Volume 38, 1978/79, No. 1

PHILIPS TECHNICAL REVIEW



PHILIPS

Philips Technical Review is published by the Research Laboratories of N.V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands, and is devoted to the investigations, processes and products of the laboratories and plants which form part of or cooperate with enterprises of the Philips group of companies. In the articles the associated technical problems are treated along with their physical or chemical background. The Review covers a wide range of subjects, each article being intended not only for the specialist in the subject but also for the non-specialist reader with a general technical or scientific training.

The Review appears in three editions: English, Dutch and German, all identical in contents. There are twelve numbers per volume, each of about 32 pages. A yearly index is given for each volume and an index covering ten volumes appears every five years (the latest one was included in Volume 35, 1975).

Editors:	Drs J. W. Miltenburg, Editor-in-chief
	Dr J. W. Broer
	Dipl.-Phys. R. Dockhorn
	Dr E. Fischmann
	Ir J. A. Klaassen
	Dr J. L. Sommerdijk
	Dr J. Ubbink
	W. van der Weel, assistant editor
English edition:	D. A. E. Roberts, B.Sc., M. Inst. P., A.I.L. (Redhill, Surrey)
German edition:	Dipl.-Phys. A. Hoyer (Hamburg)

© N.V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands, 1978.
Articles may be reproduced in whole or in part provided that the source 'Philips Technical Review' is mentioned in full; photographs and drawings for this purpose are available on request. The editors would appreciate a complimentary copy.

Philips tech. Rev. Vol. 38, 1978/79, No. 1

Contents

		Page
European Window		1
The inquiry		2
Pollution of the River Shannon in Limerick		6
Lichens		8
A hovercraft sprayer		11
Local functional equations		13
An instrument for determining the cross-section of a cave		14
Science and Education	Bernard Dixon	17
Science and Society	W. Martienssen	25
Science and World Problems	A. E. Pannenborg	33

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ x 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ x 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 17½ x 25 cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 x 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 x 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

Crystal defects in GaP

Experimental etching equipment

Frequency dividers for UHF

PTC effect

Amalgams for fluorescent lamps

Philips tech. Rev. Vol. 38, 1978/79, No. 2

Contents

Page

Influence of crystal defects on the luminescence of GaP . . . . . C. Werkhoven, C. van Opdorp and A. T. Vink 41

Experimental etching equipment . . . . . R. Kilian and M. Liehr 51

Frequency-dividers for ultra-high frequencies . . W. D. Kasperkovitz 54

Scientific publications 69

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)
Subscription rate per volume fl. 58.00

Bank F. van Lanschot
Eindhoven, The Netherlands

Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ x 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ x 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 x 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 x 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 x 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

PTC effect

Amalgams for fluorescent lamps

Phenol synthesis and photomorphogenesis

Grinding brittle materials

Real-time orthogonal transformation of CTV pictures

Ozone measurement in an aircraft

Soldering electronic components

Philips tech. Rev. Vol. 38, 1978/79, No. 3

Contents

Page

The PTC effect of barium titanate . . . . . J. Daniels, K. H. Härdtl and R. Wernicke 73

Amalgams for fluorescent lamps . . . . . J. Bloem, A. Bouwknegt and G. A. Wesselink 83

Phenol synthesis and photomorphogenesis . . . . . G. Engelsma 89

Scientific publications 101

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)
Subscription rate per volume fl. 58.00

Bank F. van Lanschot
Eindhoven, The Netherlands

Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ x 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ x 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 x 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 x 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 x 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

Grinding brittle materials

Real-time orthogonal transformation of CTV pictures

Ozone measurement in an aircraft

Soldering electronic components

Galvanic effects in etching

Hot splicing of optical fibres

Photoelectrochemical energy production

Philips tech. Rev. Vol. 38, 1978/79, No. 4/5

Contents

Page

Grinding brittle materials . . . . . A. Broese van Groenou and J. D. B. Veldkamp 105

Real-time orthogonal transformation of colour-television pictures . . . . . H. Bacchi and A. Moreau 119

Measurement of ozone in an aircraft . . . . . S. van Heusden and L. G. J. Mans 131

The thermal behaviour of electronic components during soldering . . . . . R. J. Klein Wassink 135

Scientific publications 143

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)
Subscription rate per volume fl. 58.00
Bank F. van Lanschot
Eindhoven, The Netherlands
Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ × 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ × 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 × 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 × 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 × 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

Galvanic effects in etching

Hot splicing of optical fibres

Photoelectrochemical energy production

The automatic digital multimeter PM 2517

Garnet films for fast magnetic bubbles

Plasma etching

Philips tech. Rev. Vol. 38, 1978/79, No. 6

Contents

Page

Galvanic effects in the wet-chemical etching of metal films . . . . . J.J. Kelly and G.J. Koel 149

Experimental semi-automatic machine for hot splicing glass fibres for optical communication . . . . . A.J.J. Franken, G.D. Khoe, J. Renkens and C.J.G. Verwer 158

Energy production by photoelectrochemical processes . R. Memming 160

Scientific publications 178

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)
Subscription rate per volume fl. 58.00
Bank F. van Lanschot
Eindhoven, The Netherlands
Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ × 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ × 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 × 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 × 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 × 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

The PM 2517 automatic digital multimeter

Garnet films for fast magnetic bubbles

Plasma etching

Transducer for medical echography

PHLIQA 1; questioning a data base in ordinary English

Photomultipliers with microchannel plates

High-pressure research with the PM 1100

Philips tech. Rev. Vol. 38, 1978/79, No. 7/8

Contents

Page

The PM 2517 automatic digital multimeter . . . . . M. P. van Alphen, R. E. J. van de Grift, J. M. Pieper and R. J. van de Plassche 181

A transducer for medical echography . . . . M. Auphan and G. Dale 195

Plasma etching in IC technology . . H. Kalter and E. P. G. T. van de Ven 200

Garnet films for fast magnetic bubbles . . . . . D. J. Breed, F. H. de Leeuw, W. T. Stacy and A. B. Voermans 211

Scientific publications 225

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)
Subscription rate per volume fl. 58.00
Bank F. van Lanschot
Eindhoven, The Netherlands
Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ × 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ × 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 × 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 × 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 × 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

PHLIQA 1; data-base consultation in natural English

Fast-response photomultiplier tubes with microchannel plate

High-pressure research with the PW 1100 diffractometer

The atom as a metallurgical building block

Philips tech. Rev. Vol. 38, 1978/79, No. 9

Contents

Page

PHLIQA 1, a question-answering system for data-base consultation in natural English, I. Organization and performance . . . . . PHLIQA Project Group 229

Two microchannel-plate photomultipliers with subnanosecond characteristics . . . . . J. P. Boutot and J. C. Delmotte 240

Apparatus based on Philips PW 1100 diffractometer for crystal-structure research at high pressures . . W. Denner and Heinz Schulz 246

Scientific publications 254

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)

Subscription rate per volume fl. 58.00

Bank F. van Lanschot
Eindhoven, The Netherlands

Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ × 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ × 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 × 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 × 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 × 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

PHLIQA 1, part II

The atom as a metallurgical building block

Picture processing (special issue, Vol. **38**, No. 11/12)

Philips tech. Rev. Vol. 38, 1978/79, No. 10

Contents

Page

The atom as a metallurgical building block . . . . . A. R. Miedema 257

PHLIQA 1, a question-answering system for data-base consultation in natural English, II. The artificial languages and translation operations . . . . . PHLIQA Project Group 269

Scientific publications 285

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)

Subscription rate per volume fl. 58.00

Bank F. van Lanschot
Eindhoven, The Netherlands

Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ × 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ × 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 × 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 × 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 × 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

Picture processing (special issue, Vol. **38**, No. 11/12)

Ion implantation in semiconductors

Correction plate for projection television

Electroradiography

Crystal defects in DKDP

Philips tech. Rev. Vol. 38, 1978/79, No. 11/12

Contents

Page

The challenge of picture processing . . . . . G. Groh 291

Digital image enhancement . . . . . A. Hoyer and M. Schlindwein 298

The Dot Scan CCTV, a flexible system for real-time image-processing experiments . . . . . L. H. Guildford 310

A universal instrument for digital picture processing . . G. J. Scholl 326

An experiment in flexible automation . . P. Saraga and J. A. Weaver 329

Flashing tomosynthesis . . . . . E. Klotz, R. Linde, U. Tiemens and H. Weiss 338

Computerized mammogram processing . . . . . A. Hoyer and W. Spiesberger 347

A system that can learn to recognize two-dimensional shapes . . . . . E. H. J. Persoon 356

Self-organizing systems . . . . . T. J. B. Swanenburg 364

PHILIPS TECHNICAL REVIEW
Philips Research Laboratories
EINDHOVEN (The Netherlands)
Subscription rate per volume fl. 58.00
Bank F. van Lanschot
Eindhoven, The Netherlands
Payment only after invoicing, please.

Printed in the Netherlands

OTHER PHILIPS PUBLICATIONS

Philips Journal of Research

A publication in English on the research work carried out in the various Philips laboratories. Published in annual volumes of six issues each of about 100 pages, size 15½ × 23½ cm.

Philips Telecommunication Review

A publication in English, dealing with the technical aspects of radio, television, radar, telephone and telegraph transmission and automatic exchanges. Published in volumes of four issues, about 40 pages per issue, size 20½ × 28½ cm.

Electronic Components and Applications

A publication in English, containing articles dealing with the theory and practice of electronic components and materials. Four issues per year, about 60 pages per issue, size 21 × 29½ cm.

Medicamundi

A publication in English on radiology, nuclear medicine and medical electronics. Three issues per volume, about 60 pages per issue, size 21 × 29½ cm.

Science and Industry

A publication in English, containing articles dealing with applications of systems for scientific and industrial analysis, measurement and control. A French translation, Science et Industrie, is also published. Three issues per year, about 36 pages per issue, size 21 × 29½ cm.

Forthcoming issues of Philips Technical Review will include articles on:

Ion implantation in semiconductors

Correction plate for projection television

Electroradiography

Crystal defects in DKDP

Magnetic properties of a microscopic particle

Computer calculation of magnetomechanical components

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79, No. 1

European Window

In Western Europe there are thousands of young people who, on their own initiative, identify a particular scientific problem and undertake independent research on it. They are fascinated by the many natural phenomena that are not yet fully understood or are still unknown. They investigate flora and fauna in their natural surroundings, going as far as into the high mountains. They measure the pollution of rivers. They scan the night sky in search of interesting phenomena. They do archaeological or geological excavations and make surprising discoveries. They also do mathematical, physical and chemical research.

The young scientists do their research as a school project or in their own time. In some cases they work on a topic as a team, in others alone. All are impelled by the desire to bring their work to a successful conclusion. This is evident from the contests for young scientists and inventors that are held every year in fifteen West European countries, where the contestants put their projects on show and have them judged by a jury of experts. The work displayed is often of a high scientific standard, and earns the admiration of the jury members.

Since 1968 Philips have organized an annual European contest, at which the participants are the winners of the national contests. In the first few years this European Philips Contest for Young Scientists and Inventors was held in Eindhoven, and subsequently in London, Aachen, Madrid and Paris. In the past ten years some 400 young scientists have taken part in this international event.

To mark the tenth anniversary of the European Philips Contest, which was held from 29 May to 3 June 1978 in the Evoluon at Eindhoven, a two-day symposium called 'European window' was organized. Young scientists and distinguished representatives of the world of science and education exchanged experiences and views at this symposium on three topics: Science and Education, Science and Society, and Science and World Problems.

Preparatory to the symposium all former participants in the European Contest were sent a questionnaire dealing with these three topics in the Autumn of 1977. Nearly 70% of the 'veterans' returned the questionnaire with their answers, often together with an extensive commentary. Partly on the basis of these answers three professional scientists presented their views on these topics: they were Dr Bernard Dixon, Editor-in-Chief of 'New Scientist' (Science and Education), Prof. W. Martienssen of the Physikalisches Institut der Goethe-Universität, Frankfurt am Main (Science and Society), and Dr A. E. Pannenborg, Vice-President of Philips (Science and World Problems). The participants were able to join in the discussions that followed the presentation of the results of the enquiry and the lectures.

The present issue of Philips Technical Review — on this occasion we are departing for once from the well-worn paths — is mainly devoted to the talks given on that occasion by Dr Dixon, Prof. Martienssen and Dr Pannenborg. They are prefaced by a brief discussion of the nature and results of the inquiry and by a brief account of the nature, judgement and contents of each of the five entries that won an award.



The inquiry

The questionnaire used for the inquiry comprised five to ten questions per topic, some of which were subdivided. The questions were based on a number of 'hypotheses', which are printed on the adjoining page.

Although recognized opinion-polling methods were used to draw up the questionnaire and for processing the results, this does not imply that conclusions may be drawn from the outcome about 'the opinion' of 'European youth'. The findings relate only to tendencies apparent in the views of a subgroup — the prize-winners — out of the group of young scientists.

Education

A small majority of the young scientists questioned (62%) said they were happy with the quality of their education. About one third were not. It also appeared that the young scientists living in Western Europe (Benelux, Great Britain and Ireland) were the most satisfied. Those from Southern Europe (France, Spain, Italy and Greece) were the least positive about their educational systems.

Young scientists have a preference for education in the physical sciences, which is not so surprising judging by their entries for the finals. To the question concerning the most ideal time-table for lessons, they replied: twelve hours for chemistry, physics and mathematics, three hours for biology, six hours for languages and four hours for social studies. On this matter there was hardly any difference between the nationalities.

In general it was felt that the knowledge acquired at school was not sufficient for the work of young scientists (66%). This criticism came especially from those living in Southern Europe. It was also stated that the school gave little encouragement to young scientists.

From their comments it is evident that young scientists set high standards for their schooling. 'There is no such thing as the ideal school. I would have liked to go to a school that offered me more opportunity for individual work and more freedom to develop my creativity', wrote one young scientist. Another added: 'I would have liked to do more practical work instead of all the theory I have had to wade through.' And another: 'My ideal is a free Waldorff school, where mainly the physical sciences and languages are taught'.

Other young scientists had more specific wishes with regard to the educational system. One of them wrote

The questions were formulated on the basis of the following working hypotheses

Science and Education

1. Young scientists have very positive attitudes towards their education.
2. The research activities of young scientists are stimulated from all sides.
3. Young scientists have a great interest in natural sciences during their education.

Science and Society

1. The young scientists consider that the scientist himself and no one else should choose this research activities.
2. The scientist is not responsible for the abuse of his scientific work.
3. Young scientists are seeking technical solutions to social problems.

Science and World problems

1. Young scientists pay more attention to problems in the developing countries than those in the Western industrialized countries.
2. Very little attention was paid to world problems at school.

that more attention should be paid to the technique of learning. 'As for example how I should use a library, and how to handle modern information storage systems'.

Young scientists' work demands a lot from young people. This appears from the fact that three quarters of them do little or nothing about other forms of spending leisure time, such as sport, music or watching television.



Social responsibility

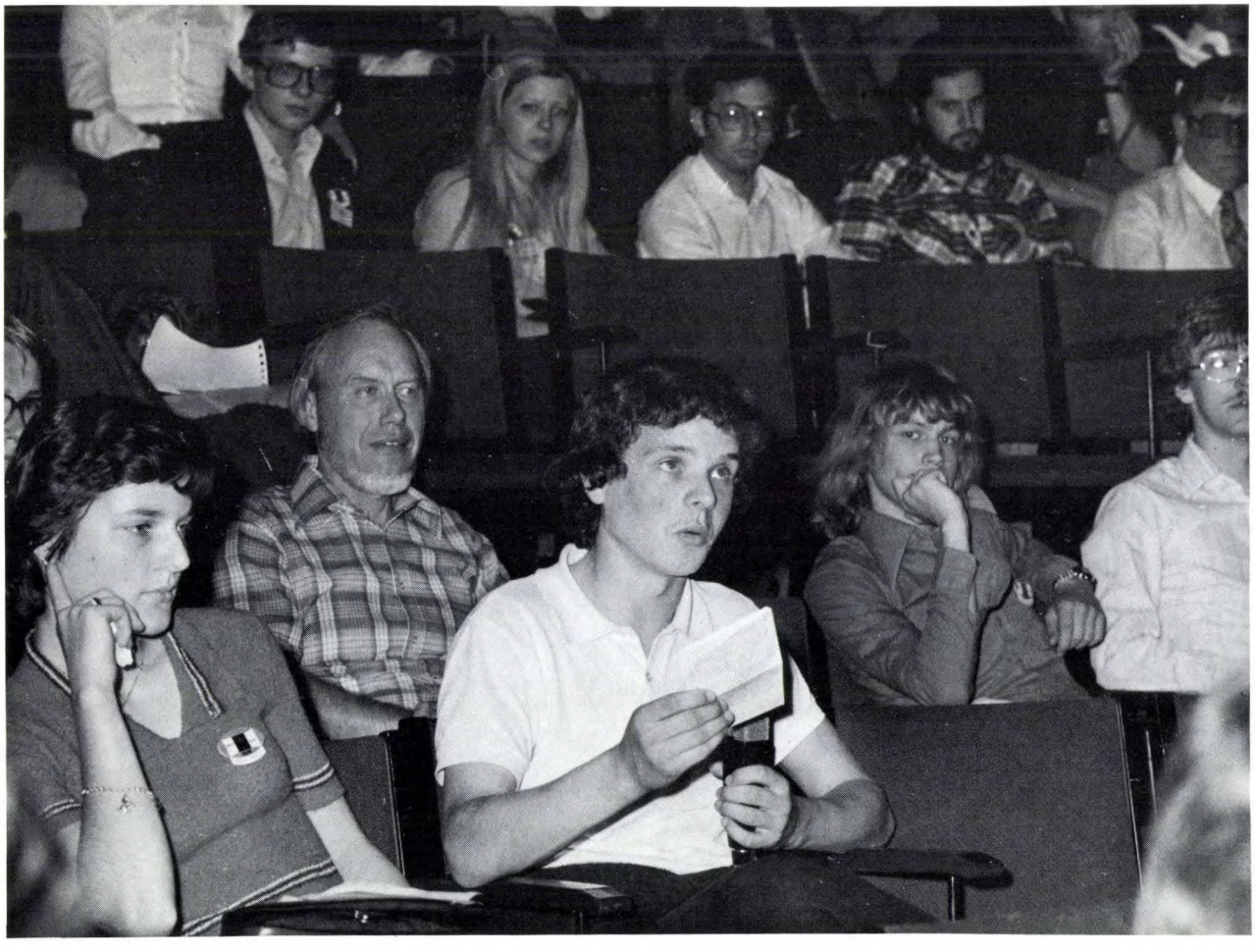
Is a scientist free in his choice of subject, and can he be held responsible for the results of his work? These were questions that provoked a great deal of comment from the young scientists. The group which considered that the scientist should be free to decide on his own field of research was almost as large as the group that was opposed to it (about 45%).

It was noticeable that young scientists who had earlier done fundamental work were prominent among those in favour of a free choice of research topic. 'If a scientist is not free to choose his field of research, he cannot put much effort and interest into the subjects he is required to study.' Another young scientist wrote: 'Only the man of science can best identify the problems that need to be studied.'

Another question related to the direction of scientific research. Half the young scientists questioned considered that scientific research should always be directed at finding solutions to social problems (38% disagreed with this view). In particular young scientists from Central Europe (West Germany, Austria and Switzerland) showed a preference for socially oriented topics of research.

More than three quarters of the young scientists felt themselves responsible for the results of their research. Some of them, however, added that it is difficult to determine in advance what use is going to be made of the results of scientific research. Most results can be used either for good or evil. 'It is a fallacy to believe that one can deliberately avoid a particular scientific field because of its negative consequences, and that such a policy would make the world safe.'

Some young scientists placed the responsibility for the use made of scientific results on society itself. 'Progress is not the work of one man but the property of us all, and therefore each one of us bears responsibility for it.' For this reason some of them considered that scientists do have a duty to be more open about their work. 'There is a an overwhelming need for proper communication between scientists and the rest of the community.'





World problems

The third set of questions related to the major world problems. Some young scientists emphasized in their comments that they considered the major world problems to be very closely interwoven. There was also a feeling that current world problems were only symptoms of even deeper ones.

In general the young scientists attached more importance to problems in the industrialized West than to those in the developing countries. There was also a distinct difference between those whose preference was for one of the two problem areas. Young scientists who were more interested in the problems of the West generally considered that not enough attention was paid at school to the major problems of the world.

In general the majority of the young scientists wished they had been taught more about these problems, even if it had been at the expense of a number of basic subjects at school. Others considered it was too late for that. The line they took was: 'Let us stop all this theorizing about world problems. The time has come to do something practical about them.'

Subgroups

There are marked individual differences between young scientists. As appeared from the inquiry, it is not easy to answer the question: 'What is a young scientist?' Nevertheless the computer succeeded in distilling from all the material submitted three subgroups of young scientists, each of which had the closest affinities.

Briefly, there is a group that is more socially oriented, a group that is oriented more towards the physical sciences, and a small group that shows a marked preference for the exact sciences. From all the comments written on their questionnaires, it is evident that young scientists feel a great sense of responsibility for the world. 'I feel myself responsible for the environment I leave behind,' wrote one of the young scientists, 'and at the same time I owe a debt to my fellow men, and also to those who will come after me.'

Or, as another young scientist put it: 'As an individual I am insignificant. That is why I call urgently upon all scientists to do everything in their power to save the Earth from heading towards its own destruction.'

Pollution of the River Shannon in Limerick

When excessive amounts of untreated waste water are discharged into a river, a situation may be reached when the natural purifying mechanisms are no longer able to cope; the river will then become polluted. Plant life and animal life both suffer the consequences. As a rule, pollution has the effect of diminishing both the number of individuals of a species and the actual number of species. To determine the degree of pollution a lengthy and fairly comprehensive programme of measurements is needed. By means of chemical analysis the content of nitrogen compounds and phosphates in the water is determined and also the concentration of various metals and anions, detergents, pesticides and oil. In addition the quantity of oxidizable substances and dissolved oxygen have to be determined. A bacteriological analysis is also necessary to ascertain the nature and amount of pathogens present.

Donald McDonnel (Limerick, Ireland) has undertaken independent research in this field, on which he reported in his entry entitled: 'The impact which discharge has on the biota of the Shannon River at Limerick'. This research won the special praise of the jury. McDonnel did not confine his work to chemical and bacteriological analyses, but also investigated, with success, methods of using the plant and animal life in the river as a means of determining the degree of pollution.

Chemical research

The river Shannon enters the town of Limerick in a fairly unpolluted state; in the town itself, however, so much waste water is discharged that the river is no longer able to cope with it. At some places there is very intensive algal growth, and mud banks are found that are distinctly anaerobic. To determine the seriousness of this pollution, analyses were made at regular time intervals of the river water during a period of five months. Every week during this period water samples were taken at six places along the approximately 3 mile long stretch of the river in the town. From these analyses, which were performed by standard methods, McDonnel determined the oxygen content, the biochemical oxygen demand (BOD), the carbon-dioxide content, the content of nitrates and phosphates, the amount of oil and grease, the concentrations of various metals, the pH and the quantities of suspended particles.

These analyses confirmed the initial impression that the water was seriously polluted. The oxygen content at many places is low and the BOD high; often only anaerobic forms of life are possible. The nitrate and phosphate concentrations are high, which can give rise sometimes to very strong growth of algae. In the summer this leads during the day to high local oxygen concentrations due to photosyn-

McDonnel in conversation with Mr H. A. C. van Riemsdijk, Chairman of the Supervisory Board of N.V. Philips.



thesis; during the night there is no oxygen production, and a shortage of oxygen results from the metabolism of the plants. In the winter the rotting of the algae reduces the oxygen content to an extremely low level, while the nitrates and phosphates released make the water unfit for animal life.

Together with the waste water various metals are also discharged into the river. Analysis showed that metals are concentrated in plants and sludge deposits. As a result some plants along the river can be harmful to the life of animals living on the banks.

The removal of the sewage discharged into the river has been delayed because a hydroelectric power station upstream of the town has reduced the outflow of the river. The discharged water may remain stagnant for a time between the mud banks at the side of the river, and reach the main stream only very slowly. It was also found that the discharge points are so close together that the river is not always able to recover sufficiently after each discharge.

Bacteriological research

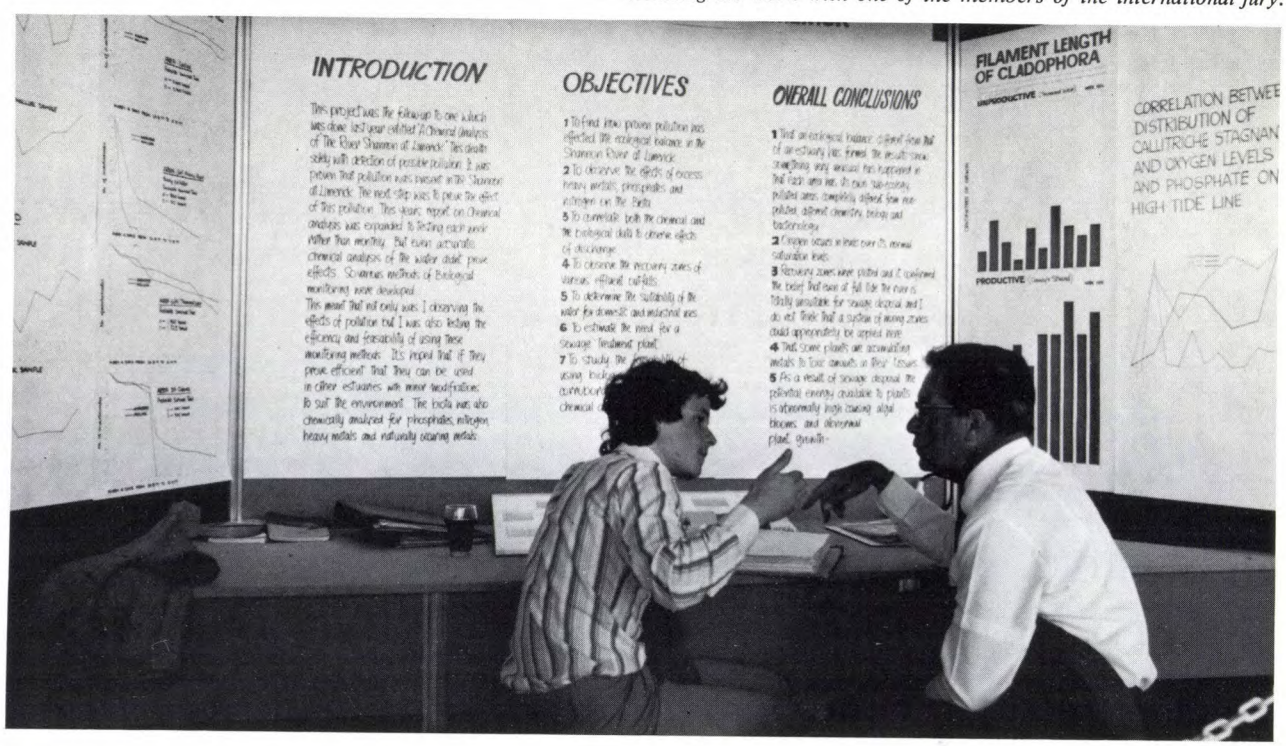
The bacteriological research revealed the presence of *Escherichia coli*. Very high concentrations were found, indicating the faecal origin of part of the pollution. It therefore seems likely that other pathogenic bacteria of faecal origin will also be present in the river water. From the amounts of *E. coli* found in freshwater shrimps and snails it was evident that the concentration of *coli* bacteria in these animals was connected with the concentration of the water. An increase of *E. coli* in the water was immediately followed by an increase in these animals. A decrease of *E. coli* in the water was followed by a decrease in the number of these bacteria found in shrimps and snails, though with some delay.

Conclusions from the chemical and bacteriological research are that the river is decidedly unsuitable for the discharge of large quantities of untreated waste water, and that the concentration of pollutants is so high that the river water downstream of the town is unfit for the production of drinking water for human or animals.

Biological methods

McDonnel looked for methods that could give some indication of the water pollution without the need to use the always tedious method of taking water samples for chemical analysis. He found that the growth rate of particular algae in the river was a good measure of the content of nitrates and phosphates. At various places along the river-bank he also immersed gauze cages containing freshwater shrimps and snails. He found that the chance of survival for these animals differed considerably from one place to another. The survival time for 50% of the numbers so exposed varied from 50 to 400 days, and was directly correlated with the degree of pollution. In this way he attempted by relatively simple means to gain an indication of the trend of the harmful pollution.

McDonnel discussing his work with one of the members of the international jury.



Lichens

Lichens are found in nearly all parts of the world where the temperature is not too high and the air is not dry for long unbroken periods. They are not 'self-supporting plants' but consist of algae and fungi growing in symbiotic association. Lichens are very interesting forms of life; for amateurs, however, they are difficult to study because the available literature is often highly specialized and rather dry. Sophie Valtat (Paris, France) has written a treatise on lichens, entitled 'Tentative d'approche méthodologique des critères biologiques pour une systematique des lichens'. The treatise, which earned the admiration of the jury, gives a very readable account of the structure of lichens and of their propagation. It ends with a description of a number of places where lichens are to be found in the mountains of the Greek mainland. The author names some species found there whose occurrence in Greece has not been recorded in the literature earlier. She also describes a lichen which contains a second, parasitic fungus.

Forms of lichens

Lichens occur in widely different forms. Often they are green or greyish crusts or small shrub-like growths adhering more or less firmly to the barks of trees, dead wood, stones or to the bare ground. They are also found as large or small patches on rocks, varying in colour from yellowish-green to brown or black. In all cases they are combinations of algae and fungi living in symbiotic association. The forms they take differ as a rule quite considerably from the separate fungi and algae. Various types are distinguished on the grounds of their appearance, such as foliaceous, crustose, fruticose, etc. The internal structure of lichens, which can only be seen under a microscope, may be just as widely different as the macroscopic form. In general, the alga and the fungus which together form a lichen can clearly be distinguished.

Metabolism

Most lichens obtain the water they need by storing rain water. The necessary nutrient salts are usually drawn from the substructure on which the lichens grow, in the same way as with mushrooms and toadstools. There are species, however, that live 'from the air' or from what they receive from rain water.

Below: the lichen Cladonia coccifera. Right: Cladonia pyxidata. Both lichens were found in the Vosges (France).





The fungoid contribution to the symbiosis consists of proteins and vitamin C, while the alga supplies sugars and various other vitamins. The relationship is not always completely symmetrical; there are cases in which one of the two partners is a parasite, although this is not always so clear.

Propagation

Propagation may be either vegetative or non-vegetative. In the simplest case vegetative propagation takes place by means of detached pieces of lichen that are carried by the wind or disseminated in some other way. A lichen may also grow special organs from which minuscule particles readily break off and are carried away by the wind to grow into complete lichens elsewhere. In non-vegetative propagation the fungoid partner of a lichen forms sporogenic organs. The spores have to come into the proximity of the appropriate algae in order for a lichen to grow from them.

Finds in Greece

Sophie Valtat gathered many species of lichens at four locations in the mountains of the Greek mainland. The places where she found them were mostly at a height of 1000 metres or more, and of the 20 to 30 species found at each location there were some that were obviously specific to a particular height above sea level. Most of the lichens, in keeping with the ground on which they grew, were more or less calciphile (calcicolous), i.e. they showed an affinity for calcareous matter, or limestone. On rocks containing only moderate amounts of calcium she found calciphile and calcifuge species growing side by side. Species with an affinity for nitrate were only found at places where animal life can supply the nitrates. Species that have to take up nitrates from the air were found only in the vicinity of tracks or roadways with occasional motor traffic; it is conjectured that the exhaust gases might supply the necessary nitrogen compounds. On a peninsula near Mount Athos a number of calcifuge species were found close to the sea. One of these species, which is not mentioned in the available literature, proved to be a host for a second, parasitic fungus.

Some meridional species were found in Greece that do not reach full development in the South of France.

A conclusion drawn from these field observations is that it is very important when classifying species of lichens to mention the characteristics of the terrain, such as type of rock, height, amount of sunlight, the presence of animal life, etc. Only in this way will it be possible in the long run to gain a better insight into the growth conditions of the various lichen species and associations.

The Chairman of the Jury handing the Award to Sophie Valtat.



A hovercraft sprayer

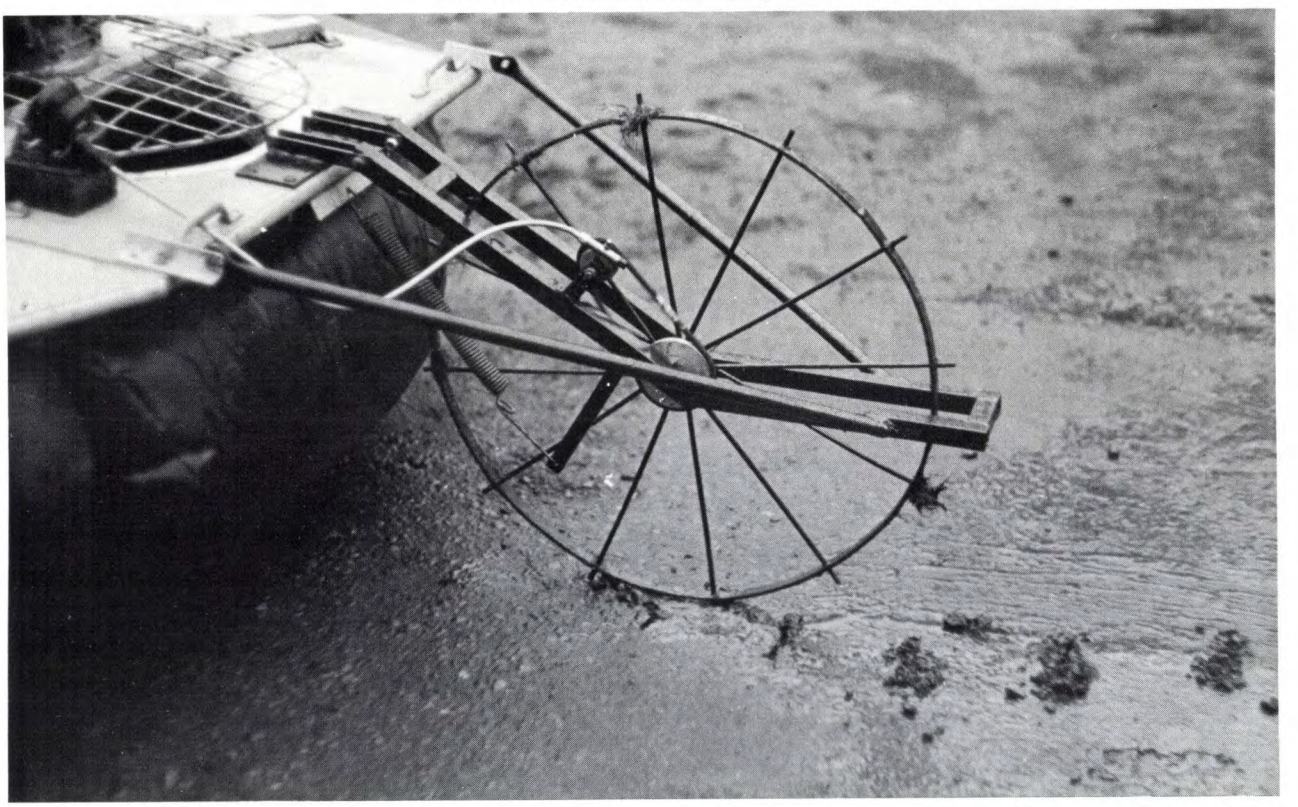
It may happen that crops have to be sprayed at a time when the ground is so wet that conventional sprayers, which are drawn by tractors, would cause too much damage to soil and crops. It occurred to Paul Brown and Alistair Wolf (Newcastle upon Tyne, England) that a small hovercraft, which they had built together with some fellow pupils as a practical exercise in mechanical engineering, would be much better suited for this purpose. The total weight of a hovercraft, a vehicle which can travel over land just as easily as over water, is distributed by an air cushion over a so much larger surface than in the case of vehicles on wheels that damage to soil or crops in wet periods would be ruled out. To be used as a sprayer the hovercraft, which was originally built to take part in races, had to be modified in some points. The modification involved a number of problems, and the jury was most impressed by the manner in which Brown and Wolf succeeded in solving them with the means they had at their disposal. The result of this work is described in their entry entitled 'The hovercraft sprayer'. In its first trials their hovercraft fully proved its usefulness as an agricultural spraying machine.

A hovercraft in its present form can only be used as a light transport vehicle and for racing. Its load-carrying capacity is relatively small. At first sight, therefore, a hovercraft would not seem to be so suitable as a spraying machine, because in order to spread pesticides homogeneously it is necessary to spray large quantities of dilute solutions. Moreover a hovercraft is not sufficiently manoeuvrable to be able to reach all corners of a field with the same ease. The great advantage of the hovercraft, however, is that, even in adverse conditions, it can still travel over crops without causing any appreciable damage.

Nowadays there are atomizing installations available that need much less liquid per hectare than conventional sprayers — about 10 litres of liquid per hectare against several hundred litres. The use of such atomizers enabled Brown and Wolf to overcome the difficulties presented by the load-carrying capacity of the hovercraft.

A hovercraft is normally steered with the aid of rudders that are fitted behind the propellors. Since there is no direct contact with the ground, however, it tends to drift sideways in the bends, like a sailing ship without a keel, and there is also the risk that it may spin around on its axis. The young scientists overcame this steering disability by providing the

Front wheel of the hovercraft sprayer showing the extended spokes that grip the ground.



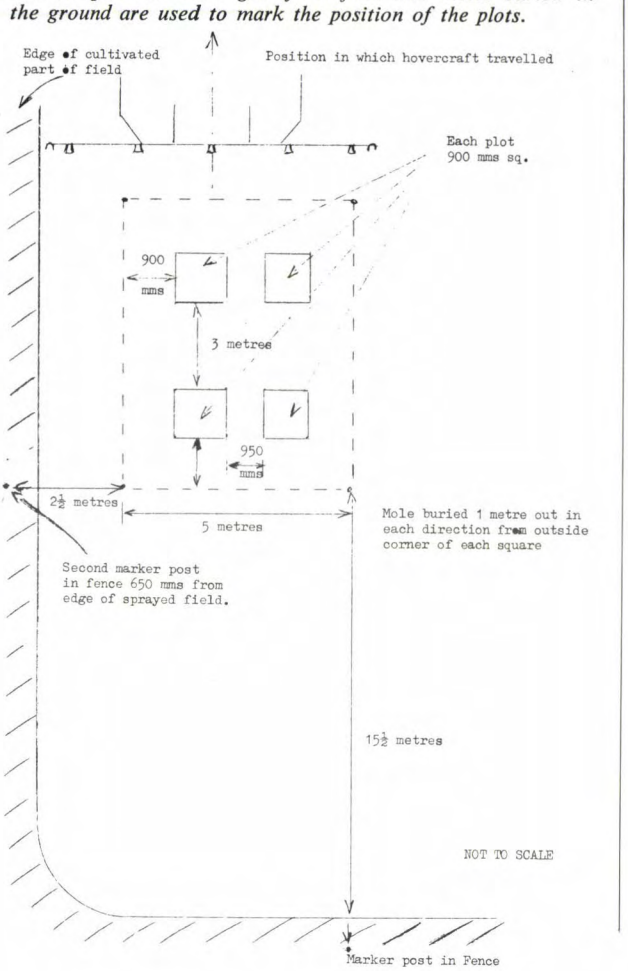
hovercraft with three light wheels with extended spokes which grip the ground and thus give the necessary stability. Two of these wheels are steerable. They are coupled with the rudders in such a way as to minimize the lateral forces on the wheels. There is one front wheel, which is provided with a tachometer so that the driver, by selecting speed, spray-fluid pressure and atomizer setting, can give the correct dosage of pesticide. All wheels are free to move vertically, and they are held on the ground by slight spring pressure.

Trials were set up to check the evenness of atomization. This was necessary because there was a possibility that air escaping from the air cushion could disturb the uniformity of atomization.

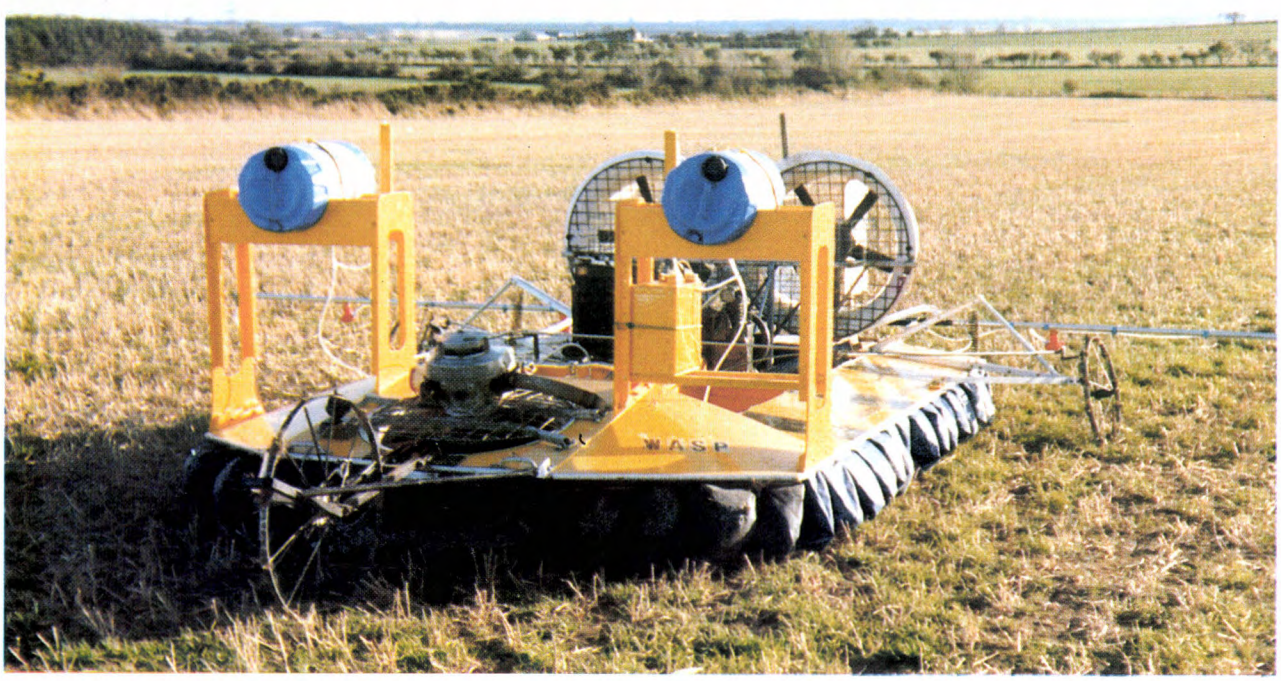
At a later stage, to ensure proper dosage of the atomized liquid, a peristaltic pump was used, coupled to the front wheel, so that the stream of liquid is directly affected by the forward speed. The front wheel was also fitted with a brake, to keep the hovercraft under better control on slopes.

Minor improvement of some features, together with various other planned improvements, mean that this hovercraft may be regarded as a prototype of a new sprayer, which has distinct advantages over conventional sprayers in difficult conditions.

Test plots laid out to check the evenness of atomization. Marker poles at the edges of the field and moles buried in the ground are used to mark the position of the plots.



The hovercraft sprayer.



Local functional equations

The entry sent in by Norbert Brunner (third-semester student of mathematics at Vienna University), entitled 'Lokale Funktionalgleichungen', treats a very specialized topic of mathematics. In such a short account as this it is therefore not possible to do justice to the contents. It is hoped, however, that the following exposé will give some idea of the nature of this entry which, in the jury's opinion, is a good contribution in this specialized field and certainly goes far beyond what one would normally expect from a third-semester student.

Norbert Brunner's entry is a study that may be counted as a contribution to topology in modern mathematical set theory. Very briefly, it is concerned with the word 'local'. In many areas of mathematics the properties of classes of functions, of spaces and the like are modified by formulating them as 'local'. A simple example of this usage is the term 'locally bounded'. A given function f for all real values of x is said to be 'bounded' if there is a number M such that $-M \leqslant f(x) \leqslant M$ for all x. The function is said to be 'locally bounded' if for every point a an interval I can be found that contains a and on which f is bounded. Thus, for example, all continuous functions are locally bounded.

In this example of the term 'bounded' the scope of the concept is extended by local formulation. There are many more functions that are *locally* bounded than are bounded without any further qualification. The converse is the case of 'surjectivity', to which the young scientist has devoted a great deal of attention; here the scope of the concept is on the contrary diminished by local formulation. A function f is said to be surjective when it can take any value, that is to say when for any number y there exists an x with $f(x) = y$. Surjectivity without any limitation is a well-nigh trivial concept: for example, the function $f(x) = x$ is surjective, and many other instances could be mentioned.

Remarkably enough there are also locally surjective functions: these have the properties that for any interval I and for any number y an x is to be found in I with $f(x) = y$. Examples of local-surjective functions are fairly complicated and to a non-mathematician seem rather far-fetched.

Brunner makes a systematic study of the local formulation of surjectivity and of various other mathematical concepts. He deals with examples from and connections with a large number of modern mathematical theories. It is not of course possible to review these here, because any treatment would have to be prefaced by a long and intricate introduction to the many mathematical concepts which Brunner uses and evidently masters so well.

The Contest Director with some of the young scientists; first from left: Paul Brown, third from left: Norbert Brunner.



An instrument for determining the cross-section of a cave

In noting the results of spelaeological research, one of the things that has to be recorded is the cross-section of the cave. Accurate determination of cave cross-sections is no simple matter and requires a great deal of measurement work.

A chance finding when photographing the interior of a cave prompted the brothers Clemens and Martin Trüssel (Rheinfelden, Switzerland) to embark upon a closer investigation, which finally led to a photographic method for measuring the cross-section of caves. They report on this method in their entry 'Fotographische Höhlenquerschnittsvermessung'.

In working out the method, so we find stated in the jury's report, the brothers made such a meticulous analysis of their method of measurement and of the errors it involves that a standard method for speleology could certainly evolve from it.

Speleological surveying

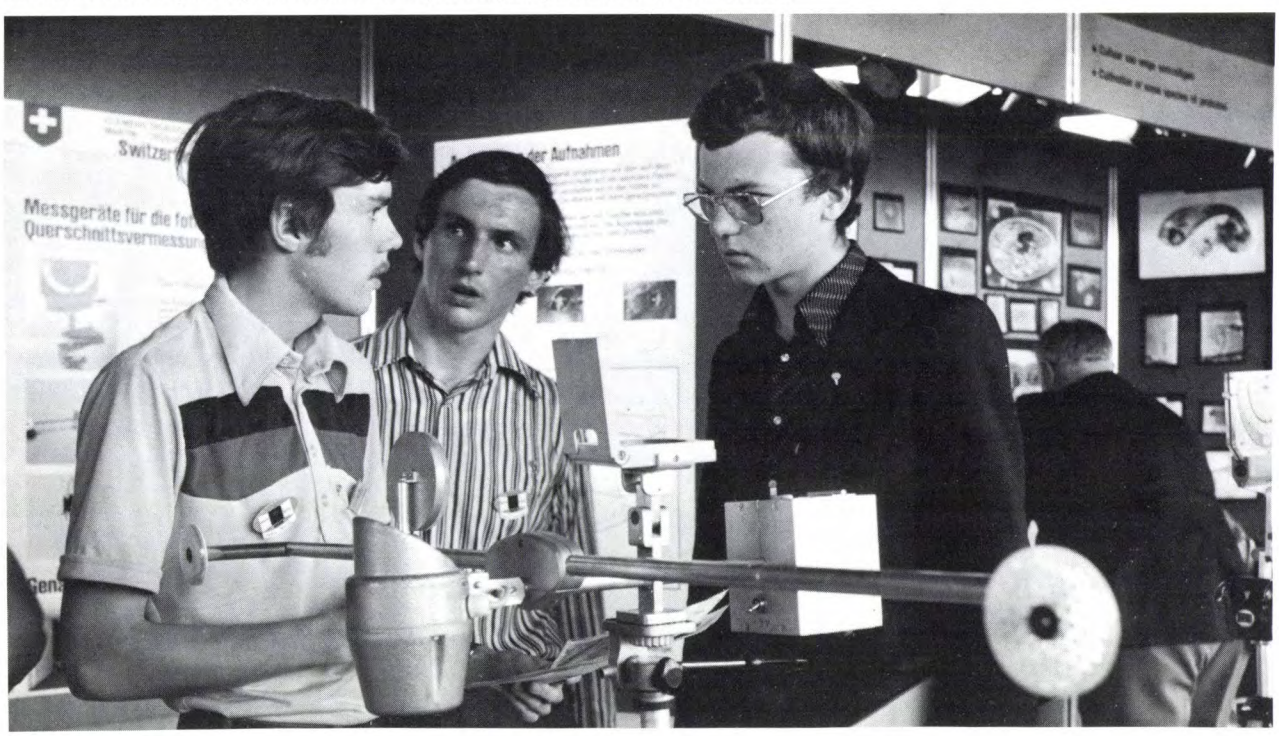
Apart from the polygon, the line that closely approximates to the centre-line of a cave, the cross-sectional profiles can also contain important information on the history of a cave's origin. Both of them, polygon and cross-sections, should therefore be recorded so that the results of such a survey can serve as the basis for further studies. For the polygon survey the young scientists submitted a theodolite, combined with a compass and a clinometer.

By dint of careful construction the brothers were able to produce an instrument capable of sufficiently high accuracy. The accuracy of the instrument was determined by measuring the position of a number of points in the free field and comparing the results with the directions and elevation differences on a topographical map.

Determining cross-sections is a much more complicated problem. The element of chance and the alertness needed to perceive the usefulness of coincidences both played a role here.

When the beam from a pocket torch is moved around in a dark cave, the places where the beam strikes the wall will be recorded as a bright line on a photographic film in a camera, whose shutter is left open during this movement. This was first observed quite fortuitously; further work on the ob-

Martin Trüssel explaining his instrument to fellow young scientists at the Contest.



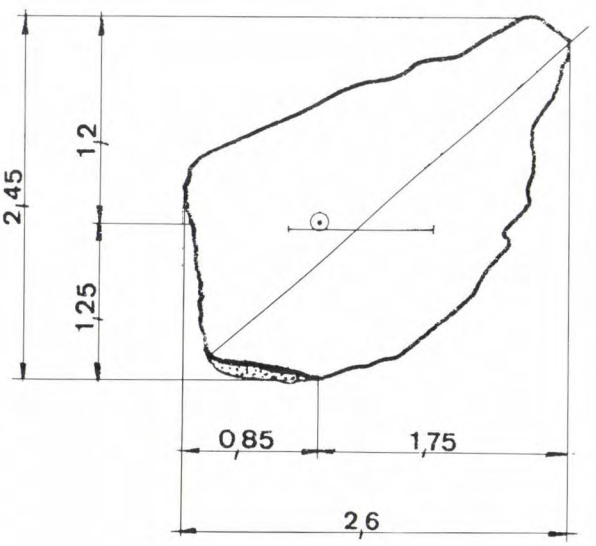
servation led to an instrument with which cross-sections of caves can be quantitatively recorded.

A small spotlight is fixed to a horizontal axis in such a way that the light beam describes a vertical plane when the axis rotates. A camera with open shutter placed some distance away records the intersection of this plane and the cave wall, giving the required cross-section.

The stands for the spotlight and the camera both carry a geological compass, so that it is possible to locate the film in the camera parallel to the plane described by the light beam. Also fixed to the spotlight stand are two reflectors, at a distance of one metre apart. After a cross-section has been recorded. both reflectors are illuminated by means of a flashbulb, thus recording on the film a measure of distance, and at the same time making details of the cave wall visible on the photograph.

In very high or very wide subterranean spaces it can be a drawback that the camera has to be situated rather a long way from the spotlight stand if the whole cross-section is to fall within the aperture angle of the objective lens of the camera. A fish-eye lens may provide a solution here.

To ascertain the accuracy of the instrument the young scientists took photographs in a room. The dimensions of the room as determined from the photographs were compared with direct measurements. The greatest deviations proved to be approximately 3%.



Below: a cave cross-section as photographically recorded; above: numerical data obtained from the photograph.





Dr Bernard Dixon, Editor-in-chief of 'New Scientist'.

Science and Education

Bernard Dixon

I want to begin by thanking Philips, not only for their invitation to speak at this event and for their generous hospitality, but also for their enterprise in sponsoring the European Contest for Young Scientists and Inventors. This annual competition, now celebrating its 10th anniversary, has become a major force in European science, and has helped to stimulate interest among young scientists and future scientists in the wider implications of their work. I wish it well, and trust that the competition will go from strength to strength.

It is, of course, the broad social and political relations of science that have brought us here today. I have been asked to speak about one aspect of the subject: science and education. I propose to do so in three separate sections. First, I shall look at the place of science and scientific ideas in education generally. Then I shall consider specifically the education and training of scientists. Thirdly, I will say a few words about the relationship between scientists and technologists on the one hand, and the media — the press, television and radio. Although science journalism is in no sense a formal part of education, it can genuinely be considered part of what we in Britan call 'further education'. It is, of course, my own particular interest and the organisers have asked me to include the subject in my talk.

* *
*

Let us begin, with the education we *all* receive — with teaching at school which acquaints us with the history, geography and culture of the world in which we live. Nearly two decades ago, the distinguished English novelist C. P. Snow (now Lord Snow) caused a national furore by delivering a lecture, the Rede lecture, in which he defined what he called 'the two cultures'. In doing so, he categorized (and some believe encouraged) the tendency to see the domain of the scientist and artist as distinctly separate. He spoke of 'literary intellectuals at one pole — at the other scientists, and as the most representative, the physical scientists. Between the two a gulf of mutual incomprehension, sometimes (particularly among the young) hostility and dislike, but most of all lack of understanding.' Snow's analysis was certainly exaggerated, as one might expect from a novelist. But from the strength of reaction it provoked, if from nothing else, we must conclude that it contained an unpalatable truth. The part of his lecture which interested me especially, and the one I want to highlight today, is that in which he cited as an example of cultural poverty the person who could not state the second law of thermodynamics.

'A good many times', Lord Snow said, 'I have been present at gatherings of people who, by the standards of the traditional culture, are thought highly educated and who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the second law of thermodynamics. The response was cold: it was also negative. Yet I was asking something which is about the scientific equivalent of: Have you read a work of Shakespeare's?'

