ll B_r/2$ [$_JH_c \ll B_r/2$] is mainly responsible for the limitation of this product in *metallic* magnets.

Ferroxdure I is very suitable for use in cases where a high coercive force is required, that is, where the material is exposed to very strong demagnetizing fields, as in focusing magnets for television tubes [4]), magnets for bicycle dynamos, yoke magnets and magnetic clutches. By virtue of its high resistivity, this material can be employed for producing a biasing field in high-frequency transformer cores of Ferroxcube [5]). Other advantages of Ferroxdure are its simple composition and low density, and also the fact that it is made of relatively inexpensive and readily obtainable materials. Moreover, the magnetization loss caused by demagnetizing fields is limited by the low permeability of this material.

The use of Ferroxdure I to sustain a strong magnetizing force in a narrow air-gap, as in a loudspeaker circuit, is confined mainly to those cases which

[4]) J. A. Verhoef, Philips tech. Rev. **15**, 214-220, 1953/54 (No. 7).

[5]) W. Six, Philips tech. Rev. **13**, 301-311, 1951/52.

favour the use of a very flat magnet. However, a complete change in the situation, particularly in this respect, has taken place as a result of a new development [6]), which has led to the production of Ferroxdure materials showing an appreciably higher induction on the demagnetization curve, and having $(BH)_{max}$ values comparable with those of modern anisotropic magnet steels.

Anisotropic Ferroxdure

In the isotropic material, the preferred directions of magnetization of the constituent particles are oriented at random. It is found that the remanence B_r of material so constituted is equal to half the saturation magnetization, i.e. about 0.21 Wb/m² [2100 gauss]. Given a material in which the preferred directions of the individual particles are aligned parallel to one another, however, the remanence in this direction will be the same as in a single crystal, viz.:

$$B_r = J_s, \qquad [B_r = 4\pi J_s].$$

Accordingly, we have as a maximum:

$$(BH)_{max} \approx \frac{1}{\mu_0}\left(\frac{0.42}{2}\right)^2 \approx 35\,000 \text{ J/m}^3$$

$$\left[(BH)_{max} \approx \left(\frac{4200}{2}\right)^2 \approx 4.4\times 10^6 \text{ gauss oersted}\right].$$

To procure this optimum value, it is necessary to ensure that in all circumstances: $\mu_0 JH_c > B_r/2$ $[JH_c > B_r/2]$. The fulfilment of this condition becomes more and more difficult as the remanence B_r increases. Moreover, the coercivity is invariably weaker in anisotropic than in isotropic material, owing to the fact that the changes in the direction of magnetization are caused by the formation of Bloch walls. The field H^* required to form a Bloch wall in a particle is a minimum when the preferred direction of magnetization of the particular particle coincides with the direction of the applied field H; also, the movement of Bloch walls across grain boundaries takes place far more readily in anisotropic materials, which in this respect resemble single crystals.

Accordingly, the problem is to procure an anisotropic orientation of the grains in Ferroxdure, at the same time avoiding any appreciable weakening of the coercive force.

Such anisotropy can be induced in several ways. For example, one approach to the problem is to take advantage of the non-spherical shape of the Ferroxdure single crystals. These crystals exhibit very strong growth along the basal plane, but relatively slight growth in the hexagonal direction coinciding with the preferred direction of magnetization. By packing the powder in a steel tube, welding the ends of this tube and passing it through rolls at a high temperature, a preferred direction of magnetization at right-angles to the plane of rolling is obtained. Similarly, it is possible to produce a small amount of anisotropy merely by compressing the Ferroxdure powder under high pressure.

However, a far superior method is to employ the magnetocrystalline anisotropy of Ferroxdure to the best advantage. When an external magnetic field H is applied, (*fig. 2*), the preferred axes of magnetization of any particles which are free to rotate will move into alignment with this field. This produces in the material a texture corresponding approximately to the magnetic state of a single crystal, that is, one preferred direction, parallel to the direction of the field.

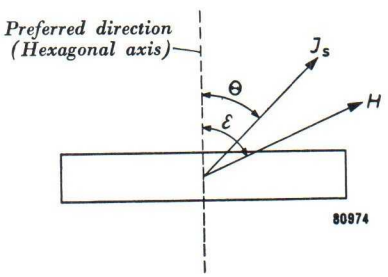


Fig. 2. The state of equilibrium of the magnetization J_s in a uniaxial magnetic particle when an external field H is applied at an angle ε to the preferred direction.

The aligning couple exerted by the magnetization J_s on the crystal is:

$$M = K \sin 2\Theta. \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

The absolute value of this couple reaches a maximum equal to K when $\Theta = 45°$ or 135°, and is zero when $\Theta = 0°$, 90° or 180°.

It will be seen, then, that the highest gain in $(BH)_{max}$ is attained when those particles whose hexagonal axes make an angle ε of about 90° with the field direction are also brought into alignment. If at $\varepsilon = 90°$ the magnetizing force exceeds $2\ K/J_s$, the magnetization will rotate in the direction of H ($\Theta = 90°$), the orientation of the crystal axis remaining unchanged; the mechanical couple is then zero. The rotation of the magnetization is instantaneous; hence the application of the field does not usually produce a simultaneous rotation of the particle. Accordingly, the latter will also be

[6]) G. W. Rathenau, J. Smit and A. L. Stuijts. Z. Phys. **133**, 250, 1952. A. L. Stuijts, G. W. Rathenau and E. W. Gorter, J. appl. Phys. **23**, 1282, 1952.

able to remain out of alignment when the magnetizing force is increased, However, if H be small enough, e.g. $H = \sqrt{2}K/J_s$, a sudden application of the field will produce the state corresponding to $\Theta = 45°$, i.e. that with the maximum couple: hence the particle will also be brought into alignment with the field. This suggests that the value of an external magnetic field suddenly applied should be approximately equal to $\mu_0 H = \mu_0 \sqrt{2}\, K/J_s \approx 1.1$ Wb/m² [$H \approx 11000$ oersted].

Method of preparation

To apply in practice the process considered here, it is necessary to fulfil a number of conditions.

Firstly, the particles subjected to the field treatment must be magnetic: the basic material is isotropic Ferroxdure. Secondly, the basic material must consist of separate particles, each having only one crystal orientation; this condition can be fulfilled by grinding the sintered material until it contains no aggregates comprising more than one crystal.

A magnetic particle introduced into a magnetic field will turn to the particular direction most favourable from a magnetic point of view, if able to overcome the local forces opposing this orientation. To procure optimum efficiency, then, it is necessary to minimize the tendency of one particle to impede the orientation of another. Since the flatness of Ferroxdure crystals is unfavourable in this respect, too close a packing of the powder subjected to the magnetic field treatment has to be avoided.

When once the particles in the magnetic field are properly aligned, they must be stabilized without any disturbance of the texture thus obtained. By mixing the powder with a little melted paraffin wax and introducing this sludge into the magnetic field, stabilization is obtained when the wax solidifies. However, this method is of little practical value owing to the fact that the induction B is proportional to the density of the magnetic material, which would be low. In the case of ceramic materials, a product combining stability with adequate density is normally obtained by press-forming the powder at high pressure and then sintering it. Experience has shown that pressing is an excellent method of stabilizing the particles, provided that it takes place in a magnetic field; although the forces exerted during this process far exceed the aligning force of the magnetic field, they do not seriously disturb the texture of the material. As regards sintering, the condition that undue particle growth must be avoided in order to maintain a high coercive force is even more important than in the case of Ferroxdure I.

Finally, the high temperature employed in the sintering process, which is well above the Curie point, must not spoil the magnetic texture of the material. This is a point to which we shall shortly return.

The powder is press-formed as an aqueous sludge. A diagram of the apparatus employed is seen in *fig. 3*. The die is made of non-magnetic material,

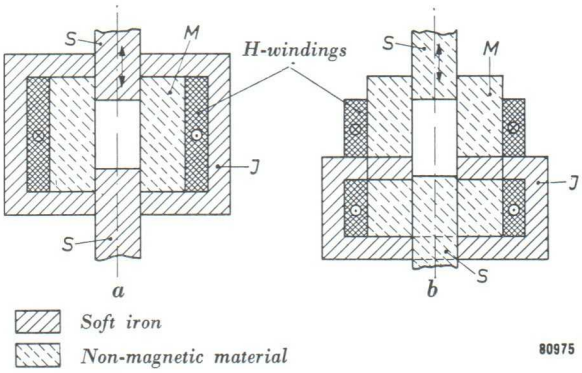


Fig. 3. Diagram of the dies M used to compress the powder in a magnetic field. S = plunger. J = iron yoke.
a) Direction of compression // to direction of field.
b) Direction of compression ⊥ to direction of field.

and iron pole pieces are provided to intensify the field within the powder. The surplus liquid is ejected along the plunger. Experience has shown that the relative directions of the compressing force and the magnetic field are immaterial.

After pressing, the powder still contains a large amount of water, and must be dried thoroughly before sintering. The technique of press-forming such an aqueous sludge differs considerably from that employed for dry powder and, in fact, the adaption of the process to large-scale production involved the solving of many technological problems.

The effect of sintering

It will be seen, then, that the texture of the compressed powder is not ideal; the particles obstruct one another during alignment, compression affects the texture, and the aligning couple decreases with angle ε (see fig. 2).

It is also necessary that the texture be preserved during sintering. However, the notable feature of this process is that it not only preserves, but actually enhances, the texture. Consider *table I*, which shows the remanence values of a particular material as measured parallel and at right-angles to the preferred direction, after sintering at different temperatures. In the case of a fully anisotropic material

Table I. Remanences of anisotropic Ferroxdure as measured parallel and at right-angles to the direction of the field in which the crystals are oriented, after sintering at different temperatures.

Sintering temp.	Density		Remanence $B_{r//}$ // to original field-direction		Remanence $B_{r\perp}$ ⊥ to original field-direction		$\frac{B_{r\perp}}{B_{r//}}$	Coercivity	
°C	$\frac{kg}{m^3}$	$\frac{g}{cm^3}$	$\frac{Wb}{m^2}$	gauss	$\frac{Wb}{m^2}$	gauss		$\mu_0 {}_JH_c$ $\frac{Wb}{m^2}$	${}_JH_c$ oersted
1250	4060	4.06	0.270	2700	0.121	1210	0.45	0.321	3210
1275	4360	4.36	0.292	2920	0.125	1250	0.43	0.285	2850
1300	4660	4.66	0.326	3260	0.117	1170	0.36	0.247	2470
1320	4840	4.84	0.362	3620	0.072	720	0.20	0.117	1170
1340	4880	4.88	0.385	3850	0.025	250	0.07	<0.020	<200

$B_{r\perp}/B_{r//} = 0$, whereas in that of the isotropic material $B_{r\perp}/B_{r//} = 1$.

From the magnetic point of view, the texture changes from an imperfect one to that of a single crystal after the Ferroxdure has been sintered at a high temperature. *Fig. 4* shows part of the hysteresis loops of such a material, again measured parallel and at right-angles to the preferred direction. *Figures 5a* and *5b* are photo-micrographs showing the surfaces of a cube of the same material. Note the distinct texture produced by sintering at a high temperature. In the upper photograph the direction of the field applied during compression is at right-angles to the plane of the paper; in the lower photograph it is parallel.

This improvement of textural alignment during sintering could be explained by supposing the small particles in the material to be eliminated in favour of the larger ones by the action of interfacial tensions at the crystal boundaries during the sintering process; this explanation, if correct, implies that the smaller particles are not properly aligned.

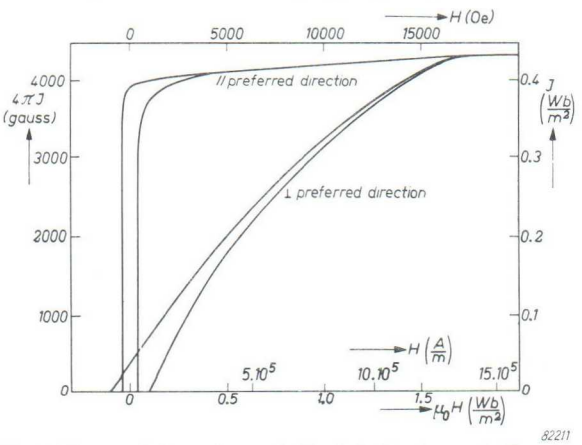


Fig. 4. Hysteresis loops (upper half only) of an isotropic Ferroxdure sample sintered at a high temperature, as measured // and ⊥ to the preferred direction. The magnetic texture closely resembles that of a single crystal.

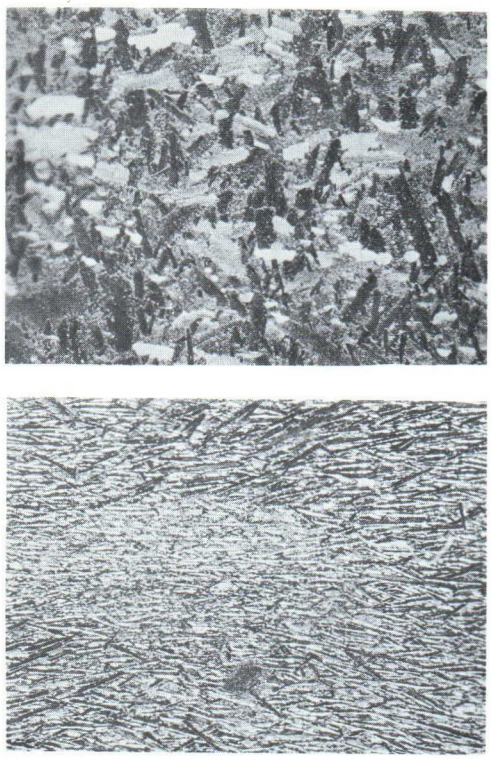


Fig. 5. Photo-micrograph of Ferroxdure with crystals aligned. In (*a*) the hexagonal basal planes are in the plane of the photograph, and in (*b*) they are at right angles to this plane. Magnification 10×.

However, experience has shown that the most significant improvement of the texture takes place when the density of the material is already about 90% of that of an ideal single crystal; it will be seen from the table that the coercive force then decreases very appreciably.

Accordingly, the following explanation is in our opinion more feasible, although, no doubt, the above-mentioned mechanism is also involved.

In the case of metals, the beneficial effect of grain growth on a preferred orientation has been closely studied. This effect can be observed direct with the aid of an emission electron-microscope [7]). It is seen that appreciable grain growth takes place only when the difference in orientation between adjacent crystals is such that the boundary energy between these crystals is high. Given a matrix of identically, or almost identically, oriented crystals surrounding one crystal differently aligned, this odd crystal will usually be eliminated, whereas the boundaries between the similarly aligned crystals will be preserved. Again, it is found that the odd crystal can grow only if it is large in relation to the crystals of the matrix.

[7]) G. W. Rathenau and G. Baas, Physica **17**, 117, 1951.

A calculation analogous to that suggested by C.S. Smith for polygonal crystals [8]) may be employed for the flat crystals of Ferroxdure. Consider *fig. 6*, which shows a cross-section of a thin, flat crystal, of thickness δ and diameter D, making an angle Θ with the neighbouring mutually aligned crystals, of thickness d. If σ_{dt} be the boundary energy per unit area between the crystals meeting at angle Θ, and σ_{tt} that between the aligned crystals, the odd crystal will be able to grow provided that this growth reduces the overall boundary energy, that is if

$$2\,D\,\sigma_{dt} < \frac{D\sin\Theta}{d}\,\frac{\delta}{\sin\Theta}\,\sigma_{tt}.$$

It has been assumed in this calculation that angle Θ is not small, and that $D \gg \delta$.

From the above formula, then:

$$\frac{\delta}{d} > \frac{2\sigma_{dt}}{\sigma_{tt}}.$$

Now the boundary energy σ_{dt} increases very sharply with the deviation from the general orientation; in most cases, then, the incorrectly aligned crystal will be eliminated.

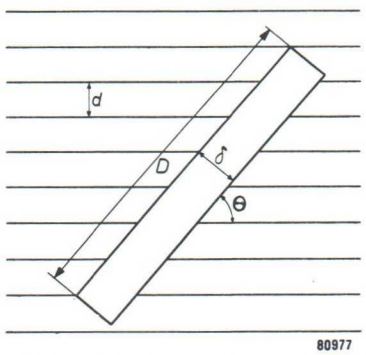


Fig. 6. Particle whose orientation deviates from that of the adjacent, mutually aligned crystals. As a general rule, such an oddly oriented crystal will be eliminated from the system.

Ferroxdure II and III

It will be seen from the above that any improvement in the anisotropy, that is, any increase in the remanence, is invariably associated with grain growth and therefore incompatible with a high coercivity (see table I). Accordingly, the theoretical maximum $(BH)_{max}$ value has so far proved unattainable in practice, owing to the fact that when the remanence is sufficiently high, $\mu_0 J H_c$ [JH_c] falls below the value corresponding to $B_r/2$. Nevertheless, the $(BH)_{max}$ values obtained by the method described above are considerably higher than those of isotropic material. It is possible to produce anisotropic materials with demagnetization curves specially suited to particular ranges of applications. At present, two types of anisotropic Ferroxdure magnet material are in production (see *fig. 7*).

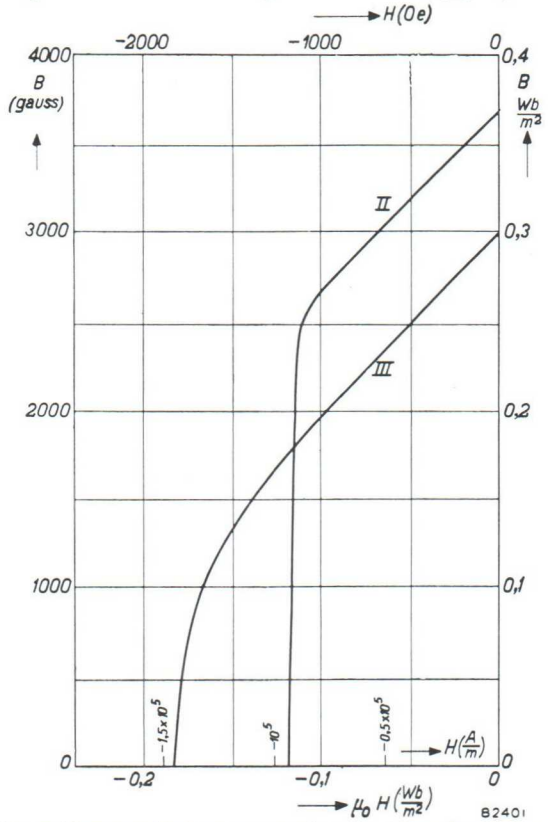


Fig. 7. Demagnetization curves of the anisotropic permanent magnet materials Ferroxdure II and Ferroxdure III.

One of them, Ferroxdure II, is intended for applications involving a fixed air-gap, in which high induction is the most important requirement, e.g. for loudspeakers. Its magnetic properties are as follows:

$B_r = 0.35 - 0.40$ Wb/m² [3500 - 4000 gauss]
$\mu_0 J H_c = 0.11 - 0.15$ Wb/m² [JH_c = 1100 - 1500 oersted]
$(BH)_{max} = 19\,000 - 24\,000$ J/m³ [2.4 - 3.0 × 10⁶ gauss oersted]

The other, Ferroxdure III, is intended for use in a strong demagnetizing field, or in circumstances where the working point may vary, as in magnetic clutches, dynamo magnets and yoke magnets. The magnetic properties of this material are:

$B_r = 0.28 - 0.32$ Wb/m² [2800 - 3200 gauss]
$\mu_0 J H_c = 0.19 - 0.25$ Wb/m² [JH_c = 1900 - 2500 oersted]
$(BH)_{max} = 16\,000 - 18\,500$ J/m³ [2.0 - 2.3 × 10⁶ gauss oersted].

[8]) C. S. Smith, Trans. Am. Inst. Mining Met. Engrs. **175**, 15, 1948.

It can be seen that in the case of Ferroxdure II, $\mu_0 JH_c < B_r/2$ [$JH_c < B_r/2$], whereas for Ferroxdure III, $\mu_0 JH_c > B_r/2$ [$JH_c > B_r/2$).

For the average properties of Ferroxdure types I, II and III at room temperature, see *table II*.

Table II. Average values of certain physical properties of Ferroxdure I, Ferroxdure II and Ferroxdure III at room temperature.

Magnetic moment per unit mass		8.75×10^{-5} Wb m/kg [70 gauss cm³/g]			
Density		4800-4900 kg/m³ [4.8-4.9 g/cm³]			
Density of ideal single crystal		5300 kg/m³ [5.3 g/cm³]			
Electrical resistivity		$> 10^6$ Ωm [10^8 Ωcm]			
Temperature coefficient of induction		—0.2 % per °C			
		Ferroxdure			
		I	II	III	
Remanent induction		0.205	0.37	0.30	Wb/m²
		2050	3700	3000	gauss
Coercivity	$\mu_0 JH_c$	0.26	0.12	0.21	Wb/m²
	JH_c	2600	1200	2100	oersted
Coercivity	$\mu_0 BH_c$	0.15	0.12	0.20	Wb/m²
	BH_c	1500	1200	2000	oersted
$(BH)_{max}$		7200	22000	16750	J/m³
		$0.9 \cdot 10^6$	$2.8 \cdot 10^6$	$2.1 \cdot 10^6$	gauss oersted
Working point	B	0.11	0.26	0.17	Wb/m²
		1100	2600	1700	gauss
	$\mu_0 H$	0.08	0.10	0.125	Wb/m²
	H	800	1000	1250	oersted
Curie point		450 °C			

A high temperature coefficient of induction for points on the demagnetization curve is still an inconvenient feature of the anisotropic materials Ferroxdure II and Ferroxdure III.

A magnetizing force equivalent to $\mu_0 H = 1.40$ Wb/m² [$H = 14\,000$] oersted is required to produce effective saturation in Ferroxdure I, whereas Ferroxdure II and Ferroxdure III can be saturated by a much weaker field.

In small-scale production, magnetic properties appreciably higher than those specified above can be attained. For example, a $(BH)_{max}$ of about 26 000 J/m³ [3.3×10^6 gauss oersted] was obtained on each type of demagnetization curve, Ferroxdure II and Ferroxdure III.

The best result obtained so far is a $(BH)_{max}$ of 28 000 J/m³ [3.5×10^6 gauss oersted] [9]); it will be seen from *fig. 8*, which shows the demagnetisation curve of the material concerned, that this is Ferroxdure II.

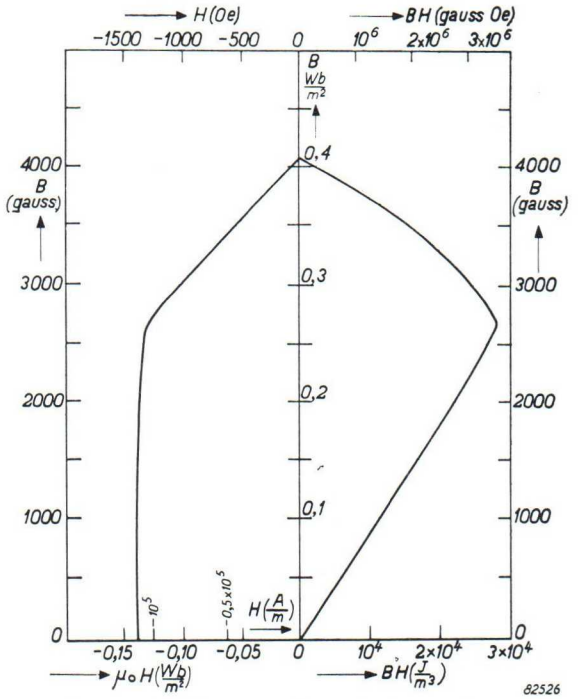


Fig. 8. Demagnetization curve of a particular sample of Ferroxdure II the $(BH)_{max}$ value of which is 28 000 J/m³ [3.5×10^6 gauss oersted].

Summary. Owing to the low remanence of Ferroxdure I, the $(BH)_{max}$ value of this material is likewise low. The preferred directions of the particles constituting Ferroxdure I, an isotropic material, are distributed at random; hence the remanence of this material is no more than half its saturation value. However, by virtue of the distinct magnetic anistropy of the constituent crystals, the remanence of such materials can be increased considerably by press forming the powder in a magnetic field before sintering it. Before being so pressed, the Ferroxdure powder is ground until the diameter of the individual grains is less than the critical value for the formation of Bloch walls. Pressing in a magnetic field aligns the hexagonal axes of the grains more or less parallel to the direction of the lines of force, thus producing in this direction $(BH)_{max}$ values 3 to 4 times as high as that of the isotropic material. It is found that grain growth taking place during the sintering process enhances the magnetic anisotropy of the material considerably, although the decrease in coercivity caused by this growth limits the possible $(BH)_{max}$ value to some extent. Two anisotropic materials, Ferroxdure II and Ferroxdure III, each having its own particular range of applications, are described.

[9]) F. G. Brockman and W. G. Steneck, Philips tech. Rev. **16**, 79-87, 1954/55 (No. 3).

STEREO REVERBERATION

by R. VERMEULEN.

534.844:534.846.4:621.395.625.3

While too long a reverberation time makes speech unintelligible, a reverberation time which is too short makes music sound "dry" and brittle. Many varieties of acoustic materials are available for shortening the reverberation time and improving the intelligibility. Lengthening of the reverberation time and — perhaps more important — making the sound diffuse, can be achieved by electro-acoustic means. Tests have shown that in this way a good theatre hall can be made suitable for concerts.

Introduction

The acoustic properties of a hall are determined by the behaviour of the sound waves, in particular those which are reflected from the walls. The importance of the reflected sound can be well understood if one has listened to an open-air speech made without an amplifier installation, or if one imagines an open-air performance by a symphony orchestra without this aid.

A simple calculation will show that in a hall at even quite a small distance from the orchestra, the direct sound can sometimes be weaker than the reflected sound. At a distance r from a sound source of power P, the energy density of the direct sound is equal to $P/(4\pi r^2 c)$, where c is the velocity of sound. In a hall with a volume V and reverberation time T (defined as the time in which the sound intensity decreases by 60 dB, after the source has ceased radiating), the energy density of the indirect sound [1]) is $PT/(13.8\ V)$. These two energy densities are equal at a distance $r_0 = (1.1\ V/cT)^{\frac{1}{2}}$. Values of r_0 corresponding to various practical values of V and T are as follows:

$V =$	100	1000	10 000 m³
$T =$	0.7	1.0	1.5 s
$r_0 =$	0.7	1.8	4.5 m.

At distances greater than r_0 the indirect sound predominates. It is seen that this can occur at distances of only a few metres.

The qualities required of a hall for speech and for music are quite different. In a theatre, the intelligibility is of primary importance. If speech is to be clearly understood, the reflected sound must reach the audience with so little delay that it reinforces the direct sound but does not overlap the sounds which follow. For the latter, the persistence of the preceeding sound must be regarded as "background noise", which adversely affects intelligibility. As a rough guide, one can say that all sound which reaches the audience within 50 milliseconds can be regarded as useful sound [2]). Erwin Meyer has formulated the idea of "clearness" or "definition" [3]), which he defines as follows:

$$\frac{\int_0^{50\ \mathrm{msec}} p^2(t)\ \mathrm{d}t}{\int_0^{t_1} p^2(t)\ dt},$$

where p is the sound pressure, t is the time (measured from the moment at which the source is silenced) and t_1 is a time much greater than 50 msec.

For a concert hall, on the other hand, the first requirement is not intelligibility, but a fine, full tone. Here it is much more difficult to specify the requirements. For speech, the reverberation must be accepted as an inevitable, disturbing accompaniment to the useful sound, because it simply is not possible to silence the sound suddenly after 50 msec. For music we know that the reverberation time not only may be, but must be, longer. The optimum value is clearly dependent on the nature of the music. It is often seen that a composer has consciously taken into account the acoustics of the space (church, concert hall, room) where he wanted his music to be played. The inclination to sing in the bathroom can probably be largely attributed to the long reverberation time of this acoustically "hard" room.

It is becoming increasingly clear that the reverberation time is not the only property governing the suitability of a hall for musical performances.

[1]) See for example A. Th. van Urk, Auditorium acoustics and reverberation, Philips tech. Rev. **3**, 65-73, 1938.

[2]) See the article referred to in [1]), pp. 72-73.

[3]) E. Meyer, Definition and diffusion in rooms, J. Acoust. Soc. Amer. **26**, 634, Sept. 1954.

One might even conjecture that here, too, reverberation is merely an inevitable subsidiary effect. Just as important, or perhaps even more so, is the "diffuse-ness" of the sound (and possibly also the nature of the fluctuations of the reverberation).

To study these phenomena more precisely, we attempted to produce an artificial diffuse reverberation in the laboratory by means of distributed loudspeakers which repeated the music played, with controllable intensity and lag. This experiment appeared to improve the acoustics to such an extent that we ventured to take the bold step of using this artificial reverberation to make a theatre suitable for concerts. We propose that this artificial diffuse reverberation be called "stereo reverberation".

Our first installation was in the Philips Theatre at Eindhoven, whose acoustical properties as a theatre were very satisfactory as a result of rebuilding in 1935, but which left much to be desired as a concert hall. In addition to this theatre, a hall known as the "Gebouw voor Kunsten en Wetenschappen" (Arts and Sciences Hall) in The Hague, is now fitted with a permanent installation for stereo reverberation [4]).

Principle of the installation for stereo reverberation

The delay wheel

The principle of the stereo reverberation installation may be explained with the help of *fig. 1*. Controllable time lags are obtained by means of magnetic recording and playback. Magnetic material such as is used for magnetic tape is coated on the rim of a wheel ("delay wheel") [5]). The music is recorded on this material via a microphone and a recording head. A number of play-back heads — in the final apparatus four, in experimental types six (*fig. 2*) or more (*fig. 3*) — are mounted around the circumference of the wheel and connected via separate channels to loudspeakers. These are installed in various places in the hall: on the ceiling, along the balustrade of the balcony, in a lighting cornice, in "dead" corners under the balcony, etc. Between the last play-back head and

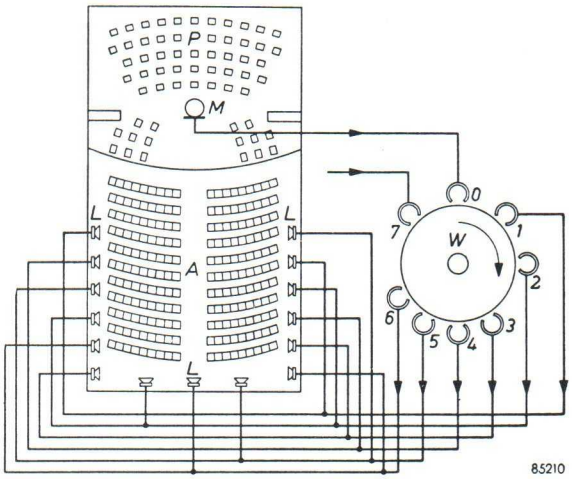


Fig. 1. Installation for simulating indirect sound with various time lags. *A* auditorium. *P* platform. *M* microphone. *W* delay wheel, coated on the edge with magnetic material suitable for magnetic sound recording. *O* recording head, *1...6* play-back heads. *7* erasing head. *L* loudspeakers.

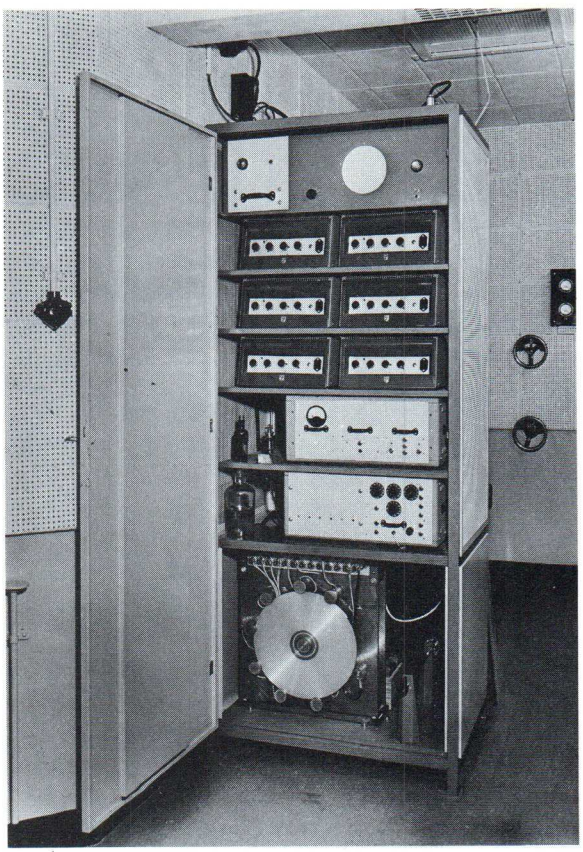


Fig. 2. Stereo reverberation installation in the Philips Theatre at Eindhoven. At the bottom of the cabinet is the delay wheel (cf. fig. 1); above it the amplifiers.

[4]) Demonstrations of stereo reverberation have also been given at the first I.C.A. Congress on Electro-acoustics, (Acustica **4**, 301, 1954), at Gravesano in Switzerland (at the invitation of the conductor W. Scherchen; see his book Musik, Raumgestaltung und Elektroakustik, Arsviva Verlag, Mainz 1955) and at the 3rd "Tonmeistertagung" of the Nordwestdeutsche Musikakademie Detmold (see D. Kleis, Elektron. Rdsch. **9**, 64-68, 1955 (No. 2)). Similar tests, but done in the open air, are described by H. S. Knowles, Acustica **4**, 80-82, 1954 (No. 1).

[5]) Others have also constructed a similar equipment, but have not created diffuse reverberation with it; see H. Schiesser, Einrichtungen zur Erzeugung künstlichen Nachhalls, Funk und Ton **8**, 361-368, 1954 (No. 7), and P. Axon and co-workers, Artificial reverberation, J. Instn. El. Engrs. **1**, 368-371, 1955 (No. 6).

the recording head is an erasing head, which ensures that the magnetic layer is blank on returning to the recording head.

For three reasons, however, the listener hears more than these seven $(1 + 6)$ reports: firstly because to each play-back head, several loud-

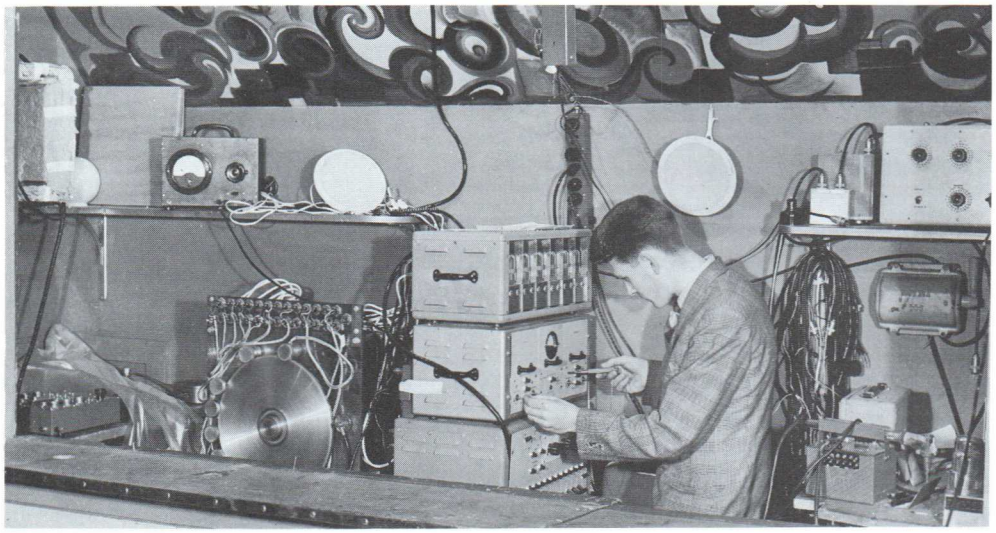


Fig. 3. Stereo reverberation installation in the "Gebouw voor Kunsten en Wetenschappen" (Arts and Sciences Hall), The Hague, in experimental form. The delay wheel was then fitted with ten play-back heads, of which six could operate simultaneously. It was possible to switch over rapidly from one set of play-back heads to another.

Let us consider the case of a sharp report produced in the hall (pulse I_0, *fig. 4*). After a certain transit time this reaches the microphone and then the artificial indirect sound begins. The latter, if we are using six play-back heads, consists in the first place of six successive reports from the loudspeakers $(I_1 \ldots I_6)$. The intensity of each of these reports can be adjusted at will by the gain controls; the time intervals are also under control, by the spacing of the heads round the wheel and by its speed of revolution.

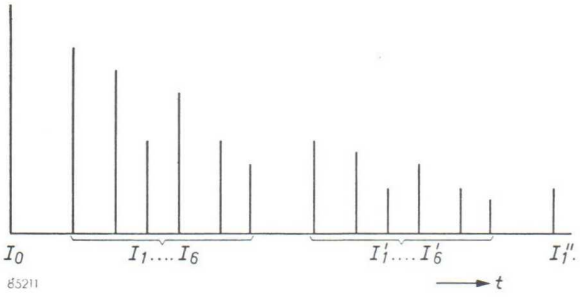


Fig. 4. If a report (pulse I_0) is produced in a hall with stereo reverberation, the six loudspeakers deliver during one revolution of the delay wheel the six reports $I_1 \ldots I_6$. If electrical feedback is applied from the sixth play-back head to the recording head, the loudspeakers give a second series of six reports $(I_1' \ldots I_6')$, a third, and so on. (Acoustic feedback is here neglected.)

speakers are connected, which are dispersed throughout the hall and are at different distances from the listener, so that the transit times are different; secondly because the walls reflect both the original sound and those coming from the loudspeakers, and thirdly because the report from each loudspeaker is picked up by the microphone, once more recorded on the wheel and, though attenuated, reproduced six-fold.

All these effects contribute to the fact that the original report is followed by a large number of others, so that the reverberation time can achieve considerable values. The third effect mentioned above — the feedback from the loudspeakers to the microphone — even involves the danger that one particular note continues to sound too long and in the extreme case howls back continuously; therefore this effect must be curtailed as much as possible (we shall return to this point presently). This acoustic feedback can be advantageously replaced by electrical feedback, in which a chosen fraction of the signal from the last play-back head is fed back to the recording head. This records the signal afresh: in our example, a second series of six reports then ensues $(I_1' \ldots I_6'$ in fig. 4); this series is once more recorded and a third series follows, and

so on. By adjusting the fraction of the output signal which is fed back, each series can be attenuated to any given degree, so that the reverberation time can be chosen at will.

Calculation of the effect of the stereo reverberation on the acoustic properties of the hall

The following calculation shows, for a simple case, the effect which stereo reverberation has on the acoustic properties of the hall. Suppose that the energy-density in the hall is $E(t)$. Then $-V\mathrm{d}E/\mathrm{d}t$ is the rate at which acoustic energy in the hall diminishes. In a hall without stereo reverberation, this must be equal to the power absorbed by the walls [6]), which is proportional to E, say equal to αE. Stereo reverberation supplies extra power proportional to the energy density at some time τ ago: $\beta E(t-\tau)$. We can thus construct the equation:

$$-V\frac{\mathrm{d}E}{\mathrm{d}t} = \alpha E(t) - \beta E(t-\tau).$$

A solution to this is:

$$E = E_0 \exp(-mt),$$

where

$$mV = \alpha - \beta \exp(m\tau);$$

the quantitity m is inversely proportional to the reverberation time T, and equal to $13.8/T$. For small values of $m\tau$, an approximate value of $\exp(m\tau)$ is given by $1 + m\tau$. Thus we have:

$$m \approx \frac{\alpha - \beta}{V + \beta\tau}.$$

From this it is seen that increasing the strength of the stereo reverberation (β) has the same effect as decreasing the absorption (α) or increasing the volume (V) of the hall. An increased lag τ also gives the effect of a hall of larger volume; we shall return later to this point.

The microphones

The above calculation was based on the assumption that one could speak of "the" energy density in the hall. In many cases the total energy (sum of potential and kinetic energy) is indeed fairly evenly distributed throughout the hall. With standing waves, however, as is well known, the potential and kinetic energy alternate with each other, and this means that the sound pressure at constant frequency changes sharply from place to place, and, conversely, that at a particular place the sound pressure is very dependent on the frequency. The microphone which picks up the signal which is fed back to the hall via the delay wheel and the loudspeakers, responds only to the pressure at the spot. The factor β in our calculation therefore varies sharply with the frequency, and the reverberation time, which is inversely proportional to $\alpha - \beta$, varies also and to a much greater degree. If β as a function of the frequency shows a peak, then increasing the amplification will cause the note at which the peak occurs to go on sounding for a long time, while for most other frequencies, the reverberation time is not yet lengthened appreciably. With even greater amplification, the note fails to decay at all (howl-back: $\alpha - \beta$ has become negative).

The obvious solution is to try and suppress the peak by a filter in the microphone channel. However, there are so many peaks in the frequency characteristic of a hall, and these peaks are so sharp, that the suppression of all of them would be a hopeless task, particularly since the peaks are modified by all changes made in the hall or on the platform. It is worthwhile, nevertheless, to attenuate those frequency ranges in which the highest peaks occur in such a way that increase of amplification causes a number of widely separated frequencies to howl back at the same time.

In this connection it can be argued that electrical feedback via the delay wheel is a better means of obtaining a long reverberation time than acoustic feedback via the hall.

In a system with delayed feedback, the feedback signal and the input signal exhibit, in general, a relative phase difference which is proportional to the product of frequency and time lag. If there are a great number of feedback channels (as there invariably are in practice, with acoustical feedback), and if we suppose that they all make equal contributions to the input signal, then in general the contributions will show fairly random phase differences, so that the average resultant increases as the root of the number of feedback channels. There will be, however, one or more frequencies for which all the contributions are nearly in phase with each other. For these frequencies the resultant will be a maximum and equal to the algebraic sum of the contributions, and thus proportional to the number of feedback channels. This reasoning makes it clear that the ratio of the maximum value of the resultant (which occurs at a particular frequency and limits the maximum amplification) to the average value (at other frequencies) increases with the root of the number of feedback channels. In the case of a microphone and loudspeakers in the same hall, this number is very large and it is to be expected that a situation can easily arise in which one note continues to sound for a long time. With feedback via the delay wheel,

[6]) See for example the article referred to in [1]), where it is shown that the proportionality factor α is equal to $^1/_4 aAc$, where a is the average absorption coefficient and A the area of the walls.

we are using only one feedback channel (from the last playback head to the recording head) and the above-mentioned danger is thus much less. It would be possible to introduce more feedback channels, e.g. from one or more of the preceeding playback heads to the recording head. Experiments have shown that this is undesirable, and this is partly explained by the above considerations.



Fig. 5. The stage in the "Gebouw voor Kunsten en Wetenschappen" in The Hague. M_1 is the line microphone (rod with ten condenser microphones). (Other microphones visible in the photograph were for broadcasting and had nothing to do with stereo reverberation.)

Instead of a microphone which responds to the average sound density in the hall, the other extreme would be one which responds exclusively to sound coming directly from the source and is insensitive to sound from the hall [7]). The danger of notes howling back would then be completely averted. This situation can be closely enough approximated to by using a microphone having a sharp directional characteristic. An arrangement for achieving this, a so-called "line microphone", consists of a group of ten condenser microphones mounted at equal intervals along a rod rather more than a metre in length (*fig. 5*). By suitably positioning this line microphone above the orchestra, we can ensure that the loudspeaker signal at the microphone, although still making an important contribution, no longer dominates. If it is not possible to cover the orchestra adequately with one line microphone, two may be used (*fig. 6*).

The use of one or more extra microphones can be desirable in order to strengthen weak instruments in the orchestra. Thus, for example, in the Philips Theatre at Eindhoven the organ was rather weak in comparison with the orchestra and choir in the annual performance of Bach's St Matthew Passion; it appeared to be an improvement when this instrument was boosted by a microphone of its own (*fig. 7*). With such cases in mind, the stereo reverberation system was provided with several microphone channels with mixing facilities. These should, however, be used with the utmost restraint, for the sound engineer must never encroach upon the conductors prerogative for the balance of the instruments in the orchestra.

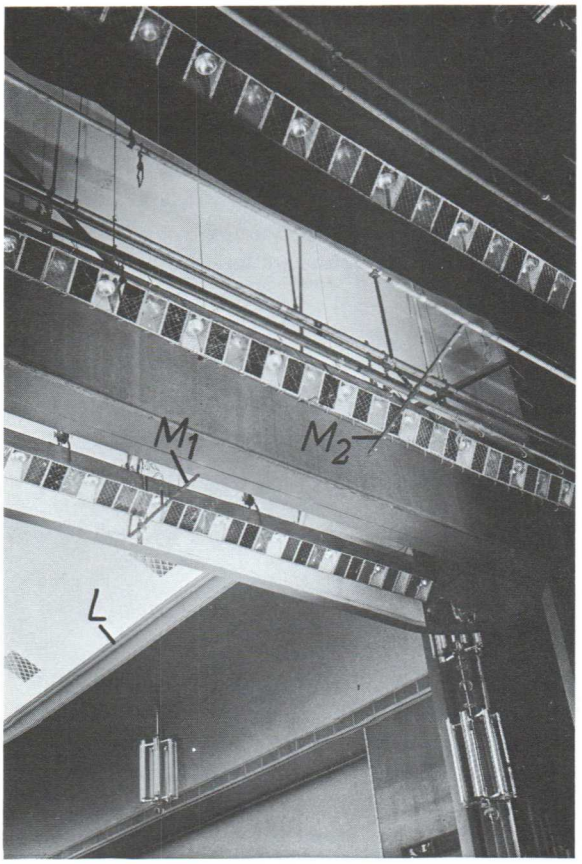


Fig. 6. In the Philips Theatre at Eindhoven two line microphones are used (M_1, M_2). The loudspeakers are mounted in a concealed position behind the lighting cornice L (compare fig. 9).

[7]) This occurs, for example, in the case of music reproduced in a hall, when it is (or was) actually performed elsewhere. We return to this point again at the conclusion.

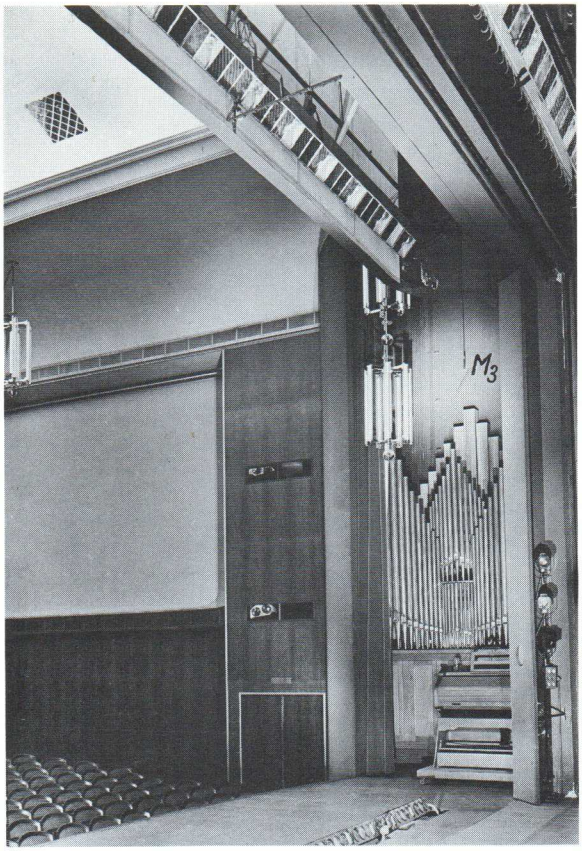


Fig. 7. The small organ in the Philips Theatre can be reinforced through an amplifier channel fed from a separate microphone M_3.

The loudspeakers

As already suggested, the diffuseness of the sound is perhaps even more important than the lengthening of the reverberation time. Diffuseness can be obtained by dispersing the loudspeakers over the hall (*fig. 8*) and connecting them to the various play-back heads. The wiring is simplified if all the loudspeakers belonging to one group (fed from the same play-back head) are connected in series. In the case of four play-back heads, one four-core cable is run around the hall, balconies, etc; where a loudspeaker is installed, the appropriate core is cut and the speaker connected in series (*fig. 9*).

The distribution of the loudspeakers between the various play-back heads should be done as randomly as possible. The only restriction is that the audience should never get the impression that the sound comes from the loudspeakers. We shall now try to explain further the general lines to follow in order to avoid this impression.

We shall use some results of the work of K. de Boer on stereophony [8]). In fact, we are here dealing

[8]) For a recapitulation of the principles of stereophony, with references to the literature, see R. Vermeulen, Philips tech. Rev. **17**, 171-177, 1955/56 (No. 6).

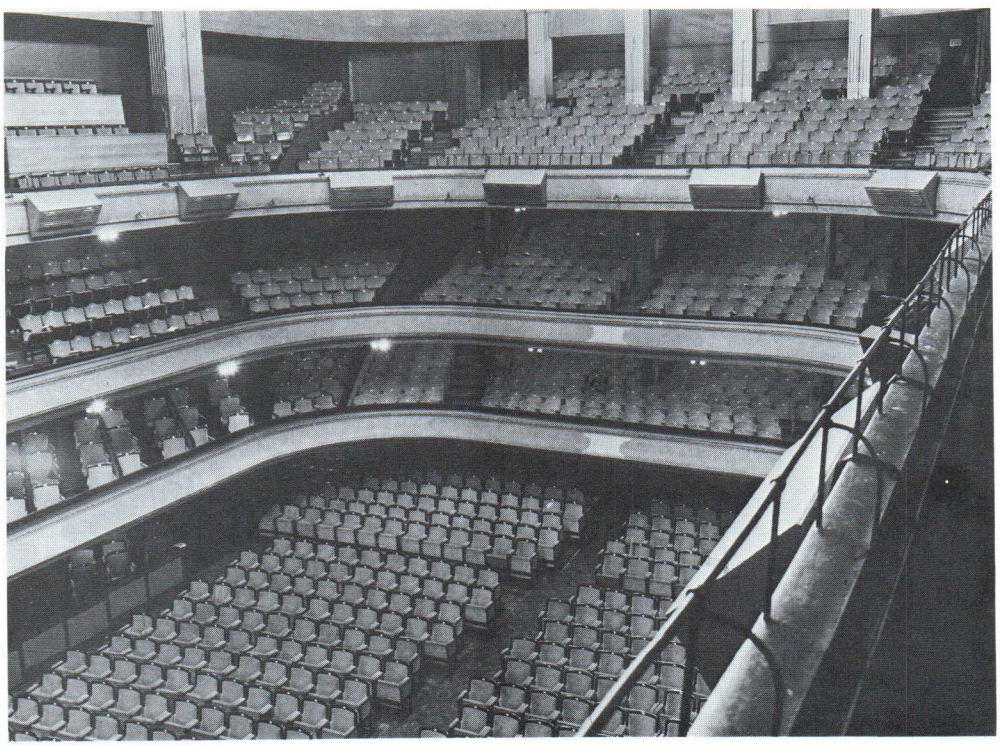


Fig. 8. In the "Gebouw voor Kunsten en Wetenschappen" the loudspeakers are mounted along the edge of the upper balcony and (not visible in the photograph) under the balconies.

with an analogous problem. The condition that nobody in the hall may consciously hear music coming from the loudspeakers, means that at all points in the hall the "sound image" must appear to be located in the orchestra. This latter can be

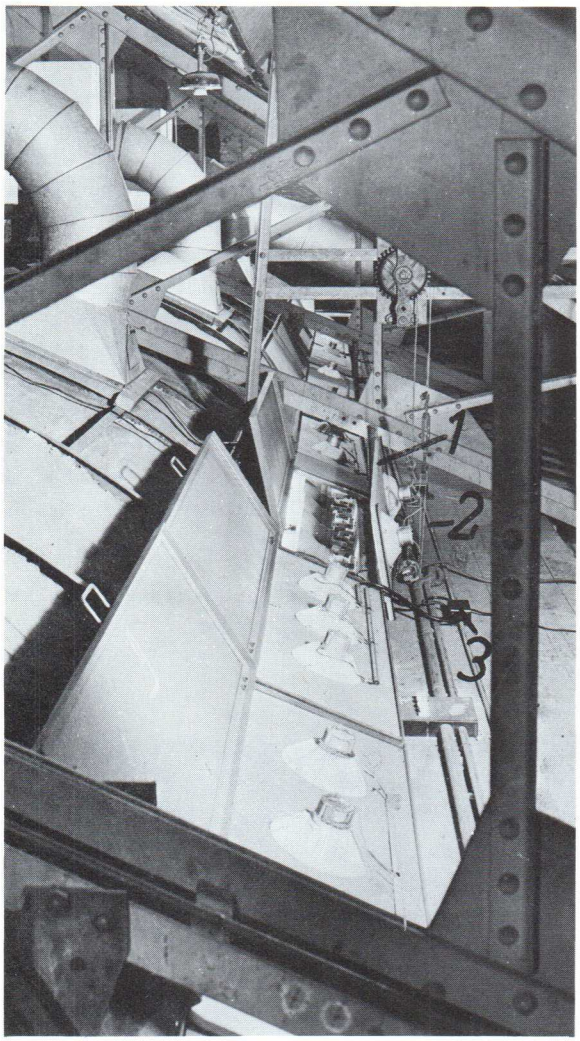


Fig. 9. The loudspeakers of the stereo reverberation installation in the Philips Theatre are mounted in groups of two or three in the panels which cover the lighting cornice (*L* in fig. 6). At *1* one of the panels has been lifted up. *2* is the (six-core) cable, *3* a junction box.

regarded as one of the two sound sources in a stereophonic installation, one of the loudspeakers being the other. For the sake of simplicity we assume that the observer is in the plane of symmetry of the two sources (if he is otherwise situated, the values to be mentioned presently must be modified by suitable amounts). We then know that he will locate the virtual sound source (the "sound image") in the orchestra if (a) the sound from the orchestra and that from the loudspeaker arrive at the same time but the intensity of the orchestra is at least 10 dB above that of the loudspeaker, or (b) orchestra and loudspeaker are equally loud but the sound from the orchestra arrives 2 msec earlier. These are the two extremes; for intermediate cases, as far as the position of the sound image is concerned, a lag of 1 msec in one sound is compensated by an increased intensity of 5 dB, according to the almost linear relationship [9]) shown in *fig. 10*. The sound image is still located in the orchestra even when the intensity of the orchestra is 5 dB less than that from the loudspeaker, provided the sound from the orchestra arrives 3 msec earlier.

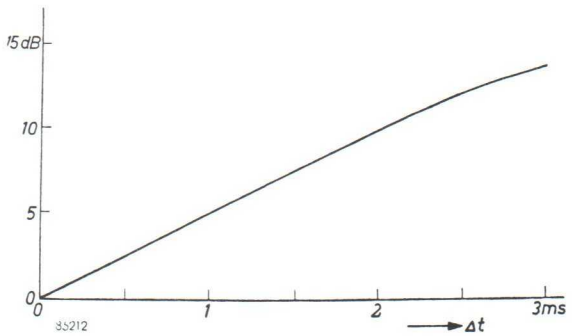


Fig. 10. Differences in intensity (in dB) plotted against the phase differences Δt which produce the same angular displacement of the sound image.

Another condition is that for no listener may the sound from the loudspeakers arrive with so great a time lag that it is no longer experienced as reverberation of the orchestra, but as a separate echo. This means that for no observer may the first loudspeaker signal which reaches him arrive more than 50 msec after the direct sound from the orchestra. This value corresponds to that found earlier when investigating the maximum time interval during which, in speech, the indirect sound contributes to the intelligibility (see Introduction). Recent investigations [10]) have confirmed this value and also that the indirect sound may be stronger than the direct without disturbing the location of the sound image.

On the basis of these data we can set down the following rules for a stereo reverberation installation:

---

[9]) K. de Boer, Stereophonic sound reproduction, Philips tech. Rev. **5**, 107-114, 1940.

[10]) H. Wallach, E. B. Newman and M. R. Rosenzweig, The precedence effect in sound localization, Amer. J. Psych. **62**, 315-336, 1949; G. Meyer and G. R. Schodder, Über den Einfluss von Schallrückwürfen auf Richtungslokalisation und Lautstärke bei Sprache, Nachr. Akad. Wiss. Göttingen, IIa, 31-42, 1952.

1) No member of the audience may receive sound from any loudspeaker before the direct sound has reached him.
2) Nowhere may the first loudspeaker sound arrive more than 40 msec after the direct sound (a safety margin of 10 msec has been deducted from the limiting value of 50 msec).
3) The intensity of the loudspeakers may nowhere be more than 5 dB above that of the direct sound [11]) (apart from the diffuseness required, this is another reason why many dispersed loudspeakers must be used).

If the same hall is also to be used for plays or lectures and improvement of the intelligibility is desirable, the stereo reverberation installation can also make itself useful for this purpose. One must then ensure that no sound is repeated later than 40 msec after the direct sound, and no feedback should be applied.

Directions for operation of the installation

Complying with the rules listed above does not necessarily ensure that the installation works satisfactorily. Because the theory lacks sufficient experimental backing, one should beware of clinging to unfounded preconceived opinions.

An example will illustrate this. In one of our first experiments with stereo reverberation, it was thought that the artificial reverberation should be built up of as many repetitions as possible, in order to obtain the smoothest possible exponentially decreasing intensity. It appeared, in fact, that though this did give the impression of a long reverberation time, this was by no means accompanied by the feeling of being in a large hall — rather that of a small "hard" room such as a bathroom. To suggest a large space, it was necessary to increase the time interval between the echos and to make the reverberation not at all smooth. By careful adjustment in a laboratory room of about 1000 m^3 volume, we could create the acoustic impression of being in a cathedral.

One general piece of advice relating to electro-acoustical intervention in musical performances is that the engineer must show considerable retraint. He must suppress his desire to make the effect *striking* and take care that he does not exaggerate. His is a thankless task: whenever his work is recognized for what it is, he will be reproached, and the better he does his work, the more natural the result will appear and the less thanks he can expect. The highest praise he can receive is probably the simple verdict: "...the orchestra sounded much better".

As we have said, not only the Philips Theatre at Eindhoven, but also the Gebouw voor Kunsten en Wetenschappen in The Hague, is fitted with a stereo reverberation installation. On November 30th, 1954, this installation (then still provisional) had its public debut during a concert given by the Residentieorkest. After the concert, the effects which can be achieved were demonstrated more emphatically.

The verdict on the operation during the concert varied from "favourable" to "very favourable". Some people, however, found it difficult to observe the effect consciously. This is illustrated by a remark from a musician in the audience: "It was remarkable that one could not consciously hear that the installation was in operation; one only felt, or experienced it. Only during the demonstration after the concert did I become consciously aware of it, and convinced that my observations during the concert were not imaginary".

Though the improvement may be appreciated only unconsciously by some of the audience, it is another matter for the performing musicians, both members of the orchestra and soloists; they experience the stereo reverberation very clearly and consciously as making the hall more "playable". This undoubtedly contributes to the attainment of a higher artistic level.

We feel that it is an important milestone in the development of electro-acoustics that leading musicians not only permit microphones and loudspeakers and all that goes with them, in the concert hall, but actually welcome their help.

Other applications of stereo reverberation

Stereo reverberation will undoubtedly find other applications apart from the conversion of theatres, acoustically speaking, into concert halls. However, the apparatus is so complicated that these applications will be limited to the professional field for the time being.

One obvious application is in broadcasting studios. Here, stereo reverberation can be a means of adjusting the acoustics of the studio to the nature of the music or the play to be performed. This can, of course, be done afterwards by adding artificial reverberation to the microphone signal; the echo chamber is a device often used for this purpose. But this deprives the musicians of the stimulus of good acoustics and if there is an audience in the studio, they too miss the full effect. Stereo

[11]) See fig. 8 in the article by Meyer and Schodder, referred to in [10]).

reverberation can overcome these difficulties [12]).

Another application that springs to mind is in the cinema. Here stereo reverberation can be obtained in the manner described above, with the difference that the microphone is discarded and the recording head on the delay wheel is now fed with the signal derived from the sound track of the film. It is even simpler if the cinema is fitted for a system like "CinemaScope", i.e. fitted with loudspeakers around the hall and projectors suitable for films with more than one sound track; the direct and the delayed diffuse sound could then be recorded beforehand on these tracks, so that the cinema has no need to be equipped with the delay wheel.

Finally we may mention the "duplication of concerts" [13]), that is the stereophonic reproduction in one or more "overspill" halls of a concert given elsewhere. Clearly, diffuse reverberation in the "overspill" halls can considerably increase the musical value of the programme presented.

Summary. A shortage of good concert halls means that music is often presented in a hall which is less suitable for this purpose, for example, in a theatre. Such halls can be given better acoustics for music by providing artificial diffuse reverberation. An installation is described with which such "stereo reverberation" can be provided electro-acoustically. It is based on a so-called "delay-wheel", the rim of which is coated with a material suitable for magnetic sound recording. A recording head records the music performed. A number of play-back heads (say, four or six) around the circumference of the rotating wheel pick up the recorded sound with predetermined time lags, and separately feed a group of loudspeakers, mounted throughout the hall. The intensity of each group is independently controllable, and the time lags can be regulated by the speed of the wheel. The sound is picked up by one or two "line microphones", each consisting of ten condenser microphones with strongly directional characteristics, to reduce the possibility of continuous "howl-back".

It has been shown that such an installation can produce a great improvement in the musical acoustics of a hall. Some of the audience experience this only unconsciously, but the performing musicians are strongly aware of it as making the hall more "playable".

Finally, some other possible applications of stereo reverberation are discussed: in broadcasting studios, in the cinema and in the "duplication" of concerts.

[12]) According to a private communication from Dr. J. J. Geluk, Head of the Dutch Broadcasting Union Laboratories, the Union is planning to build two large broadcasting studios (7500 m^3 each) in which the principles of stereo reverberation will be applied.

[13]) R. Vermeulen, Philips tech. Rev. **10**, 169-177, 1948/49 and R. Vermeulen and W. K. Westmijze, Philips tech. Rev. **11**, 281-290, 1949/50.



Research in stereophony

When experiments on stereophony were undertaken in the Eindhoven laboratories in the late thirties, it became clear that the addition of the possibility of directional hearing to a mere increase of loudness would be a useful asset for a person with impaired hearing. A pair of amplifiers fed from two microphones was built for a certain (well-to-do) patient who could then hear the two separate outputs with a pair of headphones. The microphones were to be located in an "artificial head"; the influence of the details of the human head on the variations in the sound pressure received at the inner ear was investigated for this purpose with the aid of a shop-window dummy, in whose ears the microphones were mounted.

(See: K. de Boer and R. Vermeulen, On improving defective hearing, Philips tech. Rev. **4**, 316-319, 1939.)

THE GYRATOR, AN ELECTRIC NETWORK ELEMENT

by B. D. H. TELLEGEN.

621.372.2

There is a traditional antithesis between the practical mind and the mind inclined to pure scientific pursuits: the former is mainly interested in knowledge that can be put immediate to use, while the latter ponders fundamentals and tries to introduce greater clarity and generality by rigorous considerations which the practical man probably finds superfluous. The invention of the gyrator by Prof. Tellegen is a striking illustration of the service that can be rendered to engineering by the pure scientific approach: originally regarded as a hypothetical possibility, a possibility that had to be recognized for the sake of completeness, the gyrator has subsequently and perhaps rather surprisingly become a reality in the world of microwaves.

Wide use is made in electrical engineering of networks composed of resistors, coils and capacitors, the latter being known as network elements. In use, there is conversion and exchange of energy within and between these network elements. In resistors electrical energy can be converted into heat; in coils and capacitors energy can be stored and later released. The networks are provided with terminal pairs allowing an exchange of energy to take place with the exterior. A terminal pair consists of two conductors between which a voltage of instantaneous value v can exist and through which a current of instantaneous value i can flow (see *fig. 1*). The energy

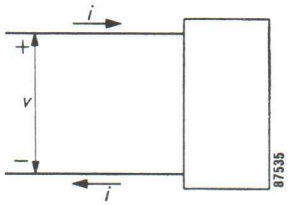


Fig. 1. Network with one terminal pair.

supplied to the network via the terminal pair in a time dt is $ivdt$; it can be positive or negative. The network determines relations between the voltages and currents at the terminals, the number of relations being equal to the number of terminal pairs. If for example the network consists of one resistor R, it has one terminal pair and the relation is $v = Ri$.

Networks made up of resistors, coils and capacitors

For the user networks are characterized primarily by the relations that hold between the terminal voltages and currents. As long as the relations between these voltages and currents are the ones desired, the actual composition of the network is, for the user, of secondary concern. It is therefore of great importance to know what sets of such relations are possible with networks consisting of resistors, coils and capacitors. All the possible sets of relations form an arsenal from which the user can take his choice.

In order to build up the whole arsenal, i.e. a complete list of the possible sets of relations, let us first look for general properties of the relations; we can then try to demonstrate that, for each set of relations having these properties, it is possible to design a network of resistors, coils and capacitors for which the said relations are valid. This can be put in another way by saying that we shall try to find necessary and sufficient conditions for the sets of relations if they are to be realizable by networks composed of resistors, coils and capacitors.

General properties of the relations are that they consist of *linear* differential equations relating the terminal voltages and currents, and that the coefficients of the equations are *constant*, i.e. not dependent on time but determined only by the magnitude of the elements composing the networks. Furthermore, the networks contain no source of energy; they are said to be *passive*. Certain properties of the coefficients of the differential equations can be deduced from the fact that the networks are passive. Let us consider, for example, a network with one terminal pair, the properties of which are governed by a differential equation of the first order, so that we can write:

$$a\frac{di}{dt} + bi = c\frac{dv}{dt} + dv, \quad . \; . \; . \quad (1)$$

where a, b, c and d are constants. From the passivity of the network it can be shown that a, b, c and d all have the same sign.

This can be demonstrated as follows. If we short-circuit the network, in other words if we keep $v = 0$, the current will be determined by:

$$a\frac{di}{dt} + bi = 0.$$

The solution of this differential equation is:

$$i = C\, e^{-\frac{b}{a}t}$$

where C is the constant of integration. In consequence of the passivity, the current in the short-circuited network cannot rise indefinitely and hence a and b have the same sign. It follows in a similar way, by considering the network to be open-circuited, i.e. by keeping $i = 0$, that c and d have the same sign.

For direct current and voltage equation (1) simplifies to

$$bi = dv,$$

so that b/d represents the DC resistance of the network. This must be positive on account of the network's passivity, and hence b and d also have the same sign.

For convenience we shall group the three properties together and say that the sets of relations characterizing networks made up of resistors, coils and capacitors are linear, constant, passive. As far as networks with one terminal pair are concerned, these properties are not only necessary but also sufficient, for Brune has demonstrated [1]) that *any* linear constant passive relation can be realized by a network made up of resistors, coils and capacitors. In particular, it can be demonstrated that any relation having the form of equation (1) can be realized by a network consisting either of two resistors and a coil (if $a/c > b/d$), or of two resistors and a capacitor (if $a/c < b/d$). The two networks are shown in *fig. 2*.

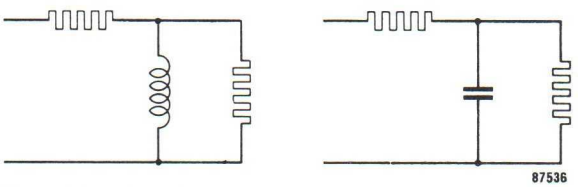


Fig. 2. Networks with one terminal pair and of the first order.

For any network consisting of resistors, coils and capacitors and having more than one terminal pair a set of relations holds that, besides being linear constant passive, has the property known as reciprocity. To explain this property we shall take a network with two terminal pairs and express the two terminal voltages in terms of the two terminal currents; for the present purpose it will be convenient to write the relationships in complex form, rather than use the instantaneous values. Accordingly, we obtain:

$$\left.\begin{aligned} V_1 &= Z_{11}I_1 + Z_{12}I_2, \\ V_2 &= Z_{21}I_1 + Z_{22}I_2, \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where I_1, I_2, V_1 and V_2 represent the complex values of the terminal currents and voltages. If we assume the sign convention for currents and voltages as indicated in *fig. 3*, it can be shown that always

$$Z_{21} = Z_{12}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

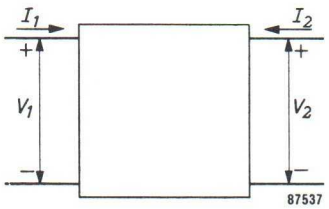


Fig. 3. Network with two terminal pairs.

Expressing the current I_1 through the first terminal pair and the voltage V_2 across the second terminal pair in terms of the two other terminal quantities, we find from equations (2):

$$\left.\begin{aligned} I_1 &= \frac{1}{Z_{11}}V_1 - \frac{Z_{12}}{Z_{11}}I_2, \\ V_2 &= \frac{Z_{21}}{Z_{11}}V_1 + \frac{Z_{11}Z_{22} - Z_{12}Z_{21}}{Z_{11}}I_2. \end{aligned}\right\} \quad . \quad . \quad (4)$$

The coefficient of V_1 in the second equation above is now equal and opposite in sign to the coefficient of I_2 in the first equation. The relation (3) between two coefficients of equations (2) and the relation just discussed between two coefficients of equations (4) are known as the reciprocity relations.

For networks with more than one terminal pair Bayard and others have demonstrated [2]) that any linear constant passive set of relations having the reciprocity property can be realized by a network consisting of resistors, coils and capacitors. Hence the said properties of sets of relations are sufficient as well as necessary.

Our arsenal of all possible sets of relations capable of realization by networks made up of resistors, coils and capacitors is thus complete.

Linear constant passive systems

The result obtained above is not entirely satisfactory, however. We have based our considerations on the resistors, coils and capacitors that had their origin in the laboratory. Electrical engineering has simply accepted these elements for constructing networks. There is something arbitrary and fortuitous about all this. Must we necessarily use these

[1]) O. Brune, J. Math. Phys. **10**, 191, 1931.

[2]) M. Bayard, Bull. Soc. franç. Elect. **9**, 497, 1949. Also B. D. H. Tellegen, J. Math. Phys. **32**, 1, 1953, which gives further references to the literature.

elements for the building-up of networks? Should an affirmative answer be found to this question, we would be faced with another one: are other elements conceivable, elements that have not been found in the laboratory?

In order to investigate these questions we must not proceed from the basis of the conventional network elements and the networks composed of them but approach the problem from another direction. We must consider "black boxes", systems with terminal pairs that are characterized only by the sets of relations existing between the terminal voltages and currents without concerning ourselves with what is inside them. In what follows we shall confine ourselves to linear constant passive systems, by which we mean systems characterized by linear constant passive sets of relations. We could of course subject ourselves to less stringent limitations: we could also consider systems containing energy sources (active systems, such as amplifiers), systems characterized by linear equations with variable coefficients (variable systems), or systems characterized by non-linear equations (non-linear systems). However, such systems have properties that depart appreciably from those of the networks considered above, and are of a more complex nature. We shall therefore leave them aside.

Let us now try to find the "simplest" kinds of linear constant passive system. These simplest systems we shall call network elements, *by definition*. In this way we shall try to arive at a set of network elements such that any linear constant passive system can be realized as a network composed of them. We can regard such a set of network elements as a complete set.

In order to give a meaning to the epithet "simplest", we shall have to classify our "black boxes". We shall do so according to the number of terminal pairs, the order of the differential equations characterizing them, and whether or not they are able to dissipate electrical energy, i.e. to transform it into heat. This leads, as a first step, to the investigation of systems with one terminal pair, of zero order and involving no dissipation. For such systems the power supplied must be zero at any instant, i.e. $iv = 0$; therefore either $i = 0$ (open terminal pair) or $v = 0$ (short-circuited terminal pair). This does not produce a network element. In accordance with the threefold classification, our next step will be to investigate three classes of systems, namely (1) systems with one terminal pair, of zero order and involving dissipation, (2) systems with one terminal pair, of the first order and involving no dissipation, and (3) systems with two terminal pairs, of zero order and involving no dissipation.

(1) Systems with one terminal pair, of zero order and *involving dissipation*

These systems are characterized by an equation having the form:

$$v = Ri, \quad R > 0 . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

The fact that R is positive is a consequence of the passivity of the network, this requiring that iv should be positive.

(2) Systems with one terminal pair, of the *first order* and involving no dissipation

Systems with one terminal pair and of the first order are characterized by an equation of the form of (1). It can be shown that this represents a system involving no dissipation either if $b = 0$ and $c = 0$ or if $a = 0$ and $d = 0$. Hence these systems are of two kinds. The first kind is characterized by an equation of the form:

$$v = L\frac{\mathrm{d}i}{\mathrm{d}t}, \quad L > 0, . \quad . \quad . \quad . \quad . \quad (6)$$

and the second kind by an equation of the form:

$$i = C\frac{\mathrm{d}v}{\mathrm{d}t}, \quad C > 0 . \quad . \quad . \quad . \quad . \quad (7)$$

The above may be derived as follows. The system to which (1) is applicable has an impedance of

$$Z = \frac{j\omega a + b}{j\omega c + d}.$$

In order that there should be no dissipation, the real part of Z must be zero at all frequencies. It follows that $ac = 0$ and $db = 0$. Since $a = 0$ and $b = 0$ means that $Z = 0$, and $c = 0$ and $d = 0$ means that $Z = \infty$, these solutions may be disregarded. Thus either c and b, or a and d, must be zero, and these two cases produce equations (6) and (7), respectively. Since the passivity of the network requires that a, b, c and d should all have the same sign (see above), it further follows that $L = a/d$ and $C = c/b$ are both positive.

Before examining the third class of simplest systems we shall take a closer look at the three systems, (5), (6), and (7), already found. As stated above, we shall regard them as network elements. The latter, then, are defined by equations, not by the physical means required to realize them. For example, we cannot infer from the equation $v = L\,\mathrm{d}i/\mathrm{d}t$ that it describes a "coil". A superconductor has the same equation when we take the effect of the mass of the conducting electrons into account. The kinetic energy of these electrons then takes the place of the magnetic energy of the coil.

The three network elements, so defined, are "ideal"; they can only be realized approximately. For example, (5) is approached by a resistor with low stray capacitance and low self-inductance (6) by a low-loss coil with a low stray capacitance, and (7) by a low-loss capacitor with a low self-inductance (hence the symbols R, L and C used in the equations).

The way in which the network elements are realized is only of secondary importance to the user; it is the external properties given by (5), (6) and (7), that are of primary interest. This remark is similar to that made earlier regarding networks: the user has little interest in their internal composition.

(3) Systems with *two terminal pairs*, of zero order and involving no dissipation

For these systems the total power supplied via the terminal pairs is zero at any instant, so that

$$i_1 v_1 + i_2 v_2 = 0,$$

where i_1, i_2, v_1 and v_2 represent instantaneous values of the terminal currents and voltages, the sign convention for these quantities being assumed to be in accordance with fig. 3. This condition results in two kinds of systems, namely those characterized by equations of the form:

$$\left.\begin{aligned} i_1 &= -n i_2, \\ v_2 &= n v_1, \end{aligned}\right\} \quad \ldots\ldots \quad (8)$$

and those characterized by equations of the form:

$$\left.\begin{aligned} v_1 &= -s i_2, \\ v_2 &= s i_1, \end{aligned}\right\} \quad \ldots\ldots \quad (9)$$

Equations (8) and (9) may be derived as follows. If, for a two-terminal-pair system of zero order, we express the two terminal voltages in terms of the two terminal currents, we can write the equations as:

$$\left.\begin{aligned} v_1 &= a_{11} i_1 + a_{12} i_2, \\ v_2 &= a_{21} i_1 + a_{22} i_2. \end{aligned}\right\} \quad \ldots\ldots \quad (10)$$

It follows from this that:

$$i_1 v_1 + i_2 v_2 = a_{11} i_1^2 + (a_{12} + a_{21})\, i_1 i_2 + a_{22} i_2^2. \quad \ldots \quad (11)$$

In order that (10) should represent a system involving no dissipation, (11) must be zero for all values of i_1 and i_2; hence $a_{11} = 0$, $a_{12} + a_{21} = 0$, and $a_{22} = 0$, which results in (9).

Starting from the equations that express i_1 and i_2 in terms of v_1 and v_2 we also arrive at (9). However, it is also possible to conceive systems in which v_1 and v_2 cannot be expressed in terms of i_1 and i_2, or vice versa. This is the case for systems that are characterized by a relation between i_1 and i_2 and a relation between v_1 and v_2, making it impossible to chosse i_1 and i_2 or v_1 and v_2 as independent variables. In order to investigate these systems we can start from equations expressing i_1 and v_2 in terms of v_1 and i_2, or vice versa. We then arrive at (8).

We shall regard the systems thus found, (8) and (9), as two further network elements. System (8) is called an ideal *transformer* and we have proposed [3]) the name ideal *gyrator* for system (9). Observations similar to those made above on (5), (6) and (7) may also be made on these two systems.

An approximation to the ideal transformer is given by two tightly coupled low-loss coils of large self-inductances. This is in fact the way in which the concept first arose. The theoretical route by which we have now arrived at the same concept shows why we should regard the ideal transformer as a separate network element, defined, in fact, by (8). From the remarks on reciprocity in connection with equations (4) we see that the ideal transformer also possesses this property. The ideal gyrator does *not* possess the property of reciprocity, as will be clear from a comparison of (9) with (2) and (3). It is therefore impossible to obtain an approximation to it by combining resistors, coils and capacitors, because such a combination always has the property of reciprocity, as we saw above. For the realization of the gyrator other physical means are necessary. One way of realizing it, making use of gyromagnetic effects in ferromagnetic materials, will be described in a further article in this Review [4]).

The addition of the gyrator to the set of network elements has extended the list of possible sets of relations between terminal voltages and currents. This extension was made possible by the fact that we started from linear constant passive systems as such, without imposing on them the condition of reciprocity. The fact that system (9), while being linear constant passive lacks the property of reciprocity, shows that the latter is not a consequence of linearity, constancy and passivity, and that imposing it does indeed constitute a limitation of the possibilities.

Some properties of the gyrator

The ideal gyrator has the property of "gyrating" a current into a voltage, and vice versa. The coefficient s, which has the dimension of a resistance, we call the gyration resistance; $1/s$ we call the gyration conductance. We shall represent the gyrator in circuit diagrams by the symbol shown in *fig. 4*.

The following properties of the ideal gyrator can be easily derived from (9).

[3]) B. D. H. Tellegen, Philips Res. Rep. **3**, 81, 1948.
[4]) H. G. Beljers, Application of ferroxcube to uni-directional waveguides, to appear shortly in this Review.

If we have the output terminals open-circuited, i.e. $i_2 = 0$, the input terminals are short-circuited,

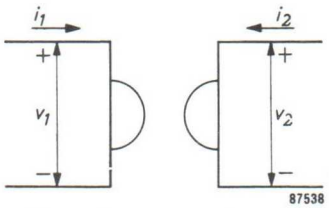


Fig. 4. Symbol for the gyrator.

i.e. $v_1 = 0$, and vice versa. If we connect an inductance L or a capacitance C across the output terminals, we find a capacitance L/s^2 or, in the second case, an inductance s^2C between the input terminals. In general, if we connect an impedance Z across the output terminals, we find an impedance s^2/Z between the input terminals. An impedance Z in series or in parallel with the output terminals has the same effect as an impedance s^2/Z in parallel or in series, respectively, with the input terminals (*fig. 5*).

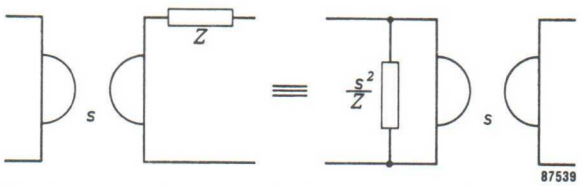


Fig. 5. An impedance in series with one terminal pair of an ideal gyrator is equivalent to another impedance in parallel with the other terminal pair.

Two ideal gyrators in cascade form an ideal transformer; an ideal gyrator and an ideal transformer in cascade form another ideal gyrator.

The combination of a gyrator with a resistor shown in *fig. 6a* gives a system to which the following apply:

$$\left.\begin{aligned} v_1 &= Ri_1 + (R-s)i_2, \\ v_2 &= (R+s)i_1 + Ri_2. \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (12)$$

Thus an input i_1 gives rise to a voltage component $(R+s)i_1$ across the output; an output current i_2 gives rise to a voltage component $(R-s)i_2$ across the input. If $R = s$, the latter component is zero, so that the input voltage is independent of the output current and dependent only on the input current. In such a case we may say that the system transmits only in the direction from input to output, and not in the reverse direction. The system shown in fig. *6b* has properties of a similar kind; this can be easily demonstrated by writing down the equations expressing the terminal currents in terms of the terminal voltages.

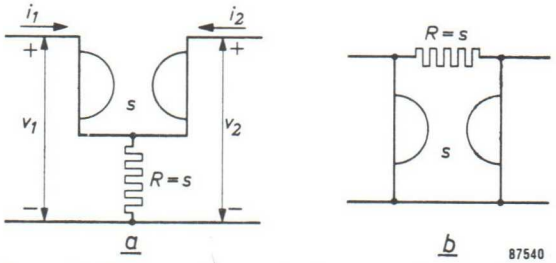


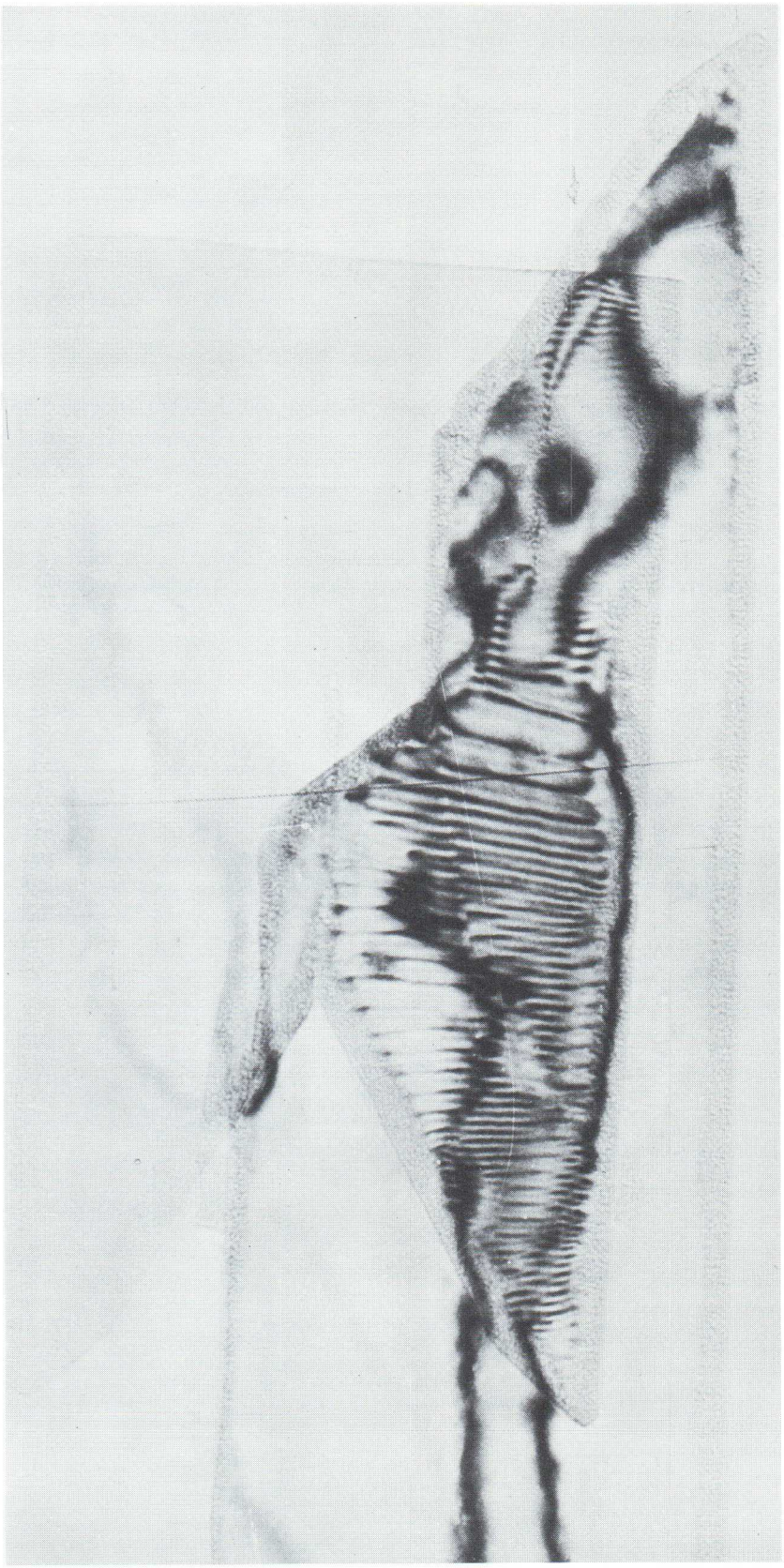
Fig. 6. Uni-directional networks, i.e. networks that transmit only in one direction.

No further elements remain to be added to the five ideal linear constant passive network elements: Oono and Yasuura have demonstrated [5]) that any linear constant passive system can be realized as a network composed of these five network elements. The five ideal network elements therefore form a complete set.

[5]) Y. Oono and K. Yasuura, Mem. Fac. Engng., Kyushu Univ. **14**, 124, 1954; also in Ann. Télécomm. **9**, pp. 73 and 109, 1954.

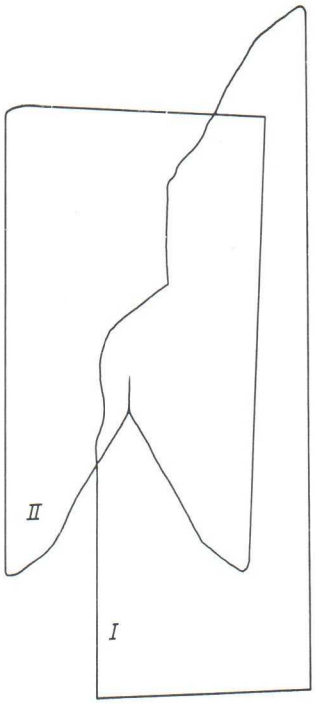
Summary. The writer considers the electric networks having one or more terminal pairs that can be built up from conventional network elements, namely resistors, coils and capacitors. The set of relations between terminal voltages and currents determined by such a network has properties that correspond to the linear, constant and passive nature of such networks and, furthermore, has the property of reciprocity. It has been demonstrated that, conversely, these properties are sufficient for any set of relation possessing them to be realizable by a network. The writer reverses this line of reasoning and raises the question: what (ideal) network elements must be introduced in order to make it possible to realize all linear constant passive systems. There are no grounds for including reciprocity amongst the properties imposed. It is then shown that, apart from the conventional elements — resistor, coil, capacitor, and the ideal transformer — a new network element has to be introduced; this new element has been christened ideal gyrator. Networks containing the new element generally lack the property of reciprocity. A brief sketch is given of some of the properties of the gyrator.

Bragg reflections in electron photomicrographs



The photograph, taken with a Philips EM 100 B electron microscope, is of *synthetic* γ FeOOH, a substance which also occurs naturally as the mineral lepidocrocite. The overall magnification is about 200 000 times. The specimen consisted of small, extremely thin crystal wafers (thickness approx. 10 nm). As a rule the wafers are slightly bent, so that the planes of atoms in the wafer differ a little in orientation from place to place. Certain regions in the wafers may then be oriented so that part of the incident electron beam undergoes Bragg reflection. Electrons reflected at wide Bragg angles are intercepted by a diaphragm, and do not reach the photographic plate, so that these places show up in the picture as dark patches. The photograph represents two partly overlapping crystal wafers *I* and *II* (see the sketch). The two black strips in the non-overlapped bottom part of wafer *I* are due to Bragg reflections of this nature. In the part where the wafers overlap the play of Bragg reflections in the two crystals produces the remarkable skeleton-like pattern shown here.

(From: Philips tech. Rev. **26**, 48, 1965.)



THE PROJECTION OF COLOUR-TELEVISION PICTURES

by T. POORTER and F. W. de VRIJER. 621.397.2:621.397.62:535.881

A projection system for monochrome television was developed in the Philips Eindhoven Laboratory over 10 years ago. Further work has now resulted in the development of apparatus for the projection of coloured television pictures on to a screen. Projectors have been built giving pictures 22 cm × 29 cm and 35 cm × 46 cm at a highlight luminance of 200 cd/m², and 2.25 m × 3 m at a highlight luminance of 20 cd/m². The latter value is about the same as that on normal cinema screens. The projectors were demonstrated before the colour-television commission of the C.C.I.R. in 1955 and 1956.

Almost all colours occurring in nature can be simulated by adding together three kinds of light, each of a suitable primary colour: red, green and blue. Colorimetric theory, as it concerns the addition of colour, was dealt with in an earlier article [1]) which will be referred to as I. It is possible to obtain a television picture in natural colour in the manner stated by building up each element of the picture out of appropriate quantities of light of the three primary colours. A practical method of doing so is to project a red, a green and a blue image simultaneously on to a common screen, in such a way that the three images coincide exactly. The luminance of each element of the resultant image is thus the sum of the luminances of the three corresponding elements of the primary-colour images, the colour of the resultant element being determined by the proportions of these luminances.

One obvious way of building such an apparatus would be to combine a projection type cathode-ray tube giving coloured images — working on the same principle as direct-viewing tubes for colour television — with a suitable optical system, but no projection tube for coloured images has so far been produced. However, projection tubes luminescing in each of the three primary colours do exist. A set of three projectors, one for each of the three primary colours is therefore taken as the basis of the system: the three projectors are equipped with projection tubes giving red, green and blue light.

It is of course necessary that corresponding elements of the primary-colour images should coincide on the projection screen. The correct superposition of the three images is therefore a matter requiring particular attention. One of the conditions for proper superposition is that the three projectors should be identical in design and that any differences arising in their construction should be kept as small as possible.

The primary-colour projectors

The primary-colour projectors can be designed on exactly the same lines as those for monochrome television. A projector of this kind, embodying the Schmidt optical system, was described in this journal [2]) and elsewhere [3]) in 1948. The Schmidt system combines three valuable properties, namely low optical distortion, low transmission losses and large aperture. In its simplest form the system comprises the arrangement shown diagrammatically in *fig. 1a*. However, as stated in the article referred to in footnote [2]), the alternative arrangement shown in fig. 1*b* has been used in home projection receivers. Here the light beam is "folded" by means of a plane mirror M_2, which is placed at

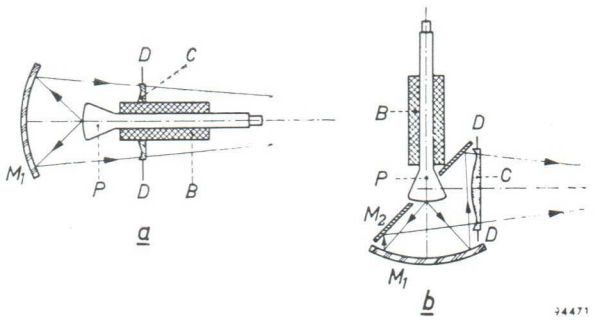


Fig. 1. Schmidt optical systems for the projection of television images; *a*) "in-line" Schmidt system, *b*) "folded" Schmidt system. *P* projection tube with focussing coil and scanning coils *B*. M_1 spherical concave mirror. *C* corrector plate mounted in diaphragm *D*. The centre of curvature of M_1 lies in the plane of *D* in the case of (*a*) and in the image of *D* given by plane mirror M_2 in the case of (*b*).

[1]) F. W. de Vrijer, Fundamentals of colour television, Philips tech. Rev. **19**, 86-97, 1957/58 (No. 3).

[2]) P. M. van Alphen and H. Rinia, Projection-television receiver, I. The optical system for the projection, Philips tech. Rev. **10**, 69-78, 1948/49.

[3]) H. Rinia, J. de Gier and P. M. van Alphen, Home projection television, I. Cathode-ray tube and optical system, Proc. Inst. Rad. Engrs. **36**, 395-400, 1948.

an angle of 45° to the principal axis. This arrangement allows the set to be constructed more compactly and to be enclosed in a dust-proof box. When space is availabe, however, there is no reason why the "in-line" optical system shown in fig. 1*a* should not be used [4]).

Superposition of the three primary-colour images

There are several possible ways of arranging for the primary-colour images to coincide on the projection screen. A general distinction may be made between arrangements whereby the three projectors, from an optical point of view, are effectively coincident (all done by mirrors), and all other arrangements.

Three projectors effectively at the same point

In theory superposition errors are absent if one can somehow place the projectors such that the three images are projected, as it were, from one point. *Fig. 2* shows how this may be done with the aid of two dichroic mirrors as discussed recently in this journal [5]). Mirror *r* reflects red light and passes green and blue; mirror *b* reflects blue and passes red and green. Green light has to go through both mirrors, the proportion transmitted being between 70% and 80%.

The crossed mirrors may be mounted with the line of intersection either horizontal or vertical. If the line of intersection is horizontal the three projectors will have to be in a vertical plane, and

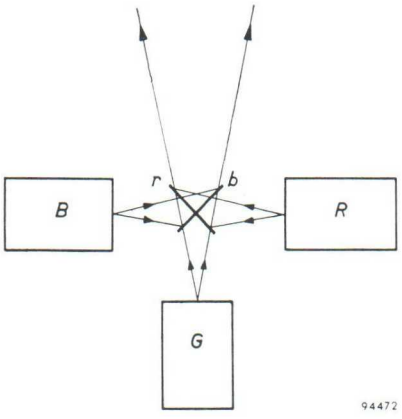


Fig. 2. Superposition of three primary-colour images by means of crossed dichroic mirrors (*r* and *b*). Mirror *r* reflects the red light from *R* and passes the green from projector *G* and the blue from projector *B*. Mirror *b* reflects blue light and passes red and green.

this has the advantage that colour variations over the area of the projected picture are reduced to a minimum. The variations arise because the light beams do not all strike the mirrors at the same average angle; with the line of intersection horizontal,

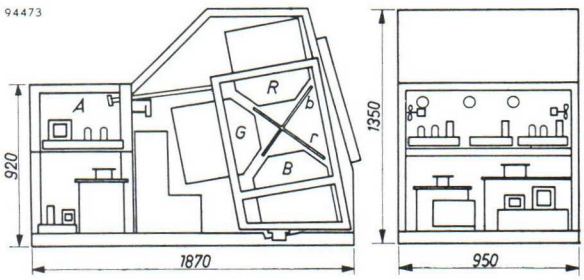


Fig. 3. Sketch of the first large-screen colour-television projector constructed at Eindhoven (see also fig. 4); it gives a picture measuring 2.25 m × 3 m (≈ 7½′ × 10′). *r* and *b* are crossed dichroic mirrors with their line of intersection horizontal. This means that *R*, *G* and *B*, the primary-colour projectors, are mounted in a vertical plane; despite this it was possible to limit the overall height of the apparatus to 1.35 m. Compartment *A* contains the supply unit. The dimensions shown are in mm.

the angle over which variations occur is subtended by the lesser dimension of the picture, namely its height. However, if the apparatus is for projecting large pictures, care must be taken not to build it too tall, otherwise it will interfere with part of the audience's view of the screen. This problem arises because in order to keep down the dimensions of the main optical system, the projectors have to be placed as close to the sreen as possible, i.e. in front of the viewers. It has proved possible to keep the height of the projection apparatus down to 1.35 m as shown in *fig. 3*, which gives the dimensions of the first of the large-screen colour television projectors built in this laboratory (photograph in *fig. 4*).

If the mirrors are mounted with their line of intersection vertical and the primary-colour projectors are accordingly placed in a horizontal plane, it is easier to keep down the height of the apparatus but there will be greater colour variations in the projected picture.

Much depends on whether the crossed mirrors are combined with "in-line" or "folded" Schmidt optical systems. In the first kind of Schmidt system the projection tube, surrounded by the focussing and scanning coils, protrudes through an opening in the corrector plate (fig. 1*a*). In order that the tube and coils should block the path of the reflected light as little as possible, they have to be located outside the crossed mirrors. These latter have therefore to be set up some distance away from the corrector plates (in fig. 1*a* they would be mounted to the right of the foot of the tube), in a place where the re-

[4]) J. Haantjes and C. J. van Loon, A large-screen television projector, Philips tech. Rev. **15**, 27-34, 1953/54.
[5]) P. M. van Alphen, Philips tech. Rev. **19**, 59-67, 1957/58 (No. 2).

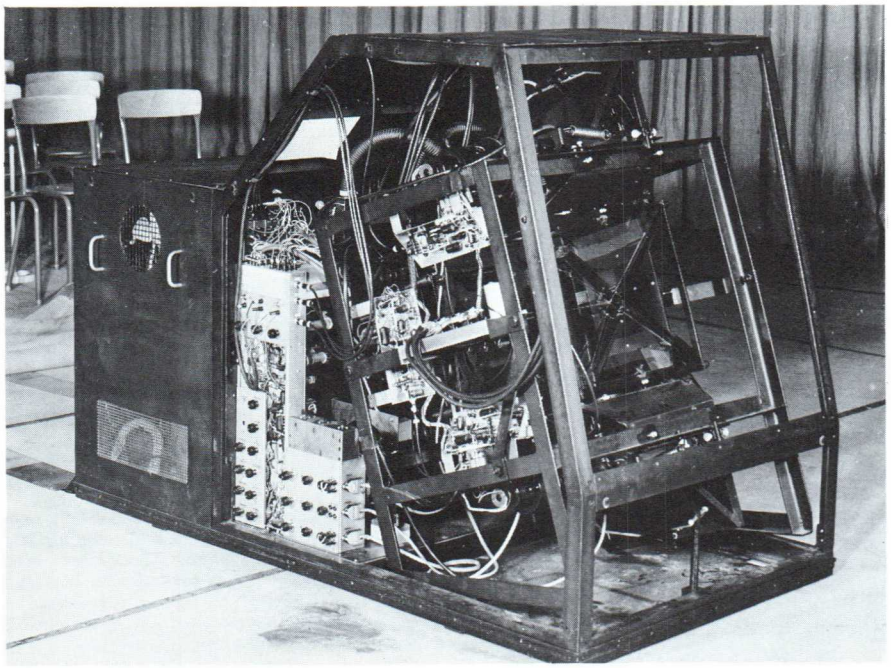


Fig. 4. Photograph of the colour-television projector sketched in fig. 3. The size of the projected picture is 2.25 m × 3 m (≈ 7½′ × 10′) and its maximum luminance is about 10 cd/m². Electrical adjustments are carried out from a control desk (not shown in the photograph).

flected beams have a fairly large diameter. In consequence, the mirrors have to be bigger than they are in "folded" systems (fig. 1*b*), where the crossed mirrors can be set up close to the corrector lenses. The "in-line" optical system also necessitates special arrangements for inserting and withdrawing the CRT; for example, it might be necessary to design the concave mirror in such a way as to allow of dismantling.

Projectors using horizontally intersecting dichroic mirrors have been built, and have given good results. In general, however, designs embodying dichroic mirrors have two disadvantages. The first is a loss of brightness and contrast, due to the large number of reflecting and absorbing surfaces in the path of the beams, especially if these are "folded". The second disadvantage is a loss of sharpness due to the difficulty in achieving a completely stress-free mounting of the mirrors. In consequence, ways have been sought of superimposing the three primary colour images without the aid of dichroic mirrors. This line of enquiry has led to arrangements whereby the three primary-colour images are *not* projected effectively from the same point. This necessarily results in certain superposition errors and, in general, some compensation of these must be sought, using special corrective measures.

Projectors not effectively at the same point

a) The projectors are placed at the apexes of an equilateral triangle. Of all the different arrangements that are possible, this involves the smallest errors of superposition; the triangular arrangement will therefore be attractive where it is not desired to undertake any corrective measures. However, such measures can only be dispensed with if the optical magnification is large — the errors diminish in relation to the picture size as the magnification increases. Correction of the errors, necessary where magnification is not so large, is more complicated in this arrangement than in those now to be discussed.

b) The projectors are arranged side by side. We may distinguish between two cases: (1) where the three projectors are parallel to one another, their axes being perpendicular to the projection screen, and (2) where the outside projectors are inclined inwards slightly.

The following observations apply equally well, *mutatis mutandis*, to arrangements whereby the three projectors are mounted one above the other. Such arrangements are less desirable, however, because they occupy a greater height.

b1) The three projectors are side by side with their axes parallel. Each throws a rectangular image on to the screen and, in theory, the projected images

can be made to coincide by displacing the luminescent images on the two outside tubes electrically. If the luminescent images are to be displaced from the central position on the tube screen it is necessary that they should not occupy the whole of its area. Thus only part of the tube screen is utilized, with the consequence that stronger magnification is necessary and that the highlight luminance on the projection screen will be lower than the maximum obtainable. In addition, vignetting will occur, owing to the fact that the central ray passes obliquely through the optical system. It is a condition for success that the deflection coils should be so designed that the images on the projection screen are completely free from distortion. This ideal cannot be approached closely enough in practice, and so the method can only be used where the magnification is very large and where, in consequence, the outside CRT screen images need only be displaced slightly.

b2) The outside projectors are pointed inwards (*fig. 5a*) in such a manner that their optical axes intersect in the centre of the projection screen. The deflection coils are adjusted such that the vertical centre lines of the separate images coincide (along *MM'* in fig. 5*b*). The result is that, for the images from the outer projectors the magnifications are not equal in the left and right halves of the picture, giving rise to trapezium shaped images. By introducing an electrical correction into both the line and frame scanning of the outside tubes (how this is done will be discussed below) the images from these two projectors can be made to coincide with the rectangular image projected by the centre projector.

Prior to correction, with the three projectors positioned in the best possible manner, the superposition errors arising in the corners of our projected pictures were as follows. In a picture measuring 2.25 m × 3 m the error was six times the inter-line separation; in a picture measuring 22 cm × 29 cm the error was thirty times the inter-line separation. Both cases refer to a raster of a 625 lines.

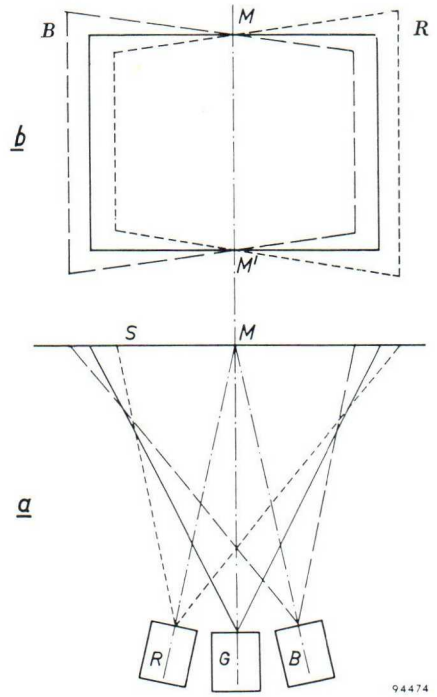


Fig. 5. *a*) Arrangement whereby the three primary-colour projectors *R*, *G* and *B* are placed side by side, with the two outer ones inclined inwards at an angle such that the vertical lines through the centres of the three images (*MM'* in diagram *b*) coincide.
b) In consequence, the red and blue images suffer from trapezium distortion.

Since the axes of projectors *R* and *B* are not now perpendicular to the projection screen (fig. 5*a*) there ceases to be a linear relationship between the horizontal movements of the red and blue spots on the screens of the outside tubes and the movements of their respective projected images. In order to give the spots on the projection screen a constant horizontal velocity, the line scanning current must be given the form indicated in *fig. 6a*. This means adding correction currents to the normal line scanning currents of the outside projection tubes. The two correction currents are, of course, equal and opposite; hence only one current has to be generated and to be passed in the appropriate

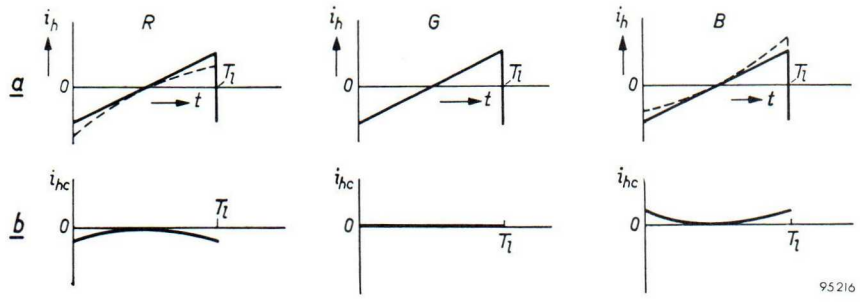


Fig. 6. *a*) Waveform required (dashed curves) for the line scanning current (i_h) in primary-colour tubes *R*, *G* and *B* (mounted as in fig. 5*a*) in order that linearity be restored to the movement of the three spots across the projection screen. T_l is the line period.
b) To obtain the above waveforms, the linear sawtooth currents in *R* and *B* (full lines in *a*) must have correction currents i_{hc} added to them; the two correction currents are equal and opposite.

direction through the scanning coil of each of the outside tubes. Prior to this correction, the three images on the projection screen have the outlines shown in *fig. 7a* after applying the above correction, they take on the shapes shown in fig. 7*b*.

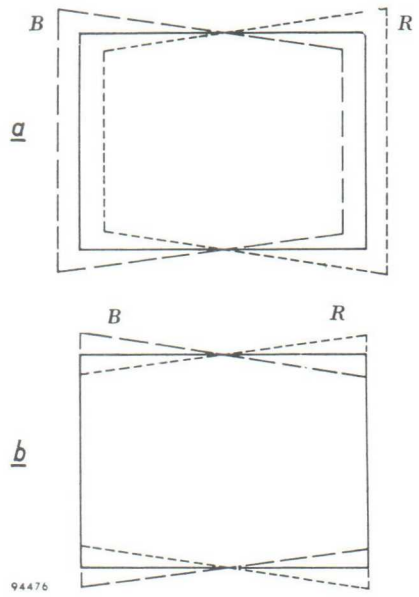


Fig. 7. *a*) Outlines of the three primary-colour images on the projection screen prior to correction (cf. fig. 5*b*).
b) Application of the correction illustrated in fig. 6 brings the vertical edges of the three images into coincidence. To get complete coincidence of the images, correction of the vertical (frame) scanning is further necessary — see fig. 8.

The red and blue images on the projection screen are still of trapezium shape; to make them coincide with the rectangular green image, the frame scanning must also be subjected to a correction. From a glance at the outline of, say, the blue image in fig. 7*b*, it will be clear that the vertical scan has to be shortened on the left and lengthened on the right, in the rhythm of the line frequency. The frame scanning current of the blue CRT must therefore have a correction current of the following description superimposed upon it.

(i) While the beam is at the top of the screen the correction current must go from negative to positive as the beam swings from left to right.
(ii) When the beam is half way down the screen the correction current must be zero.
(iii) While the beam is at the bottom of the screen the correction current must go from positive to negative as the beam swings from left to right.

(A "positive" correction current is here taken to mean a deflection towards the top of the projected image.)

The correction current must therefore have the shape drawn in *fig. 8a*, and it might be obtained by multiplying a sawtooth of frame frequency (fig. 8*b*) by a sawtooth of line frequency (fig. 8*c*); how it is actually done will be described later.

At the same time the red CRT must also be supplied with a correction current of the same magnitude but of opposite direction. For the corrections to the vertical scanning for both red and blue tubes, therefore, it will again suffice to generate a single correction current.

It should be observed for the sake of completeness that the addition of correction currents is not the only way of modifying the rasters of the outside tubes; it is possible in theory to give the scanning coils a shape different from the normal. But this involves two complications: first, electron-optical

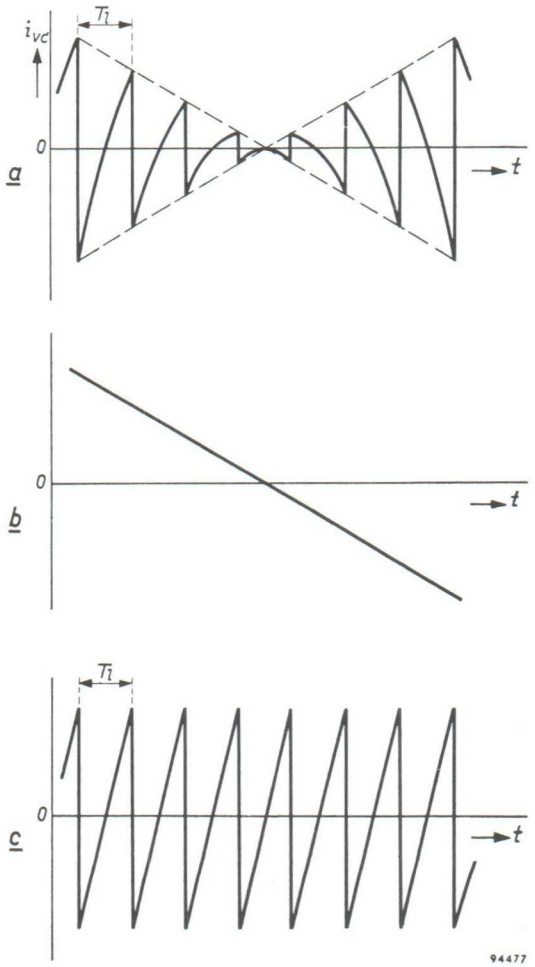


Fig. 8. Trapezium correction. *a*) Waveform of correction current i_{vc} that has to be added to the frame scanning current of the right-hand projector (*B* in fig. 5*a*); the left-hand projector requires correction with a current of the same magnitude but of opposite sense. T_l is the line period. The function drawn in diagram (*a*) is the product of the sawtooth (*b*) of frame frequency, and the sawtooth (*c*) of line frequency.

errors may be introduced and, secondly, the coils in question are difficult to manufacture reproducibly. The method involving special scanning coils is therefore less attractive than that utilizing correction currents.

Optical system of the equipment

The dimensions of the optical systems and projection tubes are determined by what is required of the projected picture as regards size and peak luminance, and also by the luminous efficiency of the phosphors used for the tube screens. Phosphors capable of taking the very high loads that occur and with colour points close to the primary colours laid down in the N.T.S.C. (U.S.A. National Television System Committee) system are listed in *Table I*. Luminous efficiencies in lumens per watt are also indicated.

Table I. Data on the phosphors used.

Primary colour	Composition of phosphor	Coordinates of the colour point		Luminous efficiency lm/W
		x	y	
Red	$(Zn,Be)_2SiO_4$-Mn (with Wratten filter 25)	0.670	0.330	$\eta_r = 2.5$
Green	Zn_2SiO_4-Mn	0.195	0.720	$\eta_g = 21$
Blue	$(Ca,Mg)SiO_3$-Ti	0.160	0.125	$\eta_b = 2.5$

Projection equipment without dichroic mirrors

For equipment not employing dichroic mirrors, the colour points of the phosphors are the same as those of the primary colours forming the projected picture. In order to obtain standard white C (see [1]), fig. 8) from a combination of these primary colours, the luminance values of the primary colours have to be in the following proportions:

$$L_r : L_g : L_b = 29 : 51 : 20.$$

Accordingly, the projection tube screens must dissipate powers in the proportions:

$$P_r : P_g : P_b = \frac{L_r}{\eta_r} : \frac{L_g}{\eta_g} : \frac{L_b}{\eta_b} = 53 : 11 : 36.$$

We see that the red tube screen has to dissipate the highest power. In a normal colour-television picture, white features are responsible for the heaviest loading of all three of the tubes; hence the maximum loading the red tube can be allowed to undergo sets a limit to the maximum brightness that can be obtained in the projected picture. If therefore the three projectors are identical in design, the design of the apparatus as a whole will have to be based on the luminance that the red light contributes to the brightest white parts of the projected picture. If this is desired to have a highlight luminance of L'_{max}, the red contribution will be

$$L_r'_{max} = \frac{L_r}{L_r + L_g + L_b} L'_{max} = 0.29\, L'_{max}. \quad (1)$$

It has been found that the mean luminance of an average television picture amounts to about a third of the peak value. The red tube must therefore supply a total luminous flux of

$$\Phi_r = \frac{\pi b' h'}{\tau_1 \tau_2 a^2 \mu} \frac{L_r'_{max}}{3} \text{ lumen}, \quad . . \quad (2)$$

where

b' is the width and h' the height (in metres) of the picture on the projection screen;
a the numerical aperture of the projector;
μ the "gain" given by the projection screen (due to directionally selective reflection);
τ_1 the masking coefficient (masking effect of CRT in optical system);
τ_2 the optical transmission coefficient of the system;
$L_r'_{max}$ the maximum luminance in cd/m² of the red image on the projection screen.

The electrical power that the red tube must dissipate in order to supply the above value of luminous flux is

$$P_r = \frac{\Phi_r}{\eta_r}. \quad . . . . . \quad (3)$$

If p is the maximum power that can be dissipated per unit area of the tube screen, the area bh of the image on the latter must satisfy

$$b\,h = \frac{P_r}{p}. \quad . . . . . \quad (4)$$

The aspect ratio $b : h$ is laid down at 4 : 3, and this allows us to work out b, h and the diagonal of the picture which should not, of course, exceed the diameter of the fluorescent screen. This gives us the magnification that has to be effected in the optical system of the projector:

$$N = \frac{b'}{b}.$$

Straightforward geometrical-optical reasoning shows that the following relationships must exist between the magnification N, the distances x and x' of object and image from the focus F', the focal length f' and the distance Δ' between the corrector plate and the screen (see *fig. 9*):

$$|N| = \left|\frac{x'}{f'}\right| = \left|\frac{f'}{x}\right| = 1 + \frac{\Delta'}{f'}. \quad . . \quad (5)$$

For a given magnification, f' would have to be

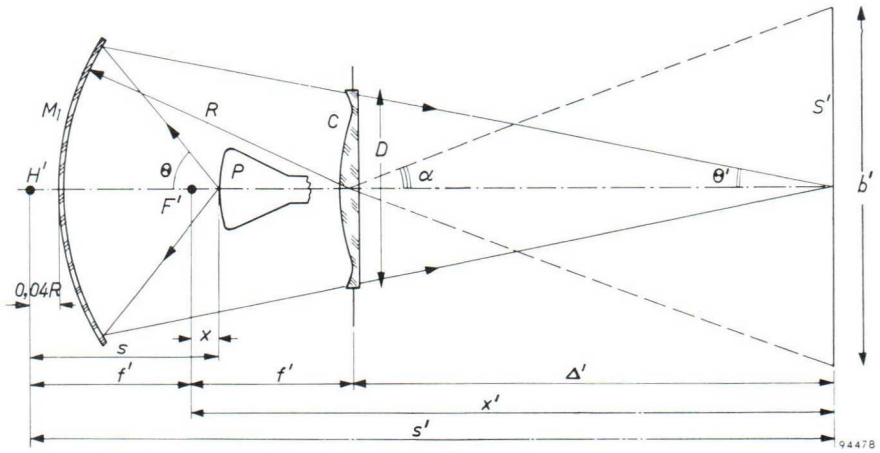


Fig. 9. Schmidt optical system of the "in-line" type. P projection tube. M_1 concave mirror of radius R. C corrector plate of diameter D. S projection screen, the image on which is of breadth b'. F' focus of mirror plus corrector plate. H' pole of mirror plus corrector plate.

increased in direct proportion to Δ', and so also would all the principal dimensions of the optical system. The distance Δ' should therefore be kept as short as possible (this is why the apparatus is set up in front of the viewers). However, there is also a lower limit to Δ', since, as the picture angle $2a$ increases (this being the angle subtended by the width of the projected picture at the centre of the corrector plate — see fig. 9), optical aberrations increase and, apart from that, an increasing fall-off in brightness between the centre and edges of the screen picture becomes evident. There is therefore a certain picture angle $2a_{max}$ at which these effects are still tolerable and which must not be exceeded. Accordingly, the minimum distance between the corrector plate and projection screen is then

$$\Delta'_{min} = \frac{\frac{1}{2}b'}{\tan a_{max}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

We can now obtain the focal length from (5) and (6):

$$f' = \frac{\Delta'_{min}}{|N|-1} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

The radius of curvature R of the concave mirror is approximately $2f'$.

The diameter D of the corrector plate is worked out as follows. We see from fig. 9 that $D = 2\Delta' \tan \Theta'$. Furthermore, $\sin \Theta' = (\sin \Theta)/|N| = a/|N|$. The angle Θ' is so small that $\tan \Theta' \approx \sin \Theta'$. Hence,

$$D \approx \frac{2a}{|N|}\Delta' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

By way of illustration, a design study will now be outlined for an equipment to project a coloured television picture measuring 2.25 m × 3 m and having a maximum luminance of 20 cd/m². We assume that the picture is projected onto a "beaded" screen having a "gain" of $\mu = 2.8$. A practical value for the numerical aperture is $a = 0.7$. The product $\tau_1\tau_2$ can be assumed to be 0.5 and we can take p to be 4000 W/m². Formula (1) gives the maximum luminance of the projected red image:

$$L_r'_{max} = 0.29 \times 20 = 5.8 \text{ cd/m}^2.$$

The following values are obtained successively from (2), (3) and (4):

$$\Phi_r = 59 \text{ lm},$$

$$P_r = \frac{59}{2.5} = 24 \text{ W},$$

$$b\,h = \frac{24}{4000} = 6 \times 10^{-3} \text{ m}^2.$$

From these we work out the dimensions of the image on the CRT screen:

$$b = 9 \text{ cm}; \; h = 6.7 \text{ cm}; \text{ diagonal} = 11.2 \text{ cm}.$$

A projection tube with dimensions of the MW 13-38 (which has a screen of diameter 13 cm) can therefore be used.

The magnification N is equal to $b'/b = 300/9 \approx 33$. In a Schmidt optical system the maximum picture angle $2a_{max}$ is round about 30°, and on this basis we can determine the minimum distance between corrector plate and projected picture from (6):

$$\Delta'_{min} = \frac{\frac{1}{2} \times 3}{\tan 15°} = 5.60 \text{ m}.$$

Formula (7) gives the focal length:

$$f' = \frac{560}{33 - 1} = 17.5 \text{ cm}.$$

The concave mirror must therefore have a radius of curvature of $R \approx 2f' = 35$ cm. The periphery of the mirror is usually given a diameter approximately equal to R; with a smaller mirror, differences of illumination over the projection screen area would be too apparent.

Finally, the diameter of the corrector plate is obtained from (8):

$$D \approx \frac{2 \times 0.7}{33} \times 560 \approx 24 \text{ cm}.$$

Projection equipment embodying dichroic mirrors

The foregoing applies to an equipment that does not embody dichroic mirrors. The design procedure for a projector that does employ such mirrors is similar, except that the design must be based on a different set of relationships between L_r, L_g and L_b, the reason being that the selective transmission and reflection of the dichroic mirrors causes the colour points of the primary colours to shift. In addition, transmission losses in the mirrors must be taken into account in the calculation of the electrical powers that must be dissipated by the projection tubes. As already stated, the "folded" optical system is preferred where dichroic mirrors are to be used; but such a system has a lower transmission coefficient. If the projectors are to have the same dimensions as in the foregoing, the overall result of all this is that highlight luminance of the projected picture will be reduced by roughly 50%. Projectors designed on the above lines and used in combination with dichroic mirrors will therefore be able to produce a maximum luminance of about 10 cd/m² on the projection screen.

Table II gives the dimensions of the four equipments built in this laboratory. Column 1 relates to a projector corresponding roughly to the example worked out above.

Mechanical details of the optical systems

The projection tube, complete with focussing and deflection coils, must be mounted in its optical system in such a way as to enable *optical focussing* to be carried out by altering the position of the tube with respect to the concave mirror. For this purpose it must be possible to move the tube forward and backward along its own axis and also to turn it about two axes perpendicular to each other and to the axis of the tube, running through the centre of the tube screen. *Fig. 10* shows how this is done in the projector depicted in fig. 4. The projection tube P, together with coils D and F, is suspended in a yoke B consisting of a flanged plate holding the tube assembly. Two threaded holes in the plate carry threaded rods I whose ends rest in recesses U and V, being retained by tension springs between the plate and the chassis. The third point of support is provided by an angle bracket fixed to the plate, resting on a bolt (W) threaded in a chassis member. The points U and V lie on the horizontal line passing through the centre of the tube screen. If the adjusting screws marked I are turned to the same extent in the same direction, the tube moves forward or back; if they are turned to the same extent in opposite directions, it turns about the vertical axis through the centre of its screen. Adjusting the knob II on W causes the tube to turn in a vertical plane, about the axis UV. These operations, each of which is substantially independent of the others, allow the whole area of the projected image to be brought into focus.

Table II. Data on the Schmidt optical systems used in the four colour-television projectors constructed.

	1	2	3	4
Dichroic mirrors	no	yes	yes	yes
Type of system	"in-line"	"folded"	"folded"	"folded"
Picture on projection screen:				
Height h' **)	2.25	2.25	0.35	0.22
Breadth b' **)	3.00	3.00	0.46	0.29
Max. luminance L'_{max} (cd/m²)	20	10	200	200
Contrast ratio	35	25	25	25
Definition *)	600	425	450	500
Screen "gain" μ	2.8	2.8	7	7
Diam. of tube screen **)	0.13	0.13	0.06	0.06
Radius of curvature R of concave mirror **)	0.40	0.40	0.20	0.20
Focal length f' **)	0.208	0.208	0.104	0.104
Diameter D of corrector plate **)	0.28	0.25	0.125	0.125
Projection distance A' **)	6	6	0.825	0.645
Magnification $\|N\|$	31	31	9.2	7.4

*) Number of lines; determined with an RMA test plate.
**) In metres.

Vignetting is an important point in the design of optical systems. The endeavour must be to keep the amount of vignetting the same to the left and right, so that both halves of the projected image are equally bright. In an apparatus embodying dichroic mirrors, the images on the three tube screens are in different orientations with respect to their optical systems, owing to the various mirror inversions; if therefore local colour variations are to be avoided, it is particularly important that the vignetting of the optical systems should be the same at the top and bottom of the picture.

We shall deal later with the arrangements for altering the position of the coils. It may be merely

observed at this point that interaction between the focussing and scanning coils may lead to an error in superposition that cannot be corrected; this error takes the form of differences in curvature between horizontal and vertical lines in the three primary-colour images. Such interaction must therefore be obviated, one way of doing so being to fit a copper ring and a ring of material that is a good magnetic conductor between the focussing and the deflection coils.

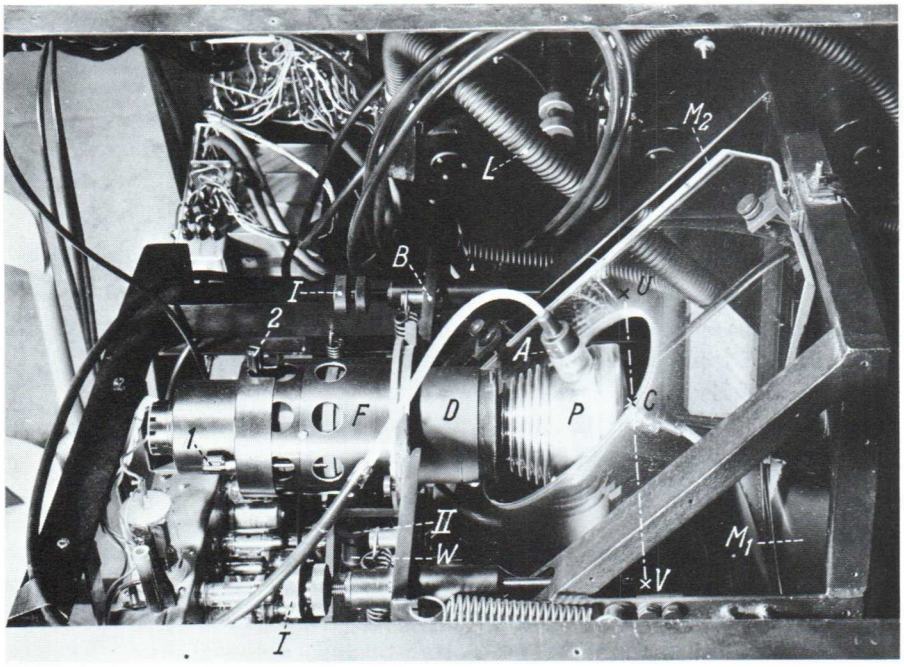


Fig. 10. One of the primary-colour projectors of the large-screen apparatus shown in fig. 4. P is the projection tube, A its anode connection, D the scanning coils and F the focussing coil. M_1 is the concave mirror, M_2 a plane mirror inclined at 45°. L is a flexible pipe through which cooling air is blown against the window of the tube. Part of the video amplifier is visible below, on the left. Tube and coils are suspended in a yoke B supported at points U, V and W. By turning the screws I-I both in the same direction the tube is moved backward or forward along its own axis; turning them in opposite directions causes the tube to swing laterally about the vertical line through the centre C of its fluorescent screen. Turning screw II causes the tube to turn about the horizontal line UV passing through C. Controls *1* and *2* are for moving the focussing coil (see below).

Circuits and adjustments for obtaining superposition of the primary-colour images

If the three primary-colour images are to be brought into correct superposition on the projection screen, the apparatus must allow of a number of adjustments and corrections, some of which have already been mentioned.

1) In an apparatus with converging projectors (fig. 5) *trapezium correction* (figs. 6 and 8) will be necessary to turn the two images with that shape into rectangular ones.

2) The three reactangular images must be of the same size; if must therefore be possible to control the *amplitude* of the scanning currents.

3) It may be necessary to correct the *linearity* of the scanning, in order to ensure movement of the three spots across the projection screen at the same speed.

4) *Displacements* of the three images will be necessary, in order to make their centres coincide.

5) *Rotational alignment* of the images will be necessary, in order to bring their longer edges into parallelism.

In addition to all this it will be necessary to take into account the *spread* in the electrical and mechanical properties of certain components: there will be differences in the deflection sensitivity of the projection tubes, in the position of the electron gun in relation to the tube axis, in the accuracy with which the angle between the two scanning directions approximates to 90° for the different sets of coils, and so on.

Most of the corrections can be incorporated in the scanning arrangements, and we shall therefore start by discussing the appropriate circuits.

The scanning circuits

Apart from the necessary corrections the deflection of the electron beams of the three projection tubes must be identical. Hence, if the three scanning units were exactly similar and if the three tubes had exactly the same deflection sensitivity, the scanning currents would likewise have to be exactly equal for each of the two scanning directions. In order to ensure from the outset that the said currents will be as nearly equal as possible, it is advisable to take them from a common generator for each scanning direction, rather than separate ones for each CRT. With separate sawtooth generators, the free oscillations that might arise in the deflection coils and transformer windings would be independent of one another, and this would render the exact superposition of the three primary-colour images impossible. Moreover, changes in the properties of the scanning generator tubes would, in the long run, give rise to discrepancies in the scanning current amplitudes, and hence to dimensional differences in the primary-colour images. Furthermore, such a method would be expensive.

The common scanning generator can be made to supply the three deflection coils connected either in series or in parallel. It might be thought that, with the series arrangement, exactly the same current would flow through the three coils. However, each coil has a stray capacitance in parallel with it, so that the arrangement in fact constitutes three oscillatory circuits in series, each able to oscillate at its natural frequency. In general, the three frequencies will not be the same and discrepancies will again arise in the scanning currents, resulting in superposition errors. Another objection is that two of the three coils carry high peak voltages with respect to earth, which might give rise to difficulties.

If the coils are in parallel, the same voltage is across all three and hence it is impossible for them to oscillate independently of one another. By inserting a variable impedance in series with each coil, the amplitude and linearity of the current through it can be adjusted to some extent. It is very easy to introduce other corrections that may be necessary, as will presently be evident. For these reasons it is only the parallel arrangement that is used.

Skew-correction

A scanning unit should deflect the beam in two directions with an angle of exactly 90° between them. In practice there is a certain amount of spread in this angle as exhibited by different deflection yokes. Any deviation from 90° manifests itself on the receiver screen as a departure from exact perpendicularity between the vertical and horizontal lines of the picture. The deviation is so slight in normal units that its effect is unnoticeable in monochrome television. In colour television, on the other hand, the deviations may be troublesome because they may differ in magnitude and direction as between the three primary-coulour images. To obviate them, either scanning units must be used that have been built with extreme care, or the fault must be corrected by electrical methods.

In order to correct skewness in the projected image electrically, the line scanning coils are injected with a correction current that is proportional to i_V, the frame scanning current. Since the frequency of i_V is so low that the scanning coils respond to i_V practically as resistances, it will suffice to introduce variable and reversible voltages, proportional to i_V, in series with the coils. These voltages should preferably be obtained from separate sources, in order to prevent the adjustments affecting one another. This skew correction only needs to be provided for in two of the three primary-colour projectors.

Adjustment of amplitude

The output circuit of the common sawtooth generator for the frame scanning is as shown in *fig. 11*. As already stated the frame scanning coils may be regarded as behaving as resistances with regard to i_V. They also possess some inductance, however, and the linearity of the sawtooth current would not be preserved if its amplitude were adjusted with the aid of a variable resistor alone, placed in series with the coil. The ratio of the resistance of the combination to its inductance would vary, and so would the shape of the current, causing errors of superposition. If therefore the shape of the sawtooth supplied to the three CRTs is to be kept uniform, a small variable inductance must be placed in series with each coil (fig. 11).

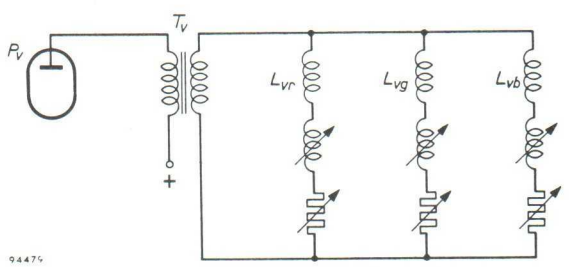


Fig. 11. Adjustment of the amplitude of the frame scanning currents. L_{Vr}, L_{Vg} and L_{Vb} are the frame scanning coils of the red, green and blue tubes respectively. P_V is the output tube of the common sawtooth generator. In series with each of the coils are a variable resistor and a variable inductance, the purpose of the latter being to keep the ratio between total resistance and inductance constant.

With the line scanning coils it is otherwise; these behave more or less as inductances with regard to i_h, the current flowing through them. Consequently, the adjustment of amplitude must be effected by means of a variable inductance in series with the coil, a small variable resistance also being present, so that the ratio of resistance to total inductance can be kept constant. It is desirable that the loading of the horizontal time-base should be kept as nearly constant as possible; accordingly, a potentiometer frequency is fed in via pentode P_{lin} and transformer T_{lin}; it is controlled by means of the cathode resistor of P_{lin}. This allows for the compensation of any lack of linearity in the current from the common sawtooth generator, and is effective for the line scanning of all three projection tubes.

The coarse adjustment of scanning amplitude is carried out elsewhere in the line scanning generator, one control serving for all three scanning coils. The same applies to the frame scanning generator.

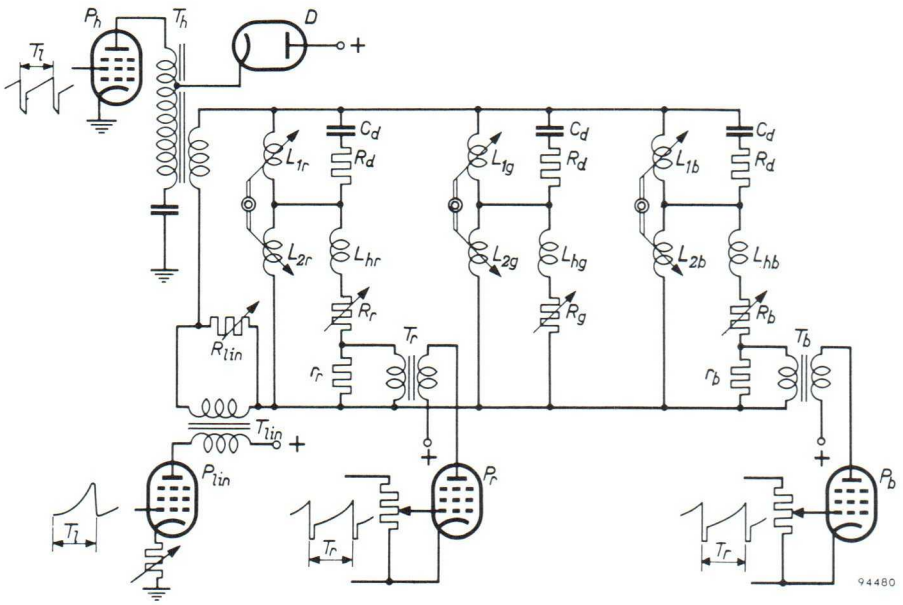


Fig. 12. Adjustment of the amplitude of the line scanning currents. L_{hr}, L_{hg} and L_{hb} are the line scanning coils of the red, green and blue tubes respectively. P_h is the output tube of the common sawtooth generator, T_h transformer. D series efficiency diode (usual in line time-bases). L_{1r}-L_{2r}, L_{1g}-L_{2g} and L_{1b}-L_{2b} are potentiometrically connected variable inductances. R_r, R_g and R_b are variable resistances allowing the ratio between total inductance and resistance to be kept constant. The R_d-C_d networks are for damping free oscillations. The sawtooth current delivered by T_h is rendered linear via pentode P_{lin}, transformer T_{lin} and resistor R_{lin}, the linearity control being the cathode resistor of P_{lin}. A correction current proportional to i_v is fed to the red and blue CRTs via P_r-T_r-r_r and P_b-T_b-r_b respectively (for skew-correction). T_l = line period. T_r = frame period.

arrangement of inductances, as shown in *fig. 12* is to be preferred.

It may also be seen from fig. 12 how the skew correction is effected in a large-screen projector. Transformers T_r and T_b are fed, independently of each other, from pentodes P_r and P_b respectively, these tubes being controlled by an adjustable fraction of the sawtooth voltage on the grid of the frame-scan output tube. Resistors r_r and r_b thus have voltages across them which have the same shape as i_v, and which, by correcting the current i_h flowing through line-scanning coils L_{hr} and L_{hb}, serve to correct skewness in the projected images.

A further voltage of sawtooth shape and line

Trapezium correction

In apparatus with the arrangement of projectors shown in fig. 5, the frame and line scanning of the two outer CRTs must be made subject to correction, as already pointed out (figs. 6, 7 and 8).

A circuit for superimposing a correction current on the *frame* scanning current appears in *fig. 13*. The uncorrected sawtooth current arrives via transformer T_v, the correction current via transformer T_{vc}. The circuit is designed in such a way as to maintain the correct relative direction of the two currents and to prevent mutual interference between the generators producing them.

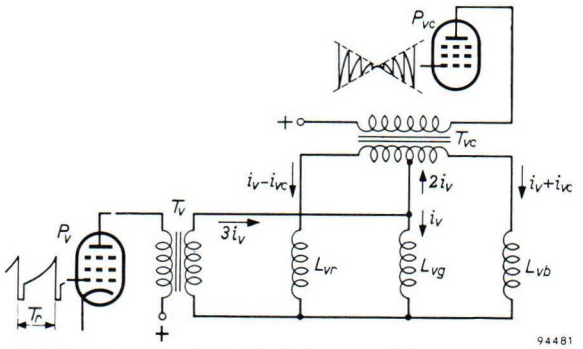


Fig. 13. Circuit for adding i_{vc}, the current for correcting trapezium distortion, to the frame scanning current i_v in the inductances L_{vr} and L_{vb}. The correction current is delivered by tube P_{vc}. T_{vc} is a centre-tapped transformer. The remaining letters have the same meaning as in fig. 11.

Fig. 14 shows the essentials of a circuit from which the correction current can be obtained. Voltages e_1 and e_2 are sawtooth waveforms of opposite sense and have the frame frequency, $1/T_r$. Switches S_1 and S_2 (in reality they are valves) are closed only during the flyback of the line sawtooth. When they are open, good approximations for the voltages u_1 and u_2 across capacitors C_1 and C_2, are:

$$u_1 \approx (E + e_1)\,\frac{t - nT_l}{RC},$$

$$u_2 \approx (E + e_2)\,\frac{t - nT_l}{RC},$$

where E is a direct voltage, R is the value of the resistors in series with each capacitor (both capacitors have the value C), T_l is the line period, n is an integer indicating the order of the line in the raster ($1 \leqslant n \leqslant 625$ for the 625-line system) and t is an instant such that

$$nT_l < t < (n+1)T_l.$$

If $\hat{e}$ is the peak value of e_1 and e_2, and $\beta = \hat{e}/E$, then

$$E + e_1 = E\left(1 - \beta + \beta\,\frac{t}{T_r}\right)$$

and

$$E + e_2 = E\left(1 - \beta\,\frac{t}{T_r}\right).$$

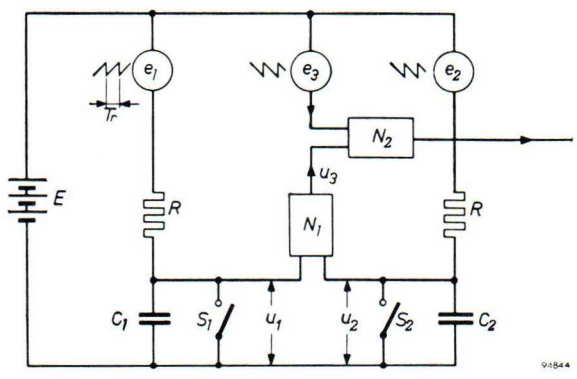


Fig. 14. Schematic diagram of the circuit for generating i_{vc}, the current for correcting trapezium distortion. Switches S_1 and S_2 (in reality they are valves) are closed only during the flyback of the line sawtooth. e_1, e_2 and e_3 are sawtooth voltages having the frame frequency. The subtraction $u_1 - u_2 = u_3$ takes place in network N_1, the addition $u_3 + e_3$ in network N_2.

This gives us the following for the voltages across the capacitors:

$$u_1 \approx E\left(1 - \beta + \beta\,\frac{t}{T_r}\right)\frac{t - nT_l}{RC},$$

$$u_2 \approx E\left(1 - \beta\,\frac{t}{T_r}\right)\frac{t - nT_l}{RC}.$$

These voltages have the waveforms shown in *figs. 15a* and 15*b*, respectively. The difference between the two is a voltage u_3 whose waveform is as shown in fig. 15*c*:

$$u_3 = u_1 - u_2 \approx \beta E\left(-1 + \frac{2t}{T_r}\right)\frac{t - nT_l}{RC}.$$

This needs only the addition of a sawtooth voltage

$$e_3 = \tfrac{1}{2}\beta E\left(1 - \frac{2t}{T_r}\right)\frac{T_l}{RC}$$

at the frame frequency (fig. 15*d*) to give a voltage of the desired shape (fig. 15*e*, cf. fig. 8*a*). The subtraction of u_2 from u_1 takes place in network N_1, the addition of e_3 to u_3 in network N_2.

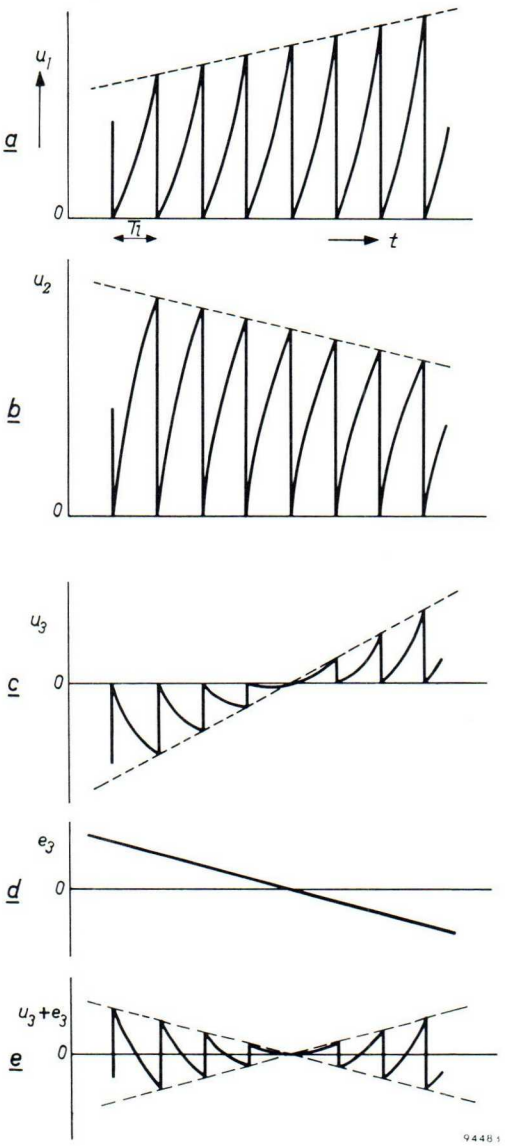


Fig. 15. Voltage waveforms u_1 and u_2, $u_3 (= u_1 - u_2)$, e_3 and $u_3 + e_3$ occurring in the circuit of fig. 14.

The *line* scanning currents of the outer CRTs must be corrected in opposite senses by currents such as were indicated in fig. 6*b*. The shape is approximately parabolic. A correction current of this kind can be obtained by introducing a sawtooth voltage e_{hc} of line frequency in series with the scanning coils. Over one line period e_{hc} can be written as

$$e_h = A\,t \quad (\text{for } -\tfrac{1}{2}T_l \leqq t \leqq +\tfrac{1}{2}T_l),$$

A being a factor of proportionality. In the parallel arrangement of coils of inductance L, this voltage produces a current i_{hc} which, if losses can be neglected, may be written as:

$$i_{hc} = \frac{1}{L}\int e_{hc}\,dt.$$

Apart from a constant of integration, this is equal to $(A/2L)t^2$ and thus corresponds to a curve of parabolic shape. The circuit is given in *fig. 16*. Like that of fig. 13, it is balanced in such a way as to prevent interaction between the two generators.

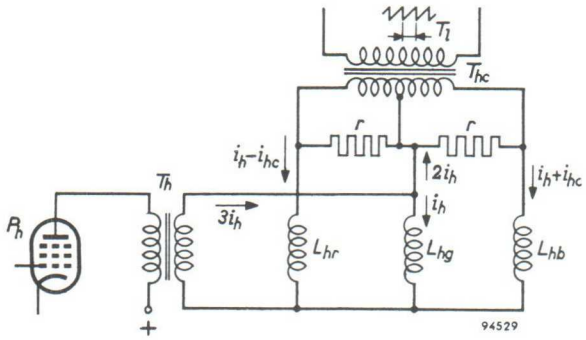


Fig. 16. Circuit for adding correction current i_{hc} to the line scanning current i_h in the inductances L_{hr} and L_{hb}. Transformer T_h, supplied from the output tube P_h, delivers an uncorrected sawtooth current $3i_h$. The centre-tapped transformer T_{hc} supplies resistors r with a sawtooth voltage, which causes i_{hc}, a correction current of parabolic shape, to flow through inductances L_{hr} and L_{hb}.

Displacement and rotational alignment of the images on the projection screen

If the three primary-colour images are to be in exact registration, it must be possible to shift them across the projection screen until their centres coincide, and then to turn them about their common centre until the lines of each raster are exactly horizontal. The latter operation is carried out by turning the deflection coils about the tube axes by means of trimming knobs.

Displacement of the images can be effected in various ways; it can be done with the aid of adjustable direct currents, by turning the focussing coils about axes perpendicular to the axis of the tube, or by altering the alignment of the primary-colour projectors.

1) If an adjustable direct current be passed through each of the six scanning coils, a steady magnetic field of adjustable strength and direction will be set up across each of the CRTs; by this means each electron beam can be biased in direction to the desired degree. The advantage of this method is that the operation can be performed by remote control. But it has a number of disadvantages; one is that the currents have to be stabilized, owing to the temperature dependence of the resistance of the scanning coils.

If the scanning currents are already being subjected to various kinds of correction, there are objections to passing an adjustable direct current through the coils in addition to the others. Separate windings can be incorporated in the coils for carrying these direct currents, but then a high impedance has to be placed in series with the extra windings, since they would otherwise act as short-circuited loops coupled to the scanning windings.

Fitting the direct-current coils in a position before the focussing coils (i.e. nearer the electron gun) is another possibility. Care has to be taken that they do not interfere with the focussing of the beam, and cause part of it to strike the tube diaphragm.

2) It is also possible to give a directional bias of adjustable strength and direction to the electron beam by altering the position of the focussing coil with respect to the axes of the tube. This calls for a mechanical contrivance enabling each of the focussing coils to be turned about two axes at right angles to each other and to the axis of the tube. The method has proved to be very stable and quite satisfactory in practice, in spite of one objection. This arises from slight differences in the position of the electron gun in different projection tubes. For correct superposition of the projected images, it is necessary that the three focussing coils should occupy different positions with respect to the nominal axis of the system. If the "general focussing control" is turned once the projected images have been made to coincide, the three focussing currents will increase or decrease to the same extent, but the images will shift arbitrarily (in directions depending on the position that the gun happens to have in each tube) and superposition will be lost. It is therefore necessary to start by focussing the three images, each of the three focussing currents being separately adjusted, and to adjust for superposition only after this has been done.

The difficulty might be avoided by first setting the focussing coils into a position such that the images would not shift when the value of the focus-

sing current was changed. However, owing to the position of the electron gun, this would generally result in part of the CRT image falling outside the tube screen, which would obviously be quite impermissible. In any case it is desirable that the CRT images should as far as possible occupy the same position on the tube screens, to obviate differences in distortion (due to both optical and electron-optical causes).

3) Altering the alignment of the primary-colour projectors is a purely mechanical operation which can easily be made very stable. For this purpose it must be possible to turn each projector about two axes perpendicular to each other and to the axis of the projector itself. Preferably the two axes should pass through the centre of the corrector plate, but this need not necessarily be the case, if using other axes simplifies the mechanical design. Movement of the projectors about the two axes will then involve a change in their distance from the projection screen, but so small a change that it will not cause any noticeable error.