I believe that Lord Snow was right in this comment, which might easily be seen as a prime example of the arrogance of the scientist. Indeed, one could go further. The significance of the second law is so pervasive, far-reaching, and fundamental to our understanding of the natural and physical world, that an appreciation of its general sense is surely as important as a knowledge of the work of the world's great creative artists — Tolstoy, Mozart, Rembrandt.

A word of warning. I do not mean to argue that we should all, scientists and non-scientists, butchers, accountants and oboe players, be capable of sitting down and producing a rigorous mathematical treatment of the second law of thermodynamics. That would be absurd, the equivalent of expecting similar detailed knowledge of the score of Beethoven's ninth symphony. What *is* important is that we should all have a grasp of the fundamental concept, the idea, that is enshrined in the second law. And here we come to the role of education. Whatever else a general education may provide in factual terms about the world, I believe that it is failing if it does not convey a sound and persistent awareness of the major principles, laws, and concepts of science. The second law of thermodynamics is just one example.

Another is that of probability. Let us consider for a moment the significance of this idea in the modern world. We live at a time of public concern about the hazards associated with man-made chemicals in the environment, at a time when more and more of us are concerned about our chances of avoiding coronary

heart disease — whether by curbing our consumption of animal fats or by gentle morning jogging. None of these issues can be even remotely understood without an appreciation of the concept of probability. Yet time and time again, we hear them being discussed, in bars, on television, and in parliament, in the most mindless way by people who manifestly cannot think in probabilistic terms.

Take another example. The concept of causality is the key to our understanding of a whole range of different diseases. The best example is lung cancer. Two or three decades ago it became strikingly clear that there was a relationship between heavy smoking and this form of malignant disease. People who smoked a large number of cigarettes appeared more likely to contract lung cancer than those who smoked very little or not at all. The common sense conclusion was that smoking did indeed cause the condition. But at that stage, at least, the evidence was far from complete. All we knew was that there was a link or association between two phenomena: tobacco smoking and carcinoma of the lung. Although health propagandists tried to use the relevant figures to dissuade people from the habit, there were other experts who were only too anxious to point out that the relationship may not be a causal one. It could be, for example, that *A* (heavy smoking) did not cause *B* (lung cancer), but that both *A* and *B* were the result of a third factor, *C*. Perhaps people with a propensity to develop this particular disease are also likely to become smokers, both phenomena being determined by heredity.

It took many years of diligent research, including studies on different populations, on the same population over period of time when smoking habits were changing, and on the altered prevalence of lung cancer among groups of smokers who gave up the habit compared with those who did not, before we came as near to the truth as we are today. What I would like to emphasize is that the essential tools for producing our current understanding include ideas such as probability and causality. These same ideas are central to any sensible discussion of climatic change, of the relationship between intelligence and heredity, and of the significance of changes in the ozone layer to life on Earth. They are beguilingly simple notions, habits of thought to scientists, but often appallingly neglected in other spheres of life. It is ideas of this sort, rather than catalogues of facts and figures, which should come from science education at school. Yet to this day, many people are able to leave school without being introduced to this conceptual approach to problems and evidence.

Some years ago, Dr Michael Bassey carried out a survey among arts undergraduates at University College London. He circulated questionnaires to some 200 people, all of whom had passed either chemistry, physics with chemistry, or general science at GCE '0' level, but not studied science subsequently. He was trying to find out how much of what they had been taught the arts undergraduates had learned and remembered. Dr Bassey received 115 replies, and the following are some examples of the comments they contained.

'I didn't like school chemistry because, at the time, it seemed to consist of a number of unrelated facts with a very limited relevance to everyday life, or indeed to anything but the general science text book.'

'I am vague about atoms and electrons. No such questions, even in the simplest form, were dealt with in the chemistry or physics I did at school.'

'There was little scope for straying outside the syllabus, which is somewhat antiquated. This I assume to be because our text books followed the same line as the examination. Looking back, chemistry seems to me to have been much more of a game with beakers and bunsen burners than a utilitarian proposition. Atomic structure was simply out for non-specialists. I have learned far more since from two or three television programmes.'

'The chemistry presented to us was purely for examination purposes: there was little attempt to stimulate any interest outside the immediate syllabus. Taking general science we were considered as poor relations, and so possibly not worth worrying about.'

'The only impression I retain is how much I disliked school chemistry. If they had taught me in what direct ways chemistry impinges upon the daily fortune of living, I might be better disposed.'

'I preferred physics to chemistry as it seemed to me to be much more practical and applicable to daily life, e.g. the working of batteries, electric bells, etc. as compared with a mixing of various chemicals in test tubes, which, although entertaining, seemed completely unnecessary to anyone but a professional chemist.'

'School chemistry was generally rather dull and seemed to have very little relation to everyday life.'

Dr Bassey used his findings to support the point made in this last example — that school science is of little help in illuminating our understanding of the world. He cited DDT and aspirin as important substances, both of which have had a major effect no human welfare, but about which many of his respondents were totally ignorant. I take the point. Knowledge of these chemicals and how they are used is certainly more significant for most of us than the scholastic junk which we still carry around from our school chemistry lessons — details of how to prepare a pure, dry sample of chlorine gas, for example. But even Dr Bassey has got the emphasis slightly wrong. Even he

seems largely concerned with the facts, rather than ideas, of science. His nearest approach to a criticism of school teaching for its failure to convey the conceptual framework of science is in a question put to his subjects about which of four materials — water, clay, air and wood — contained atoms. One in five could not answer correctly. The undergraduates concerned had studied chemistry at school for three, four, or five years. They had passed an examination in the subject. They were sufficiently intelligent to enter a university. Yet one in five had failed to appreciate the most fundamental discovery of chemical science — that all matter is atomic.

I do not believe that science education has changed radically since Michael Bassey carried out his survey. Most European countries have seen moves in recent years towards a more 'open-minded', investigative approach, in contrast to the rote learning of data required to answer exam questions. In Britain, this approach has been promoted in particular by the Nuffield Foundation. But much science teaching today still apparently fails to convey those precious ideas that form the real guts of science and scientific research. It certainly does not do so with any persuasive potency. How otherwise can we explain repeated failures — during public arguments over the safety of whooping cough vaccine or the hazards of nuclear power, for example — to appreciate elementary patterns of scientific thinking? Of course, we would all like to have black-or-white answers to questions of this sort. But where they are not possible, the meticulous, cautious scientific method gives us the nearest achievable approach to what we can call 'the truth'. It is a tragedy that the unique power of this method is so frequently unrecognized, and that anecdote, hunch, or prejudice are preferred in reaching an opinion. The underlying failing here, I believe, comes from conceptually inadequate, fact-dominated teaching about science.

I shall argue later that this state of affairs obliges both scientists and science writers to, in a sense, compensate for the manifest failings of formal education. We should, I believe, strive to proselytize on behalf of the simple elegance of the scientific method.

* *
*

For the moment, though, let me turn to my second subject: the education of scientists. Here I want to argue that training in science, whatever its excellence, is seriously deficient if it presents science in isolation from the real world. By the same token, it is strong where it does portray science in its social and international context. It is particularly interesting to find that well over half of the respondents to the Young Scientists Inquiry, sponsored by Philips, said they felt that too little attention was paid at school to important world problems. Almost 70 per cent, moreover, would have liked to have been told more about this kind of problem, even if this had been done at the cost of 'basic subjects'. I am convinced that this narrow focus carries over, and indeed is often intensified, during higher education at college and university. Despite the debate and clamour that has surrounded science in recent years, it is still all too easy to take a first degree or equivalent qualification in, say, chemistry, physics, botany, without being made to concentrate one's mind on the social and political ramifications of the subject.

Of course, many technologists receive a somewhat broader training than that of pure research scientists. Civil engineers, for example, learn economics and management techniques, because these are essential for their future work. Building a new motorway is not simply an engineering exercise. Such a project is intrinsically complex, requiring skills other than technical ones and has to be planned with sensitive regard for its social consequences and acceptability. But even such self-interested breadth of training as engineers receive is lacking in the education of many of tomorrow's scientists. Yet again, an increasing proportion of the most far-reaching developments nowadays are initiated and brought to practical realization by scientists, not technologists.

An excellent example of this is to be found in the newly emerging field of genetic manipulation. Based on the use of what have been termed recombinant DNA techniques, this new science will bring together hereditary material from disparate creatures, such as animals and bacteria, thereby creating novel organisms that may not have existed previously in nature. It is being conducted with such laudable aims as facilitating the industrial production of insulin, incorporating nitrogen-fixing genes into crop plants, and increasing our understanding of cellular metabolism (both normal and abnormal, as in cancer). We could, therefore, aptly describe it as applied science. But the molecular biologists who are pursuing this seminal line of inquiry are for the most part pure scientists. They are pure scientists whose discoveries are likely to spawn practical applications very quickly. I say this mindful of the growing constraints and controls surrounding their activities which so irritate many biomedical companies — but mindful of the enormous practical repercussions which are expected to follow from the new discipline.

I would not for one moment argue that exponents of genetic manipulation are behaving irresponsibly. Indeed, a small number of them have achieved historical fame by going out of their way to draw attention to potential dangers in their new field of research. They did so notably during a conference held at Pacific Grove,

California, in February 1975 which has since led to guidelines and regulations in several countries for the conduct of recombinant DNA work. Nonetheless, the fact remains that the 'whistle blowers' in this area have formed a numerically tiny proportion of people involved in this research, most of whom have bitterly resented public exposure and criticism of their activities. Distinguished medical microbiologists have even argued that a major failing has been the narrowness of the training received by this new brand of specialist. Most, if not all, of the potential hazards of genetic manipulation, they argue, could be contained simply by adopting the practical techniques used routinely in medical laboratories. But molecular biologists tend to treat their materials as chemicals, rather than living organisms. A more telling illustration of the folly of narrow specialization in science education would be more difficult to imagine.

For this sort of reason I find very striking the replies to another of the questions posed in the Young Scientists Inquiry. Those who characterized their research projects as 'pure' were in general less satisfied with the quality of the education they received than were researchers with projects in the 'applied' area. Assuming that this reflects a dissatisfaction with a blinkered stance in education, this is surely a very heartening sign. And of course those who have been able to recognize this deficiency have clearly gone at least part of the way towards overcoming it.

Before turning to the question of science in the media, I would like to explore one other aspect of science education, which is directly related to the conduct of scientific research — the need for flexibility. A number of factors, ranging from the sheer pace of research in some sectors of science to the practical effects of technology on patterns of employment, are making it obligatory that we reconsider the traditional pattern by which scientists are educated. That pattern is, of course, based upon the idea that an individual takes a first degree, perhaps followed by a higher degree, which provides the basic specialization upon which he or she can build an entire career in science. This is the system which, in the past, has led to the great scientific teams, their researches revolving around one or two talented individuals who have made a life's work of developing a particular speciality. Is this pattern likely to continue? I was very struck recently to read the following comments in '*Nature*' by a distinguished British physicist, Professor John Polkinghorne, who has decided to leave science and take up a totally different career in the Church:

'I've had a good run for my money for 25 years. I've played a small but useful role. But the subject is continually changing and enriching itself. I admire the way my younger colleagues can move into this sort of game. Partly through time, but more through a question of mental flexibility, I don't think I can make that sort of move. I could therefore tick over for a few more years, but I don't think that would be tremendously useful use of my time. I've done most of what I can do. If I felt I was going to stick in it and make reasonably useful contributions I would probably stay with it, and regard it as being a vocation to do so. When you're young you like to learn new things — that's part of the fun — but if I'm honest it gets slightly wearisome as you get older, particularly in a subject like particle physics where genuinely new things do come along at quite a surprising rate. Most people in my subject do their best work early, and they tend to do the same tricks later on. There are some subjects in which these tricks remain very important, subjects whose basis is well understood, but elementary particle physics, relativistic quantum mechanics, is continually changing. There's less room for the wise old man. I think his wisdom is limited.'

Aside from Dr Polkinghorne's positive reasons for entering the church, I find this an unusually honest statement from someone who, at the age of 48, has decided that he is no longer able to contribute to his own specialist field of science with the same originality that was possible 20 or even 10 years ago. I certainly cannot name names, but it is not too difficult to think of other scientists — and indeed whole research departments — continuing under their own momentum along a particular research avenue long after it has been effectively exhausted. This state of affairs, which in times of financial hardship is becoming less tolerable, has obvious implications for careers in science. It means, at the outset, that science education should seek to engender flexibility of mind rather than rigid specialization, so that people can change direction relatively painlessly later in life, and even be invigorated by the process. Secondly, it means that our educational machinery should be such that retraining and redeployment are commoner and easier to achieve than in the past.

One idea which has been mooted from time to time in Britain is that we should all, when we normally finish secondary schooling, receive a certain number of 'tickets', each representing, say, a year's higher education, that can be taken at any time. One would therefore have a choice between on the one hand 'cashing in' all of the tickets at once — taking perhaps two substantial degree courses — or on the other hand training to become an ecologist or botanist with the firm intention of retraining in middle age to do something entirely different. The idea as expressed in this form may appear fanciful, but I believe that it is in this



The jury during the announcement of the result of the European Contest.

general direction that education (and not only, of course, in science) will have to move in future.

Some years ago, when the present director was appointed to the Imperial Cancer Research Fund Laboratories in London, he made a particular point of saying that he intended to exercise an unusual degree of flexibility in organising the institute's research work. Departments would not necessarily continue to exist simply because they had always existed. Instead, the research and technical staff would find themselves changing their research goals and their groupings within the institute as the pattern of science itself evolved. Rather than have the institute's research programme shaped solely by tradition (and even by the physical structure of the building) it would be allowed to alter as and when necessary. In my experience, such ruthless but productive flexibility is rather commoner in industrial laboratories, with their commercial imperatives, than in the academic world. More's the pity.

The problem, of course, is that it is very much easier to demolish and repartition a research laboratory than to persuade a skilled researcher that he or she should abandon a line of inquiry upon which a substantial reputation has been built. That is why I have raised the subject under the heading of 'education', because if such adaptability is to be achieved, this will depend upon corresponding adjustments in the pattern of education. It has almost become a cliché to argue that a degree in, say, electrical engineering or biochemistry, followed by three years of research training, should not be seen as a 'meal ticket' for a lifetime of work in that particular specialism. Yet (human nature aside) it can still be extraordinarily difficult for someone who wishes or needs to change direction in later life, to actually do so. At the moment, the appropriate machinery to help such individuals is not available. But moves are now afoot in most European countries to improve prospects of this sort. The extent to which they are developed more fully does, of course, depend in part upon the degree to which people take advantage of those facilities.

* *
*

The last section of my talk I want to devote to that often thorny subject of science in the media. Earlier, I quoted one of Dr Michael Bassey's respondents who claimed to have learned more from two or three television programmes than from chemistry at school. One of my predecessors as editor of '*New Scientist*', Nigel Calder, has used the phrase 'intellectual entertainment' to refer to a particular type of television programme which presented the discoveries and ideas of science to non-scientists. That, clearly, is the competent and responsible end of science in the media. At the other extreme, we are all uncomfortably aware of the trashy, sensational and inaccurate coverage of scientific developments as purveyed by certain newspapers. There is much that needs to be done to improve the presentation of science to the public, and science journalists themselves — most of them, at any rate — are very conscious of this need. The question I have been asked to consider today is: what should be the attitude of laboratory scientists to the world of the media? Let me make just four basic points.

First, I believe that research workers must expect that newspaper reporters and television producers are likely to be interested in what they are doing. Inevitably, the more glamorous or spectacular the research appears to be, the greater is likely to be that interest. But we should also remember that, whether in the realm of industrial research or that funded from government agencies, the public at large feels that it has a right to know about work that may have considerable implications for their lives. What I am saying, therefore, is that there is no longer a place for the monastic attitude, still to be found in some laboratories, which resents intrusions and inquiries from journalists. I have little sympathy with the physiologist studying *in vitro* fertilization of the human egg cell, for example, who refuses to speak to science or medical correspondents. Just as the first surgeons to carry through a human heart transplant must have anticipated immense public curiosity about their work, so anyone entering a sensitive field of scientific research really must expect that there will be corresponding public concern that will have to be satisfied. Inevitably, this will mean giving up some time to handle inquiries, to give interviews perhaps, and sometimes to write for the press. Any young person contemplating a career in science, particularly in a field that is or is likely to become controversial, must reckon with this obligation. I use the word obligation very consciously, and believe that a young scientist who is unwilling to spend a small proportion of time in this way should seriously re-consider their choice of career.

My second point may sound surprising. Always be honest when dealing with the press. Most people, most companies, most research institutes, are indeed scrupulous. Others are less straightforward. There are organisations that are content to issue press releases which are, shall we say, misleading. One example I have in my possession is a statement put out a few years back by an internationally famous hospital strongly denying newspaper reports that preparations were in hand at the hospital to proceed with a human heart transplant. In fact, this type of operation *was* being planned, and was indeed carried out shortly afterwards. And the publicity which the hospital had tried to stave off by issuing its erroneous statement proved all the more inconvenient.

The lesson for the press here (alas) is that one should never automatically trust the contents of a press release. The lesson for scientists is that, in the long term, it always pays to be honest with journalists. The risks of not doing so can be seen very clearly in American press coverage of nuclear power during the 1950s and 60s. In the 50s, nuclear reactors were portrayed as the dazzling, new, clean source of future energy, contrasting splendidly with the filth, grime and heavy social costs of using fossil fuels. This picture was presented by an alliance of press officers working for the US Atomic Energy Commission, and journalists who believed everything the AEC told them. Little or nothing was heard about hazards associated with the operation of nuclear power stations. Then, almost overnight, the tone of reportage changed radically. Journalists began to suspect that they had been misled into an inordinate admiration for nuclear power, and suddenly began to *disbelieve* all that they were told. Thus the tendentious, slanted press releases of a decade earlier rebounded on to the nuclear industry, which found itself plunged into the role of public villain. As I said earlier, it pays to be honest — even, and perhaps most of all, with those sections of the press which begin by being most suspicious of what you have to say.

A third observation is that scientists who come to deal with the press could benefit by learning how it works, by appreciating the problems faced by science correspondents in their battles with deadlines, and the phrenetic machinery of a newspaper office. I often meet research workers who have felt unhappy about reports of their work but who, on close questioning, turn out to expect newspapers to report science with the same detailed, qualified precision that is achievable in learned journals. This is unrealistic. Science writers, like any other journalists, should seek the highest standards of accuracy in their work. But it is simply impossible to expect them to distil the extensive discussion and fine nuances of a research paper into a story of 400 words.

Finally, as well as urging the scientists here to be positive in their relations with the media, let me argue too that more of them should take opportunities for writing themselves for the general public. It has always seemed to me a fallacy to say that scientists cannot write. Some of the most beautiful, economical prose I have ever encountered has been written by scientists — by people such as Sir Peter Medawar, Dr Lewis Thomas and Dr Jacob Bronowski.

The reason why non-scientists have a generally pitiful opinion of scientists as writers is because of the dreadful, unnecessary syntax which research workers slip into so easily when writing learned papers. Let me give you one example. A few years ago, I was editing a manuscript for publication and came across the following words: 'The mode of action of anti-lymphocytic serum has not yet been elucidated'. I changed this to: 'We don't yet know how ALS works'. On reading his edited manuscript, the author complained. I had altered his meaning, he said. Not so, I replied — and after discussing the matter for a quarter of an hour I convinced him quite firmly that my version was clearer *and* meant precisely the same as his. I could repeat this story many times over. Often, one can improve the readability of an article dramatically merely by moving sentences out of the passive and into the active voice. 'We modified the reaction' rather than 'Modification of the reaction was effected'. 'The plant can also produce methanol' rather than 'Production of methanol may in addition be achieved at the plant concerned'. Effective writing is rarely if ever easily achieved. It is always hard work. But a simple change of mental approach can work miracles. Forget the straigjacket of the research report, and imagine instead that you are telling your story to a friend who is not a scientist. Above all, try not to seek profundity behind a smokescreen of jargon and tortuous syntax.

Writing well, as I have said, is never easy. But the person who sits down to write about science has one enormous advantage. The discoveries and ideas of science are intrinsically exciting. No other field of endeavour yields such a rich fund of marvels and insights. Sadly, though, many people who would, and indeed do, find this material fascinating are apparently inhibited by the very word 'science'. It is, I believe, no accident that the two television programmes that are most successful in Britain in putting across science do not use that word in their titles. They are called 'Horizon' and 'Tomorrow's World'. For many, education has turned 'science' into a bogey word. To some degree, scientists and science writers have to overcome the intellectual unease over their subject that 'education' has instilled in those who are not scientists. At a time when political decisions *on scientific and technical matters* are increasingly being resolved by emotion uninformed by logic, it is surely our joint responsibility to crusade for the spirit of science.



Prof. Dr W. Martienssen, of the Physikalisches Institut der Johann Wolfgang Goethe-Universität, Frankfurt am Main, West Germany.

Science and Society

Research aspects of science in present-day society

W. Martienssen

Imagine that somebody asks you: 'What are the most important and the most exciting results of the natural sciences?' Personally, I would specify three items. My first point — I am a physicist, so I shall start with physics — my first point is the fact that there is only one universal physics: the entire universe, macroscopically and microscopically, is controlled by only one set of natural laws. They are in operation in the reactions of elementary particles as well as in nuclear and atomic physics; they are effective in our laboratory experiment and they finally also determine the origin and the fading of stars in the huge dimensions of our sky. The details of all these reactions and events are multifold. It seems as if nature uses, with the greatest of ease, all the various possibilities which are considerable in this interplay. But all these variations are in accordance with the few rules governing the game. These rules, we are convinced, have always been there and have always been the same, even a thousand million years ago, and we are fairly certain that they will also be there a thousand million years from now. Fortunately, they are completely protected from any human intervention.

In physics, I feel, this uniformity is the most exciting result. In chemistry — that is my second point — it is more the opposite, it is the multiplicity of matter which strikes us. To put it into an allegory: take three caskets, fill them up with small balls, the balls in each casket will be all the same, but different in the three caskets. Now, create the world out of it, out of just those three species, which may be called protons, neutrons and electrons; create a variety out of them, which is comparable to our world. Is not that phenomenal? You simply find it almost impossible to believe that the flower of a rose, the influenza virus and the semiconductor in an integrated circuit are all made out of the same particles, just three species, protons, neutrons and electrons. Different substances are produced when they are differently arranged.

Coming to my third point: it concerns the discovery that all living beings on Earth have the same genetic code, that they are all governed by the same computer and are all programmed in the same language. This is perhaps the most fantastic fact which has ever been revealed in nature. It not only strikes our intelligence, but also arouses our feelings. Are we — the trees, the birds, the mushrooms, and us ourselves — all intertwined together? You see, mankind is not alone; rather he exists in harmony with the other inhabitants of the Earth. Certainly, then, man is no longer the pre-eminent creature on Earth. Perhaps, however, this global community of seemingly infinite entities could comprise the supreme creature — *die Krone der Schöpfung* — in the universe?

But, you may ask, what has all that to do with science and society? To put it briefly, what I mean to say is that science is always, first and foremost, a spiritual adventure. The driving force in this adventure is enthusiasm. There is no scientific progress without enthusiasm. Please take a look at the EVOLUON; you will recognize how much enthusiasm helps, even if the available facilities may be limited.

Motivation of scientists and progress of science

Research is what keeps science young. A researcher is a man who raises, investigates and eventually answers new scientific questions, who interpretes scientific results and who unveils new applications of scientific findings. Most rewarding to a scientist are the moments when he:

— observes what has never been seen before,
— discovers what has never been dwelt upon before,
— invents what has never been conceived before.

It seems to me that understanding of the world around us and recognizing the role man plays in this world — the old questions about the nature of the world and about the purpose of life — are really the most obvious aims which engage the intellectual capacity of man. Along with our task of surviving, we are free to think. The imagination that each one of us bears within ourselves, about the world and mankind in it, is the most real expression of our existence and of the intellectual freedom which characterizes the individual human being.

Progress in science is approached in numerous ways. Sometimes the discovery of only one fact or the under-

standing of only one interrelation demands centuries of research. Take for example, the never-ending research towards the discovery and the unification of the six different forces which serve to keep our world together and which prevent it from collapsing, namely celestial and terrestrial gravity, electricity and magnetism, the weak forces and the strong forces. The discovery and unification of these forces started at the historial moment when Isaac Newton, sitting in the shadow of an apple tree in his garden, noticed a falling apple. It was at this very moment that the Goddess Athene told him that celestial and terrestrial gravity were one and the same thing. But research on this subject has been continuing for over 300 years and the unification of the four non-gravitational forces with gravity to produce the most general type of interaction, recently named supergravity, is one of the hot subjects of theoretical physics nowadays, with a large number of scientists working on it.

There are scientific results, on the other hand, which have been discovered in the glance of a moment, which seem to have fallen from the sky. For example, I should like to remind you of the way in which August Kekulé discovered the valency of the carbon atom, thereby establishing structural analysis in organic chemistry. He relates [1] that on a nice, warm summer evening in 1854 he was sitting on the roof of a horse-drawn bus, driving through the rather empty city of London. In a revealing dream, carbon and hydrogen atoms appeared, first forming sets of pairs, then groups, and finally linking together to constitute long chains. Then all the atoms started to turn around in a whirling dance. This dream was the dawning of the carbon chain and carbon ring molecular structure in organic chemistry. It was a very fertile and effective dream indeed.

Not only do the time scales in which scientific results are revealed tend to be different, but the level at which progress is achieved, can also vary considerably. The majority of today's scientists work in a fairly continuous fashion, dealing with problems which have already been dealt with by others, but which, due to lack of experience or lack of equipment, could not be solved. Only very few scientists are truly able to penetrate the boundary in order to forge ahead and to establish a new research field in a previously unknown realm.

Three of the most famous names in this context are Albert Einstein, who in the first decades of this century established the modern understanding of our physical world; Charles Robert Darwin, who in the middle of the last century published his book 'The origin of species by means of natural selection', in which he developed the hypothesis that natural selection is the key to evolution, and last but not least, we have Pythagoras, who established mathematics two and a half thousand years ago. We also owe a great deal of the conceptual basis of our philosophy of science to him. (It is said that Plato concluded from Pythagoras' work that God must be a mathematician.)

For the most part, scientists, as I mentioned before, have always been quite eager to bridge the gap between the newly established fields and the more traditional ones. Rather than breaking through, they cause a gradual shift with regards to the boundary of knowledge and experience: they apply, generalize and interpret scientific results. When I talk about scientists in the following, I mean scientists within this majority.

The scientist's place in society

Let me now turn to the second part of my talk. Scientists are not alone, they are members of a society. Like other people, the scientist is a *ζῷον πολῖτικόν* as Aristotle put it — a political being. Scientists serve society; society supports scientists.

Scientists enjoy a fairly high degree of social esteem. It is felt that scientists are educated to cope with both technological and other daily issues by using their critical judgement. From there, it is sometimes deduced, that they should also have a clear judgement about problems concerning social life. This belief, I think, is dangerous. Nature's behaviour is quite different from that of social groups. Consequently scientists may not reach their goal in this new field. Continuing along this same line, it is believed that scientists could be of assistance in time of need. That is, chemists may be able to purify the gases pouring out of a smoke stack which pollute the environment. Biologists may be capable of growing bacteria able to absorb oil which covers the sea and the beaches in the wake of a supertanker disaster. Physicists may eventually find new means of uncovering new sources of energy. Scientists, on the whole, may be able to contribute to solving the unemployment problem. Last not least, it is felt that the scientist's role in society is to do something new, something of interest.

Although they enjoy this esteem, academic scientists are frequently criticized for their self-centred interests, because they offer little in the way of viable solutions to the everyday dilemmas of ordinary citizens. Furthermore, scientists are reproached for their remoteness from reality. This is sometimes a strong accusation, but sometimes the criticism surfaces in a less direct manner. As a matter of fact, it happened to the very first scientist of whom we know: Thales of Miletus. Legend [2] has it that he was rebuked for his state of poverty. If he could not meet his own basic needs, then how could science and philosophy be of any use? Through his skill in reading the stars, however, he knew that there

would be an abundant harvest of olives in the coming season. (This was more likely the result of his knowledge of weather forecasting, of course.) So, having a little money, he put down deposits for the use of all the olive presses in Chios and Miletus, and hired them at a low price, because no one bid against him. When harvest time came, and many were wanted all at once, he hired them out again at any price he chose and made a great deal of money. Thus he showed the world that scientists and philosophers can easily acquire wealth if they choose, but their ambition is of another sort. Such is the legend told by Aristotle in his work 'Politics'.

Let us turn to more pressing problems: 'what factors exert the greatest influence on the research activities being carried out today? Should the scientist himself and no-one else determine the choice of his research subjects?' You have answered these questions in the inquiry. I was pleased that nearly half of you gave a positive answer to the last question: only the scientist himself and no-one else should determine the choice of his research subjects. Nevertheless, I am rather surprised that nearly 50% of you disagree with this statement. I am with the first group. My reasons are as follows.

1. Only the scientist himself can decide whether his interest, ability and experience will be sufficient for a successful approach. Also, the scientist himself must be the one to judge the limits of his laboratory facilities.
2. I should like to repeat a statement which one of you added at the end of the inquiry: 'if the scientist is not free in the choice of his subject, he cannot generate enthusiasm and interest'. This is quite in line with my way of thinking. Since research is a manifestation of intellectual freedom, the scientist should make full use of it from the very beginning, even in selecting his subject.

To make myself understood, I wish to emphasize that free choice of the subject would by no means call for exempting it from control. There is a natural and very effective control on research activities by the authorities which grant financial support. Very few scientists are sufficiently financially independent to bear the costs of their research themselves. They have to apply for financial aid and then, of course, they have to give an account of their progress. Besides this 'official' one there is yet another kind of control: scientists are eager to communicate. At least, they want to publish their results, but most scientists want to discuss their work before publishing or even before they actually start. A scientist isolated from everybody and not communicating with anybody is a contradiction in himself. Who or what is he working for? So it is also the scientific community which automatically controls the research activities, not so much on a legal basis, but on a moral and human basis which in the end would be even more effective. Again, I should like to stress: scientists serve society and society supports scientists.

In connection with the free choice of the subject of research it might be worthwhile to draw attention to another aspect of great importance in the daily lives of researchers. Most scientists no longer work alone, they work in a group and cooperate in a group. That is especially true in the experimental sciences. Moreover, groups have to cooperate in such a manner as to fulfil common aims of the research institution that they belong to. This co-operation, of course, restricts the freedom of the individual to some extent. In reality, it is the subject matter. What I would like to emphasize, however, is that each member of the institution who is actively contributing to the research programme should also participate in discussing the choice of subjects.

Another interesting issue in the inquiry is raised by the statement: 'scientists should feel responsible for the use and abuse of their research results'. More than 75% of the Young Scientists have reacted favourably to this statement. Personally, I would check the double plus block too. I feel that this attitude is a 'must' after having voted for the free choice of the scientist's research subject. Greater freedom automatically entails greater responsibility — at least this is my understanding of liberalism.

Again, however, the statement, as it has been given, needs some interpretation. We have to keep in mind that nearly every scientific result — even in the field of fundamental research — will sooner or later lead to various applications, beneficial ones as well as hazardous ones. 'Feeling responsible' in the above statement does not mean that a scientist who discovered something twenty years ago should be persecuted because somebody else starts to misuse his earlier results. What is meant here is that scientists who choose their research subject should always keep in mind the possibility that their results might be put to harmful use. If that were the case, they should diffuse as much information as possible on the potential risks which are involved. If large-scale dangers are to be suspected they should do everything in their power to eliminate them.

There are only few questions in the inquiry which have been answered so unanimously. This reflects a general attitude of today's scientists: most of them are quite aware of the dangers of applying scientific results without reflecting on the consequences. Let us hope that this attitude will be effective in the future as well. The same demand, of course, holds true for the other

[1] Richard Anschütz: 'August Kekulé', Berlin 1929.
[2] Bertrand Russell: 'History of Western philosophy', George Allen and Unwin Ltd., London 1961.

branches of human intellectual activities, outside the natural sciences, as well. It is especially relevant in the domain of the social sciences, where the object of scientific research is man himself and his behaviour in society. We can only hope that such an inquiry amongst young social scientists would result in a similar unanimous accord.

Coming back to the general question, as to which factors exert the greatest influence on the research activities that are carried out, I should like to mention two additional factors whose importance is frequently underestimated. On the one hand, we have what is called tradition. Tradition and continuity certainly play an important role in the choice of research subjects. Predicting what the scientist will do is much the same as making a weather forecast: if you say that tomorrow the weather will be the same as today, you will be, roughly speaking, 80% right. If you say that next year the scientist will still work on the same problem as he is now working on, you may be even more accurate.

On the other hand, we have what is called fashion. Fashion plays quite a big part in the scientific community. There seems to be a strong tendency to focus one's research along the same lines as one's colleagues. If you need to apply for financial support, don't forget to mention that scientists in other countries — for example in the United States or in the Soviet Union — are planning to do work in the same area. It will help a lot. Scientists, in doing their research today, are not as self-confident as they might have been during the 19th century. They are searching for moral justification. The most elementary one, of course, is the statement: other people do it, so why can't I?

A look into the future

We live in an industrial and technological age. Scientists have acted as instigators in bringing about the enormous technological developments of the last few decades. Since the Second World War the total industrial production has been increasing at a breathtaking rate. It has been estimated that the industrial production of the entire world within a period of four or five years is now larger than the total industrial production of mankind during the period dating from Adam and Eve until 1945.

But it is not only the quantity of the production which impresses us, it is the level of sophistication which draws our attention even more. For example
— man invented the computer,
— man developed the synthetics industry,
— man learned how to handle antibiotics.

We have every reason to be proud. Science and technology have enabled us to understand our physical world and to control nature. Science and technology are at the roots of our economic well-being, thereby determining life's wealth and versatility. In short: science and technology have made our lives comfortable.

Yet there are serious doubts as to whether technological progress has always been for the good of society. Three reasons for these doubts have been analysed extensively by the British economist Ernst Friedrich Schumacher [3] in his book 'Small is beautiful'. In short, they are.

1. The enormous increase in manufactured goods entails an enormous waste of raw materials. However, as you know, the supply of raw materials is limited. It is diminishing much faster now than had been expected. Our economic well-being is based on exploiting the Earth's resources. We take the raw materials without paying for them. Continuing to do so at such an alarming rate will leave our children deprived of necessities which we, in this century, have taken for granted. Furthermore, if we live from the resources, would it then not be fair to share them with the world's population?
2. We are able to control nature, but at the same time we are spoiling it. The enormous industrial production is overshadowed by an alarming increase of pollutants. Our means of coping with it are limited. The most common procedure is to dilute and to distribute them over such large volumes that people no longer perceive them. Although this procedure does work on a small-scale basis, since nature has a way of purifying itself, it no longer works when the quantity of refuse becomes too great. One thing is certain: the disposable volume of clean air and water is limited. Of course, there is again the additional question of whether this procedure is fair to our fellow-citizens; let us confess, it is nothing else than the age-old practice of throwing the junk over the hedge into the neighbour's garden.
3. We feel that our economic system and our way of life not only spoil our environment, but spoil our basic human values as well. We no longer understand ourselves as an intrinsic part of nature. We have come to the point where we are able to govern nature. As a result, we presume that it is our right to consume nature. Man has never been so unrestrictedly convinced that he is the pre-eminent creature in the world as he is today. This pretentious attitude against nature goes hand in hand with the decline of values in today's civilization. Modern people tend to disgard nearly all values, religious and political ones as well as ethical and educational ones. Our philosophy has become very

[3] Ernst Friedrich Schumacher: 'Small is beautiful. A study of economics as if people mattered', Blond and Briggs, London 1973.



The Chairman of the Jury with the winners of the five Awards. From left to right: Alistair Wolf, Paul Brown, Donald McDonnell, Prof. Dr. J. Volger, Sophie Valtat, Norbert Brunner, Martin Trüssel and Clemens Trüssel.

restrained. The only generally accepted aims in life seem to concern:
— more money,
— more social advancement,
— less work.

What can be done, and how can scientists help overcome this situation? The two first-mentioned problems belong to the domain of world problems, which will be studied extensively this afternoon in the talk by Dr Pannenborg. The third one is part of the education problem, on which Dr Dixon gave his talk this morning. But all three problems are interdependent. I am not at all in a position to give an exhaustive answer, but I should like to contribute to the discussion by specifying two aspects which in a way are in between 'education' and 'world problems'.

My first point is to criticize the belief that economic expansion is absolutely necessary. In the past technological progress has mainly been measured in relation to its contribution to the increase in the gross national product. Economic growth has been a sort of a magic formula for our way of life. This inbred belief is in fact the basic stumbling-block separating us from a more realistic outlook on the world around us. It must be clear to each scientist that the idea of an unlimited economic growth is very short-sighted. As specified before — it is neither possible nor desirable to change the whole Earth's surface into a great industrial maze. In long-range terms, the only economic system which guarantees lasting stability is a system which approaches self-consistency. This is a system which to the greatest possible extent makes use of recycling processes such that the waste of raw materials on the one hand and the pollution of the environment on the other are reduced to the lowest level possible.

Politicians claim that economic expansion is necessary in order to maintain political stability and full employment. It is obvious that the recycling practice would initially be unwelcome, because it entails higher production costs which in turn will make all products more expensive. On the other hand, it is not necessarily true that restriction of economic growth would aggravate unemployment. On the contrary, recycling would tend to increase the work force as a whole, since new tasks will require new jobs.

It will take a lot of persuasion to change the overall outlook of people in this respect. For the time being, people think along quite different lines. Let me give you one example. When you visit an automobile dealer and observe people buying cars you can see that they are quite willing to pay out an extra 1000 guilders for some extra gadgets. But have you ever heard a purchaser arguing that he wants a special engine which would be less damaging to the environment and that he would be willing to pay an extra 1000 guilders for it? Or have you ever heard of a purchaser buying a new car and leaving the old one at the dealer's with an extra amount of money for its dismantling and recycling?

This leads me to the second aspect I wish to mention. Technological progress has made us prosperous. In a way, however, it has also made us poor. We have concentrated so much on all the fascinating avenues which have been opened up by it that we have lost track of what we are doing. In order to illustrate what I mean, I will give only one example: go to an international airport during the holiday season and ask people why they are planning to fly such a long distance for their holidays. Most of them will answer that they have already been to all the nearer places, therefore they want to go elsewhere this time. No more. It is really necessary to spend so much precious fuel for that purpose?

So, beyond all the technological innovations, we have lost touch with the essentials: what is our life good for, what is the purpose of our being? All over the world there is nowadays a strong tendency to do everything in a more scientific manner. School teaching, for example, is supposed to be much more scientific than it was before. A much larger percentage of the population has the opportunity to participate in high-school education. But it seems that the growing scientific education is paralleled by a growing helplessness in the most central branch of knowledge, philosophy, which formed the true beginnings of science. During the last meeting of the Nobel prize winners at Lindau/Lake Constance the physician and biochemist Sir Hans Krebs, talking on juvenile delinquency and unwillingness to work, referred to the fact that antisocial behaviour grew worse during the very period in which the social sciences were being heavily promoted. One could continue by saying that the assurance on what to do and what not to do got lost during the very period in which scientific education was being intensely promoted.

A solution to the technological crisis cannot be ordered by law. It demands a change of consciousness. That will take a long time. What we really need is a philosophy of work, a philosophy which brings home to us that work actually has an importance attached to it, that we do have an obligation to work. In the long run the essential goods on this Earth cannot be taken free of charge, we have to work for them. Work is not only a burden, work has a meaning of its own. It is the basis of our material and also of our spiritual well-being. Work will help us to combat the widespread social diseases of insecurity, loneliness and frustration.

The loss of a proper relation to work has been strengthened by far-reaching centralization and specialization: children cannot see nor can they understand

what their parents do professionally; industrial workers do not know what the end-product is for which they prepare specialized components; physicians lose their goal when they concentrate entirely on single organs. So it might be necessary to reconsider the organization of work. Besides the specialized work in large organizations we need a renaissance of the 'home-made' type of work. One concept is that of 'Eigenwork', a term introduced by Christine and Ernst von Weiszäcker [4]. Eigenwork is meant as work for oneself; it allows for full self-expression. Work and success go together and are experienced as one unit. It can be work in the family, in the community, in clubs; it can be manual work, agricultural work, education and training. Scientific research, for example, even in the second half of the 20th century, is still very close to the concept of Eigenwork, inasmuch as it is planned and executed by one and the same person, bringing to light his freedom of initiation and his satisfaction in accomplishment.

There are fields where Eigenwork has been the usual form of work at all times; house-keeping is a typical example, but it should be possible to transfer the concept to industrial manufacturing as well. It then will need a special technology, a human technology, which does not alienate the worker from his own product. It should be a technology which is oriented towards human satisfaction rather than economic expansion.

Whatever the future development of science and technology may be, it has to be judged on a human scale and it should be able to bring man back into harmonious relation with nature. The longing of scientists for this type of awareness was demonstrated most impressively to me during a Young Scientist's Fair in Mainz in 1972, when Neil Armstrong, who together with Edwin Aldrin in July 1969 was one of the first men to walk on the Moon, reported on their trips. Neil Armstrong did not give a scientific talk, nor did he give a technological report, but he told a beautiful fairy-tale about man and the Moon, sublimating all his scientific, technological and psychological experience into this very human form of poetry. I have rarely at any time seen an audience so fascinated. The report was perceived as a look into an unscathed world, where science and humanities again form a unit.

Ladies and gentlemen, please allow me to express my gratitude to the Philips Company for the kind invitation to attend the 10th Anniversary of the Philips European Contest for Young Scientists and Inventors. This has enabled me to enjoy with you a glimpse through the European Window, tracing the path of science and society in their evolution from the past into a distant future.

---

[4] Christine and Ernst von Weizsäcker: 'Für ein Recht auf Eigenarbeit', Technologie und Politik, Bd. 10, Rowohlt 1978.



Dr Ir A. E. Pannenborg, Vice-President of the Board of Management of N.V. Philips' Gloeilampenfabrieken and responsible for Research and Development in the Philips group of companies. Dr Pannenborg is a member of the 'Club of Rome'.

Philips tech. Rev. 38, 33-39, 1978/79, No. 1

Science and World Problems

A. E. Pannenborg

From a distance of two hundred thousand kilometres the Earth is a coloured globe. The oceans, the continents, the mountain chains and the deserts appear as separately coloured patches. I regard it as a privilege that we can see our Earth in this way. For many centuries, people thought that they lived on a flat disc from which they could easily fall off. Later, navigators proved that the Earth was a globe. These explorers also showed that unheard-of possibilities existed in distant parts of the world. But now that man — continuing the human adventure — has left the confines of the globe we are seeing for the first time the finiteness of the dimensions and the possibilities.

From a distance, nothing can be seen of human activity on Earth. It is only when you are standing with both feet on the ground that you see the silent forests, the village squares covered in flowers, our mechanized agriculture, the industrial complexes, the urban areas, the chequered landscape, the soil and water pollution, the slums, the exhaustion of our natural resources, the hunger, the poverty and the massive mortality of men, women and children.

Classification of world problems

Poverty, hunger and death are as old as mankind. A new element in human history is that in the last half century many problems have imperceptibility become interwoven with each other. That is why it is also increasingly difficult to solve big problems like the population explosion, environmental pollution, the shortage of food or energy and the social issues linked with these. The problems have become world-wide.

The expression 'world problem' is relatively new and does indeed reflect a new notion. A precursor of this notion can be found in the term 'world war', which has been used to describe two major international conflicts in this century. It implies that these wars affected the majority, if not all, of the world's population. Accordingly, a world war — or to be more precise at this moment the prevention of the next one — certainly is a world problem. If there was any doubt about this in the past, that doubt has completely vanished since nuclear weapons became available. In discussing a notion such as world problems, it seems proper to define or classify its content. This is not difficult. In the first place the term 'world problem' should only be used of an issue which is the concern of everyone or almost everyone who is alive at the present day. The definition could then be extended to cover those problems which will concern everyone at some time in the future.

The second criterion for classifying something as a world problem is the effect of the finite size of the Earth on its existence. This finite size makes itself felt in two main ways. One is the limitation it imposes on the physical resources available to man, a limitation of which man has recently become extremely conscious in the case of fossil fuels. The other is the interdependence between different problems which in former times could be considered in isolation.

Man's reaction to world problems

Now that I have outlined my definition of what are nowadays termed world problems we can examine what motives there are for man to consider these as real problems, and to apply himself, through study and analyses, to trying to understand them and to finding answers that will enable him to solve them. Here we find that this motivation is more fundamental and universal than in the case of questions of a more limited nature. It is clear that world problems are recognized as being potentially so threatening that the urge to preserve mankind as a species comes into play. This is quite distinct from '*amour propre*' which applies to an individual or even the motive of self-preservation which inspires a nation.

A world problem also brings out the natural urge to conserve what man has achieved up to now. From the outset it is clear that we cannot be completely successful in this respect and in discussing any world problem one is always confronted with the question of how to set priorities in order to conserve at least the major part and the essence of what man has already attained.

The technological background of world problems

Much is at stake, both on the positive side of the balance sheet and, unfortunately, also on the negative side. Man has subordinated the world to his will, but he has paid a high price for this in the form of world problems. Unmistakably, the driving force behind these developments is man's technical genius. Before giving you a review of the world problems of today, I would first like to take you on a trip into history in search of

the use of technology. I shall go through it touching only on the major milestones. I am mainly concerned with showing you how the availability of energy has affected — and is still affecting — all kinds of technological developments. The threat of surrounding nature was a primary stimulus for man to escape from this. In this sense, man is a technical being. In the dawn of human existence, that was already shown by his skill in working materials: stone, bronze and iron, which are even characteristic of successive ages. But in using these simple materials, the human being of those times already reached great heights. In the period of half a million years the shapes of, for example, flint hand axes evolved in such a way that they are still characteristic of the subdivisions of the Stone Age. With only five flint tools such as the hand axe, the knife, the plane, the drill and the spear, paleolithic man succeeded in conquering almost the entire Earth.

With the transition from the hunter's and collector's existence to an agricultural economy, in which the population lived in one fixed place, man's technical genius made further important contributions — for example to improving food supplies and facilities for shelter. Up to the end of the Roman empire the most important sources of energy were animals and human beings. Thanks to the great wars of conquest, the Romans had an abundance of cheap labour, so that there was little or no stimulus to apply the inventions of Roman manufacturing technology on a large scale, though these inventions did exist. As early as one century before Christ, Vitruvius invented the water-wheel, but it was only in the latter days of the empire that it was used on a very modest scale for grinding corn and sawing blocks of stone. This last fact is known from the poem 'Mosella' by Ausonius (370 A.D.) who speaks of 'screeching saws which the water (of the River Ruwe) draws through the smooth marble blocks'. In those days the great wars of conquest were past and the Romans no longer had large numbers of slaves available to do the (manual) work. In this respect, the power supplied by men and animals, which was used for centuries, restricted possible technical development which perhaps might still have saved the Roman empire.

As against this, the watermill won an early and rapid success in Western Europe as a prime mover. For example, it is striking to note the large number of watermills of the Vitruvian type which are mentioned in the English Domesday Book of 1086. In it, the compilers arrive at a count of around six thousand. This source of energy had a great influence on mediaeval society.

It is likely that the introduction of the water-wheel took place so easily and rapidly because there were no longer any — or only very few — slave workers. The main difference between this early Anglo-Norman society and that of Roman times was the greater freedom enjoyed by its members, although compulsory services to the lord of the manor remained customary for a long time.

Some time later, the use of windmills also had a stimulating effect on the early industrial development of society. Particularly in the flat country in the west of the Netherlands, thousands of windmills were built. In the sixteenth and seventeenth centuries the Netherlands even had a technological lead in comparison to other European countries. This was undoubtedly partly due to the availability of this well-developed prime mover, the windmill, which formed the basis for developments in shipbuilding, food-processing, cable-making and of course drainage and the reclamation of land. Technical progress was by and large based on an experimental approach, but the first contributions from science to technology became visible in the work of Christiaan Huygens and Simon Stevin.

Thus man, in confrontation with nature, attempted to conquer everyday problems. In fact, nowadays we can scarcely conceive what privations were formerly suffered and the abuses to which that gave rise. From the summary made by the nineteenth-century physician Dr Coronel, it can be seen how Middelburg, a small town on an island in the South-West of the Netherlands, was devastated by disasters in the past and similar facts are known about a large number of other towns. For example, in 1348, the plague reduced the population in Europe to two-thirds of its size in a period of three years.

This disease struck at various places at intervals of ten years and even in 1720, half of the total population of 90 000 died in Marseilles.

But even if there were no outbreaks of disease, there were worries. To keep the population growth within bounds, recourse was had to birth-control practices which now sound incredible to us. As recently as the middle of last century infanticide was used on a large scale in Europe as a method of regulating the birth rate. Abandoning a child also virtually always meant its certain death, even if the infant was taken into a foundling hospital.

But let us continue with the development of man's technical progress. The middle of the eighteenth century saw the discovery of the steam engine, which was to result in a tremendous expansion of technical potential, an expansion which went for a long time. Its first application was in mining where water had to be removed from the pit. That was a problem that could scarcely be solved in any other way. Initially, the steam engine was regarded as an addition to existing energy sources such as the water-wheel. If, for example, a rotary

movement was needed for hoisting loads out of the mine (which prior to that time were carried up on the workers' shoulders) that was still done by a water-wheel using the water which had been pumped up. Even when the crank and crankshaft were already in use, the combination of steam engine and water-wheel was still employed — for example, in the textile industry, because a regular speed of rotation was required there. However, in a short time the steam engine won its own firm position as a supplier of mechanical energy. As is well known — I need not go into this in detail — its use in this respect had great consequences for the contours of Western society. Centralized industrial production grew up. The expansion of transport facilities (the train and the steamship) made it easy to convey raw materials over great distances and later to carry the finished products away again. An extensive network of lines of communication gave an extra dimension to society. And so, gradually, our present industrialized world came into being, with an exponential growth of human activities as an answer to the increasing demand for products for an increasing population. This development was continually accompanied by a series of peaks and troughs, the phenomenon of the economic cycle.

Independently, the pursuit of science also expanded enormously, Even in the 19th century scientific investigation was in most cases carried out by highly gifted individuals working alone. Around the middle of last century research as an organized activity was taken up. The first subject area was organic chemistry in general and dye stuffs in particular. These investigations might be called the first example of industrial research insofar as they were carried out on behalf of and to the direct benefit of an emergent chemical industry. The electrical industry was soon to follow.

The Philips Research Laboratories were founded in 1914, first for research on problems posed by the incandescent lamp and later expanding into large physical and chemical laboratories, exploring a broad field. Thanks to this fundamental and applied scientific research, numerous inventions could be industrialized. It was the start of science-based industry. Large industrial areas grew up in the countries of the West and prosperity rose — as can be seen, for example, from the great increase in the goods consumed per person. And all this was possible thanks to the availability of abundant and cheap energy: coal, followed later by oil and gas and electricity generated from them. It should be added, however that the interaction between science and technology is not one-sided. Professor Casimir has pointed to the 'science-technology spiral' in order to describe the use of the latest technological results and products in scientific investigations. This two-sided interaction has been a major accelerator in the unprecedented advance of natural science over the last 100 years. This relates not only to physics and chemistry, but also to biomedical research.

It is the widespread use of technology on a massive scale which, as a side-effect, creates problems and it is necessary to seek ways of eliminating these. As the Greeks pointed out, it is the old case of 'Panta rhei' — everything flows and is constantly changing. The people who lived before us were perhaps unable to foresee clearly the harmful influence of the side-effects. One reason might be that in the two centuries which have passed since the start of the industrial revolution, technology has created a tightly woven network of relationships on the Earth. Life on Earth has become one large coherent system. A disturbance at one place in the world can bring about a reaction somewhere else, which still takes us by surprise. There are no longer any simple solutions. Often the problems are interlinked and can therefore no longer be solved individually. We must tackle them all simultaneously to prevent fatal reactions.

By and large, in most parts of the Western world the problem of protecting ourselves against nature has been solved and true poverty has been banished. The new challenge which faces us consists in preserving the many desirable effects achieved, whilst eliminating the undesirable side-effects that have now become so clearly evident. A good illustration is provided by the car. We all enjoy the tremendous mobility which it has given to the individual. On the other hand, we are angry or depressed (depending on one's temperament) by the fumes it creates, the noise, the amount of valuable mineral fuel used and the seemingly ever-increasing amount of land that is converted into highways. We want to get rid of all these inconveniences, but for the time being we are not prepared to give up our recently won freedom of movement.

What is a world problem?

So far we have discussed world problems in a very general and abstract way. It seems useful at this point to indicate the complex questions which have been recognized as world problems in more concrete terms. As is clear from my introduction, the threat of war and the prevention of its outbreak rank high on the list. In a way it forms a world problem apart from the rest, though in common with several other international issues war — always a lethal threat — has only become a world problem since man, through the application of technology, has created tools so powerful as to implicate the whole of mankind.

Surveying world problems as a whole I would divide them into two categories.

The first category is a direct consequence of the finite dimension of the globe and the limitations which this places on the resources available in our physical world. The resources meant here are not confined to those which command a price in the world economy, but they include such physical entities as three-dimensional space, water and air. The fact that the limitation of these resources has only recently given rise to world problems is a result of their vastly increased use by man. This increase is due to the phenomenal success of Western man in building up his wealth, and equally to the sharp growth which has lately taken place in the number of inhabitants of this Earth.

The second category of world problems seems from the very uneven distribution of most resources over the Earth. This should be compared with the distribution of mankind over the surface of the Earth — also quite uneven, but in a different way from the resources. These facts have led to a strong interdependence right across the world. The autarchy of kingdoms in the Middle Ages is far behind us.

From the multiplicity of world problems I shall select a number for brief discussion in order to try to draw up a balance sheet reflecting the situation on Earth. For this purpose I have selected the world population and the food supply, environmental pollution and the shortage of energy.

There is probably no Western technology which has affected the world situation so deeply as the development of medical know-how. Fairly abruptly in the 19th century a situation was reached in which the mortality figures in the European countries could be reduced drastically. This in turn gave rise to a sharp increase in the population of the continent, because the birth-rate remained at the traditional high level for at least a century. In step with the colonial expansion of the European nation-states, Western medical technology became available to vast regions of the world, including those which today we call 'less-developed countries'. This is one of the main underlying causes for the population explosion which we observe in these countries today.