In the projectors constructed in this laboratory use has been made of a combination of methods (2) and (3); the CRT images are moved into as nearly central a position as possible on the tube screens by turning the focussing coils (i.e. by turning screws *1* and *2*, see fig. 10). The three projected images are then made to coincide on the viewing screen by altering the alignment of the projectors.

Extra-high-tension supply

Our small-screen projection sets are supplied with EHT at 25 kV, the large-screen sets with EHT at 50 kV. It is very important that the EHT supply should be kept strictly constant in spite of fluctuations in the load and in mains voltage. Changes in EHT voltages are accompanied by the following phenomena:

a) A change in the size of the images, due to the fact that the deflection sensitivity depends on the EHT value. Small changes of dimension are permissible providing they affect the three primary-colour images simultaneously and to the same extent. In view of this, the EHT supply for the three projection tubes is provided from a common source.

b) Mutual displacement of the three primary-colour images. The directional bias imposed on the electron beam referred to in the previous section generally differs in amount and direction in the three projection tubes, being dependent on the position that the electron gun happens to have; the consequence is that the three primary-colour images move in different directions when the EHT value changes, superposition thus being lost.

c) Loss of electron-optical focus, resulting in loss of definition.

The EHT value can be rendered more or less independent of the load current by applying parallel stabilization and/or negative feedback.

The term "parallel stabilization" refers to a circuit in which a triode in parallel with the load is controlled in such a way that the total load, consisting of the triode and the three projection tubes, is kept constant. The method is suitable for all systems of generating EHT. Triodes have been developed for this purpose that are capable of standing voltages as high as 80 kV, allowing parallel stabilization to be applied in all cases arising in practice.

Another method makes use of a combination of current and voltage feedback into the EHT generator, which should preferably consist of an oscillator generating a voltage with a fairly high frequency plus a cascade arrangement of diodes and capacitors (voltage multiplier) [6]). A control voltage whose value is dependent both on the load current and on the value of the EHT acts on the oscillator — being applied to its control grid, for example — and on the control curcuit of the regulated HT supply of the oscillator. In this manner it is possible to keep the EHT value constant to within 0.5% over the whole range of load currents.

Neither stabilizing arrangement reacts quickly enough to sudden changes in the load. The circuit should include a buffer capacitor of sufficiently high value to absorb any rapid changes.

The effect of fluctuations in mains voltage can largely be neutralized by using a mains voltage stabilizer or by electronic stabilization of the oscillator HT supply.

Proportional brightness control

It is desirable that a colour-television projector should have facilities for adjusting background brightness by means of a single control. This control, called a proportional brightness control, allows the bias voltages on the control grids of the three CRTs to be altered simultaneously, and in such a way that the colour of the projection screen is always a neutral grey when no video signal is coming in, whatever the position of the control knob. The condition for this may be derived as follows.

[6]) J. J. P. Valeton, Philips tech. Rev. **14**, 21-32, 1952/53.

The luminous flux Φ emitted by a cathode ray tube is related to the grid voltage (measured from the cut-off value) by the following approximate expression:

$$\Phi = c\, V^{\gamma}, \quad \ldots\ldots \quad (9)$$

where c is a constant and the exponent γ has a value between 2 and 2.5. The condition for obtaining standard white C on the projection screen, from a mixture of the primary colours employed, is that

$$\Phi_r : \Phi_g : \Phi_b = 29 : 51 : 20.$$

The voltages on the three control grids must therefore be in the following proportions:

$$V_r : V_g : V_b = \left(\frac{29}{c_r}\right)^{1/\gamma} : \left(\frac{51}{c_g}\right)^{1/\gamma} : \left(\frac{20}{c_b}\right)^{1/\gamma}, \quad (10)$$

c_r, c_g and c_b being values of the constant c in (9), appropriate to the red, green and blue tubes respectively. Since the proportions on the right-hand side of (10) are independent of brightness, it must be possible to achieve the aim in view with linear variable resistors.

The circuit employed is shown in *fig. 17*. The brightness of the red, green and blue tubes is adjusted separately by means of control knobs 2_r, 2_g and 2_b. Each knob actuates two variable resistors (R_r' and R_r'', R_g' and R_g'', R_b' and R_b'') ganged in such a way that their total resistance always remains constant; hence the voltage across potentiometers P_r, P_g and P_b is independent of the position of controls 2_r, 2_g and 2_b. Potentiometers p_r, p_g and p_b are the proportionality controls; they select the fraction of the voltage at the sliders of P_r, P_g and P_b (ganged, and adjusted by knob *1*) that is applied to the grids of the projection tubes. Once 2_r, 2_g and 2_b, and p_r, p_g and p_b, have been correctly set, the background brightness can be adjusted by turning knob *1*, without any change in the neutral tint on the projection screen. The circuit has the attractive characteristic that, when knob *1* is in the zero position, the right settings of 2_r, 2_g and 2_b can be found irrespective of the settings of p_r, p_g and p_b. If this were not so, finding the right settings of knobs *2* and knobs *p* would involve repeated adjustment of first the one set and then the other.

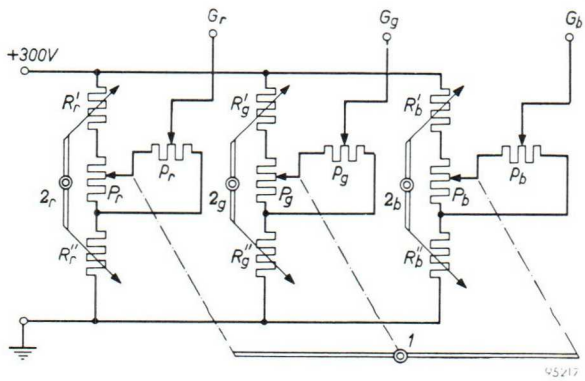


Fig. 17. Schematic diagram of the proportional brightness control. Bias voltages for the red, green and blue CRTs are derived from points G_r, G_g and G_b. *1* is the knob for adjusting general brightness level via the ganged potentiometers P_r, P_g and P_b. 2_r, 2_g and 2_b are knobs for adjusting the brightness of the red, green and blue images separately. The potentiometers p_r, p_g and p_b are for obtaining the right proportions between the brightness values of the three images.

It may be said that the separate brightness controls *2* serve to compensate differences in the cut-off voltages of the projection tubes and that the proportionality controls *p* serve to compensate differences in slope. In reality the projection tubes do not obey law (9) exactly; however, since it is generally only necessary to adjust background brightness over a small range, departures from the correct colour balance are only very slight.

The video amplifiers

Each of the three projection tubes has its own video amplifier. Besides performing all that is normally required of a video amplifier, the three amplifiers must satisfy special requirements arising from the fact that they are together responsible for the colour rendering of the picture on the projection screen.

In colour television systems, as in black-and-white systems, measures are taken at the transmitter to compensate non-linearity in the receiver picture tube (as exemplified by eq. (9) [7]). In theory, therefore, the receiver video amplifiers should be completely linear, so that the overall channel (camera to projection screen) will be linear, as it must be if colours are to be reproduced correctly. If the three amplifiers are fed with the same video signal, the amplitudes at the three inputs being in certain proportions, a monochrome image must necessarily appear on the screen. However, if the characteristics of the projection tubes differ from the nominal characteristic on which the compensation at the transmitter is based, the range of greys between black and white will no longer be faithfully reproduced. It will in fact contain regions that, instead of being neutral in tone, will exhibit a tint that depends on the differences between the actual and the nominal characteristics of the CRTs. What is more, few of the other colours will be faithfully reproduced.

[7]) See the concluding observations in article I (footnote [1])).

As we have seen earlier, to get a given grey shade on the viewing screen, it is necessary that the quantities of red, green and blue light should be present in a certain specific proportion. Because the characteristics of the CRTs conform to the law $\Phi = cV^{\gamma}$, the control voltages will again be in specific — although different — proportions. But in practice the light flux from a CRT screen does not obey the law exactly, one of the main causes being the saturation of the phosphors. There are differences in the properties of the three phosphors; moreover, the anode currents of the CRTs are not equal when adjusted to produce a grey shade on the projection screen. In consequence, deviations from the law $\Phi = cV^{\gamma}$ are not the same for the three primary colours, being greatest for the red phosphor which exhibits the greatest degree of saturation. The saturation effect is particularly strong in apparatus for large-screen projection, which involves a high specific loading of the phosphors. In practice the control voltages on the three CRTs are adjusted to values such that grey shades are properly reproduced in both the lighter and darker parts of the picture (*fig. 18*); the values chosen for the bias voltages are such that the tubes are just cut off when no signal is coming in. It will be seen from fig. 18

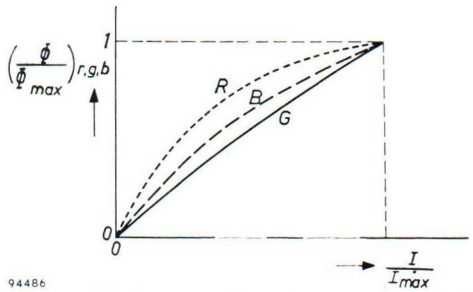
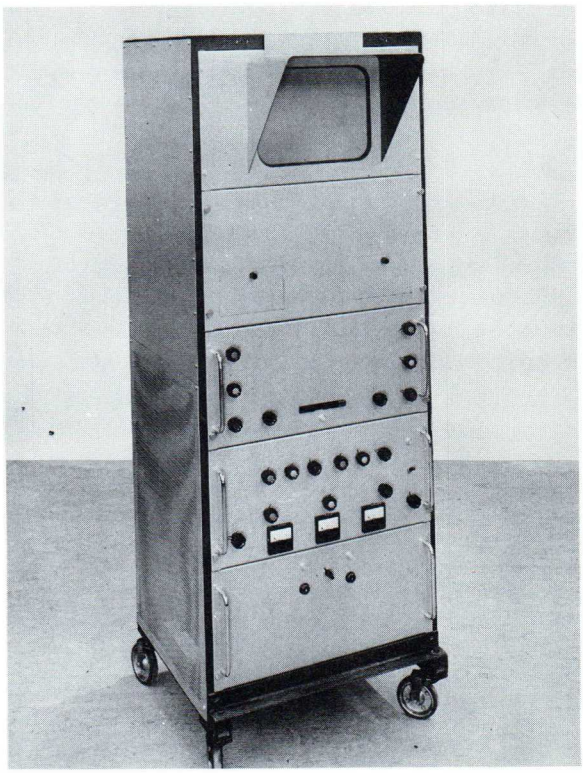
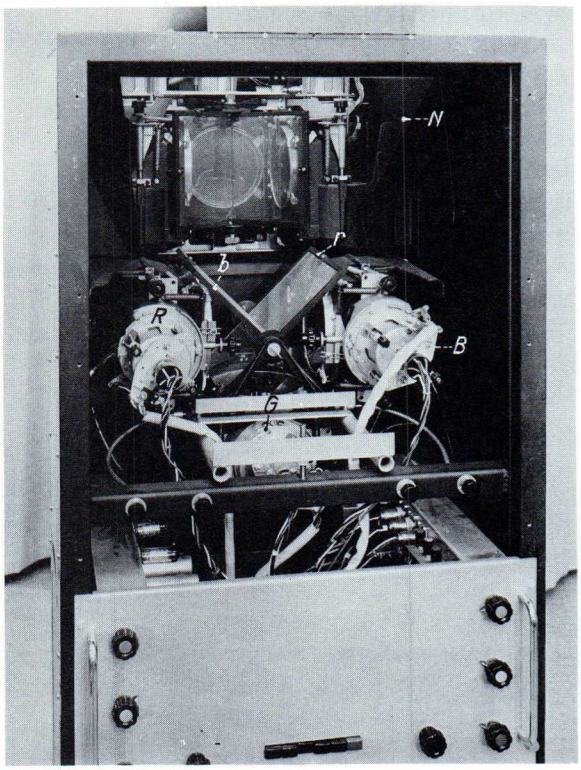


Fig. 18. Ratio of the luminous flux Φ to Φ_{max} as a function of the ratio of anode current I to I_{max}, for the red, green and blue projection tubes (R, G and B respectively). Saturation is heaviest in the red CRT and lightest in the green one.

that the proportion of red light on the projection screen is too great throughout the range of greys between black and brightest white. This effect can be compensated by introducing non-linearities into the video amplifiers such that amplification is higher in the darker parts of the picture than it is in the



a



b

Fig. 19. Colour television projection set for studio monitoring (picture size 22 cm × 29 cm, maximum luminance 200 cd/m²). *a*) The complete set, with transmission projection screen of polyvinyl chloride. *b*) View of the set with the two upper panels removed; R, G and B are the primary-colour projectors, which are on the folded-beam principle. r and b are the crossed dichroic mirrors. N is a plane mirror inclined at 45°.

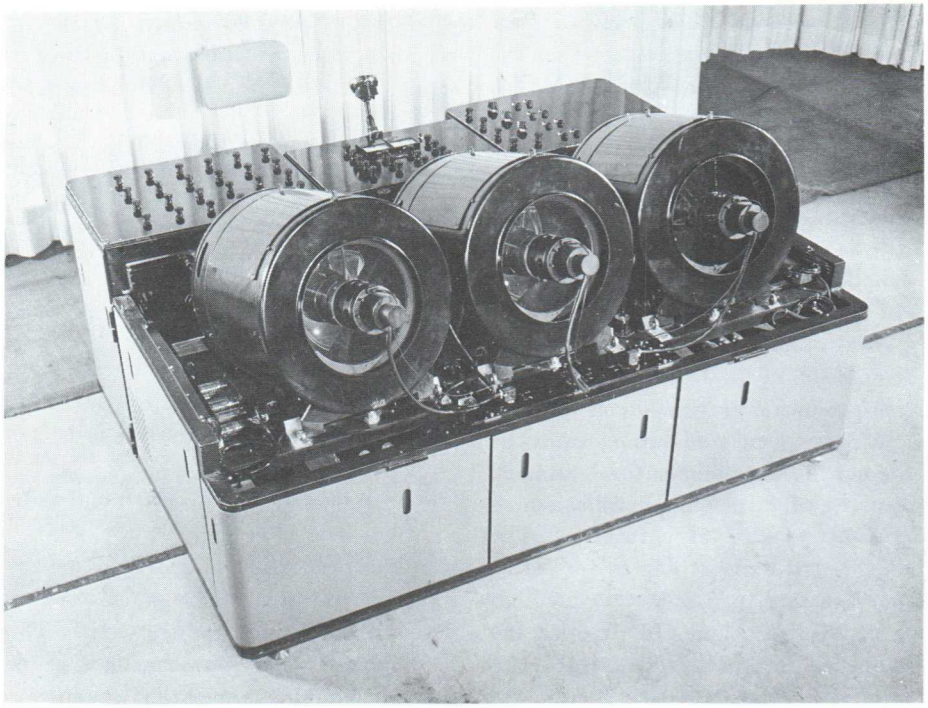


Fig. 20. Large-picture projection set with cover removed. The Schmidt optical systems, which are of the "in-line" type, are mounted side by side. See also figs. 21 and 22.

lighter. This implies a voltage characteristic that can often be approximated by two straight lines of different slope [8]).

In addition, the amount of amplification given by the three amplifiers must remain as nearly constant as possible and be independent of the properties of the electronic tubes with which they are equipped. For this to be so, it is often necessary to apply negative feedback and to stabilize the heater current.

[8]) F. H. J. van der Poel and J. J. P. Valeton, Philips tech. Rev. **15**, 221-232, 1953/54.

Of course, each amplifier must be provided with a contrast control with which the amplification can be adjusted in order to achieve correct colour reproduction.

The projectors constructed

Many of the principles dealt with above have been applied in a number of colour television projectors built in the Philips Eindhoven Laboratory.

In projectors giving relatively small pictures (35 cm × 46 cm and 22 cm × 29 cm) use is made of experimental projection tubes having the same

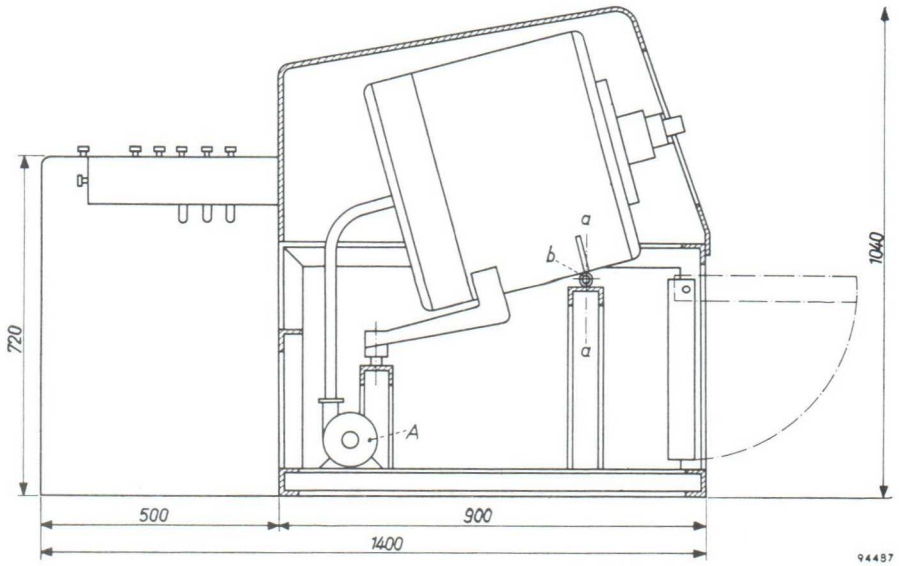


Fig. 21. Sketch showing the principal dimensions of the large-screen projector shown in figs. 20 and 22. Controls within the operator's reach allow him to turn the primary-colour projectors about a vertical axis *a-a* (in a ball and socket joint) and about a horizontal axis *b* (perpendicular to the plane of the drawing). *A* is the fan for cooling the projection tubes.



Fig. 22. The projector of fig. 20 in its housing. The picture size is 2.25 m × 3 m (≈7½′×10′) and the maximum luminance 20 cd/m².

dimensions and supplied with the same voltages as the MW 6-2 monochrome tube [9]) (screen diameter 6 cm, anode voltage 25 kV). Each tube is mounted in a Schmidt optical system with a mirror at 45°, a system of the type originally developed for the MW 6-2 [2]). The light passes through crossed dichroic mirrors, is reflected by a further plane mirror inclined at 45° and falls on to a transmission projection screen of polyvinyl chloride. Grooves on the front and back of the screen give it the desired directional characteristic. *Fig. 19* shows a projection apparatus of this kind, intended for studio monitoring purposes. The picture measures 22 cm × 29 cm and has a maximum luminance of about 200 cd/m².

In equipment for large pictures (2.25 m × 3 m) use is made of experimental projection tubes with a screen of 13 cm diameter and an anode voltage of 50 kV. A photograph of a large-screen projection apparatus equipped with these tubes and with dichroic mirrors has been shown in fig. 4, earlier in this article. A maximum luminance of about 10 cd/m² is obtained on a beaded screen giving a "gain" of $\mu = 2.8$. Electrical adjustments are made by remote control, from a desk at the rear of the hall.

In another type of large-screen projector use is made of "in-line" Schmidt optical systems mounted side by side (*figs 20* and *21*). In this apparatus the control panel forms an integral part of the projector unit (*fig. 22*). Amongst the controls on the panel are six knobs for altering the alignment of the primary-colour projectors. These can be turned about two axes, a vertical axis *a-a* (fig. 21) and a horizontal axis through the points *b*. The apparatus embodies neither dichroic mirrors nor 45°-mirrors, giving as a result a maximum luminance of about 20 cd/m² on the screen (again a beaded screen with a μ of 2.8). Moreover, definition and contrast are considerably better in this projector than in that of fig. 4, and its height (fig. 21) is a good 30 cm less.

[9]) J. de Gier, Philips tech. Rev. **10**, 97-104, 1948/49.

Summary: In the reproduction of colour television pictures by projection methods, three cathode-ray tubes of the projection type are used having respectively red, green and blue fluorescing phosphors. Each tube is mounted in a Schmidt optical system, either of the "in-line" type or of the "folded" type. The article is largely devoted to discussing two methods of obtaining superposition of the three primary-colour images. In the first method dichroic mirrors are employed in combination with folded optical systems. All three images are virtually projected from a single point; in theory, therefore, they are superimposed from the outset. This is not the case in the method where the primary-colour projectors are mounted side by side, but not parallel to one another; here the two outside tubes project images of trapezium shape. The trapezia can be converted into rectangles by making certain modifications to the scanning currents of the outer tubes. The current waveforms required for trapezium correction are discussed, as also the manner of generating them. Other matters dealt with in connection with scanning are skew correction, the adjustment of amplitude, and the displacement and rotational alignment of the images on the projection screen. Extra high tension stabilization methods, proportional brightness control and the characteristics of the video amplifiers are also discussed. Finally, some details are given of a number of colour television projectors actually constructed in the Philips laboratory at Eindhoven. Of two studio monitoring sets whose optical systems embody dichroic mirrors, one projects a picture measuring 22 cm × 29 cm, the other a picture measuring 35 cm × 46 cm; the maximum luminance is 200 cd/m². One large-screen projection set embodying dichroic mirrors gives a picture 2.25 m × 3 m with a maximum luminance of 10 cd/m²; another, in which the primary-colour projectors are mounted side by side, gives a picture 2.25 m × 3 m with a maximum luminance of 20 cd/m².



Photograph H. de Boer

The Philips pavilion at the 1958 Brussels World Fair



The pavilion, in which visitors were entertained to a "spectacle in light and sound" (with electronic music composed by Edgar Varèse), was designed by Le Corbusier and his co-worker Y. Xenakis. It aroused considerable interest amongst architects because of its unusual conception as a shell structure composed entirely of hyperbolic paraboloids, and because of the highly original and elegant method of construction devised by H. C. Duyster. Tests on scale models were made in the T.N.O. Institute at Rijswijk (Netherlands) as a preliminary to the actual building, and the inset shows a plywood model being severely overloaded.

(See: The Philips pavilion at the 1958 Brussels World Fair; III. A. L. Bouma and F. K. Ligtenberg, Model tests for proving the construction of the pavilion; IV. H. C. Duyster, Construction of the pavilion in prestressed concrete; Philips tech. Rev. **20**, 17-27 and 27-36, 1958/59.)

THE PHILIPS HOT-GAS ENGINE WITH RHOMBIC DRIVE MECHANISM

by R. J. MEIJER.

621.412-231.312

Since the first reports on the Philips hot-gas engine, published in this Review in 1946 and 1947, further research and development on the engine has taken place along two different lines. These investigations have resulted on the one hand in the construction of the cold-gas refrigerating machine, now in production for some years, and on the other hand they have led to the design of a new type of drive mechanism which is of special importance for the hot-gas engine proper (particularly large machines). By the application of this drive mechanism it has been possible to drop the idea of using one piston to perform both power and gas-transfer functions (the "double-acting" principle) and return to the thermodynamically more efficient system of separate power and transfer pistons. In this way and by the incorporation of a number of improvements in the design of regenerator, heater etc., the engine has now been given a form that promises well for future development. Measurements carried out on a 40 H.P. experimental engine built according to the new design have demonstrated that, as far as efficiency and specific power are concerned, the new engine can compete with the best amongst familiar forms of prime mover, besides possessing all the virtues inherent in the hot-gas cycle.

With the help of modern materials and with new knowledge of flow and heat-transfer phenomena, the hot-air or hot-gas cycle, which has been known since the early part of the last century, can be made to take place with high efficiency. This was made clear in a series of articles that appeared in earlier volumes of this Review [1,2,3,4,5]).

Contrary to expectation, development work on the cycle in the Philips laboratories did not in the first place lead to a hot-gas engine, but to a *gas refrigerating machine* — which, in the meantime, has come to occupy an important place in refrigeration practice [4,5]). There are several reasons for the slower development of the *engine*. The principal among these were the practical difficulties encountered when it was tried to apply what was termed the "double-acting" or single-piston principle.

This was indeed an elegant principle. The basic type of hot-gas engine, the Stirling engine, has separate power and transfer pistons, while in our "double-acting" engine one moving body performed the functions of both, this constituting a very considerable mechanical simplification. Moreover, in distinction to the small transfer-piston (or displacer-piston) engines made at the time (which worked very well), the single-piston engines did not require the crankcase to be pressurized. This promised considerable advantages for higher-power engines where a pressure crankcase would necessarily involve large-weight penalties. However, these advantages had to be paid for. In the first place an exceptionally intractable lubrication problem now arose: a piston had to act as a moving gas-tight seal between a hot and a cold space between which a large periodic pressure difference occurs. Also, both thermodynamically and aerodynamically, this type of engine was inferior to the displacer-piston engines, owing to the fact that the volume variations of the hot and cold spaces could no longer be freely chosen as regards their relative magnitude and phase.

With further development it has proved possible to drop the idea of a single "double-acting" piston and to design an engine of the *displacer-piston* type

[1]) H. Rinia and F. K. du Pré, Air engines, Philips tech. Rev. **8**, 129-136, 1946.

[2]) H. de Brey, H. Rinia and F. L. van Weenen, Fundamentals for the development of the Philips air engine, Philips tech. Rev. **9**, 97-104, 1947/48.

[3]) F. L. van Weenen, The construction of the Philips air engine, Philips tech. Rev. **9**, 125-134, 1947/48.

[4]) J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine, Philips tech. Rev. **16**, 69-78, 1954/55.

[5]) J. W. L. Köhler and C. O. Jonkers, Construction of a gas refrigerating machine, Philips tech. Rev. **16**, 105-115, 1954/55.

which also possesses the feature that the crankcase need not be pressurized. This was made possible by the embodiment of a new kind of drive mechanism. The latter offers the additional advantage that even a one-cylinder engine can be perfectly balanced. Of the various engines built according to this design, a single-cylinder 40 H.P. machine will be described here and some results of measurements of the output and efficiency will be given.

To save the reader the trouble of consulting earlier volumes of this Review, we shall now very briefly recapitulate the principles of the hot-gas engine, restricting ourselves to an engine provided with a displacer piston.

Brief account of the hot-gas cycle

An internal-combustion engine provides a surplus of work in virtue of the compression at low temperature of a certain quantity of air, to which atomized fuel is added either before or after the compression, the subsequent heating of the mixture by rapid combustion, and its expansion at high temperature.

The hot-gas engine is based on the same principle, i.e. the compression at low temperature and expansion at high temperature of a given quantity of gas. The heating takes place, however, in an entirely different manner, the heat being supplied to the gas from outside, through a wall. For this reason the description "external-combustion engine" is appropriate. Owing to the high thermal capacity of the wall, it is not of course possible to heat and cool the gas simply by rapid heating and cooling of the wall. Stirling had realized as far back as 1817, however, that the gas temperature could be made to change periodically by causing a "displacer piston" to transfer the gas back and forth between two spaces, one at a fixed high temperature and the other at a fixed low temperature — see *fig. 1*. If we raise the displacer piston in fig. 1, the gas will flow from the hot space via the heater and cooler ducts into the cold space. If now the displacer piston is moved downwards the gas will return to the hot space along the same path. During the first transfer stroke the gas has to yield up a large quantity of heat; an equal quantity of heat has to be taken up during the second stroke. The *regenerator* shown in fig. 1 is inserted between the heater duct and cooler duct in order to prevent unnecessary wastage of this heat. It is a space filled with porous material to which the hot gas yields heat before entering the cooler; when the gas streams back, it takes up the stored heat again prior to its entry into the heater.

The displacer system, which serves to heat and cool the gas periodically, is combined with a power piston that compresses the gas while it is in the cold space and allows it to expand while in the hot space (all dead spaces in cooler, heater etc. being disregarded). Since compression takes place at a lower temperature than expansion, a surplus of work results. *Fig. 2* shows four phases of the cycle through which the whole system passes if a *discontinuous* movement of power piston and displacer piston is presupposed. The displacements they are assumed to undergo are plotted as functions of time in *fig. 3*; the ordinates in band *E* represent the variation in the volume of the hot space, and those in band *C* the variation in the volume of the cold space. The volume variations are plotted separately in the lower part of the diagram. *Fig. 4* is the p,V diagram of the cycle (V is the total volume of the gas).

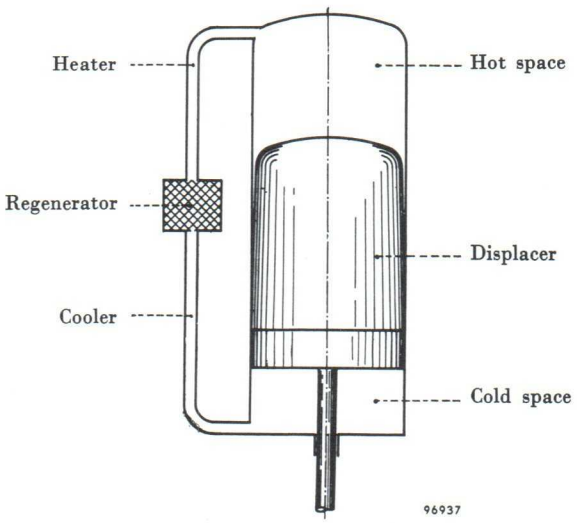


Fig. 1. Principle of the displacer-piston system. Moving this piston up and down causes the gas to be transferred back and forth between the hot and cold spaces, via heater, regenerator and cooler.

In a practical version of the engine the movements of power and displacer pistons must of course be *continuous*, not discontinuous, as they have been assumed to be in these figures; the continuous movements will be obtained with the aid of some kind of crank and connecting rod mechanism. It will not then be possible to distinguish any sharp transitions between the four phases, but this will not alter the principle of the cycle (or detract from its efficiency — see below). The movements of power piston and displacer might now be as indicated in *fig. 5*, in which the volume variations of the cold and hot spaces have again been plotted separately. The only essential condition for obtaining a surplus of work is that the volume variation of the hot space should

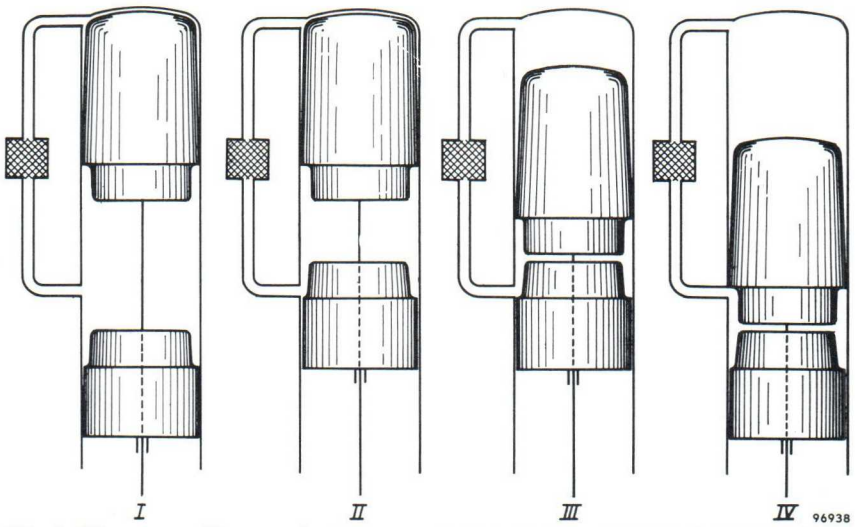


Fig. 2. Diagrams to illustrate the hot-gas cycle. For the sake of clarity the power piston and the displacer piston are supposed to move discontinuously; it is then possible to distinguish four phases that make up the complete cycle.

I Power piston is in lowest, displacer piston in highest position. All the gas is in the cold space.

II The displacer piston is still in the highest position. The power piston has compressed the gas while it is at low temperature.

III The power piston is still in its highest position. The displacer piston has pushed the gas through the cooler, regenerator and heater into the hot space.

IV The hot gas has expanded, and the power and displacer pistons have together returned to the lowest position. While the power piston remains there, the displacer piston will now push the gas through heater, regenerator and cooler into the cold space, whereupon situation *I* will have been restored.

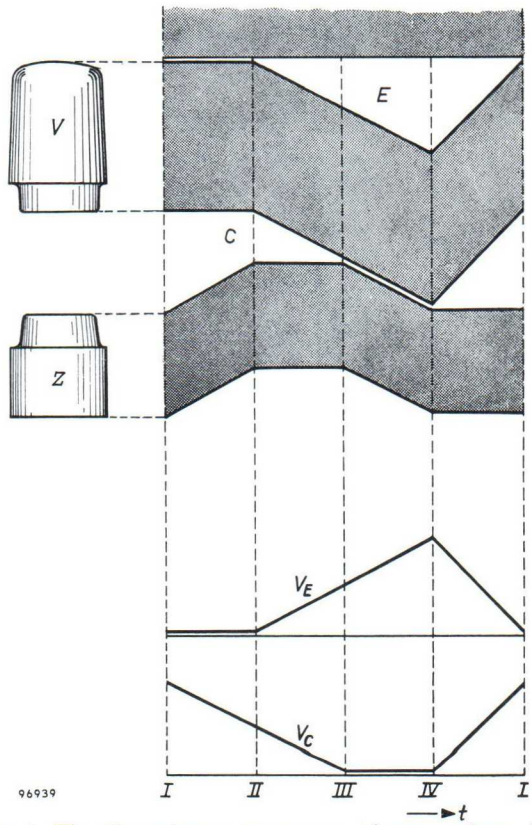


Fig. 3. The discontinuous movements of power piston (Z) and displacer piston (V) assumed in fig. 2, plotted as functions of the time. Band E represents the volume variatons of the hot space, band C those of the cold space. These variations are plotted separately lower in the diagram.

have a phase lead with respect to that of the cold space; this is equivalent to requiring that the appropriate p,V diagram, shown in *fig. 6*, should be traced out in the clockwise direction.

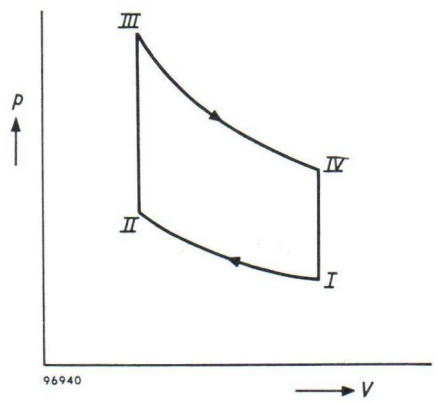


Fig. 4. The p,V diagram of the hot-gas cycle represented by fig. 3.

For the subsequent discussion of specific power and efficiency it will be necessary to recall the more important of the formulae worked out in the articles cited. They are based on the assumption that V_E, the volume of the hot (expansion) space, and V_C, the volume of the cold (compression) space, vary with the crank angle α in a purely sinusoidal fashion:

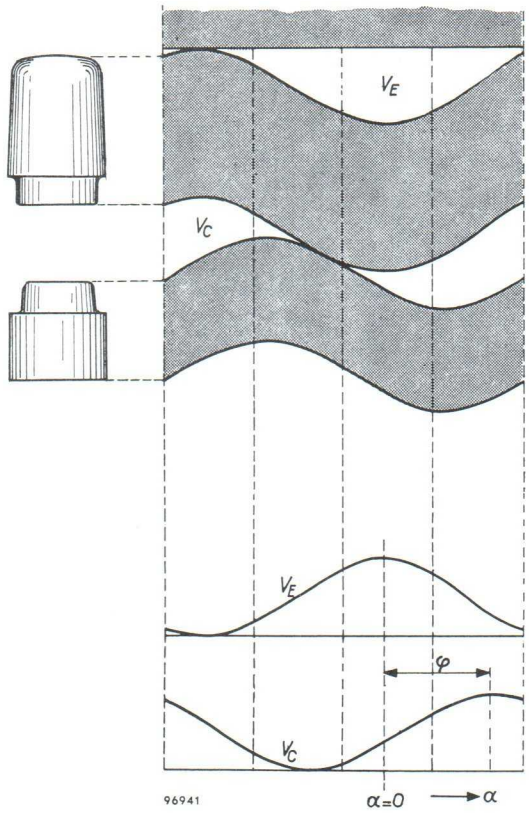


Fig. 5. As for fig. 3, except that now the piston and displacer movements are *continuous* (the abscissa represents the crank angle α). It is no longer possible to distinguish sharp transitions between the phases of the cycle.

$$\left.\begin{aligned} V_E &= \tfrac{1}{2}V_0(1+\cos\alpha), \\ V_C &= \tfrac{1}{2}wV_0[1+\cos(\alpha-\varphi)]. \end{aligned}\right\} \quad . \; . \quad (1)$$

The crank angle α (for constant angular velocity ω, proportional to the time: $\alpha = \omega t$) is measured from the position at which V_E has its maximum value V_0; φ is the phase angle between the volume variations of the hot and cold spaces, and w is the ratio between their maximum volumes. In reality the variations in V_E and V_C will certainly not be purely sinusoidal, but the calculations show that the effect of the higher harmonics can generally be neglected. A further assumption underlies the derivation of the formulae, namely that the cycle is "ideal", characterized by constant gas temperatures T_E and T_C in the hot and cold spaces respectively, and negligible flow and other losses (to which we shall return later on). Another important quantity introduced is the gas temperature ratio,

$$\tau = T_C/T_E \qquad (\tau < 1). \; . \; . \; . \quad (2)$$

The condition that the mass of the working fluid remains constant throughout the cycle now leads to the formula for the pressure p as a function of the crank angle:

$$p = p_{max}\frac{1-\delta}{1+\delta\cos(\alpha-\Theta)}, \quad . \; . \; . \quad (3)$$

in which p_{max} is the maximum pressure occurring during the cycle and

$$\delta = \frac{\sqrt{\tau^2+w^2+2\tau w\cos\varphi}}{\tau+w+2s}, \quad . \; . \; . \quad (4)$$

$$\tan\Theta = \frac{w\sin\varphi}{\tau+w\cos\varphi} \qquad (\Theta < 180°) \; . \quad (5)$$

(s is the relative volume of the "reduced dead space"). From the above we obtain the average pressure,

$$\bar{p} = p_{max}\sqrt{\frac{1-\delta}{1+\delta}}, \; . \; . \; . \; . \; . \quad (6)$$

and the power output

$$P = \tfrac{1}{2}\omega V_0\bar{p}\,(1-\tau)\frac{\delta}{1+\sqrt{1-\delta^2}}\sin\Theta. \; . \; . \quad (7)$$

We are concerned here with a reversible cyclic process in which, in accordance with the "idealized" conditions assumed (isothermal behaviour in cold and hot spaces and 100% efficient regenerator action), the supply of heat takes place at only one temperature T_E and the removal of heat at only one temperature T_C. There is a theorem in thermodynamics which states that under these conditions the efficiency with which heat is converted into work (the thermal efficiency) is that of the Carnot cycle:

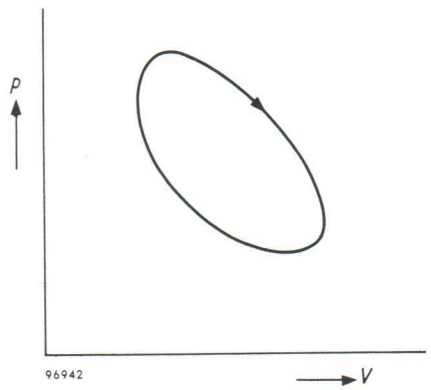


Fig. 6. The p,V diagram of the cycle represented by fig. 5.

$$\eta = \frac{T_E-T_C}{T_E} = 1-\tau. \; . \; . \; . \quad (8)$$

From this we may obtain the quantity of heat supplied per second,

$$q_E = \frac{P}{\eta} = \tfrac{1}{2}\omega V_0\bar{p}\frac{\delta}{1+\sqrt{1-\delta^2}}\sin\Theta. \; . \quad (9)$$

This recapitulation will suffice for our present purposes. We shall now go on to draw certain conclusions which have been essential in deciding the manner in which the cycle must be made to take place in a practical hot-gas engine.

Specific power, losses and overall efficiency

The main considerations governing the design of an engine will in general be to see that its specific power and overall efficiency are as high as possible (consistent, of course, with certain conditions, such as a reasonable working life). The former term refers to the power output per unit volume swept out by the power piston. Overall efficiency is defined as the ratio of work performed per unit time divided by the heat content of the fuel consumed per unit time. These two aims are in some respects contradictory.

We see from formula (7) that the output of a given engine (for which certain values of the parameters w, φ and s have been decided upon) can be stepped up by increasing the mean pressure $\bar{p}$ and the shaft revolutions n (angular velocity ω) and by reducing the ratio τ. A reduction in τ also brings about a direct improvement in the efficiency; formula (8) shows that, theoretically, η is dependent on τ alone. The coolant temperature, which determines T_C, should therefore be made as low as possible, and the temperature T_E at which heat is supplied to the heater element should be made as high as possible, i.e. as high as the properties of the heater wall material (e.g. resistance to creep and oxidation) will permit. To find out the effect on the efficiency, if any, of the other two factors, $\bar{p}$ and n, we shall have to pay attention to the various kinds of *losses*, which are not taken account of the in formulae. The following are the most important [6]).

a) Losses due to mechanical friction in the moving parts.

b) "Adiabatic losses". Within the spaces whose volumes vary, expansion and compression take place more or less adiabatically, not isothermally, with the result that T_{Ea}, the mean temperature in the hot space, will be lower than T_E, the heater temperature and that at which heat is supplied. Similarly, T_{Ca} in the cold space is higher than T_C. Internally, therefore, the engine has a temperature ratio $\tau_a = T_{Ca}/T_{Ea}$ which is greater than the external τ, the implication being a lower efficiency.

c) Fluid friction or flow losses, representing the work done in forcing the gas to flow back and forth through cooler, regenerator and heater.

d) Regenerator losses. Since in practice a regenerator can never be 100% effective, the temperature of the gas leaving it is too high on entry into the cooler and too low on entry into the heater. Thus an additional amount of heat is removed in the cooler, and lost to the engine.

e) Heat-transfer losses at the walls of the elements responsible for heat exchange between the working fluid on the one hand and the coolant and combustion gases on the other. As with the "adiabatic losses", these losses have the consequence that the ratio τ is effectively increased.

f) Flue losses, which are inevitable in any kind of burner. The hot burning gases can yield up heat to the heater body only while their temperature is yet above that of the latter. When they have cooled to this temperature the remaining heat content of the gases will — if no special measures are taken — be lost via the flue. These flue losses can, however, be considerably reduced by further cooling the flue gases by heat exchange with the incoming fresh air, which is consequently preheated.

Now, closer consideration of the influence of $\bar{p}$ and n on the losses shows that the former factor has little effect on their relative magnitude [7]). Hence it is always advantageous to make $\bar{p}$ as high as possible; the experimental engine that will be described below was designed for an average pressure $\bar{p}$ of up to about 100 kg/cm². On the other hand, an increase in the engine speed does affect efficiency adversely, because the *flow losses* increase with n at more than a strictly proportional rate. These losses, however, do not only depend on the engine speed (and, of course, on the dimensioning of the ducts); they differ considerably according to the gas used as the working fluid. Since the hot-gas cycle takes place within a closed system, the designer has a free choice as to the gas so used. We decided on *hydrogen* as the working fluid for the new engine. Owing to its low density, flow losses are not excessive even at considerable engine speeds. Moreover, with hydrogen good heat-transfer characteristics are attainable. In the machine described here the nominal running speed is 1500 r.p.m.

[6]) Losses (*a*) to (*e*) inclusive were discussed at rather greater length with reference to the cold-gas refrigerating machine [5]), in which they likewise occur; the losses under (*f*) are of course absent in the refrigerating machine.

[7]) However, the relative magnitude of the losses (*e*) does depend somewhat on the power developed by the engine and therefore also on $\bar{p}$ (and n too): the greater the rate at which it is desired to pass heat through a given area of heat-exchanger wall, the greater will be the temperature drop across the wall, and hence the smaller is the internal temperature difference, i.e. the greater is the ratio τ.

Helium may also be used as the working fluid instead of hydrogen. The overall efficiency is then found to be somewhat lower.

The further details of the engine we shall now proceed to describe are mainly concerned with the new drive mechanism to which reference has already been made, the design of regenerator, heater, cooler and air pre-heater and finally a complete control system for regulating the power of the engine.

The drive mechanism of the displacer-piston machine

Earlier versions

There are various ways of making the power and displacer pistons perform the required movements. *Fig. 7* shows a drive mechanism of a kind that used to be embodied in small hot-gas engines. The power piston is linked to the crankshaft K by a connecting rod (con-rod) D in the normal way; the movement of the displacer is derived from that of the power piston with the aid of the rocker T, which is linked by arm S_1 to the displacer-piston rod V and by arm S_2 to a certain point on the power-piston con-rod. The latter has to be forked in order to leave free passage for the arm S_1 and the displacer-piston rod, which passes centrally through the power piston. A different form of drive mechanism has been adopted for the Philips gas refrigerating machine — see *fig. 8*. There are three cranks in the shaft, and the two outside ones, which lie at the same angle, are linked to the piston by two parallel con-rods; the displacer piston is actuated from the middle crank, which is offset by a certain angle from the other two.

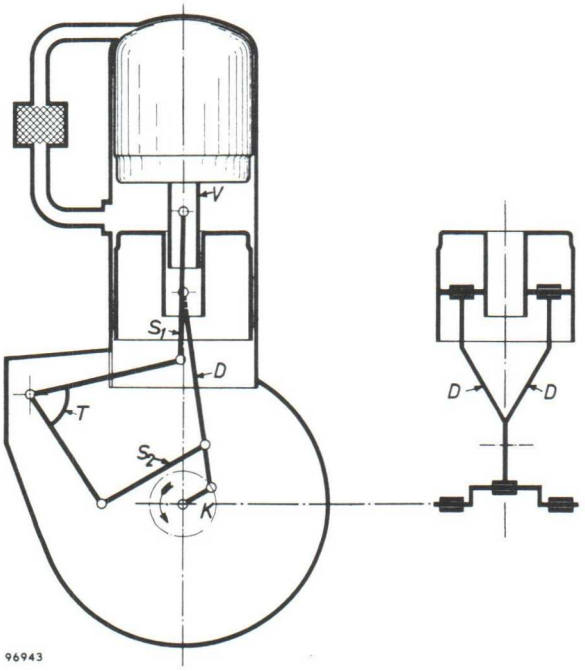


Fig. 7. Drive mechanism of the kind formerly used in small hot-gas engines. D = forked piston con-rod. K = crankshaft. T = rocker linked by arms S_1 and S_2 to the displacer-piston rod V and to a certain point on D.

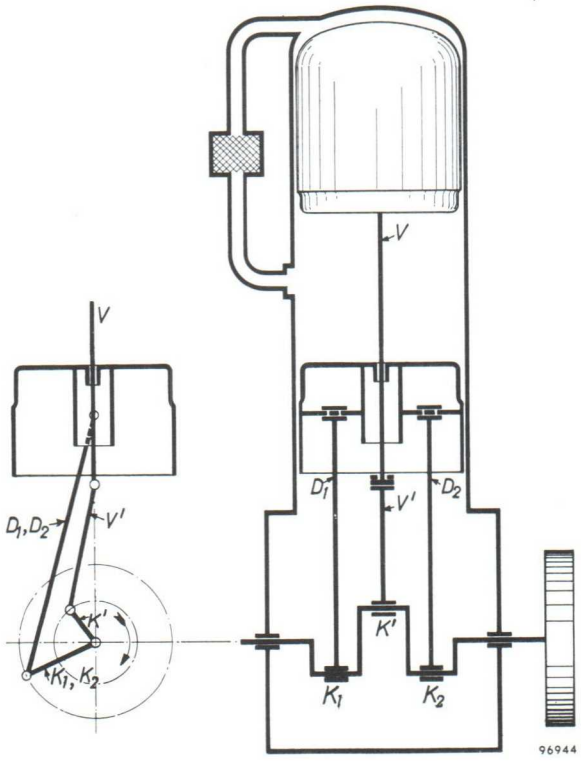


Fig. 8. Drive mechanism used in the gas refrigerating machine. The piston has two parallel con-rods D_1 and D_2 which link it to cranks K_1 and K_2; between the latter is a third crank K', which con-rod V' links to the displacer-piston rod V.

We alluded in the introduction to the drawback of mechanisms such as those of figs. 7 and 8, especially for large engines: the crankcase has to be filled with the working gas under high pressure. This is necessary for two reasons: first, to limit the leakage of gas past the piston; second, to preclude unnecessarily large downward forces on the drive mechanism. In the absence of any compensating pressure (buffer pressure) in the crankcase there is a difference of pressure across the piston fluctuating between p_{min} and p_{max}, the minimum and maximum pressures occurring within the cylinder during the cycle. Introducing a buffer pressure of value p_{min}, for example, reduces the maximum pressure difference across the piston to $p_{max} - p_{min}$.

It is clear that with such a pressurized crankcase considerable weight penalties are involved. Particularly troublesome in large engines, this drawback becomes even more of an embarrassment if it is

desired to operate at very high pressures in order to increase the specific power in the way explained above.

The new drive mechanism (rhombic drive)

The above-mentioned drawbacks do not arise in the new drive, which we now propose to describe. Briefly, it comprises twin cranks and con-rod mechanisms, identical in design and offset from the central axis of the engine; the cranks rotate in opposite senses and are coupled by two gearwheels.

Fig. 9 is a schematic diagram of the system. Fixed to the power piston *1* by way of piston rod *2* is a yoke *3*. One end of the yoke is linked by con-rod *4* to crank *5*, the other end by con-rod *4'* to crank *5'*. The displacer piston is actuated by a precisely similar arrangement: the displacer-piston rod *7*, which passes through the hollow rod *2*, has fixed to it a yoke *8* which is linked to cranks *5* and *5'* by con-rods *9* and *9'* respectively. If *9* and *9'* are given the same length as *4* and *4'*, the two pairs of con-rods will form a rhombus, only the angles of which vary

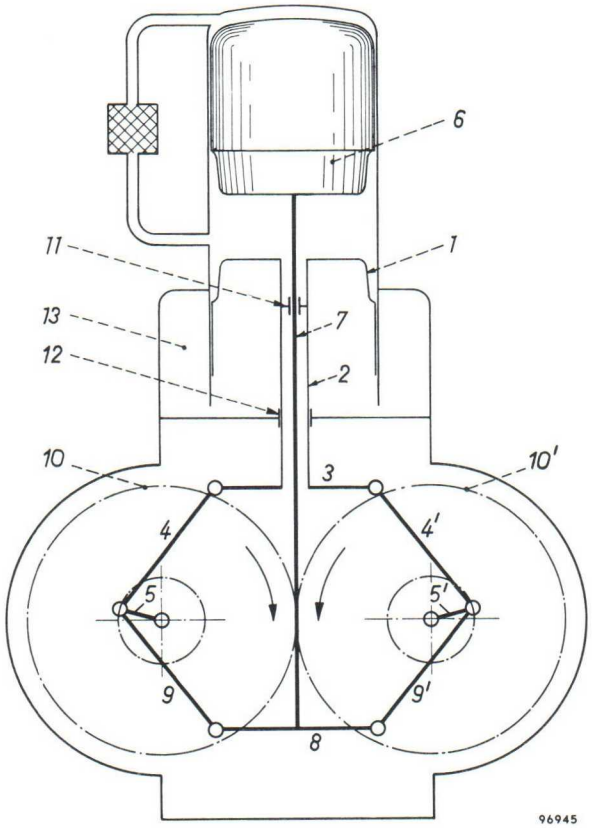


Fig. 9. Schematic diagram of rhombic drive mechanism. *1* = power piston. *6* = displacer piston. *5-5'* = cranks in two shafts rotating in opposite senses and coupled by gears *10-10'*. *4-4'* = con-rods pivoted from ends of yoke *3* fixed to the hollow power-piston rod *2*. *9-9'* = con-rods pivoted from ends of yoke *8* fixed to displacer-piston rod, which runs through the hollow power-piston rod. *11* and *12* = gas-tight stuffing-boxes. *13* = buffer space containing gas at high buffer pressure.

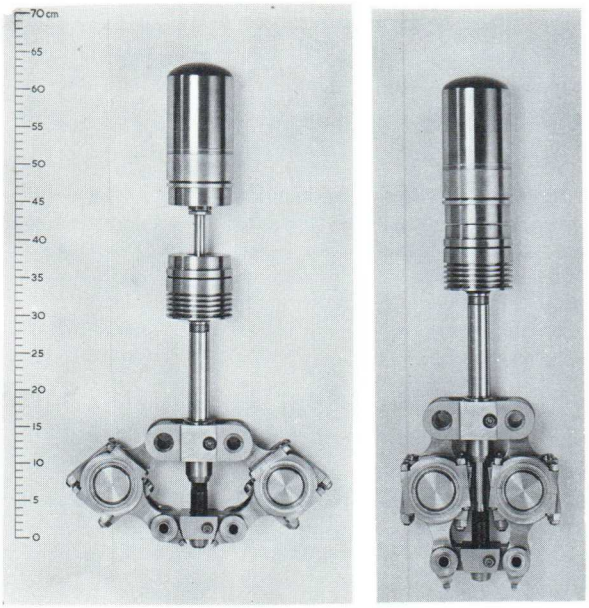


Fig. 10. Model of an actual rhombic drive mechanism, in two positions (cranks and gearwheels not present in this model).

when the system is in motion; it is for that reason that we have adopted the name "rhombic drive". Gearwheels *10-10'* ensure exact symmetry of the system at all times. The two crankshafts being geared together, the entire shaft output can be taken off either.

Fig. 10 shows a practical form of such a crank mechanism. The motion is illustrated by photographing the mechanism in two different positions (the cranks and the gearwheels are absent in this model).

The symmetry of the system and the coaxial arrangement of power-piston and displacer-piston rods make it an easy matter to avoid putting the crankcase under high pressure. The stuffing-box *11* for the displacer-piston rod is inside the hollow power-piston rod. One more seal, namely the stuffing-box round the power-piston rod (*12* in fig. 9), is all that is necessary to form a comparatively small cylindrical chamber *13* under the piston, separate from the crankcase; this "buffer space" can be filled with gas at the desired buffer pressure. The success of this arrangement depends essentially on the fact that the power-piston rod stuffing-box is subject to no lateral thrust, as *the horizontal components of the forces exerted by each pair of con-rods are exactly balanced at each yoke.* (That they do so, and that frictional losses are low in consequence, has the further advantage of enhancing the mechanical efficiency of the system.) The minimum permissible volume of the buffer space is determined

only by the range within which it is desired to limit the pressure variations inside the chamber. In a multi-cylinder engine the buffer chambers can be interconnected; this allows the volume of the individual spaces to be made even smaller.

In principle a space of this kind can be walled off under the piston in engines embodying the crank systems shown in figs. 7 and 8, but it is then necessary to have the rods running in a crosshead guide in order to prevent their gas-tight stuffing-boxes being acted upon by the horizontal components of the connecting-rod forces. This leads to considerable structural complications.

The power piston and displacer movements resulting from the new drive are displayed graphically in *fig. 11*, in the same way as in figs. 3 and 5. It will be seen that if the direction of rotation is as indicated in fig. 9, the volume variation of the hot space

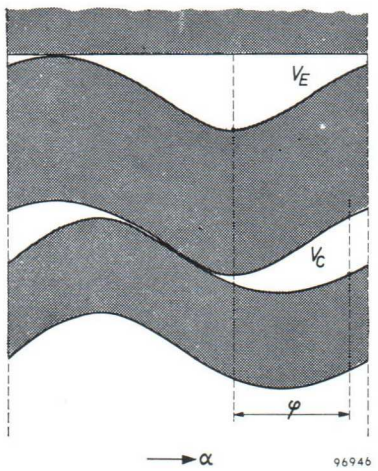


Fig. 11. Curves similar to those in figs. 3 and 5, showing the displacements undergone by the power and displacer pistons in a rhombic drive system such as that of fig. 9.

will have a phase lead with respect to that of the cold space, as is required. The power and displacer pistons do not by any means move in simple harmonic motion, but it is found that a very good version of the Stirling cycle is obtainable. This might seem surprising, for on the face of it the use of the rhombic drive may appear to have restricted the freedom of choice with regard to parameters w and φ (the amplitude ratio and relative phase of the variations in volumes V_E and V_C — see above), which play a large part in the design of a hot-gas engine. However, more exact analysis of the new drive shows that it is in fact possible, by altering the offset of the crankshafts, the proportions of cranks and con-rods and the ratio of power and displacer piston diameters, to vary these parameters over quite a wide range.

The balancing of the drive mechanism

The rhombic drive has the important property of allowing the designer of an engine even with only *one* cylinder to obtain complete dynamic balancing of the forces due to the inertia of moving parts, and of the moments of these forces — "complete" in the sense that the fundamentals *and* all the higher harmonics are balanced. We shall demonstrate this for the simple case where the con-rods are of equal length (as in fig. 9).

The configuration is given in *fig. 12*. On grounds of symmetry it is clear that the sum of all inertial forces acting horizontally is zero at any given instant. The same applies to all inertial-force moments about an axis perpendicular to the plane of the drawing. Hence all we need consider are the inertial forces acting in the vertical direction. We begin by resolving the circular motion of each crankpin T-T' into a vertical and a horizontal component. The vertical movements that the power and displacer pistons make as a result of the *horizontal* movement of the crankpins are equal and opposite because the con-rods are of equal length (symmetrical deformation of the "rhombus"). Hence, if the masses of the two pistons (including the rods etc. attached thereto) are made equal:

$$m_z = m_v, \quad \ldots\ldots \quad (10)$$

the sum of the vertical inertial forces corresponding to horizontal crankpin movement is always zero. There remain the vertical movements of power and displacer pistons as a result of the *vertical* component of the crankpin movement. These movements are exactly equal and, moreover, identical with the vertical movement of the crankpins. One can therefore imagine the moving mass of the two pistons,

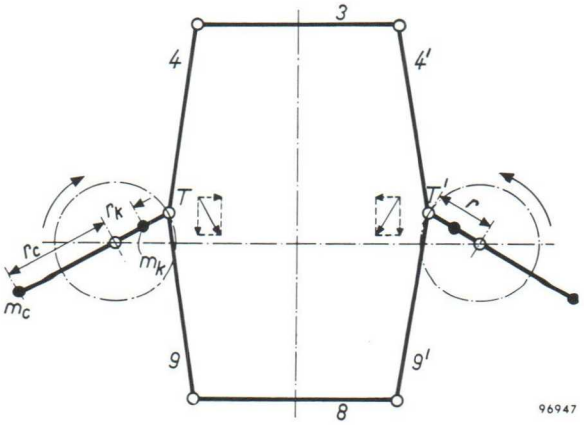


Fig. 12. Configuration of rhombic drive in which yokes *3* and *8* are of equal length, and all four con-rods *4-4'* and *9-9'* are equal. For discussing the inertial forces, the circular motion of crankpins T-T' can be resolved into vertical and horizontal components.

$m_z + m_v$, as localized in the crankpins, $\frac{1}{2}(m_z + m_v)$ in each. By means of two counterweights m_c mounted opposite each crankpin at a suitable radius r_c, vertical inertial forces can be created which exactly balance those of $(m_z + m_v)$ — and this is true even if the rotation of the crankshafts is not exactly uniform.

In fact, the mass and offset of each counterweight are chosen such that the latter also serves to balance the crank and other eccentrically positioned moving parts. If m_k is the effective mass of all these parts, and r_k the distance of their centre of gravity from the crankshaft axis, we have the following condition for balancing:

$$r_c m_c = \tfrac{1}{2} r (m_z + m_v) + r_k m_k . \quad . \quad . \quad (11)$$

From the foregoing it will be clear that the reason why it is possible in this way to balance the inertial forces in a displacer type of engine is that, even for single-cylinder machines, there are *two* reciprocating masses whose movements differ in phase [8]). In order to satisfy the condition $m_z = m_v$ it will generally be necessary to make the masses of the two pistons themselves different. This should not involve any difficulty, for the two bodies have different shapes and different functions and are not interchangeable.

The configuration of fig. 12 is a special case of a much more general form of the new drive. We obtain the more general form by abandoning the restriction that the pairs of connecting rods are equal in length, and by allowing each displacer-piston con-rod to hinge not on the crankpin but about any arbitrary fixed point on the power-piston con-rod (see *fig. 13*). In this general form, which design considerations may render more attractive than the special one of figs. 9 and 11, the power and displacer piston rods continue to move coaxially without the presence of any lateral thrust. This means that the expedient of the buffer chamber with gas under pressure can again be applied without complications. Moreover — and perhaps surprisingly — it can be shown that the inertial forces of the moving parts of the general form of this drive mechanism can be completely balanced as before, provided certain straightforward conditions are fulfilled. We shall state these conditions here without proof. For balancing to be possible the parameters of the mechanism as indicated in fig. 13 must satisfy the equations:

$$\left.\begin{aligned} \Delta e &= 0 , \\ c &= \sqrt{1 - 2\frac{a}{l}\cos\beta + \left(\frac{a}{l}\right)^2} , \\ m_1 - m_4 &+ \frac{a}{l}(m_3 + 2m_4)\cos\beta = 0 . \end{aligned}\right\} \quad . \quad . \quad . \quad (12)$$

The net inertial force can then be completely balanced by fitting counterweights whose positions and masses are given by

$$\gamma = \pi - \tan^{-1} \frac{\frac{a}{l}(m_3 + 2m_4)\sin\beta}{m_1 + m_2 + m_3 + m_4 + \frac{r_k}{r} m_5}$$

and

$$m_6 = \frac{r}{r_c} \sqrt{\left(m_1 + m_2 + m_3 + m_4 + \frac{r_k}{r} m_5\right)^2 + \left\{\frac{a}{l}\left(m_3 + 2m_4\right)\sin\beta\right\}^2} \quad . \quad . \quad . \quad (13)$$

respectively. It will be seen that when a is zero, as it is in the special case first discussed, these conditions transform into the simple ones given earlier.

The particular version of the hot-gas engine with which we shall concern ourselves in this article has the simple form of crank mechanism shown in fig. 9.

In addition to inertia forces, an *alternating moment* acts on the foundation of every piston engine when under load as a consequence of the periodic variation of the torque produced on the shaft. With the rhombic drive mechanism (either the special or the general form) a relatively simple expedient makes it possible to balance out a large part of this alternating moment. It is only necessary to fit two flywheels instead of the usual one, one on each of the two oppositely-rotating crankshafts.

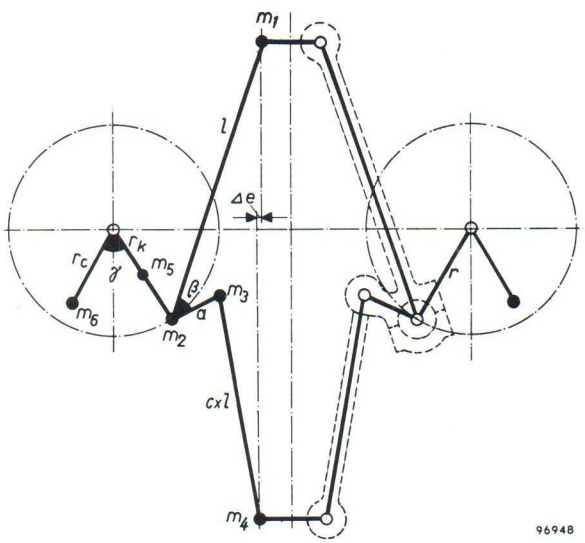


Fig. 13. General form of the new drive. $m_1 \ldots m_6$ are the masses of the moving parts, imagined localized at various points for the purpose of deriving the conditions for the balancing of inertial forces and moments.

Further constructional details of the engine

The general appearance and layout of the experimental engine may be seen in *fig. 14*. *Fig. 15* is a cross-sectional sketch, burner and pre-heater not being shown; for the sake of clarity some of the parts have not been drawn exactly to scale. The notation of the drive mechanism is the same as in preceding diagrams; the counterweights are also shown (*14*).

[8]) It is clear that the rhombic drive could also be used for attaining complete dynamic balancing in internal-combustion engines, but it would probably be a proposition only in engines with at least two cylinders, i.e. in which at least two pistons are in reciprocating motion.

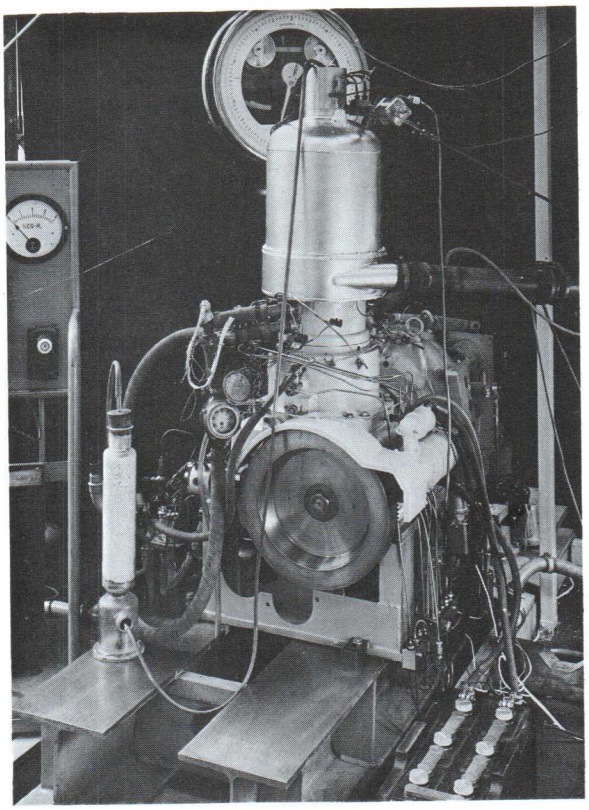


Fig. 14. Experimental model of the Philips hot-gas engine with rhombic drive, in position on the test bench.

As in previous designs described in this Review [3]), the cooler, regenerator and heater all have an annular configuration, and are mounted around the expansion-compression cylinder. This arrangement makes for compactness, and low gas-flow losses. The displacer piston consists of the piston base *6a* and a thermally insulating dome *6b*. The piston base fits like an ordinary piston into the cylinder *16* at the level of the cooler. There is a small clearance between the dome *6b* and the cylinder wall *15* of the hot space, just sufficient that the two never come into contact. Underneath the power piston the buffer chamber *13* can be seen.

The design of the *regenerating casing* has undergone a good deal of modification. In the first version of the experimental engine the regenerator casing was a comparatively flat annular body (*fig. 16*); the cylinder wall *15* formed the inner wall, the outer wall *a* being joined to the cylinder via perforated flange *b*. Considerable stresses were set up in the walls of the casing owing to expansion of the upper side as its temperature rises to the high value T_E, the lower side remaining at the low temperatur^ T_C. Efforts to step up the output of the hot-gas engine were soon brought to a halt by the thermal stresses arising in this way, these becoming excessive especially as the working pressure of the engine was raised. The difficulty has been avoided by dividing up the annular regenerator space into a number of

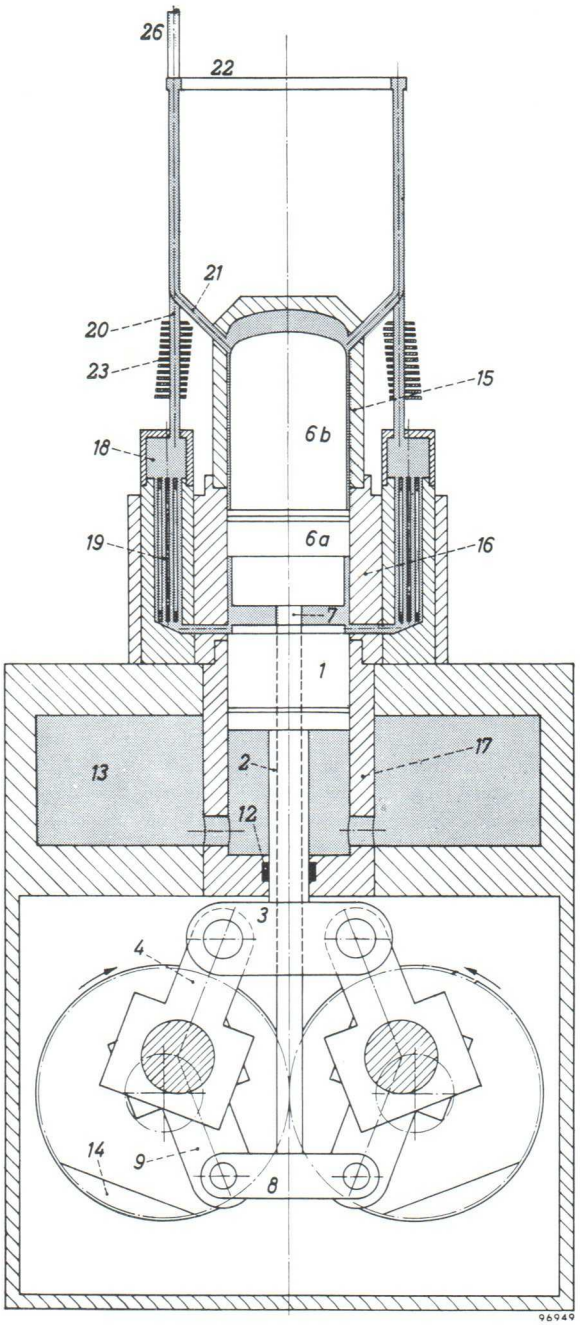


Fig. 15. Cross-section of the engine shown in fig. 14, much simplified and, for the sake of clarity, not entirely drawn to scale. The burner and air pre-heater are absent; they are shown separately in fig. 18. Spaces filled with hydrogen are shaded. *1* = power piston. *6a* and *6b* = displacer piston and insulating dome. *12* = power-piston stuffing-box. *13* = buffer chamber. *14* = counterweights. *15*, *16*, *17* = cylinder in which power piston and displacer piston move. *18* = regenerator compartment. *19* = cooler compartment, *20*, *21*, *22* = heater tubes. *23* = fins. *26* = tube for temperature probe. Other symbols as in fig. 9.

small chambers ("cups") not contiguous with the cylinder wall; see *fig. 17*. Thermal stresses in the regenerator wall then ceased to be a problem. Furthermore, it was now possible to ensure that the lines of principle stress ran substantially straight from the head of the cylinder *15* via the cylinder

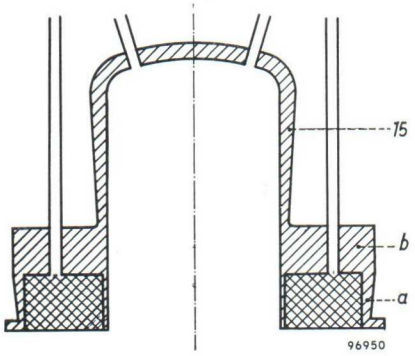


Fig. 16. Earlier design of regenerator compartment, with cylindrical outer wall *a* attached to cylinder wall *15* by means of perforated flange *b*.

wall *16*, *17* to the base of the latter cylinder (see fig. 15). In earlier designs embodying the flange mentioned above, these lines tended to curve outwards, which meant that the flange had to withstand a strong bending moment. The new design allows the whole assembly to be kept comparatively light despite the high maximum pressure of the gas.

The design of the cooler has been adapted to that of the regenerator. The cooler housing *19* contains a number of tubes bunched together in groups (cooler "units"), the end of each unit protruding into one of the regenerator cups. The tubes of each cooler unit are mounted with a *sliding fit* in the cooler housing; in this way the regenerator cups are left free to undergo small longitudinal displacements consequent on the expansion of the heater tubes.

The heater consists of two sets of pipes, *20* and *21*, in close-packed configuration but with small clearances. The hot gases from the burner are blown between these tubes in a direction perpendicular to their length. The tubes of set *20* terminate inside the regenerator cups (see figs. 15 and 17); gas from the regenerator streams upward through these tubes into the annulus *22*, whence it is carried downwards by the tubes of set *21*, which are brazed into the hot end *15* of the cylinder. To improve the heat transfer between the combustion gases and the working gas, a series of fins *23* are brazed to the pipes of set *20*.

The heater is surrounded by the air *pre-heater* (*27* in *fig. 18a*), which functions in the following way. The air necessary for combustion enters the

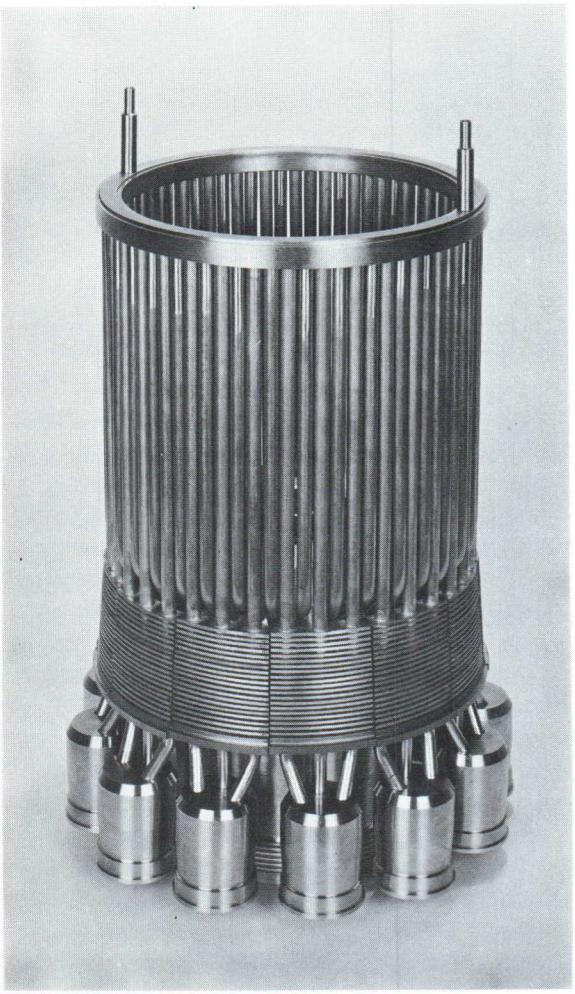
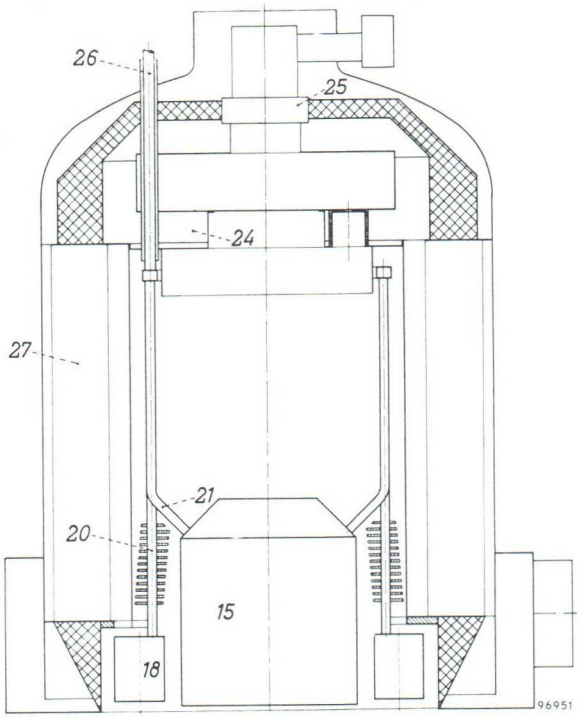
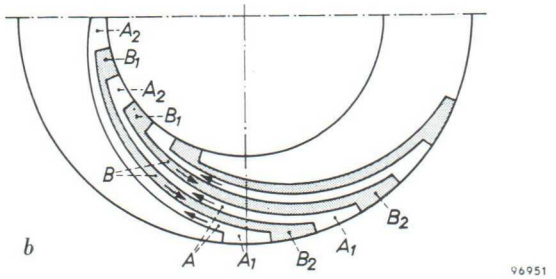


Fig. 17. New design of regenerator compartment and heater. The regenerator (below) is now split up into a series of cups surrounding but not touching the displacer cylinder. The new design solved the thermal-stress problem. It will be seen that three tubes issue from each regenerator cup, their extensions forming the vertical heater tubes that carry the upward-streaming gas. Alternating with these, and connected to them via the annulus at the top, are downward-flow tubes that lead to the head of the cylinder. The lower part of this tube system is fitted with fins in order to improve the heat transfer from the combustion gases to the tubes.

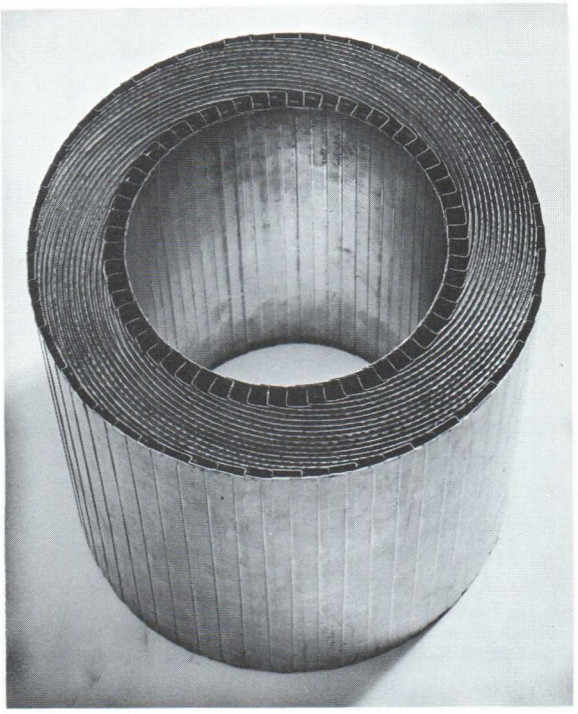
lower end of channels A_1 situated around the outer perimeter of the air pre-heater (fig. 18*b*) and then flows via the spaces A between the spirally curved partitions into the channels A_2 around the inner perimeter and thence, upwards, to the burner. The exhaust gases, which leave the burner with a temperature somewhat higher than that of the heater, enter channels B_1 around the inner perimeter and flow via B, a second set of spiral spaces lying between the A spaces, in a direction opposite to that of the incoming air, to the channels B_2 around the outer perimeter of the air pre-heater. It has proved pos-



a



b



c

Fig. 18. *a*) Cylinder head of engine, surmounted by burner *24*, atomizer *25* and surrounded by air pre-heater *27*. Other symbols as in fig. 15.
b) Cross-section through air pre-heater, showing the narrow *A* and *B* spaces which lie between the spirally curved partitions and through which pass, in opposite directions, fresh air for and spent gases from the burner. The air enters via the lower end of the channels A_1 and leaves via the upper end of A_2, both sets of channels running parallel to the cylinder axis. Similarly the burnt gases enter via the lower end of channels B_1 and leave via the upper end of channels B_2, likewise parallel to the cylinder axis.
c) View of air pre-heater, which fits over the heater shown in fig. 17.

sible on this principle to build a very compact pre-heater (fig. 18*c*) with low resistance to air flow. Moreover, the pre-heater "insulates itself"; all the channels around its outer perimeter (A_1 and B_2) are near the ambient temperature.

The *burner* (*24* in fig. 18*a*) is of the "swirl-chamber" type, which can be used in conjunction with a large number of fuels possessing widely differing properties, such as coal gas, propane, butane, kerosene, diesel oil and light fuel oil. If liquid fuels are to be used, the burner is fitted with an atomizer *25*, which sends a fine spray of the liquid into the upper part of the burner.

Regulation of engine output

It will be seen from formula (7) that the power output of the hot-gas engine can be controlled very conveniently by altering $\bar{p}$, the mean pressure of the working gas; changing $\bar{p}$ has no effect on the thermal efficiency (eq. 8) of the engine, provided the cooler and heater temperatures remain unaltered. To fulfil this proviso a separate automatic regulator is required. (The overall efficiency is in fact somewhat affected by a pressure change because, in addition to the effect mentioned under [7]), the absolute magnitude of some — less important — losses remains practically constant; the effect can be deduced from fig. 21 below.)

Accordingly, the control system of the hot-gas engine falls into two parts:

1) A system for admitting gas to and withdrawing it from the engine in order to adjust the pressure to give the output required (i.e. the braking torque), while maintaining the (pre-set) speed of the engine constant.
2) A system for controlling the fuel supply to the burner in such a way that the temperature of the heater remains steady when the power output (and hence the rate at which heat has to be

supplied) is changed. For this purpose, control action is exercised by a thermostat, the temperature-sensitive element of which is inserted, for example, in the stream of gas through one of the heater pipes (through tube *26* in fig. 15; in the heater of fig. 17 two such tubes are incorporated).

The control system for regulating the pressure will be discussed here.

Regulation of the pressure in the engine described here is effected by a set of hydraulically operated valves. A diagram of the system is given in *fig. 19*. The governor *1*, which is driven by the engine shaft, ensures that, at the nominal speed to which the engine has been adjusted to run, a certain oil pressure is maintained in pipes *2* and *3*. If the engine slows down slightly in consequence of an increase in the braking torque, the governor raises the oil pressure in those pipes, with the result that valve *4* in the feed device *5* opens up, and gas from tank *6* passes via feed valve *7* into the space *8* which represents the cylinder of the hot-gas engine. Pressure inside the engine having risen accordingly, its power output increases. The injection of additional gas continues until the engine speed has reverted to the original setting (whereupon valve *4* closes). If the speed increases slightly in consequence of a decrease in the braking torque, the governor causes the oil pressure in pipes *2* and *3* to drop, with the result that valve *9* of the regulator *10* opens up, whereby allowing gas from the cylinder to flow via non-return valve *11* to the inlet of a small auxiliary compressor *12*, whence it passes into the tank *6*. The release of gas from the engine continues until the nominal speed is attained (whereupon valve *9* closes).

As stated above, the buffer space *13* under the piston (see also fig. 15) is filled with gas at a buffer pressure; this pressure must also undergo adjustment in accordance with changes in the load. The engine under discussion has been designed for a buffer pressure equal to $\bar{p}$, the mean pressure in the cylinder. In principle this buffer pressure is obtained automatically, by means of a capillary tube connecting the buffer space with the cylinder. We shall return presently to other provisions made for this purpose.

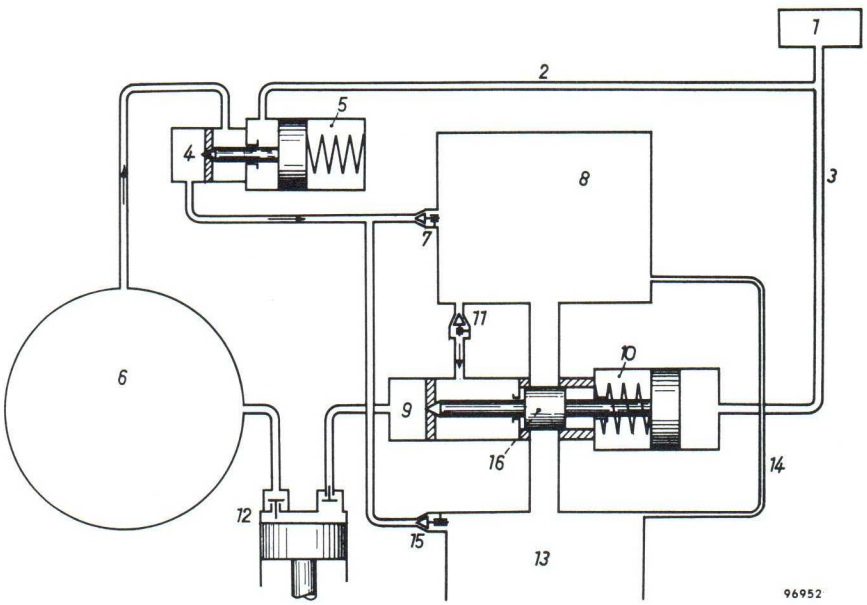


Fig. 19. Hydraulic mechanism for regulating the output of the hot-gas engine. By altering the oil pressure in pipes *2* and *3*, governor *1* controls the operation of valves *4* and *9*. If the engine speed falls below the pre-set value, gas from the tank *6* flows via valves *4* and *7* into the cylinder *8*; if the speed setting is exceeded, gas passes from *8* via valves *11* and *9* and compressor *12* back to *6*. The function of the slide throttle *16*, which connects the cylinder with the buffer space *13*, is to ensure a rapid response when the load is reduced suddenly.

The auxiliary compressor gives a gas-tank pressure rather higher than the minimum pressure in the cylinder under conditions of full load. For increasing the engine output, therefore, it is possible to inject additional gas via valve *7* until the full power is developed. This adjustment takes place very rapidly. When it is a question of reducing the output, for example from full power, the auxiliary compressor initially performs no function

because gas from the cylinder returns to the tank via valves *11* and *9* under the excess pressure in the cylinder; this continues until the maximum pressure in the cylinder has dropped to the tank-pressure value. This too happens very quickly. If, however, the cylinder pressure is to be lowered still further, then the auxiliary compressor must function, and thereafter it is this that determines the speed of response. If a *sudden* decrease in the load (e.g. sudden removal of the whole braking torque) is not to result in racing of the engine, the compressor required would have to be very large. This difficulty has been overcome by incorporating an additional regulating mechanism. Besides opening and closing valve *9*, regulator *10* actuates a slide throttle *16* which offers the gas a direct path between cylinder and buffer chamber; the escape of gas will be large or small in accordance with the magnitude of the change in the load. It is clear that this gas escape path is equivalent to a leak past the power piston: it causes a certain decrease of the useful power delivered by the engine (which is now indeed associated with a decrease in efficiency; whence the term "loss-regulation", coined for this control). Because the action of the slide throttle has an almost instantaneous effect, a very small auxiliary compressor is adequate [9]).

In consequence of the direct connection between cylinder and buffer space when the power of the engine is being reduced, the buffer pressure likewise adjusts itself almost immediately to the correct value. This quick readjustment is highly desirable, since it limits unnecessary asymmetrical thrust on the drive mechanism. A similar purpose is served by valve *15*. It prevents the buffer pressure lagging too much when output is suddenly *raised* to a much higher value: if the value from which it is raised is so low that the mean pressure in the cylinder is less than the tank pressure, gas from the tank will pass directly into the buffer chamber via valves *4* and *15*.

Engine data and test results

We now give the more important data relating to a single-cylinder hot-gas engine designed and built in Philips Research Laboratories, Eindhoven, along the lines described above.

Cylinder bore	88 mm
Power-piston stroke	60 mm
Nominal engine speed	1500 r.p.m.
Mean piston velocity at 1500 r.p.m.	3 m/sec
Highest permissible value of p_{max}	140 kg/cm² (≈140 atm)
Mean pressure $\bar{p}$ corresponding to the above value of p_{max}	105 kg/cm² (≈105 atm)
Compression ratio p_{max}/p_{min}	2.0
Nominal cooling-water temperature	15 °C (≈60 °F)
Nominal heater temperature	700 °C (≈1290 °F)
Working fluid	Hydrogen
Fuel	Light fuel oil

We have already alluded to the advantages of hydrogen as the working fluid; the flow losses are relatively small and good heat transfer is attainable, the efficiency being thereby greatly enhanced. In addition, the speed of response of the engine to changes in the load or controls, which is ultimately determined by the effective cross-sectional areas of ducts and valves in the regulating system, is improved by using a low-density gas.

The results of measurements of the brake horsepower and specific fuel consumption of the engine will now be dealt with. The efficiencies derived from these figures relate to the engine without auxiliary devices [10]) and are based on a fuel of calorific value (low heat value) 10000 kcal/kg (≈18000 B.T.U./lb). The measurements were made at four different pressure levels corresponding to p_{max} values of 50, 80, 110 and 140 kg/cm², and at six different engine speeds for each of these pressure levels, namely at 250, 500, 1000, 1500, 2000 and 2500 r.p.m. *Fig. 20* is a general view of the equipment used for the measurements. In *fig. 21* the measured power and overall efficiency are plotted as functions of the engine speed in r.p.m. It will be noted that a power of up to 40 H.P. can be developed and that the overall efficiency has a maximum value of 38%. In addition, the measured torque is plotted

[9]) In a multi-cylinder engine we arrange for the leakage to take place in a rather different manner: the slide throttle opens a gas path from one cylinder to the next. This is equally effective because neighbouring cylinders must in any case differ in phase. Thus the one throttle provides instantaneous "loss regulation" for two cylinders.

[10]) Apart from the usual pumps and fans for forced oil feed and cooling, the auxiliaries comprise the regulating compressor already mentioned and a blower for the burner. Of all these the latter absorbs the most power, viz. 1 to 2% of the engine output.

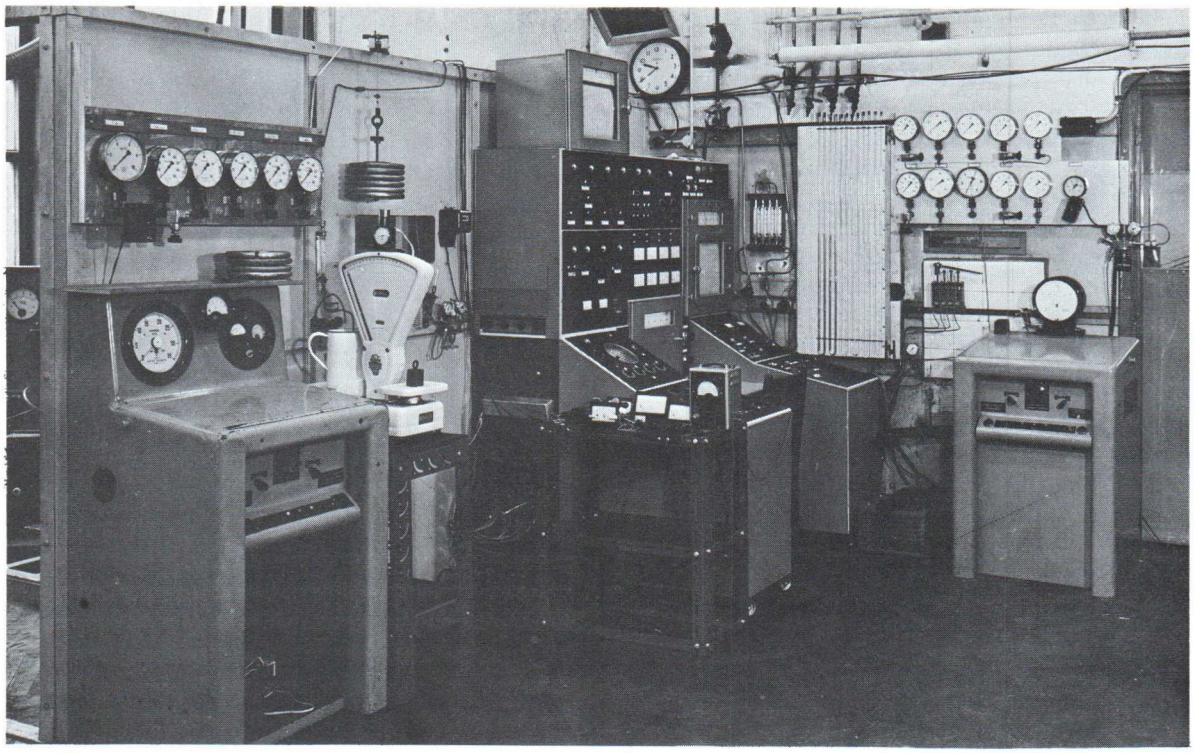


Fig. 20. Equipment used for measuring the torque, shaft power and efficiency of the hot-gas engine under widely differing conditions.

in *fig. 22* for the highest working pressure, for which $p_{max} = 140$ kg/cm². The unusual flexibility of the engine is apparent from the small variation in the torque (11.5 to 15 kg.m or 80 to 110 lb.ft) at greatly different speeds. Equally remarkable is the fact that the overall efficiency varies only slightly over a considerable range of speeds and pressures.

It is the usual practice to display the results of

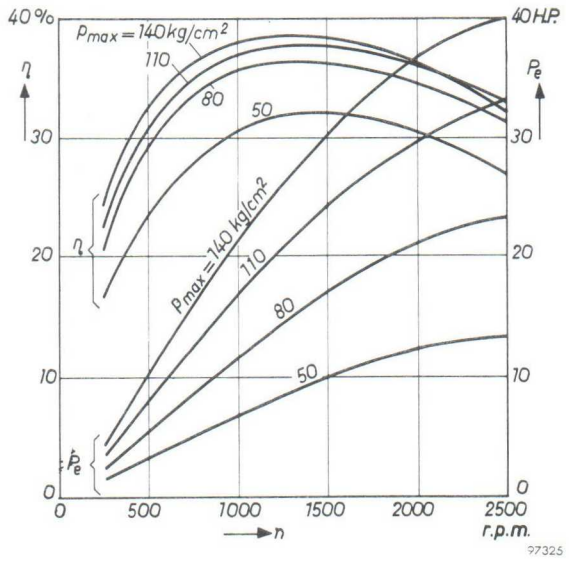


Fig. 21. Measured shaft power P_e and efficiency η of the experimental version of the new Philips hot-gas engine with rhombic drive, plotted as functions of the engine speed for various values of p_{max}.

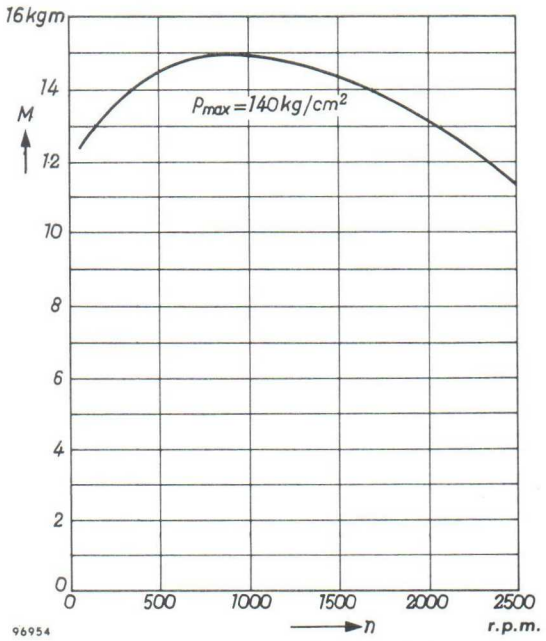


Fig. 22. Measured torque M as a function of the engine speed, for the highest admissible pressure level.

engine tests in a graph such as that of *fig. 23*. Here the coordinates are the engine speed and p_e, the "mean effective pressure" (m.e.p.). This latter is not the mean gas pressure actually occurring in the engine but merely a formal ratio, being defined as the work done by the shaft per revolution divided by the volume swept by the piston; hence the product of p_e and n, the speed in r.p.m., is the specific power output. For a given engine, therefore, every point in the p_e,n diagram corresponds to a definite value of the power developed P_e, and the curves of constant P_e are hyperbolae, in view of the fact that

$$p_e . n \propto P_e . \quad . \quad . \quad . \quad . \quad (14)$$

In such a diagram contours of constant specific fuel consumption (grams of fuel per horsepower-hour, output measured at the shaft) are now drawn.

Fig. 23*a* shows these curves for the experimental engine. The upper boundary to this family of curves, indicated by a broken line, is due to the ceiling on the gas pressure in the cylinder; in the present case the highest permissible pressure is 140 kg/cm², and consequently P_e for a given engine speed cannot exceed the value corresponding to that pressure.

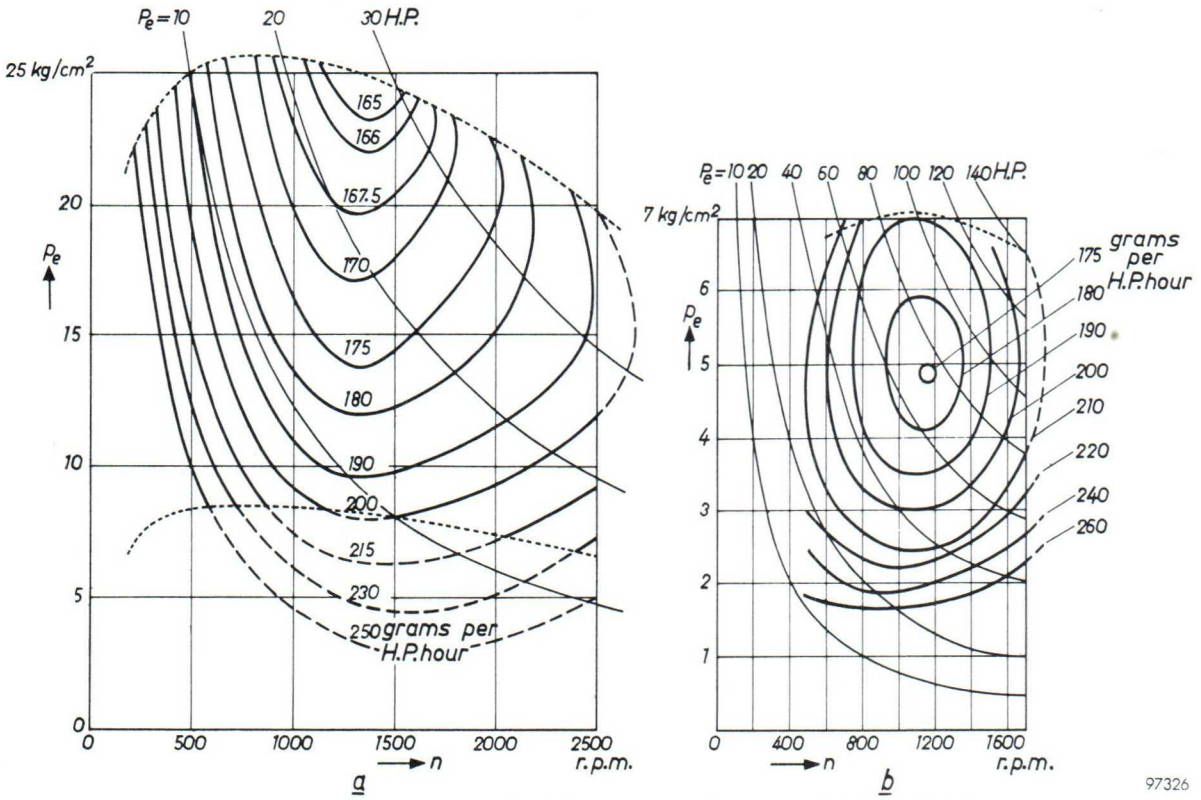


Fig. 23. *a*) Curves of constant specific fuel consumption (in grams of fuel per H.P. hour; for lb per H.P. hour divide by 453) are drawn here in the p_e,n diagram for the hot-gas engine. p_e is the mean effective pressure (m.e.p.) defined as the useful work done by the shaft per revolution divided by the volume swept out by the piston. Accordingly, $p_e . n \propto P_e$, which means that, for a given engine, every point in the diagram corresponds to a definite value of the power P_e and that the curves of constant P_e (lightly drawn in the diagram) are hyperbolae. The broken line representing the upper limit to the measurements corresponds to the pressure ceiling of $p_{max} = 140$ kg/cm² (this curve is derived from the P_e curve for this pressure, given in fig. 21, simply by dividing each ordinate by the corresponding abscissa). The broken line representing the lower limit corresponds to the lowest pressure level at which measurements were made, corresponding to $p_{max} = 50$ kg/cm².
b) p_e,n diagram for a diesel engine (after J. M. Kuijper, Dieselmotoren, Stam, Haarlem 1954, page 81, fig. 52).

(The lower boundary to the curves is not an actual restriction on the engine; it corresponds to the lowest pressure at which measurements were carried out, namely $p_{max} = 50$ kg/cm². The broken-line portions of the curves of constant specific fuel consumption, which lie beneath the lower limit, have been extrapolated.)

The striking thing about the diagram is that the curves do not form complete loops, as they do in the corresponding diagram for a diesel engine, for

example — see fig. 23*b*. This means that the hot-gas engine achieves its highest efficiency at an m.e.p. which lies close to the highest permissible value. This highest permissible value is here not determined by a "smoke-limit" as in a diesel engine, but simply by the strength of the engine casing. In making such a comparison, however, it should be emphasized that *m.e.p. values alone are quite useless* (and so is the specific power output) *as a yardstick for comparing the hot-gas engine with internal-combustion types* from the standpoint of *bulk*. The fundamental differences are too great. A better yardstick is the specific weight, that is, the weight per horsepower.

All the above values were measured with the nominal heater and cooler temperatures given earlier. The effect of changes in these temperatures on output and efficiency may be seen in *figs. 24a* and *b*. The curves given are appropriate to an engine speed of 1500 r.p.m. and a p_{max} of 140 kg/cm².

Finally, *figs. 25a* and *b* are graphs showing the heat balance of the engine plotted as a function of p_e (at $n = 1500$ r.p.m.) and as a function of n (at $p_{max} = 140$ kg/cm²) respectively. The narrowness of the band representing flue losses shows the effectiveness of the air pre-heater: over a large range of n and p_e values the burner has an efficiency of

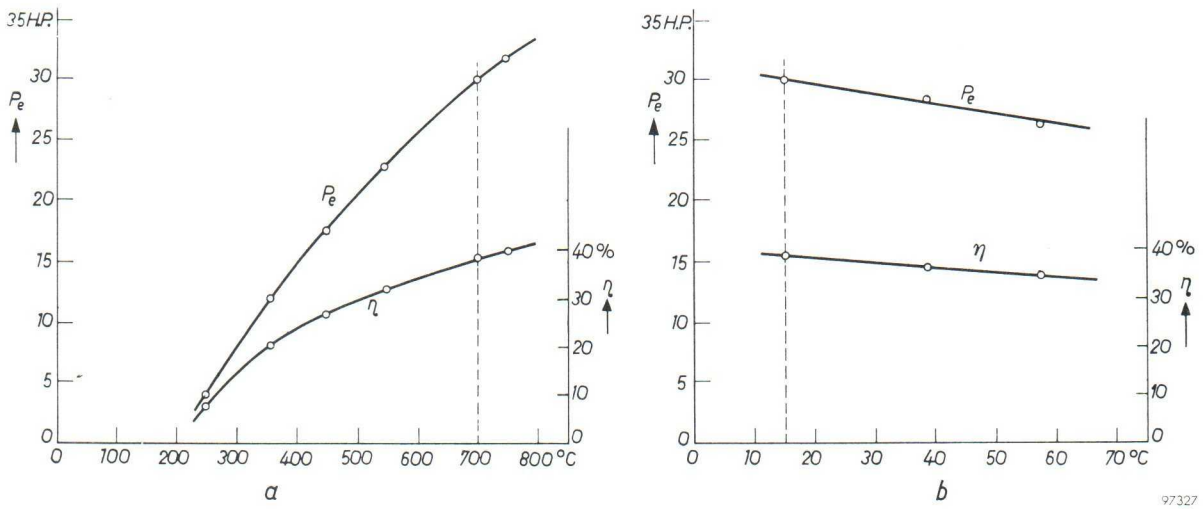


Fig. 24. Shaft power and efficiency of the experimental hot-gas engine as functions of (*a*) the heater temperature and (*b*) the cooler temperature. The curves are those appropriate to an engine speed of 1500 r.p.m. and a p_{max} of 140 kg/cm².

Although few special measures were taken to keep the weight of the present experimental engine low, the specific weight is in fact only 5 kg per H.P. (11 lb per H.P.).

The following are the extreme values found for various quantities.

Maximum shaft output P_e *)	40 H.P.
Maximum mean effective pressure p_e **) . . .	26 kg/cm² (~ 26 atm)
Maximum torque **) .	15 kg.m (~110 lb.ft)
Maximum specific power output *) (H.P. per litre swept volume of piston)	120 H.P./litre
Minimum specific fuel consumption ***) . .	165 grams/H.P.hour (0.36 lb/H.P.hour)

*) Occurs at $n = 2500$ r.p.m. and $p_{max} = 140$ kg/cm².
**) Occurs at $n =$ approx. 800 r.p.m. and $p_{max} = 140$ kg/cm².
***) Occurs at $n =$ approx. 1350 r.p.m. and $p_{max} = 140$ kg/cm².

over 90%. Chief amongst what are shown as "other losses" are those due to mechanical friction. The item may also include losses due to incomplete fuel combustion ("non-detectable" flue losses). One cannot expect these "other losses" to be specified with any great accuracy, but from the good reproducibility of the results, we may reasonably conclude that the mechanical efficiency of this rhombic-drive engine is high.

Concluding remarks

To end with, we may recall here the essential features of the hot-gas engine in general (cf. [2])):

a) Liquid and gaseous fuels of greatly differing properties can be used, little modification of the engine being necessary.

b) Lubricating-oil consumption is low and there is little wear, owing to the absence of corrosive substances and to the relatively low piston velocity.

c) Noise level is very low.

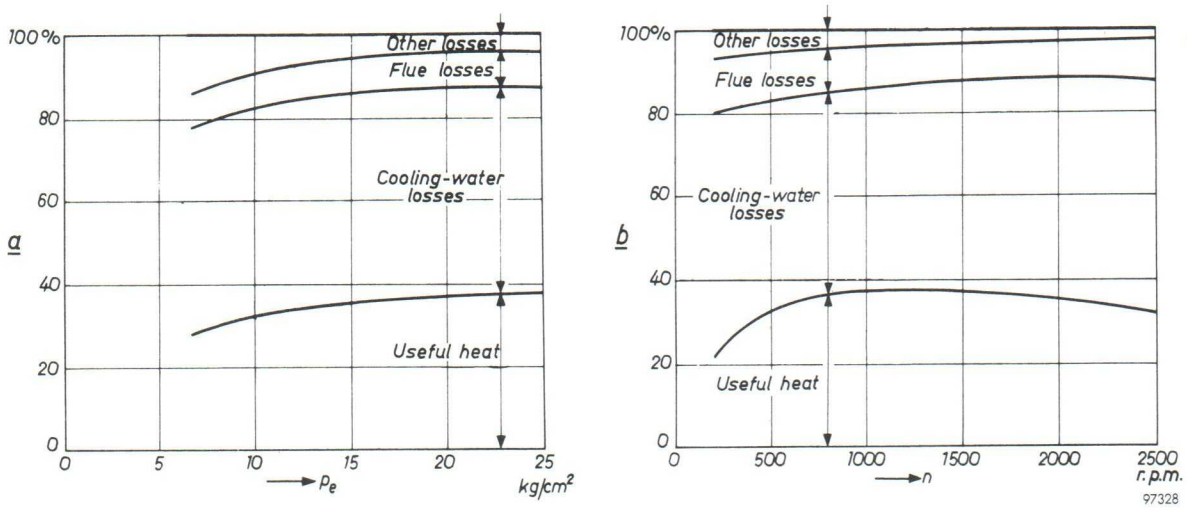


Fig. 25. Heat balance for the experimental hot-gas engine (a) as a function of p_e at $n =$ 1500 r.p.m., and (b) as a function of n at $p_{max} = 140$ kg/cm².

d) The torque varies only slightly over a large range of engine speeds.

e) The torque variation as a function of the crank angle is relatively small. In this respect a single-cylinder hot-gas engine is comparable with a four-cylinder internal-combustion engine.

In addition to these general features, the experimental engine described here converts heat into mechanical energy with a high efficiency, as the results of our measurements illustrate. Moreover it has been shown that good values of the weight per horsepower are attainable. These favourable qualities are mainly to be attributed to three features of the design: the use of a displacer piston with its attendent aerodynamic and thermodynamic advantages, the use of hydrogen as the working fluid, and the use of high working pressures. A further advantage of the newly developed engine is its freedom from vibration, in consequence of the complete dynamic balancing of the inertial forces of moving parts.

Summary. After a brief recapitulation of the hot-gas cycle as applied to an engine provided with a displacer piston, certain new design features of the Philips hot-gas engine are described. An account is given of a new kind of drive mechanism in which the required power and displacer piston movements are obtained by means of twin crank and con-rod mechanisms offset from the engine axis; the crankshafts are coupled by two gear-wheels, and rotate in opposite senses. Where such a system is used, the buffer pressure that is desirable under the power piston can be contained in a relatively small cylindrical buffer chamber — there is no need to pressurize the whole crankcase. The new drive mechanism also permits complete dynamic balancing of the inertial forces of moving parts, even in a single-cylinder engine. Other innovations concern the design of the regenerator compartment, the cooler and the heater: an entirely new approach has cleared up the problem of thermal stress in these components, whose weight can be kept quite low despite the high gas pressure. The pre-heater has also been redesigned, with the result that burner efficiencies of up to 90% are now reached. In addition, a fast-acting system has been developed for regulating the engine output, this being controlled by altering the mean working pressure of the gas in the cylinder.

These various features have been applied in the design and construction of an experimental single-cylinder engine of about 40 H.P., containing hydrogen (helium could also be used) as the working fluid and burning light fuel oil. The highest pressure permissible in the cylinder is 140 kg/cm². The nominal speed of the engine is 1500 r.p.m. Extensive tests on this engine, the results of which are given in the article, show that with the heater temperature at 700 °C it attains an overall efficiency of up to about 38% (engine alone, without auxiliaries). The maximum value of the specific power was found to be 120 H.P. per litre swept volume and the best specific fuel consumption 165 grams (0.36 lb) per H.P. hour. The specific weight of the engine is 5 kg (11 lb) per H.P.

MOTION-PICTURE PROJECTION WITH A PULSED LIGHT SOURCE

by P. HOEKSTRA and C. MEYER. 778.554.4:621.327.534.3

Present-day motion-picture projectors are still almost invariably equipped with a continuously burning light-source — usually a carbon arc — the light from which is interrupted twice per frame period by a rotary shutter in order to obtain a stationary picture free from flicker. This article describes an illumination system in which the light source is a super-high-pressure mercury-vapour lamp operated by current pulses, and which dispenses with the need for a shutter.

As long ago as 1938 Philips introduced a motion-picture projector in which the light source was a discharge lamp, viz. a water-cooled super-high-pressure mercury-vapour lamp [1]). This lamp, which was fed by direct current and consumed 1000 watts (2 A, 500 V), replaced the conventional carbon arc. Both types of light source combine small dimensions of the radiant element with great brightness, a characteristic which is essential for application in film projectors [2]). All cinema-goers know that satisfactory projection can be achieved with the carbon arc. Nevertheless, this light source has certain serious drawbacks. In the first place, it requires constant attention from the projectionist [3]). The arc must be allowed to burn for several minutes before use, and the carbon electrodes must be long enough to last for the duration of the film. The incandescent ball of gas in the crater of the positive electrode, which is the main source of light emission, shows a tendency to shift its position with respect to the mirror focus. If the projectionist does not intervene, the result is a noticeable diminution of luminous flux on the screen and a change in the colour of the light. Furthermore, an exhaust system is required for removing the combustion products, but even this cannot prevent deposits forming at undesired places. The carbon arc also causes considerable heating of the film, at least of the picture areas; the edges remain cold. As a result the film buckles slightly, an effect which is accentuated by the fact that it is chiefly the photographic emulsion that absorbs the heat, and not so much the transparent base. This buckling adversely affects the sharpness of the picture — a drawback which is more noticeable nowadays than formerly. There are two reasons for this. Firstly, present-day films are generally made of non-inflammable material, and this buckles more than the material previously used. Secondly, there is a tendency to use faster projection lenses, which have less depth of focus. The fact that the film buckles is therefore more serious and it is more difficult to obtain a picture of uniform definition on the screen.

Compared with the carbon arc the super-pressure mercury lamp has appreciable advantages. It requires much less attention from the projectionist; in fact, without much trouble its operation can be made largely automatic. With the present shortage of skilled personnel, this is an attractive feature. A discharge lamp of this kind produces no unwanted deposits, thus dispensing with the need for an exhaust system, and it heats the film much less then a carbon arc with the same luminous flux, for one reason because the lamp produces much less infrared radiation. The lamp house, too, is much smaller than that for a carbon arc. Water-cooling is required, but this complication is more than offset by the advantages mentioned. In any case, the

[1]) Described by G. Heller in Philips tech. Rev. **4**, 2-9, 1939.
[2]) See e.g. T. J. J. A. Manders, Incandescent lamps for film projection, Philips tech Rev. **8**, 72-81, 1946, and also the article cited in footnote [1]).
[3]) See, for example, Mitchell's manual of practical projection, International Projectionist Publishing Co., New York 1956, pp. 112-117 and pp. 177-206.

use of high amperage carbon arcs also calls for water-cooling of the film gate. Still higher currents require that the positive carbon contact, too, be cooled.

The fact that the mercury lamp was unsuccessful in its first attempt to rival the carbon arc was mainly due to the deficiencies at that time in its luminous flux, in its rendering of red colours and in its reliability. The life of a mercury lamp ends fairly suddenly, and just when it will end cannot be predicted with any certainty. Consequently the lamp could not be changed in good time and its possible failure during a performance was something that had to be accepted. With the projector of 1938 it took some tens of seconds before the stand-by lamp came into operation. In the following we shall describe a new film illumination system using a super-pressure mercury lamp which overcomes these drawbacks whilst preserving the favourable features. The new system in fact possesses certain important additional advantages, including a very flicker-free picture and an improved efficiency. Some particulars of the lamp itself will also be mentioned. The illumination system can be incorporated into a projector of a new design, discussed elsewhere in this issue [4]). A photograph of this projector, type FP 20 S, equipped with the new illumination system, is shown in *fig. 1*.

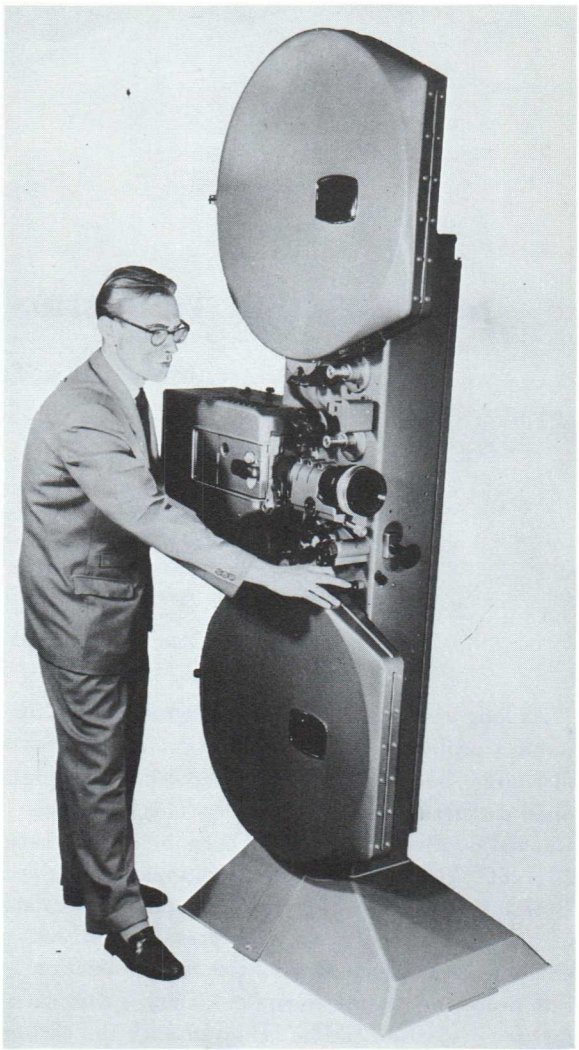


Fig. 1. The new Philips motion-picture projector, type FP 20 S, provided with the new illumination system which is the subject of this article. The projector itself is discussed elsewhere in this issue [4]).

The illumination system

As mentioned, the illumination system contains a water-cooled super-pressure mercury lamp as light source. This discharge lamp consists of a small tube of fused silica (length 8 cm, max. diameter 5.6 mm) in which a linear discharge is maintained between the electrodes (spacing 17 mm). The light output of the lamp increases more than proportionally with the power consumed, since the efficiency of high-pressure mercury lamps increases with the power supplied per unit length of discharge [5]). Higher loading improves the colour rendering at the same time, the reason being that the share of the continuum in the emitted light increases with respect to that of the line spectrum. Unfortunately the life of the lamp is thereby shortened, mainly because of the higher temperature of the fused-silica wall. In the projector brought out in 1938 this set a limit to the quantity and the quality of the light output. At the time that this projector was developed an attempt was already made to solve the heat problem by loading the lamp intermittently instead of continously. In motion-picture projection this is an attractive proposition in as much as the illumination of the film must in any case be interrupted twice during each frame period: once in order to pull down the film from one frame to the next without this being visible on the screen, and a second time in order to increase the number of light pulses per second sufficiently to prevent the picture from flickering [6]). The energy consumed by the lamp during the interruptions yields no useful light. If the lamp is loaded only when the generated light is usefully employed, the instantaneous load can be raised whilst maintaining the same

[4]) See the following article in this issue: J. J. Kotte, A motion-picture projector of simplified design, Philips tech. Rev. **21**, 83-77, 1959/60.

[5]) See e.g. W. Elenbaas, Fifty years of the high-pressure mercury vapour lamp, Philips tech. Rev. **18**, 167-172, 1956/57.

[6]) For a detailed discussion of this phenomenon, see: J. Haantjes and F. W. de Vrijer, Flicker in television pictures, Philips tech. Rev. **13**, 55-60, 1951/52.

mean load. The result is more useful light, and moreover the changed spectral distribution of the light provides better colour rendering. The luminous intensity of a mercury lamp follows current variations virtually without inertia. When fed with current pulses, then, the lamp flashes on and off synchronously. This makes it possible to dispense with the rotating shutter which, in the conventional continuously-burning carbon arc (or other source), intercepts the light twice per frame.

Tests with a pulsed mercury lamp were initially disappointing. The life of the lamp was found to decrease sharply when the instantaneous load was raised, even though the mean load remained unchanged. Only after a great deal of development work on the lamp, and careful matching of the pulse generator to the lamp, was it possible to achieve an acceptable life with an instantaneous load high enough to ensure good colour rendering. The final version of the lamp represents a compromise between useful life on the one hand and light output (determined by the *mean* load) and colour rendering (determined by the peak amplitude of the current pulses) on the other. The choice fell on a mean load of 800 W with current pulses of about 15 A peak amplitude. The mean life is then 33 hours. The luminous flux incident on the screen is equal to that from a 60 A carbon arc, and is therefore adequate for the majority of cinemas; the colour rendering, too, is entirely satisfactory. The lamp has been given the type designation SPP 800 W.

Increased projection frequency

We have already mentioned as a particular advantage the very flicker-free picture obtained. This is due to the light source flashing *three* times per frame (*fig. 2*). The frequency is thus 72 flashes per second instead of the usual 48 (the number of frames per second remains unchanged of course). The sensitivity of the eye to flicker increases with the brightness of the picture, and moreover the periphery of the retina is more sensitive to it than the central part. The lack of flicker in the picture from the new projector is therefore particularly noticeable in the brighter projection of wide-screen pictures, such as "CinemaScope."

The higher frequency of 72 flashes per second is in principle also possible with a continuously burning light-source if a three-bladed shutter is used. The resultant loss of light, however, would be unacceptable. With a film-transport mechanism employing the usual four-slot Maltese cross, the shutter would then intercept 75% instead of 50% of the light output (see *fig. 3*).

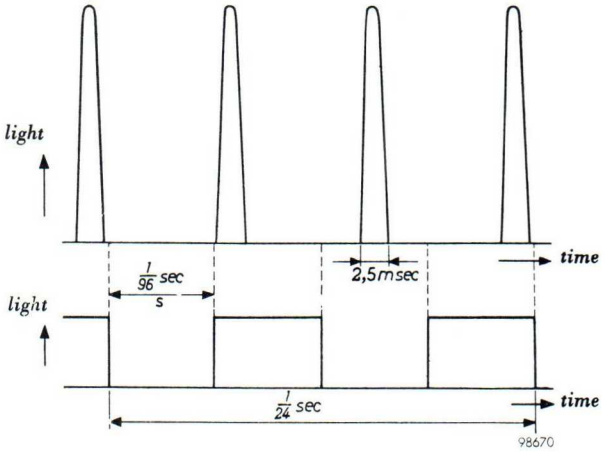


Fig. 2. Distribution over one frame period (1/24 sec) of the three light pulses from the mercury lamp type SPP 800 W (above), and (below) distribution of light periods over a frame period in a normal projection system with shutter. The film is transported during the interval denoted by *s*.

Lamp holder and turret

Fig. 4 shows a dismantled lamp holder together with a type SPP 800 W lamp. The components are so made that they mate together correctly without any adjustments during assembly. The round glass window in front of the lamp is a filter which absorbs ultraviolet radiation. The cooling water that flows over the lamp also passes over the filter, thus dissipating the heat generated in it. Of the little infra-red radiation emitted by the lamp the major part is absorbed by the coolant. The heating of film and film gate is therefore slight.

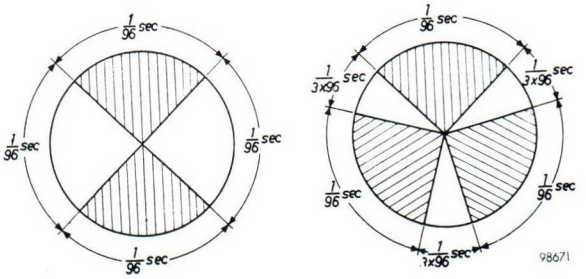


Fig. 3. The theoretical loss of light caused by a shutter that completes one revolution per frame period is equal to the fraction of 360° covered by the blades jointly. Since the fundamental frequency of the illumination cycle is mainly responsible for the nuisance of flicker, the shutter blades must be identical; otherwise the fundamental frequency of the illumination cycle would be equal to the frame frequency (24 per second). The film-transport mechanism is almost invariably a normal four-slot Maltese cross [7]). The frame-shift period must then be $\frac{1}{4}$ of the total period per frame, so that each shutter blade must cover 90°. With a two-bladed shutter (left) the theoretical light-loss is thus 50%, and with a three-bladed shutter (right) 75%.

[7]) The Maltese cross is described in J. J. Kotte, A professional cine projector for 16 mm film, Philips tech. Rev. **16**, 158-171, 1954/55, in particular fig. 7.

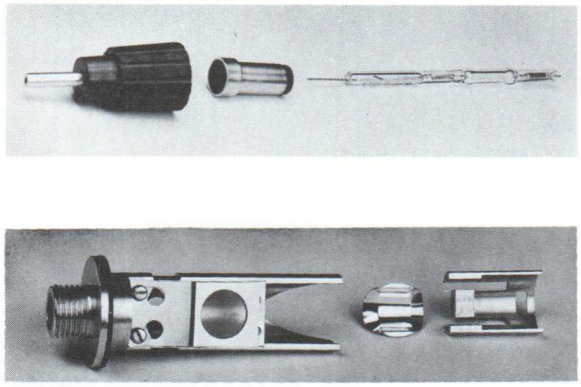


Fig. 4. Dismantled lamp holder. From left to right, above: the head of insulating material with silver contact peg, the centring bush (also acting as coolant seal) and the lamp. Below: the holder with filter for absorbing ultra-violet radiation, the cylindrical reflector, and its holder. The assembled lamp holder is shown in fig. 5.

Fig. 5 shows how two lamp holders are fitted in a "lamp turret". The lower holder, shown removed from the turret, contains the lamp normally in use, the upper holder contains the stand-by lamp. If the lamp in the lower holder becomes defective, the upper one automatically takes its place, the lamp turret pivoting about a horizontal spindle under its own weight. This is effected by the lower lamp being kept in its operating position by a pawl which is tripped when the lamp current cuts out. The cooling water flows in along the spindle around which the lamp turret rotates, and reaches the operative lamp holder via corresponding holes in the spindle and the turret. The cooling water is thus automatically switched to the stand-by lamp as soon as this is called into operation. The same applies to the electric supply. The stand-by lamp takes over so quickly as to be imperceptible to the audience watching the screen. The projectionist notices it, however, from the colour of a transparent sector on the front of the turret. If the projector is operating normally the green sector is visible; when the stand-by lamp enters into operation the red sector appears as a warning to the projectionist. He can now unscrew the lower lamp holder from the turret and replace the lamp whilst the projector remains in operation. Later, when there is a suitable opportunity, he can turn the lamp turret by hand back to its normal position. In this way the useful life of each lamp is exploited to the full, and the performance is insured against the unlikely event of a double lamp failure.

The above implies that, as far as the lamp is concerned, there is no limit to the lengths of film that can be used. This advantage promises to gain in importance in the future. In connection with the risk of fire, film spools and drums were formerly designed to accommodate films not longer than 600 metres. Now that inflammable films are no longer commonly used and have in fact been prohibited in many countries, it is likely that longer lengths of film (up to 1800 metres) will come increasingly into use.

When a lamp holder and its lamp are loaded into the turret, the exposed end of the lamp (see lamp-holder in fig. 5) slides into a hole surrounded by an annular nozzle inside the turret. The electrode pin at this end then presses against a spring-loaded contact, which also acts as a valve shutting off the water supply to the nozzle when no lamp is mounted. The other lamp electrode pin is clamped inside the contact peg at the head of the lamp holder (fig. 4, upper photo). When the lamp holder is in the operating position, this peg lies on a contact consisting of a metal flat partly encapsulated in a block of insulating material. This block is fixed to the lamp house in which the lamp turret is incorporated (*fig. 6*).

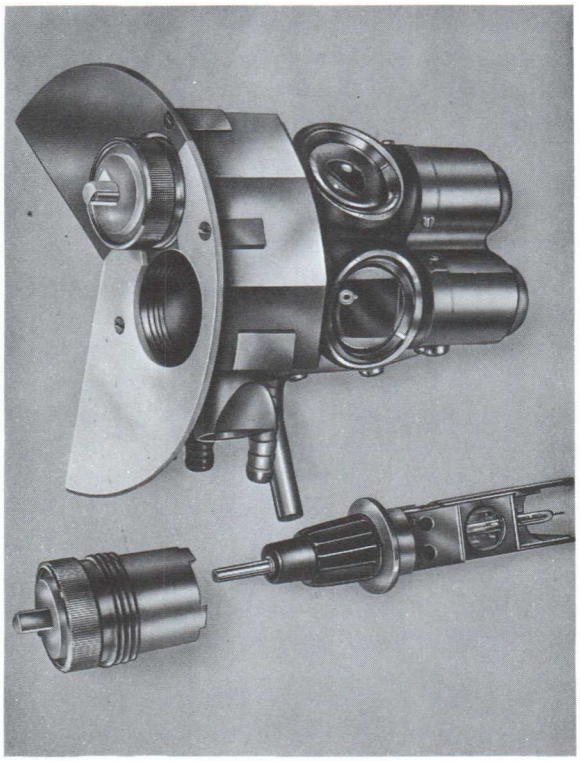


Fig. 5. Lamp turret with two lamp holders. The lower one, which is removed from the turret, contains the lamp normally in use; the upper lamp holder contains the stand-by lamp. The windows in front of the lamps also act as the first condenser lenses. The turret is provided with a red and a green sector, which indicate to the projectionist at a glance which lamp is in operation.

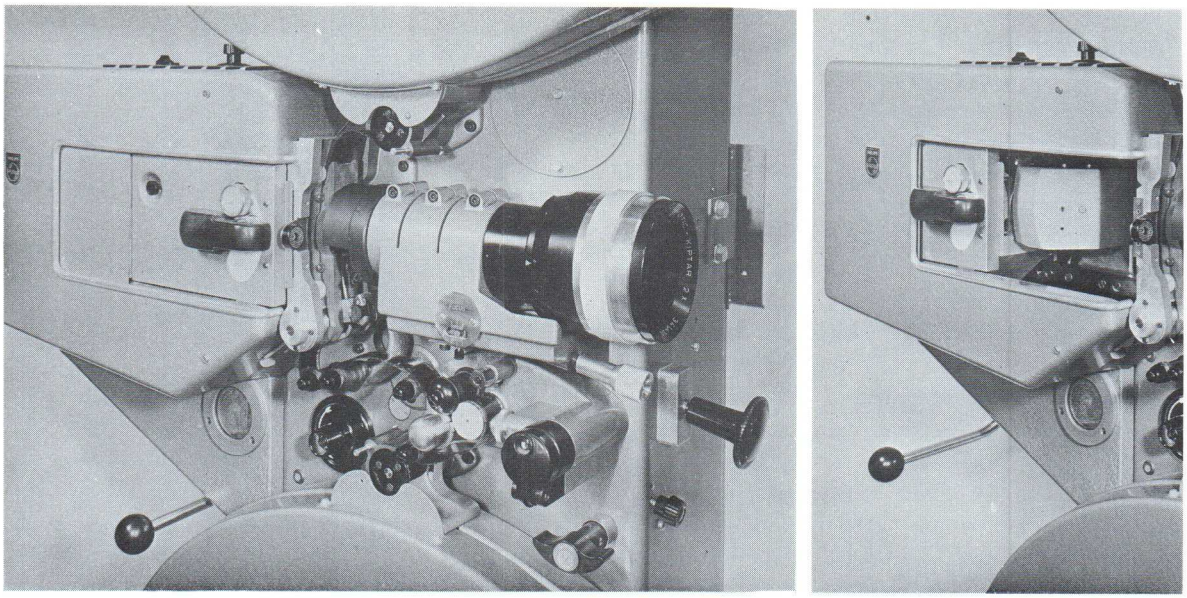


Fig. 6. Part of cine-projector fitted with the pulsed light-source illumination system. Above the block of black insulating material, which carries the positive contact, can be seen the cap which encloses the lamp holder with the stand-by lamp, in the turret. The photograph shows on the left the lamp house in the position for film projection. By means of the lever (bottom left) the lamp house can be moved backwards. The same lamp then serves for projecting lantern slides (right).

The optical system

The optical system is shown schematically in *fig. 7*. Seen from the light source, the film and aperture are located directly behind the condenser lenses. The effective diameter of the condenser is therefore only slightly larger than the diagonal of a film frame (26 mm in standard films). Since this implies a small condenser, the light source must be close to it in order to ensure that the condenser receives a sufficient part of the luminous flux. In this way a very compact assembly is obtained. The fact that it is technically feasible is due in the first place to the small dimensions of the lamp and in the second place to the simple way in which the ultraviolet and infra-red radiation is intercepted and rendered harmless. This radiation would otherwise be largely absorbed by the condenser and would damage it. Precisely because it is so small, the condenser is in any case severely heated merely by the high flux of visible light, of which it absorbs a very small fraction [8]).

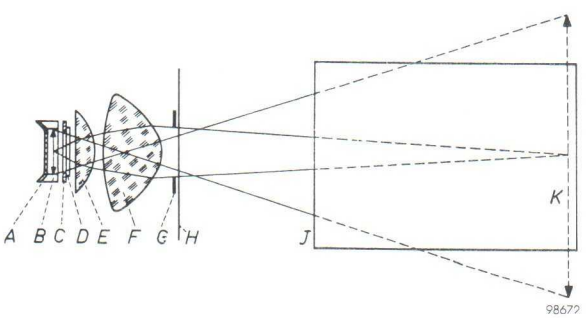


Fig. 7. The optical system. *A* cylindrical mirror, *B* linear light-source, *C* diaphragm, *D* ultraviolet-absorbent filter, *E* first condenser lens, *F* second condenser lens, *G* film aperture, *H* film, *J* projection lens, *K* image of light source projected by condenser.

The condenser system consists of two lenses, the first of which also acts as a water-tight window for the lamp turret. Behind the light source there is a cylindrical reflector of aluminium sheet (0.3 mm thick). Since this mirror provides more than 60% of the projected light, it is essential that its reflective power should not be diminished. For this reason a new reflector is supplied with every new lamp.

In a projection system as in fig. 7 the condenser is always arranged to form an image of the light source at the position of the projection lens. It is not the image formation itself that is important (the image is usually unsharp) but the fact that a pencil of light emanating from an arbitrary point of the light source has a small cross-section *at* the projection lens. This being so, if the central ray of

[8]) With conventional light-sources a different and less compact system is employed for film projection. The two systems are compared in the article mentioned in footnote [1]) and more recently in: P. M. van Alphen and M. Bierman, A mirror condenser lamp for 8 mm projectors, Philips tech. Rev. **19**, 233-235, 1957/58.

a pencil from any point of the source passes through the projection lens, all or almost all of that pencil of light does so. The optical centre of the condenser falls virtually in the film and hence virtually in the focal point of the projection lens (the screen is a considerable distance away). Consequently, the relative aperture of the projection lens determines what length of the light source is effective in contributing to the projected light. In the pulse-operated projector the central rays emanating from the ends of the discharge make such large angles with the axis that the projection lens should have at least a relative aperture of f 1.4 in order to transmit them. A normal value for the relative aperture of a projection lens is f 2. If such a projection lens is used, the extremities of the light source do not contribute to the luminous flux on the screen. This is done deliberately, for it precludes that the luminous flux incident on the screen gradually diminishes with the age of the lamp, as a result of the darkening of the fused-silica wall (a re-crystallization or devitrification phenomenon) which begins at the extremities of the discharge.

To avoid needless heating of condenser and film gate, a diaphragm is interposed between the lamp and the condenser. This is matched to the projection lens and intercepts the unused light. This diaphragm also serves as holder for the ultraviolet filter; it can be seen in fig. 4.

Projection of lantern slides

The same illumination system is employed for the projection of lantern slides. The slide holder is located at the other side of the projector. In order to convey the light to that side, the lamp house is mounted on rails. By means of a lever (see fig. 6) the lamp house can be shifted backwards. A plane mirror is thereby moved into the path of the rays at an angle of 45° (right, fig. 6). (When the lamp house is returned into position for film projection, this mirror folds upwards again.) The light reaches the lantern slide via the lens E (*fig. 8*), a second 45° mirror F and the lens G, whence it is transmitted to the lantern slide projection lens. In *fig. 9* the lamp house is shown in an intermediate position, in which the mirror has not yet descended.

The lamp is also pulsed for the projection of lantern slides, in order to benefit from the improved colour rendering resulting from pulsed operation.

Luminous flux and efficiency; light distribution

With a standard film gate, but with no film in it, and using a projection lens of the Petzval type

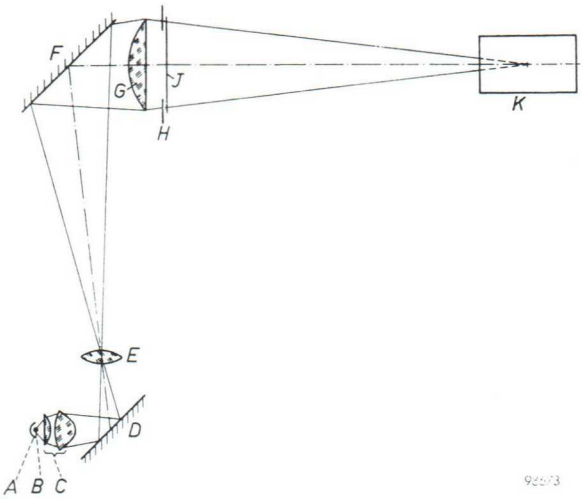


Fig. 8. Path of rays in the projection of lantern slides. A cylindrical mirror, B light source, C condenser, D plane mirror, E lens, F plane mirror, G lens, H mask, J lantern slide, K projection lens for lantern slides.

having a relative aperture of f 2, and a pulsed light-source, operated at 800 W, the system delivers a luminous flux on the screen of 4800 lumens. With a projection lens of the same type, but having a relative aperture of f 1.6, the luminous flux on the screen is as high as 6000 lumens [9]). The efficiency in these two cases is thus 6 and 7.5 lumens per watt, respectively. The earlier-mentioned 1938 projector, with a lens of relative aperture f 2 and

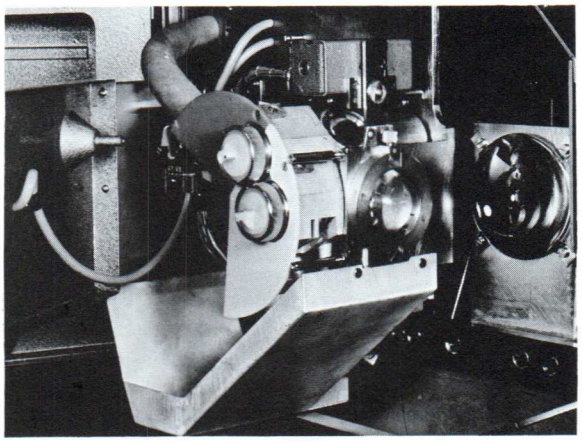


Fig. 9. Lamp-house opened to position between that for film projection and that for lantern slide projection. The 45° mirror has not yet descended into the path of the rays. Below can be seen the tray for collecting any leakage water. The cooling water is fed in through the thick hose visible above.

[9]) A projection lens system of the Petzval type consists of four lenses. Systems having a relative aperture of f 1.6 are usually more complicated and contain six or seven lenses; the gain due to the greater relative aperture is then partly lost as a result of higher losses in the lens system.

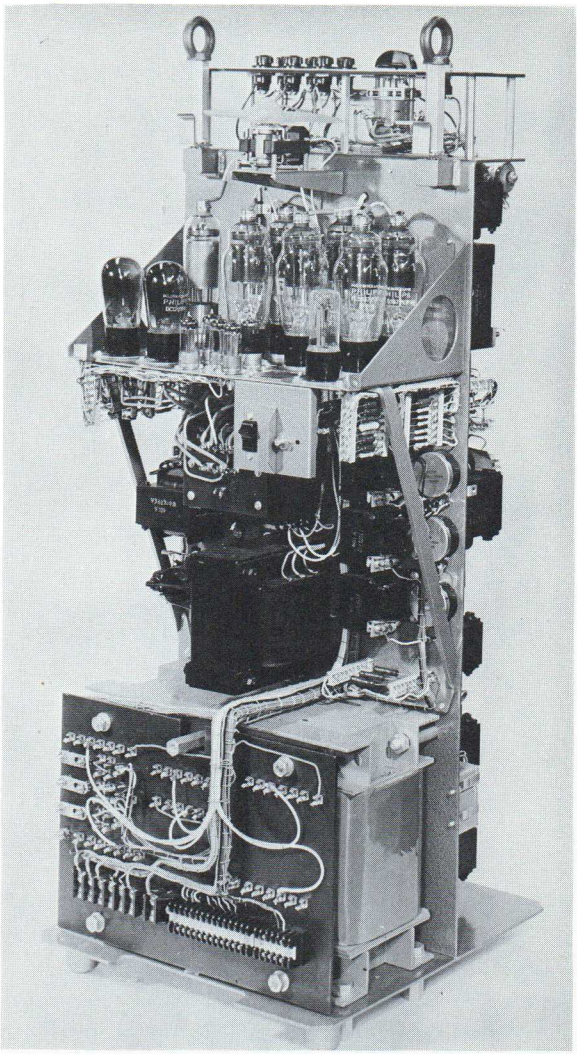


Fig. 10. Power pack (pulsator) for the pulsed-source illumination system, with protective hood removed.

a rotary shutter, gave a total flux of 2900 lumens [10]). The lamp consumed 1000 W, so that the efficiency was 2.9 lm/W. The more than double efficiency of the pulsed illumination system is, of course, mainly attributable to the fact that light is only produced when it is usefully employed. For carbon arcs of 45 and 60 amperes the flux on the screen is respectively 3500 and 5500 lumens. The efficiency of a carbon arc is approximately 2.5 lm/W.

According to the practice in projection technique, the fluxes mentioned are based on measurements of the luminous intensity in the middle of the screen. Since the luminous intensity always decreases from the middle towards the edges, the actual flux values are lower. In the case of the pulsed lamp the decrease towards the edges amounts at the most to 10% in the horizontal direction and to between 15 and 20% in the vertical direction. (The smaller horizontal decrease is due to the elongated form of the horizontally positioned lamp.) This is appreciably less than in the case of the carbon arc, where the decrease for the same flux values is 25% horizontally and about 20% vertically [11]). For the same values of luminous flux, as given according to the normal practice, the true luminous flux of the pulsed system here described is therefore actually greater than that of a carbon arc.

Power pack (pulsator)

The power pack which supplies the current pulses for the lamp (the "pulsator") is shown in *fig. 10*. In the simplified block diagram, given in *fig. 11*, *A* is a six-phase rectifier which charges the capacitor *C* to a voltage higher than the operating voltage of the lamp. At the moments when the lamp is required to flash, a voltage pulse is applied to the grid of thyratron *T*, making the latter conduct. The capacitor *C* thereupon discharges through the lamp.

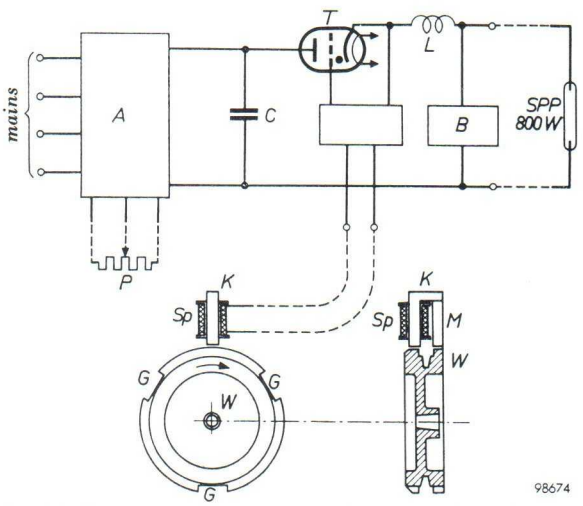


Fig. 11. Simplified block diagram of power-supply system for the pulsed super-pressure mercury lamp, type SPP 800 W. *A* six-phase thyratron rectifier. *P* potentiometer for varying the voltage on the thyratron grids. *C* capacitor which supplies the current pulses. *T* thyratron to whose grid the synchronizing pulses are supplied which are generated in the coil *Sp*. This coil is wound on the soft-iron core *K* and is located in the projector. *W* flywheel of Maltese cross. The grooves *G* in *W* produce in *Sp* three voltage pulses per revolution of the flywheel. *L* inductance which gives the condenser discharge the nature of a damped oscillation, and which together with *C* governs the flash duration. *B* holding-current rectifier; this keeps a small current flowing through the lamp during the intervals when the lamp emits no light.

[10]) In the article cited in footnote [1]) a figure of 2500 lumens is given. This referred to a lens system that was not provided with the low-reflection coated lenses now commonly employed.

[11]) With a carbon arc the decrease towards the edges can certainly be reduced, but only at the expense of the illumination level in the middle of the screen. Some decrease is desirable, as otherwise contrast effects would make the edges of the screen appear brighter than the middle.

The inductance L then tends to produce an oscillation in the circuit. At the moment, however, that the current tries to change direction, the thyratron becomes non-conducting and the lamp is extinguished. The rectifier charges up the capacitor, and upon the arrival of the next voltage pulse on the grid of T the lamp flashes again. At the same time the lamp is fed by the "holding current" rectifier B. This ensures that in the period between the flashes a small current continues to flow through the lamp, thereby keeping the gas sufficiently ionized to ensure that it flashes readily upon the arrival of the succeeding pulse. (If the lamp is cold, the gas pressure is low and so therefore is the ignition potential; switching on the lamp when no holding current is yet flowing therefore presents no difficulty.) The holding current is very low and produces no light of any significance.

The duration of the flash depends on the period of the oscillation produced in the circuit, and this period can be regulated by varying L or C, or both.

Rectifier A operates with six thyratrons. The voltage delivered by the rectifier, and to which C is charged up, can be adjusted by means of a variable bias on the grids of these thyratrons. In this way it is possible to control the amplitude of the current pulses. The variation is effected with potentiometer P (fig. 11), which can either be mounted on the projector or placed at any other desired position, e.g. in the cinema auditorium. Flash duration and peak current together determine the energy per flash and also — at a given number of flashes per second — the average lamp-load. If this load varies from 600 to 800 W, the power pack consumes from 1.3 to 1.6 kW. A comparable carbon-arc installation consumes approximately 3 kW.

Synchronization and protection

The three light flashes per frame must, of course occur outside the pull-down period. The synchronizing pulses for the grid of the thyratron T are therefore derived from the flywheel of the Maltese cross in the projector. This steel flywheel, in which three grooves G are milled (fig. 11) forms part of a magnetic circuit containing, in addition, a permanent magnet M and a soft-iron core K around which a coil Sp is wound. Every time one of the grooves passes this assembly, the magnetic flux is sharply attenuated. The resultant voltage pulse in the coil Sp is applied to the grid of T. The exact moments at which the flashes occur can readily be adjusted by shifting the position of the magnet-core-coil assembly in relation to the flywheel. This can be done whilst the projector is operating.

Precautions are taken to ensure that the lamp, in the event of a fault, cannot burn continuously directly from the rectifier A, and also to prevent it operating only from the holding-current rectifier.

When the apparatus is switched over for lantern-slide projection, the projector is not running. The synchronizing pulses for the thyratron T are then derived directly from the mains. Pulses are then delivered with twice the mains frequency. On 50 c/s mains this means that there are 100 instead of 72 pulses per second as for film projection. To avoid overloading the lamp the capacitor voltage is then at the same time automatically reduced such that the average load remains the same.

Automatic change-over

A motion-picture film often consists of two or more reels, and for this reason most cinemas have two projectors installed. When one reel is at an end, the next reel, which is already loaded on the other projector, is switched on without interrupting the projection. A correct "change-over" calls for vigilance on the part of the projectionist. With a projector equipped with the new illumination system the change-over can easily be made automatic. For this purpose an automatic change-over device has been designed. It consists of a base plate on which several microswitches are mounted, and these are operated by cams on a common motor-driven shaft. When one reel is nearly at an end, a warning sign appears in the top right corner of the projection screen, which is the signal for the projectionist to start the motor for the other projector. To automatize the change-over a contact strip is affixed to the film, adjacent to the motor-starting sign, between the edge and the perforations. This contact strip actuates a relay which starts the idle projector turning and switches on at the same time the motor that drives the automatic change-over device. The cams then start to rotate, whereupon 1) picture and sound are switched over after 8 sec, 2) some seconds later the "run-off" projector is stopped and 3) the cam-drive motor is switched off. The automatic device is now ready for the next change-over cycle. This system relieves the projectionist of quite a number of routine operations.

The lamp

The D.C. lamp type SP 1000 W used in the 1938 projector was not the immediate starting point for the development of the SPP 800 W lamp used in the new illumination system. The SP 1000 W type contains a surplus of mercury, that is to say so much that, even when the lamp is burning, only a very

small fraction of the mercury is vaporized. This lamp has a number of drawbacks which become particularly apparent when used for film projection and which will be touched on presently [12]). A start was therefore made on the development of a so-called *dosed* D.C. lamp, that is to say a type containing so little mercury that practically all of it is vaporized when the lamp is burning (*fig. 12*). This work was subsequently diverted to the development of a dosed lamp for *pulsed* operation.

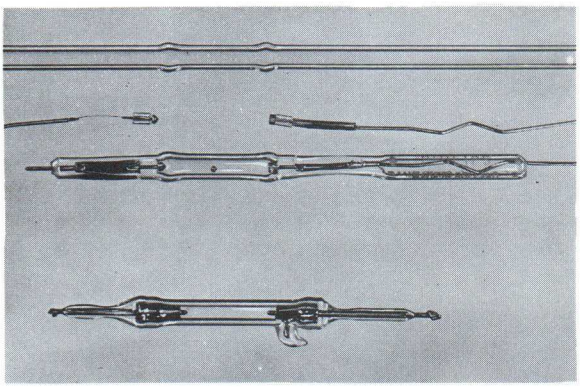


Fig. 12. The dosed super-pressure mercury lamp SPP 800 W for pulsed operation. The photograph shows from top to bottom: the pre-shaped tube of fused silica in which the electrode systems are sealed; the electrode assemblies; and the complete lamp. A fused silica capillary fits over each electrode stem. In the complete lamp the dosed quantity of mercury can be seen in the form of a droplet. Below, the surplus-mercury lamp type SP 1000 W. The mercury can be seen around the electrodes.

There are two chief disadvantages attaching to the lamp containing a surplus of mercury. Before igniting the lamp it is desirable to shake and tap it in order to distribute the mercury uniformly around the two electrodes which project some way into the discharge space. This may prove rather troublesome, particularly in a lamp that has already burnt for several hours. Furthermore, the lamp sometimes burns unsteadily owing to the abrupt rising or falling of the arc voltage. This phenomenon is probably connected with changes in the position of mercury surfaces.

The dosed lamp

A dosed lamp cannot be made by simply taking a lamp identical to the old surplus-mercury type and providing it with a dosed quantity of mercury. The electrodes would then project too far into the discharge space. The distance from the point of an electrode where the discharge is initiated — and where the heat is produced — to the seal would be so great, so that the temperature at this seal would be considerably below the temperature in the discharge. The latter temperature is limited by what the fused silica can stand, and the lowest temperature in the lamp (which determines the mercury vapour pressure) would therefore be lower than in the case when the surplus mercury *is* present. The mercury vapour pressure would therefore not reach the required high value. To produce a dosed lamp in which the mercury-vapour pressure rises just as high as in a surplus-mercury type, the end-wall must be located at approximately the position taken up by the mercury surfaces in the surplus-type lamp (see, fig. 12, below). This represents a problem, however. In the surplus-type lamp the electrodes are tungsten wires. Since these cannot be directly sealed in fused silica, the wires are first sealed in an intermediate glass which, in its turn, is sealed in the fused silica. This form of intermediate seal is not suitable for the dosed lamp because the spaces behind the ends of the electrodes cannot be made small enough, and moreover the intermediate glass seal cannot withstand the necessary high temperature. For this reason a molybdenum-foil seal was adopted, thereby eliminating the need for an intermediate glass. Very thin foils of molybdenum can quite readily be sealed in silica [13]). The foils used here are 10 to 12 microns thick. The lead-in wire is welded to one end of the foil and the tungsten wire electrode to the other end (fig. 12). However a foil seal is itself too weak to secure an electrode with sufficient mechanical strength. On each electrode wire, therefore, is a closely fitting capillary of fused silica. The whole electrode assemblies are then introduced into the fused silica lamp tube and sealed in. The Mo foils then fuse into the silica tube, and likewise the pieces of capillary. This leaves a tiny annular gap around each electrode wire, so narrow that the capilliary still firmly holds the wire.

Care must be taken that the temperature developed in the gaps round the electrodes is not below that of the wall near the discharge. With this in mind the diameters of the electrode wires are chosen so as to ensure that the lamp current generates the right amount of heat in the wires for this purpose. Another problem is the dissipation of heat. In the surplus-type lamp the mercury around the electrodes is responsible for considerable heat dissipation by conduction; in the dosed lamp the heat must be dissipated mainly by radiation. Each of the

[12]) The SP 1000 W lamp is well suited to other applications, for example in shipyards for apparatus in which the lamps are not switched on and off very frequently.

[13]) D. Gabor, D. R. P. 573 448, 1931. See also J. L. Ouweltjes, W. Elenbaas and K. R. Labberté, A new high-pressure mercury lamp with fluorescent bulb, Philips tech. Rev. **13**, 109-118, 1951/52, especially p. 115.

electrodes is therefore fitted at the end with a head to obtain a sufficiently large heat-radiating surface. The edges of the electrode heads are so close to the wall of the lamp that it is necessary to widen the bulb slightly at this position. Otherwise the bulb temperature here would become so high as to cause rapid devitrification of the silica. In fig. 12 it can be seen that the widening of the bulb at these positions is effected before the electrodes are sealed in.

As stated, the dosed lamp was developed with an eye on pulsed operation. DC pulses impose a heavier load on a discharge lamp than ordinary DC operation. The life of a lamp intended for DC operation is therefore almost invariably very much curtailed if the lamp is pulsed at the same mean power. One reason for this is the heat generated in the electrode seals. The heat generated at any given instant is, of course, proportional to the square of the current, and this means that, at the same mean power, the heat generated is greater the shorter the duration of the pulses. One cannot reverse the argument however, and conclude that a lamp which has a good useful life under pulsed operation will have a much longer life if operated on direct current. In order to obtain the most favourable conditions for DC operation various factors make it necessary to adopt, among other things, different dimensions for the electrodes.

One of the most important points in the development of the pulsed lamp was the control of manufacturing conditions. For motion-picture projection there is a constant demand for more light, and lamps for this purpose are always loaded to the utmost. This means that any manufacturing deviation of the lamp from the exact design almost invariably means a lamp of considerably shorter life. Mechanization of the manufacturing process was therefore just as necessary to obtain a uniform product as it was to lower the production costs. The average life of 33 hours mentioned at the beginning of this article applies under average conditions encountered in motion-picture projection, that is to say with the lamp being switched on or off every quarter of an hour (this period was determined by counting the number of occasions the lamp was actually switched on and off in an arbitrary cinema). In spite of all precautions, the spread in lifetimes is considerable. This presents no practical difficulties, however, because of the automatic change-over to the stand-by lamp.

At a mean load of 800 W the lamp gives a light output of about 40 000 lumens. The fact that the SPP 800 W lamp gives good colour rendering can really only be demonstrated by viewing a colour film projected with this light source. The improvement compared with the SP 1000 W is in fact much greater than might be expected from a comparison of the light-flux in spectral bands for the two types of lamp as shown in the *table*. (This spectral-band method was devised by Bouma [14]) who divided the whole visible region of wavelengths into eight sections, in a specified way, taking into account the properties of the eye.) It can be seen from the table that the SPP 800 W lamp, as compared with the SP 1000 W, shows appreciably less luminous output at the blue end of the spectrum, and a greater output at the red end.

Table. Relative luminous-flux of spectral bands for the lamps SPP 800 W and SP 1000 W. The visible spectrum is divided into eight bands after Bouma [14]). For comparison, the distributions are also given for daylight and for the high intensity carbon arc. The colours are indicated at the foot of the table.

Section	1	2	3	4	5	6	7	8
Wavelength range mμ	400-420	420-440	440-460	460-510	510-560	560-610	610-660	660-720
SPP 800 W	0.025	0.28	0.50	4.24	43.7	46.6	4.3	0.28
SP 1000 W	0.042	0.53	0.87	4.6	52.6	37.6	3.4	0.25
Daylight	0.025	0.26	0.91	11.1	40.8	36.2	9.9	0.73
Arc	0.050	0.27	0.97	10.2	43.7	33.2	10.6	0.94

436 495 566 589 627

violet blue green yellow orange red

[14]) See e.g. P. J. Bouma, Colour reproduction in the use of different sources of "white" light, Philips tech. Rev. **2**, 1-7, 1937.

Summary. The water-cooled super-pressure mercury lamp has many properties that make it more attractive as a light source for motion-picture projection than the conventional carbon-arc. In a normal projector, however, in which the shutter intercepts about 50% of the light, the light yield is rather on the low side for large theatres, whilst in colour projection less than justice is done to the red colours. This article describes an illumination system in which the mercury lamp is pulsed, thus producing light only when it can be usefully employed and rendering a shutter superfluous. At the same power consumption, this gives twice as much light on the screen as obtained from a continuously burning lamp, half of whose output is intercepted by a shutter. With a projection lens having a relative aperture of *f* 2, the luminous flux on the screen is 4800 lumens, and with a lens of relative aperture *f* 1.6, as much as 6000 lumens. The lamp consumes 800 W, so that the efficiency is respectively 6 and 7.5 lumens/watt on the screen, compared with 2.5 lumens per watt in the case of a carbon arc. Owing to the heavier instantaneous loading during the pulses the continuum in the spectrum is intensified with respect to the spectral lines, resulting in good colour rendering. Three light flashes of about 2.5 milliseconds duration are emitted per frame. This gives a strikingly flicker-free picture, even when the picture is very bright and wide. Apart from the illumination system proper, the power supply equipment is also discussed. The mercury lamp employed, type SPP 800 W, is a dosed type and was specially developed for pulsed operation in motion-picture projectors. Some particulars of its development are given.

WAVEGUIDE EQUIPMENT FOR 2 mm MICROWAVES

by C. W. van ES, M. GEVERS and F. C. de RONDE. 621.372.8

Summary. For measurements in the 2 mm wavelength region a comprehensive range of waveguide components has been developed, the principal of which are reviewed in this article. The 2 mm waves are produced by using a silicon diode to double the frequency of the 4 mm waves generated by a reflex klystron, type DX 151. The detector used is also a silicon diode. The matching is effected by means of waveguide plungers and a pivoting screw tuner (a variant of the sliding screw tuner). Use is made either of synchronous detection or of a selective indicator, the 2 mm waves being modulated in amplitude, e.g. by superposing an audio-frequency voltage of about 1 V on the frequency multiplier. Without a frequency multiplier the *PIN* modulator may be used, which is an electrically controlled attenuator consisting of a germanium wafer having a *P*, an *I* and an *N* region. The waveguides are generally rectangular in cross-section (0.83 × 1.66 mm, or 0.0325″ × 0.065″) and gold-plated inside and out. They are joined together by claw flanges, which allow particularly compact assemblies. Three components are described in some detail: the rotary attenuator, the variable impedance and the rotary directional coupler. The properties of these components are entirely or largely independent of frequency. The rotary directional coupler gives a coupling which is continuously variable from −3 to −30 or −40 dB, and can be read from a dial. The variable impedance gives a direct reading of the modulus and argument of the reflection coefficient.

Microwave techniques are nowadays the subject of considerable interest, for general scientific reasons as well as in connection with the development of technical applications. The main branches of science in which microwave techniques are of importance are plasma physics, microwave gas spectroscopy and solid-state research.

In plasma physics, interest is focused on the electron density and temperature of plasmas. The electron density may be derived from the propagation characteristics of millimetre waves which are passed through the plasma. One way of determining the temperature is to compare the energy radiated by the plasma in a certain millimetre wave region with that of a radiation source of known intensity (noise generator) in the same waveband. Such measurements are of particular importance in experiments relating to nuclear fusion [1]).

Microwave gas spectroscopy is concerned in particular with the study of molecular rotational energy. In principle a microwave gas spectrometer consists of a microwave generator, a vacuum-tight waveguide filled with the gas under investigation, and a detector. The equipment measures the absorption occurring at specific frequencies [2]).

A third application is in the field of solid-state research. When an electromagnetic wave is directed upon a solid, such as germanium or silicon, in which a magnetostatic field is present perpendicular to the electric field, cyclotron resonance will occur at certain frequencies [3]), yielding information, e.g., on the charge carriers. Millimetre waves are also used in research on superconductivity [4]).

For investigations of this kind, suitable equipment is needed for generating the microwaves, for propagating them and performing measurements with them. As regards microwave generators, the Philips 4 and 2.5 mm reflex klystrons have proved very satisfactory [5]). The reflex klystron used for the present range of equipment is the 4 mm DX 151, the wavelength being reduced to 2 mm by means of a frequency doubler (point-contact diode).

For propagating the 2 mm waves, waveguides of rectangular cross-section are generally used (0.83 × 1.66 mm or 0.0325″ × 0.065″).

The measurements involved usually relate to frequency, phase, power, absorption and reflection coefficient. Components for measuring these quantities in the centimetre wave region were described in this journal ten years ago [6]). Since then, comprehensive series of components have been developed for wavelengths of 3 cm, 8 mm and 4 mm. At the latter wavelength the utmost care is needed to keep the dimensions within the required tolerances. Even so, it has proved possible, with the same relative tolerances, to develop a range of components for 2 mm waves [7]), the more important of which will be reviewed in this article. Some are scaled-down versions of comparable components for longer waves; others differ mechanically from the familiar designs, and three are entirely new, namely: the *PIN modulator*, the *variable impedance* and the *rotary directional coupler*.

All components for 4 and 2 mm waves are fitted with claw flanges (to be discussed presently), by

[1]) M. A. Heald, Microwave measurements in controlled fusion research, Nat. Conv. Rec. Inst. Radio Engrs. **6**, No. 9, 14-18, 1958.

[2]) W. C. King and W. Gordy, One-to-two millimeter wave spectroscopy, Phys. Rev. **93**, 407-412, 1954.
M. Cowan and W. Gordy, Further extension of microwave spectroscopy in the submillimeter wave region, Phys. Rev. **104**, 551-552, 1956.

[3]) C. J. Rauch, J. J. Stickler, H. J. Zeiger and G. S. Heller, Millimeter cyclotron resonance in silicon, Phys. Rev. Letters **4**, 64-66, 1960 (No. 2).

[4]) M. A. Biondi and M. P. Garfunkel, Millimeter wave absorption in superconducting aluminum, Phys. Rev. **116**, 853-867, 1959 (No. 4).

[5]) B. B. van Iperen, Reflex klystrons for wavelengths of 4 and 2.5 mm, Philips tech. Rev. **21**, 221-228, 1959/60 (No. 8). Another method of generating millimetre waves is described by M. Yéou-Ta, Carcinotrons du type O fonctionnant sur une longueur d'onde de 2 mm, Onde électr. **39**, 789-794, 1959 (No. 391).

[6]) A. E. Pannenborg, A measuring arrangement for waveguides, Philips tech. Rev. **12**, 15-24, 1950/51.
A. E. Pannenborg, Some aspects of waveguide technique, Communic. News **11**, 65-75, 1950.

[7]) We are particularly indebted to the instrument-maker J. van Ostade for his contribution to this work.

which their respective waveguides are joined together. This construction makes for compactness and permits ready assembly and dissembly of microwave systems.

Generation, detection and modulation of 2 mm waves

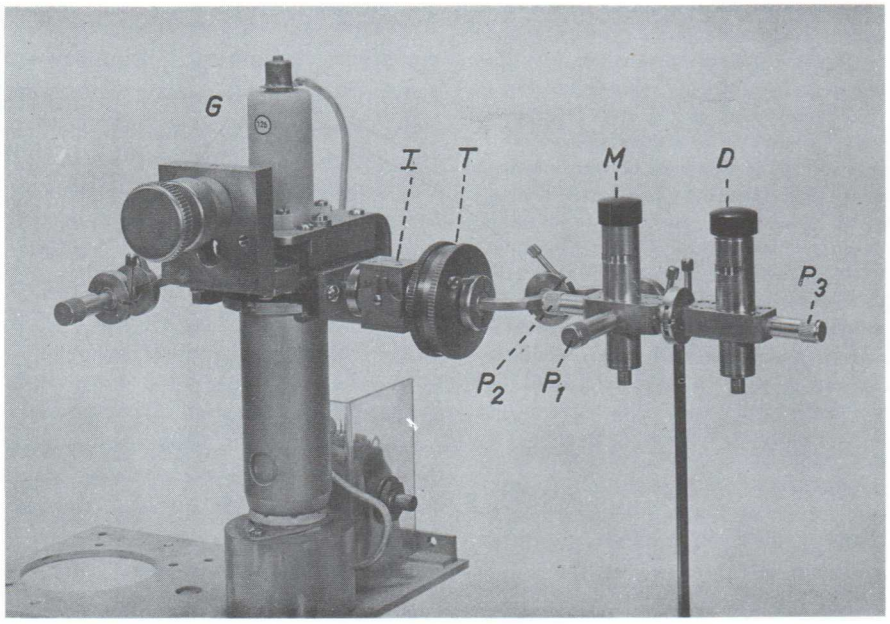
The generator, the frequency multiplier and the detector

As mentioned above, the 2 mm waves are generated in our equipment by doubling the frequency of a reflex klystron, type DX 151, designed for 4 mm waves. Mounted between the klystron and the frequency multiplier are an isolator *I* and a tuner *T* (*fig. 1*). The isolator prevents feedback from the load to the generator, whose frequency and power output would otherwise vary with the load. The principle of this isolator depends on gyromagnetic resonance [8]). The tuner, combined with a shorting plunger P_1, ensures that virtually the whole power output of the klystron reaches the frequency multiplier. The design of the tuners used will be described later.

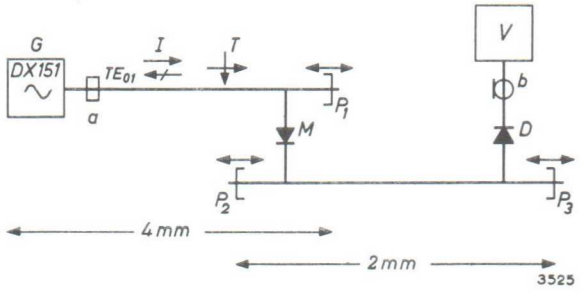
[8]) H. G. Beljers, Resonance isolators for millimetre waves' Philips tech. Rev. **22**, 11-15, 1960/61 (No. 1).

The frequency multiplier consists of a point-contact diode situated at the junction between the 4 mm waveguide *1* and the 2 mm waveguide *2* (*fig. 2*). Projecting through these two waveguides is a tungsten catswhisker *3* which, together with a silicon crystal *4*, forms the point-contact diode. The lower end of the catswhisker is welded to an insulated pin *5*. The other end is etched to a fine point. The silicon wafer, which can be finely adjusted in the vertical direction by a differential screw, is brought into contact with the whisker. The part of the catswhisker running through the two waveguides acts as a coupling probe. The resistance of the diode (and also the capacitance) depends on the applied voltage [9]). As a result of this non-linearity the current induced in the coupling probe by the incident 4 mm waves contains a component with twice the frequency. This produces a 2 mm wave in

[9]) H. C. Torrey and C. A. Whitmer, Crystal rectifiers, M. I. T. Radiation Lab. Series, **15**, McGraw-Hill, New York 1948, in particular pages 97-107, Rectification at high frequencies. R. S. Ohl, P. P. Budenstein and C. A. Burrus, Improved diode for the harmonic generation of millimeter and submillimeter waves, Rev. sci. Instr. **30**, 765-774, 1959 (No. 9).



b



a

Fig. 1. Block diagram (*a*) and photograph (*b*) of a set-up for generating 2 mm waves. *G* generator of 4 mm waves (reflex klystron DX 151). *I* isolator. *T* tuner. P_1, P_2, P_3 adjustable waveguide plungers. *M* frequency multiplier. *D* detector for 2 mm waves, with millivoltmeter *V*. The sign *a* means "rectangular waveguide" and the sign *b* means "coaxial lign".

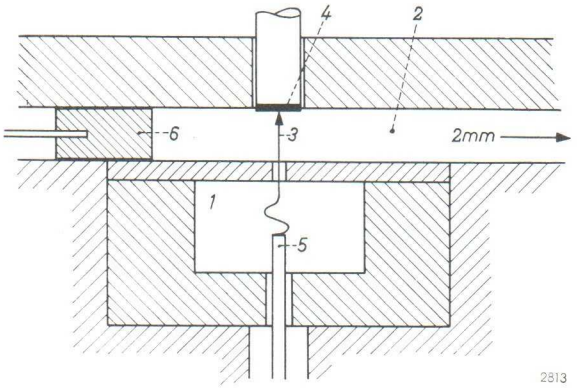


Fig. 2. Cross-section of frequency multiplier. *1* is the 4 mm waveguide, *2* the 2 mm waveguide. *3* catswhisker and *4* silicon crystal, together forming a point-contact diode. *5* insulated pin carrying the catswhisker. *6* shorting plunger in the 2 mm waveguide (P_2 in fig. 1).

the upper waveguide. *Fig. 3* is a photograph of the frequency multiplier.

To verify that the equipment is functioning properly up to the frequency multiplier, the latter can be used temporarily as a 4 mm-wave detector. For this purpose a DC voltmeter (e.g. type GM 6020 electronic millivoltmeter) is connected between the central pin 5 and earth — i.e. in parallel with the point-contact diode. If the meter reading is satisfactory, a crystal detector D (fig. 1) is then connected to the 2 mm waveguide for the purpose of detecting the 2 mm waves. The detector is virtually identical with the frequency multiplier, except of course that it has no 4 mm waveguide.

The pivoting screw tuner

The matching for the frequency multiplier and the 2 mm detector is effected by waveguide plungers (P_1, P_2 and P_3 in fig. 1) in conjunction with a tuner of special design. At longer wavelengths a tuner of the sliding screw type is ordinarily used. This consists of a metal stub *1* (*fig. 4a*), whose depth of penetration into the waveguide *2* can be adjusted by turning the screw, and which can also be moved along the waveguide, in a slot, over a distance of about $\frac{3}{4}\lambda_g$ (the wavelength in the waveguide, λ_g, is greater than the corresponding wavelength λ in free space).

The relation between λ_g and λ is given by:

$$\frac{1}{\lambda^2} = \frac{1}{\lambda_g^2} + \frac{1}{\lambda_c^2},$$

where λ_c is the cut-off wavelength. (See p. 17, formula (2) of the first article mentioned in reference [6].) For the mode used in the waveguide — the TE_{01} mode — λ_c is equal to twice the width a of the rectangular waveguide. For $\lambda = 2.00$ mm and $\lambda_c = 2a = 3.32$ mm, the wavelength λ_g is 2.51 mm.

The depth to which the stub penetrates into the waveguide regulates mainly the extent to which the stub reflects the wave; displacement of the stub along the guide varies the phase of the reflected wave. Reflections in the equipment can thus be cancelled by generating with the sliding screw tuner a reflection of equal magnitude but opposite phase. In this way a travelling wave is produced.

The 2 mm-wave screw tuner is based on a new design. The stub does not slide along the waveguide, but turns about an axis *3* (fig. 4*b*), so that it describes a pivoting motion, hence the term "pivoting screw tuner". In the millimetre wave region this design has special mechanical and electrical advantages:

1) The objection to the sliding screw tuner, that the stub may move slightly up and down as a result of surface irregularities, is entirely avoided here; there is less friction and adjustment is smoother.
2) The axial length of the tuner, from flange to flange, is limited to 12 mm, as a result of which

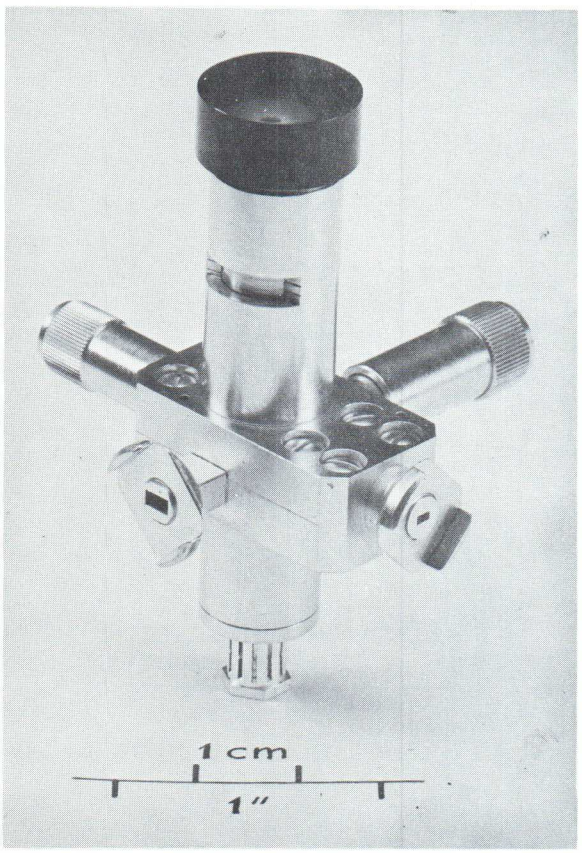


Fig. 3. The frequency multiplier. Foreground, left: connection for the 4 mm waveguide; right, connection for the 2 mm waveguide. The arms on the opposite side are fitted with waveguide plungers. The vertical arm contains the point-contact diode, and is surmounted by the adjusting knob. At the bottom is the coaxial connection for the millivoltmeter. The conversion loss is 15 to 20 dB.

the pivoting screw tuner causes particularly little attenuation (no more than 0.1 dB).

At first sight it might seem a drawback that, since the stub is pivoted about the axis *3*, its depth as well as its position varies in the waveguide, implying that the amplitude of the reflected wave will also vary slightly with the phase. This effect is so small, however, that it gives no trouble in practice.

A photograph of the pivoting screw tuner appears in *fig. 5*.

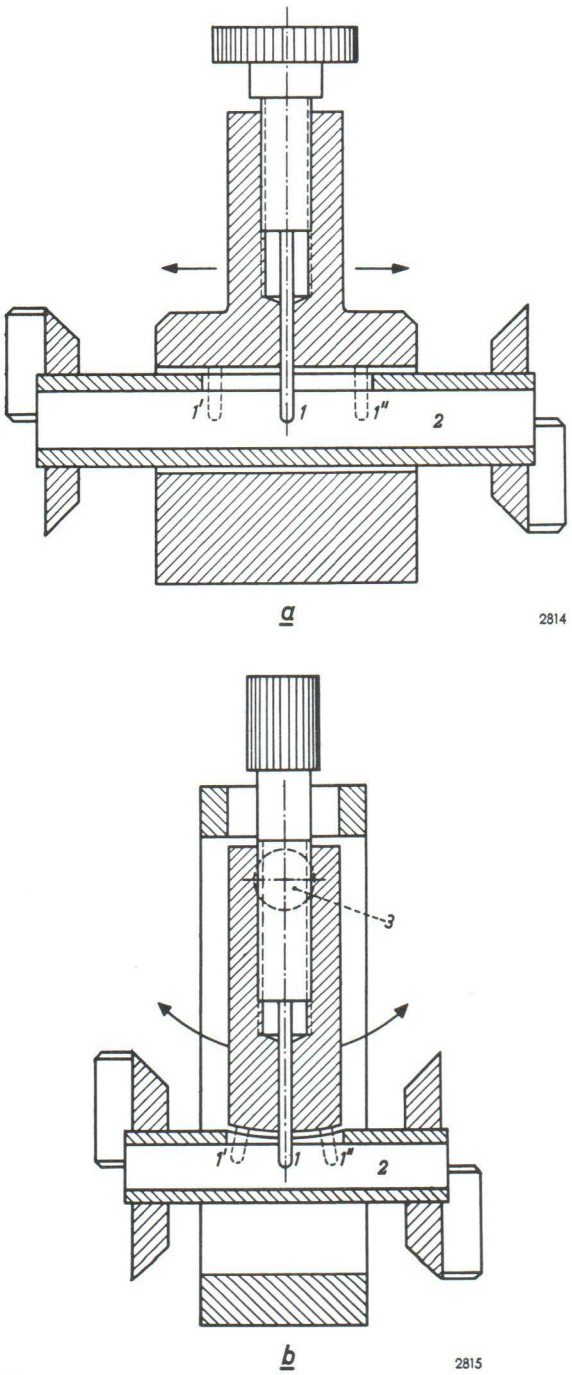


Fig. 4. Schematic representation of screw tuners.
a) Sliding screw tuner. The dept of the stub *1* in the waveguide *2* can be adjusted by turning a screw, and its position along the waveguide can be adjusted between *1'* and *1''*.
b) Pivoting screw tuner. Here too the stub *1* can be screwed in and out, but instead of sliding lengthwise it now pivots about spindle *3*. The extreme positions are again *1'* and *1''*.

Methods of modulation

At an input power of, say, 100 mW the frequency multiplier delivers a 2 mm output of up to 2 mW. To detect the 2 mm wave at various points in the equipment, where the power may be much lower still, a highly sensitive method of measurement is therefore required.

For this purpose, in accordance with a widely used principle, the wave is modulated in amplitude and a selective indicator is employed behind the detector. The indicator consists of an amplifier tuned to the modulation frequency (e.g. 800 c/s), a low-frequency detector and a DC voltmeter. With a modulation depth of 100% the meter will then give a full scale deflection at a microwave power of the order of 1 μW. The sensitivity can be raised still further by the use of synchronous detection; in that case, to keep the noise level down, the modulation frequency should be fairly high (e.g. 8000 c/s).

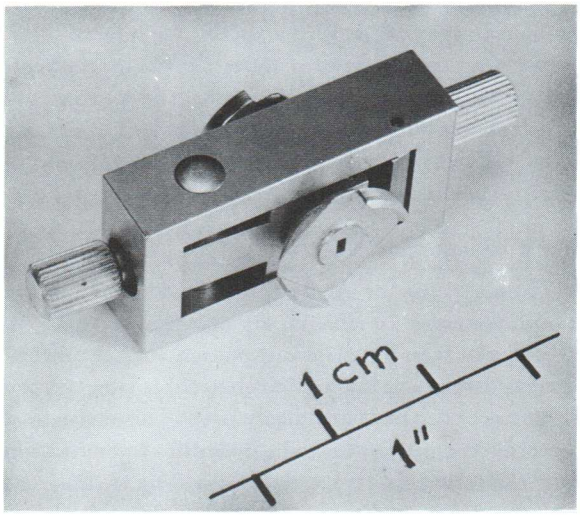


Fig. 5. Pivoting screw tuner for 2 mm waves. Left, the knob for screwing the stub in and out of the waveguide; right, the knob for altering its position (by pivoting about *3* in fig. 4*b*). This component is capable of effecting a match in all cases encountered in practice. The attenuation is no more than 0.1 dB.

We shall now consider briefly some of the modulation methods used. The most common method is to periodically vary the repeller voltage V_r of the reflex klystron between the value for optimum oscillation and a value at which the klystron does not oscillate. An objection to this method is that the oscillation

frequency also depends on V_r [10]); to avoid unwanted frequency modulation one would have to give V_r a square waveform. In practice, however, it is very difficult to make the edges sufficiently steep and to keep V_r constant enough in the oscillation interval. Some frequency modulation therefore remains, which can be particularly undesirable where bridge circuits are involved.

In our case, where a frequency multiplier is used, the modulation can be simply effected, without frequency modulation arising, by applying a low-frequency alternating voltage across the diode of the frequency multiplier, in series with any biasing voltage present. An alternating voltage of about 1 V is quite sufficient to modulate the 2 mm wave 100% in amplitude.

Without a frequency multiplier, pure amplitude modulation is possible by making use of a new device called the *PIN* modulator [11]). This is an electrically controlled attenuator which is inserted in the waveguide. It consists of a germanium wafer in which a *P* region and an *N* region are separated by an *I* region of pure germanium, possessing extremely low "intrinsic" conductivity. Only the *I* region is contained inside the waveguide. In the normal state it behaves like a low-loss dielectric. When, however, a voltage is applied between the *P* and *N* regions in the forward direction, holes and electrons are injected into the *I* region, which thereby becomes an absorbing medium. In this way, using a control current of 15 mA (control power approx. 10 mW), the losses in the *I* region can be made so high as to increase the attenuation by 25 dB. If an alternating control current is used, the wave is thus modulated in amplitude. A *PIN* modulator for 2 mm is illustrated in *fig. 6*.

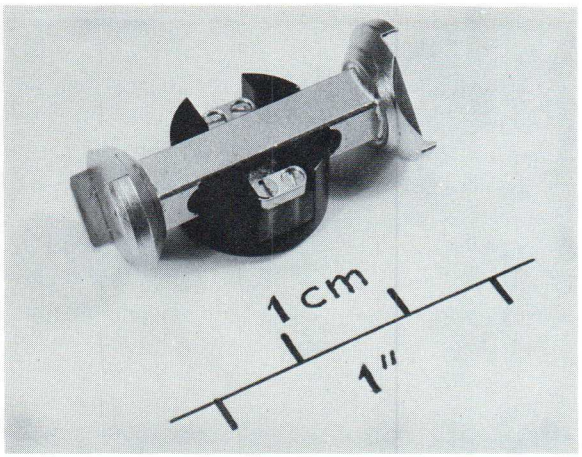


Fig. 6. *PIN* modulator for 2 mm waves. The attenuation is adjustable from about 5 dB to about 30 dB. The modulation frequency at which the modulation depth has dropped by 3 dB is approximately 50 kc/s.

Waveguide and claw flanges

Except in the rotary attenuator, the variable impedance and the rotary directional coupler, where circular waveguides are used for reasons presently to be discussed, the waveguides in the various components are all rectangular.

The rectangular waveguides are constructed of a U-section, made by milling a groove in rectangular bar stock, to which a flat cover plate is screwed or soldered. In most cases the material used is brass; the waveguides are gold-plated inside and out in order to keep the surface properties constant. Silver is used only where minimum attenuation is required.

The internal dimensions of the 2 mm waveguides are 0.83×1.66 mm, with a tolerance of 0.01 mm. The outside dimensions are determined solely by the mechanical considerations applicable to the part in question. The internal cross-section must be uniform, otherwise a discontinuity can arise between coupled components, giving rise to reflections. But even when the cross-sections are uniform, reflections may still occur if the walls of the coupled sections of waveguide are not well-aligned (*fig. 7a, b* and *c*) [12]). Such misalignment points to shortcomings of the waveguide coupling (the flanges and the means by which they are fixed together). An essential requirement of a waveguide coupling is that the walls should be in true alignment with one another.

Other desirable properties of a waveguide coupling are:

1) The axial length should be small in order to mini-

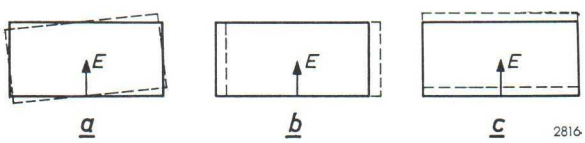


Fig. 7. Excessive tolerances between the flanges mutually or of the flange with respect to the waveguide may result in misalignment of the walls, e.g. the walls may be slightly turned in relation to one another (case *a*), or displaced at right angles to the electric field *E* (case *b*), or displaced in the direction of *E* (case *c*).

[10]) See article by Van Iperen in reference [5]), page 224 (fig. 6 and note [4])).

[11]) F. C. de Ronde, H. J. G. Meyer and O. W. Memelink, The *PIN* modulator, an electrically controlled attenuator for mm and sub-mm waves, Trans. Inst. Radio Engrs. on microwave theory and techniques MTT **8**, 325-327, 1960 (No. 3). See also Dutch patent application No. 229 531 of 11th July 1958.

[12]) U. von Kienlin and A. Kürzl, Reflexionen an Hohlleiter-Flanschverbindungen, Nachr.techn. Z. **11**, 561-564, 1958.

mize attenuation and frequency-dependence [13]). This applies both to the flange and to the space between flange and component. The construction should be such that no tools need be used between flange and component; for this reason the joint should preferably be secured from a direction perpendicular to the waveguide axis.

2) The losses should be low and also reproducible, i.e. the losses should not be dependent on the force applied to draw the two flanges together. This force should preferably be applied by means of no more than one screw. The construction illustrated in *fig. 8*, where the flanges are fixed together by four screws, is not only unpractical but fundamentally wrong, because it is not possible to check whether the screws have all been tightened equally.

3) The coupling should be "universal", i.e. it should be possible to reverse the components in the waveguide system and also to connect them in reversed positions about the axis of the waveguide.

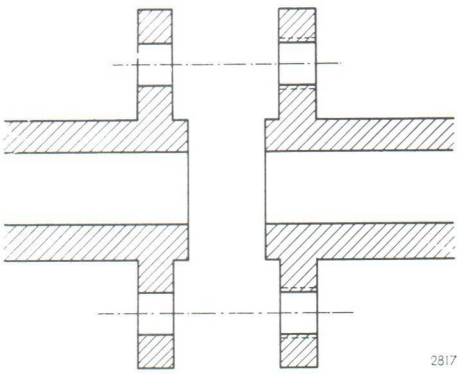


Fig. 8. Conventional waveguide joint. The flanges are joined by four bolts. The objections to this system are:
1) It is impossible to check whether the bolts have all been equally tightened.
2) The waveguide section between flange and component is necessarily relatively long, owing to the presence of the clamping bolts.

[13]) The frequency range in which a rectangular waveguide operating in the TE_{01} mode can be used is from about 1.2 to 1.9 times the cut-off frequency $f_c = c/2a$ (c = velocity of light, a = width of waveguide).

A flange that meets all these requirements is the *claw flange*, which is fitted to all our 2 mm components. *Fig. 9* shows two claw flanges and the clamping ring consisting of two hinged halves, by which the flanges are drawn tightly together by means of a single screw perpendicular to the waveguide axis. The whole coupling — the two flanges and the clamping ring — has a total length of twice the flange thickness, namely 5 mm. As appears from the various photographs in this article, the flanges are fitted close up against the components, making for a particularly compact assembly.

Description of some 2 mm components

Some components of the microwave equipment will now be discussed in somewhat more detail, viz. three adjustable components, provided with dials: the rotary attenuator, the variable impedance and the rotary directional coupler.

The rotary attenuator

It is sometimes necessary in microwave measurements to adjust the power level of the transmitted wave. There is also often a need to introduce a known attenuation. In both cases, use is made of a variable attenuator. One of the oldest types is the vane attenuator, the principle of which is illustrated in *fig. 10*. An insulating plate V (the vane), coated with an absorbent resistive layer, is inserted in the waveguide, to an adjustable depth, parallel to the direction of propagation and to the electric field E. The absorption caused by the vane is greater the deeper the vane extends into the waveguide and the greater its axial length l with respect to the wavelength. The drawback of the latter fact is that the attenuation is partly dependent on frequency. If l is not large with respect to the wavelength, the attenuation will moreover be dependent on the standing-wave ratio in the waveguide.

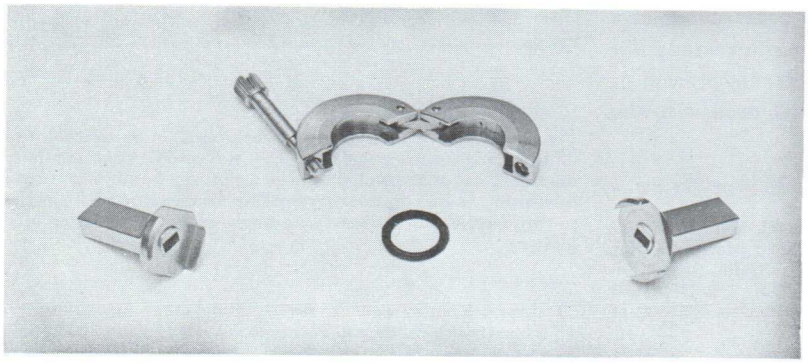


Fig. 9. Left and right, sections of waveguide with claw flanges; top centre, the hinged clamping ring which joins the flanges together; below, a rubber ring. (True size.) The clamping ring is tightened round the flanges with a single screw, perpendicular to the axis of the waveguide. The rubber ring ensures an airtight seal; the air in the whole waveguide system can then be put under pressure to prevent the entry of moisture.
The modulus of the reflection coefficient of the claw-flange coupling is smaller than 0.005.

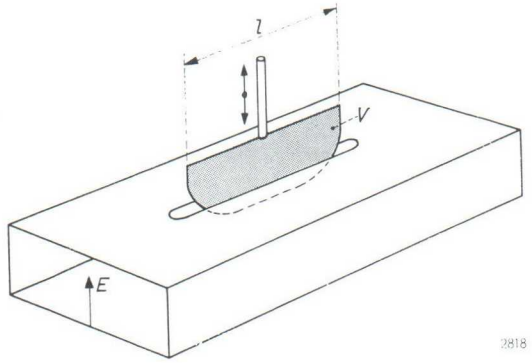


Fig. 10. Principle of the vane attenuator. An absorbent vane V, parallel to the electric field E, can be inserted to a variable depth in the waveguide through a narrow slot. The attenuation increases with the depth of insertion, but also depends on the ratio of the length l of the vane to the wavelength, i.e. it is frequency-dependent.

These drawbacks are entirely absent in the *rotary attenuator*, a device invented in America during the war by Bowen [14]). A further advantage of this attenuator is that it is an absolute instrument, that is to say the scale giving the attenuation in dB obeys a mathematical law, and thus can be engraved at once to the instrument without previous calibration.

A schematic cross-section of the rotary attenuator is given in *fig. 11a*. The ends *1* and *2* are rectangular, and the middle section *5-7-6* is circular. The transition between the rectangular and the circular cross-sections takes place in the regions *3* and *4*. In the circular section the centre piece *7* is rotatable about the waveguide axis. In each of the (fixed) regions *5* and *6*, an absorbing vane (*8*, *9*) is mounted perpendicular to the electric field, i.e. parallel to the long sides of the rectangular waveguide section (fig. 11*b*). The central, rotatable part *7* of the waveguide also contains an absorbing vane, *10*. This can be rotated from a position parallel to *8* and *9* to a position perpendicular thereto; as we shall see, the extreme positions correspond to the minimum and maximum attenuation, respectively.

The attenuation as a function of the angle α between the plane of the vane *10* and that of the vanes *8* and *9* is found as follows. The electric field E of a wave entering at *1* is perpendicular in the fixed section *5* to the vane *8*; the wave here therefore passes through unattenuated. In the rotary section *7* we resolve E into a component $E \sin \alpha$ in the plane of vane *10* and component $E \cos \alpha$ perpendicular thereto (fig. 11*c*). The component $E \sin \alpha$ is entirely absorbed in the vane, leaving only the component $E \cos \alpha$. Finally, in the fixed section *6* the field is $E \cos \alpha$, whose component $E \cos \alpha \sin \alpha$, in the plane of *9*, is absorbed by this vane so that only component $E \cos^2 \alpha$, perpendicular to *9*, remains (fig. 11*d*).

The emergent wave thus has an electrical field-

[14]) G. C. Southworth, Principles and applications of waveguide transmission, Van Nostrand, New York 1950, p. 374.
B. P. Hand, A precision waveguide attenuator which obeys a mathematical law, Hewlett-Packard Journal **6**, Jan. 1955.

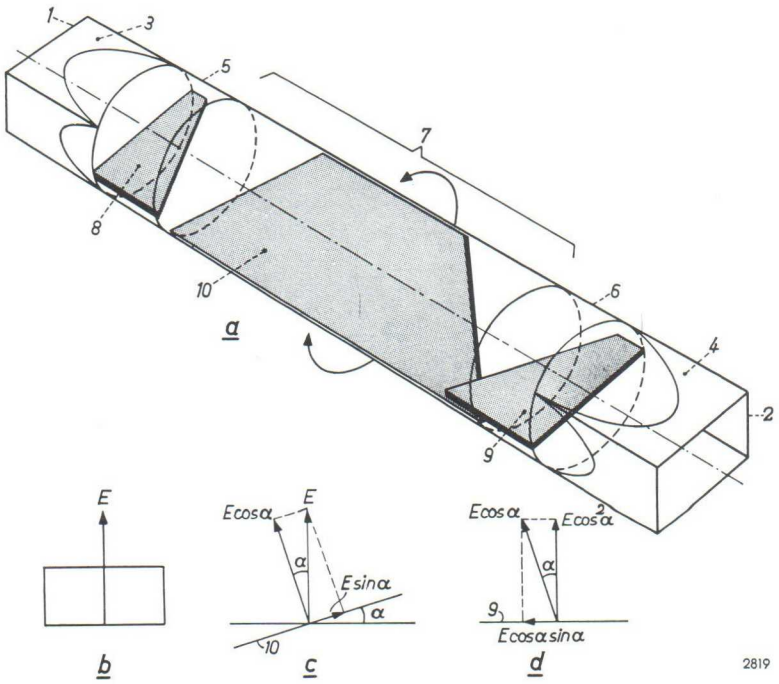


Fig. 11. *a*) Schematic diagram of the rotary attenuator. *1*, *2* rectangular ends for connection to rectangular waveguides. *3*, *4* transitions to circular waveguide. *5*, *6* fixed sections of circular waveguide, *7* rotary section. *8*, *9* and *10* vanes. *b*) Rectangular cross-section at *1* and *2* (long sides parallel to the vanes *8* and *9*). E electric field.
c) The component $E \sin \alpha$ of the electric field of a wave entering at *1* is absorbed in vane *10*, leaving only the component $E \cos \alpha$ (α is the angle made by the plane of *10* with that of the fixed vanes *8* and *9*).
d) In section *6* the component $E \cos \alpha \sin \alpha$ of the remaining wave is absorbed by vane *9*, leaving only the component $E \cos^2 \alpha$. The total attenuation is thus $\cos^2 \alpha$, or $-20 \log \cos^2 \alpha$ dB.

strength which is $\cos^2 \alpha$ times that of the incident wave. The attenuation is therefore solely a function of the angle α and amounts to $-20 \log \cos^2 \alpha$ dB. Without calibration, then, a dial can be fitted whose scale gives an absolute reading of the attenuation.

It is important that the vanes *8*, *9* and *10* should reflect as little energy as possible. For this purpose they are tapered at the ends to form trapezia and are extremely thin, consisting of metallized slivers of mica. The thickness of the metal coating is small compared to the skin depth, so that the vane acts as a resistive layer which strongly absorbs the wave. Due to the tapering of the ends and the thinness of the vanes, the reflection coefficient of the attenuator is less than 0.02.

The 2 mm rotary attenuator is illustrated in *fig. 12*.

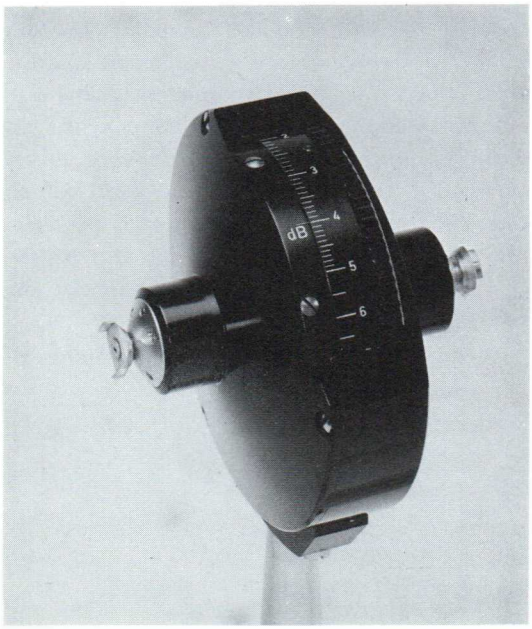


Fig. 12. Rotary attenuator for 2 mm waves. The attenuation, which is independent of the frequency, can be read on a scale covering a range from 0 dB to 50 dB. The insertion loss is 3.5 dB. The standing-wave ratio is less than 1.05.

The variable impedance

At the end of a waveguide there is generally some reflection of energy [6]). The incident and the reflected waves give rise to a standing wave, as a result of which the electric field-strength in the waveguide shows a minimum (E_{min}) at certain places and a maximum (E_{max}) at others. The standing-wave ratio, E_{max}/E_{min}, is a measure of the mismatch (the extent to which the terminating impedance differs from the characteristic impedance), and in many cases this is an important indication of the quality of a microwave component.

In the centimetre range the standing-wave ratio is measured in waveguides in the same way as in lecher lines, that is with a standing-wave indicator. In the millimetre wave range, however, it is difficult to produce a standing-wave detector having the required accuracy. Instead of the standing-wave ratio we can measure the *reflection coefficient*, which is simply related to the standing-wave ratio (see page 18 of the first article mentioned in reference [6])). With the aid of a variable impedance, which gives a direct reading of the reflection coefficient, this measurement can be readily performed with a form of bridge circuit. The bridge is balanced by making the reflection coefficient of the variable impedance equal to that of the unknown impedance.

The device employed here as the bridge is a hybrid T, a schematic representation of which is shown in *fig. 13a*. A wave entering the arm *1* divides equally at the branching point into two components, which enter the symmetrical arms *2* and *3*. There is no direct transfer of energy from arm *1* to arm *4*. Arm *2* is terminated by the unknown impedance Z_2, and arm *3* by the variable (known) impedance Z_3. The waves reflected from Z_2 and Z_3

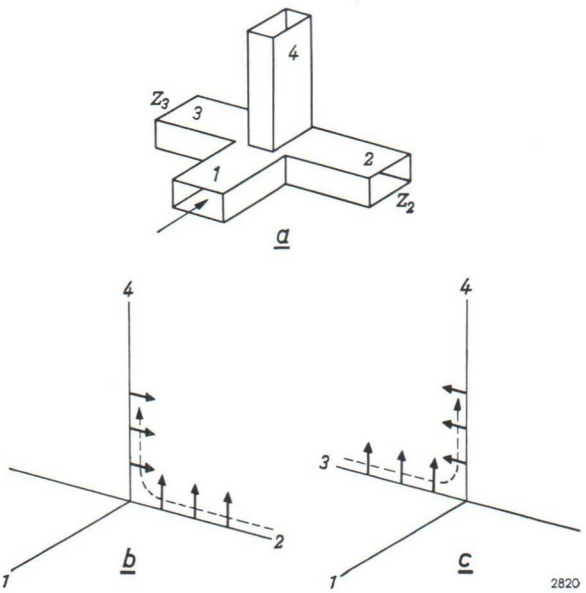


Fig. 13. *a*) Schematic representation of a hybrid Tee. A wave entering arm *1* divides into equal components which enter arms *2* and *3*. These waves are reflected from the impedances Z_2 and Z_3 terminating these arms, and return partly into arm *1* (not considered here) and partly into arm *4*. Here the waves cancel each other if Z_3 is equal in modulus and argument to Z_2. *b*) The short arrows represent the electric field of the wave which is reflected from Z_2 and enters arm *4*; the dashed arrow denotes the direction of propagation. *c*) The same for the wave reflected from Z_3. When $Z_3 = Z_2$, the two waves cancel out in arm *4*.

return to the branching point and enter arm *4*. When the modulus and the argument of Z_3 are made equal to those of Z_2, the electrical field-strength of the wave reflected from Z_2 into arms *2* and *4* will be as shown in fig. 13*b*, while that for the wave reflected from Z_3 is as shown in fig. 13*c*. In arm *4* the field-strengths are equal and opposite, and the waves thus cancel each other. An indicator connected to a detector mounted in arm *4* will therefore give no deflection when the bridge is balanced, but does give a deflection as long as Z_3 is not equal to Z_2.

Fig. 14 shows a hybrid T for 2 mm waves. A particularly important requirement for hybrid T's is that the common planes of symmetry of the Tee *1-2-3* and of the Tee *4-2-3 exactly* coincide.

The principle of the variable impedance is illustrated in *fig. 15* [15]). At the position of flange *1* the waveguide has a rectangular cross-section; via a transition *2*, this becomes a circular cross-section *3*. The latter contains two absorbing vanes, *4* and *5*, which again consist of thin metallized slivers of mica. Vane *4* is fixed, perpendicular to the electric field E of the incident wave; vane *5* is mounted on a metal plunger *6*, which can be both rotated and displaced axially. If the planes of *4* and *5* subtend an angle α, the component $E \sin \alpha$, parallel to vane *5*, is absorbed by this vane; the component $E \cos \alpha$, perpendicular to *5*, is reflected by the plunger *6*. Now $E \cos \alpha$ has a component $E \cos \alpha \sin \alpha$, which is absorbed in vane *4*, and a component $E \cos^2 \alpha$, which emerges from *1* as

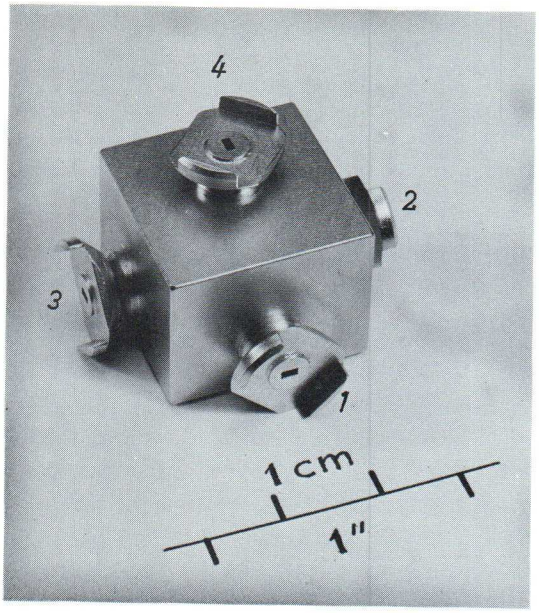


Fig. 14. Hybrid T for 2 mm waves. Numbering of the arms as in fig. 13. The decoupling between arms *1* and *4* is greater than 30 dB.

a reflected wave, so that the modulus of the reflection coefficient is equal to $\cos^2 \alpha$. This modulus is varied by turning the vane *5* by means of the plunger spindle. The argument of the reflection coefficient is varied by displacing the reflecting surface, i.e. by moving the plunger axially, in or out. A displacement δl alters the phase φ of the reflected wave by

$$\delta\varphi = 2 \times \frac{\delta l}{\lambda_{gc}} \times 2\pi \text{ radians},$$

where λ_{gc} is the wavelength in the circular waveguide; the factor 2 is due to the fact that the wave

[15]) See also F. C. de Ronde, A simple component for impedance measurements at cm and mm waves: the direct-reading variable impedance, Communic. Congrès internat. Circuits et antennes hyperfréquences, Paris 1957, Part I (Suppl. Onde électr. **38**, No. 376 bis), pp. 294-295, 1958.

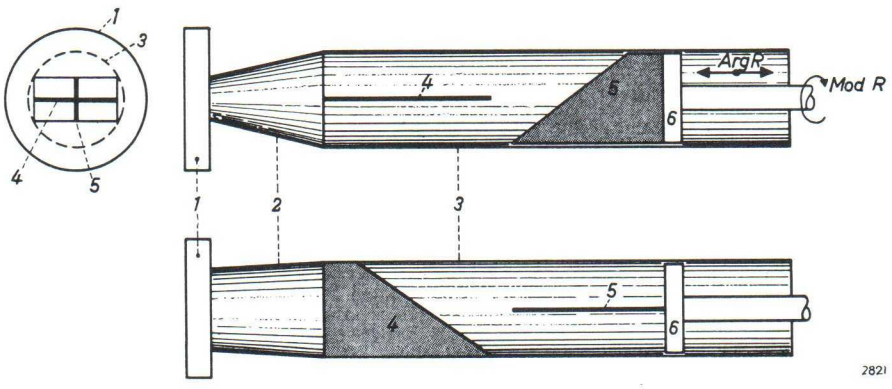


Fig. 15. Axial sections of the variable impedance along two mutually perpendicular planes, and end view. *1* flange with rectangular opening. *2* transition from rectangular to circular cross-section. *3* circular waveguide. *4* fixed vane. *5* vane capable of being rotated and axially displaced, fixed to plunger *6*. When *5* is rotated, the modulus of the complex reflection coefficient is changed; when *5* is displaced axially, the argument is changed.

traverses the waveguide twice (incident and reflected waves).

As can be seen in *fig. 16*, the instrument gives a direct reading of both modulus and argument. The modulus scale, like the attenuation scale on the rotary attenuator, requires no previous calibration.



Fig. 16. Variable impedance as in fig. 15, for 2 mm waves. The modulus and argument of the reflection coefficient can be read from separate scales. If reflection losses are eliminated, the absolute error in the modulus reading is less than 0.05.

The rotary directional coupler

A directional coupler is a device that makes it possible to sample either the forward or backward wave in a waveguide, and thus to demonstrate, e.g., the presence of reflection. The most familiar type contains coupling holes between the main and a subsidiary waveguide. Its operation is illustrated in *fig. 17*. A wave travelling to the right through the main waveguide W_1 in fig. 17*a* produces no response in the detector mounted in the subsidiary waveguide W_2; the detector does respond, however, to a wave travelling in the opposite direction (fig. 17*b*). In this way one can detect the presence of a reflected wave.

A drawback of such directional couplers is that they function only in the frequency range where the (fixed) distance between the coupling holes is roughly equal to $\frac{1}{4}\lambda_g$. Instead of two large holes, it is the usual practice for various reasons to employ a series of small coupling holes, whose spacing must be extremely accurate; this involves considerable difficulties in manufacture. A third drawback is that the coupling is not variable. None of these disadvantages apply to the new directional coupler which we shall now describe.

The *rotary directional coupler* [16] (*fig. 18a*) consists of a circular waveguide W_1 (only part of which is shown in the figure, for simplicity) which terminates at both ends in a rectangular cross-section. The circular part, at end *I*, contains a metal strip *Pol*, the polarization strip, which can be regarded as lossless. At the left end (near *I*), this strip is exactly parallel with the broad faces of the waveguide, but it can be twisted into a helical surface by means of a rotatable section of the circular waveguide. The polarization strip terminates in a tapered absorption vane *V*, of the type already mentioned. This vane lies in exactly the same plane in which the polarization strip ends. Beyond the vane the circular waveguide contains a thin rod *A*, which acts as a coupling probe, and situated a quarter wavelength further on is a metal plate *R*, which acts as a reflector (see below). The probe and the reflector pass through the axis of the circular waveguide and are parallel to the broad sides of the rectangular ends. Transverse slots in the polarization strip and in the reflector prevent the occurrence of higher modes of oscillation.

The probe couples the circular waveguide with a rectangular subsidiary waveguide W_2. To direct

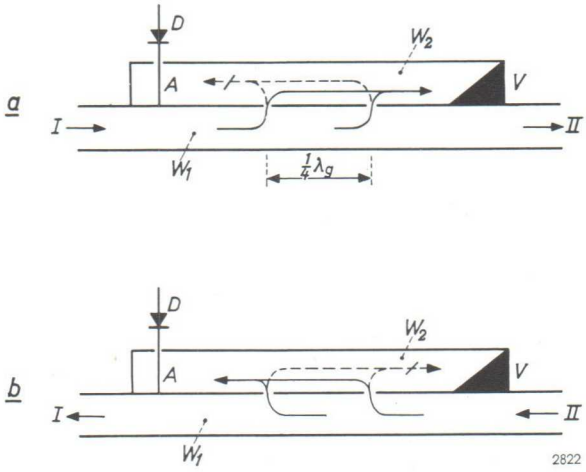


Fig. 17. Two-hole directional coupler. The common wall of the main waveguide W_1 and the subsidiary guide W_2 contains two holes one-quarter wavelength apart. *A* denotes a coupling probe, *D* a detector and *V* a dissipative vane or wedge.
a) A part of a wave in W_1 travelling in the direction of the arrow passes through the coupling holes into W_2, where it generates both backward and forward waves. The waves denoted by the solid arrow are in phase and thus reinforce one another, but they are absorbed by the vane *V*. The waves denoted by the dashed arrows differ in path-length by $2\times\frac{1}{4}\lambda_g$, so that they cancel. No current is therefore induced in the probe.
b) A wave travelling in the opposite direction in W_1 again produces a wave in W_2 which travels in the same direction as in W_1, and now induces a current in the probe.

[16]) Belgian patent application No. 464 685 of 16th December 1959.

all energy in this waveguide to one end — where there is usually a detector — the other end of W_2 is terminated by an adjustable plunger P.

The rotary directional coupler works as follows. We consider first a wave entering the coupler at I. The electric field E is thus perpendicular to the beginning of the polarization strip (fig. 18*b*). The wave then passes along this strip, which is twisted through an angle α. The direction of polarization

The reflector R ensures that the component $E \sin \alpha$ is coupled as fully as possible to the waveguide W_2. For this purpose, R is situated a quarter wavelength behind the probe, and consists of a vertical metal plate, its function being to reflect vertically polarized waves back to the probe and let horizontally polarized waves pass through. The joint effect of the reflector R and the plunger P is to couple the component $E \sin \alpha$ almost entirely to W_2.

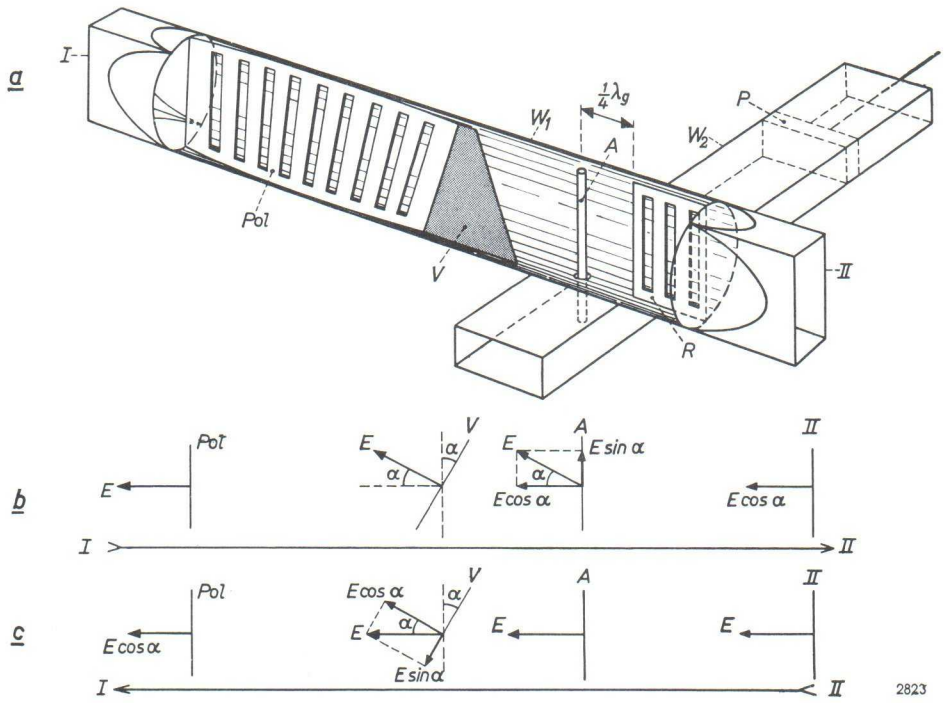


Fig. 18. *a*) Rotary directional coupler (schematic). W_1 circular waveguide terminating at both ends in rectangular openings I and II. *Pol* polarization strip which can be twisted from 0 to 45° by a central rotatable section of the waveguide wall. V vane. A coupling probe. R reflector. W_2 subsidiary waveguide with rectangular cross-section. P movable plunger. A wave passing from I to II is partly transferred to W_2, the coupling being continuously variable and independent of frequency. No energy is coupled to W_2 in the case of a wave passing from II to I.
b) Electric field of a wave travelling from I to II, at the beginning of the polarization strip *Pol*, at the absorption vane V, at the probe A and in the opening II. The angle of twist is α.
c) The same for a wave from II to I; from right to left: the field at II, at A, at V, and at the end of *Pol*.

thereby undergoes a rotation α, since the vector E remains perpendicular to the lossless strip. At the end of the strip, then, the field is unattenuated and is at right angles to the absorption vane, so that no energy is absorbed by the latter. Beyond the strip the field possesses a vertical component $E \sin \alpha$ and a horizontal component $E \cos \alpha$. The component $E \sin \alpha$ is coupled, via the probe, with the waveguide W_2, whilst the component $E \cos \alpha$ passes on to the rectangular opening II.

A wave entering II in the opposite direction, with E again perpendicular to the broad sides of the rectangle (fig. 18*c*, right), passes the reflector, induces no current in the probe (so that energy in this direction is not transferred to W_2) and arrives at the absorption vane at a certain angle. The component $E \sin \alpha$ is absorbed by the vane, and the component $E \cos \alpha$ perpendicular to the vane reaches the twisted polarization strip, after which, still perpendicular to the strip and unattenuated,

it is turned back through an angle α to the position perpendicular to the broad sides of the rectangular guide, in which position it emerges from the rectangular opening I.

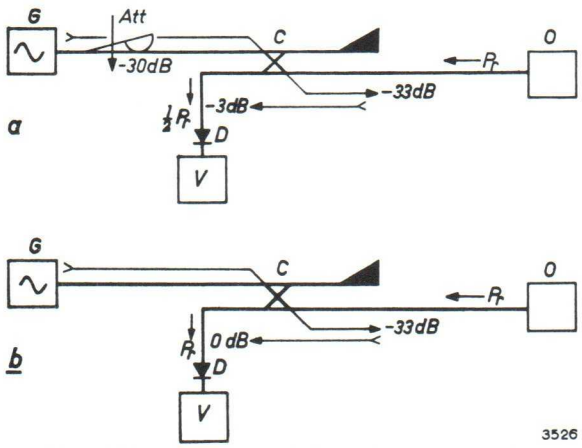


Fig. 19. *a*) Illustrating the principle of a gyromagnetic resonance measurement. *G* microwave generator. *Att* variable attenuator. *C* fixed directional coupler. *O* specimen (solid in resonant cavity, in variable magnetostatic field). *D* detector. *V* millivoltmeter.

The directional coupler gives a coupling of, e.g., −3 dB in the path from *G* to *O*. To avoid saturating *O*, an extra attenuation of 30 dB is introduced, bringing the total attenuation to 33 dB. Of the total reflected power P_r, only half enters the detector via *C*.

b) The same layout, now with a rotary coupler. The required attenuation of 33 dB in example (*a*) can now be produced without a variable attenuator by adjusting the coupling from *G* to *O* to −33 dB. The coupling from *O* to *D* is then strong enough for almost the entire reflected power P_r to enter the detector.

The result is thus as follows.

a) When waves travel in the direction from I to II, the component $E \sin \alpha$ is transferred to W_2, whilst the component $E \cos \alpha$ passes through.

b) When waves travel in the direction from II to I, no energy is transferred to W_2, the component $E \sin \alpha$ is absorbed and the component $E \cos \alpha$ passes through.

The factor that indicates what fraction of the field of the waves travelling from I to II is transferred to the subsidiary waveguide is referred to as the coupling. The power coupling in this case is thus $\sin^2 \alpha$ or, in decibels, 10 log $\sin^2 \alpha$ dB. The coupling effected by the rotary directional coupler is therefore continuously variable and moreover solely a function of the angle α, irrespective of the frequency. The value of the coupling can be read from a scale. In all other existing directional couplers the coupling is not variable and is to some extent frequency-dependent.

The usefulness of a continuously variable coupling may be illustrated by an example, in this case measurements of paramagnetic resonance in solids [17]). (Such measurements are ordinarily performed in the centimetre wave region, but this does not detract from the advantages of a variable coupler over a fixed type.)

The principle of the measurement is illustrated in *fig. 19a*. Microwave energy from a generator *G* is conducted through a directional coupler *C* to the sample under investigation *O*, which is contained in a resonant cavity, itself situated in a variable magnetostatic field. Through the same directional coupler a part of the reflected power P_r is transferred to the detector *D*. When the magnetostatic field is varied, resonance is observed as a sharp dip in the reflected power. The sensitivity is optimum when the coupling is −3 dB. This value is obtained with a hybrid T (preferably a magic T, i.e. a hybrid T free from reflection). For this reason, and also because of its good directivity (this term is explained below), the magic T is often used as a directional coupler in this kind of measurement.

A coupling of −3 dB, however, may give rise to saturation

[17]) See e.g. J. S. van Wieringen, Paramagnetic resonance, Philips tech. Rev. **19**, 301-313, 1957/58.

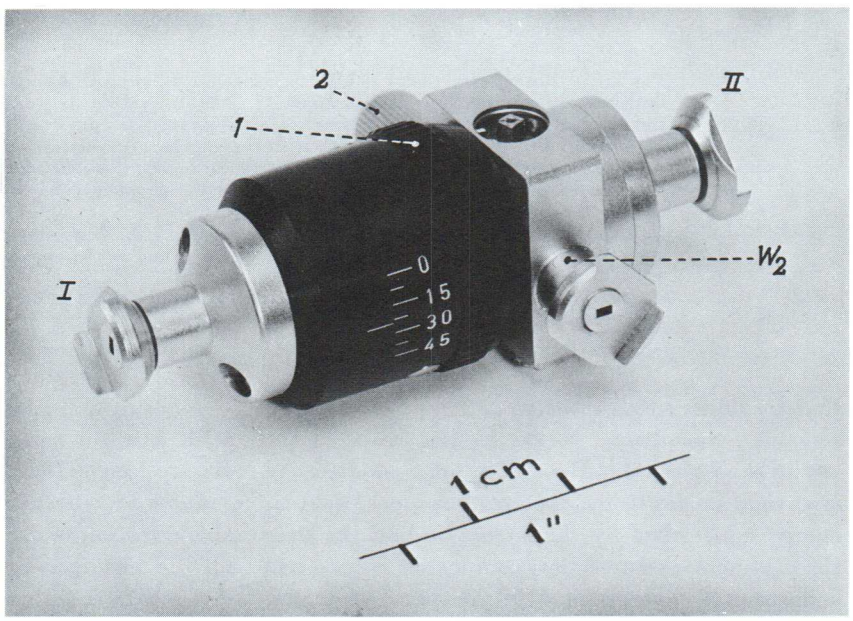


Fig. 20. Rotary directional coupler for 2 mm waves. I, II and W_2 as in fig. 18*a*. *1* rotatable ring for twisting the polarization strip; the angle of twist α is read from the scale. *2* knob for adjusting the plunger in W_2. Directivity >25 dB, standing-wave ratio <1.16.

in the specimen O, that is to say the incident power is so high that the reflected power is virtually independent of it. To remain below the saturation region one must therefore reduce the incident power. The most obvious method is to replace the fixed directional coupler by one which gives a weaker coupling, but this is cumbersome and entails stepwise regulation. Another method is to insert a variable attenuator (*Att*, fig. 19*a*) between the generator and the directional coupler. Suppose that the power must be attenuated by 30 dB to remain below the saturation limit. The attenuation between the generator and the specimen is then 33 dB (fig. 19*a*). This can be achieved more simply by substituting for the attenuator and the fixed 3 dB directional coupler a variable rotary coupler adjusted to −33 dB (fig. 19*b*). This dispenses with the variable attenuator. A further advantage is that, with a weak coupling from G to O, the coupling from O to D is strong enough to cause virtually all the reflected power P_r to arrive in D. Where a fixed directional coupler of −3 dB is used (fig. 19*a*), half of P_r is lost.

In fig. 19*b* the rotary coupler is connected as follows (cf. fig. 18*a*): the generator at the free end of the subsidiary waveguide W_2, the object at side I and the detector at side II.

For the rotary coupler it is found to be sufficient if α is variable from 0 to 45°. Theoretically, the coupling is then variable from $-\infty$ dB to −3 dB. Owing to mechanical imperfections, the weakest coupling in practice is between −30 and −40 dB.

For the same reason W_2 still receives a small fraction of the energy of a wave passing through the rotary coupler from II to I. This imperfect directional effect — found in all such devices — is expressed quantitatively in the *directivity*. This is the ratio of the power coupled to the subsidiary waveguide as a result of a wave from I to II, to that of an equally strong wave from II to I. The directivity of the rotary coupler is thus better the tighter the coupling.

A photograph of the 2 mm rotary directional coupler can be seen in *fig. 20*.

Some measuring arrangements using the components described in the foregoing will be the subject of Part II of this article. This will also provide an opportunity to describe various other components not dealt with here.

II. MEASURING SET-UPS

When designing components for a new frequency range, the microwave technician requires good measuring equipment in order to investigate the properties of the developed components. In this context, his preference will be for measuring instruments which do not need to be calibrated, i.e. instruments which are "absolute". The rotary attenuator is one such instrument. This latter can be used for the measurement of attenuations such as the dissipative loss in microwave components, which will later be discussed. The rotary attenuator itself was treated in Part I of this article [1]); likewise, the variable impedance with which reflection coefficients can be measured. This instrument is also absolute as far as the modulus of the reflection coefficient is concerned. It is used in the second measuring set-up discussed below for measuring the reflection coefficient of an unknown impedance in a bridge circuit (with a hybrid T as the bridge element). Here, the variable impedance serves as a reference impedance.

The third measuring set-up to be discussed is for the investigation of absorption spectra in gases. An example is carbonyl-sulphide gas (COS), which exhibits absorption at about 146 Gc/s. If the gas is irradiated at the correct frequency, the energy supplied allows the COS molecules to move to a higher rotational level. This transition is accompanied by absorption. The frequency at which absorption occurs can be determined with the set-up to be described.

Measurement of the dissipative loss in microwave components

When a microwave component is inserted in a waveguide set-up which has been matched to both the load and generator sides, the losses will in general increase. This increase, which is called "insertion loss", is caused partly by reflections and partly by dissipation in the four-terminal network constituted by the inserted component. Reflection and dissipation are both characteristic properties of a four-terminal network. The dissipative loss can be measured in the manner now to be described; the reflective loss is derived from an impedance measurement using a second set-up which will be dealt with presently.

It is almost always desired to keep the dissipative loss as small as possible, and for that reason an effort has been made to keep the length of our components to a minimum. The need for short components becomes more evident as the frequency increases, since the loss per unit length increases as the $\frac{3}{2}$ power of the frequency. As mentioned in Part I, the claw flange with its associated lock ring is one of the devices used to achieve short constructions

[1]) Philips tech. Rev. **22**, 113-125, 1960/61 (No. 4).

with correspondingly low dissipative losses. Moreover, at high frequencies the surface roughness has a considerable influence: the higher the frequency, and therefore the smaller the depth of penetration, the greater the resistance which irregularities and impurities will offer to the current.

The dissipative loss can be determined by two power measurements, one *with* the component being investigated in the set-up and one *without* it. The power can be measured by a water calorimeter; this is an absolute instrument but is cumbersome to use. In its place, use can be made of a thermistor, i.e. a thermal detector, which has been calibrated by a water calorimeter or (somewhat less accurately) by DC [2]). Also, a crystal detector (*fig. 1*) with a DC meter can be used as a power indicator. The crystal is a more or less square-law detector, so that the deflection of the meter is approximately proportional to the microwave power. The dissipation loss can be found from the ratio of the readings given by the two power measurements. Use of the rotary attenuator has the advantage that the two measurements can be made at the same crystal power, the measurements then being completely independent of the detection characteristic. The difference in position of the rotary attenuator when the component is present from that when it is removed from the set-up thus gives the dissipative loss directly.



Fig. 1. Crystal detector for 2 mm waves. Left, claw flange, right, shorting plunger. The differential screw which is used to bring the silicon crystal into contact with the catswhisker is situated beneath the black cap; the displacement can be monitored through the window. Below: the coaxial connection for the millivoltmeter. The sensitivity is better than 10 mV per mW.

[2]) See Philips tech. Rev. **21**, 228 (Note [7])), 1959/60 (No. 8).

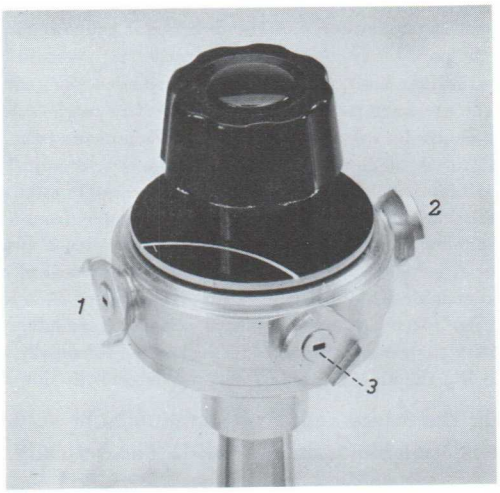


Fig. 2. Waveguide switch for 2 mm wavelength. A rotatable piece of waveguide, in the form of a quadrant, connects waveguide *3* to waveguide *1* in one position of the switch, and waveguide *3* to waveguide *2* in the other position. The standing-wave ratio is less than 1.02.

In these measurements, it is desirable to be able to switch rapidly from one condition to another since, in the interim period, the supplied power may vary for all sorts of reasons: the output power from the klystron, the conversion of the frequency multiplier and the sensitivity of the detector are subject to irregular fluctuations. Rapid switching-over is made possible by means of a waveguide switch (*fig. 2*). This contains a rotatable part consisting of a curved waveguide of quadrant form so that, according to requirements, a connection can be made between the waveguides *3* and *1* or *3* and *2*. It will be clear that particularly high demands are placed upon the mechanical construction and finish, in order to obtain a reproducible connection with a very low reflection coefficient (e.g. $|R|$ smaller than 0.01) between the rotatable and fixed parts of the waveguide.

The circuit with crystal detector and waveguide switch is reproduced in *fig. 3*; a photo of the whole set-up is given in *fig. 4* and a photo of the 2 mm part in *fig. 5*. Two branches are connected to the waveguide switch *S*. One consists of the component to be measured, in this case a piece of waveguide of

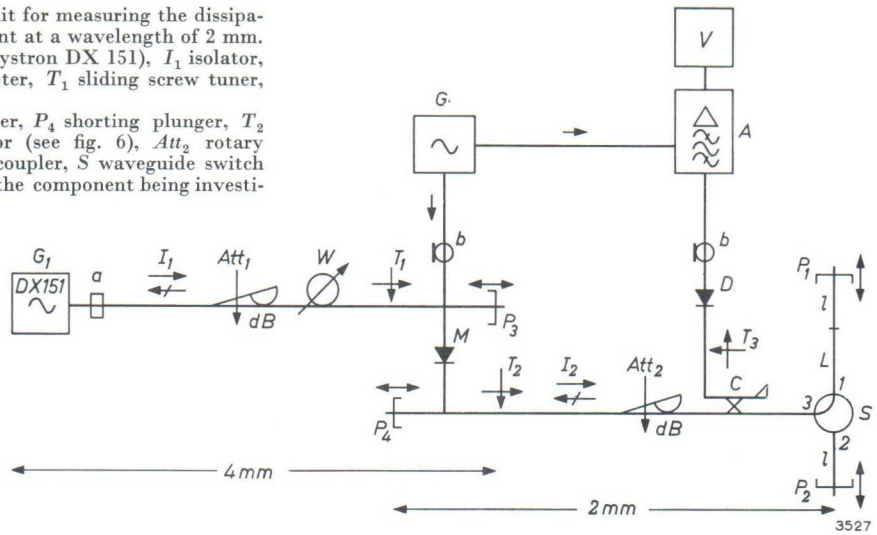


Fig. 3. Block diagram of the circuit for measuring the dissipative loss in a microwave component at a wavelength of 2 mm. *4 mm part*: G_1 generator (reflex klystron DX 151), I_1 isolator, Att_1 vane attenuator, W wavemeter, T_1 sliding screw tuner, P_3 shorting plunger.
2 mm part: M frequency multiplier, P_4 shorting plunger, T_2 pivoting screw tuner, I_2 isolator (see fig. 6), Att_2 rotary attenuator, C rotary directional coupler, S waveguide switch (see fig. 2). Branch *1* consists of the component being investigated (length L) combined with a line of length l adjustable with plunger P_1. Branch *2* consists of a similar length l adjustable with plunger P_2. T_3 pivoting screw tuner. D detector (see fig. 1).
Other equipment: G generator (8 kc/s) for synchronous detection, A selective amplifier, V millivoltmeter.
The sign *a* means "rectangular waveguide" and the sign *b* means "coaxial line".

length L terminated at a distance l from the flange by a shorting plunger P_1. The other branch is terminated directly by a plunger P_2, this latter being adjusted to the same distance l. By means of the waveguide switch, either the branch $L + l$ or the branch l can thus be connected to the rest of the circuit. The circuit contains, in addition to a rotary attenuator, a rotary directional coupler C — a component described in Part I. The rotary directional coupler is used to measure the reflected wave which comes back from the branch switched in by S.

The measurement of the dissipative loss is carried out as follows. With the waveguide switch in position *1*, adjustments are made until the meter V shows full-scale deflection, and the rotary attenuator is set to zero. This done, S is switched to position *2*, i.e. to the opposite branch from that carrying the component to be measured. The attenuation will

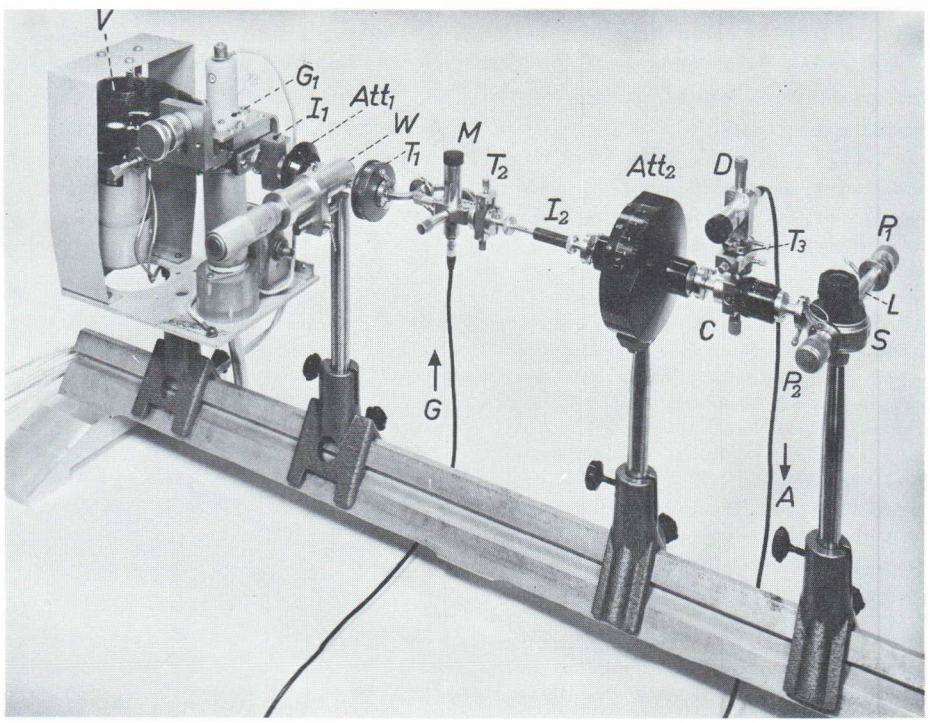


Fig. 4. Set-up for measuring the dissipative loss in components at a wavelength of 2 mm (refer to the diagram in fig. 3). V cooling fan for the reflex klystron. Notation otherwise as in fig. 3.

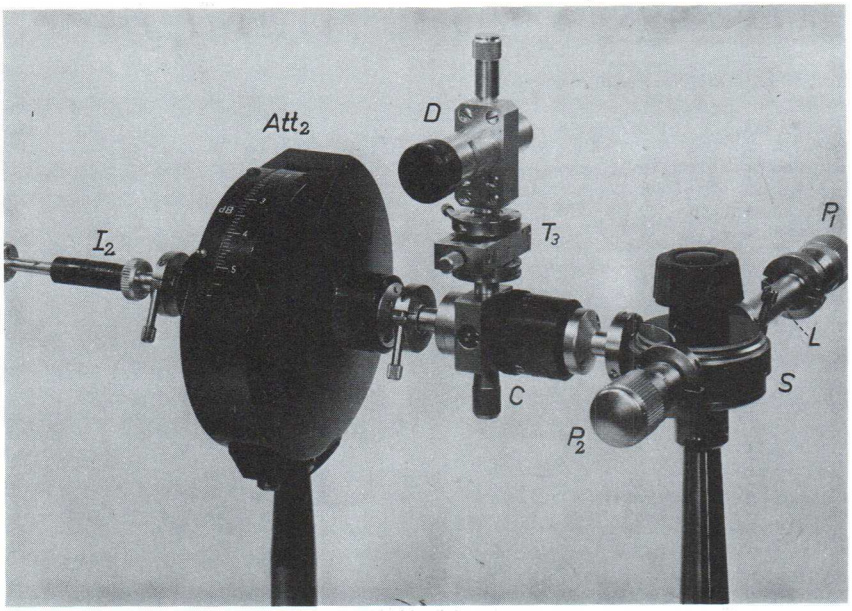


Fig. 5. The 2 mm part of the set-up reproduced in fig. 4. Notation as in fig. 3.

then be less, hence the reflected wave stronger and the meter will tend to deflect further. Now, using the rotary attenuator, the signal is attenuated sufficiently to restore the meter to its original deflection. We then read off from the rotary attenuator the indicated attenuation. Half this attenuation gives the dissipation in the component (half, since the wave traverses the length L of the component twice).

In order to make the measurement as accurate as possible, a few precautions must be taken, one of which is to include a directional isolator.

Looking towards the generator, the transmission line will not generally be reflection-free. For this reason, on displacing the plunger P_1 or P_2, the deflection of the meter will vary. In order to keep this effect as small as possible, an isolator is connected between the attenuator and the multiplier (I_2 in fig. 3). This device, whose operation depends on the Faraday effect [3]), is depicted in *fig. 6*. However, even with these precautions some reflection will remain so that, in order to adjust the meter to the maximum or minimum deflection, the plungers P_1 and P_2 must both be adjustable.

At the beginning of the measurement, and before the component of length L has been inserted in one of the branches, it must be ascertained that the two branches (length l) are identical, i.e. that, for both positions of the waveguide switch, the meter deflections are the same. When the component L is subsequently fitted into the branch l, a small phase change will generally appear (unless the length L is precisely a whole multiple of $\frac{1}{2}\lambda_g$, where λ_g is the wavelength in the guide). This phase change can be corrected by adjusting the plunger P_1. In its turn, this adjustment causes some change in the dissipative loss, but only to a negligibly small degree.

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

Since the use of a directional coupler means that *reflections* from the component are also measured, it must first be ascertained that these are indeed negligibly small. If this is not the case, the reflection must first be compensated.

The measures taken make a very accurate measurement possible. The circuit described is particularly suited to the measurement of losses in sections of waveguide and residual losses in components, e.g. the losses in a vane attenuator (see Part I) in the zero position. With this method, losses of the order of 0.1 dB can easily be measured.

The dissipative loss in the components discussed in Part I was measured with the circuit of fig. 3. This does not, of course, mean that the method is only suitable for measurements on microwave components; it can be used equally well for measuring attenuations caused by any sort of physical

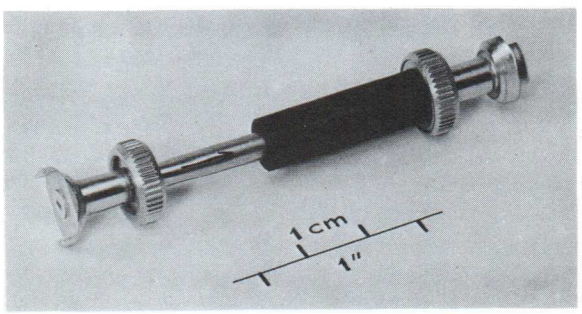


Fig. 6. Isolator for 2 mm wavelength, based on the Faraday effect [3]). The black rings are ferroxdure magnets. Attenuation in the forward direction (in the direction of the arrow): 2 dB; in the reverse direction: 15 dB.

phenomena. It is, for example, very useful for measuring the absorption in superconductors [4]).

Impedance measurements using the bridge method

In general, an impedance reflects some of the applied microwave energy when the impedance differs from the characteristic impedance of the line. A measure of this "mismatch" is either the standing-wave ratio or the reflection coefficient [5]). In order to determine the impedance by means of the standing-wave ratio, a standing-wave detector is necessary. However, it is extremely difficult to make a standing-wave detector for millimetre waves with reasonable accuracy. A more direct measurement of the impedance is possible by inserting it in a bridge circuit and comparing it with a standard impedance. The variable impedance discussed in Part I can serve this latter purpose. This instrument is absolute and the value of the reflection coefficient of the unknown impedance can be read off directly.

The circuit is reproduced in *fig. 7*. It consists of a 4 mm and a 2 mm part, the latter containing a hybrid T (see Part I), *HT*, as the bridge element. The input arm *1* of the T is connected to the frequency multiplier *M* via a pivoting screw tuner T_1. On the arm *4*, via another pivoting screw tuner T_3, a crystal detector *D* is connected, and on the arms *2* and *3* the variable impedance Z_v and the unknown impedance Z_3, respectively. The latter consists of a pivoting screw tuner T_2 in conjunction with a matched load. Using this assembly, any required impedance can be made up.

The energy coming from the multiplier divides in the hybrid T into two equal parts, which travel along the arms *2* and *3*. Z_3 will cause a certain amount of reflection in arm *3*; the reflected wave returns to the branching point of the hybrid T and distributes itself over the arms *1* and *4*. If the reflection from Z_v is now made equal in phase and amplitude to the reflection from Z_3, then a reflected wave of equal magnitude will likewise be split into two equal parts at the branching point and travel along the arms *1* and *4*. As explained in Part I, the geometric configuration is such that the waves issuing from the arms *2* and *3* cancel each other in arm *4;* therefore, the detector *D* receives no signal and the meter does not deflect. This is thus an indication that $Z_v = Z_3$, and the modulus of the reflection coefficient can be directly read off on the variable impedance, whilst its argument can be rapidly determined (see Part I, page 121).

A detail photo of the 2 mm part of this set-up is shown in *fig. 8*.

The sensitivity of this method is very high. Another property of the bridge circuit is that the sensitivity is considerably greater if the measurement is made after the bridge has been slightly unbalanced. This property is useful in the measurement of very weak reflections, such as those which occur at flanges, when values of $|R|$ less than 0.01 can still be fairly reliably measured. For this purpose, one arm of the hybrid T is terminated by a matched load and

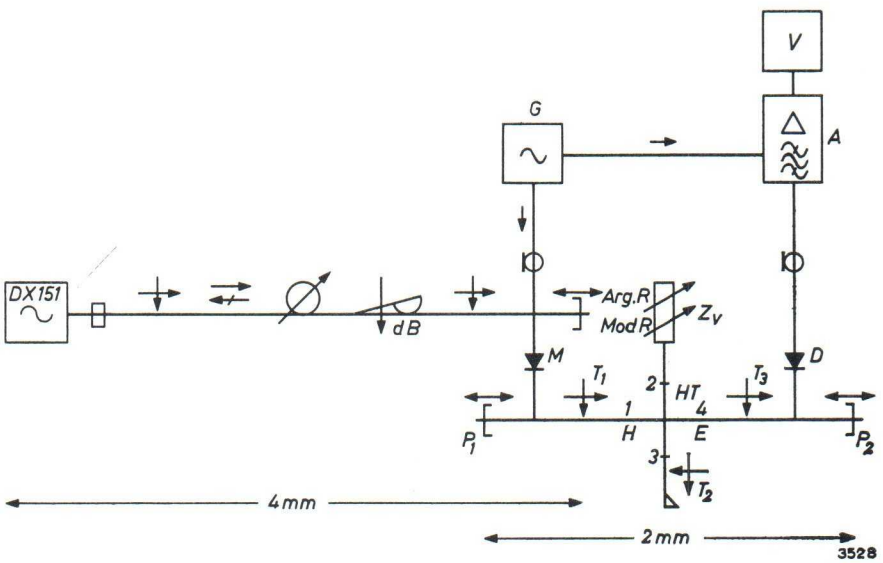


Fig. 7. Block diagram of the circuit for measuring the reflection coefficient of impedances at a wavelength of 2 mm.
4 mm part, from left to right: generator, sliding screw tuner, isolator, wavemeter, vane attenuator, sliding screw tuner, shorting plunger.
2 mm part: *M* frequency multiplier, P_1 shorting plunger, T_1 pivoting screw tuner, *HT* hybrid T, Z_v variable impedance, T_2 pivoting screw tuner (with matched load), T_3 pivoting screw tuner, P_2 shorting plunger, *D* detector.
Other equipment: *G* generator (8 kc/s), *A* selective amplifier (tuned to the frequency of *G*), *V* millivoltmeter.

[4]) M. A. Biondi and M. P. Garfunkel, Millimeter wave absorption in superconducting aluminum, I and II, Phys. Rev. **116**, 853-867, 1959 (No. 4).

[5]) For the relationships between the quantities impedance, reflection coefficient and standing-wave ratio, see e.g. A. E. Pannenborg, A measuring arrangement for waveguides, Philips tech. Rev. **12**, 15-24, 1950/51.

a small reflection is introduced in the arm containing the variable impedance, so that the bridge is slightly out of balance. By altering the $|R|$ of the variable impedance by a certain amount with respect to the adjusted value (readable upon the scale "mod R", see fig. 16 of Part I), a specific variation in the deflection of the meter is obtained. The introduction of a flange coupling between the hybrid T and the matched load likewise causes a change in the deflection. The reflection coefficient can be found from the two changes in deflection. The fact that there are still other reflections of the same order of magnitude as those which are being measured, actually makes the measurement somewhat more complicated than is here described. It is necessary, for example, to compensate as well as possible for the reflections at other flange connections in the arms and for reflections at the transition, and the reflection at the matched load can no longer be neglected. When reasonably accurate measurements are required of very small reflection coefficients such as occur at claw-flange couplings, all this must be taken into account.

The accuracy with which the reflection coefficient of an impedance can be determined is dependent upon the quality of two components: *a*) the hybrid T and *b*) the variable impedance. As has already been shown in Part I, as far as the hybrid T is concerned the accuracy is determined solely by the mechanical construction and the finish. In the variable impedance, still other factors play a part. The detrimental influence of all these factors upon the accuracy are particularly noticeable in the two extreme positions, mod $R = 1$ and mod $R = 0$. In the latter position, there is always some residual reflection, namely at the transition from the rectangular waveguide to the circular waveguide. If extremely accurate measurement is required, this residual reflection must be neutralized by means of a tuner.

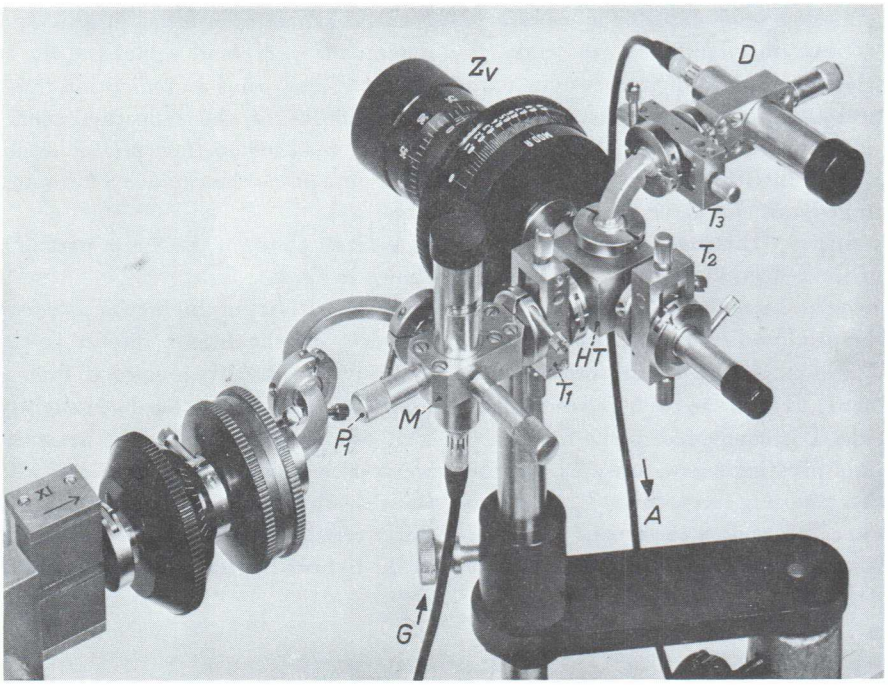


Fig. 8. The 2 mm part of the set-up shown in the diagram of fig. 7. Notation as in the latter. At the extreme left, a few 4 mm components: isolator, vane attenuator and sliding screw tuner.

In the maximum reflection position, $|R|$ is not quite 1 because of the losses in the waveguide. Here, again, is an advantage of the bridge circuit, in that these losses can be compensated by inserting between the impedance Z_3 and the hybrid T a section of waveguide which causes identical losses. For this purpose, the variable impedance is set to the position mod $R = 1$ and, in the other arm of the hybrid T, Z_3 is replaced by a short-circuited piece of waveguide of such length that, by adjustment of the phase, equilibrium can be established. Having balanced the bridge in this way for mod $R = 1$, we can then determine the dissipative loss a in components having low reflection. For this purpose the component is connected between the short-circuit

and the added piece of waveguide. If balance is subsequently re-established, then:

$$a = -10 \log |R| \text{ dB}.$$

Thus, the dissipative loss in microwave components can be determined by the method just described as well as by the circuit of fig. 3. The measurement of large losses is, in both cases, limited by reflections.

Measurement of an absorption line of COS at 2 mm

The binding forces existing between particles which together form a "system" (such as between the atoms in a molecule, between the nucleus and the electrons in an atom, between the protons and the neutrons in an atomic nucleus) give rise to a series of energy states which the system can occupy. On absorbing electromagnetic energy of certain frequencies, such a system can change from a given energy state to one of higher energy. A transition from the n^{th} to the $(n+1)^{\text{th}}$ energy level is made possible by absorption of radiation whose frequency ν is proportional to the energy difference $E_{n+1}-E_n$ between the two levels:

$$\nu = \frac{E_{n+1}-E_n}{h}.$$

Here, h is Planck's constant ($= 6.6\times10^{-34}$ joule seconds). The energy levels E_n are determined by the nature of the forces and the kinds of particles. Thus, absorption in the gamma ray region is caused by changes of state within the atomic nucleus; absorption in the optical region is predominantly concerned with the binding forces between nuclei and electrons, and absorption in the infra-red and microwave regions are related to the force which the atoms of the molecule exert upon one another.

Thus, microwave spectroscopy can provide information concerning the chemical bonds in molecules (vibrational and rotational states). Here, infra-red spectroscopy is also important, but spectroscopy in the microwave region has the advantages that the radiation is purely monochromatic and that the sensitivity, the accuracy of determination of the absorption frequencies and the resolving power are greater. The latter is such that fine and hyperfine structures can be observed, enabling conclusions to be drawn concerning the behaviour of the nuclei of individual atoms. For further particulars the reader is referred to the literature [6]).

The third set-up, now to be discussed, is a microwave gas spectrometer used for determining the rotational spectra of gas molecules. The diagram of the circuit is reproduced in *fig. 9*. The main components in the 2 mm part — shown separately in *fig. 10* — are a gas cell *GC* and a 2 mm crystal detector D_1. The gas cell consists of a piece of 8 mm waveguide (this being used since it has a greater volume and smaller wall-losses per unit length) with at each end a transition to 2 mm guide. The

[6]) See, for example: W. Gordy, W. V. Smith and R. F. Trambarulo, Microwave spectroscopy, Wiley, New York 1957.
D. J. E. Ingram, Spectroscopy at radio and microwave frequencies, Butterworth's Scientific Publications, London 1955.
C. A. Burrus, Stark effect from 1.1 to 2.6 millimeters wavelength: PH_3, PD_3, DI, and CO, J. chem. Phys. **28**, 427-429, 1958.
M. Cowan and W. Gordy, Precision measurements of millimeter and submillimeter wave spectra: DCl, DBr, and DI, Phys. Rev. **111**, 209-211, 1958.

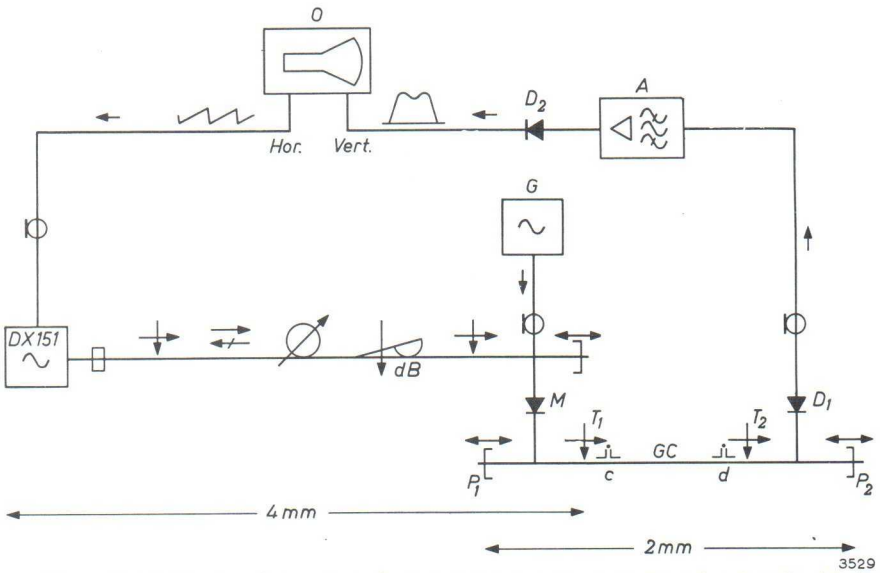


Fig. 9. Block diagram of the circuit for measuring the absorption line of carbonyl sulphide at about 146 Gc/s.
4 mm part, from left to right: generator, sliding screw tuner, isolator, wavemeter, vane attenuator, sliding screw tuner, shorting plunger.
2 mm part: *M* frequency multiplier, P_1 shorting plunger, T_1 pivoting screw tuner, *GC* gas cell (with mica windows *c* and *d*), T_2 pivoting screw tuner, P_2 shorting plunger, D_1 detector.
Other equipment: *G* generator (45 kc/s) which amplitude-modulates the 2 mm wave, *A* selective amplifier tuned to the frequency of *G*, D_2 detector, *O* oscilloscope. The sawtooth time-base voltage of *O* frequency-modulates the 4 mm generator. The form of the detected microwave is indicated at the input of *A*; the form of the signal after low-frequency detection by D_2 is shown at the input of *O*.

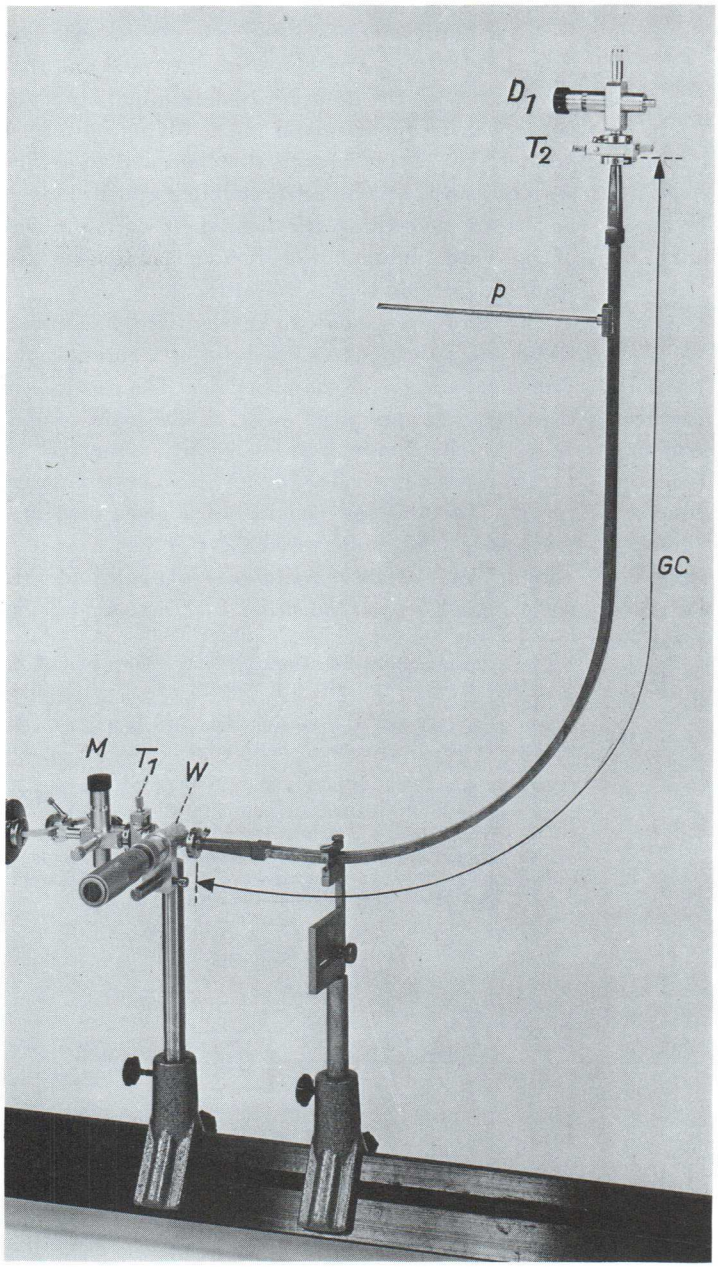


Fig. 10. The 2 mm part of the set-up, the diagram of which is reproduced in fig. 9. p is the pump connection to the gas cell GC (8 mm waveguide). Notation otherwise as in fig. 9. For the purpose of monitoring a wavemeter (W) (TE_{01n} mode) is included in the 2 mm part.

cell is made gas-tight by providing the transitions with mica windows (c and d in fig. 9). In fig. 10 the pump connection p is visible, this being used to evacuate the gas cell before the admission of the gas sample.

One gas that exhibits an absorption line in the neighbourhood of 2 mm wavelength is carbonyl sulphide (COS). The COS molecule is a linear rotator, i.e. the three constituent atoms lie on a straight line (S = C = O), and the molecule rotates about an axis perpendicular to this line. The transitions between the rotational levels lie within the microwave region. Thus, for COS, the following transitions in the 2 mm region have been calculated:

transition from level 9 to level 10 at 121.625 Gc/s,
transition from level 11 to level 12 at 145.947 Gc/s,
transition from level 13 to level 14 at 170.267 Gc/s.

The method of measuring these absorption frequencies will now be described [7]). For this purpose, we choose the line in the neighbourhood of 146 Gc/s. The repeller voltage of the klystron is modulated by the sawtooth voltage which is used for the time-base of the oscilloscope O. The klystron oscillates only when the instantaneous value v_r of the repeller voltage lies between the limits V_{r1} and V_{r2}, and in this region the frequency f varies more or less linearly with v_r (*fig. 11*). In this way, a certain frequency region

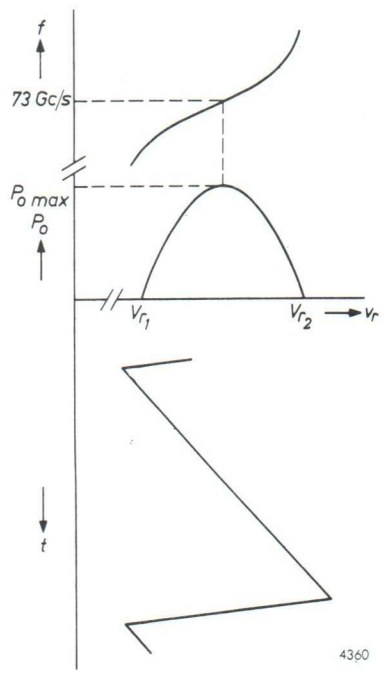


Fig. 11. The repeller voltage v_r of the klystron in fig. 9 has a sawtooth waveform. The klystron oscillates between the limits V_{r1} and V_{r2}. The output power P_o and frequency f vary as shown. At $P_o = P_{o\,max}$ the value of f must be approximately 73 Gc/s.

[7]) C. G. Montgomery, Technique of microwave measurements, M.I.T. Radiation Lab. Ser., Vol. 11, McGraw-Hill, New York 1947, p. 24-33.

is covered [8]). The frequency sweep has a value of, for example, 0.1 Gc/s and the sweep must take place about the value 73 Gc/s approximately; this frequency must occur in the neighbourhood of the maximum output power $P_{o\,max}$.

A check on whether the latter conditions are fulfilled can be made by switching off the generator G and connecting the oscilloscope to the frequency multiplier M, which now acts temporarily as a 4 mm detector. This will give an oscillogram of the type shown in *fig. 12*, which represents the power characteristic of the klystron — P_o as a function of v_r (and therefore of f) — with a superimposed dip caused by the absorption in the 4 mm wavemeter. The latter is tuned to 73 Gc/s. If the klystron is properly tuned, then the peak of the curve occurs at the same frequency as the dip. If this is not so, then the klystron must be mechanically adjusted until the maximum *does* coincide with the dip.

Using the circuit of fig. 9, when *no* gas is present in the cell, an oscillogram of the type shown in *fig. 13a* is obtained. When the cell contains COS gas, the picture obtained is as shown in fig. 13*b*, where the absorption line is visible. The sharper this line is, the more accurately the frequency can be determined. At a given cell volume, the line width increases as the gas pressure rises; this is the result of the interaction between the molecules. An excessive microwave power has a similar detrimental effect because of the saturation which then occurs at the higher level.

The determination of the frequency is achieved in the first instance with the aid of a 4 mm or 2 mm wavemeter. With, for example, a carefully calibrated

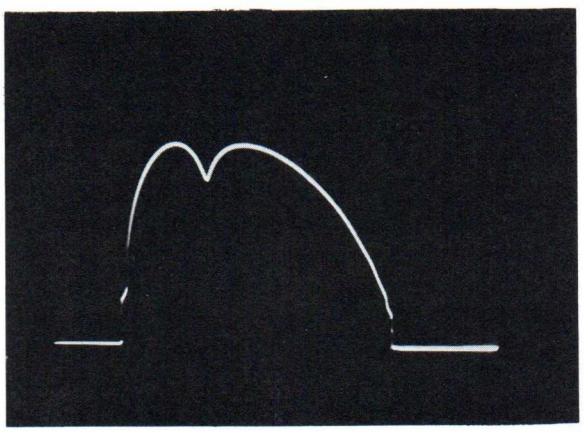
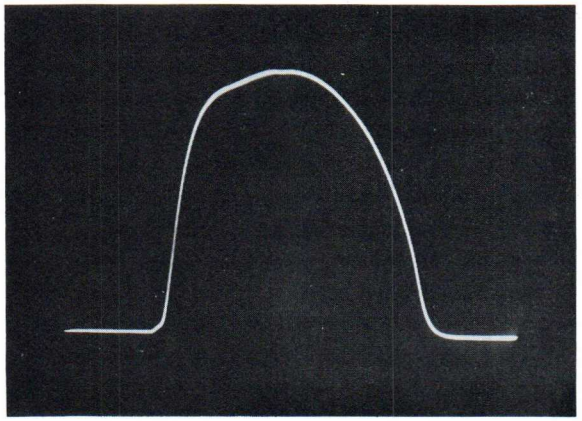
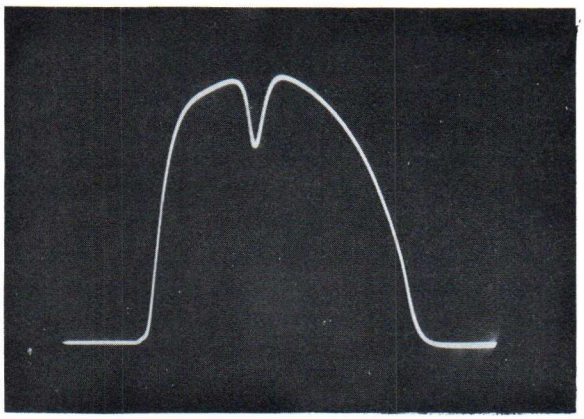


Fig. 12. Oscillogram of the detected output power P_o from the klystron; the abscissa represents both time and frequency. The dip marks the frequency to which the wavemeter is tuned (73 Gc/s), the klystron being mechanically adjusted so that the peak of P_o occurs at this frequency.

[8]) Use is thus made of the "electronic tuning range" of the reflex klystron; see Philips tech. Rev. **21**, 224 (fig. 6, Note [4])), 1959/60 (No. 8).



a



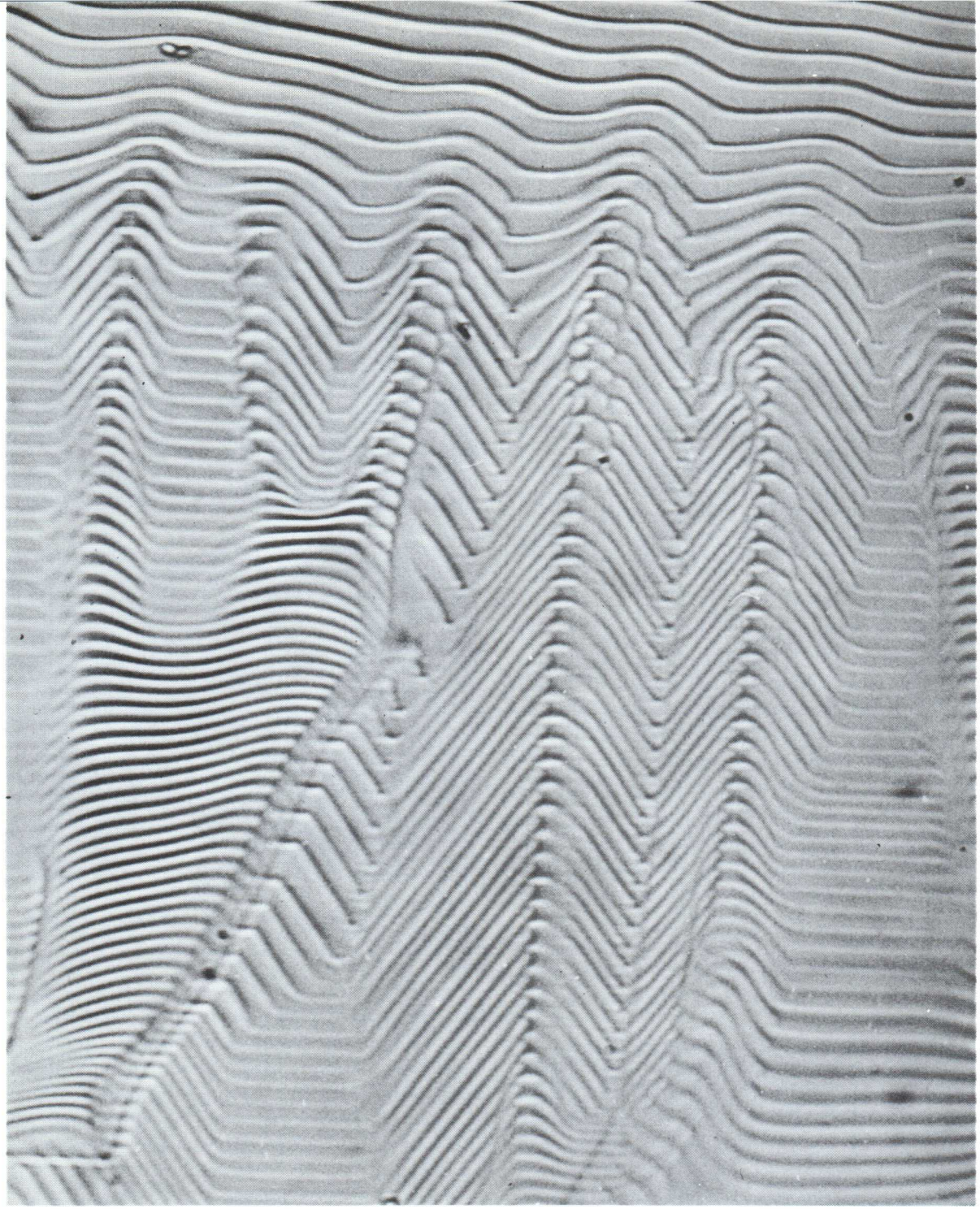
b

Fig. 13. Oscillograms obtained from the oscilloscope O of fig. 9, *a*) evacuated gas cell, *b*) gas cell filled with COS gas. The absorption line in (*b*) occurs at a frequency of approximately 146 Gc/s.

TE_{01n} wavemeter of good construction, the measurement can be made with an accuracy of up to 0.01%. For greater accuracy, more elaborate equipment is necessary [9]), including a frequency standard which is regularly checked. However, the discussion of this does not fall within the scope of this article.

[9]) O. R. Gilliam, Ch. M. Johnson and W. Gordy, Microwave spectroscopy in the region from two to three millimeters, Phys. Rev. **78**, 140-144, 1950.

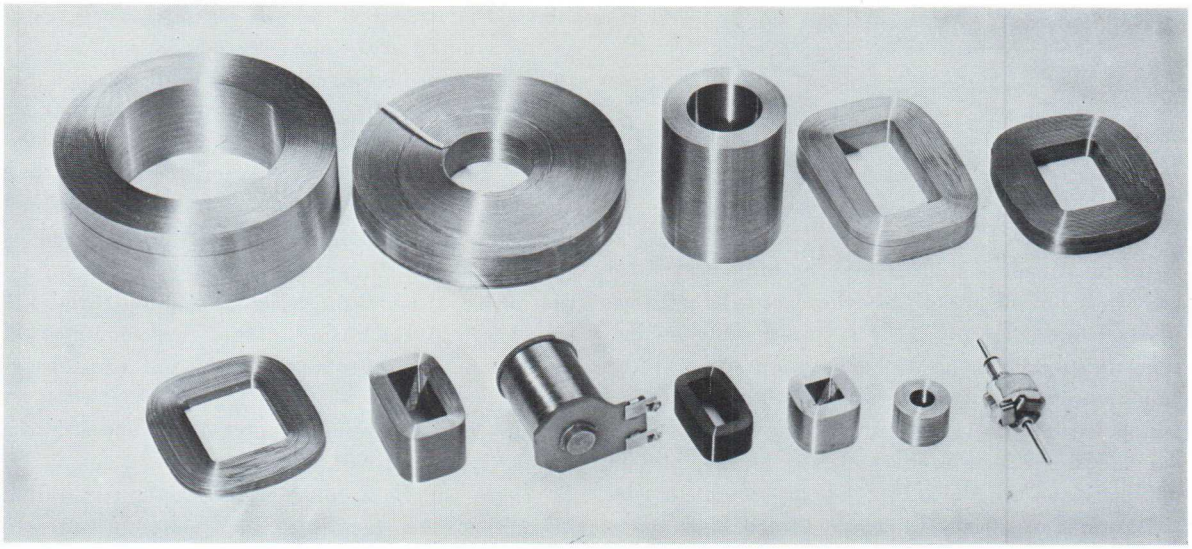
Summary. In continuation of Part I of this article, in which waveguide components for 2 mm wavelength were discussed, Part II describes three measuring set-ups. The first of these is for measuring dissipative losses of the order of 0.1 to 3 dB in microwave components. In the second set-up, the reflection coefficients of unknown impedances are determined by means of a bridge circuit using a hybrid T. With certain precautions, voltage reflection coefficients lower than 0.01 can be measured with reasonable accuracy. The third set-up is used for the experimental determination of absorption lines in gases. The gas carbonyl sulphide (COS), which exhibits an absorption line at about 146 Gc/s, is chosen as an example. Some 2 mm components are also discussed which were not treated in Part I: a crystal detector, a waveguide switch and an isolator.



Inhomogeneities in a doped germanium crystal

This is not a modern painting but a pattern obtained in a germanium crystal heavily doped with gallium (10^{20} Ga atoms per cm^3). The crystal was grown by Czochralski's pulling method; it was pulled in a ⟨100⟩ direction under conditions of constitutional supercooling. These conditions are produced in the melt by suitably raising the pulling rate. A cell structure is formed on the growth interface, with projections which in this case are of pyramidal shape. The concentration of the doping material in the vicinity of a projection is increased. Such changes during the history of the grown crystal are revealed by irregularities in the "striations" which a pulled crystal often exhibits owing to slight thermal or other asymmetries present during the pulling process and which can be made visible for instance by etching. Magnification 400 times.

(See: J. A. M. Dikhoff, Inhomogeneities in doped germanium and silicon crystals, Philips tech. Rev. **25**, 195-206, 1963/64.)



THE ORTHOCYCLIC METHOD OF COIL WINDING

by W. L. L. LENDERS *).

621.3.045.16

Orthocyclic winding is a method developed by Philips for obtaining coils whose turns are stacked in the most compact fashion possible. Such coils have certain particularly good properties including good heat conduction, even distribution of electric field strengths, and the highest possible space factor (which means that the coil dimensions, and hence also the statistical spread therein, are reduced to a minimum).

Any device exploiting electromagnetism contains one or more coils or windings made of insulated copper wire. The performance of the device and certain of its more important characteristics, such as size, weight and price, are to some extent dependent on the success with which the basic ingredients of coil design have been mixed. We are referring to the number of turns, the gauge of the wire, the shape and volume of the coil and the method employed for winding it. This hetrogeneous collection of variables opens the door to a great variety of coil properties. The range of possibilities becomes even greater where, as is usually the case, the magnetic flux produced by the coil passes through a ferromagnetic material, since the magnetic properties of these materials differ considerably and there is further a wide choice in the dimensioning of the magnetic circuit.

One aim that is generally striven for in the coil design is to get the highest possible *space factor*. In this context the space factor F means the ratio between the aggregate cross-sectional area of the turns inclusive of insulation and the area of the rectangle enclosing them ($ABCD$ in *fig. 1*). F is dependent on the shape of the wire (though we shall only concern ourselves here with *round* wire) and on the way in which it is wound. Inevitably there are interstices between the turns, and these represent wasted space, so that F is less than unity even

*) Radio, Television and Record-playing Apparatus Division, Philips, Eindhoven.

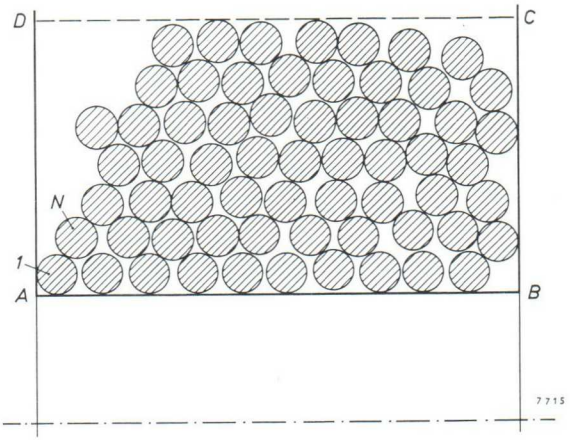


Fig. 1. Cross-section of irregularly wound coil.

in an ideal coil, though it may be only very little less. In a non-ideal coil, in other words in an irregularly wound one (fig. 1), F is considerably smaller than unity. The theoretical maximum space factor can actually be attained with a winding method which has been worked out at Philips and which we call the *orthocyclic* method. The method and its applications form the subject of the present article.

But first of all the importance of a high space factor will be demonstrated: let us take as example a telephone relay or selector embodying a coil which has not been wound by the orthocyclic method, and which provides a required number of ampere-turns at a certain rated voltage. Now let us suppose it proves possible, by adopting a better winding method to accommodate (say) 1.7 times as many turns of the same wire in the same space; the adoption of orthocyclic in place of "wild" winding does in fact result in an improvement of this order. The resistance of the new coil is 1.7 times that of the old; at the same voltage, the current is $1/1.7 \approx 0.6$ times what it was before. The number of ampere-turns remains the same, so that this design requirement continues to be satisfied; but the power consumption of the new coil is only 0.6 times that of the old one. In a telephone exchange, where the number of relays and selectors is large, the resulting economies are appreciable: less power-supply capacity is required, lighter-gauge wiring and cables can be used, and less energy is dissipated as heat.

Coils wound "wild" and coils with interleaved insulation

We shall first discuss two common ways of winding coils. *Fig. 2* shows a coil with the first layer completed: the turns cross the axis of the coil at an angle $90° - \alpha$. They have been wound from left to right, and form a left-handed helix. The second layer will run from right to left; it cannot be wound with the same neatness as the first. It ought to form a right-handed helix. In practice, however, its regularity will continually be disturbed because the wire has a tendency to lie in the grooves between the turns of the first layer, but since the two helices do not interlock, the wire of the second layer can only follow the grooves for short distances. Hence the irregularities will become ever greater as the winding of the coil proceeds, to such a point that it is soon no longer possible to distinguish one layer from the other. It is for that reason that we call this "wild winding".

The advantages of this primitive coil-winding method are that it is quick, that it can be done on a very simple machine, and that little time need be spent on the training of the operator. But examination of fig. 1, which represents a section through a coil wound by the "wild" method, will soon make it plain that this procedure has the following drawbacks:

1) The coil has a much smaller space factor than a neatly wound coil.
2) There will be a considerable random variation in the space factors of such coils. In order to reduce the number of rejects, the design will have to be based on almost the lowest space factor likely to be obtained. The variation in the space factor is associated with a variation in the average length of wire constituting a turn, and hence in the resistance of a coil.
3) Once the winding process starts to "run wild", some turns, particularly those at the sides, will inevitably end up nearer the axis than they

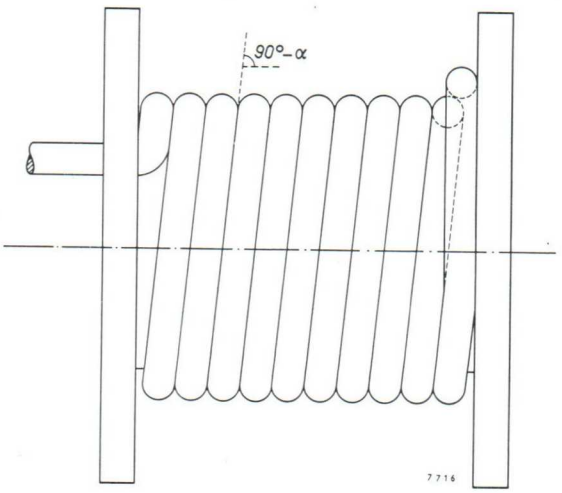


Fig. 2. Single-layer coil wound from left to right and having the form of a left-handed helix. The turns cross the axis of the coil at an angle of $90° - \alpha$. Irregularities will develop when the second layer is wound from right to left, and it will not be long before the winding process "runs wild".

ought to be, so that there will be relatively high voltages between turns lying in close proximity.

In consequence of (3) there is no way of knowing U_n, the highest voltage arising between contiguous turns of an irregularly wound coil with a voltage U across its ends; nor, therefore, is there any way of assessing the risk of breakdown in a coil carrying AC and of corrosion in a coil carrying DC. If the coil were regularly wound, all this would be known: for a winding consisting of l layers,

$$U_n = 2\,U/l. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

(U_n is the voltage between the first turn of one layer and the last turn of the next.) Fig. 1 illustrates the worst case that can arise, that being when the last turn (N) lies on top of the first one (1).

To overcome these difficulties, it is a common practice to lay a strip of paper or other insulating material on each layer as it is completed; sometimes the insulation is inserted every two layers. This can be done automatically, so no loss of time is entailed. The great advantage of using interleaved insulation is that even and uniform layers are obtained: the paper prevents the layer which is being wound from being disturbed by the pattern of the layer underneath. The effect is as if a new start is made with each layer. Eq. (1) holds in this case and, moreover, the interleaved paper improves the breakdown voltage. Yet the space factor remains small: the gain obtained by regularity is cancelled out because the paper takes up some space, and because spaces also have to be left on either side of the coil (*fig. 3*). If the layers were wound over the whole breadth of the paper, the end turns would be liable to slip off.

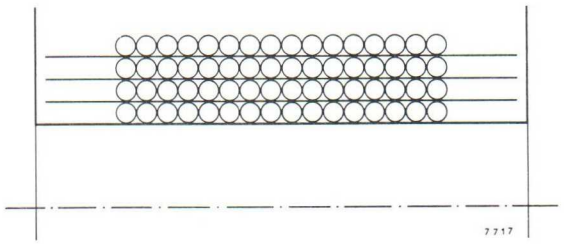


Fig. 3. Cross-section of coil with interleaved insulation. The turns are arranged in an orderly manner but the space factor is still small because the insulation takes up some space, and because some space has to be left free on either side of the layers.

Orthocyclic coils

We have seen that in "wild" coil winding it is impossible for regularity to be maintained after the first layer. The reason is that the tendency of the wire to lie in grooves formed by the preceding layer conflicts with the need for adjacent layers to have opposite helical configurations. The condition relating to orientation can be expressed in another way: if the turns of the first layer lie at an angle of $90° - \alpha$ to the coil axis (fig. 2), then those of the second layer should lie at an angle of $90° + \alpha$. It follows that the difficulty disappears if α can be reduced to zero, in other words, if the first layer can be wound in such a way that each turn, or at least the greater part of each turn, crosses the axis *orthogonally*.

Fig. 4 shows the pattern the first layer would have to assume in this case. Now, with a little care

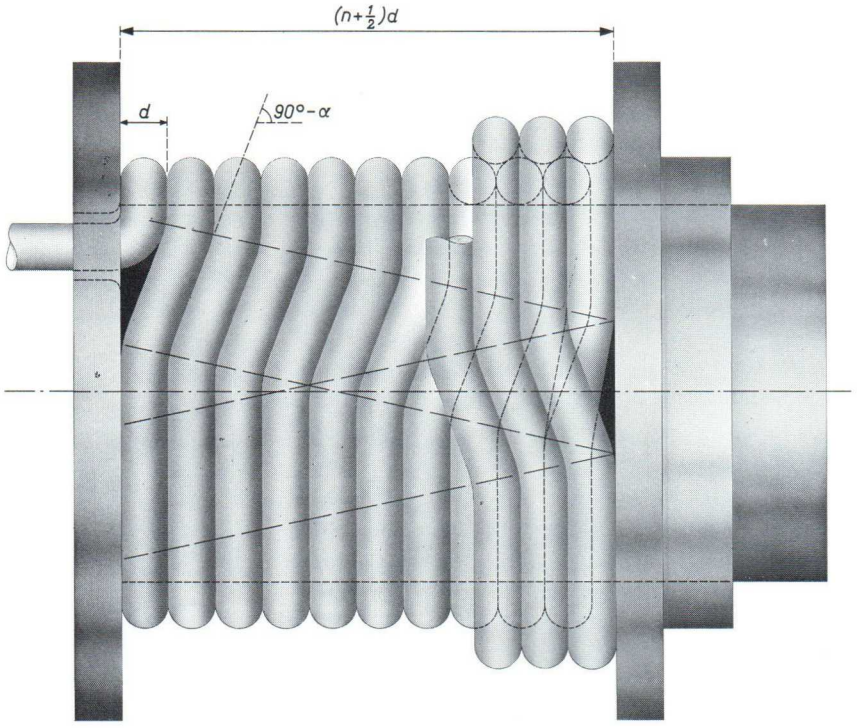


Fig. 4. A round orthocyclic coil at the stage where the first layer and a few turns of the second layer have been wound. Over about 90% of its length, each turn crosses the axis orthogonally. The remaining, inclined portion is known as the crossover; together the crossovers of one layer form a crossover line. Except for their crossover portions, the turns of the second layer lie neatly in the grooves formed by those of the first layer.

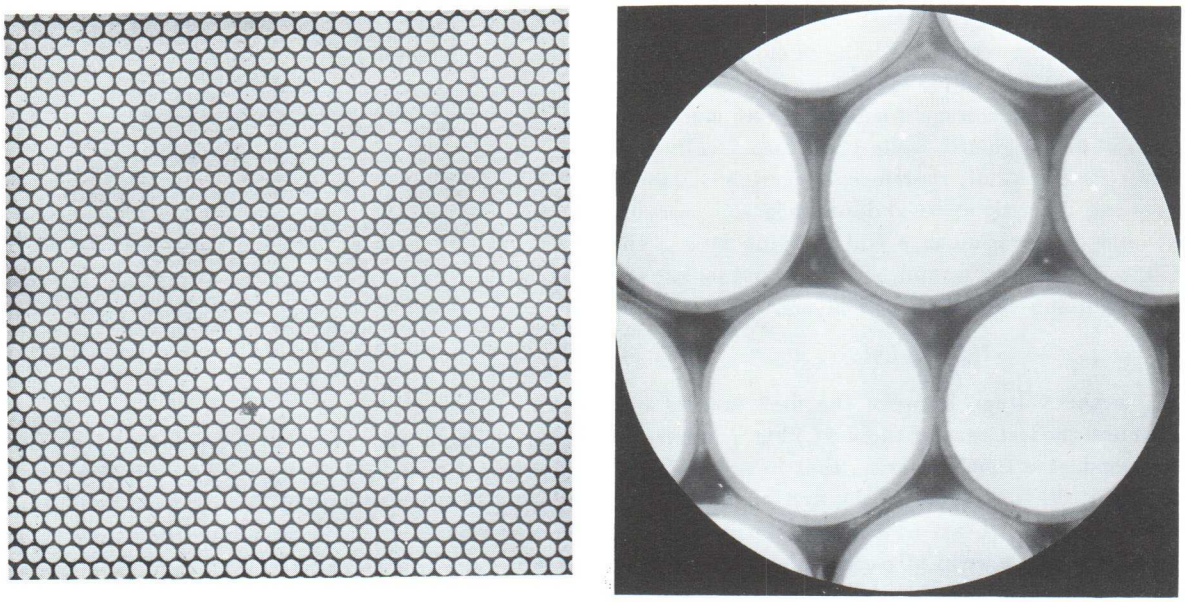


Fig. 5. Section revealed by sawing through an orthocyclic coil wound with copper wire of diameter 50 μm; the photograph on the left is 48 times, that on the right 600 times actual size. In virtue of the perfect orderliness and extreme compactness of the stacking pattern, the highest possible space factor has been achieved.

and ingenuity it is in fact possible to wind the first layer in this fashion, and it has been found that thereafter no difficulty is encountered in stacking succeeding layers evenly; this optimum stacking is the characteristic feature of what we call *orthocyclic winding*, by reason of the fact that the turns must lie orthogonal to the axis. The remarkable regularity and compactness of orthocyclic coils will be evident from the photographs (*fig. 5*) of a section revealed by sawing through a coil of this kind.

Crossovers

Obviously, the orthogonality requirement cannot be satisfied over the whole length of the turn. We use the term *crossover* to denote that section of each turn which lies out of the orthogonal (see fig. 4). Together, the crossovers of one layer form a *crossover line.*

In coils with a circular cross-section the crossover occupies about 10% of the total turn length. The crossover lines run through the coil in a zigzag fashion and their ends, as seen in the cheek of a finished coil, form an Archimedes' spiral (*fig. 6*). In coils with rectangular cross-section, crossovers in the lowest layers occupy one side of the rectangle (*fig. 7*).

Coil thickness and space factor

The thickness of orthocyclic coils is slightly greater in places where crossovers have been made. Let us first consider the places where no crossovers are present. Here it is only the first layer that contributes the full diameter of the wire d to the thickness of the coil. Because they lie in the grooves of the preceding layer, all layers other than the innermost contribute only $\frac{1}{2}\sqrt{3}d$ (*fig. 8*). Accordingly, the crossover-free portion of an orthocyclic coil comprising l layers has a depth of

$$[1 + \tfrac{1}{2}\sqrt{3}(l-1)]d\,,$$



Fig. 6. Round orthocyclic coil. The ends of the crossover lines form an Archimedes' spiral in the cheek of the coil.

which, for a large number of layers, is roughly $\frac{1}{2}\sqrt{3}ld = 0.87ld$.

In considering the portion in which crossovers are present, round and rectangular coils must be dealt with separately. In *round* coils, the number of successive layers whose crossovers overlap does not exceed three; the local increase in coil thickness due to crossovers does not therefore exceed $3(1 - \frac{1}{2}\sqrt{3})d = 0.4d$. Whether this is important or not depends entirely on circumstances. In *rectangular* coils the length of the crossovers, those of the lower layers at least, is the same at that of the side on which they lie (fig. 7*b*). Here, then, the crossovers are stacked one above the other, and the increase in thickness they occasion will be greater than the corresponding increase in a round coil with the same number of layers.

We shall take the symbol F_0 to denote the "pure" space factor of an orthocyclic coil, in which no account is taken of the crossovers and of a second

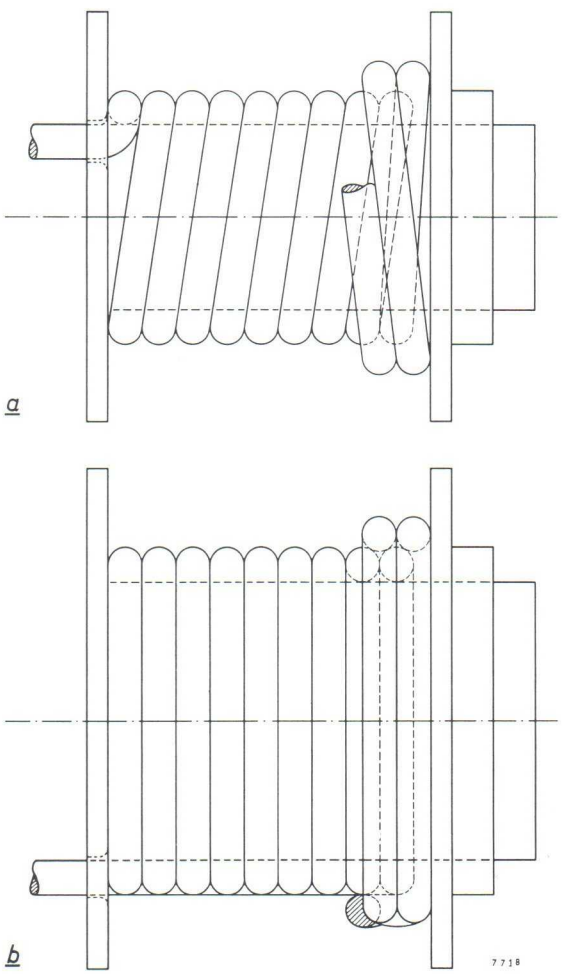


Fig. 7. Two faces of a rectangular orthocyclic coil, (*a*) that on which the crossovers are made, and (*b*) one of three faces on which the turns lie orthogonally.

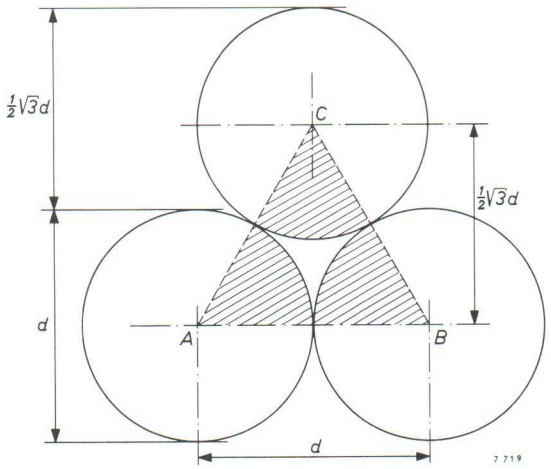


Fig. 8. Section through three contiguous turns in an orthocyclic coil wound with wire having an outside diameter *d*. If the crossovers and the "edge effect" be neglected, the space factor is the proportion of triangle *ABC* occupied by the hatched areas, i.e. 91%.

effect that will be dealt with below. F_0 is easily calculated: it is the ratio of the total area of the three hatched sectors in fig. 8 to the area of the equilateral triangle *ABC*:

$$\frac{\frac{3}{6} \times \frac{1}{4}\pi d^2}{\frac{1}{4}\sqrt{3}\, d^2} = \frac{\pi}{2\sqrt{3}} = 0.91.$$

But so far we have ignored the fact that certain turns of the coil are not completely surrounded by other turns; the grooves in the innermost and outermost layers and in the cheeks of the coil are not utilized. The magnitude of this "edge effect" — to which, of course, non-orthocyclic coils are also subject — depends on the shape of the coil's cross-section. It may be important in the case of long, thin coils (in which the number of turns per layer is large relative to the number of layers) and still more so for flat, disc-shaped coils (in which the number of layers is large relative to the number of turns per layer). Still, even in the latter case, the true space factor F is unlikely to fall short of F_0 by more than one or two percent.

Because the space factor remains much the same from one coil to another, orthocyclic coils exhibit a particularly small random variation in external dimensions. The size of the coils is determined entirely by the outer diameter of the wire, the number of turns per layer and the number of layers.

The coil mandrel

Generally coils are wound on a former or bobbin made of impregnated paper or plastic; the bobbin affords mechanical protection and additional electrical insulation. Such bobbins are not as a rule

up to the standard of dimensional exactitude required for orthocyclic winding. That greater precision is called for by this coil-winding method will be clear from the fact that each layer must consist of exactly the same number of turns n; the distance between the cheeks of the coil former must therefore be $(n + \frac{1}{2})d$, and experience has shown that deviations from this dimension must not exceed $\pm 0.2d$. This requirement is a severe one, particularly where it is a matter of winding very thin wire (with a diameter of 50 μm, for example).

Orthocyclic coils are not wound on a bobbin, then, but on a rigid metal mandrel that has been carefully machined to close tolerances. It is only after removal from the mandrel that the completed coil is mounted on a bobbin and provided with additional insulation.

The wire

In fig. 4 it has been tacitly assumed that in orthocyclic winding no clearance is left between turns of the same layer. This tight packing means that d must remain within particularly close limits. For example, if at a certain point in the wire its outside diameter should increase by 1%, only 50 turns later it would be riding on top of a turn in the preceding layer, instead of lying in the groove; the next turn would slip off the turn beneath, leaving a gap of $\frac{1}{2}d$. Thus disturbed, the regularity of the coil would be irremediably lost, and the disorder would worsen progressively.

The above tolerance of $\pm 0.2d$ for the layer breadth imposes a tolerance of $\pm 0.2d/n$ for the outside wire diameter. If for example d is nominally 100 μm and there are $n = 120$ turns per layer, then $0.2d/n = 0.17$ μm. In wire manufacture, the bare copper is reduced to the desired gauge by drawing through a die. Consequently the spread in the diameter of bare copper wire on a given reel is negligibly small [1]). An "enamel" covering is the only kind of insulation suitable for wire that is to be used for orthocyclic coil winding. The deposition of the "enamel" (which is really a type of varnish) is an easily disrupted process which must take place under strictly defined electrical, chemical and mechanical conditions [2]). It may be regarded as something of an achievement that ways and means have been found of keeping the spread in d to within $\pm 1\%$ for a given reel of enamelled wire. Yet this is still six times more than the tolerance required in the above example.

There are two expedients whereby enamelled wire with an excessive spread in d can nevertheless be utilized for orthocyclic winding: a clearance can be left between turns, or the wire can be redrawn.

1) The variation in d can be allowed for by winding the first layer with a small clearance between the turns. It has been found in practice that the clearance must not exceed $0.03d$, otherwise the turns of the second layer will push those of the first layer too far apart. As can be imagined, the maintenance of an inter-turn clearance of a few microns calls for a rather complicated winding machine which accurately controls the pitch of the winding. We shall return to this problem later.
2) Wire exhibiting an excessive spread in d can be rendered suitable for orthocyclic winding *without* inter-turn clearances by redrawing it, i.e. by passing the enamelled wire through another die. The aperture of the die used must have a diameter equal to the minimum value of d, so that all thicker portions of the wire are reduced to this particular outside diameter.

Redrawing involves a perhaps somewhat surprising phenomenon: the variations in d are transferred, with a change of sign, to the copper core; in places, then, the diameter of the core is reduced in the same proportion as the outside diameter of the enamel covering (*fig. 9*), and this has two undesirable consequences:

a) The copper hardens on account of plastic deformation and so loses some of its pliability; consequently the crossovers tend to lengthen, and it is less easy to control their positioning.

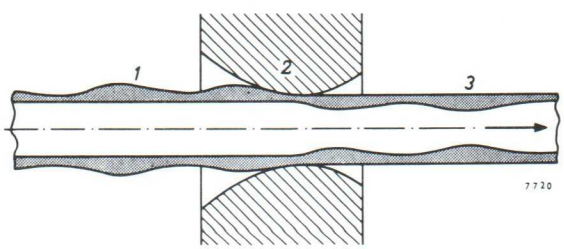


Fig. 9. Schematic representation of enameled wire undergoing the redrawing process. The aim is to get insulated wire whose outer diameter is everywhere the same. Besides producing the desired effect, redrawing transfers to the copper core the variations that were originally present in the outer diameter. *1* represents the wire still to be redrawn, which has a constant core diameter but exhibits variations in outer diameter, *2* is the drawing die, and *3* is the redrawn wire, which has a constant outer diameter but exhibits variations in core diameter. The thickness of the insulation and the variations therein have been exaggerated for the sake of clarity.

[1]) Slightly bigger differences of diameter can naturally be expected in wire from different reels, which has been drawn through different dies.

[2]) R. J. H. Alink, H. J. Pel and B. W. Speekman, Manufacture and testing of enamelled wire, Philips tech. Rev. **23**, 342-351, 1961/62 (No. 11).

b) Since the diameter of the copper has been reduced in all the places where d was in excess, the electrical resistance per metre will have increased; if d was on average $a\%$ too large, the resistance increase will be about $2a\%$.

Attention is being devoted, in wire manufacture, to making the spread in d even smaller than it is at present, so that redrawing will involve less deformation, the above-mentioned disadvantages of the treatment thus being minimized.

In consequence of the variation in the core diameter of redrawn enamelled wire, orthocyclic coils wound from such wire exhibit a certain variation in resistance values. This, however, is generally smaller than the variation in the resistance of coils wound by the "wild" method. In the latter case the lack of uniformity in coil resistance is due, as stated above, to the considerable spread in the space factor and hence in the total length of wire that goes into the coil.

Orthogonal wire feed arrangements in coil winding

As we have seen, the characteristic feature of an orthocyclic coil is that the greater part of each turn lies in a plane at right angles to the axis. Accordingly, feed arrangements for the winding machine must be such that the wire coming off the reel likewise remains at right angles to the coil axis, at any rate within close limits. This aim might be achieved by dispensing the wire via a sheave wheel set up at a sufficient distance from the machine; in practice, the wheel would have to be a good many metres away. Apart from the inconvenience of the set-up, vibration in the long stretch of wire between the sheave and the machine could interfere with the regularity of the winding process. For that reason we make use of a guiding sheave which is free to move laterally, and which maintains the orthogonal feed direction automatically without having to be set up at a considerable distance from the machine. The device is shown schematically in *fig. 10*.

"Thermoplac" wire

The finished orthocyclic coil would fall apart on removal from the winding mandrel if it were not for the fact that "Thermoplac" wire [3]) is used in orthocyclic winding. The insulation of this wire has a thin thermoplastic coating. Before the coil is taken off the mandrel, it is warmed in order to fuse and bond the coatings on contiguous turns. Cooling results in a sturdy self-supporting coil. The thermoplastic film need only be about 2 μm thick. It can withstand the drawing process.