It is estimated that there are at present more than four thousand million people living on Earth. Shortly after the end of the century that number will have risen to eight thousand million. In order to cope with this, the number of provisions for living and for working will have to be at least doubled, which also means an increasing demand for food, capital, and all kinds of materials and energy. An additional problem is constituted by the fact that many of the countries which are growing rapidly are already overpopulated now.

The countries of the West take a more respectful view of the value of any human life than they did a hundred years ago. In many respects, we feel responsible for the fate of the countless people living in conditions which the West itself knew until not so long ago. But at the same time we also see the problems which will arise if the developing countries achieve a significant increase in their standard of living. In spite of the gloomy prospects, there are nevertheless some brighter features. There are official figures which show that in some of the emerging countries the population growth is levelling out. This is a result of intensive information campaigns about birth control.

What will be the number of people at which a state of equilibrium will be reached on Earth? There are calculations which work out at around fifteen thousand million people. Will the Earth be able to support this number? A great deal of work will still have to be done on this. Highly developed agricultural areas will not be able to raise their production much further. The total area in the world devoted to food production can be further extended by only a modest fraction. The production per hectare in underdeveloped countries has to increase. In this connection the improvement of varieties of grain — an activity carried out under the auspices of the FAO under the very promising name of 'The Green Revolution' is very important, but will lose its value if steps are not taken to combat pests at the same time. Rats alone eat up twenty per cent of the world's harvest each year.

The increase in agricultural output is counteracted by the loss of good land. Deserts are still expanding. This is a typical example of a world problem. It is beyond the power of an underdeveloped country to try to hold back the advancing desert. If the neighbouring countries do nothing, the work would in any case have little effect. An international approach is essential in this respect. But many deserts are situated in those countries which have an underdeveloped economy and by and large lack the political power, organization and capital to tackle the problems. The interwoven nature of these problems again emerges here when it is remembered that deserts interact strongly with the climate well beyond their own geographical limits. Other than saying that great changes will occur, it is completely impossible to predict with certainty what exactly the consequences for the Earth's climate will be if deserts are cultivated and irrigated on a large scale.

The prosperity of the West is best characterized by the level and growth in industrial production per head of the population. There are two negative aspects to this: the environmental pollution which occurs in the

large-scale production of goods and later in the use and destruction of the product, and, secondly, the vast increase in the raw materials and energy required for the production of the desired product.

As far as environmental pollution is concerned, I would make a distinction between the *local* pollution resulting from industrial or agricultural production (think, for example, of fertilizing the fields) and the *world-wide* impairment of the living environment — for example by the burning of coal, oil and gas resulting in an increase in the amount of carbon dioxide in the atmosphere. The first type of pollution is at present being combated by all kinds of actions in conjunction with legal regulations. These can achieve a great deal in preventing the pollution that can be observed every day. A question of mentality is also involved here. The second kind of attack on the environment — i.e. on a world-wide scale — is more difficult. Here again, international agreements should be made which of course will only be effective if they are observed. In this respect, there are still many uncertainties about the influence of human activities on the environment.

The most important raw material for human use is at present fossil fuel. We have seen how, in the course of the centuries, the available energy determined the level and scope of technology. This in turn has led to a close correlation between economic growth and the growth of energy consumption. If this correlation were to continue, we would in the not too distant future be forced to an economic stand still by the depletion of mineral energy resources. It therefore seems that one of the great challenges that faces mankind is to develop patterns of economic growth which are not concurrent with an increase of energy consumption. Even if this can be done the existing sources of energy will be insufficient within a few decades to meet the demand. We shall therefore have to do specific research on new energy sources, whose freedom from pollution must be a prime consideration. One of these sources is of course the Sun, which can be used both for the direct conversion of light into electricity (solar cells) and for heating (solar collectors). We are still only at the start of these technological developments. At present, it is still not possible to foresee what part they will ultimately play in the energy supply, though it is certain that in this century their role will be modest.

Simultaneously with the search for new forms of energy we will therefore have to realize that we must make use of the existing energy sources more carefully. Three-quarters of the energy required in a household in Western Europe, for example, is used for heating. This can be reduced appreciably. The solutions for this reduction, however, require appreciable investment and will come gradually.

The role of science

We shall now come back to the role of science in the study of, and approach to, world problems. The major portion of today's technology is science-based. As far as world problems are linked with the widespread use of the results of technology, part of their solution will again be technological in nature. Accordingly, progress in the pure sciences will make a broader spectrum of technological possibilities available to man. In this way science, through the intermediary of technology, is visibly linked to world problems and their possible solutions.

Quite a different contribution of a scientific nature is provided by the recently developed disciplines of systems dynamics. As was pointed out above, the majority — if not all — of the problems which face the world today are interlinked. The interdependence and interaction between the various problem segments is complex and manifold.

Even before the Second World War a general control theory had been developed in which the concept of feedback has introduced. After the war, this was developed further into cybernetics. Under the direction of Professor Forrester, a group at MIT in the United States has played a decisive role in the consolidation of this new field, particularly by applying these scientific methods to urban and world dynamics.

The complexity of world problems easily surpasses the limit of what man can grasp through intuitive reasoning. One way to approach this complexity is to describe the situation — i.e. the problem — in a simplified way. This is a method which has been used by science for centuries, in the sense that the behaviour of physical reality is described by mathematical equations which connect a limited number of parameters which we consider to be relevant. The equations always represent an idealized picture that leaves out the non-essential qualities. When there is a high degree of simplification, which is usually necessary to permit a description of a complex situation like a world problem, we tend to call that simplified description a model. The model should contain the main parameters and their interrelationships, expressed in quantitative or semi-quantitative terms. The model as such — which is often not easy to create — brings order into the complex problem but does not necessarily provide deep insight, and certainly does not as yet indicate any solutions to the problem described. For this we have to study how changes in certain parameters will affect all the others. In present-day electronic computers man has created a tool which can carry out the multitude of calculations this requires. It is this method of systems dynamics which allows man to study the consequences of his own possible future decisions.

It will be clear that not all the problems which face the world today can be described in a model which is sufficiently quantitative in its relations between the relevant parameters to enable computer calculations to contribute to a solution. It has, however, been a great victory for human intelligence to demonstrate what a surprisingly large variety of problems is open to the methods of systems dynamics.

It is interesting at this point to look at the opinions of young scientists about the relative importance of research on world problems of a different nature. The questionnaire offered a choice of eight different world problems to be studied. The young scientists selected pollution in industrialized countries and the supply of raw materials and energy to these countries as the best problems to study. The population explosion in South Asia and the related problem of the food supply there came next, whereas they considered that the threats resulting from over-intensive urbanization and the question of rising unemployment were less interesting subjects for research. I think that what becomes clear from these results of the questionnaire is the fact that young scientists have a keen perception of the limitations of present-day science. They favour doing research on those problems for which the available scientific knowledge and methods give a fair likelihood of obtaining results. This is probably quite contrary to the opinions prevailing among the uninformed public, who today all too readily assume that in view of the unprecedented success of the natural sciences during this century the social sciences have also developed to a similar extent.

If we look again at the various categories of world problems and their component parts and try to envisage which disciplines from the realms of science can be applied, we find that they can broadly be divided into two parts.

One part consists of all those problems which are related to the physical environment of man on Earth. These can largely be dealt with in terms of natural sciences and systems dynamics. This can be supplemented by notions and relations from economic theory. A number of studies published as reports to the Club of Rome have been carried out on this basis. Of course, the limitations of man's present-day knowledge of natural phenomena again become evident here. A chapter which, in comparison to pure physics and chemistry, seems to have been somewhat neglected is climatology. Because all life is confined within a thin layer on the surface of the Earth, climatology plays a dominant role in several world problems with pollution in its various aspects as the conspicuous example. Climatology is also connected with the possibilities of recycling resources, for which nature provides us with the splendid example of the water cycle through evaporation and precipitation.

The second broad part of world problems as described above could be called 'man and his behaviour'. To study these we draw on emerging disciplines such as sociology, anthropology, political science and of coures economics, which are much less firmly rooted in historical development than the natural sciences. At present, with the exception of economics, these disciplines have not yet developed the basic theories and parameters that allow the construction of models and the derived conditional predictions which are possible with models for the physical world.

As yet, too little is understood, *a priori*, about the reactions of communities to physical or social changes in their environment. Anthropology, however, provides us with numerous indications that human communities react conservatively to innovations. And that is certainly the case when the change does not stem from a community's own culture. Ideas or techniques foreign to them are certainly copied, but not necessarily readily integrated into their traditional socio-cultural pattern of life. This leads to great tensions in assimilating such innovations. We looked briefly at an example in connection with the drop in the mortality rate in the developing countries as a result of Western medical knowledge.

With regard to man as a member of an organized community it is clear that scientific theories about decision processes and their mechanisms are lacking. Human decisions can, however, be taken into account in the models describing the physical parts of world problems in an interactive way, so that the consequences of an assumed decision can be studied in some detail.

Conclusion

In the above I have tried to give a description of what we nowadays call world problems and the tools which man has developed for their analysis and potential solution so far. The role of science, both in its possibilities and its limitations, is quite clear. Science provides the framework of concepts and laws which enables us to describe a fair number of world problems in the simplified form which is also customary in scientific investigations. Because of the overwhelming complexity of true world situations the degree of abstraction and aggregation is, however, considerably larger than is customary in scientific investigation and thus introduces a corresponding degree of inaccuracy. The creation of systems dynamics as a systematic method has become important in conjunction with computers as a tool for calculations which are so extensive that they

could not be carried out previously.

Let us come back to the Earth. Some of the older professional scientists and other invited guests present here today have found a task there. A great task still awaits the young scientists. I hope — and now I am addressing the young scientists — that I have scared you sufficiently to make you realize that never before in human history has such a call been made on human creativity as it has in our time. Perhaps you might say that the Earth has to be conquered all over again. The tools available to you for this are considerably more powerful than what primitive man had when he started his journey over the surface of the Earth. You also know what can go wrong. Whatever the direction in which your lives may develop, either in science or along some other course, I hope that you will have the opportunity always to take the decisions which can prevent negative consequences for the Earth and its inhabitants. I have complete confidence in your efforts to achieve this goal.



Volume 37, 1978/79, No. 1 pages 1-40 Published 8th December 1978

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79, No. 2

Influence of crystal defects on the luminescence of GaP

C. Werkhoven, C. van Opdorp and A. T. Vink

Light-emitting diodes (LEDs) are nowadays widely used in such devices as pocket calculators and digital watches, and as indicator lamps. In recent times there has been a growing demand for green LEDs, for which single-crystal GaP is the usual basic material. Although the control of this material has been greatly improved, crystal defects in GaP still cause so much non-radiative recombination that they seriously limit the efficiency with which electrical energy is converted into green light. The research described in the article below has yielded a greater understanding of the individual contribution of the various types of crystal defect to the non-radiative recombination in GaP.

Introduction

Single crystals of the III-V semiconductor gallium phosphide can be used for making diodes that emit red, yellow or green light when an electric current is passed through them [1]. In the fabrication of a GaP diode an *N*-type layer (about 30 μm thick) is grown epitaxially on a substrate (about 350 μm thick), which likewise consists of single-crystal *N*-type GaP. Next a *P*-type layer is added, either by epitaxial growth (about 30 μm) or by zinc diffusion in a part (about 5 μm) of the epitaxial *N*-type layer. Epitaxial growth may be performed by liquid-phase epitaxy (LPE) or by vapour-phase epitaxy (VPE).

The luminescence of a light-emitting diode is caused by recombination of electrons injected into the *P*-type layer and holes injected into the *N*-type layer by means of an externally applied voltage. Electrons and holes can also recombine without generating the desired visible radiation. In GaP diodes the luminescence efficiency is limited to a great extent by non-radiative recombination near or at crystal defects, which are always present in the *N*- and *P*-type layers and in the substrate.

The most important crystal defects that occur in an epitaxial GaP layer are shown schematically in *fig. 1*. The defects are classified by their spatial extent. Point defects may consist, for example, of foreign atoms in the GaP lattice, of vacant lattice sites (vacancies) or of Ga or P atoms that occupy interstitial sites in the lattice. Complexes of two or more point defects also belong to this category. Linear defects such as line dislocations usually originate in the substrate and extend up to the surface of the epitaxial layer. Planar 'defects' are formed by boundary surfaces such as the surface of the epitaxial layer and the interface between the epitaxial layer and the substrate. Three-dimensional defects include precipitates and also very small dislocation loops. When their dimensions are of the order

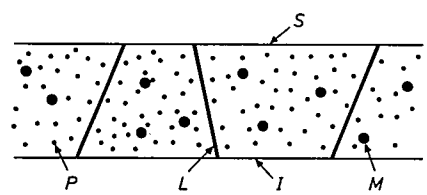


Fig. 1. Diagram representing an epitaxial GaP layer with the three main types of crystal defect. *S* surface of the layer. *I* interface with the GaP substrate. *P* point defects. *L* line dislocations. *M* microdefects.

Dr C. Werkhoven, Dr Ir C. van Opdorp and Dr Ir A. T. Vink are with Philips Research Laboratories, Eindhoven.

[1] The luminescence of GaP has been described earlier in this journal by H. G. Grimmeiss and H. Koelmans, Philips tech. Rev. **22**, 360, 1960/61, by H. G. Grimmeiss, W. Kischio and H. Scholz, Philips tech. Rev. **26**, 136, 1965, and by R. N. Bhargava, Philips tech. Rev. **32**, 261, 1971. General information on light-emitting diodes can be found in the book by A. A. Bergh and P. J. Dean, Light emitting diodes, Clarendon Press, Oxford 1976, and in the book by J. I. Pankove (ed.), Electroluminescence, Springer, Berlin 1977.

of 1 μm or less, these imperfections are also referred to as microdefects. They often occur when the solubility limit of point defects is exceeded during cooling of the crystal from the preparation temperature down to room temperature.

The nature and amount of crystal defects depend to a great extent on the conditions during the growth of the *N*- and *P*-type layers. The aim will be to choose these conditions in such a way as to minimize the influence of crystal defects on the luminescence efficiency of a GaP diode. To make a suitable choice it is necessary to have an understanding of the recombination processes that take place at crystal defects.

A study of these processes in diodes is complicated by the presence of the *P*-*N* junction. Our investigation was therefore directed at determining the ways in which the various types of crystal defect contribute to the non-radiative recombination in epitaxial *N*-type layers. By applying theoretical considerations the results thus obtained can be used for determining the influence of crystal defects on the luminescence efficiency of GaP diodes.

To ascertain the nature, concentration and distribution of the crystal defects [2] the defects were as far as possible made visible with the aid of X-ray diffraction, electron microscopy and chemical etching. The next step was the optical excitation of a large number of different layers so as to be able to measure the luminescence intensity as a function of time after the excitation was ended. The luminescence decays faster the more a particular type of crystal defect contributes to the non-radiative recombination, the influence of the other crystal defects remaining unchanged. From a careful analysis of the data we were able to find in broad outline the various contributions made to the total non-radiative recombination. It was found that line dislocations generally gave the largest contribution. By means of a special method of preparation the formation of line dislocations could largely be avoided. In a layer with relatively few line dislocations point defects play an important role.

The experiments were mainly carried out on LPE layers, but the results are also applicable to other GaP layers. Presumably the understanding gained can partly be carried over to other widely used III-V semiconductors, such as GaAs. The radiative recombination in this material gives rise to infrared luminescence, which is many times more efficient than the luminescence of GaP. Owing to the smaller influence of crystal defects on the efficiency, a study of the non-radiative recombination in GaAs is rather difficult.

In this article we shall first deal with the preparation of GaP and the characterization of the crystal defects. We shall then consider the recombination that takes place at the various defects. Finally, we shall discuss the reduction of the non-radiative recombination that can be obtained when the number of line dislocations is reduced without at the same time increasing the number of point defects.

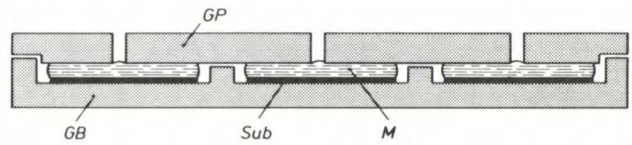


Fig. 2. Cross-section of a 'boat' for the preparation of epitaxial GaP layers by liquid-phase epitaxy (LPE). *GB* graphite boat. *Sub* GaP substrate, thickness about 350 μm. *M* gallium melt, thickness about 2 mm. *GP* graphite plate with perforations through which dopant can be supplied via the gas phase.

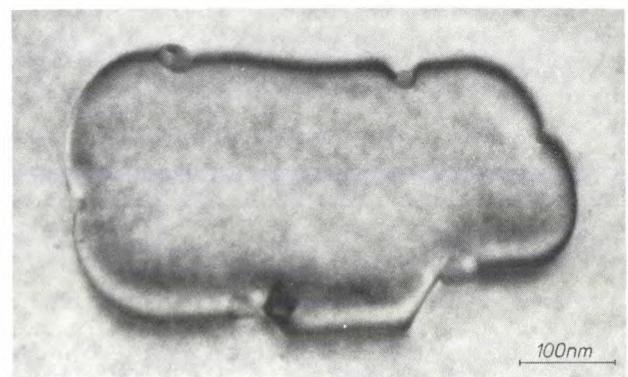


Fig. 3. Image produced with a transmission electron microscope (by A. J. R. de Kock and H. B. Haanstra) of a dislocation loop in a GaP substrate. The substrate contains no line dislocations.

Preparation of GaP

The GaP substrates were obtained from wafers cut from a large single crystal, which was made by pulling from a stoichiometric melt. This pulling is done at such a high temperature (about 1473 °C) that a relatively large number of point defects are incorporated in the lattice during the preparation process. The material obtained in this way consequently shows little or no luminescence and is not suitable for use as an *N*- or *P*-type layer in a diode.

The preparation of epitaxial layers by the VPE or LPE method takes place at lower temperatures (750-1100 °C). In a widely used VPE method hydrogen-chloride gas is passed over gallium metal, to produce $GaCl_3$. This is allowed to react at a suitable temperature with phosphine (PH_3), so that GaP can be deposited on the substrate.

In the LPE method of preparation that we used, a GaP substrate is covered with liquid gallium; see *fig. 2*. Part of the substrate is dissolved during heating. For a gallium layer 2 mm thick on a substrate 350 μm thick a final temperature of 1000 °C is suitable. After cooling

again an epitaxial GaP layer is obtained with a thickness of between 50 and 75 μm.

The experiments described in this article were mainly carried out on *N*-type LPE layers with a net donor concentration of about $10^{17}/cm^3$. The layers give green luminescence upon optical excitation or under cathode-ray bombardment.

Characterization of crystal defects

Points defects are too small to be made individually visible. There are various methods of characterizing spatially extensive defects, such as X-ray diffraction, electron microscopy and selective etching. The X-ray-diffraction methods have a limited resolution (about 2 μm), and microdefects are often too small to give sufficient contrast. The transmission electron microscope (TEM) has a very high resolving power (about 1 nm) and also a large focal depth. With this instrument it is therefore possible to identify even very small defects. As an example, *fig. 3* shows a TEM image of a small dislocation loop in a GaP substrate. A drawback of this method is that the crystals first have to be made very thin so that they are transparent to the electron beam. The crystals of GaP have to be thinner than 1 μm. This requires a complicated preparative technique, and also means that only crystal defects with a sufficiently high density ($> 10^6/cm^2$) have any great chance of falling within the transparent region. Added to this is the fact that the defect density may depend to a great extent on the location in the crystal, which limits the information that can be obtained with this method.

A rapid and simple method of making spatially extensive crystal defects visible is scanning electron microscopy (SEM) using cathodoluminescence [3]. In this method the surface of the crystal is scanned by the electron beam and the luminescence generated by the electron beam is detected. Non-radiative recombination occurs at the crystal defects, so that the intensity of the luminescence there is lower and the defects show up as dark spots in the cathodoluminescence image; see *fig. 4*. The resolution is limited to about 1 μm owing to the diffusion length of the minority carriers.

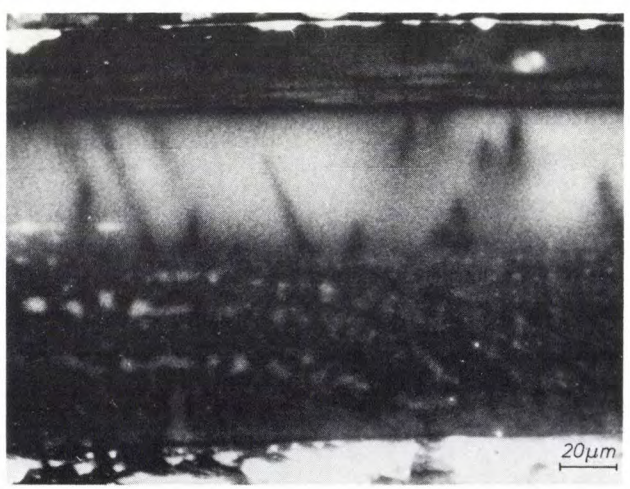


Fig. 4. Cathodoluminescence image, obtained with a scanning electron microscope, of an epitaxial GaP layer on a GaP substrate. The dark lines in the lighter part of the image indicate the presence of line dislocations in the epitaxial layer. This photograph was made by J. H. T. Hengst.

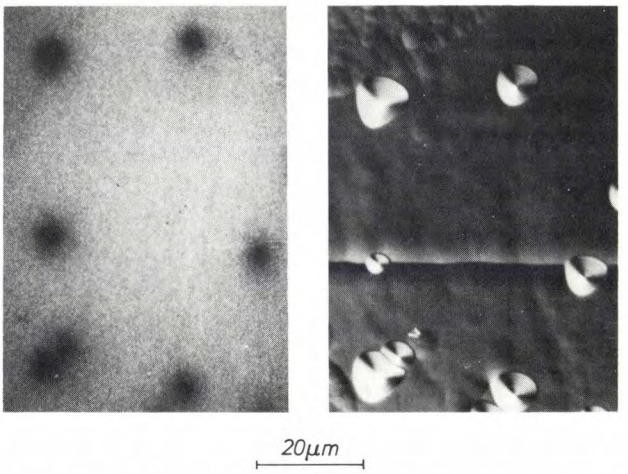


Fig. 5. Comparison of an etch-pit image (*right*) with a cathodoluminescence image, obtained before etching, of the same part of an epitaxial GaP layer (*left*). The etch pits show a clear one-to-one relation with the dark spots in the luminescence image.

A particularly fast method of making spatially extensive crystal defects visible is selective etching, using for example an aqueous solution of hydrochloric acid, nitric acid and silver nitrate. This method is based on the effect that at the location of a crystal defect the material dissolves faster in the etchant than at places without crystal defects. The appearance of an etch pit then indicates the presence of a crystal defect. In *fig. 5* it can be seen that each etch pit does indeed correspond to a dark spot in the cathodoluminescence image, which was obtained before etching. With the aid of the etching method large areas can be studied, so that even defects with a relatively low density can be found. In addition, information can be obtained about the nature of the defects if the correct etching time and an optical microscope of high magnification (about 2500×) are used. The resolution that can be achieved in this way is about 0.5 μm.

Line dislocations and microdefects can easily be distinguished by following the image of the etch pits with increasing etching times. This is because microdefects, unlike line dislocations, give rise to etch pits that dis-

[2] We shall henceforth refer to this as the 'characterization' of the crystal defects.

[3] D. B. Wittry and D. F. Kyser, J. appl. Phys. **35**, 2439, 1964. This method has also been described by C. Schiller and M. Boulou, Philips tech. Rev. **35**, 239, 1975.

appear again after a number of etching steps. We were thus able to establish that epitaxial GaP layers that are not too strongly doped contain hardly any microdefects. Substrates of GaP, on the other hand, always contain large numbers of microdefects. As an example of microdefects in a substrate *fig. 6* shows a photomicrograph of a substrate with dislocation loops after about 4.5 minutes of etching. The etch behaviour of a dislocation loop depends on the etching time. This is illustrated schematically in *fig. 7*. In the first instance a dislocation loop manifests itself as a pair of pits. After prolonged etching the dislocation loop vanishes from the material and the pair changes into a single saucer-like etch pit, which steadily becomes shallower and finally disappears. On the other hand, after longer etching new etch-pit pairs keep on appearing as a result of deeper-lying dislocation loops. The different stages can clearly be distinguished in a 'snapshot' as in fig. 6.

The line dislocations in epitaxial GaP layers are found to appear not only singly but also as pairs of closely adjacent dislocations called 'dislocation dipoles'. The etching behaviour of a dislocation dipole, like that of a dislocation loop, depends on the etching time; see *fig. 8*. After a short period of etching an etch-pit pair appears. *Fig. 9* shows a photomicrograph of an epitaxial layer with dislocation dipoles after an etching time of one minute. The etch pits clearly occur in pairs, each pair corresponding to a dislocation dipole. After a longer period of etching each pair becomes a single etch pit, which becomes steadily deeper and broader. By comparing photomicrographs of dislocation dipoles

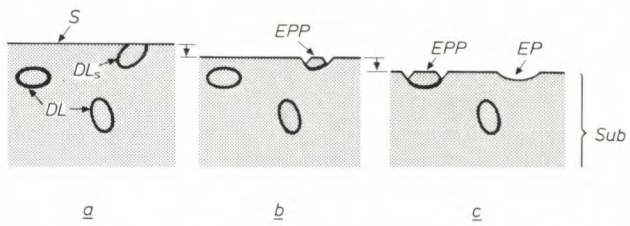


Fig. 7. Schematic representation of the etching behaviour of dislocation loops *DL* in a GaP substrate *sub*. The dislocation loop DL_s is intersected by the surface *S* (*a*). After a short period of etching, DL_s appears as an etch-pit pair *EPP* (*b*). After longer etching this pair changes into a single shallow pit *EP* (*c*). In addition, a new etch-pit pair arises from another dislocation loop *DL*.

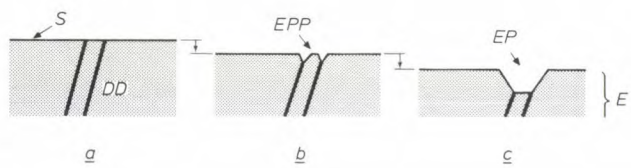


Fig. 8. Schematic representation of the etching behaviour of a dislocation dipole *DD* in an epitaxial GaP layer *E* with surface *S* (*a*). After a short period of etching, *DD* changes into the etch-pit pair *EPP* (*b*). After longer etching, *EPP* becomes a single pit *EP* (*c*), which becomes increasingly deeper and broader.

in epitaxial layers with those of dislocation loops in substrates we discovered that a dislocation dipole arises from a dislocation loop that is intersected at two places by the interface. This is illustrated schematically in *fig. 10*. The average distance between the two dislocations of a dipole is about 0.5 μm, which corresponds to the average diameter of the dislocation loops in the substrate.

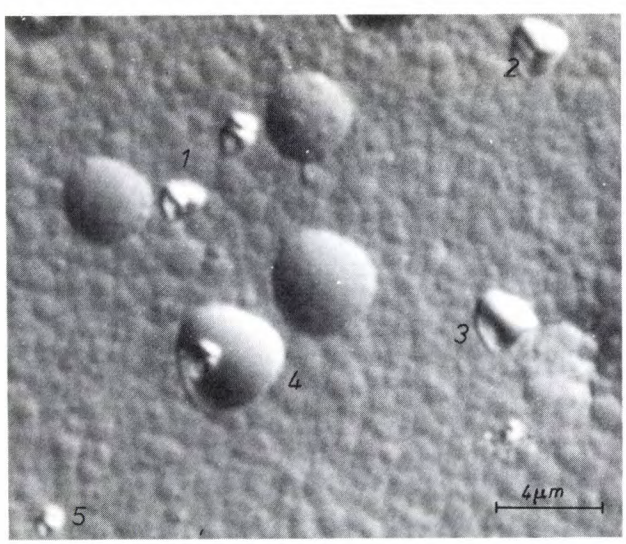


Fig. 6. Etch-pit image of dislocation loops in a GaP substrate after 4.5 minutes of etching. Various stages can be distinguished in the development of the etch-pit pairs as a result of the dislocation loops present: *1* small, well-separated pair of etch pits. *2* larger etch-pit pair. *3* combined etch pit. *4* shallow etch pit. *5* new, unresolved etch-pit pair.

Recombination at crystal defects

In epitaxial GaP layers we are mainly concerned with recombination at point defects, at line dislocations, at the surface and at the interface between epitaxial layer and substrate [4]. We shall now show first of all how it is possible, through the decay of the green luminescence after the excitation has been stopped, to obtain information about the contribution of each of these kinds of defect to the non-radiative recombination.

The mathematical basis of our approach is the continuity equation of the excess Δp of the minority charge carriers generated by the excitation. This differential equation states how Δp changes at every point in the layer during the course of time when the spatial pattern and duration of the excitation as well as all the recombination processes are given. In general the solution is a very complicated function of place and time. The parameters in this solution are the absorption coefficient α of the material for the light causing the excitation, the diffusion coefficient D of the minority carriers, the layer thickness W and the recombination parameters of the

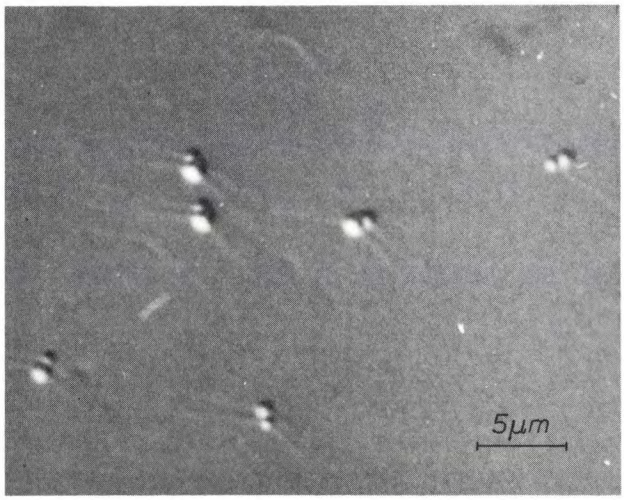


Fig. 9. Etch-pit image of part of an epitaxial GaP layer after one minute of etching. The clearly separated etch-pit pairs correspond to dislocation dipoles.

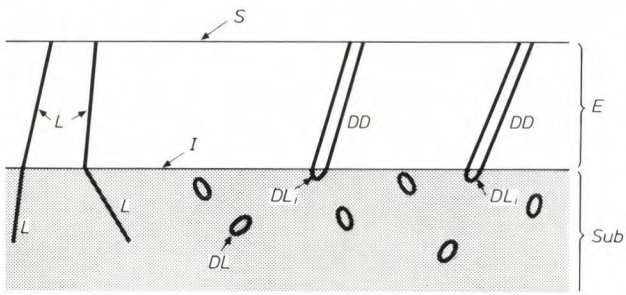


Fig. 10. Schematic representation of the formation of line dislocations and dislocation dipoles in an epitaxial GaP layer. A line dislocation *L* in the substrate *Sub* continues in the epitaxial layer *E* as a line dislocation. A dislocation loop DL_i, which is intersected at two places by the interface *I*, changes into a dislocation dipole *DD*, which extends up to the surface *S* of the epitaxial layer.

crystal defects. We shall first consider, as a simple case, a large thin plane-parallel layer with no dislocations but containing point defects. The light producing the excitation is uniformly incident on one of the surfaces ($x = 0$), and irradiates it long enough for a steady-state distribution $\Delta p(x)$ to occur in the layer. The excitation is then stopped. The minority carriers disappear mainly by two processes: recombination at point defects that are present in the layer in a high concentration and homogeneous distribution, and recombination at the two surfaces. A feature of the first process is that the rate of recombination at every point is proportional to the local excess concentration itself, and can thus be written as $\Delta p(x,t)/\tau_0$, where τ_0 is a constant with the dimension of time. It can easily be demonstrated that τ_0 is equal to the average lifetime of a minority carrier in an infinitely thick layer ($W \to \infty$). For the second process we make the realistic assumption that the recombination process at both surfaces is very much faster than the supply of minority carriers to the surface, which has to take place by diffusion [5].

Fig. 11a presents the calculated curves of Δp as a function of x at various times t after ending the excitation. The calculations were carried out for the case of an absorption coefficient α of 2000 cm^{-1} (for GaP this corresponds to an excitation wavelength of 458 nm) and for a layer thickness W of 75 μm. For the diffusion coefficient D we took a value of 3 cm^2/s; it will presently be shown that this corresponds to the true value. It can be seen that the strongly asymmetrical initial steady-state profile gradually changes during the decay to a symmetrical shape. From the moment when t/τ_0 has become equal to about 2.0, the profile continues to collapse while maintaining this symmetrical shape. The quantity measured during the decay is not $\Delta p(x,t)$ but the total intensity of the luminescence $L(t)$. Because the contribution of each volume element to the luminescence is directly proportional to $\Delta p(x,t)$, the value of $L(t)$ is proportional to the area under the curve of $\Delta p(x)$ in fig. 11*a* at the time t. The curve of $L(t)/L(0)$ as a function of t/τ_0 calculated in this way is shown in fig. 11*b*. Owing to the asymmetrical initial profile of $\Delta p(x)$, the light at first decays faster than exponentially. The deviation from exponential behaviour is much less pronounced than the deviation from the symmetrical shape of $\Delta p(x)$. From a t/τ_0 value of only about 1.0 the decrease of $L(t)$ is virtually exponential, and can therefore be characterized by an effective decay time τ_{eff}, given by the equation

$$L(t > t_1) = L(t_1) \exp [-(t - t_1)/\tau_{eff}], \qquad (1)$$

where t_1 is the time at which the decay curve becomes a straight line; see fig. 11*b*.

It also follows from the calculations that τ_{eff} can be expressed very simply in the parameters involved [6]:

$$1/\tau_{eff} = \pi^2 D/W^2 + 1/\tau_0. \qquad (2)$$

The contributions of $1/\tau_{eff}$ of the point defects and of the two surfaces together are thus given by two separate terms. The first term in eq. (2) is sometimes given in abbreviated form as $1/\tau_s$, so that

$$1/\tau_{eff} = 1/\tau_s + 1/\tau_0. \qquad (3)$$

This equation has a simple physical significance. Differentiation of eq. (1) gives

$$dL/dt\ (t > t_1) = -L/\tau_{eff}. \qquad (4)$$

Recalling that L is proportional to the total excess of minority carriers, we can then express eq. (3) as follows: the decrease of the total number of minority carriers

[4] C. Werkhoven, C. van Opdorp and A. T. Vink, in: Gallium arsenide and related compounds (Edinburgh), 1976 (Inst. Phys. Conf. Ser. No. 33a), p. 317, 1977.
[5] C. van Opdorp, Philips Res. Repts **32**, 192, 1977.
[6] W. Shockley, Electrons and holes in semiconductors, Van Nostrand, New York 1956, p. 323.

via point defects and via both surfaces can be described during the final phase of the decay as a removal through two parallel channels operating entirely independently of each other.

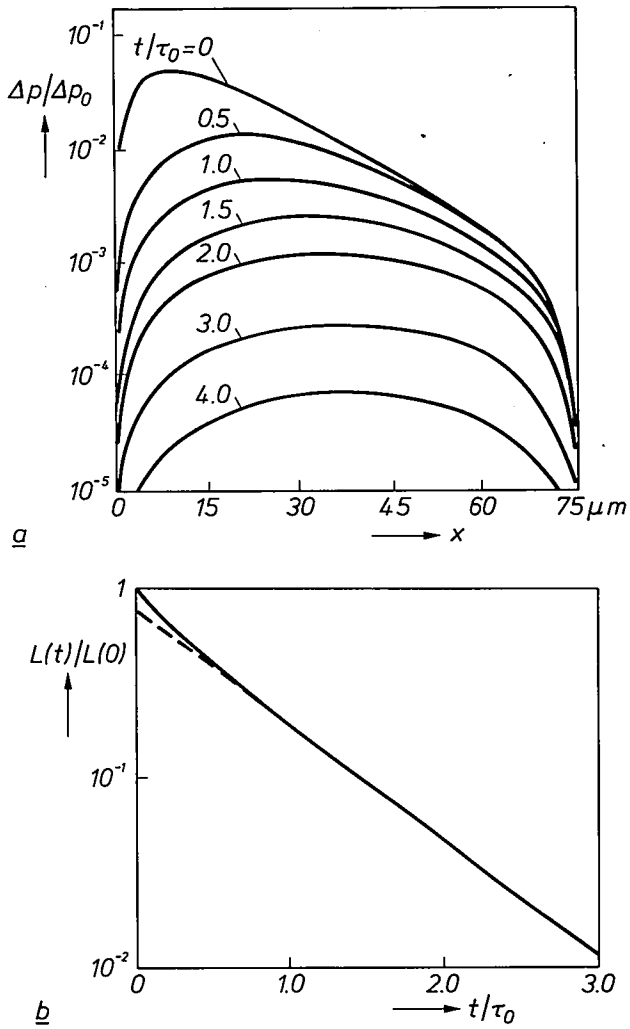


Fig. 11. Some calculated concentration profiles of the minority carriers (*a*), and a calculated decay curve of the luminescence in a plane-parallel GaP layer (*b*), after the excitation has been stopped. In (*a*) the excess Δp of minority carriers is plotted on a logarithmic scale as a function of the distance x from the surface. This has been done for different values of t/τ_0, where t is the time that has elapsed after the ending of the excitation and τ_0 is the characteristic decay time for the recombination at point defects. It has been assumed that the recombination at the two surfaces takes place very much faster than the minority-carrier diffusion. It has also been assumed that the distribution of Δp had reached the steady state before the excitation was stopped. The calculations relate to a GaP layer of 75 μm with an absorption coefficient α of 2000 cm^{-1}, in which the diffusion coefficient D of the minority carriers is 3 cm^2/s and τ_0 has a value of 0.75 μs. Values are normalized to Δp_0, the excess of minority carriers that would prevail under homogeneous excitation at the time $t = 0$ in a layer without surface recombination. As can be seen, as t increases, the distribution of Δp becomes more and more symmetrical. In (*b*) the calculated ratio of the luminescence intensity $L(t)$ at the time t to that at the time $t = 0$ is plotted on a logarithmic scale as a function of t/τ_0. The deviation from exponential behaviour arising because of the asymmetrical initial profile of $\Delta p(x)$ is relatively small and has practically vanished at a t/τ_0 value of about 1.0.

Now let us look at the case where there are also line dislocations present in the layer. The decay curve is then different, of course. In this case there is again an exponential decrease in the intensity of the luminescence after a certain time. It can be proved mathematically that, given regularly distributed dislocations that are perpendicular to the surfaces (a fairly reasonable approximation to the actual situation), equation (3) should be extended as follows [7]:

$$1/\tau_{\text{eff}} = 1/\tau_s + 1/\tau_0 + 1/\tau_d, \qquad (5)$$

where τ_d is the characteristic decay time for the recombination at line dislocations. Equation (5) implies that this recombination occurs in the form of a third independently operating channel during the final phase of the luminescence decay.

The contribution of microdefects to the recombination is negligible, since until now these have hardly been observed in not too heavily doped epitaxial GaP layers. From this it follows that eq. (5) can find general application in the analysis of the experimental decay curves.

It is possible to obtain information about the various processes more rapidly by measuring the decay times than by measuring the luminescence efficiencies, since the interpretation of the measured results is much easier. In fig. 11*a* the area under the curve for $t/\tau_0 = 0$ is a very complicated function of τ_0, D, W and α. This means that the emission intensity L and the luminescence efficiency during steady-state excitation are also very complicated functions of those parameters. The effective decay time for this case, on the other hand, can be expressed simply in terms of the parameters τ_0, D and W; see eq. (2). In addition it is often difficult to determine the *internal* luminescence efficiency from luminescence measurements.

For the excitation of the layers we used the 488 nm radiation from an argon-ion laser. *Fig. 12* shows by way of example the experimental decay curve for an arbitrary GaP layer. This curve has the same shape as the theoretical curve in fig. 11*b*, calculated for a special case. A curve of this type is representative of the GaP layers investigated. We shall now deal with the determination of the three separate terms in eq. (5) by means of τ_{eff} measurements.

Contribution from surface and interface

The recombination at the interface between the epitaxial layer and the substrate closely resembles the recombination at the surface of the epitaxial layer [8]. Even without additional crystal defects, the interface behaves like a plane on which the recombination rate is almost as high as at the surface. The high effective recombination rate at the interface is due to the great difference in concentration of crystal defects between

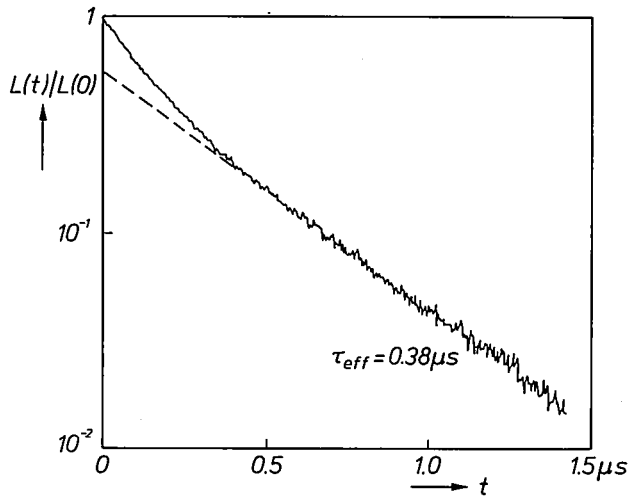


Fig. 12. Experimental decay curve of the luminescence of an epitaxial GaP layer after steady-state excitation with an argon-ion laser. After about 0.5 μs the curve is practically straight. From the slope of the straight part the value of the effective decay time τ_{eff} can be calculated with the aid of eq. (1). The dashed line is an extrapolation to $t = 0$ of the straight part of the decay curve.

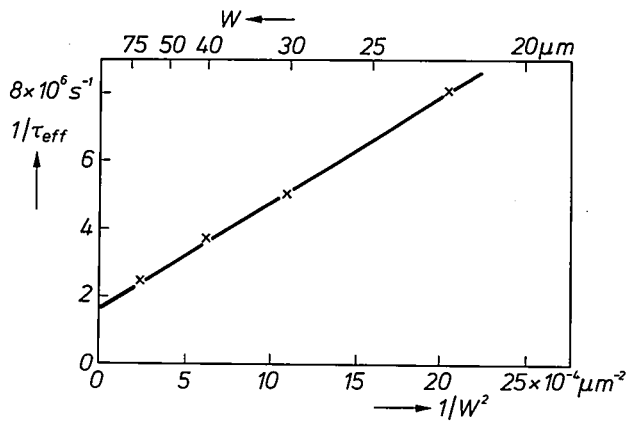


Fig. 13. Reciprocal value of the effective decay time τ_{eff} of the luminescence of an epitaxial GaP layer, as a function of the reciprocal value of the square of the layer thickness W.

epitaxial layer and substrate. Because of this the recombination rate in the substrate is many times higher than that in the epitaxial layer. Consequently the charge carriers are hardly able to penetrate into the substrate, and the interface may be formally regarded as a plane with very fast recombination. In the total process of the diffusion of charge carriers and the subsequent recombination at the surface or interface, the diffusion is therefore the rate-determining step. As we have just seen, the effective decay time is then given by

$$1/\tau_{eff} = 1/\tau_s + 1/\tau_0 + 1/\tau_d = \pi^2 D/W^2 + 1/\tau_0 + 1/\tau_d. \quad (6)$$

It follows from this that the contribution $1/\tau_s$ can be determined by measuring the effective decay time of epitaxial layers that have different thicknesses W but otherwise the same properties. A convenient way of varying the layer thickness without at the same time changing the other properties too much is to dissolve a part of the same layer in successive steps. In this way we reduced the layer thickness from about 75 μm to about 20 μm. The layers chosen for these experiments contained relatively few dislocations (about $4 \times 10^4/cm^2$). *Fig. 13* gives the results of a series of measurements on a particular layer, where $1/\tau_{eff}$ is plotted against $1/W^2$. It can be seen that a linear relation exists between these two quantities, in agreement with eq. (6). By extrapolating the straight line in fig. 13 to $1/W^2 = 0$ (infinitely thick layer), the collective contribution of point defects and line dislocations can be estimated. For the layer under consideration this gives a value of about 1.8×10^6 s^{-1}. The contribution $1/\tau_s$ from the recombination at surface and interface predominates as soon as the layer thickness is less than about 40 μm. For a layer thickness of 20 μm the value of $1/\tau_s$ is in fact nearly a factor of four greater than the sum of the other contributions. From the slope of the straight line obtained the value of the diffusion coefficient D can be derived. This is found to be about 3 cm^2/s, which agrees well with the value obtained from Hall measurements of hole mobility in *P*-type GaP [9]. It appears that the type and amount of crystal defects in the layer do not greatly affect the value of $1/\tau_s$, and hence of D; in all cases we found a value for D of about 3 cm^2/s.

Contribution from line dislocations

In the past the recombination at line dislocations has received much less attention than the recombination at the surface and at point defects. Partly because of this, relatively little is known about the influence of line dislocations on the luminescence. As in the case of recombination at the surface and interface, the recombination at a line dislocation is determined by the diffusion to the defect and by the recombination at or near the defect. In this case, however, it is much more difficult to determine which of these two is the rate-determining step of the whole process.

A useful expression for the decay time τ_d, characteristic of the recombination at line dislocations, can only be obtained after introducing a number of simplifications. These consist in assuming that the dislocations are perpendicular to the surface and that they are equally spaced from each other with a density N_d. It is assumed that each dislocation is surrounded by a cylinder of radius r_d, on the surface of which the

[7] C. van Opdorp, A. T. Vink and C. Werkhoven, in: Gallium arsenide and related compounds (St Louis), 1976 (Inst. Phys. Conf. Ser. No. 33b), p. 317, 1977.
[8] C. van Opdorp, C. Werkhoven and A. T. Vink, Appl. Phys. Lett. 30, 40, 1977.
[9] H. C. Casey, Jr, F. Ermanis and K. B. Wolfstirn, J. appl. Phys. 40, 2945, 1969.

recombination takes place with a characteristic velocity s. For such a system $1/\tau_d$ is given to a good approximation by [7]

$$1/\tau_d = [(1/2\pi N_d D) \ln (0.31/r_d \sqrt{N_d}) + 1/2\pi N_d s r_d]^{-1}. \quad (7)$$

The first term between square brackets corresponds to the diffusion, the second to the recombination of minority carriers (series process). If the diffusion takes place much faster than the recombination, the first term can be disregarded, so that

$$1/\tau_d = 2\pi N_d s r_d. \quad (8)$$

The speed of the total process is then determined by the recombination velocity, and $1/\tau_d$ then increases linearly with increasing N_d.

In the opposite case the second term between the square brackets in eq. (7) can be neglected so that

$$1/\tau_d = 2\pi N_d D/\ln (0.31/r_d \sqrt{N_d}). \quad (9)$$

In this case the diffusion is the rate-determining step, and $1/\tau_d$ no longer increases linearly with N_d, although the deviation from the linear dependence will not be large.

If we try to experimentally verify the relation of eq. (7) between $1/\tau_d$ and N_d by measuring τ_{eff} for layers of different N_d, we come up against the difficulty that there is not very much variation in the average dislocation density in the various GaP crystals. What is more, the contribution of other crystal defects (point defects) can differ quite considerably from one crystal to another, which tends to make the determination very inaccurate.

A better method is to perform the measurements on the same layer and to make use of the strong local variation of N_d within that layer, making the assumption, however, that the point defects are homogeneously distributed. The inhomogeneous distribution of the line dislocations can be made visible by treating the surface with the etchant mentioned earlier. With an optical microscope it is then possible not only to distinguish areas of different dislocation densities but also to excite them selectively. We did this by focusing the laser beam on these areas through the microscope objective. By detecting the generated luminescence via the microscope we were then able to determine the value of τ_{eff}. The diameter of the excitation area is normally 50 to 75 μm, but if desired it can be reduced to a few μm. This makes it possible to determine the effective decay time τ_{eff} of the luminescence from a piece of the layer that may happen to contain no line dislocations. In this case we measure, from eq. (5):

$$1/\tau^0_{eff} = 1/\tau_s + 1/\tau_0. \quad (10)$$

For areas that do contain dislocations, the contribution to the recombination can be expressed as

$$1/\tau_d = 1/\tau_{eff} - 1/\tau^0_{eff}. \quad (11)$$

Fig. 14 gives a plot of $1/\tau_d$ as a function of N_d for a large number of epitaxial GaP layers with different τ^0_{eff} values. As can be seen, the experimental value of $1/\tau_d$ increases more or less linearly with increasing N_d. This result may be regarded as a reasonable confirmation of

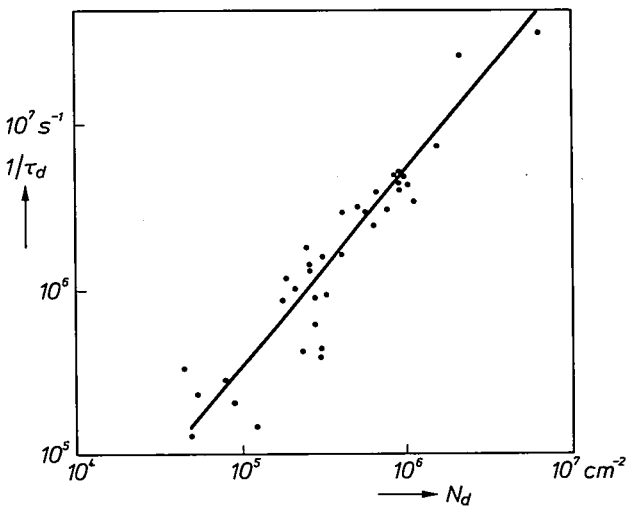


Fig. 14. Reciprocal of the decay time τ_d, characteristic of the recombination at line dislocations, as a function of the dislocation density N_d for a large number of epitaxial GaP layers. The value of $1/\tau_d$ increases more or less linearly with the value of N_d. The spread in the measured $1/\tau_d$ values is fairly large.

eq. (7). It is not possible to estimate the separate contributions of the diffusion and recombination terms from fig. 14, because their independencies on N_d do not differ greatly and because there is a fairly wide spread in the measured $1/\tau_d$ values. The spread in the measured points in the area with low densities is due to the small difference between the measured values of τ_{eff} and τ^0_{eff}. The presence of the dislocation dipoles mentioned earlier also gives rise to a spread in the values of $1/\tau_d$, because the recombination rate at a dislocation dipole will be considerably greater than at an ordinary line dislocation. Also, the distributions of the line dislocations within the excitation area are not homogeneous.

From the results given in fig. 14 it is possible to make an estimate of the average relative contribution of line dislocations to the recombination. The thickness of the layers is usually about 50 μm, so that the contribution $1/\tau_s$ of the recombination at the surface and the interface is about 1×10^6 s^{-1} (fig. 13). The average dislocation density in the layers is usually between 3×10^5/cm^2 and 5×10^5/cm^2. From fig. 14 we see that the average value of $1/\tau_d$ is then about 2×10^6 s^{-1}, which is about twice the value of $1/\tau_s$.

Contribution from point defects

The dimensions of point defects are of the same order of magnitude as lattice atoms. For this reason there is no method of directly displaying or identifying indidual point defects. It is however possible in special cases to demonstrate the presence of some point defects by indirect means [10]. The contribution $1/\tau_0$ of the point defects to the non-radiative recombination can be indirectly determined from the value of τ_{eff} for layers with known values of τ_s and τ_d:

$$1/\tau_0 = 1/\tau_{eff} - (1/\tau_s + 1/\tau_d). \qquad (12)$$

When this is done for a number of layers with different thickness and dislocation density, it is found that the contribution $1/\tau_0$ increases considerably the more the mean dislocation density decreases. This suggests that on average there are many point defects in a material with on average few dislocations. The reason for this is probably that point defects have less chance to precipitate on dislocations during the cooling from the preparation temperature to room temperature. In layers with a thickness of about 50 μm, and with the usual dislocation density of $3 \times 10^5/cm^2$ to $5 \times 10^5/cm^2$, point defects play a subsidiary role in the total non-radiative recombination: the value of $1/\tau_0$ is then at the most 5×10^5 s^{-1}, which at the least is a factor of two lower than the value of $1/\tau_s$ and a factor of four lower than that of $1/\tau_d$.

Reduction of non-radiative recombination

From the investigations described so far, it follows that line dislocations generally give the main contribution to the undesired non-radiative recombination in epitaxial GaP layers.

Any appreciable reduction in the non-radiative recombination would thus have to be brought about by trying to reduce the recombination at line dislocations, which means that the dislocation density would have to be reduced. This in turn implies reducing the number of line dislocations and dislocation loops in the substrate, since these are the origin of the line dislocations in the epitaxial layer.

By means of a special method of preparation it has proved possible to make GaP substrates that contain hardly any line dislocations [12]. Unfortunately, these substrates contain a large number of small dislocation loops with an average diameter of about 0.5 μm [13,14]. These are caused by the fact that during cooling from growth temperature to room temperature point defects cannot now be drawn away to dislocations. The resultant excess concentration of point defects gives rise to precipitates which, in their turn, may form dislocation loops. This effect of an increasing concentration of microdefects with decreasing dislocation density is also known in silicon and germanium crystals [15].

Epitaxial GaP layers grown on substrates with few line dislocations contain mainly dislocation dipoles; their density is approximately equal to that of the loops in the substrate, from which they originated. It has been found that it is then better to start from substrates with more line dislocations but with fewer loops [14]. The lowest average dislocation density that we have achieved in this way with epitaxial GaP layers is about $4 \times 10^4/cm^2$. The non-radiative recombination is then indeed found to be somewhat reduced (*Table I*).

The reduction is by no means as great, however, as one would expect from the strongly reduced dislocation density. As already noted, the reason for this is that the epitaxial layer contains a large number of point defects which, during growth, are no longer able to precipitate on neighbouring dislocations, which gives rise to an increase in $1/\tau_0$. This effect can be minimized by generating extra line dislocations, but then only *in* the interface. This is possible by introducing a large difference in lattice spacing between epitaxial layer and substrate, for example by means of very different doping. Given a sufficiently large difference in lattice spacing, the resultant stress is partly removed by the formation of line dislocations in the interface. Owing to the high density of these line dislocations these are to a large extent able to absorb the point defects from the thin epitaxial layer. In this way $1/\tau_0$ remains low, even when the dislocation density is low. This causes very little change in the contribution $1/\tau_s$, because the interface may be regarded, even without these line dislocations, as a plane with a high recombination rate. We have thus

In a continuously operating diode the line dislocations will be even more dominant than is the case in the layers under consideration with no *P-N* junction during the final phase of the luminescence decay. The reason for this is that the concentrations of both kinds of minority carriers in a diode are highest at the *P-N* junction; the concentrations decrease approximately exponentially with the distance from the junction. As a result, the line dislocations can draw away a large number of the injected minority carriers at the intersection with the plane of the *P-N* junction [11], while the surface and interface with the substrate are far from the *P-N* junction.