All the following methods are suitable and are in fact employed for heating the coil while still on the mandrel.

a) Baking in an oven.
b) Induction heating in an HF magnetic field (most of the heat is generated in the mandrel and passes to the coil by conduction).
c) Heating the coil by passing through it a current from an external source.
d) Short-circuiting the coil and inducing in it an alternating current (normally of mains frequency).

In methods (c) and (d), as against (b), heat is developed in the coil itself and the mandrel acts as a sink. It is therefore necessary, if either of these

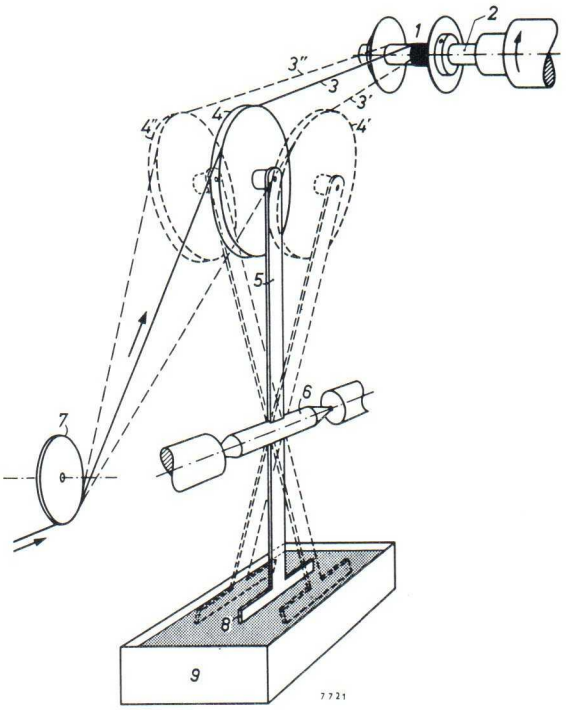


Fig. 10. Automatic arrangement to ensure that wire coming off a reel (not shown) will at all times be orthogonal to the coil axis, irrespective of the position of the turn being wound. The coil *1* is being wound on a mandrel turning about axis *2* and fed with wire *3* via a sheave wheel *4* on an arm *5* pivoting on a pin *6* that is orthogonal to *2*. Wire from the reel goes to *4* via a fixed sheave *7*. The point at which it leaves *7* is in line with *6*. By reason of the tension in the wire, sheave *4* automatically takes up position 4′ when a new layer is started, and moves to position 4″ as the layer is completed. If the small deviation due to the finite length of arm *5* be discounted, the wire dispensed to the mandrel remains orthogonal to axis *2* as it repeatedly passes from position *3′* via position *3* to position *3″*.

An extension to arm *5* carries a vane *8* which is immersed in an oil bath *9*. The oil provides enough damping to prevent transverse oscillations of the arm.

[3]) See the article cited in footnote [2]), p. 347.

two methods is used, to ensure adequate heating of those parts of the coil which are in contact with the mandrel while guarding against overheating of the remainder. In practice this means that the desired temperature must be attained within a few seconds. This is quite feasible with method (c) provided the resistance of the coil is low enough; fast heating of a coil having a high resistance would require a voltage involving the risk of breakdown between the first and last turns via the metal mandrel. Method (d) does not endanger the insulation of the coil because the terminals are shorted. It does, however, necessitate a core carrying a primary winding marrying with the type of coil being manufactured, and the cost of this will only be acceptable if the coils are to be turned out in sufficient quantity.

A closer look at certain aspects of orthocyclic winding

The exact manner in which orthocyclic winding can best be done depends on various circumstances, among these being the shape of the coil (round or rectangular) and the batch size (a small number of coils or enough to justify continuous production). Here we shall divide orthocyclic coil-winding procedures into two classes, those in which the first layer is wound and the pattern established by hand, and those in which this initial phase is done by machine. The two classes correspond more or less to the two cases of production on a small and on a large scale respectively.

Patterning the first layer by hand

If done by hand, orthocyclic patterning of the first layer calls for a certain amount of skill on the part of the operator. This is not a serious drawback so long as the number of coils to be made is small. However, only redrawn wire can be used, since it would be impossible, by hand, to maintain the small inter-turn clearance that is necessary when winding with wire whose outside diameter has not been corrected.

In these circumstances a simple winding machine will suffice, consisting of a spindle running easily in good bearings, a source of mechanical power and a revolution counter. The mandrel must have one fixed and one movable flange, the position of which can be accurately adjusted; the whole must have been carefully machined. Orthogonal wire feed arrangements are necessary (fig. 10). Redrawing of the wire can be combined with the winding operation.

Once the first layer has been wound and the movable flange correctly set, winding can be continued mechanically; it has been found in practice that the subsequent turns automatically adopt the correct "lie", the same regular pattern repeating itself from layer to layer. But care must be taken to see that the regularity of the winding process is not disturbed by incidental troubles such as vibration in the wire coming off the reel and — when very thin wire is being wound — the intrusion of dust particles. Insofar as it is not limited by such interfering effects, the winding speed can be stepped up to about 10 000 turns per minute from the second layer onward.

Forming the pattern mechanically

Mass-production requirements have led to the mechanization of the process whereby the orthocyclic pattern is established in the first and repeated in subsequent layers of the coil. The resulting winding machine is considerably more complicated, as will be clear from the account that now follows.

The separation of the mandrel flanges is so adjusted that there is just enough room for the prescribed number of turns per layer when the wire has its maximum diameter. By leaving a small clearance between turns it is possible to achieve even distribution, over the same breadth, of wire which has not been redrawn and whose diameter is (say) 3% below the maximum value. Even spacing of the turns is effected automatically by a shuttle arrangement that moves intermittently along a line parallel to the axis of the coil.

During the winding of the first layer, the shuttle is motionless for part of each revolution of the mandrel; it is during this interval that the orthogonal portion of the turn is produced. Oversimplifying for a moment, we can say that immediately a crossover is due to begin, the shuttle moves abruptly over a distance s equal to the wire diameter d plus the inter-turn clearance t.

This simplified description only holds good for a shuttle that is close to the mandrel — between its flanges, in fact. But if this was so in practice it would be impossible for the layer to extend from flange to flange, since the breadth of the greave must necessarily be greater than the diameter of the wire it guides on to the mandrel. Accordingly, the distance between the shuttle and the mandrel axis must be at least half the mandrel flange diameter. But if it is placed further away from the mandrel, each time the shuttle advances it must temporarily overstep its cumulating displacement $s = d + t$ by an amount s_e; see *fig. 11*, which shows the crossovers on a rectangular coil. If the shuttle did not travel this extra distance it would pull the wire against the preceding turn, the inter-turn clearance would be locally reduced to zero (along a line running

transversely over the layer), and this part of the layer would be "waisted", the desired regularity thus being lost. The result, in a round coil, would be that the crossover became longer with every successive winding; in a rectangular coil with all the crossovers on one side of the rectangle, the layers would be narrower along one edge than along the other three.

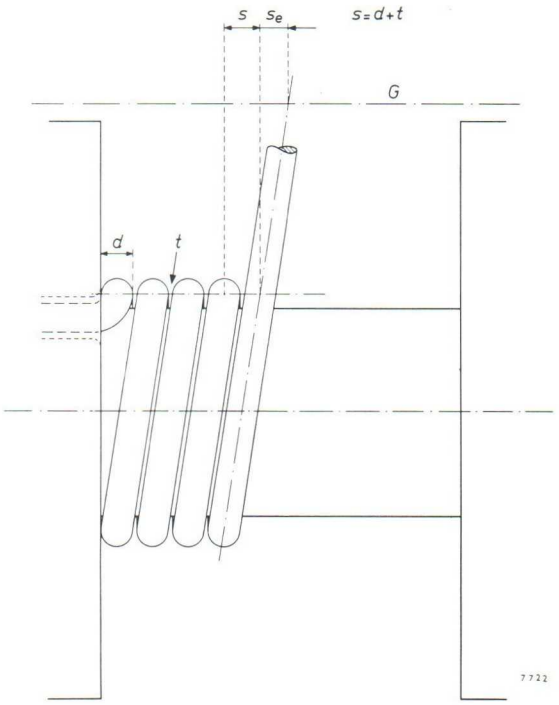


Fig. 11. Winding of the first layer of a coil with an inter-turn clearance t. Shuttle movement takes place along line G. Each time a crossover is begun, the shuttle, having undergone displacement through an interval $s = d + t$, must continue its forward movement, advancing over a further distance s_e, through which it will return as soon as the crossover is completed.

On completion of each crossover the shuttle must move back over the distance s_e.

There is no need for this extra shuttle displacement during the winding of the second and subsequent layers, the "lie" of their turns being determined by the grooves in the preceding layers.

While winding each turn of the innermost layer, then, the shuttle has to perform a somewhat complicated movement. The time available for its performance is the time taken by the mandrel to turn through the angle within which the crossover is made. This sets a limit to the winding speed. The crossover angle is 90° for square coils, and rather more or rather less for rectangular coils, depending on whether the crossovers are being made on the long or short side of the rectangle; for round coils the crossover angle is only 30° to 40°. It follows that the first layer of rectangular coils can be wound more quickly than the first layer of round ones.

Reasons why crossovers in rectangular coils must be confined to one side of the rectangle

Fig. 4 shows how the crossover lines run in zigzag fashion over the surface of a round coil. In rectangular coils the crossover lines have a tendency to do the same thing; but it is necessary to confine them to one side of the rectangle, as in fig. 7*a*, for otherwise the coil cannot remain orthocyclic. This is due to the fact that the length of the crossover, and hence also the inclination of the crossover line with respect to the coil axis, is dependent on the radius of curvature of that part of the turn which constitutes the crossover. If the crossover line should turn the corner, as it were, invading an adjoining face of the coil, it would do so at a point where the radius of curvature of the wire changes abruptly, and the angle of the crossover line with respect to the axis would also alter. The effect is as if the crossover line is bent by the corner of the rectangle, with the result that one turn is missed out. Starting at this point, the irregularity worsens progressively as further layers are added. Finally the winding process "runs wild".

There is a second reason why it is desirable to keep all crossovers on one face of a rectangular coil. Take for example the winding of a transformer with a core made up of E and I-shaped laminations: two sides of the rectangular winding, which we may call A and C, have to be accommodated inside the openings in the core, and the winding depth should therefore be as small as possible on these two sides. The endeavour will accordingly be to confine the crossovers to B or D, the sides of the rectangle lying outside the core.

A special winding mandrel, designed to provide effective control of crossover lines, enables all crossovers to be confined to one side of rectangular coils. The ends of the crossover lines form a pattern in the cheek of the coil, revealing whether the crossovers do in fact lie on one side of the rectangle, or whether they have invaded an adjoining side. They have done so in coil *a* in *fig. 12*, but not in coil *b*.

On sides A and C the winding — we still have a transformer coil in mind — will bulge slightly, to an extent dependent on the tension in the wire coming off the reel, and will therefore take up more space than necessary in the core openings. The bulge can be evened out by compressing the sides in question either during or after the heat treatment. There is no risk of damaging the insulation because

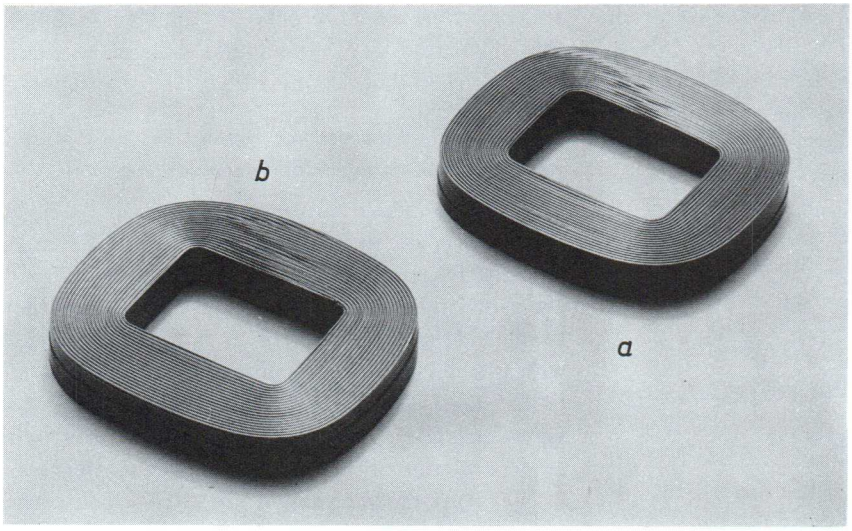


Fig. 12. Rectangular orthocyclic coils. The patterns traced in the cheeks of the coils by the ends of crossover lines reveal that in coil *a* the crossovers have "turned the corner", but in the case of coil *b* have been confined to one side.

sides *A* and *C* are free of crossovers. Coil *a* in *fig. 13* has not been compressed; coil *b* has been squared up on three sides.

Orthocyclic coils compared with types wound by other methods

The foregoing will have made it clear that orthocyclic coils possess two big advantages as compared with other kinds:

1) In orthocyclic coils the space factor attains the theoretical maximum value. In consequence, the random variation in the space factor, and hence in the overall dimensions of the coil, is extremely small.
2) Since each turn of an orthocyclic coil lies in its appointed place, the voltages between given pairs of turns are well defined; moreover, the greatest voltage between contiguous turns is as given by eq. (1). It is therefore possible to ascertain in advance whether U_n will be too high for the wire-covering and whether, accordingly, interleaved insulation will have to be used (in which case there will not of course be any point in winding the coil orthocyclically).

Further important properties which distinguish orthocyclic coils from other kinds will now be enumerated.

3) *Permanence of shape*

Thanks to the thermoplastic bonding agent the turns of an orthocyclic coil coalesce into a very sturdy unit, much stronger than a coil that has been wound "wild" out of wire having a thermoplastic coating of the same thickness. This greater robustness is due to the fact that the turns bind one another over their whole length, and it enables orthocyclic coils to withstand considerable mechanical stresses without loosening of the turns. In virtue of this property orthocyclic coils are particularly suitable for applications in which they are subject to severe vibration, as is the case with the rotor windings of electric motors and moving coils for loudspeakers. We shall revert to the subject of orthocyclic loudspeaker coils below, when dealing with individual applications.

4) *Good heat conduction*

The heat developed in a coil has to be dissipated via its outside surfaces. There is therefore a temperature gradient in the coil. With an eye to the life of the insulation, a ceiling must be fixed for temperatures arising in the coil [4]; but normally the point where the highest temperature prevails is

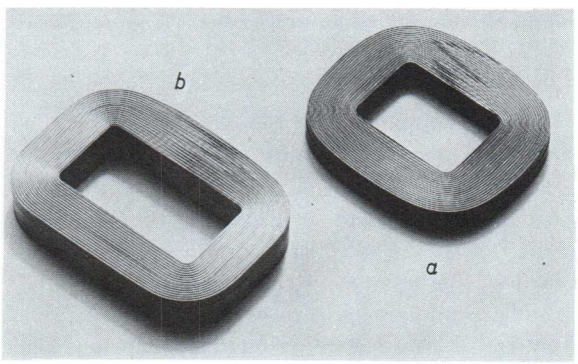


Fig. 13. Rectangular orthocyclic coils in which crossovers have been confined to one side of the rectangle. The faces of coil *a* bulge slightly. The three faces of coil *b* not containing crossovers have been pressed flat, with some saving of space as a result.

[4]) In regard to the connection between life and temperature, see for example T. Hehenkamp, The life of ballasts for gas-discharge lamps, I. Transformers and chokes, Philips tech. Rev. **20**, 59-68, 1958/59.

inacessible for measurement, and in practice, by way of compromise, the average temperature is determined, this being worked out from the easily measured increase in electrical resistance resulting from the warming-up of the coil. The highest temperature arising in the interior of a non-orthocyclic coil may easily be 10 °C above the temperature of the outside surface and 5 °C above the average temperature.

Orthocyclic coils conduct heat so well that in normal use no perceptible difference of temperature arises between the interior and the outside. The thermal conductivity of orthocyclic coils, measured perpendicular to the wire, was found to be 1 W/m^2 per °C/m (or 0.85 kcal/mh°C), which is 20 times the value for coils with interleaved paper insulation. Hence, if a non-orthocyclic coil exhibiting the temperature differences just cited is replaced by an orthocyclic coil, the insulation of the latter will not be endangered if its temperature is allowed to rise 5 °C above the average temperature of the first coil. That higher temperatures are permissible in orthocyclic coils is fortunate in view of the fact that, on account of their better space factor, they occupy a smaller volume and are thus likely to have a smaller surface area than non-orthocyclic coils. By reason of better heat conduction an orthocyclic coil is warmer on the outside than a non-orthocyclic coil with the same internal temperature, and is thus able to dissipate heat at a given rate through a smaller surface area.

5) *High quality factor*

A direct consequence of the high space factor of orthocyclic coils is that less wire of a given gauge is needed to wind a coil with a given number of turns. On account of the smaller winding length the ratio of resistance R to inductance L is smaller than in non-orthocyclic coils; the quality factor $Q = \omega L/R$ is therefore relatively high.

6) *Small variation in inductance and self-capacitance*

In orthocyclic coils each turn has its exactly defined position, and in consequence of this the inductance and self-capacitance differ very little from the nominal values. This is of importance in connection with centre-tapped transformers, which, in telephony and other fields, have to satisfy high standards of electrical symmetry. If such transformers are wound by ordinary methods, exact balance can only be achieved by having recourse to trimming capacitors. These are generally superfluous if the transformer has been wound by the orthocyclic method.

Disc-shaped coils, consisting of narrow layers stacked to a considerable height, are called for in cases where very low self-capacitance is essential.

The random variation in the resistance of the coils has already been mentioned (page 371).

Applications of orthocyclic coils

Our factories have now been employing the orthocyclic winding method for several years. In the early days, of course, knowledge of certain factors affecting the product — tolerances for wire thickness and mandrel dimensions, the correct pattern for the first layer, and the like — was still incomplete. Because these matters had not yet been fully investigated, more had to be asked of the coil-winder in the way of skill and insight into the process than was required by the old, familiar method. For that reason a long training period was necessary, output was inclined to be small, and the rejection rate was sometimes high. As a consequence, at that time orthocyclic coils were comparatively expensive to make, and there could be no question of producing them in quantity. Initially, then, applications of the technique were limited to a few products in the professional class. At present all factors influencing the orthocyclic winding process are well under control, so much so that the method is being adopted in mass production.

A few examples of applications in both categories will now be given.

Professional applications

1) The 600 MeV CERN synchrocyclotron at Geneva is equipped with a modulator that causes the frequency of the accelerating voltage to sweep periodically (55 times per second) over a range extending from about 29 to 16.5 Mc/s. The frequency-modulating device is a gigantic "tuning fork" that acts as a vibrating capacitor [5]). The fork is excited by the field of a set of orthocyclic coils carrying current with a frequency of 27.5 c/s. The whole assembly is in a vacuum.

The main reasons for choosing orthocyclic coils were their high space factor (space was very limited), good heat dissipation and ability to stand up to vibration. During evacuation of the vacuum chamber it was found that the coils released very little gas, the pumping time thus being particularly short.

2) In incandescent lamp manufacture, the bulbs are filled with gas via an electromagnetic valve.

[5]) B. Bollée and F. Krienen, The CERN 600 MeV synchrocyclotron at Geneva, III. The tuning-fork modulator, Philips tech. Rev. **22**, 162-180, 1960/61 (No. 5). — The excitation system is shown in figs 11 and 12 of that article.

For constructional reasons it was highly desirable that the magnet dimensions be reduced to a minimum; an orthocyclic solenoid was therefore indicated.

3) For monitoring steam turbines, electronic equipment has been developed which keeps a continuous record of vibration amplitudes and the dimensional changes affecting various parts of the turbine [6][7]). For example, the eccentricity of the shaft is measured with the aid of a displacement pick-up working on the inductive principle, as illustrated in *fig. 14*. The pick-up windings form part of an AC bridge circuit. For zero eccentricity the bridge should be in balance; any deviation from dead true running should give rise to a signal proportional to the eccentricity of the shaft. It often happens in practice, however, that the temperatures and hence also the resistances of the different windings differ considerably. Consequently the bridge circuit is not always exactly in balance for zero eccentricity, and the signal it yields is not dependent on shaft eccentricity alone. The lower the resistance of coils having a given inductance value — the higher their Q factor, in other words — the smaller the error. In this respect, as we have seen, orthocyclic coils and windings have the edge over non-orthocyclic ones.

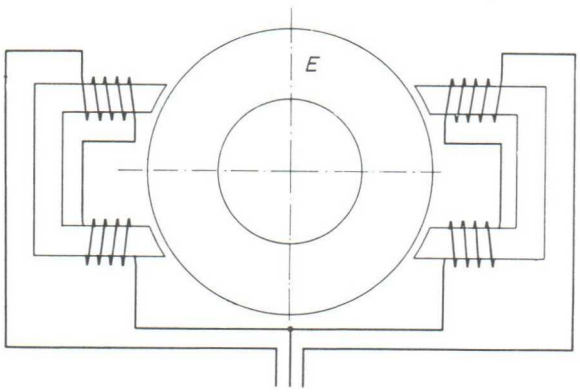


Fig. 14. Inductive pick-up for the continuous detection of eccentricity in a turbine shaft [6]). A disc E of nickel-iron alloy is mounted on the shaft. On either side of the disc is an iron yoke carrying windings that form part of an AC bridge circuit. Provided the two pairs of windings have the same resistance, the signal originating in the bridge will be proportional to the eccentricity of the shaft. A disparity in their resistances, due to temperature differences, gives rise to a measuring error. The higher the Q factor of the windings, the smaller this error is, and that is one of the reasons why the windings are made by the orthocyclic method.

The same considerations apply to other inductive types of displacement pick-ups used on turbines (see figs 9 and 10 of the above-mentioned article [6])).

4) Focusing coils for travelling-wave tubes. In this form of electron tube — one of the kinds used to generate or amplify centimetre waves — an electromagnetic wave propagates itself along a metal helix, through the centre of which passes an electron beam focused by a magnetic field [8]). The field must possess a certain distribution in the axial direction, and it must satisfy stringent requirements in regard to rotational symmetry. In most travelling-wave tubes the field is supplied by permanent magnets. However, in some high-power types a very strong field is required, extending over a considerable distance along the axis, so that an electromagnet is the only practicable means of focusing the beam. The desired distribution of field strength along the axis can be obtained by setting up three differently shaped coils side by side, two disc types being placed at either end of a long tubular coil. Besides offering the familiar advantages of small dimensions with a negligible statistical spread, round orthocyclic coils, by reason of their regular structure, satisfy the rotational symmetry requirement much better than do coils wound by the "wild" method.

Mention may be made here of one of the few cases in which orthocyclic winding was not a success. During design studies for a large X-ray apparatus in which the HT transformer was to be combined with the tube in a single block, it became evident that a smaller and lighter transformer would be desirable. Accordingly, the secondary winding with its interleaved paper insulation was replaced by an orthocyclic winding. However, this secondary had the drawback of excessive self-capacitance. It is true that this difficulty could be got over by substituting for the single secondary a set of flat orthocyclic disc windings connected in series, but the voltage between neighbouring windings was then so high (of the order of 10 kV) that corona and flashover gave rise to further difficulties. Adequate insulation between the disc sections would have meant spacing these so far apart as to cancel out completely the gain obtained by winding them orthocyclically.

Applications in mass production

Over recent years our coil winders have acquired such a degree of expertise in the orthocyclic tech-

[6]) C. von Basel, H. J. Lindenhovius and G. W. van Santen, Electronic equipment for the continuous monitoring of turbines, Philips tech. Rev. **17**, 59-66, 1955/56.

[7]) C. von Basel, Messanlage zur Überwachung von Dampfturbinen, Arch. techn. Messen V 8232-2, 221-224, October 1956.

[8]) J. G. van Wijngaarden, A travelling-wave tube for the frequency band of 3800 to 5000 Mc/s, Le Vide **15**, 36-40, 1960.

nique that there is a growing tendency for it to be employed on a large scale in mass production. A few examples will now be given.

1) Selectors for telephone exchanges. The type U 45a high-speed uniselector (*fig. 15*), made by N.V. Philips' Telecommunicatie-Industrie, Hilversum [9]), embodies one electromagnet which couples the contact wipers to a continuously rotating shaft, and another which arrests them at the desired place. Both the coupling and the stopping magnets are energized by orthocyclic windings, which, as we saw at the beginning of this article, have the great advantage of saving power and so reducing the amount of heat developed, the numerous selectors and relays usually being the main source of unwanted heat in telephone exchanges.

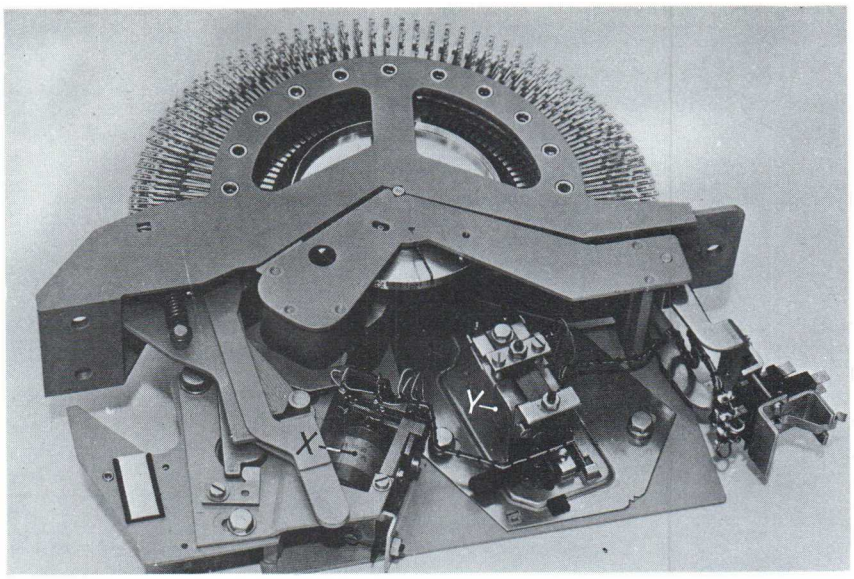


Fig. 15. Type U 45a high-speed uniselector [9]). Both coupling magnet *X* and stopping magnet *Y* embody orthocyclic windings.

2) Moving coils for loudspeakers are wound on a miniature bobbin that is subsequently glued to the cone. Thin-walled as the bobbin may be, it still takes up a certain amount of space in the air gap of the magnet, where even fractions of a millimetre count. Furthermore, the dimensioning of the gap has to allow for the deviations from exact circular shape which are exhibited by most bobbins in consequence of their lack of rigidity in the radial direction.

Thanks to a method which has been developed for attaching self-supporting coils to loudspeaker cones, and which provides a particularly strong bond between the two, an immediate advantage of winding the moving coils by the orthocyclic technique is that the bobbin can be done away with. If the coil consists of several layers, as it is bound to do if destined for a high-impedance speaker [10]), then orthocyclic winding will also reduce its thickness appreciably. In addition, the radial rigidity of orthocyclic coils is much greater than that of coils of older design. For various reasons, then, a narrower air gap suffices for an orthocyclic moving coil, and this pays off either in enhanced efficiency or in reduced magnet size. Another advantage is that orthocyclic coils are better able to stand up to vibration.

3) Self-supporting (but non-orthocyclic) moving coils have long been used in electrodynamic microphones. If the moving coil is wound orthocyclically the air gap can be narrowed, with a resulting increase in the sensitivity of the microphone.

4) In the winding of field coils for moving-iron voltmeters and ammeters, the change-over to the orthocyclic technique has allowed so many additional turns to be accommodated that the sensitivity of the meters is 2 to 2.5 times as great as before.

5) Being battery-fed, portable electronic products must be small in size and have the lowest possible power consumption. When an electric motor for battery gramophones and tape recorders was being designed, it was found that an orthocyclic rotor winding saved so much space that a much lighter rotor could be made if the "wild" method were abandoned in favour of the orthocyclic one. Also, because less iron was present, losses due to eddy currents were smaller, resulting in a good 25% saving in power consumption. A second advantage of orthocyclic rotor windings is the negligible variation in their weights, in virtue of which they are unlikely to be responsible for unbalanced distribution of rotor masses. *Fig. 16* shows the rotor with its three orthocyclic windings.

6) Ballasts for "TL" lamps in aircraft. By reason

[9]) J. M. Unk, A high-speed uniselector for automatic telephone exchanges, Philips tech. Rev. **18**, 349-357, 1956/57.

[10]) J. Rodrigues de Miranda, Audio amplifiers with single-ended push-pull output, Philips tech. Rev. **19**, 41-49, 1957/58.

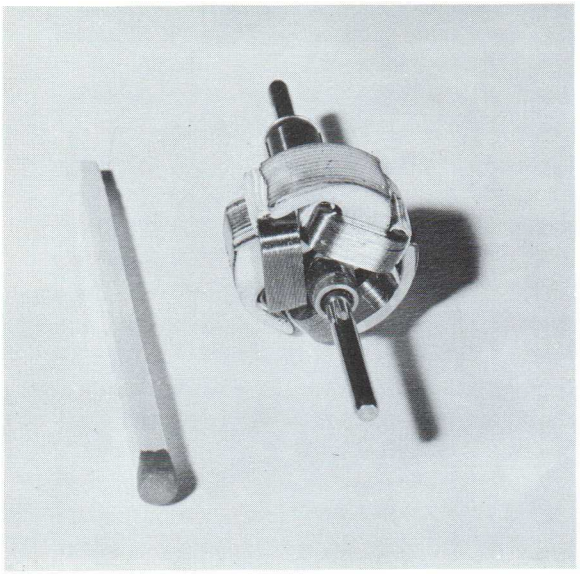


Fig. 16. Rotor of the electric motor incorporated in battery-operated portable gramophones and tape recorders. Employment of the orthocyclic technique for the three armature windings has allowed the size of the motor and its power consumption to be reduced.

of their high efficiency tubular fluorescent lamps are being employed more and more in the public transport sector [11]), including aviation. Passenger-carrying aircraft can now be equipped with (say) twenty 40 W "TL" lamps. In common with everything else that goes to equip aircraft, the weight of the ballasts must be cut to an absolute minimum. Employment of the orthocyclic technique for winding the choke coil has made it possible to bring down to 195 grams the weight of the ballast (consisting of a choke coil and a capacitor) for a 40 W lamp, rated for 115 V at 400 c/s, the standard aircraft electricity supply. The ballast in question may be seen in *fig. 17*.

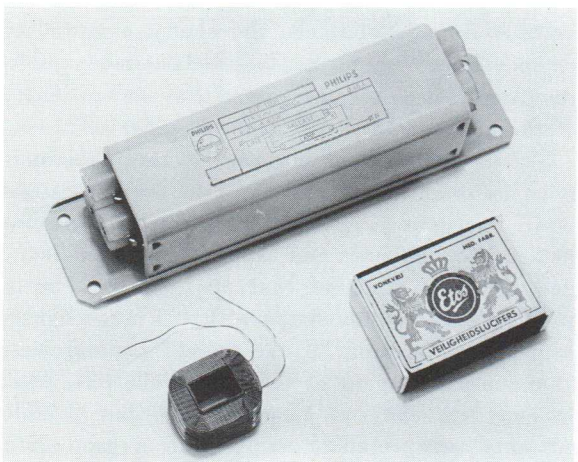


Fig. 17. Ballast for a 40 W "TL" lamp to work on the standard aircraft supply of 115 V at 400 c/s. Altogether, the choke coil and associated capacitor weigh only 195 g. Below, left, the orthocyclic choke coil.

7) Electromagnetically-controlled air valves. A

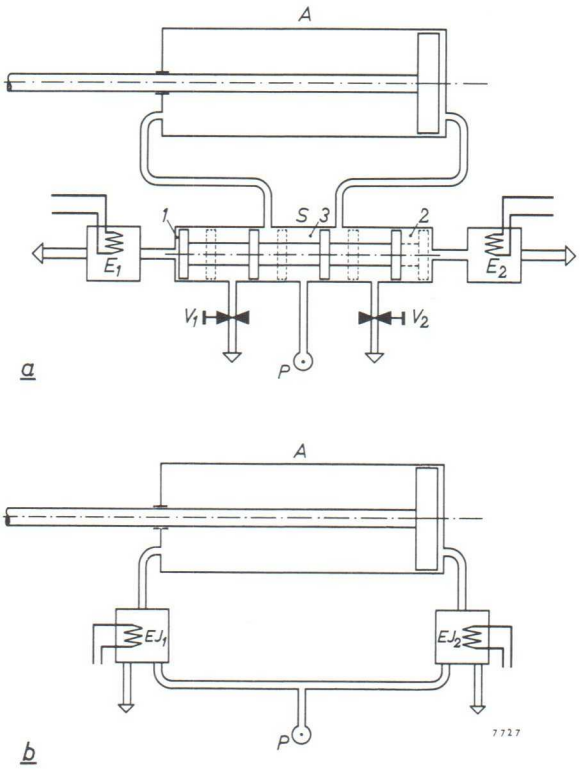


Fig. 18. *a*) Air cylinder A and associated pneumatic equipment. S piston valve, E_1-E_2 solenoid-operated air-control valves, V_1-V_2 adjustable bleed chokes. Compressed air enters the system from pipe P. When the pistons in valve S occupy the positions indicated by fully-drawn rectangles, the space in A to the left of the main piston is connected to P, and the air to the right of the main piston can escape via V_2. In these circumstances the cylinder piston is at the extreme right of its stroke. Spaces *1* and *2* in S are connected (by ducts not shown in the diagram) to space *3*, the pressure in which is at all times that of the compressed-air supply. So long as control valves E_1 and E_2 are closed, then, the air in *1* and *2* is at the same high pressure as that in *3*.

To shift the cylinder piston to the left, control valve E_2 is opened, whereupon air escapes from *2*, the valve pistons move into the positions indicated by the broken-line rectangles, compressed air from P is free to pass via *3* into the space on the right of the cylinder piston and the air in the left-hand side of A exhausts through V_1. The speed with which the cylinder piston travels to the left can be regulated by adjusting V_1. To return the cylinder piston to the right, E_2 is closed and E_1 is opened.

b) In the new system, control of air cylinder A is effected solely by means of EJ_1 and EJ_2, two "Electrojet" valves made by the Martonair Company. These supersede air control valves E_1 and E_2 as well as piston valve S, and accordingly their bore must be at least as great as that of S in (*a*). Having adjustable outlets, the "Electrojet" valves are further able to take over the function of bleed chokes V_1 and V_2. They embody orthocyclic coils, these having been chosen on account of their small size and good heat dissipation.

[11]) L. P. M. ten Dam and D. Kolkman, Lighting in trains and other transport vehicles with fluorescent lamps, Philips tech. Rev. **18**, 11-18, 1956/57.

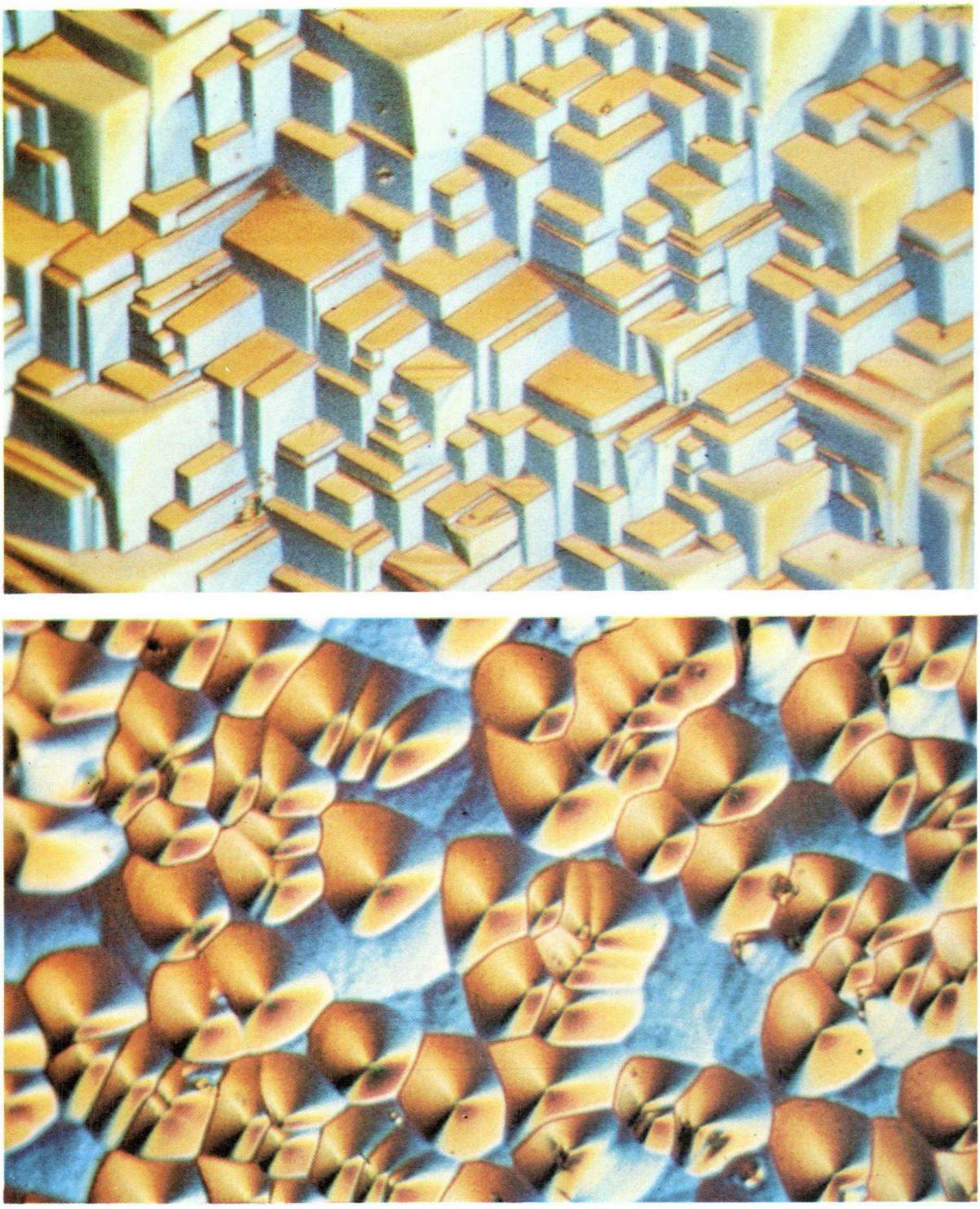
favourite practice in machine building is to fit an air cylinder to supply mechanical power for certain movements which, as a rule, are initiated by switching an electric circuit. Until recently the air cylinder, essentially a simple device, had to be used in conjunction with a pneumatic circuit comprising, in addition to a filter and lubricating unit, a piston valve controlled by one or two solenoid-operated air control valves, one or two adjustable bleed chokes and the associated piping and fittings. An example of a circuit of this kind (many variations are possible) is given in *fig. 18a*.

These arrangements can be greatly simplified by using the "Electrojet" system designed and marketed by the Martonair Company of Amsterdam. Fig. 18*b* shows the "Electrojet" version of the circuit in fig. 18*a*: it will be noted that in addition to the air cylinder itself, only two electromagnetic valves of the new type are necessary, these having taken over the functions of the piston valve, the two air control valves and the two bleed chokes. The overall dimensions of the "Electrojet" valve are no greater than those of the air control valve in the original circuit. The new valve must however have a much bigger bore in order to do the work of the piston valve, and the electromagnet actuating it must therefore develop a stronger pull. One of the ways in which increased pull has been obtained is by using an orthocyclic solenoid which, thanks to good heat conduction, can be heavily loaded without overheating, and is thus able to supply the required higher number of ampere-turns.

The bore of the "Electrojet" valve is large enough for most of the air cylinders on machines used at Philips. It passes enough air at 5 atm to displace a 4″ diameter cylinder piston at a speed of about 17 cm/s against a force of 300 kg, or to enable a $1\frac{1}{4}$″ piston to travel at about 170 cm/s while overcoming a force of 30 kg.

The above examples will doubtless have sufficed to demonstrate the advantages of the orthocyclic winding method, and to illustrate the variety of its applications. It need hardly be added that the method can profitably be employed in many other cases.

Summary. The term "orthocyclic" is applied to a coil-winding method whereby the greater part of each turn is made to lie orthogonally with respect to the coil axis. Provided the correct pattern is established in the first layer and certain precautions are taken, subsequent layers will automatically fall into the same regular pattern, each turn fitting into its appointed place, a groove in the layer beneath. The turns are stacked in the most compact fashion possible, in consequence of which (1) the space factor attains its theoretical maximum value, (2) the highest voltage existing between contiguous turns is fully defined, and (3) the dimensions, inductance and self-capacitance of individual coils of the same type fall within very close limits. It is possible by winding with "Thermoplac" wire to obtain strong self-supporting coils which have invariable dimensions and stand up well to vibration. The thermal conductivity of orthocyclic coils is so high (1 W/m °C, or 20 times that of coils with interleaved paper insulation) that they display no appreciable temperature gradient, and their temperature can safely rise to a value higher than the average permissible in non-orthocyclic coils. In conclusion, various applications of orthocyclic coils (to focusing coils for travelling-wave tubes, electromagnet windings in uniselectors, field coils in moving-iron meters, rotor windings in small motors, etc.) are discussed and the resulting benefits enumerated.



Etch pits on a zinc-sulphide crystal

The cubic modification of ZnS has a polar axis in the ⟨111⟩ direction: the crystal is built up of alternate layers of zinc and sulphur atoms, stacked in that direction. This polarity appears in a surprisingly beautiful way in the macroscopic-chemical behaviour of such a crystal: etching with hydrochloric acid of the opposite end faces perpendicular to the polar axis of one ZnS-crystal results in two distinctly different pictures, although the pattern of etch pits in both cases is basically the same. Each pit is bounded by {100} planes (rather rounded in the lower photograph) from which the triaxial symmetry of the ⟨111⟩ axis can easily be recognized. Photographs taken by Nomarski's interference-contrast method, magnification approx. 800 times.

(See: A. J. Eland, Etch pits on a zinc-sulphide crystal, Philips tech. Rev. **24**, 61-63, 1962/63.)

THE ALLOY-DIFFUSION TECHNIQUE FOR MANUFACTURING HIGH-FREQUENCY TRANSISTORS

by P. J. W. JOCHEMS *).

621.382.333

The highest frequency at which a transistor can provide a reasonable signal gain depends amongst other things on its dimensions; the higher the frequency, the smaller the transistor has to be. The manufacture of transistors for operation at 100 Mc/s necessitates strict control of dimensions which are of the order of tens of microns and even of microns in the case of base width. Simplicity and reliability are conditions for success in the mass-production of transistors of this kind. The requirements are very largely satisfied by the alloy-diffusion technique, which Philips have now been using for some years in the manufacture of HF transistors.

In connection with microminiaturization, the technique has also been used on an experimental scale for making an almost complete three-stage amplifier out of a single germanium crystal.

The invention of the transistor in 1948 in the Bell Laboratories [1]) initiated a tremendous surge of development in the semiconductor field. The first transistors were of the *point-contact* type, and consisted of a thin plate or wafer of crystalline germanium with two wires touching it at points about 50 μm apart (*fig. 1*). One wire acted as emitter, the other as collector. The name "base" was given

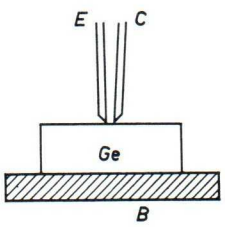


Fig. 1. Schematic representation of a point-contact transistor. *Ge* germanium crystal. *E* emitter. *C* collector. *B* base contact.

to the crystal itself. Attempts to account for the action of the point-contact transistor (even now, a completely satisfactory explanation has still to be found) brought certain facts to light that formed a foundation for the theory of *P-N* junctions published in 1948 by Shockley — also of the Bell Laboratories [2]). In the same paper Shockley described the structure and properties of the *junction* type of transistor. Transistors made in accordance with these theoretical considerations were found to possess the properties predicted for them. It is fortunate that the theory of the junction transistor is well understood [3]), for this knowledge reveals what possibilities there are of giving a transistor the properties that are desirable in a given application.

Almost all transistors being made at the present time are of the junction type. They consist of a small block, disc or rod of monocrystalline material, generally germanium or silicon, containing the requisite *P* and *N* regions. To perform the various functions allotted to them in radio and television receivers, switching circuits, computers and so on, they have to meet all sorts of requirements as to frequency range, output power, etc. These require-

*) Philips Research Laboratories, Eindhoven.

1) J. Bardeen and W. H. Brattain, The transistor, a semiconductor triode, Phys. Rev. **74**, 230-231, 1948.

2) W. Shockley, The theory of *P-N* junctions in semiconductors and *P-N* junction transistors, Bell Syst. tech. J. **28**, 435-489, 1949.

3) See for example F. H. Stieltjes and L. J. Tummers, Simple theory of the junction transistor, Philips tech. Rev. **17**, 233-246, 1955/56. The influence of frequency is not dealt with in the article.

ments lead to differences in the dimensions of the transistor and in certain characteristics of the P and N regions of which it is made up, notably their specific resistance, and also in the manner of mounting the crystal. For example, transistors designed to handle high powers are invariably mounted in such a way as to ensure the closest possible thermal contact with the surroundings.

A great deal of effort has been devoted to finding and developing methods which allow P-N junctions to be made at the desired places in monocrystalline material, and which at the same time enable the characteristics and the geometry of the P and N regions to be kept under strict control. In the present article we shall start by discussing some of these methods. They were originally developed for making transistors operating at frequencies below 10 Mc/s. To render them suitable for making transistors that would work satisfactorily at still higher frequencies, these methods have had to undergo all kinds of refinements. In this article we shall deal in some detail with one of the more elaborate techniques. It is a version, developed in Philips Research Laboratories, of the "alloy-diffusion" technique and which for some years now has been employed for mass-producing various types of transistors to operate in a frequency range extending up to 200 Mc/s. These types have become known as "pushed-out base" transistors, for reasons that will shortly become evident.

The alloy-diffusion technique is also suitable for making what are known as "solid-state circuits", a form of microminiaturization. The experimental construction of a three-stage solid-state amplifier will be described by way of example.

Methods of making P-N junctions and junction transistors [4])

The addition of impurities in crystal-pulling

The earlier methods entailed forming the P-N transitions while the monocrystal was being pulled. In one such method the starting material is a germanium or silicon melt in which donor material (Sb, As or P) has been dissolved; naturally, a crystal pulled from this melt will be of type N. However, at a given instant enough acceptor material (Ga, In or B) is added to the melt to make the acceptor concentration greater than the donor concentration. Accordingly, the next section of crystal to grow is of P type.

[4]) For a fuller review, containing many references to the literature, see the article by W. C. Dunlap, Methods of preparing P-N junctions, which appears as Chapter 7 in L. P. Hunter, Handbook of semiconductor electronics, 2nd impression, McGraw-Hill, New York 1962.

In the making of transistors by this method, the P material is only allowed to grow to the point necessary to give a base of the desired thickness. The melt is then dosed with a sufficient amount of donor to start growing N-type material again. A large number of transistors can be obtained by sawing up the resulting crystal (*fig. 2*). Leads still have

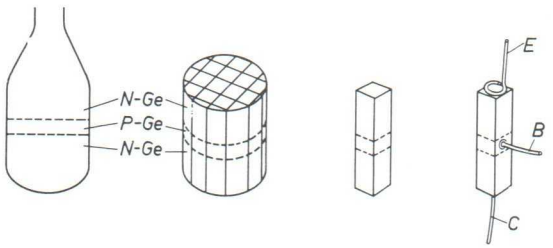


Fig. 2. Diagrams illustrating the manufacture of "grown" transistors. A P-N and an N-P junction have been created in a large monocrystal while it was being pulled from the melt (left). The central portion of the crystal, containing the junctions, can be sawn up into a large number of transistors. To these, electrode leads E, B and C have to be attached. It is no easy matter to attach lead B to the thin base layer (extreme right).

to be fitted to emitter, base and collector of such a "grown" junction transistor, and it is no easy matter to attach a wire to the base layer, which may be only 40 μm thick. For one thing, the neighbouring P-N junctions are easily damaged.

Transistor manufacture took a big step forward when methods were developed for making P-N junctions *after* the crystal had been grown. It now became possible to saw up a large homogeneous block into crystals of suitable dimensions, and to make the required P-N junctions in these smaller units.

The alloying method

The widely used alloying method is one way of producing P-N junctions in a homogeneous germanium crystal. The crystal can either be of P-type or of N-type germanium. If it is of N-type, a pellet of acceptor material — usually indium — is placed upon it, and the whole is heated to 500 or 600 °C. Having a melting point of about 150 °C, the indium liquefies while the germanium (melting point about 950 °C) remains in the solid state. The molten indium takes up germanium into solution and consequently eats into the solid material (see the hatched area in *fig. 3*). The amount of germanium taken up is obviously proportional to the amount of indium and to the solubility of germanium in liquid indium. This solubility, which increases with temperature, can be read from the Ge-In phase diagram. The molten indium stops advancing as soon as it has become saturated with germanium. Various measures can be taken to

ensure that its advance will be even and regular (i.e. over a plane front, see fig. 3), one being to apply the indium to a (111) face of the germanium crystal. The area in contact with the indium pellet can be restricted with the aid of a jig. By a suitable choice of contact area, indium quantity and temperature, one can determine in advance the depth to which the indium will penetrate (h in fig. 3). As the system cools, the solubility of germanium in indium decreases, and the germanium recrystallizes. It is redeposited on the undissolved germanium, refilling the hole in the monocrystal. However, the deposited germanium now contains a little indium, only a tiny residue, it is true, the concentration being of the order of 10^{-4} at. % [5]), but this is enough to make

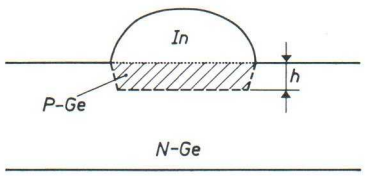


Fig. 3. A P-N junction formed in N germanium by the alloying method using indium (In) as acceptor element. In this and subsequent figures of the same kind, a dashed line represents a P-N junction and a dotted line represents normal electrical contact.

the newly formed recrystallized layer a well conducting P region. A P-N junction has therefore been created.

A transistor can be made by alloying into both sides of the germanium wafer (*fig. 4*). The base of the transistor is formed by the original N germanium lying between the two P regions. The thickness of the base can be controlled via the indium penetration depths and the thickness of the wafer. It is usual to make the collector region rather more extensive than the emitter region, as otherwise some of the minority charge-carriers injected by the emitter might miss the collector and recombine at

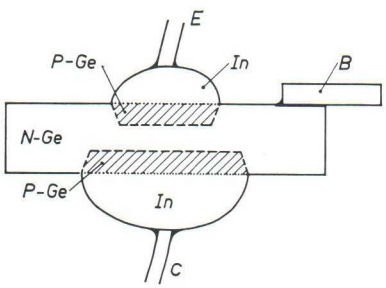


Fig. 4. Section through a transistor made by the alloying method.

[5]) Graphs showing the solubility of various metals in solid germanium and in solid silicon may be found in F. A. Trumbore, Bell Syst. tech. J. **39**, 205-233, 1960.

the opposite face of the crystal without having made any useful contribution to the action of the transistor.

A great advantage of the alloying method is that attachment of the electrode leads to the indium presents no difficulty. The indium pellets are still present as bulges on the surface of the wafer, offering a convenient means of contact with the emitter and collector regions. It is an easy matter to solder leads to these bulges; the base lead can be attached elsewhere on the crystal. The alloying method is used for many of the transistor types manufactured by Philips.

The diffusion method

A second method of creating P-N junctions in a ready-grown crystal is based on the principle of acceptor diffusion into an N crystal or, alternatively, that of donor diffusion into a P crystal. Generally, donors diffuse through germanium at a much faster rate than acceptors; in silicon it is just the other way about. Accordingly, the starting material usually taken for germanium transistors is P germanium, into which donors are allowed to diffuse. *Fig. 5* shows, by way of example, how the donor element antimony diffuses into P germanium held at 780 °C, in the presence of antimony vapour in equilibrium with its solid phase at 600 °C. The P-N junction occurs at the place where the donor concentration is equal to the acceptor concentration. Under the conditions to which these diagrams relate, an N layer 5.5 μm thick takes a quarter of an hour to form.

Diffusion is a slow process dependent on temperature and time. The accurate control of the diffusion process that is possible, makes this method very suitable for obtaining thin layers reproducibly. Characteristic of such layers is the gradual fall-off in impurity concentration from the crystal surface inwards.

Transistors for high frequencies

The higher the frequency at which a transistor must provide a reasonable gain, the smaller are the values acceptable for the various capacitances in the structure — the diffusion capacitance and the collector- and emittter-barrier capacitances [6]). One way of obtaining small capacitances is to make the transistor smaller. Very small size is accordingly a feature of transistors for high frequencies; for example, a transistor designed to give adequate

[6]) The factors limiting the frequency range of transistors are dealt with in an article by M. Beun and L. J. Tummers, on transistor behaviour with increasing frequency, which is to appear in this review in the near future.

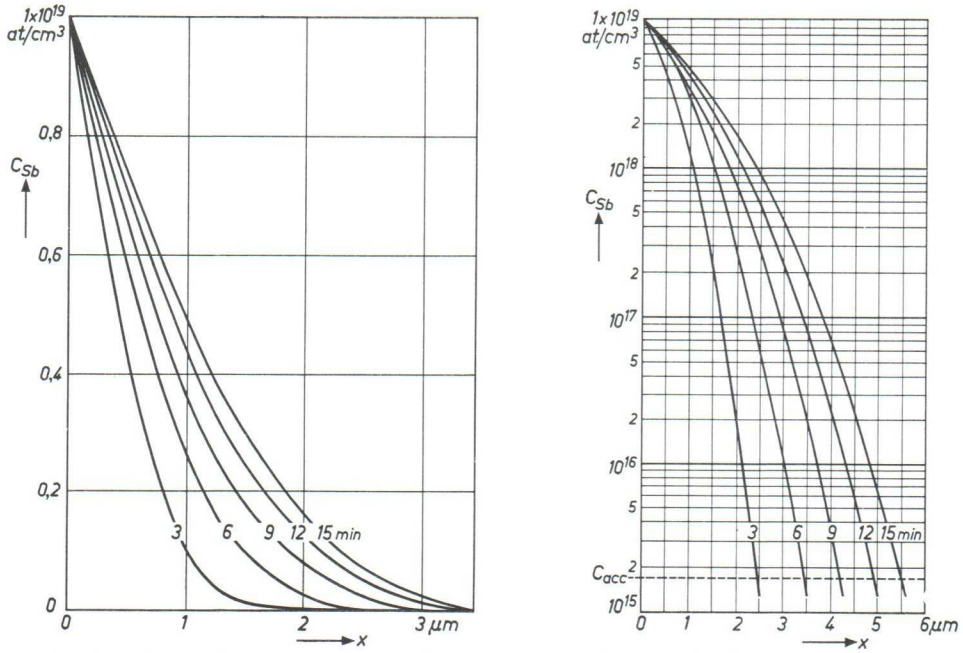


Fig. 5. Diffusion of the donor element Sb into germanium. The germanium is held at 780 °C in Sb vapour that is in equilibrium with solid Sb kept at 600 °C. Under these conditions the Sb concentration at the surface has a value close to the maximum possible at 780 °C, (about 10^{19} Sb atoms per cm³). The Sb concentration C_{Sb} has been plotted as a function of x, the depth in the germanium crystal, after 3, 6, 9, 12 and 15 minute minute periods of exposure to the Sb vapour. The concentration scale is linear in the left-hand and logarithmic in the right-hand diagram. The latter reveals that in germanium with an acceptor (e.g. In) concentration of 1.7×10^{15} atoms per cm³, corresponding to a resistivity of 2 ohm cm, the N layer attains a thickness of 5.5 µm after 15 minutes.

amplification at frequencies as high as 100 Mc/s must have a base no more than a few microns thick. Since the base current flows *laterally* through the base layer, it will be obvious that a thin base is unfavourable towards the low base resistance that is also a requirement for HF transistors.

Transistors for frequencies above about 10 Mc/s cannot be manufactured satisfactorily by the alloying method described above. It is no easy matter to control the penetration depth of the emitter and collector regions (see fig. 4) so accurately that one can rely on getting a reproducible value of a few microns for the thickness of the N region remaining sandwiched between collector and emitter; deviations from nominal in the wafer thickness and the two penetration depths, will be reflected in the residual base thickness. Another unfavourable aspect of alloying is that the penetration depth would have to be large compared with the base thickness, because wafers thinner than about 50 µm cannot be handled.

But even if these difficulties could be overcome, the designer's difficulties would still not be at an end. In an alloyed transistor the resistivity ϱ_B of the base material is always large with respect to that of the collector material. This leads to conflicting requirements regarding the base resistivity of a transistor to operate at high frequencies. ϱ_B has to be large to satisfy certain conditions, and small to satisfy others. Thus a compromise has to be found, but the best possible compromise on ϱ_B will still not raise the limit of the usable frequency range beyond about 10 Mc/s.

The way round this is to make the resistivity of the collector material a great deal higher than that of the base. We shall see below that this can be done by using a diffusion technique to produce the base region.

If the resistivity ϱ_C of the collector material is much smaller than ϱ_B, the collector barrier extends mainly into the material of the base. The thickness of the barrier is proportional to the square root of the voltage across it. Since the base is extremely thin there is a danger of "punch-through" occurring at a relatively low voltage, the collector barrier moving through the base and coming into contact with the emitter barrier. To obviate this danger, and also in order to keep down the base resistance, ϱ_B must be made small. But the advantage of so doing is nullified because the smaller ϱ_B becomes, the lower will be the voltage at which the barrier breaks down and the higher its capacitance [7]).

[7]) More light is cast on these questions in the article cited in [6]). See also the article by B. N. Slade, Device design considerations, which appears as Chapter 10 in L. P. Hunter's book [4]), and in which a numerical example is worked out.

However, if the material of the collector has a resistivity very much greater than ϱ_B, the collector barrier will extend mainly into the collector region. In these circumstances it is possible to choose ϱ_C as large as is necessary to raise the collector barrier breakdown voltage, and to lower the collector barrier capacitance to the appropriate safe levels. ϱ_B can then be made small enough for an acceptable value of the base resistance to be obtained.

Diffusion and alloying in separate stages

One possible procedure is to take a wafer of *P* germanium whose resistivity makes it suitable collector material, and to allow donors (Sb or As) to diffuse into it, thereby forming a surface layer of *N* germanium (*fig. 6a*). The *N* layer is etched away at the place where the collector contact is to be made (fig. 6*b*). The emitter region and the base contact are then produced by the alloying method, on the other side of the crystal (fig. 6*c*). Two indium pellets can be used for this purpose, one containing gallium in solid solution and the other containing antimony (or arsenic). Alloying is done in the same way as for the alloyed type of transistor. The germanium that recrystallizes on cooling contains impurities additional to indium: gallium is present in the germanium deposited under one pellet, antimony in that deposited under the other. Like indium, gallium is an acceptor element, but the solubility in solid germanium is about one hundred times greater for gallium than for indium. Consequently a strongly doped *P* region, which can serve as emitter, forms under the gallium-containing indium pellet [8]). In the region under the antimony-containing pellet, acceptor atoms (of indium) are overcompensated by donor atoms (of antimony) and this is accordingly an *N* region which will ensure good electrical contact with the *N* layer.

To reduce the capacitance of the collector barrier, as much as possible of the *N* layer is etched away. The result is a crystal of the original *P* material carrying a remnant of the diffused *N* layer and topped by an emitter and a base pellet (figs. 6*d* and *e*). Leads are attached to the pellets by soldering.

Regarding the dimensions, it is found that with a base thinner than 3 μm and a collector and an emitter barrier which extend over areas smaller than 0.05 and 0.02 mm² respectively, it is possible to obtain a transistor with a good amplifying action at frequencies as high as 100 Mc/s. These requirements as to internal dimensions are satisfied by a transistor whose emitter and base pellets, if equal in size, have the diameter and separation indicated

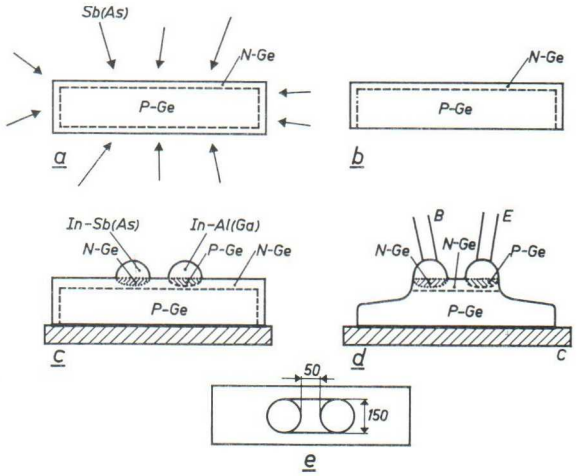


Fig. 6. Manufacture of an HF transistor in separate diffusion and alloying stages. *a*) Diffusion of a donor (Sb or As) into a *P*-type germanium crystal causes an *N* layer to form on the surface of the crystal. *b*) The *N* layer is etched away from the underside of the crystal. *c*) The emitter region (which also serves as emitter contact) and the base contact are made by alloying; a collector contact, which may take the form of a nickel strip, is soldered to the bottom of the crystal. *d*) Electrode leads *B*, *E* and at *C* are then attached, and all the unwanted parts of the diffused *N* layer are removed by etching, leaving an island which constitutes the active zone of the transistor. *e*) View of the finished transistor from above the dimensions, which are in μm, are appropriate to a type operating at frequencies up to 100 Mc/s.

in fig. 6*e*. HF transistors manufactured in the manner just described are known as diffused base transistors.

Mention was made of a technological difficulty arising in the manufacture of alloyed transistors, namely the strict control of the alloying process that is necessary if base layers of the required width are to be obtained in a reproducible way. The same difficulty is encountered in the making of diffused base transistors though to a much lesser degree. There is only one reason for uncertainty about the base width, that being the depth to which the alloying front penetrates; the surface *N* layer can be diffused to exactly the prescribed depth. A cross-section through the emitter region, drawn to scale as in *fig. 7*, reveals how severe are the require-

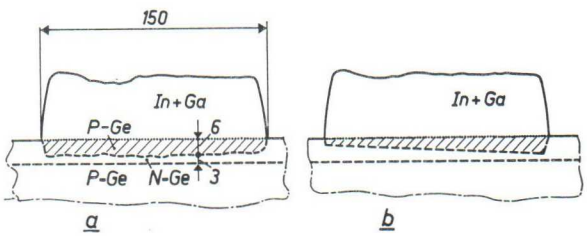


Fig. 7. Cross-section, drawn to scale, of the emitter region of a transistor made by the method illustrated in fig. 6. The drawings reveal the severe requirements to ensure that the alloying front is flat (*a*), and that this front — a (111) plane — is parallel to the face of the crystal (*b*).

[8]) F. H. Stieltjes and L. J. Tummers, Behaviour of the transistor at high current densities, Philips tech. Rev. **18**, 61-68, 1956/57, in particular p. 67 and 68.

ments that have to be satisfied by the alloying front, not only regarding depth of penetration but also the flatness and parallel alignment with the underlying *P-N* junction. Failure to meet these requirements causes a high reject rate in mass production.

Mesa transistors

The difficulties just mentioned are largely due to the fact that the distance over which the alloying front travels is several times greater than the required residual base width. With normal alloying methods this is inevitable: the alloying must be carried out at a temperature considerably above the melting point of indium in order to ensure adequate wetting of the germanium by the liquid indium. But at this relatively high temperature so much germanium becomes dissolved in the indium that an appreciable depth of penetration is inevitable. There is, however, a technique available in the shape of vacuum deposition of the alloying materials. The emitter can be made by depositing aluminium and the base contact by depositing gold containing a certain amount of antimony. Very small quantities of material can be transferred in this way, so that the depth to which the alloy fronts penetrate are restricted to a fraction of the residual base width. Fluctuations in the alloying front likewise diminish and thus cease to be dangerous. Devices made by this method are called mesa transistors [9]). The emitter and base leads are attached to the alloy films by a special technique known as thermocompression bonding.

The p.o.b. transistor

It has proved possible to get around the above difficulties by adopting what is known as the alloy-diffusion technique [10]). In the resulting p.o.b. transistor, the typical transistor layer structure is formed in a single process, contact areas for easy attachment of electrode leads being produced at the same time. The method lends itself extremely well to mass production, and many millions of p.o.b. transistors for frequencies up to 200 Mc/s are now being produced every year. Here we shall restrict ourselves to describing the principle underlying the manufacturing process.

The starting material, as in the case of the diffused base and mesa transistors that have just been described, is a supply of *P* type germanium wafers whose resistivity of 1 to 2 ohm cm is that required for the collector region. Two lead pellets of diameter 150 μm are placed about 50 μm apart on each germanium wafer. Both pellets contain the donor element antimony, and one of them further contains the acceptor element aluminium. The alloy-diffusion process takes place at 780 °C. The situation is as shown in *fig. 8*. Both pellets have taken up germanium, antimony was already present in solution, and one pellet additionally contains aluminium. At 780 °C antimony has a fairly fast rate of diffusion in solid germanium whereas aluminium, under the same conditions, has negligible diffusion. Accordingly, the donor element antimony invades the solid *P* germanium and turns it into *N* germanium. The temperature of 780 °C is maintained long enough for the *N* layer under the lead pellets to attain the required base width of 3 μm. But at the same time antimony vapour is escaping from the molten lead and penetrating the exposed part of the germanium crystal (fig. 8). The result is a continuous *N* layer that entirely encloses the crystal.

The holes made by the invading lead fill up again with germanium as the crystal cools. All the deposited germanium contains antimony, and that deposited under the pellet with aluminium contains aluminium as well. This last, an acceptor metal, has a solubility in solid germanium very much greater than that of the donor antimony; this means that the material which recrystallizes from the alumi-

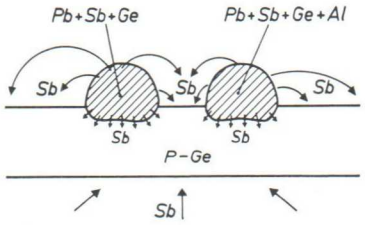


Fig. 8. Situation in a p.o.b. transistor immediately after it has reached the temperature of 780 °C at which the alloy-diffusion process takes place. The lead beads have melted and, in taking germanium into solution, they have eaten into the crystal. The temperature of 780 °C has to be maintained for some time to enable the Sb to diffuse into the crystal. Diffusion takes place in two ways, via the vapour phase and through the contact planes between the liquid lead and the solid germanium.

[9]) The name "mesa" (the Spanish word for "table") is given to this type of transistor because of the shape of the crystal after the etching treatment (see figs. 6*d* and *e*). The nomenclature is not altogether apt: transistors of the "diffused base" type, mesa transistors and the p.o.b. transistors discussed below all have a base layer that is formed by diffusion, and all three have an active region of the table shape indicated in figs. 6*d* and *e*.

[10]) P. J. W. Jochems, O. W. Memelink and L. J. Tummers, Construction and electrical properties of a germanium alloy-diffused transistor, Proc. I.R.E. **46**, 1161-1165, 1958. Similar investigations have been carried out by J. R. A. Beale, Proc. Phys. Soc. (London) **B 70**, 1087-1089, 1957. The alloy-diffusion process has been employed previously by J. J. A. Ploos van Amstel of our laboratories, who by this means produced alloyed transistors with a drift field, and by R. L. Longini, who used the technique to make "hook-collectors" (British patent No. 754404).

nium-containing pellet is *P* type germanium having the low resistivity required for the emitter. The emitter region is thus formed under the Al-containing pellet, whereas under the other pellet an *N* region is formed, ensuring good non-rectifying contact with the base layer produced by diffusion (*fig. 9*).

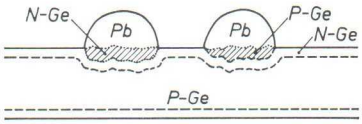


Fig. 9. The situation in a germanium wafer subsequent to alloy-diffusion treatment. The transistor structure has formed automatically. Evaporating from the lead beads, the donor element Sb has diffused into the *P* type germanium crystal, forming a surface *N* layer. This *N* layer is connected to the *N* layers that have formed under the two beads — which have now solidified again — constituting the base layer of the transistor. The hatched area on the right has solidified as *P* germanium, since here Sb atoms outnumber those of the acceptor metal Al; this constitutes the emitter. The hatched region on the left has solidified as *N* germanium and provides good electrical contact with the base. The above sketch is not to scale, the thickness of the diffused *N* layer being greatly exaggerated (cf. fig. 7).

The choice of a good "support metal" — the lead — is all-important. Designedly, the alloy-diffusion process is carried out at a temperature at which the diffusion rate for antimony is fast enough for the base layer to be formed within a few minutes. At this temperature the solubility of germanium in the support metal must be such as to ensure a reasonable penetration depth. Lead satisfies these requirements very well.

In the single-stage alloy-diffusion process, base and emitter contacts are produced as well as the required *P-N* junctions. The effect is as if the base layer were pushed out of the two lead pellets (hence the name "pushed-out base" transistor). Consequently neither the depth of the alloying front nor its flatness or orientation is of much significance and the thickness of the base layer can be accurately controlled. Thus the alloy-diffusion technique offers a neat way round all the difficulties presented by diffusing and alloying in separate stages. *Fig. 10* is a microphotograph of a p.o.b. transistor; in *fig. 11* the transistor can be seen mounted on its mechanical base, and *fig. 12* shows the encapsulated transistor ready for use.

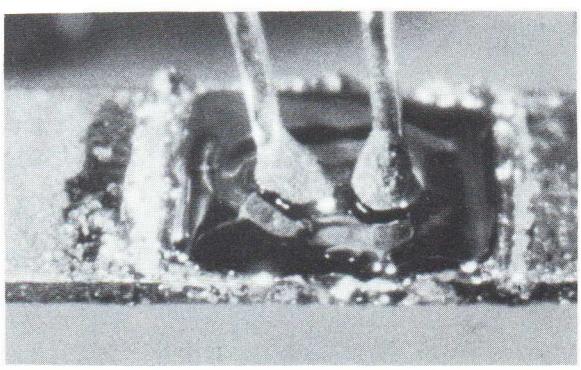


Fig. 10. Microphotograph of a p.o.b. transistor for operation at 100 Mc/s. Magnification 62×.

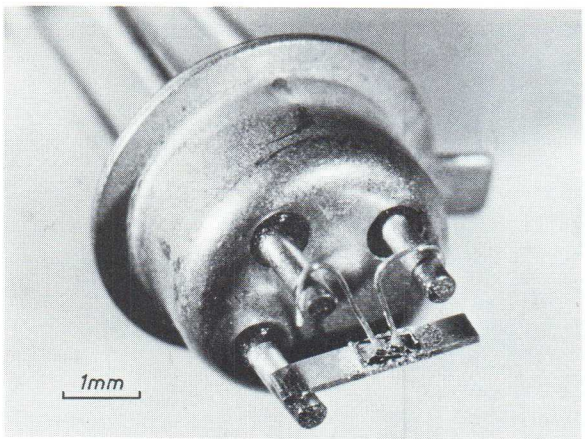


Fig. 11. A p.o.b. transistor mounted on its mechanical base.

A refinement to the alloy-diffusion process is to carry out a preliminary diffusion treatment. As in the process by which diffusion and alloying are done in separate stages, the *P* germanium crystal is heated in antimony vapour so that a surface layer of *N* germanium is formed upon it. It then undergoes the alloy-diffusion treatment. The Pb from the pellets passes right through the existing *N* layer, so that the thickness of the base layer formed under the pellets is not affected by the pre-diffusion treatment. The *N* layer formed between the pellets is now thicker and consequently it is a better conductor. The overall result is a reduction in base resistance that improves the properties of the transistor at high frequencies.

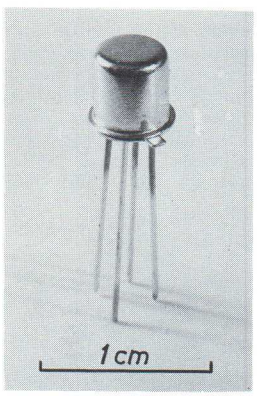


Fig. 12. The same transistor as in fig. 11, ready for use.

A microminiaturized solid-state circuit

In electronics there is a constant endeavour to reduce the size of components and complete circuits, and to increase their reliability by limiting the number of soldered joints. The advent of crystal diodes and transistors has greatly strengthened the trend towards microminiaturization. Not only are the diodes and transistors small in themselves, but

they develop very little heat because they embody no filaments. Heat dissipation is a big obstacle to more compact electronic equipment [11]).

What is known as "solid-state circuitry" is one approach to microminiaturization: instead of being made up of discrete elements, the solid-state circuit takes the form of a single semiconducting crystal which has been processed into resistors and capacitors as well as diodes and transistors. Resistors of the desired value can be provided by parts of the crystal with the right shape, size and resistivity. Capacitors consists of *P-N* junctions of suitable dimensions. It is possible in this way to accommodate in one crystal an almost complete circuit having a given function, which may be amplification, pulse inversion, switching, the generation of oscillations or the like. For this purpose use can be made of the alloy-diffusion process of which an account has just been given. By way of example we shall describe here the making, under laboratory conditions, of a solid-state amplifier by means of the alloy-diffusion technique. The good HF properties available with this technique are not exploited in this particular circuit [12]).

The circuit in question, shown in *fig. 13*, is that of a three-stage amplifier. We shall not discuss the way it functions [13]), but merely show how the solid-state version is arrived at. The circuit elements outside the chain-dotted line in fig. 13 are not embodied in the crystal, but are assembled on it as individual components.

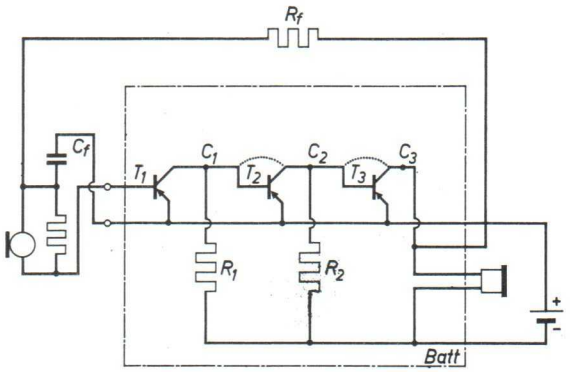


Fig. 13. Circuit diagram of a 3-stage amplifier [13]) which has been made by a "solid-state" microminiaturization technique. The part of the circuit enclosed by the chain-dotted line is accommodated in a single germanium wafer. The dotted lines between C_1 and C_2 and between C_2 and C_3 indicate shorting paths that have been broken by breaking the *P-N* junction at the places marked *I*, *II* and *III* in fig. 15. R_f and C_f provide DC feedback.

The starting material for the solid-state circuit is a germanium monocrystal in which a *P-N* junction has been produced during the pulling process (see p. 232). The section containing the junction is sawn up into slices and these "blanks" are cut to the required shape (*fig. 14*) with an ultrasonic drill [14]). It is important that the grown *P-N* junction should occupy the position indicated in fig. 14.

The damaged surface layer caused by the cutting operations is removed by etching, and three transistors are then made by the alloy-diffusion technique at the places indicated in *fig. 15*. In the case of transistors T_2 and T_3 no base pellet is applied since, for reasons that will shortly become evident, no base connection is required. The surface *N* layer formed

11) A systematic review of the various approaches to miniaturization, illustrated with many examples of circuits that have actually been built, may be found in E. F. Horsey and P. J. Franklin, Status of microminiaturization, I.R.E. Transactions on component parts **CP-9**, 3-19, 1962 (No. 1).

12) The necessary experiments were carried out by H. G. Kock.

13) D. L. Jones, Directly coupled transistor hearing aid, Mullard tech. Comm. **5**, No. 41, 2-9, 1959.

14) E. A. Neppiras and R. D. Foskett, Ultrasonic machining I. Technique and equipment, II. Operating conditions and performance of ultrasonic drills, Philips tech. Rev. **18**, 325-334 and 368-379, 1956/57.

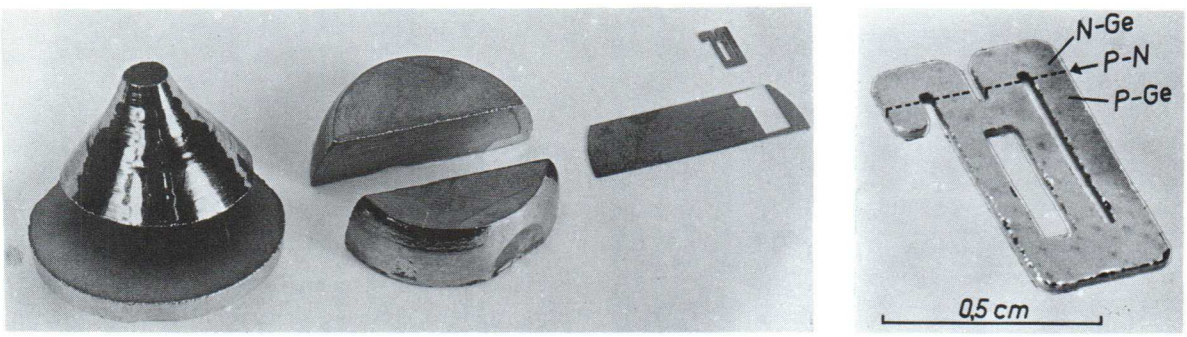


Fig. 14. The wafers are sawn out of a germanium crystal containing a *P-N* junction, and cut to the required shape with an ultrasonic drill. The *P-N* junction must occupy the position indicated in the enlarged photograph on the right.

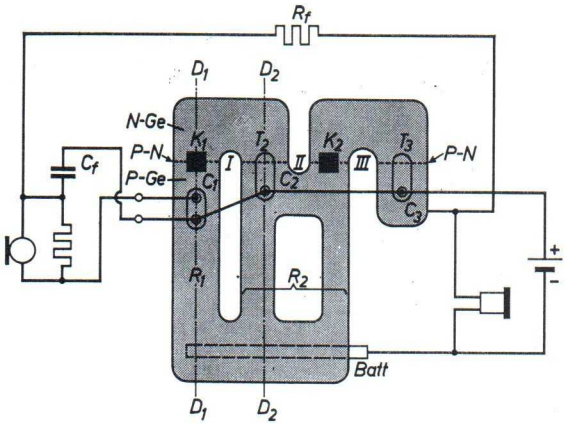


Fig. 15. Diagram to show how the circuit of fig.13 is built into the germanium wafer. Figs 13 and 15 have been drawn in such a way that resistors R_1 and R_2 and transistors T_1, T_2 and T_3 occupy roughly the same positions relative to one another. K_1 and K_2 represent short-circuits across the grown P-N junction (broken line). (Unfortunately, the letter T_1 has been omitted in the diagram.)

as a result of the alloy-diffusion process is etched away except in small areas around the lead pellets. A cross-section taken along D_1 in fig. 15, cutting through transistor T_1, is shown in *fig. 16a*. As is clear from the circuit diagram in fig. 13, the collector of T_1 has to be connected to the base of T_2 without resistance; accordingly, a conducting strip is applied across the P-N junction near T_1. In practice it has been found that the junction can be short-circuited just as effectively by making a few scratches across it with a diamond. The base layer of transistor T_2 extends over the neighbouring grown P-N junction (as can be seen also in fig. 16*b*, which is a cross-section taken along D_2 cutting through transistor T_2). By these means a connection without any intervening P-N junction has been provided between the collector C_1 of T_1 and the base of T_2. The direct connection between the collector of T_2 and the base of T_3 is established by short-circuiting the grown P-N junction between

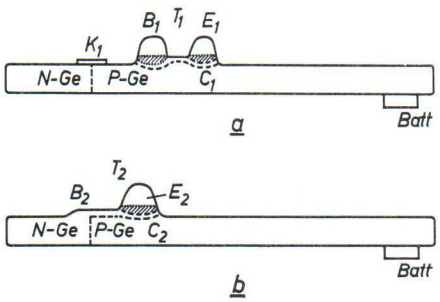


Fig. 16. *a*) Cross-section through transistor T_1, taken along line D_1 in fig. 15. K_1 short-circuits the grown P-N junction produced in the original crystal.
b) Cross-section through transistor T_2, taken along D_2.

these two transistors (K_2 in fig. 15). Like that of T_2, the base layer of T_3 extends over the neighbouring grown P-N junction. It will now be clear that the base pellets can be omitted when T_2 and T_3 are made. The grown junction had to be broken at the places marked *I*, *II* and *III* in fig. 15 in order to prevent short-circuits between the collectors (along the dotted lines in fig. 13; cuts *I* and *II* break the connection between C_1 and C_2, cut *III* breaks that between C_2 and C_3).

The battery contact, marked *Batt*, consists of a strip of metal soldered to the germanium crystal. The connection between the emitters has to be made by attaching wires to the appropriate pellets. *Fig. 17* is a photograph of the finished solid-state circuit.

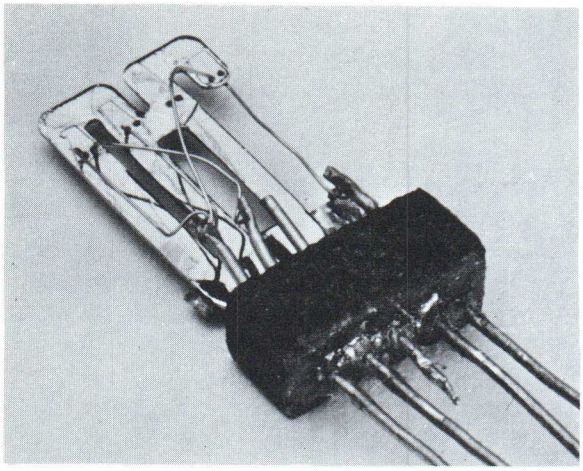


Fig. 17. Experimental version of a three-stage amplifier miniaturized by the solid-state technique. Additional lead beads have been deposited to facilitate measurements. Wire tails, going to a common emitter lead in the circuit mounting, have been attached to each of the emitter beads.

Summary. Almost all transistors nowadays are of the junction type. One way of producing the P-N transitions in junction transistors is to add suitable impurities while the germanium (or silicon) monocrystal is being pulled from the melt. However, the alloying and the diffusion methods are the ones most commonly employed. The higher the frequency at which the transistor is required to provide a reasonable signal gain, the more difficult it becomes to make good P-N junctions efficiently, mainly because extending the frequency range necessitates reducing the dimensions of the transistor. In the alloy-diffusion technique, the P-N junction between emitter and base is formed as a result of an alloying process, and that between base and collector is formed as a result of the diffusion process which takes place at the same time; this, in essentials, is the method developed in Philips Research Laboratories for mass-producing HF transistors. Many millions of transistors capable of operating at frequencies up to 100 and 200 Mc/s are now being manufactured by this method every year.

The alloy-diffusion technique has also been used in experimental work on microminiaturization. As an example the processing of a germanium wafer into an almost complete three-stage amplifier is described.



Glass fracture surface

Fracture surface of a fragment of glass from a television picture tube, taken with the electron microscope a few hours after fracture. Magnification approx. 20 000×. Within about an hour of fracture, traces of chemical attack are visible: they are probably local swellings caused by the action of water (layers of silica gel). After a few hours a kind of frost-work tracery as seen in the photograph often appears.

(From: Philips tech. Rev. **26**, 85, 1965.)

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.

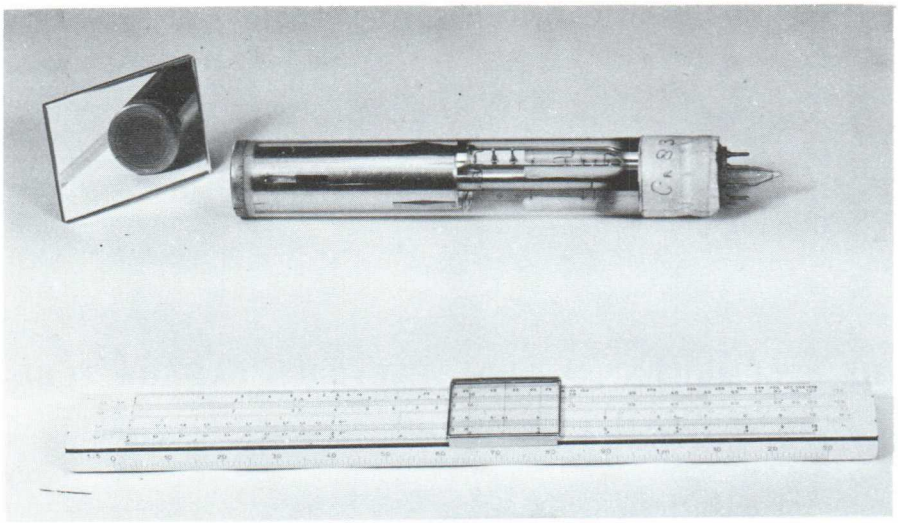


Fig.1. A pick-up tube of the "Plumbicon" type. The tube is in the form of a cylinder having a diameter of 3 cm and a length of 19 cm. The target diameter is 2 cm.

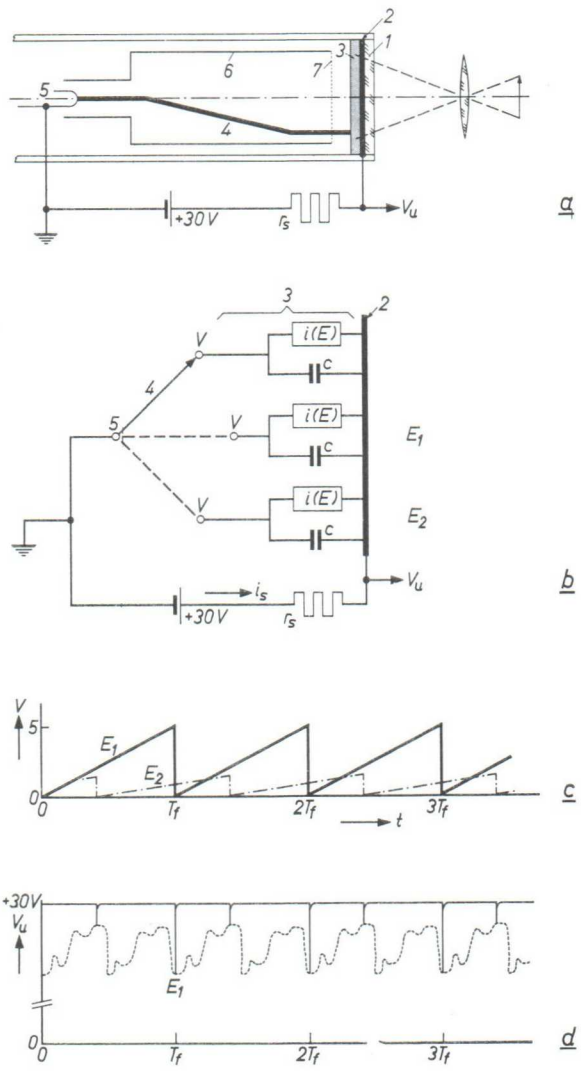


Fig. 2. *a*) Electrode layout of the "Plumbicon" (schematic). In the front of the tube (on the right) is a glass window *1*, on the inside of which have been applied, in that order, a transparent, conductive SnO_2 layer *2* and a photoconducting layer of PbO constituting the target *3*. An image of the 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_s, proportional to ΔV, flows each time the beam completes the circuit containing an element, and this current causes a corresponding difference of potential to arise across signal resistor r_s. Consequently the potential V_u varies and an output signal is obtained. If E is changing rapidly, ΔV will have a value roughly corresponding to the mean of the intensity of illumination on the element during the frame period in question.

c) The variation over time t in the potential V on the free surface of two picture elements, one exposed to a high intensity of illumination E_1 and the other to a low intensity of illumination E_2. Since, under normal operating conditions, i_f does not depend on the difference of potential $V_u - V$, between scans the variation in V is linear with respect to time.

d) The variation in V_u over time t, corresponding to (*c*). When the contributions of *all* the picture elements are taken into account, then V_u varies e.g. as shown by the dotted line.

many ways to be a kind of vidicon, there is a characteristic difference however between the "Plumbicon" and the present vidicons, and this concerns the photoconducting layer. Not only has this layer been made from a different photoconductor — Sb_2S_3 and sometimes As_2S_3 or Se have so far been used in ordinary vidicons — but what is more important, the PbO layer together with the SnO_2 layer form a unit consisting of three sublayers, each of differing conduction type. The inner sublayer consists of almost pure PbO, which is an intrinsic semiconductor. The PbO in the layer struck by the electrons is doped to make it a *P*-type semiconductor (hitherto not possible with Sb_2S_3). The SnO_2 signal plate is strongly *N*-type. The contact between the PbO and the SnO_2 may also give rise to a thin *N*-type layer in the PbO. The *P*-type and *N*-type layers are relatively thin, so that the inner (intrinsic) layer, the *I* layer, takes up most of the overall thickness of the PbO layer. For simplicity it is assumed in the following that the *N*-type PbO layer is always present [2]).

Examination under the electron microscope shows the PbO layer to be porous in structure; it is built up of crystallites having dimensions of about $1.0 \times 1.0 \times 0.1\,\mu m$. The filling factor ranges from 30% to 50%. The dimensions of the crystals are small compared with the line spacing (20 μm). They are therefore too small to be detrimental to the resolving power. The overall thickness of the photoconducting layer ranges from 10 to 20 μm.

The fact that the tube has many favourable properties for television broadcasting is thanks to the multilayer structure of its photoconducting target. Its ability to satisfy two of the main desiderata, namely low dark current and high sensitivity, is easily explained. When the tube is in operation its photoconducting layer — in contrast with a conventional vidicon — constitutes a reverse-biased diode. The dark current is the (small) inverse current through this diode. The tube owes its high sensitivity to the *I* layer sandwiched between the *P* and *N* layers. Conduction electrons and holes generated by light cannot contribute to a photo-current unless they originate in a region where a relatively high field-strength prevails. If the diode in question were simply a *P-N* device, the requisite field-strength would only be available in the immediate vicinity of the junction, and a large proportion of the charge-carriers generated in the PbO would be ineffective. In the "Plumbicon", however, there is a high field-strength throughout the *I* layer, in consequence of the fact that this is a relatively poor conductor, and since the PbO layer consists almost entirely of *I* material, almost all the charge-carriers generated in the PbO contribute to the photo-current [3]).

In principle, an excessive dark current can be compensated by electrical means. In practice, however, electrical compensation is scarcely ever adopted: unless the dark current has the same value within close limits at all points on the screen, compensation leads to an uneven image signal which is undesirable and in colour television is quite unacceptable. An additional difficulty is that the dark current is strongly dependent on temperature. If, on the other hand, the dark current has a very low value (say 10^{-9} A, or about 1% of the signal current), local variations, even though quite large (the values differing by a factor of 2, for example) will not perceptibly affect the uniformity of the image.

In most other important respects — spectral characteristic, definition, speed of response and service life — the favourable properties of the tube are mainly dependent on a suitable choice of the parameters governing the properties of the sublayers, such as their thickness, the doping substance and its concentration, and so on. Although some of the demands of studio use give rise to conflicting requirements, it has not been necessary to make compromises between the various parameters. On the contrary, there is so much play that certain properties can be varied quite widely without interference with the others. This makes it possible to manufacture camera tubes of the "Plumbicon" type with properties very closely fitted to the demands of a given application. Left out of discussion are the very large variations encountered when the choice of photoconducting material is not restricted to PbO, but when e.g. PbS is added [4]).

Various points mentioned above will now be elaborated [5]). First however the relevant physical and chemical properties of PbO will be discussed; some observations will be made on the process of deposition of the PbO layer, and a brief account given of the way in which the potential of the free surface of the photoconducting layer varies during a frame period.

The properties of PbO; deposition of the photoconducting layer

Two modifications of PbO are known: the red (tetragonal) modification, which is stable at temperatures below 488 °C, and the yellow (orthorhombic),

[2]) The characteristics of the "Plumbicon" are compared with those of other tubes in an article by A.G. van Doorn and S.L.Tan, shortly to be published in this Review.

[3]) For a succinct explanation of the physical principles underlying photoconductivity see e.g. L. Heijne, Philips tech. Rev. **25**, 120-131, 1963/64 (No.5).

[4]) See the article by E.F. de Haan, F.M. Klaassen and P.P.M. Schampers, shortly to be published in this Review.

[5]) See also L. Heijne, thesis Amsterdam 1960, Ch. 8.

which is stable at higher temperatures. The red PbO is built up from "sandwiches" consisting of a plane occupied by O atoms, on either side of which is a plane occupied by half as many Pb atoms. These sandwiches have a thickness of 2.38 Å, and the spacing between the O planes is 4.99 Å.

The structure of the yellow modification differs considerably from that of the red. This too is built up from sandwiches, with Pb atoms on the outside, but the "filling" is more complicated. Accordingly, the thickness of the sandwich is rather greater (2.72 Å). Various investigators have demonstrated that oxygen can be inserted between the sandwiches (especially those of red PbO) without drastically modifying the crystal structure of the compound; in other words, departures from stoichiometric ratio are possible. It is in virtue of this important property, among others, that PbO can be turned into either an *N*-type or a *P*-type semiconductor. It has been found that PbO becomes a *P*-type semiconductor when an excess of oxygen is present, or when it has been doped with Tl, Cu or Ag. The compound becomes an *N*-type semiconductor when an excess of lead is present, or when it has been doped with Bi.

The energy gap ΔE of the forbidden zone between the valency band and the conduction band, the quantity that determines the upper limit of the range of wavelengths within which photo-excitation can occur, is 2.0 eV for red PbO and 2.7 eV for the yellow modification. The band gap of red PbO gives rise to a cut-off wavelength of about 6200 Å, that of yellow PbO gives about 4500 Å.

The production of the PbO layer in a tube of the "Plumbicon" type is by vapour deposition. PbO contained in a small platinum crucible is evaporated by inductive heating, and condenses on a window previously coated with SnO_2. The crucible temperature is held at about 900 °C; at this temperature evaporation proceeds at a reasonable rate. During the deposition process the window is likewise kept, within fairly close limits, at a certain temperature. A high temperature is especially undesirable because the crystals of the deposited PbO become too large to give the required image resolution. Deposition takes place not in vacuum, but in a certain gas atmosphere.

Owing to the considerable difference in the cut-off wavelengths of red and yellow PbO, the red-yellow ratio has an important effect on the spectral response of the "Plumbicon". An X-ray diffraction study, in which the diffraction pattern of PbO layers deposited by the normal process was compared with those of certain mixtures of yellow and red PbO powders, has revealed that these layers consisted of about 90% of the red modification and about 10% of the yellow. At vapour pressures lower than the normal one, the proportion of yellow PbO was greater. Moreover, the red crystals were found to have a preferred direction.

Potential of the target surface; stabilization

As already stated, the free surface of the PbO target has a potential V that varies within a range close to the potential of the cathode. At the instant the scanning beam leaves an element of this surface, the potential of this element is roughly equal to that of the cathode; in the time elapsing before the beam returns, V rises by a few volts only. The actual range over which the surface potential varies is determined by the requirement that in the steady state, the amount of (negative) charge removed per unit time, in consequence of the flow of photo-current, must be equal to that supplied by the electron beam. (Since the negative charge is supplied intermittently, this equality is only valid if an interval of time is considered containing an integral number of frame periods.)

Fig.3 shows, schematically, how i_a, the net current flowing to a surface element of a PbO target under electron bombardement, depends on the potential V of the target surface. The various curves refer to different values of beam current, i.e. the current formed by the electrons leaving the electron gun. When V is the same as the cathode potential (which is zero), the deceleration of the electrons in the final part of their path is equal to their previous acceleration, with the result that only a few of them actually reach the target. The *net* flow of current to the target is even smaller, owing to secondary emission; it is on account of increasing secondary

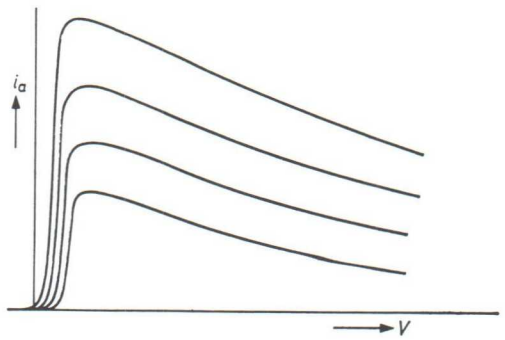


Fig.3. The net flow of current i_a to a surface element of a PbO target under continuous electron bombardment as a function of the potential V of the target surface (the cathode of the electron gun is assumed to have zero potential). As V increases, i_a also initially increases but subsequently falls off on account of increasing secondary emission. The above curves, which are schematic, relate to various values of beam current (i.e. the current constituted by the electrons leaving the gun).

emission that the net current to the target, after attaining a maximum for a certain potential value, falls off again as V is raised further. *Fig. 4* gives an overall impression of the way the photo-current that flows through the target layers, and so removes negative charge from its free surface, depends on the potential difference U between the two faces of the PbO

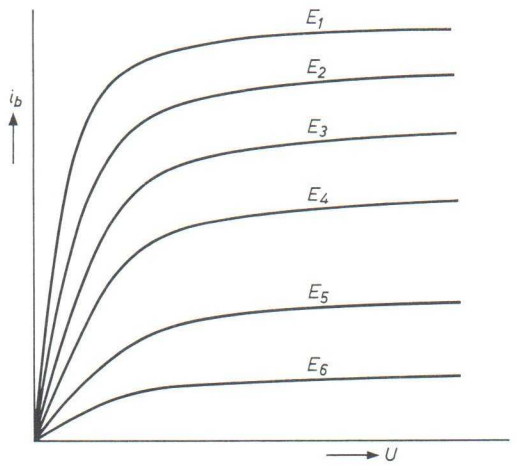


Fig. 4. The variation of i_b, the photo-current that removes (negative) charge from an elementary area of the free target surface, with the potential difference U across the target. The curves, which are schematic, hold for different values E_i of the illumination. The stronger the illumination, the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

target ($U = V_u - V$; see fig. 2). The plotted quantity i_b is again the value of the current per surface element.

In *fig. 5* one of the curves of fig. 3 has been combined with the family of curves in fig. 4. That, in fact, corresponds to the situation prevailing in the tube: the intensity of illumination may assume almost any value, but unless altered from outside the beam current always remains the same. For reasons that will shortly become clear, the ordinate values for the curves representing the current flowing through the target layers have been multiplied by a factor N representing the number of surface elements into which the surface is understood to be divided. Accordingly, the quantities plotted are i_a and Ni_b.

Consider first the imaginary case in which the beam electrons continuously bombard the same spot on the target surface, i.e. a spot whose area is $1/N$th the total target area (this is what we call a picture element), and in which a constant photo-current Ni_b is flowing at this spot. The potential V of the spot then assumes a value given by the projection on the abscissa of one intersection point of the i_a curve with the relevant Ni_b curve: the loss of charge is then exactly balanced by the rate of supply. The points of intersection representing stable states have been marked in fig. 5 with a dot. It will be noted that these points occur both on the rising and on the falling branch of the i_a curve. In the former case the value of V is close to zero, and the voltage U across the target is therefore roughly equal to V_u. The points of intersection on the falling branch represent states in which the value of V is close to the signal-plate potential, and U is very small.

Let us now consider the true situation, in which the beam scans the target and the current supplying a given picture element only flows for $1/N$th of the time. On account of the surge-like character of the charge-supply process V does not assume a steady value; it can however be said that the range over which V varies must extend on either side of one of the stable-state values referred to above. Since the range of variation ΔV is normally small compared with V_u, these stable-state values nevertheless give a fairly good indication of the operating conditions that are possible in the tube: there are again *two* possibilities when the tube is operational. Operating conditions such that $V \approx V_u$, U accordingly being small, are extremely unfavourable, however: in the first place the equilibrium value of i_a differs very little with different intensities of illumination, with consequent loss of contrast, but apart from that, special forms of sluggish response are liable to occur at low U values.

Therefore in practice the beam-current value is chosen high enough for an intersection to be available on the *rising* branch of the i_a curve even at the highest intensity of illumination likely to occur on the PbO target during a broadcast. In this way a third, but no less important, disturbing effect is

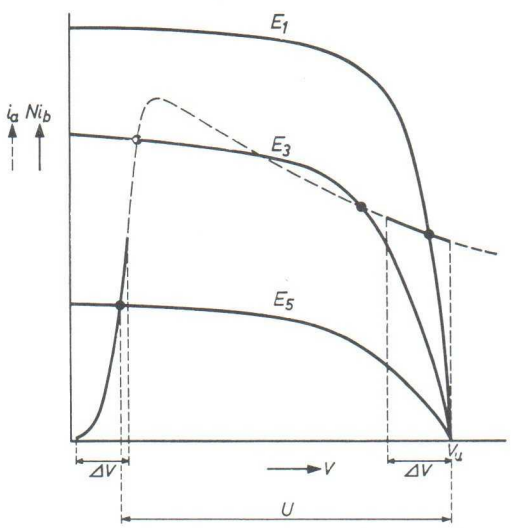


Fig. 5. To explain the fact that the interval ΔV can fluctuate in the neighbourhood of the cathode potential of the electron gun ($V \approx 0$), as well as in the neighbourhood of the potential of the signal plate ($V \approx V_u$).

avoided that occurs when one surface element has a potential close to V_u and the other a potential close to zero. In these circumstances the element with the higher potential will start to attract electrons towards it when the beam approaches, and continue to do so after the beam should have passed. In consequence, lighter-coloured parts of the scene will be "blown up", i.e. appear bigger than they really are.

The fact that V moves within an interval ΔV, contained between the abscissa values of the intersections of the i_a and Ni_b curves, can be proved as follows. During the short time T_p in which the beam passes a surface element, then for the (negative) charge:

$$dQ = (i_a - i_b)dt \approx i_a dt = c dV .$$

Here dQ is the charge supplied in time dt and c is the capacitance of the layer per surface element. It follows from this that $dV/i_a = dt/c$. In the time T_p, V decreases from V_2 to V_1 and so:

$$\int_{V_2}^{V_1} \frac{dV}{i_a} = \frac{T_p}{c}.$$

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

$$\int_{V_1}^{V_2} -\frac{dV}{i_b} = \frac{T_f}{c},$$

or, as $T_f = NT_p$:

$$\int_{V_2}^{V_1} \frac{dV}{Ni_b} = \frac{T_p}{c}.$$

If $1/i_a$ and $1/Ni_b$ are both set as functions of V, then the areas under the curves between the ordinates $V = V_1$ and $V = V_2$ are equal. This implies that the curves must intersect at least once in the interval $V_1 - V_2$. So the same holds for the curves for i_a and Ni_b.

The target considered as a *P-I-N* diode

To obtain the smallest possible dark current and the greatest sensitivity possible in combination with it, then theoretically one must use a layer of an intrinsic photoconductor fitted with two contacts one of which, when current is flowing in a given direction, will hinder electron supply and the alter the hole supply ("blocking contacts"; see *fig.6*). In the "Plumbicon", the P layer acts as the contact hindering the entry of electrons, and the obstacle to the entry of holes is formed by the SnO_2 or by the PbO immediately adjoining it, when this region of the target has become an N-type conductor [6]). The fact that blocking contacts have been attached to the photoconducting layer in the "Plumbicon" constitutes one of the most striking differences with the conventional vidicons.

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

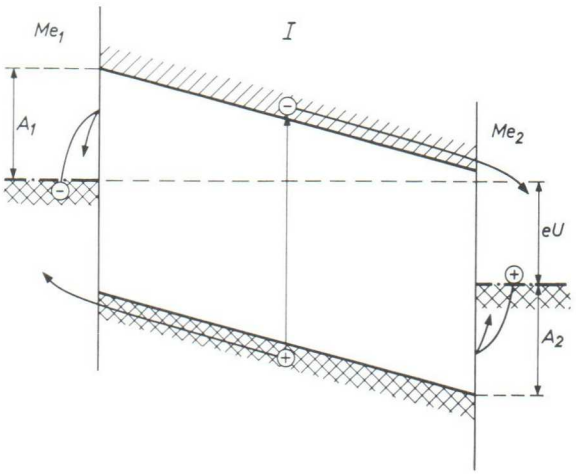


Fig. 6. Energy-band diagram (greatly simplified) of an intrinsic semiconductor (I) fitted with two metal contacts (Me_1 and Me_2). Me_1 has such a high work function (A_1) that electrons contained in it have very little chance of crossing the interface and entering the conduction band of the semiconductor. In the same way holes arriving from the metal Me_2 are unable to reach the valence band. On the other hand the charge-carriers that the incident light has liberated in the semiconductor, having moved over to the appropriate contact under the action of the electric field due to applied voltage U, encounter no obstacle hindering entry to that contact. (For simplicity, A_1 and A_2 have been chosen equal to $\frac{1}{2}\Delta E$.)

The energy-band diagram of a *P-I-N* diode is indicated in *figs.7a* and *b*. Fig.7*a* shows the details and fig.7*b* is a simplified diagram. It differs from the drawings in fig.6 in that the bands in the I layer are curved. This curvature can be explained as follows. In reality, the I layer can never be completely free from impurities; donor and acceptor centres in certain number are always present. The diagrams apply to the case where both concentrations are relatively high and roughly equal. When the concentrations are low the curved parts merge and the band diagram has the appearance indicated in fig.7*c*. Fig.7*d* shows how this diagram alters when a voltage is applied to the diode. The central layer (I), having by far the greatest resistance, shows the steepest fall-off in potential. However, because of the curvature of the bands the field-strength is not everywhere the same. *Fig.8* shows how the potential variation in the middle layer differs from that of fig.7 when the donor and acceptor concentrations are unequal, so that the layer is either slightly P-type or slightly N-type. In the former case the steepest fall-off in potential (and the higher field-strength) is on the N-contact side; in the latter case, the steepest fall-off is on the P-contact side. The practical significance of all this will be discussed below.

Fig. 9 shows the band diagram of a single contact, along with the more important quantities that allow the shape of the potential barrier to be described,

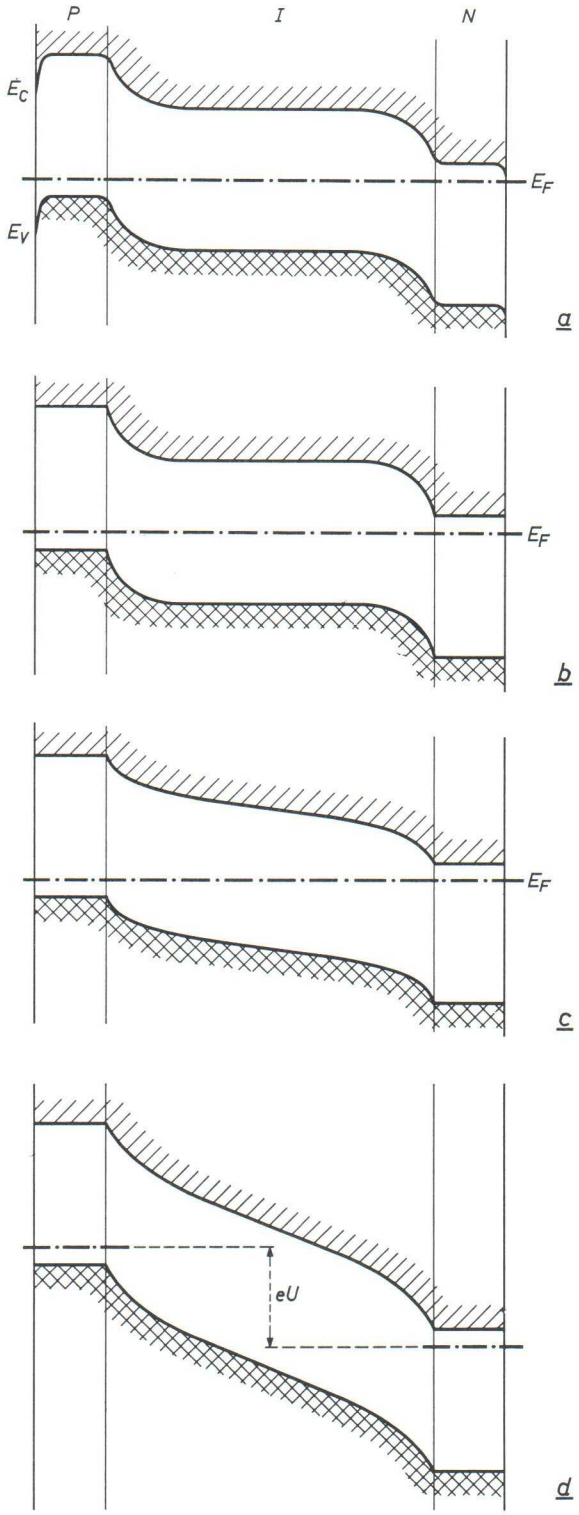


Fig. 7. *a*) Energy-band diagram of *P-I-N* diode. The curvature shown by the bands in the *I* layer is due to the fact that in practice this layer can never be completely free of impurities. In the case illustrated here, the concentration of impurities is relatively high. Diagram (*b*) is a simplified version of (*a*), the marked curvature exhibited in the two junctions having been converted into an abrupt change of slope. Diagram (*c*) is similar to (*b*) but relates to a diode with an *I* layer of such a high purity that the bands are nowhere horizontal and the curved portions meet. Diagram (*d*) shows how (*c*) is modified when a voltage *U* is applied to the diode. It must be made plain that this band diagram is itself a simplification since it implies that the target material has a homogeneous structure while in reality the PbO is made up of a large number of small crystallites. However, the diagram is quite adequate for the purpose of explaining the action of the diode and its more important properties.

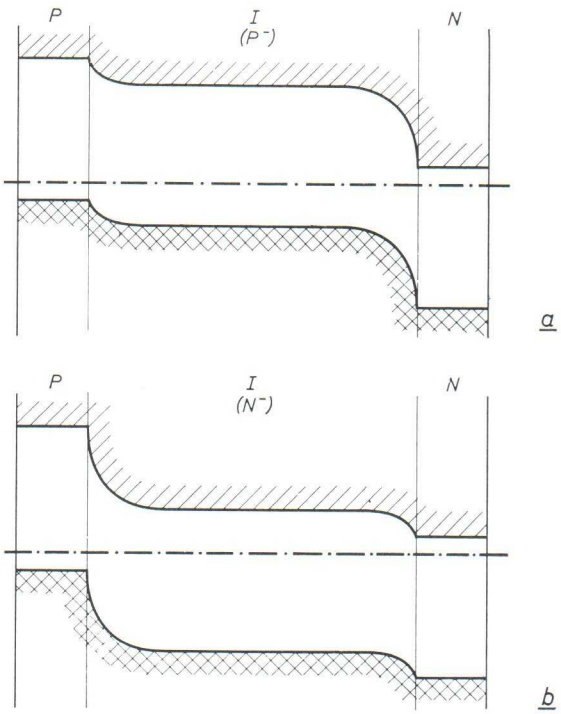


Fig. 8. Band diagrams simplified in the same way as fig.7*b* and relating to *P-I-N* diodes whose *I* layers are slightly *P*-type (*a*) and slightly *N*-type (*b*). In these cases there is a fairly steep fall-off of potential close to one contact and a region with a gradual potential fall-off (and consequently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

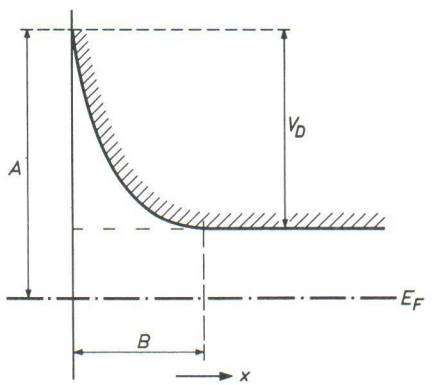


Fig.9. Near the interface between two regions of different conduction type a potential barrier arises which can be described in terms of the work function *A*, the diffusion potential V_D, and the width *B* of the barrier [7]).

namely the work function (barrier height) A, the width B of the region in which the bands are curved, and the diffusion potential V_D [7]). B is given by the formula

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

and the capacitance c' per unit area by

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

Here ε and e are the dielectric constant and the absolute value of the electric charge respectively, N_D is the (small) donor concentration in the I region, and U is the applied voltage. Further, for the field-strength F in the barrier,

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

in which x is the distance to the junction. Use is made of these formulae in the following section.

The dark current

As we have seen, the dark current is mainly determined by the inverse current through the P-I and I-N junctions. We shall now consider the magnitude of this current, taking only one of the contacts and only the kind of charge-carrier obstructed by that contact — the P-I junction, say, and the electron current. For such a contact, when reversed biased, the density j_n of the electron current is given by

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

where μ_n is the mobility of the electrons and $n(0)$ is the electron concentration in the boundary plane. The value of $n(0)$ depends on the absolute temperature T, the work function A and the constant N_c (sometimes called the effective density of the states in the conduction band), being connected by the formula

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

Given the values of U, T, N_D, ε and μ_n, one can use (4) and (5) to calculate the minimum value A must have for the dark current to be smaller than, say, 10^{-8} A (roughly 10% of the signal current; this corresponds, since the target area is a good 3 cm², to a current density of $j_n \approx 3\times10^{-9}$ A/cm²). The results of a number of such calculations have been collected in *Table I*. These are based on the assumption that $U = 50$ V, $T = 300$ °K and $\varepsilon = 12\,\varepsilon_0$ (for non-porous PbO, $\varepsilon = 26\,\varepsilon_0$).

Table I. Minimum values of A (eV), the relative work function of a contact on a photoconducting layer, appropriate to certain combinations of mobility μ_n (cm²/Vs) and donor concentration N_D (cm⁻³). The minimum values are obtained from the requirement that the dark current should not exceed 10^{-8} A when a voltage of 50 V is applied to the diode. (The relative dielectric constant is assumed to be 12, the absolute temperature 300 °K, and the area of the layer 3 cm². The value of V_D has been put equal to zero.)

μ_n \ N_D	10^{14}	10^{15}	10^{16}	10^{17}	10^{18}
1	0.79	0.81	0.84	0.87	0.90
10	0.84	0.87	0.90	0.93	0.96
100	0.90	0.93	0.96	0.99	1.02

As can be seen, over a wide range of μ_n and N_D values, the required value of A lies around 0.9 eV (it should be noted that the N_D value of 10^{14} cm⁻³ is for very pure material, that of 10^{18} cm⁻³ for heavily doped material). We can infer from fig. 7*b* that the band gap ΔE must be equal to or greater than A; in other words, the band gap too must be at least about 0.9 eV. At higher temperatures a bigger work function and a wider band gap will be required (see eq. 2); but smaller A and ΔE values suffice at lower temperatures. Or expressed the other way round: the use of a material with a band gap of less than 0.9 eV is not impossible but then cooling is required to get a sufficiently low dark current. It will now be clear why an extremely low dark current can be achieved with a layer of PbO to which blocking contacts have been attached: the band gap of red PbO is no less than 2.0 eV. It can be seen in *fig. 10* how the dark current in the "Plumbicon" varies with the potential difference across the target.

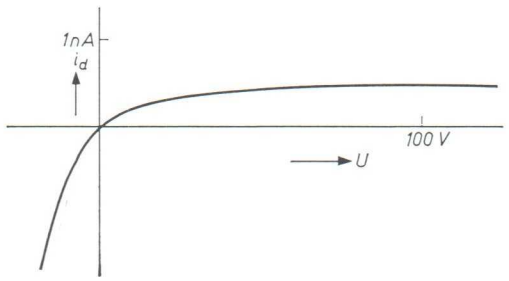


Fig. 10. The dark current i_d as a function of the potential difference U across the target. At no value of U likely to occur in practice does the dark current exceed 0.5×10^{-9} A. As might be expected, the curve resembles a diode characteristic. (In fact, the signal-plate potential V_u has been plotted along the horizontal axis; in practice V_u is nearly equal to U. The same holds for figs. 11 and 12.)

[7]) See the article cited under [3]). Since the thermal work function solid-vacuum does not appear in this article, for simplicity the thermal work function contact-semiconductor A will be referred to as the work function.

In the foregoing it has been tacitly assumed that there is negligible thermal generation of charge-carriers in the *I* region. Calculation confirms that this effect can, in practice, safely be neglected. (Only in the most unfavourable case, viz, that where the impurity levels lie at about half-height in the forbidden zone, can too strong a dark current be obtained when using a material with a band gap of 0.9 eV.)

Sensitivity

We shall now discuss the way in which the photo-current flowing through the PbO depends on the potential difference between the *P* and *N* layers, on the character of the light and on the intensity of illumination, and examine how these relations are affected by the thickness and other characteristics of the sublayers composing the target. *Fig.11* is a graph of the photo-current i_f versus the applied voltage *U* 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 that the zones just referred to extend pretty well through-

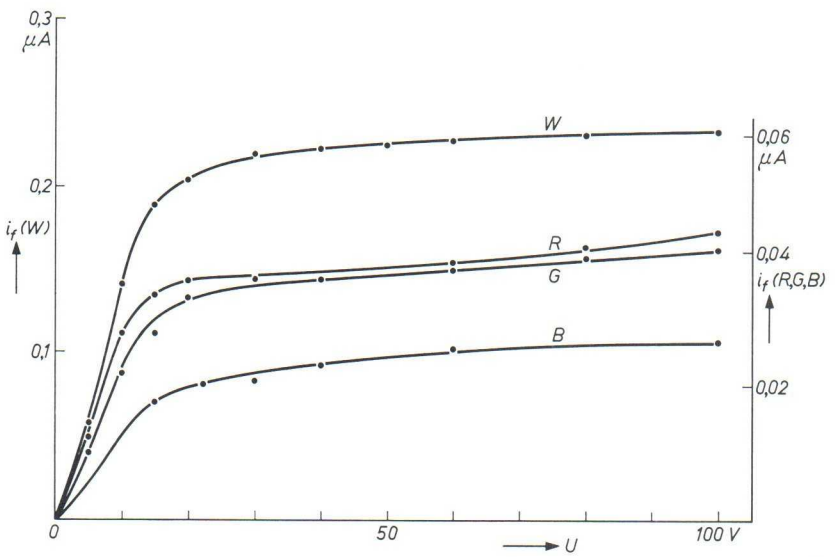


Fig.11. The variation of the photo-current i_f as a function of the signal-platevoltage V_u for different kinds of light, the intensity of illumination remaining the same (measured on a randomly selected tube). Curve *W*, which is to be read in conjunction with the i_f scale on the left, relates to white light with a colour temperature of 2640 °K. Curves *R*, *G* and *B*, to be read in conjunction with the i_f scale on the right, relate to red, green and blue light respectively; these latter were obtained by placing coloured filters in the path of the light, no other detail of the experimental set-up being altered. The filters in question have transmission characteristics similar to those of the red, green and blue filters used in a colour-television camera. (They are Gevaert R 586, G 537 and B 488 respectively.) For all four kinds of light, i_f already attains its saturation value at *U* values of the order of some tens of volts. The curves are slightly concave at the extreme low-voltage end.

out the thickness of the layer, that the field-strength is high enough for all the charge-carriers liberated by the incident light to be able to contribute to the flow of current to the fullest extent. Now, in the "Plumbicon" the donor and acceptor concentrations in the *I* layer are so small that the band curvature extends through the whole thickness of the layer, even when the applied voltage is very small.

Moreover the impurity that is inevitably present is of a type that least affects the properties of the tube. To a very limited extent the middle sublayer is a *P*-type semiconductor. The region where field-strength is lowest and where, accordingly, most of the recombination takes place — from now on we shall call it the "field-free" region — is therefore to be found in that part of the *I* layer which lies next to the *P* layer (cf. fig.7*a*). Here the field-free region does much less harm than if it were close to the *N* layer (fig.7*b*), owing to the way in which PbO absorbs light. Red light is absorbed rather gradually as it passes through the target, but most of the blue is absorbed in the first 5 μm of the target thickness. In a target containing a wide field-free region immediately behind the *N* layer, charge-carriers generated by blue light make no contribution whatsoever to the photo-current. It is true that the field-free region shrinks as the applied voltage is increased, but those parts of it which lie immediately behind the window, and which absorb the greatest amount of blue light, are the last to be affected. In these circumstances the characteristic for blue light changes at the low-voltage end (*fig. 12*), and a far higher value of applied voltage is required for saturation than when the central layer has *P*-conducting properties and the field-free region is on the gun side.

From the point of view of service life, it is even an advantage for the *I* layer to be somewhat *P*-type, as is shown later.

In approximation, the field-strength can be equated with U/d in cases where the curvature of the bands is only slight, d being the thickness of the layer. The velocity of the charge-carriers is then $\mu U/d$, and for the transit time to be smaller than the mean life τ of the charge-carriers it is required that $d^2/\mu U < \tau$. For a U value of 10 V and a layer thickness of 10 μm the requirement becomes $\mu\tau > 10^{-7}$ cm²/V.

The way in which i_f, the photo-current flowing through the PbO layer, depends on the incident luminous flux L may be found represented graphically in *fig. 13*. Plotted on log-log paper, this function appears as a straight line; that is to say, it can be expressed by a formula of the type $i_f \propto L^\gamma$. The value of the exponent is close to unity. Generally, γ is found to have a value between 0.8 and 1.0, which means that the photo-current is more or less proportional to the incident luminous flux; a single value, uniform at all light levels, can therefore be quoted for the sensitivity of the "Plumbicon" [8]).

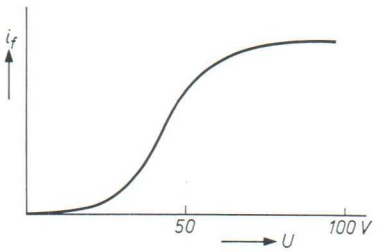


Fig. 12. When the *I* layer is too strongly *N*-type, the low-voltage end of the i_f,U characteristic will be depressed, and a high applied voltage will be required for i_f to attain its saturation value. This effect is strongest for blue light.

The significance of this will be further discussed in the article quoted in [2]). For the tube whose characteristic appears in fig. 13, this value is 210 μA/lm. (For the conventional vidicons one value only cannot be quoted; γ differs widely from unity — roughly $\gamma = 0.5$ — and is even dependent on L.)

[8]) Also when i_f is below the saturation value the photo-current is still proportional to the incident luminous flux, and from this fact we can probably conclude that the recombination is monomolecular.

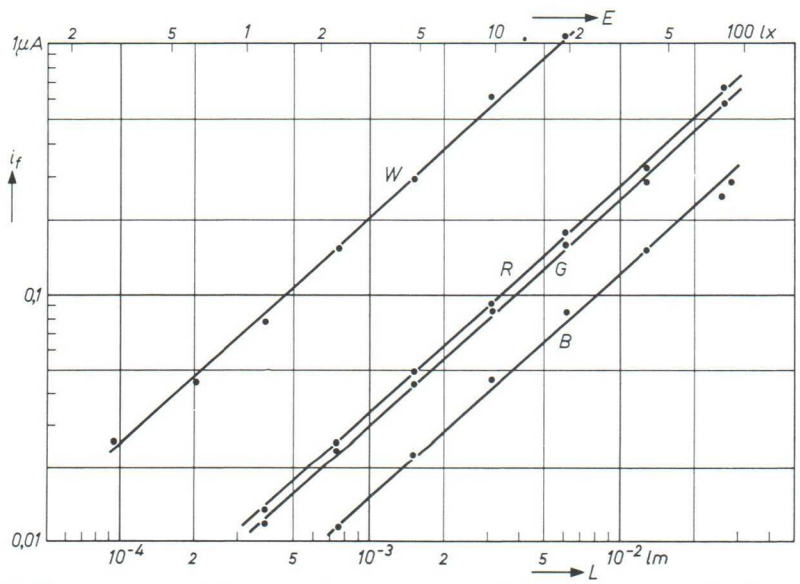


Fig. 13. The measured variation in photo-current i_f as a function of luminous flux L (or illumination E) for white, red, green and blue light (curves *W*, *R*, *G* and *B*), the applied voltage having been kept at a fixed value. In contrast to conventional vidicons, in the "Plumbicon" L and i_f are roughly proportional ($\gamma \approx 1$) for all kinds of light. Curves *R*, *G* and *B* were obtained in the same manner as the corresponding curves in fig.11. They have been plotted against the luminous-flux values for white light, i.e. the flux incident on the tube before the filter was interposed.

It has been found that tubes made in the same way show very little spread in sensitivity, the widest variation being of the order of 10 μA/lm. Another point of interest is that higher sensitivity values can be achieved, where necessary, e.g. 400 μA/lm, by modifying the deposition process or by depositing a thicker layer of PbO.

When sensitivity to different kinds of light is measured, γ still has a value around unity.

The spectral sensitivity distribution

It has already been pointed out that PbO does not absorb different kinds of light to the same degree. The shorter the wavelength, the higher is the degree of absorption; roughly speaking, blue light is almost completely absorbed after passing through the first 5 μm of PbO, but quite a high proportion of red light passes right through the target (*fig.14*). Clearly, then, the spectral sensitivity of the target can be varied between rather wide limits: a higher relative sensitivity to red light can be obtained by making the *I* layer thicker; maximum sensitivity to blue can be obtained by making the *N*-type layer as thin as possible, and by ensuring (for the reasons explained above) that there is no field-free region in the part of the *I* layer directly adjoining the *N* contact.

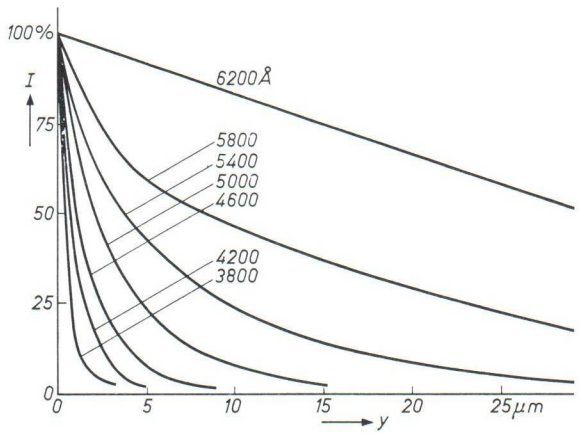


Fig. 14. The absorption of monochromatic light of different wavelength by PbO. The variation of the intensity *I* is plotted against the distance *y* the light has travelled.

The upper limit to the range of wavelengths within which the tube is sensitive is roughly the same as that for red PbO. A band gap of 2.0 eV corresponds, as already mentioned, to a cut-off at 6200 Å. The lower limit depends on the thickness of the *N* layer and the distribution of potential in the neighbouring part of the *I* layer. (The upper limit can be shifted to considerably longer wavelengths, without cooling being required, by making the target of a material having a smaller band gap. See the article quoted in [4].)

Fig. 15 shows the spectral sensitivity distribution of two tubes whose PbO targets were intentionally made in a different way. The dashed line represents the response of the human eye. Curve *1* is that of a tube whose PbO target was made by the standard process. By modifying the process to reduce sensitivity to blue — and possibly increase sensitivity to red — the peak of the characteristic can be shifted towards the right. It is even possible to shift it to the right-hand side of the response curve of the eye (curve *2*).

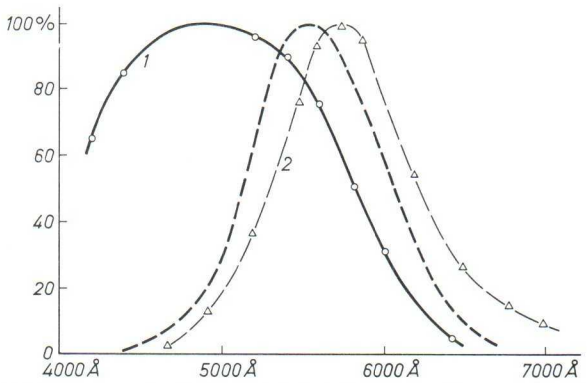


Fig. 15. The spectral sensitivity distribution of two tubes of the "Plumbicon" type. The thick dashed line represents the spectral sensitivity of the eye. Curve *1* is that of a tube whose PbO target was made by the standard process. The peak of the curve can if desired, be shifted to the right (curve *2*), to the other side of the eye sensitivity curve.

Since the (standard) "Plumbicon" has a spectral sensitivity distribution much closer to that of the human eye than the corresponding characteristics of the ordinary vidicon (Sb_2S_3) or the image orthicon (Ag-Bi-O-Cs), it requires no filters when used in monochrome television, yet gives a far better gradation of colours than the tubes just named. Also for *colour* television the somewhat smaller sensitivity to red of the "Plumbicon", compared to the human eye, does not appear in practice to be a serious objection [2]). In such a case, however, a tube with greater sensitivity to red offers a solution [4]).

Resolution

A light-to-dark transition in the image of the scene projected on the window is not reproduced in the video signal with exactly the same abruptness. Apart from the properties of the electronic equipment to which the pick-up tube is connected, and the properties of the electron beam — factors we shall not be discussing here — this is due to the fact that the corresponding transition in the charge image formed on the free surface of the photoconducting

layer between two scans is less sharp. Two effects are responsible for this: in the first place a certain amount of the light is scattered in the target, and in the second place some transport of charge takes place in the PbO parallel to the target surface (cross-conduction). The latter effect can be subdivided into cross-conduction in the P layer and cross-conduction in the I layer.

It was found that the lack of definition caused by blue light is much less than that caused by red. If the target thickness is reduced, blurring of red outlines becomes less but that of blue remains the same. The reason will be clear if it is remembered that blue light does not penetrate so far into the target as does red (*fig.16*).

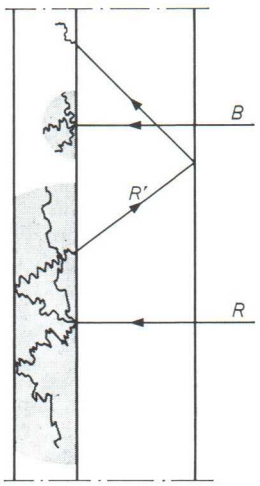


Fig. 16. PbO has a lower absorption for red light than for blue, and consequently the scattering of red light has a more adverse effect than that of blue on the resolving power of the PbO target. The shaded areas indicate, schematically, the extent of the zones penetrated by the diffused light originating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

A certain amount of diffused light escapes from the target by way of the front face of the target, but is then reflected back by the front face of the glass window; it is again the less rapidly absorbed red light (R') that is mainly responsible for the contribution of this effect to the loss of definition.

Consequently from the standpoint of definition it is desirable to make the target as thin as possible. As we have seen, this will involve a reduction in sensitivity to red light; furthermore, the capacitance of the target will be increased and, for reasons which will be explained below, only a limited increase in target capacitance can be accepted. Thus there is not complete freedom of choice of the target thickness. (The difficulty disappears if the PbO is replaced by a basic material with a much greater absorptivity for red light [4].)

Cross-conduction in the P layer can naturally be limited by making the layer as thin as possible; and further, the less the doping of the layer, the less cross-conduction there will be. However, as we have seen, if the layer is too lightly doped the dark current will be excessive. Also, as will be explained below, from the viewpoint of service life it is advantageous to make the P layer thick and dope it heavily. Accordingly, here again the choice is not entirely free.

Cross-conduction occurs in the I layer when, owing to insufficient purity of the material, this layer contains a "field-free" zone. In such a zone the charge-carriers are liable to fan out instead of crossing the layer by the most direct route. This does not matter if the charge-carriers in question are electrons, since the point where these arrive on the signal plate has no bearing on the resolving power. But the place where the holes arrive is of course important: a hole that has not crossed the target by the most direct route alters the distribution of potential on the free target surface in a way that does not correspond to the light pattern of the broadcast scene. So from the viewpoint of resolving power also, it is desirable that the I layer be as pure as possible.

Before the values of resolving power relevant to the "Plumbicon" are quoted, first a word about how this resolving power is expressed in figures. Suppose that a pattern like that in the upper half of *fig.17* is being projected on the screen of the tube. The pattern consists of alternate vertical light and dark stripes of the same width. In some parts of the screen the width of the stripes is such that 20 light-and-dark pairs would completely fill the picture height (in the language of the television engineer this is called 40 "lines"); elsewhere this number is 200 (400

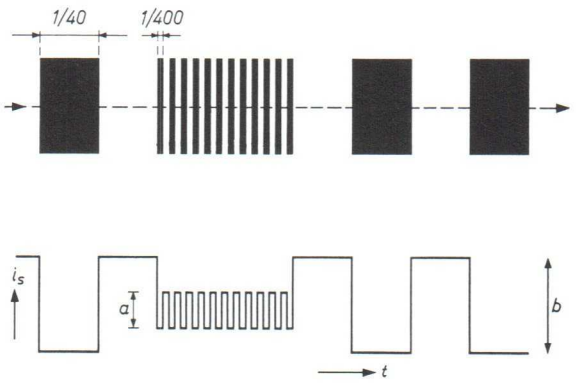


Fig. 17. Explanation of how the resolving power of a television camera tube is expressed in figures. A pattern of alternate light and dark vertical stripes, some having a breadth of 1/40th of the picture height and others a breadth of 1/400th of the picture height, is projected on the screen. This pattern is scanned in the direction of the dashed line. The broad and the narrow stripes give rise to alternating voltages with fundamental frequencies of 0.5 Mc/s and 5 Mc/s respectively. The ratio between amplitudes a and b, expressed as a percentage, and known as modulation depth, provides the required measure for resolving power.

"lines"). If an electron beam scans the corresponding charge image in the direction of the broken line, the signal current (fig.17, lower half) will have the form of an alternating current with fundamental frequencies of 0.5 and 5 Mc/s respectively (in the case of a 625-line system with a frame period of 1/25th second). Parts of the signal current corresponding to the broad dark stripes will have approximately the dark-current value, but the narrow dark stripes will give rise to higher current values. Parts of the signal current arising from the broad light stripes will have the same value as if the window were illuminated over its whole surface; the narrow stripes will yield lower current values. Let the letters a and b denote the difference between the light and dark values of i_f in the region of fast and slow alternations respectively. The ratio a/b expressed as a percentage, and known as the modulation depth, is commonly adopted as a measure of resolving power.

A more detailed impression of the tube properties is obtained if, instead of restricting the measurement of a to a pattern with 400 lines per picture height, the variation of a/b is investigated when a number of patterns with different stripe breadths are used. An example of this kind of measurement, done on a random-selected tube of the "Plumbicon" type, is shown in *fig.18*. It will be noted that at 400 lines the tube under investigation had an a/b ratio of about 45%. In general a/b is found to be not less than about 35%.

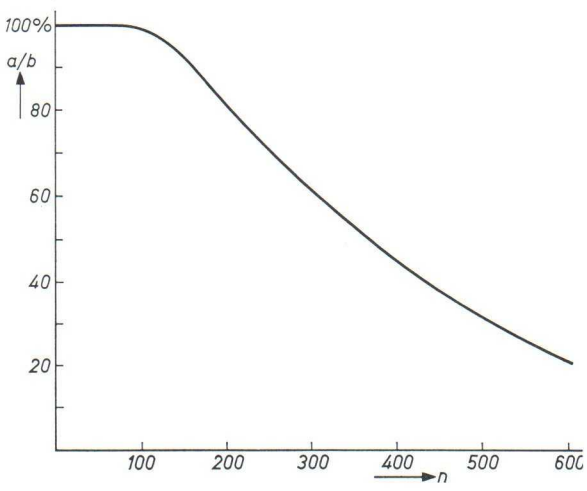


Fig. 18. A detailed description of resolving power can be obtained by plotting the variation in a/b against the breadth of the narrow stripes in the pattern appearing in fig. 17, the stripe breadth being expressed as the number n of black-white pairs of such stripes that would be required to fill one picture height. The figure shows the results of measurements done with white light on a random-selected tube of the "Plumbicon" type having a PbO layer 20 μm thick. As is inevitable, the ancillary equipment causes some interference; so normally such measurements give a figure that is lower than corresponds to the real performance of the tube. By adopting a very careful procedure, the difference is here reduced to a minimum.

Speed of response

When a sudden change occurs in the luminous flux incident on the photosensitive layer of a pick-up tube, the signal current does not immediately reach its new equilibrium value. When the target of a tube undergoes the variations of illumination shown in *fig.19a*, in the signal current the forms of response

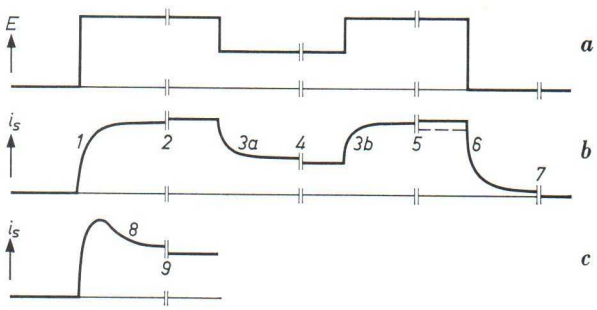


Fig. 19. The various forms of delayed response which, in principle, a "Plumbicon" may exhibit. In a good tube they are too slight to cause any trouble, and some are absent altogether. Diagram (*a*) shows the programme of varying intensity of illumination presented to the tube: the light is switched on, decreased in intensity after an appreciable time has elapsed, then brought back to its full intensity and finally switched off. The various forms of delayed response are represented in diagram (*b*). They are: 1) black-white inertia, 2) black-white trailing (a slight increase over a long period), 3a) and 3b) intermediate inertia, 4) intermediate trailing, 5) persistence (which may also manifest itself in *attenuation* of the signal), 6) white-black inertia, and 7) white-black trailing. Instead of effects (1) and (2) the tube may exhibit, see diagram (*c*): 8) fatigue and 9) fatigue trailing.

illustrated in fig.19*b* may be encountered. The names that we have given these inertial effects may be found in the caption to that figure.

From the practical point of view one of the most important forms of response is "intermediate inertia". (Inertia effect occurring when the light intensity changes from white to grey or grey to white.) Black-white inertia only becomes noticeable under the most unfavourable conditions, and white-black trailing can be compensated electrically. Intermediate inertia is always fully manifest in the received picture, however, and the same applies to persistence and fatigue effects.

All the above forms of response can be regarded as resultants of two components, that are due to 1) the electron beam being incapable of supplying an unlimited amount of charge during the brief time in which it is directed on a given picture element; this component is accordingly called *beam-current-* or *discharge lag*, and 2) the presence of traps in the *I* layer. In the "Plumbicon" the contribution made

by both these factors have been reduced to a satisfactorily low level.

Both the discharge lag and that due to the presence of traps decrease with increasing target voltage U, although it should be observed that no further appreciable increase in the speed of response is achieved above a certain value of the applied voltage. The traces in *fig. 20* show this effect. It will be noticed that the highest speed of response is attained at the relatively low U value of 30 V. This speed of response is very satisfactory; the change in signal level resulting from a transition from strong to less strong illumination (and involving intermediate inertia) is 95% complete after only 3/50ths of a second.

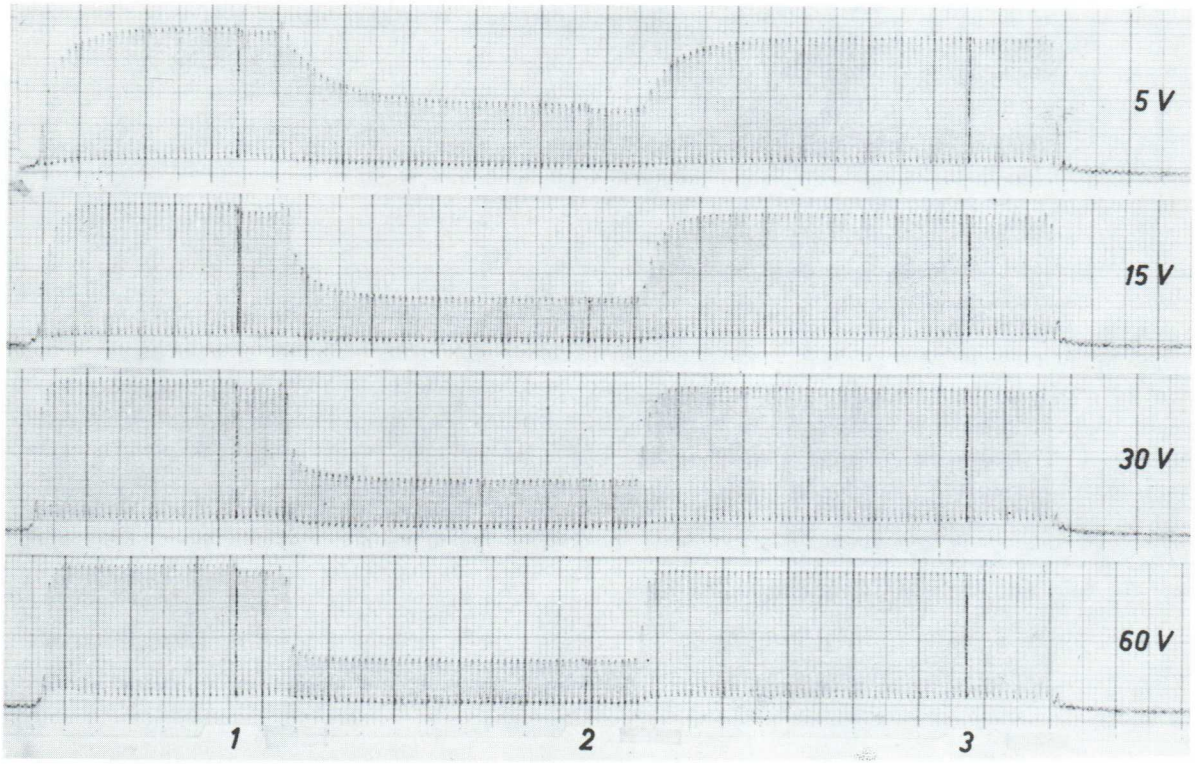


Fig. 20. The response of a randomly selected "Plumbicon" tube at four values of U, the voltage across the PbO target. At the relatively low U value of 30 V the response of the tube is already quite fast, and there is scarcely any persistence. The traces were obtained by illuminating the top half of the screen only, so that the signal current consisted of a train of pulses spaced at intervals of 1/50 s (*half* a frame period because of the interlaced scanning). The programme of illumination was equal to that of fig. 19 and began in each case with a dark period lasting one minute. At the points *1*, *2* and *3* the reading was interrupted for 10, 10 and 30 seconds respectively.

The discharge lag

When we were discussing how the potential of the free surface of the PbO target adjusts itself to an equilibrium value, we found that the current flowing towards that surface was dependent on its potential V (fig. 3). The deflection and subsequent fall-off in the curve of current versus potential was explained as a result of an increase in secondary emission. The initial rise (V low) is due to the fact that the electrons leaving the cathode do not all have the same velocity in the axial direction, the axial components of velocity having a quasi-Maxwell distribution. To a good approximation, the rising portion of the curve can be described by an expression of the form $i_a = ae^{b\Phi}$, where a is a constant connected with the beam-current and b is inversely proportional to the cathode temperature; Φ differs from V by a small constant amount.

Over the range of illumination values within which V remains small and the above formula accordingly remains valid, the discharge lag has the following properties:

1) It is independent of the beam-current.
2) It decreases with decreasing target capacitance C.
3) When the light intensity drops, the higher the new intensity the shorter the discharge lag.
4) The lower the cathode temperature, the shorter the lag.

Property (2) explains why the discharge lag initially decreases as voltage U is increased. At low U values the I layer contains a field-free region which,

in effect, forms part of the adjoining contact. This region becomes narrower and narrower as U increases. Consequently the distance between the capacitor "plates" becomes greater and the target capacitance decreases, giving a shorter discharge lag; see formula (2).

At higher intensities of illumination, corresponding to the region in which the i_a versus V curve starts to change direction, (1) ceases to apply: the discharge lag decreases with increasing beam-current when the light level is high. Property (2), relating to the low target capacitance desirable, is retained at high intensities of illumination; properties (3) and (4), for obvious reasons, are not.

In practice, for the decrease of the discharge lag, only the decrease of the target capacitance can be considered. If a time constant shorter than about 1/25th of a second is required for the response to a change from dark to light (to a moderate intensity of illumination giving rise to a photo-current of 10^{-8} A, say), then the target capacitance must not exceed about 2000 pF. Also at high intensities of illumination the discharge lag will continue to be negligibly short, provided it is possible to keep the target capacitance below 2000 pF.

These requirements can be fulfilled very comfortably with a target having the *P-I-N* structure described above. For example, suppose that the PbO layer is 10 μm thick and that it has an I layer so pure that it contains no field-free region at the U values employed in practice — when discussing sensitivity we saw that this degree of purity can actually be attained — then the target will have a capacitance ranging from 1000 to 1500 pF.

The four properties of the discharge lag at relatively low illumination levels, as listed above, can be inferred from the following. As before, we shall first assume that the beam is supplying charge *continuously*. The behaviour of V, the potential of the free surface of the target, is in accordance (cf. figs. 2*b*, 3, 4 and 5) with the differential equation

$$\frac{dV}{dt} = \frac{1}{C}(i_f - i_a).$$

(The whole target is imagined to be evenly illuminated; but the above equation applies equally to a single picture element, the capacitance C and the (negative) charge supply then being smaller in the same proportion.) Further:

$$i_a = ae^{b\Phi}.$$

Now suppose that a change in illumination abruptly raises the photo-current i_f from the value i_1 to the value i_2; the problem is to ascertain how i_a, initially equal to i_1, arrives at its new value of i_2.

On solving the above set of equations and eliminating V, we obtain:

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

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 10 V is placed on the variation of V and that a photo-current of up to 10^{-7} A is required, then C must not be smaller than about 800 pF.

The capacitance of the PbO layer in the "Plumbicon" has, it appears, a value such that on the one hand the lag is sufficiently short and on the other hand the influence of the electron beam on neighbouring picture elements can be neglected. The signal given by a picture element depends solely upon the intensity of the incident light, and is independent of its position, of its history, and of the situation in the surrounding picture elements. This is why the tube is so very suitable for colour television [2]).

Lag due to traps

Let us now go a little more deeply into the type of inertial response caused by traps. It is known from photoconductor theory [3]) that the presence of traps does not in the first instance affect the relationship between intensity of illumination and the steady-state electron concentration in the conduction band, or similar relationship; it does however have a bearing on the *speed* with which a new situation supervenes when the intensity of illumination is altered. Generally there are many more electrons in the traps than in the conduction band, the respective concentrations being c_t and c_c (much the same thing applies to the holes). If the illumination E is increased, with a consequent increase in c_c, the c_t/c_c ratio must nevertheless remain the same, and this implies a large absolute increase in the number of trapped electrons. Initially the demand is largely supplied by electrons liberated by the light incident on the target; c_c cannot therefore jump directly to the value corresponding to the change in E. The c_t/c_c ratio is proportional to N_t, the concentration of traps; the greater N_t is, then, the greater will be the (absolute) deviation of c_c at a given time after the increase in E. Similar reasoning applies if E is reduced: when this happens a large number of filled traps have to "dry out", and the higher the concentration of such centres, the greater will be the absolute deviation of c_c.

Since the concentration in the conduction band

falls off with increasing U, it is understandable that these inertial effects should become less noticeable at higher values of applied voltage.

Apart from these *direct* consequences of the presence of traps, which will not be further analysed, the traps have an *indirect* effect which at low U values may be clearly manifest in the response of the tube. The capture of charge-carriers in traps can modify the curvature of the energy bands in the I layer so drastically that the characteristics of the tube are affected. Any such modification is purely temporary, of course, beginning subsequent to a change in E and proceeding at the same rate as the establishment of a new value of space charge (c_c+c_t), so that for the observer it has the character of an inertial effect. We shall now look a little more deeply into this effect; for simplicity we take as an example the phenomena of *fatigue* and *black-white inertia*. To obviate misunderstanding it should be pointed out that in the discussion which follows, cases will only be considered in which the applied voltage U is chosen so small that a field-free region is present in the I layer.

Fatigue, a slow decay in photo-current following an abrupt increase, occurs because charge-carriers are trapped in such a way that the field-free region expands. Here a distinction must be made between cases in which the I layer is slightly N-type, and those in which it is slightly P-type.

As we have seen, if the I layer is slightly N-type there will be a field-free zone on the N-layer side when U is small. In consequence of hole capture this zone widens (see *fig. 21a*) with the result that the sensitivity of the target is modified, particularly its sensitivity to blue.

If the I layer is slightly P-type there will be a field-free zone on the P-contact side, which widens owing to electron capture (fig. 21*b*). This affects the tube's sensitivity to red light.

Under normal operating conditions U is high enough to eliminate the field-free zone or at least to reduce it to such small dimensions that the effects just described are of little importance.

In the case of *black-white inertia* there is, in addition to the direct effect of charge-carrier capture, a side-effect that is the opposite of that just discussed; here the field-free zone *shrinks*, causing the sensitivity of the target to increase (see *fig. 22*).

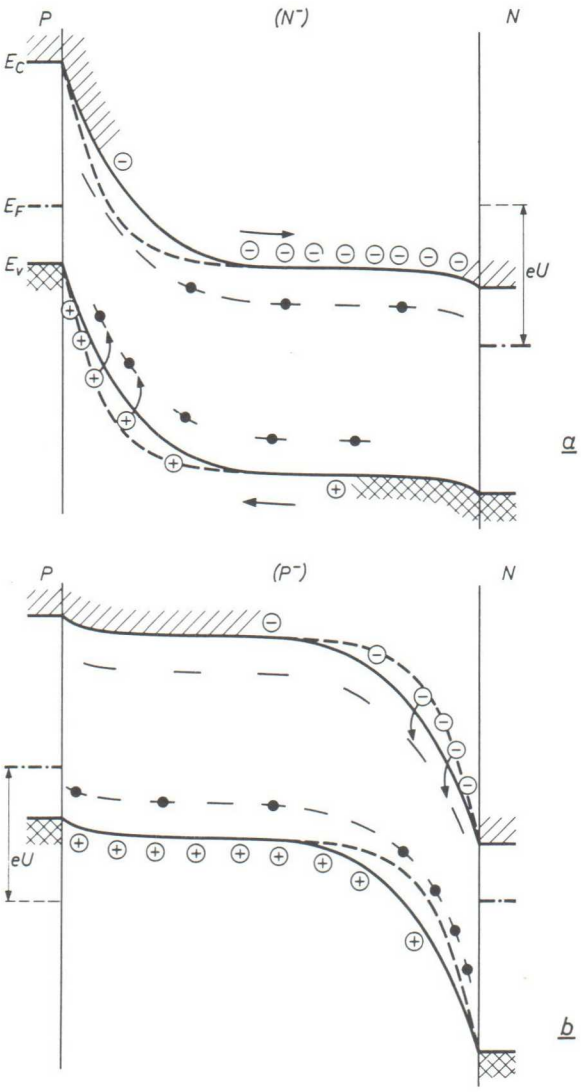


Fig. 21. To explain "fatigue".

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.

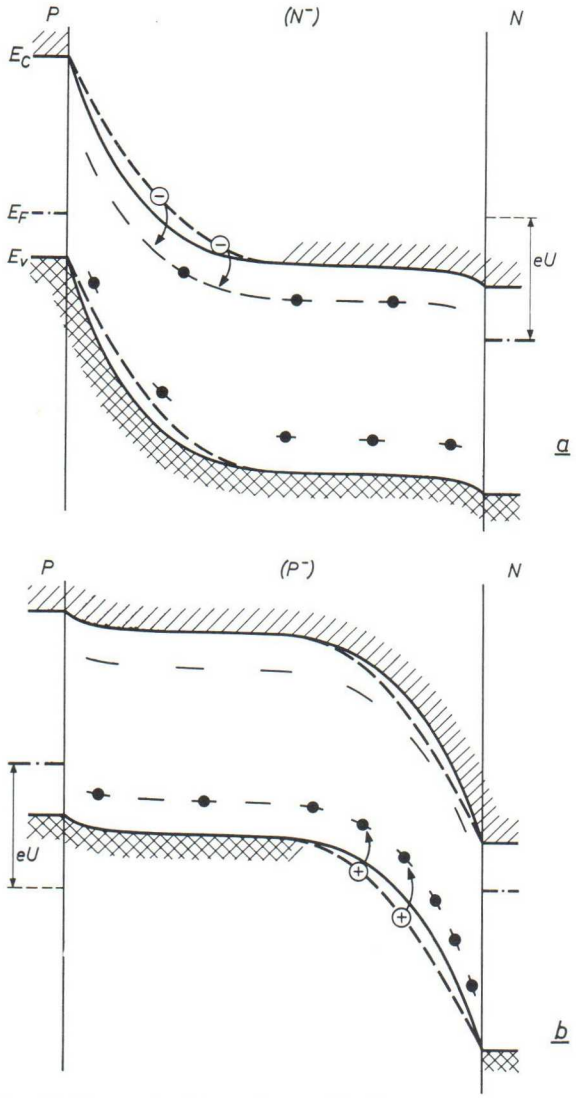


Fig. 22. To explain "black-white trailing".

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

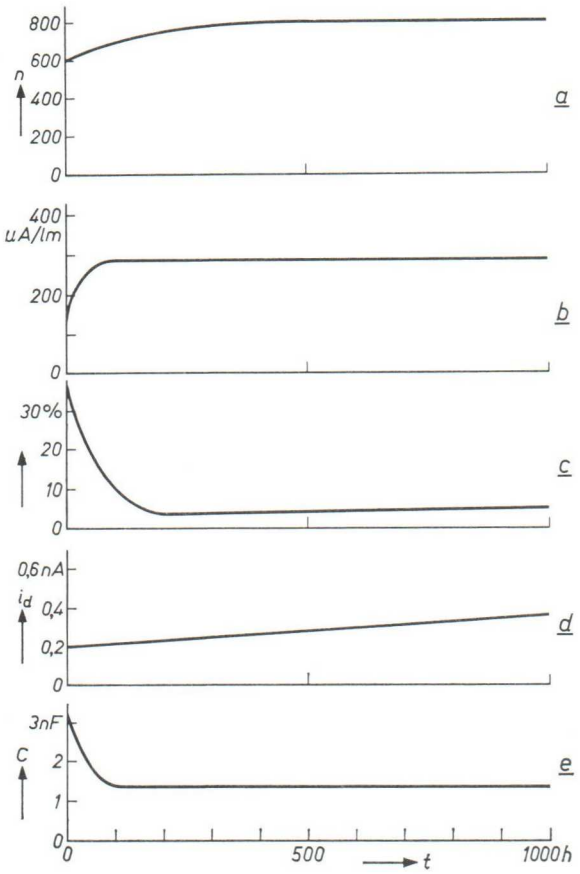


Fig. 23. Changes in various tube characteristics and in the capacitance of the PbO target, measured on a randomly selected tube of the "Plumbicon" type, during the first 1000 hours in operation. *a*) Resolving power, shown as the number of lines that can just be resolved by the eye (the corresponding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

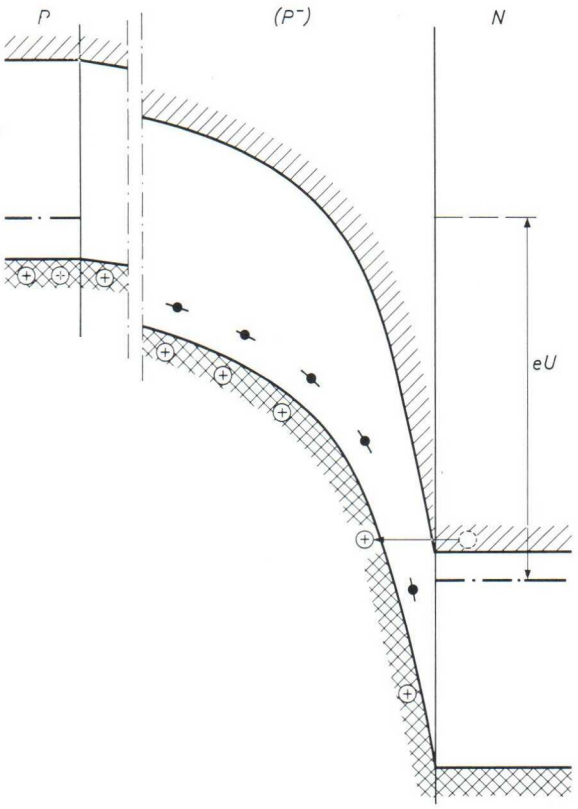


Fig. 24. To explain the occurrence of a strong dark current of holes, when the inner sublayer contains a limited region of high *P*-conductivity. The barrier can then be so narrow that tunnelling is possible (see arrow).

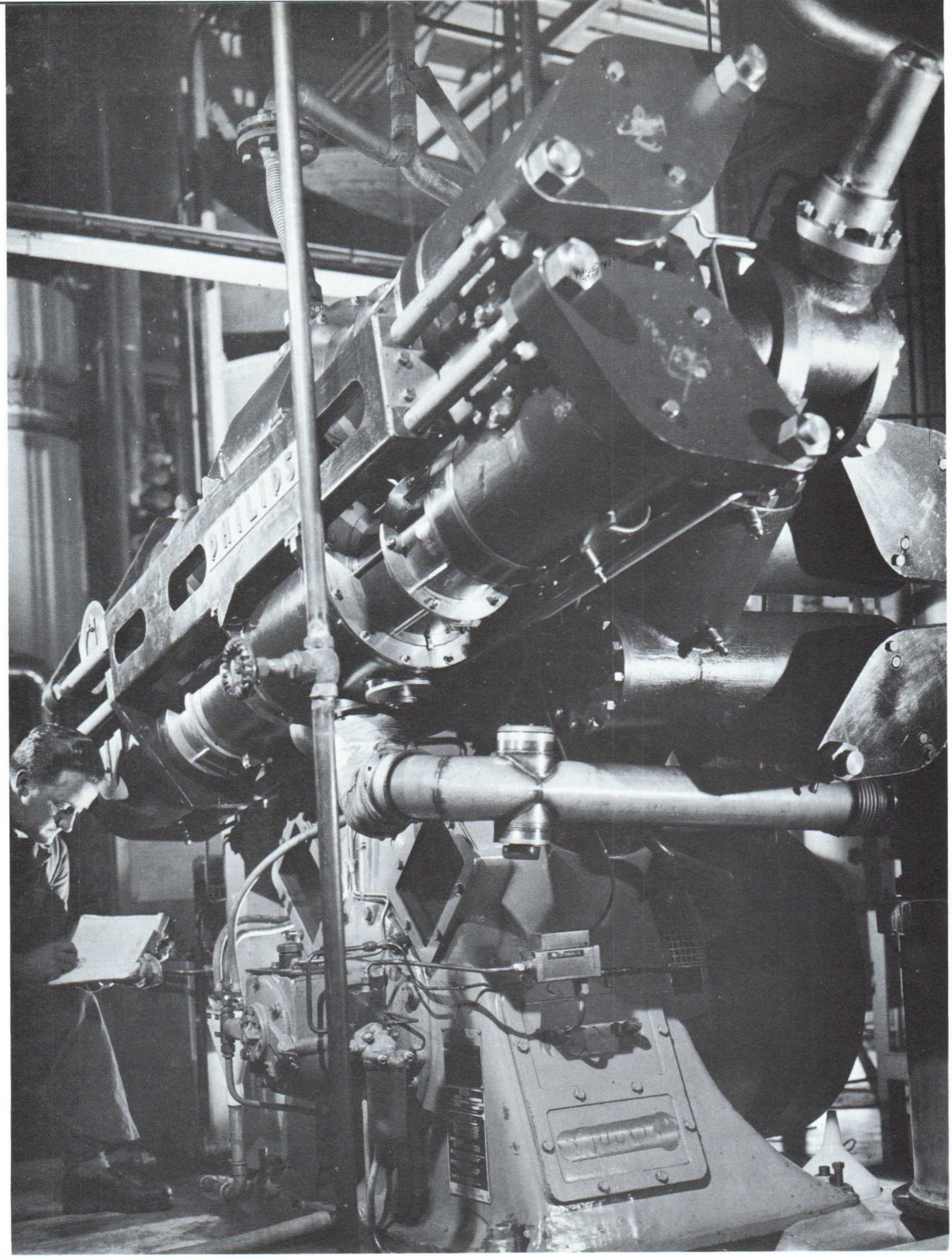
By arranging for fabrication of the tube under carefully controlled conditions in dust-free rooms, each part being subjected to rigid inspection, it has been possible to eliminate most of the causes of these defects, and the PbO targets now being produced are accordingly of consistently high quality 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).

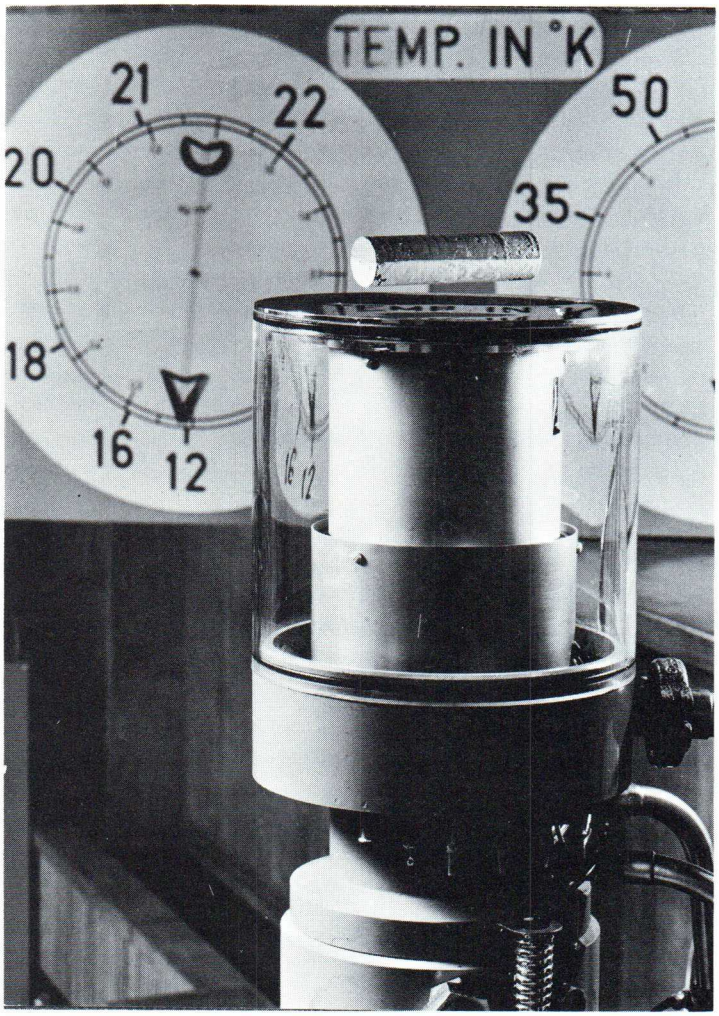


Industrial gas refrigerating machine

This four-cylinder machine (type C), designed for a cold output of 20 kW at a temperature of 77 °K and developed by Philips Industrial Equipment Division, is based on a rather unusual concept. The two pistons of each process cylinder are driven *hydraulically*, the pistons of adjacent cylinders being operated in tandem from plungers on the arms of a normal Vee-drive. Each piston is hermetically sealed by means of a rolling diaphragm (see this book, page 423). Using hydrogen as a working fluid, the machine attains an efficiency (compared to the Carnot-value) of more than 41 %.

(See: A. A. Dros, An industrial gas refrigerating machine with hydraulic piston drive, Philips tech. Rev. **26**, 297-308, 1965.)

Scientific research and technical development work at temperatures around the boiling points of hydrogen and helium — e.g. in some branches of solid state research, the development of masers, certain aspects of computer component research, etc. — are still restricted to a considerable extent by the size and complexity of the machines required to produce these extremely low temperatures. Consequently this work has until now been mainly confined to specially equipped cryogenic laboratories. The appearance of a small, easily operated gas refrigerating machine that can work with good efficiency at 20 °K, can even reach temperatures in the region of 12 °K, and can produce about 5 litres of liquid hydrogen per hour, will therefore undoubtedly be welcomed by many research workers.



A gas refrigerating machine for temperatures down to 20 °K and lower

G. Prast

621.573

For some considerable time now, Philips Research Laboratories have been investigating the possibilities which the Stirling cycle offers as a refrigeration process. It soon became clear that this process has very attractive properties, particularly when the temperature at which the "cold" must be supplied is substantially lower than that of the environment (i.e. as a rule room temperature). This led to the construction of a very compact cryogenerator — the well-known *gas refrigerating machine* [1] — which can work with good efficiency in

Ir. G. Prast is a research worker at the Philips Research Laboratories, Eindhoven.

[1] See J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine, Philips tech. Rev. **16**, 69-78, 1954/55, and J. W. L. Köhler and C.O. Jonkers, Construction of a gas refrigerating machine, Philips tech. Rev. **16**, 105-115, 1954/55.

the temperature range from —100 °C to —200 °C (173 °K to 73 °K). This range includes the boiling points of oxygen and nitrogen. With this machine — of which there is now a four-cylinder version in addition to the original single-cylinder type — installations have been made for producing liquid air and for the separation of nitrogen from air, which are now in fairly wide use. The gas refrigerating machine has also found numerous other applications [2].

The following are the principal advantages of the gas refrigerating machine over other types:

1) The machine operates with a fixed, small quantity of gas, either hydrogen or helium, which is contained in an enclosed space where it cannot be contaminated.
2) The machine has no valves.
3) Considering its large refrigerating capacity, the dimensions of the machine are very small.
4) The machine has a high efficiency.
5) The machine requires hardly any attendance or maintenance.

After this initial success, one of the main objects of research was to discover whether the Stirling cycle could be used to attain much lower temperatures, e.g. 10 °K, still with a good efficiency. The temperature range of a machine capable of this would then also contain the boiling points of neon (27 °K) and hydrogen (20.4 °K) — that of helium lies at 4.2 °K. This continued research proved successful and indeed led to the construction of a machine of this kind. This machine, which was designed for 20 °K but is capable of operating with good efficiency in the region between 14 °K and 60 °K, is the subject of the present article. The new machine is not simply an improved version of the familiar gas refrigerating machine, but operates on a principle that may be described as a modification of the Stirling cycle. It retains, however, all the merits of the original gas refrigerating machine as summarized above.

It may be recalled that in a machine based on the "normal" Stirling cycle, the compartment containing the working gas consists of two spaces in open communication with each other via a regenerator (*fig. 1a*). One part is at the temperature T_E at which the cold is produced, and the other is at the ambient temperature T_C. The working gas is compressed in the latter space — the heat of compression being removed by cooling water in the "cooler" — after which it passes through the regenerator where it is cooled to T_E and then expanded in the other space. In this expansion the cold is generated and supplied to the "freezer". The gas then returns to the first space — the compression space — and in passing through the regenerator it is again heated to T_C, by reabsorbing the heat it left behind in its previous passage through the regenerator. The lowest temperature that can be reached is determined by the losses, which cannot of course be entirely reduced to zero.

In the final form of the gas refrigerating machine there are not *two* pistons but only one, and the gas is transported from the one space to the other by what is termed the displacer [1]. In this case the regenerator, like the cooler and the freezer, is annular in shape (fig. 1*c*).

The modified cycle in the new machine consists essentially of two such cycles combined, whereby one cycle compensates partly for the losses of the other, so that a lower temperature can be reached than is possible with a single cycle [3]. *Fig. 2* gives a schematic representation of this combined process. As can be seen, there are now *three* spaces: one compression space and *two* expansion spaces; there are also *two* regenerators. Upon expulsion from the compression space, part of the gas goes to the first expansion space and another

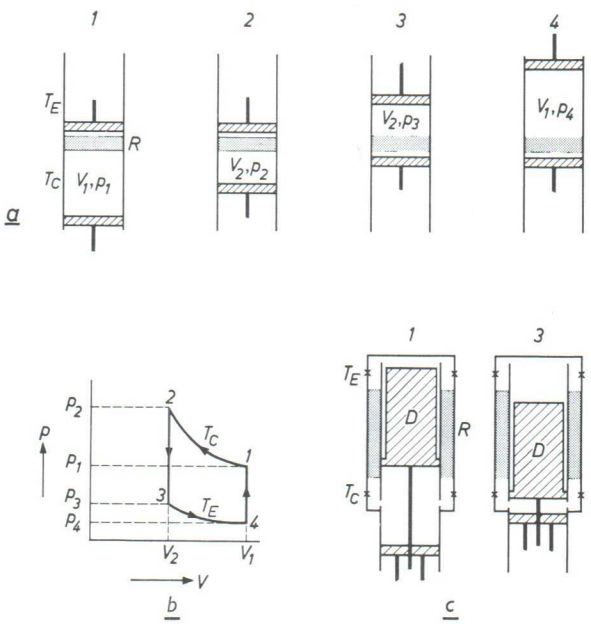


Fig. 1. *a*) Principle underlying the use of the Stirling cycle for refrigeration. Two pistons move in a cylinder which contains one space at room temperature T_C and one at low temperature T_E, and between them a regenerator R. The regenerator is a space filled with a mass of finely divided metal (e.g. thin wires). The filling has a high heat capacity and excellent heat-transfer properties. The cycle comprises four phases, whose initial states (1-4) are shown in the diagram.
b) Variation of the pressure and volume of the working gas. The phases 1 and 2 occur isothermally, phases 3 and 4 are isochoric.
c) Diagram of the actual construction of a conventional gas refrigerating machine. There is only one piston, and the cylinder space is divided into two by a displacer D. The regenerator is annular in shape. The crosses on the walls indicate schematically the location of the cooler (below) and the freezer (above) which are also annular. Piston and displacer are driven by a crankshaft and have an almost harmonic motion. Consequently the gas cycle is not exactly identical with that shown in fig. 1*b*.

part simultaneously to the second expansion space (the splitting into two fractions is represented schematically by the dashed line). The latter part passes through both regenerators. The spaces a_1 and b, together with the part of the regenerator R_1 situated below the dashed

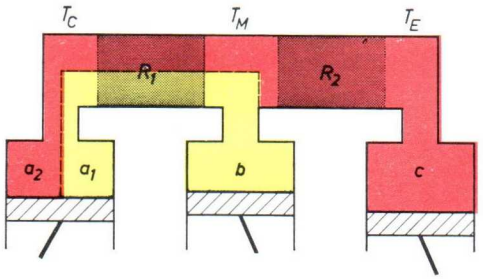


Fig. 2. Diagram of a machine with *two* expansion spaces (three-space machine). The process here can be represented as divided into two normal Stirling cycles; the first occurs in the yellow part of the space and the other in the red part. In principle the pistons in the two expansion spaces b and c need not move in phase; in our machine they do, however. The cold supplied by the first process is entirely or partly used to compensate the losses of R_1.

line (yellow), can as a whole be regarded as the space in which the one part of the process takes place, and the spaces a_2 and c together with R_2 and the section of R_1 above the dashed line (red) are where the other part of the process occurs. The first part of the process is a normal Stirling cycle, which supplies cold at the temperature T_m of the middle space. This cold is not removed but compensates the losses of R_1.

The lowest temperature we have hitherto been able to reach with the machine described in this article (designed for 20 °K) is 10.5 °K. Obviously, the only eligible working gas is helium. With a machine specially designed for the purpose it would certainly be possible to reach an *even lower* temperature. On the other hand, it is evident that the boiling point of helium cannot be reached directly with such a machine, simply because no gas exists that has a lower boiling point under atmospheric pressure.

The fact that the machine can attain the extremely low temperatures mentioned above may be elegantly demonstrated by a disc of niobium-tin on the head of the machine. At a temperature lower than 18 °K, NbSn is a superconductor, and the superconductivity is demonstrated, as shown in the title photograph, by the magnet floating over the disc [4]. For the purpose of the experiment, the normal insulation space of the head was replaced by one in which the distance between the NbSn disc and the outer wall was only 1 millimetre.

The new machine was designed entirely on a theoretical basis. Our starting point was the theory of the normal gas refrigerating machine, as earlier developed by us. Although this theory takes account of all important effects — such as the various kinds of losses, imperfect heat transfer in the regenerator, the non-ideal behaviour of the working gas, the almost harmonic motion of the piston, etc. — it is simple enough for practical application. The results have been found in all cases to be within 10% of measurements on the actual models. This theory was now extended to the case of a machine with three spaces. The extended theory is of course more complicated, and in order to use it for designing a machine it is necessary to resort to a computer. Once the required programmes are available, however, the design of a three-space machine does not take much more time. We shall return to this point later. A discussion of the theory itself does not enter into the scope of this article.

In the following we shall give a somewhat more detailed description of the modified Stirling cycle on which the operation of the new machine is based, and briefly consider the difference between the use of such a machine as a "pure refrigerator" and as a liquifier; furthermore we shall give a description of the machine and compare its performance with that of other installations, and finally review the principal possibilities for application. We shall begin by considering the normal Stirling cycle at somewhat greater length, with particular reference to the lowest temperature that can be reached with it.

The minimum value of T_E for the normal Stirling cycle

To determine the lowest temperature that can in practice be reached using a normal Stirling cycle, we have to find how the (gross) cold production and the cold losses vary with T_E. The lowest attainable value of T_E is of course that at which the losses are equal to the cold production. These cold losses include those due to regeneration, conduction and insulation; the regeneration loss is the most important of the three. We can do the calculation by representing the process in a highly schematic form. We assume that a given mass m of gas is isothermally compressed at room temperature to a pressure p_1, then isobarically cooled in a regenerator, and finally expanded again isothermally at low temperature (T_E) from the pressure p_1 to a pressure p_2. In this process the gross cold production Q_E is given by:

$$Q_E = \int_{p_1}^{p_2} p\,dV = mRT_E \ln p_1/p_2 . \quad . \quad . \quad (1)$$

[2] For a concise review, discussing the construction and various applications of the gas refrigerating machine, see J. W. L. Köhler, Progress in Cryogenics **2**, 41, 1960. An air fractionating plant recently put into operation and using the four-cylinder machine was shown in Philips tech. Rev. **25**, 340, 1963/64 (No. 11/12).

[3] See also G. Prast, Cryogenics **3**, 156, 1963. The idea of modifying the Stirling process in this way is due to H. Fokker and J. W. L. Köhler, both of Philips Research Laboratories (see e.g. British patent No. 749 815).

[4] See for example A. H. Boerdijk, Levitation by static magnetic fields, Philips tech. Rev. **18**, 125-127, 1956/57.

The gross cold production is thus proportional to the (absolute) expansion temperature T_E and is therefore smaller the lower the value of T_E.

The amount of heat which the regenerator must extract from this quantity of gas in order to cool the gas from T_C to T_E is $mc_p(T_C - T_E)$. In practice this does not happen completely; the regenerator extracts only a part η_r, and there is thus a cold loss ΔQ_R, called the regeneration loss, whose magnitude is given by:

$$\Delta Q_R = (1 - \eta_r)mc_p(T_C - T_E) \quad . \quad . \quad . \quad . \quad . \quad (2)$$

The quantity η_r is called the regenerator efficiency. In practice η_r is in the region of 0.99, so that the loss is roughly 1 %. Since η_r and c_p are nearly independent of temperature, the regeneration loss is proportional to $(T_C - T_E)$. For $T_E = T_C$ the loss is zero and becomes greater the lower the value of T_E. *Fig. 3* presents this variation of Q_E and ΔQ_R in the form of a graph. The lowest obtainable temperature is given by the abscissa of the point where the two lines intersect. Its value can be calculated by equating the two expressions just arrived at for Q_E and ΔQ_R. Taking $R = 2$J/g, $c_p = 5$ J/g (the value for helium), ln $p_1/p_2 = 0.7$, $T_C = 300$ °K and $\eta_r = 0.99$, we arrive at approximately 10 °K as the minimum value of T_E. As can be seen, the result depends markedly on the value of η_r. In practice there are other losses than those mentioned and the compression and expansion are not in fact isothermal. The lowest temperature we have hitherto reached with the normal system is therefore higher, being in the region of 20 °K.

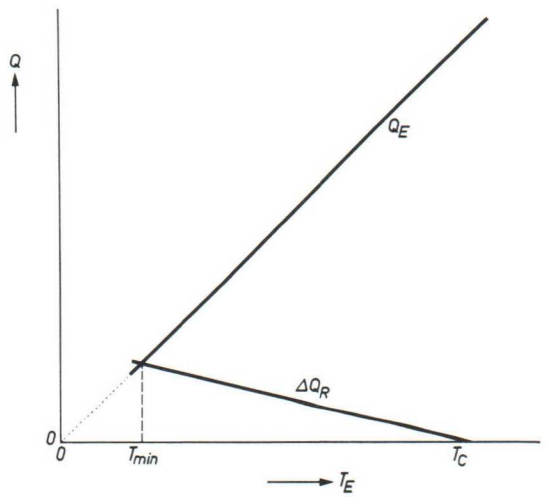


Fig. 3. Variation of the gross cold production Q_E and the regeneration loss ΔQ_R as a function of the temperature T_E of the expansion space in a normal Stirling cycle. Both curves are straight lines; extended, the first passes through the origin, the other passes through the point $T_E = T_C$. The lowest temperature T_{min} attainable is given by the abscissa of the point where the two lines intersect. At this temperature the entire cold produced is needed to compensate the losses.

From (1) and (2) it can be deduced that the minimum temperature obtainable with an ordinary Stirling cycle is lower:

a) the higher is the efficiency η_r of the regenerator,
b) the greater the pressure ratio p_1/p_2, and
c) the lower the compression temperature T_C.

Scarcely any practical use can be made of the first two properties for lowering the minimum value of T_E. In general, to increase the efficiency of the regenerator we need to increase its volume: this brings us into conflict with b), which calls for a large pressure ratio and hence for a small dead space. The machine that can reach 20 °K has in fact been designed with an almost optimum combination of η_r and p_1/p_2.

The third possibility, that of lowering the compression temperature T_C, presents difficulties in connection with the removal of compression heat. Where T_C is below room temperature, a second refrigerating machine is then needed to remove the heat of compression. This can of course be done, and we have thus been able to produce a temperature as low as 8 °K by using a cascade system of two gas refrigerating machines. A system of this kind, however, is not nearly as easy to operate, and does not have such a high efficiency as a single machine. Prospects are better with a multi-space machine, and this has been confirmed in practice.

Operation and characteristics of a multi-space machine

To give a better insight into the merits of the modified Stirling cycle underlying the new gas refrigerating machine, we shall return for a moment to fig. 2. We saw there that the regenerator losses of the section composed of a_2-R_1-R_2-c are partly compensated by the cold production of section a_1-R_1-b, which produces cold at the temperature T_M between T_C and T_E. It is known from thermodynamics that cold is produced most efficiently at the temperature at which it is needed. This means that theoretically the best way of compensating the regeneration losses of a Stirling cycle taking place between T_C and T_E is by using an infinitely large number of cold sources of different temperatures, together spanning the whole interval T_C-T_E. By using one or two in practice, however, a substantial proportion of the possible gain is already achieved. *Fig. 4* shows the diagram, analogous with that of fig. 2, of a machine which has *three* expansion spaces, called a four-space machine; here the cold required to compensate the regeneration losses is thus produced at *two* of the temperatures lying between T_C and T_E. In the following we shall confine our considerations to the three-space machine.

Referring to *fig. 5*, we shall now try to explain how and to what extent the regeneration losses of a three-space machine are compensated by the cold production

of the middle cylinder (b in fig. 2) and the manner in which this shifts the minimum attainable temperature to a lower value. For this purpose we assume the same schematic simplifications as used in the reasoning that

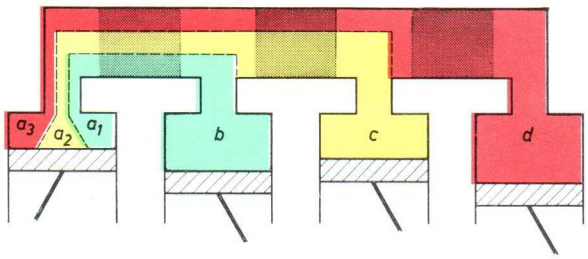


Fig. 4. Representation, analogous to fig. 2, of the operation of a machine with *three* expansion spaces (four-space machine). The process can be regarded as divided into three normal Stirling cycles.

led to equations (1) and (2) and to fig. 3. Both solid lines in fig. 5 are the same as those in fig. 3 and thus represent the variation with T_E of the gross cold production Q_E and of the regeneration losses ΔQ_R of a normal Stirling cyle. Looking again at the situation at $T_E = 75$ °K, we see that the losses there are about 20 % of the total production of cold. These losses, then, can be compensated by the cold production of only 20 % of the total quantity of gas. Let us now turn to the process illustrated in fig. 2 and assume that this 20 % is exactly the quantity of gas taking part in the (normal) cycle in the spaces a_1-R_1-b with T_M being 75°K. The remaining 80 % of the gas is similarly cooled in the regenerator R_1 to 75 °K, and we shall now see what

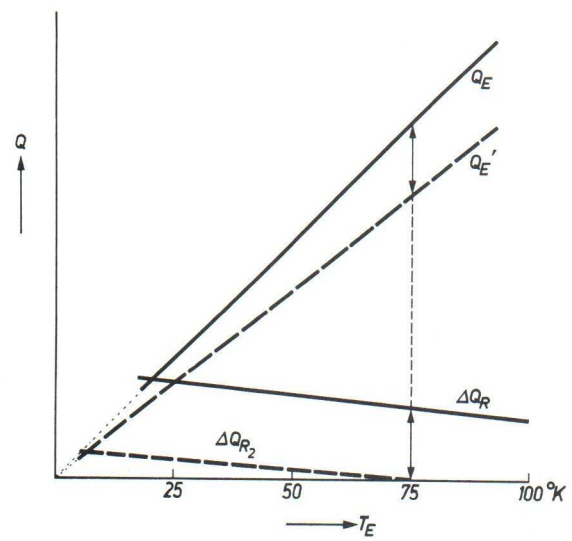


Fig. 5. Illustration of the fact that, using a three-space machine, a lower minimum value of T_E can be reached (abscissa of the point of intersection of the dashed curves) than with a machine in which the same quantity of gas goes through a single Stirling cycle (intersection point of the solid curves, cf. fig. 2).

can be done with this in the second (normal) cycle, into which we imagine the modified cycle to be split, i.e. the cycle taking place in a_2-R_1-R_2-c. The curve representing the gross cold production Q_E' of this cycle, like the curve of Q_E, is a straight line through the origin, but of course its slope is 20 % less steep. In the second cycle the curve of losses cuts the abscissa at 75 °K — the starting temperature for the part of the cycle taking place in R_2 — and has a slope which is 20 % less than that of the ΔQ_R line. Expressed mathematically (see equation 2):

$$\Delta Q_{R_2} = 0.8(1 - \eta_r) m\, c_p (T_M - T_E). \quad . \quad . \quad (3)$$

As can be seen, the point of intersection of the two dashed curves is at a lower temperature than that of the curves relating to the normal cycle; at the lowest temperature attainable with the normal cycle, the modified cycle still has a reasonably high net yield.

Refrigerator or liquefier

So far we have spoken only of the possibility of using the cold production in the middle space for the compensation of regeneration losses; the production of externally useful cold was considered as taking place only in the end space at the temperature T_E. Where a three-space machine is employed in this way we shall say it is used as a "pure refrigerator". We can also, however, let more gas expand in the middle space than is needed for compensating the regeneration losses, so that in this space, too, useful cold is available for other purposes. This can be of practical importance, for example in the liquefaction of gases, where it is necessary, prior to the actual condensation, first to cool the gas down from room temperature to the temperature of condensation. In this case, cold is needed at all temperatures between room temperature and the condensation temperature. To effect this cooling most advantageously we should again, as noted in regard to the compensation of the regeneration losses, need to have a series of sources of cold, each operating at one of the temperatures within the interval. If a Carnot cycle took place in all these sources, the refrigeration needed for liquefaction would be obtained in the most efficient way. It is evident that here, too, we can benefit by arranging for the cold needed for cooling the gas to be produced not only at the lowest temperature of the interval, but also in the middle stage of the machine. Calculation shows, for example, that the cooling and condensation of hydrogen — condensation temperature 20.4 °K — requires:

a) with an infinite series of Carnot cycles . . . . . 14×10^6 J/kg;
b) with an ideal, pure refrigerator . . 62×10^6 J/kg;
c) with one intermediate temperature 27×10^6 J/kg.

If we calculate the intermediate temperature at which the useful effect is greatest, we find for $T_E = 20$ °K that this should be roughly 80 °K both when the machine is used as a "pure refrigerator" or as a liquefier. From the equality of both T_M values it is obvious that the design must depend on the purpose for which the machine is to be used; the middle stage of one and the same machine, with a given T_E and T_M, obviously cannot have two different values of cold production.

When cooling by means of an infinite series of Carnot cycles, the energy Q_{min} needed per gram to liquefy a gas with specific heat c_p, heat of condensation r and boiling point T_{liq}, is given to a good approximation by the formula:

$$Q_{min} = -c_p \int_{T_0}^{T_{liq}} \frac{T_0 - T}{T} \, dt + r \frac{T_0 - T_{liq}}{T_{liq}} . \quad . \quad . \quad . \quad (4)$$

In this expression $(T_0 - T)/T$ is the reciprocal value of the efficiency of a Carnot cycle taking place between the temperature T_0 (in our case the ambient temperature) and T. (The approximation employed consists in considering c_p as a constant; at temperatures just above T_{liq} this is not the case.) Denoting T/T_0 by τ and integrating, we find from (4):

$$Q_{min} = c_p(T_0 - T_{liq}) \left(\frac{\tau}{\tau - 1} \ln \tau - 1 \right) + r(\tau - 1) . \quad . \quad (5)$$

Using a pure refrigerator, however — which must produce all the cold at T_{liq}; this is the most unfavourable situation —, cooling plus condensation costs per gram an energy Q_{max} given by:

$$Q_{max} = \{c_p(T_0 - T_{liq}) + r\} (\tau - 1).$$

With one intermediate temperature the energy Q_1 required per gram for cooling plus liquefaction is:

$$Q_1 = c_p(T_0 - T_m) \frac{T_0 - T_m}{T_M} + c_p(T_M - T_{liq}) \frac{T_0 - T_{liq}}{T_{liq}} + r \frac{T_0 - T_{liq}}{T_{liq}} . \quad . \quad (6)$$

Calculating the value of T_M at which this expression is minimum, we find:

$$T_M = \sqrt{T_0 T_{liq}} . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

For this value of T_M we have:

$$Q_1 = c_p \left(2 T_0 \sqrt{\frac{T_0}{T_{liq}}} - 3T_0 + T_{liq} \right) + r \frac{T_0 - T_{liq}}{T_{liq}} . \quad . \quad (8)$$

Applying the above to the liquefaction of hydrogen ($T_{liq} = 20.4$ °K) we then find the numerical values for the three examples discussed.

Since the cooling and condensation take place under constant pressure, the production of cold required from both stages in a given case can easily be computed from the mass m and the relevant enthalpy change of the gas to be condensed (according to the thermodynamic definition, the change in enthalpy H is equal to the heat supplied under constant pressure). The middle stage, which is at a temperature of about 80 °K, must therefore deliver $m(H_{300} - H_{80})$, and the end stage: $m(H_{300} - H_{liq})$. For hydrogen the first amount is roughly three times as high as the second, i.e. for cooling and condensing hydrogen the net cold production of

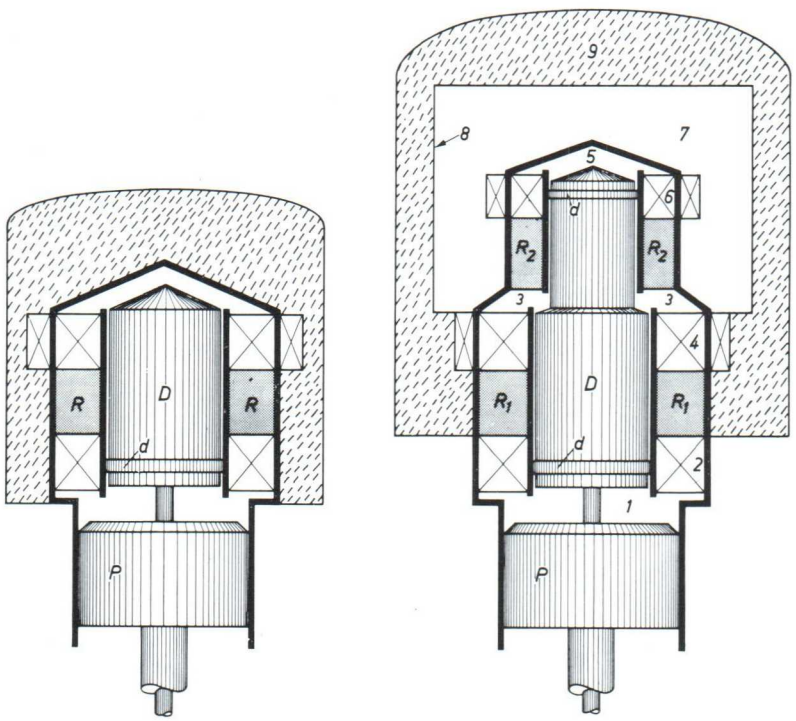


Fig. 6. Schematic cross-section of the new machine. A cross-section of the normal gas refrigerating machine is shown on the left for comparison. *P* piston. *D* displacer; in the new machine this consists of two sections with different diameters. *d* piston rings of *D*. *1* compression space. *2* appertaining cooler. R_1 first regenerator. *3* intermediate expansion space. *4* appertaining freezer; this is used only when the middle stage has also to supply cold. R_2 second regenerator. *5* expansion space. *6* appertaining freezer. *7* high-vacuum space which insulates R_2, *5* and *6*. *8* screen at temperature T_M to restrict radiation losses. *9* low-vacuum space filled with powder for thermal insulation.

the middle stage at 80 °K should be roughly three times as high as that of the end stage at 20 °K.

Design and features of the new machine

In designing the prototype of the three-space machine the working temperature of the end stage was taken to be 20 °K, the requirement being that the machine should have a high efficiency as a pure refrigerator and at the same time meet reasonable demands as a hydrogen liquefier.

We have just seen that these two applications call in principal for different machines if the maximum of efficiency is wanted. A compromise therefore had to be found. The procedure for this purpose was first of all to design, with the aid of a computer, a pure refrigerator having the highest possible efficiency at 20 °K. To this purpose we started from a design produced on the basis of estimates, and successively changed all principal parameters in the direction of higher efficiency. The characteristics of the design thus obtained were then computed. These are the curves that show the net cold production q_E and q_M of both spaces as a function of the temperature of the middle space.

From the considerations discussed in the previous section it may be deduced that the characteristics can be modified by changing the dimensions of the intermediate expansion space. Taking now the optimized case as our basis, we altered the dimensions so as to obtain a q_E characteristic which was almost horizontal. The net cold of the end stage of the new machine is thus hardly influenced at all by the value of T_M, that is to say by the amount of cold removed from the middle stage.

The construction of the machine is quite different from that possibly suggested by fig. 2. There is only one piston and one displacer, but the latter consists of two parts of different diameter. In this way both expansion spaces could be obtained using a single displacer body. *Fig. 6* shows side by side the schematic cross-section of the new machine and that of the single expansion-space machine with displacer.

The virtues of the single-space machine with the displacer construction, i.e. compactness and a low pressure drop over the displacer seal, are also present in the new machine. A further advantage of this design is that, since there is only one displacer, the driving mechanism is the same as that of the former machine.

The other details of the construction, with the two freezers and insulation spaces around the colder freezer, can clearly be seen in the drawing.

Characteristics

In order to be able to measure the characteristics, we fitted both freezers in the prototype with an electrical heating wire. This device makes it possible to measure the net cold production in a wide temperature range. In fact one really does the opposite: the choice of the current in each of the heater wires establishes q_E and q_M, and it is then a matter of determining the corresponding values of T_E and T_M. With this method of measurement a relatively high degree of accuracy can be achieved, because the freezers can in principle be perfectly insulated.

The results of the measurements are presented in *fig. 7*. It can be seen that the $q_E(T_M)$ curves are almost horizontal; the cold produced by the end stage is virtually unaffected by that of the middle stage.

From these characteristics it is also easy to find what the cold production is when the machine is used as a "pure refrigerator". All one has to do is to find from the graph, for a given value of T_E, what the value of q_E is at the T_M value at which q_M is zero — the left end of the q_E curves in fig. 7. In *fig. 8* these q_E values are plotted versus T_E. At 12 °K the value of q_E is zero, which is evidently the lowest temperature attainable. Above this the curve is fairly steep, and at 20 °K as much as 100 W of cold is already being produced. Because of this high production the time the machine takes, after being switched on, to reach a certain low temperature is extremely short.

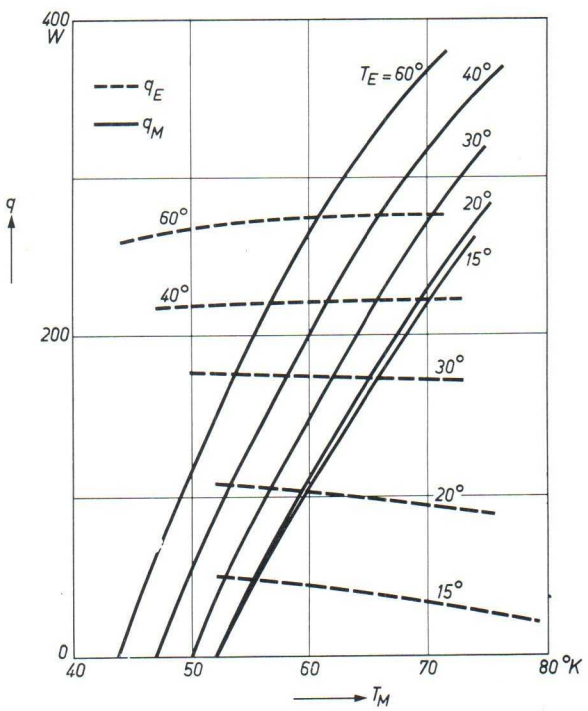


Fig. 7. Characteristics of the machine. The solid lines represent the variation of q_M at the indicated values of T_E, the dashed lines the variation of q_E. The latter curves are almost horizontal, which is an advantage for the liquefaction of gases. The curves were obtained with the working gas (helium) at an average pressure of 30 atm and with a crankshaft speed of roughly 1500 r.p.m.

When the characteristics for each adjusted value of q_E were measured the shaft power N of the machine was measured at the same time, and therefore the efficiency η could also be calculated and from this the ratio of η to the efficiency η_C of a Carnot cycle operating between the same temperatures. This relative efficiency is given by:

$$\frac{\eta}{\eta_C} = \frac{q_E}{N} \times \frac{T_0 - T_E}{T_E}.$$

The results of these calculations are also plotted in fig. 8 (dashed curve). It can be seen that at the working temperature on which the design was based (20 °K), the machine supplied the above-mentioned 100 W of cold with a relative efficiency of 17 %.

Finally, it should be noted that the curves in fig. 7 and fig. 8 were obtained with the working gas at an average pressure of 30 atm and a crankshaft speed of 1500 r.p.m. If these values are reduced somewhat the efficiency is slightly higher but the production lower.

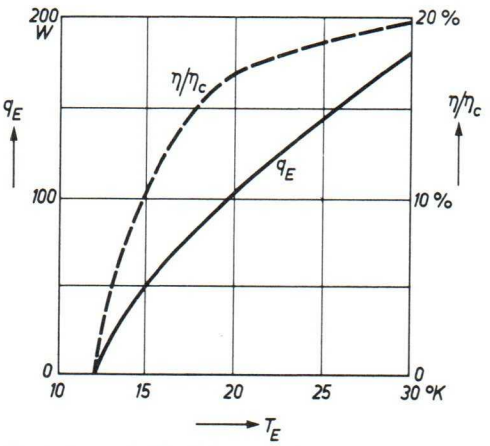


Fig. 8. Variation with T_E of the net cold production q_E and the coefficient of performance η (as a fraction of the efficiency of the Carnot cycle η_C) when cold is taken from the middle freezer (the machine being used as a pure refrigerator). The net production is zero at $T_E = 12$ °K, which is thus the lowest attainable temperature. Above this value the cold production increases rapidly with rising temperature.

Comparison with other plants

The main advantage of the new machine *(fig. 9)* compared with other refrigeration plants operating in the same temperature range is its simplicity. It requires none of the various components that make other refrigeration plants complicated, such as compressors, counterflow heat-exchangers, expansion engines, precooling baths and throttle valves, that always need control. In our machine the compression, expansion and heat exchange are combined in one unit, all processes thus being automatically adjusted to one another. Finally, for the liquefaction of hydrogen it is not necessary to compress the gas very much, or if so wished not at all, which is a great advantage.

The merits summarized above have not been acquired at the expense of the efficiency. As mentioned, the efficiency of our plant working as a pure refrigerator is 17 % of the efficiency of the Carnot cycle at 20 °K; if we had not chosen the above-mentioned compromise, but had designed the machine optimally as a refrigerator, our calculations show that the efficiency would have been about 21 %. These figures compared favourably with those of similar plants. For example, a refrigerating machine has been described [5] which was specially built for compensating the heat losses of a hydrogen bubble chamber and gives 300 W of cold at 27 °K. Calculations, which the designers themselves refer to as rather optimistic, put the relative efficiency at 20 %. The relative efficiency of our machine at this temperature is 19 %. A three-stage cooling system has also been described [6] which has a relative efficiency of 4.1 % at 14 °K. At this temperature our machine has a relative efficiency of 8 %.

As a hydrogen liquefier too, our machine compares very favourably. The previously built liquefiers have a power consumption 8 to 10 times the theoretical minimum [7]. This amounts to a value from 1.85 to 2.3 kWh/l. The consumption of our machine is 1.9 KWh/l.

Applications

We have shown in the foregoing that the new gas refrigerating machine constitutes an exceptionally simple refrigerator that can operate at temperatures down to about 12 °K, and has a particularly high efficiency in the region from 14 to 60 °K. In principle a machine of this kind can find application where a temperature is to be maintained in this region or where it is required to cool something to that temperature.

With regard to the use of the new machine as a pure refrigerator, an application that springs to mind is the recondensation of hydrogen (or neon) which has evaporated from a cooling bath, a storage tank, a bubble chamber, etc. In the case of hydrogen this can be of particular importance where during the liquefaction the ratio of the concentrations of orthohydrogen and parahydrogen have not been brought into line with the new temperature (at 300 °K 25 % para-, at 20 °K 99 %). The conversion then continues for hours in the liquid produced, in which process heat is liberated [8]. The machine can be used as a recondenser at any temperature between 14 and 30 °K. The cold production at a given temperature can be read from fig. 8.

In low temperature research it is not absolutely necessary to work with a cryostat kept at the temperature required by the evaporation of a liquid. It is also possible to use a cryostat mounted on the head of the ma-

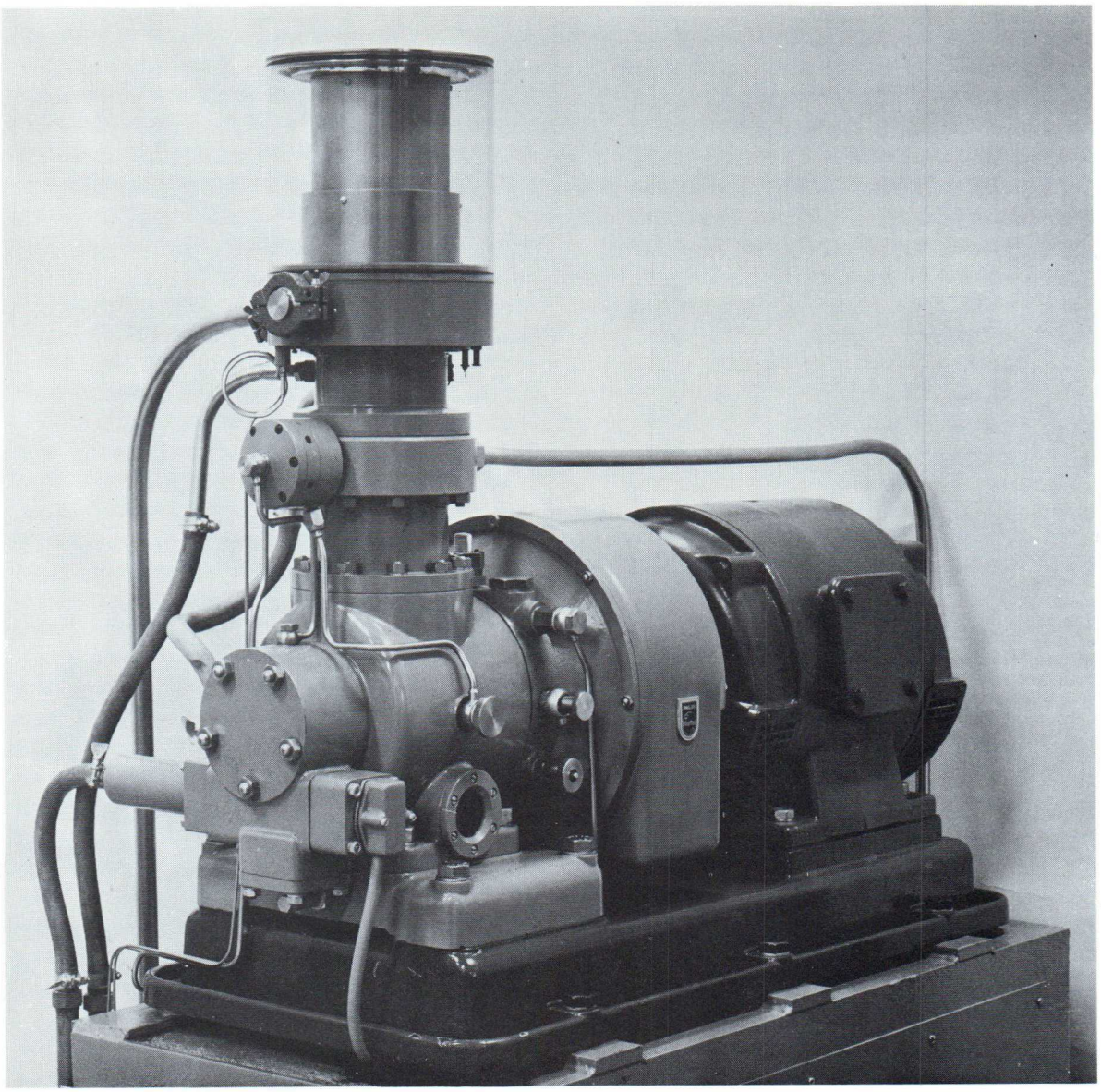


Fig. 9. The new machine. The insulating wall (*9* in fig. 6) of the evacuated space which surrounds the colder expansion space has been replaced by a wall of Perspex so that the stepped portion of the cylinder can be seen.

chine itself and kept cold by conduction. The advantages of this are that the cryostat can be used throughout the entire range from 12 to 300 °K and that the cooling is very much faster. A drawback is that the cryostat is rigidly mounted on the machine and is thus subjected to mechanical vibrations. There will be many measurements, however, where this will not be an objection. A cryostat of this type is shown schematically in *fig. 10*, and a photograph is to be seen in *fig. 11*. Other important advantages are that the installation is very simple, that the cold space is easily accessible, because there is no need to be so sparing with the cold, and that the temperature can quickly be varied, e.g. from 80 to 12 °K in 11 minutes.

A relatively recent low-temperature application is what is termed "cryopumping", i.e. the evacuation of an enclosure by freezing gases. For this purpose, too,

[5] D. B. Chelton, J. W. Dean and B. W. Birmingham, NBS Technical Note 38.
[6] H. O. McMahon, Cryogenics **1**, 65, 1960.
[7] See e.g. Cryogenic Technology, edited by R. W. Vance, Wiley, New York 1963, Chapter II.
[8] This effect is described by R. B. Scott in Cryogenic Engineering, Van Nostrand, Princeton 1959, and also by S. Flügge, Handbuch der Physik XIV, Springer, Berlin 1956.

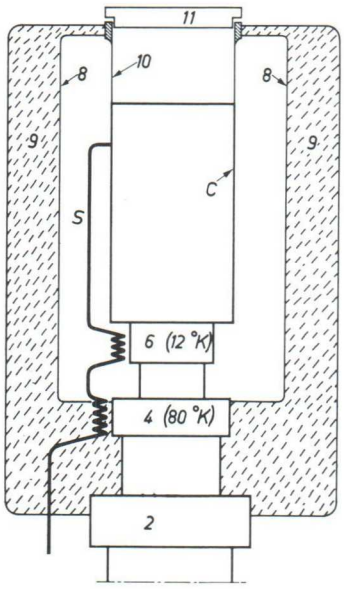


Fig. 10. Schematic cross-section of a cryostat for measuring purposes, mounted directly on the machine. The figures have the same meaning as in fig. 6. The measuring space *C* is formed by a copper vessel of 11 cm diameter and 30 cm length. *S* gas supply line (hydrogen or neon). *10* cylinder of material with low thermal conductivity. *11* cap.

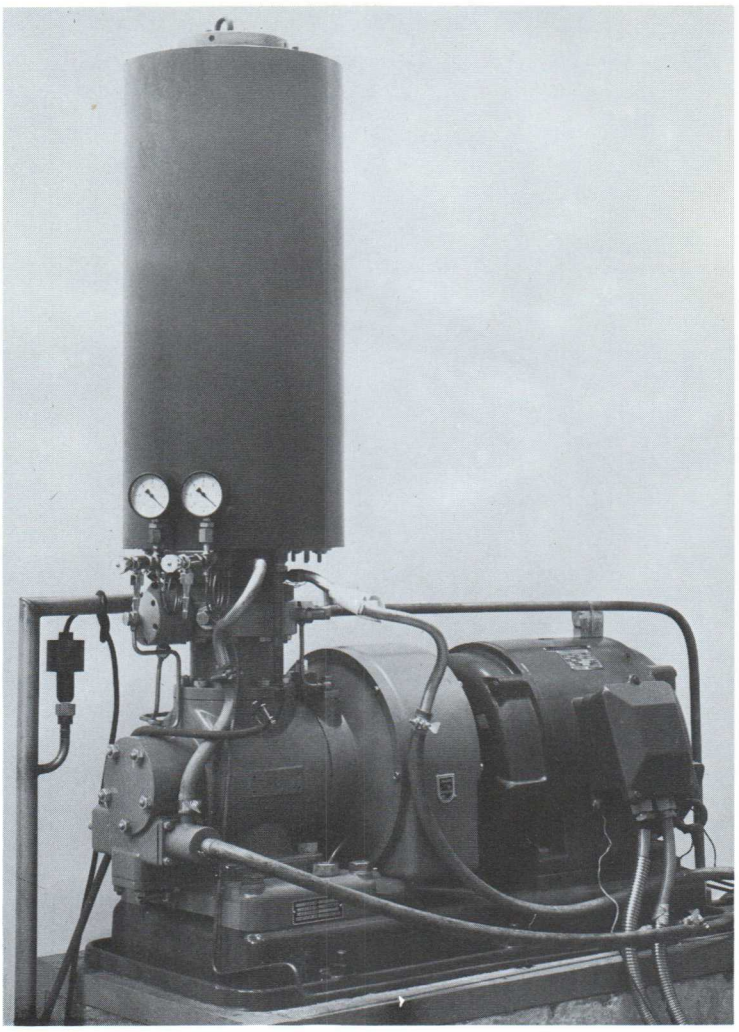


Fig. 11. The cryostat of fig. 10 mounted on the three-space machine.

the new machine is eminently suitable. Since nitrogen and many other gases have a vapour pressure at 20 °K of about 10^{-9} mm Hg or less, the machine can be used to produce a very high vacuum. Cryopumping is superior where the evacuation should be fast, or where large quantities of desorbed gas are to be removed while maintaining a pressure of 10^{-3} to 10^{-6} mm Hg; the pumping speed required for the latter purpose cannot normally be reached using conventional pumping methods. A great advantage of cryopumping is that the "pump" does not introduce any contamination (such as oil or mercury) into the space to be evacuated.

Mention has already been made of the machine's application as a hydrogen liquefier, and of the considerable importance in this connection of the middle stage. Since the machine can supply cold at 20.4 °K, the hydrogen can be condensed at a pressure of 1 atm; thus that compression is not needed, as has already been pointed out. The machine can supply a quantity of about 5 litres of liquid hydrogen per hour. It can also, of course, condense neon gas.

Another application is as a precooler for a helium liquefier. The two stages of the machine can then be used instead of the nitrogen and hydrogen baths of the conventional helium liquefiers (*fig. 12*).

An aggregate of this kind can also be used as a refrigerating plant, as distinct from a liquefier. It is difficult to indicate how much cold a plant can supply at 4.2 °K, because this depends primarily on the helium compressor and on the quality of the heat exchangers. In theory, however, at 4.2 °K almost as much cold can be produced as is supplied by the head of the machine, the reason being that in principle a Joule-Thomson system is simply a system for transport of cold. If the head is held at 20 °K, then, the production is 100 W. Since, however, the maximum integral Joule-Thomson

effect at 20 °K is 11.4 J/g, it is necessary for this purpose to pump the helium around the circuit at a rate of about 9 grams per second. To do this a fairly large compressor is required.

A simple experimental arrangement on this principle has been made using a small compressor and very simple heat exchangers. This apparatus produces about 4 W of cold at 4.2 °K. The relative efficiency is 3.75 %, which is a most satisfactory value for such a simple set-up [6]. This simple helium system was designed in the form of a cryostat and is used for keeping masers, cryotron memory devices, superconducting coils, etc. at the required low temperatures.

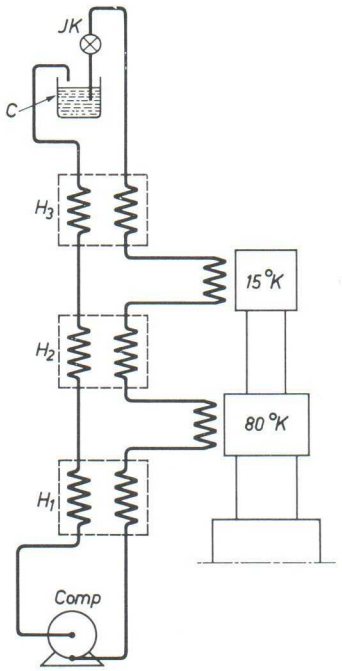


Fig. 12. Diagram of a helium liquefier — or refrigerating plant for 4.2 °K — consisting of a three-space gas refrigerating machine (right), several heat exchangers *H*, a Joule-Thomson system (throttle valve) *JK* with helium vessel *C* and a compressor *Comp*. In the compressor helium gas is compressed to e.g. 20 atm and cooled when passing through the two freezers — by the first to about 80 °K, by the second to about 15 °K. In the throttle valve *JK* the gas expands to 1 atm, in which process it is further cooled so that a part condenses. This part arrives in the vessel *C*. The gaseous part flows back to the compressor where, in the heat exchangers H_1 to H_3, it largely transfers its cold to the gas which is on its way in the other part of the circuit o the throttle valve.

Summarizing, it can be said that a new type of refrigerating machine has been developed which extends to 12 °K the temperature region covered by Philips refrigerating machines operating on the Stirling-cycle principle. Round about 20 °K the machine has a high refrigeration output and a relatively high efficiency; it also takes up very little room. The main advantage, however, is the remarkable simplicity of the system, which makes maintenance almost superfluous and reduces attendance to switching on and off.

The new machine opens up a temperature region which has hitherto only really been accessible to workers in cryogenic laboratories.

Summary. The efficiency at low temperature of a refrigerating machine based on the Stirling cycle can be improved and the minimum temperature lowered by compensating the regeneration losses by means of a series of additional expansion spaces whose temperature lies between room temperature and that of the last expansion space. The working gas, after compression, flows through a regenerator into the first auxiliary expansion space, but part of it passes through a second regenerator to the next expansion space, and so on. A machine is described which has one extra expansion space. The most favourable operating temperature of this "three-space machine", as it is called, depends on whether the extra space is required to supply cold, e.g. for the liquefaction of gases, or not. The machine is designed to produce cold (100 W) at 20 °K with good efficiency (17 % of the efficiency of the Carnot cycle), while at the same time meeting reasonable requirements as a hydrogen liquefier (production 5 l/h). The machine has a cylinder with a narrow and a wide section, in which a displacer moves similarly consisting of two sections having different diameters; the intermediate expansion space is the part of the wide section of the cylinder in which the narrow section of the displacer moves. The lowest temperature that can be reached is about 12 °K.

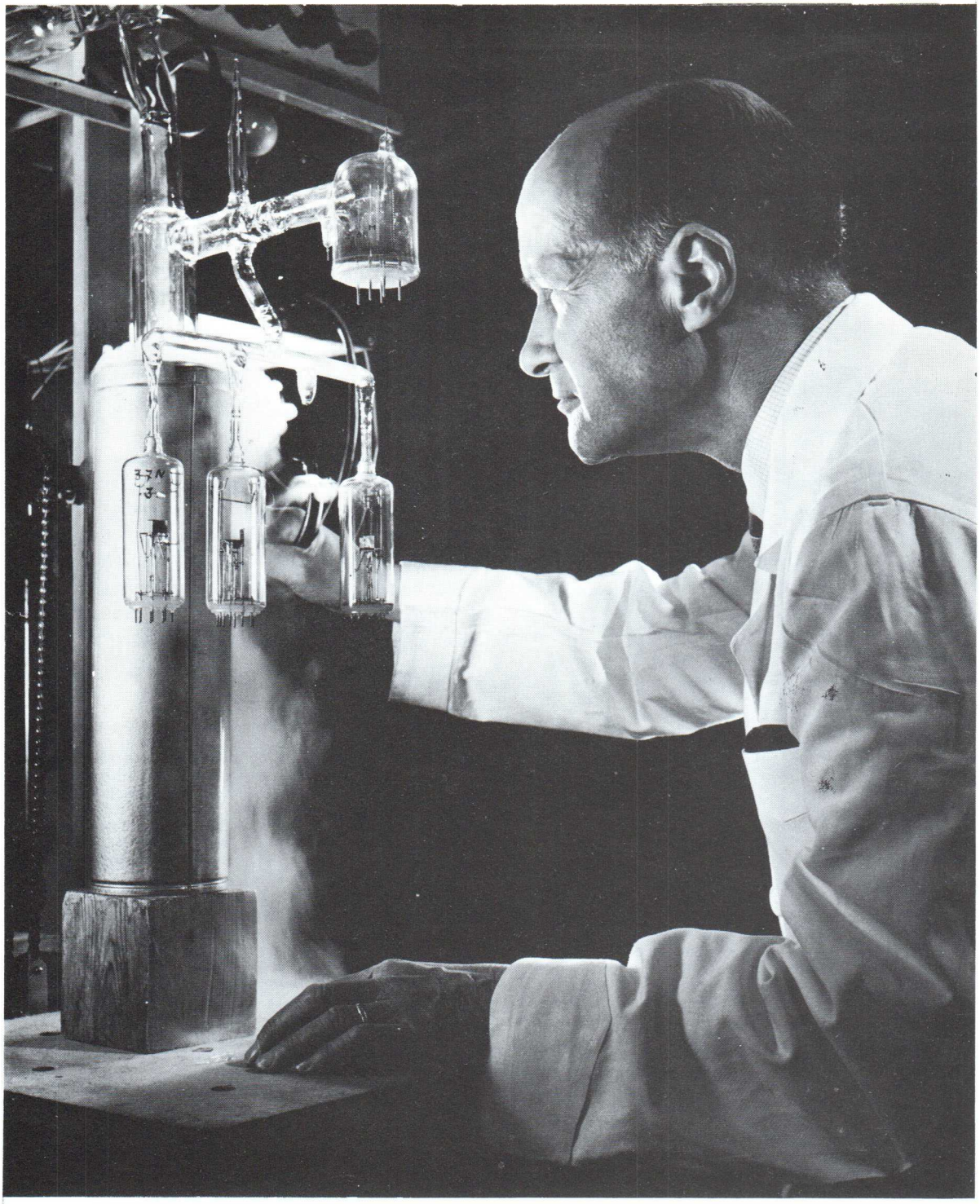


Photo Maurice Broomfield

Thermionic-cathode testing

The emission properties of thermionic cathodes are usually investigated in simple test-valves, containing only an anode besides the cathode. The photograph shows three such valves being evacuated on a mercury-diffusion pump. Behind them is the cold trap, which is kept at about −190 °C by liquid nitrogen and condenses the mercury vapour from the diffusion pump. Above the test valves can be seen an ionization gauge. In situations like this it is necessary to allow for the fact that the pressure in the ion gauge is not the same as in the valves.

These preliminary tests under greatly simplified conditions are followed at a later stage by tests with the cathodes mounted in the valves for which they are intended.

(From: Philips tech. Rev. **24**, 228, 1962/63.)

Osmium dispenser cathodes

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

621.3.032.213.2:621.385.6

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

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

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

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

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

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

The development of microwave tubes, such as disc-seal triodes, klystrons and magnetrons for higher and higher frequencies and output powers has benefited from these characteristics of dispenser cathodes, and

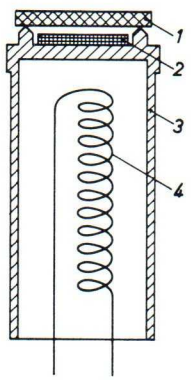


Fig. 1. Basic construction of an L cathode. *1* porous tungsten body. *2* chamber containing a reserve supply of barium-calcium-aluminate. *3* molybdenum cylinder. *4* filament.

has in its turn stimulated further investigations into these cathodes. Dispenser cathodes are also being used with success in many other types of thermionic valve. In all cases, however, the valves have been for professional applications.

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

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

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

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

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

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

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

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

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

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

Possible ways of reducing the work function

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

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

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

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

In order to lower further the work function of our dispenser cathodes, one might choose a different adsorbate or a different substrate. It has been found that in practice a different adsorbate does not lead to better results. Although Cs(Cs_2O) lowers the work function more than Ba(BaO), it is not as a rule suitable for use in thermionic valves because of its high volatility: this would so reduce the maximum permissible temperature of the cathode that the emission density would drop to a lower value than found with a Ba(BaO) layer.

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

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

Work function theory based on a simple model

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

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

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

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

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

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

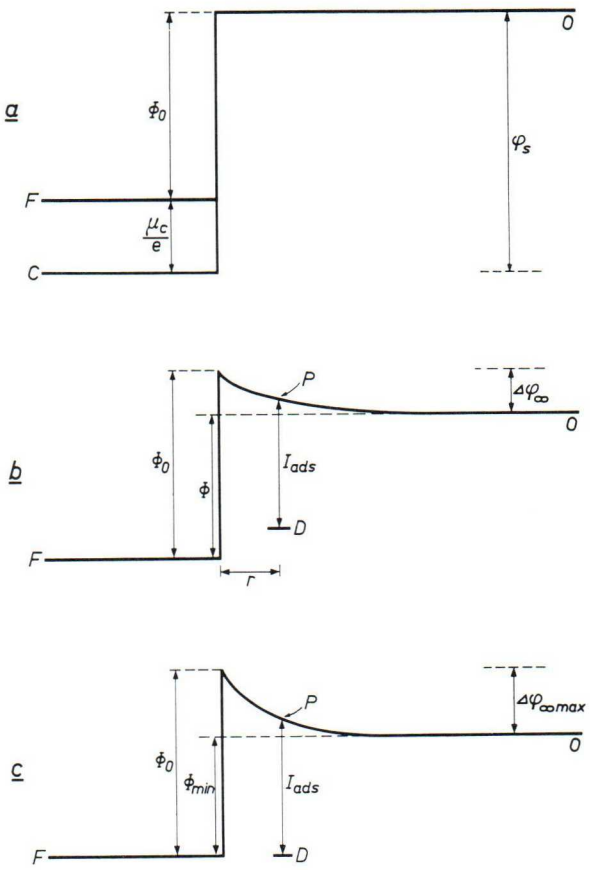


Fig. 2. Potential curve at the surface of an emissive body. As is customary, the graph shows the positive direction of the potential downwards. F Fermi level. O potential level in vacuum. Φ_0 work function.

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

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

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

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

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

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

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

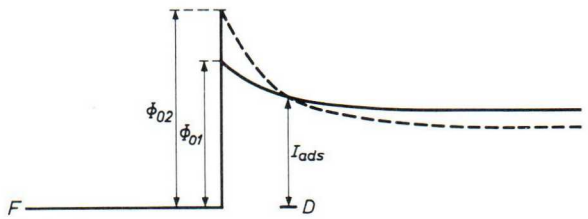


Fig. 3. Starting from a *higher* Φ_0 ($\Phi_{02} > \Phi_{01}$), a *lower* Φ_{min} (dashed curve) may be expected when the surface is optimally covered.

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

The calculation is as follows.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Measurements on various substrates using a barium adsorbate

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

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

Rhenium, ruthenium, iridium and osmium are metals that meet these requirements. The work function of polycrystalline rhenium is 5.1 V. The work functions of

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

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

Osmium-coated cathodes

Current density

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

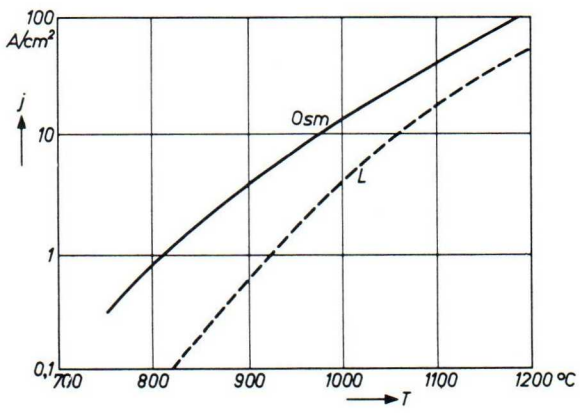


Fig. 4. Current density j as a function of cathode temperature T, for a normal L cathode (dashed curve) and for an L cathode with an osmium coating 0.5 μm thick (full curve).

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

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

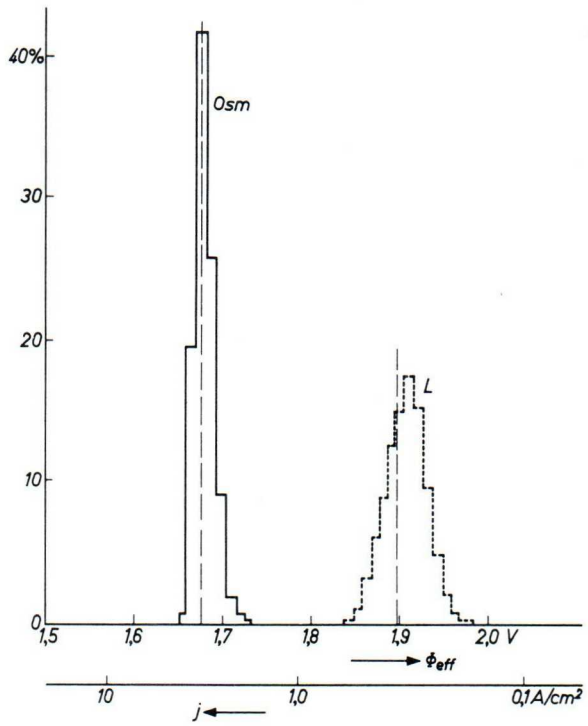


Fig. 5. Emission from the same two cathodes as in fig. 4, at 800 °C. The abscissa represents the effective work function Φ_{eff} (and also the corresponding current density), and the ordinate gives, for each value of Φ_{eff}, the fraction of the total cathode surface at which that value is found. (The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

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

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

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

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

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

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

Life

The life of an L cathode is proportional to the quantity of barium-calcium-aluminate in the storage chamber. For a given porosity of the tungsten body, the quantity of Ba (and BaO) lost by evaporation per unit time and per unit cathode surface depends only on the cathode temperature (*fig. 6*). The life of a normal L cathode operating at 1110 °C is in the region of 15 000

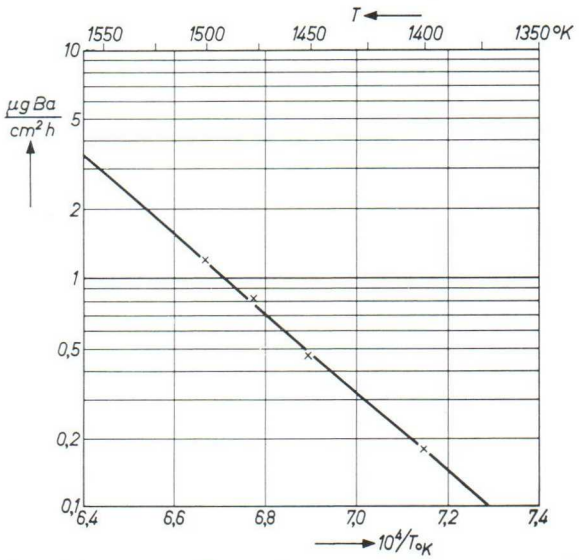


Fig. 6. Evaporation of Ba as a function of cathode temperature T at a given porosity of the tungsten body (after unpublished measurements performed by C. A. M. van den Broek and A. Venema).

hours. An osmium-coated cathode with the same reserve supply of barium-calcium-aluminate may be expected to have the same life if operated at 1110 °C; if, however, the Os cathode is operated at a temperature of 1010 °C, which is sufficient for the same emission, the cathode should have a 6 to 10 times longer useful life.

Two series of cathode life-test experiments have been started to check this prediction, using Os cathodes of 8 mm² surface area, with 0.5 μm osmium, mounted in diodes with molybdenum anodes and operated at 100 mA, i.e. 1.3 A/cm², at a cathode temperature of 1110 °C in one series and at 1010 °C in the other. These experiments have now been in progress for 14 000 hours — there are unfortunately only 9000 hours in the year. During this time the saturation emission has remained virtually constant. This means that the advantage of the lower operating temperature has been brought into effect: whether the extremely long life predicted will be achieved in these experimental valves is not yet known. In fact, it is conceivable that the osmium coating will eventually disappear as a result of evaporation, diffusion or migration. Nothing of this has yet been noticed at temperatures lower than 1050 °C, even with the thinnest coatings of 0.1 μm. At temperatures higher than 1150 °C, however, there are some indications of coating loss. A separate series of tests has been started to investigate this effect as a function of cathode temperature and coating thickness, but these tests will take a considerable time to complete because of the long lifetimes involved.

Osmium-coated impregnated cathodes are more subject to coating loss. At a cathode temperature of 1110 °C the "osmium properties" have disappeared after only about 5500 hours of operation, and the cathode then behaves in the same way as an uncoated impregnated cathode. At a cathode temperature of 1060 °C this happens after about 7500 hours, but at a temperature of 1010 °C this type of cathode also shows no signs of osmium loss, even after 14 000 hours.

The exhaustion of the reserve of barium-calcium-aluminate, which we have mentioned as the primary cause of limitation of life, is of course not the only possible cause. The filament may break, the porous tungsten body may become deformed, the valve characteristics may be affected by the evaporated Ba and BaO (insulation faults, grid emission). At the lower operating temperature of the osmium cathodes, all these factors as well will only make their effect felt after a much longer running time than with normal L cathodes.

Secondary electron emission

In some applications, for example in magnetrons, the contribution of the secondary electron emission to the total current density is very important. Although no particular attention was paid to this point in the development described here, it is nevertheless a welcome incidental feature that the secondary emission coefficient of the osmium-coated cathodes is appreciably higher than that of normal dispenser cathodes. It is worth noting that in this respect as well, osmium is by far the most suitable of the metals investigated.

In *fig. 7* the results of measurements of the secondary emission coefficient are presented as a function of the energy of the primary electrons, for an osmium-coated L cathode, for a normal L cathode and for a normal impregnated cathode. Although the differences in the secondary emission coefficient are evidently not very large, they are nevertheless important. In a 4 mm magnetron type DX 164, for example, the maximum current density is 190 A/cm² with an uncoated impregnated cathode, but 320 A/cm² with an Os cathode.

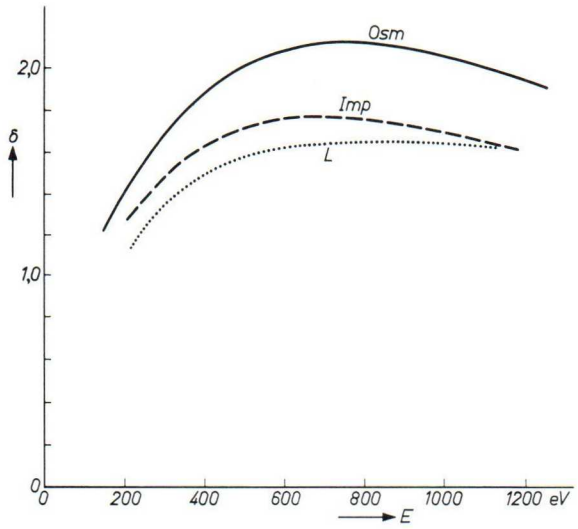


Fig. 7. Secondary emission coefficient δ as a function of the energy E of the primary electrons (in eV) for a normal L cathode (*L*), an impregnated cathode (*Imp*) and an osmium-coated L cathode (*Osm*).

Other practical data

As shown in the foregoing, the reduction in the temperature required for operation increases the possibility of using dispenser cathodes in non-professional valves, where the expense of initial purchase and replacements is such an important consideration. The development of such applications is still in the initial stage. As indicated, however, the new cathodes with their higher performance may also be useful in various professional valves, in particular in microwave valves. We have already mentioned magnetrons. The use of Os cathodes is also being investigated in disc-seal triodes [14] and in various reflex klystrons.

In the *disc-seal triodes* osmium-coated cathodes are under test both at 1150 °C (the usual temperature for uncoated cathodes in these tubes) and at 1020 °C. In both cases the cathode loading is 0.75 A/cm² and up to now the useful life is longer than 10 000 hours. Because of the low barium evaporation and the imperceptibly small deformation of the tungsten body, the valve characteristics, particularly for cathodes operated at 1020 °C, are remarkably constant compared with those of valves with uncoated cathodes. From these data the expected life of Os cathodes at an operating temperature of 1020 °C (970 °C black-body temperature) may be estimated at several times 10 000 hours — more than sufficient for valves to be used in interplanetary missions.

The use of Os cathodes makes it possible to make *reflex klystrons* for even shorter wavelengths than have up to now been possible. Normal uncoated L cathodes have a reasonably long life in continuous operation at loadings up to 8 A/cm². This is just sufficient, as B.B. van Iperen has shown, for a 2 mm reflex klystron to be feasible [15], albeit with marginal characteristics. Using an Os cathode, G. H. Plantinga of this laboratory has succeeded in producing a practical 2 mm reflex klystron (type DX 247) in which the current density is 20 A/cm². With the same type of cathode he was also able to develop a 1.5 mm reflex klystron. In this case continuous loading at about 25 A/cm² is required.

[14] On disc-seal triodes see J. P. M. Gieles, Philips tech. Rev. **19**, 145, 1957/58.

[15] B. B. van Iperen, Philips tech. Rev. **21**, 221, 1959/60.

Summary. To widen the useful scope of dispenser cathodes, the work function of the emissive tungsten surface should be reduced to a value lower than that already achieved with the coating of adsorbed barium. This has been made possible by the paradoxical effect that the work function of a metal surface optimally covered with barium is lower for a *higher* work function of the *uncoated* metal. This effect is explained with the aid of a simple model, and by means of an approximate calculation equations are derived for estimating the work function to be expected. The predictions are found to agree very reasonably with experiments. With a view to the practical application of the effect, dispenser cathodes were made with the porous tungsten body variously coated with a thin layer (0.1 to 1 μm) of rhenium, ruthenium, iridium or osmium. The best results, i.e. a lowering of the work function from 1.95 V to 1.60 V, were obtained with osmium. This means that an osmium-coated L cathode operated at 800 °C has a current density 10 times higher than a normal L cathode. This improvement can be utilized either for operating the cathode at a lower temperature, thus making more economical designs and longer lifetimes possible (e.g. in disc-seal triodes several times 10 000 hours) or for achieving even higher current densities (e.g. 25 A/cm² in a 1.5 mm reflex klystron).

C. MISCELLANEOUS CONTRIBUTIONS

		page
1	**Some modern methods of arc-welding steel** P.C. van der Willigen (Vol.24, p.14)	329
2	**The evolution of the permanent magnet: a brief review** H.J.Meerkamp van Embden (Vol.18, p.358)	344
3	**A radar sonde system for upper air measurements** N.E.Goddard and H.A.Dell (Vol.19, p.258)	348
4	**A klystron multiplier for generating 0.8 mm waves** B.B. van Iperen and W.Kuypers (Vol.26, p.61)	355
5	**A dynamo for generating a persistent current in a superconducting circuit** J.Volger (Vol.25, p.16)	362
6	**The Foucault pendulum in the United Nations Building in New York** J.A.Haringx and H. van Suchtelen (Vol.19, p.236)	366
7	**Lightning and aircraft** A.C. van Dorsten (Vol.16, p.296)	372
8	**New forms of bearing: the gas and the spiral groove bearing** E.A.Muijderman (Vol.25, p.253)	375
9	**Direct observation of Weiss domains by means of the Faraday effect** C.Kooy (Vol.19, p.286)	398
10	**A method of growing dislocation-free germanium crystals** B.Okkerse (Vol.21, p.340)	402

11 **A method of measuring the resistivity and Hall coefficient on lamellae of arbitrary shape** 410
L.J. van der Pauw
(Vol.20, p.220)

12 **A small, stable gas laser** 415
J.Haisma, S.J. van Hoppe, H. de Lang and J. van der Wal
(Vol.24, p.95)

13 **A simple method of determining the thermal conductivity of solids** 419
J.Schröder
(Vol.21, p.357)

14 **A positive rod or piston seal for large pressure differences** 423
J.A.Rietdijk, H.C.J. van Beukering, H.H.M. van der Aa and R.J. Meijer
(Vol.26, p.287)

15 **Luminescent *P-N* junctions in gallium phosphide** 434
W.Glässer, H.G.Grimmeiss and H.Scholz
(Vol.25, p.20)

16 **In search of a measure of perceptual work** 436
J.M.Westhoff
(Vol.25, p.56)

17 **Fluorescent pigments as an artistic medium** 446
J.L.H.Jonker and S.Gradstein
(Vol.11, p.16)

18 **Generation of musical intervals by a digital method** 454
D.Gossel
(Vol.26, p.170)

19 **Nitrogen in silicon iron** 461
J.D.Fast
(Vol.26, p.81)

20 **Crystal growth of silicon carbide** 465
H.B.Haanstra and W.F.Knippenberg
(Vol.26, p.187)

SOME MODERN METHODS OF ARC-WELDING STEEL

by P. C. van der WILLIGEN *).

621.791.8

Electric arc welding is nowadays the most widely used method of obtaining a mechanically strong joint between metal parts. The term arc welding covers a very wide variety of techniques. In this article the author confines himself to the arc welding of steel, including the various mild (unalloyed) steels as well as low-alloy types. Three Philips developments are discussed: Contact electrodes, zircon-basic electrodes and CO_2-shielded welding.

The extensive use made of the arc welding of steel can be judged from the annual consumption of electrodes. In the Netherlands some 275 million electrodes were consumed in 1955 [1]). The Netherlands consumption of steel in that year was 2.5 million metric tons, so that per ton of steel 110 electrodes were used. The average weight of an electrode at that time being 40 grams, this means an electrode consumption of 4.4 kg per ton of steel.

The corresponding figure in most other countries is somewhat lower, a relatively large proportion of Dutch steel consumption being attributable to ship building, where arc welding is booming. In the United States, on the other hand, a big steel consumer is the automobile industry, where more use is made of resistance welding; nevertheless, U.S. consumption of welding electrodes in 1955 was no less than 2.9 kg per metric ton of steel.

In view of the enormous scale on which the arc welding of steel is finding application, it is not surprising that numerous investigations have been undertaken in this field and a wide variety of welding methods developed. We shall discuss here three developments that have taken place in the Philips Research Laboratories at Eindhoven: Contact electrodes, zircon-basic electrodes (both types essentially overlap each other, as we shall see) and CO_2-shielded welding. A machine for automatic CO_2 welding can be seen in *fig. 1*; particulars are given in the relevant section of this article.

To place these developments in the right setting, we shall first give a survey of welding methods classified in accordance with one particular point of view, without however attempting to be complete and without being concerned too much with chronological order. We shall return to our classification later when comparing the various methods in regard to such points as economy, quality of the weld, and the possibility of automatic operation.

*) Philips Research Laboratories, Eindhoven.

[1]) W. Gerritsen, Smit Mededelingen **14**, 125, 1959 (in Dutch).

Survey of arc-welding methods

In the first years after the invention of arc welding (carbon electrodes in 1886, consumable metal electrodes in 1892) it was not yet realized that to obtain a mechanically sound weld it is essential to exclude the air from the vicinity of the arc; if the droplets of the weld metal, which are heated in the arc to a temperature several hundred degrees above the melting point, are exposed to the atmosphere they absorb some tenths of a percent of nitrogen and oxygen, resulting in a brittle and porous weld.

All the welding methods to be mentioned here are therefore designed to shield the weld metal from the atmosphere. The methods are shown in *fig. 2*, classified according to the means of shielding adopted.

An effective way of protecting the molten metal against the atmosphere is to weld in vacuum. This is in fact done in some cases, the heat for welding being generated by an electron beam. This method cannot of course be counted among the forms of arc welding, and it is obviously impracticable for the great majority of arc-welding applications.

A generally useful solution to the problem of shielding the weld metal was found in about 1910 with the development of the *coated* (or covered) *electrode*. During the melting process the coating produces *slag*, which envelopes the droplets with a molten shell, and it also evolves *gas*, which excludes the air from the vicinity of the arc. Initially, up to about 1937, the only known coated electrodes produced an acid slag and a gas mixture consisting mainly of carbon monoxide and hydrogen [2]). These electrodes are eminently satisfactory for welding mild (unalloyed) steels, but in the case of low-alloy steels, which are nowadays gaining increasing

[2]) Particulars of the numerous kinds of acid electrodes will be found, for example, in the Welding Handbook, published by the American Welding Society, New York 1960, 4th Edn. This work also contains detailed information on many arc-welding questions that can only be touched upon here in passing.

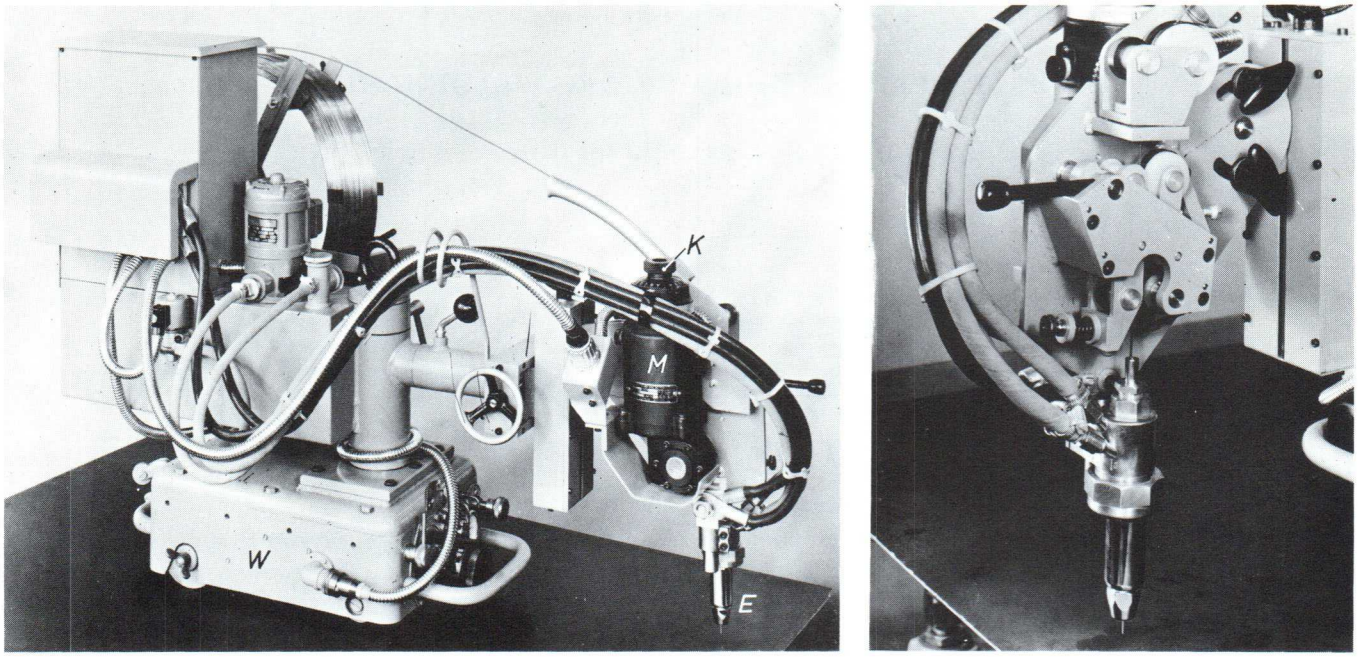


Fig. 1. Machine for fully automatic CO_2 welding. The carriage *W* travels at an accurately adjustable speed over a flat plate or on rails, so that the welding head with nozzle *E* moves over the joint to be welded. The welding head (shown separately on the right) can be adjusted to any desired position. With this machine welding currents up to 700 A and rates of travel up to 200 cm/min are possible.

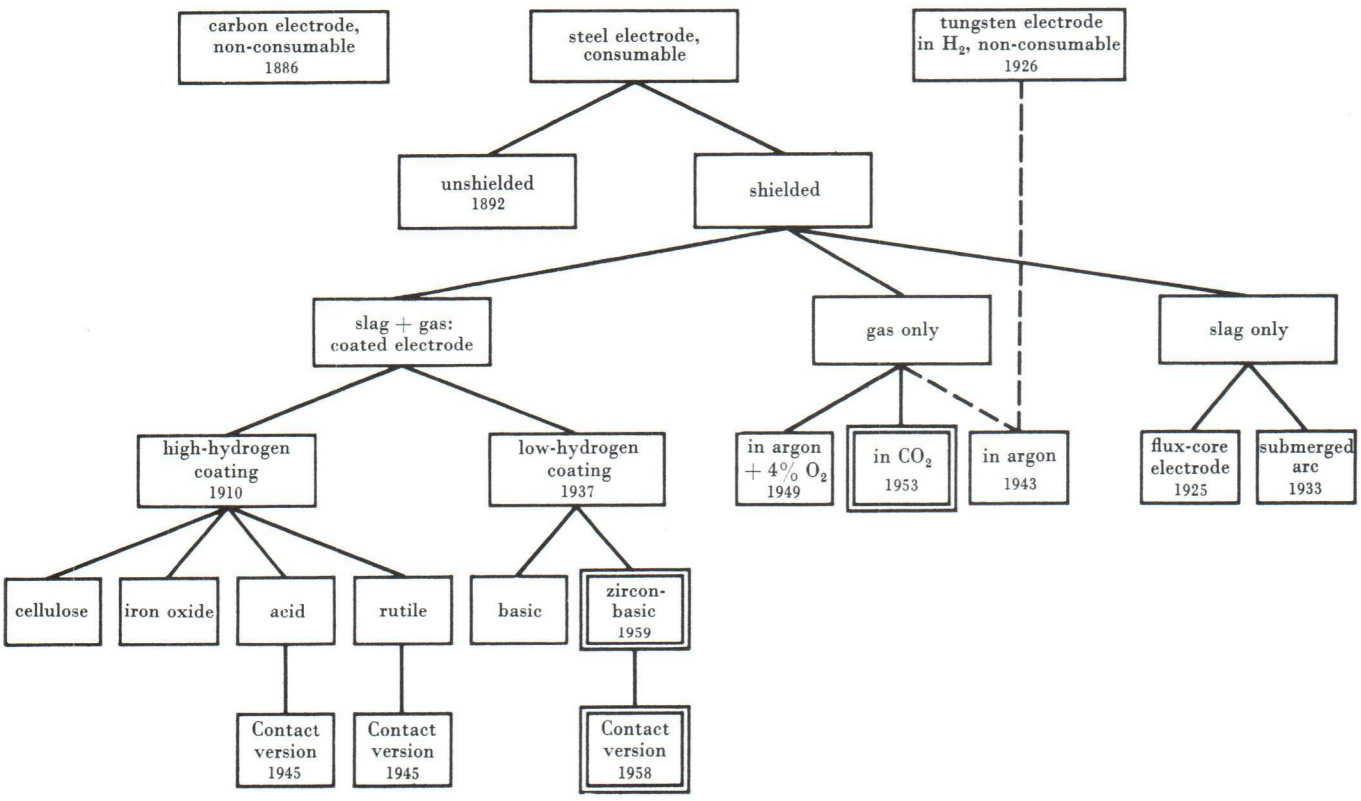


Fig. 2. Diagram of the development of various methods of arc-welding steel, classified according to the manner in which the molten metal is shielded from the atmosphere. The methods framed by double lines are discussed further in the article. The coated electrodes provide shielding by slag and gas combined. Numerous methods, not shown in this diagram, have been proposed in recent years for combined gas and slag shielding.

importance, they tend in unfavourable circumstances to cause cracks in the welded joints and other difficulties. This led to the development of *basic* (low-hydrogen) *electrodes*, the coating of which yields a basic slag and a gas which is largely CO and contains very little hydrogen. An independent investigation in about 1943 at Eindhoven led to the development of the *Contact* (iron-powder) *electrodes*, which were marketed in 1947 as a special version of the acid-coated electrodes [3]). From these, with a view to applying the same principle to the basic electrodes, the *zircon-basic coating* was developed [4]).

Meanwhile, however, other means were found of shielding the weld metal from the air: whereas the coated electrode does this by producing slag *and* gas, it proved possible to protect the weld either by slag or gas alone (fig. 2).

Before going into this point, we shall examine the question whether gas shielding alone is fundamentally sound. For if the deposited metal is shielded without slag, we sacrifice several additional functions which slag fulfils or is supposed to fulfil: 1) it provides for metallurgical purification of the weld metal; 2) since some of its constituents readily emit electrons, it has a stabilizing effect on the arc, thereby facilitating welding with alternating current; 3) it protects the weld, as the arc moves away, against too rapid cooling by the air; 4) it continues to protect the cooling metal against the action of atmospheric oxygen and nitrogen.

The first function mentioned may indeed be important on occasion, but it can be dispensed with if the deposited metal has the same (or higher) purity as the parent metal, i.e. the steel to be welded. The second function can be fulfilled by arc-stabilizing substances, added in small quantities to the steel itself. The other functions, at least as far as steel is concerned, have no real significance: cooling by the air is immaterial, in view of the very high heat dissipation through the workpiece itself; and experience shows that the air does no harm to the cooling weld in the case of steel [5]).

As long ago as 1925 an electrode was brought out, designed to provide shielding by the slag alone: this was the "flux-core electrode", in which substances that form slag and also contribute to the stabilization of the arc are contained in an axial cavity of the electrode wire. This did not, however, provide adequate protection of the weld metal, and these electrodes have therefore virtually dropped out of use.

Good results, on the other hand, were obtained with another method of shielding by slag alone; this was called "submerged-arc welding", and came out in about 1933 in the United States [6]). The arc in this case burns in a cavity inside a coarse-grained layer of powder (flux). The arc is therefore not visible, but this is not a serious objection as the method was in fact primarily developed for automatic welding, and has so far remained the most widely employed for this purpose. The flux is made by melting a mixture of slag-forming substances and, after solidification, grinding it into grains of a particular size. The mixture may be sintered instead of melted; in both operations, the temperature is so high as to remove all substances that give off gas, including possible sources of hydrogen. As we shall see, the absence of hydrogen is an important characteristic.

In the welding methods with gas shielding only, the absence of hydrogen is also essential. One would not think so, however, judging from the first method of this kind, *atomic-hydrogen welding*, dating from 1926, which uses an arc between two tungsten electrodes in a hydrogen atmosphere. The hydrogen atoms formed by dissociation in the white-hot arc recombine outside the arc and thus, together with the radiation from the arc, melt the workpiece and the added weld metal. The method is still used here and there for metals and alloys other than steel. Where steel is concerned, however, the appropriate method in this category is argon-arc welding, developed about 20 years ago in the United States. The weld metal is shielded here by an inert gas — which in America may be helium as well as argon, but in Europe is always argon — which is directed around the arc from a gas cylinder. In its original form, in which the arc is struck between a non-consumable tungsten electrode and the parts to be welded, the method is still employed on a wide scale for special alloy steels and other metals. Since 1949 a method derived from inert-gas welding has gained popularity for welding normal steel, the tungsten electrode being replaced by a consumable bare steel wire, which is fed from a motor-driven reel. In the classification given in fig. 2, this method is denoted as "argon + 4% O_2", the addition of 3 to 5% oxygen to the inert gas having proved desirable to avoid porous welds.

The latter method was extended by the process of CO_2 welding, which, in 1953, was conceived and

[3]) P. C. van der Willigen, Welding J. **25**, 313S, 1946 and Philips tech. Rev. **8**, 161 and 304, 1946.

[4]) P. C. van der Willigen, Welding News No. 89 (March 1958), No. 123 (July 1961) and No. 124 (Sept. 1961).

[5]) This is not the case in the welding of metals that have a great affinity for oxygen and nitrogen, e.g. aluminium or titanium. When such metals are welded by a gas-shielded method, extra gas shielding may be necessary during cooling.

[6]) See the book mentioned in reference [2]); further L. Wolff in "Werkstoff und Schweissung" (Ed. F. Erdmann-Jesnitzer), Akademie-Verlag, Berlin 1951, p. 458.

developed in the Philips Research Laboratories [7]): the argon, which is rather costly, is replaced here by carbon dioxide, which is cheap. Although this is not an inert gas, it provides complete protection for the weld metal when properly used. The method also has many other attractive features, and has therefore gained considerable ground in the course of a few years [8]).

The classification in fig. 2 is, of course, far from complete. In particular, combined shielding by slag and gas is found not only in the above-mentioned coated electrode, but also in numerous welding methods subsequently developed, which were in fact based on shielding by gas or by slag alone. A method of CO_2 welding has been introduced, for example, in which a little powdered flux is added just in front of the arc by means of a magnetic field. There are also various techniques for special cases, such as stud welding and spot arc welding, which might also be included in this diagram. Passing these over, we shall now turn to the three above-mentioned Philips developments, prefacing our discussion with a short account of the basic electrodes.

Basic electrodes

With *acid*-coated electrodes, which were the only covered electrodes known up to 1937 (and are still used in enormous numbers), the arc atmosphere is roughly 40% hydrogen, the rest consisting of water vapour, carbon dioxide and carbon monoxide. The latter two gases originate from carbonates and carbohydrates in the coating. H_2 and H_2O come from substances such as water-glass and kaolin, added to the coating as binders and moulding compounds. For ordinary mild steel these electrodes are very suitable, but when they were used for free-cutting steel (which contains a fairly high percentage of sulphur) they gave porous welds. They proved equally unreliable for low-alloy steel, the welds sometimes showing microcracks. Other difficulties were also encountered, such as under-bead cracking, reduced tensile strength of the deposited metal (noticeable in the occurrence of "fish-eyes" in tensile test specimens taken from this metal) and flaking of enamel from the weld in subsequently enamelled parts.

[7]) Netherlands patent application No. 176 664, March 1953, and U.S. patent No. 2 824 948 in the name of P. C. van der Willigen and H. Bienfait. See also P. C. van der Willigen and L. F. Defize, Schweissen und Schneiden **9**, 50, 1957. An English translation of the latter article has appeared in Welding News, No. 83, 2-14, 1957.

[8]) For the original form of argon-arc welding the recommended international designation is TIG (tungsten inert gas), for the variant with consumable bare wire MIG (metal inert gas) and for CO_2 welding MAG (metal active gas).

At a time when no answer had yet been found to these undesired effects, efforts were being made to make electrodes that would yield a *basic* type of slag (with $CaCO_3$ and CaF_2 as the characteristic constituents of the coating). At first the results were negative, and porous welds were obtained. Sound welds were not produced in this way until a start was made, in France, with "baking" the coated electrodes at high temperatures (about 450 °C). The basic electrode then quickly came into use. In the Netherlands round about 1942 these electrodes were referred to as "St 52" electrodes, since they were found particularly useful for a low-alloy steel of that name. As mentioned, this high-quality structural steel could not always be satisfactorily welded using acid-type electrodes. After 1942 it gradually became clear that the above-mentioned difficulties were attributable to hydrogen [9]), and only then was it understood that it was necessary to bake the electrodes at high temperature in order to remove residual hydrogen sources such as water. For this reason, basic electrodes are still frequently called low-hydrogen electrodes to indicate their essential feature. This, incidentally, is a typical example of a successful technique being applied long before its underlying principles had properly been understood.

Basic electrodes, with their excellent mechanical properties, fell short in one important respect: they were not easy to weld with. The slag was not easily removable, the surface of the bead was not smooth, the arc was not very stable (particularly when welding with alternating current and low open-circuit voltage), and so on. The latter difficulty was due to the fluoride in the coating; because of its high electron affinity, the fluorine accelerated the deionization of the arc and thus hampered its periodic re-striking. This snag was overcome by adding to the coating magnesium powder, which acts as an arc stabilizer, or, in the case of another basic electrode, by applying two coatings one on top of the other, the inner coating containing no fluorine. The other difficulties, however, remained.

The chance of further improving the welding properties came unexpectedly in 1943 from an invention made in another connection.

Contact electrodes

In that year we were working on experiments which it was hoped would improve the thermal

[9]) G. L. Hopkins, Trans. Inst. Welding **7**, 76, 1944. A. E. Flanigan, Welding J. **26**, 193S, 1947. See also J. D. Fast, Causes of porosity in welds, Philips tech. Rev. **11**, 101-110, 1949/50; J. D. Fast, Low-hydrogen welding rods, Philips tech. Rev. **14**, 96-101, 1952/53.

efficiency of welding electrodes, i.e. reduce the heat generated in proportion to the deposited quantity of metal. Possible means to that end seemed to be to reduce the depth of penetration (and hence waste less heat on needless melting of the material of the workpiece) by not subjecting *all* the metal to the passage of the current. With this in mind we decided, after various experiments, to distribute uniformly in the coating a quantity of iron in the form of iron filings (later iron powder). This resulted in a very thick coating with a high iron content. When we started welding with these electrodes, we noticed that after striking an arc it was not necessary to hold the electrode at a certain distance from the workpiece, but that it could be rested upon it, i.e. the electrode could be touch-welded. During welding with the heavily coated electrode a deep cup forms at the end of the coating which automatically keeps the arc length constant (*fig. 3*). Moreover, we

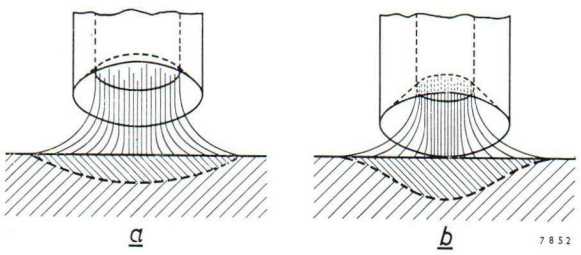


Fig. 3. *a*) Welding with a normal coated electrode and free arc. *b*) Welding with a Contact electrode. The coating being thicker, a deeper cup is formed than in (*a*). The rim of the cup can thus rest on the surface of the workpiece. The arc here is concentrated on a thinner core wire, so that the penetration in the middle is somewhat deeper than in (*a*).

found that given a suitable percentage of iron powder with the proper grain size, the coating has sufficient electric conductivity to restart the arc automatically by contact with the workpiece. Since these electrodes are in mechanical as well as in electrical contact with the workpiece via the coating we gave them the name "Contact electrodes" [3]).

Further experiments showed that existing types of electrodes could be adapted to the Contact principle, by transferring part of the core iron in a finely powdered form to the coating, in which it is intimately mixed with the other constituents. The iron-to-slag ratio is then unchanged; the slag has the same composition and the weld is given roughly the same mechanical properties as with the original electrode. The advantages are not merely greatly increased ease of welding, but also improved shielding (due to the deep cup) against the nitrogen in the atmosphere, and above all greater welding speed. Because of the shielding effect of the cup, the heat radiation by the arc is used more effectively, so that for a given power more metal is deposited per unit time — the effect aimed at in the original experiments, although in a different direction — and less of the molten metal is lost by spatter. As a result (and because the arc voltage is somewhat higher) Contact electrodes give an average rate of deposition of 0.24 grams of metal per ampere-minute compared with 0.17 grams with conventional coated electrodes (with a free arc).

In about 1947 Philips put on the market Contact versions of *acid* electrodes. In the following years similar types were produced by manufacturers in other countries, e.g. Russia. In Great Britain and the United States they first came into use in 1952 [10]): the welders there were initially opposed to the introduction of these electrodes which made welding easier and faster. The iron content in the coating of present-day Contact electrodes generally weighs roughly half of the iron core. The "deposition efficiency" of the electrode, i.e. the ratio between the weight of the deposited metal and the consumed core, is consequently about 150%.

Zircon-basic electrodes

Soon after 1943 we attempted to make a Contact version of the basic type of electrode, whose welding properties were in greatest need of improvement. Our efforts at the time were frustrated by the inadequate *viscosity* of the slag formed.

It will be evident that the viscosity of the slag is of considerable importance from the point of view of welding in various positions, e.g. the welding of flat grooves, vertical upward and overhead welding, and standing-fillet welding. In vertical upward welding, for example, a fairly thin, fluid slag is permissible or even desirable, as it easily runs off downwards; for standing-fillet welds, on the other hand, a thin, fluid slag is unsuitable because it collects at the lower edge of the bead, displacing the metal there, and thus weakening the final joint (*fig. 4*).

At the melting point of steel, which is about 1500 °C, it is very difficult to measure the viscosity of the slag exactly with the available instruments because the instruments are attacked. Data on this subject are therefore scarce and unreliable. Nevertheless it can be said that the viscosity of acid slag at 1500 °C is in the region of 10 poise, and that of basic slag is

[10]) In English-speaking countries Contact electrodes are often referred to as iron-powder electrodes. In Germany the term "Hochleistungselektroden" is used.

roughly half this value. (The viscosity of both is thus much higher than that of the molten steel itself, which is 0.025 poise.) Every welder knows that basic slag is much less viscous than acid slag: he would say that the latter has the consistency of treacle and the former that of water. If now, in order

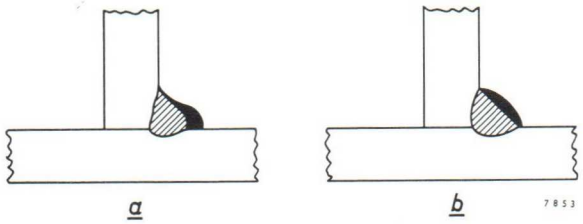


Fig. 4. For depositing a standing-fillet weld an electrode is required that forms a viscous slag. If the slag is too fluid, it pushes the molten metal at the lower side of the bead to one side; this can be seen in (*a*), while (*b*) shows the desired shape.

to make a Contact version, iron powder is added to the coating of a basic electrode, the already thin slag becomes even more fluid. Attempts to thicken the slag by adding, for example, a high percentage of rutile (TiO_2) to the coating, yielded the desired result but reduced the basicity of the slag [11]), the consequence being poorer mechanical properties and a greater tendency for the weld to show porosity.

After many years of research, a good basic Contact electrode was at last produced in 1957 [4]) by adding to the coating about 20% zirconium oxide (ZrO_2). It proved impracticable to measure the viscosity of the slag from this coating with any reliability, not only because of the difficulty mentioned but also because of the added complication that the ZrO_2 particles in the "zircon-basic" slags tend to settle out during the measurement (the specific gravity of ZrO_2 is 5.6 g/cm³, that of the basic slag about 3 g/cm³). This very fact, however, indicates that the slag must have a higher viscosity, for it is evident that the ZrO_2 crystals, whose melting point is 2700 °C, have not reacted with the other constituents and are present as such in the fluid slag. A suspension of solid particles in a liquid has a viscosity η_s which is greater than the viscosity η_0 of the liquid itself. This is expressed by the Einstein viscosity equation:

$$\eta_s = \eta_0 \left(1 + \frac{5}{2}\varphi\right),$$

where φ is the total volume fraction of the dispersed particles.

[11]) The basicity is given by the molecular ratio of basic and acid oxides; in our case these are mainly CaO and SiO_2 respectively, with $CaO : SiO_2 \approx 1.5$.

The viscosity-enhancing effect of the ZrO_2 proved useful both for making a Contact version of the basic electrode and also for the normal versions. With the new "zircon-basic" electrode, type Ph 86, standing-fillet welds can easily be made. The Contact version of this electrode, type C 6, which has a deposition efficiency of about 160% and is now marketed alongside the free-arc version, is particularly suitable for the rapid filling of Vee joints. Owing to the greater viscosity of the slag, zircon-basic electrodes cannot however be used for vertical upward welding.

For welding in various positions an important point besides the viscosity is the *melting range* of the slag (i.e. the difference in temperature between the softening point and the melting point, which lies between 1200 and 1400 °C for most slags) [12]). This melting range can be measured by heating pieces of slag on a platinum sheet in a furnace and observing the change of shape through an optical pyrometer. The value found for various acid slags is 100-200 °C, and only about 40 °C for normal basic slags. The zircon-basic slag is found to have a much longer melting range, more like that of acid slags.

Returning for a moment to the ZrO_2 particles present in suspension, a side-effect of their presence is that the slag is light grey in colour where it has not been exposed to the air; this is, as it were, an identification mark of the zircon-basic electrode. See *fig. 5*, which shows photographs of the underside of two kinds of slag.

The presence of ZrO_2 crystals in the slag has also been demonstrated by X-ray diffraction diagrams. At the normal basicity of 1.5, ZrO_2 evidently behaves neutrally; at a higher basicity it appears that the compound $CaZrO_3$ is formed.

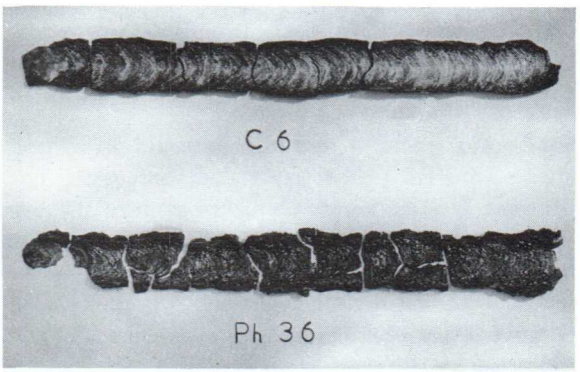


Fig. 5. Slag formed by a zircon-basic and a basic electrode (C 6 and Ph 36, respectively). The photo shows the underside of the slag, which was in contact with the bead. The slag of C 6 appears light grey in the middle where it was not exposed to the air; this is due to the dispersed ZrO_2 particles (which are also concentrated near the surface, due to sedimentation of the particles). If the bead is deposited not on a flat plate, as here, but in a groove, the air has less access to the slag and the light-grey colour is perceptible up to the edges.

[12]) G. J. Pogodin-Alexejew, Theorie der Schweissprozesse, Verl. Technik, Berlin 1953, p. 188.

Notable features of the C 6 and Ph 86 electrodes are the ease with which the slag is removable even from narrow grooves, and the smoothness of the weld. The surface of the weld metal is less ridged, and the transition between weld and parent metal is smoother than with the basic electrode (*fig. 6*). It is due to this that fatigue tests of welded joints in the

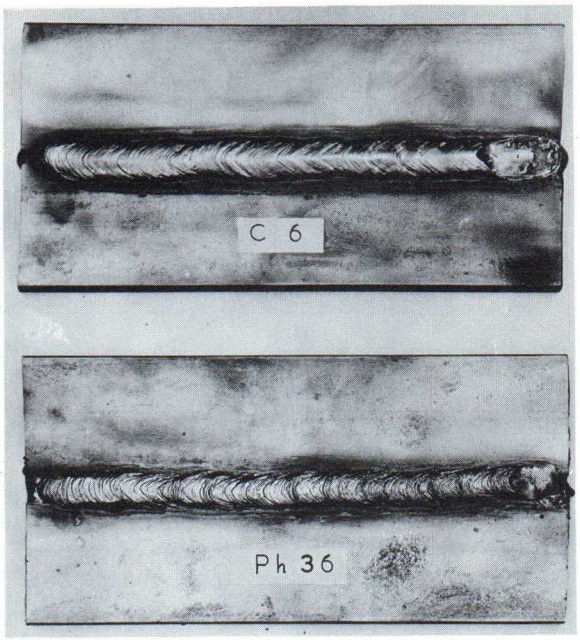
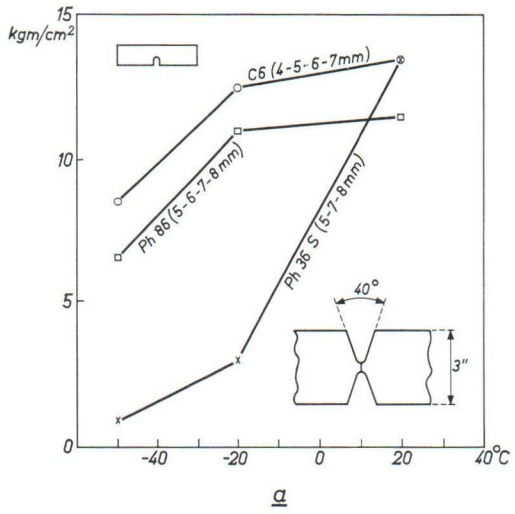


Fig. 6. Appearance of an unprepared butt weld in 13 mm plate with 2 mm gap; one made with a zircon-basic Contact electrode C 6 and the other with a normal basic electrode Ph 36. The surface of the latter weld is uneven and the weld metal does not spread out very well.

as-welded condition (i.e. not smooth-ground or stress-relieved by heat treatment) have shown values of fatigue strength about 30% higher than similar welds made with normal basic electrodes.

As mentioned at the beginning, the introduction of basic electrodes was stimulated by the increasing use of low-alloy steel, which makes lighter constructions possible but at the same time imposes high demands on the mechanical properties of the welds. Another tendency in modern engineering is the use of thicker and thicker plates, e.g. for reactor pressure vessels, for very large ships and in the chemical industry. For welding such plates, too, basic electrodes have proved equally successful. We have carried out extensive tests to determine the notch toughness of welded joints in very thick plate (76 mm) made with type C 6 and Ph 86 electrodes. At room temperature these electrodes were found to give practically the same high notch-toughness value as the normal basic electrode Ph 36 S. At low temperatures, on the other hand, the zircon-basic electrodes give appreciably better values; see the graphs in *fig. 7a* and *b*, which also give sketches of the type of impact bars used and some data on the welds. The plates measured $400 \times 400 \times 76$ mm. The total number of passes in each welded joint was 24. After each pass we reversed the workpiece and waited until the temperature had dropped below 100 °C, after which it was cooled with water and dried. This method gave a better approximation to the situation during the welding of very thick plates in practice.

In many modern methods of welding the welder must take precautions to avoid inhaling too much of the fumes. This also applies to the basic



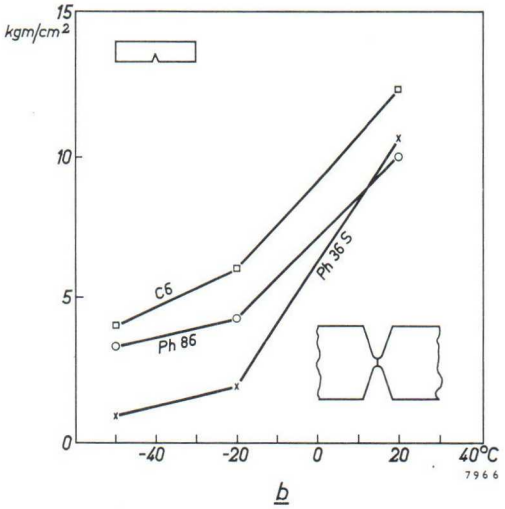


Fig. 7. *a*) Average notch toughness of a welded joint made with three types of electrode, as a function of temperature. The notch toughness was measured on Charpy test specimens 7×10 mm² of the shape shown top left (U notch), taken from a welded joint in 3 inch thick plate, the preparation being as shown bottom right.
b) As (*a*) using Charpy test specimens 8×10 mm² with a Vee notch as shown top left.

electrodes, the vapours from which contain a small amount of fluoride. *Table I* shows the total amount of fluorine per 100 grams of deposited metal measured in the welding vapours from basic and zircon-basic electrodes. The quantities found for zircon-basic electrodes are considerably smaller, which was to be expected since the coating of zircon-basic electrodes contains much less fluoride.

Table I. Measured quantity of fluorine in welding vapour from various electrodes.

Electrode, type and thickness	mg F in welding vapour per 100 g deposited metal
C 6 (4 mm)	55-56
Ph 86 (5 mm)	75-77
Ph 36 S (5 mm)	97-139

CO_2-shielded welding

The development now to be described constitutes an entirely different approach to the problems of arc welding.

In 1952 we started experimental welding in mixtures of CO and CO_2, based upon our knowledge of the composition of the arc atmosphere in welding with basic electrodes. We soon stopped adding CO, which is relatively expensive and is moreover poisonous. Welding with an arc atmosphere consisting entirely of CO_2 had already been tried in 1926, but the result at that time was a brittle and porous weld. We shall describe below how this problem was finally solved, but first it will be useful to outline the whole method as at present employed.

Main features of CO_2 welding

As mentioned at the beginning of this article, CO_2 welding is based on the use of consumable bare wire which is unwound from a reel and fed by a motor-drive to the arc. This method makes it possible to weld long joints uninterruptedly, and can readily be made automatic or semi-automatic.

The principal difference between welding with separate electrodes and welding with consumable bare wire consists in the way in which the current is supplied. The current is fed to the wire close to the tip by means of a feed tube as sketched in *fig. 8*. This makes it possible to work with much higher current densities, and hence at greater speed, than when using electrodes where the current is normally applied to the rear end, i.e. at a distance of 35 or 45 cm from the arc. An electrode of 2 mm core-wire diameter, for example, is welded with a current of at the most 70 A; a higher current would make the electrode too hot, perhaps even red-hot, giving rise to decomposition of the coating and other difficulties. In CO_2 welding with consumable wire of the same thickness, the welding currents can be as high as 700 A.

A machine for fully automatic CO_2 welding was shown in fig. 1. The motor (*M* in this figure) feeds the wire from the reel to the welding arc through two hardened-steel rollers. The wire is clamped with a variable pressure between the rollers, and the speed of the motor can be continuously varied by a manual control (*K*). The nozzle (*E*), in the axis of which is located the feed tube and out of which the CO_2 gas flows through an annular opening, is double-walled and is fitted with closed-circuit water cooling.

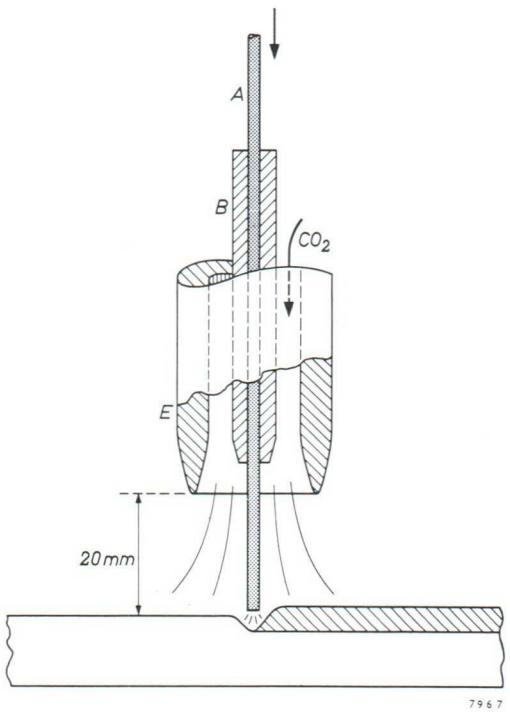


Fig. 8. Sketch of the bottom section of the welding head. *E* is the water-cooled nozzle from which the CO_2 is supplied. The consumable bare wire *A* is fed through tube *B*. The current is supplied to the wire via the same tube, the mouth of which is only about 25 mm from the wire tip. To ensure good contact it is essential not to straighten the wire beforehand.

The wire projecting from the tube *B* is pre-heated by the current it carries. This effect, which is extremely important with the high current used and the small diameter of the wire, can largely be controlled by varying the length of the wire protruding (here about 25 mm). This effect was recognized in 1943 by T. Hehenkamp and described in U.S. patent No. 2 475 835.

Complete equipment for semi-automatic CO_2 welding can be seen in *fig. 9*.

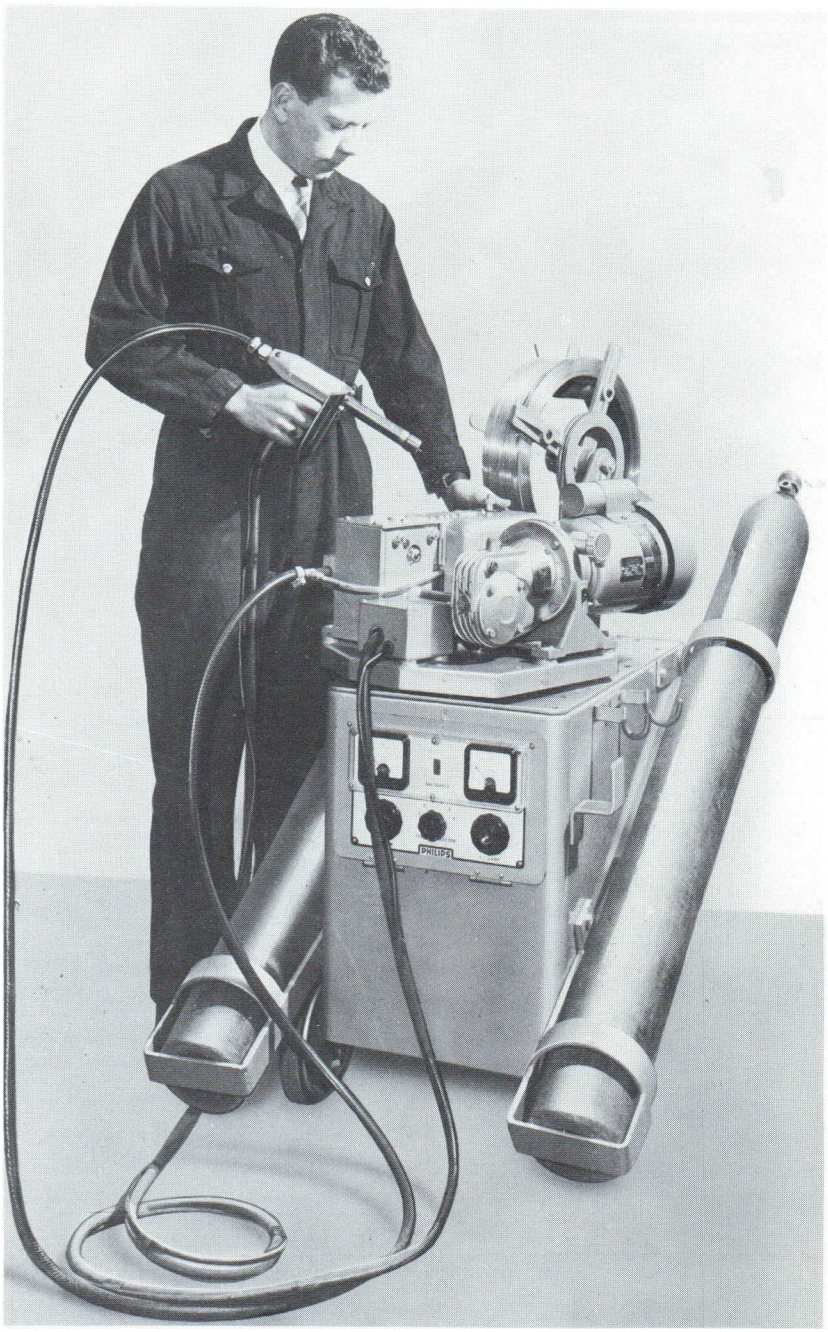


Fig. 9. Complete equipment for semi-automatic CO_2 welding. An electric motor feeds the consumable bare wire from the reel through the long flexible pipe to the welding gun, which the operator passes along the joint to be welded. The current (DC up to 300 A in this machine) is supplied near to the wire tip; the carbon dioxide (from one of the two gas cylinders) flows through a nozzle in the gun and envelopes the wire electrode and the arc.

The heat generated by the current is very efficiently used in all gas-shielded methods of welding, since no slag needs to be melted. Consequently the deposition rate (at a given current) is considerably higher in these methods than, for example, when using the electrodes discussed above: in CO_2 welding the deposition rate is between 0.30 and 0.40 grams per ampere-minute. A further advantage of CO_2 welding is the particularly favourable shape of the penetration: whereas argon-shielded welding produces a V-shaped penetration (*fig. 10a*), CO_2-shielded welding produces a U-shaped penetration (fig. 10*b*) which is moreover, using the same current, several millimetres deeper than with argon. The upshot is that CO_2 welding, generally speaking, permits less bevelling of the parts to be welded, and thus narrower grooves that can be filled in fewer passes.

All this makes CO_2 welding a very fast process. The gain in welding speed compared with methods employing separate electrodes is of course enhanced by the fact that it is no longer necessary to stop every minute or so to insert a new electrode. Compared with the continuous method of submerged-arc welding (described at the end of this article) a favourable feature of CO_2 welding is that no time is wasted in removing slag from the bead.

Apart from the welding speed it is important to consider the mechanical properties of the welds obtained. We shall presently examine in some detail the manner in which the metal is shielded in CO_2 welding, but it may be mentioned at this stage that the deposited metal has roughly the same mechanical properties as obtained when using basic electrodes. As to the properties of the weld, if good use is made of the deep penetration which is a feature of

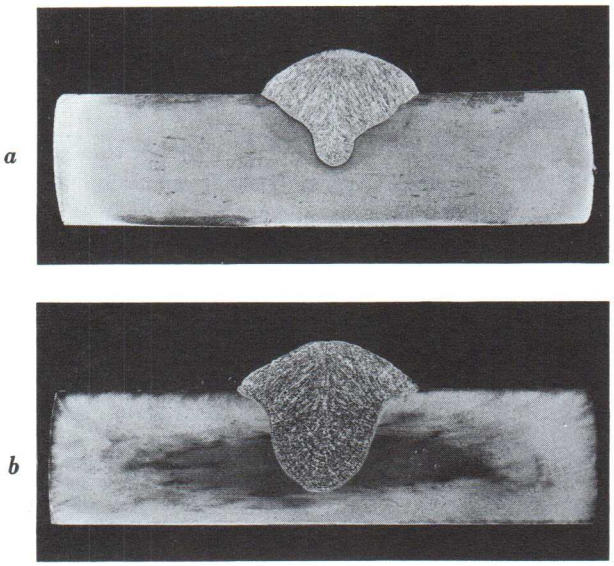


Fig. 10. Etched cross-section of weld beads, deposited with 2 mm thick bare wire on a flat plate 13 mm thick, at a current of 480 A and a rate of travel of 40 cm per minute: (*a*) welded in argon + 5% oxygen, (*b*) welded in CO_2. In (*a*) the penetration is V-shaped, 7 mm deep; in (*b*) it is U-shaped, 10 mm deep.

CO_2 welding, the deposited metal will be mixed with a high percentage of molten parent metal. This is one of the main points governing the mechanical properties of the weld. Also of importance is the number of layers: several layers entail repeated annealing of the previous beads, and this as a rule improves the structure and increases the toughness of the weld. The deep penetration in CO_2 welding does, however, call for extra precautions in some cases. If the width of the penetration is less than the depth, which may happen if the arc is very short and the current high, internal shrinkage cracks (hot cracking) may occur. This phenomenon is also known in other welding methods. The remedy in all cases is to increase the arc length and/or to widen the groove to be welded.

Shielding the metal in CO_2

When the molten metal in the arc is enveloped by CO_2, it is completely shielded against the nitrogen in the atmosphere. Although the metal droplets transferred in the arc are also thereby shielded against oxygen in the air, they are nevertheless highly susceptible to oxidation: the CO_2 itself dissociates at a relatively low temperature into CO and oxygen, and at the temperature of the arc there is likely to be almost complete dissociation, i.e. a high partial oxygen pressure (see fig. 13).

The oxidizing action of the dissociated CO_2 was evidently the reason for the porosity of the welds which caused the failure of the experiments in 1926. The experiments begun in 1952 were based on the idea that the oxidizing action could be compensated by adding deoxidizing agents to the steel wire, i.e. alloying elements such as titanium, silicon, manganese, etc., the oxides of which possess a high heat of formation and which therefore largely bind the "available" oxygen by, e.g., the following reactions:

$$Si + 2\,CO_2 \rightleftarrows SiO_2 + 2\,CO,$$

$$Mn + CO_2 \rightleftarrows MnO + CO.$$

In view of the high partial oxygen pressure expected (even higher than 50% in the case of complete dissociation in the arc) we thought at first that fairly high percentages of additions would be necessary. The experiments soon showed, however, that the CO_2 had little oxidizing action, less even than the air (partial oxygen pressure 20%) — a surprising fact to which we shall return below. Under favourable conditions the addition of 0.3% Si + 0.3% Mn was found to be sufficient to prevent porosity in the weld metal [7]). The results of a representative experiment, illustrating the relatively low oxidation, are summarized in *Table II*. In this experiment a particular kind of weld was made successively in three different gas atmospheres: CO_2, air, and argon + 20% O_2. In the latter atmosphere oxygen was present with roughly the same partial pressure as in air, so that this experiment could show whether the nitrogen in the air had anything to do with the oxidation. The welding wire had roughly the same composition as that now used in the CO_2 welding process; it contained 0.9% Si and 1.6% Mn, together with various other elements. The tabulated results of the analysis of the deposited metal show that in CO_2-shielded welding relatively little Si disappears (the "burn-off" of Si is by far the smallest). Further experiments demonstrated that, as regards the burn-off of Si and Mn, the CO_2 behaves like argon plus about 9% O_2.

Table II. Burn-off of silicon from consumable steel wire, thickness 1.6 mm, welded in three different gas atmospheres. In all three cases the wire contained 0.9% Si, together with 0.1% C, 1.6% Mn, 0.02% P and 0.02% S. The metal was deposited in a wide Vee joint in 16 mm thick steel plate, using direct current of 340 A (wire positive) at 32-33 V arc voltage and a travel speed of 40 cm/min. The joint was filled in eight passes, with interpass cooling to 100 °C. Gas feed at nozzle 25 litres per minute.

Shielding gas	Si in deposited metal %	Si burn-off in consumed wire %
CO_2	0.46	48
Air	0.30	67
Argon + 20% O_2	0.27	70

This favourable behaviour makes it possible to use relatively low percentages of Si and Mn, which

can readily be added as alloying elements to the steel wire. Of course, one cannot expect the added Si and Mn to give a 100% shielding of the iron (for reasons of reaction kinetics), even though the Si, for example, is by no means completely consumed (see Table II): a small quantity of Fe is oxidized, and all reaction products together form silicates, which are found in the form of a thin film of slag on the weld bead. If a second bead is to be applied to the first, this film is in many cases not brushed off before welding.

In regard to the composition of the steel of the weld wire, it is also necessary to take account of the *carbon* which is always present. A very low carbon content in the deposited metal cannot be realized — even if it were desirable — since, when welding with a wire containing very little carbon (e.g. 0.03%), some carbonizing of the deposited metal always occurs (up to about 0.05%). The explanation is that the transferred droplets are enveloped in a blanket of almost pure CO, from which it is possible for iron to take up carbon.

In view of its toxicity, we cannot be indifferent to the further fate of this CO, produced by dissociation of CO_2 and also by the above-mentioned reactions with Si and Mn etc. One might be inclined to suppose that the CO would be oxidized by the oxygen in the surrounding air. Measurements of the quantity of CO present, with and without the admission of air to the surroundings of the CO_2 arc, have shown [7]) that this is only partly the case: because of the marked dilution of CO with CO_2, some of the CO does not undergo combustion and remains in the welding vapours. At a supply of 20 litres of CO_2 per minute, the fumes from the welding site are found to contain 0.5 litre of CO per minute. Air conditioning is therefore just as essential with CO_2 welding as with most other welding processes, at least in small workshops.

Droplet transfer in CO_2 welding

When developing a welding process it is desirable to have a clear idea of the way in which the molten metal is transferred in the arc to the workpiece. Right from the beginning of our work on CO_2 welding we therefore studied the droplet transfer with the aid of a high-speed 16 mm cine camera [13]). As a result we were able to give a reasonably satisfactory explanation for the relatively slight oxidizing action discussed in the foregoing.

The camera used can take up to 3000 exposures per second; projecting the film at normal speed (24 frames per second), the welding process is thus seen in slow motion, being slowed down more than 100 times. *Fig. 11* shows the set-up employed. A weld bead is deposited on a flat steel plate. The arc remains stationary and the plate is moved, so that the cine camera and the illumination can be kept in a fixed position. When filming we used a long arc,

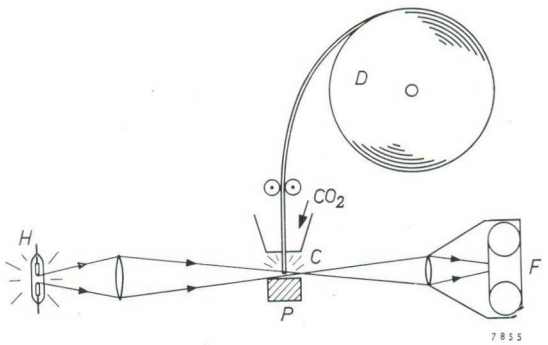


Fig. 11. Set-up for filming droplet transfer in CO_2 welding. *C* arc with tip of consumable wire automatically fed from the reel *D*. The workpiece *P* moves in the direction of the observer. *F* 16 mm cine camera capable of 3000 exposures per second. *H* water-cooled high-pressure mercury lamp, which, with the aid of a lens, provides a sufficiently brilliant background for the arc during filming.

the melting tip of the wire being about 7 mm above the surface of the plate, with the object of observing as much as possible of what was happening in the arc. In actual welding practice a short arc is always preferable, in order among other things to reduce spatter loss (this applies to all arc-welding methods). The tip of the wire is then located about 2 mm or even less above the surface of the workpiece. This is possible owing to the fact that, at the enormous current density used in CO_2 welding (about 20 000 A/cm^2), the "arc pressure" blows, as it were, a cavity into the parent metal immediately after it has been melted, thus creating its own arc length. The arc then burns largely in the cavity, thereby reducing spatter loss and also energy losses due to radiation.

The slow-motion films show that the electrode material is transferred in a CO_2 atmosphere in the form of *very coarse droplets*, the diameter of which is between 3 and 4 mm, independent of the diameter of the wire used; see *Table III*. The droplets grow asymmetrically on the tip of the wire towards the side of the bead already deposited. This oblique melting of the wire appears to be bound up with the

[13]) See the articles mentioned in reference [7]). The same set-up was used for studying the droplet transfer of coated electrodes: P. C. van der Willigen and L. F. Defize, Philips tech. Rev. **15**, 122, 1953/54. See also: L. F. Defize and P. C. van der Willigen, Droplet transfer during arc welding in various shielding gases, Brit. Welding J. **7**, 297-305, 1960. (The Sir William J. Larke medal of the Institute of Welding for 1961 was awarded to Messrs. Van der Willigen and Defize for the latter article. — *Ed.*)

Table III. Average size of transferred metal droplets in CO_2 welding with wire of various thicknesses.

Wire diameter mm	Current A	Current density A/cm²	Droplet diameter mm
2.5	490	10 000	3.7
2.0	315	10 000	4.0
1.2	110	10 000	3.6

fact that at the fairly rapid relative displacement of arc and workpiece (e.g. 110 cm/min) the arc is in contact with the hottest part of the workpiece, i.e. the metal last melted, and that the tip of the wire receives radiation mainly from that side. At a given moment the growing droplet separates, the arc jumps over from the droplet to the wire, and the droplet falls spinning (this can clearly be seen in the film) along a curved path into the weld pool. Since the arc creates its own length, as we have seen above, the coarse droplets do not give rise to short-circuiting. *Fig. 12* gives some frames from a colour film taken with the camera pointed obliquely downwards, showing the cavity blown by the arc, and the oblique melting of the wire.

It was known from a previous investigation [14]) (and confirmed by our experiments) that the droplet transfer in argon is quite different, particularly when a small percentage of oxygen has been added to the argon: *small* droplets are formed at the tip of the melting wire and split off in the direction of the wire. Increasing the current reduces the size of the droplets, until at last they are no longer separately distinguishable [15]).

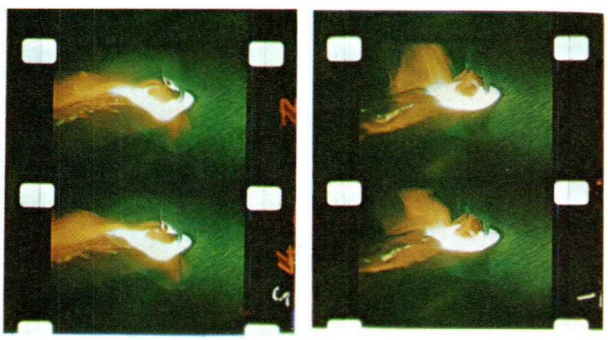


Fig. 12. Frames from a 16 mm colour film of CO_2 welding, using a short arc (as in practice). The film speed was 2000 frames per second, the camera being directed obliquely downwards, giving a view of the cavity blown by the arc in the workpiece. On the left can be seen the growing end of the weld bead; right, above the weld pool, the molten tip of the wire and the droplet passing in the arc. (This is particularly clear in the photos on the right, which also show under the weld pool the reflection of the electrode wire in the gleaming surface of the workpiece.)

[14]) H. T. Herbst and T. McElrath, Welding J. **30**, 1084, 1951.

[15]) Our investigation related to helium as well as argon: as far as droplet transfer is concerned, He resembles A at high currents, and CO_2 at low currents.

The slow-motion pictures reveal another remarkable difference between argon and CO_2: with argon the arc plasma envelopes the tip of the wire, including the hanging droplet, but with CO_2 only part of the droplet is covered, no more than half its surface area. Without going too much into the details, some of which have still not been explained, we should mention here that this difference in plasma coverage is probably attributable to a difference in the *contraction* of the plasma [16]). Both with argon and CO_2 there is the familiar contraction due to magnetic attraction of the current paths ("pinch effect"), but the contraction is increased by local cooling of the gas [17]), especially in the neighbourhood of the relatively cold wire tip, and this effect will be most marked in gases with the highest thermal conductivity — a property which is particularly noticeable in gases which dissociate at the high arc temperatures, such as H_2 and N_2 and above all CO_2 (see *fig. 13*).

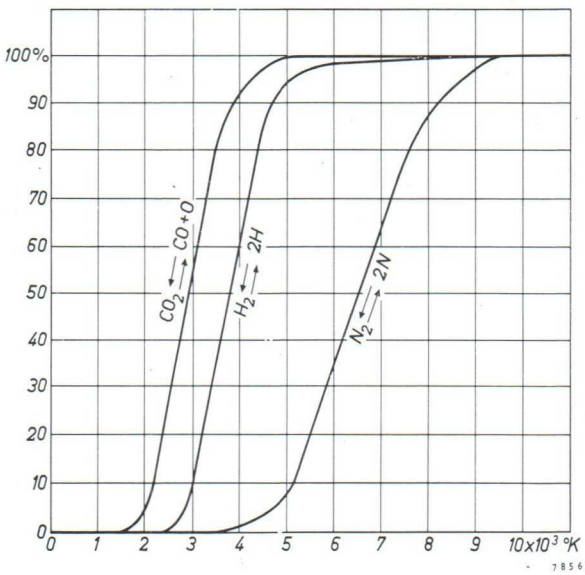


Fig. 13. Dissociation of various poly-atomic gases, as a function of temperature, at a pressure of 1 atm. (After: H. C. Ludwig, Welding J. **38**, 296S, 1959.)