[10] D. V. Lang, J. appl. Phys. **45**, 3023, 1974.
[11] A. T. Vink, C. Werkhoven and C. van Opdorp, Proc. Topical Conf. on Characterization techniques for semiconductor materials and science, Electronics division, Vol. 78-3, p. 259; Electrochemical Society Spring Meeting, Seattle 1978.
[12] P. J. Roksnoer, J. M. P. L. Huijbregts, W. M. van de Wijgert and A. J. R. de Kock, J. Crystal Growth **40**, 6, 1977.
[13] A. J. R. de Kock, W. M. van de Wijgert, J. H. T. Hengst, P. J. Roksnoer and J. M. P. L. Huijbregts, J. Crystal Growth **41**, 13, 1977.
[14] C. Werkhoven, J. H. T. Hengst and W. J. Bartels, J. Crystal Growth **42**, 632, 1977.
[15] A. J. R. de Kock, Thesis, Nijmegen, Philips Res. Repts Suppl. 1973, No. 1.

been able to make a layer with a dislocation density of about $4 \times 10^4/cm^2$, in which the sum of the contribution of $1/\tau_d$ and that of $1/\tau_0$ was reduced by about a factor of seven by the introduction of additional line dislocations in the interface (Table I).

Table I. Comparison between the average dislocation density in the epitaxial layer and the sum of $1/\tau_d$ and $1/\tau_0$, characteristic respectively of the recombination at line dislocations and at point defects, for three different combinations of epitaxial GaP layer and GaP substrate.

Number of line dislocations per cm^2 in substrate	Number of dislocation loops per cm^2 at the substrate surface	Average dislocation density in cm^{-2} in epitaxial layer	$1/\tau_d + 1/\tau_0$ (s^{-1})
0	5×10^5	5.5×10^5	2.7×10^6
4×10^4	$\approx 10^4$	4×10^4	1.4×10^6
4×10^4 [*]	$\approx 10^4$	4×10^4 [*]	2×10^5

[*] This combination contains in addition a large number of line dislocations *in* the interface between epitaxial layer and substrate, causing a reduction of the number of point defects in the epitaxial layer.

The work described here has given a good idea of the influence of the various kinds of crystal defect on the luminescence of epitaxial GaP layers, and of the means by which non-radiative recombination can be reduced. Further research will have to show the extent to which the knowledge thus obtained can be used to increase the luminescence of GaP diodes.

Summary. Single-crystal GaP is a suitable basic material for 'green' light-emitting diodes (LEDs). In the manufacture of these LEDs the GaP is grown as an epitaxial layer on a substrate that also consists of single-crystal GaP. The efficiency of the luminescence is limited to a great extent by the presence of crystal defects, which give rise to non-radiative recombination. The investigation described here set out to determine the magnitude of the contributions made by the various kinds of crystal defect. The decay curves of the luminescence of a large number of GaP layers were therefore measured and analysed. In addition the crystal defects of the same layers were characterized, as far as possible, by means of X-ray diffraction, electron microscopy and selective chemical etching. The investigation revealed that the non-radiative recombination in epitaxial GaP layers is in most cases mainly due to line dislocations, which extend up to the surface of the epitaxial layer. These occur both as single dislocations, originating from dislocations in the substrate, and pairs, originating from dislocation loops. Reducing the average dislocation density by reducing the number of line dislocations and dislocation loops in the substrate does not in itself result in an appreciable reduction in the non-radiative recombination. This is because the number of point defects increases with decreasing dislocation density. An appreciable improvement can however be achieved when, as an additional measure, extra line dislocations are only introduced *in* the interface between epitaxial layer and substrate, which has the effect of keeping the number of point defects in the epitaxial layer relatively low, even when the dislocation density there is low.

Experimental etching equipment

In planar technology, which is applied in the manufacture of integrated circuits, the etching operation is an important stage of the process. Most of the etching methods now used employ a fluid etchant. One obvious way of bringing the specimen to be etched into contact with the etchant is by dipping. For routine applications, however, the dipping method can be rather laborious. Moreover, the reproducibility often leaves something to be desired, mainly because the etching rate tends to vary from place to place.

Since commercially available products offered no suitable alternatives, we had to design a simply operated etching equipment, giving reproducible results, that could be used in the research programme at Philips Forschungslaboratorium (PFH). Particular attention had to be given to the nature of the movement of the specimen and also to the way in which the etchant was applied. In addition it was of course also necessary in the design to take account of the chemical resistance of some of the components to common etchants, such as hydrochloric, nitric, acetic, sulphuric and hydrofluoric acids.

The equipment is shown in the photograph of *fig. 1* and in cross-section in *fig. 2*. It is essential that the specimen *Sp*, which is attached to the specimen support *Sup* by suction, should be given a tumbling or rotating motion, or both, and that the etchant should be sprayed on to the specimen.

The specimen support *Sup*, which is made of 'Teflon' [*] (PTFE), is located in the etching chamber *EC*, whose wall and base also consist of 'Teflon'. This material had been shown in an extensive preliminary investigation to be exceptionally resistant to the etchants named above. The support *Sup* has a raised sloping spring-loaded edge, which provides a seal between the support and the specimen. This allows the specimen to be held in place by suction. The suction (or underpressure) is produced by a vacuum pump, which is con-

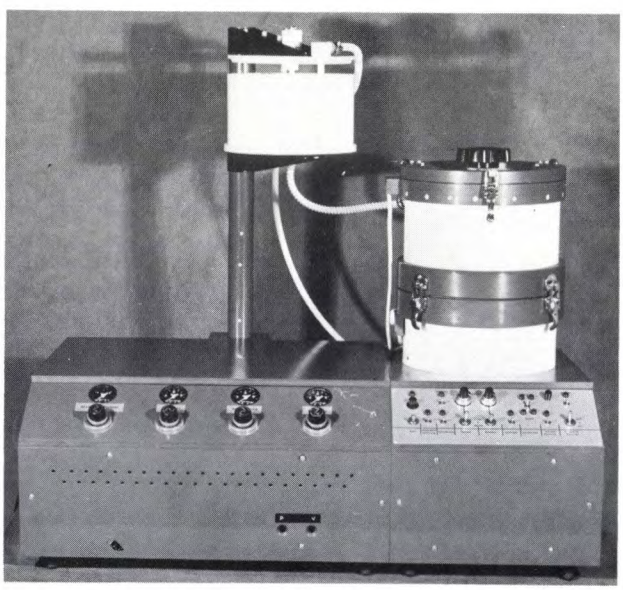


Fig. 1. Photograph of the complete etching equipment. The reservoir for the etching fluid is at the top, and the etching chamber is at the right. The motors, pumps and reducing valves etc. are all mounted in the base behind the control panel.

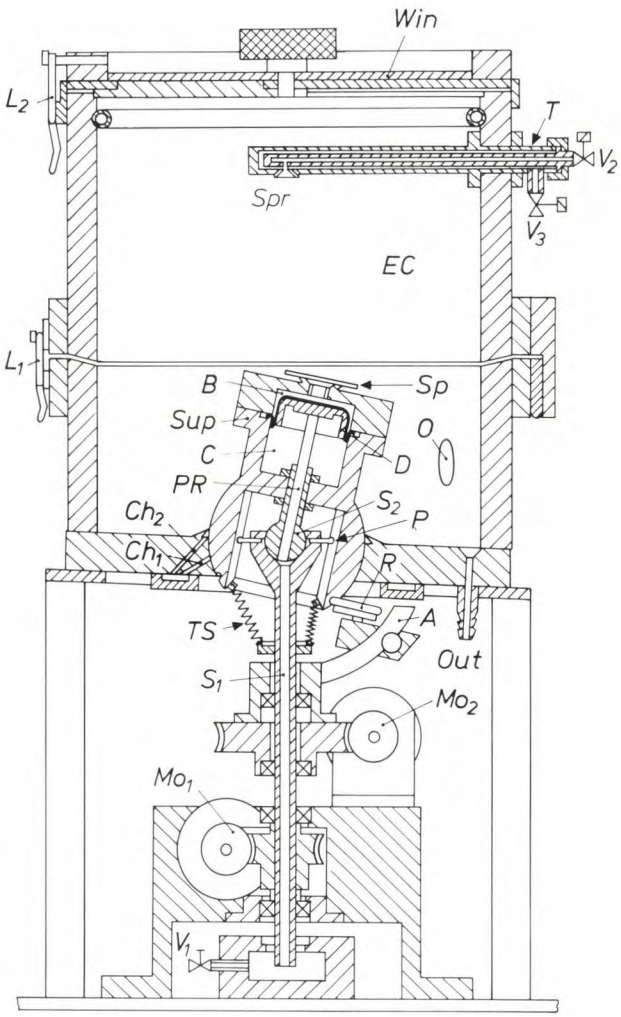


Fig. 2. Longitudinal cross-section of the etching unit. The etchant is sprayed over the specimen *Sp* by the spray head *Spr*; the specimen is mounted on a support *Sup* in the etching chamber *EC*. Both support and etching chamber are made of 'Teflon'. The specimen is held in position by suction. It can perform a rotating movement (driven by the motor Mo_1) or a tumbling movement (driven by the motor Mo_2, via the rollers *R* attached to the arc *A*), or both movements together. The bearing formed by the spherical part of *Sup* and the cup-like opening in the base of *EC* is lubricated with gas (nitrogen). The gas is supplied under pressure via the channels Ch_1 and — for sealing — via the channels Ch_2.

[*] Trade Mark, E.I. du Pont de Nemours & Co.

nected with the chamber C via a valve V_1, hollow shafts S_1 and S_2 and an open piston rod PR. An acid-resistant diaphragm D ensures that the underpressure in the space B connected with the specimen is the same as in the chamber C. The underpressure is adjusted by means of the valve V_1 so that the specimen is not bowed, which would have adverse results on the etching.

To allow the specimen support *Sup* to rotate and make tumbling movements its base is shaped like a ball, and this shape fits into a cup in the base of the etching chamber EC. A bearing is thus produced,which is found to have excellent characteristics because of the use of 'Teflon'. This bearing is lubricated with nitrogen gas at a pressure of about 3 bars, supplied via a number of channels Ch_1 that open into the hollow part of the bearing. The base is provided with a raised sloping spring-loaded sealing ring around this part of the bearing. Nitrogen is also supplied to the space sealed off by this ring, via the channels Ch_2. This prevents etching fluid from penetrating through the bearing to the region underneath it.

The rotating motion is produced by a motor Mo_1, which is connected to the shaft S_1 by a worm gear. Radial drive pins P transfer the rotating motion of the conical upper part of S_1 to the specimen support *Sup*. The outer ends of these pins, which can move freely in the radial direction, engage with an axial groove on the inside of the spherical part of the bearing. This arrangement forms a universal joint that allows the specimen support to be rotated about its own axis irrespective of its position in relation to S_1.

The tumbling motion is produced by a motor Mo_2, which causes an arc A to rotate about the axis of S_1, again by means of a worm gear. This arc carries two displaceable pressure rollers R, which engage with a ring-shaped groove in the spherical part of the bearing. Since the pressure rollers run round in this groove, the specimen support *Sup* makes a circular tumbling movement, in which the axis of the support describes an inverted cone about the vertical axis. The apex angle of this cone depends on the position selected for the rollers on the arc. A number of tension springs TS are provided to prevent backlash in the tumbling motion.

The rotating motion and the tumbling motion can be performed separately or at the same time.

The directions of rotation of Mo_1 and Mo_2 can each be reversed, so that the two motions can be arranged to be in the same or in opposite directions at will. *Fig. 3* shows a photograph of the mechanism.

As we said above, the etchant is sprayed on to the specimen *Sp*. The spray head *Spr* is attached to a coaxial tube T (see the detail diagram in *fig. 4*). The etching fluid flows through the inner tube, while nitrogen is supplied through the outer tube. The etching fluid and the nitrogen are supplied via the valves V_2 and V_3. The outward flow of nitrogen sets up an underpressure in *Spr* which automatically ensures that etchant is drawn out of a supply reservoir. The etchant is mixed with the nitrogen in the spray head, giving a homogeneous cone of spray with an apex angle of 40°. An acid-resistant pump removes the etchant via the outlet *Out* and recirculates it to the reservoir. These mechanical operations of the etching process can be continued for an unlimited time. When the etching has

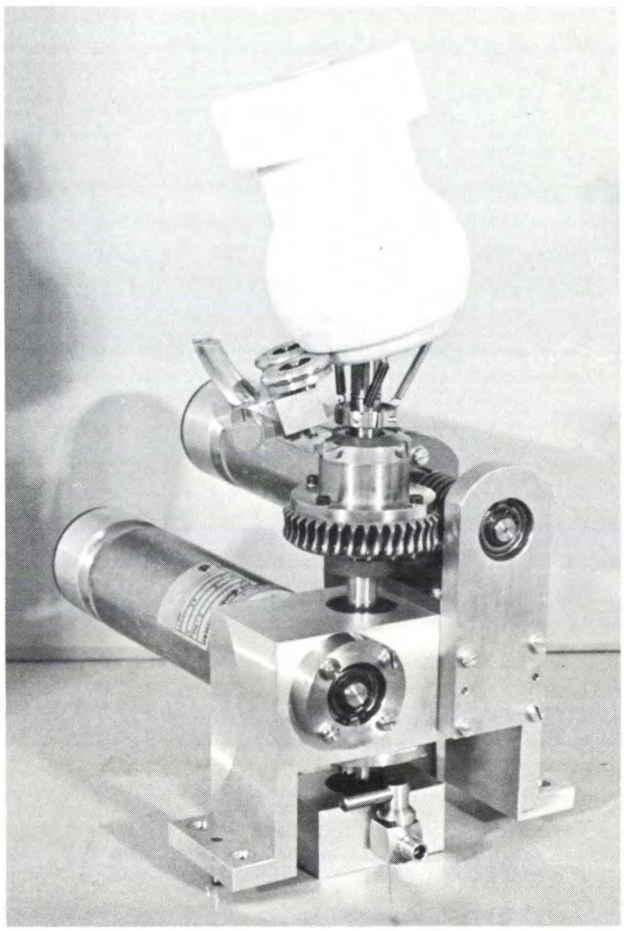


Fig. 3. Photograph of the motion mechanism. The 'Teflon' substrate support and the two motors for the rotating and tumbling motions are clearly visible.

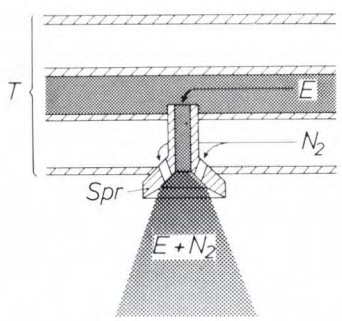


Fig. 4. The spray head (*Spr* in fig. 1) for the etchant E. A 40° cone of spray is produced. T coaxial tube. N_2 nitrogen.

been completed the valve V_2 is closed and another valve is opened for the supply of distilled water, which is used for rinsing out the etching chamber and the specimen. The cover of the etching chamber, which includes the window *Win*, can also be cleaned with distilled water by means of the ring-shaped spray head at the top of the chamber. A vent hole *O* ensures that an overpressure does not arise in the etching chamber during all these operations.

The etching chamber is constructed in two parts to make it accessible; the upper part hinges open and can be hermetically closed again by means of the locking mechanism L_1. There is a similar arrangement for the cover, which has a locking mechanism L_2.

The height, including the reservoir for the etching fluid, is 75 cm, the width 79 cm and the depth 50 cm. The unit contains the operating panel, the switching and control circuits, the motors and the pumps. Connections are also provided for water supply, 220-volt a.c. mains and nitrogen gas at a pressure of 5 bars; the nitrogen pressure is reduced where desired by means of reducing valves. The electrical circuit is designed in such a way that incorrect or unsafe operation is as far as possible prevented. For example, the motion mechanism can only be switched on when there is sufficient gas pressure for the lubrication and sealing of the bearing. Similarly, the valves for the etching and rinsing liquids can only be opened when the spray pressure is sufficiently high and the etching chamber is hermetically sealed. The programme selected for a particular etching process, as well as the parameters, can be read directly from the operating panel.

The equipment is suitable for specimens with a cross-section between 15 and 80 mm. The speed of the rotating motion can be continuously adjusted from 0 to 400 revolutions per minute, in either direction. The speed of the tumbling motion can be continuously adjusted from 0 to 200 revolutions per minute, also in either direction. The tumbling angle can be continuously adjusted from 0° to 19° with the vertical (i.e. a total of 38°). It is possible to have either separate or combined rotating and tumbling motion. The etchants most commonly used are hydrochloric, acetic, nitric, sulphuric and hydrofluoric acids. An etchant with a viscosity of 1 cP can be circulated at a rate of 3.5 litres per hour.

The experience we have gained with this equipment looks promising, and further experiments are continuing.

R. Kilian
M. Liehr

Ing. grad. R. Kilian and M. Liehr are with Philips GmbH Forschungslaboratorium Hamburg, Hamburg, West Germany.

Frequency-dividers for ultra-high frequencies

W. D. Kasperkovitz

To derive from a periodic electrical signal another periodic signal with twice the period a logic circuit is often used. At high frequencies — 1000 MHz and above — however, the transitions between the high and low voltage levels of a logic signal take up a not inconsiderable portion of the period. One way of explaining this is to say that, mainly because of the parasitic capacitances and the delay times in the transistors, the circuit is unable to pass the higher harmonics of the ideally rectangular signals. A reduction in the number of transistors and other components to the absolute minimum helps to shift this limitation as far as possible towards the higher frequencies, but even then there is still a transition region at the highest frequencies in which the operation of the circuit can be described equally well as analog or digital. The author gives a series of examples of circuits that he has reduced to an often elegant simplicity, which together form a complete progression from digital to analog. The 'travelling-wave divider' is applied in a digital tuning system for television receivers.

Introduction

In various branches of electronics there is a need for circuits that will operate as frequency-dividers at ultra-high frequencies — and even above 1000 MHz — or at any rate at frequencies approaching the transition frequency of the transistors now available (the frequency at which the current gain has fallen to unity). Such circuits would be useful in professional measurement and control equipment, digital tuning systems for television and radio and in quartz-controlled electronic watches. Each application has its own special requirements for the highest frequency to be divided, the supply voltage, the current required and the input sensitivity. The temperature range is also of significance. The degree to which the circuit can be integrated is very important. Sometimes the circuit must operate with sinusoidal input voltages even though it forms a part of a digital system; this is the case where the minimum current is required, for instance in watches.

There is no established method for translating the requirements for a particular application into a circuit. Neither is it possible in most cases to tell beforehand whether the requirements can be met. It is however relatively simple to give some general rules for the components of the circuit. For example, it seems very reasonable — though unproved — that the transition frequency of the transistors should be higher than the highest frequency to be divided, particularly for circuits of high input sensitivity. Similar semi-quantitative reasoning applies for the parasitic capacitances and inductances present in any circuit: the RC time constant associated with each parasitic capacitance and the L/R time constant for the parasitic inductance should be less than the shortest period or lowest pulse-width of the input signal. R represents the internal impedance of the circuit, which varies strongly from node to node. It is often difficult or even impossible to make an estimate of this impedance, because it depends on the state of the circuit.

In addition to keeping the number and value of the frequency-limiting parasitic elements as small as possible, it is desirable, where possible, to choose the internal impedance levels such that the maximum input frequency is not determined by a single RC or L/R time constant that dominates the circuit. Where this is the case it is often possible to increase the sensitivity or the maximum frequency by changing resistance values or bias currents.

The rules given above have to serve in the absence of a proper design procedure. They allow the designer considerable freedom in the topological arrangement of the circuit. This is more so with integrated circuits than with circuits made from discrete components, since these are interconnected by wires at least a few millimetres in length, which have an inductance of a

Dr.-Ing. W. D. Kasperkovitz is with Philips Research Laboratories, Eindhoven.

few microhenries. In integrated circuits the parasitic inductance of the interconnections is usually negligible on account of the small dimensions; but here again it is desirable to keep the number of interconnections small, because beyond a certain degree of complexity the number of crossovers increases, and so does the number of parasitic capacitances (usually of a few picofarads).

In both integrated and discrete circuits, keeping the number and value of the parasitics as low as possible thus corresponds to keeping the number of interconnections as low as possible. A small number of interconnections in general means a small number of components, particularly of components with three or more terminals. In frequency-dividers, which consist of resistors, capacitors and transistors, these are the transistors. A comparison of the parasitics in the different components also shows that the transistors should be limited in number, since the Miller capacitance and the base-emitter capacitance of the transistors are usually much more significant than the parasitics in the resistors and capacitors in both integrated circuits and those with discrete components.

A number of dividers for high frequencies will now be described. One approach that will be followed is that of the reduction in complexity, in particular in the number of transistors [1]. Another approach shows that there is no sharp distinction between digital and analog: the operation can no longer be completely described by referring to high and low logic levels, but it is essentially determined by the nature of the transitions via intermediate levels. Either approach can lead us to the 'travelling-wave divider', a cyclic circuit with a circulating periodic signal at half the frequency of the input signal. In other circuits deliberate use is made of the volatile-memory operation that is present in parasitic elements such as the base-emitter capacitance of a transistor. This gives the circuits a dynamic character and they do not necessarily operate at low frequencies.

To make the operation of the various frequency dividers easier to follow the survey will start with the description of a conventional frequency-divider for binary signals.

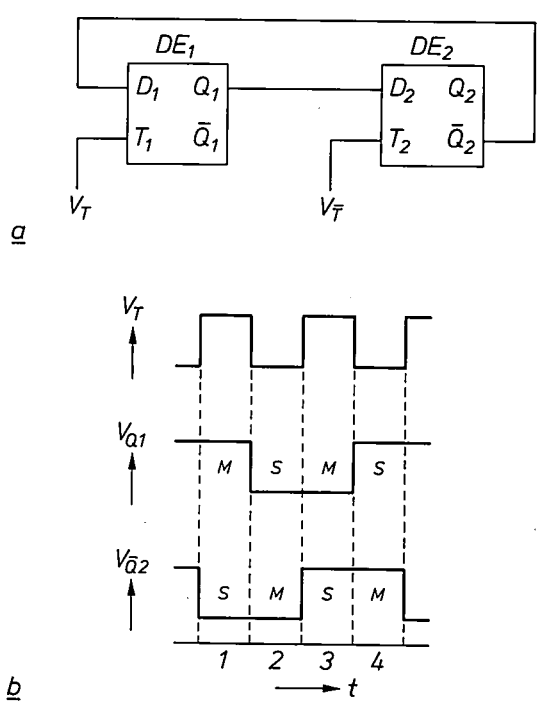


Fig. 1. *a*) Block diagram of a binary divider-by-two using the master-slave principle. This consists of two latches with single excitation DE_1 and DE_2. These are bistable when the voltage at input T is at the high level (master state), but only pass the level of input D to the output Q if the voltage at T is low (slave state). Complementary trigger voltages $V_T, V_{\bar{T}}$, which serve as input signals to the divider, are applied to the inputs T_1 and T_2. *b*) Waveforms of the trigger voltage V_T and the output voltages V_{Q1} and $V_{\bar{Q}2}$. The period of the output voltages is doubled. M master state. S slave state. t time.

Binary frequency-divider using the master-slave principle

The block diagram of a frequency-divider using the master-slave principle [2] is shown in *fig. 1a*. It consists of two latches with single excitation or D-type flip-flops — bistable circuits that can be in either the 'master state' or the 'slave state'. This depends on the voltage at the trigger input T. If this voltage is at the high level the latch is in the master state; the output voltage at Q then only depends on the logic state of the element. If the voltage at T is low, the latch is in the slave state; output Q then has the same logic level as input D.

The two latches in fig. 1*a* are triggered by two complementary voltages V_T and $V_{\bar{T}}$, so that one is always in the master state and the other in the slave state. Fig. 1*b* shows the effects of this on the output voltages V_{Q1} and $V_{\bar{Q}2}$. In the time interval *1* V_T is high and $V_{\bar{T}}$ low; latch 1 is therefore in the master state (M) and latch 2 in the slave state (S). If we now assume that Q_1 is high, then D_2 — and consequently Q_2 — is high also. The complementary output $\bar{Q}_2$ is then low, and so is D_1.

In interval *2* the roles are reversed. Latch 2 is now in the master state. It is still at the high level. Latch 1 has gone into the slave state and output Q_1 has therefore followed D_1 to the low level. This change-over took place while V_T was falling.

While $V_{\bar{T}}$ is falling — and hence V_T is rising — there is a similar changeover in latch 2, which then goes into the slave state and remains there during time interval *3*. As shown in fig. 1*b*, the reciprocal interchange of in-

[1] Improvement of performance by reduction in the number of transistors is also discussed by Z. E. Skokan, Emitter function logic — logic family for LSI, IEEE J. **SC-8**, 356-361, 1973.

[2] A. Richardson and R. C. Foss, New binary counter circuit, Electronics Letters **1**, 273, 1965.

formation continues and after two time intervals or two periods of V_T the output state is reached again. Two periods of the input signal correspond to one period of the output signals V_{Q1} and $V_{\bar{Q}2}$ and the circuit therefore works as a divider-by-two.

A more detailed circuit diagram of the divider is shown in *fig. 2*. The circuit consists entirely of differential amplifiers, and all the signals are therefore applied in complementary form. Each of the latches consists of a cross-coupled differential amplifier T_1,T_2 functioning as a bistable circuit, a buffer amplifier T_3,T_4 and an input amplifier T_5,T_6. The input amplifier switches the current I supplied by a current generator either as I_{bi} to the bistable differential amplifier T_1,T_2 or as I_s to the buffer amplifier T_3,T_4. The current goes to T_1,T_2, when V_T is high and T_5 conducts, and to T_3,T_4 when $V_{\bar{T}}$ is high and T_6 conducts. In the first case the bistable differential amplifier is activated ('enabled') and the buffer amplifier is not operating ('inhibited'); the output signal depends only on the logic state of the bistable differential amplifier, and the element is therefore in the master state. If $V_{\bar{T}}$ is high, the buffer amplifier T_3,T_4 is enabled; the signals at the inputs D and $\bar{D}$ are immediately passed on to the outputs while the bistable pair T_1,T_2 is inhibited. The latch is now in the slave state.

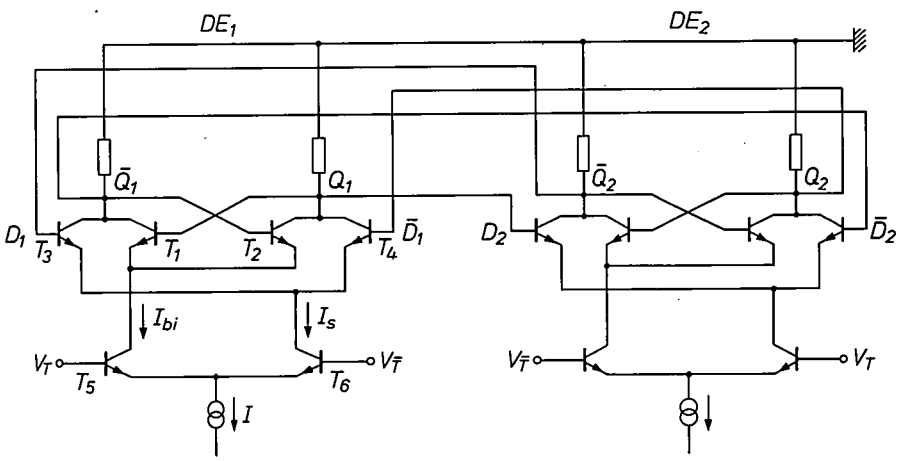


Fig. 2. Circuit of a binary divider-by-two using the master-slave principle. An output signal at half the frequency of the control signal $V_T, V_{\bar{T}}$ can be taken in complementary form from the outputs Q_1, $\bar{Q}_1$ or Q_2, $\bar{Q}_2$. Each latch circuit DE_1, DE_2 consists of a cross-coupled differential amplifier (T_1,T_2) that functions as a bistable memory element when the common emitter current I_{bi} is equal to I (master state), and a buffer differential amplifier (T_3,T_4) connected to the same collector resistances, which is enabled when $I_s = I$ (slave state).

As can be seen from fig. 2, this circuit based on the master-slave principle is rather complicated; the many interconnections between the nodes make it difficult to make a practical version with few parasitic elements. This is due in part to the rather complicated structure of the latch circuit, and also to the use of two separate latch elements that control one another alternately. However, by applying other principles it is possible to make a divide-by-two circuit with only one bistable element, and this implies a considerable simplification. One such approach will now be described. This is a dynamic circuit, where unlike the previous one deliberate use is made of the short-term memory effect present in the base-emitter capacitance of the transistors.

Dynamic frequency-divider with only one bistable element

Fig. 3a shows a dynamic divide-by-two circuit that contains only one bistable element. This is the cross-coupled differential amplifier T_1,T_2 with a common emitter current I_{bi}. The circuit also contains a differential amplifier T_3,T_4, which has negative feedback via the external base resistances R_B. Its common emitter current is I_s. Both differential amplifiers have the collector resistances R_C in common, so that the output signals are added. The trigger voltage V_T and its complement $V_{\bar{T}}$, which also form the input signal of the divider, are presented to the common emitters via the coupling capacitances C.

In the absence of the trigger voltage $I_{bi} = I$ and $I_s = I'$; I and I' are supplied by current generators. Their magnitudes are chosen such that the gain of the cross-coupled differential amplifier is higher by more than one than that of the differential amplifier with negative feedback: this gain is proportional to the emitter current through the transistors.

This condition is satisfied if

$$I - I' > \frac{2kT}{Rce},$$

where k is Boltzmann's constant, T is the absolute temperature and e the electronic charge. This can be seen from the following analysis.

The current i through the base-emitter junction of a transistor is given by

$$i = I_R\left(\exp\frac{v_{BE}e}{kT} - 1\right),$$

which approximates to

$$i \approx I_R \exp\frac{v_{BE}e}{kT},$$

where I_R is the reverse-bias current and v_{BE} the voltage between base and emitter, see *fig. 4*. If the base current is neglected, the collector current is also i. The change in collector current i with the base-emitter voltage v_{BE} is given by

$$\frac{di}{dv_{BE}} = \frac{e}{kT} I_R \exp\frac{v_{BE}e}{kT} = \frac{e}{kT} i.$$

The voltage gain A of the transistor stage is the ratio of the voltage drop v_C across the collector resistance to the base-emitter voltage:

$$A = \frac{dv_C}{dv_{BE}} = R_C \frac{di}{dv_{BE}} = \frac{Rce}{kT} i.$$

We can consider the transistor stage as half of a differential amplifier. In fig. 3 the half of a differential amplifier has a bias current of $\frac{1}{2}I$ or $\frac{1}{2}I'$.

If the bistable pair in fig. 3 is to have a gain A_{bi} that is greater by more than one than A_s for the pair with negative feedback, then this leads to the condition

$$\frac{Rce}{kT}(\tfrac{1}{2}I - \tfrac{1}{2}I') > 1,$$

which corresponds to the relation stated at the beginning.

With the bias-current settings indicated above, and in the absence of a trigger signal, the gain of the two coupled amplifiers is greater than unity and the circuit is therefore bistable.

A general picture of the voltage variations is shown in fig. 3*b*. The transition of the output voltage $V_Q - V_{\bar{Q}}$ from one level to the other always occurs at the trailing edge of V_T. In the master-slave principle of fig. 1 the switching process is actuated by a second bistable circuit, which does not change state during the trailing edge. The function of this second circuit is taken over here by the base-emitter capacitances of the transistors T_3,T_4; the state previous to the trailing edge is stored temporarily in these capacitances [3].

The effect of these capacitances will be shown in a detailed description of the operation with the aid of *fig. 5*. The time axis is expanded in this figure; fig. 5*a* shows the trailing edge of the trigger signal. During the falling edge of the trigger signal the charging and discharging currents of the coupling capacitors C are added to I_{bi} and I_s; the coupling capacitors are of the same order of magnitude as the base-emitter capaci-

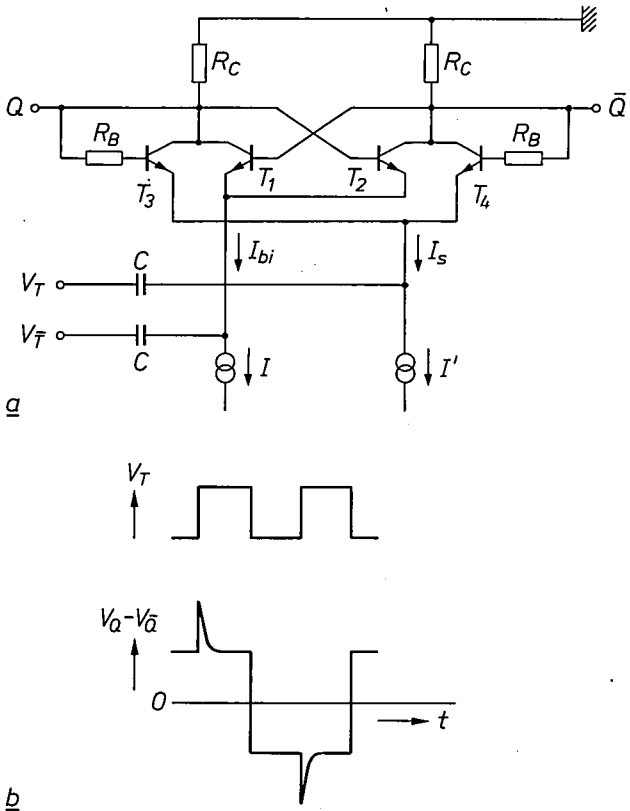


Fig. 3. *a*) Dynamic divider-by-two with a single bistable circuit instead of two. The bistable circuit is the cross-coupled differential amplifier T_1,T_2. The differential amplifier T_3,T_4 has negative feedback via the base resistances R_B. The base-emitter capacitances of T_3,T_4, in conjunction with the resistances R_B, function as a dynamic memory, which makes a second bistable circuit unnecessary. I' is so much smaller than I that the gain of the combination of the two differential amplifiers, one cross-coupled and the other with negative feedback, is greater than 1. *b*) Waveforms of the trigger voltage V_T and the output voltage $V_Q - V_{\bar{Q}}$. The output is at half the frequency of V_T and the voltage only changes polarity at the falling edge of V_T. At the rising edge a peak appears in the output voltage; this peak decays exponentially with the time constant of the base-emitter circuit of T_3,T_4.

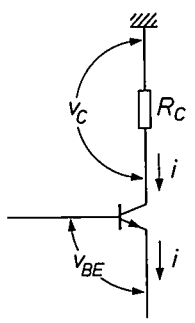


Fig. 4. Voltage gain of a transistor stage. v_{BE} base-emitter voltage. v_C collector voltage. i collector current and emitter current, taken as equal here, with the base current neglected. R_C collector resistance.

[3] In H. H. Muller and W. J. Wu, A GHz IC counter using Isoplanar II, 1974 IEEE Int. Solid-State Circuits Conf. Dig. tech. Papers, pp. 20-21, a fast divider-by-two that only uses a single bistable circuit is described. This circuit operates in the 'race mode', in which the output from Q is inverted in a separate amplifier, then coupled directly to D to provide feedback. The inverter stage also functions here as a short-term memory. This circuit contains about three times as many transistors as the one in fig. 3*a*, partly because it was necessary to derive pulses of fixed width from the trigger signal.

tance of T_3 and T_4. The current I_{b1} becomes smaller because of this superposition, and I_s becomes larger by the same amount. The effect of this is that the condition that the gain in the bistable pair should be greater by one than the gain in the pair with negative feedback is temporarily no longer satisfied, and the circuit is temporarily monostable instead of bistable (fig. 5*b*).

Let us assume that in the initial state before switching $I_{b1} = I$ flows through T_2 and $I_s = I'$ through T_3; the voltages at the outputs are then $V_Q = -I'R_C$ and $V_{\bar{Q}} = -IR_C$ (fig. 5*c*). The addition of the charging and discharging currents from the coupling capacitors C produces immediate and corresponding changes in the collector currents. The collector current of T_2 falls and $V_{\bar{Q}}$ consequently rises; the current through T_3 increases and consequently V_Q falls. Because of the increase in $V_{\bar{Q}}$ transistor T_1 starts to conduct, but after a small delay owing to the presence of the base-emitter capacitance. The fall in V_Q will stop T_3 from conducting, but here the presence of R_B is responsible for a much longer time constant and hence a much longer delay. For the moment T_3 remains conducting, so that the sum of I and I' flows in the left-hand collector resistance and V_Q assumes a minimum value. In a similar way T_4 still remains non-conducting for a short time interval, so that no current flows in the right-hand collector resistance and $V_{\bar{Q}}$ reaches a maximum.

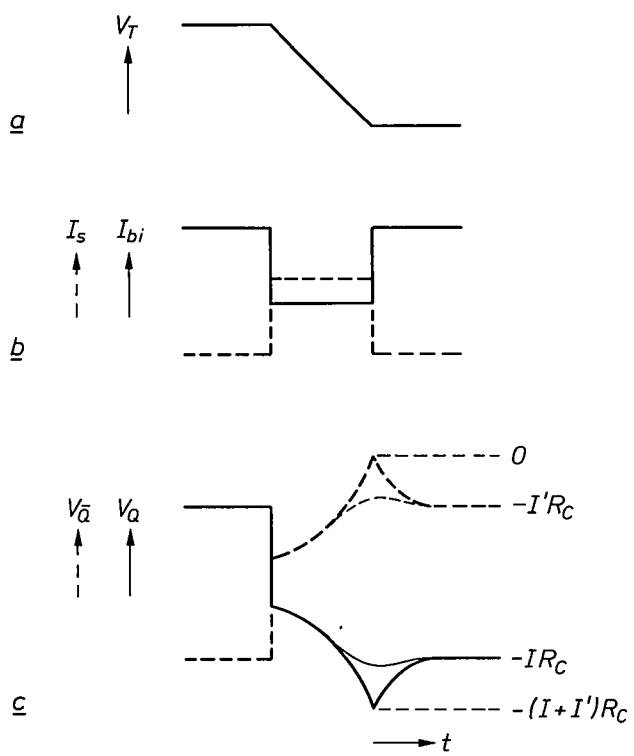


Fig. 5. Diagram showing the variation of the currents and voltages in the dynamic divider-by-two (fig. 3) during the falling edge of the trigger voltage V_T. The time axis has been greatly expanded. *a*) Falling edge of V_T. *b*) Variation of the common emitter currents I_{b1} and I_s. For a short time I_{b1} is smaller than I_s and the circuit is not dominated by the bistable differential amplifier; it is then monostable. *c*) Variation of the collector voltages V_Q and $V_{\bar{Q}}$ when it is assumed that the collector currents of T_3 and T_4 have not yet changed. Since the current in T_2 changes over to T_1 and T_4 was not conducting, the current through the right-hand collector resistance R_C is zero for a short time and $V_{\bar{Q}}$ reaches the value 0. The pattern of events for V_Q is the mirror image of this. In fact the switching of the current I_s from T_3 to T_4 starts immediately, although it is slower, and the actual voltage variation is shown by the thin lines.

After a delay, however, the bases of T_3 and T_4 go to their new potential. This is higher for T_4 than for T_3, so that I_s eventually passes through T_4. This makes the situation exactly the opposite of that at the start.

In fact the base-emitter voltages of T_3 and T_4 really begin to change to their new values right at the start of the switching time. This causes the peaks in fig. 5*c* to be rounded off, as shown by the thin lines.

It has been shown above how, starting from a divider-by-two based on the master-slave principle, a simplification can be made by leaving out one of the bistable circuits and making use of dynamic storage of information. However, this is not the only simplification possible. The next section shows how a simplification can also be obtained, while retaining the master-slave principle and static operation, by changing the coupling between the two bistable circuits.

Static frequency-divider with negative feedback

Fig. 6 shows a divider-by-two that is based on the master-slave principle like the circuit in fig. 2, but has a fixed coupling between the two bistable circuits that is independent of the trigger signal [4]. The bistable elements here are the cross-coupled differential amplifiers T_1,T_2 and T_3,T_4; these amplifiers are enabled in turn at a periodic trigger signal V_T by the collector currents I_{b112} and I_{b134} of the differential amplifier T_7,T_8. This brings the bistable circuits T_1,T_2 and T_3,T_4 alternately into the master or the slave state. 'Slave state' does not mean quite the same here as in the circuit of fig. 2, since there are no latches. In the slave state of the transistor pair T_3,T_4 the collector voltages are simply the collector voltages of T_1,T_2 reduced by a voltage divider.

The two bistable elements are coupled symmetrically via the resistances R_{CC}; both elements affect one another in the same way through this path. There is also a directional coupling from the pair T_3,T_4 to T_1,T_2 via differential amplifier T_5,T_6 with common emitter current I_s. Seen from T_1,T_2 this directional coupling looks like a negative feedback from the output signal ΔV_{12}, in which the feedback signal is also dependent on the state of T_3,T_4.

The two forms of coupling work in the opposite sense. In schematic terms the symmetrical coupling with R_{CC} tries to make the polarity of the output

voltages ΔV_{12} and ΔV_{34} the same, while the directional coupling via differential amplifier T_5,T_6 tries to force the pair T_1,T_2 into a state in which the output voltage ΔV_{12} has the opposite polarity to ΔV_{34}. The special feature of this configuration is that the asymmetrical coupling between T_3,T_4 and T_1,T_2 brings the differential amplifier T_5,T_6 to a state such that it helps to make T_3,T_4 change over when this pair goes into the slave state.

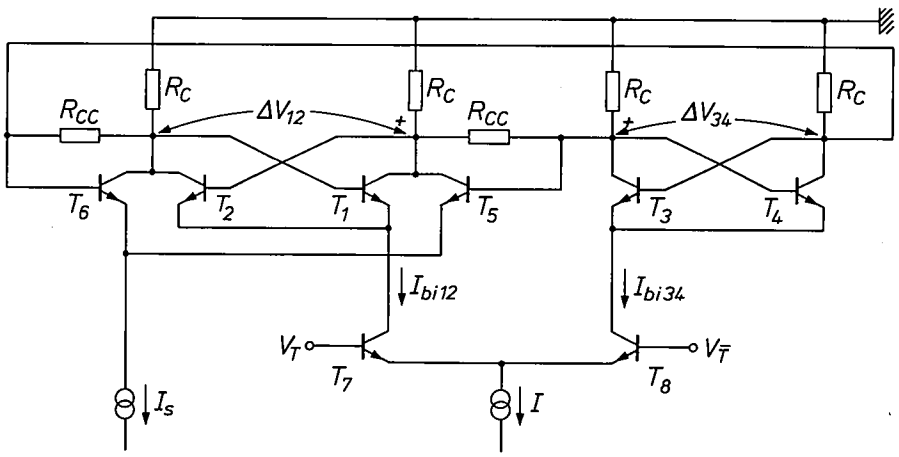


Fig. 6. Simplified static divider using the master-slave principle with two bistable differential amplifiers T_1,T_2 and T_3,T_4. The amplifiers are symmetrically coupled via the resistances R_{CC}. Further coupling from T_3,T_4 to T_1,T_2 is provided via differential amplifier T_5,T_6. The circuit has fewer transistors than the one in fig. 2 and has better high-frequency performance.

The voltage variation is shown in *fig. 7*. When the trigger voltage V_T has the low logic level, $I_{bi12} = 0$ and $I_{bi34} = I$, so that T_1,T_2 are in the slave state and T_3,T_4 are in the master state. Let us assume that in the starting condition ΔV_{34} is negative. Because of the effect of the directional coupling via T_5,T_6 the voltage ΔV_{12} is then positive. This situation is stable until V_T starts to rise. I_{bi34} is then reduced to zero and I_{bi12} increases from zero to I. While T_6 was already in the conducting state, T_2 now starts to conduct, so that ΔV_{12} increases. The rapid increase in ΔV_{12} to a high value makes the bistable pair T_3,T_4 switch over quickly, before I_{bi34} has quite fallen to zero. However, once this has happened the differential amplifier T_5,T_6, which reacts with some delay owing to the base-emitter capacitances, also switches over. Finally, T_5 and T_2 become conducting. The difference voltage ΔV_{12} decreases at this last changeover, but remains positive.

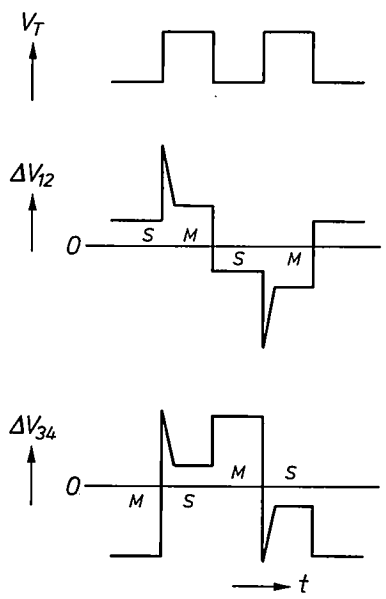


Fig. 7. Voltage waveforms in the simplified divide-by-two circuit (fig. 6). The slightly delayed reaction of differential amplifier T_5,T_6 is the cause of the peaks in the voltage ΔV_{12}; these peaks help the bistable pair T_3,T_4 to switch over quickly.

At the falling edge of V_T, I_{bi12} decreases. When I_{bi12} has fallen to about I_s the left-hand half of the circuit becomes monostable. The transistor pair T_1,T_2 now switches to a state determined by differential amplifier T_5,T_6 and hence opposite to that of T_3,T_4.

The circuit described here has been made up experimentally with BFR92 microwave transistors on a ceramic substrate with evaporated resistors. The sensitivity to sinusoidal trigger signals has been measured as a function of frequency for this experimental version; here again the trigger signal is the input signal of the divider. The results are shown in *fig. 8*. The increase with frequency of the smallest input amplitude for correct operation is partly due to parasitic elements, but is of course also related to the transition frequency of the transistors, which was between 2 GHz and 3 GHz for the bias current used. The circuit also works at low

[4] W. D. Kasperkovitz, U.S. Patent No. 3 996 478.

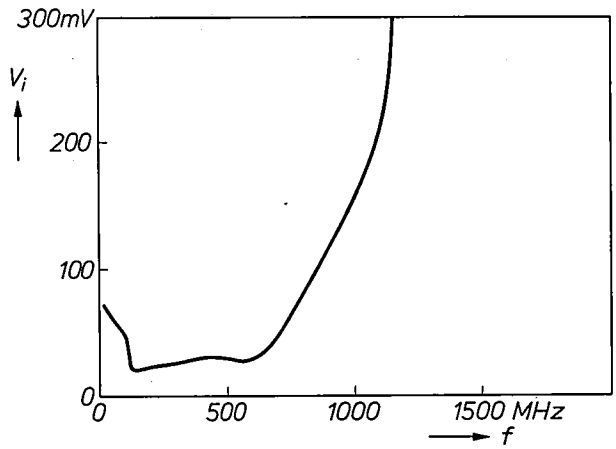


Fig. 8. The minimum input voltage V_i for correct operation of the simplified static divider-by-two, as a function of the input frequency f. The curve was measured for an experimental circuit using BFR92 microwave transistors. The rise of the curve at the higher frequencies is due to the parasitic elements in the circuit but is also connected with the transition frequency f_T of the transistors, which is between 2 GHz and 3 GHz for the operating point used.

Like the logic-circuit frequency-divider with two latches, the divider with negative feedback described here is based on two bistable circuits that are alternately in the master and the slave state. If circuit 1 is in the master state, circuit 2 is consequently brought to a slave state of similar polarity, but if circuit 2 is in the master state, circuit 1 is brought to a slave state of opposite polarity. Both frequency-dividers therefore possess an essential asymmetry in the coupling between the two bistable circuits, and this asymmetry is particularly noticeable in the divider with negative feedback. However, in optimizing the many parameters of such a circuit it is particularly desirable to make the circuit as symmetrical as possible. In general an increase in the symmetry not only reduces the number of free parameters, but also gives more certainty that the maximum input frequency will not be determined by a single RC or L/R time constant and could therefore be increased by a better choice of components.

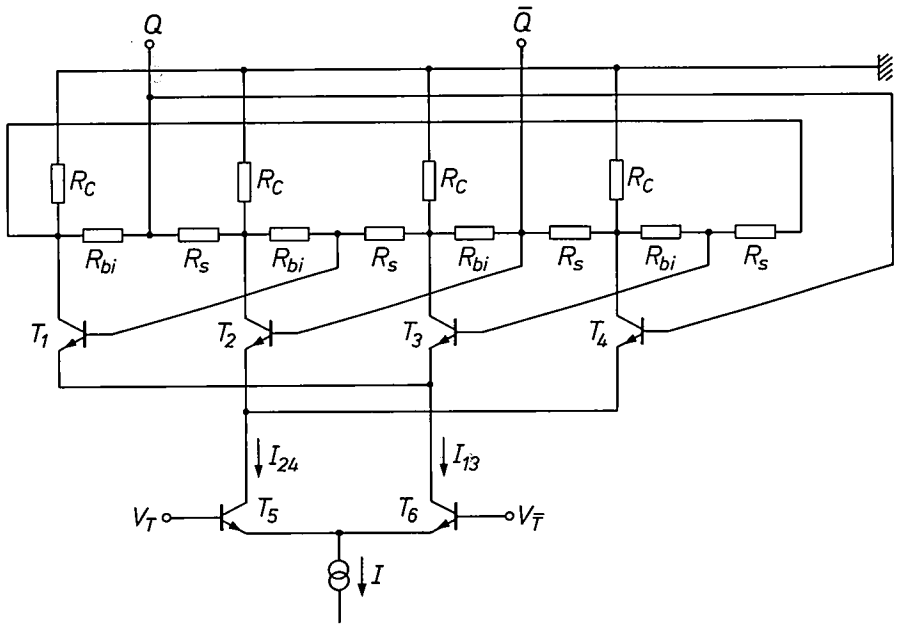


Fig. 9. Travelling-wave divider. The two differential amplifiers T_1,T_3 and T_2,T_4 are 'interwoven' and are coupled by the resistances R_{bi} and R_s. The circuit forms a closed-loop structure in which there is a circulating collector-voltage wave that has half the frequency of the trigger signal $V_T, V_{\bar{T}}$.

frequencies, as would be expected for a static circuit that alternates between stable states. The small increase in input signal required on going towards the lower frequencies can be explained by the decrease of the change in voltage with time in a sinusoidal input voltage of decreasing frequency. Since the operation of the circuit is in part dependant on certain built-in time constants, too slow a change in the currents I_{bi12} and I_{bi34} results in an incorrect transfer of information. This can be prevented by making the amplitude of the sinusoidal signal larger on decreasing the frequency.

In the next section a frequency-divider is described that has a maximum of symmetry with a drastic reduction in the number of transistors.

Divider-by-two based on the travelling-wave principle

The principle on which the frequency-divider shown in *fig. 9* depends can be called the travelling-wave principle [5]. The circuit consists of three differential amplifiers: the trigger or input amplifier T_5,T_6 and two differential amplifiers T_1,T_3 and T_2,T_4 with complemen-

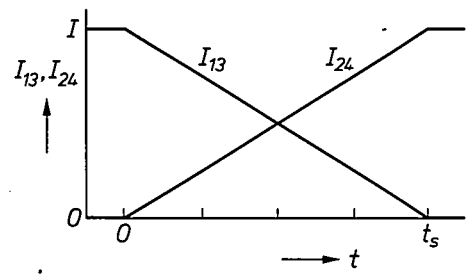


Fig. 10. In the explanation of the operation of the travelling-wave divider (fig. 9) it is assumed that the common emitter currents I_{13} and I_{24} decrease linearly from I to zero and increase linearly from 0 to I during the time t_s associated with the switching.

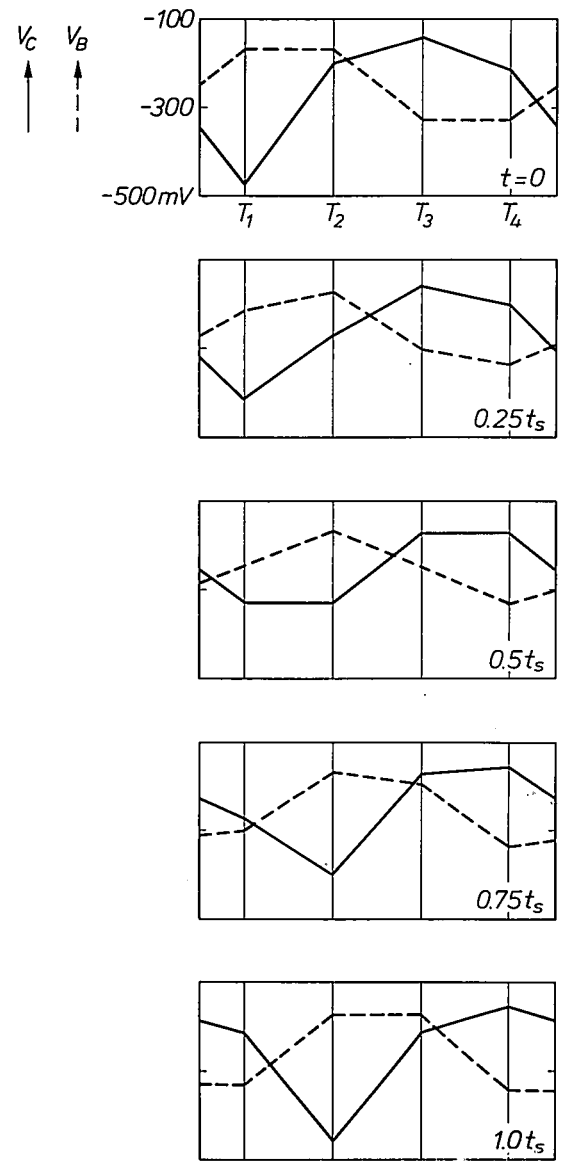


Fig. 11. The collector voltage V_C and the base voltage V_B of the transistors T_1, T_2, T_3 and T_4 of the travelling-wave divider during one step of the switching process. After four steps the starting position is reached again. R_{b1} and R_s are assumed equal. *a*) Start of the switching process; $t = 0$. The total current I flows through T_1. *b*) After a quarter of the switching time; $t = 0.25\ t_s$. T_2 starts to take over the current. *c*) $t = 0.5\ t_s$. A current $\frac{1}{2}I$ flows in T_1 and T_2. *d*) $t = 0.75\ t_s$. The greater part of the current has transferred to T_2; T_2 also has the highest base voltage. *e*) End of the switching step; $t = t_s$. The total current I flows through T_2. Transistor T_3 has the same base voltage as T_2 and is thus prepared to take over the current in its turn.

tary control. The collector and base terminals of transistors T_1 to T_4 are connected to one another in a cyclic manner via coupling resistances R_{b1} and R_s.