The differences found between argon and CO_2 in regard to droplet size and plasma contraction are undoubtedly related to one another. If the cross-section of the contact surface between plasma and droplet is smaller than the cross-section of the neck of the droplet, as in CO_2, the result is an electro-

[16]) See last article of [13]).

[17]) In the "plasma burner", which is now in use for producing exceptionally high temperatures (e.g. 10 000 °C), a water-cooled copper wall is employed for constricting the arc plasma. See, for example, Welding and Metal Fabrication **27**, 287, 1959.

dynamic force directed upwards, which is increased by the pressure of the gas in the arc. This enables the droplet to grow to quite a size until the pinch effect in the neck, together with the force of gravity and surface tension, cause the droplet to separate and the arc to jump back to the wire (see above). If the contact surface is larger than the cross-section of the droplet neck, however, as in argon(*fig. 14*), the resultant force is directed downwards. The droplets then break off and fall into the weld pool without having a chance to grow large.

We have dwelt on these effects at some length because, in our opinion, they make the relatively slight oxidation of the steel in the CO_2 atmosphere to some extent understandable. For owing to the large volume of each droplet transferred, and also to the fact that the droplet is only partially in contact with the extremely hot plasma, the surface exposed to oxidation is relatively small. It is probable, however, that other factors have a bearing on the oxidation, as for example the very short time in which the droplet forms and passes in the arc, and the violent perturbation of the molten material. Our hypothesis is not yet capable of providing a quantitative explanation of the processes involved, especially the measured burn-off of Si and Mn.

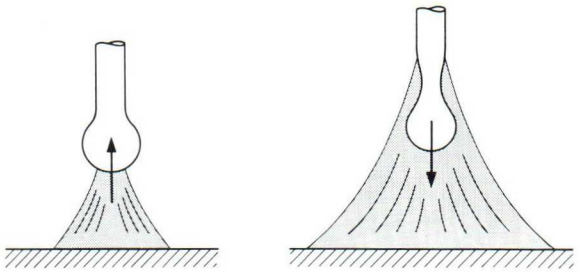


Fig. 14. Droplet formation on the wire tip, *a*) CO_2 welding, *b*) welding in argon + 5% oxygen. In (*a*) the arc plasma (shaded area) covers only a relatively small part of the droplet surface. Electrodynamic and other forces produce a resultant upward force, enabling the droplet to grow to an appreciable size; in (*b*) the resultant force is in the downward direction, and the droplets remain small.

Comparison of CO_2 welding with other methods

In the engineering industry there are many machining operations that resemble one another to some extent but which exist in their own right side by side in fields that often overlap, as for example boring, drilling, broaching, punching, milling, reaming, etc. One cannot say that any one of these methods is "the best", and the same applies to the many and various methods of welding. Any comparison between them should rather serve as an attempt to delineate their useful scope.

If we first compare CO_2 welding with the methods using basic and zircon-basic electrodes, we note that while they are all suitable for mild steel and low-alloy steel, CO_2 welding entails higher investment costs and therefore will not be considered where the volume of welding work is small. Moreover, zircon-basic electrodes give a smoother weld and will accordingly be preferred where a particularly trim finish is required. Another limitation of CO_2 welding, at least in the present state of the art, is the need to use direct current, which again calls for more expensive welding equipment. Alternating current is less suitable because CO_2 has a high ionization potential, which means that a very high open-circuit voltage would be required from the welding transformer to ensure reliable re-striking. (This drawback can be eliminated by using special wire containing as additive or coating a substance having a high thermionic emission, e.g. caesium.)

In the case of mass production, however, or where very long joints have to be welded, the investment costs are outweighed by the greater speed of CO_2 welding. This process can be used to best advantage for automatic or semi-automatic work, to which it is ideally adapted. In fact, where fully automatic work is concerned, there is only one other method that can compare with CO_2 welding, and that is submerged-arc welding, which has hitherto been the most widely adopted automatic welding process.

Submerged-arc welding was described very briefly at the beginning of this article. *Fig. 15* gives a clearer idea of its operation. Bare wire is continuously fed from a reel and melted in an arc which is blanketed by the uninterrupted deposition of a

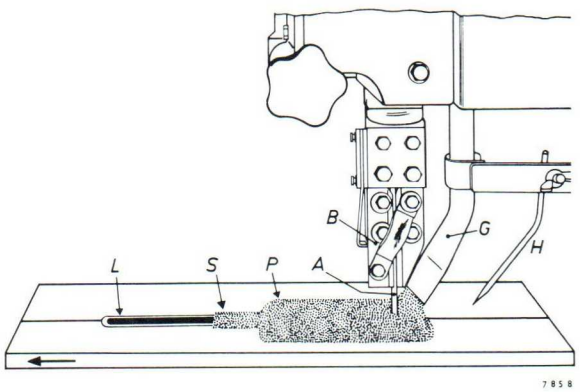


Fig. 15. Submerged-arc welding. *A* consumable bare wire. *B* current supply. *G* tube through which the slag-forming granular flux *P* is continuously deposited over the arc. The arc itself is therefore invisible and burns in a cavity inside the layer of flux. *S* weld bead with layer of slag. *L* bead from which slag has been removed. *H* adjustable pointer for directing the arc to the groove to be welded.

granular slag-forming flux. The fact that the arc is thus submerged largely governs the characteristics of the process: 1) very high currents can be used since the intense radiation is not troublesome; 2) relatively good use is made of the arc heat, which is retained inside the fusible material; 3) the method lends itself well only to fully automatic welding, for if the operator wanted to control the welding head manually he would have to work by feel, being unable to see the arc. This is the price to be paid for the advantage of not having to shield the eyes. As regards point (1) currents of 1000 A and higher were initially used, permitting very high welding speeds and often making it possible to produce thick welded joints in a single pass. It was later established that the structure of such a single-pass weld was too coarse, to the detriment of the mechanical properties. Extremely high currents are therefore no longer the rule, but the welding speeds are still very high.

In comparing CO_2 welding and submerged-arc welding, then, it must be stated that neither method differs much as far as welding speed is concerned. Preference for one method or the other will therefore depend rather on secondary factors, some of which we will mention here.

In the CO_2-shielded method the nozzle through which the gas streams gradually becomes constricted by spatter during welding, and therefore has to be regularly cleaned. Submerged-arc welding, being free from spatter, does not have this drawback. On the other hand, time is wasted with this method in removing the surplus flux and the slag formed around the weld, particularly in the case of narrow joints, where the slag has to be chipped away. This has the further disagreeable consequence of making it difficult to keep the working area clean. One of the virtues of gas-shielded welding is precisely the fact that the gas, after having served its purpose, disappears of its own accord or can easily be continuously exhausted.

We have already pointed out that the melting of the slag costs heat. This reduces the rate of deposition in grams per ampere per minute. *Table IV*, which includes some deposition rates mentioned earlier in this article, shows that the highest values are obtained in CO_2 welding.

The profile of the weld bead in CO_2 welding is often quite convex. This is especially noticeable in square butt welds; in submerged-arc welding the surface of the weld is smoother.

Submerged-arc welding sometimes gives trouble in a humid atmosphere, because the flux absorbs moisture. CO_2 welding, on the other hand, may be hampered by side winds, making it necessary to fit up a screen or curtain, or to increase the gas supply.

Finally, as the flux in submerged-arc welding is held in place by gravity, vertical upwards and overhead welding are of course impossible. These positions present no problems in CO_2 welding.

Table IV. Deposition rate of metal in grams per ampere per minute (approximate values) in various welding methods.

Coated electrodes with free arc	0.17
Contact electrodes	0.24
Submerged-arc welding	0.24-0.30
CO_2-shielded welding	0.30-0.40

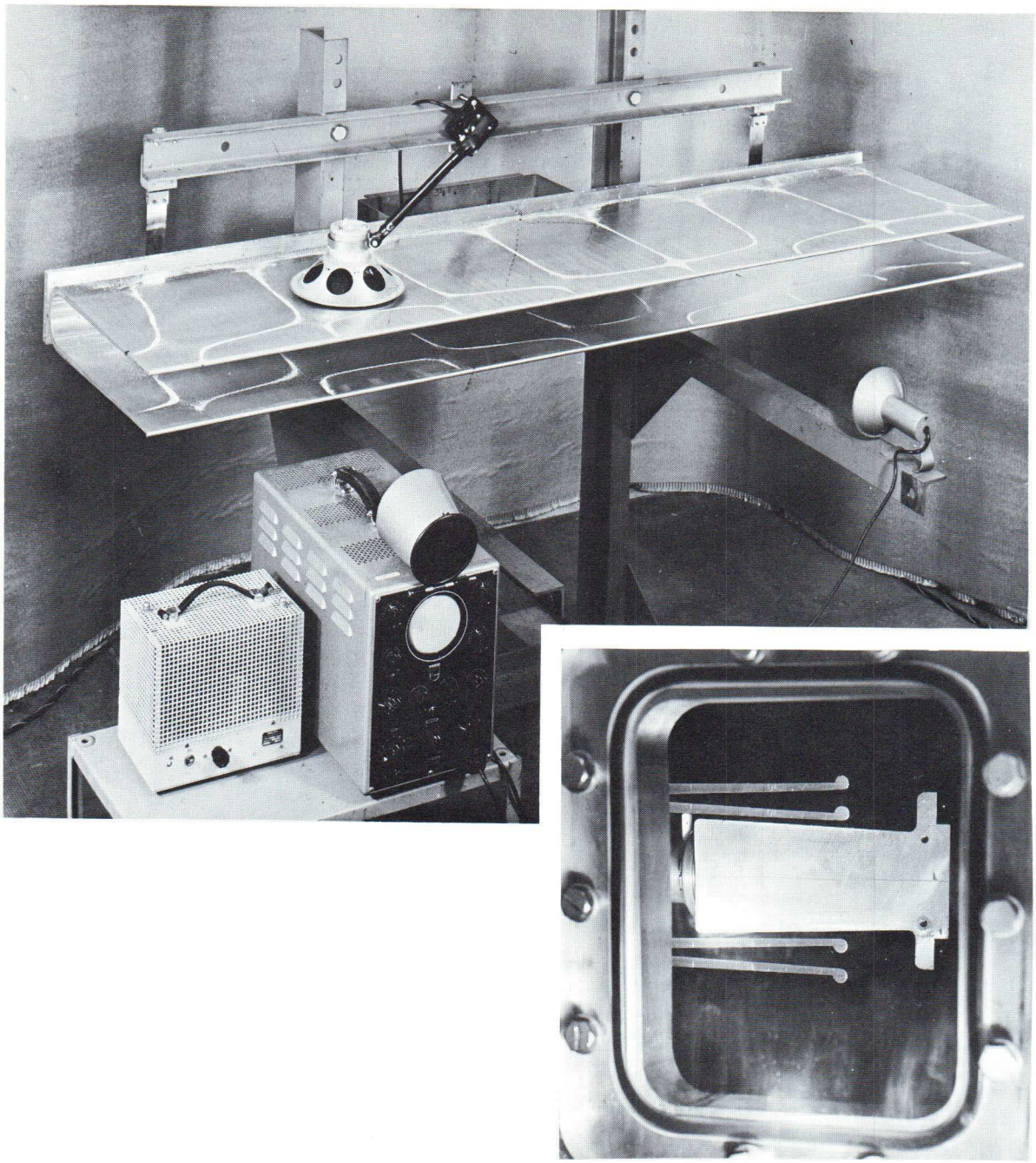
Although the CO_2 method of welding has made headway all over the world in a remarkably short time, and is still gaining ground, it will be clear from the survey presented here that we are not predicting that it will eventually supersede submerged-arc welding. It is rather to be assumed that engineering will have room for both methods, just as there will continue to be scope for coated welding electrodes and for the numerous variants of processes in which the weld metal is shielded by slag, by gas or by combinations of both.

Summary. The many existing methods of arc-welding steel can be classified according to the way in which the molten metal is shielded from the atmosphere. This can be done by slag (as in submerged-arc welding), by gas (as in argon-arc or CO_2 welding), or by slag and gas combined (e.g. using coated electrodes). Three developments are discussed in the latter two categories, due to Philips Research Laboratories in Eindhoven: Contact electrodes, zircon-basic electrodes and CO_2 welding.

Contact (iron-powder) versions of acid-coated electrodes have been widely used since 1947 for their ease and speed of welding and other useful properties. A Contact version of basic (low-hydrogen) electrodes, which are especially suitable for low-alloy steels, was difficult to make because of the low viscosity of the slag. The addition of ZrO_2 to the coating brought an improvement, and the "zircon-basic" electrodes developed on this basis (type Ph 86, and the Contact type, C 6) produce excellent results in down-hand welding, both as regards the appearance of the weld and its mechanical properties.

In CO_2 welding, consumable bare wire, containing about 0.9% Si and 1.6% Mn as deoxidizing additives, is fed by an electric motor from a reel to the arc, which is shielded by carbon dioxide gas issuing from a nozzle around the tip of the wire. Although a very high partial oxygen pressure is to be expected, due to dissociation of the CO_2 in the arc, the above percentage, or even lower, of Si and other additives is surprisingly enough capable of almost entirely preventing oxidation of the deposited metal. This is attributed to the large size of the droplets transferred in the arc, as observed with a high-speed cine camera, and which is probably bound up with the marked contraction of the arc plasma (in argon, for example, the contraction is much less marked and the droplets are therefore smaller).

Like submerged-arc welding, CO_2 welding lends itself very well to automatic operation. In a comparison of the various methods of arc welding a rough indication is given of their appropriate fields of application.



Chladni's figures on giant tuning-fork

Frequency modulation of the CERN 600 MeV synchrocyclotron at Geneva (whose high-frequency system was designed and built by Philips, Eindhoven in cooperation with CERN) is achieved by means of a vibrating capacitor, which has a cross-section like a tuning-fork. In developing the system, the possibility of interfering resonances occurring during normal operation of the synchrocyclotron had to be prevented. Such resonances are revealed by beautiful Chladni figures like those shown in the upper photograph.

The lower photograph shows a stroboscopic picture of the "tuning-fork" vibrating in its vacuum enclosure.

A nearly identical system has been built at Eindhoven for the NASA 600 MeV synchrocyclotron in the U.S.A. (See: B.Bollée, Chladni's figures on the vibrating capacitor of a synchrocyclotron, Philips tech. Rev. **19**, 84-85, 1957/58. Further reading: Philips tech. Rev. **22**, No. 5, 1960/61, special issue on the CERN synchrocyclotron, with contributions by W.Gentner; K.H.Schmitter and S.Kortleven; B.Bollée and F.Krienen.)

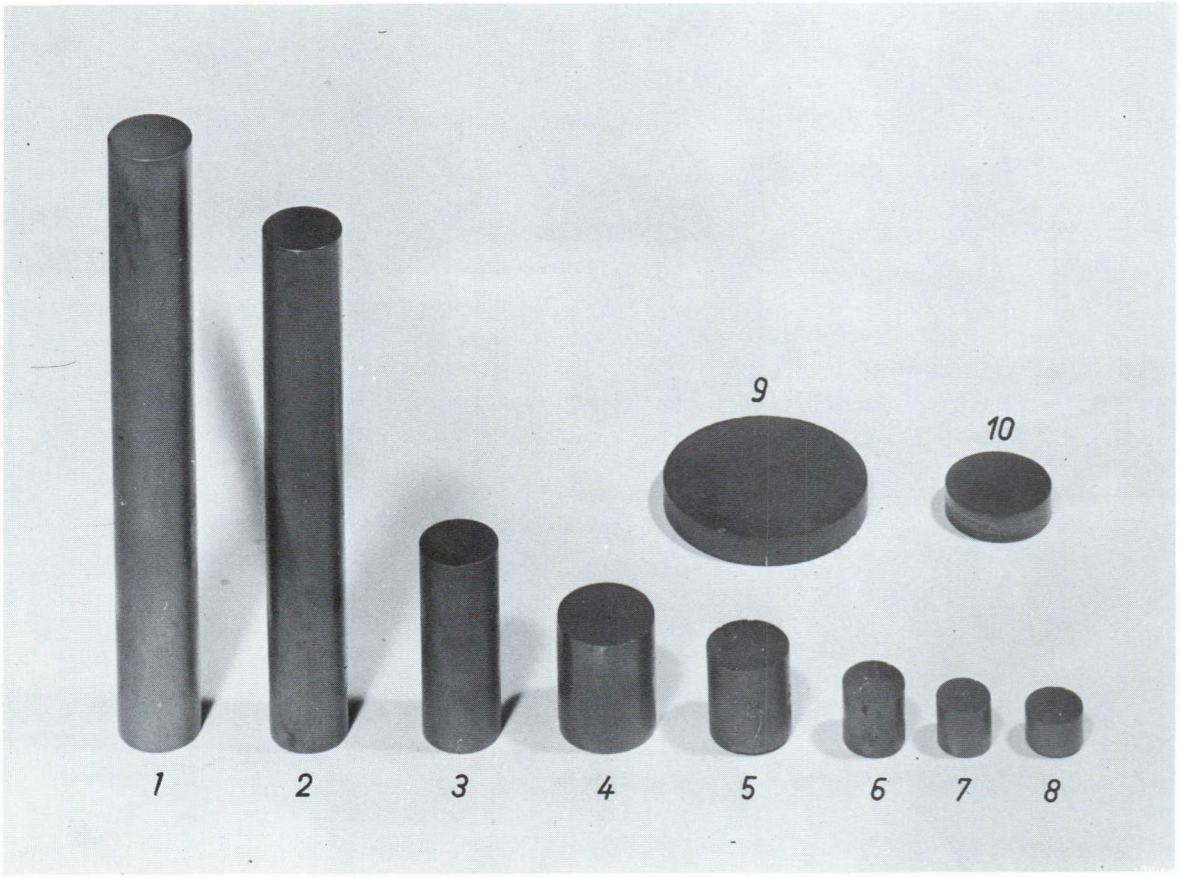
THE EVOLUTION OF THE PERMANENT MAGNET: A BRIEF REVIEW

621.318.2

In evaluating a material for making permanent magnets it is the magnetic properties that are of primary interest. Many other properties play a secondary part, such as workability, resistance to oxidation, the stability of the magnetic properties and, finally, the availability of raw materials and the price of the magnet.

The quality of a magnetic material as regards its magnetic properties can be characterized fairly well by the maximum value, as occurring in the demagnetization curve, of the product of induction B and the corresponding field strength H. The product $B \times H$ defines the magnetic energy obtainable per unit volume of the magnetic material when placed in a magnetic circuit to give a field in an air gap. The greater $(BH)_{max}$, the smaller are the dimensions with which the magnet can be made in order that it should possess a specified total magnetic energy (i.e. create a specified magnetic energy in an air gap). It is then, of course, necessary to design the magnet and the magnetic circuit in such a way that the operating point on the demagnetization curve is precisely that point at which the BH product has its maximum [1] $(BH)_{max}$.

The photograph will serve to illustrate the evolution of the permanent magnet from the point of view of the $(BH)_{max}$ value. It shows a series of magnets made of different materials, each magnet being of such a volume that they would all give the same total magnetic energy in a (hypothetical) air gap. Each magnet has in fact a cross-section and a length such that, when used at their $(BH)_{max}$ point, all magnets supply the same flux ($B \times$ area of cross-section) and the same magnetomotive force ($H \times$ length). A well-designed loudspeaker assembly served as basis for the choice of these values of flux and magnetomotive force. Needless to say, the new magnets designed to give those values would not fit into the magnetic circuit of the original assembly and, were the original circuit adapted to



[1]) For the design of permanent magnet circuits, see for example A. Th. van Urk, Philips tech. Rev. **5**, 29-35, 1940.

the new magnets, the complication of a different stray field would arise (see the article cited in footnote [1]). If we leave this correction aside, however, the photograph gives a striking picture of the progress that the development of new materials has rendered possible, and also of the effect of their development on the proportions of magnets. Some details of the materials used for the magnets in the photograph will now be given (see also the table below).

On the extreme left is a magnet made of the oldest kind of magnetic steel, namely carbon steel having a $(BH)_{max}$ of 0.26×10^6 gauss.oersted [2]). Next to it is a magnet of tungsten steel (almost identical with chrome steel, both being developed before the first world war) having a $(BH)_{max}$ of 0.36×10^6. In 1917 Honda steel appeared; it contained cobalt and had a much higher $(BH)_{max}$ value, viz. 0.9×10^6 for the magnet shown, which is of steel with 35% cobalt. This steel differed to such an extent from the types then known, both in the attainable magnet dimensions and in price (cobalt is about one hundred times dearer than iron), that years passed before it came into general use; the then youthful radio industry, in which stronger yet smaller and lighter magnets were in demand for loudspeakers, was largely responsible for its introduction. The fourth magnet in the photograph is of Mishima alloy, a magnetic alloy based on nickel-aluminium-iron, which was developed in Japan in 1933 and which has the somewhat higher $(BH)_{max}$ of 1.05×10^6. The price of such a magnet is about a third of that of an equivalent cobalt-steel magnet. Experiments on the addition of other metals to the alloy, mainly cobalt, titanium and copper, and on strictly controlled heat treatment, were successful in further improving the properties of Mishima alloy. In this way "Ticonal" II was developed, this having a $(BH)_{max}$ of 1.8×10^6; it was taken into production about 1936.

In 1938 an important step in the evolution of magnet steels was made in the Philips laboratories, when Jonas introduced heat treatment in a magnetic field [3]) for certain well-defined groups of

Table. Further data on the materials from which the permanent magnets shown in the photograph were prepared. Magnet 7a does not appear in the photo, but its material is discussed in the text.

No. in photo	Year	Designation	Magnetic properties					Dimensions of magnets		
					At optimum working point					
			B_{rem} (gauss)	$H_{coërc}$ (oersted)	B (gauss)	H (oersted)	$(BH)_{max}$ (10^6 gauss × oersted)	Diameter (mm)	Length (mm)	Weight (gm)
1	1880	Carbon steel	10000	50	5400	48	0.26	19	136	294
2	1900	Tungsten steel	11000	70	6000	56	0.34	18	115	225
3	1917	Cobalt steel 35% (Honda)	9200	240	6000	150	0.90	18	43	83
4	1938	Ni-Al alloy (Mishima)	6100	480	4050	260	1.05	22	25	69
5	1936	"Ticonal" II	6300	780	5300	340	1.80	19	19	37
6	1937	"Ticonal" G (anisotropic)	13000	580	10000	500	5	14	13	14.5
7	1949	"Ticonal" GG *) (orient.cryst.struct.)	14500	720	13000	640	8.3	12	10	8.5
7a	1953	"Ticonal" X (great coercivity)	9000	1250	5000	900	4.5	19.5	7.2	16
8	1956	"Ticonal" XX *)	11800	1315	9200	1200	11.0	14.7	6	7
9	1952	Ferroxdure I	2000	1800	1000	900	0.9	44	7.2	53
10	1954	Ferroxdure II	3750	1300	2640	1100	2.9	27	6	17.2

*) Material prepared on laboratory scale.

[2]) The values of the $(BH)_{max}$ product are given in gauss oersted in accordance with established practice, rather than in rationalized Giorgi units commonly used in this Review. In the rest of the article units will be omitted when $(BH)_{max}$ values are quoted.

[3]) Netherlands Patent No. 53734; B. Jonas and H. J. Meerkamp van Embden, Philips tech. Rev. **6**, 8-11, 1941. A short time previously Oliver and Shedden (Nature **142**, 209, 1938) had carried out similar tests on magnetic alloys already in use, producing a very slight improvement in their properties. Owing to an incorrect interpretation of the phenomenon, these investigators did not at the time continue their observations.

alloys containing nickel, cobalt, aluminium, copper and iron (and in certain cases other elements such as titanium). In this way the $(BH)_{max}$ value was raised to over 5×10^6 (in "Ticonal" G, the sixth magnet from the left). So important was this step that, even today, about a half of the world production of magnets is carried out according to the process.

In the years following 1933 much work was devoted to the problems of magnetism all over the world. Understanding of magnetic phenomena was deepened by investigators such as Néel, Snoek, Kersten and many others, and various new materials were discovered, such as the iron-cobalt-molybdenum alloys of Sykes and Köster, the cunico and cunife of Dannöhl et al., the cobalt-platinum alloys of Jellinghaus, the "Vicalloy" of Nesbitt et al., the powdered magnetic materials of Néel and the "Bismanol" (bismuth-manganese alloy) of Guillaud and the Naval Ordnance Laboratory in Washington. Considerations of cost price or practical difficulties, or both, prevented these materials being applied except in rare cases, however.

It was not until 1949 that $(BH)_{max}$ was further stepped up. The normal "Ticonal" alloy was so dified from the melt in a special manner so t'at the crystals of the material obtained had a certain orientation. By subjecting magnets of this kind to the heat treatment previously devised [4]), very high values of $(BH)_{max}$, rising to 8.4×10^6, were attained in the laboratory. It was found possible to produce these materials on a larger scale with $(BH)_{max}$ values of 6.5 to 7.5×10^6.

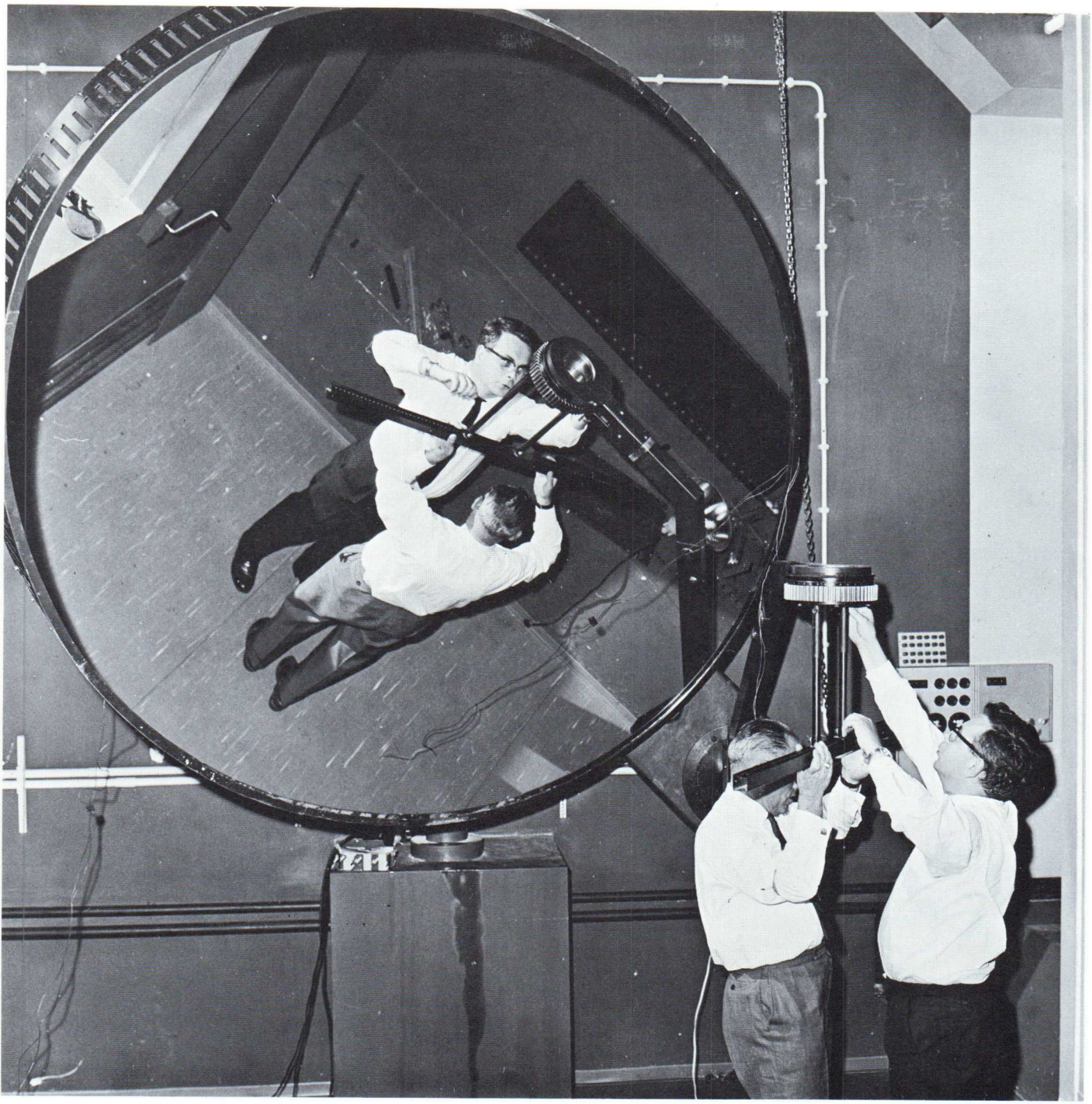
In the meantime efforts have continued to raise $(BH)_{max}$ still higher. In recent years the record has been held by the American General Electric Co. with a cobalt-platinum alloy. The alloy had been described by Jellinghaus much earlier; however, in 1955, by subjecting it to a heat treatment which ordered the crystal lattice, Martin and Geisler [5]) raised its $(BH)_{max}$ value to 9×10^6. Its price naturally prohibits the extensive employment of this material.

The record has now been broken again in the Philips laboratories at Eindhoven. Earlier, a method had been discovered of increasing the coercive force of "Ticonal" alloys. The method consists of the addition of titanium and the raising of the cobalt content [6]), combined with a special kind of heat treatment [7]). It is possible thus to prepare materials ("Ticonal" X) with a coercivity H_{coerc} of 1300 Oe and a $(BH)_{max}$ of 4.5 to 5×10^6. In the course of attempts to improve this material still further by solidifying it in the desired orientation, Luteijn and De Vos succeeded a short time ago in obtaining, in the laboratory, magnets having crystal orientation with a $(BH)_{max}$ of 11×10^6 [8]); for this purpose they worked with very pure constituents, fusing them in an inert atmosphere. In the row of magnets shown in our photograph, the one on the extreme right, the smallest of all, is made of this material.

In addition, the photograph shows two permanent magnets made of ferroxdure, a ceramic (non-metallic) material; these two have the same flux and same magnetomotive force, and hence the same magnetic energy, as the other magnets appearing in the picture. Although the development of ferroxdure [9]) was governed by quite other considerations than the attainment of extremely high $(BH)_{max}$ values, it is instructive to include these two magnets in the series; the relatively low remanence but extraordinarily high coercive force of ferroxdure result in the rod form of the magnet we started with becoming a disc. The $(BH)_{max}$ of ferroxdure I lies in the vicinity of 0.9×10^6; in the meantime this value has been raised to above 3×10^6 (maximum 3.8×10^6) for ferroxdure II, this having been achieved by aligning the particles of the basic material in a magnetic field prior to sintering. It may be remarked that both the great coercive force of this material, and its non-metallic character, has resulted in a wide range of special applications.

H. J. MEERKAMP van EMBDEN.

[4]) Netherlands Patent No. 71925.
[5]) D. L. Martin and A. H. Geisler, J. appl. Phys. **24**, 498, 1953.
[6]) Netherlands Patent Nos. 70092 and 70424.
[7]) Netherlands Patent applied for.
[8]) A. I. Luteijn and K. J. de Vos, Philips Res. Rep. **11**, 489-490, 1956.
[9]) J. J. Went, G. W. Rathenau, E. W. Gorter and G. W. van Oosterhout, Philips tech. Rev. **13**, 194-208, 1951/52; A. L. Stuijts, G. W. Rathenau and G. H. Weber, Philips tech. Rev. **16**, 141-147, 1954/55.



Rotating mirror for examining lighting fittings

This rotatable mirror is used in the lighting engineering department of the Aachen laboratory of Philips Zentrallaboratorium GmbH for measuring the distribution of light from large lighting fittings. The principles of the measuring arrangement used here go back as far as 1952 (see L. de Wit, Philips tech. Rev. **14**, 200, 1952/53). The photoelectric cell which measures the light is placed in a fixed position at a considerable distance. The lighting fitting whose intensity in different directions is being measured always remains in the normal operating position, although it is rotated (together with the mirror) for scanning all possible directions of light emission.

(See: Philips tech. Rev. **26**, 98, 1965.)

A RADAR SONDE SYSTEM FOR UPPER AIR MEASUREMENTS

621.396.969: 551.508.8

The measurement of upper air conditions is an important branch of meteorology because information is required for research, for weather forecasting and for aircraft operation. Although the first soundings were made in the 18th century, using kites and manned balloons, the systematic measurement of air conditions at great heights has become possible only since the development of radio and radar. The development of radio and radar sonde equipment is therefore one of the most significant advances in meteorological instrumentation.

In the following note, a short description will be given of a radar sonde system which has been developed by the Mullard Research Laboratories, Salfords (England), in conjunction with the Royal Radar Establishment, Malvern, for measurements of wind speed, wind direction, temperature, pressure and humidity at heights up to at least 24 km (80 000 ft) [1]). The equipment is automatic in operation and provides continuous records at a ground station of the wind vector as a function of height. Temperature, pressure and humidity are measured and recorded at the ground station at approximately 100 m intervals of the height of the sonde.

For each sounding, an airborne unit (which in a large percentage of soundings is not recovered) is carried into the upper atmosphere by a free, hydrogen-filled balloon. This unit (*fig. 1*) contains a small combined transmitter and receiver with auxiliary equipment and instruments for measuring temperature, pressure and humidity. The transmitter allows the telemetering of these data during the whole ascent, until the bursting of the balloon, which may occur after it has reached a height greater than 80 000 feet and travelled a distance of over 100

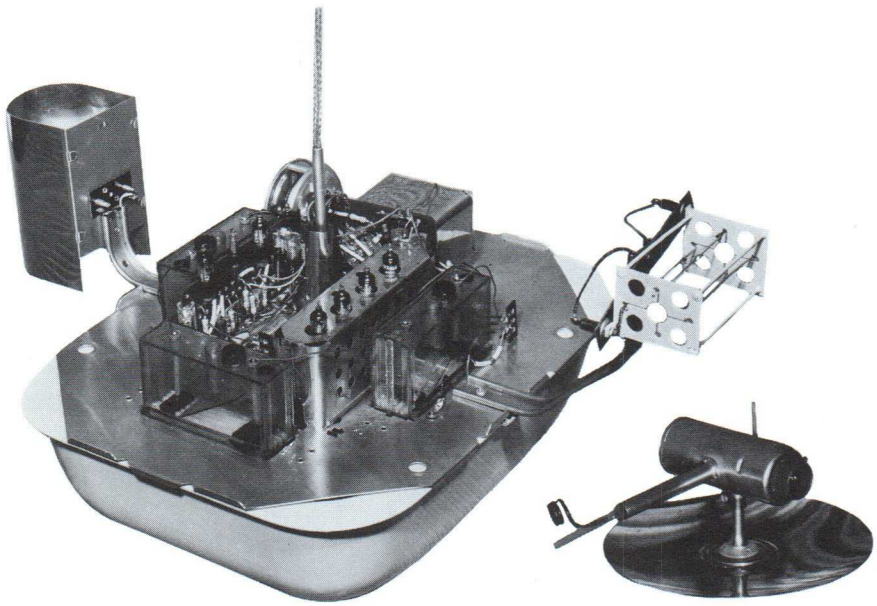


Fig. 1. Airborne unit of the radar sonde (cover removed). The humidity and temperature measuring elements are carried on external arms; the main part of the unit is enclosed in a thermally insulating case. The aerial, receiver, telemetering circuit, switch, barometer and batteries are seen mounted on the chassis plate. The 10 cm wave transmitter and aerial assembly, which are mounted underneath the chassis plate, are shown separately on the right-hand side of the photograph.
The airborne unit is suspended from a hydrogen-filled balloon which is approximately 2 m in diameter at the ground and lifts the unit at 360 m per minute (1200 ft/min). A parachute is inserted in the suspension to allow a gentle descent after the balloon has burst (at a height greater than 80 000 feet).

[1]) See also: F. E. Jones, J. E. N. Hooper and N. L. Alder, The radar sonde system for the measurement of upper wind and air data, Proc. Instn. El. Engrs. **98**, II, 461-469, 1951.
A. L. Maidens, Sounding the upper atmosphere, Discovery **18**, 156-160, 1957 (No. 4).

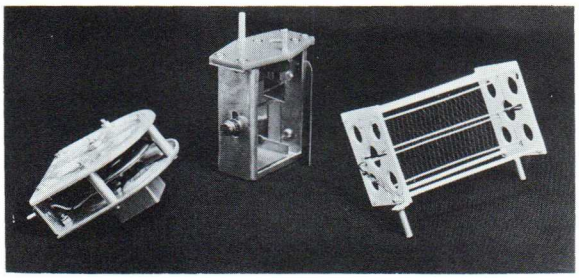


Fig. 2. *Left*: the aneroid barometer; *right*: the fine-wire resistance thermometer; *centre*: the skin hygrometer. The aneroid and hygrometer are conventional elements, but the thermometer is a considerable improvement over earlier types. It is constructed with fine, coiled-coil, tungsten wire, has a very short time constant and a very small radiation error at high altitude.

miles from the release point. Tracking the flight of the balloon by radar at the ground station provides the data necessary for determining the wind vector and height at each moment. Conventional radar methods, which depend on a pulse signal transmitted by the ground station being *reflected* by the target, would involve a very high transmitting power in order to enable the balloon to be tracked during its whole flight. Therefore in this radar sonde a system known as "secondary radar" has been employed. The airborne transmitter is made to send a pulse (or in fact 2 pulses, see later) every time the airborne receiver picks up a pulse signal transmitted by the ground station. The time delay measured at the ground station between the departure of the "interrogating pulse" and the arrival of the "responding pulse" is a direct measure of the range of the balloon; the azimuth and elevation angles of the transmitting aerial, which is automatically kept aligned on the balloon, convey the information concerning its direction.

The interrogating signal has a frequency of 152.5 Mc/s and is transmitted at a peak power of 50 kW with a pulse repetition rate of 404 pulses per second. The responding signal of the airborne transmitter has a frequency of 2850 Mc/s and is radiated at a peak power of 0.03 kW.

For the sensing of pressure, humidity and temperature, the airborne unit carries an aneroid barometer, a goldbeater's skin hygrometer and a resistance thermometer (*fig. 2*). The changes in these elements under varying atmospheric conditions are converted into voltage variations which, in turn, control an "encoding" circuit. This circuit is triggered by each of the interrogating pulses, derived from the receiver, and generates a pulse of its own with a time delay dependent on the control voltage. The airborne transmitter is operated both by this delayed pulse and by the interrogating pulse and thus transmits back to the ground a pair of pulses for each interrogating pulse, the time delay between the two pulses of a pair representing the reading of the meteorological instrument. The control voltages of the three instruments are brought into operation in sequence by a motor-driven switch, so that each instrument reading is encoded once every 17 seconds.

The airborne centimetric transmitter contains a triode transmitting valve, controlled by a blocking-oscillator modulator and mounted in a coaxial cavity designed to facilitate large-scale production. The transmitter aerial is an unipole with counterpoise (fig. 1). The power supply of the sonde comprises three primary cells and a vibrator.

The ground equipment is divisible functionally and physically into two groups. The *radar* group is concerned with the interrogation of the airborne unit, reception of the return signal and determination of the basic positional data (azimuth angle, elevation angle and slant range of the balloon as a function of time). The *computer* group continuously translates the positional data into wind speed and direction, decodes the telemetering signal and records all the results in graphical and printed form.

The radar equipment is illustrated in *figs 3* and *4*. The transmitting aerial (a Yagi array) and the

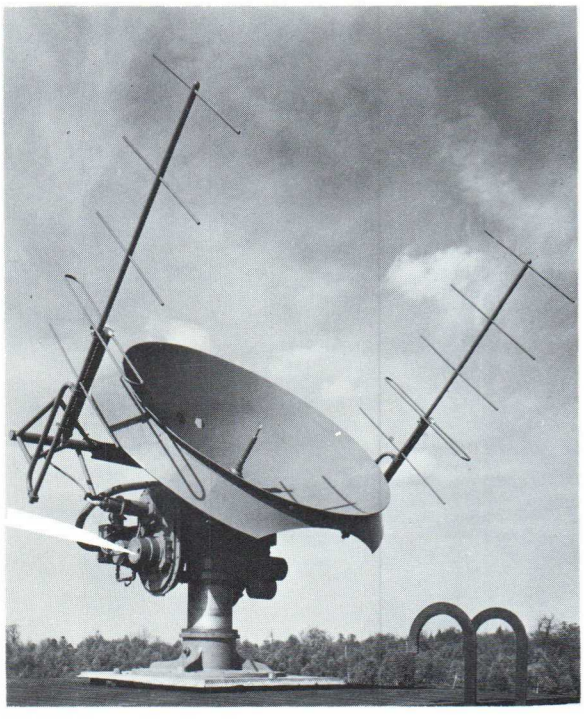


Fig. 3. The Yagi array transmitting the 50 kW interrogating pulses of 2 m waves is mounted on a common pedestal with the receiving aerial, a 5 ft diameter parabolic reflector with nutating dipole. They are automatically kept aligned on the balloon.

receiving aerial (a parabolic reflector) are mounted on a common pedestal driven by a motor servosystem. The receiving aerial is fitted with a nutating dipole which produces a conical scan of the narrow aerial beam and superimposes a modulation on the received signal when the aerial is incorrectly aligned. This error signal steers the servosystem so as to give accurate automatic following of the balloon in direction. The microwave receiver, which is equipped with automatic gain and frequency control, amplifies the signal from the balloon and delivers it to the automatic ranging system and the telemetering system. The ranging system is of a novel type in which the rotation of a motor-driven, phase-shifting transformer is related to the change in distance of the balloon, so that the speed of rotation is an accurate measure of the radial velocity of the balloon (radial wind component). A rotation of 360° is equivalent to a change in distance of one tenth of a nautical mile. The phase-shifting transformer, therefore, permits fractions of a tenth of a mile to be measured; full tenths of a mile are determined by counting the number of oscillations of a crystal calibrator occurring between transmitted and responding pulse. The calibrator controls the pulse repetition frequency of the transmitter and oscillates with a time period exactly corresponding to one tenth of a nautical mile in radar range. The received signal is monitored and readings of range R, azimuth Θ and elevation E are provided by a display unit (fig. 4). This unit is also fitted with the manual controls for range and aerial position which are used at the beginning of a flight before the equipment is finally switched over to fully automatic operation.

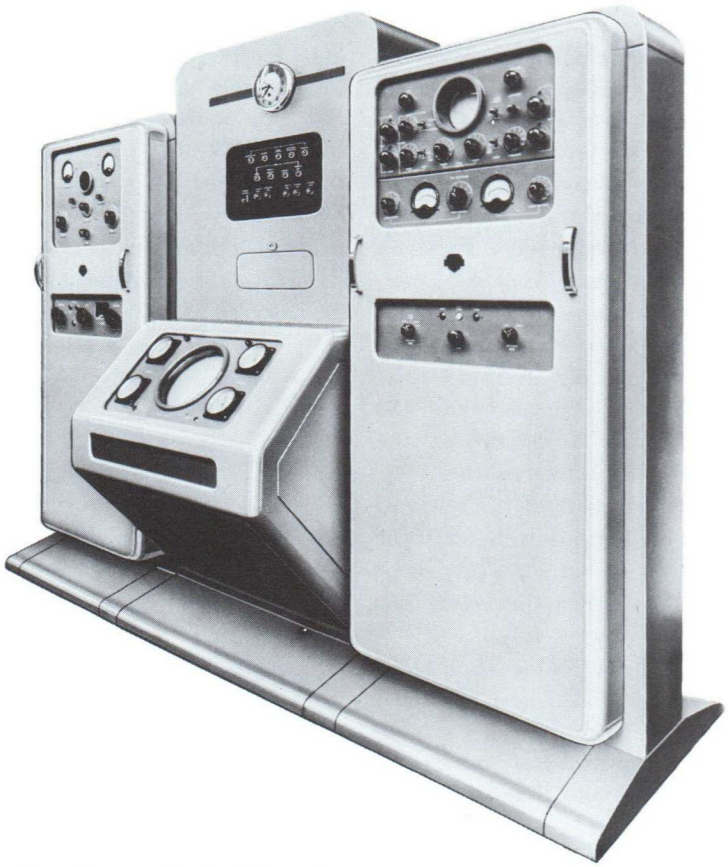


Fig. 4. The radar console. The left-hand unit contains the 50 kW transmitter, the right-hand unit the 10 cm wave receiver and the automatic ranging and aerial-alignment equipment. The centre unit is the display and control system.

350

The continuous computation of the wind vector (which is parallel to the surface of the earth) is effected from the tangential and radial wind components, given by the equations (in spherical polar coordinates):

$$V_T = R\,(\mathrm{d}\Theta/\mathrm{d}t)\cos E,$$

$$V_R = (\mathrm{d}R/\mathrm{d}t)\cos E - R(\mathrm{d}E/\mathrm{d}t)\sin E.$$

An important feature of these equations is the occurrence of the derivatives of (slant) range R, azimuth Θ and elevation E with respect to time t, i.e. the radial and angular velocities. The auto-ranging and auto-alignment circuits are so designed as to provide signals directly proportional to these velocities. The wind computer which determines the vector resultant of V_T and V_R is illustrated in *fig. 5*. It is an analogue computer employing precision potentiometers for multiplication and magslips (synchros) for trigonometrical computation. The vector amplitude and direction are derived from V_T and V_R in a "triangle-solver" magslip whose rotor shaft assumes a position corresponding at every moment to the wind direction. A 4 ft diameter recording table is directly coupled to this shaft and a recording pen is driven at a constant speed in a radial direction across the table. The pen thus traces out a continuous graph of instantaneous wind direction against time. A voltage proportional to the vector amplitude, i.e. the wind speed, is induced in the magslip rotor coil and is recorded by a conventional recording meter. The height of the balloon, corrected for earth curvature, is computed by similar analogue methods and recorded as a function of time.



Fig. 5. The wind computer. A magslip turns the rotatable 4 ft diameter table so that at every instant it has a position corresponding to the computed wind direction. A recording pen moving at constant speed along the fixed radial arm traces out a graph of wind direction as a function of time (cf. fig. 7).

The ground telemetering equipment, shown in *fig. 6*, is designed to identify the coded signals from the balloon, to measure accurately the time delay between the ranging and telemetering pulses and to encode the results in a form suitable for operating a standard teleprinter. The time delay is measured by allowing crystal-controlled timing pulses, with a one microsecond separation, to pass through a gating circuit during a period initiated by the ranging pulse and terminated by the telemetering pulse. The total number of timing pulses corresponding to five hundred pairs of signal pulses is counted by an electronic counter and the average delay is transferred to the teleprinter encoder. The figures are printed by the teleprinter in five columns, each of four digits, corresponding to the three meteorological parameters and two reference signals which are also transmitted by the airborne unit. Coding sig-



Fig. 6. The telemetering console. The left-hand unit contains wind-speed and height recorders, the right-hand unit encoders and temperature, pressure and humidity recorders and the centre unit a digital computer and teleprinter.

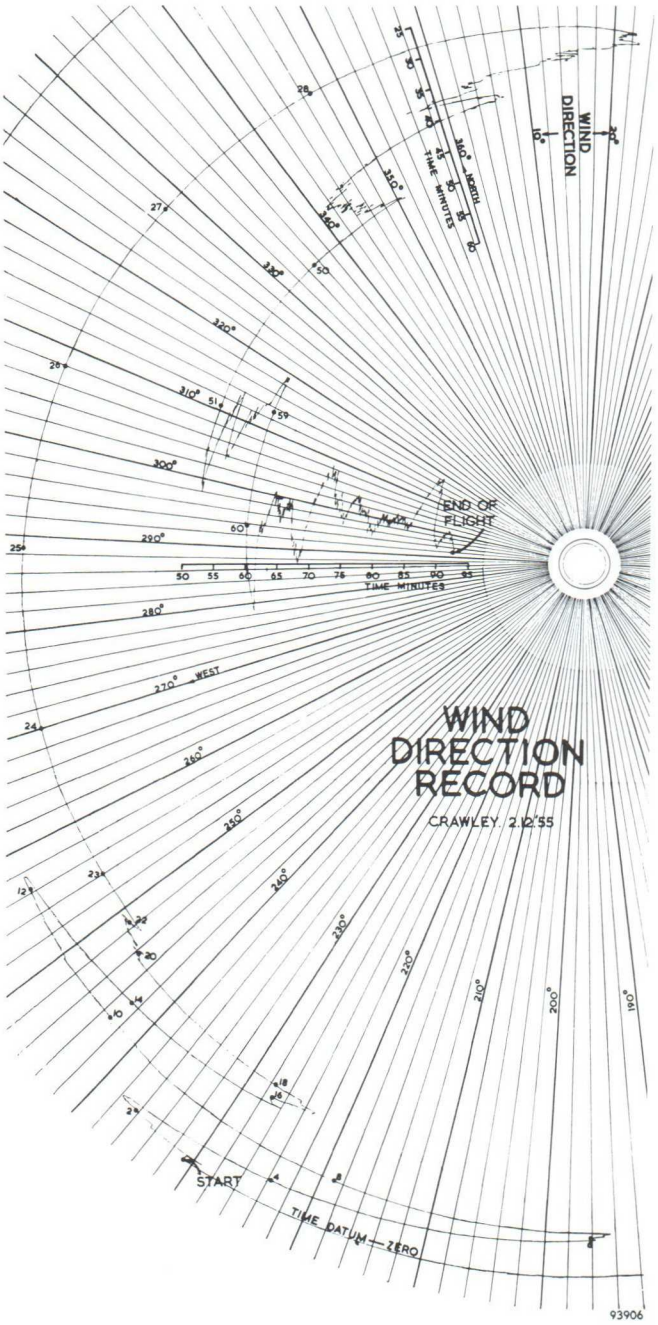


Fig. 7

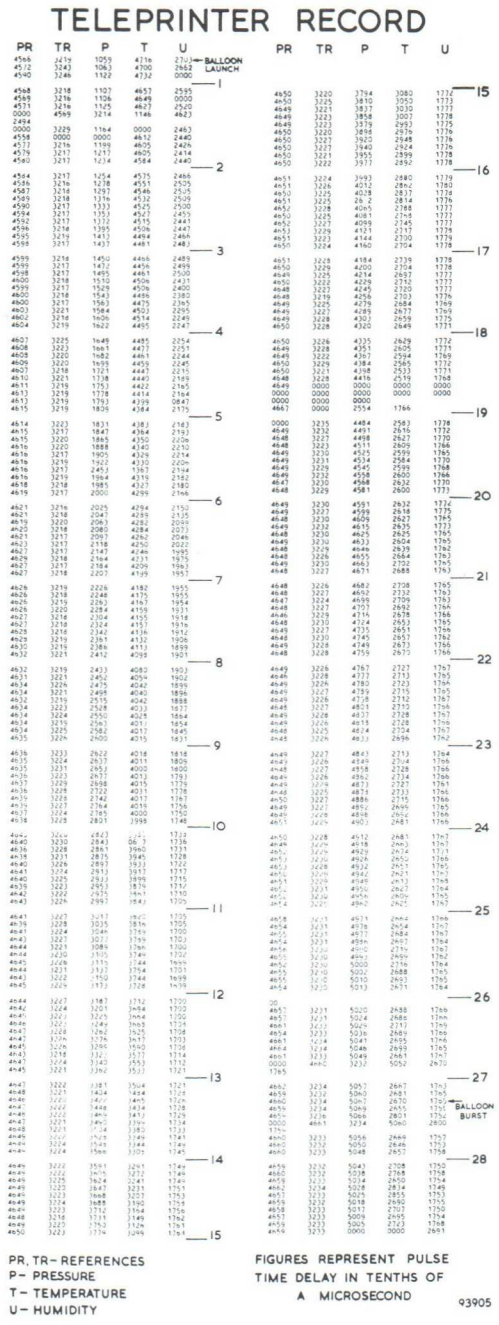


Fig. 8

Fig. 7. A typical wind direction record obtained from the radar sonde system.
Fig. 8. Typical teleprinter record obtained from the radar sonde system.

nals interposed in the airborne unit switching sequence are used to ensure that the results are printed in the correct columns.

Some typical flight records, which are self-explanatory, are shown in *figures 7, 8* and *9*. The equipment is designed to measure wind speed with an error not exceeding 5 km/hr. The telemetering channels are designed for an error of the order 0.1%, but at present the meteorological instruments themselves contribute somewhat larger errors.

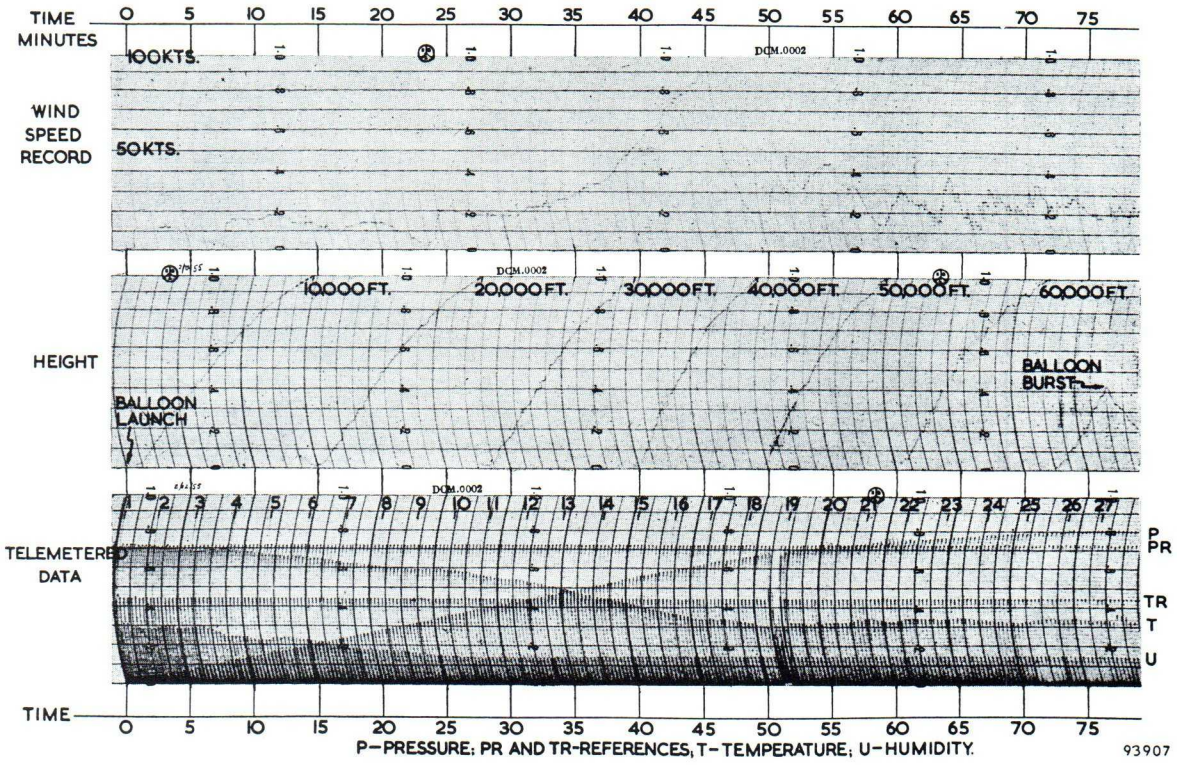


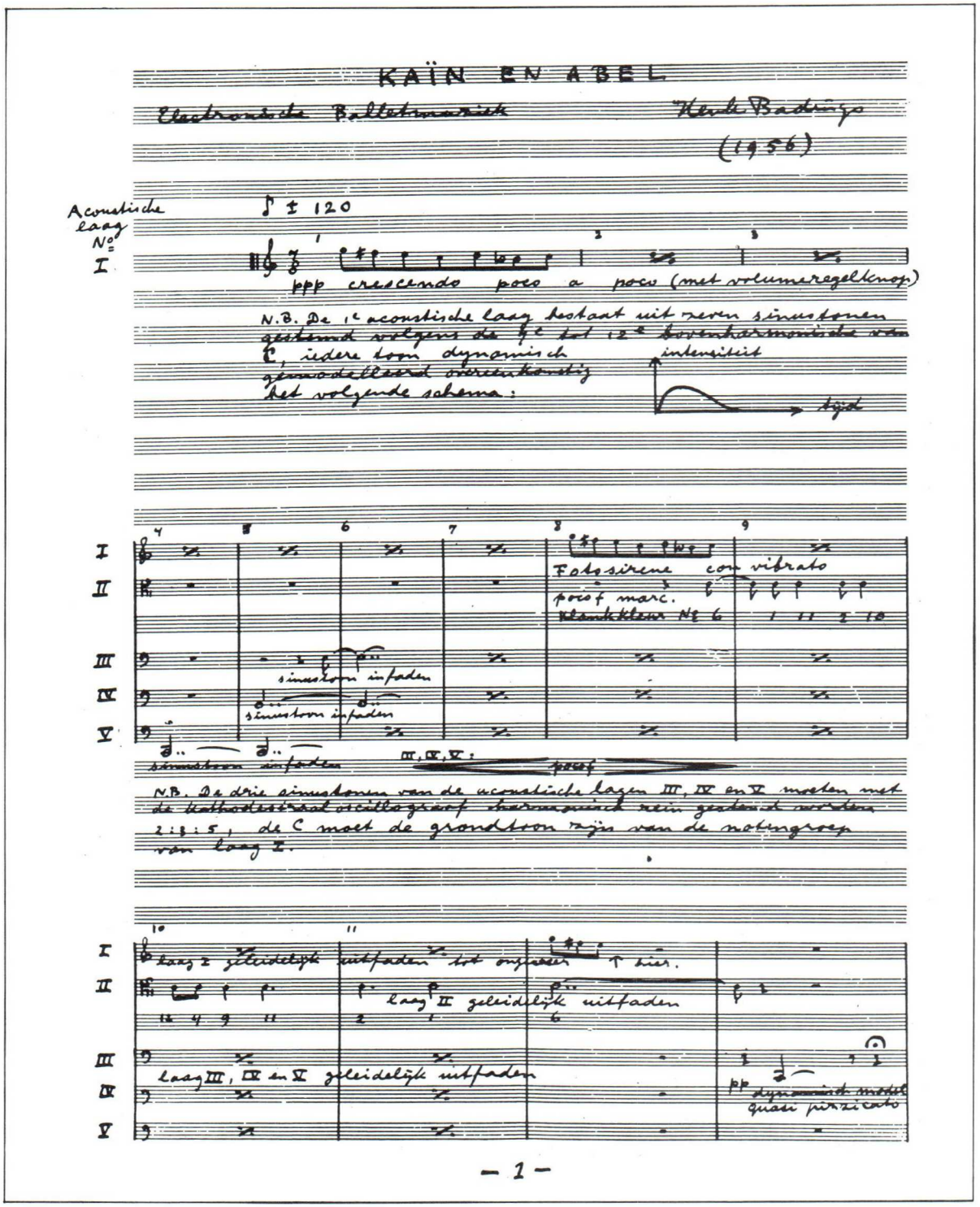
Fig. 9. Example of wind-speed, height and telemetering records obtained from the radar sonde system.

The prototype equipment is at present undergoing flight trials to assess its suitability for routine soundings and research.

The authors are indebted to the Director of the Mullard Research Laboratories, the Ministry of Supply and the Meteorological Office for permission to publish this note.

N. E. GODDARD*) and H. A. DELL*).

*) Mullard Research Laboratories, Salfords (England).



Electronic music

This is page 1 of the score of the electronic ballet music "Cain and Abel", which the Dutch composer H. Badings wrote and realized using equipment of an electronic studio in Philips Research Laboratories, Eindhoven. It is seen that the sound is composed of five "acoustic layers" (each of which is recorded on magnetic tape and treated separately). The composer must give many explanations and precise descriptions in the score in order to define the operation of unorthodox sound sources and the treatment of the recorded tape.

(See: H.Badings and J.W. de Bruyn, Electronic music, Philips tech. Rev. **19**, 191-201, 1957/58. The article was supplemented by a gramophone record which among other things contained a shortened version of "Cain and Abel". — We regret that we are now out of stock of this record.)

A klystron multiplier for generating 0.8 mm waves

B. B. van Iperen and W. Kuypers

621.385.623.5

At the beginning of the 'fifties, when the most vigorous period in the development of microwave tubes was over, the view was generally held that it would be difficult, if not impossible, to generate wavelengths shorter than a few millimetres by means of the principles employed at that time. In most research centres concerned with the problem of generating shorter waves, other ways and means of doing this were therefore sought and studied.

In recent years, however, the "classical" principles have again been gaining ground. Thanks to improved theoretical insight and new technological possibilities, it has proved feasible to use the same principles for generating submillimetre waves. An example of this new development is discussed in this article. Here a tube is described that is capable of continuously generating waves of 0.8 mm with a power output of some tens of mW.

Introduction

One of the most widely employed sources for generating cm and mm waves is the reflex klystron. Tubes of this type are simple to use and deliver a continuous output which is amply sufficient for many purposes [1]. It has not previously been possible, however, to make continuous-wave reflex klystrons for the submillimetre region. The shortest wavelength achieved with these tubes is about 2 mm. The main reason for this limitation is that the reflex klystron is itself an oscillator and must therefore satisfy the starting condition for oscillation [2]. According to this condition the sum of electronic admittance and resonator admittance must be negative. Since the resonance admittance of the cavity resonator is inversely proportional to the square root of the wavelength, it becomes increasingly difficult at shorter wavelengths to satisfy the oscillation condition. In addition, there are also other serious problems of a less fundamental nature, such as the extremely small dimensions dictated by the wavelength, and the cathode loading and heat dissipation. It does not seem likely, therefore, that it will be possible in the foreseeable future to generate wavelengths of less than 1 mm with the reflex klystron.

In recent years research has been undertaken at the Philips Research Laboratories, Eindhoven, into the possibility of generating microwaves with the aid of a two-cavity klystron that does not work as an oscillator but as a frequency multiplier. This method has the great advantage that it is *not* necessary to satisfy the oscillation condition. The designer of the tube then has more freedom in the choice of certain parameters, so that other problems can be solved more easily than in the case of a reflex klystron.

The method is based on the fact that the velocity-modulated electron beam of a klystron can contain in addition to the fundamental frequency of the driving signal a large number of its harmonics. If the driving signal and the transit time between the two resonant cavities is suitably chosen, these higher harmonics can be relatively strong [3]. According to a simple theory, which disregards the interaction forces between the electrons, the amplitude of the tenth harmonic in the beam current can even be greater than half the magnitude of the first harmonic. In reality the situation is even more favourable. If the output cavity (catcher) is tuned to the tenth harmonic of the driving frequency,

Drs. B. B. van Iperen and Ir. W. Kuypers are research workers at Philips Research Laboratories, Eindhoven.

[1] B. B. van Iperen, Reflex klystrons for wavelengths of 4 and 2.5 mm, Philips tech. Rev. **21**, 221-228, 1959/60.
[2] See the article in reference [1], p. 225 et seq. (eq. (3)).
[3] This was first noted by D. L. Webster, J. appl. Phys. **10**, 501-508, 1939.

the beam delivers energy at ten times the frequency of the driving signal (*fig. 1*).

Some years ago tubes working on this principle were described, which were suitable for the generation of cm- and mm-waves [4]; the shortest wavelength reached was 5 mm, with an output power of 0.1 mW using the second harmonic. We have found, however, that submillimetre

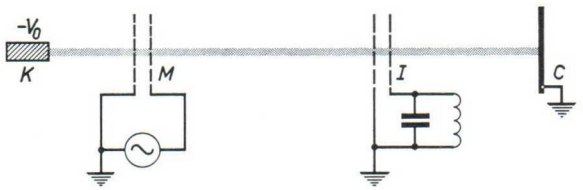


Fig. 1. Principle of klystron operation. K cathode. V_0 electron accelerating voltage (beam voltage). M two modulation (buncher) grids, between which the electron beam is velocity-modulated as a result of the alternating voltage — the driving signal — applied to M. In the drift space between M and I the velocity-modulation gradually gives rise to density modulation, because the electrons accelerated in M catch up with the earlier electrons that have been slowed down: the electron beam, which was a "direct current" at M, has become at the induction grids I an "alternating current" with the frequency of the modulating signal. In certain circumstances this alternating current contains many higher harmonics. The tube can be made to work as a *frequency-multiplier* by tuning the resonant circuit connected to I to the frequency of one of these higher harmonics. The tenth harmonic is the one used in the 0.8 mm tube. In reality, M and I are not pairs of grids with separate resonant circuits connected to them, but gridless interaction gaps in resonant cavities.

waves can also be generated with a klystron multiplier provided the velocity of the electrons is drastically increased [5]. In the tube described in this article, the beam voltage V_0 — 2 to 2.5 kV in reflex klystrons — is not less than 25 kV. This tube generates 0.87 mm waves having a continuous output power of 35 mW [6].

In the following section we shall explain the significance of raising the beam voltage, and say something about the factors that govern the optimum value of the beam current. In conclusion we shall discuss the construction and the principal characteristics of the new 0.8 mm tube.

Theoretical considerations

As stated, the greatest obstacle to the construction of tubes for very short wavelengths can be overcome by using a frequency multiplier instead of an oscillator. At the same time this makes the above-mentioned increase in the beam voltage V_0 possible. This considerably reduces the other difficulties, such as the fact that at very short waves the dimensions of the resonant cavities and their distance apart become extremely small, the limitations in regard to the permissible loading of the cathode, etc.

By increasing V_0 one increases what is termed the electronic wavelength, i.e. the distance travelled by the electrons in one cycle of the driving or output signal; this distance is proportional to $\sqrt{V_0}$. We will now explain the advantages to be gained by increasing the electronic wavelength.

a) If V_0 is increased the electron beam can be given a larger diameter, thus reducing the cathode loading. This is immediately apparent from the following consideration. To ensure good interaction between beam and resonant cavity the transit time of an electron in the alternating electric field of the interaction gap (see *fig. 2a*) must not be much longer than half a cycle of this field. So in the first place, the gap width d must not be much larger than half an electronic wavelength. This has not yet met the above-mentioned requirement, however, for even if d is relatively small the electric field extends along the axis of the resonant cavity over a length which is of the same order of magnitude as the diameter $2r$ of the interaction space. For good interaction, therefore, this diameter too should not be much greater than half an electronic wavelength. The same consequently applies to the beam diameter. (In our tube the beam diameter is governed by the electronic wavelength related to the *output signal*, this wavelength being much smaller than the one related to the input signal.)

b) The output power increases in proportion to V_0. This is bound up with the characteristics of the catcher cavity. At wavelengths below about 1 mm there are disadvantages in using resonant cavities as drawn in fig. 2*a*. In the first place, the re-entrant parts are then difficult to make because of their small dimensions. Moreover, there is a danger that these parts will be overheated by electron bombardement. We therefore

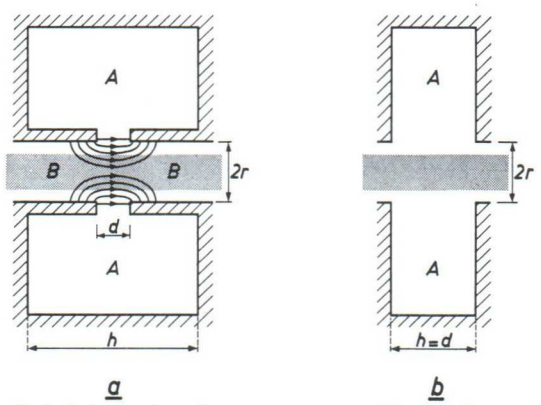


Fig. 2. *a*) Cross-section of a resonant cavity of the type frequently used in klystrons (re-entrant cavity), showing an electron beam (the shaded strip) in interaction with it. The beam passes the resonant cavity A through the hole B, in the wall of which a slit of width d is cut. The part of the hole near the slit, where the electric field is relatively strong, is the actual interaction space (see the sketched paths of the electric lines of force). For good interaction between resonant cavity and beam the diameter $2r$ of the interaction space and the gap width d should at the most be roughly equal to half the "electronic wavelength" of the beam. *b*) Example of a flat resonant cavity, as used in the tube for 0.8 mm waves. The letters have the same meaning as in (*a*). The re-entrant parts are missing here, so that the gap width d is equal to the height h. A resonant cavity of this kind can be made for extremely short wavelengths and gives better heat dissipation.

use what is called a "flat cavity" (fig. 2*b*). The resonant frequency of such a cavity is primarily determined by the dimensions perpendicular to the beam. In the first instance the height h can be freely chosen. It is advantageous to choose it as high as possible, because this increases the resonance impedance (which is proportional to h^2) and thus the output power also. The extent to which h can be increased, however, is limited by the fact that h here is equal to the gap width, and must not therefore be greater than about half an electronic wavelength. Thus, in order to obtain a reasonable output power it is equally important to have a high V_0, for as stated the output power is proportional to the resonance impedance and hence to V_0.

c) When the beam voltage is raised, the optimum distance between the centres of the two resonant cavities also increases; at constant input power the increase is more than proportional. By choosing the beam voltage in the region of 25 kV one ensures that there is sufficient space to accommodate resonant cavities of optimum dimensions together with their tuning devices and connections, and also that the heat generated in the cavities is adequately dissipated.

The beam current

The extent to which the electron beam current in a klystron can be increased in order to step up the power output is limited by the influence of mutual repulsion between the electrons. If the density of the beam current is too high, this repulsion tends to oppose electron bunching in the beam and the formation of harmonics is disturbed. Calculation of the influence of this effect is a complicated problem that can only be solved with the aid of an electronic computer, and its treatment does not come within the scope of this article [7].

Our calculations have shown, remarkably enough, that to some extent a higher current density initially favours the production of higher harmonics; the adverse effect referred to only appears later on. Moreover, the optimum beam current is found to depend, among other things, on the rank number of the relevant harmonic (i.e. on the factor by which the frequency is multiplied) and on the driving power; the voltage V_0, on the other hand, has little influence. These calculations show that in our tube the optimum beam current for a driving power of about 10 W is roughly 20 mA.

An experimental tube for 0.8 mm

A simplified cross-section of the tube is shown in *fig. 3*. From left to right can be seen successively the electron gun K, the buncher cavity M with tuning device and input waveguide, the block I containing the catcher cavity and the output waveguide, and the collector C. The catcher cavity itself is too small to be shown in this figure. Also to be seen are the pole pieces P of the electromagnet which encloses the tube

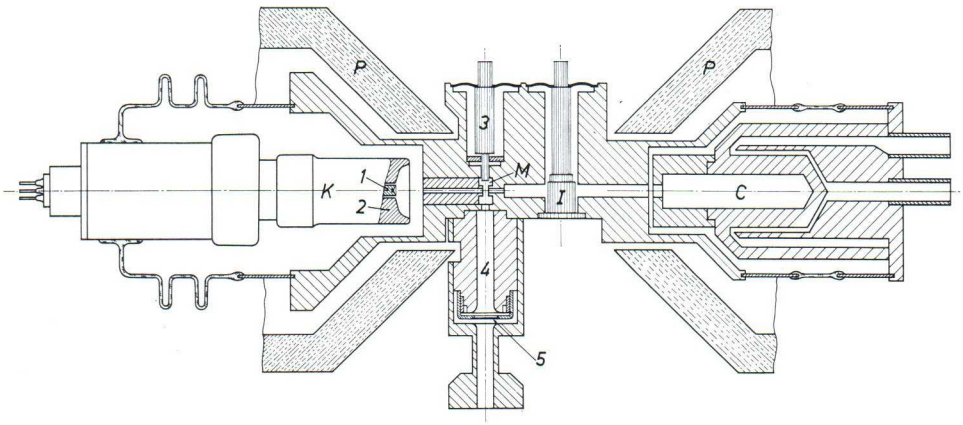


Fig. 3. Simplified cross-section, 1.5 times full size, of the klystron multiplier for 0.87 mm. *K* electron gun with cathode *1* and focussing electrode *2*. *M* modulation cavity (buncher) tuned to 8.7 mm. *3* tuning plunger, movable by means of a diaphragm in the outside wall of the tube. *4* input waveguide with glass vacuum-tight window *5*. *I* the block containing the catcher cavity with the output waveguide and accessories. (A cross-section of the catcher cavity on a larger scale can be seen in fig. 6.) *C* collector (water-cooled). *P* pole pieces of focussing magnet.

[4] See for instance W. H. Cornetet, Jr., IRE Trans. on Electron Devices **ED-6**, 236-241, 1959.

[5] See B. B. van Iperen, Proc. IEEE **51**, 935-937, 1963. In this article a 2.5 mm tube is described in which this increase in beam voltage has been applied for the first time.

[6] See also B. B. van Iperen and W. Kuypers, Experimental CW klystron multiplier for submillimetre waves, Philips Res. Repts **20**, 462-468, 1965 (No. 4).

[7] See B. B. van Iperen and H. J. C. A. Nunnink, Harmonics in velocity-modulated cylindrical electron beams, Philips Res. Repts **20**, 432-461, 1965 (No. 4). See also B. B. van Iperen and H. J. C. A. Nunnink, C.R. 5me Congrès int. tubes hyperfréquences, Paris 1964.

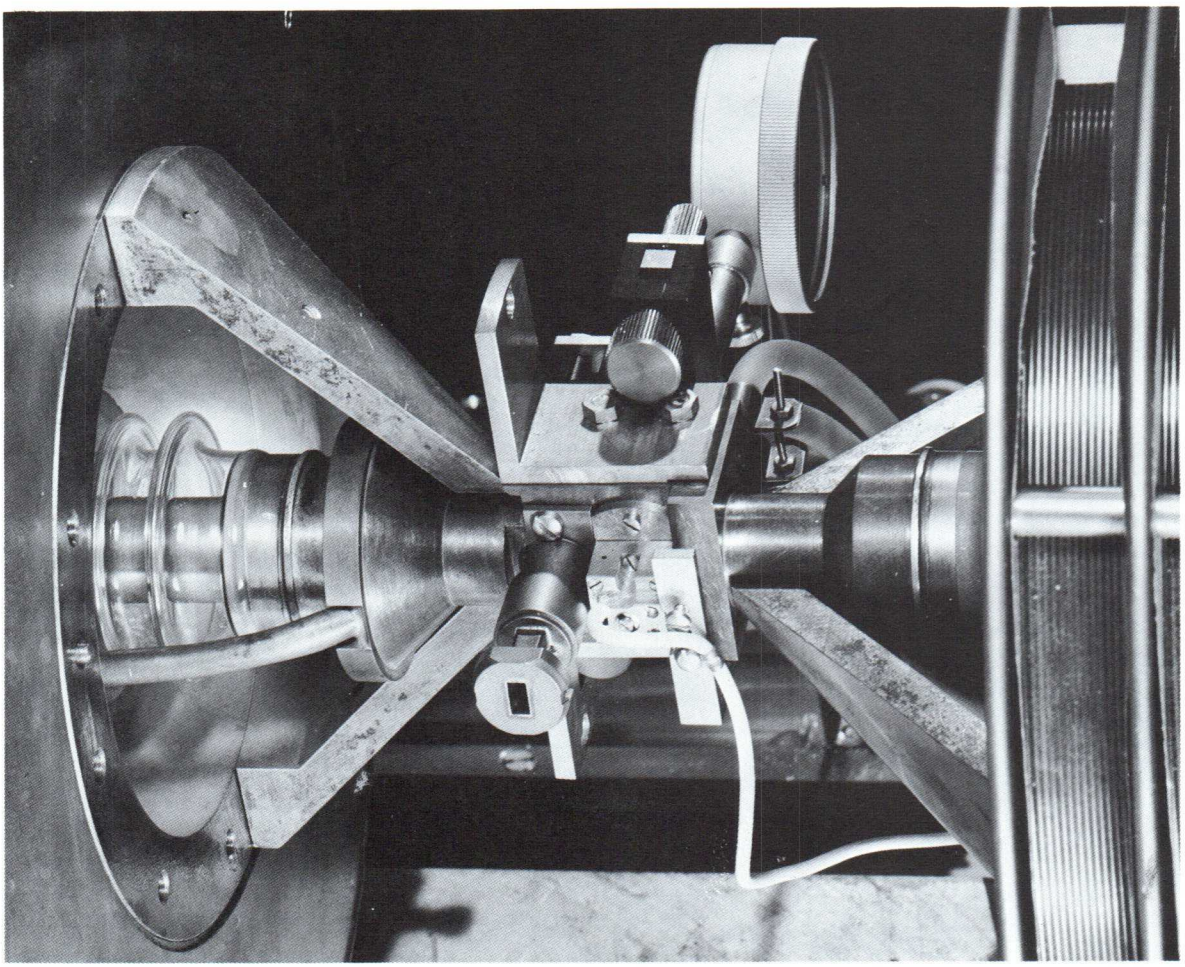


Fig. 4. The 0.8 mm klystron in the magnet. To make the ends of the tube visible, half of each of the tapering pole pieces has been removed; on the left is the gun section. The connection for the input waveguide can be seen at the front in the centre. On the right, partly behind the white cable, is a bolometer employing a thermistor and matching transformer, for measuring the output power. In the centre, above the tube, is the tuning knob for the catcher cavity. In the background is a micrometer for reading the position of the tuning plunger in the catcher cavity.

and which supplies the magnetic field needed for focussing the electron beam. *Fig. 4* shows a photograph of the tube inside the magnet.

The electron gun is a Pierce gun, which gives an electron beam current of 20 mA at 25 kV. The beam diameter is about 0.2 mm. Since the beam converges, the current density at the cathode surface is limited to 7 A/cm^2; this value is well below the permissible maximum for a modern L-cathode. To obtain a well-shaped beam only a circular part of the front face of the cathode is used for emission, the remainder having been made non-emissive by a special treatment [8]. This non-emissive portion thus acts as a focussing electrode at cathode potential.

The driving signal is supplied by an 8 mm klystron. The catcher cavity is a flat resonant cavity of rectangular form, which oscillates in the TE_{012} mode. Because of this shape and resonator mode it is possible to make the catcher cavity tunable. This can be understood as follows.

In the TE_{012} mode the alternating electric field is in the x direction (see *fig. 5*) while the alternating magnetic field in this direction is zero. Both fields are independent of x but vary sinusoidally in the y and z directions, in such a way that a half wave is along the y axis and a full wave along the z axis inside the cavity. The beam aperture is located at a position of maximum electric field, i.e. at a distance $l/4$ from one of the end faces (l being the length of the cavity along the z axis). At a

[8] The process used for this was found and developed by Dr. R. Levi, Philips Metalonics, Mt. Vernon, N.Y., U.S.A. It is described by E. S. Rittner and R. Levi in J. appl. Phys. **33**, 2336, 1962.

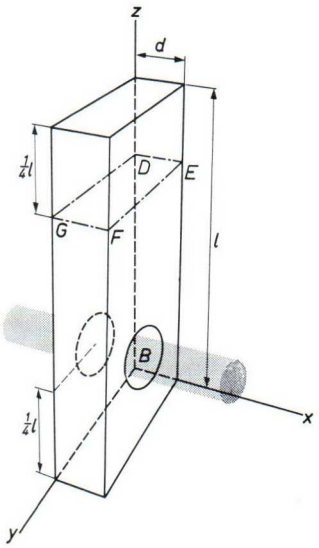


Fig. 5. Rectangular flat cavity, used as the catcher cavity for 0.87 mm waves. The cavity is excited in the TE_{012} mode, the alternating electric field then being in the x direction and the alternating magnetic field in this direction being zero. *B* apertures. The cavity can be cut along the lines *DEFG* without disturbing the resonator mode. This fact is turned to use for making the cavity tunable.

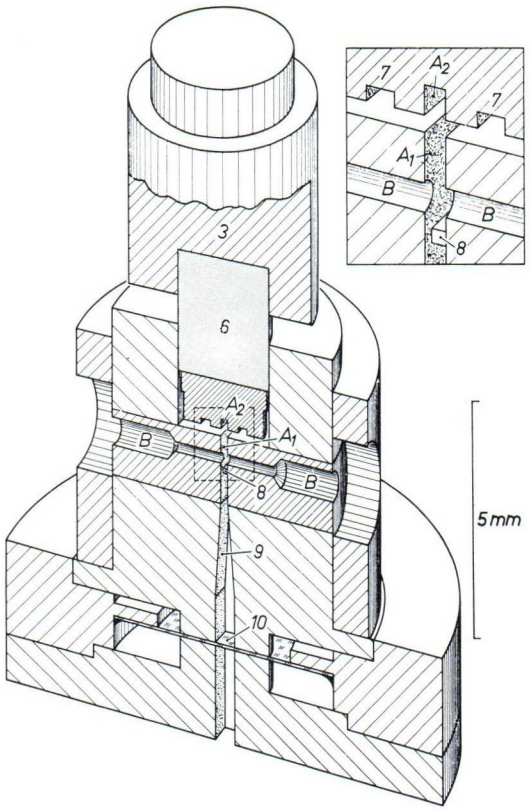


Fig. 6. Perspective sketch of the catcher cavity for 0.87 mm. The cavity has been cut along a plane through the beam axis and parallel with the xz plane (see fig. 5), and the front half removed. A_1 rigid section of the cavity. A_2 movable part of the cavity, which is fixed to the tuning plunger by means of a sapphire rod *6*. The plunger can be moved axially via a diaphragm in the tube wall. *7* choke grooves to prevent leakage of the high-frequency field through the gap between A_1 and A_2. *8* partition between resonant cavity and output waveguide. The coupling hole is between this partition and the one in the cut-away half. *9* tapered output waveguide. *10* mica window, 5 μm thick, for vacuum seal. *B* beam aperture.

distance $l/4$ from the end faces the y component of the magnetic field is zero. Since the current in the wall is at right-angles to the adjacent magnetic field, the z component of the wall current is also zero at that position. The resonant cavity can thus be cut through at that point without upsetting the resonator mode.

The latter fact is turned to use for making the catcher cavity tunable. For this purpose it is composed of two parts, one of which is $3l/4$ long and the other $l/4$. The first part contains the beam aperture and is a fixed part of the tube. The other can be moved in the z direction by means of a diaphragm in the wall of the tube. In this way the cavity can be tuned by about 5% without significantly changing the impedance.

The construction of the catcher cavity is shown in the perspective drawing in *fig. 6*. The figure also shows the output waveguide, which is separated from the actual resonant cavity by two partitions, with a coupling hole between them. The other end of this tapering waveguide is made vacuum-tight by a mica window 5 μm thick. The small dimensions — the volume of the resonant cavity is about 0.1 mm^3 — and the high demands made on the dimensional accuracy and on the quality of the surfaces, make it necessary to use special fabrication techniques such as "hobbing", "optical" milling and diffusion welding. By these means it proves possible, however, to achieve a quality factor and an impedance which are not far below the theoretical values.

The most important characteristics

To conclude we shall mention some of the principal characteristics of the new tube. *Fig. 7* shows the result of measurements of the output power P_o at 0.87 mm as a function of the driving power P_i (wavelength 8.7 mm).

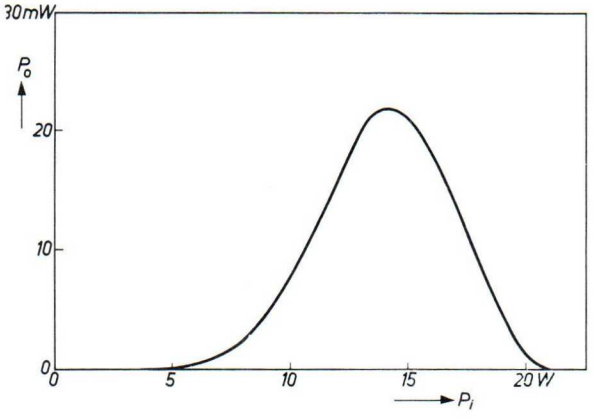


Fig. 7. The continuous wave output power P_o at a wavelength of 0.87 mm as a function of the driving power P_i at 8.7 mm. For this plot the beam voltage was 25 kV, the beam current 20 mA and the focussing magnetic field 0.37 Wb/m^2 (3700 gauss). The curve is rather broad, and therefore P_i does not have to be stabilized so drastically in order to obtain a reasonably constant output power.

The curve relates to a beam voltage of 25 kV and a beam current of 20 mA. This plot is in reasonable agreement with the calculations.

If one compares the maximum output power attainable at this beam voltage and current with the driving power which it calls for, it can be seen that the conversion of 8 mm waves into 0.8 mm waves involves a conversion loss of about 26 dB. In view of the wavelength and the high order of harmonic involved, this is a very satisfactory value.

As can be seen, the curve in fig. 7 is fairly broad; to obtain a reasonably constant output power it is therefore not necessary to stabilize the driving power too rigorously.

There are two reasons, which apply to every klystron, why the output power cannot go on rising indefinitely with the input power. In the first place, the amplitude of each of the harmonics in the beam is never larger than twice the DC value of the unmodulated beam. In the second place, the position of the point where the bunching is strongest, and thus where the maximum output power can be expected, depends on the input power: there is only one value of the input power where that point coincides exactly with the location of the catcher cavity. This explains why the curve shows a maximum.

It has been found that, with a low driving power, at the most 0.5% of the beam electrons are lost on the way; the rest reach the collector. As the driving power is raised the loss increases, and at the optimum value of the driving power it amounts to roughly 7%. These figures relate to a focussing magnetic field of 0.37 Wb/m² (3700 gauss).

Fig. 8 shows a plot of $P_{o\,max}$, the maximum output power at 0.87 mm, and $P_{i\,max}$, the associated driving power at 8.7 mm, as functions of the beam voltage at constant perveance — i.e. the ratio of the beam current to $V_0^{3/2}$. It can be seen from the figure that an output power of 1 mW is already achieved at a beam voltage of 16 kV, where the beam current is 10 mA and the driving power 8 W. At 28 kV, the highest voltage at which measurements were taken, the output power is 35 mW at a driving power of only 15 W. The measured values of the driving power are within 20% of the values calculated using the theory given in the article referred to [7]. This theory also seems to be confirmed by the measurements as far as the output power is concerned. The difference in this case was estimated at no more than 30%. An accurate comparison here is not possible at present, since the necessary measuring methods in the submillimetre range are not yet sufficiently advanced.

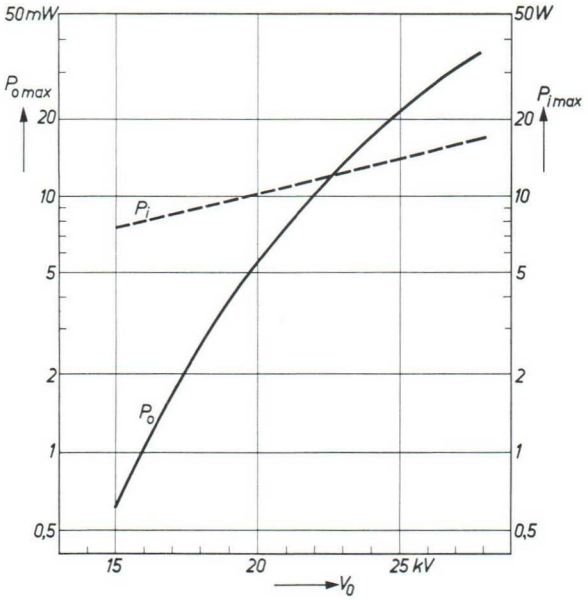


Fig. 8. The maximum output power $P_{o\,max}$ and the appertaining driving power $P_{i\,max}$ as a function of the beam voltage V_0 at constant perveance (in this case 5×10^{-9} A/V$^{3/2}$).

All in all it may be said that it has proved possible to generate 0.8 mm waves with reasonably high continuous power by means of a klystron frequency multiplier. There are reasonable prospects that it will be feasible with this method to penetrate still further into the submillimetre range within the foreseeable future.

Summary. Submillimetre waves (wavelength 0.87 mm) can now be generated with a continuous power of some tens of mW by means of a klystron frequency multiplier (the catcher cavity being tuned to the tenth harmonic of the driving frequency). This result was mainly obtained by choosing the beam voltage (25 kV) much higher than in reflex klystrons and earlier klystron multipliers (2-3.5 kV). This has made it possible, for example: a) to increase the beam diameter (this is about 0.2 mm), which reduces the cathode load; b) to use a flat resonator as the catcher cavity (better cooling and easier to make) without substantially sacrificing output power; c) to increase the distance between the buncher and catcher cavities. The latter is rectangular, oscillates in the TE_{102} mode, and can be mechanically tuned by about 5%. The beam (20 mA) is focussed by a magnetic field of 0.37 Wb/m² (3700 gauss). The maximum power output achieved (35 mW) required an input power of only about 15 W.



Coloured X-ray pictures

Image subtraction as proposed for X-ray diagnostics more than 30 years ago can conveniently be achieved by closed-circuit television techniques. A normal X-ray photograph of a skull can be subtracted by this method from a photograph taken of the same skull in an identical position after the injection of a contrast liquid in the artery: the resulting picture will not contain the bone shadows which are such a nuisance in examinations of the brain.

As a further step, subtraction is applied to several photographs showing distinct phases of the movement of the contrast liquid through the arterial and venous systems, and the subtracted television signals are supplied to suitable channels of a colour TV monitor. A picture showing simultaneously the arterial and venous systems in different colours is then obtained on the screen.

(See: W.J.Oosterkamp, A.P.N. van 't Hof and W.J.L. Scheren, Coloured X-ray pictures, Philips tech. Rev. **27**, in preparation.)

A DYNAMO FOR GENERATING A PERSISTENT CURRENT IN A SUPERCONDUCTING CIRCUIT

by J. VOLGER *).

621.313.291:537.312.62

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

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

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

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

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

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

*) Philips Research Laboratories, Eindhoven.

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

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

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

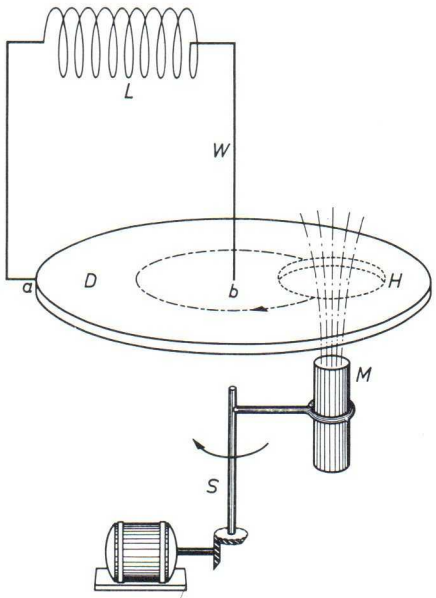
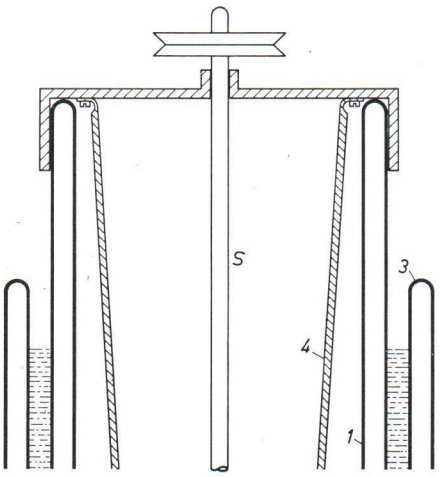


Fig. 1. Schematic diagram of the dynamo for generating a persistent current in a superconducting circuit. *D* lead disc. *W* circuit of "hard" superconducting material, connected to *D* at the rim (*a*) and in the centre (*b*). The shaft *S* mounted below the centre of *D* carries on an arm the bar magnet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

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

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



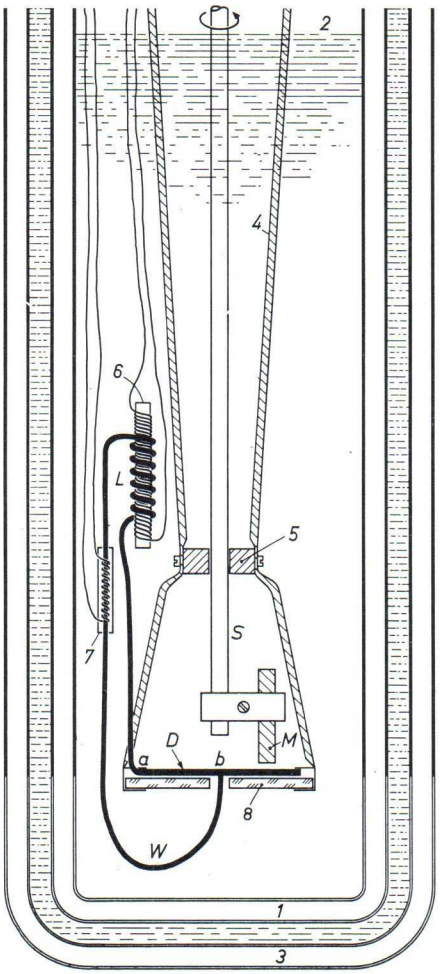


Fig. 2. Sketch showing the construction of the new dynamo and its positioning in a cryostat. The letters have the same meaning as in fig. 1. Other symbols are: *1* Dewar vessel of inner cryostat. *2* surface of the liquid helium. *3* Dewar vessel of outer cryostat; the space between *1* and *3* is filled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

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

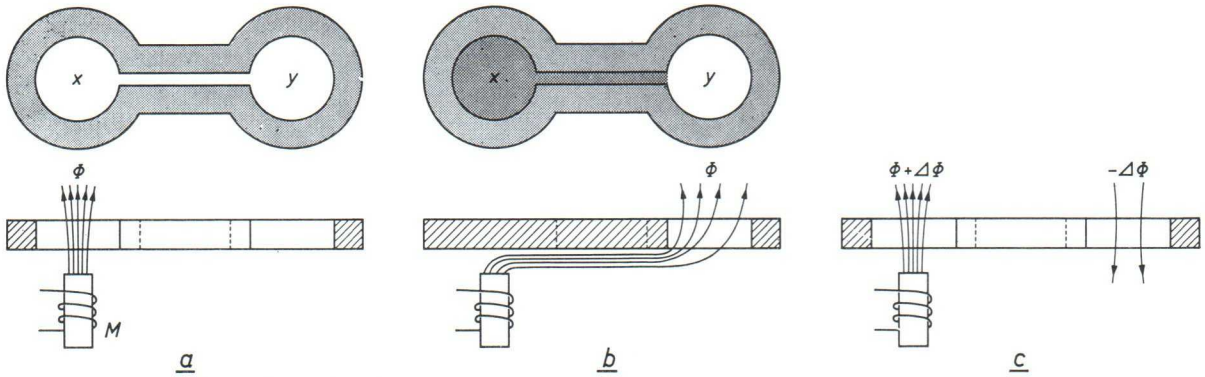


Fig. 3. Illustrating the properties of a superconducting ring circuit of complicated shape, in one part of which a flux is enclosed.

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

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

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

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

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

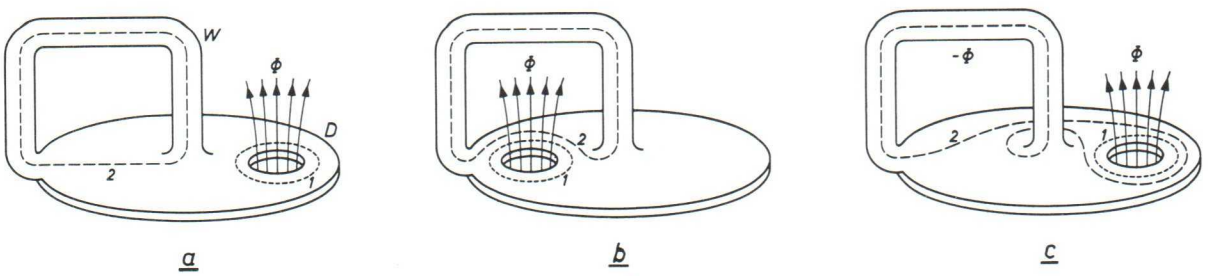


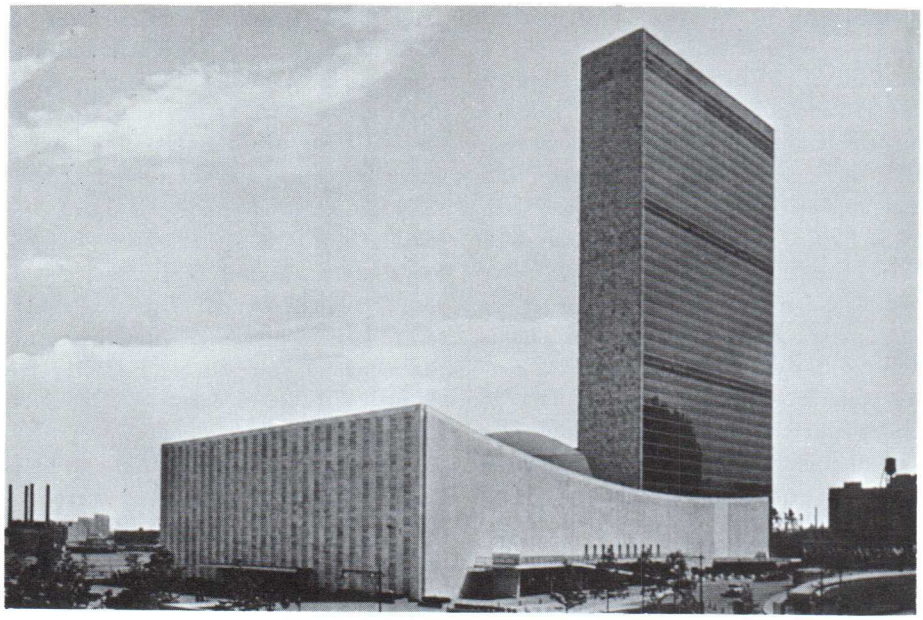
Fig. 4. Illustrating the operation of the new dynamo.

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

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

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

Summary. The extremely strong magnetic fields (of the order of 10 Wb/m²) 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.



THE FOUCAULT PENDULUM IN THE UNITED NATIONS BUILDING IN NEW YORK

by J. A. HARINGX and H. van SUCHTELEN.

525.36.002.2

Among the gifts presented by member states for embellishing the United Nations building in New York is a large Foucault pendulum from the Netherlands. The pendulum, which is suspended in the entrance hall leading to the main Assembly Hall, continuously demonstrates by its movement that the earth is in rotation.

The article below briefly describes the suspension and driving systems of this pendulum. The design was based on the requirement that the pendulum should function well for a very long time without supervision and maintenance.

On 7 December 1955, Mr. H. Luns, Dutch Foreign Minister, presented to the Chairman of the United Nations General Assembly a Foucault pendulum on behalf of the people of the Netherlands. The pendulum is suspended above the central column of the staircase in the entrance hall of the United Nations building in New York (see *fig. 1*). The pendulum was specially designed in the Research Laboratory of N.V. Philips' Gloeilampenfabrieken, the guiding consideration being that the pendulum should function uninterruptedly for many years without requiring supervision or maintenance.

A Foucault pendulum is in principle merely a weight (which we shall henceforth refer to as the "bob") suspended by a long wire, swinging in a vertical plane. Owing to the manner of suspension the plane of swing is not, however, fixed with respect to the earth but, in consequence of the rotation of the earth, turns about the vertical through the point of suspension. In the northern hemisphere the plane of swing deviates in the clockwise direction, in the southern hemisphere in the opposite direction. It can be shown [1]) that for an idealized case (very long pendulum, very small amplitude) the time T' in which the plane of swing rotates 360° is given by the formula:

$$T' = \frac{T}{\sin \varphi}, \quad \ldots\ldots \quad (1)$$

in which T is the period of revolution of the earth and φ is the geographic latitude. At the poles $T' = T$, while at the equator $T' = \infty$, i.e. there is no rotation of the plane. In New York ($\varphi = 40°45'$), time $T' = 36$ hours 50 minutes. It further appears from the theory that the angular velocity with which the plane rotates with respect to the meridian is constant.

[1]) See e.g. A. Sommerfeld, Vorlesungen über theoretische Physik, Part I, Dieterich Verl., Wiesbaden 1952.



Fig. 1. The entrance hall in the United Nations building in New York. The Foucault pendulum, a gift from the Netherlands, is mounted above the central column of the staircase.

In practical conditions, the time of rotation is found from the formula:

$$T'' = \frac{T}{\sin \varphi}\left(1 - \frac{3}{8}\frac{a^2}{l^2}\right), \quad \ldots \quad (2)$$

in which a is the amplitude and l the length of the pendulum [2]). In the case of the pendulum in the United Nations building, l is about $17\frac{1}{2}$ m and a about 80 cm. This length gives a period of swing of approximately $8\frac{1}{2}$ sec. The weight of the pendulum bob is about 90 kg.

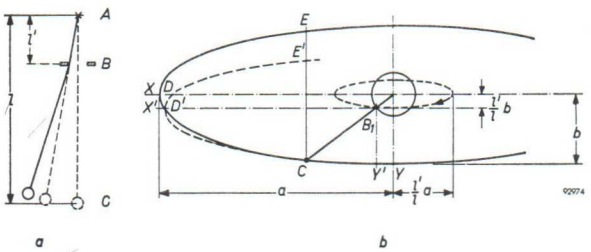
If we try to construct such a pendulum and repeat the celebrated experiment carried out by Foucault in 1851, a number of disturbing effects — due, for example, to imperfections in the rotational symmetry of the wire support — cause the pendulum, after some time, to execute an elliptical or even a circular motion. This had to be avoided with the present pendulum, which was required to swing continuously

[2]) Handbuch der Physik, Part V, p. 339, Springer, Berlin 1927.

without supervision. There were two other technical problems. In the first place there had to be the certainty that the wire would not break after some time of continuous operation, and in the second place a driving system was needed that would be able to provide the pendulum with sufficient energy to compensate for the loss of energy due to air resistance.

A simple method of overcoming the "ellipsing" problem has been described by Charron [3]). At a distance l' below the point of suspension A (see *fig. 2a*) he fixed a ring B having an internal diameter slightly larger than the thickness of the pendulum wire, leaving an annular spacing of d. As soon as the deflection of the bob exceeds the value ld/l', the wire touches the ring and the point of contact then functions as the "point of suspension". The consequence is that the minor axis of an elliptical orbit which may have been forming is rapidly diminished. With reference to fig. 2*b*, this can be roughly explained as follows.

Assuming that the wire, as long as it does not touch the ring, is straight, it will describe at the level of the ring an ellipse geometrically similar to the elliptical orbit of the bob reduced in the ratio l'/l. This is represented in the figure by the small broken ellipse; it is traversed, for example, in the direction of the arrow. The circle represents the limit set by the ring to the deflection of the wire. The moment the wire touches the ring, point B_1 functions as the point of suspension. Assuming that the velocity of the ball in the Y direction is small at that moment, the ball will thereupon proceed to describe the ellipse $CD'E'$, the axes of which lie along the lines B_1X' and B_1Y'. We see, then, that the amplitude



a b

Fig. 2. *a*) Suspension of pendulum according to Charron. The point of suspension is at A. The ring B, whose internal diameter is only slightly larger than the diameter of the wire, prevents the bob C from describing an ellipse.
b) Functioning of Charron's ring. When the pendulum describes an ellipse with semi-axes a and b, the wire touches the ring e.g. at point B_1. If there is sufficient friction between ring and wire, B_1 then functions for a while as the effective point of suspension and the bob describes the ellipse $CD'E'$ instead of the ellipse CDE, which it would describe in the absence of the ring. The amplitude in the Y direction is thus diminished in a half period by the length EE', which is equal to $2bl'/l$.

[3]) F. Charron, Bull. Soc. Astr. Fr. **45**, 457, 1931.

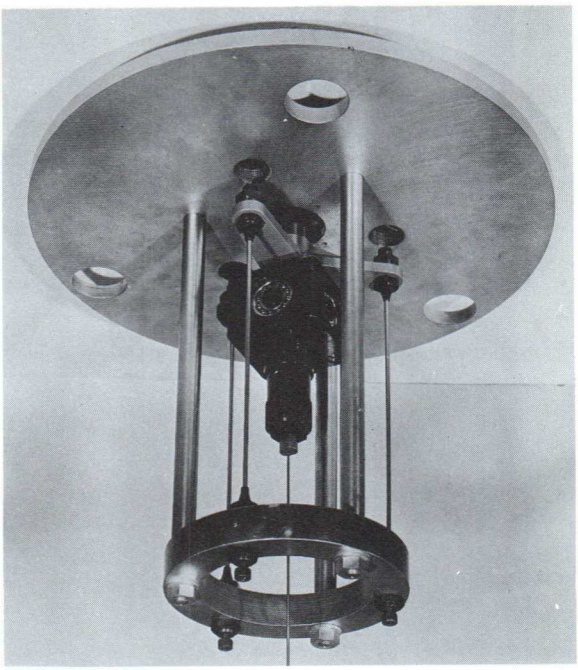


Fig. 3. The actual suspension system.

in the Y direction at the beginning of the following half swing has been reduced by the distance EE', which is approximately equal to $2bl'/l$. The relative reduction for each full swing amounts then to $\Delta b/b = 4l'/l$.

It has been tacitly assumed in the foregoing that while the bob describes the orbit $CD'E'$ the wire stays pressed against the ring at point B_1. As a rule the contact friction will not be sufficient for this to occur. In reality, therefore, $\Delta b/b$ is smaller than calculated here, but this does not alter the fact that the mounting of the ring is a very effective means of combating the elliptical motion.

Charron deduced that the use of this construction would slightly shorten the period of rotation and derived the following expression for the period:

$$T''' = \frac{T}{\sin \varphi}\left(1 - \frac{3}{8}\frac{a^2}{l^2} - \frac{4d}{\pi a}\right). \quad . \quad . \quad . \quad (3)$$

For the pendulum installed in New York a suspension was designed, wich, while equivalent to that of Charron, differs considerably in its construction. *Fig. 3* shows a photograph of the suspension and *fig. 4* a diagram indicating its construction. The mounting plate A, fixed on two beams, carries via three rods B a ring C. This ring supports by means of three flexible rods D a three-armed yoke E fitted with a universal joint F. The wire is suspended from F. By the bending of the flexible rods the yoke can move radially over a distance equal to the radial

play s between pin G and the periphery of a hole in the mounting plate. The flexible rods can be moved vertically and their length accurately adjusted; in this way the system as a whole can be given the required lateral stiffness and the pin made to lie exactly in the centre of the hole in the position of rest.

If we compare this arrangement with that of fig. 2*a*, we notice that the universal joint corresponds to the point of the wire which, in the Charron system, is at the height of the ring B. The hole in the mounting plate fulfils here the function of the ring, so that in the new construction the play s is equivalent to the play d in the old. The use of the universal joint avoids the friction and bending which the wire undergoes in Charron's construction at the level of the ring, thus greatly reducing the risk of wear or fatigue failure. The three flexible rods, like Charron's wire length AB, bear the weight G' of the pendulum bob etc. and, via the yoke, exert on the joint a centrally directed force which is proportional to the radial displacement. At a displacement u the above-mentioned point of Charron's wire is subjected to a force of magnitude Gu/l'. By appropriately dimensioning the flexible rods in the new construction, the same magnitude of restoring force can be produced.

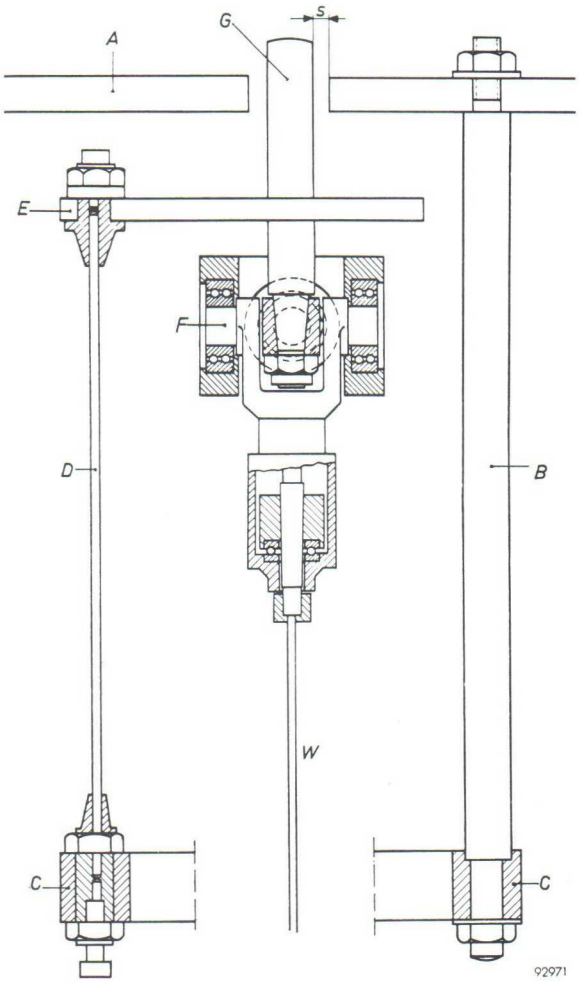


Fig. 4. Schematic diagram of the suspension system of the Foucault pendulum in the United Nations building. Mounting plate A carries three rigid rods B to which a ring C is fixed. Three flexible rods D, clamped on the latter, support a yoke E, which is fitted with a universal joint F. Pin G mounted on E, lies centrally in a hole in the mounting plate, and limits the lateral displacement of the yoke to the small distance s. The suspension wire W is fixed to the joint by clamping in a bifurcated steel cone. The dimensions of the flexible rods are such as to ensure the required lateral stiffness and to ensure that the bending stress in these rods is below their fatigue bending strength.

In connection with the dimensions of the flexible rods it should be borne in mind that the lateral stiffness c is determined not only by the dimensions and the modulus of elasticity but also to a large extent by the compressive force $\frac{1}{3}G'$. The length l' no longer corresponds to any physical quantity such as the length of the flexible rods, but is equal to $G'/3c$.

The lateral stiffness for this case (parallel-constrained ends) is given by the formula

$$c = \frac{P}{l_1} \frac{\frac{1}{2}ql_1}{\tan\frac{1}{2}ql_1 - \frac{1}{2}ql_1}, \quad \ldots\ldots (4)$$

where $q^2 = P/EI$ and $P = \frac{1}{3}G'$. Further, l_1 is the length of the flexible rods, E their modulus of elasticity and I their second moment of area in bending.

The suspension wire is of hard drawn stainless steel with a diameter of 2.5 mm. The wire is clamped at both ends by a truncated cone of hardened steel with internal teeth. The cone has diametric saw cuts at its narrower end and is drawn fast into a ring by the weight of the bob.

In the design of the driving mechanism, which is necessary to keep the pendulum swinging without any reduction in amplitude, account had to be taken of the fact that access to the point of suspension is extremely difficult once the pendulum is mounted. The drive is therefore provided at the bob of the pendulum. The method adopted was to repel the bob by means of eddy currents generated in it by a coil mounted on the pedestal (*fig. 5*), the coil being energized at suitable moments by alternating current. If a copper plate is placed above such a coil at right angles to its axis, voltages will be induced in the plate which are 90° out of phase with the field, and currents which are in their turn almost 90° out of phase with these voltages. The eddy currents are therefore almost 180° out of phase with the current in the coil, thus giving rise to repulsion. If a not too large, horizontal, round plate is placed with its centre-point exactly in the axis of the vertical coil, the resulting force acting on the plate is a vertical one, but if the plate is slightly eccentric to the coil axis, the force will also have a horizontal component which can be used for driving the pendulum. A plate



Fig. 5. The pedestal, designed by architect G. Rietveld, on which the drive coil is mounted. The pedestal bears the inscription:

It is a privilege to live
This day and to-morrow.
Juliana

of this kind is contained in the lower half of the pendulum bob. In order to concentrate the magnetic field on the plate, 18 U-shaped yokes of ferroxcube are fitted around the coil.

This method, then, avoids the use of iron in the ball, and thus prevents the pendulum from being influenced by the magnetic field of the earth or by the steel structure of the building.

Some remarks may be added on the transfer of energy in this driving system. Eddy currents in the copper plate react on the coil with the result that the self-inductance of the latter is smaller than when the plate is absent. When the plate moves away this self-inductance increases and the energy transferred to the plate is approximately equal to $I^2 \Delta L$, where ΔL is the change in the self-inductance and I the current. This does not take into account that I itself changes slightly, but, partly because of the circuit arrangement (see below), this relation holds fairly accurately.

The moment at which the coil is energized is determined by the moment at which the rod in the yoke (G, fig. 4) looses contact with the mounting plate [4]). For this purpose, rod and hole were designed to constitute an electrical contact. A delay circuit, which also controls the duration of energization, ensures that the current in the coil is switched on at a specific time after the moment the rod leaves the mounting plate, viz. at approximately the same moment at which the centre of the bob passes the axis of the coil. The anode load of tube II in the circuit (see *fig. 6*) includes the D.C. winding of a transductor T [5]). The A.C. winding L_1 of this transductor is connected in series with the drive coil L_2 and a capacitor C''. In addition, another capacitor C' is connected in parallel with L_1. When the D.C. coil of the transductor is not energized, the circuit L_1-C' is in resonance and represents such a high impedance that the current through L_2 is only 85 mA.

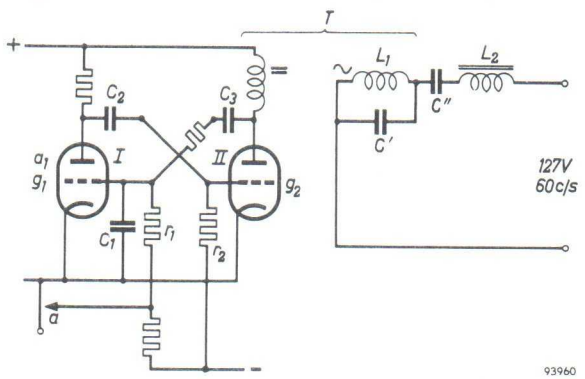


Fig. 6. Control and drive circuit for driving the pendulum. The control circuit (left) causes the required alternating drive current from the mains to be supplied to the drive coil L_2 only after a certain delay after the opening of contact a; the circuit sustains the current for a certain fixed time, even if contact a is again closed in the meantime. Contact a is formed by rod G (fig. 4) and the hole in the mounting plate A. In the quiescent state, tube I is conducting and tube II cut off. When a opens, the potential of g_1 falls and the potential of a_1 rises: at a certain point the circuit triggers, tube II now becoming conducting (via C_2) and tube I being cut off. The delay between the opening of a and the triggering of the circuit depends on the product C_1r_1. The anode current of tube II flows through the D.C. winding of transductor T and reduces the value of L_1 such that the drive circuit (right) comes into series resonance. After an interval, which depends on the products C_2r_2 and C_3r_1, the control circuit returns to the quiescent state and L_1 rises again. The value of C' is such that resonance now occurs in the parallel circuit L_1-C'. The latter then represents such a high impedance in the circuit that the current is reduced to an ineffective value.

[4]) R. Stuart Mackay, Amer. J. Phys. **21**, 180, 1953, describes a pendulum driven by eddy currents, the synchronization being derived from the pendulum bob.

[5]) A transductor (saturable reactor) is a choke with an iron core, the self-inductance of which can be varied by changing the magnetization of the iron core with the aid of a D.C. winding.

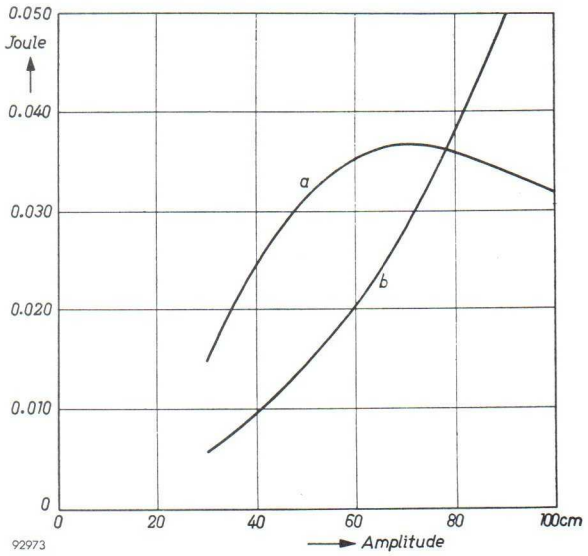


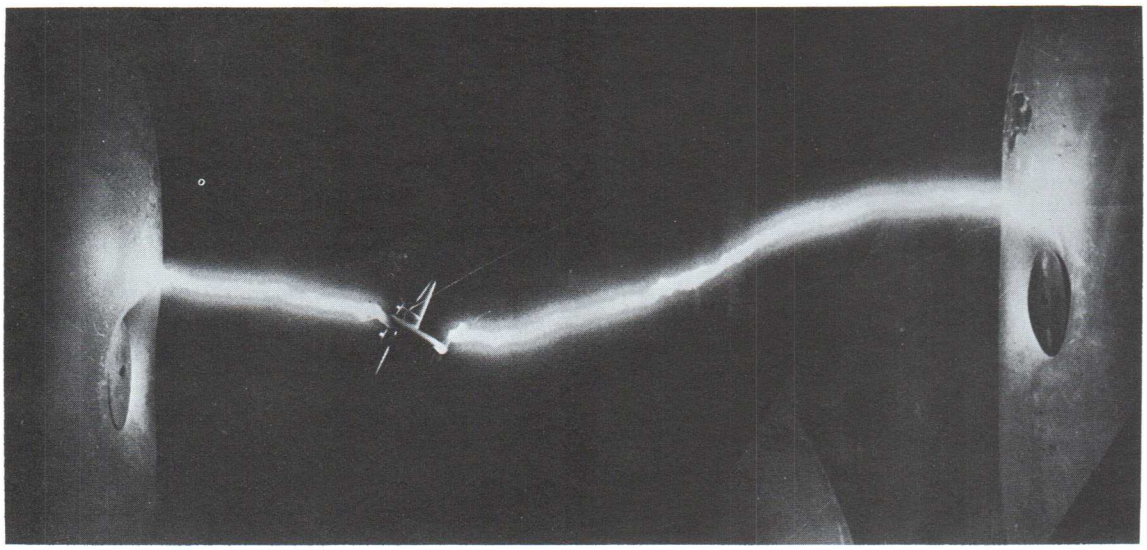
Fig. 7. The energy supplied per period to the pendulum (curve a) and that lost by air resistance (curve b) as a function of amplitude of swing. At an amplitude of approx. 80 cm the curves intersect. At smaller amplitudes the energy supplied is greater than that lost, i.e. the point of intersection represents a stable state of equilibrium.

When the transductor is energized by the anode current of tube II, the self-inductance of L_1 is reduced by about a half. The capacitance of C'' is such that the whole circuit now comes into series resonance, whereupon the current through L_2 rises to 240 mA. The ratio between operating and quiescent currents does not seem particularly large, but it must be remembered that the energy transferred is proportional to the square of the current. In this way about 0.035 joule is supplied to the bob in each period, which is sufficient to provide an amplitude of swing of the required value, viz. approximately 80 cm (see *fig. 7*).

Summary. A description is given of the suspension and driving systems for the Foucault pendulum suspended in the hall of the United Nations building in New York. The pendulum was presented on 7th December 1955 to the Chairman of the General Assembly by the Dutch Foreign Minister H. Luns on behalf of the people of the Netherlands. The main problem of preventing the pendulum bob from describing an ellipse instead of swinging in a flat plane was solved by utilizing the principle described by Charron. The construction is so designed as to reduce to a minimum the risk of the wire breaking owing to wear or fatigue. The drive, which is necessary to compensate for the energy losses, caused mainly by air resistance, is effected by means of a magnetic coil with ferroxcube core placed centrally under the pendulum. This coil is energized by alternating current and induces eddy currents in a copper plate contained inside the lower half of the bob. The energizing current is controlled by an electronic relay, which ensures that the current is switched on some time after the pendulum itself has broken an electrical contact in the suspension system, and keeps it switched on for a certain period. The coil is mounted on a pedestal designed by the architect G. Rietveld.

LIGHTNING AND AIRCRAFT

551.594.221:629.135



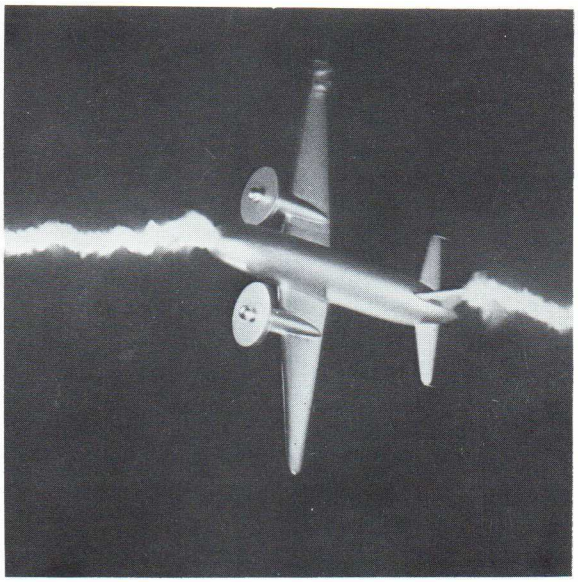
Thunderstorms are violent atmospheric phenomena, and on such a large scale that it hardly seems feasible to similate them in small-scale laboratory experiments.

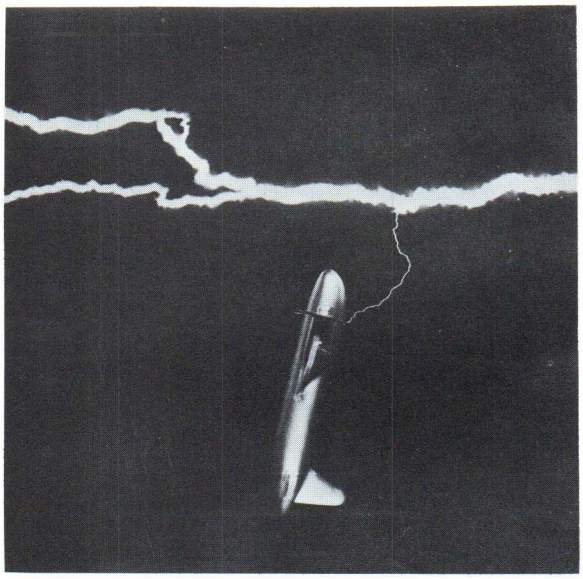
However, it is possible to study certain aspects of these phenomena by means of scale experiments, in those cases in which the effects are essentially unchanged by the scale. An example is the impulse testing of insulators and transformers at high tension, to simulate the effects of lightning striking a power line.

The photographs reproduced in this article are illustrative of quite a different investigation undertaken recently in Eindhoven. The problem is concerned with the question as to what parts of an aircraft flying through a thunderstorm region are most likely to be struck by lightning. With the increasing speed and size of modern aircraft, it has been found that incidents of this kind are becoming more and more frequent. Although accidents only rarely result from aircraft being struck by lighting, some damage, often only slight, usually results. In order to maintain full airworthiness the damage will probably have to be repaired before the next flight; this means loss of time which can have serious effects on the economics of operations.

To investigate which parts of an aircraft are most susceptible to a lightening stroke, a model aircraft, scale 1 : 100, was suspended from insulating threads between two high-tension electrodes (of dimensions of some yards, see photograph at the head of this article). One electrode is connected to earth, the other to the positive high-tension terminal of a million-volt cascade generator [1]). The separation of the electrodes is such that spark-over would occur spontaneously at any smaller separation. Only a faint brushing (corona discharge) occurs. This is caused by the inevitable roughness of the electrodes, dust particles on the surface, etc; brushing is also sometimes observed on the model aircraft. The total value of such corona currents is very small (a fraction of 1 mA). In the space between the

[1]) See e.g. A. Kuntke, Philips tech. Rev. **2**, 161, 1937; and A. B ouwers and F. A. Heyn, Philips tech. Rev. **6**, 46, 1941.





electrodes, around the model airplane, however, a critical condition of electric stress prevails, which may break down into a disruptive discharge by any slight change. A similar condition, on the verge of a discharge, may also prevail around an aircraft in a thunderstorm. A discharge can then be triggered by any adventitious factor causing a momentary increase in field strength at some point. As the experiments involved a large number of bolts of lightning — sufficient, that is, for a statistical evaluation of the results — it was obviously impracticable to wait each time for a chance disturbance. Instead, a disturbance was induced by shooting a pellet from an airgun against the high tension electrode. Once the voltage had been adjusted near enough to the critical value, virtually every shot would trigger a full discharge [2]).

Not all, but a large number of the discharges thus triggered, chose a path via the model aircraft. On one or two occasions the model was struck by a secondary discharge branching off from the main discharge (see photograph), a phenomenon which has also been observed with natural lightning.

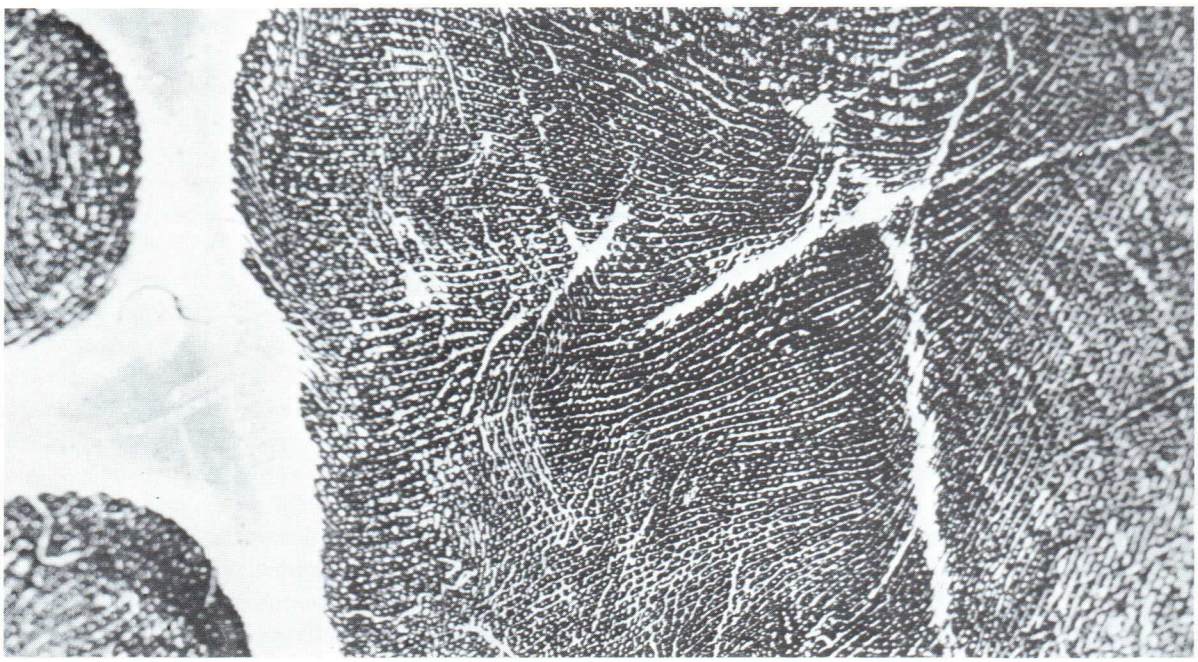
The place at which the model aircraft was struck by the discharge was photographically recorded. The aircraft model was placed in every conceivable spatial attitude with respect to the electric field. From the many hundreds of photographs is was possible to form some idea of the parts of the model most likely to be struck by the discharge.

On the basis of experiments of this kind it may well be possible in the future to provide certain points on aircraft with lightning arresters to prevent or minimize damage.

A. C. van DORSTEN.

[2]) The phenomenon that a rapid mechanical disturbance may cause a discharge across a gap when the applied voltage is just below the critical value was, as far as is known, first observed and studied by A. Kuntke in the Laboratory of C. H. F. Müller in Hamburg; see VDE-Fachberiche **12**, 157-164, 1948.





Fingerprinting via total internal reflection

"The Rays of Light in going out of Glass into a Vacuum, are bent towards the Glass; and if they fall too obliquely on the Vacuum, they are bent backwards into the Glass, and totally reflected; . . . . if the farther Surface of the Glass be moisten'd with Water or clear Oil, or liquid and clear Honey, the Rays which would otherwise be reflected will go into the Water, Oil or Honey; and therefore are not reflected before they arrive at the farther Surface of the Glass, and begin to go out of it. If they go out of it into the Water, Oil, or Honey, they go on . . . ."

The phenomenon of frustrated total internal reflection thus described by Newton in his "Opticks" has been put to good use in Philips Laboratories at Irvington-on-Hudson, U.S.A. (now at Briarcliff Manor, N.Y.) for making clearly visible and photographing the pattern of skin ridges on fingertips etc. without having recourse to the age-old ink process. This method allows very fine details to be recorded.

The lower photograph shows a "print" of a portion of an infant's foot; such prints are frequently used in hospitals for identification.

(See: N.J.Harrick, Fingerprinting via total internal reflection, Philips tech. Rev. **24**, 271-274, 1962/63.)

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.

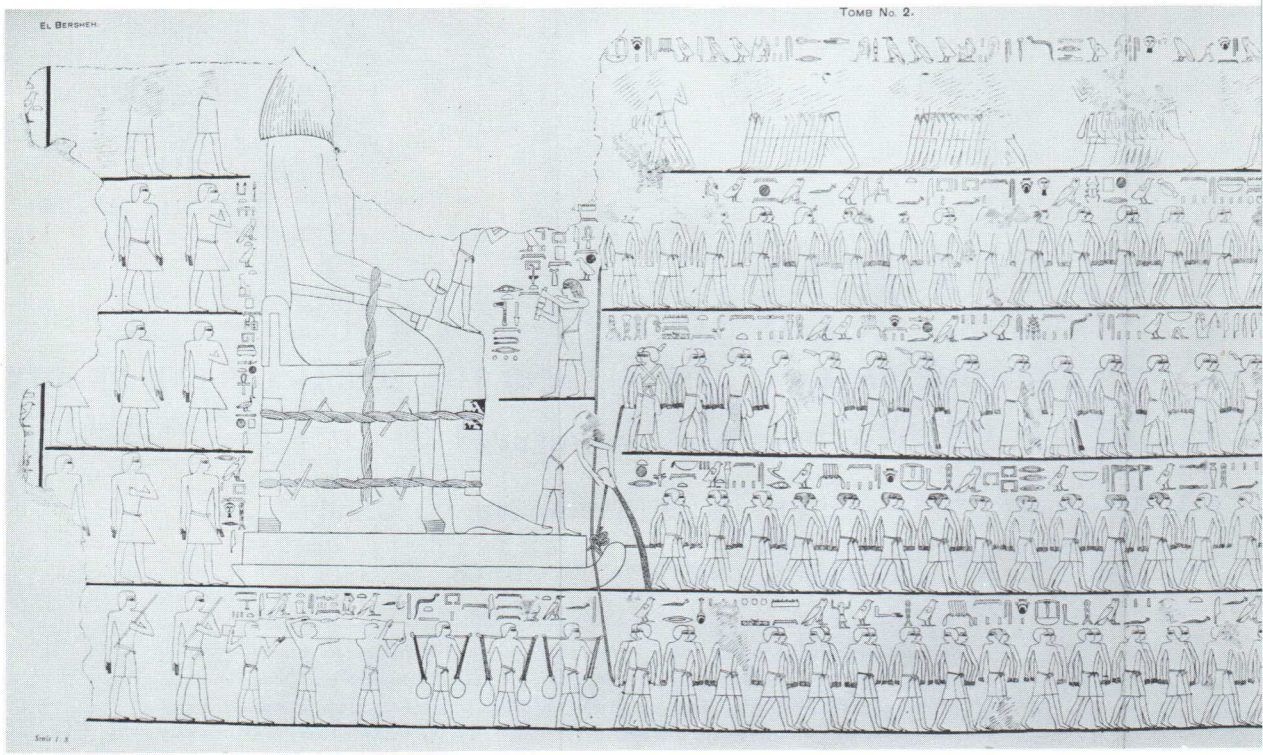
I. THE CONTACTLESS BEARING PRINCIPLE

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.

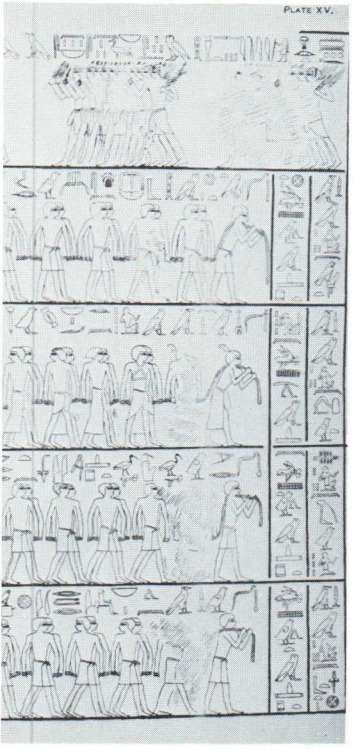
Two categories will now be discussed in turn: contactless bearings with an external pressure source ("pressurized bearings"), and "self-acting" contactless bearings. Since spiral groove bearings fall into the latter category, it will be accorded all the greater attention.

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.



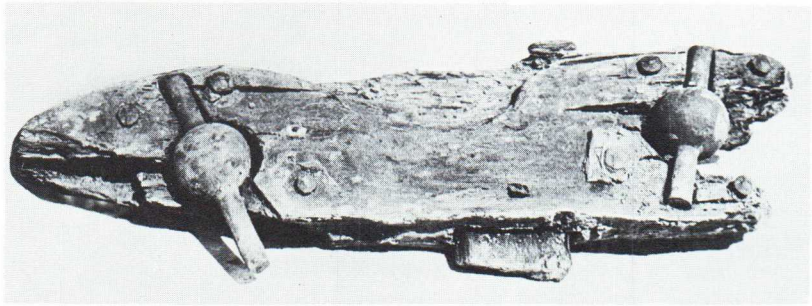


Fig. 2. Remains of a ball bearing found in one of two Roman ships that were raised round about 1930 out of the lake of Nemi. On the evidence of coins and inscriptions discovered in and near the ships it is fairly certain that they were built by the emperor Caligula (about 40 A.D.) and that they sank 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.

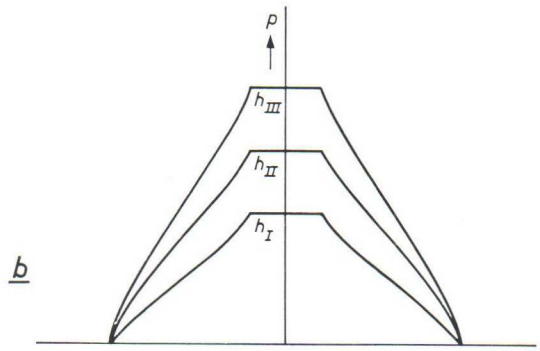
←

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-



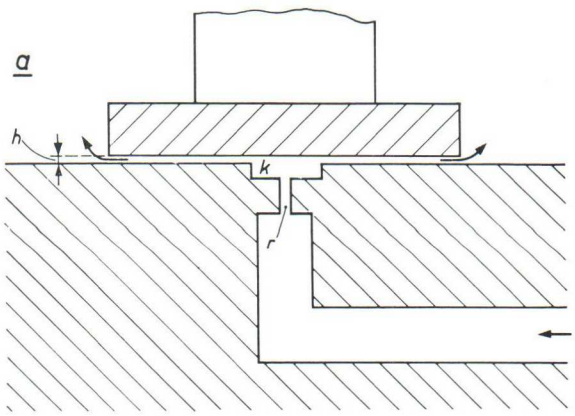


Fig. 3. (*a*) Principle of a contactless bearing with an external pressure source, the first application of which was in an exhibit called "*le chemin de fer en glace*" demonstrated at the 1878 Industrial Exhibition in Paris. This was a vehicle on "lion's feet" that was able to skim over a smooth steel 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 ($h_I > h_{II} > h_{III}$). 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*).

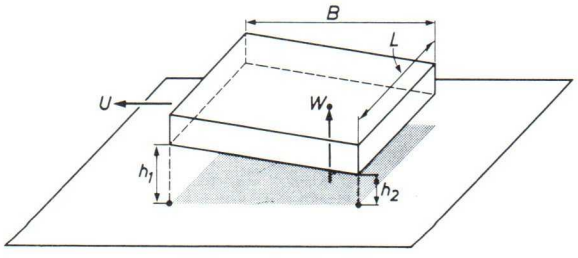


Fig. 4. To illustrate the principle of a self-acting bearing, based on the convergence or wedge effect. A flat block inclined to a base-plate in such a way that their opposed surfaces converge (in the case shown here, the gap in between is strictly cuneiform — a right triangular prism) is displaced through 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.

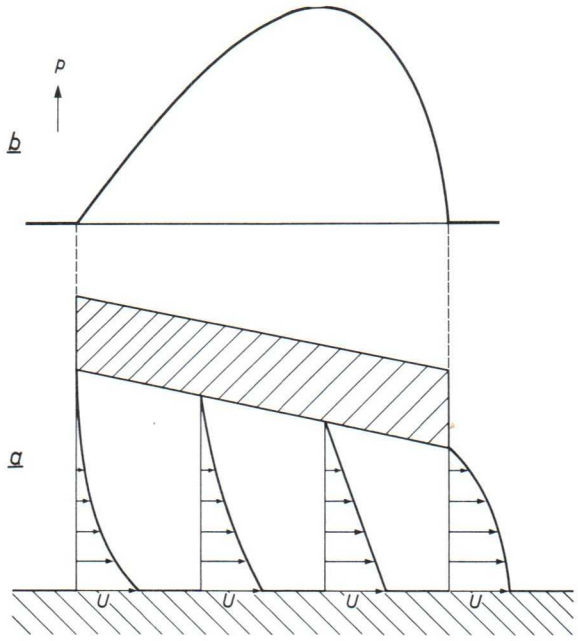


Fig. 5. (*a*) Velocity distribution and (*b*) pressure distribution in the incompressible viscous fluid filling the wedge-shaped space in fig. 4. The flat block is at rest and the base-plate is moving at a velocity U, dragging with it the fluid in the neighbouring boundary layer. The pressure is at a 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]).

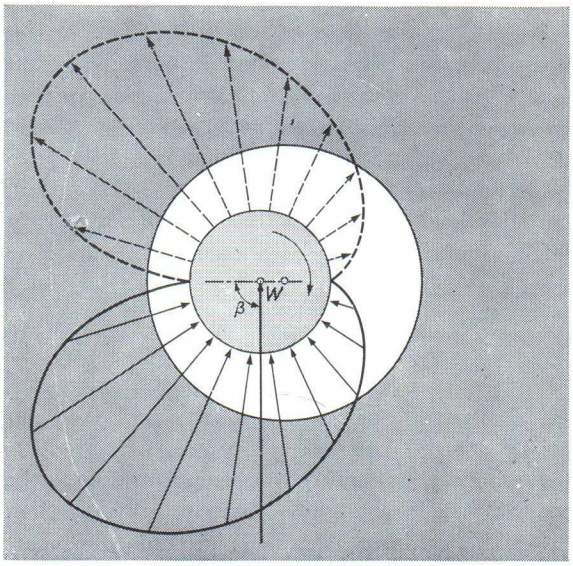


Fig. 6. Build-up of pressure inside a journal bearing lubricated with an incompressible fluid (schematic representation with bearing clearance greatly exaggerated). The fully-drawn arrows represent pressures above atmospheric, the dashed arrows pressures below it. Excess pressure arises where the bearing 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

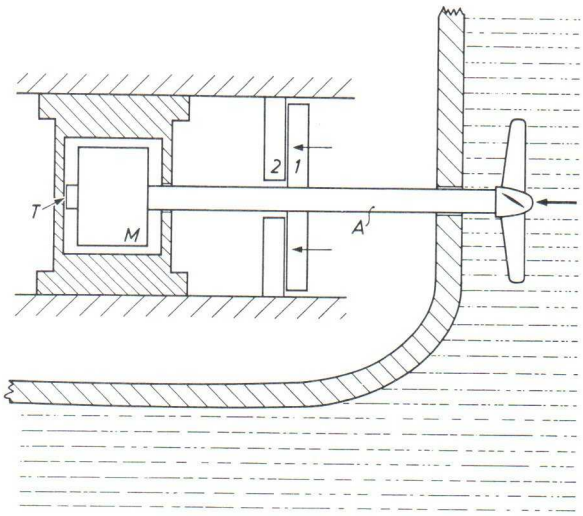


Fig. 7. A marine engine *M* drives the ship's screw via a long shaft *A* passing through a "tunnel" in the hull. The thrust exerted by the water on the screw, when this is turning, cannot of course be passed on to the hull of the ship by the bearing *T* supporting the crankshaft end in the engine housing. 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.

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

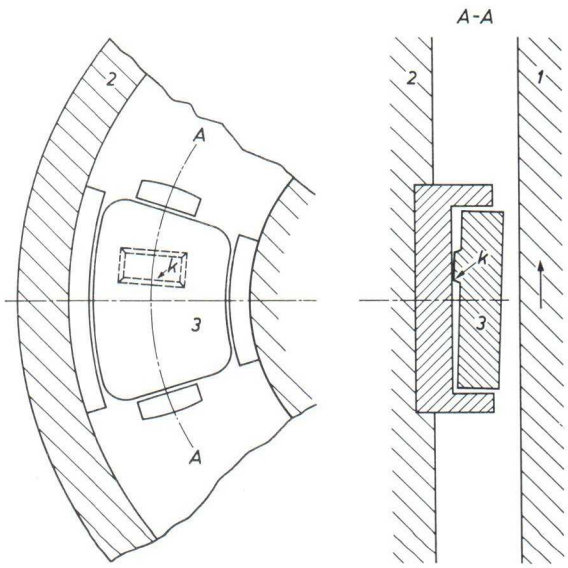


Fig. 8. The Michell bearing consists of a plane collar *1* and a circular plate *2* around whose edge are fitted slipper blocks *3* able to tilt about a pivoting edge *k*. The pivoting edge is given a position such that each block automatically adjusts itself for maximum load-carrying capacity (see the 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:

$$W = \frac{C_1 \eta U B^2 L}{h_2^2}, \quad \ldots\ldots \quad (1)$$

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:

$$F = \frac{C_2 \eta U B L}{h_2}, \quad \ldots\ldots \quad (2)$$

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

$$f = \frac{C_2}{C_1} \frac{h_2}{B}, \quad \ldots\ldots \quad (3)$$

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:

$$W = \frac{K_1 \eta \omega r^3 L}{(\Delta r)^2}, \quad \ldots\ldots \quad (4)$$

$$P = \frac{K_2 \eta \omega^2 r^3 L}{\Delta r}. \quad \ldots\ldots \quad (5)$$

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 the 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.

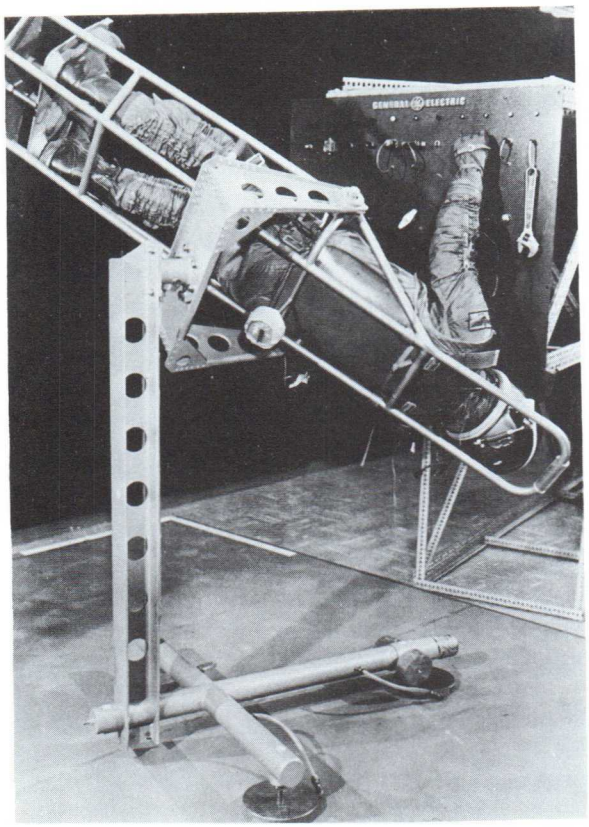


Fig. 9. Space travel simulator built by General Electric. The trainee astronaut is strapped into a frame mounted in a universal joint and supported on air-lubricated pads which glide almost frictionlessly over the floor (the principle being exactly the same as that of the "skating railway" exhibited 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.

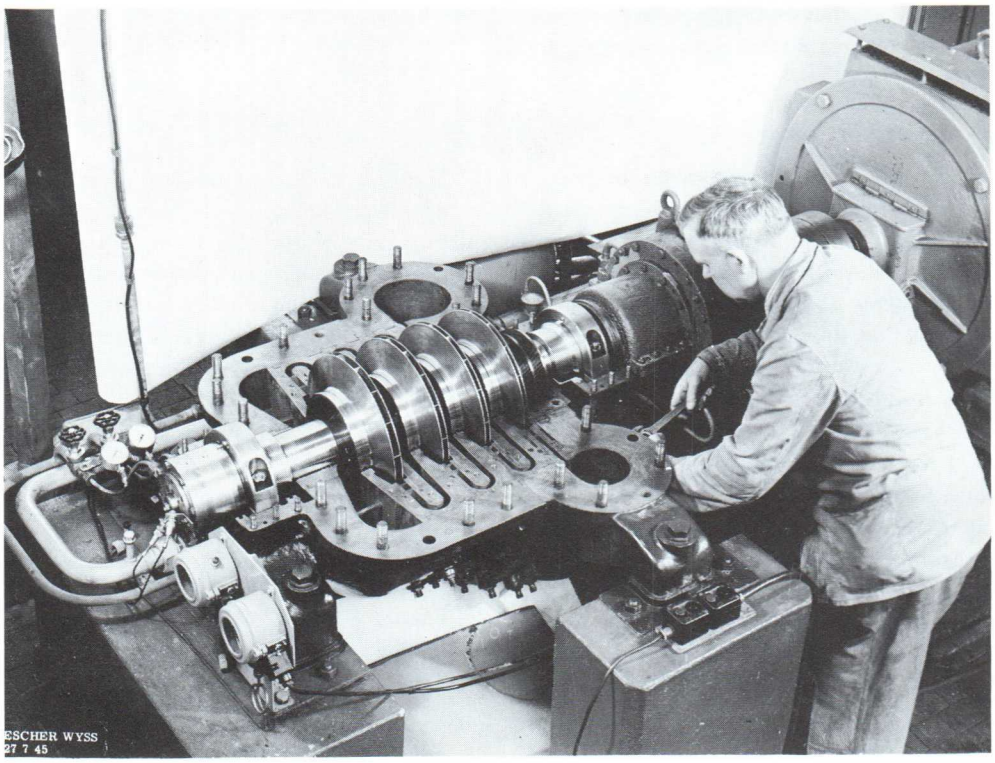


Fig. 10. Experimental four-stage centrifugal compressor made by Escher Wyss, partly dismantled. Gas bearings afford both axial and radial support for the shaft. Pipes and pressure gauges for the incoming gas supply for these bearings may be seen on the left. This compressor runs at higher speeds than 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]).

[12]) E. Loch, Escher Wyss Mitteilungen **33**, No. 1/2/3, p. 118, 1960.

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-

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.

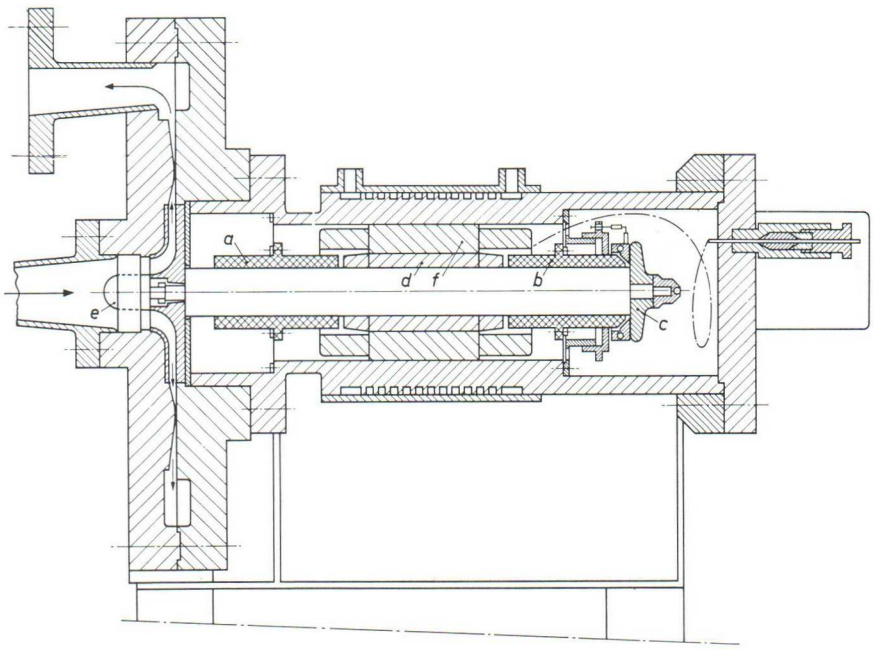


Fig. 11. CO_2 circulator for a nuclear reactor. The carbon dioxide gas used to cool the reactor (or, to put it more accurately, used to recover the heat developed in the reactor for exploitation elsewhere) is forced through the circuit by a radial fan *e* driven by an electric motor (stator *f*, rotor *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.)

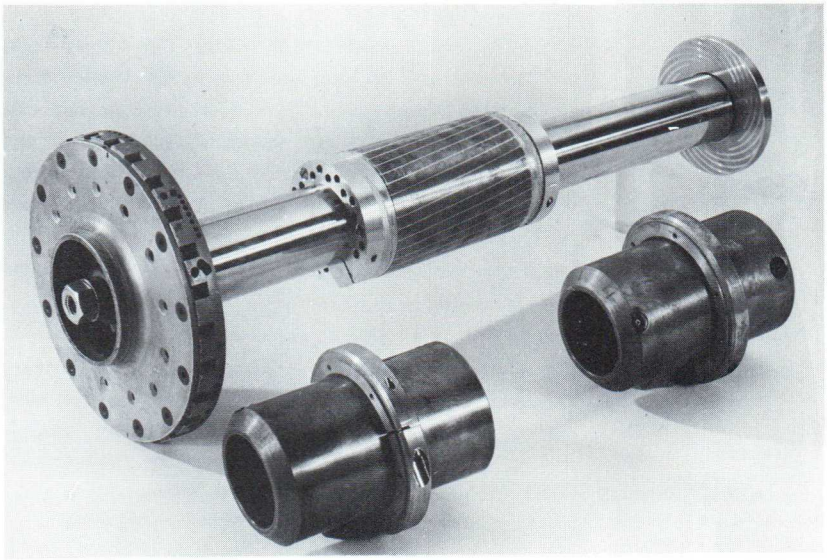


Fig. 12. Rotor assembly of the CO_2 circulator drawn in fig. 11, in which the gas circulated by the machine serves to lubricate its thrust and two journal bearings. In the foreground, the bushes for the journal bearings.

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,

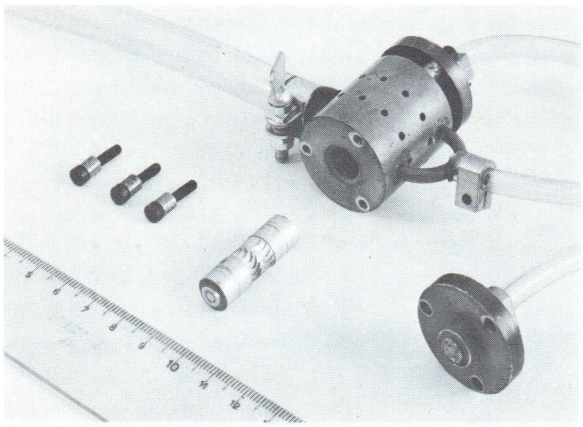


Fig. 13. Small experimental rotor which is driven and cushioned by compressed air. It can attain a speed of about 450 000 r.p.m. Recesses acting as turbine "buckets" have been milled in the rotor, which may be seen in the foreground. The split pipe entering the housing on the right supplies air for lubrication 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.

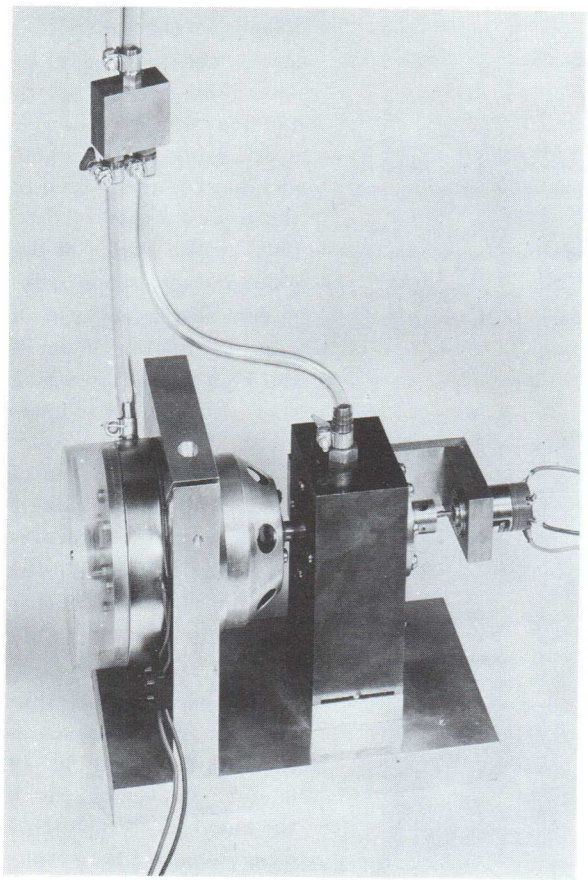


Fig. 14. Arrangement for measuring the torque delivered by a very small electric motor, mounted in the holder on the extreme right. The measuring instrument (left) contains a disc which is coupled to the motor spindle and which undergoes braking due to eddy currents; the braking couple is adjustable. 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.

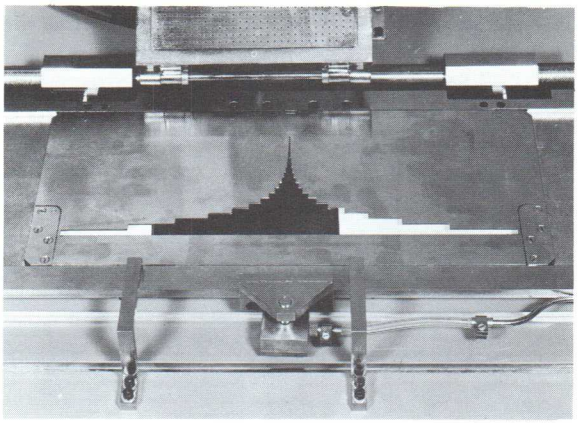


Fig. 15. Sensitometer for determining the sensitivity and other characteristics of photographic material. A masking pattern moves slowly across the instrument, with the result that the photographic material is exposed stepwise to light of known spectral composition and intensity from a source (not visible 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

$$\Delta\hat{y} = \frac{(x\omega)^2/\omega_n^2}{1-(x\omega)^2/\omega_n^2}\,\hat{h},$$

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.

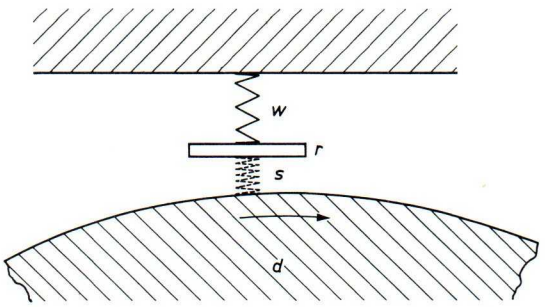


Fig. 16. Schematic drawing to show suspension of a magnetic head r for writing on to or reading off from a drum memory d such as is employed in the PASCAL computer. Spring w, which has a low stiffness value, tends to press the head against the rotating drum. In virtue of the convergence effect the air 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.

[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.

With this example we close the part of the article concerning gas bearings, and turn our attention to bearings of the spiral groove type.

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.

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).

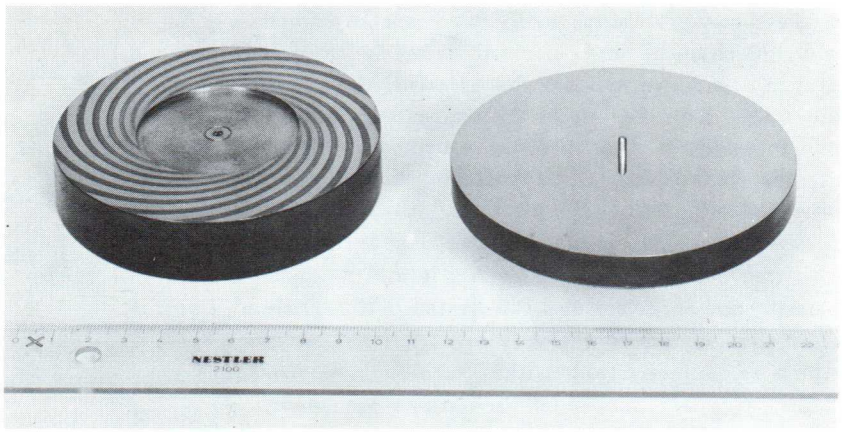


Fig. 17. Spiral groove bearing made in the Research Laboratories, consisting of a smooth and a grooved circular plate. When the plates are assembled and one is turned by hand with respect to the other (it must be turned in the right direction) complete separation occurs at a speed as low as one revolution per second, the air layer between the bearing surfaces then being about 11 μm thick.

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:

$$W = \frac{K_1'\eta\omega r^4}{h_2^2} \geqslant 10 \text{ (newton)}, \quad \ldots \quad (1a)$$

$$P = \frac{K_2'\eta\omega^2 r^4}{h_2} \leqslant 1.5 \text{ (watt)}. \quad \ldots \quad (2a)$$

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

$$h_2 \leqslant 2 \ \mu\text{m}.$$

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:

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} = 0 . \quad . . . . . \quad (7)$$

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_d above a ridge and the other giving the pressure p_g above a groove.

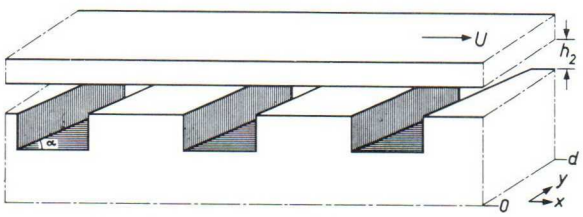


Fig. 18. Rectilinear groove model used in calculations for optimum spiral groove bearing.

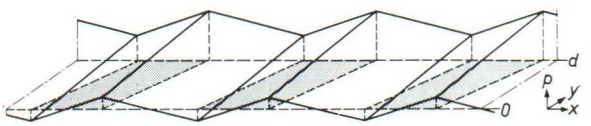


Fig. 19. First approximation to the distribution of pressures in the model appearing in fig. 18. This approximation still does not satisfy all the boundary conditions of the differential equation. By superimposing a correction which has been derived elsewhere, it is possible to obtain a distribution in which pressures are uniform along edges $y = 0$ and $y = d$, as they must necessarily be in reality.

$$p_d = p_1 + A_1 x + B_1 y , \quad . . . . \quad (8)$$

$$p_g = p_2 + A_2 x + B_2 y . \quad . . . \quad (9)$$

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:

$$\Delta p = g(\alpha,\delta,\gamma)\frac{\eta U d}{h_2^2} \quad . \quad . \quad . \quad . \quad (10)$$

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 $\delta = 0.4$.

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

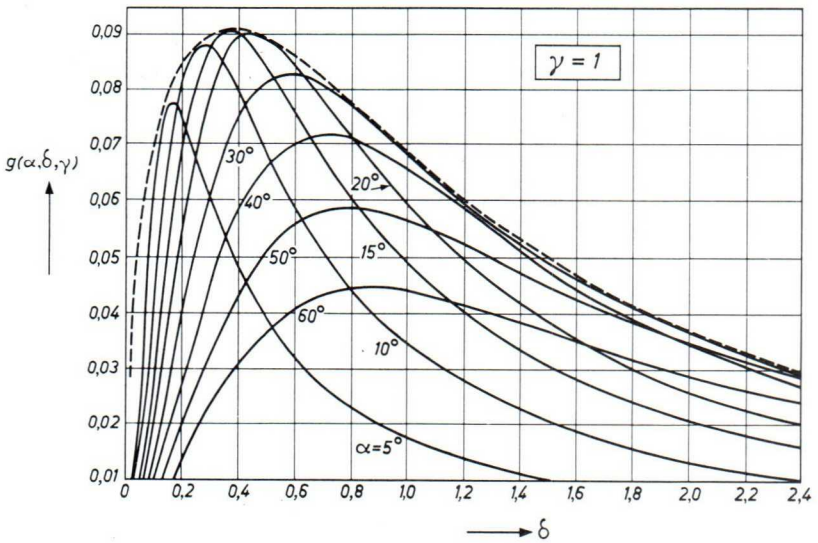


Fig. 20. Function $g(\alpha,\delta,\gamma)$ plotted against δ for $\gamma = 1$ and a set of different α values. An envelope enclosing this family of curves has been drawn as a dashed line. Inspection shows that the highest peak is that of the curve for $\alpha \approx 16°$, this maximum being attained at $\delta = 0.4$.

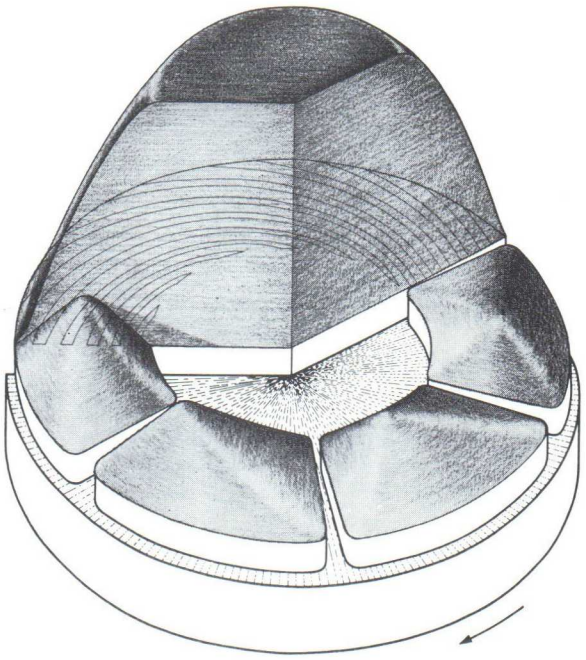


Fig. 21. Three-dimensional sketch showing calculated pressure distributions in an optimum spiral groove bearing (at the rear) and in a Michell bearing of the same size, equipped with a total of six slipper blocks (in front). A smooth upper member, rotating in the sense indicated by the arrow, is to be 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

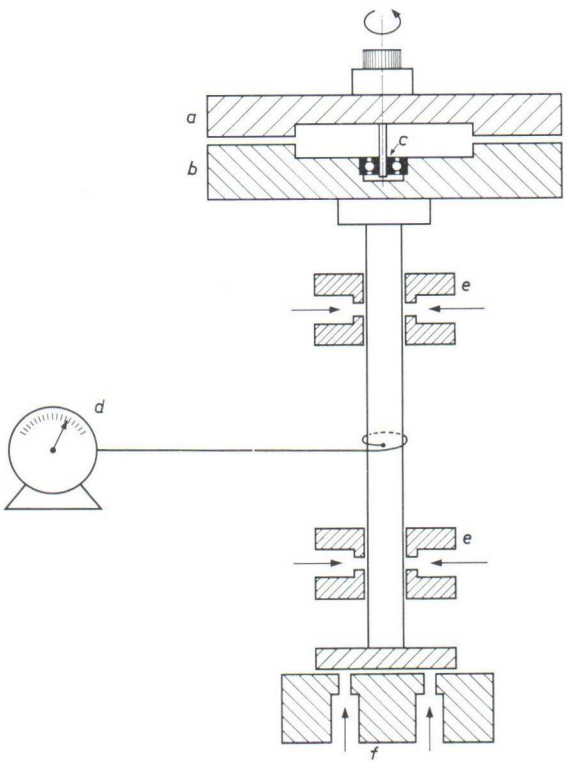


Fig. 22. Arrangement for measuring friction in a spiral groove bearing (schematic). The upper, grooved member *a* is caused to rotate at a given speed relative to the smooth plate *b*. The frictional couple exerted on member *b* is measured with a small dial-reading dynamometer *d*. A very small ball 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 Tiefe 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 einzigen 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 Zälgeln 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.

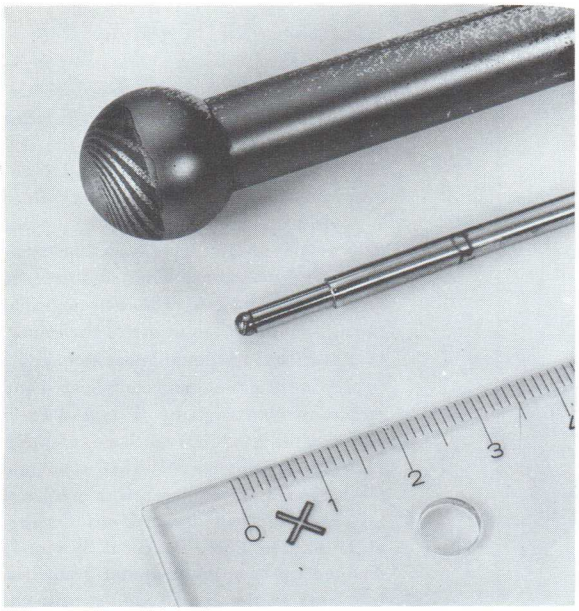


Fig. 23. Spindle-ends for spherical-type spiral groove bearings with diameters of 18 mm and 3 mm. The grooves are cut into hardened and polished steel balls (such as are used in ball-bearings) by an etching process. The surface of the metal is coated with light-sensitive lacquer, and the desired spiral 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.

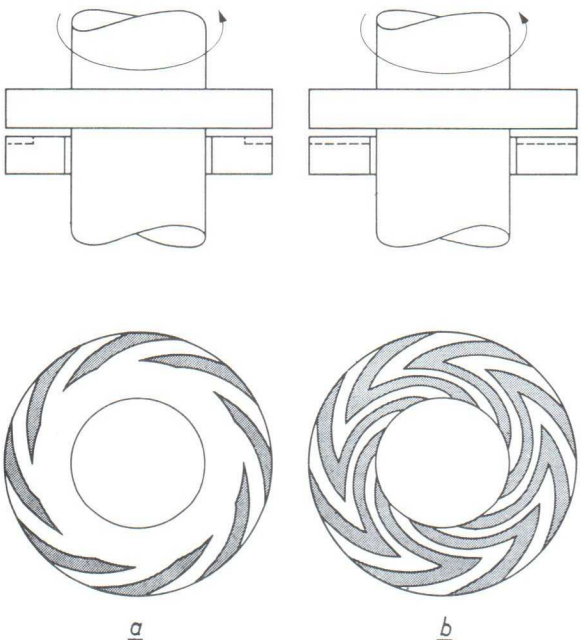


Fig. 24. Two types of spiral groove bearings suitable for taking up axial thrust from a shaft that passes through the bearing. (*a*) Partial spiral groove bearing, in which the grooved area is restricted to the outer margin of the plate. (*b*) Herringbone bearing.

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*.

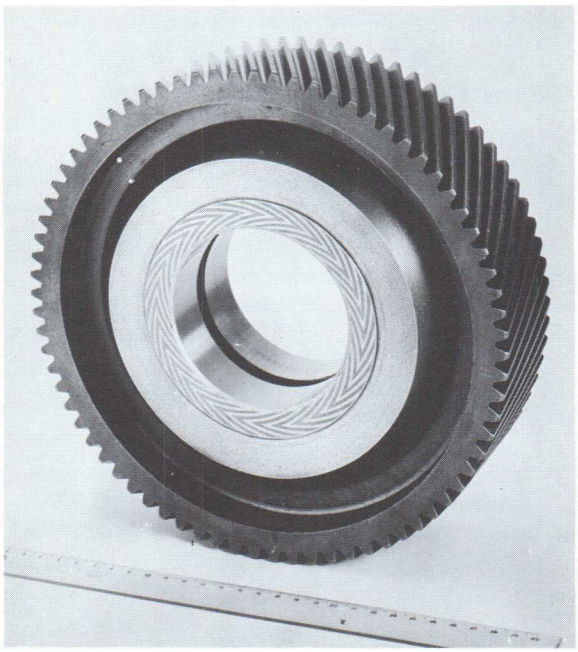


Fig. 25. A herringbone bearing employed in a hot gas engine developed in the Research Laboratories [21]).

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

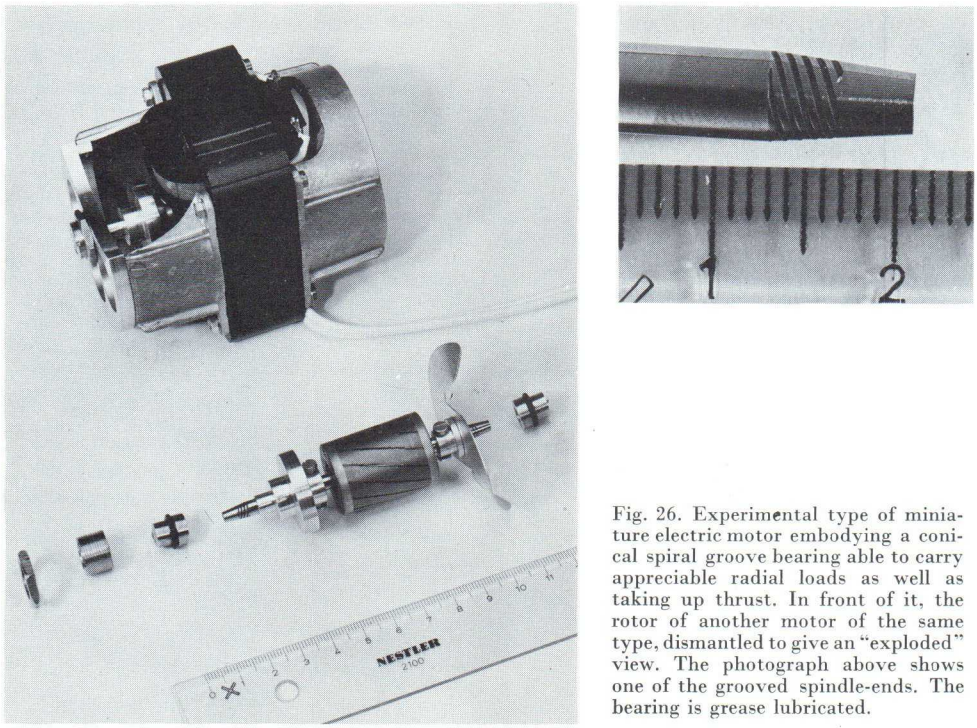
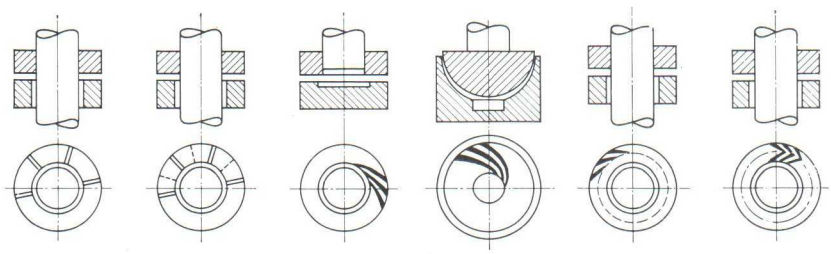


Fig. 26. Experimental type of miniature electric motor embodying a conical spiral groove bearing able to carry appreciable radial loads as well as taking up thrust. In front of it, the rotor of another motor of the same type, dismantled to give an "exploded" view. The photograph above shows one of the 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.

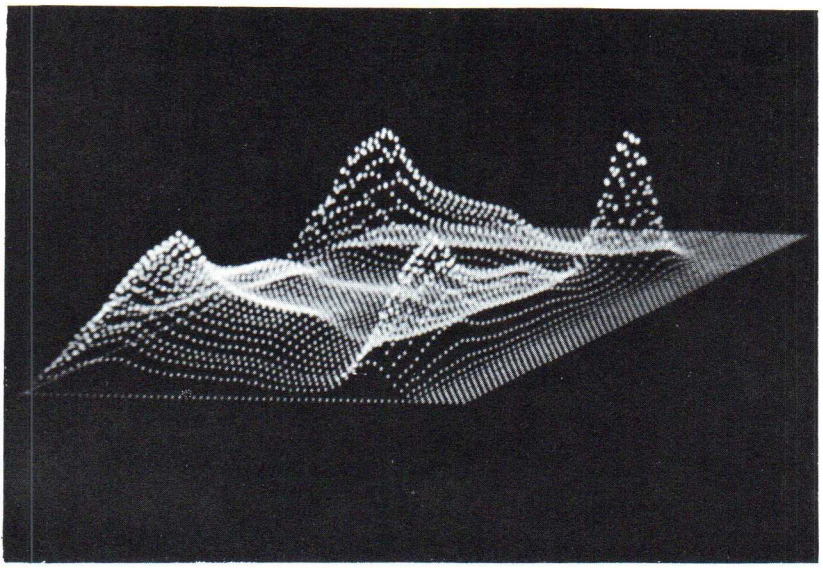
	Michell bearing	step bearing	spiral groove bearings			
			flat type for shaft-end	spherical cap	for mid-shaft support	
					partial	herringbone
						
K_1'	0.042	0.047	0.366	1.11	0.124	0.106
K_2'	22.0	ca. 25	2.9	2.8	8.1	10.2

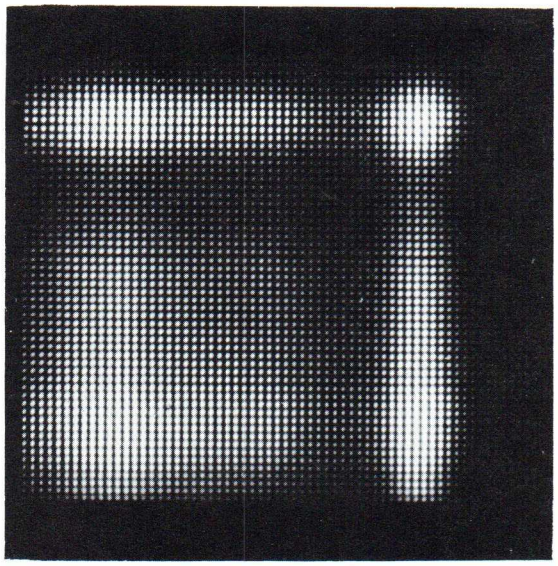
Though they will still have to undergo the test of time in practical applications, we personally feel that these miniature grease-lubricated contactless bearings have good future prospects.

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.

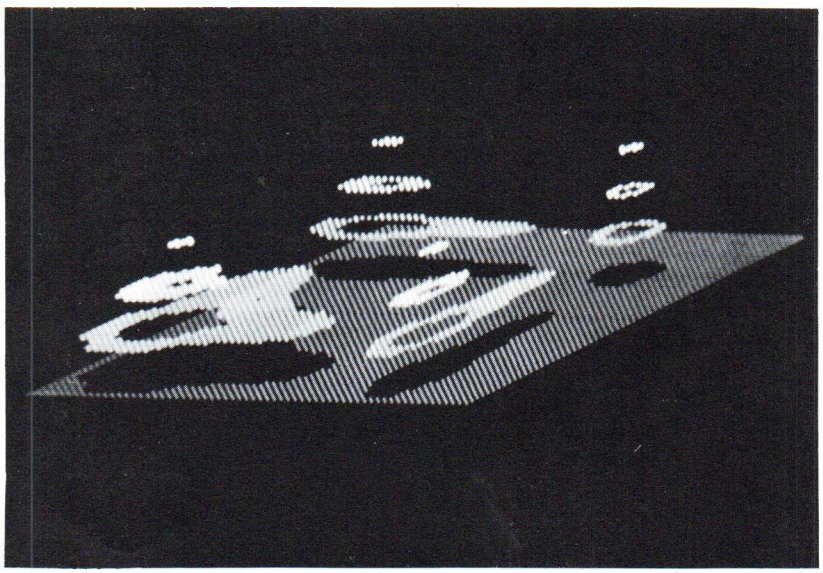
On the basis of the self-acting contactless bearing a new type — the spiral groove bearing — has been developed, this work having largely been done in Philips Research Laboratories at Eindhoven. The main advantage of the new type is its very low coefficient of friction. A flat thrust bearing of the spiral 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.

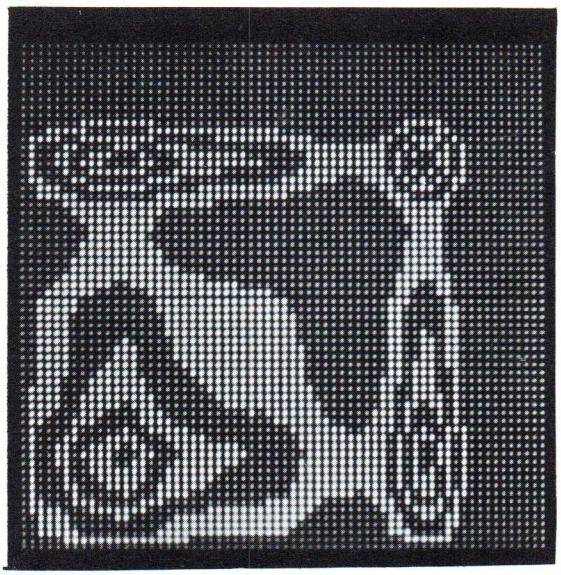
Display of nuclear radiation spectra





b





d

These pictures were obtained from an experimental 4096-channel kick sorter designed in the Eindhoven Research Laboratories for analysing nuclear radiation as a function of *two* variables, one of them being the energy of an emitted particle, the other, say, its angle of emission, or instead the energy of a simultaneously emitted second particle. The resulting three-dimensional spectra can be displayed on the screen of a cathode ray tube in a variety of ways, for instance *a*) as a perspective view of the surface; *b*) as a projection from above, height being translated into intensity; *c*) as (*a*) but reducing the surface to a kind of step function; *d*) as a projection of (*c*) from above, successive "steps" being shown as alternately light and dark.

(See: F. Bregman, Display of three-dimensional energy spectra, Philips tech. Rev. **27**, 132-133, 1966.)

DIRECT OBSERVATION OF WEISS DOMAINS BY MEANS OF THE FARADAY EFFECT

538.614

By cooling a nearly eutectic melt, large plate-like single crystals of the hexagonal compound $BaFe_{12}O_{19}$, the main constituent of ferroxdure, can be obtained. In this process, now in use for some years in the Philips Research Laboratories [1]), extremely thin crystal plates may be obtained under certain conditions. Their thickness is less then a micron and their surface area of the order of one square millimetre. X-ray diffraction photographs of these tiny platelets have shown that they too are single crystals of $BaFe_{12}O_{19}$, with the hexagonal axis perpendicular to the flat surface.

[1]) A. L. Stuijts, to be published elsewhere.

The special interest of these platelets is that they are reasonably transparent to red light. (This is related to the high resistivity of $BaFe_{12}O_{19}$, a characteristic which is due to the fact that all iron ions present in this compound have valency 3. The presence of even small quantities of iron ions of valency 2 besides those of valency 3 will in general cause a much lower resistivity and a strong absorption of light.) The transparency of the platelets opens up an elegant method for the direct observation of the Weiss domains in $BaFe_{12}O_{19}$. The platelet is observed in a polarizing microscope with nearly-crossed nicols. One then sees an image such as those depicted here in *figs. 1-3*.

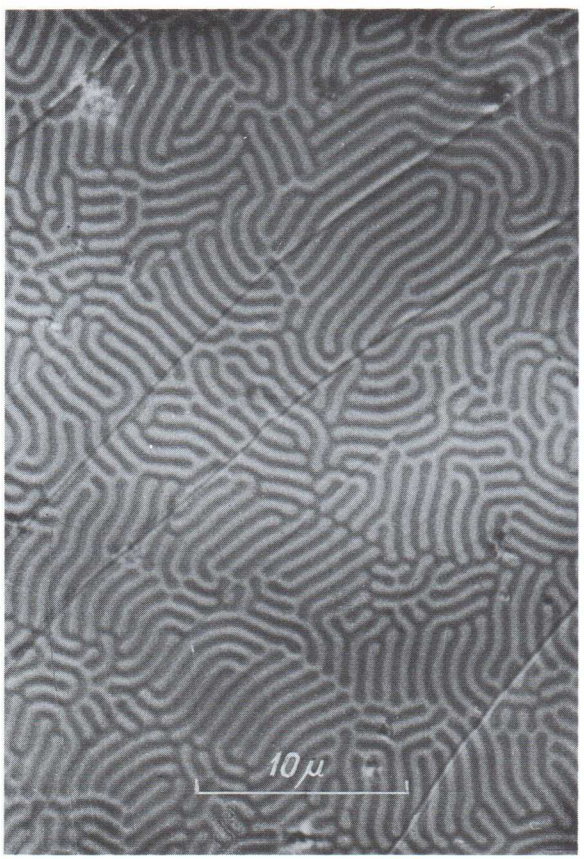


Fig. 1.

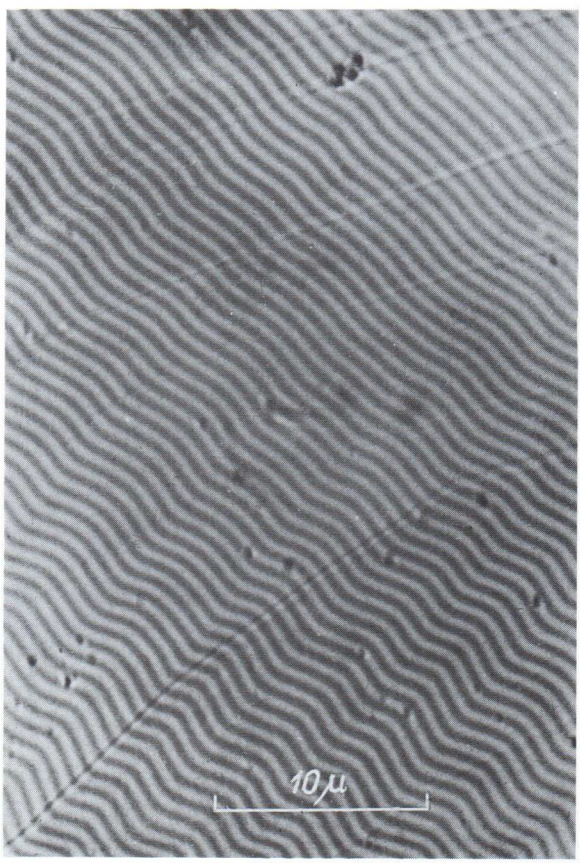


Fig. 2.

Fig. 1. Single-crystal platelet ($< 1\ \mu$ thick) of $BaFe_{12}O_{19}$ in the virgin state, as seen under a polarizing microscope with the nicols not quite exactly crossed. The light and dark bands are elongated Weiss domains.

Fig. 2. The same platelet as in fig. 1, after magnetizing. The domain pattern has assumed a much simpler form. (From scratches and spotty surface faults one can see that this photo represents almost the same region as that in fig. 1.)

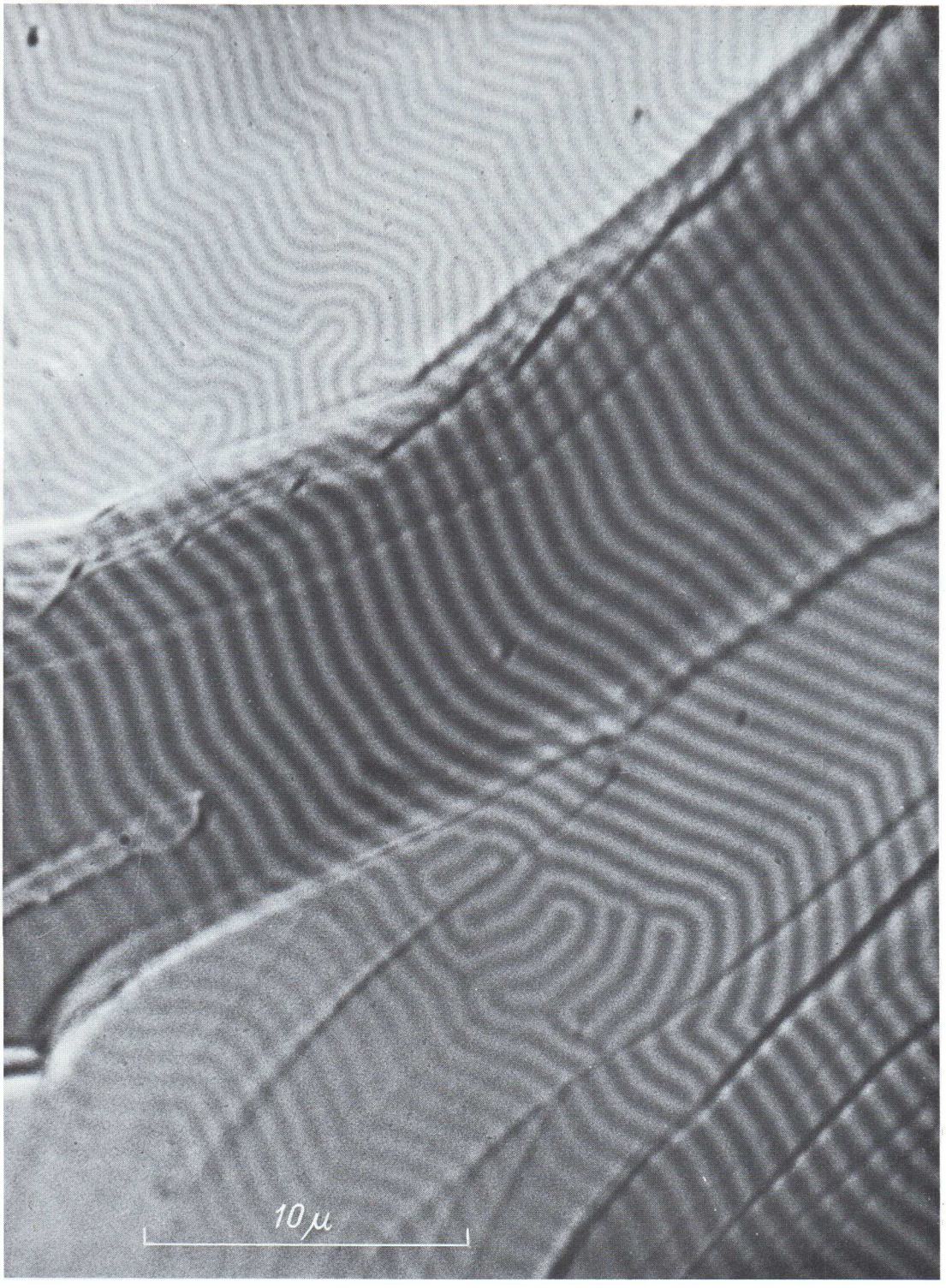


Fig. 3. Similar photo to that of fig. 2, showing another region of the same platelet. Terraced thickening of the specimen can be seen in this region by the dark broad bands across the pattern. The Weiss domains are seen to be wider within these dark bands. The "contraction" of the pattern on passing the boundary from a thick to a thinner place is at some points taken up by dislocation-like imperfections in the pattern.

A brief explanation of the occurrence of these patterns and an indication of some of the details to be seen may be appropriate here.

In each Weiss domain in a ferromagnetic material the magnetic moments of the atoms or ions are parallel aligned, so that in each domain the satura-

tion magnetization exists. The direction of magnetization, which depends on the crystal-shape- or stress-anisotropy, differs from one Weiss domain to another. In $BaFe_{12}O_{19}$ the direction of the magnetization is entirely determined by the very large crystal anisotropy; the preferred direction of magnetization is here the hexagonal axis (c-axis). In the crystal platelets therefore, the magnetization of all Weiss domains is perpendicular to the flat surface, in spite of the high demagnetization factor of the plate form. The Weiss domains are thus magnetized as sketched in *fig. 4*, with the domain walls all perpendicular to the flat surface. Because the magnetization is so strongly bound to the hexagonal axis in $BaFe_{12}O_{19}$, no "closing" domains occur at the surface (i.e. Weiss domains with magnetization parellel to the surface), in contrast with the usual situation in other materials.

If a beam of plane-polarized light is now projected parallel to the hexagonal axis on to such a platelet, a Faraday rotation occurs, of magnitude and sign dependent on the magnetization. Suppose that for one domain the plane of polarization undergoes a rotation described by a right-handed screw; the light passing through an adjacent domain, owing to the oppositely-directed magnetization, will then undergo an equal rotation of its plane of polarization in the opposite direction, i.e. corresponding to a left-handed screw. An analysing nicol is placed in the emerging beam. By placing this not exactly perpendicular to the plane of polarization of the incident beam, the intensity of the rays passing through adjacent, oppositely-magnetized domains will differ slightly; see *fig. 5*. In this way the whole domain pattern becomes visible [2]).

The pictures shown in figs. 1-3 were made in this manner. The hexagonal axis of the crystal lies perpendicular to the plane of the photographs. The Weiss domains, with their magnetizations directed alternately upwards and downwards out of the plane of the picture, are seen as light and dark bands. The remarkable labyrinth structure of fig. 1 is the domain pattern in the *virgin state*, as occurs after cooling the material from a temperature above the Curie point. Fig. 2 is a photo of practically the same region as that in fig. 1, taken after magnetizing the specimen in a field of 48×10^4 A/m [6000 oersteds] parallel to the hexagonal axis. The domain pattern has here assumed a much simpler form; the ribbon-like domains are seen as wavy lines. The remanent magnetization of the platelet is practically zero as a result of the very large demagnetization factor with such a thin specimen. This can be seen in the photo since the light and dark bands are equally wide; in the total volume, therefore, their is no surplus of one direction of magnetization over the other.

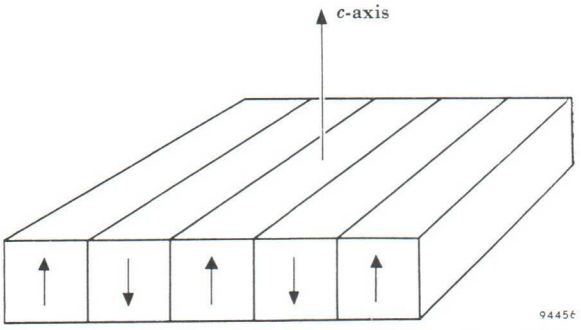


Fig. 4. Schematic diagram showing the configuration of Weiss domains in the single-crystal platelets of $BaFe_{12}O_{19}$. The domain walls are perpendicular to the flat surface of the platelets, parallel to the hexagonal c-axis. The arrows show the directions of the magnetization.

The *width* of the Weiss domains depends on the thickness of the crystal plate. The formation of the domain walls requires a certain energy, proportional to the wall area; on the other hand the magnetostatic energy of the whole body is reduced by the formation of the domain walls. With increasing thickness of the plate the magnetostatic energy becomes smaller, whereas the domain walls would take up more energy owing to their greater area (greater height) if their spacing remained the same. The total energy

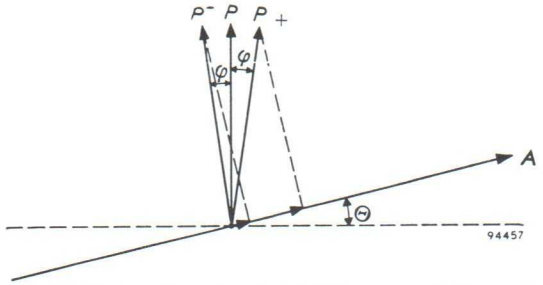


Fig. 5. If P is the direction of polarization of a light beam (here considered incident perpendicularly to the paper), the Faraday effect in one Weiss domain gives rise to a rotation φ to $P+$, while in adjacent oppositely-magnetized domains the same rotation occurs in the opposite sense, to P^-. An analyser set with its transmission direction at A, $90° - \Theta$ from P, attenuates the light with the polarization P^- more than that with the polarization $P+$.

The transmitted intensities are related to each other as $\sin^2(\Theta - \varphi) : \sin^2(\Theta + \varphi)$. Maximum contrast is obtained at $\Theta = \varphi$, but for a small value of φ the transmitted intensity is too small when $\Theta = \varphi$. One therefore takes a somewhat larger value of Θ at which the image gives a good compromise between intensity and contrast.

[2]) For methods for the *surface* observation of Weiss domains (Bitter patterns, magneto-optical Kerr effect, etc.) see, for example, R. M. Bozorth, J. Phys. et Rad. **12**, 308, 1951 and B. W. Roberts, Conf. Magnetism and Mag. Mat., Pittsburgh, Pa., 14-16 June 1955, p. 192.

The method using the Faraday effect has also been used by P. A. Miles, who employed iron oxides of garnet structure as the transparent ferromagnetic (Prog. Rep. M.I.T. Lab. Insul. Res. **21**, 17, 1957).

is therefore minimized by a reduction of the number of domain walls per cm length, i.e. the domains become broader. This effect is strikingly demonstrated in fig. 3. This photo is analogous to fig. 2 but represents another region of the specimen. The platelet has here a terraced change in thickness which can be seen (owing to the increased absorption) as broad dark bands passing through the whole domain pattern. Inside the dark bands the Weiss domains are seen to be broader than in the lighter regions.

The elongated Weiss domains run continuously through the light and dark regions and apparently are not affected by the "steps" on the surface. To enable this continuity to occur, with a different number of domains per cm length in adjacent regions, some geometrical rearrangement is necessary. One can see how this occurs in the pattern; at the boundary of *A* and *B*, "extra" Weiss domains can be seen to end at various places on the boundary, showing a certain resemblance to edge dislocations in a crystal lattice. At one place on the boundary of *B* and *C* a region of meandering Weiss domains occurs, which serves to fill in a space which has arisen owing to the contraction of the domains in *C*.

Finally, a remark concerning the Faraday effect which is responsible for making the domain pattern visible. The Faraday effect in ferromagnetics has been known for some years, viz. in ferrites which are transparent to electromagnetic waves of centimetre and decimetre wavelengths. From the theory of the propagation of these waves in ferromagnetics, due to Polder [3]), a rotation of the plane of polarization is to be expected in the presence of a suitable external magnetic field; this is a result of the interaction between the precessing electron spins (aligned by mutual interaction) and the rotating *magnetic* field of a circularly polarized wave with the right sense of rotation. This is the same interaction as causes ferromagnetic resonance. The Faraday effect in ferrites has been usefully applied in ferrite isolators for microwaves [4]). One is tempted to attribute the Faraday rotation in $BaFe_{12}O_{19}$ at optical wavelengths, with which we are concerned here, to the same mechanism as that in ferrite isolators at microwave frequencies. This is not borne out quantitatively, however. A calculation of the rotation to be expected on the basis of magnitude of the spin-field interaction (known from other effects), results in a value about 100 times smaller than that observed. The Faraday effect in the present case must in fact be attributed to the interaction between the orbital electrons (whose orbits are aligned by the spin-orbit interaction in a ferromagnetic) and the rotating *electric* field of the circularly polarized light beam [5]). This is essentially the same mechanism as that responsible for the normal Faraday effect with light waves in diamagnetics. In the ferromagnetic $BaFe_{12}O_{19}$ the effect is so strong that even a thickness of 1 micron provides an observable effect.

C. KOOY.

[3]) D. Polder, On the theory of ferromagnetic resonance, Phil. Mag. **40**, 99-115, 1949. See also H. G. Beljers and J. L. Snoek, Gyromagnetic phenomena occurring with ferrites, Philips tech. Rev. **11**, 313-322, 1949/50.

[4]) C. L. Hogan, The ferromagnetic Faraday effect at microwave frequencies and its applications. The microwave gyrator; Bell Syst. tech. J. **31**, 1-31, 1952.
A. G. Fox, S. E. Miller and M. T. Weiss, Behavior and applications of ferrites in the microwave region, Bell Syst. tech. J. **34**, 5-103, 1955. See also H. G. Beljers, The application of ferroxcube in unidirectional waveguides and its bearing on the principle of reciprocity, Philips tech. Rev. **18**, 158-166, 1956/57 (No. 6).

[5]) H. R. Hulme, Proc. Roy. Soc. **A 135**, 237, 1932.

A METHOD OF GROWING DISLOCATION-FREE GERMANIUM CRYSTALS

by B. OKKERSE.

548.571:546.289

The importance of research into the properties and technology of the semiconductors germanium and silicon is evident when one considers their application in the manufacture of diodes and transistors. The article below describes the method by which single crystals of germanium entirely free of dislocations have been made in Philips Eindhoven laboratories. This article links up with a previous article in this journal, which dealt with the purification and doping of germanium and silicon.

An important aspect of solid-state research at the present time is the study of the way in which the macroscopic physical and physico-chemical properties of crystals depend on the physical imperfections or defects of the crystal lattice. Principal among these defects are: 1) the absence of atoms at sites in the crystal lattice that should be occupied (*vacancies*), 2) the presence of atoms at places that are not lattice sites (*interstitial atoms*) and 3) linear lattice defects or *dislocations*.

The presence of dislocations in the lattice of a single crystal may affect, for example, the mechanical properties, the diffusion of foreign atoms, the rate of alloying and dissolution and also, in the case of semiconductors such as silicon and germanium, the life time of injected charge carriers. Further, depending on the circumstances, the dislocation may act as a sink or as a source of both vacancies and interstitial atoms, and in this way too it may detract from the perfection of the crystal. As far as the semiconductors referred to above are concerned — other materials will be left out of account in this article — research on the relation between crystal imperfections and physical properties is important not only to fundamental science but also to technology, in connection with the manufacture of diodes and transistors.

For the purpose of such research it would be a great help to have a material in which one particular type of lattice defect is entirely absent. This requirement, however, can be fulfilled only in respect of one type of defect: since point defects are thermodynamically stable, only the absence of dislocations enters into consideration.

In this respect germanium is a promising material. In germanium crystals, grown by conventional methods, the dislocation density, i.e. the total length of the dislocation lines in one cubic centimetre, measured in cm/cm^3 or in cm^{-2}, is about 1000 times less than in the most perfect *metal* crystals, the respective orders of magnitude being 10^4 cm^{-2} and 10^7 cm^{-2}. Some time ago it in fact proved possible to make single crystals of germanium entirely free of dislocations. We shall first describe the procedure adopted and the physical principles on which it was based. We shall then discuss the methods by which the crystals made were examined for the absence of dislocations, and we shall present the arguments which lead us to believe that the crystals were indeed free from dislocations. Finally we shall comment briefly on the perfection of the dislocation-free crystals thus obtained.

Method of preparing dislocation-free germanium crystals

The method evolved for preparing dislocation-free single crystals of germanium is simply a special method of crystal-pulling. The pulling apparatus can either be used with an ordinary crucible or with the recently developed floating crucible [1]). After the rotating seed crystal has been brought into contact with the liquid germanium and the pulling process started, the diameter of the growing crystal is gradually reduced to 1 or 2 mm. This is done by gradually raising the temperature of the melt a few degrees. The new situation is thenceforth maintained over a length of growth of about 20 mm. Finally, in a similar way, the diameter is again gradually increased to the required value (*fig. 1*) and the pulling operation is continued until the crystal is the right length. The pulling rate chosen is less than about 1 mm/min, compared with 2 or 3 mm/min in normal crystal-pulling. The flow rate of the inert gas circulating around the crystal must be kept carefully constant. It is also important that there should be no particles of solid matter floating

[1]) The method of pulling a single crystal from a melt was recently described in this journal in an article by J. Goorissen, Segregation and distribution of impurities in the preparation of germanium and silicon, Philips tech. Rev. **21**, 185-195, 1959/60 (No. 7), in which references to other literature on the subject will be found.

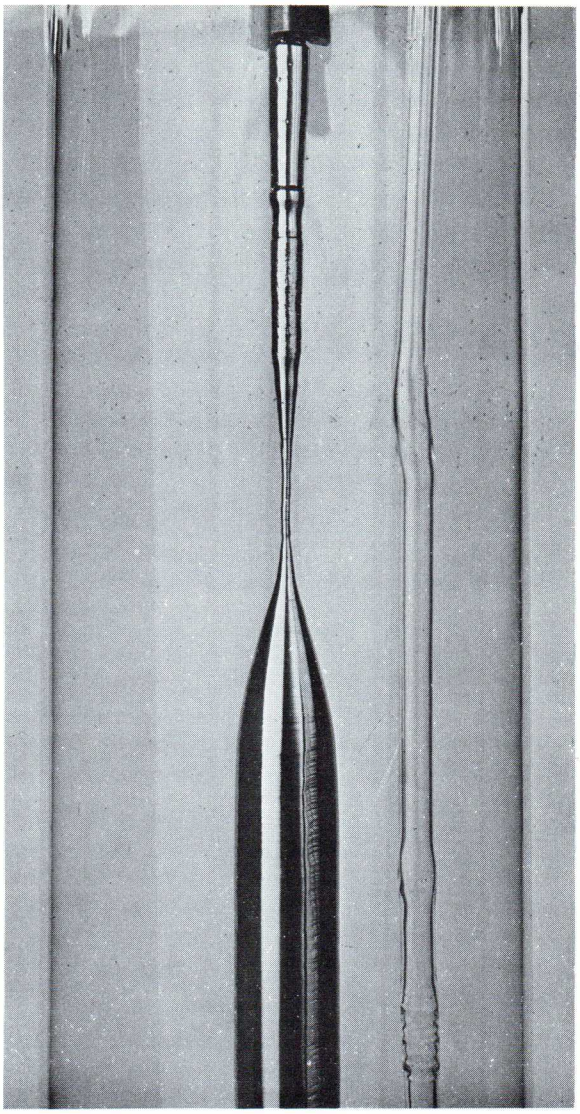


Fig. 1. Dislocation-free single crystals of germanium can be made by the pulling method if the diameter of the first part of the growing crystal is reduced to 1 or 2 mm and this diameter is maintained over a length of about 20 mm. The photograph shows from top to bottom: the seed holder, the seed, the narrowed portion, and finally the dislocation-free portion whose diameter is first gradually increased and subsequently held fixed. (The irregularities in the contours, particularly of the thin portion, are mainly due to the fact that the crystal was photographed through the wall of the glass envelope.)

on the surface of the molten germanium (e.g. germanium oxide or graphite from the crucible).

Physical considerations

The occurrence of dislocations in a crystal presupposes the presence of dislocation sources and mechanical strains. The dislocation sources may be inside the crystal as well as on the surface. A dislocation pinned at two points in the crystal can act as an internal source. A source of this kind can, in principle, give rise to large numbers of ring-shaped dislocations [2]). The nature of the surface sources is not yet clear.

With regard to the mechanical stresses required for the generation of dislocations, it may be noted that in a crystal grown by pulling from the melt these can only be caused by the presence of a temperature gradient. It can be demonstrated [3]) that a temperature gradient produces no stresses only when 1) the heat flux through a crystal is constant, 2) there are no radial heat-losses, and 3) the rate of growth is zero. Obviously, these conditions cannot be fully satisfied in practice. Fortunately, this is not necessary, since the occurrence of growth stresses leads firstly to a merely elastic deformation of the crystal lattice, plastic deformation setting in, i.e. dislocations being created, only when the stress exceeds a certain value [2]).

The reason why the method of pulling described above produces dislocation-free crystals may now be explained as follows. By initially reducing the diameter to 1 or 2 mm, and by pulling the crystal slowly, we approach the conditions that must be satisfied for the thermal stresses to be zero. Evidently, this is done in such a way that the lattice undergoes only elastic deformation and the generation of new dislocations from the existing sources is prevented. The dislocations present in the seed crystal are prevented from growing further down the crystal by making the thin section of the crystal long enough: those dislocations which intersect the solid-liquid interface at a given moment will normally grow further into the subsequently solidifying part of the crystal, but the thin neck of crystal increases the chance that they will intersect the crystal surface and thus "grow out" of the crystal (cf. *fig. 2*).

In the next part of the pulling operation, during which the diameter of the crystal is again enlarged and finally held constant at a relatively large value, the thermal stresses are certainly not small. It is difficult, however, to generate dislocations in a dislocation-free germanium crystal. Even if such a crystal is thermally "mishandled", e.g. by pulling it abruptly out of the melt during growth,dislocations are still not produced in it [4]). If, on the other hand, this experiment is done on a crystal having a dislocation density of about 10^4 cm^{-2}, the density increases to 10^6 or 10^7 cm^{-2}.

A remarkable effect is found if the melt contains

---

[2]) The mechanism of this process is described e.g. by H. G. van Bueren, Philips tech. Rev. **15**, 255, 1953.

[3]) P. Penning, Philips Res. Repts. **13**, 79, 1958. See also P. Penning, The generation of dislocations by thermal stresses, Philips tech. Rev. **19**, 357-364, 1957/58.

[4]) The same was found by Dash for silicon: W. C. Dash, J. appl. Phys. **30**, 459, 1959.

floating solid particles and these collide against the growing crystal. The part of the crystal produced after the moment of contact is found to contain a continually expanding network of dislocations, originating at the point of collision. It appears that contact with a foreign particle creates dislocation sources in the crystal surface. As the crystal goes on growing, these sources, as a result of thermal stresses, give rise to more dislocations, which later act in their turn as new sources, and so on. The existence of dislocation sources of this type has only recently become evident. From the experiment referred to in the previous paragraph, in which the

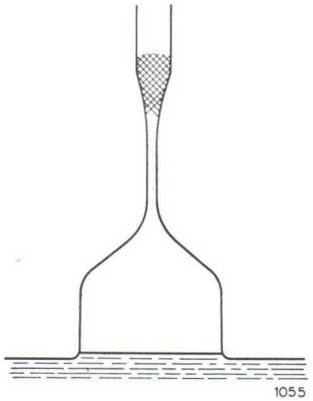


Fig. 2. In the necked portion of the crystal the dislocations as they grow with the crystal are bound to intersect the surface, where they must necessarily end; new dislocations cannot form because the thermal stresses are too small. In the lower portion, of large diameter, the thermal stresses are considerable, but now they create no dislocations because there are no longer any dislocation sources present.

crystal was suddenly cooled by withdrawing it from the melt, it appears that surface sources of this nature do not occur in a crystal grown from a melt whose surface is clean.

Evidence for the absence of dislocations

In order to demonstrate that a crystal is free of dislocations, reliable methods are needed for detecting the presence of dislocations. One of the best available methods is that based on etch pits. If a germanium crystal is cut parallel to a (111) plane and etched with a suitable agent [5]), it is found to dissolve quicker at places where a dislocation intersects the surface than elsewhere, so that after some time "etch pits" form at these places. This phenomenon is attributed to the fact that the crystal lattice around the dislocation is deformed and thus

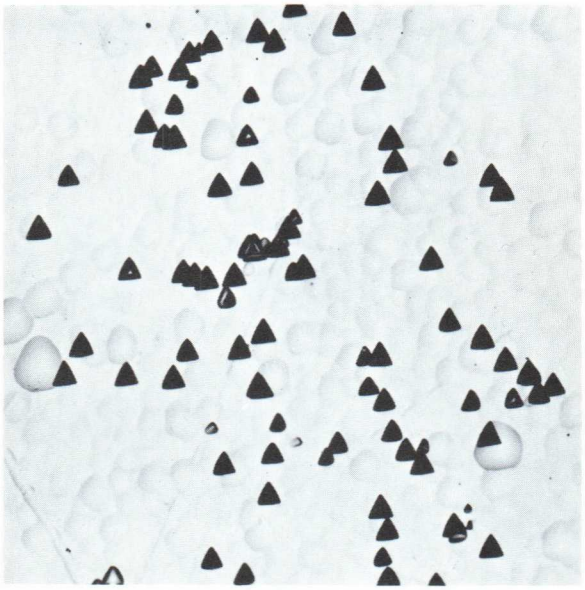


Fig. 3. If a germanium crystal is cut parallel to a (111) plane and etched with a suitable agent, the crystal dissolves quicker at the places where dislocations intersect the crystal surface then elsewhere. Each resultant etch pit corresponds to a dislocation. The sides of the triangular periphery of an etch pit are lines of intersection of the other (111) planes with the crystal surface. (Magnification 100 times.)

possesses at such locations a higher energy per unit volume. The etch pits can easily be observed and counted under a microscope (*figs. 3* and *4*).

This etching procedure is simple to use and has proved very reliable. By various methods it has been shown that in Ge, at least, all dislocations are

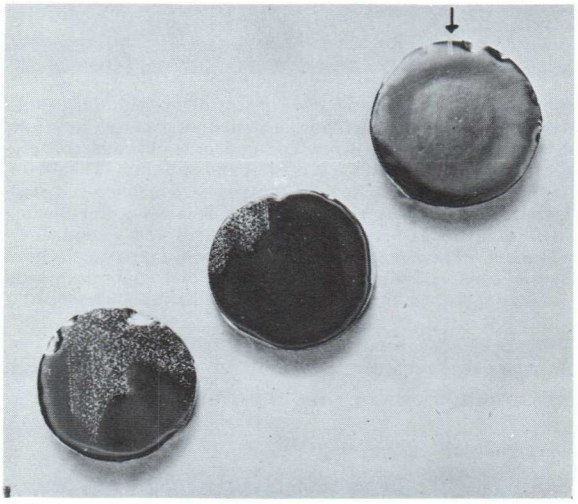


Fig. 4. Etched surface of three discs cut from an originally dislocation-free germanium crystal which, during the pulling process, came into contact with a piece of graphite on the melt. The disc on the right was cut at the point of collision; the two others were cut successively 5 mm further in the subsequently solidified part of the crystal. At the point of collision (see arrow) dislocation sources were evidently produced, from which stems a growing network of dislocations.

[5]) Use can be made of the etchant described by E. Billig, Proc. Roy. Soc. **A 229**, 346, 1955, or of the etchant known as CP4, the composition of which is described in the article by Vogel et al. under footnote [6]).

shown up in this way [6]). Various experiments carried out by us have also led to the same conclusion: the dislocation densities determined by the etching method on germanium crystals were compared with the results of two other methods. The first is based on the anomalous transmission of X-rays — an effect found only in perfect or nearly perfect crystals — and the second consists of determining the rate at which copper precipitates in a germanium crystal supersaturated with copper; this precipitation rate is governed solely by the dislocation density and the temperature (see appendix). At all the dislocation densities measurable by these methods — the lower limit of detection in our case was about 50 cm^{-2} for both methods, the upper limit for the X-ray method is some thousands of cm^{-2} — the results showed excellent agreement with those of the etch-pit method.

After etching, nearly all the germanium crystals grown by the method described above were found to be entirely free of etch pits (*fig. 5*). For this reason we believe we may conclude that these crystals are free of dislocations. In the few that were not dislocation-free, the dislocation density is several times

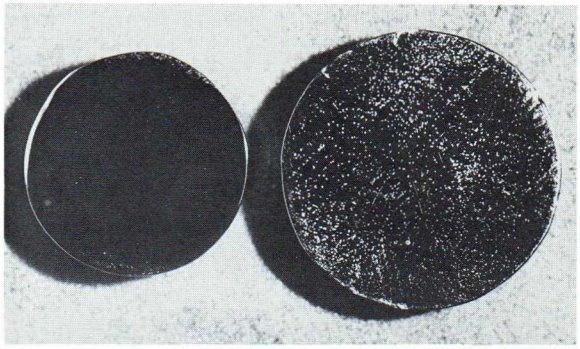


Fig. 5. Etched surface of a germanium crystal pulled by the method described (left) and of a crystal made without special precautions during the pulling process. The left crystal shows no etch pits, and this warrants the conclusion that it contains no dislocations.

10^3 cm^{-2}. The negative result in these cases is most probably due to a trivial cause, such as the presence of an overlooked piece of solid matter on the surface of the melt.

It might be asked whether the above conclusion is indeed justified, and whether it is not conceivable that the crystals appear to be dislocation-free only because the dislocations present do not happen to intersect the etched surface. The answer is that the chance of this happening in the crystals made by the method described is exceptionally small. As already noted, the thermal stresses in a crystal of normal diameter (1 to 3 cm) are considerable. As a result, the dislocation density, unless it is zero, will always be high (10^3 to 10^4 cm^{-2}) even though there was originally only one source. The chance that no single one of these dislocations will intersect the etched surface may be considered to be zero — indeed, even at a dislocation density of, say, 10 cm^{-2}, this chance would be negligible. If not a single etch pit is found in a crystal, it is therefore highly probable that the crystal is entirely free of dislocations.

In passing it may be remarked that it follows from this reasoning that, in order to demonstrate the absence of dislocations in these crystals, it is not necessary that the detection method should detect every dislocation. In practice the etch-pit method is commonly used, not so much because of its sensitivity but because it is simple.

Various investigations have shown that dislocation-free germanium crystals are not perfect. During etching, most samples do not dissolve at a uniform rate at all parts of a cross-section (see fig. 4). Moreover, the anomalous transmission of X-rays through bent crystals is found to depend strongly on position. These effects indicate that the material is not homogeneous. This is probably attributable to a non-uniform distribution of the vacancies (or vacancy agglomerates) with which these crystals are in fact supersaturated. Since there are no dislocations that can function as sinks (see appendix) the vacancies in dislocation-free crystals must diffuse all the way to the crystal surface in order to disappear. Large numbers of vacancies have insufficient opportunity to do so during the cooling of the crystal and are therefore frozen-in.

Appendix: Determining the dislocation density from the anomalous transmission of X-rays and from the precipitation rate of copper

Anomalous transmission of X-rays [7])

The transmission of X-rays incident at the Bragg angle on a perfect crystal, i.e. at the angle Θ that satisfies the Bragg law, can be described by resolving the wave phenomenon in the crystal into two components (*fig. 6*). In a direction parallel to the lattice planes that reflects the X-ray beam, both components have the character of a travelling wave; in the direction perpendicular thereto they behave like a standing wave having a period equal to the lattice spacing d. Upon emerging from the crystal the two components split into a transmitted

[6]) See F. L. Vogel, W. G. Pfann, H. E. Corey and E. E. Thomas, Phys. Rev. **90**, 489, 1953 and W. C. Dash, J. appl. Phys. **27**, 1193, 1956.

[7]) See L. P. Hunter, J. appl. Phys. **30**, 874, 1959, where references are given to earlier literature.

and a diffracted beam. The amplitude pattern of the one component (A) is now such that the minima coincide with the atomic planes and the maxima fall in between. There is therefore virtually no absorption of this component. On the other hand the maxima of the other component (B) coincide with the atomic planes, so that this component is strongly absorbed.

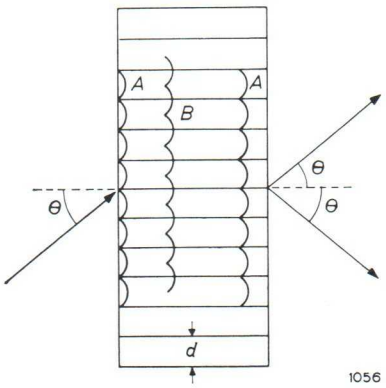


Fig. 6. Illustrating the anomalous transmission of X-rays incident at the Bragg angle on a group of parallel lattice planes of a nearly perfect crystal.

If the crystal is not perfect, e.g. owing to the presence of dislocations, then the first component, too, can suffer a degree of absorption. It is found that the intensity of the transmitted and of the diffracted beam both decrease markedly as the dislocation density of the crystal increases. The lowest dislocation density that could be demonstrated in this way with the apparatus used was about 50 cm^{-2}; it should certainly be possible, however, to lower this limit.

Calculating the dislocation density from the precipitation rate of copper

The possibility of determining the dislocation density in germanium by measuring the precipitation rate of copper is based on the interaction between dislocations and vacancies on the one hand and between the dissolved copper atoms and vacancies in the germanium on the other [8]).

Copper atoms can be found in two kinds of site in the germanium lattice; they can replace a germanium atom in the lattice (substitution) or they may occupy an interstitial site. We shall denote atoms of the first kind by Cu_s, and atoms of the second kind by Cu_i. The concentrations of these atoms are not independent of each other; there exists an equilibrium $Cu_s \rightleftarrows Cu_i + V$. In this equation V represents the vacancy produced when a substitutional copper atom abandons its lattice site for an interstitial site. The substitutional copper atoms are strongly bound to their sites (the diffusion constant is approximately 10^{-11} cm^2/s) as are the vacancies. The interstitial atoms, on the other hand, are extremely mobile (diffusion constant approximately 10^{-5} cm^2/s).

If we now saturate a germanium crystal at 750 °C with copper and then, after cooling, we raise the crystal to a temperature of 500 °C, the crystal at the latter temperature will be about 1000-fold supersaturated with substitutional copper atoms. (The solubility of substitutional copper at 750 °C and 500 °C is respectively 10^{16} and 10^{13} atoms per cm^3. The concentration of the interstitial atoms can be disregarded, since, owing to the high mobility of the interstitial atoms, it never differs significantly from the equilibrium value.) The excess substitutional copper atoms can only disappear from the crystal by first becoming interstitial. Since the vacancies thereby produced can diffuse but slowly to a site capable of functioning as a "sink" for vacancies — such as a dislocation, or the surface of the crystal — a state of quasi-equilibrium arises in which the ratio between the concentrations of the substitutional copper atoms and the vacancies is constant. The rate at which the excess copper disappears, and thus at which the solution changes from a supersaturated to the normal saturated condition, is therefore determined by the rate at which the vacancies can be removed. The latter rate depends in its turn on the number of sinks present, i.e. on the dislocation density.

Conversely, then, one can determine the dislocation density by measuring at a certain temperature T (in our case 500 °C) the time constant τ of the exponential time-variation of the substitutional Cu concentration towards its equilibrium value. In practice this is done by examining the way in which the resistivity of the crystal varies with time during the period in which the crystal is held at the temperature T. Clearly, the value of τ thus found also depends on this temperature.

It is found that the product of τ and the dislocation density N has a constant value at a given temperature. The relation between τ and $1/T$, as found experimentally by Tweet [8]), is shown in *fig. 7*. It is seen from this figure that at 500 °C the product $\tau \times N$ has roughly the value 10^9. Thus, if we find at a temperature of 500 °C a τ value of, say, 10^4, we may conclude that the dislocation density is approximately 10^5 cm^{-2}. The smallest density at which the presence of dislocations can be detected by this method is about 50 cm^{-2}, as in the previous case.

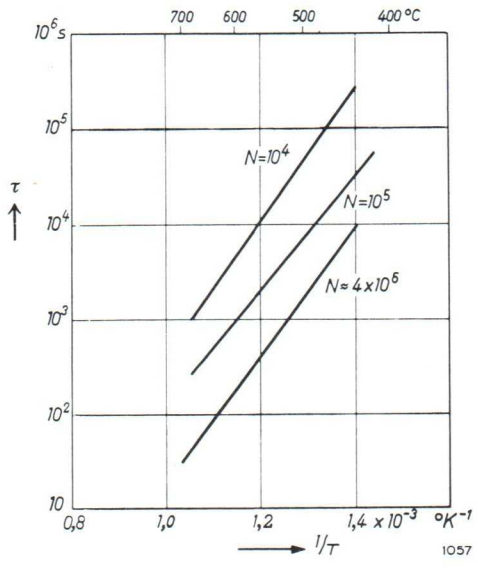


Fig. 7. The time constant τ of the rate of precipitation of copper in supersaturated germanium, plotted as a function of the reciprocal of the absolute temperature T, for three germanium crystals of differing dislocation densities N (cm^{-2}). (After Tweet [8]).)

Finally, a word about the way in which a dislocation can function as a sink for vacancies. It is known that an edge dislocation may be thought of as arising from the insertion of an extra atomic plane, the edge of which is located inside the

[8]) A. G. Tweet, Phys. Rev. **106**, 221, 1957, and P. Penning, Philips Res. Repts. **13**, 17, 1958.

crystal. Now it is very well possible that an atom at the edge of this plane may break away and occupy the site of a neighbouring vacancy site. This is particularly the case if the dislocation contains a jog: the latter then moves along the dislocation by one lattice site (*fig. 8*). A vacancy moves along a dislocation relatively fast, and the time needed to reach such a jog is equal, to a first approximation, to the time needed for a vacancy to diffuse to the dislocation.

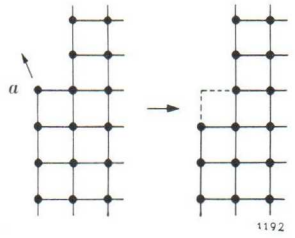


Fig. 8. An edge dislocation with a "jog" in it readily serves as a sink for vacancies. The drawing shows a part of the extra atomic half-plane of the dislocation containing a jog (left). If the atom *a* breaks away to occupy the site of a neighbouring vacancy, the jog shifts laterally along the dislocation by one lattice site as shown on the right.

Summary. Dislocation-free germanium crystals can be produced by the pulling method if the diameter of the crystal grown from the seed is initially reduced to 1 or 2 mm. This diameter is maintained over a length of about 20 mm, after which it is gradually raised to the desired value. In the thin neck of crystal thus produced the dislocations present tend to intersect the surface somewhere, where they end; thermal stresses are too small to generate fresh dislocations. The thermal stresses generally increase as the crystal diameter is increased but no dislocations are created because of the absence of sources. However, should a piece of solid material floating on the melt come into contact with the growing crystal, dislocation sources are produced in the surface. Freedom from dislocations is demonstrated by the etch-pit method. Dislocation-free single crystals still contain point defects.

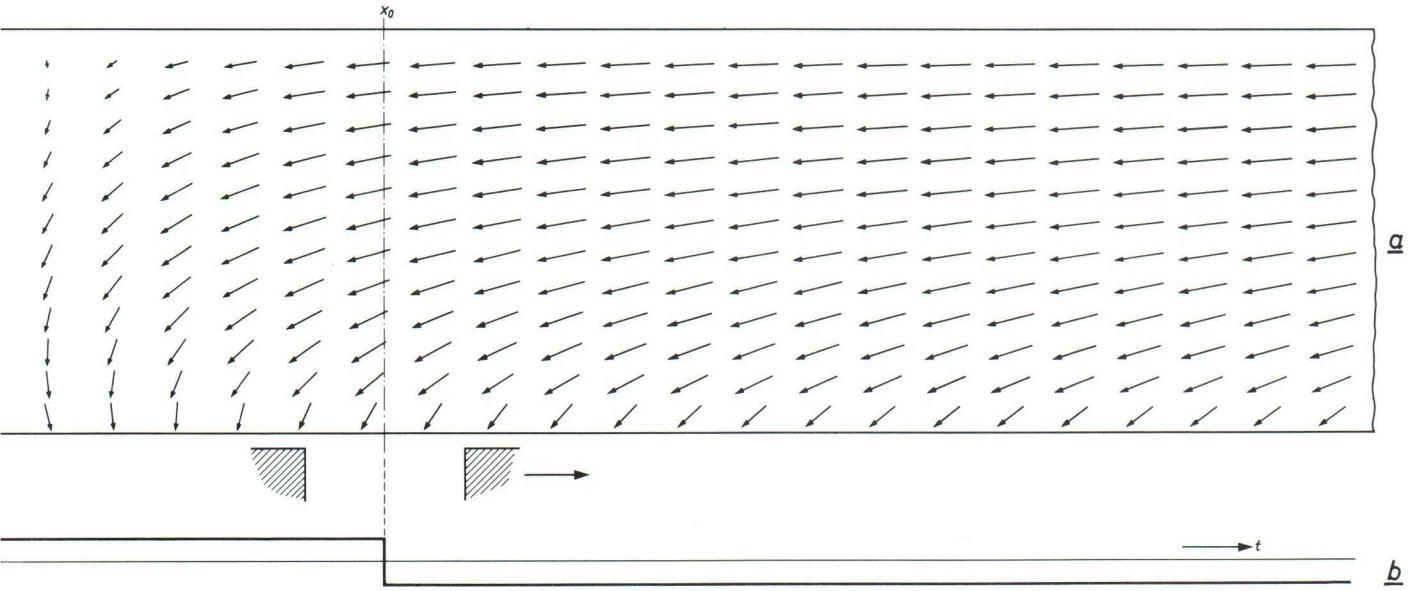
Magnetic recording in a 5000 : 1 scale model

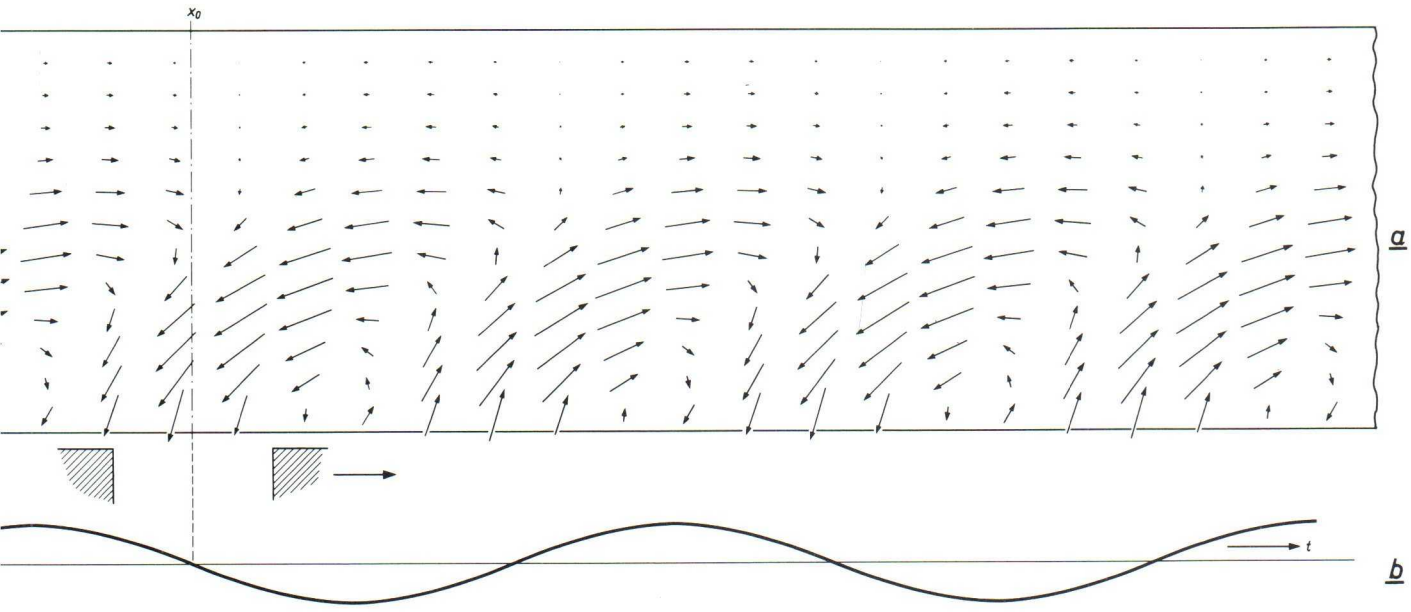
The vector fields shown in these drawings represent states of magnetization in a *longitudinal cross-section* of a magnetic tape. The state represented in the upper strip was obtained when a step-function signal, shown in (*b*), was recorded on the tape. The discontinuity occurred when the recording head was at the location x_0 in relation to the tape. The state represented in the lower strip was obtained by the recording of a sinusoidal signal, shown in (*b*) (note the larger scale used in this case for plotting the magnetization M).

The data for these drawings were obtained from actual measurements on samples taken from a magnetic "tape" 5 cm thick (i.e. scaled up 5000 times), which had been fed through a "recorder" with a recording-head whose relevant dimensions were scaled up by the same factor.

(See: D.L.A.Tjaden and J.Leyten, A 5000 : 1 scale model of the magnetic recording process, Philips tech. Rev. **25**, 319-329, 1963/64.)




x_0
a
t
b


x_0
a
t
b

A METHOD OF MEASURING THE RESISTIVITY AND HALL COEFFICIENT ON LAMELLAE OF ARBITRARY SHAPE

621.317.331:538.632.083

Resistivity and Hall-coefficient measurements at different temperatures play an important part in research on semiconductors, such as germanium and silicon [1]), for it is from these quantities that the mobility and concentration of the charge carriers are found.

Such measurements are commonly carried out with a test bar as illustrated in *fig. 1*. The resistivity is found directly from the potential difference and the distance between the contacts O and P, the current i and the dimensions of the bar. To determine the Hall coefficient the bar is subjected to a magnetic field B applied at right angles to the direction of the current and to the line connecting the diametrically opposite contacts O and Q. From the potential difference thus produced between these latter contacts the Hall coefficient is derived. (The relation between the Hall coefficient and the change in electric potential distribution due to a magnetic field will be explained presently.)

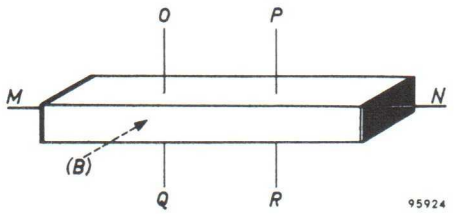


Fig. 1. Classical form of sample used for resistivity and Hall-coefficient measurements. The test bar is provided with current contacts M and N and voltage contacts O, P, Q and R. The fourth voltage contact R serves for check measurements.

In measurements performed at low temperatures — e.g. in liquid nitrogen — point contacts possess resistances of the order of megohms, in consequence of which the voltages cannot be determined with sufficient accuracy. In such cases "bridge-shaped" samples are used as illustrated in *fig. 2*. The voltage and current contacts here have a relatively large surface area, and hence the contact resistances are low.

The methods referred to are based on the fact that the geometry of the sample ensures a simple pattern of virtually parallel current stream-lines. Formulae have been devised to correct for the deviation from parallelism in fig. 2, caused by the finite width of the arms. A drawback of the bridge-shaped sample is that it is rather difficult to make, having to be cut out of the brittle semiconductor material with an ultrasonic tool. There is therefore a considerable risk of breakage, particularly when the arms are made narrow.

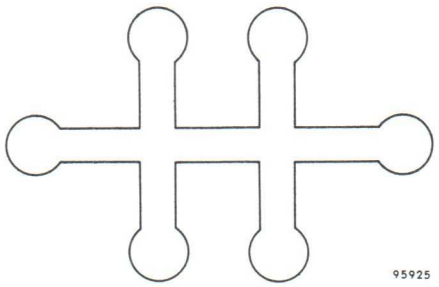


Fig. 2. The bridge-type sample, which is provided with relatively large contact faces to reduce contact resistances. This form is of special importance in measurements at low temperatures.

In the following we shall describe a method of performing resistivity and Hall-coefficient measurements on lamellae of arbitrary shape [2]). The lamella must not, however, contain any (geometrical) holes.

New method of measuring resistivity

We take a flat lamella, completely *free of holes*, and provide it with four *small* contacts, M, N, O and P, at arbitrary places *on the periphery* (*fig. 3*). We apply a current i_{MN} to contact M and take it off at contact N. We measure the potential difference $V_P - V_O$ and define:

$$R_{MN,OP} = \frac{V_P - V_O}{i_{MN}}.$$

Analogously we define:

$$R_{NO,PM} = \frac{V_M - V_P}{i_{NO}}.$$

The new method of measurement is based on the theorem that between $R_{MN,OP}$ and $R_{NO,PM}$ there exists the simple relation:

$$\exp\left(-\frac{\pi d}{\varrho} R_{MN,OP}\right) + \exp\left(-\frac{\pi d}{\varrho} R_{NO,PM}\right) = 1, \quad (1)$$

where d is the thickness of the lamella and ϱ the

[1]) See e.g. C. Kittel, Introduction to solid state physics, 2nd edition, Wiley, New York 1956, Chapter 13, p. 347 et seq.

[2]) L. J. van der Pauw, A method of measuring specific resistivity and Hall effect of discs of arbitrary shape, Philips Res. Repts. **13**, 1-9, 1958 (No. 1).

resistivity of the material. If d and the "resistances" $R_{MN,OP}$ and $R_{NO,PM}$ are known, then (1) yields an equation in which ϱ is the only unknown quantity.

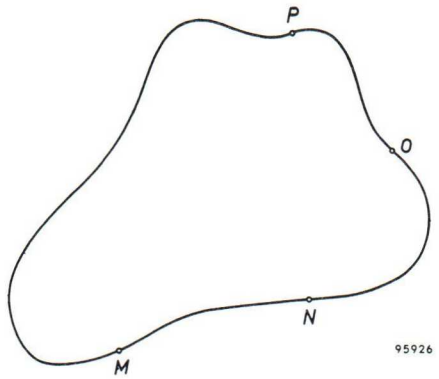


Fig. 3. A flat lamella of arbitrary shape, with four contacts M, N, O and P on the periphery, as used in the new method of measuring resistivity. The Hall coefficient can also be measured on a sample of this kind.

The situation is particularly straightforward if the sample possesses a line of symmetry. In that case, M and O are placed on the line of symmetry, while N and P are disposed symmetrically with respect to this line (*fig. 4*). Then:

$$R_{NO,PM} = R_{MN,OP}, \quad . \quad . \quad . \quad . \quad . \quad (2)$$

which may be seen as follows. From the reciprocity theorem for passive fourpoles, we have quite generally that $R_{NO,PM} = R_{PM,NO}$ (interchange of current and voltage contacts), and it follows from the symmetry that $R_{PM,NO} = R_{MN,OP}$. Hence we arrive at (2); ϱ can then easily be found from (1):

$$\varrho = \frac{\pi d}{\ln 2} R_{MN,OP}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

In this case a single measurement suffices.

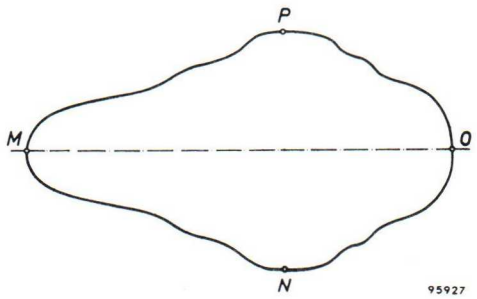


Fig. 4. The resistivity measurement is simplified if the sample has a line of symmetry. If two of the contacts are situated on the line of symmetry and the two others are symmetrically situated with respect to this line, one measurement is sufficient to give the required resistivity.

In the general case it is not possible to express ϱ explicitly in known functions. The solution can, however, be written in the form

$$\varrho = \frac{\pi d}{\ln 2} \frac{R_{MN,OP} + R_{NO,PM}}{2} f, \quad . \quad . \quad . \quad (4)$$

where f is a factor which is a function only of the ratio $R_{MN,OP}/R_{NO,PM}$, as plotted in *fig. 5*. Thus, to determine ϱ, we first calculate $R_{MN,OP}/R_{NO,PM}$, read from fig. 5 the corresponding value of f and then find ϱ from (4).

Photographs of samples as used for the old and for the new method are shown in *fig. 6*.

The complete proof of the theorem underlying the measurement of ϱ is given in the paper quoted in footnote [2]). The proof consists of two parts. First of all, relation (1) is developed for a special case, the case of a lamella in the form of an infinite half-plane, provided with four contacts at the periphery. It is then shown that the relation must also apply to a lamella of any shape. This is done by means of conformal mapping of the arbitrarily shaped plate on the infinite half-plane with the aid of complex functions.

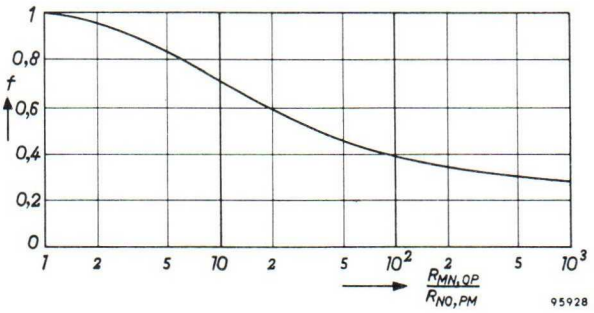


Fig. 5. Between the factor f in formula (4) and the ratio $R_{MN,OP}/R_{NO,PM}$ there exists the relation:

$$\cosh\left\{\frac{(R_{MN,OP}/R_{NO,PM}) - 1}{(R_{MN,OP}/R_{NO,PM}) + 1} \frac{\ln 2}{f}\right\} = \tfrac{1}{2} \exp \frac{\ln 2}{f},$$

which is represented here graphically.

We shall consider the first part of the proof in more detail, since it reveals the origin of the exponential functions in (1).

We first consider a lamella which extends to infinity in all directions. To a point M we apply a current $2i$, which flows away from M with radial symmetry into infinity. Let d again be the thickness of the lamella and ϱ the resistivity. Then at a distance r from M the current density is

$$J = 2i/2\pi rd.$$

The field-strength E is radially oriented and according to the generalized form of Ohm's law has the value:

$$E = \varrho J = \varrho i/\pi rd.$$

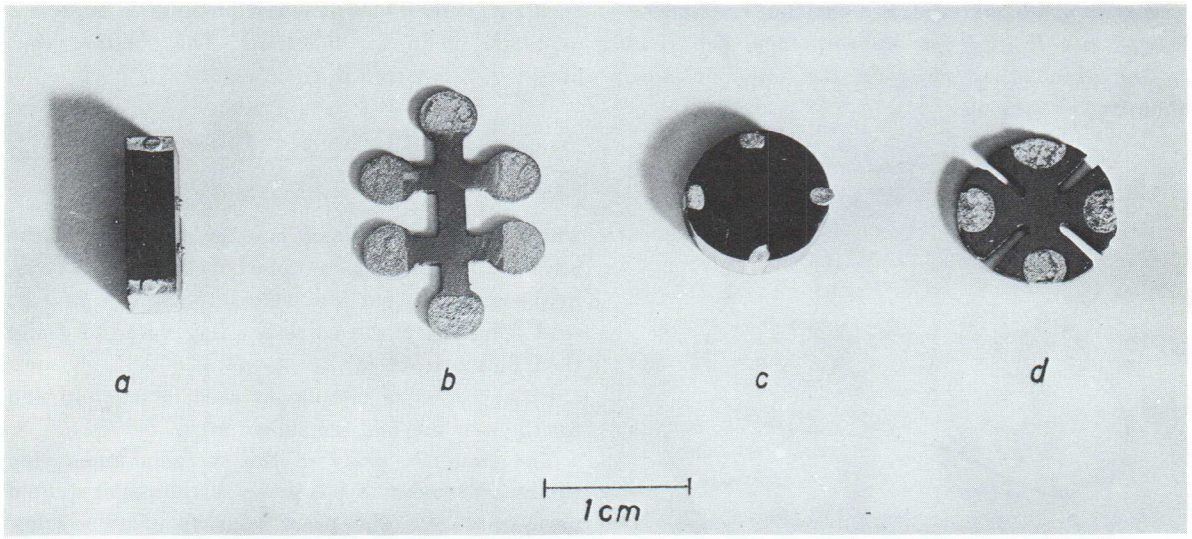


Fig. 6. Some samples of silicon used for resistivity and Hall-coefficient measurements. Samples *a* and *b* correspond respectively to figs. 1 and 2. Measurements on samples *c* and *d* are possible only by the new method. The incisions in sample *d* serve to reduce the error caused by the contacts not being infinitely small.

The potential difference between two points O and P lying on a straight line with M (*fig. 7a*) is:

$$V_P - V_O = \int_P^O E\,\mathrm{d}r = \frac{\varrho i}{\pi d}\int_P^O \frac{\mathrm{d}r}{r} = -\frac{\varrho i}{\pi d}\ln\frac{a+b+c}{a+b}.$$

Since no current flows in the direction perpendicular to the line through M, O and P, the result obtained remains valid if we omit the part of the lamella at one side of this line — yielding a half-plane — and if at the same time we halve the current (fig. 7*b*).

Next we consider the case of *c*) in fig. 7, where a current i now flows *out* from a point N, that again lies on the same straight line with OP, viz. on the edge of the infinite half-plane. We now find:

$$V_P - V_O = +\frac{\varrho i}{\pi d}\ln\frac{b+c}{b}.$$

Superposition of the cases *b*) and *c*) in fig. 7 yields the case *d*), the current i being introduced at M and taken off at N. The value now assumed by $V_P - V_O$ is found by adding together the two previous results. After dividing by i we then find:

$$R_{MN,OP} = \frac{\varrho}{\pi d}\ln\frac{(a+b)\,(b+c)}{(a+b+c)b},$$

or

$$\frac{(a+b+c)b}{(a+b)\,(b+c)} = \exp\left(-\frac{\pi d}{\varrho}R_{MN,OP}\right).$$

Similarly we find:

$$\frac{a\,c}{(a+b)\,(b+c)} = \exp\left(-\frac{\pi d}{\varrho}R_{NO,PM}\right).$$

Addition of the last two equations yields (1).

We shall now explain how formula (4) follows from (1). For simplification we put:

$$\left.\begin{aligned}\pi\,d\,R_{MN,OP} &= x_1,\\ \pi\,d\,R_{NO,PM} &= x_2.\end{aligned}\right\} \quad \ldots\ldots\ldots \quad (5)$$

Formula (1) can then be written:

$$\mathrm{e}^{-\frac{x_1}{\varrho}} + \mathrm{e}^{-\frac{x_2}{\varrho}} = 1. \quad \ldots\ldots\ldots \quad (6)$$

Next we write:

$$x_1 = \tfrac{1}{2}\{(x_1 + x_2) + (x_1 - x_2)\}$$
$$\text{and}\quad x_2 = \tfrac{1}{2}\{(x_1 + x_2) - (x_1 - x_2)\},$$

whereby (6) takes the form:

$$\mathrm{e}^{-\frac{x_1+x_2}{2\varrho}}\left(\mathrm{e}^{-\frac{x_1-x_2}{2\varrho}} + \mathrm{e}^{+\frac{x_1-x_2}{2\varrho}}\right) = 1.$$

This is the same as:

$$\mathrm{e}^{-\frac{x_1+x_2}{2\varrho}}\cosh\frac{x_1-x_2}{2\varrho} = \tfrac{1}{2}. \quad \ldots\ldots\ldots \quad (7)$$

The exponent of e in (7) is now written as $-(\ln 2)/f$, i.e. we put:

$$\frac{x_1+x_2}{2\varrho} = \frac{\ln 2}{f}. \quad \ldots\ldots\ldots \quad (8)$$

Formula (7) then becomes:

$$\mathrm{e}^{-\frac{\ln 2}{f}}\cosh\left\{\frac{(x_1/x_2)-1}{(x_1/x_2)+1}\times\frac{\ln 2}{f}\right\} = \tfrac{1}{2}.$$

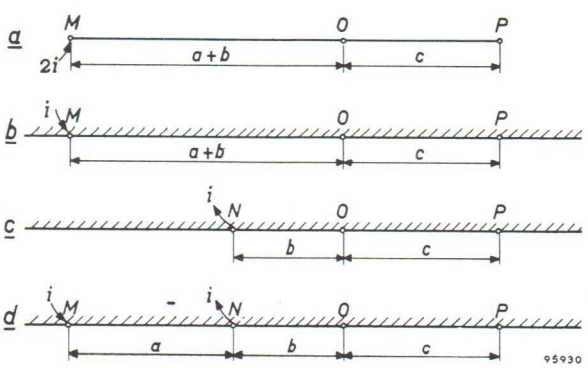


Fig. 7. Illustrating the derivation of formula (1).

This expression represents a relation between f and x_1/x_2, and hence also between f and $R_{MN,OP}/R_{NO,PM}$ (see 5). The relation is shown graphically in fig. 5. By re-writing (8) to give ϱ and substituting for x_1 and x_2 from (5), we find formula (4).

Method of measuring the Hall coefficient

The Hall coefficient, too, can be measured on an arbitrary lamella as in fig. 3. We then apply the current to one of the contacts, say M, and take it off at the contact following the succeeding one, i.e. in our case at O. We measure $R_{MO,NP}$, after which we set up an uniform magnetic induction B at right angles to the surface of the lamella. This changes $R_{MO,NP}$ by an amount $\Delta R_{MO,NP}$. We shall now denote the Hall coefficient R_H and show that it is given by:

$$R_H = \frac{d}{B}\,\Delta R_{MO,NP}, \quad \ldots\ldots \quad (9)$$

provided that:

a) the contacts are sufficiently small,
b) the contacts are on the periphery,
c) the lamella is of uniform thickness and free of holes.

The validity of formula (9) depends on the distribution of current stream-lines not changing when the magnetic field is applied. With samples of the classical shape of figs. 1 and 2, where the current stream-lines are always parallel to the edges of the sample, there is evidently no change. From the properties of the vector field representing the current density it follows that the same also applies to lamellae of arbitrary shape, provided the above conditions are satisfied [3]).

Under the magnetic induction B, the charge carriers, with charge q, are subjected to a force perpendicular to the stream-lines and perpendicular to the lines of magnetic induction. The magnitude of this force is $F = qvB$, where v is the velocity of the charge carriers. Between v, the concentration n of the charge carriers and the current density J there exists the relation $v = J/nq$. Dividing the force exerted on the charge carriers by their charge q, we see that the effect of the magnetic field is equivalent to an apparent electric field E_H, the Hall electric field, for which we can write [4]):

$$E_H = \frac{1}{nq} J B.$$

E_H is proportional to J and to B; the proportionality constant ($= 1/nq$) is called the Hall coefficient R_H.

Since q is known, one can calculate from R_H the concentration n of the charge carriers.

The fact that the current stream-lines are not affected by the magnetic field implies that after application of the magnetic field the electric field is no longer in the same direction as the current stream-lines, but has acquired a transverse component E_t which exactly compensates the apparent Hall electric field E_H (*fig. 8*). The change $\Delta(V_P - V_N)$ in the potential difference between P and N is therefore given by integrating E_t from P over a path orthogonal to the current stream-lines to N' across the lamella (*fig. 9*), and thence along the periphery — i.e. along a stream-line — from N' to N. The last portion of the path makes no contribution to the integral; hence

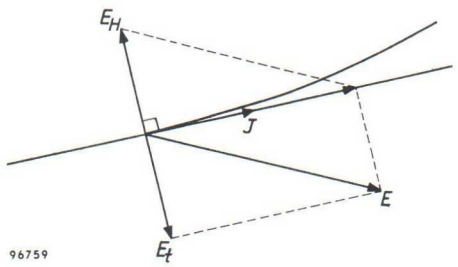


Fig. 8. The resultant of the electrical field-strength E and the Hall field-strength E_H lies in the direction of the current density J. Resolving E in directions perpendicular and parallel to J therefore yields a perpendicular component E_t which in magnitude is equal to E_H.

$$\Delta(V_P - V_N) = \int_P^{N'} E_H\,ds = R_H B \int_P^{N'} J\,ds = R_H B \frac{i_{MO}}{d},$$

where d is again the thickness of the lamella. This expression leads directly to (9).

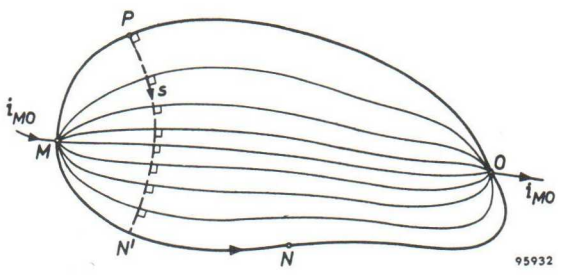


Fig. 9. To calculate by how much the potential difference between points P and N changes when a magnetic field is applied at right-angles to the plane of the sample, the transverse electric field E_t produced by the magnetic field is integrated along the path s which runs from P, orthogonal to the current stream-lines, to N' and thence along the periphery from N' to N.

[3]) The proof of this statement is also indicated in the paper quoted under [2]).

[4]) This formula is not entirely exact because, apart from their ordered motion with velocity v, the electrons also move randomly owing to thermal agitation. More precise calculation shows, however, that the formula given here is a good approximation.

Estimation of errors

In the foregoing we have assumed the contacts to be "sufficiently" small and to be situated on the periphery. To gain an idea of the error made in the calculations when these conditions are not exactly satisfied, we have worked out the error for three cases. For simplicity we consider a circular disc of diameter D with the contacts mutually 90° apart. We assume further that only one of the contacts is not ideal. The three cases are represented in the adjoining *table*, together with the formulae for the relative errors in the resistivity and the Hall coefficient. The cases are:

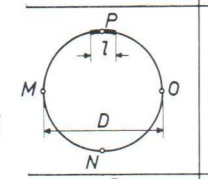
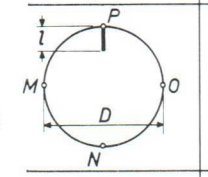
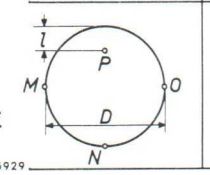
a) One of the contacts has a length l along the periphery.

b) One of the contacts has a length l perpendicular to the periphery.

c) One of the contacts, although a point, is situated at a distance l from the periphery.

In practice, none of the contacts will be ideal. To the first approximation the total error is then equal to the sum of the errors per contact.

Table. The relative errors $\Delta\varrho/\varrho$ and $\Delta R_H/R_H$ in the calculated values of the resistivity and the Hall coefficient for a circular disc of diameter D on which one contact P is non-ideal, in the senses indicated in the sketches.

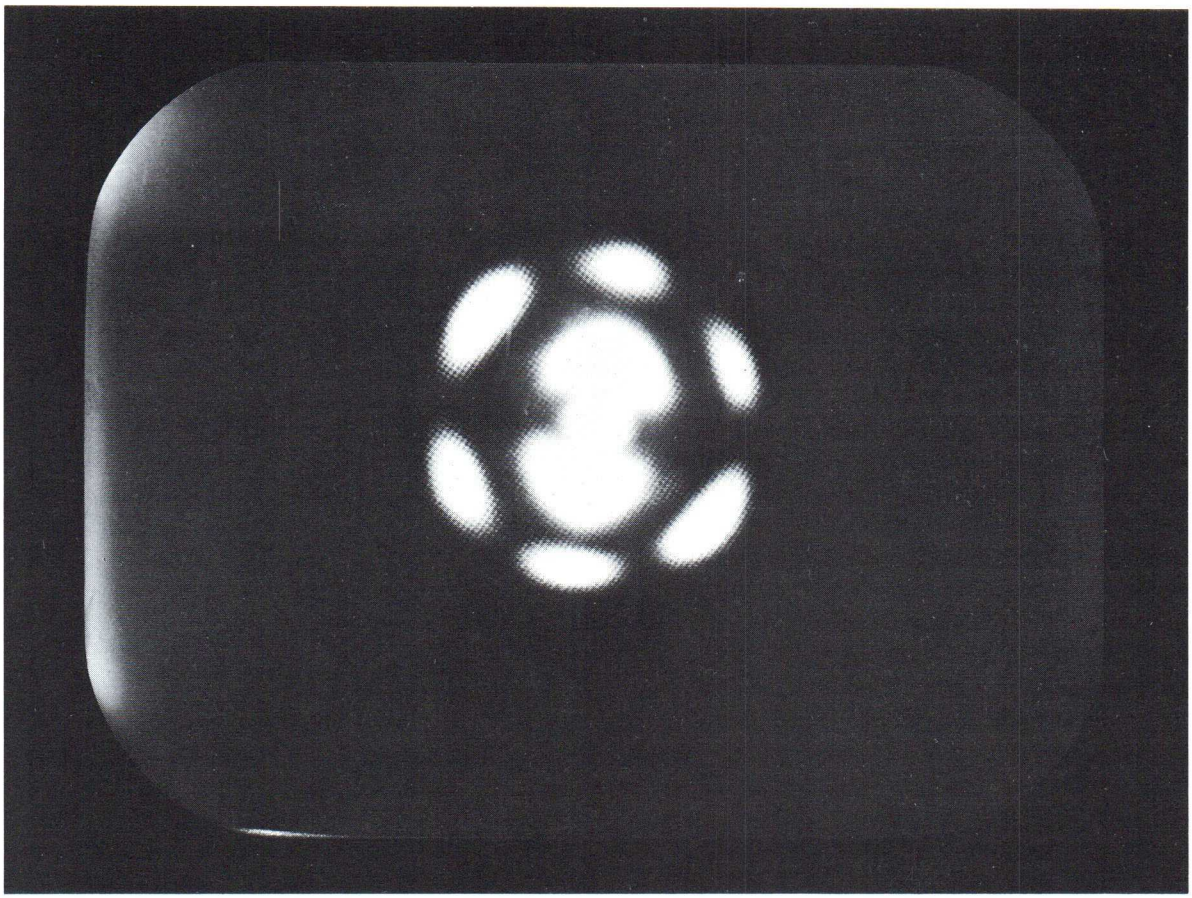
	$\Delta\varrho/\varrho$	$\Delta R_H/R_H$
a 	$\approx \frac{-l^2}{16D^2 \ln 2}$	$\approx \frac{-2l}{\pi^2 D}$
b 	$\approx \frac{-l^2}{4D^2 \ln 2}$	$\approx \frac{-4l}{\pi^2 D}$
c 	$\approx \frac{-l^2}{2D^2 \ln 2}$	$\approx \frac{-2l}{\pi D}$

95929

The value of the method described here lies in the fact that, in many cases, the material under investigation is already available in the form of small lamellae (e.g. thin discs sawn from a crystal rod); these samples now need no further preparation and can therefore be used for other purposes after the measurement.

If very small contacts are undesirable, having regard to contact resistances in measurements at low temperature, use can be made of "clover-leaf" samples (see fig. 6*d*), the incisions in which substantially reduce the error due to the finite dimensions of the contacts. The clover-leaf sample thus replaces the bridge-type sample (fig. 6*b*) which is used for the same purpose in the classical method. The clover-leaf sample is easier to make than the bridge-type sample and is also less susceptible to breakage.

L. J. van der PAUW.



A SMALL, STABLE GAS LASER

537.525:535.339

The electromagnetic radiation emitted by an unmodulated *radio transmitter* has the form of a single, practically continuous sine wave. It is thus coherent and monochromatic. The radiation emitted by a normal *light source*, e.g. a sodium lamp, on the other hand, consists (even if we only consider one single spectral line) of innumerable waves, emitted with random phase by the atoms involved. It is thus incoherent, and the spectral lines are relatively broad.

Now some years ago a new type of light source was introduced which, like a radio transmitter, emits a coherent and monochromatic radiation: the LASER. This name is formed from the initials of "Light Amplification by Stimulated Emission of Radiation". (One also sometimes speaks of an "optical MASER", where M stands for Microwave.) The above-mentioned properties, and yet others due to the special design of this light source (e.g. a sharp beaming of the emitted power), make the laser of great scientific interest and open wide perspectives for its practical application [1]).

[1]) A. L. Schawlow and C. H. Townes, Infrared and optical masers, Phys. Rev. **112**, 1940-1949, 1958. A good review, with extensive references, is to be found in W. Kaiser, Der optische Maser, Physica status solidi **2**, 1117-1145, 1962 (No. 9).

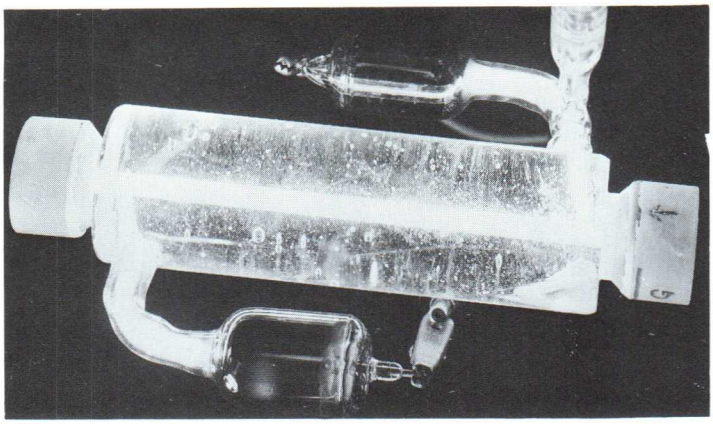


Fig. 1. The small gas laser. It consists of a heavy cylindrical block of fused-quartz 12 cm long, whose ends are polished flat to within 0.1 μm, with a hole only 3 mm in diameter bored along the axis. Two fused-quartz blocks which are polished flat to within even closer limits and are provided with infrared-reflecting interference mirrors (reflection coefficient 99%) are placed against the ends of the cylinder, where they give a vacuum-tight seal by adhesion alone. The mirrors thus do not need to be sealed on by the application of heat. The great thickness of the cylinder (35 mm) is necessary to give the system sufficient mechanical rigidity (no deformation due to e.g. gravity) and to ensure that the temperature distribution in the "resonant cavity" of the laser has very good cylindrical symmetry.

The electrical discharge in the gas mixture with which the tube is filled (85% He + 15% Ne, pressure 3 torr) is produced by applying a DC voltage between two electrodes which are sealed into side tubes which connect with transverse borings in the fused-quartz cylinder. The length of the active gas column between the two transverse borings is 10 cm.

Many research teams are therefore at present engaged in investigations on and development of the laser.

Our own work in this direction has recently led to a result which we would like to describe here briefly [2]). We hope to make it the subject of a more extensive article in due course.

The work in question concerns the *gas laser* (as opposed to the solid-state laser, which we shall not consider here). The stimulated-emission effect, on which the operation of the laser is based, consists in the fact that the probability of an atom making the transition from an excited state to a state of lower energy, and at the same time emitting radiation of the corresponding wavelength, is increased if the atom in question is situated in a field of radiation of the same wavelength; moreover, the transition then occurs at such a moment that the emitted radiation is *in phase* with the stimulating radiation. The latter is thus *amplified*. The amplification which can be obtained in a gas is generally small, e.g. a few percent per metre path length of the radiation in the gas. Very large amplifications however can be obtained in a relatively short gas column if parallel plane *mirrors* are placed at each end, and the distance between the mirrors is adjusted to exactly a whole number of half wavelengths of the radiation: the radiation can then pass to and fro many times through the gas, thus giving a kind of positive feedback exactly in phase; the resulting amplification may even be so great that self-excitation of the radiation becomes possible. We then have not a light *amplifier* (analogous to the maser for microwaves [3])) but an *oscillator* for light waves, i.e. a light *source* which can continuously emit radiation.

The stimulated-emission effect can only be perceived if there are enough atoms in the required excited state. This is ensured by means of a (usually fairly complicated) excitation mechanism with its own energy source ("optical pump"). In the gas laser in question, a mixture of neon and helium is used as the medium. The neon atoms give stimulated emission at a wavelength of 1.153 μm, i.e. in the near infrared.

[2]) See also: H. G. van Bueren, J. Haisma and H. de Lang, Physics Letters (Amsterdam) **2**, 340, 1962 (No. 7).

[3]) See e.g. R. W. DeGrasse, D. C. Hogg, E. A. Ohm and H. E. D. Scovil, Ultra-low-noise measurements using a horn reflector antenna and a traveling-wave maser, J. appl. Phys. **30**, 2013, 1959.

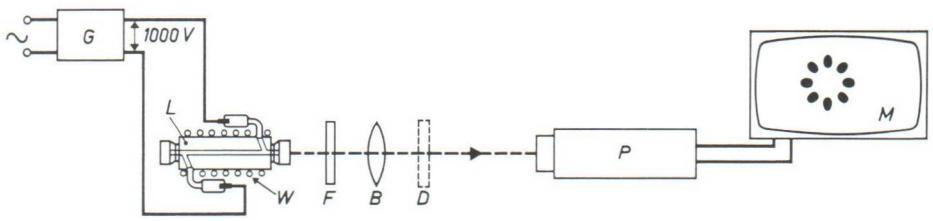


Fig. 2. The whole set-up. *L* laser tube, with voltage supply *G*. The laser radiation emerging from one of the end windows is focused on the photocathode of the infrared-sensitive lead-oxide vidicon *P* by means of the lens *B*. The filter *F* cuts off the visible radiation of the gas discharge, which is not due to laser action. *D* is a rotatable polarizer which is used to investigate the state of polarization of the emitted radiation. *M* television monitor on which the spatial distribution of the emitted laser radiation is made visible. *W* heater wire.

The helium atoms can take care of the pump effect if they in their turn are excited by an electric field. A mixture of neon and helium of suitable composition (other necessary conditions being fulfilled) can therefore be made to exhibit laser action if it is simply placed between parallel mirrors as described above and a gas discharge is produced in it [4]).

Originally, gas columns 50 to 100 cm in length were used in order to obtain sufficient amplification. There is much to be said for using shorter columns: the set-up then becomes much more manageable and less sensitive to mechanical vibrations, and it is easier to satisfy the condition that the mirrors should be accurately parallel. We have now succeeded in making a laser with a gas column only 10 cm long, owing among other things to the use of a narrow channel for the gas discharge, and to the adoption of a type of mirror with reflection and other losses of less than 1% and with deviations from flatness of not more than about 0.03 μm and of not more than 1 second of arc from the proper alignment. This tube is shown in *fig. 1*; some constructional details are given in the caption. The construction is very simple, and no difficult adjustments are required to obtain the laser action. The accurate adjustment of the distance between the two mirrors (which can also be regarded as bringing the electromagnetic resonant cavity, formed by the laser tube, into resonance) is carried out by means of a heating wire wound round the tube, which gives an adjustable expansion. The continuously radiated power amounts to some tenths of a milliwatt.

Fig. 2 shows the whole experimental set-up. The DC supply for the laser is very simple, as it only has to deliver about 10 watts at a voltage of about 1000 volts. The whole set-up is therefore easily transportable. The infrared laser radiation obtained is made visible by means of closed-circuit television using a lead-oxide vidicon ("Plumbicon") which has been developed in these laboratories and whose photoconductive layer is sensitive to infrared [5]). This layer is placed in the focal plane of a lens which is situated in line with the laser. An image is then produced on the screen of the TV picture tube e.g. like that shown in the title photograph. The regular pattern of spots of light seen here may be ascribed to narrow beams of light leaving the laser tube at very small angles, and which probably correspond to various standing waves in the tube (various resonance states due to different modes of vibration in the resonant cavity). The different beams of light also differ in their polarization direction. A detailed interpretation of these phenomena cannot be given here. The strong dependence of the laser effect on the length of the tube appears most clearly if one observes the image on the screen for some time during the warming-up or cooling-down of the tube: the pattern will then be seen to change gradually, owing to the appearance and disappearance of the various modes. This has been recorded on film, and we show a few frames of such a film in *fig. 3*.

J. HAISMA *),
S. J. van HOPPE *),
H. de LANG *),
J. van der WAL *).

[4]) A. Javan, W. R. Bennett, Jr. and D. R. Herriott, Population inversion and continuous optical maser oscillation in a gas discharge containing a He-Ne mixture, Phys. Rev. Letters **6**, 106-110, 1961 (No. 3).
[5]) E. F. de Haan *et al.*, not yet published.
*) Philips Research Laboratories, Eindhoven.

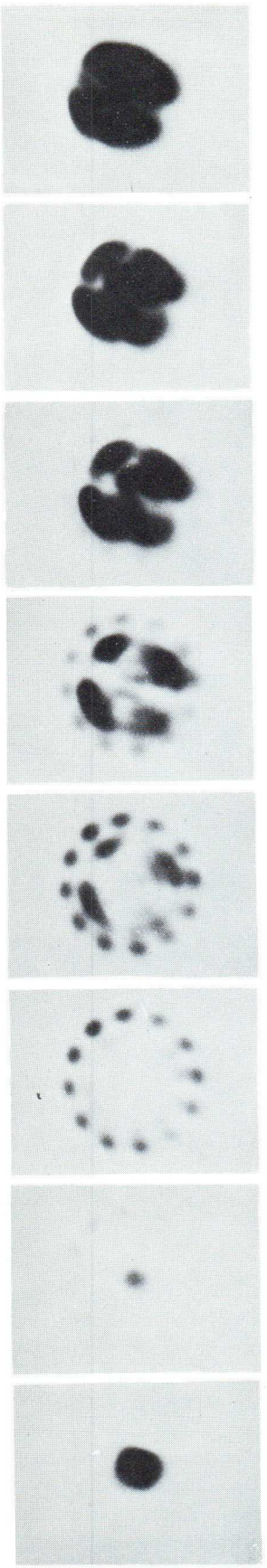
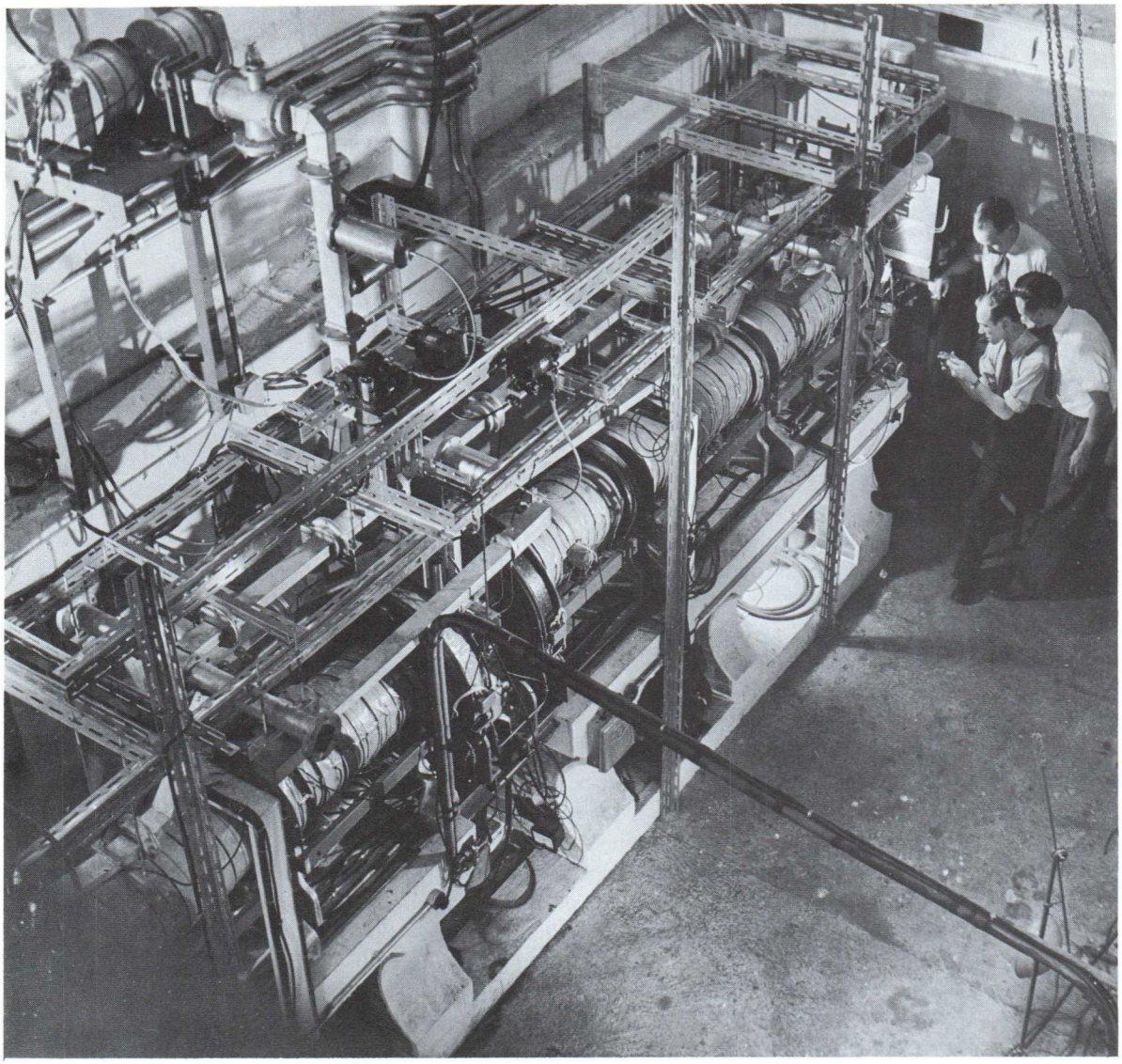


Fig. 3. Series of 16-mm film frames, as recorded from the screen of the TV tube. Time interval between successive frames 3 seconds.

15 MeV linear electron accelerator during construction



Photograph Walter Nurnberg

This electron accelerator, at the time of construction one of the largest of its kind, was designed and built by a team of scientists at the Mullard Research Laboratories, Redhill (England). The installation was ordered by the Atomic Energy Research Establishment at Harwell and was handed over to them during August 1952, after several months of testing. The photograph pictures the accelerator, whose length is more than 6 metres, during trial assembly in an underground shelter built especially for this purpose at M.R.L. The 10 cm waves which impart their energy to the particles injected into the accelerator are produced by a magnetron which is just visible in the upper left hand corner.

(From: Philips tech. Rev. **14**, 235, 1952/53. Further reading: C.F.Bareford and M.G.Kelliher, The 15 MeV linear electron accelerator for Harwell, A. Choice of parameters and general design, B. Detailed description of main components and circuits, C. Testing and performance, Philips tech. Rev. **15**, 1-26, 1953/54.)

A SIMPLE METHOD OF DETERMINING THE THERMAL CONDUCTIVITY OF SOLIDS

by J. SCHRÖDER *).

536.21.08

The remarkable feature of the method described is that it involves no temperature or quantitative heat measurements. The procedure, using a small cylindrical sample, consists merely in taking a stop-watch reading (of a few minutes) and finding a value on a calibration chart.

Unlike many other material properties, the thermal conductivity has been measured on only relatively few solid materials. Moreover, the values reported for one and the same material often show considerable discrepancies, sometimes 50% or more. These discrepancies are partly explained by differences in the composition, degree of purity, pretreatment, etc., of the samples measured. They stem largely, however, from the difficulties involved in measuring this property reliably and reproducibly. These difficulties may also explain why relatively so few solids have been subjected to thermal conductivity measurements, in spite of the growing importance of this property for scientific as well as technical reasons. It is of interest scientifically because no comprehensive theory has yet been put forward to explain the mechanism of (non-metallic) thermal conduction in solids. It is of technical importance in connection with all solid-state applications involving the conversion of thermal energy. In some applications, such as Peltier cooling and thermo-electric energy production, heat conduction in fact governs the efficiency that can be achieved.

The methods of determining thermal conductivity, λ, described in the literature, are too numerous to summarize here [1]). Common to all of them is the measurement of the temperature difference ΔT across a sample of length l — conveniently done with thermocouples — and of the quantity of heat Q flowing through the sample in a given time t. The thermal resistance R is then given by

$$\frac{1}{R} = \frac{A}{l}\lambda = \frac{Q}{t\Delta T}, \quad \ldots . \quad (1)$$

where A is the cross-section of the sample. The value of Q is usually derived from the electric power used for the heat supply, but it can also be measured calorimetrically. The most accurate values of λ are obtained by measurements in the steady state, to which (1) applies. This involves a lengthy measurement, however, since it is necessary to wait until this state is reached, and particularly since, to minimize heat losses, the sample and the heat source must be shielded and the shielding must also be brought to the requisite steady temperature. Moreover, these precautions entail the use of rather cumbersome equipment. In many methods, too, a good deal of time is spent on preparing the sample, so that all in all no more than one sample a day can be investigated.

In the following a new method is described, which uses simple equipment and makes it possible to measure the thermal conductivity of solids at room temperature quickly and accurately. An incidental advantage of the method is that the samples required are not only easy to make but can also be fairly small — a most important requirement where many solids are concerned and one which is certainly not met by certain older methods. The same method, with some supplementary equipment, can also be used for determining λ values at temperatures well above and below ambient, viz. between −200 and +400 °C.

The basis of the new method is the maintenance of a fixed temperature difference between the ends of the sample, the temperature of each end being held constant by contact with two boiling liquids. What is measured is the time in which, in the steady state, a certain quantity of heat passes through the sample; this quantity is established quite simple by the quantity of liquid made to evaporate at the "cold" end of the sample.

The principle of the method is illustrated in *fig. 1* [2]). In the lower vessel A a pure liquid a is brought to the boil. The vapour flows over the silver plate S_1, is condensed in the cooler K_1, and the condensate returns through the overflow pipe L into the vessel. In this way, plate S_1 is very effectively kept at a constant temperature, the boiling

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

[1]) See e.g. M. Jakob, Heat transfer, Wiley, New York 1949, Part 1, Chapter 9. Also Archiv für technisches Messen V 9213 1 to 4.

[2]) The apparatus will shortly be put on the market by Colora Messtechnik GmbH, Lorch/Würtemberg, Germany.

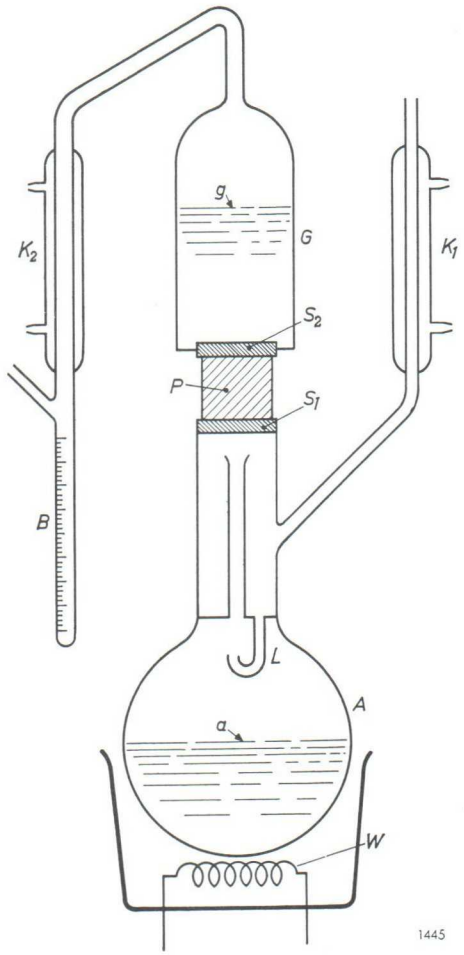


Fig. 1. Principle of the new method of determining the thermal conductivity of solids. P sample of material under investigation. A and G vessels containing two different liquids (a and g). W heating element that keeps a on the boil. S_2 and S_1 silver plates, ensuring good thermal contact of P with the liquid g and the vapour of a, respectively. K_1 condenser and L overflow pipe for returning the condensate of a. The liquid g, whose boiling point is roughly 10 °C lower than that of a, is brought to the boil by the heat transmitted through the sample P. K_2 condenser and B tube, graduated in millilitres, for collecting and indicating the vaporated quantity of liquid g.

point T_a of the liquid, for if S_1 drops only slightly below that temperature, vapour condenses on the plate and transfers to it its heat of condensation, causing the temperature of the plate to rise again to T_a. The upper vessel G contains another pure liquid, g, whose boiling point is, say, 10 °C lower than that of liquid a. At the bottom of this vessel there is another silver plate, S_2. Fitted between the two plates, S_1 and S_2, is a cylindrical sample P of the material under investigation. The heat that flows through this sample from S_1 to S_2 brings the liquid g to the boil, thereby keeping the temperature of S_2 constant at the boiling point T_g of g. Between the two silver plates there is therefore a constant temperature difference $T_a - T_g$. The vapour from liquid g is condensed in the cooler K_2 and the condensate is collected in a tube B, marked with a millilitre scale.

As soon as the steady state of heat flow is reached — which is reached when about 0.1 ml of condensate has collected in B — the time t taken for a certain liquid volume V, say 1 ml, to flow into B is determined with a stop watch. Let S be the heat of vaporization of the liquid g per unit volume, then eq. (1) gives:

$$\lambda = \frac{l}{A}\frac{1}{R} = \frac{l}{A}\frac{VS}{t(T_a - T_g)}. \quad . \quad . \quad . \quad (2)$$

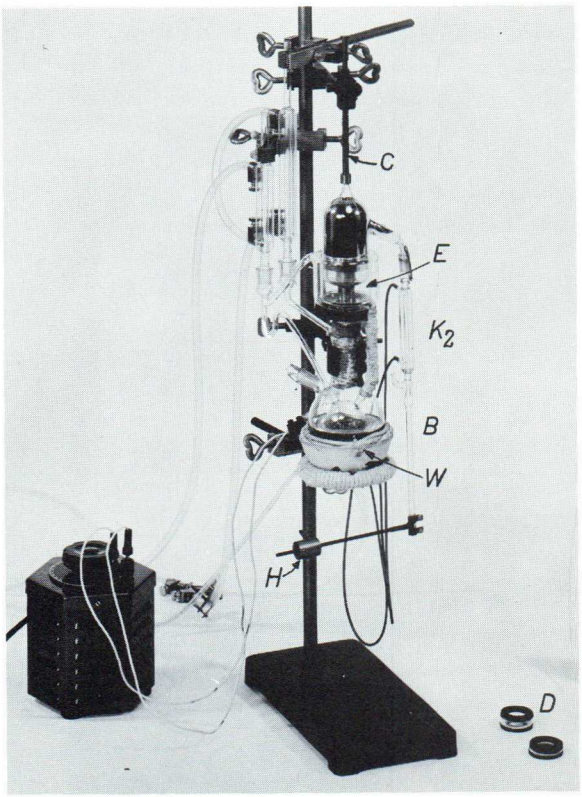
Given the data on the liquids and the dimensions l and A of the sample we can calculate λ from the measured values of V and t.

The method shows a remarkably high degree of accuracy, owing to the fact that the effect of heat losses between S_1 and the liquid g is kept relatively very low. This is done primarily by choosing the dimensions of the sample so as to ensure a reasonably large heat flow through the sample. For instance, if the substance is a poor conductor of heat, the sample taken will be in the form of a thin disc. For a better conductor, assuming the same temperature difference, a thicker sample will be taken, for an unduly large heat flow would cause the liquid g to boil too turbulently, with the risk of liquid splashing over into tube B. Furthermore, if the appropriate quantity of liquid distils too quickly into B, the time measurement is less accurate. The time t should preferably be between 100 and 1000 seconds, and this can be arranged not only by a suitable choice of the thermal resistance of the sample but also by using suitable liquids (fixing the values of S and T_a and T_g).

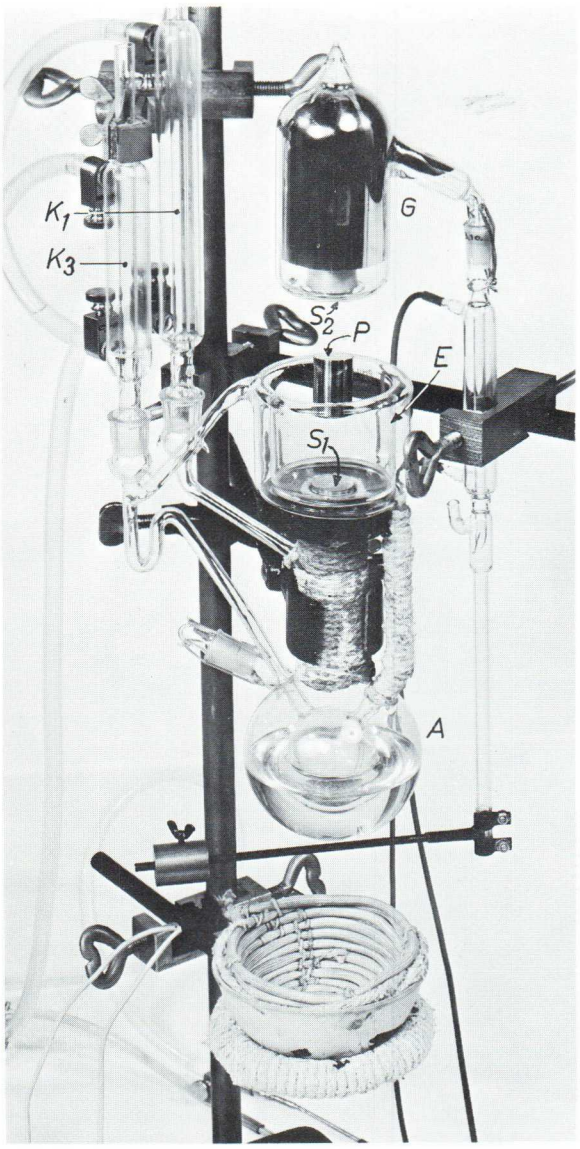
Heat may be lost from the upper vessel as well as from the sample, causing errors in V through insufficient evaporation or premature condensation. These errors can be minimized by enclosing the upper vessel G in a vacuum jacket (dewar flask); see the photographs of the set-up in *fig. 2*.

The method as described enables absolute determinations of λ to be made. If a number of calibrated samples are available, λ can be determined even more easily by a comparative measurement [3]). The times t found for the calibrated samples are plotted in a diagram against the known values of the thermal resistance R. On the calibration curve thus

[3]) The standard samples were prepared for us by K. H. Bode and W. Fritz of the Physikalisch-Technische Bundesanstalt, Brunswick. The values of thermal resistance determined by them were accurate to within ± 1% (Z. angew. Phys. **10**, 470-479, 1958). We take this opportunity to thank them for their valuable assistance.



a



b

Fig. 2. The apparatus, *a*) ready for use, *b*) partially dismantled (for this photograph the sample P was suspended by a thread between the two silver plates S_1 and S_2).

The upper vessel G is in the form of a dewar flask, with the sealed-off pinch at the top and the condenser K_2 and graduated tube B at the side. It is placed over the sample and pressed down by the long helical spring C. K_2 and B are balanced by an arm with the weight H. One of the rings D is placed around the sample to give thermal insulation. A refinement not shown in the arrangement in fig. 1 is a double-walled glass vessel E surrounding the sample; between the walls the vapour can be circulated of a liquid whose boiling point is between those of a and g. The vessel containing this liquid can be seen inside vessel A, and associated with it is a condenser K_3 with return pipe.

obtained we can then, after having measured the time t, read off directly the value of R for a given sample, and from this value, together with the dimensions of the sample, we can calculate the value of λ.

It should be noted that for these relative measurements it is not necessary to know the exact values of V, T_a and T_g. This at once eliminates the errors that can creep into a calculated λ value if *a*) the boiling points T_a and T_g should differ slightly from the assumed values owing to impurity of the liquids, and if *b*) the ends of the sample should not attain the exact temperatures T_a and T_g owing to contact resistances to the heat transfer between the silver plates and the sample. It should of course be ensured that these resistances are equal for all measurements, in order to obtain reproducible results. The same applies to the magnitude of the small residual heat loss, to which we shall return presently.

The calibration curve, as can be seen in *fig. 3*, is very closely a straight line. The negligible spread of the points is evidence of the high degree of reproducibility of which the method is capable. According to eq. (1) the curve should be a straight line through the origin. In the diagram this is not so, the line cutting off a section R_0 on the ordinate axis. This corresponds to the thermal resistance of the various heat-transfer surfaces. For each pair of liquids used the calibration line must be plotted afresh, but since it is a straight line it is sufficient to measure two standard samples, possibly with a third for verification.

By this comparative method, a measurement, including the preparation, lasts no more than 5 to 15

minutes. The error of measurement is $\pm$ 3% at the most. No temperatures need be measured, and it is not necessary to know the exact values of the boiling point and heat of vaporization of the liquids used. The liquids need not even be pure, provided the same liquids are used for calibration and measurement. The ease and speed with which a sample can be measured by this method open up the practical possibility of investigating systematically large series of specimens, as for example mixed crystals of varying composition.

Some particulars follow concerning the equipment and the samples.

Various measures are taken to ensure that the temperatures of the silver plates S_1 and S_2 will be accurately reproducible. To avoid any delayed-boiling effect in liquid g, capillary holes are made in the plate S_2. To increase the surface area for the heat transfer, both plates S_1 and S_2 are provided with grooves on the side in contact with the vapour and liquid, respectively. At the other side the plates are ground flat to obtain the best possible contact between sample and plates; also, the sample length is thereby precisely defined. To reduce the contact resistances a heat-transfer fluid is applied between the faces, and the plates are held against the sample under considerable (and constant) spring pressure (see fig. 2*a*). In this way the variations in the thermal resistance across the heat-transfer surfaces are reduced to such an extent that their effect in all measurements remains below the limits of accuracy determined by other factors.

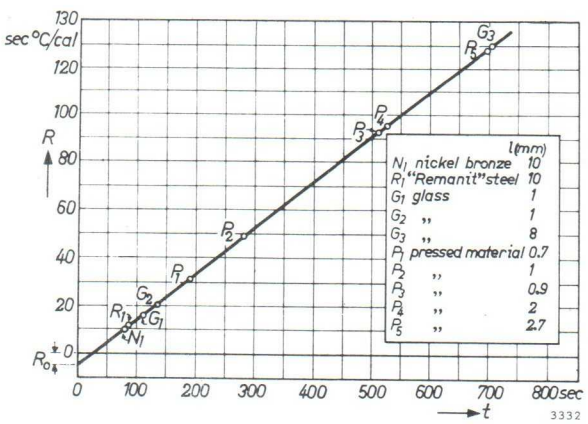


Fig. 3. Calibration chart for an apparatus using carbon disulphide and ether as boiling liquids (boiling points 46.3 and 34.5 °C, respectively). For ten calibrated samples (see inset) for which the thermal resistances R were known — values differing by a factor 10 for these samples — R is plotted against the measured time t taken for 1 ml of condensate to collect in tube B. The calibration points lie on a very good straight line, demonstrating the accuracy of the method.

Extrapolation of the line to $t = 0$ gives the correction R_0 to be applied in an absolute measurement of thermal resistance (R_0 takes account of contact resistances as well as heat losses).

It has already been mentioned that the upper vessel G is enclosed in a dewar flask to minimize heat losses. The sample must also, of course, be provided with heat insulation; a series of rings are available to surround the sample (two such rings can be seen at the right of the stand in fig. 2*a*). In order to carry out thermal conductivity measurements at temperatures very different from room temperature, both vessels and the sample must be surrounded by shields, held at the appropriate temperatures by baths in which the same liquids as in A and G are kept at the boil. For measurements below room temperature, where G and A thus contain liquids that are gaseous at room temperature, the vapour from G need not be condensed in a cooler but can be collected and measured in a gas meter.

As regards the liquids themselves, pairs can be found for the whole temperature range from —200 to +400 °C having suitable boiling points differing by 5 to 15 °C. If necessary, the same liquid can be used for both vessels, and the boiling point in vessel A raised slightly by a small and constant increase of pressure, controlled by a pressurestat.

The samples used for both calibration and measurement are cylindrical in form, ranging in diameter from 16 to 18 mm and in length from 0.5 to 30 mm, depending on the thermal conductivity of the material. Samples of material of high thermal conductivity (pure metals) can be made in the form of hollow cylinders, say 10 mm in length and with a wall thickness preferably not less than 1.5 mm.

Summary. The two ends of a cylindrical sample, about 18 mm in diameter and 0.5-30 mm in length, are kept at constant temperatures, by contact with two boiling liquids of suitable boiling points, viz. differing by about 10 °C. The *time* is measured in which a quantity of heat flows, in the steady state, through the sample. This quantity is fixed very simply by the evaporation of a fixed quantity of liquid at the "cold" end, collected as condensate. In practice, two calibrated samples are measured first and a calibration line is plotted from which the thermal resistance of subsequent samples can be found at a glance. It is then not necessary to know the exact boiling points of the two liquids, nor the absolute amount of evaporated liquid, nor its heat of vaporization. A measurement takes 5 to 15 minutes and the error is not more than $\pm$ 3%. With some additional equipment the method can also be used for measurements at other than room temperature, between —200 and +400 °C.

A positive rod or piston seal for large pressure differences

J. A. Rietdijk, H. C. J. van Beukering, H. H. M. van der Aa and R. J. Meijer

621-762.649

The moving seal between the gas spaces represents one of the most difficult design problems in hot-gas engines and gas refrigerating machines based on the Stirling cycle. It now appears that this problem can be satisfactorily solved, through the discovery that a specially shaped rubber diaphragm can be rolled up and down more than 2×10^9 times before failing — provided the rolling is done in a suitable way. This discovery and the device based upon it may well also prove to be of great value in other types of machine.

From the beginning of the work by Philips on the Stirling cycle, much attention has been given to the investigation of moving seals. Stirling machines consist of a driving mechanism and a piston-fitted cylinder system in which a working agent goes through a continuous cycle of temperature and pressure changes; two basic design variations appear in *fig. 1*. At the moment, hydrogen and helium under high pressure are virtually the only practicable working agents. Working agent pressures may be as high as 140 atm in hot-gas engines, and as high as 60 atm in gas refrigerating machines. The mode of functioning of these machines has been described in earlier articles in this review [1] [2], and a brief account of the principle of the Philips-Stirling gas refrigerating machine may be found in the present issue [3].

In the design appearing in fig. 1*a* two moving seals are necessary to shut off the space containing the working agent from the space containing the driving mechanism. These seals have to satisfy stringent requirements. Helium is expensive and hydrogen is inflammable, so neither must be allowed to escape, the more so as leakage of the working agent would lower the efficiency of the machine. Conversely, the seals must prevent any leakage of lubricating oil or other contaminants into the working space, where they might seriously interfere with the functioning of the machine by, amongst other things, blocking of the regenerator. A certain amount of lubrication, however, is essential, to prevent wear on the bearing surfaces of pistons or rods and their guides.

The design in fig. 1*b* has two rod seals whose function is as described above. In addition, there are two seals that might be described as "internal"; their function is not to shut off the space containing the driving mechanism, but to separate spaces which each contain working agent. Any leakage past these points likewise has an adverse effect on the performance of a Stirling machine, and the need to keep the working spaces uncontaminated makes the lubrication of the bearing surfaces a particularly difficult problem.

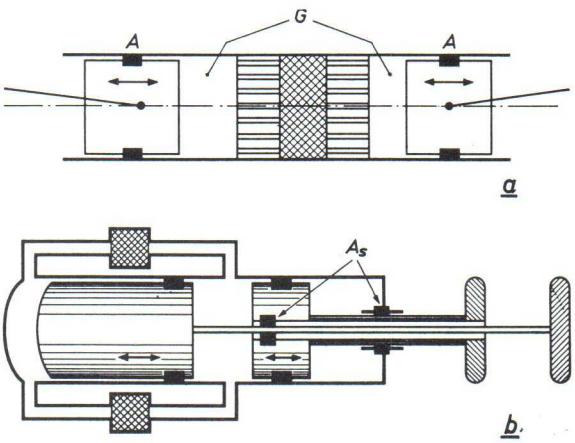


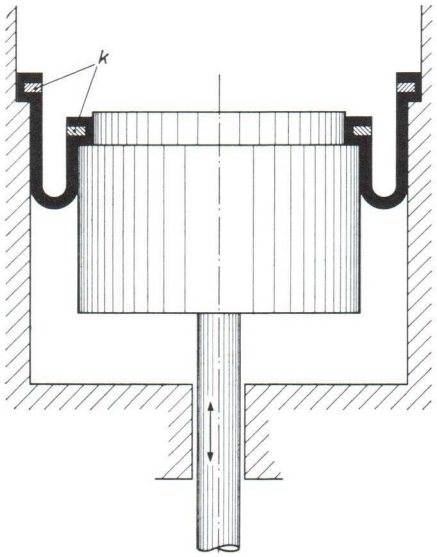
Fig. 1. *a*) Stirling machine working on the two-cylinder principle. Two seals *A* are necessary between the working space *G* and the driving mechanism.
b) Stirling machine working on the displacer principle. Two "external" seals A_s for the piston rods are necessary to isolate the working space; in addition, two further seals are required for separation of spaces at the same mean pressure.

Dr. Ir. J. A. Rietdijk, Ir. H. C. J. van Beukering, H. H. M. van der Aa and Dr. Ir. R. J. Meijer are research workers at Philips Research Laboratories, Eindhoven.