The operation of the circuit shown here can be described by considering it as two closely coupled and 'interwoven' bistable differential amplifiers that are brought alternately to a master and a slave state. However, there are advantages in using a description that puts less emphasis on the digital switching between master and slave states, but pays more attention to the smooth transitions of the currents in the various branches of the circuit. This will show more clearly that the circuit is partly analog in nature.

Since the circuit has a cyclic configuration it is only necessary to describe the operation during a quarter of the period of the output voltage $V_Q - V_{\bar{Q}}$, and to discuss the variation of the collector and base voltages during this quarter-period. The voltages during the remaining three-quarters of the period can be derived from the voltages during the first quarter by cyclic interchange of the numbers of transistors T_1 to T_4.

In many cases the input signal to the circuit will be a sinusoidal control voltage. To simplify the description of the operation, however, the variation of the drive currents I_{13} and I_{24} during the switching of differential amplifiers T_5,T_6 will be approximated by a linear variation (*fig. 10*). The switching time t_s is determined by the rise time of V_T.

Fig. 11 shows the collector and base voltages of the transistors T_1 to T_4 at five successive instants during the switching time; the length of time between the successive 'snapshots' is always $\frac{1}{4}t_s$. The following values are assumed: $R_C = 100\ \Omega$, $R_{b1} = R_s = 50\ \Omega$, $I = 10$ mA.

Fig. 11*a* illustrates the situation at the start of the switching process: $t = 0$, $I_{13} = I$, $I_{24} = 0$ (see fig. 10). It is assumed that I_{13} flows only through T_1. The greater part of the collector current of T_1 flows through its collector resistance R_C. A certain fraction flows via the coupling resistances R_{b1} and R_s through the collector resistances of T_2 and T_4, however, and an even smaller fraction through the collector resistance of T_3. The result is that T_1 has a low collector voltage (about -450 mV), T_3 a high collector voltage (about -150 mV) and T_2 and T_4 intermediate collector voltages (about -200 mV).

The points representing these voltages in fig. 11*a* are connected by straight lines, which can be considered as the representation of the voltage drop across the coupling resistances R_{b1} and R_s. Since these are assumed to be equal, the base voltages of T_1 to T_4 are half-way

[5] D. Kasperkovitz and D. Grenier, Travelling-wave dividers: a new concept for frequency division, Microelectronics and Reliability **16**, 127-134, 1977 (No. 2).

between the collector voltages. The points representing the base voltages in fig. 11 are joined by dashed lines.

At $t = 0.25\ t_s$ (fig. 11*b*) $I_{13} = 0.75\ I$ and $I_{24} = 0.25\ I$. The current I flows only through T_2, since this transistor has a higher base voltage than T_4. The collector voltage of T_2 has therefore decreased. The collector voltage of T_1 has increased because of the decrease in I_{13}.

Half-way through the switching time ($t = 0.5\ t_s$) both drive currents are equal ($I_{13} = I_{24} = 0.5\ I$). I_{13} still flows only through T_1 and I_{24} through T_2 giving equal base voltages for T_1 and T_3 (fig. 11*c*). The fact that I_{13} nevertheless still continues to flow only through T_1 can be explained by the rapidity of the switching process; the transistors have an internal switching time of the same order of magnitude as t_s.

Fig. 11*d* shows the situation at $t = 0.75\ t_s$. It has been assumed here that the greatly reduced common emitter current $I_{13} = 0.25\ I$ of differential amplifier T_1,T_3 still continues to flow through T_1, although the base voltage of T_1 is about 100 mV lower than the base voltage of T_3. Such a situation only occurs for very short switching times t_s. If t_s is of the same order of magnitude as the internal switching time of the transistors, some of the common emitter current I_{13} switches from T_1 to T_3. This decreases the collector voltage of T_3, and increases the collector voltage of T_1, by the same amount. The result of this is an increase in the base voltage of T_4 and a decrease in the base voltage of T_2. This makes the conditions less favourable for conduction of the current I_{24} through T_2 only, which should after all be the result at the end of the switching process. We have therefore reached the most critical moment in the transfer of the information. If the switching time t_s is greater than a critical switching time t_{sc}, the base voltages of T_2 and T_4 will approach one another too closely and the circuit will no longer work properly as a divider-by-two.

Fig. 11*e* shows the situation at the end of the switching time t_s. Current $I_{24} = I$ flows only through T_2. The base voltage of T_2 is 150 mV higher than that of T_4; the situation is therefore stable.

The curves in fig. 11 can be considered as a collector-voltage wave and a base-voltage wave; in a single switching time they travel one position to the right. After four switching times, i.e. two periods of the trigger voltage V_T, they will have gone through one cycle; the circuit therefore works as a divider-by-two.

As can be seen from fig. 11*a* the two differential amplifiers T_1,T_3 and T_2,T_4, which together form the circuit, have the same input voltages ($V_{B1} - V_{B3} = V_{B2} - V_{B4}$) at the start of the switching process. This comes about because of the assumption that R_{b1} and R_s were equal. However, for optimum circuit behaviour

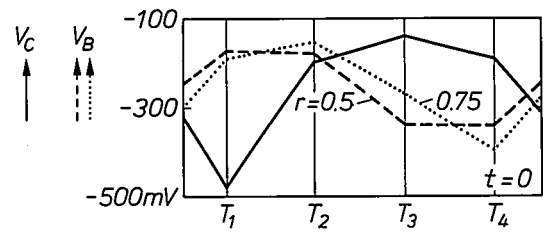


Fig. 12. The base voltages at the time $t = 0$ for two values of $r = R_s/(R_{b1} + R_s)$. For higher r the input voltage $V_{B2} - V_{B4}$ of differential amplifier T_2,T_4 is higher, so that the current transfers more readily from T_1 to T_2; the switching is therefore faster.

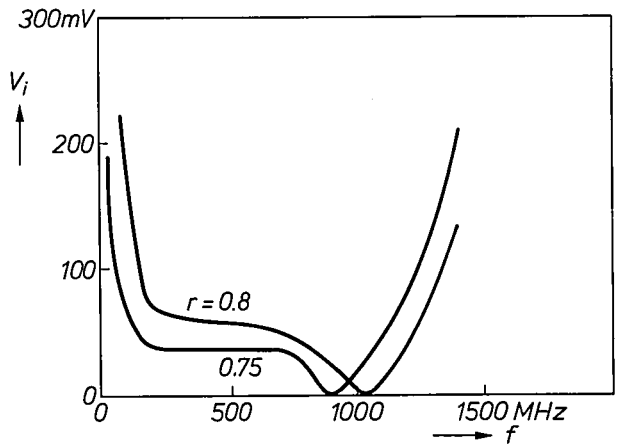


Fig. 13. The minimum input voltage V_i for correct operation of the travelling-wave divider, as a function of the frequency f. The curves were measured for two integrated versions with $r = 0.75$ and $r = 0.8$. There is a distinct minimum in the curves at the frequency of maximum sensitivity; for $r = 0.8$ the minimum is at a higher frequency but the sensitivity at low frequencies is reduced.

the two input voltages should be different. The input voltage of differential amplifier T_1,T_3 should be as small as possible because it must change polarity during t_s. On the other hand the input voltage of differential amplifier T_2,T_4 should be as large as possible, since the common emitter current I_{24} must eventually flow through T_2 only.

Both requirements can be satisfied by changing the ratio of R_{b1} to R_s. In the example discussed these two resistances were the same, so that $r = R_s/(R_{b1} + R_s) = 0.5$.

Fig. 12 shows how the base voltages in the starting condition at $t = 0$ change if r is made equal to 0.75 while the sum $R_{b1} + R_s$ remains constant. The result is twofold: the input voltage $V_{B1} - V_{B3}$ of differential amplifier T_1,T_3, which passes current, has been reduced from 150 mV to 90 mV, whereas the input voltage $V_{B2} - V_{B4}$ of differential amplifier T_2,T_4, which does not pass current, has increased from 150 mV to 260 mV. The decrease in the input voltage of differential amplifier T_1,T_3 and the increase in the input voltage of differential amplifier T_2,T_4 causes both differential ampli-

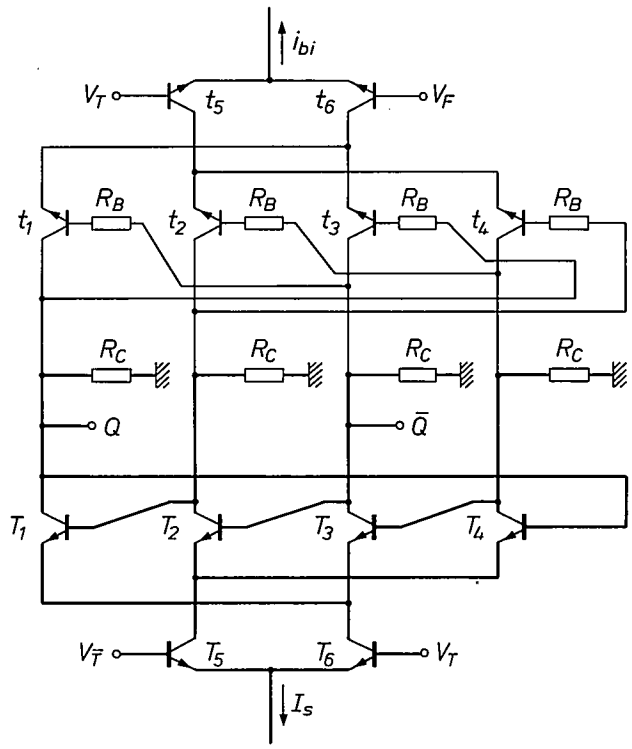


Fig. 14. Adaptive travelling-wave divider. Two travelling-wave dividers are connected to the same collector resistances R_C and operate in parallel. The upper divider (t_1 to t_6) represents the limiting case $r = 0$ and is optimum for low frequencies; the lower one (T_1 to T_6) represents the limiting case $r = 1$ and is optimum for high frequencies. The upper travelling-wave divider consists of two separate cross-coupled differential amplifiers. The two common emitter currents i_{bi} and I_s, and hence the share of each half in the operation of the complete circuit, are independently adjusted. This allows the band of highest sensitivity to be shifted along the frequency scale.

fiers to switch more rapidly to their new states. An increase in r thus gives an increase in the input sensitivity at high frequencies and an increase in the highest divisible frequency.

On the other hand the danger of an incorrect transfer of information with the drive currents changing slowly becomes greater because of the reduction of the input voltage $V_{B1} - V_{B3}$ of differential amplifier T_1,T_3. There is a critical switching time t_{sc} that must not be exceeded. This critical switching time is made shorter by the increase of r. At low frequencies the amplitude of sinusoidal input voltages has to be increased; the circuit is thus less sensitive at low frequencies.

The sensitivity to sinusoidal input voltages has been measured for two versions of the travelling-wave frequency-divider described here, both made in integrated-circuit form for application in a digital tuning system for television receivers. In one version r was 0.75, and in the other r was 0.8. It can be seen from *fig. 13* that the frequency responses of the two versions are appreciably different; the small change in r has shifted the maximum sensitivity from 850 MHz to 1050 MHz. At the same time the sensitivity at low frequencies has decreased.

A comparison of fig. 13 with fig. 8 shows that the sensitivity of the travelling-wave divider at high frequencies is considerably higher than that of the simplified static divider-by-two. However, this advantage is obtained at the expense of a lower sensitivity at lower frequencies. The time a signal takes to travel around the cyclic structure determines a preferred frequency to which the travelling-wave divider can be considered to be tuned (see the minima in the curves of fig. 13). The circuit does in fact tend to oscillate at these frequencies, but oscillation is impossible while a trigger signal of sufficient strength is present; the divider then follows this trigger signal.

Adaptive travelling-wave divider

In the travelling-wave divider just described the parameter r allows the designer to locate the maximum-sensitivity part of the response anywhere within a wide range. This choice, which also determines the sensitivity at low frequencies, cannot be altered once the circuit has been integrated. However, the ability to alter the sensitivity would be very useful in a number of applications. If this variation in sensitivity automatically followed the input frequency it would be possible to obtain a high sensitivity over a wide frequency band and at the same time to increase the maximum input frequency.

A circuit whose frequency response can be varied by making an external adjustment of currents is shown in *fig. 14*. It consists of a combination of the two limiting cases of the travelling-wave divider; the lower half in the figure is the optimum limiting case for high frequencies with $r = 1$, the upper half is the low-frequency limiting case with $r = 0$. The upper half consists of nothing more than two bistable cross-coupled differential amplifiers. Both halves have the collector resistances in common and are coupled in this way. The emitter currents i_{bi} of the upper half and I_s of the lower half can be independently adjusted.

The collector currents and base voltages of the eight transistors are plotted in *fig. 15* in a similar way to that used in fig. 11 for the simple travelling-wave divider. Here, however, the collector currents are plotted instead of the collector voltages, because these are always the same for each pair of transistors. Fig. 15*a* shows the situation in which V_T is high, so that T_6 and t_5 are conducting. I_s flows only through T_1, i_{bi} through t_4. In addition, the two common emitter currents I_s and i_{bi} are assumed to be equal. In this condition, and considering the centre of the flat top of the base-voltage wave as its maximum, the base-voltage wave is 45°

ahead in phase of the collector-current wave. From fig. 11 it can be seen that the same phase difference arises in a simple travelling-wave divider if the ratio of the coupling resistances is $r = 0.5$; this is another situation midway between $r = 0$ and $r = 1$, though in another way.

As was shown earlier, increasing r gives an increase in the frequency for maximum sensitivity. An increase in r can be brought about by reducing the share of the half with $r = 0$ in the behaviour of the complete circuit. Fig. 15*b* shows a situation in which $i_{bi} = 0.25\, I_s$. The phase difference has increased from 45° to about 70°.

Besides the facility of adjustment by current variation, the circuit includes a mechanism that gives a further automatic increase in the phase advance of the base-voltage wave with respect to the collector-current wave, provided that the input frequency is high enough. This frequency-dependent mechanism depends on the base resistances R_B in the upper half of the circuit. The resistances R_B are either added to the transistors t_1 to t_4, or, in an integrated version, included in them by a suitable choice of their surface-area ratios.

The base resistances together with the base-emitter capacitances of the transistors cause a delay in the signal. At a sufficiently high frequency this delay interferes with the cyclic switching of the common emitter current i_{bi} from t_1 to t_2, etc. Instead, each transistor takes a quarter of i_{bi}, so that the collector-current wave of the upper half degenerates to a straight line. The collector voltages and hence the base voltages of the complete circuit are now determined by the lower half only, the limiting case of a travelling-wave divider with $r = 1$. In this situation the phase difference between base voltage and collector current has increased to 90°.

The minimum input signal required for an integrated adaptive travelling-wave divider has been measured as a function of frequency. Two curves are shown in *fig. 16*, one for $i_{bi} = 0.5\, I_s$ and one for $i_{bi} = 0.2\, I_s$. As might be expected, the curve for $i_{bi} = 0.2\, I_s$ has a minimum at higher frequencies, but gives a lower sensitivity at low frequencies.

The transistors T_1 to T_8 in this integrated circuit have two emitter fingers and three base contacts, while the transistors t_1 to t_8 have only one base contact and one emitter contact of the minimum permitted length. With this arrangement, not only is the major part of the base resistances R_B integrated in a compact way, but at the same time the parasitic load on the collector contacts of T_1 to T_4 is kept to a minimum because the collector-base and collector-emitter capacitances of t_1 to t_4 are kept small.

An examination of the frequency-dividers that have been discussed so far clearly reveals a gradual transition from purely digital circuits with the associated digital input and output signals (fig. 2, fig. 3) to circuits with

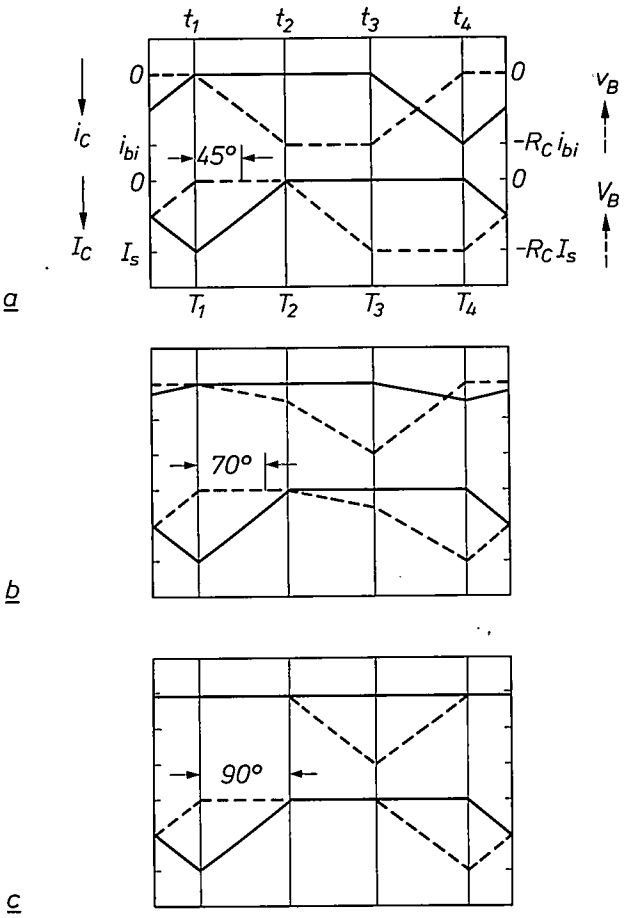


Fig. 15. Collector currents and base voltages in the adaptive travelling-wave divider. i_C collector current, v_B base voltage of transistors t_1 to t_4 (upper half of the circuit in fig. 14); I_C collector current, V_B base voltage of transistors T_1 to T_4 (lower half of circuit). *a*) Bias currents equal, $i_{bi} = I_s$. The base-voltage wave is 45° ahead of the collector-current wave in phase. *b*) $i_{bi} = 0.25\, I_s$. The base-voltage wave is about 70° ahead of the collector-current wave; the circuit is faster. *c*) The upper, slower half of the circuit cannot follow at higher frequencies. The four collector currents then all assume the same value, $\frac{1}{4} i_{bi}$. The resultant base-voltage wave is 90° ahead of the lower half of the circuit in phase.

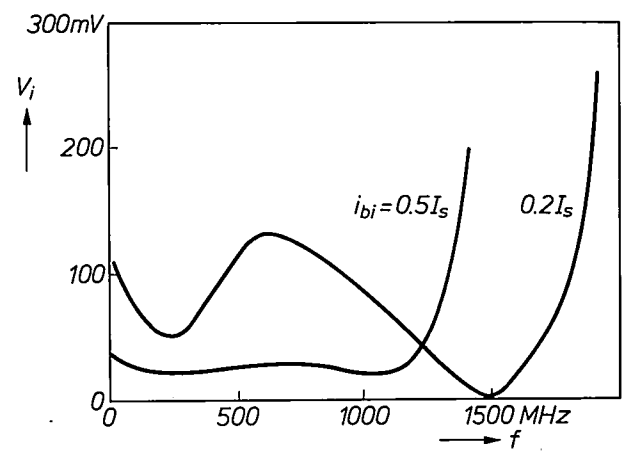


Fig. 16. The minimum input voltage V_i for correct operation of the adaptive travelling-wave divider as a function of the frequency f. The two curves were measured for an integrated circuit. The curve for $i_{bi} = 0.2\, I_s$ has a higher sensitivity at high frequencies and a lower sensitivity at low frequencies than the curve for $i_{bi} = 0.5\, I_s$.

input voltages that are sinusoidal and output voltages that rather resemble sine waves (fig. 9, fig. 14). An extrapolation of this trend suggests that it might well be possible to complete the pattern by designing frequency-dividers with sinusoidal input and output signals.

Dynamic frequency-divider with mixer stage

The block diagram of a frequency-divider for sinusoidal signals is shown in *fig. 17a*. The circuit includes a mixer stage, in which the input signal is multiplied by a 'pump signal' at twice the frequency. This is followed by a wide-band amplifier and a lowpass filter. There is a closed loop in this arrangement; the output signal from the lowpass filter, which is also the output signal of the divider-by-two, serves as the input signal for the mixer.

Fig. 17*b* shows the waveforms. The multiplication of the pump signal, which we shall represent by $\cos 2\omega t$, and the fundamental $\cos \omega t$ of the input signal produces the signal $\frac{1}{2}(\cos \omega t + \cos 3\omega t)$. The first term has the desired half frequency. This signal is now amplified, by a factor of two in the figure. The production of higher harmonics as a result of distortion must be prevented here. The lowpass filter attenuates the unwanted component at angular frequency 3ω; it has been assumed for the figure that the attenuation of the filter is zero for signals at frequencies up to ω and is proportional to the square of the frequency for signals at higher frequencies (the response has a slope of -12 dB per octave above ω).

The output signal is now fed back to the input of the mixer. The attenuated third harmonic in this signal gives rise to a fifth harmonic in the mixing process, but this is attenuated even more strongly by the filter than the third harmonic.

The assumption that the voltage gain is equal to two is only valid for the steady state. When the divider is switched on the gain is higher than two and the loop gain is therefore greater than unity for the signal $\cos \omega t$. This signal is present in the noise and increases exponentially until limiting in the mixer or amplifier brings the gain back to unity. This happens only for the signal $\cos \omega t$ with the phase shown in fig. 17*b*; for other signals the phase relations are not correct and there is no positive feedback with a loop gain greater than unity.

Two experimental versions of the divider in fig. 17*a* have been made. One has a mixer that makes use of Schottky diodes and will operate up to 4 GHz. This is combined with an amplifier of bandwidth 1.3 GHz. The highest divisible frequency is 2.9 GHz. This is exactly twice the frequency at which the specifications indicate that the conversion loss of the mixer is equal to the gain of the amplifier and the loop gain is therefore unity. In this version there was no lowpass filter in the loop. The filtering was provided by the bandwidth-limiting elements in the mixer and the amplifier. By using an amplifier of greater bandwidth the highest divisible frequency can of course be increased further.

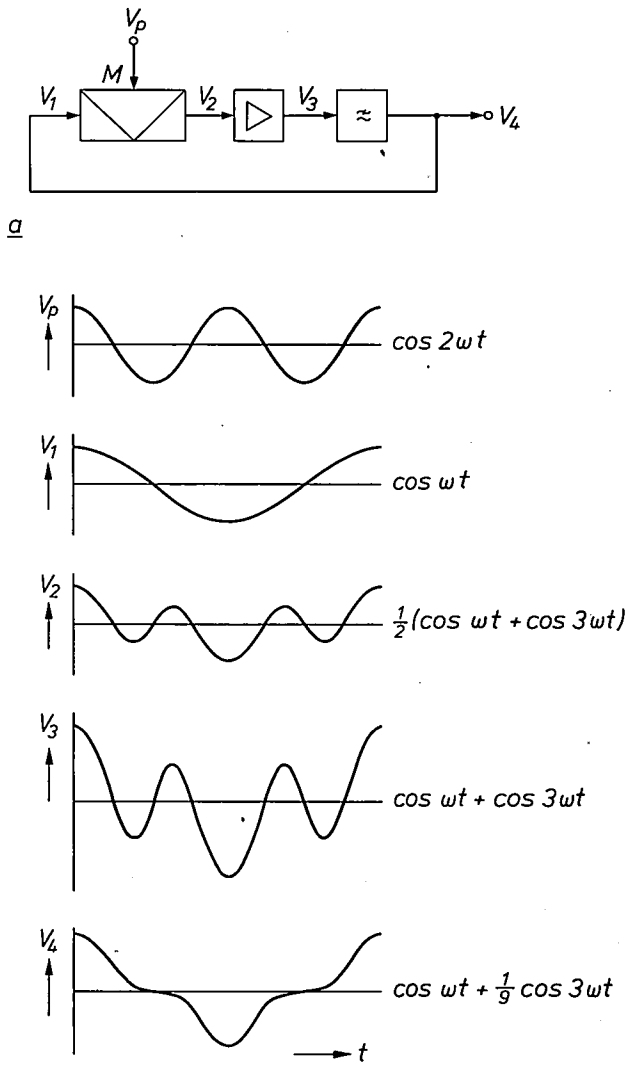


Fig. 17. *a*) Block diagram of a dynamic frequency-divider with mixer stage. The circuit operates with sinusoidal signals and consists of a closed loop containing a mixer M, a wide-band amplifier and a lowpass filter. *b*) Waveforms at various points in the circuit. When a 'pump signal' $V_p = \cos 2\omega t$ is multiplied by a signal $V_1 = \cos \omega t$ a product $V_2 = \frac{1}{2}(\cos \omega t + \cos 3\omega t)$ is produced. This is amplified twice, and the third harmonic is then attenuated in the lowpass filter. For simplicity it is assumed in this diagram that V_1 only contains the fundamental $\cos \omega t$ of V_4.

The other experimental version was based on a double-balanced modulator, shown in the schematic circuit of *fig. 18*. A circuit of this type also has gain, and since the base-emitter capacitances of the transistors give sufficient attenuation of the higher har-

monics the output can be directly connected to the input to form the closed loop of fig. 17*a*. This gives the very compact circuit of *fig. 19*.

This circuit was built as an experiment from discrete resistors and transistors (the microwave transistor BFR90). *Fig. 20* shows the minimum input voltage as a function of frequency. Comparison with similar measurements made on the circuits discussed earlier revealed two notable features: this 'analog' divider-by-two does not work at low frequencies, and the minimum input voltage increases linearly with the input frequency over practically all of the operating band.

This effect can be explained directly from the block diagram of fig. 17*a*. Since the loop gain decreases with increasing frequency because of the action of the filter, the pump signal from the mixer has to increase to produce the critical loop gain of unity.

The lower limit of the operating band is also determined by the filter. If the pump or input frequency 2ω decreases to below the cut-off frequency of the filter, the attenuation of the component at frequency 3ω becomes progressively smaller with respect to that of the component at frequency ω. Below a certain critical pump frequency two frequencies with comparable loop gain are present at the same time. The preference for the frequency ω is then lost and the circuit no longer operates properly as a divider-by-two.

It is interesting to compare the circuit described here with the circuit of fig. 3*a*, which also depends on the base-emitter capacitances of the transistors for its correct operation. Except for the base resistances R_B in fig. 3*a* the connections between the transistors T_1 to T_4 are the same in both circuits. The variations in the currents I_{b1},I_8 and I_{12},I_{34} also correspond: one grows larger at the expense of the other and vice versa 'in time' with the input frequency. There is however a marked difference in the average values of these currents; in fig. 3*a* they are unequal to make the amplifier with positive feedback dominant, whereas in fig. 19 both currents are equal. This difference corresponds to the difference between static and dynamic behaviour. The circuit of fig. 3*a* operates at frequencies down to 0 Hz, while the circuit of fig. 19 cannot operate at low frequencies because it contains nothing that can function as a static memory.

To summarize, then, a particularly simple and symmetrical divider-by-two for sinusoidal input and output signals has been described that will operate at very high input frequencies. However, the circuit cannot be used at low frequencies. In the next section an extension of the same principle will be described that has also been derived by assuming sinusoidal input and output voltages, yet will operate as a divider-by-two at low frequencies because of a gradual change in the waveform.

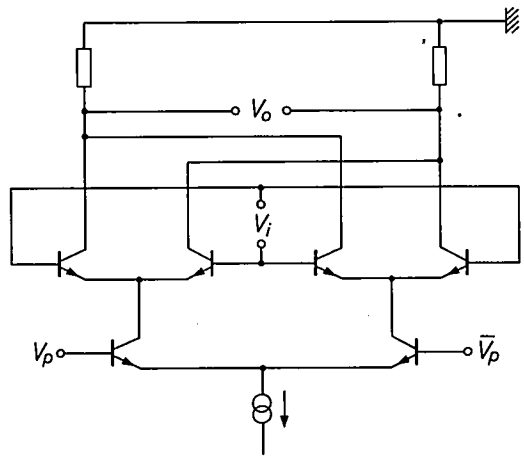


Fig. 18. Schematic circuit of a double-balanced modulator. The output signal V_o is the product of the input signal V_i and the pump signal V_p. All signals are balanced to earth.

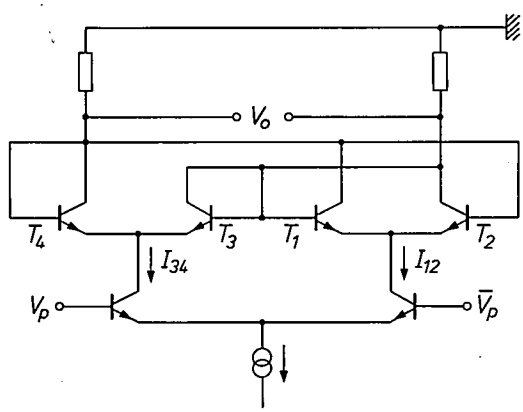


Fig. 19. Compact dynamic divider-by-two with mixer stage, derived from the double-balanced modulator (fig. 18) by directly connecting the output to the input.

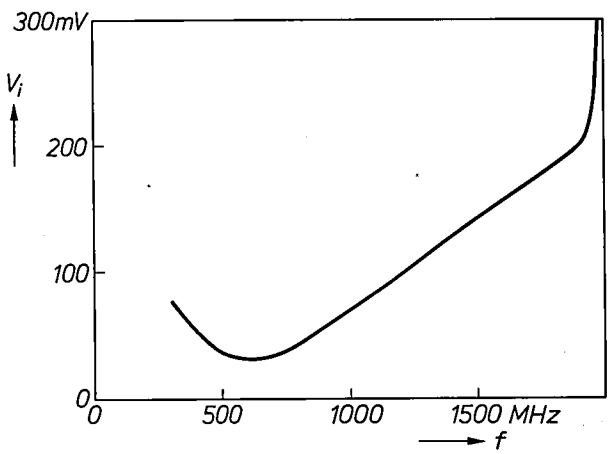


Fig. 20. The minimum input voltage V_i for correct operation of the dynamic divider-by-two with a mixer stage as a function of the frequency f. The curve was measured for an experimental version with BFR90 microwave transistors. The linear increase with frequency of the input voltage required is due to the low-pass nature of the circuit. If on the other hand the input frequency is so low that the higher harmonics are no longer adequately suppressed operation of the divider is no longer possible.

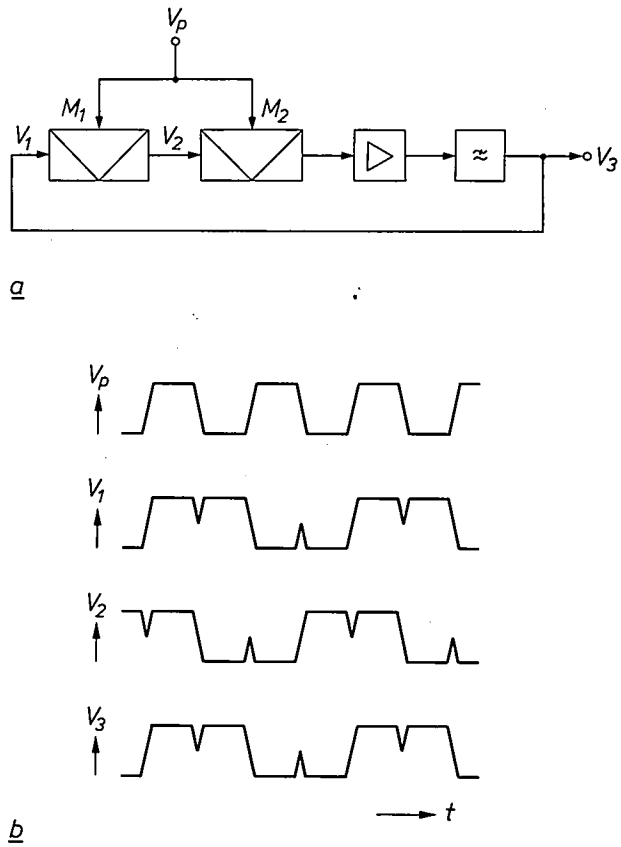


Fig. 21. *a*) Block diagram of a dynamic frequency divider with two mixer stages M_1 and M_2. The addition of the second mixer produces the correct phase relation for the division-by-two of approximately square-wave signals at low frequencies and thus gives an extension of the operating range to lower frequencies. The signal is multiplied by the same pump signal V_p in both mixers. *b*) Schematic representation of the waveforms. V_1 input signal to the first mixer. V_2 output signal from the first mixer and input signal to the second mixer. V_3 output signal of the divider, equal to V_1.

Frequency-divider with two mixer stages

Fig. 21a shows the circuit of fig. 17*a* extended by the addition of a second mixer stage. The new divider-by-two consists of a cascaded arrangement of two mixers with a common pump signal, a broad-band amplifier and a lowpass filter, together forming a closed loop. It is not difficult to see that the operation of this circuit is essentially the same as that of the earlier circuit with a single mixer, provided that the pump and loop signals are sinusoidal. Let us assume that the pump signal for the two mixers is $\cos 2\omega t$ and the input signal to the first mixer is $\cos \omega t$. The output signal from the first mixer then contains the frequencies ω and 3ω. Further mixing with the pump signal then produces a signal including the frequencies ω, 3ω and 5ω at the output of the second mixer. The components at frequencies 3ω and 5ω are attenuated with respect to the component at frequency ω by the lowpass filter. When the loop is closed the dominant component in the circulating signal is therefore the one at frequency ω.

The great difference between the two circuits becomes evident when the pump signal is given a frequency significantly lower than the cut-off frequency of the filter. The loop signal then consists of a number of components spaced at frequency intervals of 2ω, since each component is shifted by $\pm 2\omega$ in each mixing process. To show simply that the circuit of fig. 21 will also work at low pump frequencies it is easiest to start by assuming waveforms that approximate to a square wave, since such waveforms only contain the fundamental ω and the odd harmonics at a frequency spacing of 2ω. This provides a natural transition from analog to digital; the square-wave voltages can now be treated as signals in which it only matters whether they are at the high level or the low level.

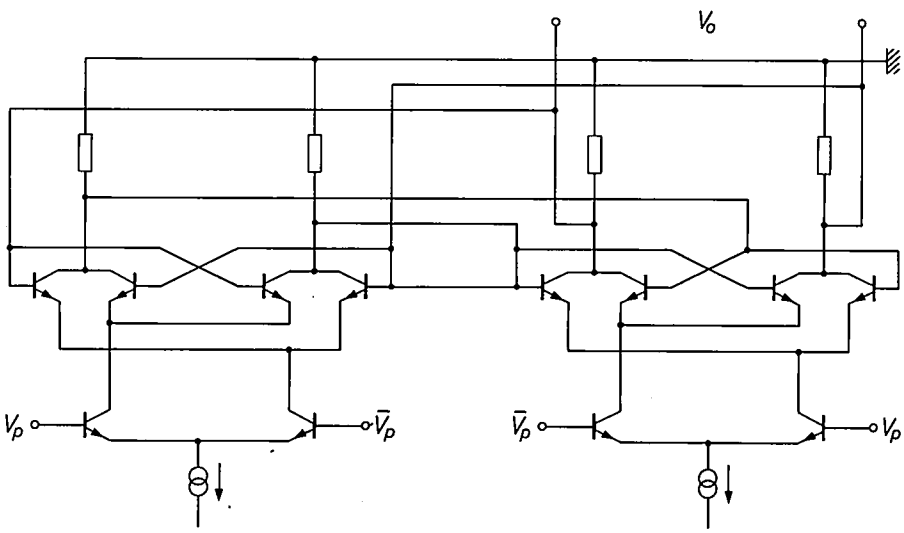


Fig. 22. Schematic circuit diagram of a dynamic divider-by-two with two mixer stages, each taking the form of a double balanced modulator. This circuit, when built with BFR90 microwave transistors, operates as a divider-by-two for frequencies between several MHz and 1.2 GHz. V_p pump signal. V_o output signal.

Fig. 21*b* shows schematically how a double multiplication of an input signal of 'nearly-square' waveform by a pump signal of similar waveform and twice the frequency gives an output voltage of the same waveform and phase as the input signal. The figure can be immediately verified by assigning the value '1' to the high level, for example, and the value

'0' to the low level and then performing the multiplications. If the successive signals are to reach the same level, energy must be supplied; in other words there must be gain in the loop. A lowpass filter plays no part in the process illustrated in fig. 21*b*; but the figure does clearly show that a second mixer is essential for the attainment of the correct phase relation.

A practical version, consisting of two double-balanced modulators in a cascaded arrangement with feedback, is shown in *fig. 22*. A version with BFR90 microwave transistors and discrete resistors was found to operate at pump frequencies from a few MHz to 1.2 GHz. Compared with the circuit with a single mixer with feedback the operating range has been considerably extended towards the lower frequencies. On the other hand the high-frequency performance is not so good because of the increased complexity.

Finally, it is interesting to look at the similarities between fig. 22 and fig. 2, which shows a divider-by-two based on the master-slave principle. Both circuits not only have the same number of transistors and resistors, but are also built up in the same way from six differential amplifiers, with the input signal applied to the two lower ones, while the collector currents provide complementary control of the two remaining pairs of differential amplifiers. The difference is to be found only in the connections to the collectors of the upper differential amplifiers. As would be expected with a comparable complexity, the high-frequency behaviour of the two circuits is very similar. At low frequencies both circuits will operate with signals of 'nearly-square' waveform; this is quite remarkable for a circuit like that of fig. 21*a*, since it consists entirely of components for analog signal processing that possess no kind of memory function.

Comparison of fig. 22 and fig. 2 shows, like the earlier comparison of fig. 19 and fig. 3*a*, that the difference between dividers based on digital principles and dividers that depend on an analog approach can in the end be remarkably small. As was said at the beginning of the article, the distinction between digital and analog is sometimes rather blurred.

Summary. Circuits for the division of high frequencies (about 1 GHz) are used in equipment for measurement and control and in the digital tuning of television receivers. In designing these circuits it is extremely important to keep parasitic capacitances and inductances small and few in number. The number of transistors and crossovers must therefore be as small as possible. The optimum operation is obtained at a maximum of symmetry. By using base-emitter capacitances as a dynamic memory it is sometimes possible to manage without a static memory in the form of a bistable transistor pair. In the 'travelling-wave divider' four transistors are connected in a closed loop in which there is a circulating voltage wave at half the input frequency. A circuit for sinusoidal input and output signals operates with a mixer stage. Modified versions of both these circuits can be made with an extended range of low-frequency operation.

Scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

E. Arnold: Conduction mechanisms in bandtails at the Si-SiO_2 interface.
Surface Sci. **58**, 60-70, 1976 (No. 1). *R*

H. M. J. M. van Ass, R. G. Gossink & P. J. W. Severin: Preparation of graded-index optical glass fibres in the alkali germanosilicate system.
Electronics Letters **12**, 369-370, 1976 (No. 15). *E*

H. Baudry: Le contrôle rhéologique des pâtes pour déterminer leur aptitude à la sérigraphie.
Electronique & Microél. ind. No. 226, 38-41, 1976. *L*

V. Belevitch: On the theory of cross-talk between twisted pairs.
Philips Res. Repts. **32**, 365-372, 1977 (No. 5/6). *B*

V. Belevitch, R. R. Wilson (NKF Kabel, Delft) **& G. C. Groenendaal:** The capacitance of circuits in a cable with twisted quads.
Philips Res. Repts. **32**, 297-321, 1977 (No. 4). *B, E*

C. Belouet, J. J. Brissot & R. Martres: Croissance de couches de silicium polycristallines sur substrat en carbone pour application aux cellules solaires.
Proc. Coll. Int. sur l'Electricité solaire, Toulouse 1976, pp. 191-198. *L*

J. W. M. Biesterbos, M. Brouha & A. G. Dirks: Pressure dependence of magnetic properties of amorphous RE-TM thin films.
AIP Conf. Proc. **29**, 184-185, 1976. *E*

G. W. Blackmore (Admiralty Materials Laboratory, Poole), **J. B. Clegg, J. S. Hislop** (AERE, Harwell) **& J. B. Mullin** (Royal Signals and Radar Establishment, Malvern): Concentrations of carbon and oxygen in indium phosphide and gallium arsenide crystals grown by the LEC technique.
J. electronic Mat. **5**, 401-413, 1976 (No. 4). *R*

J. van den Boomgaard & A. M. J. G. van Run: In-situ grown magneto-electric composites.
2nd Conf. on In situ composites, Bolton Landing, Lake George, N.Y., 1975, pp. 433-440; 1976. *E*

P. W. J. M. Boumans & F. J. de Boer: Studies of a radio frequency inductively coupled argon plasma for optical emission spectrometry, III. Interference effects under compromise conditions for simultaneous multi-element analysis.
Spectrochim. Acta **31B**, 355-375, 1976 (No. 6). *E*

F. J. A. den Broeder: Metallische glazen: een nieuwe materiaalklasse.
Chem. Weekbl. Mag. 1976, m 324-326 (juni). *E*

J. W. Broer: Presentatie, een 'foto-effect' van hoger orde.
Verslag Symp. Fotonica, Eindhoven 1975, pp. 276-285. *E*

E. Bruninx: A study of different geometries for the Seemann X-ray spectrometer equipped with a position-sensitive detector.
Philips Res. Repts. **32**, 253-265, 1977 (No. 4). *E*

T. M. Bruton: The growth of 'thick' single crystals of $Bi_4Ti_3O_{12}$ from bismuth borate solutions.
J. Crystal Growth **36**, 36-40, 1976 (No. 1). *R*

K. H. J. Buschow & M. Brouha: Narrow Bloch walls in RCo_5-type rare earth cobalt compounds.
AIP Conf. Proc. **29**, 618-619, 1976. *E*

W. A. Cense & W. Albers: In situ growth of aligned composites from the metastable state. The $Ca(NO_3)_2.RbNO_3$-$RbNO_3$ system.
2nd Conf. on In situ composites, Bolton Landing, Lake George, N.Y., 1975, pp. 59-65; 1976. *E*

V. Chalmeton, C. Patanchon* & C. Mesnage* (* S.N.P.E., Saint-Médard-en-Jalles, France): Radioscopie industrielle télévisée à haute énergie appliquée aux moteurs à carburant solide.
8th World Conf. on Nondestructive testing, Cannes 1976, paper 3E 13, 8 pp. *L*

P.-J. Courtois & G. Louchard (Université Libre de Bruxelles): Approximation of eigencharacteristics in nearly-completely decomposable stochastic systems.
Stoch. Proc. Appl. **4**, 283-296, 1976 (No. 3). *B*

H. T. van Dam: The conductance of heptyl viologen dibromide in water and methanol.
J. Electrochem. Soc. **123**, 1181-1184, 1976 (No. 8). *E*

M. Davio, J.-P. Deschamps & J.-C. Liénard: Optimization of strictly non-blocking and rearrangeable interconnections.
Philips Res. Repts. **32**, 266-296, 1977 (No. 4). *B*

Ph. Delsarte: Pairs of vectors in the space of an association scheme.
Philips Res. Repts. **32**, 373-411, 1977 (No. 5/6). *B*

G. Dittmer, A. Klopfer & J. Schröder: Heterogeneous reactions and chemical transport of tungsten with oxygen, fluorine, and fluorides of several metalloids.
Philips Res. Repts. **32**, 341-364, 1977 (No. 5/6). *A*

H. Dötsch: Stability and dynamics of microwave generated ring domains.
AIP Conf. Proc. **29**, 78-83, 1976. *H*

H. Durand: Les développements français des photopiles au silicium.
Proc. Coll. Int. sur l'Electricité solaire, Toulouse 1976, pp. 43-56. *L*

G. P. Edwards, T. Preston (both with Pye TMC Ltd, Malmesbury, U.K.), **L. D. J. Eggermont & M. H. H. Höfelt:** System and hardware design considerations for a single-channel analogue-to-PCM via-HIDM encoder.
1976 Int. Zürich Seminar on Digital communications, pp. B3.1-B3.5. *E*

W. G. Essers: New process combines plasma with GMA welding.
Welding J. **55**, 394-400, 1976 (No. 5). *E*

E. Fabre & R. Tijburg: Hétérostructures pour cellules solaires au silicium.
Proc. Coll. Int. sur l'Electricité solaire, Toulouse 1976, pp. 493-498. *L, E*

H. Figiel, A. Oppelt, E. Dormann (all with Technische Hochschule Darmstadt) **& K. H. J. Buschow:** Transferred hyperfine fields at the Y sites in Y-Co compounds.
Phys. Stat. sol. (a) **36**, 275-283, 1976 (No. 1). *E*

S. R. Fletcher*, E. T. Keve & A. C. Skapski* (* Imperial College, London): Structural studies of triglycine sulphate: Part I. Low radiation dose (Structure A), Part II. After X-irradiation/field treatment (Structure B).
Ferroelectrics **14**, 775-787, 789-799, 1976 (No. 3/4). *R*

R. C. French: Error performance of p.s.k. and f.f.s.k. subcarrier data demodulators.
Radio and electronic Engr. **46**, 543-548, 1976 (No. 11). *R*

A. D. Giles & F. F. Westendorp (Mullard Magnetic Components, Southport, Lancs., U.K.): The effect of cobalt substitutions on some properties of manganese zinc ferrites.
J. Physics D **9**, 2117-2122, 1976 (No. 14). *R*

J. J. Goedbloed & J. Joosten: Responsivity of avalanche photodiodes in the presence of multiple reflections.
Electronics Letters **12**, 363-364, 1976 (No. 14). *E*

J.-M. Goethals: Codes as elements in a group algebra.
Proc. Int. Symp. on Mathematical systems theory, Udine 1975, pp. 277-283; 1976. *B*

H. C. de Graaff & J. W. Slotboom: Some aspects of LEC transistor behaviour.
Solid-State Electronics **19**, 809-814, 1976 (No. 9). *E*

G. C. Groenendaal, R. R. Wilson (NKF Kabel, Delft) **& V. Belevitch:** Calculation of the proximity effect in a screened pair and quad.
Philips Res. Repts. **32**, 412-428, 1977 (No. 5/6). *E, B*

G. Groh: Future aspects of tomography from a physicist's point of view.
The new image in tomography, Proc. Symp. Actualitatis Tomographiae, Genoa 1975 (Exc. Med. Int. Congr. Ser. No. 392), pp. 71-75; 1976. *H*

G. J. van Gurp, D. Sigurd (Research Institute for Physics, Stockholm) **& W. F. van der Weg** (Philips Research Labs., Amsterdam Division): Tungsten as a marker in thin-film diffusion studies.
Appl. Phys. Letters **29**, 159-161, 1976 (No. 3). *E*

M. R. de Haan & C. H. F. Velzel: Intermodulation and moiré effects in optical video recording.
Philips Res. Repts. **32**, 436-459, 1977 (No. 5/6). *E*

J. C. M. Henning & J. H. den Boef: Strain-modulated electron spin resonance of Co^{2+} in MgO: a comparison of extensional and flexural modes.
Phys. Rev. B **14**, 26-34, 1976 (No. 1). *E*

A. van Herk & D. L. A. Tjaden: The magnetic field near the side edge of narrow magnetic recording heads.
Proc. Conf. on Video and data recording, Birmingham 1976, pp. 223-225. *E*

I. D. Higgins: Easy and accurate method for the characterisation of dielectric materials at X-band.
Electronics Letters **12**, 573, 1976 (No. 22). *R*

I. D. Higgins: Performance of self-oscillating GaAs m.e.s.f.e.t. mixers at X-band.
Electronics Letters **12**, 605-606, 1976 (No. 23). *R*

B. Hoekstra, J. M. Robertson & G. Bartels: Variations of magnetic anisotropy within epitaxial films of $Y_{2.85}La_{0.15}Fe_{3.75}Ga_{1.25}O_{12}$ obtained from spin wave resonance.
AIP Conf. Proc. **29**, 111-112, 1976. *E, H*

L. A. H. van Hoof & W. Albers: Conversion of X-rays into visible light in the NaCl-$PbCl_2$ composite system.
2nd Conf. on In situ composites, Bolton Landing, Lake George, N.Y., 1975, pp. 407-413; 1976. *E*

H. Ihrig: On the polaron nature of the charge transport in $BaTiO_3$.
J. Physics C **9**, 3469-3474, 1976 (No. 18). *A*

A. W. de Jager-Veenis & A. Bril: Vacuum ultraviolet excitation spectra of phosphors for use in gas discharge display panels.
J. Electrochem. Soc. **123**, 1253-1254, 1976 (No. 8). *E*

E. Klotz & H. Weiss: Short-time tomosynthesis.
The new image in tomography, Proc. Symp. Actualitatis Tomographiae, Genoa 1975 (Exc. Med. Int. Congr. Ser. No. 392), pp. 65-70; 1976. *H*

A. J. R. de Kock: Silicium-monokristallen zonder imperfecties.
Chem. Weekbl. Mag. 1976, m 329-330 (juni). *E*

J. Köhler & B. Schiek: FM-Rauschmeßplatz für Mikrowellen-Oszillatoren.
Mikrowellen-Magazin 4/76, pp. 276, 278 & 281-282, 1976. *H*

E. Kooi, J. G. van Lierop & J. A. Appels: Formation of silicon nitride at a Si-SiO_2 interface during local oxidation of silicon and during heat-treatment of oxidized silicon in NH_3 gas.
J. Electrochem. Soc. **123**, 1117-1120, 1976 (No. 7). *E*

A. M. van der Kraan, J. N. J. van der Velden, J. H. F. van Apeldoorn, P. C. M. Gubbens (all with Interuniversitair Reactor Instituut, Delft) **& K. H. J. Buschow:** Mössbauer effect study and magnetization measurements of RFe_3 intermetallic compounds with R = Sm, Gd, Tb, Dy, Ho, and Th.
Phys. Stat. sol. (a) **35**, 137-151, 1976 (No. 1). *E*

J. van Laar & A. Huijser: Contact potential differences for III-V compound surfaces.
J. Vac. Sci. Technol. **13**, 769-772, 1976 (No. 4). *E*

P. K. Larsen, R. Metselaar & B. Feuerbacher (European Space Agency, Noordwijk): UV photoemission studies of yttrium iron garnet.
AIP Conf. Proc. **29**, 668-669, 1976. *E*

P. E. J. Legierse (Philips Philite- en Metaalwarenfabrieken, Eindhoven): Fotochemische metaalbewerking.
Metaalbewerking **42**, 301-306, 1976 (No. 14).

G. M. Loiacono, J. Ladell, W. N. Osborne & J. Nicolosi: Phase transitions in $Cs(D_xH_{1-x})_2AsO_4$.
Ferroelectrics **14**, 761-765, 1976 (No. 3/4). *N*

F. J. M. J. Maessen, J. W. Elgersma (both with University of Amsterdam) **& P. W. J. M. Boumans:** A systematic and rigorous statistical approach for establishing the accuracy of analytical results and its application to a comparison of alternative d.c. arc procedures for trace analysis of geological materials.
Spectrochim. Acta **31B**, 179-199, 1976 (No. 4). *E*

W. F. G. Mecklenbräuker (I, II), **R. M. Mersereau*** (I, II) **& T. F. Quatieri, Jr.*** (I) (* M.I.T., Cambridge, Mass.): McClellan transformations for two-dimensional digital filtering: I. Design, II. Implementation.
IEEE Trans. **CAS-23**, 405-414, 414-422, 1976 (No. 7). *E*

D. Meyer-Ebrecht, J. Dittrich & J. Guldberg: Tomosynthesis.
The new image in tomography, Proc. Symp. Actualitatis Tomographiae, Genoa 1975 (Exc. Med. Int. Congr. Ser. No. 392), pp. 58-64; 1976. *H*

J. Michel, E. Fabre & M. Mautref: Cellules solaires minces au silicium à haut rendement.
Proc. Coll. Int. sur l'Electricité solaire, Toulouse 1976, pp. 161-168. *L*

A. R. Miedema: On the heat of formation of plutonium alloys.
5th Int. Conf. on Plutonium and other actinides 1975, Baden-Baden, pp. 3-20; 1976. *E*

J. H. Neave & B. A. Joyce: The origin of spurious peaks in mass spectra.
J. Physics D **9**, 2195-2200, 1976 (No. 15). *R*

A. Oppelt (Technische Hochschule Darmstadt) **& K. H. J. Buschow:** NMR investigation of the hyperfine interactions in $Y(Fe_{1-x}A_x)_2$ (A = Al, Co, Pt).
Phys. Rev. B **13**, 4698-4704, 1976 (No. 11). *E*

J. B. H. Peek: Het Nederlands URSI-comité.
T. Ned. Elektronica- en Radiogen. **41**, 61-63, 1976 (No. 2). *E*

J. M. Poate, W. L. Brown, R. Homer, W. M. Augustyniak (all with Bell Laboratories, Murray Hill, N.J.), **J. W. Mayer** (California Institute of Technology, Pasadena, Calif.), **K. N. Tu** (IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.) **& W. F. van der Weg** (Philips Research Labs., Amsterdam Division): The sputtering of PtSi and NiSi.
Nucl. Instr. Meth. **132**, 345-349, 1976.

H. Rau: Homogeneity range of high temperature $Ni_{3\pm x}S_2$.
J. Phys. Chem. Solids **37**, 929-930, 1976 (No. 10). *A*

H. Rau: Range of homogeneity and defect energetics in $Co_{1-x}S$.
J. Phys. Chem. Solids **37**, 931-934, 1976 (No. 10). *A*

J. M. Robertson: Improvement of lead-free flux systems for the growth of bismuth-substituted iron garnet films by liquid phase epitaxy.
J. Electrochem. Soc. **123**, 1248-1249, 1976 (No. 8). *E*

P. J. Roksnoer, W. J. Bartels & C. W. T. Bulle: Effect of low cooling rates on swirls and striations in dislocation-free silicon crystals.
J. Crystal Growth **35**, 245-248, 1976 (No. 2). *E*

C. L. Sam: Small-size discrete-capacitor N_2 laser.
Appl. Phys. Letters **29**, 505-506, 1976 (No. 8). *N*

H. Schomberg: Parallel solution of a nonlinear elliptic boundary value problem.
Simulation of systems, ed. L. Dekker, North-Holland Publishing Co., Amsterdam 1976, pp. 461-470. *H*

J. Schröder: Thermal energy storage using fluorides of alkali and alkaline earth metals.
Proc. Symp. on Energy storage, published by Electrochem. Soc., pp. 206-220, 1976. *A*

M. Sintzoff: Composing specifications of information structures.
New directions in algorithmic languages 1975, ed. S. A. Schuman, published by IRIA, Rocquencourt 1976, pp. 207-216. *B*

F. W. Smith: Surface-acoustic-wave parametric amplifier.
Electronics Letters **12**, 545-546, 1976 (No. 21). *R*

W. T. Stacy & W. Guse (Universität Hamburg): X-ray topographic study of Czochralski grown mullite.
J. Crystal Growth **35**, 153-158, 1976 (No. 2). *E*

F. J. van Steenwijk, H. Th. Le Fever, R. C. Thiel (all with Kamerlingh Onnes Laboratorium, Leiden) **& K. H. J. Buschow:** Mössbauer effect and magnetic properties of $EuCu_5$.
Physica **79B**, 604-609, 1975 (No. 6). *E*

A. L. N. Stevels: Luminescentie van europium(II)-geactiveerde aluminaten.
Chem. Weekbl. Mag. 1976, m 331-332 (juni). *E*

A. L. N. Stevels & A. D. M. Schrama-de Pauw: Theoretical and experimental efficiencies of X-ray screens.
J. Electrochem. Soc. **123**, 886-888, 1976 (No. 6). *E*

T. Thalhammer: Enkele algemene aspecten van energie-opslag.
Ingenieur **86**, 871-877, 1974 (No. 45). *E*

J. B. Theeten: Les ions lents: un outil (presque) idéal pour l'analyse des surfaces.
La Recherche **7**, 770-772, 1976 (No. 70). *L*

J. B. Theeten, F. Hottier & H. Paradan: Appareillage et méthodologie d'étude des surfaces de GaAs en cours de croissance en épitaxie phase vapeur.
Rev. Phys. appl. **11**, 587-595, 1976 (No. 5). *L*

M. J. J. Theunissen, J. Snel & P. H. M. Willemse: Measurements on bulk-channel MOS capacitors for dark current characterization of peristaltic charge-coupled devices.
Philips Res. Repts. **32**, 429-435, 1977 (No. 5/6). *E*

J. H. N. van Vucht & K. H. J. Buschow: Note on the occurrence of intermetallic compounds in the lithium-palladium system.
J. less-common Met. **48**, 345-347, 1976 (No. 2). *E*

H. W. Werner & N. Warmoltz: The influence of selective sputtering on surface composition.
Surface Sci. **57**, 706-714, 1976 (No. 2). *E*

M. V. Whelan & L. A. Daverveld: Thirty LED pointer-display RIGFET chip.
Philips Res. Repts. **32**, 322-337, 1977 (No. 4). *E*

C. W. White, N. H. Tolk, J. Kraus (all with Bell Laboratories, Murray Hill, N.J.) **& W. F. van der Weg** (Philips Research Labs., Amsterdam Division): Continuum optical radiation produced by low-energy heavy particle bombardment of metal targets.
Nucl. Instr. Meth. **132**, 419-425, 1976.

P. Wiedijk (Philips Lighting Division, Eindhoven): High-temperature extraction of gas from ceramics.
Anal. Chem. **48**, 1095-1096, 1976 (No. 7).

H. Zijlstra: Permanente magneten.
Natuur en Techniek **44**, 362-379, 1976 (No. 5). *E*

H. Zijlstra: Permanent magnets in magnetic field calculations.
Proc. COMPUMAG, Oxford 1976, pp. 164-167. *E*

Contents of Philips Telecommunication Review **35**, No. 1, 1977:

F. A. van Haaff: Programming an Air Traffic Control system (pp. 1-23).
C. G. Engelfriet: Building design for SPC telephone exchanges (pp. 25-32).
S. Friedrich & I. Reichelt: Z4, an economic four-channel carrier frequency system (pp. 34-43).
W. Bax: 18 MHz coaxial system with 12 MHz parameters proved feasible (p. 44).
J. de Boer: Overflow traffic of a group of lines connected to one or two link-networks (pp. 45-54).

Contents of Electronic Applications Bulletin **34**, No. 3, 1977:

G. van Schaik: An introduction to switched-mode power supplies in TV receivers (pp. 93-108).
The TCA580 integrated gyrator — principles and properties (pp. 109-122).
Notes on SMPS transformer design nomograms (pp. 123-136).

Contents of Mullard Technical Communications **14**, No. 133, 1977:

M. C. Gander, D. J. Beakhust & R. P. Gant: RGB colour decoder using three ICs from the TDA2500 range (pp. 86-109).
D. W. Parker, R. G. Pratt, F. W. Smith & R. Stevens: Acoustic surface-wave bandpass filters (pp. 110-124).
J. A. Garters: A low-power v.h.f. paging or pocket portable receiver (pp. 125-132).

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79, No. 3

The PTC effect of barium titanate

J. Daniels, K. H. Härdtl and R. Wernicke

In 1955 research workers at Philips discovered a new, spectacular property of barium titanate, a material that was then mainly used for its piezoelectric and dielectric properties. It was found that when this material in polycrystalline form was suitably doped its electrical resistivity near the Curie point increased dramatically with rising temperature (up to 100 % per °C). This high positive temperature coefficient prompted extensive investigations of the material in many laboratories. Philips are currently one of the largest European manufacturers of PTC resistors (thermistors), whose operation is based on this effect.

An explanation for the PTC effect of $BaTiO_3$ was first given by W. Heywang, and later by G. H. Jonker. The article below describes an investigation which has led to a model that offers a fuller explanation of the PTC effect of $BaTiO_3$ than has previously been possible.

After having originally been known primarily for its ferroelectric properties [1], barium titanate has attracted growing attention since 1955 as a material with a high positive temperature coefficient [2]. When $BaTiO_3$ is doped with acceptors or with donors the material is made either *P*-type or *N*-type. In a particular temperature region the electrical resistivity of *N-type* and *polycrystalline* $BaTiO_3$ can increase to 100 % per degree Celsius with rising temperature (*fig. 1*). This exceptionally strong temperature dependence with positive coefficient (known briefly as the PTC effect) is connected, as will be shown later, with the ferroelectric properties of $BaTiO_3$, which will therefore first be recapitulated.

The ferroelectric properties of $BaTiO_3$

Ferroelectricity is an effect related to the spontaneous occurrence of electric polarization in the material. In the absence of external influences, such as pressure or an electric field, the positive and negative ions in the material move in relation to one another in a way that results in local dipole moments. This spontaneous polarization is only found in a relatively small group of materials whose crystalline structure is characterized by the absence of a centre of symmetry and by the presence of a polar axis. Unless special measures are taken, the displacements referred to are accompanied by the formation of a domain structure, comparable to

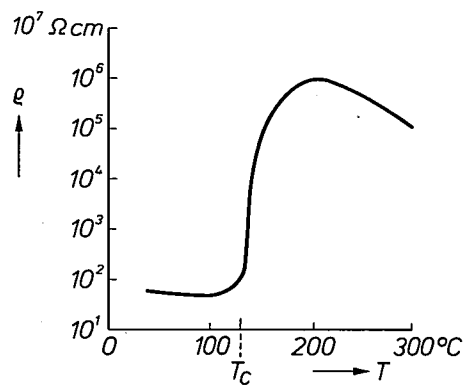


Fig. 1. The resistivity ϱ of $BaTiO_3$ doped with 0.3 % La as a function of temperature. T_C Curie point.

Dr J. Daniels, now with the Fachhochschule, Kiel, West Germany, was formerly with Philips GbmH Forschungslaboratorium Aachen (PFA), Aachen, West Germany. Dr K. H. Härdtl and Dr R. Wernicke are with PFA.

[1] See for example G. H. Jonker and J. H. van Santen, Philips tech. Rev. **11**, 183, 1949/50.

[2] P. W. Haaijman, R. W. Dam and H. A. Klasens, West German patent 929350, 23 June 1955; see also E. Andrich and K. H. Härdtl, Philips tech. Rev. **26**, 119, 1965, and G. H. Jonker and A. L. Stuijts, Philips tech. Rev. **32**, 79, 1971.

that found in ferromagnetics. The result is that the dipole moment of the material as a whole is zero. When the orientations of the local electric dipole moments can be reversed in certain crystallographic directions by applying an electric field, the material exhibits effects typical of ferroelectrics, such as hysteresis and anomalous dielectric behaviour.

Ferroelectricity was first observed in 1920 in Rochelle salt by J. Valasek, who also saw a formal analogy with ferromagnetic behaviour [3]. The complex structure of Rochelle salt (NaK tartrate) was for a long time an obstacle to a proper understanding of the effect. The first real breakthrough in understanding came with the discovery of the ferroelectric properties of $BaTiO_3$, which has a relatively simple structure (*fig. 2*).

In the temperature region above 130 °C barium titanate has a cubic crystal structure, and because of the presence of a centre of symmetry and the absence of a polar axis the material is not ferroelectric. When the material is coooled to below 130 °C the structure changes from cubic to tetragonal, and the ferroelectric properties come to the fore. This structure, shown in fig. 2, is called the perovskite structure, named after the related mineral perovskite ($CaTiO_3$). The transition temperature of 130 °C is called the Curie point T_C. As in the case of ferromagnetics, many important changes take place as the Curie point is approached and passed. Thus, for example, with rising temperature the dielectric

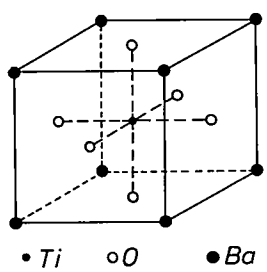


Fig. 2. Elementary cell of $BaTiO_3$ at a temperature above 130 °C. At 130 °C (the Curie point) this cubic perovskite structure undergoes a small deformation and is changed into a tetragonal structure, with an axial ratio c/a of 1.01.

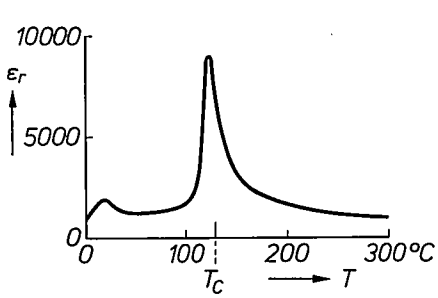


Fig. 3. The dielectric constant ε_r of $BaTiO_3$ as a function of temperature. The curve shows a peak at the Curie point T_C.

constant reaches a pronounced maximum near the Curie point (*fig. 3*). Upon a further increase of temperature the value of the dielectric constant, obeying the Curie-Weiss law, decreases as

$$\varepsilon \propto \frac{1}{T - T_C} \quad [*].$$

It should be added here that ferroelectric properties are always associated with pyroelectric and piezoelectric properties, but the converse is not necessarily true [4].

Relation between the ferroelectric properties and the PTC effect of $BaTiO_3$

What is the connection between the ferroelectric properties of $BaTiO_3$ and the high positive temperature coefficient of the material when it is *N*-type and polycrystalline? That a connection exists seems a reasonable assumption from the fact that the PTC effect only occurs near the Curie point. Investigations on mixed crystals have shown that a change in the Curie point is always accompanied by a corresponding change in the temperature region in which the PTC effect occurs.

Since the PTC effect does not occur in single-crystal $BaTiO_3$, the cause has been sought in the grain boundaries of the ceramic material. W. Heywang assumed the presence of surface states with an acceptor character at the grain boundaries, due for example to oxygen adsorption [5]. These surface states take up conduction electrons from the immediate vicinity, thus giving rise to a negatively charged boundary layer with, on both sides, a positive space charge that forms a symmetrical potential barrier for the remaining conduction electrons (*fig. 4*). The resistivity is therefore proportional to $\exp(e\phi/kT)$, where ϕ is the height of the potential barrier. According to the theory [5]:

$$\phi \propto N_s^2/\varepsilon n,$$

where N_s is the density of surface states, ε is the dielectric constant and n is the concentration of conduction electrons. Since ϕ is inversely proportional to the dielectric constant, the height of the potential barrier rapidly increases at a temperature above the Curie point, where the dielectric constant sharply decreases, and this would therefore explain the steep rise in the electrical resistivity above the Curie point. The fact that the low value of the dielectric constant below the Curie point does *not* give rise to such an effect is connected with the spontaneous electrical polarization that occurs in this temperature region, and is characteristic of a ferroelectric. As a consequence the surface states are wholly or partly compensated, and the formation of the potential barriers is suppressed [6].

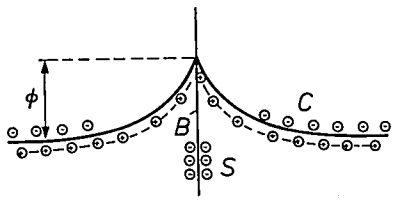


Fig. 4. A potential barrier caused by surface states S at a grain boundary B of donor-doped $BaTiO_3$ (schematic). ϕ is the height of the potential barrier. C conduction band.

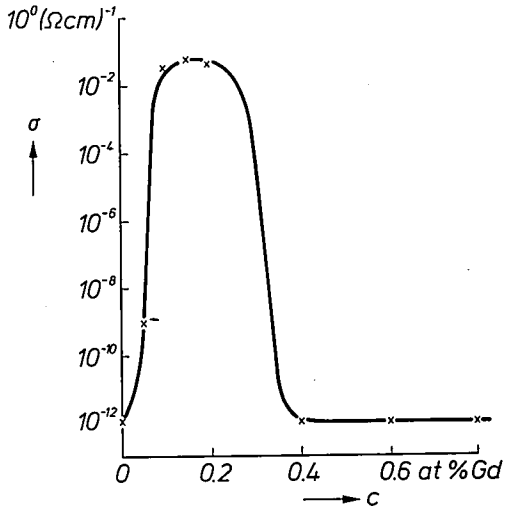


Fig. 5. Conductivity σ of semiconducting $BaTiO_3$ as a function of donor concentration c at room temperature. The donor in this case is Gd [7]. Other donors, such as La and Sb, have a similar effect on the conductivity. The ceramic material is sintered in air at 1400 °C.

Our investigation in general terms

The Heywang model gives a good explanation of the PTC effect, but it leaves certain questions open. More particularly, no satisfactory answer had been found until recently to the following four questions.

1) What is the precise nature of the acceptor-like surface states assumed in the model?

2) The PTC effect only occurs in material that has been made N-type by doping with donors. Why does it not occur in material that has been made N-type by reduction and not by doping?

3) Why is the PTC effect so greatly affected by the manner in which the $BaTiO_3$ is cooled after sintering and by its after-treatment?

4) What is the explanation for the curve of the electrical conductivity as a function of donor concentration, shown in *fig. 5*?

An answer to these questions is particularly important with a view to applications of the PTC effect. For, of course, it depends on the nature of the application — current stabilizer, temperature meter, heating element, etc. — which PTC characteristic is the most suitable.

With the object of achieving better control of the characteristics of PTC resistors (thermistors) we have tried to acquire a more complete knowledge of the nature and distribution of the defects (the 'defect chemistry') of the $BaTiO_3$ used.

The PTC effect is the result of the distribution of defects that arise in the course of two successive processes: firstly a sintering process at a temperature of about 1400 °C, and secondly a process of cooling to room temperature.

It may be assumed that during the sintering process all the defects reach their equilibrium concentration. In the next section we shall describe our investigation of the defect chemistry in equilibrium conditions.

During the cooling from about 1400 °C to about 1100 °C the cooling rate is of the same order of magnitude as the speed with which the state of equilibrium tries to restore itself. Since it is not feasible to attempt a direct approach to the complicated kinetic processes that occur in this temperature range, we calculated — as a first step — the defect concentrations that occur at room temperature after infinitely fast quenching. We took as our starting point various well-defined equilibrium states based on the model of the equilibrium situation described in the next section. This enabled us to estimate the trend of the changes that take place during cooling. We used these results as the starting point for the second step, which was to make a better estimate of the changes that actually take place, by using the true cooling rates to give a more exact calculation.

Before describing these two steps in the kinetic investigation, we shall outline the method we used for measuring the diffusion rates of the defects that occur in the material. It is obvious that without a fairly accurate knowledge of the magnitude of these diffusion rates it is scarcely possible to obtain any understanding of the kinetic processes involved.

Using the picture of the defect chemistry at room temperature that follows from these calculations, we shall present in the last part of the article a qualitative explanation of the PTC effect.

In order to indicate more or less the direction in which we were able to find the explanation, one im-

[3] J. Valasek, Phys. Rev. (2) **17**, 475, 1921.

[4] See for example F. Jona and G. Shirane, Ferroelectric crystals, Pergamon Press, Oxford 1962, especially pp. 10 and 11.

[5] W. Heywang, Solid-State Electronics **3**, 51, 1961; see G. H. Jonker, Solid-State Electronics **7**, 895, 1964, and G. H. Jonker, Mat. Res. Bull. **2**, 401, 1967.

[6] See the articles by Jonker, note [5].

[7] See T. Murakami, T. Miyashita, M. Nakahara and E. Sekine, J. Amer. Ceram. Soc. **56**, 294, 1973.

[*] Strictly speaking T_C should not appear in this relation, but the Curie-Weiss temperature Θ, which is always less than or equal to the Curie point T_C and is generally within a few degrees of T_C.

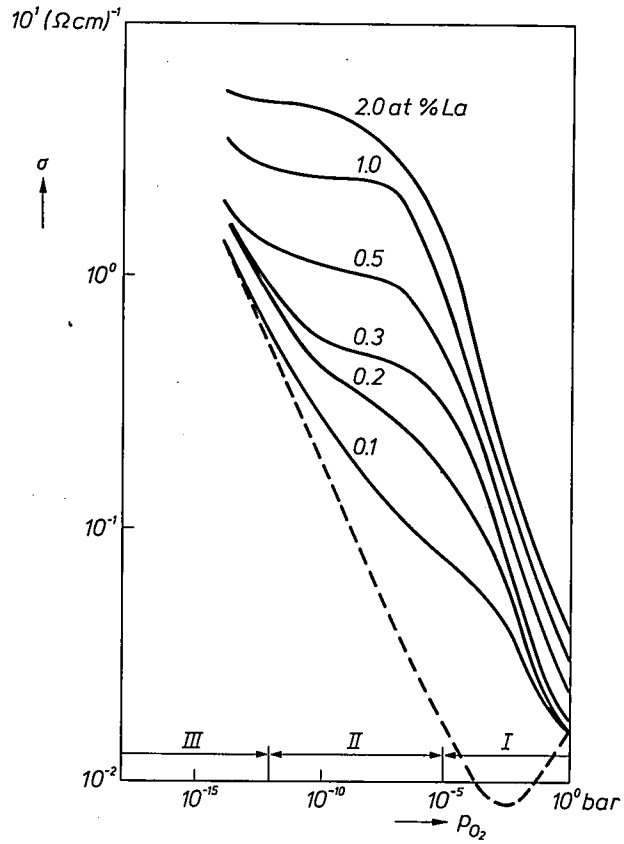


Fig. 6. Conductivity σ of undoped $BaTiO_3$ (*dashed curve*) and $BaTiO_3$ doped with the indicated percentages of La as a function of the oxygen partial pressure p_{O_2} in the ambient gas atmosphere, measured in equilibrium conditions at 1200 °C. The undoped material shows as a function of p_{O_2} a transition from *P*-type conduction (*right:* conductivity decreasing with decreasing oxygen pressure) to *N*-type conduction; in the doped material only *N*-type conduction occurs. Different mechanisms are operative in the three pressure regions *I*, *II* and *III*; see fig. 8.

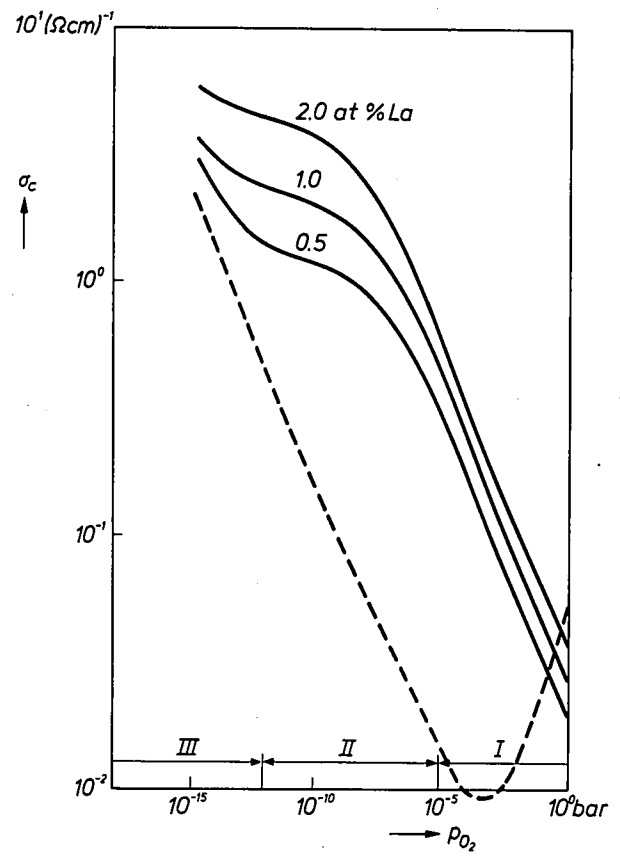


Fig. 7. The conductivity σ_c, calculated on the basis of our equilibrium model, as a function of the oxygen partial pressure p_{O_2}, see fig. 6.

portant result should be mentioned now. In the model resulting from our investigations, the potential barriers are *not* the consequence of a *two*-dimensional, negatively charged boundary layer at the grain boundary, as was assumed by Heywang, but occur in a *three*-dimensional, negatively charged zone at the grain boundary, the width of which depends on factors such as the cooling rate. The negative charge in this zone is due to the formation there of a large concentration of barium vacancies, resulting in a kind of *NIN* structure at the grain boundary. These data provide an important key to the control of the PTC effect.

The two-dimensional Heywang model may be regarded as a limiting case of our three-dimensional model, so that the models are not contradictory.

The thermodynamic investigation

N-type $BaTiO_3$ can be made in one of two ways, either by doping the material with donors, or by treating it in a reducing atmosphere. Although the PTC effect is only found in donor-doped $BaTiO_3$, it was desirable for a proper understanding of the effect to include the undoped material in the investigation.

It was found that the choice of the donor has no effect on the conductivity of the doped material. The donor mostly used in our investigations was La^{3+}. This ion is of about the same size as the Ba ion, and during doping it almost invariably replaces the Ba ion and not the much smaller Ti ion. This limitation of the number of possibilities simplifies the investigation.

The excess positive charge resulting from the substitution of an La^{3+} ion for a Ba^{2+} ion has to be compensated in one way or another in order to preserve the state of electroneutrality. In the present case this can be done by electrons that find a place in the conduction band (electron compensation) or by the formation of metal vacancies (vacancy compensation).

Fig. 6 shows the conductivity σ of undoped $BaTiO_3$ and of La-doped $BaTiO_3$ as a function of the oxygen partial pressure p_{O_2} in the ambient gas atmosphere at 1200 °C [8]. At high oxygen partial pressures the undoped material exhibits *P*-type conductivity, which gradually changes into *N*-type conductivity as the oxygen partial pressure is lowered [9]. The doped material only has *N*-type conductivity, but the shape of the conductivity curve seems to indicate that in this case there are three different mechanisms operative,

broadly corresponding to the three pressure regions *I*, *II* and *III*.

We have tried to find a model capable of explaining the shape of the curves in fig. 6. In doing so we have assumed that there are no defects in the systems considered other than neutral and singly or doubly ionized (positively charged) oxygen vacancies (V_O, $V_O^{\cdot}$, $V_O^{\cdot\cdot}$), and neutral as well as singly or doubly (negatively) charged barium vacancies (V_{Ba}, V_{Ba}', V_{Ba}''). It is also to be assumed that the presence of titanium vacancies in these systems may be treated as negligible [10]. We have also disregarded the presence of interstitial defects, defect clusters and all effects originating from interaction of the defects at high concentrations [8].

With these simplifications the electroneutrality equation reads:

$$n + [V_{Ba}'] + 2[V_{Ba}''] = p + [V_O^{\cdot}] + 2[V_O^{\cdot\cdot}] + [La^{\cdot}], \quad (1)$$

where n represents the concentration of conduction electrons, p the concentration of holes, and $[La^{\cdot}]$ the concentration of the La atoms substituted at Ba sites and subsequently ionized.

The defects to which we have confined ourselves in this investigation take part in the reaction equations listed in the left-hand column of *Table I*, to which the equilibrium conditions given by the law of mass action are applicable (right-hand column) [8]. Relation (8) applies to any semiconductor, and relation (9) is the modified Schottky equation which can be derived for these systems [11]. Combination of these nine equations yields a relation between the electron concentration, the oxygen partial pressure, the La concentration and the equilibrium constants. On the basis of a number of independently performed measurements, and making some reasonable assumptions, we have calculated the values of these equilibrium constants K [8]. We then used these to calculate the values of n, and from these values we found the conductivity ($\sigma = en\mu_n$, where μ_n is the value, known from the literature [12], of the conduction electron mobility). The results of these calculations are given in *fig. 7*. The large measure of agreement between the calculated and the measured curves indicates that our assumptions concerning the defects present are in general terms correct.

With the aid of the various equations in Table I we have subsequently calculated the concentrations of all the defects considered in the doped material (*fig. 8*).

Considering in each case only the largest concentrations of these defects, we then arrive at the following findings [8]. Whereas the conductivity of the undoped *N*-type material is determined mainly by the oxygen

Table I. Reaction equations (*left*) and equilibrium conditions (*right*) for the defects considered in La-doped $BaTiO_3$. Reaction equation (9) is derived in the article of note [11]; see also equation (13), p. 79.

Reaction equation	Equilibrium condition
(2) O (lattice) $\rightleftarrows$ $V_O + \frac{1}{2}O_2$ (Gas)	$[V_O]p_{O_2}^{\frac{1}{2}} = K_1 = N_1 \exp(-E_1/kT)$
(3) $V_O \rightleftarrows V_O^{\cdot} + e$	$[V_O^{\cdot}]n/[V_O] = K_2 = N_2 \exp(-E_2/kT)$
(4) $V_O^{\cdot} \rightleftarrows V_O^{\cdot\cdot} + e$	$[V_O^{\cdot\cdot}]n/[V_O^{\cdot}] = K_3 = N_3 \exp(-E_3/kT)$
(5) $V_{Ba} \rightleftarrows V_{Ba}' + h$	$[V_{Ba}']p/[V_{Ba}] = K_4 = N_4 \exp(-E_4/kT)$
(6) $V_{Ba}' \rightleftarrows V_{Ba}'' + h$	$[V_{Ba}'']p/[V_{Ba}'] = K_5 = N_5 \exp(-E_5/kT)$
(7) $La \rightleftarrows La^{\cdot} + e$	$[La^{\cdot}]n/[La] = K_D = N_D \exp(-E_D/kT)$
(8) $e + h = 0$	$np = K_i = N_c N_v \exp(-E_g/kT)$
(9) see eq. (13)	$[V_O][V_{Ba}] = K_s$

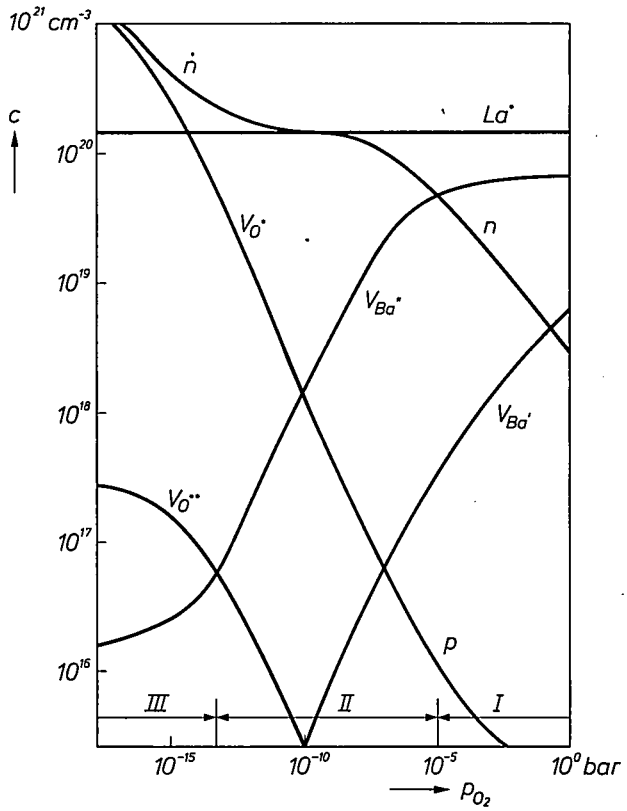


Fig. 8. The concentrations c of all defects taken into account in our equilibrium model, calculated as a function of the oxygen partial pressure p_{O_2} at 1200 °C, on samples of $BaTiO_3$ doped with 1% La. With decreasing oxygen pressure the dominating effects are, successively, the barium vacancies V_{Ba}'' (pressure region *I*), the doped La (region *II*) and the oxygen vacancies $V_O^{\cdot}$ (region *III*).

[8] A more detailed description of this part of our research, also giving experimental details, is given in J. Daniels and K. H. Härdtl, Philips Res. Repts **31**, 489, 1976.

[9] The method that makes such conclusions possible is dealt with by F. A. Kröger and H. J. Vink, in Solid State Physics **3**, 307, 1956.

[10] See T. Y. Tien and F. A. Hummel, Trans. Brit. Ceram. Soc. **66**, 233, 1967, and D. Hennings and G. Rosenstein, Mat. Res. Bull. **7**, 1505, 1972.

[11] See R. Wernicke, Philips Res. Repts **31**, 526, 1976, in particular pp. 538 and 539.

[12] A. M. J. H. Seuter, Philips Res. Repts Suppl. 1974, No. 3.

vacancies, the following three different neutrality conditions apply to the doped material in the pressure regions *I*, *II* and *III*:

$$2[V_{Ba}''] \approx [La^{\cdot}] \text{ in region } I, \quad (10)$$

$$n \approx [La^{\cdot}] \text{ in region } II, \quad (11)$$

$$\text{and } n \approx [V_O^{\cdot}] \text{ in region } III. \quad (12)$$

In the pressure region *III*, where the concentration of the oxygen vacancies is greater than that of the La dopant the conductivity curve of doped $BaTiO_3$ corresponds mainly to that of undoped $BaTiO_3$. In regions *I* and *II* of doped $BaTiO_3$ another effect is dominant, and this is the compensation of the excess positive charge introduced by the La^{3+} ions: via *vacancy* compensation in region *I* and *electron* compensation in region *II*. The virtually constant level of the conductivity in region *II* is determined by the concentration of the La dopant. In region *I* the barium vacancies acting as acceptors lower the electron concentration.

By carrying out the above calculations at different temperatures we were also able to make an estimate of the position of the energy levels corresponding to the various defects of $BaTiO_3$ (*fig. 9*).

This diagram shows why the *P*-type conduction found at room temperature is so much lower than the *N*-type conduction. Whereas the acceptor-like energy levels of the defects, such as the barium vacancies, are located about half-way across the band gap, the levels of the donor-like defects, such as the La ions and the singly charged oxygen vacancies, are situated close to the bottom of the conduction band.

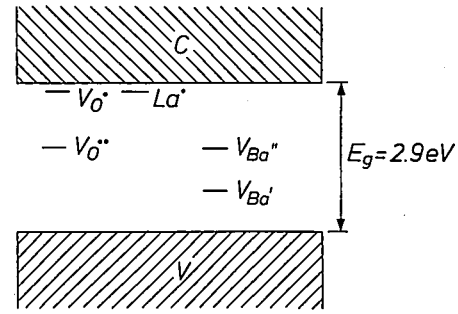


Fig. 9. Energy diagram of La-doped $BaTiO_3$. *C* conduction band. *V* valence band. E_g band gap, equal to 2.9 eV. The energy levels of $V_O^{\cdot}$ and $La^{\cdot}$ cannot be accurately determined, but the energy difference from the bottom of the conduction band amounts in any case to less than 0.1 eV.

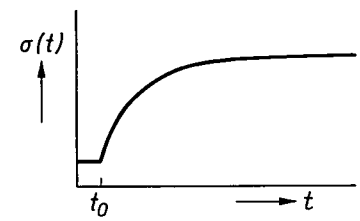


Fig. 10. Change of conductivity $\sigma(t)$ as a function of time after a stepwise change in the oxygen partial pressure at t_0.

The kinetic investigations

Determination of the diffusion rate [13]

To measure the diffusion rate of the various defects present we subjected barium titanate specimens which were in equilibrium with the surrounding, oxygen-containing, gas atmosphere to an abrupt change in oxygen pressure. The resultant changes in conductivity were measured as a function of time (*fig. 10*). From the observed time constant the diffusion coefficients were calculated. By using samples of different thickness and different grain size, we were able to determine in each case which process determines the rate of restoration of the equilibrium — diffusion processes or surface reactions — and whether the diffusion takes place along the grain boundaries or not.

The values of the diffusion constants found in this way are presented in *figs 11* and *12*. They relate to both undoped and doped $BaTiO_3$ at the oxygen pressures that are used in the manufacture of PTC resistors ($10^{-4} < p_{O_2} < 1$ bar).

In the undoped material the rate-determining step of the equilibrium restoration is the diffusion of oxygen vacancies. The diffusion coefficient of these vacancies has a relatively high value (10^{-4} cm²/s at 1000 °C), which is to be explained from the special character of the perovskite structure, in which the oxygen ions are located very close to each other. Similar results have been found for $SrTiO_3$ [14]. While the equilibrium is being established oxygen vacancies are formed (or annihilated) at the surface of the sample. From here they diffuse inwards until a new equilibrium state has been reached in the whole sample.

The behaviour of doped $BaTiO_3$ is quite different. In this material, under the conditions considered (region *I*) the *barium vacancies* occur in the largest concentration, and hence their behaviour determines the rate of equilibrium restoration. In figs 11 and 12 it can be seen that the diffusion coefficients of the barium vacancies are many orders of magnitude smaller than those of the oxygen vacancies. Our investigations also revealed the somewhat surprising fact that the barium vacancies, unlike the oxygen vacancies, reach their new state of equilibrium by diffusion *from the grain boundaries*, and not by diffusion from the surface. This is due to the special manner in which these vacancies are formed.

---

[13] See R. Wernicke, Philips Res. Repts **31**, 526, 1976.
[14] A. E. Paladino, J. Amer. Ceram. Soc. **48**, 476, 1965.
[15] F. S. Stone and R. J. D. Tilley, in: Reactivity of solids, Proc. 7th Int. Symp., Bristol 1972, p. 262.

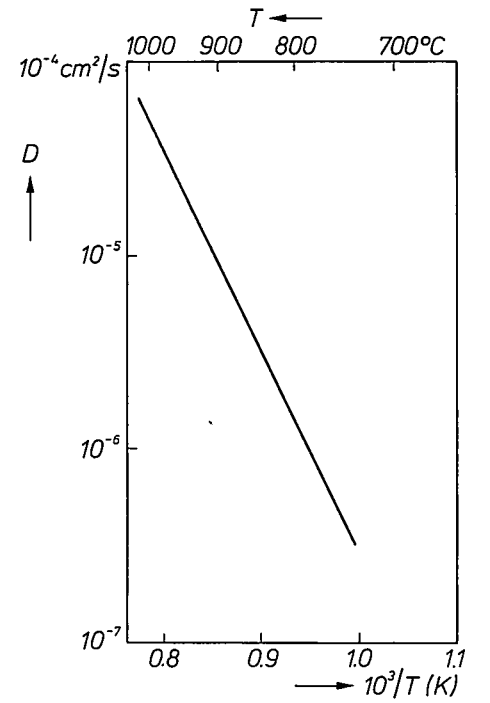


Fig. 11. Diffusion coefficients D of the oxygen vacancies in undoped $BaTiO_3$, calculated from observed time constants. The measurements were carried out at an oxygen partial pressure between 10^{-4} and 1 bar.

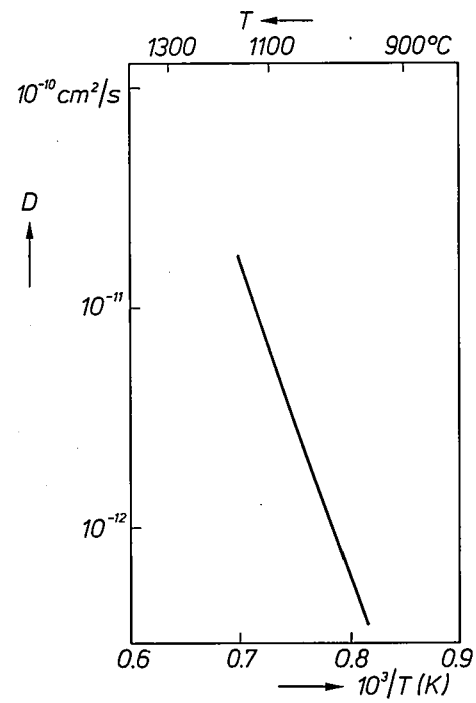


Fig. 12. Diffusion coefficients D of barium vacancies in $BaTiO_3$ doped with 0.5 % La, calculated from the time constants observed in the same oxygen-pressure region as in fig. 11.

Formation of barium vacancies

Since we have now arrived at one of the fundamental causes of the PTC effect in donor-doped $BaTiO_3$, we shall now try to picture the mechanism by which barium vacancies are formed.

In view of the low vapour pressure of barium oxide we can rule out the evaporation of barium during this process. This implies that a barium vacancy can only arise if a new site is created for a barium ion somewhere else in the lattice. One can imagine this being done by expansion of the existing $BaTiO_3$ lattice, for example at a grain boundary. For this purpose it is necessary, however, that in addition to oxygen atoms, titanium atoms must also be able to diffuse at a sufficiently high rate, and this does not seem very probable [15]. It is much more likely that, as a consequence of the excess TiO_2 always added to the starting materials of barium titanate ceramics to promote the sintering process, there is a second titanium-rich phase present at the grain boundaries (e.g. with the composition $BaTi_3O_7$) in which the barium ions can find a site to occupy. A conceivable reaction scheme could then be the following:

$$BaTi_3O_7 + 2Ba\ (\text{lattice}) + 2O\ (\text{lattice}) \rightleftarrows 3BaTiO_3 + 2V_{Ba} + 2V_O. \qquad (13)$$

The adjustment to the abrupt change in the oxygen partial pressure could then come about in the following way. First of all, the oxygen vacancies diffusing very rapidly through the entire material adapt themselves to the new equilibrium state. By means of reactions such as those given by equation (13) barium vacancies could then form at the grain boundaries and then diffuse from there to the interior until the new equilibrium state is reached throughout the grain [11].

What now happens during the cooling of sintered $BaTiO_3$ to room temperature? In the introductory section we have already noted the complexity of the processes that take place here, and for this reason we had to adopt an approach *in two steps*. Let us first see what happens on infinitely rapid quenching, as simulated by a computer.

Effects with infinitely fast cooling

Fig. 13 shows the conductivity σ_c of doped $BaTiO_3$ at room temperature after infinitely fast cooling from the equilibrium temperature T^E to 23 °C. This was calculated on the basis of the equilibrium model described in the previous section, for three different values of the oxygen partial pressure. It may be deduced from the figure that if the equilibrium temperature is chosen below a particular value, called the inversion temperature T_i, the cooled product exhibits low conductivity and that at an equilibrium temperature above the inversion temperature the end-product has a high conductivity. At the same time the inversion temperature varies with the oxygen partial pressure, as can also be seen from the figure.

The explanation for the occurrence of the inversion temperature is found in the change of the total concentration of barium vacancies as a function of the equilibrium temperature, shown in *fig. 14*. During infinitely fast cooling *the sum of the concentrations* of each type of atomic defect (e.g. $[V_{Ba}^{tot}] = [V_{Ba}] + [V_{Ba'}] + [V_{Ba''}]$) remains constant, and only a redistribution of the charges takes place (e.g. the conversion from $V_{Ba'}$ to $V_{Ba''}$). Above the inversion temperature, where $[La^{tot}] > 2[V_{Ba}^{tot}]$, there are not enough barium vacancies present to capture (compensate) all the electrons originating from the donors, as a result of which the conductivity is high, whereas below the inversion temperature complete compensation by the barium vacancies does take place, so that the conductivity is low.

Effects at practical cooling rates

Let us now consider the kinetic processes that take place at practical cooling rates, found in a sample sintered in air at 1400 °C, which was in equilibrium with the ambient atmosphere at the start of cooling. If this sample is cooled at the same rate at which the furnace cools, it is only during the initial period that the equilibrium restoration will be able to follow the temperature decrease in all parts of the sample. This implies that during this period the total concentration of barium vacancies will increase; see fig. 14. But as the temperature decreases further, restoration of the equilibrium will not everywhere be possible for kinetic reasons: the atomic defects are then no longer able to diffuse fast enough, and hence become 'frozen'. From that moment onwards, some parts of the material will no longer locally have the conductivity of the temperature prevailing there, but of the temperature down to which the atomic equilibrium has 'kept pace' with the cooling; see fig. 13.

Our calculation procedure is based on the assumption that the equilibrium state during cooling initially kept pace with the temperature decrease down to a critical temperature at which the equilibrium adjustment was *abruptly* ended by 'freezing'. The conductivity value obtained when this freezing temperature is varied may be deduced from figs 13 and 14. The actual cooling conditions can be approximated by means of a suitable choice of the freezing temperature, which should be chosen higher the faster the cooling rate.

In the foregoing we have seen that the rate at which equilibrium is reached is determined by processes which involve barium vacancies, in particular their formation at the grain boundaries and their diffusion from these boundaries to the bulk of the grain.

The process may be pictured such that during the cooling a diffusion front rich in barium vacancies penetrates the bulk of the grain from the grain bound-

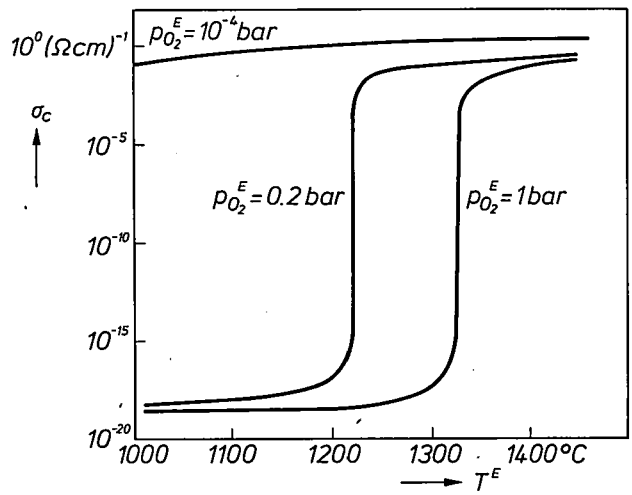


Fig. 13. Calculated conductivity σ_c of donor-doped $BaTiO_3$ at 23 °C as a function of the equilibrium temperature T^E, for three different values of the oxygen partial pressure. In each case σ_c is taken as occurring after infinitely fast quenching.

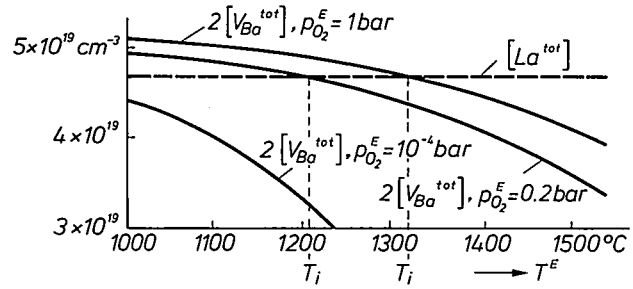


Fig. 14. Twice the total concentration of barium vacancies $2[V_{Ba}^{tot}]$ as a function of the equilibrium temperature T^E, at the three indicated values of the oxygen partial pressure. At the inversion temperature T_i the value of $2[V_{Ba}^{tot}]$ is equal to the total concentration of La donors, $[La^{tot}]$. At this temperature the electrons originating from the donors are exactly compensated by the barium vacancies acting as acceptors.

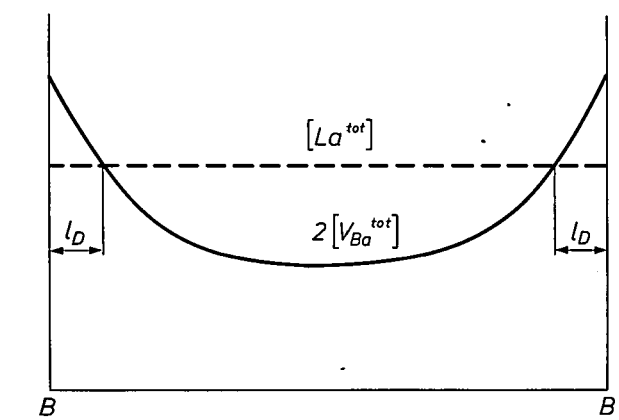


Fig. 15. Schematic profile of the defect concentrations relevant to the PTC effect, as found in a grain of La-doped $BaTiO_3$. Where the total concentration of lanthanum $[La^{tot}]$ is smaller than twice the total concentration of barium vacancies, $2[V_{Ba}^{tot}]$, all the conduction electrons introduced by the lanthanum are captured by barium vacancies. l_D width of the low-conductivity zone thus formed (depletion zone). *B* grain boundary.

ary, but that after some time the front loses its speed of propagation and comes to a standstill. The result is a heterogeneous distribution of the barium-vacancy concentration within each grain (*fig. 15*).

This means that while the barium vacancies retain their equilibrium concentration in the region *near the grain boundaries* down to fairly low temperatures, this is no longer the case in the *bulk* of the grain when the temperature has dropped to below about 1300 °C.

The distance l_D over which the diffusion front of the barium vacancies penetrates the grain depends mainly on the cooling rate and on the diffusion coefficient. In this way one arrives at an insulating boundary layer, with a thickness of l_D, while in the interior of the grain a high conductivity is maintained (10^{-1} to 10^{-2} $\Omega^{-1}cm^{-1}$). The total conductivity is then a function of the ratio of the width of the boundary layer to the grain size. The data calculated in this way are in qualitative agreement with the experimental findings.

With the model outlined in the foregoing the condition for the existence of the PTC effect is satisfied, on a broader basis than in the Heywang model [16]. In the following section we shall see more in detail which effects can be explained with the aid of this new model.

Explanation of the PTC effect with the new model

Complete compensation of the conduction electrons by barium vacancies in the outer layer of the grain gives rise to a potential barrier for the other conduction electrons which, as in the Heywang model, offers an explanation for the PTC effect. We shall now deal in turn with the four questions which could not readily be answered with the Heywang model (see page 75).

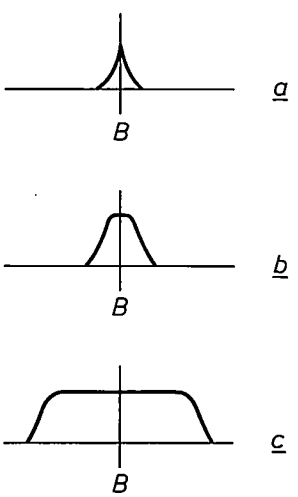


Fig. 16. Schematic shape of the potential barrier at the grain boundary *B* of donor-doped $BaTiO_3$, for increasing width of the boundary zone (*a*, *b* and *c*).

First question

The first question related to the nature of the acceptors postulated by the Heywang model, which trap the conduction electrons and thus give rise to the potential barrier. From the foregoing it appears that the barium vacancies in the outer layer of the grain function as such. The narrower this zone at the grain boundary becomes, the closer the limiting case postulated by the Heywang model is approached.

Second question

The second question was: why is the PTC effect not found in undoped $BaTiO_3$ that has been made *N*-type by a reducing treatment? The answer to this is: under these conditions the material has become *N*-type due to the formation of oxygen vacancies, and there are no or hardly any barium vacancies present, whose behaviour is a necessary condition for the formation of grain-boundary layers.

Third question

The answer to the third question, concerning the considerable influence of the cooling rate on the PTC effect, has already partly been touched upon in the introductory section: the cooling rate determines to a large extent the width of the insulating zone at the grain boundary, and the slower the cooling rate the wider this zone becomes (*fig. 16*).

Although a potential barrier changing in this way has a fairly complicated effect on the electrical characteristics of a PTC resistor, a general picture of its effect can nevertheless be formed. Roughly speaking, the wider the zone the higher is the concentration of barium vacancies at the grain boundary [17]. This implies that the wider the zone the less able will the ferroelectric polarization be to compensate these acceptors, which leads to an increase in the resistivity of the grain boundary layer below the Curie point.

If the cooling period is sufficiently long, the grain-boundary layer will in fact become fully insulating below the Curie point. Materials of this type are used in the manufacture of intergranular capacitors. Their (effective) dielectric constant can have a value as high as 150 000; this is due to their heterogeneous structure, in which the grain-boundary layers are insulators while the interior of the grain is semiconducting. With the aid of the new model it is thus easy to explain a continuous transition in the material properties of $BaTiO_3$, making it suitable both for PTC resistors and for intergranular capacitors.

[16] See J. Daniels and R. Wernicke, Philips Res. Repts 31, 544, 1976.

[17] See H. Ihrig and W. Puschert, J. appl. Phys. 48, 3081, 1977.

Fourth question

At first sight one would expect the conductivity of $BaTiO_3$ to increase with increasing donor concentration, for of course this has the effect of introducing more conduction electrons. The shape of the conductivity curve in fig. 5 has therefore never been fully explained before. With our model the shape of this curve can be understood if one also takes into account the changes taking place in the grain size [18]. In *fig. 17* it can be seen how the grain size and the conductivity vary as a function of the donor concentration. When the donor concentration increases to a value of more than about 0.3% the grain size becomes smaller, as also does the conductivity. At practical cooling rates the insulating boundary layer has a width of 1-3 μm [16]. If the donor concentration has increased to such an extent (to about 0.5%) that the grain size has acquired about the same value, the material has then, just as expected, become a complete insulator.

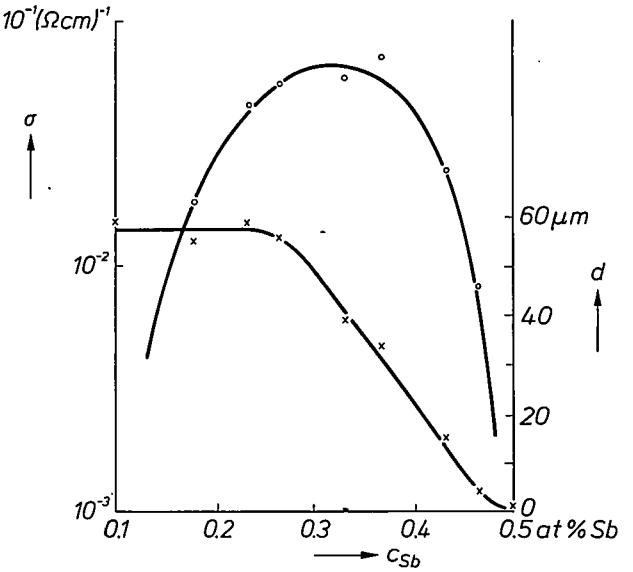


Fig. 17. Grain size d (*crosses*) and conductivity σ in Sb-doped $BaTiO_3$ as a function of the Sb concentration c_{Sb}.

Finally, a few words about the significance of this model for the practical applications of $BaTiO_3$. To begin with, the model has proved its value in practice by enabling us to achieve better control of the PTC properties of $BaTiO_3$. With the aid of this model a more methodical approach has proved possible and successful. The PTC properties can be controlled, for example, by varying the cooling conditions and using acceptors for doping.

We have just seen that as the width of the grain-boundary layer increases, a continuous transition takes place from PTC properties to dielectric properties. This means that the new model can also be used to explain and control the characteristics of intergranular capacitors made with $BaTiO_3$.

[18] R. Wernicke, Phys. Stat. sol. (a) **47**, 139, 1978.

Summary. From an investigation of the thermodynamic and kinetic behaviour of the defects in donor-doped ceramic $BaTiO_3$ a model has been developed to explain the PTC effect of this material. The model may be seen as an extension of the Heywang model. In the new model the surface states postulated by Heywang have the character of an insulating layer at the grain boundary caused by diffusion of barium vacancies. In this layer the concentration of barium vacancies is so great that the donors are fully compensated, which leads to an *NIN* structure at the grain boundary. The new model, which establishes a qualitative relation between the method of preparation and the potential barriers formed at the grain boundaries, makes it possible to control the PTC properties more methodically. In particular it provides a good explanation for the effect of the cooling conditions on the properties of this material. It shows, for instance, that an increase in the width of the diffusion layers, caused by slow cooling, produces a continuous transition from the characteristics of a PTC resistor to those of an intergranular capacitor. The Heywang model may be regarded as a limiting case of the model proposed here, where the diffusion layer is extremely narrow.

Amalgams for fluorescent lamps

J. Bloem, A. Bouwknegt and G. A. Wesselink

The luminous flux of a fluorescent lamp (low-pressure mercury type) depends to a considerable extent on the mercury-vapour pressure present in the tube. The pressure is determined by the temperature of the coolest part of the tube, which is usually the wall. The maximum luminous flux is reached when the wall temperature is about 40 °C, which for many fluorescent lamps corresponds to an ambient temperature of 25 °C. The wall temperature of lamps in closed luminaires or of special lamps with high power densities can be very much higher. In such conditions a high luminous flux can still be obtained by using a suitable amalgam in place of the pure mercury. This has the effect of lowering the mercury pressure and also of keeping it more or less stable in a broad temperature range. A difficulty with amalgam lamps is that they are generally relatively slow in reaching their maximum luminous flux after ignition. Some amalgams have now been found that can be used for making lamps that not only give a relatively high light output at a high wall temperature but also reach their maximum luminous flux quickly.