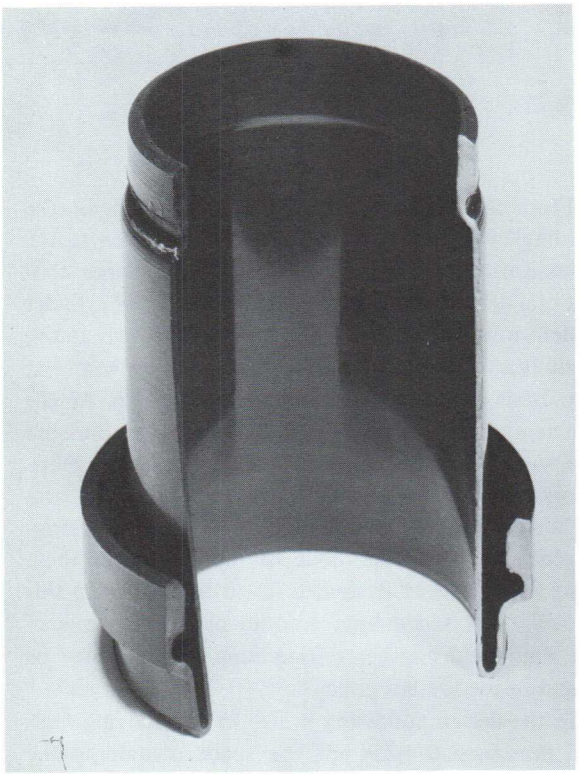
[1] J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine; Construction of a gas refrigerating machine, Philips tech. Rev. **16**, 69-78 and 105-115, 1954/55.
[2] R. J. Meijer, The Philips hot-gas engine with rhombic drive mechanism, Philips tech. Rev. **20**, 245-262, 1958/59.
[3] A. A. Dros, An industrial gas refrigerating machine with hydraulic piston drive, Philips tech. Rev. **26**, 297-308, 1965.

It would be too much to attempt to deal with all the types of seal that are in principle suitable for the above functions. The chief types can however be enumerated — piston rings, O-rings and sleeves, "small gaps" (with or without labyrinth), and glands (only suitable for rods). None of these types fully meets the above requirements; all, in fact, can be regarded as mere "leakage limiters". Their compromise nature has become more and more keenly realized in the course of the development of hot-gas engines and gas refrigerating machines, one reason having been the increasing requirements for good operating life.

A device of what has been called the "positive seal" type has been under development in our laboratory since 1960. This appears also likely to be of value in machines other than those based on the Stirling cycle. The present article is a detailed account of the new seal and of the manner in which it is applied in Stirling machines. Its application in oil-free compressors will also be discussed. These compressors play an important part in many cryogenic and other processes, and figure in sequential systems embodying the Philips-Stirling gas refrigerating machine [4]. In these compressors it is absolutely essential that there should be no contamination of the compressed medium.



a



b

Fig. 2. *a*) Rolling diaphragm, fastened to piston and cylinder-wall. Attachment is by closed clamping rings *k*. The diaphragm is made of a rubber-like material and a small pressure difference holds it snug against the piston and cylinder-wall, so that it rolls off these surfaces without creasing as the piston moves up and down. *b*) Cut-away rolling diaphragm.

Rolling diaphragms

The new seal makes use of a long-known principle, that of the rolling diaphragm. A seal of this kind is shown in *fig. 2a*; fig. 2*b* is a photograph of the diaphragm itself. The diaphragm, which has to be of a rubber-like material, is alternately unrolled from the piston and the cylinder-wall [5]. The movement it performs is rather like that of a stocking or sock; and indeed in the laboratory, the term "sock" has become commonplace. The *guide arrangements* for a piston fitted with one of these "socks" are much the same as for a conventional sealed piston.

Up till now rolling diaphragms have only been suitable for comparatively slow movement, and for *small* pressure differences, such as those in hydraulic and pneumatic control and regulating equipment, pneumatic springs, and so forth. A certain minimum difference of pressure is necessary across the diaphragm in order to prevent creasing. To raise the upper limit of the permitted pressure difference, which was about 5 atm, the diaphragms have generally been reinforced with some kind of fabric that is inextensible except around the circumference, and which performs much the same function as the canvas in a car tyre. Diaphragm life, in terms of the number of strokes before failure; has however been limited, particularly for reinforced diaphragms.

In the following we shall see how, with comparatively simple methods, a rolling diaphragm can be applied at *very high* pressure differences. It has been found that, with a suitable choice of the conditions under which the diaphragm operates, remarkably good life can be achieved.

Rolling diaphragm supported on a fluid cushion

The problem of adapting the rolling diaphragm for high-pressure seals was solved by supporting the diaphragm with a fluid cushion — see *fig. 3*. This solution implies acceptance of the limitations of the rolling diaphragm in regard to permissible pressure difference; across the diaphragm itself there is only a small pressure difference, of about 5 atm. The function of the diaphragm is confined to separating the oil in the cushion from the working agent [6] above the piston. The real difference of pressure (which may range between 50 and 100 atm) is borne by a second seal of conventional type, which separates the oil-filled space under the diaphragm from the space containing the driving mechanism.

To safeguard against rapid wear of the diaphragm, and fatigue due to alternating stresses, it is necessary to ensure that the diaphragm is at all times in close contact with the fluid and *has the same length* throughout its rolling cycle. Once these conditions are realized the reinforcement is no longer necessary and indeed has been found to shorten the life of the diaphragm.

The requirement of constant length is equivalent to one of *constant pressure difference* across the diaphragm. The graph in fig. 3 shows the desired variation in the cushion fluid pressure, for a given pressure variation in the gas. The case illustrated here is that of a "concave diaphragm" occurring when the fluid pressure is lower than the gas pressure. Alternatively, the direction of the pressure difference across the diaphragm can be reversed, so that the fluid pressure is higher than the gas pressure. The diaphragm is then called a "convex diaphragm". This is illustrated in *fig. 4*; the surface from which the diaphragm is rolled is not lubricated here, and this arrangement has in fact been found to be not so favourable.

In general, the tendency will be to employ the lubricating oil of the machine as fluid for the supporting cushion, especially since it has to lubricate the surface of the piston. Accordingly, the cushion will from now on be referred to as an "oil support". Other fluids may in special instances be suitable.

The requirement that the difference of pressure across the diaphragm is kept constant can be satisfied in two ways: by means of the stepped system, or by means of a floating piston-cap. The two methods will now be discussed in detail.

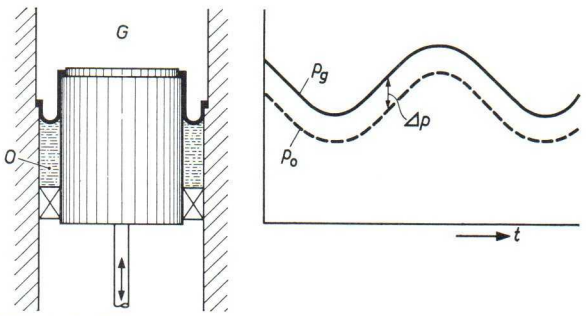


Fig. 3. Rolling diaphragm supported on a cushion of fluid. The fluid O takes up the whole of the pressure exerted by the working agent G except for a small difference in pressure Δp carried by the diaphragm. The pressure p_0 of the fluid must at all times be lower than the pressure p_g of the working agent by a small constant amount Δp. The diaphragm shown here is of the concave type. In the diagram on the right the pressures just referred to are shown as functions of time t.

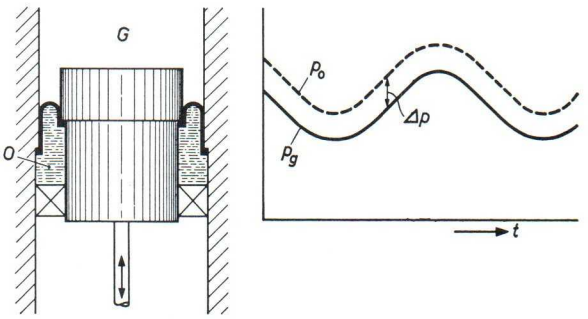


Fig. 4. A rolling diaphragm of the convex type. Here the pressure of the supporting fluid must at all times be *higher* by a certain small amount than the pressure of the working agent to be sealed off. The diagram on the right shows pressures as functions of time.

Methods of keeping the pressure difference constant

Stepped system

The oil space below the diaphragm can be shaped in such a way that the volume it encloses remains constant throughout the cycle for constant diaphragm length. This is ensured by *stepping* both piston and cylinder wall as in *fig. 5*; the area of the step in the piston must be equal to that of the step in the cylinder wall. This condition will be satisfied if the piston and cylinder diameters satisfy the relationship:

$$d_2^2 = \tfrac{1}{2}(d_1^2 + d_3^2). \quad \ldots\ldots\ldots \quad (1)$$

[4] G. Prast, A gas refrigerating machine for temperatures down to 20 °K and lower, Philips tech. Rev. **26**, 1-11, 1965 (No. 1), particularly pp. 10-11.

[5] The new seal is suitable for rods or for fitting to a piston in a cylinder. Only the cylinder and piston combination will be discussed in this article, but what is said may be considered to apply equally well to a rod passing through a cylinder base.

[6] The medium above the diaphragm will normally be a gas. However, the seal is also suitable, without modification, for use in a pump handling a liquid, for example. In the rest of the article it will be assumed for the sake of simplicity that the space above the seal is occupied by a gas.

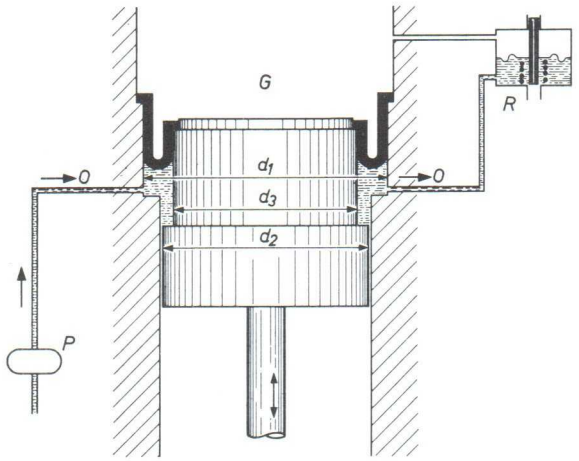


Fig. 5. Rolling diaphragm with fluid support in stepped system. Provided the amount of oil does not change as the piston moves, the length of the rolling diaphragm will not change. Oil *O* is fed by a pump *P* and removed by a control device *R*. The device embodies a spring-loaded metal diaphragm which senses the difference between the mean gas and mean oil pressures, and controls two valves. One is the safety valve which releases gas when the pressure difference becomes too large; the other is an oil escape valve which, in normal operation, is at the "cut-off point". This method of keeping the diaphragm seal length constant is suited to both the convex and the concave types of diaphragm.

A proof that the enclosed volume remains constant can be obtained by considering a certain displacement of the piston. If the length of the diaphragm remains the same, the fold in it travels half this distance; the step in the piston — which occupies half the area of the gap between piston and cylinder — travels the *whole* distance. Clearly, then, there has been no change in the overall volume enclosed. The constant pressure difference requirement is therefore met if the space is filled with a quantity of oil such as to produce the required stressing of the diaphragm, which may be convex or concave.

In practice there will be slight deviations from the ideal behaviour just described. Amongst these are the compressibility of the oil, the fact that the walls of the space are not perfectly rigid, and within-tolerance deviations in the various diameters. There will, therefore, be some slight fluctuation in the length of the diaphragm. The above deviations can be kept within limits close enough that there is no difficulty with non-reinforced rolling diaphragms, despite the small changes in their length. In reinforced diaphragms, however, very severe stresses would be set up. In short, the stepped system can only be employed in conjunction with *non-reinforced* rolling diaphragms.

The oil support itself calls for some special measures. Oil will escape from the cushion space past the oil-seal, which has to be one of the leakage-limiting types already referred to. The loss is made up by feeding in an excess of oil by means of a miniature high-pressure pump [7]. The excess is removed via a *control device* that maintains a constant difference between the oil and gas mean pressures.

The control device can take the form of a regulating valve which is located away from the diaphragm and which also serves as a safety valve. An arrangement of this kind is shown in fig. 5; its mode of action is explained in the caption. Another possibility is to use the diaphragm itself as a control element, as in *fig. 6*. Here an escape orifice is provided in the cylinder wall; this is uncovered during the short interval in which the diaphragm is in its topmost position. In this position the rate of loss is strongly dependent on very small changes in the length of the diaphragm. This arrangement makes it possible to remove oil at a varying rate while the length of the diaphragm — and hence also the pressure difference — remains very nearly constant.

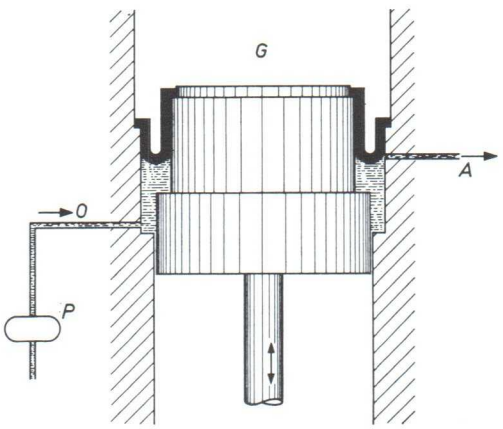


Fig. 6. Self-regulating rolling diaphragm in a stepped system. The diaphragm clears orifice *A* and allows oil to escape during an interval that depends on the rate at which fluid is fed into the support. This means that the length of the diaphragm remains almost constant. This method is only suitable for the concave diaphragm.

The floating piston-cap

Fig. 7 illustrates the second of the two methods, in which the primary object is to maintain a constant difference of pressure across the rolling diaphragm. It is only suitable for piston seals. The piston in question is equipped with a cap that is free to move in the axial direction. The cap is connected by a spring to the body of the piston. The rolling diaphragm (which can be convex or concave, as before) extends between the cap and the cylinder wall. The part of the piston body on which the cap rides is so constructed as to allow a free passage of oil between the space under the diaphragm and that between the piston and the cap.

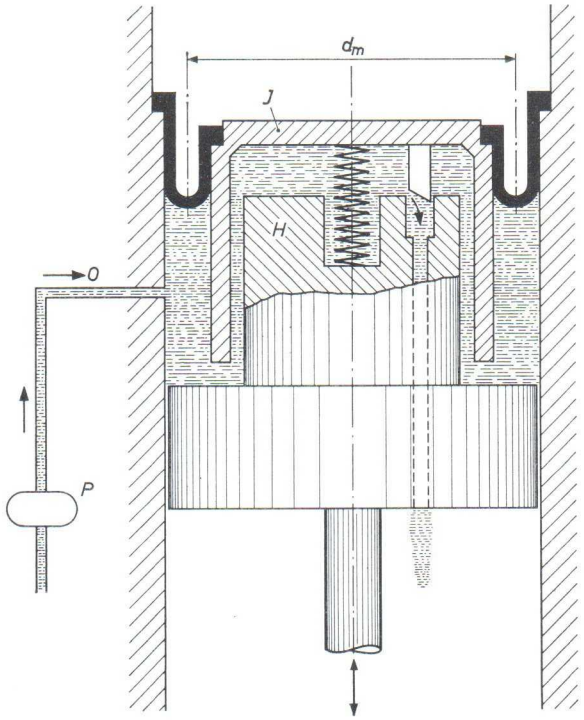


Fig. 7. Rolling diaphragm fitted on a piston with a "floating cap". The free cap J and the main piston H act together as a regulating system. The difference of pressure across the diaphragm is kept constant by means of a spring between piston and cap. A spring is chosen whose stiffness is so small that the force is more or less unaffected by small movements of the cap relative to the body. This construction is suitable both for concave and convex diaphragms.

Oil is supplied to the support in exactly the same way as in the first method. The excess is removed by an arrangement that comes into action as soon as the cap has reached a certain peak height above the piston body. The arrangement can consist either of a port in the piston wall or a dipping pin in the base of the piston, as in fig. 7. The spring is so dimensioned that the force K it exerts in its topmost position satisfies the relation:

$$K = \Delta p \frac{\pi}{4} d_m^2, \quad \ldots\ldots (2)$$

where Δp denotes the desired difference of pressure across the diaphragm. Thus the piston-cap combination functions at the same time as a regulating valve. There is no need here for the steps cut into the piston and cylinder wall to satisfy closely the condition expressed by eq. (1), since a small periodic displacement of the cap with respect to the piston body is enough to keep the oil volume constant. Provided the spring is not too stiff, the extensions it undergoes will not be accompanied by any appreciable variation in force K or, therefore, in the difference of pressure across the diaphragm. If a sufficiently weak spring is employed, the stepped geometry can be done away with altogether, as in fig. 7. At the same time this widens the choice of oil feed arrangements; a pulsating supply can be used if desired, the oil being fed in at a high rate by a plunger-type pump.

When discussing the oil-free compressor we shall see that this method also enables simple automatic control of the dead space volume in piston machines.

Both methods imply circulation and hence renewal of the oil, and this has two advantages. Gas may diffuse through the diaphragm and into the oil support, which as a result ceases to be incompressible; this may interfere with its action, especially in the stepped system. By circulating the oil the diffused gas is removed. At the same time the oil acts as coolant, and this lengthens the life of the seal.

It may also be noted that the oil supporting the rolling diaphragm ensures as near perfect lubrication of the piston surface as possible.

The foregoing will have made it clear that the rolling diaphragm has reduced the problem of sealing off the gas-filled space to one of sealing off an oil-filled space from which a certain amount of leakage is permissible. Conventional seals can be employed for this, and these are quite capable of standing up to pressure differences of hundreds of atmospheres.

Results

It will be clear that the rolling diaphragm seal depends completely on the availability of a suitable rubber-like material (elastomer). To be suitable, the material must have a high fatigue strength and creep resistance and not be subject to chemical attack by the oil or the working gas.

Fortune was with us when we tried out the first material, a polyurethane rubber recommended by the Philips Plastics Laboratory, and which proved to satisfy these requirements quite well. While investigating the material we discovered that its endurance is largely dependent on three parameters — temperature, the pressure difference across the diaphragm, and the ratio δ between $(2d_0)$, the double thickness of the diaphragm, and the piston-cylinder wall clearance (s). Some of our results are displayed in *figures 8* and *9*.

It will be seen from fig. 8 that the life of the diaphragm falls off sharply with rising temperature. This effect may be connected with the marked loss of the

[7] An elegant solution to the oil feed problem is also offered by the "pumping ring" invented by H. J. Verbeek of Philips' Industrial Equipment Division. The primary function of the ring is to seal off the oil space, but its shape is such that the piston movement sets up a hydrodynamic pump action, which pumps oil directly into the support space from the space containing the driving mechanism.

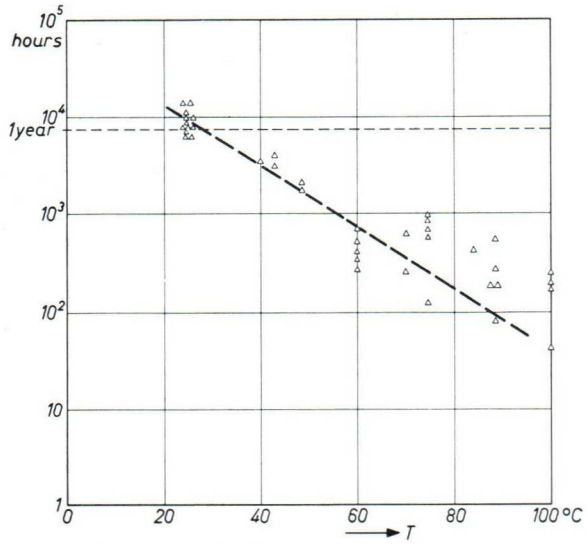


Fig. 8. Life in hours of rolling diaphragms made of polyurethane rubber, as a function of temperature T. Pressure differences across the diaphragms thus tested ranged from 4 to 6 atm. At zero pressure difference the diaphragms had a thickness of $d_0 = 0.5$ mm; the piston-cylinder clearance was $s = 2$ mm, the shaft speed 1500 r.p.m., and the piston stroke 65 mm. Raising the temperature clearly shortens the life of the diaphragms markedly. At room temperature almost all the diaphragms tested lasted for more than a year.

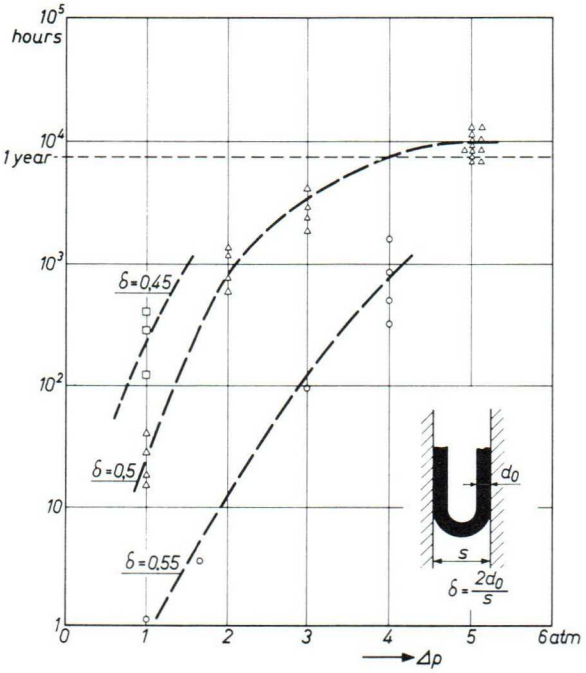


Fig. 9. Life in hours as a function of pressure difference Δp and ratio δ. Both these quantities affect the amplitude of the periodic elongation which the material undergoes during the cycle (see appendix) and hence the life of the diaphragms.

tensile strength of the material at higher temperatures. At 100 °C it has only 20% of the tensile strength it possesses at room temperature. Fig. 9 reveals that up to a certain limit, the life of the diaphragm at a given temperature can be lengthened by stepping up the pressure difference Δp. It will also be seen that for a given pressure difference, the life of the diaphragm is strongly dependent on δ. The life of the seal has been found to be independent of the absolute pressure level.

With a view to explaining the influence of Δp and δ we endeavoured to set up a simple theory and carry out calculations that would cast light on the variation in the stresses set up in the rolling diaphragm, and on the deformation it undergoes. This theoretical study is reported in the *appendix* to this article.

The results displayed in figures 8 and 9 relate to rolling diaphragms fitted to a piston with a 65 mm stroke. The machines on which the diaphragms were tested had a shaft speed of 1500 r.p.m. The longest-lived diaphragm so far tested stood up to 13 000 hours of operation (about 18 months), during which the rubber was flexed more than two thousand million times.

The abandonment of reinforcement of the diaphragms has removed a limitation on the ratio between stroke and diameter, a limitation inherent in conventional rolling diaphragms. The dimensions of the diaphragms so far tested have ranged from 6 mm diameter and 0.1 mm thickness, for a 0.4 mm piston-cylinder clearance, to 175 mm diameter and 0.7 mm thickness for a 4 mm clearance.

Research is now in progress on materials which stand up to various chemically aggressive substances and to high temperature operation. Both properties are especially important in compressors.

Applications

Gas refrigerating machine

First mention should be made here of the large gas refrigerating machine that has been developed by the Cryogenics Department of the Philips Industrial Equipment Division. This machine, known as type C, combines the rolling diaphragm seal with hydraulic piston drive in a highly elegant way; it is discussed at length in another article appearing in this number, which has already been referred to [3]. This machine has the stepped oil space system.

Hot-gas engine

A single-cylinder hot-gas engine equipped with rolling diaphragms on a fluid support has been built in our laboratory. The stepped system was again used. The engine is shown schematically in *fig. 10*. The thermodynamic part of the machine, i.e. the working space in which the Stirling cycle takes place, is shut off by a piston from a buffer space filled with gas at a pressure equal to the mean pressure of the working agent. The purpose of the buffer space is to take up

as much as possible of the thrust on the driving mechanism.

The engine incorporates four rolling diaphragm seals, two of which can be classed as "external". One of these is fitted to the displacer rod, the other to the piston rod. The other two are fitted to the piston itself, as shown in fig. 10. As can be inferred from the diagram all these seals operate at room temperature. The displacer rod seal and the upper piston seal are acted upon by the pressure of the working agent, which varies during the cycle; the one on the piston rod and the lower piston seal are both exposed to the pressure of the gas in the buffer space. The four seals thus fall into two groups of two, and there is communication between the two oil spaces in each group. Each group further has its own oil-pump and regulating valve.

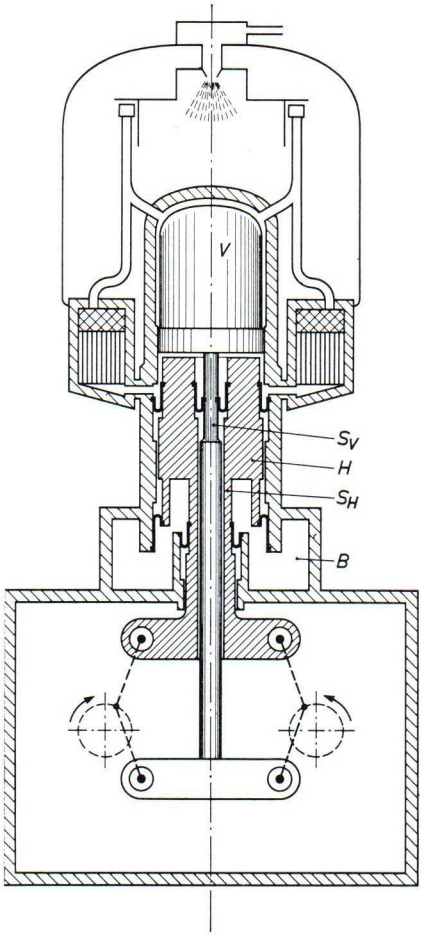


Fig. 10. Experimental model of a hot-gas engine embodying four rolling diaphragms and developing a shaft power of 80 h.p. at 1500 r.p.m.

During the cycle the working agent (hydrogen) attains a peak pressure of 140 atm. *H* piston. *V* displacer. S_H piston rod. S_V displacer rod. *B* buffer space.

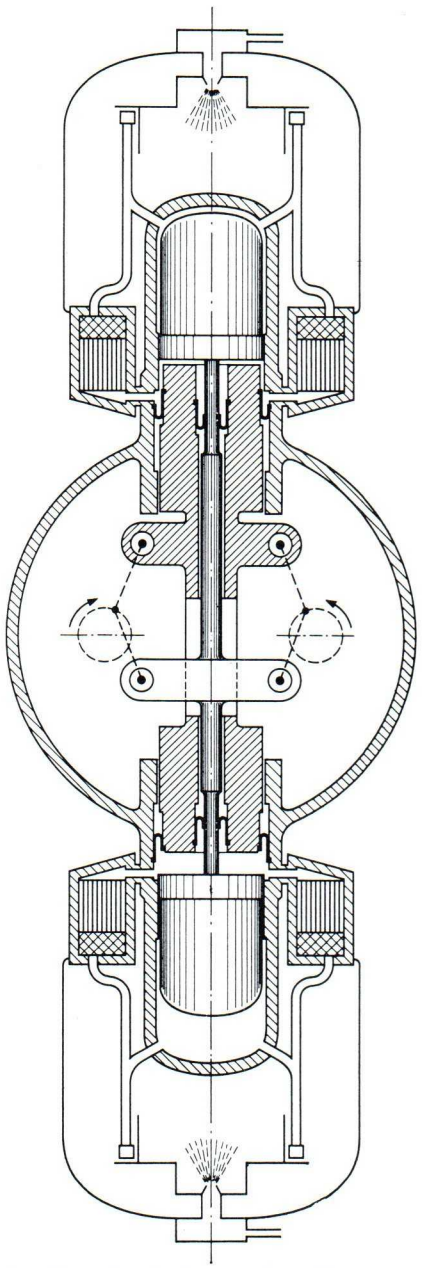


Fig. 11. Basic design of a horizontally-opposed type of hot-gas engine. Here the function of the buffer space in fig. 10 (which is to take up the mean thrust on the piston) has been taken over by a second working space; the cycle taking place in the second working space is exactly the same as that in the first. The pistons are mounted on the same rod. The number of rolling diaphragm seals per cylinder is reduced to two.

In this particular design the oil spaces belonging to the two piston seals are adjacent but separate; transference of oil from one to the other has to be prevented because there is a periodically alternating difference of pressure between the two oil supports, the peak value of which is in the region of 40 atm. The seal between the two spaces is provided by a piston ring. Although there are special problems, which we will not go into here, in the use of piston rings in this application, the arrangement is acceptable.

The buffer spaces in hot-gas engines are generally difficult of access, and in fact represent an undesirable complication. For this reason work has begun on a new type of engine in which the function of the buffer space will be performed by a second working space. The two pistons of this "horizontally opposed" engine are fitted to a common piston rod, so that each compensates for the mean thrust of the other. The arrangement is shown in *fig. 11*. The number of sealing diaphragms per working space has been reduced to two, and all are easily accessible. This design would not be practicable were it not for the positive seals fitted to the two pistons; a conventional type of seal would allow too much leakage of gas into the space containing the driving mechanism. Conventional seals only allow small diameter constructions in which the pistons have to be linked to the driving mechanism by relatively thin rods.

Oil-free compressors

The rolling diaphragm seal can also be applied in oil-free compressors, and in view of the importance of these machines, particularly in cryogenics, a great deal of attention is being paid to this application of the new seal. Two experimental piston-type compressors, fitted with rolling diaphragms, have been built. One is a single-stage machine supplying a pressure of 4 kg/cm² and having an output of about 45 normal m³/h, and the other is a three-stage machine supplying a pressure of 150 kg/cm² and having an output of about 30 m³/h. The employment of convex diaphragms makes both machines suitable for a subatmospheric gas intake, and both embody pistons of the floating-cap type. They differ only in constructional details and the performance data given above. Accordingly, we shall only deal here with the single-stage compressor, which is shown schematically in *fig. 12*.

Piston *1* is equipped with a cap *2* which, in this particular case, has a large displacement with respect to the piston. The freedom of movement possessed by the cap makes it possible, during operation of the machine, to regulate the volume of the dead space (the free space remaining in the cylinder when the piston is at the top of its stroke), and hence the output of the compressor. The cap is linked to the piston body by a spring *3*. Oil is fed into the intervening space by a plunger-type pump *4* which is built into the piston body. It flows along the inside of the cap towards the space under the rolling diaphragm, and this gives very effective cooling of the cap. At *5* there is an overflow which opens when the cap has reached a certain maximum height above the piston body. Alternatively, an excess of oil can escape by duct *6* to control device *7*. The action of this device is governed in the first instance by the pressure $\hat{p}$ in the compression vessel *8*, the valves closing as soon as pressure $\hat{p}$ falls below a certain value which is adjustable, and opening as soon as this value is exceeded. The effect of this is that the volume of the dead space is continuously adjusted as required. This control action must not interfere with maintenance of the correct pressure difference across the rolling diaphragm, and control device *7* therefore also contains an arrangement that prevents the pressure difference falling below a certain value (the latter function is performed by the upper part of *7*). It is naturally possible to arrange for the control device to respond to some quantity other than pressure. The device is to be regarded as a refinement that can be left out of the design if the regulating function is not required.

Immediately the machine is stopped, oil starts to leak past the piston body into the crankcase, and the cap gradually drops back into a starting position as a result of the pressure exerted by spring *3* and that of any gas present in the compression space. An O-ring *9* is so dimensioned that it clears the cylinder wall when the machine is running; when the machine is at rest the ring is pressed flat, and seals off the gap between the piston body and cylinder wall. This means that enough oil remains in the support space to keep the diaphragm distended, so that the machine is ready for the next start. The machine therefore starts at maximum dead space, i.e. more or less unloaded.

This single-stage compressor has so far been running continuously for a good 4000 hours at a speed of 1500 r.p.m. A life-test of the three-stage compressor has only just begun.

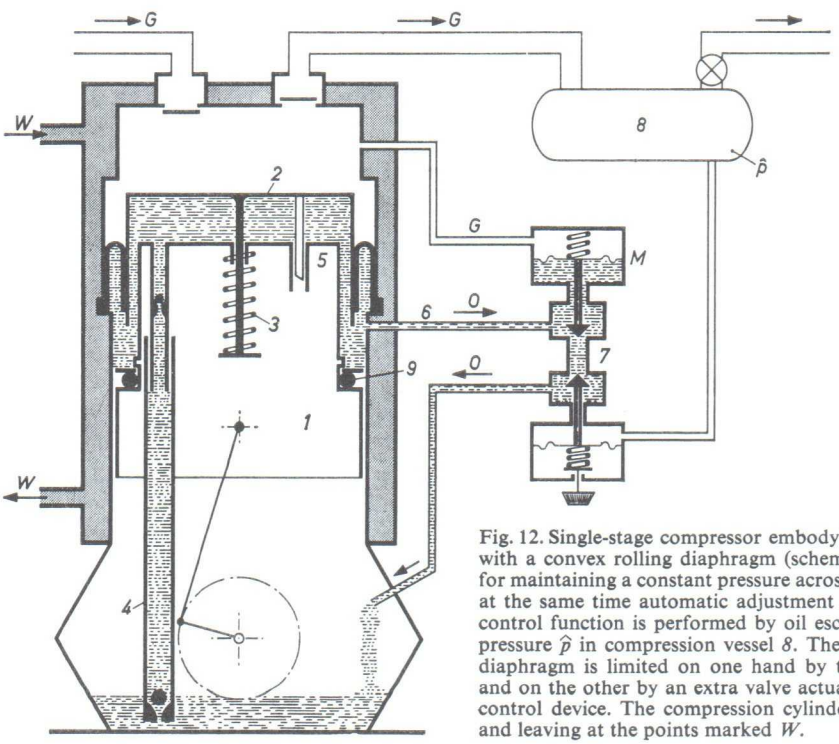


Fig. 12. Single-stage compressor embodying a piston with a floating cap, fitted with a convex rolling diaphragm (schematic). The control system responsible for maintaining a constant pressure across the diaphragm is designed to provide at the same time automatic adjustment of the dead space volume. The latter control function is performed by oil escape valve *7* which is governed by the pressure $\hat{p}$ in compression vessel *8*. The pressure difference across the rolling diaphragm is limited on one hand by the escape of oil through overflow *5*, and on the other by an extra valve actuated by the metal diaphragm *M* in the control device. The compression cylinder is water-cooled, the water entering and leaving at the points marked *W*.

Appendix: stresses in the rolling diaphragm

From the fundamental theory of elasticity any stressed state may be represented by a system of principal stresses acting along orthogonal axes, as in *fig. 13*. We shall take σ_x, σ_y and σ_z to denote the principal stresses, the subscripts referring to the directions in which they act. Here we define stress as force per unit area of an element of the body undergoing deformation; the planes on which the principal stresses act are distinguished by their lack of shear stress. We shall take λ_x, λ_y and λ_z to denote the relative deformations in the x, y and z directions (i.e. the dimensions after deformation divided by the original dimensions). If we started with a unit cube before loading, then λ_x, λ_y and λ_z will represent its new dimensions under load.

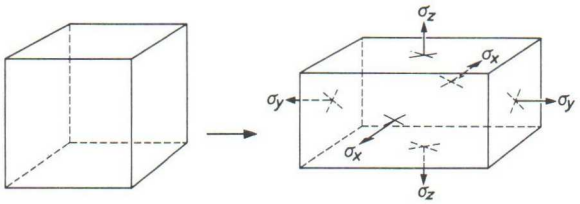


Fig. 13. Deformation of unit cube acted on by principal stresses σ_x, σ_y, σ_z.

The behaviour of a stressed elastomer can be described, to a fair degree of approximation, by the following equations:

$$\lambda_x\lambda_y\lambda_z = 1, \quad \ldots\ldots\ldots\ldots \quad (3)$$

$$\left.\begin{aligned} \sigma_x &= Ce^{\lambda x} - p, \\ \sigma_y &= Ce^{\lambda y} - p, \\ \sigma_z &= Ce^{\lambda z} - p. \end{aligned}\right\} \quad \ldots\ldots\ldots \quad (4)$$

Eq. (3) takes account of the extremely low compressibility of this kind of material. C is a material constant which in our case has a value of 9.2 kg/cm² [8]. p is taken as the "hydrostatic" pressure of the material, and is a function of σ_x, σ_y and σ_z. When $\sigma_x = \sigma_y = \sigma_z = 0$, $p = Ce$.

We shall start off by considering the stresses in the straight part of a rolling diaphragm supported by a fluid in which a pressure of p_2 prevails, the other side of it being acted upon by a higher pressure p_1 (see *fig. 14*, which shows the dimensions of the system and the axes along which the stresses act). We shall assume that $s \ll D$, in order to be able to neglect circumferential extension. (The results of the calculation will show that this is justified, for in practice the circumferential extension is very much less than that in other directions.) Hence we may put $\lambda_z = 1$. Now let us consider a bar of unit width, perpendicular to the plane of the drawing. If the vertical forces are in equilibrium we can write:

$$2d\sigma_x + p_2 s = p_1 (s - 2d),$$

or

$$\Delta p = p_1 - p_2 = \frac{2d}{s}(\sigma_x + p_1). \quad \ldots \quad (5)$$

Here it has been assumed that the wall exercises no shearing force on the rolling diaphragm. It is not difficult to see that the principal stresses in the straight portions of the rolling diaphragm act in the x, y and z directions shown in the figure and, in the case under consideration, that $\sigma_y = -p_1$. We can also infer that $d = \lambda_y d_0$, where d_0 is the initial or unloaded thickness of

[8] The above theoretical approach does not take thermal effects and creep into account. The figures quoted relate to room temperature. A different version of eq. (4) is to be found in the literature, namely $\sigma_i = C'\lambda_i^2 - p$ (where $i = x, y, z$). This formula is only true for strains up to $\lambda = 2$ (i.e 100% extension).

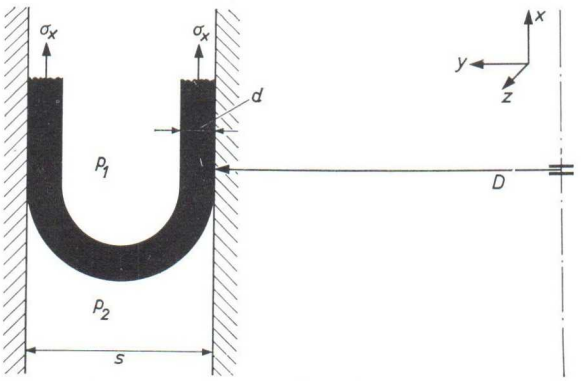


Fig. 14. Diagram to illustrate the derivation of stresses and strains operative in the straight part of a rolling diaphragm.

the diaphragm. Using (3) and (4), taking $\lambda_z = 1$, then (5) becomes:

$$\Delta p = \delta C \frac{e^{\lambda x} - e^{1/\lambda x}}{\lambda_x}, \quad \ldots\ldots \quad (6)$$

where δ, as already defined, is

$$\delta = \frac{2d_0}{s}. \quad \ldots\ldots\ldots \quad (7)$$

Measurements on a rolling diaphragm at room temperature, where $\delta = 0.5$, have confirmed that (6) corresponds closely with practical results.

Only the pressure difference figures in (6), not the absolute pressure level. The situation has been found rather analogous for deformation of the curved part of the diaphragm, which we shall now consider. Thus the elastic behaviour of a rolling diaphragm does not depend on the pressure level at which the machine or engine cycle is taking place — a conclusion that has been confirmed by the experiments.

In dealing with the curved section of the diaphragm (see *fig. 15*) we shall make two assumptions. Firstly, we neglect the circumferential extension, as before ($s \ll D$). We shall further assume that the curvature is strictly circular. The main axes x, y and z, along which the stresses are concentrated, will therefore be as shown in fig. 15. We can again say that $\lambda_z = 1$, so that $\lambda_x\lambda_y = 1$. Let us now consider a small volume whose dimension perpendicular to the plane of the drawing is unity, and whose other dimensions are $r\,d\varphi$ and dr. If forces in the y direction (r direction) are in equilibrium, we can write:

$$\sigma_x dr\, d\varphi = \frac{d(\sigma_y r\, d\varphi)}{dr} dr,$$

or

$$\sigma_x - \sigma_y = r\frac{d\sigma_y}{dr}. \quad \ldots\ldots\ldots \quad (8)$$

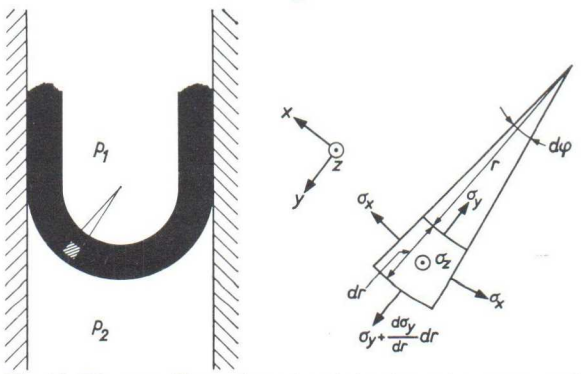


Fig. 15. Diagram illustrating the derivation of stresses and strains in the curved part of a rolling diaphragm. The volume element in the sketch on the left is shown on a larger scale on the right. The principal stresses are along the main axes x, y, z.

This being so:

$$\left.\begin{aligned}\lambda_x &= \pi r/l_0\,,\\ \text{so that}\quad \lambda_y &= l_0/\pi r\,.\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots \quad (9)$$

The quantity l_0 is the initial "fibre-length" of the material in the curved portion; it is independent of r, and in fact it depends only on the boundary conditions of the problem. From (4), (8) and (9), we get:

$$\frac{\mathrm{d}\sigma_y}{\mathrm{d}r} = \frac{C}{r}\left(\mathrm{e}^{\pi r/l_0} - \mathrm{e}^{l_0/\pi r}\right). \quad \ldots\ldots \quad (10)$$

The boundary conditions are:

$$\left.\begin{aligned}\sigma_y &= -\,p_2 \quad \text{for} \quad r = \tfrac{1}{2}s,\\ \sigma_y &= -\,p_1 \quad \text{for} \quad r = r_i,\end{aligned}\right\} \quad \ldots\ldots\ldots \quad (11)$$

where r_i is the radius of curvature of the inside surface. In the simple situation represented in fig. 14, $d_0 = d/\lambda_y$, but this does not now apply; instead:

$$d_0 = \int_{r_i}^{\frac{1}{2}s} \frac{1}{\lambda_y}\,\mathrm{d}r = \int_{r_i}^{\frac{1}{2}s} \lambda_x\,\mathrm{d}r\,,$$

and from this and (9) and (7) it follows that

$$r_i = \tfrac{1}{2}s\sqrt{1 - \frac{4\delta l_0}{\pi s}}\,. \quad \ldots\ldots \quad (12)$$

We are now in a position to calculate a number of unknowns by numerical integration of (10), using (9), (11) and (12). In this way we can calculate λ_x and λ_y, and if desired, σ_x, σ_y and σ_z, as functions of Δp, δ and r.

Strains λ_x in the x direction calculated in this manner for the straight part of the diaphragm, confirmed, as we have already stated, by measurements, and also the strains for the inside and outside surfaces of the curved part, have been plotted in *fig. 16* as functions of Δp for two values of δ. As the diaphragm rolls up and down, a volume element on the inside of the diaphragm will undergo a strain (extension) in the x direction whose magnitude swings periodically between the curve appropriate to the straight part and the curve appropriate to the relevant point in the fold. An analogous situation applies for any volume element on the outside of the diaphragm. Thus the spacing of these curves determines any fatigue effect manifested by the rolling diaphragm (volume elements *inside* the diaphragm in every case undergo less deformation than those at the surface).

The first conclusion we can draw from fig. 16 is that the smaller δ is made, the smaller will be the periodic deformation undergone by both the inside and outside surfaces of the diaphragm at a given difference of pressure Δp. This conclusion is in accord with the experimental evidence displayed in fig. 9 that diaphragms with smaller δ last longer.

As Δp increases, the periodic deformations undergone by the outside surface gradually become less, while those undergone by the inside surface either do not change at all or increase slightly. This theoretical inference must be linked up with two experimental facts. Firstly, rolling diaphragms subjected to endurance tests have always failed at the *outside surface* first; irrespective of whether the diaphragm has the convex or the concave form, the first tear appears on the outside and the plane of the break is always perpendicular to the x direction. Secondly, the life of the diaphragm increases with Δp, as is apparent from fig. 9. We must conclude that *the deformation limits do not form the only factor determining the life of the diaphragm.* The truth of this can be seen as follows. At a certain small pressure difference $(\Delta p)_1$ the strain in the outside surface varies between the extremes indicated by the double arrow on the left of fig. 16. In due course, a tear will form in this surface. At a much greater pressure difference $(\Delta p)_2$ the strain in the *inside* surface varies between almost the same limits, but *no* tear will form here within

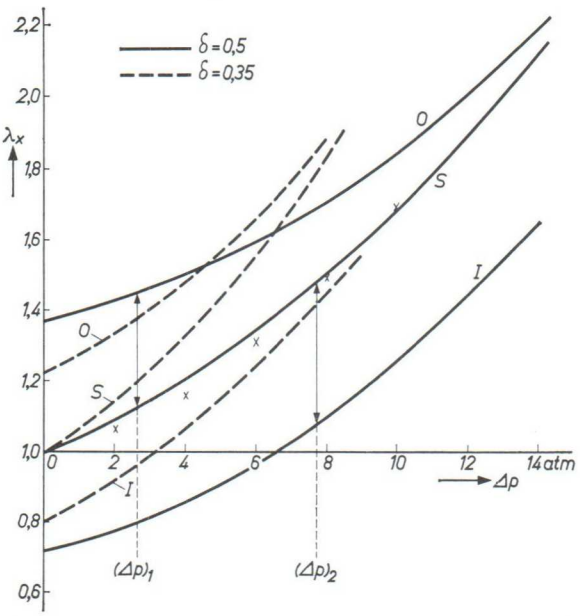


Fig. 16. Theoretical strains in the x direction as a function of pressure difference Δp for two values of δ. The curves relate to deformation λ_x in the straight part (S) and at the inside and outside surface (I and O respectively) of the curved part of a rolling diaphragm at room temperature. The crosses represent measured values obtained from the straight part of a diaphragm with $\delta = 0.5$.

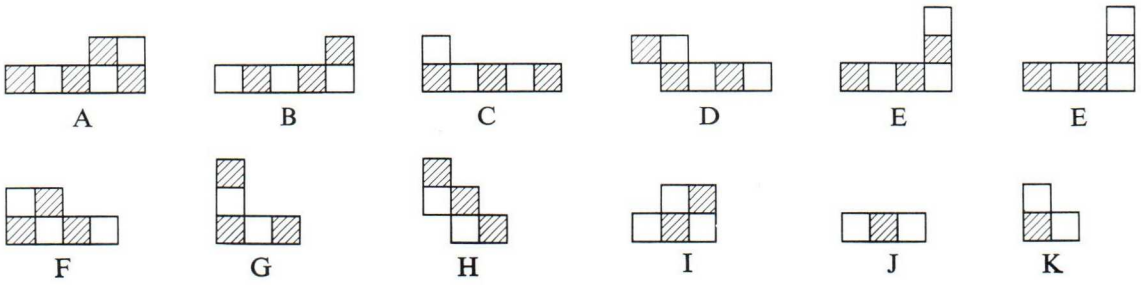
the same period of time; the life of the diaphragm at this Δp value is much longer, and failure will finally occur at the *outside* surface, as before.

What really governs the life of the rolling diaphragm is not yet known with certainty. The fact that failure invariably occurs on the outside surface first would suggest that a part is played by the difference in the *character* of the deformation. In the course of the rolling cycle, the permanent strain at the outside surface is increased for a brief interval; that at the inside surface is briefly diminished. There is yet another point of difference between the outside and inside surfaces. For the outside surface, the periodic changeover from the straight to the curved condition and back again is associated with a making and breaking of contact with the piston or cylinder wall. This inevitably involves some local sliding movement of the diaphragm over the wall, which will certainly have an adverse effect upon its life. The greater the periodic change in the strain value, the greater will be this sliding movement with respect to the wall. As fig. 16 shows, the movement decreases as Δp increases, and this is in agreement with the results displayed in fig. 9: the greater Δp, the longer is the life of the diaphragm.

Summary. It has been found that the rolling diaphragm, a familiar type of positive seal for pistons or rods, can be employed at large pressure difference when supported by a cushion of fluid. The seal must be designed in such a way that the length of the diaphragm remains very nearly constant during the rolling cycle. Two possible designs are described, one embodying a stepped oil space, the other a floating piston-cap. Both are equipped with control devices that keep the pressure difference across the diaphragm (and hence also its length) constant to within very close limits. Diaphragms made of polyurethane rubber, used in seals of this type, have proved to have extremely long life, standing up to more than 2×10^9 rolling cycles at a rate of 1500 cycles per minute. The influence of various parameters is investigated and a theoretical study is given. Applications of this kind of seal in a gas refrigerating machine, a hot-gas engine and an oil-free compressor are dealt with. The last-mentioned machine is described in rather greater detail.

Solving a chessboard puzzle with a computer

A A A A A E E E A A B B B B B E I I B J G G G E I I F J G H H E E I F J G K H H E F F D D K K H E F F C D D D D E E E C C C C C	C C C C C I I I E E E E C I I B E F F B B B B B E F F F F G G G E J A A A A A G E J A A K H H G E J D D K K H H E E E D D D D H	F F A A A A A J F F A A E E E J F B B B B B E J F B H H I I E G E H H I I I E G E H K D D G G G E K K C D D D D E E E C C C C C	H K K E E E E E H H K E J J J E G H H E E E E E G C C C C C A A G G G F F C A A I D D F F F F A I I D D D D B A I I B B B B B A
A A A A A E E E A A B B B B B E K K B H H I I E K D H H I I I E D D H F F E E E D G C F F F F E D G C C C C C E D G G G J J J E	C C C C C K B B E E E E C K K B E F F F F A A B E D D F F A A B E J D D D D A B E J I I I H A G E J I I H H A G E E E H H G G G	G G G D C E E E G F F D C E J E G F F D C E J E B F D D C E J E B F D C C E E E B A A A A A H H B A A I I H H K B B I I I H K K	K K B B G G G J K I I B A A G J H I I B A A G J H H I B E A D C E H H B E A D C E E E E E A D C E F F F F D D C E E E F F D C C
A A A A A E E E A A F F J J J E C C F F F F H E C E I I I H H E C E I I H H K K C E D D D D K G C E E E B D D G B B B B B G G G	F F A A A A A E F F A A J J J E F B D D E E E E F B I D D D D C E B I I G G G C E B I I H H G C E B B H H K G C E E E H K K C C	H H B B G G G J I H H B A A G J I I H B A A G J I I K B E A D C E K K B E A D C E E E E E A D C E F F F F D D C E E E F F D C C	



With a view to exploring the possibilities of a computer, it is useful to attempt to programme it for widely differing problems, including strictly non-arithmetical ones. This was apparently as good an excuse as any for some of the staff of the Philips Computing Centre to use their largest computer for solving one of the puzzles of the Peter Pan series. In this puzzle a chessboard has been divided into 12 checkered pieces as shown in the inset; these 12 pieces should be put together again to form the chessboard. After suitable programming of the problem and feeding the programme into the computer, the solution consisting of 11 essentially different arrangements was obtained in the printed form shown here. "Computing" time: 8 minutes.

(See: A.J.Dekkers and A.J.W.Duijvestijn, Solving a chessboard puzzle with the PASCAL, Philips tech. Rev. **24**, 157-163, 1962/63.)

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.

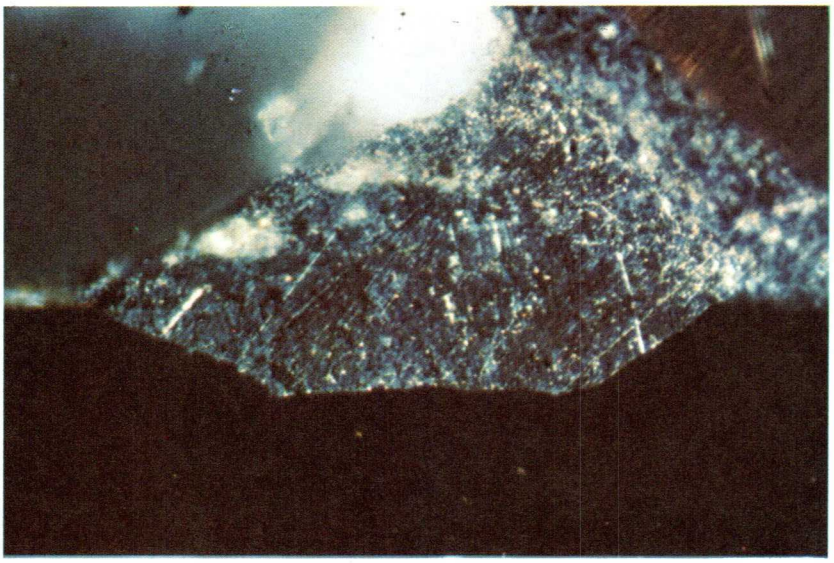


Fig. 1. Cross-section of a *P-N* junction in GaP with Au contact, photographed in polarized light on Kodachrome film (artificial-light reversal film), exposure 30 seconds. The *P-N* junction is roughly 0.5 mm long.



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.

IN SEARCH OF A MEASURE OF PERCEPTUAL WORK

by J. M. WESTHOFF *).

159.937:65.015.14

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

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

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

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

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

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

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

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

Measuring the amount of information

The bit

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

*) Institute for Perception Research, Eindhoven.

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

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

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

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

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

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

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

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

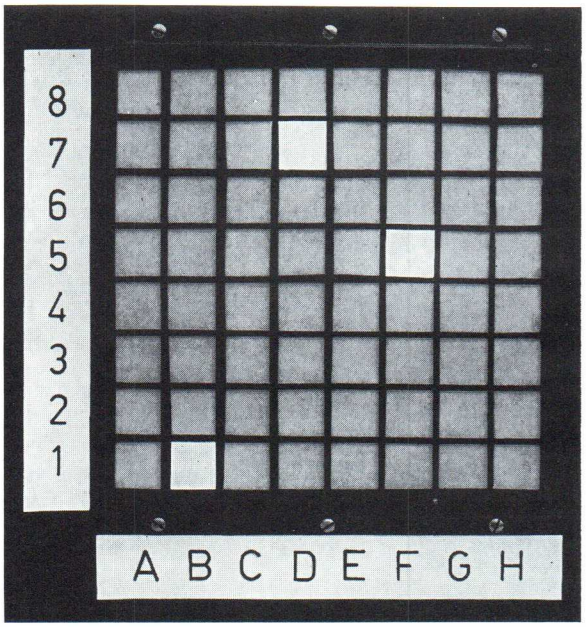


Fig. 1. The optical square. The subject must give the coordinates of the smaller squares which are illuminated for a short time (e.g. 0.1 s) by the experimenter. In this situation, the identification of one square corresponds to the intake of an amount of information equal to $\log_2 64 = 6$ bits; two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

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

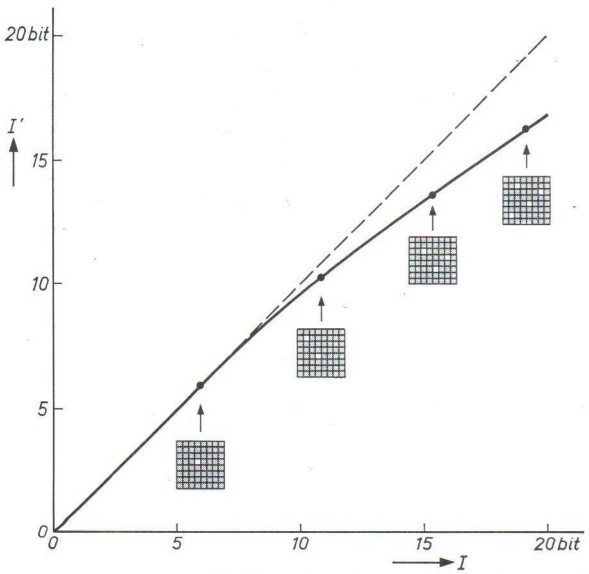


Fig. 2. The amount of information I presented in the optical square and the amount of information I' taken in by the subject are set out along the horizontal and vertical axes respectively. When more than about 9 bits are presented in 0.1 s, a steadily increasing proportion of the information is lost.

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

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

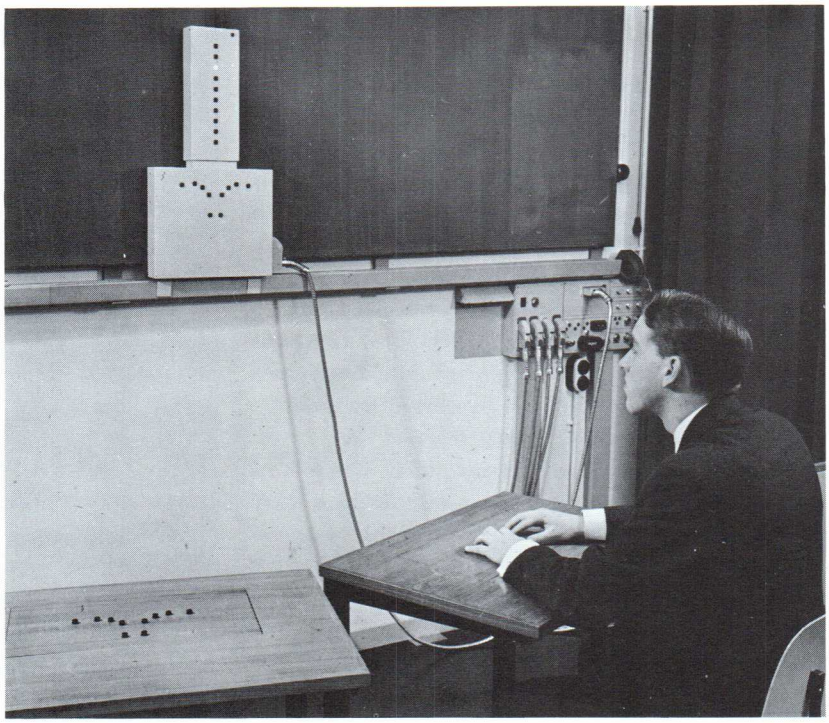


Fig. 3. Set-up which can be used to investigate the influence of the compatibility (degree of similarity between stimulus and response) on the reaction time. The subject must react as quickly as possible to the flashing of one of the lights on the stimulus panel (top left) by depressing the appropriate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

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

Compatibility

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

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

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

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

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

Time measurements

The performance time

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

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

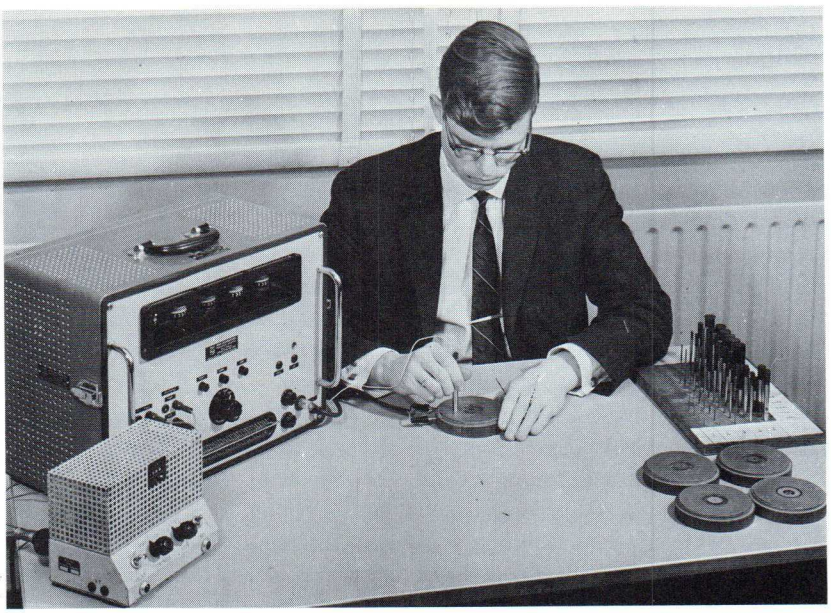


Fig. 4. Set-up for measuring the performance time for pin mounting (sticking a peg in a hole). The time taken for the operation is measured by the electronic counter on the left of the subject. This counter is set into action as soon as the pin leaves the starting circle (thus breaking an electric circuit),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

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

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

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

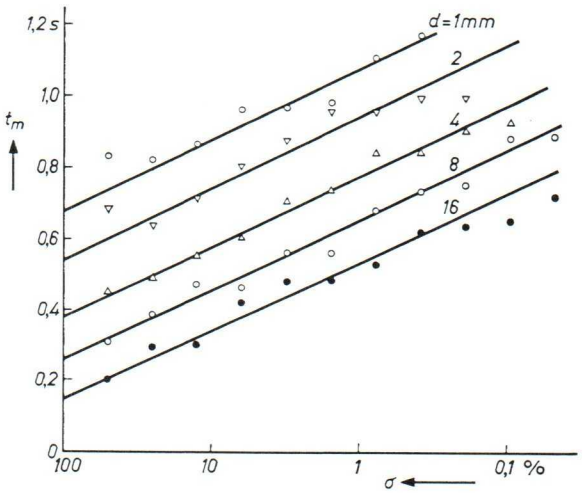


Fig. 5. Relationship between the assembly time t_m and the relative tolerance σ of the pins for various values of the diameter d of the hole. It may be seen that a logarithmic relationship exists both between t_m and σ and between t_m and d.

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

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

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

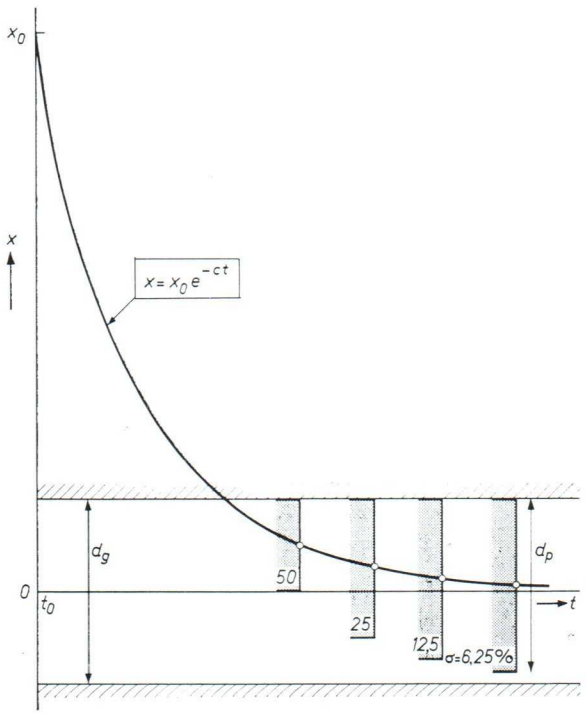


Fig. 6. In explanation of the logarithmic relationship between the assembly time t_m and the tolerance σ. If the error x (i.e. the distance between the axis of the hole and that of the pin) always decreases by the same factor in the same length of time, the pin (diameter d_p) moves to the hole (diameter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



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

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

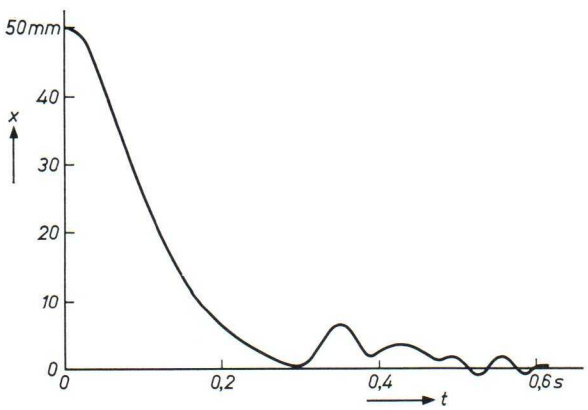


Fig. 8. The distance-time diagram for the assembly operation of fig. 7.

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

The reaction time

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

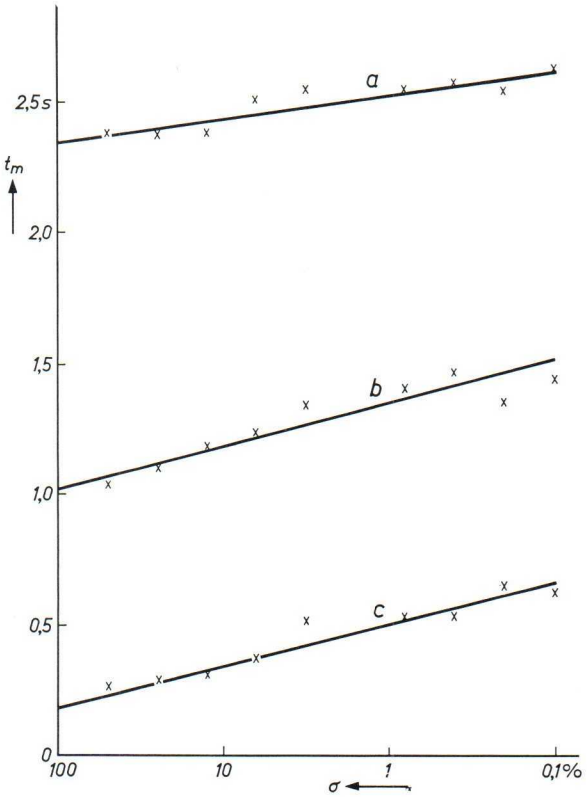


Fig. 9. Relationship between the assembly time t_m and the tolerance σ when the pin mounting is carried out *a*) slowly, *b*) at a medium rate, *c*) as fast as possible. The intercepts made by the various lines on the vertical axis represent the times needed to bring the pin from the starting circle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

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

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

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

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

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

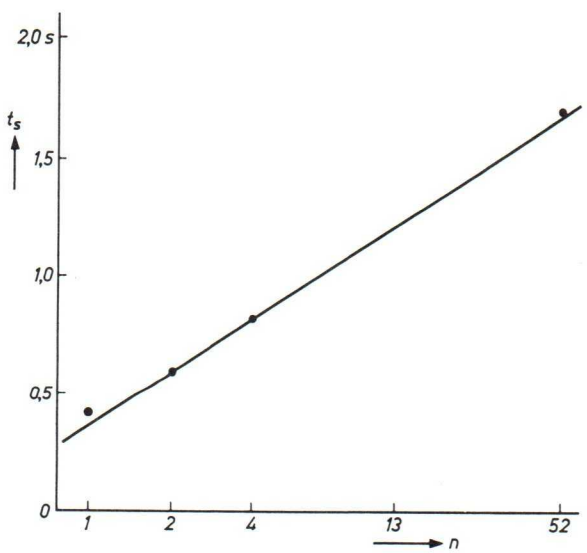


Fig. 10. Average dealing times t_s as a function of the number n of piles into which a pack of cards must be dealt. A linear relationship is found between the mean dealing time and the logarithm of the number of piles.

Dual-task situations

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

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

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



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

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

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

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

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

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

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

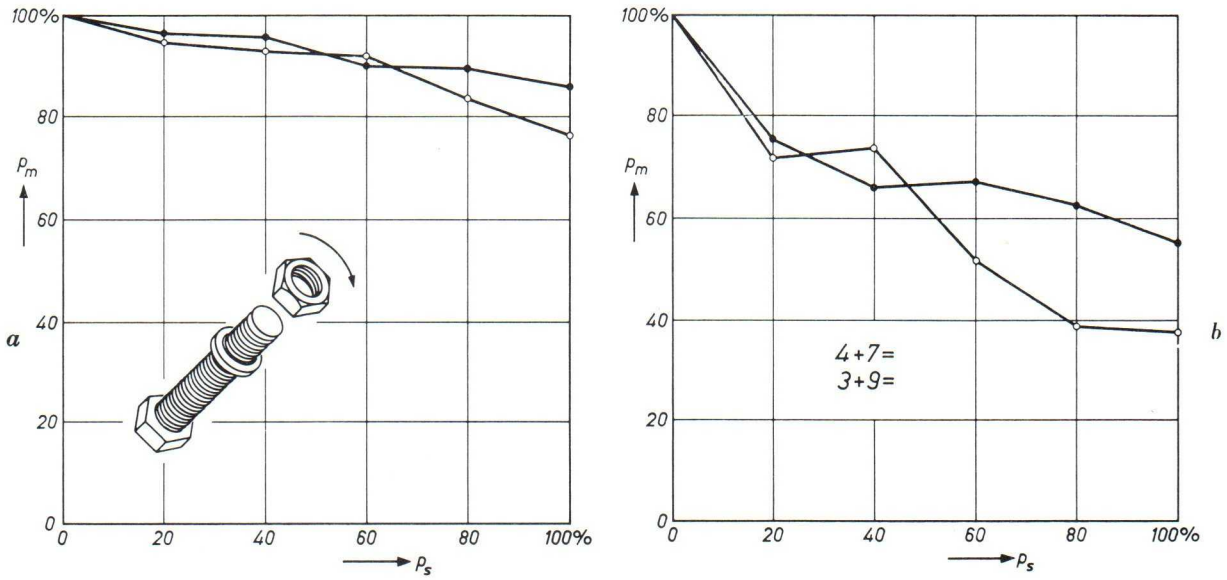


Fig. 12. Performance p_m in two perceptual tasks, *a*) assembly and *b*) simple arithmetic, as a function of the performance p_s in the standard task, in each case for two subjects. The performance is measured as the number of operations which the subject can carry out per unit time under the given conditions,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

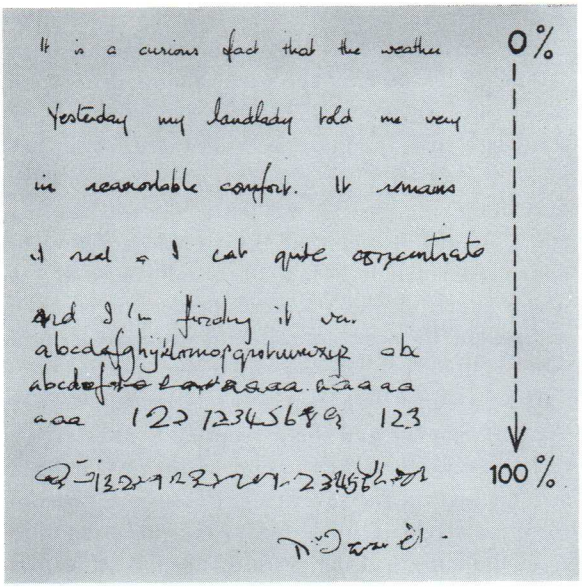


Fig. 13. Example of spontaneous writing (taken as the "trial" task) as the rate of execution of the standard task is gradually increased. Each line in the figure is taken from a whole sheet written by the subject while carrying out the standard task at a certain rate. It may be clearly seen that as the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

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

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

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



Tele-cardiotachometry

The apparatus worn by the girl in the photograph is the sensing element of an apparatus devised for continuously measuring the heart rate of persons moving freely while engaged on their normal work. Such measurements are an important means of investigation into the physiological aspects of, for instance, industrial production processes. So as not to hamper the subject in his work, the heart beat is sensed by a tiny lamp-and-phototransistor combination which is clipped to the ear and which picks up the light transmitted through the lobe: the changes in light transmission accompanying the periodic changes in blood flow will result in voltage pulses which can be transmitted to a distant recorder by a miniature radio transmitter worn by the subject. The apparatus was developed by the Radio, Gramophone and Television Division for the Philips Health Centre in Eindhoven.

(See: G.A.Harten and A.K.Koroncai, A transistor cardiotachometer for continuous measurements on working persons, Philips tech. Rev. **21**, 304-308, 1959/60.)

FLUORESCENT PIGMENTS AS AN ARTISTIC MEDIUM

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

667.65:667.624.48:535.612

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

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

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

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

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

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

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

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

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



a *b*

Fig. 1. Reproduction of the fluorescent painting
a) when exposed to ultra-violet radiation,
b) when exposed to daylight.
In the case *a*) two "Biosol" mercury lamps were used for irradiation. Filters of nickel-oxide glass absorbing practically all the visible rays were placed in front of the lamps. In front of the lens of the camera a filter was used which absorbs the ultra-violet rays. The latter was necessary because the painting reflects a large part of the incident ultra-violet light, to which the photographic plate is sensitive.
The specific impression that the fluorescent painting under ultra-violet irradiation makes upon the observer cannot, fundamentally, be fully represented by the reproduction; the reason for this is explained in the text.

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

Technique	Base	Brightness of white	Brightness of black	Brightness range
Water-colour	white paper	4.58	0.211	22
Pastel	grey paper	4.35	0.112	39
Gouache	white paper	5.85	0.130	45
Oil painting	painters' canvas	4.45	0.096	46

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

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

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

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

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

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

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

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

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

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

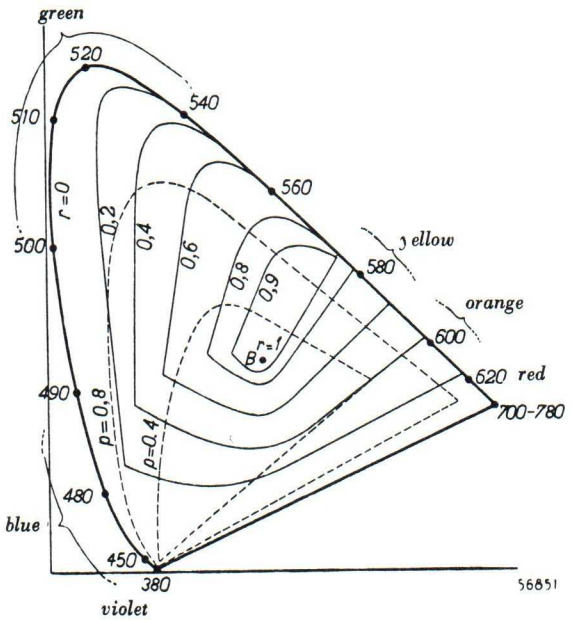


Fig. 2. In the chromaticity triangle, for each point characterized by the hue and the degree of saturation there is a maximum total reflection factor r that can be theoretically reached. (The pigment with that colour point must possess a spectral reflection curve of a certain shape to obtain the maximum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

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

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

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

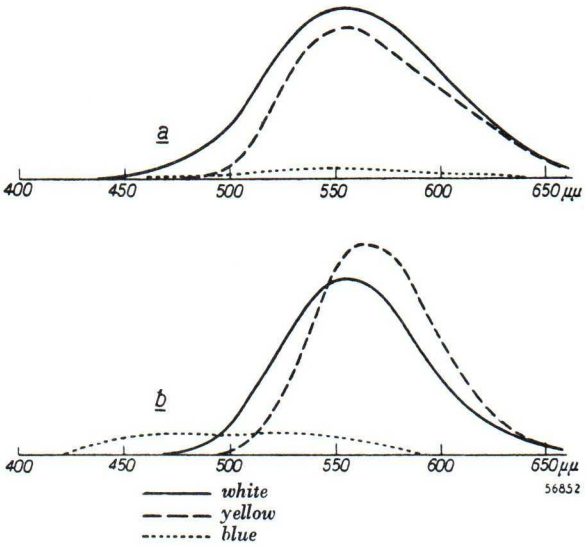


Fig. 3. *a*) The spectral distribution of the reflection has been measured for three normal oil paints, white, yellow and blue. For each wavelength the product of the spectral reflection coefficient and the relative visibility factor has been plotted. Thus the area underneath each curve gives a direct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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

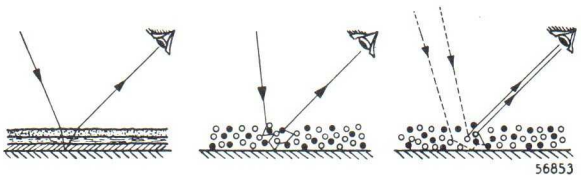


Fig. 4. *a*) In water-colour painting a mixed colour is obtained by applying different transparent layers of paint one on top of the other, each of which reduces a specific part of the spectrum in the transmission of the light: subtractive mixing. (Each layer consists in fact of innumerable grains of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

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

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

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

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

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

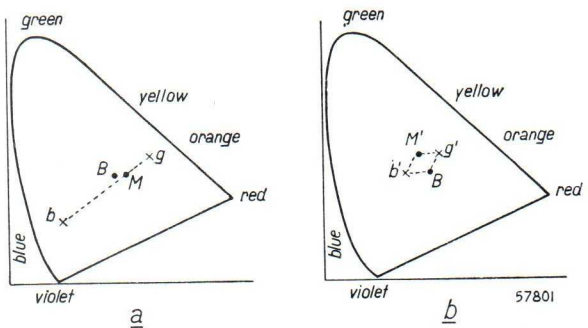


Fig. 5. *a*) Additive mixing of two colours represented in the chromaticity triangle by the points *g* and *b*. The colour point *M* of the mixture lies on the line through *g* and *b*; in the case drawn here (*g* a yellow, *b* a blue colour) by mixing in suitable proportions one can get the mixed colour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

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

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

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

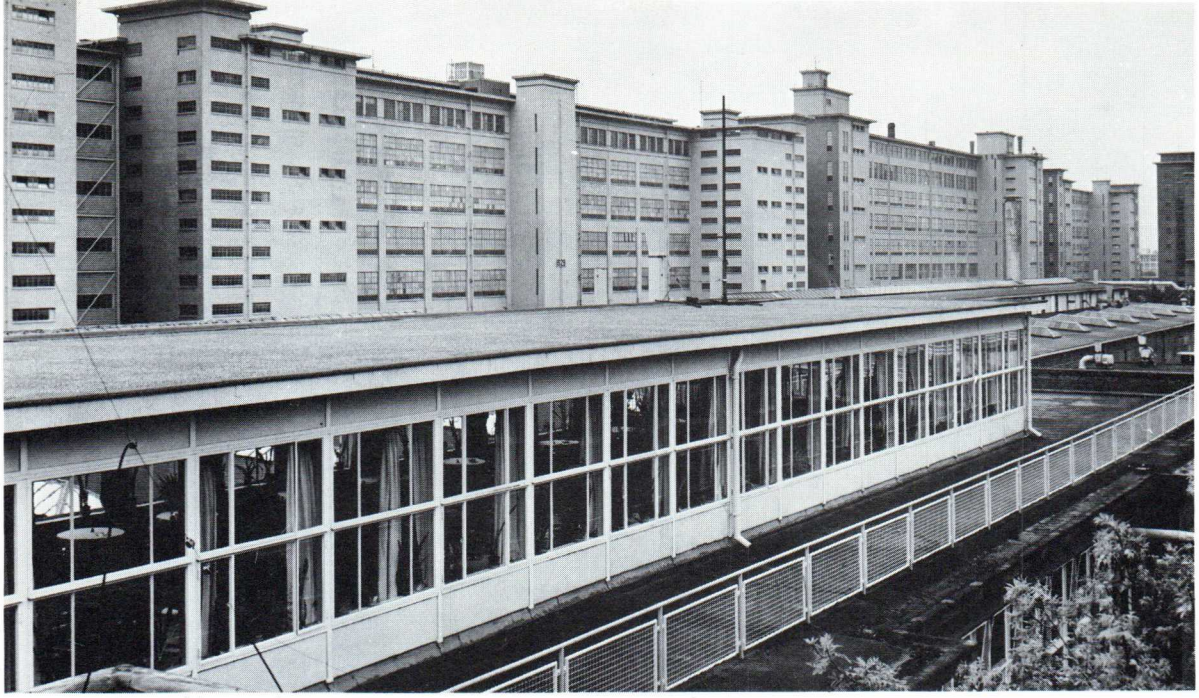
Finally a few words must be said about other practical questions which arise in the new technique. In order to avoid photo-chemical dissociation, as a result of which the "luminosity" of the picture would in course of time diminish, it is important that the crystals should be well protected against moisture.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

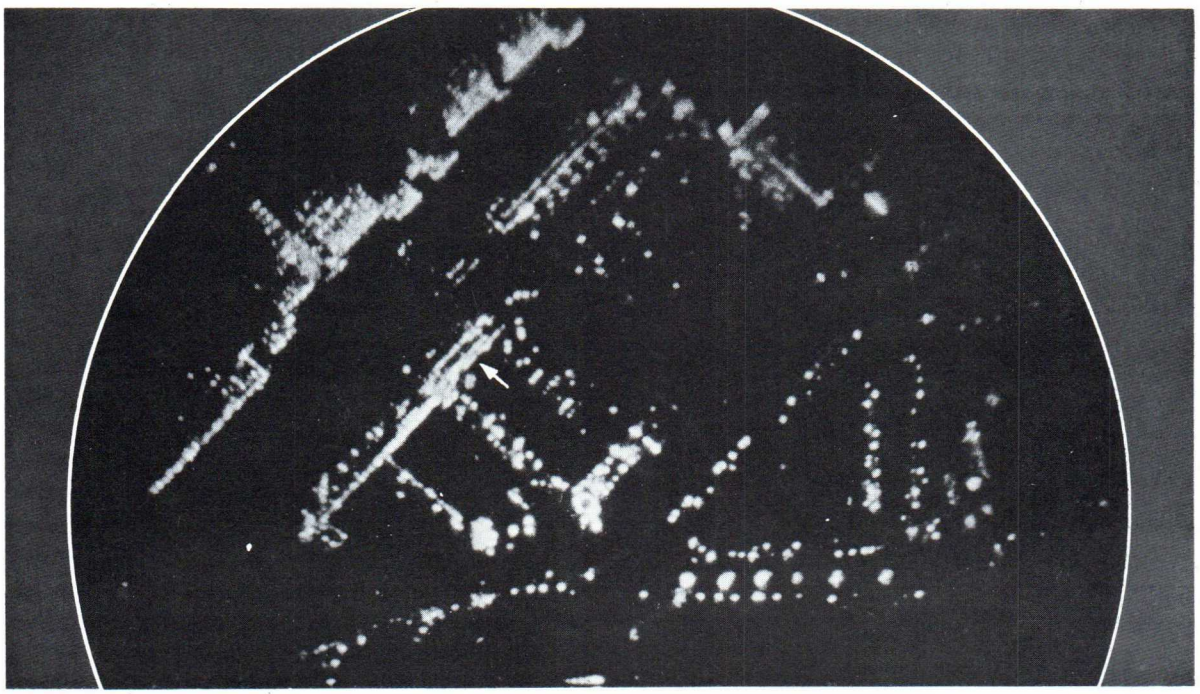
It almost goes without saying that while one is painting the canvas may exclusively be exposed to the light from the mercury lamp (with its filter). The subject, on the other hand, — if one is painting from nature — must be normally lighted. To keep the lighting of the canvas separate from the lighting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

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

May the experiment described and the experience communicated here induce painters to try out themselves the possibilities of the new technique.

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





4 mm radar

Radar installations operating at a wavelength of 8 mm are now widely employed, e.g. for maneuvering in busy water-ways and in other situations where a high degree of detail discrimination is required. Radar at a wavelength of 4 mm, using magnetrons, klystrons etc. now available, should provide even better resolution. This was corroborated by an experimental installation designed and built in the Eindhoven Research Laboratories mainly for exploring the technical problems involved (pulse duration as short as 5 nanoseconds, pulse repetition frequency as high as 12.5 kc/s for ensuring "hits" on the smallest relevant objects while scanning, etc.).

The lower photograph shows part of the radar display covering the Philips factories immediately surrounding the site of the radar installation. For this picture the range (whose maximum value was found to be no greater than about 1 km) was set at 300 m. The arrow in the display points to a canteen wall and a roof balustrade 3 m in front of it (see upper photograph) which are clearly resolved.

(See: H.Allaries, A 4 mm radar installation, Philips tech. Rev. **25**, 274-286, 1963/64.)



Generation of musical intervals by a digital method

D. Gossel

534.321.2:621.389

Introduction

The familiar kinds of musical instrument can be divided into two classes:

a) Instruments producing notes whose pitch is not decided upon until the instant of playing: bowed string instruments and certain wind instruments are examples.
b) Instruments possessing a store of notes, from which in the course of playing a selection is made in accordance with a programme. All keyboard instruments belong to this class.

Instruments in class (b) can only be endowed with a limited store of notes for constructional reasons, and because the technique of execution might otherwise be rendered too difficult; also, the access time for whatever notes are available must be compatible with practical requirements for playing the instrument. This implies the existence of some fixed rule or instruction for selecting individual tones from the continuum of pitch.

Several such rules have been laid down at various times in the history of music [1], they find practical expression in the various tonal or tuning systems. The four most important will now be briefly explained and discussed.

Tonal systems

A tonal system has been defined [2] as a scheme for dividing the octave into a progressive sequence of tones, the principle underlying the division being consistently adhered to and designed to produce musically acceptable intervals.

1. Pythagorean tuning

This system dates back to the philosopher who lived during the 6th century BC. It is based upon the fifth,

Dipl.-Ing. D. Gossel is a research worker at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

which represents a frequency ratio of 3 : 2 and is the simplest musical interval of all with the exception of the octave, with its frequency ratio of 2 : 1. A succession of musically new notes is produced when the gamut is transversed at intervals of a fifth, and in this respect the fifth differs from the octave. The notes of the Pythagorean scale are arrived at by superimposing n fifths in this way and then returning through m octaves to end up in the starting octave; the Pythagorean-type interval I_{Pv} is thus the result of multiplying by $(3/2)^n$ and dividing by 2^m. Accordingly, the underlying law is

$$I_{Pv} = \frac{(3/2)^n}{2^m} \; ; \; 1 \leqslant I_{Pv} < 2 . \quad . \; . \; . \quad (1)$$

In the Pythagorean system all fifths are true 3 : 2 intervals. The thirds, having a frequency ratio of 81 : 64, do not represent a straightforward harmonic interval and are classed as dissonant.

A property shared by Pythagorean tuning with all scales having "just" intonation, and some tempered scales, is that it does not form a closed system. It is impossible in principle to arrive exactly at an octave (say) by superimposing a "pure" interval like a true fifth upon itself, because the relevant frequency ratio represents a simple fraction that cannot yield a whole number when raised to a higher power (which is what the stacking process amounts to) [2]. In fact the octave is not one of the intervals that can be derived from eq. (1), the relation underlying the Pythagorean system.

Nowadays Pythagorean tuning possesses only historical interest; on account of its dissonant thirds, it can only serve as vehicle for a single melodic line.

2. Natural harmonic tuning

This system is the result of introducing a new interval of the simple harmonic type, the major third with its frequency ratio of 5:4, to supplement the fifth (3 : 2) and the fourth (4 : 3); the fourth is that interval which, added to a fifth, is needed to complete the octave. *Fig. 1* shows how all the other harmonic tones are engendered by the octave, fifth, fourth, major third and minor third (6 : 5), this last being the interval which, added to a major third, forms a fifth.

The natural harmonic system is quite a practical proposition for instruments in class (a) as defined above, but is unsuitable for tuning those in class (b) because too few of the resulting fifths are true to the harmonic series, and this fact limits the possibilities of modulation.

3. Mean-tone tuning

The system, proposed by Arnold Schlick in 1511, was quite commonly employed in the past, especially

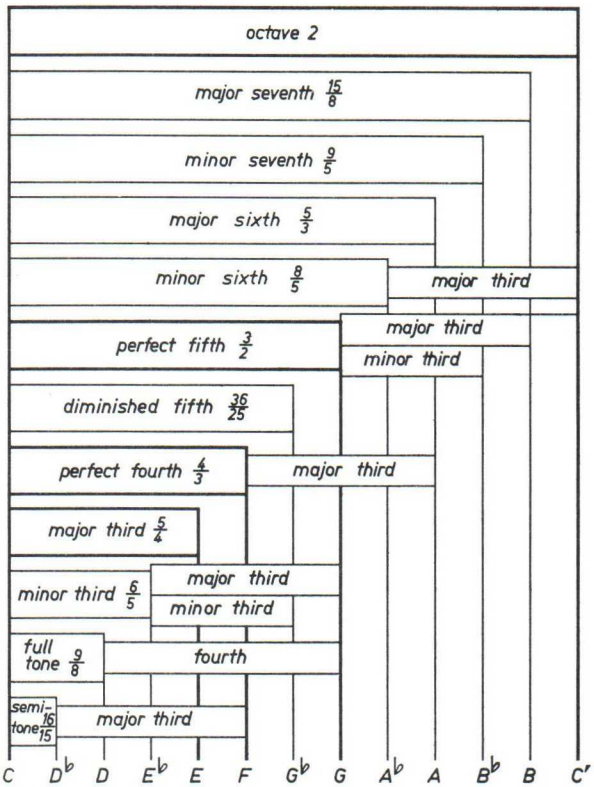


Fig. 1. Natural harmonic tuning: all the intervals of the scale arise out of the octave, fifth, fouth and third.

for organs. The resulting scale contains harmonically true thirds and tempered fifths. The stacking of four perfect fifths engenders the dissonant Pythagorean third (81 : 64), which exceeds the true major third (5 : 4) by an interval known as the syntonic comma (81 : 80). As compared with true perfect fifths, the tempered fifths of the mean-tone system are too flat by a quarter of a syntonic comma, i.e. by $\sqrt[4]{81/80} \approx 1.003$; the difference is too small to be disturbing, and four of these mean-tone fifths engender a true major third. The mean-tone system provides as many as eleven musically acceptable fifths, but the twelfth fifth needed to close the circle is far too wide and scarcely tolerable to the ear (it used to be known as a "wolf").

4. Scale of equal temperament

Practically all kinds of keyboard instrument nowadays are tuned to the scale of equal temperament (or "well tempered" tuning system). Essentially, the system is due to Andreas Werckmeister and Georg Neidhard (c. 1700) although earlier mention of it is to be found in the works of Ramis de Pareja (1440-1500). The system closes back on itself via twelve equally

[1] W. Dupont, Geschichte der musikalischen Temperatur, Bärenreiter-Verlag, Kassel 1935.
[2] Adapted from W. H. Westphal, Physikalisches Wörterbuch, Part II, Springer, Berlin 1952, pp. 546 and 547.

tempered fifths. The tempering consists in the division between these twelve fifths of the ditonic comma (531 441 : 524 288), the interval by which the octave is exceeded when twelve harmonically true fifths are superimposed. This means that the equally tempered fifth is only a twelfth of a ditonic comma (or about 1.001) flatter than a true perfect fifth. To put it another way, the scale of equal temperament divides the octave into twelve exactly equal semitones, or intervals having a magnitude of $\sqrt[12]{2}$. The advantage of having a closed tuning system is that one can modulate successively from one key to another *ad libitum*, but this is only at the price of major thirds that are a little too sharp and minor thirds that are a little too flat. However, as experience has shown, the ear becomes accustomed to this accentuation of the major and minor character of the relevant modes.

Fig. 2 is a representation of the four tuning systems just discussed. The octave covers an angle of 2π in this diagram. The magnitude of any interval can be expressed in "cents", a logarithmic unit divised by H. Bellerman and H. J. Ellis round about 1880. The multiplication of frequency ratios is thereby reduced to the addition of the corresponding cent values; the formula is

$$\frac{i}{\text{cent}} = {}^{1200}\sqrt{2}\ \log I. \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Thus the octave, an interval of $I = 2$, has a cent value of $i = 1200$ and the equally tempered semitone, for which $I = \sqrt[12]{2}$, is equivalent to 100 cents.

On converting Eq. 2 to Naperian logarithms we obtain:

$$\frac{i}{\text{cent}} = 1730 \ln I, \quad . \quad . \quad . \quad . \quad . \quad (2a)$$

and the following approximate expression for small intervals such that $I = 1 + \varepsilon$:

$$\frac{i}{\text{cent}} \approx 1730\,\varepsilon. \quad . \quad . \quad . \quad . \quad . \quad . \quad (2b)$$

It will be seen from fig. 2 that the major third in the scale of equal temperament lies intermediate between the Pythagorean and the natural harmonic third.

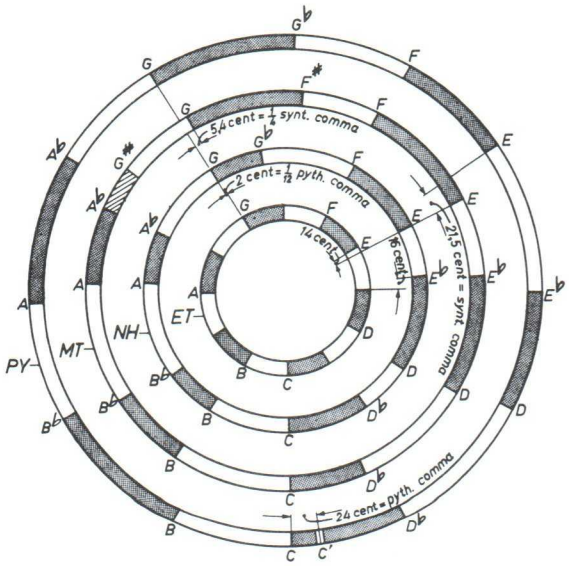


Fig. 2. The four most important tuning systems compared. The octave, having a cent value of 1200, extends over the full circumference of the circle (360°); accordingly, an angle of 1° in the above diagram represents a difference in pitch amounting to 3 1/3 cents.

ET = scale of equal temperament
NH = natural harmonic tuning
MT = mean-tone tuning
PY = Pythagorean tuning

Tone production and the tuning of musical instruments

Of particular interest within the framework of the present article are instruments belonging to class (b). With the exception of certain electromechanical organs (the Hammond Organ for example) they all have an independent oscillator for each note, or at least one for each of the twelve notes of the chromatic scale, these oscillators being switched on and off if required [3].

For reasons that are well known the oscillators (which may take the form of strings, wires, air columns, reeds or electronic LF generators) have to be "tuned", i.e. retuned, from time to time, and for instruments employing the system of equal temperament, the operation usually consists in correcting the pitch of a chain of fifths [4]. The tuner's professional skill lies in an ability to find the right tempering for the fifths; only then will the circle close in twelve steps. In principle, tuning could alternatively be done in a sequence of fourths, major sevenths or semitones. In fact the use of the semitone as a tuning interval would do away with the need to work back continually to the starting octave; this must invariably be done if any other interval is employed. But it is almost impossible to judge the tempering of a dissonant interval by ear, and so the consonant fifth and fourth are preferred.

The successful tuning of a musical instrument, in fact, calls for concentration, time and a trained musical ear. Specialists able to do the work are nowadays becoming fewer and fewer.

In the next section a small electronic device is described which permits of exact tuning to the scale of equal temperament in the shortest possible time. Adjustment of a note to the right pitch is done with a visual aid, the needle of a measuring instrument for example, and does not in any way necessitate a trained musical ear. Operation of the device is so simple that it might be worthwhile considering the desirability of

simplifying it together with certain keyboard instruments, occasional tuning of the instrument thus being left to the user. The device is also likely to be widely adopted for tuning church organs, a job that has to be done every so often on account of seasonal temperature changes. This routine is particularly laborious and time-consuming in the case of the larger instruments with their multiplicity of pipes and registers.

Generation of intervals by digital means

As is well known it is possible in digital technique to divide a given frequency f_0 by any desired whole number z. Where the value of z is on the large side the usual practice is to feed f_0 into a scaler which has been adjusted to recognize a preselected z value; having counted this number of input pulses, the scaler emits a zeroing pulse and starts to count anew. The zeroing pulses form a train with a recurrence frequency of $f = f_0/z$. It is an easy matter to select divisors z_i such that the corresponding recurrence frequencies f_i represent musically acceptable intervals.

If for example we choose $z_1 = 2$ and $z_2 = 3$, we shall obtain a true perfect fifth:

$$\frac{f_1}{f_2} = \frac{f_0/z_1}{f_0/z_2} = \frac{z_2}{z_1} = \frac{3}{2}. \quad . \quad . \quad . \quad . \quad (3)$$

All the other intervals are obtainable in a similar way. The great advantage of the digital method of generating musical intervals is that the intervals produced are independent of f_0. The absolute position of an interval in the gamut can thus be changed by varying f_0, and this fact can be exploited for the purpose of transposition. Digital equipment and methods could conceivably be used for carrying out investigations into musical aesthetics, enabling concords with various degrees of tempering to be quickly and conveniently produced and compared, and appraised as a function of absolute pitch.

A digital tuner

The intervals between notes in the scale of equal temperament are given by the law underlying the system:

$$I_{G\nu} = \sqrt[12]{2^{\nu}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Now, ν is always a positive integer, so the above relation yields irrational numbers which do not correspond exactly to any interval that can be produced by digital means (the only exception is the octave, for which $\nu = 12$). However, W. Schott of this laboratory found that the quotient of 196 : 185 approximates very closely to the equal-tempered semitone, whose value is $\sqrt[12]{2}$, the error being only 5×10^{-6}. It is on this, and on the fact that the absolute pitch of intervals generated by frequency division can be varied at will,

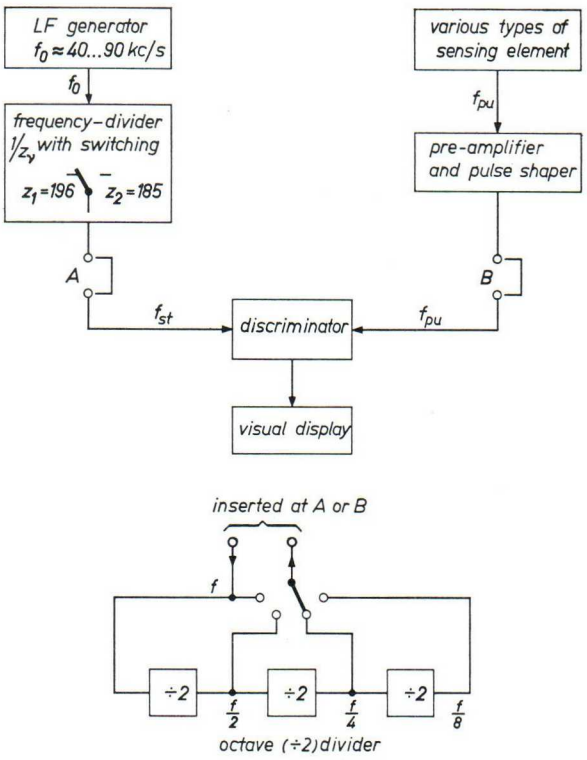


Fig. 3. Arrangement for tuning a musical instrument to the scale of equal temperament. The sensing element can take the form of a microphone, for example, or a magnetic pick up; none is needed for tuning electronic musical instruments. A moving coil measuring instrument, an electric lamp, a magic eye or the like can serve as a visual display device.

that the tuner represented schematically in *fig. 3* is based.

The output waveform from an LF generator whose frequency f_0 is continuously variable over the range between 40 kc/s and 90 kc/s, which covers little more than an octave, is fed to a frequency-dividing stage which has facilities for division by $z_1 = 196$ or by $z_2 = 185$, as desired. Thus two frequencies, $f_1 = f_0/196$, and $f_2 = f_0/185$, are alternatively available from the output of frequency-divider A; and provided f_0 is constant, the separation between them is almost exactly equivalent to an equally-tempered semitone. The frequency f_{pu} of the note to be tuned is picked up by a sensing element whose nature depends on the musical instrument being dealt with, and after amplification and conversion into a pulse train it is applied to a discriminating circuit in which it is compared with either f_1 or f_2, one of these serving as a standard or test frequency f_{st}.

The difference-indicating arrangements are phase-sensitive; they should preferably take the form of a visual display. As the picked-up frequency is gradually

[3] D. Wolkov, Electronic organ tone generators, Audio **46**, No. 2, 34-44, and No. 3, 30, 32, 65, 1962.

[4] O. Funke, Theorie und Praxis des Klavierstimmens, published by Das Musikinstrument, Frankfurt a.M. 1958.

adjusted to exact equality with the standard frequency a fluctuation in the visual display becomes slower and slower and finally stops altogether; the fluctuation may appear in the movement of a needle against a scale, ceasing when the needle finally comes to rest, or in the dimming and brightening of a small electric lamp, which finally acquires a constant brightness level.

The tuning procedure is as follows.

a) Adjust the LF generator to give its standard frequency of 81 400 c/s, which can if desired be controlled by a built-in quartz crystal. In switch position $z_2 = 185$ a standard test frequency of $f_{st} = 440$ c/s will be obtained, and can be used to correct the A above middle C on the musical instrument being tuned. It is scarcely necessary to point out that this standard A can be adjusted to any other desired pitch — 435 c/s for example.

b) Switch now to $z_1 = 196$, with the result that the standard frequency is lowered by a semitone. This new standard can be used to tune A flat above middle C on the instrument.

c) Switch back to $z_2 = 185$ and decrease f_0 until f_{st} is in unison with A flat on the instrument, as just corrected.

d) Switch to $z_1 = 196$ and tune G above middle C.

e) Switch back to $z_2 = 185$ and decrease f_0 until f_{st} is in unison with G on the instrument, as just corrected.

f) Switch to $z_1 = 196$ and tune F sharp above middle C, and so on.

In the course of twelve downward semitonal shifts, carried out in the manner described above, all the required intervals can be found and one finishes up an octave below the note first tuned. Taking account of the systematic error involved in each semitonal shift, which is 5×10^{-6}, the octave thus arrived at is true to within about 6×10^{-5}. This may be compared with the smallest deviation from unison that the ear is capable of perceiving in the most sensitive range of its response curve, around 1 kc/s; this smallest detectable difference is 4×10^{-3}, or two orders of magnitude greater than the error in the octave.

Where facilities are required for tuning the parallel octaves in bass or treble along with the twelve notes of the middle register, it is an easy matter to incorporate a chain of octave-dividers at *A* or *B*, i.e. on the standard-frequency or pick-up side of the tuner, the lower registers being covered in the former case and the upper ones in the latter. These octave-dividers are straightforward bistable circuits ("flip-flops") that divide the incoming frequency by two.

One type of frequency comparator circuit is shown in *fig. 4a*; its mode of functioning is explained in figs. 4*b* and *c*.

It is assumed that both U_{st} and U_{pu}, the standard and picked-up voltages, can each assume either of two values only, namely zero or the negative maximum. Further, the ratio between the pulse duration and the recurrance period is assumed to be constant. The circuit functions as a NOR gate: the presence of either of the two negative voltages on the transistor input suffices to switch it on. The transistor only switches to the non-conducting state when both circuit inputs carry zero potential.

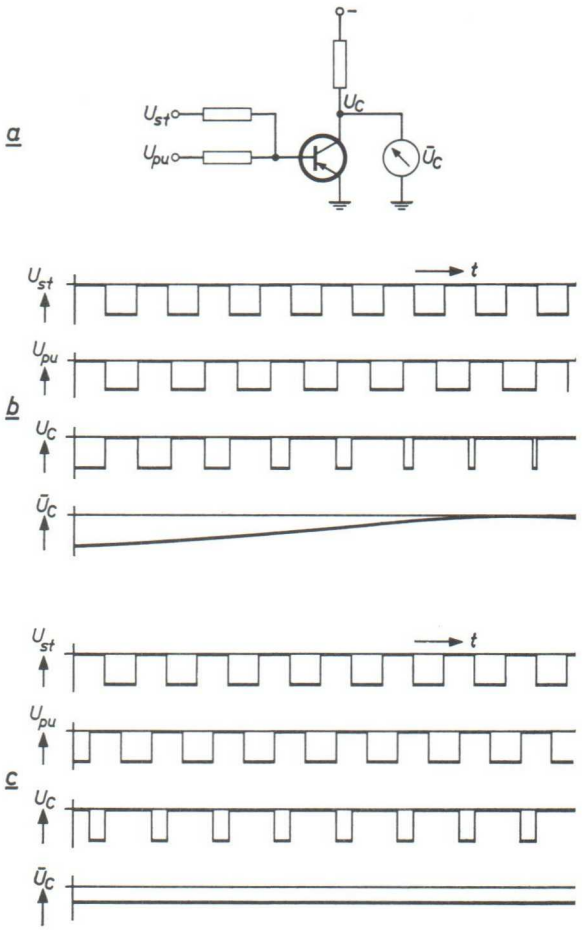


Fig. 4. *a*) Discriminator circuit.
b) Pulse trains operative in the case $f_{st} \neq f_{pu}$.
c) Shapes assumed by the same voltages when $f_{st} = f_{pu}$.

If the standard and picked-up frequencies differ (fig. 4*b*) there will arise at the collector of the transistor a train of pulses U_c whose breadth fluctuates cyclically. The corresponding mean voltage $\bar{U}_c$ also fluctuates cyclically, but is a continuous function of time; it is this quantity that is visualized. If on the other hand f_{st} is equal to f_{pu}, voltage U_c will be a train of pulses whose breadth is uniform, though dependent on the purely fortuitous phase relationship between the two incoming signals. The result will be a steady reading on the measuring instrument. The visual display method makes it possible to detect very slow fluctuations (down to a frequency of 0.01 c/s) that cannot be perceived by the ear.

The type of sensing element depends on the kind of musical instrument to be tuned. A microphone is to be preferred for picking up organ tones; for piano tuning a special magnetic pick-up is proposed, fitted with permanent magnet pole pieces which allow it to be attached to the wires on either side of the one being tuned, in such a way that it has a damping effect on vibrations in these neighbouring wires. The pick-up also embodies a small transistor amplifier coupled to a feedback circuit, which serves to keep the wire being tuned in a state of sustained vibration. For correct frequency discrimination it is necessary that the resonator under test should supply a continuous train of oscillations. Experiments have shown that damped vibrations, such as are produced by percussion of a piano wire, are useless for this purpose.

An electronic musical instrument like the "Philicorda" is the least demanding as regards the ancillary equipment required for tuning. No electro-acoustic transducer is required, because alternating voltages at the frequencies under test are available from the instrument anyway. In experiments on a "Philicorda" which had first been completely detuned in a random manner, it was possible with the aid of the digital tuner to bring the instrument back to exact conformity with the equal-temperament scale in a matter of barely ten minutes.

Musical instruments based on digital interval generation

Musical instruments of the purely melodic type (*fig. 5*) or of the type equipped for harmony (*fig. 6*) can be devised on the basis of this digital method of generating musical intervals; the "polyphonic" models are naturally dearer, the price depending on the number of melodic lines required.

Both types have much the same fundamental design as the digital tuner. One essential difference is that the purely melodic type incorporates one frequency divider with facilities for selecting 12 different divisors $z_1 \ldots z_{12}$, whereas the harmonizing type is equipped with n frequency dividers up to a maximum of 12, for each of which at least one divisor z_ν, and possibly a set of m such divisors, is available.

A feature common to all such instruments is that it is impossible in principle for them to go out of tune, since all the notes are produced by numerical division of a single master frequency f_0; and all can be transposed at will, simply by altering f_0.

The all-important factor governing the choice of divisors is the type of tonal system desired and, in some cases, the exactness or truth the individual intervals are required to have. To finish up, we shall briefly describe a possible choice for the natural harmonic tuning.

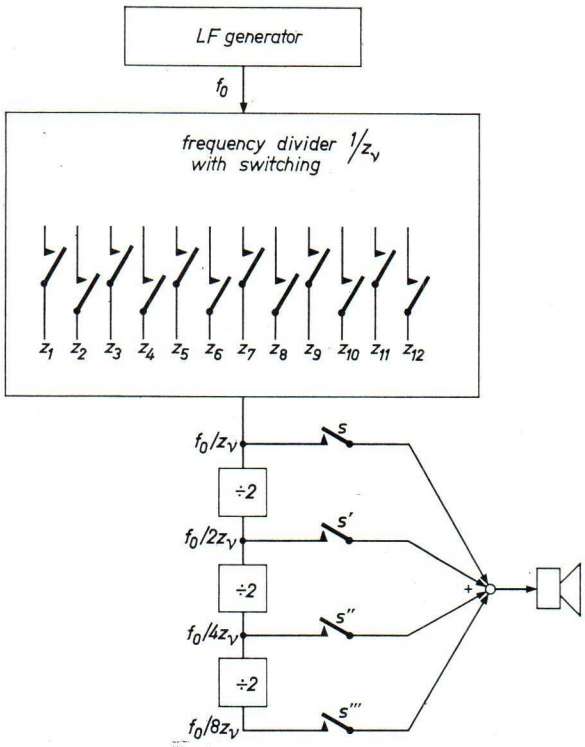


Fig. 5. Purely melodic musical instrument using digital interval generation.

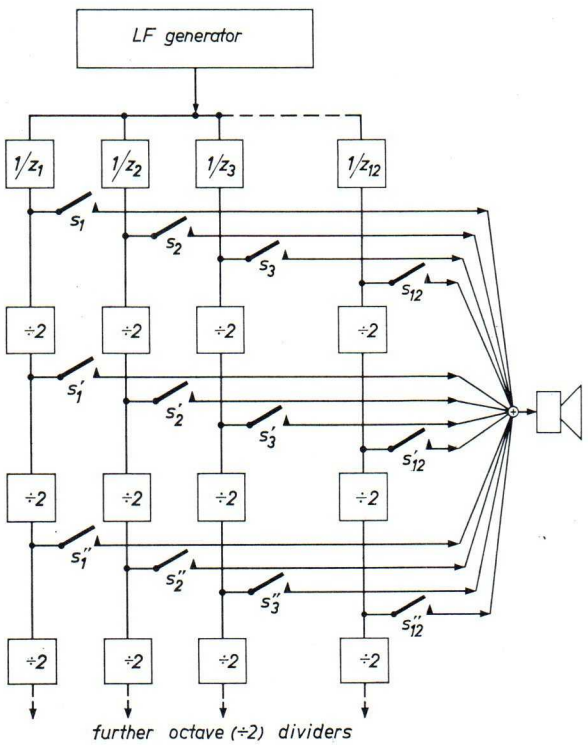


Fig. 6. "Polyphonic" musical instrument using digital interval generation.

The lowest note to be produced is associated with the highest divisor z_1:

$$f_1 = \frac{f_0}{z_1}. \quad \ldots\ldots\ldots \quad (5)$$

All the other tones are given by

$$f_v = \frac{f_0}{z_v}. \quad \ldots\ldots\ldots \quad (6)$$

where $z_v < z_1$. This means that the divisors must be

$$z_v = z_1 \frac{f_1}{f_v} = \frac{z_1 b_v}{a_v}, \quad \ldots\ldots \quad (7)$$

where a_v represents the numerator and b_v the denominator of any of the fractions appearing in fig. 1; these fractions stand for cancelled-out frequency ratios. Since all z values must be integers and no b_v value is a factor of its a_v, a sensible move will be to make z_1 the lowest common multiple of all the values assumed by a_v:

$$z_1 = 2\times2\times2\times2\times3\times3\times5 = 720$$

All the other smallest whole-number values of z_v can be found on inserting $z_1 = 2\times2\times2\times2\times3\times3\times5 = 720$ in eq. (7) (they are set out in *Table I*). The choice of

Table I. Divisors required for producing the intervals of a natural harmonic scale by the digital method, shown against the notes of the scale.

Note	f_v/f_1	z_v
C	1	720
D flat	16/15	675
D	9/8	640
E flat	6/5	600
E	5/4	576
F	4/3	540
G flat	36/25	500
G	3/2	480
A flat	8/5	450
A	5/3	432
B flat	9/5	400
B	15/8	384
C′	2	360

720 as highest divisor entails a certain outlay of circuit elements — 10 bistables, say, and the associated decoding circuits. This outlay can be greatly reduced if the designer confines himself to producing only the consonant intervals at their true values, accepting slight deviations in the dissonant ones.

For example, let us suppose that only the octave, fifth fourth and the major and minor thirds need be taken into account in fixing z_1. The highest divisor then becomes

$$z_1 = 2\times2\times3\times5 = 60.$$

The new set of divisors and the errors they involve are displayed in *Table II*; a positive sign means that the tone in question is too sharp. The quantity of

Table II. Divisors required for, and errors involved by, an approximation to natural harmonic tuning.

Note	z_v	Error	
		percentage	in cents
C	60	—	—
D flat	56	+ 4.4	+ 7.7
D	53	+ 6.3	+10.8
E flat	50	—	—
E	48	—	—
F	45	—	—
G flat	42	— 8.0	—13.8
G	40	—	—
A flat	38	—13.3	—23.1
A	36	—	—
B flat	33	+10.0	+17.3
B	32	—	—
C′	30	—	—

components required is now greatly reduced. Six bistables will be ample, and even so, seven of the intervals produced are true as against five which are not. Some of the deviations are rather large; for example, that in the minor sixth amounts to almost the fourth part of a semitone.

However, a better approximation can be obtained by choosing z_1 three times as large. The result of so doing is that the minor seventh and diminished fifth all become true intervals. With a highest divisor of

$$z_1 = 2\times2\times3\times3\times5 = 180$$

the number of bistables required is eight, and the tuning that results is as shown in *Table III*. The scale arrived at in this way would be quite acceptable musically.

Table III. Divisors and errors entailed by a closer approximation than that in Table II.

Note	z_v	Error	
		Percentage	in cents
C	180	—	—
D flat	169	—1.5	—2.6
D	160	—	—
E flat	150	—	—
E	144	—	—
F	135	—	—
G flat	125	—	—
G	120	—	—
A flat	112	+4.4	+7.7
A	108	—	—
B flat	100	—	—
B	96	—	—
C′	90	—	—

Summary. A method familiar from digital techniques, enabling a given frequency to be divided by any desired whole number, can be exploited for producing musically acceptable intervals. The more important tuning systems are briefly reviewed. A description follows of a device suitable for quick and accurate tuning of keyboard instruments; it produces an equally-tempered semitone, which can be transposed through the gamut at will. The true semitonal interval, whose value is $\sqrt[12]{2}$, can be closely approximated by performing the division 196 : 185, the error being only 5×10^{-6}. Also described is a procedure for arriving at a complete scale of notes for a melodic instrument, or one affording facilities for harmony. By way of example, numerical values are given for the intervals of natural harmonic scales obtainable by the digital method.

Nitrogen in silicon iron

J. D. Fast

546.17:621.318.13

Professor Fast has been a frequent contributor to our journal. The article below is a chapter from part I of his recently published book "Interaction of metals and gases" [*]. This chapter deals with methods of bringing about in silicon iron a special crystal orientation which gives the material favourable magnetic properties. New insight was obtained from experiments carried out by the author — about ten years ago — in which he added for the above mentioned purpose small quantities of a second phase to the metal. His experiments gave the impetus to extensive investigations in various countries, which have led among other things to better control of the production of grain-oriented silicon steel, used for making transformer cores.

Introduction

Nitrogen can play an important part in silicon steel that is used as the "iron" core of transformers, electric motors and electric generators. The steel which is used in transformers generally contains about 3% silicon.

In the earliest transformers the core was made from unalloyed steel which was soft-annealed to make it as far as possible free of internal stresses. The coercivity and hysteresis losses in this material were relatively large as a result of the many inclusions it contained, especially inclusions of iron carbide. Also the eddy current losses were relatively high due to the small electrical resistivity. It was especially unfortunate that the coercivity and hysteresis losses spontaneously increased with the course of time. This "magnetic ageing" was caused by the slow precipitation of nitrogen in the form of iron nitride [1].

The silicon steel used in transformers suffers smaller eddy current losses than unalloyed steel due to its greater resistivity. Also the hysteresis losses are smaller because silicon encourages the formation of graphite which is magnetically less harmful than iron carbide, since for the same number of carbon atoms the total volume of inclusions is much smaller. It is also of importance that silicon steel exhibits no ageing phenomena because the nitrogen which is present as an impurity occurs, after a suitable annealing treatment, in the form of a very stable silicon nitride [2]. The composition of this nitride is Si_3N_4 [3]. However, this precipitate too must be considered undesirable since, like all other precipitates, it has an unfavourable effect on the coercivity and the hysteresis losses. In the following, however, we shall see that by the deliberate addition of nitrogen to silicon steel one can profit from the presence of this element to obtain a magnetically favourable crystal orientation.

The role of Si_3N_4 in making grain-oriented silicon steel sheet

Until comparatively recently virtually all the silicon iron sheet for transformers was obtained by hot-rolling The directions of easy magnetization, <100>, of the separate crystals in the sheet are then almost randomly distributed over the various directions in space (*fig. 1*), so that the hysteresis losses are relatively large. In principle it would be most desirable to make the cores

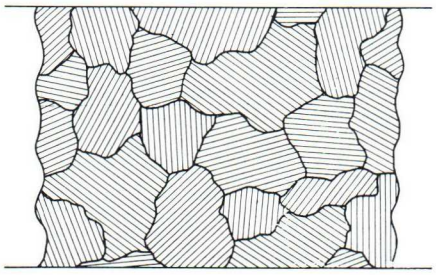


Fig. 1. The figure shows schematically that in hot-rolled silicon steel sheet the crystals show no preferential orientation with respect to the plane and direction of rolling.

Prof. Dr. J. D. Fast is a research worker at Philips Research Laboratories, Eindhoven, and a Professor Extraordinary of Physical Chemistry at the Technical University, Eindhoven. On 9th January 1965 Prof. Fast received from the Technical University of Delft the degree of Doctor honoris causa "for his outstanding services in the field of scientific and technical research on metals".

[*] J. D. Fast, Interaction of metals and gases I, published by Philips Technical Library, Eindhoven, and Academic Press, New York, 1965.

[1] J. D. Fast, Philips tech. Rev. **13**, 165, 1951/52; J. D. Fast and L. J. Dijkstra, Philips tech. Rev. **13**, 172, 1951/52.
[2] J. D. Fast, Philips tech. Rev. **16**, 341, 1954/55.
[3] W. C. Leslie, K. G. Carroll and R. M. Fisher, Trans. AIME **194**, 204, 1952; H. A. Sloman, J. Iron Steel Inst. **182**, 307, 1956.

from large laminar single crystals (bounded by cube faces) in such a way that the magnetic flux always follows a direction of easy magnetization (*fig. 2*).

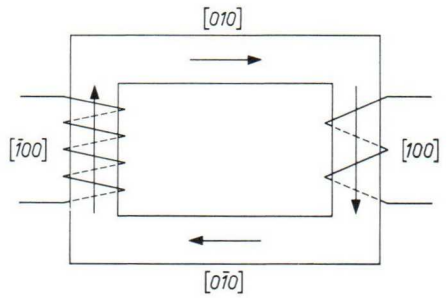


Fig. 2. Symbolic representation of a transformer, the core of which is built up of single crystal sheets in such a way that the magnetic flux can everywhere follow a direction of easy magnetization, a <100> direction.

Technically it is impossible to make single crystal sheet in large quantities. But it has been found possible to make on a large scale (in quantities of thousands of tons each month) polycrystalline silicon iron sheet, in which all the crystals have nearly the same orientation (*fig. 3*). This orientation is such that the crystals lie with a (110) plane approximately parallel to the surface of the sheet and with a [001] direction, a direction of

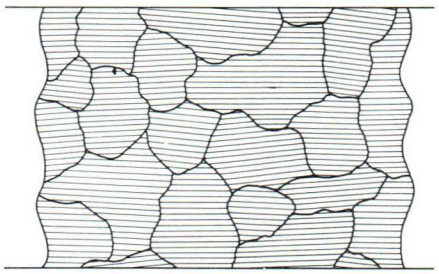


Fig. 3. Schematic representation of the crystals of cold-rolled 3% silicon iron sheet in which all the crystals have about the same orientation, viz. the orientation (110)[001] (cf. fig. 4).

easy magnetization, approximately parallel to the direction of rolling (*fig. 4*). This texture is often referred to as Goss texture after its inventor, Goss, but also as (110)[001] or cube-on-edge texture [4]. The latter name is illustrated by *fig. 5*, in which the orientation under discussion is demonstrated with the help of a number of cubes, which symbolize the unit cells. The Goss texture is obtained by cold-rolling silicon iron from a certain thickness and by subjecting it to certain heat-treatments.

For many years the way in which the crystal orientation in the Goss sheet is brought about was not understood and the production of the sheet did not always give the desired results. Our own experiments showed that it is impossible to obtain the Goss texture in silicon iron sheet made from pure iron and transistor-quality silicon. From this we concluded that the presence of impurities in the material is of essential importance [5]. In further experiments we added measured quantities, in each case of one element, to pure silicon iron alloys.

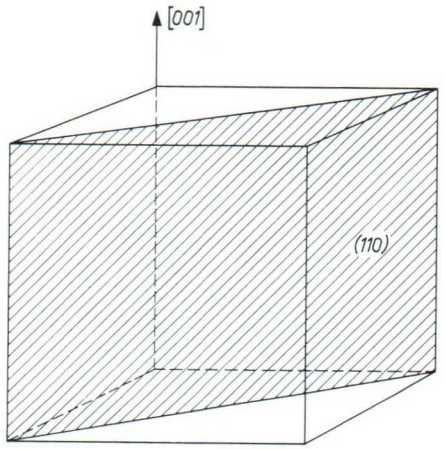


Fig. 4. In silicon iron sheet having Goss texture the crystals are oriented in such a way that they lie with a (110) plane approximately parallel to the rolling plane and with a [001] axis (cube axis) approximately parallel to the direction of rolling (cf. fig. 5).

It was found that the desired texture is readily obtained by introducing nitrogen (in quantities of a few hundredths of a percent) and heat-treating the metal before cold-rolling in such a way that it contains a finely divided precipitate of Si_3N_4, which is mainly present at the grain boundaries [6]. *Fig. 6* shows an electron microscopic photograph of 3% silicon iron, in which a precipitate of this type is present.

After cold-rolling and after primary recrystallization at 600° to 800 °C, both the pure silicon iron sheet and

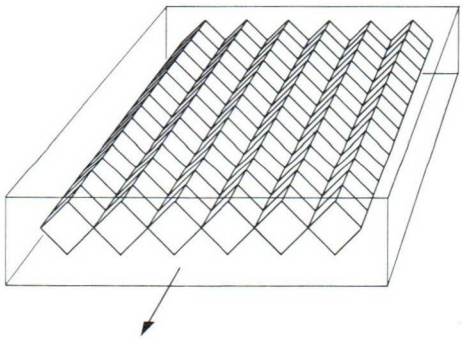


Fig. 5. Schematic representation of the crystal orientation in magnetic steel with Goss texture (cube-on-edge texture). The arrow indicates the direction of rolling.

the sheet containing Si_3N_4 contain only very few crystals with the orientation (110)[001]. In the pure alloy the primary recrystallization is followed at high temperature (e.g. 900 °C) by normal grain growth which exhibits no preference for a particular orientation. In silicon iron containing nitrogen the normal grain growth is inhibited by the Si_3N_4 precipitate. If the metal is heated in the appropriate temperature range in pure hydrogen these inhibiting inclusions slowly coagulate and go into solution, and at a certain point the few favourably oriented crystals begin to grow but not, as yet, the others. In order words secondary recrystallization (exaggerated grain growth) occurs, by means of which a few primary crystals with orientation (110)[001] grow at the cost of the other crystals to many times the sheet thickness. The driving force for growth of of these grains is the low gas-metal interfacial energy of the (110) surfaces in an atmosphere of pure hydrogen [7]. This surface energy is less than that of any (*hkl*) plane different from (110).

The selective grain growth under discussion occurs in a particular temperature range. If the nitrogen-bearing silicon iron is heated immediately before or after the primary recrystallization at too high a temperature (e.g. 1250 °C), then the active inclusions dissolve very rapidly, so that one obtains mainly normal grain growth, which results in a poor texture. Therefore one must first heat at a lower temperature (900 °-1000 °C). As soon as the (110)[001] crystals have a sufficient start in size on the other crystals, the temperature can be raised to accelerate further growth.

The Si_3N_4 particles, which are so useful for producing the Goss texture, are unfavourable for the final magnetic characteristics of the material. They must therefore be removed from the silicon iron after they have accomplished their grain-growth function. This takes place automatically in the final heat-treatment in an atmosphere of pure hydrogen, since virtually all the nitrogen then leaves the metal [5][8].

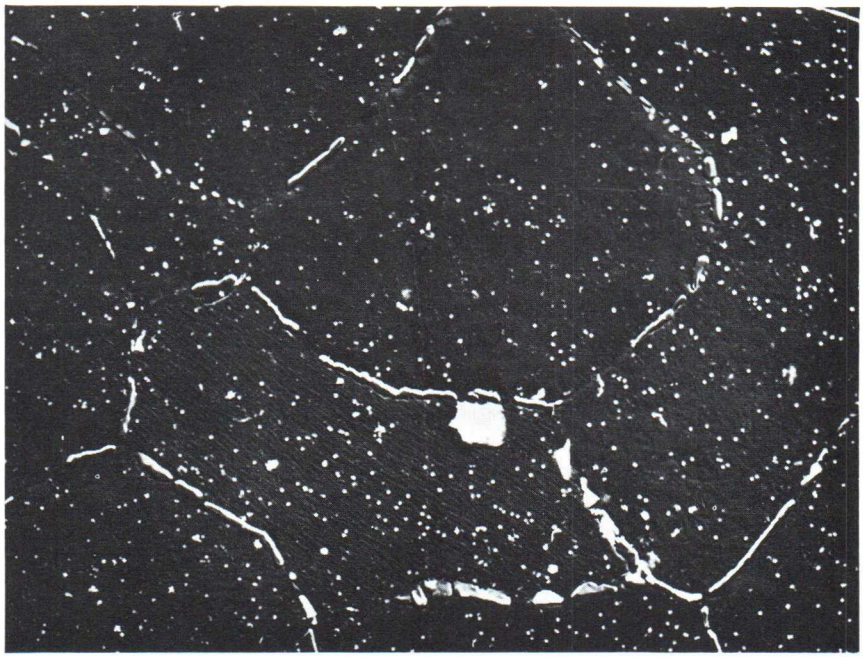


Fig. 6. Electron micrograph of 3% silicon iron in which a precipitate of Si_3N_4 is present along the grain boundaries. The micrograph corresponds to an area of 17×19 microns.

In the commercial 3% Si-Fe alloys, MnS is the most important impurity inhibiting normal grain growth after primary recrystallization [9]. Various other inclusions can also perform this task [10]. Here again it is of primary importance that they be present in the metal in

[4] N. P. Goss, Trans. Amer. Soc. Met. **23**, 511, 1935; R. M. Bozorth, Trans. Amer. Soc. Met. **23**, 1107, 1935; C. G. Dunn, Cold working of metals, Amer. Soc. for Metals, Cleveland 1949, pp. 113-120.
[5] J. D. Fast, Philips Res. Repts. **11**, 490, 1956.
[6] See also: J. D. Fast and J. J. de Jong, J. Phys. Radium **20**, 371, 1959.
[7] J. E. May and D. Turnbull, Trans. AIME **212**, 769, 1958; J. L. Walter and C. G. Dunn, Trans. AIME **215**, 465, 1959 and **218**, 1033, 1960.
[8] J. D. Fast and H. A. C. M. Bruning, Z. Elektrochemie **63**, 765, 1959.
[9] See the first article mentioned in note [7].
[10] H. C. Fiedler, Trans. AIME **221**, 1201, 1961; M. J. Markuszewicz, J. Iron Steel Inst. **200**, 223, 1962.

the desired size and distribution. For inclusions of MnS this has been shown most convincingly [11] by Fiedler. He experimented with a 3.3% silicon iron alloy which contained a little less than 0.1% MnS. This compound could be completely dissolved by heating the metal at 1325 °C. The most effective degree of dispersion could then be obtained in two ways: (*a*) by correct choice of the cooling rate, (*b*) by drastic quenching followed by precipitation heating at a lower temperature (1000 °C).

An advantage of Si_3N_4 over other grain-growth inhibitors is that after the final heat-treatment no impurities remain behind in the metal, so that the magnetic properties are particularly good.

The characteristics of silicon iron sheet with Goss texture deviate in some respects very little from those of single crystals. A disadvantage, however, is that the unfavourable [110] directions of the crystals lie in a direction perpendicular to the direction of rolling (cf. fig. 4). One can therefore only take full advantage of this material if the magnetic flux is everywhere parallel to the direction of rolling, i.e. if it is used in the form of ring cores wound from sheet and not, for example, in the form of E sheets.

For many applications it would be very desirable to have available silicon iron sheet with cube texture [12], i.e. sheet in which the crystals are so oriented that not only the direction of rolling, but also the direction perpendicular to it is a direction of easy magnetization ((100)[001] texture, *fig. 7*). In Germany and the U.S.A. research workers have already succeeded in making this material with (100)[001] texture on a small scale. In the production, the interaction between the metal surface and the surrounding gas atmosphere plays a very important part. If the oxygen activity of the gas exceeds a particular value, then it is no longer the (110) planes but the (100) planes which have a smaller surface energy than all other (*hkl*) planes, in other words, the driving force for growth of (100) grains is then greater than that of (110) grains.

A convincing demonstration of the above is given by experiments of Walter and Dunn [13] on the migration of (100)/(110) boundaries, i.e. boundaries between two grains, one of which has a (100) plane and the other a (110) plane parallel to the surface of the 3% silicon iron sheet. At 1200 °C the (100)/(110) boundaries advance into (100) grains in a good vacuum, then reverse their direction and migrate into (110) grains in an atmosphere of impure argon. The direction of migration reverses once again with (110) grains growing at the expense of (100) grains in a second vacuum anneal. These results are explained by the authors in terms of a change in concentration of oxygen atoms at the gas-metal interface during the anneals. The addition of oxygen atoms to the surface during the anneals in impure argon results in a decrease of the specific surface energy of the (100) oriented grains to a value lower than that of the (110) oriented grains. In a good vacuum or in pure hydrogen, however, the oxygen concentration at the surface is lowered to the point where the surface energy of the (110) grains has the lowest value.

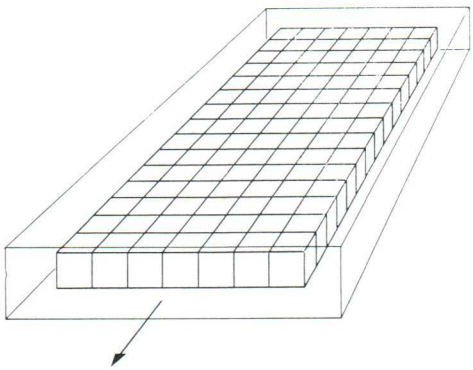


Fig. 7. Schematic representation of the cube texture (cf. fig. 5). The arrow indicates the direction of rolling.

The development of the cube texture in 3% silicon iron is, however, much more difficult and complicated than would be supposed from the foregoing. Control of the gas atmosphere in the final heat treatment is a necessary, but not sufficient condition for success. Up to now it has not been found possible to produce silicon iron sheet with cube texture economically in large quantities. The main difficulty seems to be getting the alloy into such a condition that, after primary recrystallization of the sheet, there is a sufficient number of crystals present with the required (100)[001] orientation. According to patents of the General Electric Company (U.S.A.) this aim can be achieved by starting with ingots having favourably oriented columnar crystals obtained by controlled directional solidification.

Summary. The article is a chapter from part I of the author's recently published book "Interaction of metals and gases". The chapter deals with the influence of nitrogen on crystal orientation and the associated magnetic properties of cold-rolled and recrystallized silicon iron, which is generally used for transformer cores. Given the right conditions, Si_3N_4 precipitates as a finely divided second phase which promotes the growth of crystals with a (110)[001] orientation ("cube-on-edge texture"). The MnS present in commercial types of silicon steel fulfils a function similar to that of Si_3N_4. Mention is also made of experiments aimed at producing silicon iron with a (100)[001] orientation ("cube texture"), which would be even better suited for the same purpose.

[11] H. C. Fiedler, Trans. AIME **230**, 95, 1964 (No. 1).
[12] F. Assmus, R. Boll, D. Ganz and F. Pfeifer, Z. Metallk. **48**, 341, 1957; F. Assmus, K. Detert and G. Ibe, Z. Metallk. **48**, 344, 1957; J. L. Walter, W. R. Hibbard, H. C. Fiedler, H. E. Grenoble, R. H. Pry and P. G. Frischmann, J. appl. Phys. **29**, 363, 1958; G. Wiener, P. A. Albert, R. H. Trapp and M. F. Littmann, J. appl. Phys. **29**, 366, 1958.
[13] J. L. Walter and C. G. Dunn, Acta metallurgica **8**, 497, 1960.

Crystal growth of silicon carbide (II)

H. B. Haanstra and W. F. Knippenberg

548.52:546.281'261

Some time ago [1] we discussed here crystal forms which appear in the initial stage of the growth of SiC crystals during the reaction between an intimate mixture of carbon and quartz powder at a temperature of 1500 °C in an argon atmosphere. The observations were made with an electron microscope. The crystals grow as "whiskers" and are in general very long. They are formed mainly on the carbon particles.

and there are no marked differences between individual whiskers.

When comparing the crystal forms produced by the two different reactions it occurred to us that the local thickening on the whiskers grown from the mixture of carbon and quartz powder might have something to do with the presence of globules of SiO_2. A simple experiment showed that this supposition is very probably right.

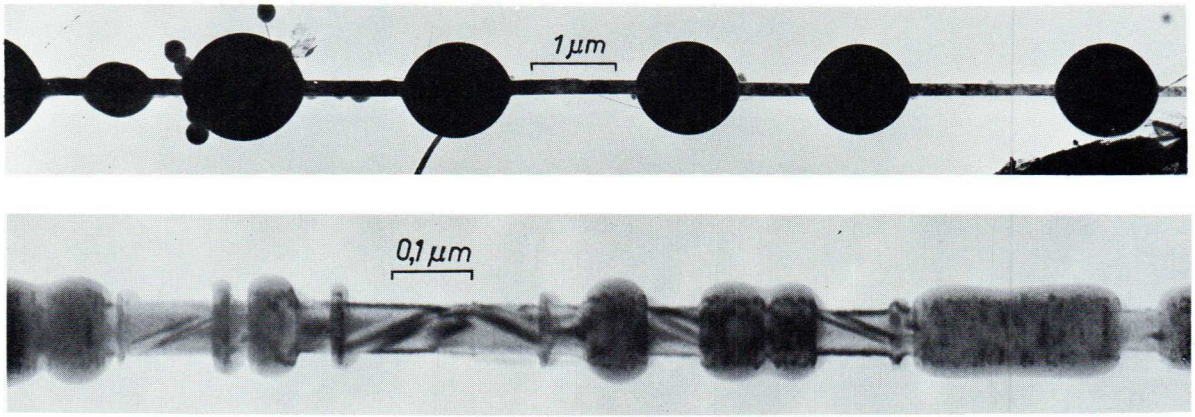


Fig. 1. Electron-photomicrographs of two types of whiskers grown in the reaction between carbon and quartz powder.

Some of these whiskers show local thickening, which may be globular, giving the crystal the appearance of a string of beads (*fig. 1a*), or may take the form of "sleeves", in which can be seen a fine structure of dark lines perpendicular to the long axis of the crystal; these lines do not continue right through to the surface of the sleeve (fig. 1*b*). The rounded outer part of the sleeve gives the impression of being amorphous. The micrograph does not show any internal structure of the beads, the latter being so thick as to be opaque to electrons of the energy used.

Recently we also found crystal growth in the form of whiskers in the reaction between silicon *vapour* and carbon at a temperature of about 1500 °C. Here again, very long and whiskery crystals grow in the first stage (*fig. 2*); the striking thing here, however, is that no local thickening is found on these whiskers. The crystals have a constant diameter over their whole length

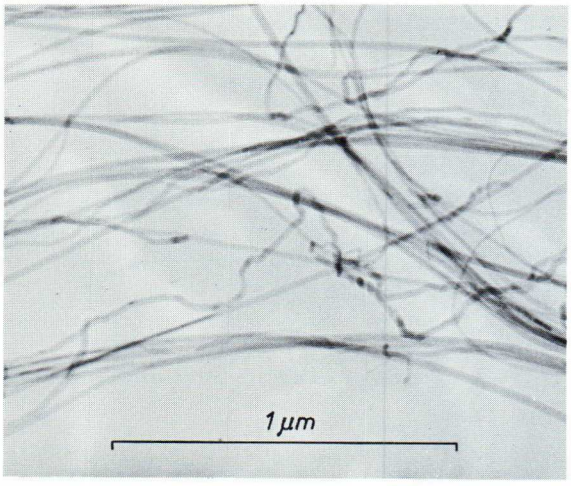


Fig. 2. Electron-photomicrograph of whiskers grown in the reaction between carbon powder and silicon vapour.

H. B. Haanstra and Dr. W. F. Knippenberg are research workers at Philips Research Laboratories, Eindhoven.

[1] W. F. Knippenberg, H. B. Haanstra and J. R. M. Dekkers, Crystal growth of silicon carbide, Philips tech. Rev. **24**, 181-183, 1962/63.

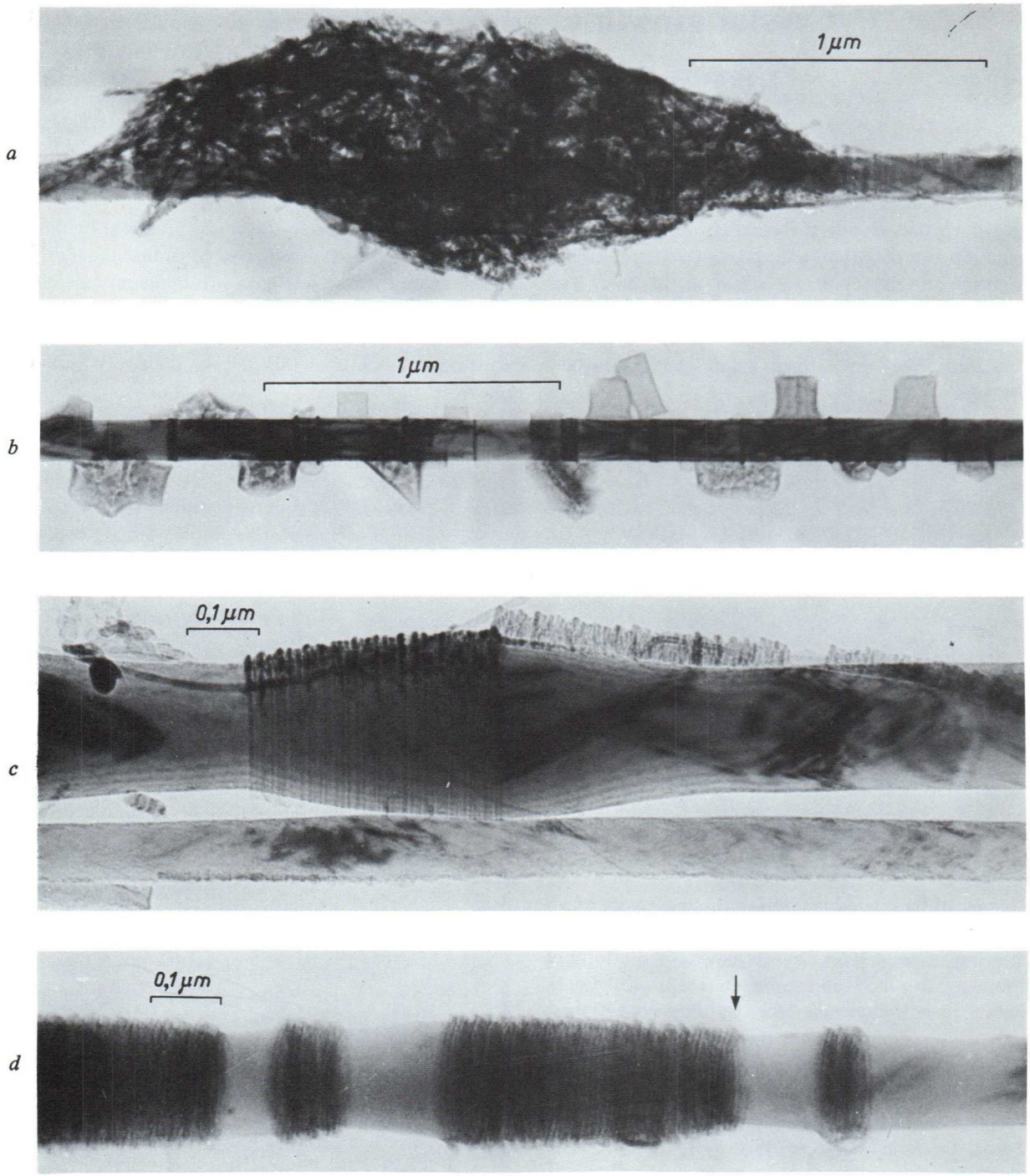


Fig. 3. Some whiskers obtained from carbon and quartz powder after exposure to the action of hydrofluoric acid.

In this experiment the whiskers were exposed for some time to the action of hydrofluoric acid vapour; the reaction of this acid with SiO_2 results in the formation of gaseous SiF_4, while it does not attack SiC. This treatment affects only the thickened parts of the whiskers. The beads have disappeared, and in their place is seen as a rule a tangled mass of whiskery crystallites around the continuing body of the whisker (*fig. 3a*); in some cases outgrowths in the form of platelets are found (fig. 3*b*). On the sleeves the amorphous-looking part disappears, while the cross-lines become much more distinct (fig. 3*c*). The experiment makes it seem likely that the parts which disappeared consisted of SiO_2. On a whisker whose long axis is oriented

obliquely to the direction of the incident electron beam, it can be seen that the bands are due to the presence of rings around the continuing body of the crystal (fig. 3*d*). At these places the crystal itself is usually slightly thickened. Sometimes one has the impression that the rings are hexagonal on the outside (the hexagon can be best seen at the part marked by the arrow in fig. 3*d*).

In view of these observations the following hypothesis to explain the growth seems reasonable. Small globules of SiO_2 on a growing SiC whisker react locally with carbon diffusing over the crystal surface [2]. The SiC thereby produced first forms rings around the whisker and then grows epitaxially on the crystal already present. This assumption is supported by the fact that, on whiskers not yet exposed to hydrofluoric acid, an amorphous layer is visible around each ring (see fig. 1*b*). It is not yet clear why the growth in the beads takes place in such a disordered manner; it may be bound up in some way with the size of the beads, the distance between the surface of the bead where the reaction takes place and the whisker being relatively large.

In many cases a globule was also found at the *end* of a whisker (*fig. 4*). This might be a confirmation of the hypothesis of Wagner and Ellis [3], who assume that the whiskers originate in a globule. Whether, however, *every* whisker has such a globule we were not able to ascertain.

According to another hypothesis, put forward in the article [1] already cited, the growth of a whisker might be the result of a lattice defect in the long axis of the crystal. In the material discussed at the time we did not observe any crystals that had such a defect. Very many more whiskers have now been investigated, and a few specimens have indeed been found in which a line is visible in the long axis of the crystal, possibly

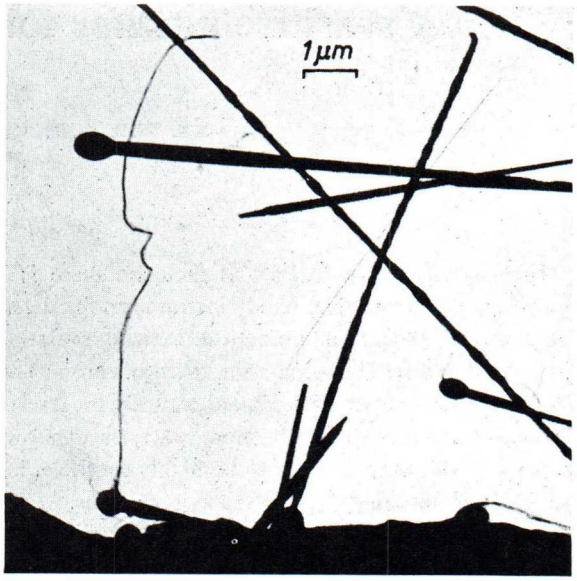
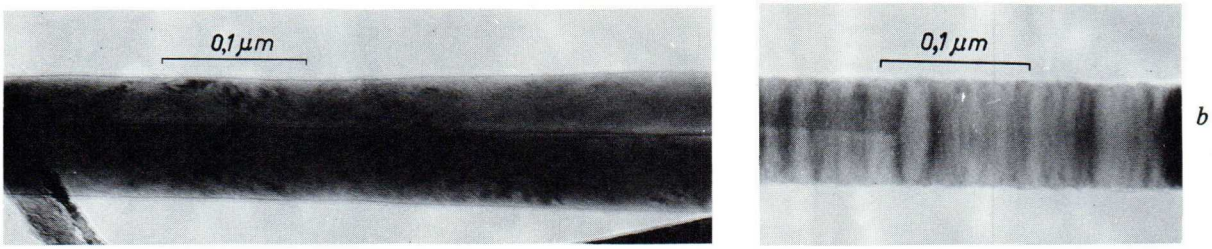


Fig. 4. Whiskers with globules at the end.

indicating the presence of an axial defect (*fig. 5a*). The remarkable thing is that in some cases these lines do not continue to the end of the crystal (fig. 5*b*). The number of whiskers showing a line of this kind is extremely small compared with those that seem to grow without an axial lattice defect.

The photomicrographs reproduced were selected from a large number made by Mrs. Gijsbers-Dekkers with the Philips electron microscopes EM 100-B and EM 200.

[2] W. F. Knippenberg, Growth phenomena in silicon carbide, thesis Leiden, 1963. This work has also been published in Philips Res. Repts. **18**, 161-274, 1963 (see page 235).
[3] R. S. Wagner and W. C. Ellis, Appl. Phys. Letters **4**, 89, 1964 (No. 5).



a *b*

Fig. 5. Electron-photomicrographs of two whiskers showing an axial lattice defect.

LIST OF SHORT ITEMS

page

Picturesque early transmitting aerial 74

Aerial for telemetering satellite-launcher 79

Development of an isochronous cyclotron 80

Manufacture of integrated circuits 93

Ancient incandescent lamps 124

Fizeau fringes observed with an interference mirror 144

First of a long series (of publications on ferrites) 156

Electron microscope photograph of ferroxdure 168

Research in stereophony 185

Bragg reflections in an electron photomicrograph 191

The Philips pavilion at the 1958 Brussels World Fair 210

Inhomogeneities in a doped germanium crystal 260

Etch pits on a zinc-sulphide crystal 276

Glass fracture surface 286

Industrial gas refrigerating machine 306

Thermionic-cathode testing 318

Chladni's figures on giant tuning-fork 343

Rotating mirror for examining lighting fittings 347

Electronic music 354

Coloured X-ray pictures 361

Fingerprinting via total internal reflection 374

Display of nuclear radiation spectra 397

Magnetic recording in a 5000 : 1 scale model 408

15 MeV linear electron accelerator during construction 418

Solving a chessboard puzzle with a computer 433

Tele-cardiotachometry 445

4 mm radar 453