Introduction

A fluorescent lamp of the low-pressure mercury type is a discharge lamp containing a mixture of inert gases and a little mercury vapour. The discharge excites the mercury atoms to a higher energy state, from which they return to the ground state, producing ultraviolet radiation. This radiation is converted into visible light by fluorescent powder applied to the tube wall. The mercury-vapour pressure (p_{Hg}) is an important parameter in a fluorescent lamp. It is usually determined by the temperature of the wall, the coolest part of the lamp. If the pressure is too low, too few mercury atoms are excited, which means that not enough ultraviolet radiation falls on the fluorescent powder. If the pressure is too high, the mercury atoms absorb much of their own radiation. This causes an increase in the number of mercury atoms in excited states, resulting in a greater probability of interactions involving non-radiative transfer of some of the excitation energy. The optimum mercury-vapour pressure is approximately 0.8 Pa (6×10^{-3} torr) [1], a value reached when the wall temperature of the lamp is about 40 °C. In most types of fluorescent tubes in open luminaires this is the case at an ambient temperature of about 25 °C.

The wall temperature of lamps in closed luminaires or of lamps with higher power densities [2] can be substantially higher than 40 °C. To prevent the mercury-vapour pressure from becoming too high in such lamps, the mercury can be introduced in the form of an amalgam. Apart from lowering the vapour pressure, the substitution of an amalgam for pure mercury has another useful effect. In the temperature range in which the solid and liquid phases of the amalgam coexist with the gas phase, the mercury-vapour pressure above the amalgam is much less temperature-dependent than that above pure mercury. *Fig. 1* shows the mercury-vapour pressure as a function of temperature for In-Hg, the amalgam that was first used in fluorescent lamps [3]. In the temperature range in which stabilization occurs the mercury-vapour pressure above the amalgam does not differ much from the optimum value. Consequently the luminous flux does not differ greatly from the maximum value over a relatively wide temperature range.

Lowering the mercury-vapour pressure — which is necessary at wall temperatures higher than 40 °C — adversely affects the starting of the lamp, the amalgam still being at room temperature. When the vapour pressure above the amalgam is much lower than that above pure mercury, lamp ignition is difficult and a relatively long time elapses before the maximum luminous flux is reached. We have therefore tried to find amalgams which, while appreciably lowering and stabilizing the mercury-vapour pressure during operation, do not have an unduly low vapour pressure at room temperature.

Prof. Dr J. Bloem and Drs G. A. Wesselink are with Philips Research Laboratories, Eindhoven; Dr A. Bouwknegt is with the Philips Lighting Division, at Roosendaal.

[1] J. F. Waymouth, Electric discharge lamps, M.I.T. Press, Cambridge Mass. 1971, p. 23.
[2] J. Hasker, J. Illum. Engng. Soc. 6, 29, 1976.
[3] C. J. Bernier and J. C. Heffernan, Illum. Engng. 59, 801, 1964.

Alloys of mercury with more than one other metal were included in our investigations. The best results were obtained with $Bi_{0.53}In_{0.47}$-6% Hg and, to a somewhat lesser extent, with $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg.

We shall now first briefly discuss the temperature dependence of the mercury-vapour pressure above amalgams. We shall then examine the reasons for choosing the new amalgams, and finally deal with the behaviour of these amalgams in experimental lamps.

Mercury-vapour pressure above amalgams

The solid and liquid phases of an alloy of a metal with a little mercury coexist in a particular temperature range. This coexistence region lies below the melting point of the pure metal. The liquid phase is in equilibrium with both the solid and the gas phase. When the partial vapour pressure of the metal is negligibly low, the gas phase above the amalgam consists almost entirely of mercury vapour. The vapour pressure of the mercury depends on the mercury content in the liquid (x_l) and on that in the solid (x_s). On an increase of temperature within the coexistence region both x_s and x_l become smaller because less mercury is dissolved in the solid and in the liquid; see *fig. 2*. On the other hand, the same temperature rise causes an increase in the ratio between the number of mercury atoms in the gas phase and the value of x_l. Thus, with rising temperature in the coexistence region the change in the mercury vapour is determined by two opposing effects. The result is that the mercury-vapour pressure in this region changes relatively little.

The behaviour of the mercury-vapour pressure above an amalgam as a function of temperature is shown schematically in *fig. 3*. The stabilization begins at T_1, the temperature at which the first droplet of amalgam melt is formed, and ceases at T_2, the lowest temperature at which the amalgam is completely melted. In the example given in fig. 3 the vapour pressure between T_1 and T_2 goes through a maximum. It follows from an analysis that for such a maximum to occur the mercury content of the amalgam must be about 5% or lower. The maximum then lies at about 20 °C below the melting point of the metal without mercury [4].

New amalgams

Since indium has the relatively high melting point of 156 °C, the stabilization above In-Hg occurs at such high temperatures that the mercury-vapour pressure in the stabilization range is on the high side (fig. 1). A more serious problem with In-Hg is the relatively high solubility of mercury in solid indium. This has the effect of narrowing the stabilized region and of giving the mercury-vapour pressure at room temperature an undesirably low value.

An improvement over In-Hg is obtained by substituting bismuth for part of the indium [4]. This has the

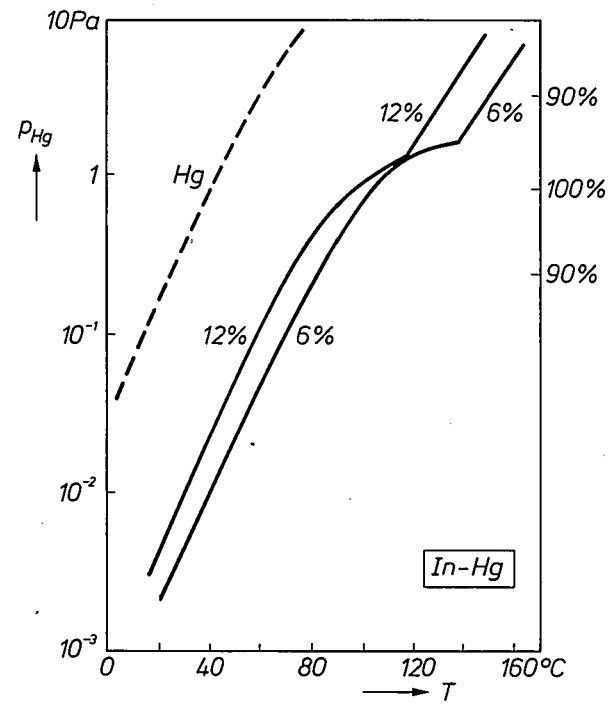


Fig. 1. Mercury-vapour pressure p_{Hg}, as a function of temperature T, above the amalgams In-6% Hg and In-12% Hg and above pure mercury (*dashed line*). The mercury-vapour pressure at which the relative luminous flux of a fluorescent lamp is at a maximum is shown at the right, and also the pressure at which it reaches 90% of the maximum value.

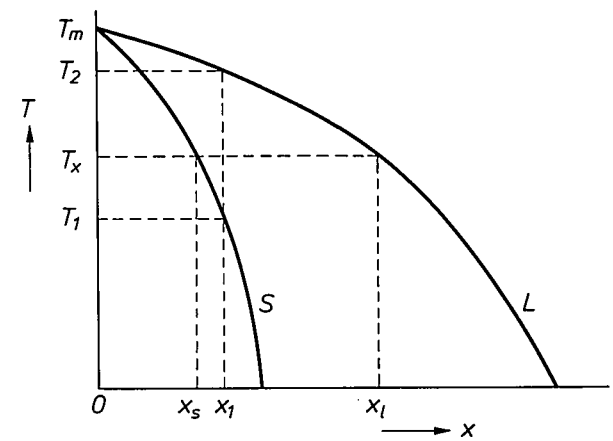


Fig. 2. Schematic T-x diagram for an amalgam; x mercury content, T temperature, T_m melting point of the metal without mercury. The solidus S indicates the temperature at which the amalgam begins to melt, and the liquidus L indicates the lowest temperature at which the amalgam is completely melted. An amalgam with a mercury content x_1 begins to melt at T_1 and becomes completely molten at T_2. Between these two temperatures the solid and liquid phases coexist. For an arbitrary temperature T_x in this coexistence region the mercury content in the solid (x_s) and that in the liquid (x_l) can be found by drawing a horizontal line and finding the points at which they intersect the curves S and L. With rising temperature the liquid-to-solid ratio increases, while x_s and x_l both decrease.

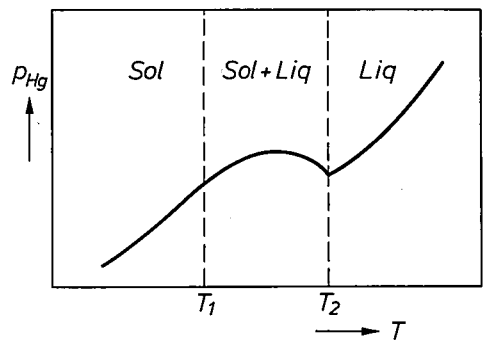


Fig. 3. Schematic curve of the mercury-vapour pressure above an amalgam as a function of temperature. Stabilization of the vapour pressure occurs between the temperatures T_1 and T_2, the coexistence region of the solid (*Sol*) and the liquid phase (*Liq*) of the amalgam. In the situation drawn here, the mercury-vapour pressure in the coexistence region reaches a maximum. The mercury-vapour pressure may also continue to increase with rising temperature in the coexistence region.

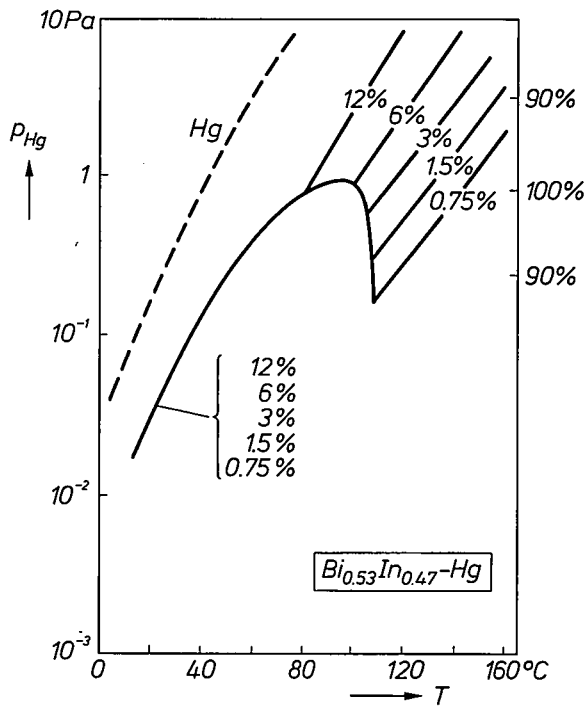


Fig. 4. The mercury-vapour pressure above $Bi_{0.53}In_{0.47}$-Hg for five different mercury contents as a function of temperature. For the amalgam with 6% mercury there is a relatively wide temperature range within which the vapour pressure differs only little from the optimum value (0.8 Pa). The vapour pressure above this amalgam at room temperature lies much closer to that above pure mercury than in the case of In-Hg (fig. 1), which improves the starting of the lamp.

effect of reducing the melting temperature and of lowering the solubility in the solid. An alloy with the composition $Bi_{0.53}In_{0.47}$ is a eutectic system of BiIn containing a small percentage of free Bi, with a melting point of 110 °C. The solubility of mercury in solid $Bi_{0.53}In_{0.47}$ is very low, because Bi and Hg atoms repel each other very strongly. *Fig. 4* shows the temperature dependence of the mercury-vapour pressure above $Bi_{0.53}In_{0.47}$-Hg for five values of the mercury content between 0.75% and 12%. From about 20 °C to about 80 °C the mercury-vapour pressure is independent of the mercury content. It is evident that the coexistence region of the solid and liquid phases of these amalgams is already reached at room temperature. Just above 80 °C the amalgam with the highest mercury content (12%) is completely melted, as a result of which the pressure-temperature curve differs from the coexistence curve, and the mercury-vapour pressure subsequently shows a marked increase with rising temperature. At about 100 °C the same occurs for the amalgam with 6% Hg, and also, at somewhat higher temperatures, for amalgams with the lower mercury contents. We see that for the amalgam with a mercury content of 6% there is a wide temperature range around 80 °C in which the mercury-vapour pressure differs very little from the optimum value. In addition, the mercury-vapour pressure with this amalgam is reasonably high at room temperature, being in fact about a factor of eight higher than that for In-12% Hg.

To determine whether there might perhaps be another candidate for use in fluorescent lamps, we theoretically compared the possibilities of a large number of amalgams [4]. These contained one or more of the metals Pb, Sn, Bi, In, Cd, Ga and Tl, most of which have a not unduly high melting point and a low vapour pressure. The criteria we used for the comparisons were the melting point of the metal, the intermetallic compound or the eutectic alloy without mercury, and the affinity between mercury and its possible alloy partner or partners [5]. These criteria made it possible to decide whether the stabilization of the mercury pressure would occur at suitable temperatures and whether a reasonably high mercury-vapour pressure would be expected at room temperature.

It was found that no improvement was to be expected from the amalgams with one or two other metals as compared with In-Hg and $Bi_{0.53}In_{0.47}$-Hg. Among the amalgams with three or four other metals the system Bi-Pb-Sn-Hg seemed to be the most promising. A eutectic mixture with the composition $Bi_{0.47}Pb_{0.29}Sn_{0.24}$ melts at a temperature as low as 96 °C ,which means that the mercury-vapour pressure above such an amalgam will stabilize at relatively low

[4] J. Bloem, A. Bouwknegt and G. A. Wesselink, J. Illum. Engng. Soc. 6, 141, 1977.

[5] The data were obtained from R. Hultgren, P. D. Desai, D. T. Hawkins, M. Gleiser and K. K. Kelley, Selected values of the thermodynamic properties of binary alloys, American Society for Metals, Metals Park, Ohio, 1973; A. R. Miedema, R. Boom and F. R. de Boer, J. less-common Met. 41, 283, 1975; R. Boom, F. R. de Boer and A. R. Miedema, J. less-common Met. 45, 237, 1976 and 46, 271, 1976.

temperatures. In addition, an amalgam with this alloy may be expected to have a higher mercury-vapour pressure at room temperature than $Bi_{0.53}In_{0.47}$-Hg, because the attractive force of tin and lead on mercury is weaker than that of indium on mercury.

Experiments confirmed these expectations; see *fig. 5*. In the case of the amalgam with 6% Hg the vapour pressure differs little from the optimum value over a wide temperature range around 60 °C. The mercury-vapour pressure above this amalgam at room temperature is only a factor of 1.7 lower than that above pure mercury.

We shall now explain how the amalgams $Bi_{0.53}In_{0.47}$-Hg and $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-Hg behave in an operating lamp.

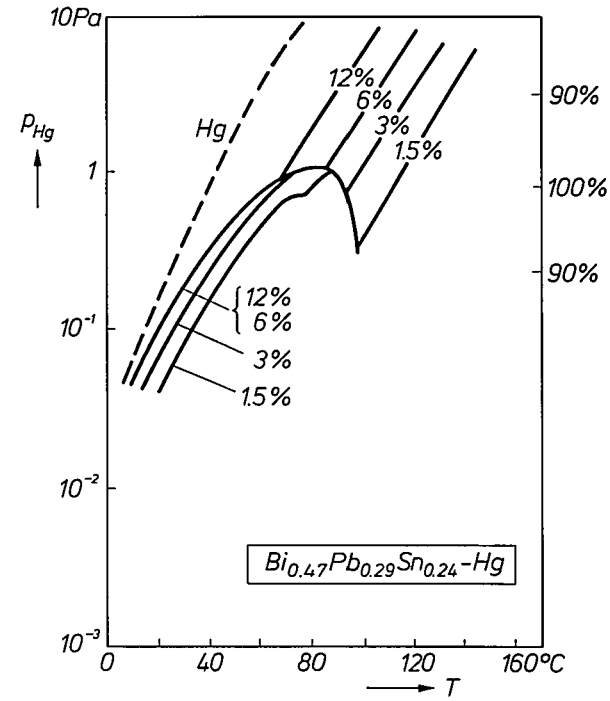


Fig. 5. Mercury-vapour pressure above $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-Hg for four different mercury contents as a function of temperature. Here again the amalgam with 6% mercury gives considerable stabilization of the vapour pressure near the optimum value. It can also be seen that the vapour pressure above this amalgam at room temperature is not much lower than that above pure mercury.

Experimental lamps with the new amalgams

The relative luminous flux Φ_{rel} of a fluorescent lamp at a particular amalgam temperature depends on the value of p_{Hg} at this temperature and on the relation between Φ_{rel} and p_{Hg}. The temperature T_{opt} at which Φ_{rel} is at a maximum is higher for an amalgam than for pure mercury, the more so the lower the value of p_{Hg} above the amalgam. The stabilization of p_{Hg} causes a broadening of the temperature range within which Φ_{rel} differs little from the maximum value. As an arbitrary measure of this broadening we have taken ΔT_{90}, the temperature range within which Φ_{rel} is more than 90% of the maximum value. In *Table I* the values of T_{opt} and ΔT_{90} for the new amalgams are compared with those for In-Hg and pure mercury. The Table also gives the values of the mercury-vapour pressure at room temperature, which are important for ignition of the lamp.

The relative luminous flux at a given ambient temperature T_{amb} is determined not only by the relation between p_{Hg} and the amalgam temperature and that between Φ_{rel} and p_{Hg}, but also by the relation between the amalgam temperature and T_{amb}. *Fig. 6* shows Φ_{rel} as a function of T_{amb} for lamps with $Bi_{0.53}In_{0.47}$-6% Hg, $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg, In-12% Hg and pure mercury. Comparing the curves of the amalgams with that of pure mercury we note in particular the broadening obtained with $Bi_{0.53}In_{0.47}$-6% Hg. The broadening obtained with $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg is about the same as that obtained with In-12% Hg.

To ensure prompt ignition of the amalgam lamp the amalgam should be situated near an electrode, where a suitable working temperature can quickly be reached. Normally the amalgam is applied at two or three places, indicated by *A*, *B* and *C* in *fig. 7*. The main amalgam is situated at position *A* behind the electrode, where the temperature rises relatively slowly after the lamp is switched on. A small amount of subsidiary amalgam is

Table I. Temperature T_{opt}, at which the relative luminous flux of the lamp is at a maximum, the temperature range ΔT_{90} within which the relative luminous flux is more than 90% of the maximum value, and the mercury-vapour pressure p_{Hg}^{20} at room temperature for three types of amalgam and for pure mercury.

System	T_{opt} (°C)	ΔT_{90} (°C)	p_{Hg}^{20} (Pa)
$Bi_{0.53}In_{0.47}$-6% Hg	81	52-120	0.03
$Bi_{0.53}In_{0.47}$-12% Hg	81	52-102	0.03
$Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg	64	38-100	0.09
$Bi_{0.47}Pb_{0.29}Sn_{0.24}$-12% Hg	64	38- 85	0.09
In-6% Hg	103	83-150	0.002
In-12% Hg	95	72-130	0.004
Hg	40	27- 58	0.16

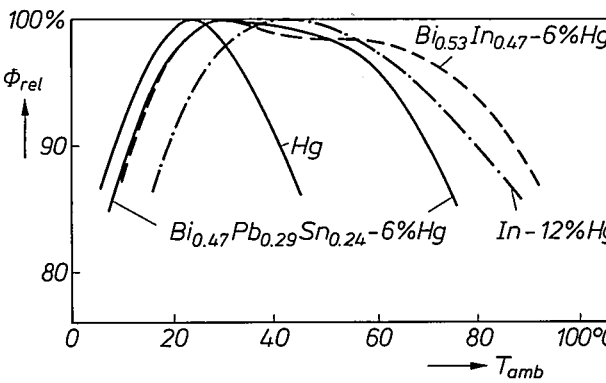


Fig. 6. The relative luminous flux Φ_{rel} as a function of ambient temperature T_{amb} for fluorescent lamps with $Bi_{0.53}In_{0.47}$-6% Hg, $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg and In-12% Hg and with pure mercury. The broadest curve is obtained with $Bi_{0.53}In_{0.47}$-6% Hg.

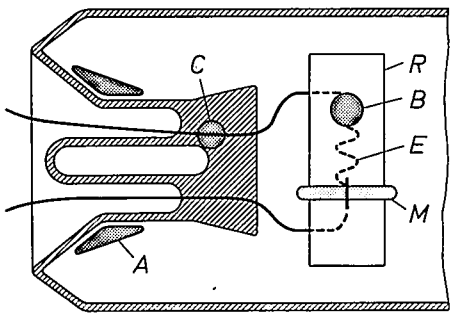


Fig. 7. Suitable locations for amalgams in a fluorescent lamp. At position *A* behind the electrode *E* there is a relatively large quantity of amalgam (the 'main amalgam'), which determines the mercury-vapour pressure in a lamp operating in the steady state. To obtain an acceptable mercury-vapour pressure soon after the lamp is switched on, a small amount of subsidiary amalgam is applied at position *B* to a metal ring *R* close to the electrode; the amalgam generally used for this purpose is In-Hg. If necessary, some subsidiary amalgam may also be applied at position *C*, on the glass between *A* and *B*. Shortly after the lamp is switched on, the subsidiary amalgam reaches such a high temperature that it is virtually depleted of mercury. After about 6 to 10 minutes the mercury-vapour pressure in the lamp is fully controlled by the main amalgam. After the lamp has been switched off, the mercury redistributes itself among *A*, *B* and *C*. The time needed for complete redistribution is shorter if the main amalgam has a high mercury-vapour pressure at room temperature. *M* capsule for dosing the mercury.

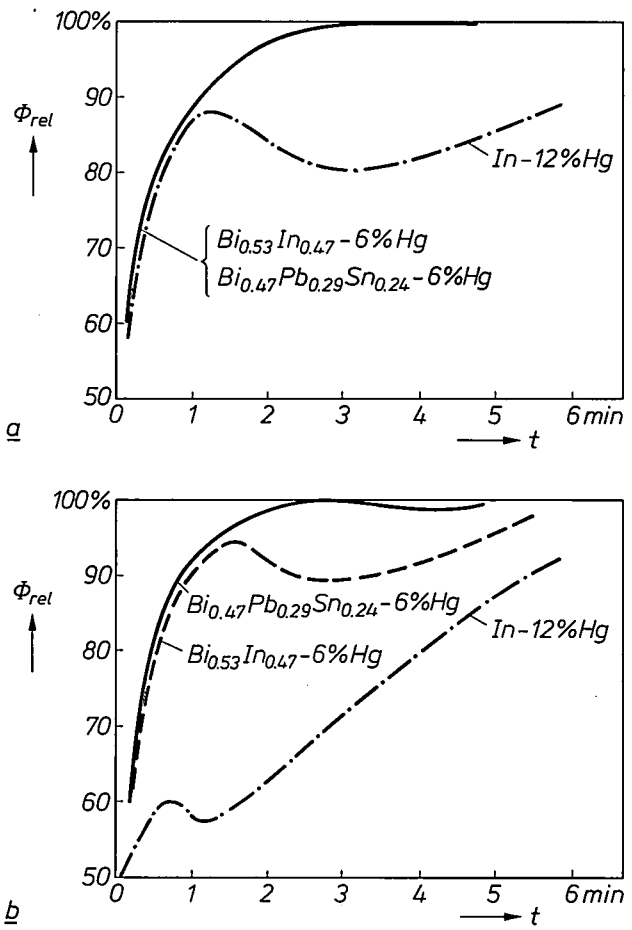


Fig. 8. Relative luminous flux Φ_{rel} as a function of time t after the ignition of amalgam lamps with previous off-periods of 15 hours (*a*) and 1 hour (*b*). A lamp with $Bi_{0.53}In_{0.47}$-6% Hg or $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg needs less time to reach an acceptable luminous flux than a lamp with In-12% Hg. The difference is greater the shorter the off-period of the lamps.

applied at position *B* to a metal ring close to the electrode, where the temperature rises very quickly. If necessary an additional quantity of a subsidiary amalgam may be introduced at position *C* on the glass between *A* and *B*. The temperature of *C* lies in between that of *A* and that of *B*. The subsidiary amalgam generally used is In-Hg.

In practice the metal or the alloy is first applied without the mercury. The appropriate amount of mercury is added usually by means of a capsule fixed to the metal ring close to the electrode; a thin electrically conducting wire is stretched over the capsule. R.F. heating of this wire breaks the capsule open, thus releasing the mercury to form the desired amalgam.

Immediately after the lamp is switched on the mercury in the discharge tube originates mainly from the subsidiary amalgam. After 6 to 10 minutes the main amalgam has warmed up and begins to control the mercury-vapour pressure. Owing to the high temperature, the subsidiary amalgam then has hardly any mercury left. When the lamp is switched off after some time, the mercury is gradually redistributed between the two amalgams, and it is not until some considerable time later that the original mercury distribution is restored. If the lamp is switched on again before that point has been reached, the subsidiary amalgam will contain too little mercury. Until the main amalgam has again warmed up sufficiently, not enough mercury will have been released in the lamp to produce an acceptable luminous flux. The redistribution of the mercury between the two amalgams takes place faster if the main amalgam has a high mercury-vapour pressure at room temperature. In this respect the new amalgams give a considerable improvement compared with In-Hg. The improvement is particularly noticeable when the lamp is switched on again after having been switched off for a short time. As an example, *fig. 8* shows the relative luminous flux of our experimental lamps as a function of time after ignition, for different switched-off intervals of 15 hours (*a*) and 1 hour (*b*). In the first case the maximum luminous flux is reached fairly quickly for $Bi_{0.53}In_{0.47}$-6% Hg and $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg. In the case of In-12% Hg a slight drop occurs after about one minute. This is an indication that the subsidiary amalgam has soon lost its reserve of mercury and that the main amalgam has not yet warmed up enough to deliver a sufficient supply. The difference is even greater when the off-period is not 15 hours but one hour. With the new amalgams an appreciable luminous flux is still reached quite soon after ignition, but this is no longer the case with In-12% Hg.

We may therefore conclude from the results obtained with the new amalgams that there are promising applications in fluorescent lamps. An additional advantage is that they use less mercury and also contain less of the expensive metal indium.

Summary. The luminous flux of a fluorescent lamp is at a maximum when the mercury-vapour pressure is 0.8 Pa. This value is reached when the temperature of the tube wall is about 40 °C, which is the case with many fluorescent lamps at an ambient temperature of 25 °C. When an amalgam is used in the lamp instead of pure mercury the mercury-vapour pressure is lowered and it also shows little variation within a fairly wide temperature range. As a consequence the temperature range within which the vapour pressure exhibits an acceptably small deviation from the optimum value can be shifted to higher temperatures and also widened. This makes amalgams suitable for use in special fluorescent lamps of the low-pressure mercury-vapour type in which the wall temperature is appreciably higher than 40 °C. The known mercury alloy In-12% Hg has the drawback that the mercury-vapour pressure above the amalgam at room temperature is much lower than that above pure mercury. As a consequence the maximum luminous flux is reached only very slowly after the lamp is switched on. The new amalgams $Bi_{0.53}In_{0.47}$-6%Hg and $Bi_{0.47}Pb_{0.29}Sn_{0.24}$-6% Hg give a substantially higher and stabilized luminous flux at higher ambient temperatures, without the mercury-vapour pressure being too low at room temperature.



Phenol synthesis and photomorphogenesis

G. Engelsma

Phenols are organic compounds, which are synthesized in the plant with the aid of enzymes and which, in their turn, also affect the synthesis of these enzymes. When a plant is exposed to light after a period in the dark, the synthesis of phenols is increased. Since phenols absorb ultraviolet radiation, the plant thus receives protection from radiation damage. The same mechanism that leads to increased phenol synthesis under the action of light is probably also operative in the control of plant elongation and flower induction.

The author and the editor (E. Fischmann) would like to dedicate this article to Prof. Dr. E. Havinga of the University of Leiden, who guided their studies in organic chemistry and is retiring in June of this year.

Introduction

The effect of light on plant development, photomorphogenesis, has been the subject of an earlier article in this journal, which will be called article I [1]. The gherkin seedling grown in the dark was taken as an example to illustrate some of the effects of illumination, showing for example how it inhibits the elongation of the stem (the hypocotyl) and causes the cotyledons, the original seed lobes, to grow into organs with a leaf function; see *fig. 1*. The title photograph shows another example of photomorphogenesis, which we shall be looking at in the article below: it shows three *Salvia occidentalis* plants; only the one on the left has started to flower, as a result of the way in which it was illuminated. This plant was illuminated for eight hours a day

Dr G. Engelsma is with Philips Research Laboratories, Eindhoven.

[1] G. Engelsma, Philips tech. Rev. **28**, 101, 1967.



Fig. 1. Gherkin seedlings five days old. *Left:* plant grown in the dark. *Right:* plant exposed to light for 12 hours a day.

cinnamic acids *p*-coumaric acid and ferulic acid — in the hypocotyl of the gherkin seedling, and considerable attention was paid to the control of the enzymes involved in the synthesis, phenylalanine ammonia-lyase (PAL) and cinammic-acid hydroxylase (*fig. 2*). The possible relation between the stimulation of phenol synthesis and the inhibition of the elongation of the seedling under the influence of light was also discussed in the article. It seemed that this relation could be the result of phenols involved in the control of the concentration of the growth hormone indoleacetic acid (IAA).

The article below will deal in more detail with this mechanism that regulates phenol synthesis in the seedling and with the manner in which light is involved in this process. It will also discuss the way in which temperature differences are compensated by this mechanism. In addition it will examine the effects of light on phenol synthesis in the leaves of a fully grown plant, *Salvia occidentalis*, an effect that may induce flowering, the link again being the growth hormone indoleacetic acid.

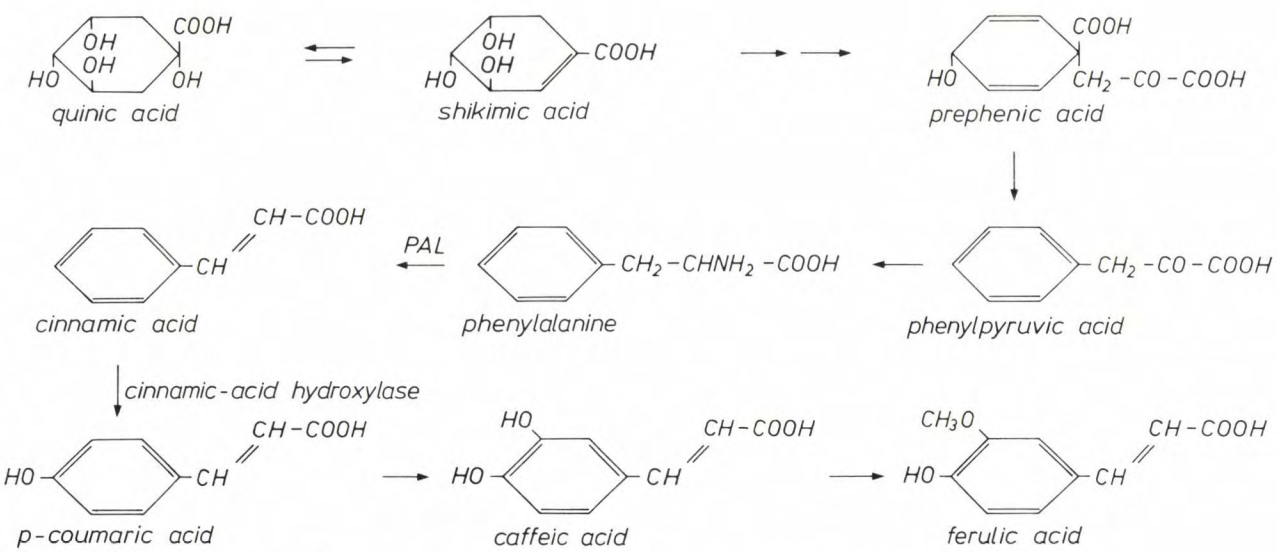


Fig. 2. Synthetic pathway of the hydroxycinnamic acids *p*-coumaric acid, caffeic acid and ferulic acid. The synthesis is mainly brought about by the enzymes phenylalanine ammonia-lyase (PAL) and cinnamic-acid hydroxylase. A number of products of this synthetic pathway, such as cinnamic acid, *p*-coumaric acid and ferulic acid, in turn affect the synthesis of the enzyme PAL. This forms the basis of an important regulating mechanism in the plant.

during a period of several weeks, with each exposure followed by 16 hours in the dark. The plant in the centre was illuminated for 16 hours a day, followed by eight hours in the dark, while the plant on the right was subjected to the same treatment as the one on the left, with the only difference that every 16-hour dark period was interrupted half-way through for 15 minutes of exposure to red light.

The main subject of article I was the effect of light on the synthesis of phenols — in particular the hydroxy-

Investigations on the gherkin seedling

In article I it was shown how the phenols in the plant are produced from phenylalanine, a key role being played by the enzyme PAL. This enzyme determines how much phenylalanine is used up for the phenol synthesis and how much is left for other processes, including protein synthesis. It was also shown how, in the hypocotyl of a dark-grown gherkin seedling exposed to light for the first time, the PAL concentration at 25 °C first remains constant for an hour and a half,

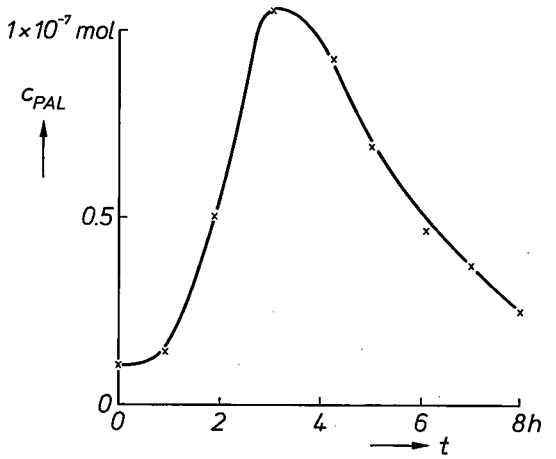


Fig. 3. The concentration c_{PAL} of PAL in the hypocotyl of the gherkin seedling as a function of the exposure time t to illumination with blue light at 25 °C. In the first few hours the enzyme concentration increases steeply, followed by a decrease. The concentration c_{PAL} is expressed here, as elsewhere in the article, as the quantity of cinnamic acid produced *in vitro* per hour under standard conditions.

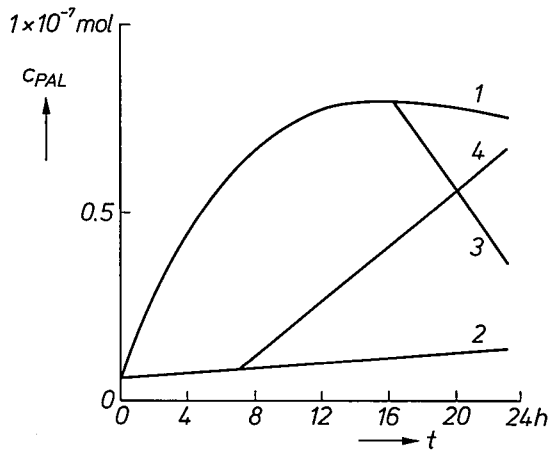


Fig. 4. The result of experiments with hypocotyl segments cut from the gherkin seedling. When the segments are floated on water, so that compounds like cinnamic acid, *p*-coumaric acid and ferulic acid leak out into the water, the concentration of PAL increases (curve *1*). This happens only to a very slight extent when the segments float on a 10^{-3} M solution of cinnamic acid in water (curve *2*). When cinnamic acid is added after 16 hours under the conditions of the first experiment the PAL concentration decreases again (curve *3*). When the cinnamic-acid solution is replaced by water after some time (7 hours) under the conditions of the second experiment, however, the PAL concentration is observed to increase (curve *4*).

then quickly increases until about three hours after the beginning of exposure, when it starts to decrease again (*fig. 3*). A curve like this is found for many other plants.

It has been found that a similar PAL-concentration curve can also be obtained without exposing the seedling to light. With a razor blade a number of small sections across the hypocotyl of a dark-grown gherkin seedling were made, and the segments were floated on water. Measurements showed that the PAL level in these segments had increased after some time in the dark [2]. An analysis of the liquid on which the segments had floated for a few hours also showed that cinnamic acid, *p*-coumaric acid and ferulic acid had diffused from the segments. When one or more of these compounds were added to the water before the segments were floated on it, it was found, depending on the concentration, that the PAL level increased very little, if at all. When they were added later, the increase in the PAL level that had taken place up to that moment was found to stop, and if the concentration of the compound added was sufficiently high, the PAL concentration could even be decreased (*fig. 4*). The same result was thus obtained as upon illumination of the intact seedling. These experiments give a clear indication that the PAL level is controlled by the reaction product (cinnamic acid) formed with this enzyme and the hydroxycinnamic acids derived from this reaction product (*p*-coumaric acid and ferulic acid). These compounds evidently inhibit the production of PAL.

If this hypothesis is correct, then the treatment of intact plants with agents that remove the cinnamic acid or hydroxycinnamic acids must also give an increased PAL level. More or less fortuitously it was found that manganese salts produce such an effect (*fig. 5*) [3]. Gherkin seedlings, whether grown in the dark or in the light, which were sprayed with a concentrated solution of manganese salts (3×10^{-2} M) were found after some time to have reddish hypocotyls. It was possible to show that the colouration was due to tannin-like compounds which are not formed under normal conditions in this plant. Tannins are addition products of cinnamic acid and compounds derived from it. Accordingly it was found that the concentration of *p*-coumaric acid and of ferulic acid in the hypocotyl of seedlings treated with a manganese salt was exceptionally low, and that on the other hand, as was to be expected on the grounds of the hypothesis, the PAL level was much higher than in the untreated seedlings. The effects of many other metal salts have been investigated, but so far these effects have only been observed after the addition of manganese salts.

Next, compounds were sought that bring about other conversions of cinnamic acid or hydroxycinnamic acids in the plant. A link was found in results obtained by J. Daams of the Boekesteyn Agrobiological Laboratory of Philips-Duphar B.V. He found that treatment of seedlings with the herbicide dichlobenil (2.6-dichlorobenzonitrile) leads to intensified lignification. Lignin is a polymer of a number of cinnamic-acid derivatives. We observed that the increased lignification after treatment with dichlobenil in the hypocotyl of the gherkin

[2] G. Engelsma, Planta **82**, 355, 1968.
[3] G. Engelsma, Plant Physiol. **50**, 599, 1972.

seedling was associated with a decreased concentration of hydroxycinnamic acids [4]. As expected, it turned out in this case, just as in the case of plants treated with a manganese salt, that the PAL level was much higher than in the untreated plants (*fig. 6*).

Effects similar to those described above are also observed after exposing the plant to ultraviolet radiation [5]. Cinnamic acid and its derivatives have a double bond in their side chain, as a result of which they can occur either as a *cis*- isomer or as a *trans*- isomer (*fig. 7*). In the plant only the *trans*- isomer is produced. On irradiating the plant with UV the *trans*- isomer compound is converted into the *cis*- isomer, as observed *in vitro*, until an equilibrium state is reached between the two isomers. From experiments with hypocotyl segments floating on solutions of hydroxycinnamic acids, it was found that the *trans*- isomers inhibit the production of PAL much more than the corresponding *cis*- derivatives (*Table I*). The increased PAL level found in the hypocotyls of gherkin seedlings that had been irradiated with ultraviolet might thus find its explanation in this *trans-cis* conversion of cinnamic acid and its derivatives. Not only do the *cis*- derivatives give less inhibition, but the *cis*- compounds are broken down in the plant much faster than the corresponding *trans*- compounds, and this will enhance the effect still further.

In the same context, experiments have been carried out by others [6] who have treated gherkin seedlings with α-amino-oxy-β-phenylpropionic acid, a compound that reduces the activity of PAL. Their starting assumption was that if the formation of the end-product is inhibited, the PAL level must rise. Their results confirmed this assumption.

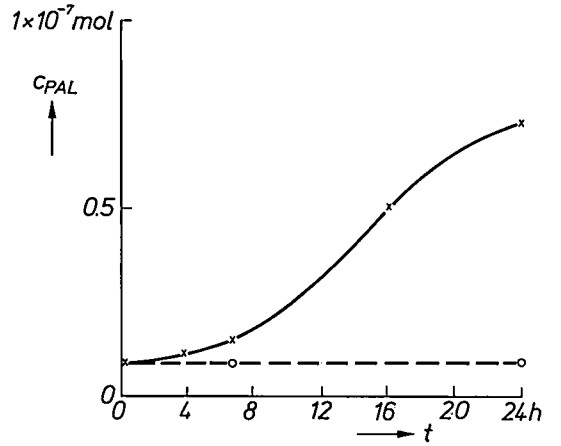


Fig. 5. The result of experiments in which gherkin seedlings, in the dark at 25 °C, were sprayed with a concentrated solution of a manganese salt (0.03 M $MnSO_4$ in water). The PAL concentration in the hypocotyl then shows a marked increase, compared with that of a gherkin seedling sprayed with distilled water (*dashed curve*). The increase is attributed to the conversion of hydroxycinnamic acids into tannin-like products under the action of manganese salts.

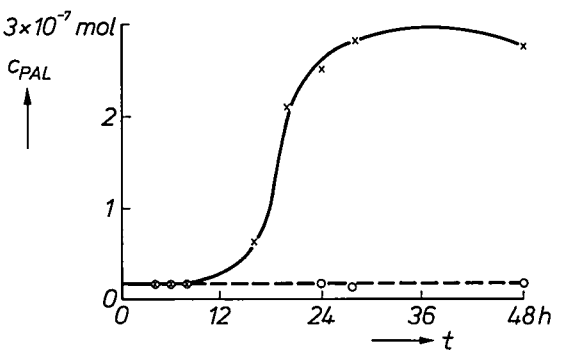


Fig. 6. The result of experiments similar to those in fig. 5, this time with a 10^{-4} M solution of the herbicide dichlobenil. Treatment with dichlobenil leads to lignification, i.e. the formation of lignin-type polymers from cinnamic-acid derivatives.

The mechanism of the photoinduction of phenol synthesis

The data gathered so far now enables a hypothesis to be put forward that will explain the mechanism behind the photoinduction of phenol synthesis.

As already mentioned in I, the phenols are synthesized in the cytoplasm of the cell and in its organelles, such as the chloroplasts. Phenols are transported from these to the vacuoles, so that they have to pass through certain membranes. It was also shown in I that the photoinduction involves at least two pigment systems: the pigment phytochrome, which absorbs in the red and infrared parts of the spectrum, with maxima at 660 and 730 nm; and a pigment that absorbs in the blue, with an absorption maximum at 460 nm, presumably a flavin-type pigment. In recent years it has been increasingly recognized that these pigments affect the permeability of the membranes [7].

The hypothesis advanced here is that the illumination causes in the first instance a reduction in the concentration of cinnamic acid and hydroxycinnamic acids, and it does so in those compartments of the cell where these compounds are synthesized and their synthesis controlled. This change in concentration can be brought about by an increased transport of these compounds from the cytoplasm to the vacuoles under the action of the pigments mentioned above or by the conversion *in situ* of these compounds into others, again possibly under the influence of a pigment.

Fig. 8 shows that on illumination in the UV region the *trans*- isomers of the phenols considered here are converted into the corresponding less inhibitory and

[4] G. Engelsma, Acta bot. neerl. **22**, 49, 1973.
[5] G. Engelsma, Plant Physiol. **54**, 702, 1974.
[6] N. Amrhein and J. Gerhardt, in press.
[7] See for example H. Borthwick, in: K. Mitrakos and W. Shropshire Jr. (ed.), Phytochrome, Academic Press, New York 1972, p. 27.

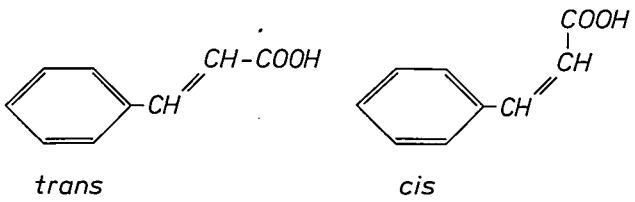


Fig. 7. The *trans-* and the *cis-* isomers of cinnamic acid. The various derivatives of cinnamic acid can also occur in both forms.

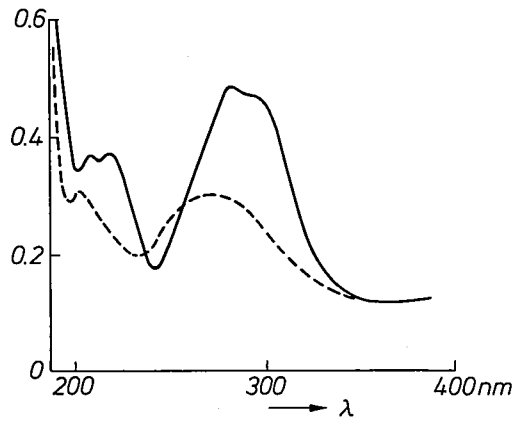


Fig. 8. Absorption spectrum of a solution of *trans-p-*coumaric acid before (*solid line*) and after 15 minutes of UV illumination at 25 °C. The maximum at 310 nm, where the *trans-*hydroxycinnamic acids absorb, has been noticeably reduced by the illumination, owing to conversion into the *cis-* isomer.

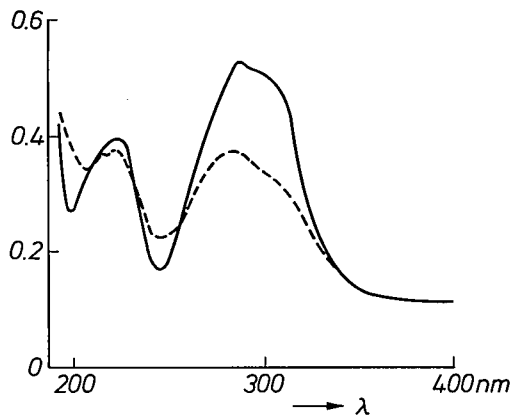


Fig. 9. Absorption spectrum of a solution of *trans-p-*coumaric acid illuminated with blue light, one solution containing riboflavin (*dashed curve*) and the other not. As can be seen, conversion to the *cis-* isomer only takes place in the first case.

less stable *cis-* isomers. It has been demonstrated by experiments *in vitro* that this also happens in the blue part of the spectrum (400-500 nm), where the hydroxycinnamic acids do not absorb, provided a pigment has been added in the form of riboflavin (*fig. 9*) [5]. On illumination with ultraviolet or blue light this initial

Table I. The effect of the conversion of *trans-p-*coumaric acid into the *cis-* isomer on the concentration c_{PAL} of PAL in segments of the hypocotyl of the gherkin seedling. The segments floated on water to which *trans-p-*coumaric acid had been added or a mixture of isomers that was obtained from a preceding UV irradiation of a *p-*coumaric acid solution (at 25 °C a mixture forms of 20% *trans-* isomer and 80% *cis-* isomer). The PAL concentration is given as a percentage of that found in the control experiment in which the segments were floated on distilled water.

Concentration of *p-*coumaric acid in water (M)	Duration of experiment (h)	PAL concentration in hypocotyl	
		for 100% *trans-* isomer	for 20% *trans-* and 80% *cis-* isomer
10^{-5}	6	38	66
2×10^{-4}	16	39	49
5×10^{-4}	24	17	30

reduction in the amount of *trans-*hydroxycinnamic acids in the hypocotyl of the gherkin can also be observed directly with the aid of *in vivo* spectrophotometry. At 310 nm, where *trans-*hydroxycinnamic acids have an absorption maximum, the absorption decreases noticeably in this initial period (*fig. 10*).

On irradiating with red or infrared light, on the other hand, we see with the same method that there is no change in the phenol absorptions during the first hour of illumination. Since cinnamic acid and hydroxycinnamic acid do not absorb in this region of the spectrum, no photochemical conversion would be expected (apart from effects that arise via a pigment). In this case there would only be a migration of phenols from one cell compartment to another, and this cannot be observed by spectrophotometry *in vivo*.

Before giving a more complete description of what happens when a dark-grown seedling is exposed to light, we should note that in addition to the inhibitory effect described above that cinnamic acid and hydroxycinnamic acids have on the production of new PAL, these compounds also have an effect on the *activity* of the PAL already present. This effect is invariably a reduction in activity, called product inhibition, which in all cases is greater in the *trans-* isomer than in the corresponding *cis-* isomer (*Table II*).

When a seedling grown in the dark is exposed to light, the first effect will be the disappearance of *trans-*cinnamic acid and *trans-*hydroxycinnamic acids from the cytoplasm, as described above. In the short term this will lead to increased activity of the PAL present and, after some delay, to the production of more PAL. In the next phase the quantity of PAL is then seen to increase appreciably, accompanied by accelerated production of cinnamic acid and hydroxycinnamic acids. A third phase is then observed in which the accumulated phenols reduce the PAL activity, inhibit the synthesis

of PAL, and in addition set in motion a mechanism, not yet mentioned, in which part of the PAL present is eliminated. The result of all this is that the rate of phenol synthesis now sharply decreases.

One of the functions of the phenols in the plant is probably that, as a result of their absorption in the UV part of the spectrum, they protect vital constituents of the cell, such as nucleic acids and enzymes, from being damaged by UV irradiation. As long as the seedling remains in the dark it produces relatively little phenol — protection from radiation is not in fact necessary in this phase. Illumination causes in the first place a reduction in the phenol concentration in the cytoplasm, and this is the signal for the plant to start producing more phenol. The reaction has the character of an overcompensation; a very large quantity of phenol now accumulates in the seedling.

After a number of hours of illumination so much phenol has accumulated that the plant will synthesize no more phenol on further illumination. In the hypocotyl of the gherkin this insensitivity to light after exposure is also observed, and this persists even after a long period in the dark following the exposure, since little or no phenol is broken down in the dark in the hypocotyl of the gherkin. There is a difference here from the processes that, as we shall presently see, take place in the leaves of *Salvia*: phenol is broken down in these leaves during the night. The sensitivity to light, which decreases during exposure in the course of the day, consequently increases again during the night.

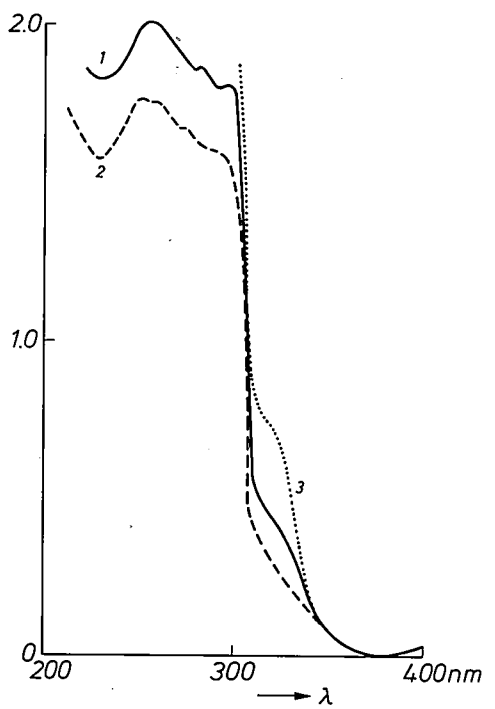


Fig. 10. Changes in the absorption spectrum as a result of ultraviolet irradiation at 25 °C of a hypocotyl of a dark-grown gherkin seedling. Curve *1* is the spectrum before irradiation. Curve *2* is the spectrum immediately after irradiation. Curve *3* is the spectrum 7 hours after irradiation. The decreased absorption at 310 nm shown by curve *2* as compared with curve *1* is the result of a conversion from *trans*-hydroxycinnamic acids into *cis*-hydroxycinnamic acids. The increased absorption at 310 nm shown by curve *3* compared with the two other curves must be attributed to *trans*-hydroxycinnamic acids newly formed after the irradiation.

Table II. The effect of *trans-cis*- isomerization of *p*-coumaric acid, under the influence of UV irradiation, on the activity of PAL, measured *in vitro*. The PAL activity is given as a percentage of that found in a control experiment in which no *p*-coumaric acid was added.

Concentration of *p*-coumaric acid (M)	PAL activity *in vitro*	
	for 100% *trans*- isomer	for 20% *trans*- and 80% *cis*- isomer
10^{-6}	90	97
10^{-5}	76	92
10^{-3}	43	71

This article has so far only dealt with the enzyme PAL in relation to phenol synthesis. In general, apart from cinnamic acid and its derivatives there are many other phenols produced in a plant, some of which have a very complex structure. The synthetic pathway of these phenols may be considered to contain a large number of other enzymes in addition to PAL. Insight into the synthesis of the more complicated phenols has largely been obtained from investigations of plant-tissue cultures [8]. On illumination of tissue cultures it was found that nearly all the other enzymes occurring in addition to PAL in the synthetic pathway of the phenols undergo changes in concentration similar to those of PAL. The manner in which these changes take place as a function of time, however, differs among the enzymes involved in the phenol synthesis.

The effect of temperature on phenol synthesis

A special aspect of the regulating mechanism outlined here is that, to a certain extent, it makes the light-induced phenol synthesis in the plant independent of the ambient temperature [9]. Normally chemical reactions take place twice as fast for every 10° increase of temperature: the value of the temperature coefficient Q_{10} is 2. In the plant the value of Q_{10} is in the region of 1 for many effects.

The regulating mechanism described here for the PAL enzyme consists of a series of sub-processes, where a particular process is started by a preceding one and is terminated by a succeeding process. Increasing the temperature has the effect of speeding up the chemical reactions in these sub-processes, but at the same time it shortens the duration of each sub-process.

Fig. 11 shows the effect at various temperatures of both changes on the quantity of PAL produced in the

hypocotyl of the gherkin seedling. As can be seen, the highest maximum is obtained at the lowest temperature (12.5 °C), although after a longer period of time. Set against this is the fact that at 12.5 °C the phenol synthesis catalysed by PAL takes place much more slowly than at 32 °C. *Fig. 12* shows the effect of temperature variations on the synthesis of hydroxycinnamic acids during continuous illumination and *fig. 13* shows the effect during alternating illumination. In both cases the accumulation is initially fastest at 32 °C, but larger quantities are ultimately formed at the lower temperatures. In both cases, however, the variations in the accumulated quantities are considerably smaller than one would expect from the temperature differences. Particularly in the region between 18 and 32 °C the quantities of phenol accumulated in twenty-four hours differ very little with temperature; in the ratio of 2½ hours of light to 21½ hours of darkness the effect of temperature is in fact completely compensated.

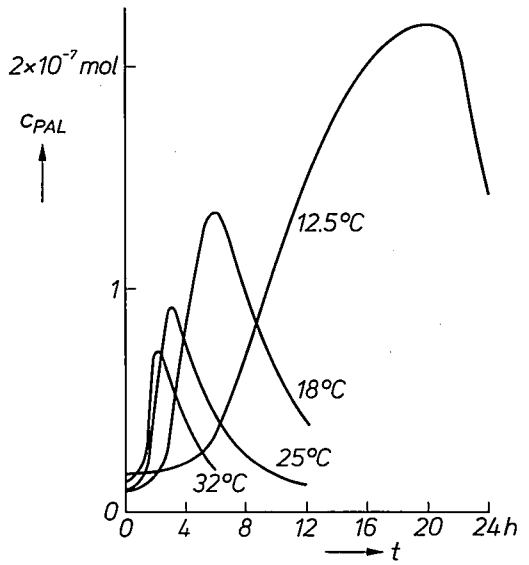


Fig. 11. The concentration c_{PAL} of PAL in the hypocotyl at four different temperatures, as a function of time under continuous irradiation with blue light (150 μW/cm²). The measurements were made on dark-grown gherkin seedlings that were three days old at the beginning of the exposure. The higher the temperature the higher the maximum, but it is not reached until after a longer time.

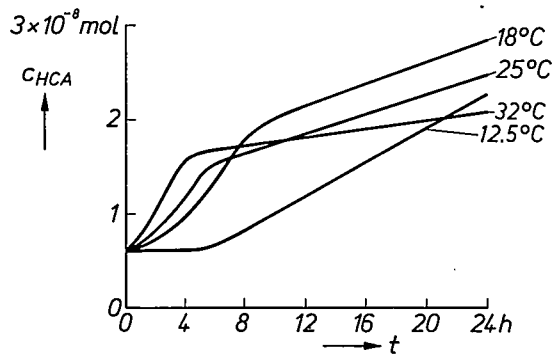


Fig. 12. The concentration c_{HCA} of hydroxycinnamic acids given in moles per hypocotyl under the conditions of fig. 11.

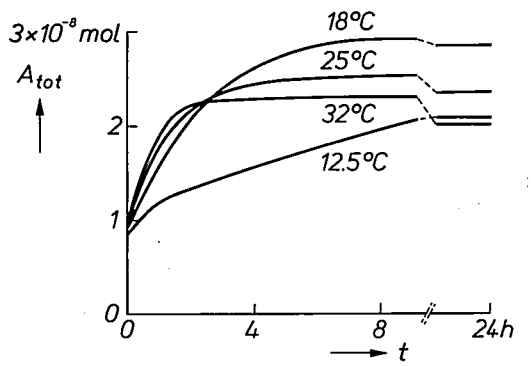


Fig. 13. The total quantity A_{tot} of hydroxycinnamic acids accumulated per day (24 h) in a hypocotyl of a gherkin seedling, illuminated for part of the day and kept in the dark for the rest of the time, as a function of the exposure time t. The experimental conditions were the same as in fig. 11.

Experiments with *Salvia occidentalis*

All the experiments described above were carried out with gherkin seedlings grown in the dark. Compared with the fully grown plant a seedling is a fairly simple system. Even so, it would also be desirable to have some understanding of the manner in which phenol synthesis is controlled in the fully grown plant, since it has a more direct bearing on the practical problems of agriculture and horticulture. One of the problems of the fully grown plant is that it has leaves of different age and life history, which implies differences in the composition of the group of phenols contained in them. The phenols occur in a very wide variety, and in addition some of them exist freely in the cytoplasm and others are bound to certain cell structures. The classical method of phenol determination, consisting of extraction, chromatography and spectrophotometry, requires a relatively large amount of material and is time-consuming. It is therefore not so suitable when the objective is to study changes in phenol composition as a function of time, especially when it is to be done for each separate leaf of the plant. A faster method was therefore sought that would also use up as little plant material as possible. This method is described below in the section headed 'phenol determination *in situ*'. All the experiments discussed were performed on *Salvia occidentalis*, a short-day plant which has already been the subject of extensive studies at Philips Research Laboratories, particularly in connection with the effect of different kinds of light on flower induction [10].

[8] J. Ebel and K. Hahlbrock, Eur. J. Biochem. **75**, 201, 1977, and the earlier literature mentioned there.
[9] G. Engelsma, Acta bot. neerl. **17**, 499, 1968.
G. Engelsma, Planta **90**, 133, 1970, and **91**, 246, 1970.
[10] G. Meijer, Acta bot. neerl. **6**, 395, 1957; **7**, 801, 1958; **8**, 189, 1959.
G. Meijer and R. van der Veen, Acta bot. neerl. **6**, 429, 1957, and **9**, 220, 1960.

Phenol determination in situ

From a chromatographic investigation on aqueous extracts of leaves of *Salvia occidentalis* it was found that the leaves contained some twenty different phenols, most of them in very small quantities. At least 90% of the total amount consists of chlorogenic acid and two isochlorogenic acids. These are esters of caffeic acid, i.e. an *o*-dihydroxycinnamic acid, with quinic acid (*fig. 14*).

These compounds are particularly interesting because of their effect on the metabolism of the growth hormone indoleacetic acid, which will be examined in more detail later. At first attempts were made to determine by direct spectrophotometry the changes in concentration which these compounds undergo in a leaf as a result of a particular illumination scheme. This proved to be not very feasible, because these phenols absorb in a region in which plant pigments, such as chlorophyll, carotene and riboflavin, also exhibit strong absorption.

The property of *o*-dihydroxycinnamic acid derivatives that in a basic environment they oxidize in the air to from reddish compounds was therefore put to use. The absorption spectra of these compounds are quite different from those of the plant pigments mentioned above. The combination of these oxidation products makes the leaf greyish-black; the shade is darker the greater the phenol concentration in the leaf (*fig. 15*). In the experiments now described the oxidation was

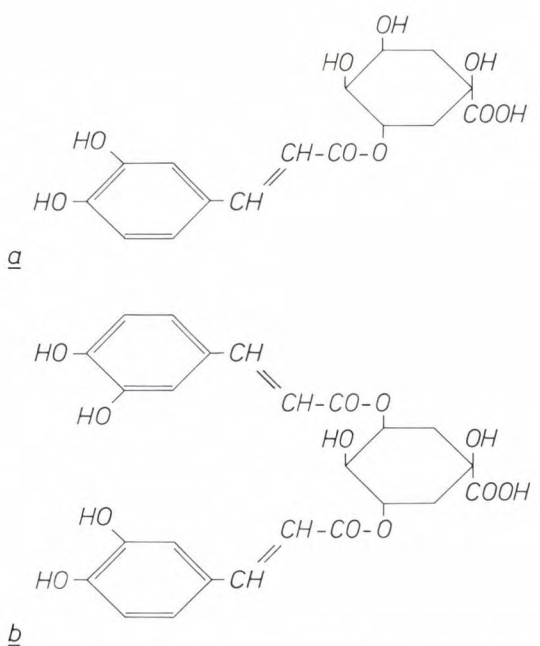


Fig. 14. Structural formulae of chlorogenic acid (*a*) and an isochlorogenic acid (*b*). Compounds of this kind are studied more closely in the article because of their possible effect on flower induction.

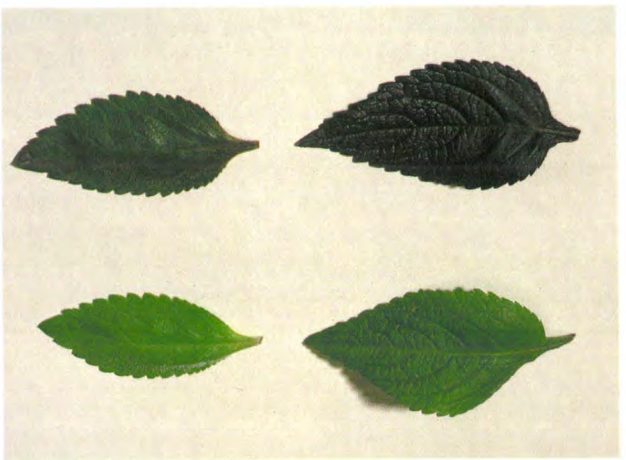


Fig. 15. The visual effect of keeping *Salvia* leaves in ammonia vapour for some time after different kinds of irradiation treatments. *Left:* two leaves of a leaf pair from a *Salvia* grown in SD; the one at the top was treated with ammonia. *Right:* similar experiments with a leaf pair of *Salvia* grown in LD. The LD leaf has a much higher concentration of phenol (rank number *8* in fig. 17) than the SD leaf (rank number *2*). This difference becomes visible as the difference in the blackening of the leaves after the ammonia treatment.

produced by exposing a small part of the leaf to ammonia vapour for two hours. The part of the leaf thus treated was then measured in a spectrophotometer against another part removed at the same time and kept for two hours in an atmosphere saturated with water vapour. The result of the measurement is a difference spectrum which, in principle, shows only the absorption of the phenol oxidation products.

Fig. 16 shows examples of such spectra, which are the result of an experiment with a *Salvia* plant that had been subjected to a diurnal treatment consisting of 8 hours of exposure to light followed by 16 hours in the dark. During the day the absorption of the oxidation products increases: under the action of the illumination phenol is synthesized and accumulates in the leaf. During the following dark period the absorption decreases: a certain amount of phenol is broken down.

In later experiments the *in situ* determination of phenols was extended on the basis of the principle described above. The results showed that the difference spectrum for the phenol oxidation products in a young leaf only has a peak at 395 nm, but the spectrum increases in complexity with the age of the leaf (*fig. 17*). This is related to an increasing phenol concentration in the leaf, the effect of which is to change the composition of the oxidation products. New peaks are seen successively at 450 and 500 nm, and 'shoulders' appear at 565 and 700 nm. No matter what the conditions are in which the plant is grown, this is always the sequence of the changes in the difference spectrum of a leaf treated with ammonia vapour. Comparison of the chromato-

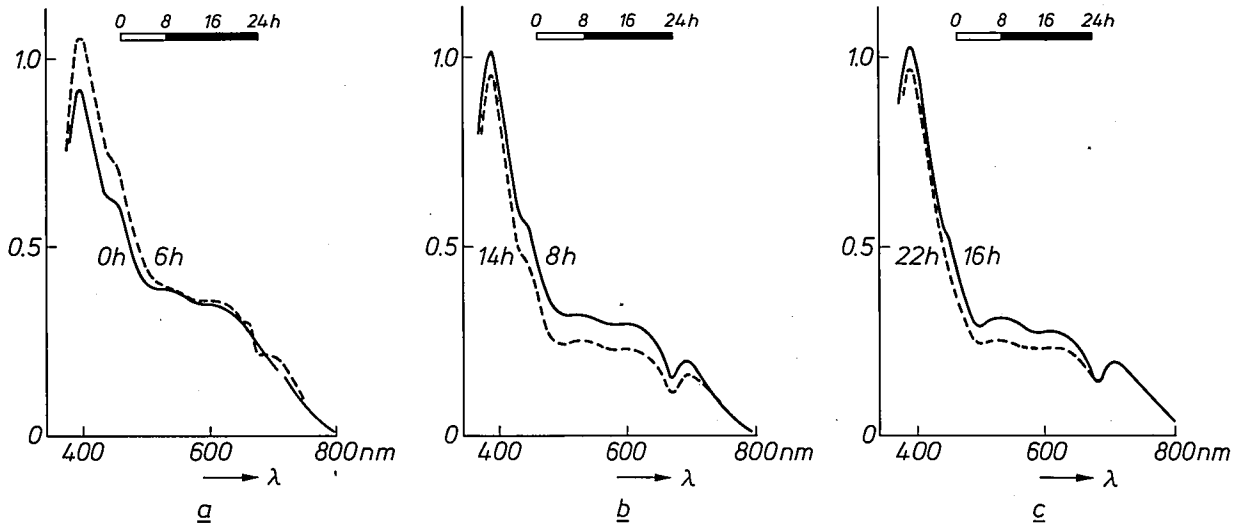


Fig. 16. Results of experiments with young *Salvia occidentalis* plants which were alternately illuminated from time zero for 8 hours and then kept in the dark for the next 16 hours. On the sixth day of this treatment, absorption spectra were measured *in situ* by the method described in the text, to determine the concentration of chlorogenic acid and isochlorogenic acids, after the following time intervals: after 0 and 6 h (*a*), after 8 and 14 h (*b*) and after 16 and 22 h (*c*). The height of the absorption maximum at 450 nm corresponds to the concentration of the phenolic compounds mentioned. It appears from the figures that the concentration of these phenols increases during the day and decreases during the night.

gram of a leaf extract with the difference spectrum of a small part of the same leaf that has been treated with ammonia vapour reveals that a particular spectrum always corresponds to a particular phenol concentration in the leaf. Instead of carrying out a complete analysis, it is therefore sufficient to record a difference spectrum. To allow the results to be presented numerically, eight 'standard' spectra were selected from among a large number of difference spectra of leaf segments treated with ammonia vapour. They were chosen in such a way as to make their rank number an increasing linear function of the total quantity of chlorogenic acid and isochlorogenic acids. A particular spectrum number thus corresponds to a particular stage in the development of the leaf with regard to its phenol concentration [11].

Phenol synthesis in short-day and long-day illumination

The critical day-length for *Salvia occidentalis* is 12½ hours, which is to say that when this plant is illuminated for less than 12½ out of every 24 hours, and the ensueing dark period is thus longer than 11½ hours, it will begin to flower after some time [10]. After an uninterrupted series of 12 such short days the flower induction can no longer be arrested — for example by means of longer light periods. After 21 days the emergence of the flower primordia can be seen under the microscope — the first signs at the growth apices that the plant is about to flower.

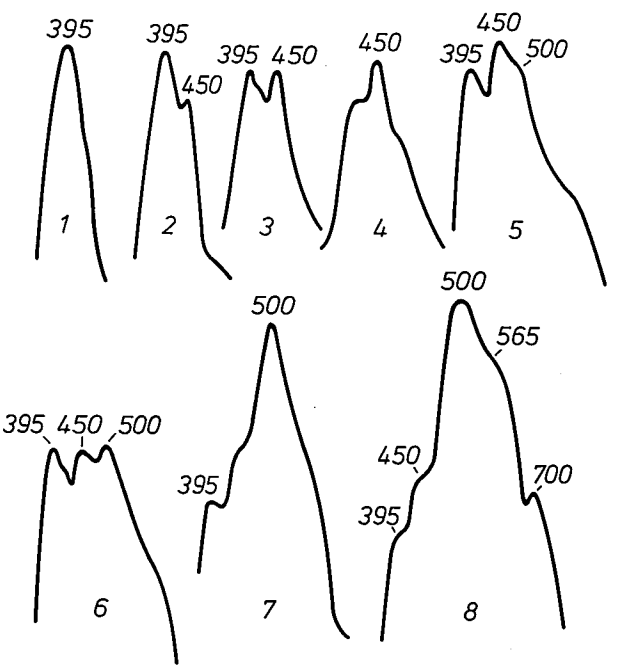


Fig. 17. During the growth of *Salvia occidentalis* changes take place in the concentration of phenols in the leaves. These changes always follow the same pattern. The figure shows characteristic features of eight absorption spectra of the phenol oxidation products determined *in situ*, marking different stages of growth (wavelengths in nm). These 'standard spectra' were chosen such that in the given sequence the concentration of chlorogenic acid and of the two isochlorogenic acids (fig. 14) exhibits a linear increase. Specification of one of the rank numbers from *1* to *8* is therefore sufficient to indicate the concentration of these three compounds.

[11] Technical details of the *in situ* phenol determination are described in G. Engelsma, Plant Physiol., in press.

Short-day or long-day illumination also makes a difference to the accumulation of phenols in the leaves of *Salvia occidentalis*. *Fig. 18* shows that with long-day illumination (LD) substantially more phenol is accumulated per leaf than with short-day illumination (SD), the phenols mainly being the three derivatives of caffeic acid mentioned earlier.

In SD a relatively high phenol concentration is found only in the lower pairs of leaves. The reason is to be found in the conditions under which the experimental plants are grown. This is done by keeping shoots in LD until the fourth leaf pair begins to develop. The lower leaf pairs have thus been longest in LD.

In SD in particular the phenol concentration in the leaves remains relatively low up to the beginning of flowering. This does not imply that no phenol is being synthesized: if the night is sufficiently long, nearly all the phenol produced in the plant as a result of the illumination is broken down again. *Fig. 19* shows that this is the case up to day-lengths of about 12 hours. Up to that day-length the phenol concentration at the end of the night is practically independent of the length of the daily light period. If the plant is illuminated for longer than 12 out of every 24 hours, the quantity of phenol per leaf then increases with the duration of the illumination. The night is then evidently too short for the whole quantity of phenol that has accumulated during the day to be removed. The possible bearing that this may have on flower induction will be examined later.

It is also found that the phenol concentration begins to increase again from the moment the plant starts to flower (in SD) or when the growth slows down (in LD). In this latter case the high phenol concentration in the leaves might also be the *cause* of the slower rate of plant growth.

The effect of a brief illumination during the night

It was stated above that *Salvia occidentalis* will start to flower when the daily period of illumination is shorter than 12½ hours, and the night thus longer than 11½ hours. If, however, the night is interrupted half-way through by a short period of illumination (15 minutes), the effect is to inhibit the flower induction, especially if the illumination is with red light [10].

It was found accordingly that when the night is interrupted in this way the phenol concentration in the leaves of *Salvia occidentalis* increases in the second half of the night (*fig. 20*), whereas when the night is not interrupted, and the other conditions are the same, the reduction that began in the first half continues in the second half of the night (see fig. 16*b* and *c*).

This effect of red light on both induction of flowering and phenol synthesis can largely be eliminated by treating the *Salvia* plants with infrared radiation during the next 15 minutes. This is an indication that phytochrome is the pigment that mediates this reaction (see I). It was also shown in I that a short illumination with red light is very effective in its action on phenol synthesis, and that lengthening the period of illumination makes little further contribution to the phenol synthesis.

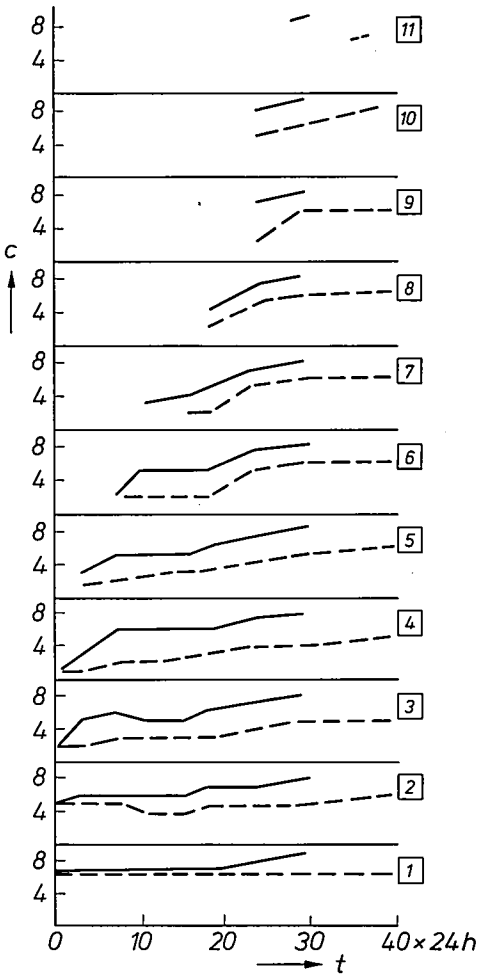


Fig. 18. The concentration *c* of the three phenols under consideration plotted in the form of rank numbers as a function of time *t* for 11 leaves, each from a different leaf pair of *Salvia occidentalis* (numbered *1* to *11* from the lowest leaf pair). During the experiment the plants were illuminated for 8 out of every 24 hours and kept in the dark for the remaining 16 hours (*dashed line*), or illuminated for 16 hours out of every 24 and then kept in the dark for 8 hours (*solid line*). The concentration of the phenols was consistently determined at the end of the night. As can be seen, the short-day treatment results in a lower phenol concentration in the leaves than the long-day treatment. It is assumed that a low phenol concentration at the end of the night is a prerequisite for the flowering of the short-day plant *Salvia occidentalis*.

We should add here that the spectacular effect of a night interruption is not only found with a short-day plant like *Salvia*. For instance, a long-day plant that receives a shorter daily illumination than its critical day-length can still be brought to flower by a similar interruption of the night [12].

Phenol synthesis and flower induction

The growth hormone indoleacetic acid (IAA) plays an important role not only in the elongation of the plant but also in its flowering. Under inducing conditions the flowering of many SD plants can be prevented by treating the plants at the beginning of the night with IAA. In the case of the soya-bean plant, like *Salvia occidentalis* an SD plant, it has been found that the day-length has a significant effect on the IAA concentration in the leaves [13]. This is higher in LD than in SD. It was also observed that the concentration was high during the light period and that it decreased in the course of the dark period. Another finding was that a night interruption also led to a higher IAA concentration. This is a picture that completely agrees with what has been said above in connection with the caffeic-acid derivatives. Since it is known that these *o*-dihydroxy compounds have a protective effect on the IAA — they inhibit the breakdown of IAA by peroxidases — then it seems that there is a causal connection here: the increased concentration of the *o*-dihydroxy compounds would lead to an increased concentration of IAA, and this in turn would inhibit the induction of flowering [14].

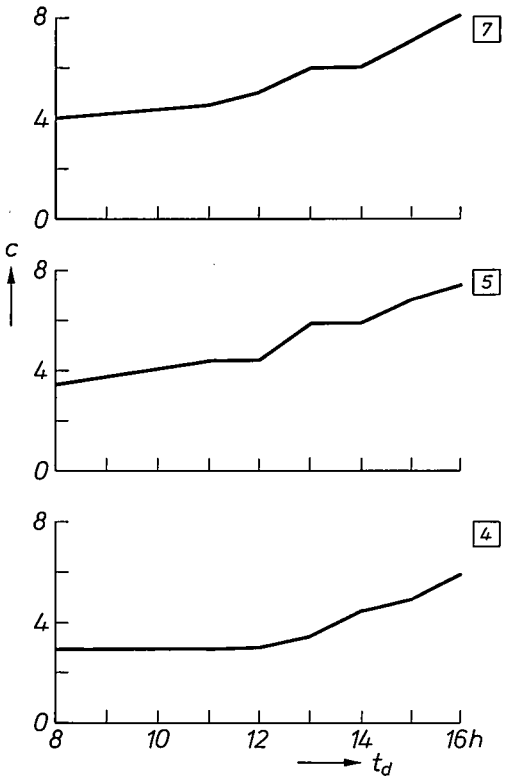


Fig. 19. The phenol concentration is plotted for leaves *4*, *5* and *7* in the same way as in fig. 18, but now as a function of the day-length t_d. At a day-length longer than about 12 hours the phenol concentration is seen to start increasing. This supports the assumption that a low phenol concentration at the end of the night is a prerequisite for the flowering of *Salvia occidentalis*.

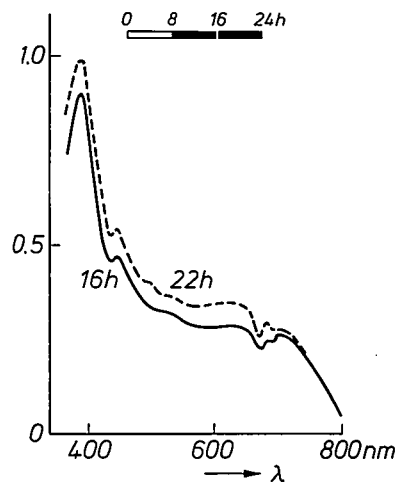


Fig. 20. Illustrating the effect of a night interruption for 15 minutes by irradiation with red light. The experiment was carried out with *Salvia occidentalis* treated for 5 days with 8 hours of exposure to light out of every 24, in each case followed by 16 hours in the dark. The interruption in the night took place after 8 hours of darkness. Compare the phenol concentrations measured at the times 16 and 22 h of a plant treated in this way with those found under otherwise identical conditions but without night interruption (fig. 16*c*). As a result of the interruption the phenol concentration increases in the second half of the night, whereas it decreases in the uninterrupted darkness.

It is not claimed that the foregoing provides a *complete* explanation for the effect of the day-length on flowering in short-day plants. It should however be pointed out that the mechanism outlined here offers a possible explanation for certain other aspects of flower induction. Light-sensitive and light-insensitive phases play a role, for example in photoperiodism in plants [12]. As already indicated on page 94, a connection does indeed exist between the sensitivity to light and the phenol concentration in the cytoplasm. Also, the critical day-length is relatively independent of temperature [12]. Since it has been found that the temperature dependence of phenol synthesis in the leaves of *Salvia occidentalis* is the same as in the hypocotyl of the gherkin seedling, it can be assumed that the same mechanism of temperature compensation also functions in the fully grown plant.

Prospects

Programmed illumination is a method of influencing the development of a plant without unduly disturbing the environment. Interruption of the dark period opens up particularly interesting prospects. Not only does it provide a means of bringing about faster and greater production of those plant organs which man has most need of — in the one plant the leaf, in the other the flower or the fruit — it also makes it possible to in-

[12] See for example D. Vince-Prue, Photoperiodism in plants, McGraw-Hill, London 1975.

[13] S. Kiyosawa, Proc. Crop Sci. Soc. Japan **29**, 163, 1960.

[14] Earlier investigations concerning the relation between phenol synthesis and flower induction are described in: M. Zucker, C. Nitsch and J. P. Nitsch, Amer. J. Bot. **52**, 271, 1965; A. O. Taylor, Plant Physiol. **40**, 273, 1965; G. Engelsma, Acta bot. neerl. **18**, 347, 1969.

fluence the composition of certain organs. The experiments with *Salvia occidentalis* described here demonstrate the manner in which the phenol content in the leaves can to some extent be manipulated. For practical purposes changes in the balance between fats, carbohydrates and proteins come most naturally to mind. The scope for further investigation here is very considerable.

Summary. During the first few hours in which a seedling is exposed to light there is a greatly accelerated accumulation of phenols. The first part of this article deals with the mechanism underlying this effect, as demonstrated by experiments with gherkin seedlings. It is shown that the production of the key enzyme of the phenol synthesis, phenylalanine ammonia-lyase, PAL, is determined by the concentrations of the reaction product cinnamic acid and its phenolic derivatives *p*-coumaric acid and ferulic acid. Treatments of the seedling that bring about a further conversion of these phenols, such as spraying with solutions of a manganese salt or with the herbicide dichlobenil, or irradiation with UV or blue light, all result in a higher PAL level. These observations have led to the hypothesis that visible light causes local reductions in the phenol concentration by the conversion of phenols or by influencing the internal transport of the phenols within the cell, causing a local increase in the PAL level. This in its turn gives rise to increased phenol synthesis, the effect of which is to protect the plant from radiation damage. The first part of the article concluded by showing that the regulating mechanism of phenol synthesis functions in such a way as to largely compensate for temperature differences.

The second part deals with the way in which the day-length affects the synthesis of phenols — in particular, three caffeic-acid derivatives: chlorogenic acid and two isochlorogenic acids — in the leaves of *Salvia occidentalis*, and discusses its possible relation to the induction of flowering in this short-day plant. First a new method is discussed by which the phenol concentration in a leaf can be determined without prior extraction. It is based on oxidation *in situ* of the phenols in an ammonia-vapour atmosphere, followed by difference spectrophotometry. It is shown with this method that the caffeic-acid derivatives and other phenols accumulate in the leaves as a result of illumination, and that they are broken down again during a succeeding dark period. The quantity synthesized increases with increasing day-length. If the night is shorter than 12 hours, not all of the phenol synthesized during the day is broken down, so that the concentration increases daily. If the night is interrupted half-way through by a short illumination, the phenol concentration during the second half of the night increases instead of decreasing further, so that here again the phenol concentration increases daily. The relation to the flower induction is thought to lie in the fact that *o*-dihydroxy derivatives, such as caffeic acid, have a protective effect on the growth hormone indoleacetic acid, and that an increased concentration of this hormone has the effect of inhibiting flower induction. This could explain why *Salvia occidentalis* does not flower under long-day illumination, nor on short-day illumination with a brief interruption in the night.

Scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands *E*
Philips Research Laboratories, Redhill, Surrey, England *R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany *H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium *B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. *N*

G. A. A. Asselman: Thermal energy storage unit based on lithium fluoride.
Energy Conversion **16**, 35-47, 1976 (No. 1/2). *E*

E. F. Assmus, Jr. (Lehigh University, Bethlehem, Pa., U.S.A.), **J.-M. Goethals & H. F. Mattson, Jr.** (Syracuse University, N.Y., U.S.A.): Generalized *t*-designs and majority decoding of linear codes.
Information and Control **32**, 43-60, 1976 (No. 1). *B*

S. H. A. Begemann & P. Jansen: Relative cost-performance of various solar based power supply packages.
Acta Electronica **20**, 197-204, 1977 (No. 2). *E*

J. Bellugue: Considérations générales sur les concentrateurs simples pour cellules photovoltaïques.
Acta Electronica **20**, 153-164, 1977 (No. 2). *L*

C. Belouet: Review of European proposals for vapour growth.
Material sciences in space, Proc. 2nd Eur. Symp., Frascati 1976, pp. 245-248. *L*

C. Belouet: Fluid-dynamics problems connected with crystal growth.
Material sciences in space, Proc. 2nd Eur. Symp., Frascati 1976, pp. 283-289. *L*

G. Bergmann: Die Zündung von Gasentladungslampen.
Lichttechnik **28**, 493-496, 1976 (No. 12). *A*

K. Bethe: Novel precision thermistor from intrinsic germanium.
IEEE Trans. **IECI-23**, 420-424, 1976 (No. 4). *H*

J. W. M. Biesterbos & A. G. Dirks: Amorfe metallische systemen.
Polytechn. T. Werktuigbouw **31**, 145-150, 1976 (No. 3). *E*

G. M. Blom: Native defects and stoichiometry in GaAlAs.
J. Crystal Growth **36**, 125-137, 1976 (No. 1). *N*

R. D. Boehnke: Anwendung und Kontrolle des Ionenätzens in Hochfrequenz-Zerstäubungsanlagen.
Vakuum-Technik **25**, 195-199, 1976 (No. 7). *H*

J. van den Boomgaard, R. A. J. Born, G. Frens & J. W. A. Nelissen: The λ-R relationship in the directional decomposition of ε-FeAl.
J. Crystal Growth **35**, 311-314, 1976 (No. 3). *E*

J. van den Boomgaard, A. M. J. G. van Run & J. van Suchtelen: Piezoelectric-piezomagnetic composites with magnetoelectric effect.
Ferroelectrics **14**, 727-728, 1976 (No. 1/2). *E*

M. R. Boudry: A simple conversion formula for type 'T' (copper-constantan) thermocouple readings.
J. Physics E **9**, 1064-1065, 1976 (No. 12). *R*

D. J. Breed & R. P. Kramer: Stable and unstable surface state charge in thermally oxidized silicon.
Solid-State Electronics **19**, 897-907, 1976 (No. 11). *E*

R. Brehm: Het ultrasoon boren van gaatjes met een diameter kleiner dan 200 μm.
Mikroniek **16**, 91-95, 1976 (No. 5). *E*

J.-J. Brissot: Silicium pour photopiles solaires.
Acta Electronica **20**, 101-116, 1977 (No. 2). *L*

J. Brokken-Zijp & H. J. L. Bressers: Sensitization of alkylaryldiazosulphides.
Mol. Photochem. **7**, 359-388, 1976 (No. 3). *E*

K. H. J. Buschow, J. H. N. van Vucht & W. W. van den Hoogenhof: Note on the crystal structure of the ternary rare earth-3*d* transition metal compounds of the type RT_4Al_8.
J. less-common Met. **50**, 145-150, 1976 (No. 1). *E*

F. M. A. Carpay: Reply to the comment of J. D. Livingston on 'The λ-R relationship in the directional decomposition of ε-FeAl' by J. van den Boomgaard, R. A. J. Born, G. Frens and J. W. A. Nelissen.
J. Crystal Growth **35**, 316-317, 1976 (No. 3). *E*

J. E. Curran, R. V. Jeanes & H. Sewell: A technology of thin-film hybrid microwave circuits.
IEEE Trans. **PHP-12**, 304-309, 1976 (No. 4). *R*

P. Delsarte: Properties and applications of the recurrence $F(i+1, k+1, n+1) = q^{k+1}F(i, k+1, n) - -q^kF(i, k, n)$.
SIAM J. appl. Math. **31**, 262-270, 1976 (No. 2). *B*

P. A. Devijver: Error and reject tradeoff for nearest neighbor decision rules.
Preprints NATO ASI Signal processing, Portovenere 1976, Part 2, pp. 22/1-22/24. *B*

P. A. Devijver: Error-reject relationships in nearest neighbor decision rules.
3rd Int. Joint Conf. on Pattern recognition, Coronado, California, 1976, pp. 255-259. *B*

C. Dierieck: Simultaneous approximation of vector-valued functions.
J. Approx. Theory **18**, 74-85, 1976 (No. 1). *B*

J. A. W. van der Does de Bye & L. Blok: Room temperature minority carrier lifetime and efficiency of p-type $GaAs_{1-x}P_x$.
J. Luminescence **14**, 101-113, 1976 (No. 2). *E*

H. C. Donkersloot & J. H. N. van Vucht: The crystal structure of IrLi, Ir_3Li and $LiRh_3$.
J. less-common Met. **50**, 279-282, 1976 (No. 2). *E*

H. Durand: Introduction *(to issue on Solar energy)*.
Acta Electronica **20**, 97-99, 1977 (No. 2). *(In English and in French.)* *L*

J. M. L. Engels & A. J. M. van Kimmenade: The mobility of excess electrons in liquid methane.
Chem. Phys. Letters **42**, 250-252, 1976 (No. 2). *E*

E. Fabre: Les structures des photopiles: homojonctions, hétérostructures ou hétérojonctions?
Acta Electronica **20**, 117-131, 1977 (No. 2). *L*

A. Farrayre, J. Grau, B. Kramer & J. Magarshack: Design integrated amps with hi-lo Impatt diodes.
Microwaves **15**, No. 10, pp. 42, 44, 46 & 47, Oct. 1976. *L*

A. J. Fox & T. M. Bruton (Morgan Thermic Technological Centre, Stourport-on-Severn, Worcs., England): Measurements of the electro-optic and electrogyratory coefficients in the optically active compounds $Bi_{12}TiO_{20}$ and $Bi_{40}Ga_2O_{63}$.
Electro-Optics/Laser International '76 UK, Proc. Conf. Brighton 1976, pp. 131-135. *R*

A. J. Fox, E. D. Fletcher & T. M. Bruton (Morgan Thermic Technological Centre, Stourport-on-Severn, Worcs., England): Bismuth titanate optical storage plane.
Electro-Optics/Laser International '76 UK, Proc. Conf. Brighton 1976, pp. 209-219. *R*

G. Frank, L. Brock & H. D. Bausen: The solubilities of Sn in In_2O_3 and of In in SnO_2 crystals grown from Sn-In melts.
J. Crystal Growth **36**, 179-180, 1976 (No. 1). *A*

J. A. G. Halé: Dot spacing modulation for the production of pseudo grey pictures.
Proc. S.I.D. **17**, 63-74, 1976 (No. 2). *R*

J. A. G. Halé: Detection of elementary features in a picture by nonlinear local numerical processing.
3rd Int. Joint Conf. on Pattern recognition, Coronado, California, 1976, pp. 764-768. *R*

W. Hoekstra & V. H. C. Evers: Een geautomatiseerd microfiche-archiefsysteem voor persoonlijk gebruik.
Informatie **18**, 494-501, 1976 (No. 9). *E*

L. P. J. Hoogeveen, F. W. Willmott & R. J. Dolphin: An automatic monitor for the analysis of organochlorine compounds in milk.
Z. anal. Chemie **282**, 401-406, 1976 (No. 4). *E, R*

A. P. Hulst & J. O. Zwolsman (Philips Elcoma Division, Eindhoven): Adaptieve regeling bij het weerstandlassen van kleine componenten.
Polytechn. T. Werktuigbouw **31**, 383-388, 1976 (No. 6). *E*

E. Kauer & T. Thalhammer: De mogelijkheden van zonne-energie.
Polytechn. T. Elektr. **32**, 62-76, 1977 (No. 2). *A, E*

J. L. W. Kessels: Parallel programming concepts in a definitional language.
SIGPLAN Not. **11**, No. 10, pp. 19-31, Oct. 1976. *E*

W. L. Konijnendijk & J. M. Stevels: Raman scattering measurements of silicate glasses and compounds.
J. non-cryst. Solids **21**, 447-453, 1976 (No. 3). *E*

W. L. Konijnendijk (Philips Lighting Division, Eindhoven) **& J. M. Stevels** (Eindhoven University of Technology): Electrical conduction and internal friction of borosilicate glasses in relation to their structure.
Verres Réfract. **30**, 519-526, 1976 (No. 4).

P. K. Larsen & R. Metselaar: Electrical properties of yttrium iron garnet at high temperatures.
Phys. Rev. B **14**, 2520-2527, 1976 (No. 6). *E*

P. K. Larsen & J. M. Robertson: Changes in optical absorption in iron garnet films due to impurity incorporation.
Appl. Phys. **11**, 259-263, 1976 (No. 3). *E*

D. A. Lucas & R. P. Vincent: A precision approach monitor.
The future of aircraft all-weather operations?, Int. Conf. London 1976 (IEE Conf. Publn No. 147), pp. 80-83. *R*

H. H. van Mal, K. H. J. Buschow & A. R. Miedema: Hydrogen absorption of rare-earth (3*d*) transition intermetallic compounds.
J. less-common Met. **49**, 473-475, 1976. *E*

J. Matthys: *A*-stable linear multistep methods for Volterra integro-differential equations.
Numer. Math. **27**, 85-94, 1976 (No. 1). *B*

D. Meyer-Ebrecht: Trends in der elektronischen Röntgenbilderzeugung und -verarbeitung.
Röntgenstrahlen **35**, 2-10, 1976. *H*

J. Michel: Les photopiles au silicium monocristallin à usage terrestre: état de l'art et perspectives d'avenir.
Acta Electronica **20**, 133-151, 1977 (No. 2). *L*

A. R. Miedema, K. H. J. Buschow & H. H. van Mal: Which intermetallic compounds of transition metals form stable hydrides?
J. less-common Met. **49**, 463-472, 1976. *E*

A. Mircea, A. Mitonneau, L. Hollan & A. Brière: Out-diffusion of deep electron traps in epitaxial GaAs.
Appl. Phys. **11**, 153-158, 1976 (No. 2). *L*

A. Molenaar, G. H. C. Heynen & J. E. A. M. van den Meerakker: Physical development by copper complexes using ferrous-ferric ions as a redox system.
Photogr. Sci. Engng. **20**, 135-139, 1976 (No. 3). *E*

A. E. Morgan & H. W. Werner: On the abundance of molecular ions in secondary ion mass spectrometry.
Appl. Phys. **11**, 193-195, 1976 (No. 2). *E*

B. J. Mulder: Preparation of ultra-thin mica windows.
J. Physics E **9**, 724-725, 1976 (No. 9). *E*

G.-J. Naaijer: Problèmes d'adaptation des photopiles en vue d'applications terrestres.
Acta Electronica **20**, 165-185, 1977 (No. 2). *L*

D. Paterson, A. R. Turner-Smith, J. A. G. Halé & P. Saraga: A television-controlled printed circuit board drilling machine.
Proc. 16th Int. Machine Tool Design and Res. Conf., Manchester 1975, pp. 137-141; 1976. *R*

D. Polder & M. F. H. Schuurmans: Resonance fluorescence from a $j = 1/2$ to $j = 1/2$ transition.
Phys. Rev. A **14**, 1468-1471, 1976 (No. 4). *E*

G. Renelt: Easily decodable runlength code (e.d.r.c.) for source encoding of black-and-white facsimile pictures.
Electronics Letters **12**, 633-634, 1976 (No. 23). *H*

E. D. Roberts: Electron-sensitive film-forming materials and their uses in semiconductor technology.
Vacuum **26**, 459-467, 1976 (No. 10/11). *R*

P. Saraga & D. R. Skoyles: An experimental visually controlled pick and place machine for industry.
3rd Int. Joint Conf. on Pattern recognition, Coronado, California, 1976, pp. 17-21. *R*

W. Schäfer, G. Will (both with Universität Bonn) **& K. H. J. Buschow:** The symmetries of magnetic structures in rare earth tetraborides.
J. Magnetism and magn. Mat. **3**, 61-66, 1976 (No. 1/2). *E*

C. W. J. Schiepers (Institute for Perception Research, Eindhoven): Global attributes in visual word recognition: Part 1. Length perception of letter strings, Part 2. The contribution of word length.
Vision Res. **16**, 1343-1349 & 1445-1454, 1976 (Nos. 11 & 12).

R. J. Seymour & F. Zernike: Infrared radiation tunable from 5.5 to 18.3 μm generated by mixing in $AgGaS_2$.
Appl. Phys. Letters **29**, 705-707, 1976 (No. 11). *N*

M. Sintzoff: Eliminating blind alleys from backtrack programs.
Automata, languages and programming, 3rd Int. Coll., Edinburgh 1976, pp. 531-557. *B*

J. W. Slotboom & H. C. de Graaff: Measurements of bandgap narrowing in Si bipolar transistors.
Solid-State Electronics **19**, 857-862, 1976 (No. 10). *E*

J. L. Sommerdijk, J. A. W. van der Does de Bye & P. H. J. M. Verberne: Decay of the Ce^{3+} luminescence of $LaMgAl_{11}O_{19}:Ce^{3+}$ and of $CeMgAl_{11}O_{19}$ activated with Tb^{3+} or Eu^{3+}.
J. Luminescence **14**, 91-99, 1976 (No. 2). *E*

A. L. N. Stevels & A. D. M. Schrama-de Pauw: Effects of defects on the quantum efficiency of Eu^{2+}-doped aluminates with the magnetoplumbite-type crystal structure.
J. Luminescence **14**, 147-152, 1976 (No. 2). *E*

A. L. N. Stevels & A. D. M. Schrama-de Pauw: Luminescence spectra of non-stoichiometric aluminates doped with Eu^{2+} ions.
J. Luminescence **14**, 153-157, 1976 (No. 2). *E*

A. L. N. Stevels & J. M. P. J. Verstegen (Philips Lighting Division, Eindhoven): $Eu^{2+} \rightarrow Mn^{2+}$ energy transfer in hexagonal aluminates.
J. Luminescence **14**, 207-218, 1976 (No. 3). *E*

M. Tasto: Maximum likelihood reconstruction of random objects from noisy objects.
3rd Int. Joint Conf. on Pattern recognition, Coronado, California, 1976, pp. 551-555. *H*

K. Teer: LSI is meer dan LSI.
T. Ned. Elektronica- en Radiogen. **41**, 147-155, 1976 (No. 5). *E*

R. Tijburg: Advances in etching of semiconductor devices.
Phys. in Technol. **7**, 202-207, 1976 (No. 5). *E*

J. F. Verwey: Nonvolatile semiconductor memories.
Adv. in Electronics & Electron Phys. **41**, 249-309, 1976. *E*

L. Vriens & M. Adriaansz: Near-resonant light scattering and fluorescence in a dense high-temperature plasma.
Appl. Phys. **11**, 253-257, 1976 (No. 3). *E*

W. Wagner: Reconstruction of object layers from their X-ray projections: a simulation study.
Computer Graph. & Image Proc. **5**, 470-483, 1976 (No. 4). *H*

K. R. Whight, P. Blood & K. H. Nicholas: Implanted high value resistors.
Solid-State Electronics **19**, 1021-1027, 1976 (No. 12). *R*

J. P. Woerdman: Doppler-free two-photon transitions of the sodium molecule.
Chem. Phys. Letters **43**, 279-282, 1976 (No. 2). *E*

C. E. C. Wood: Molecular beam epitaxial GaAs layers for MESFET's.
Appl. Phys. Letters **29**, 746-748, 1976 (No. 11). *R*

D. J. Zwanenburg & W. A. P. Reynen: An improved synthesis of salicylaldehydes. No influence of steric hindrance.
Synthesis 1976, 624-625 (No. 9). *E*

Contents of Optica Acta **24**, No. 4, 1977, special issue on optics at Philips GmbH Forschungslaboratorium Hamburg:

A. W. Lohmann (University of Erlangen): Editorial: Special issue — optics at Philips in Hamburg (p. 285).

G. Groh: Optics at the cross-roads (pp. 287-291). *H*

D. Meyer-Ebrecht & H. Weiss: Tomosynthesis — 3-D X-ray imaging by means of holography or electronics (pp. 293-303). *H*

H. Weiss, E. Klotz, R. Linde, G. Rabe & U. Tiemens: Coded aperture imaging with X-rays (flashing tomosynthesis) (pp. 305-325). *H*

G. Kowalski & W. Wagner: Generation of pictures by X-ray scanners (pp. 327-348). *H*

P. Lux: Redundancy reduction in radiographic pictures (pp. 349-365). *H*

U. Bertram & H. Weiss: MTF of X-ray systems for 3-D imaging (pp. 367-383). *H*

H. Dammann: Model MTF for conventional X-ray imaging systems (pp. 385-390). *H*

M. Tasto: Parallel array processors for digital image processing (pp. 391-406). *H*

H. J. Schmitt: Data handling and recording by laser (pp. 407-412). *H*

R. Pepperl: A solid-state digital light deflector with DKDP polarization switches (pp. 413-425). *H*

R. Pepperl & I. Sander: High density direct read after write (DRAW) recording (pp. 427-431). *H*

K. P. Schmidt: A 16×18 bit CdS spatial modulator (page composer) (pp. 433-440). *H*

U. Killat & D. R. Terrell: Performance and limitations of photothermoplastic devices (pp. 441-452). *H*

U. Killat: Holographic microfiche of picture-like information (pp. 453-462). *H*

H. Kurz: Photorefractive recording dynamics and multiple storage of volume holograms in photorefractive $LiNbO_3$ (pp. 463-473). *H*

E. Krätzig & H. Kurz: Photorefractive and photovoltaic effects in doped $LiNbO_3$ (pp. 475-482). *H*

H. Heitmann, J.-P. Krumme & K. Witter: Magneto-optic memory materials (pp. 483-494). *H*

B. Hill, I. Sander & G. Much: Magneto-optic memories (pp. 495-504). *H*

H. Dammann & E. Klotz: Coherent optical generation and inspection of two-dimensional periodic structures (pp. 505-515). *H*

Contents of Philips Telecommunication Review **35**, No. 2, 1977:

A. J. Lambell & R. P. Vincent: PAIR — Precision Approach Interferometer Radar (pp. 57-69).
Y. Schifres: Digital microwave link for local networks (pp. 70-80).
G. C. M. de Koning: How the software for Philips SPC telephone systems is assembled (pp. 81-94).
P. Kiddle: Radio telecommunication links in the Channel Islands (pp. 95-104).
C. W. M. Barrow: PRX exchange in Jersey (pp. 105-106).

Contents of Mullard Technical Communications **14**, No. 134, 1977:

R. S. Babbs: Application of fast-recovery epitaxial diodes in switched-mode power supplies (pp. 134-142).
H. Q. N. Davies: Survey of teletext transmission systems and Mullard decoders (pp. 143-158).
J. S. Malcolm: The TDA1071 IC in a.m./f.m. radio receivers (pp. 159-167).
P. J. Jackson: Product information: System design aids for the 2650 microprocessor (pp. 168-172).

Contents of Valvo Berichte **20**, No. 2, 1977:

J. A. den Ouden: Digitalvoltmeter mit der integrierten Schaltung GZF 1200 (pp. 45-61).
H. Sauermann & M. Wahl: Ein Wolfram-Metallisierungssystem für Halbleiterbauelemente (pp. 62-80).
R. Lauer: PXE-Beschleunigungsaufnehmer (pp. 81-92).

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79 No. 4/5

Grinding brittle materials

A. Broese van Groenou and J. D. B. Veldkamp

When a metal is ground the abrasive particles 'plough' through the material to form well-defined, 'clean' grooves. If the material is a brittle one, however, there is chipping and cracking in the neighbourhood of the abrasive particles. This is an advantage if the material is to be removed as efficiently as possible, but a disadvantage if a good surface finish is required. In setting up a grinding programme a better understanding of the formation of grooves, cracks and fractures is therefore highly desirable. The investigation described below contributes to this understanding. Since grinding is a 'multi-point scratching process' it will only give a simplified picture of the effects to be studied; 'single-point' experiments were found to be of greater value in the investigation.

Introduction

Grinding is necessary as the final process operation for many ceramic and glass products. It is often difficult to produce shapes accurately by the ceramic process of pressing and firing, and grinding is then the only way of bringing the product to the required shape and size. Let us look at two examples. In d.c. motors for windscreen wipers the stator magnet consists of segments of Ferroxdure (*A* in *fig. 1*), which are ground to the correct radii of curvature. In television receivers the magnetic deflection system includes a Ferroxcube yoke (*B*) fitted around the picture tube; this yoke is ground to the correct shape. Sometimes layers more than 1 mm thick have to be removed in this 'coarse grinding'.

'Finish grinding' is generally used where the optical or magnetic quality or the strength of the surface is important. For example, the magnetic quality of the pot cores used in the *LC* circuits for carrier telephony (*C*) is strongly dependent on the nature of the surface at the air gap. In magnetic heads for a video-tape recorder (*D*) a good surface is also a first requirement, both for the magnetic quality and to keep the spacing between head and tape as small as possible. Television picture-tube screens are ground to give a good optical finish, but the strength aspects are also important here. The strength of a brittle object is often dependent on the depth of surface cracks; finish grinding removes the cracks and increases the strength.

Grinding is a costly operation; the machines are expensive and the grinding takes a long time. The cost of grinding the magnet segments in fig. 1 is 20% of the production costs, and for the picture-tube yokes the corresponding figure is no less than 40%.

Detailed analysis shows that grinding is a complicated process. As the abrasive particles scratch the surface of the material several different microprocesses come into play (flow, cracking, fracture), and the result depends on a number of factors, including the shape, the force and the velocity of the particles, the humidity of the environment, the hardness and the crack resistance of the material and the distribution of microcracks, cavities and other flaws in the material. In view of the importance and the expense of grinding it seemed useful to make a closer analysis of this complicated process. This article gives a general picture of our investigations of the process. They mainly relate to the grinding of ferrites and glass; because these materials are very hard diamond was always used for the abrasive particles. In general terms the objective of the

Dr A. Broese van Groenou is with Philips Research Laboratories, Eindhoven; Dr Ir J. D. B. Veldkamp, formerly with these Laboratories, is now with the Philips Glass Division, Eindhoven.

investigation was to find out how the *specific energy*, i.e. the energy required for the removal of unit volume of the material, and the *damage to the surface* of the workpiece depend on the setting of the grinding machine and the properties of the material.

grinding conditions formed an important part of our investigations.

An essential feature of the grinding process is the force F that the wheel exerts on the workpiece (see fig. 2). This force largely determines the amount of

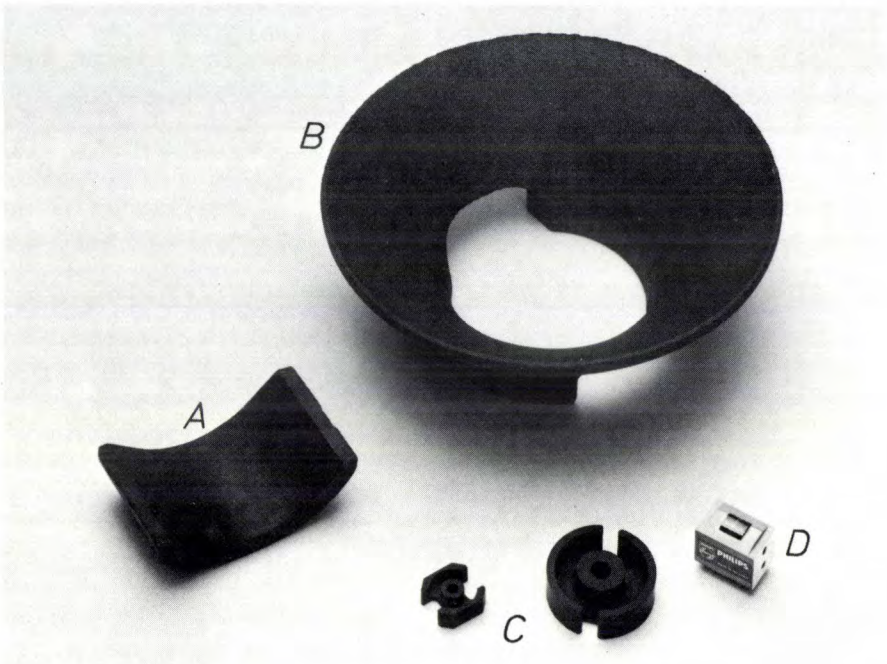


Fig. 1. Some examples of ground ceramic products. *A* Segment of the stator magnet in a windscreen-wiper motor; Ferroxdure. *B* Yoke for the magnetic deflection system of a television receiver; Ferroxcube. These products are brought to the required shape by coarse grinding. *C* Half pot cores for tuning circuits in carrier telephony; Ferroxcube. The surface of the central cylinder and the corresponding surface of the other half form an air gap whose purpose is to reduce the temperature coefficient to a low value; these two surfaces are finish ground to give good magnetic quality. *D* Magnetic head for video tape recorder; Ferroxcube. The gap width is $\frac{1}{2}$ μm. Finish grinding is required here to give good magnetic quality and a well-defined spacing between head and tape.

The characteristic quantities of the grinding system

Let us consider a surface grinder as a representative example of a grinding system (*fig. 2*). The rate of volume removal Z — the volume ground away per second — is the product of the depth of cut a, the width b_c of the contact zone between the grinding wheel and the workpiece, and the speed v_w of the workpiece (the 'rate of feed'):

$$Z = ab_c v_w. \tag{1}$$

The specific energy e is the ratio of the power P required for grinding to the rate of volume removed:

$$e = P/Z. \tag{2}$$

The specific energy appears on initial consideration to be a constant of the material (roughly speaking e increases with the hardness of the material), but it can after all be affected by the grinding conditions, sometimes very strongly. Indeed, a study of the effect of the damage to the surface and the wear of the wheel. The force ratio

$$f = F_t/F_n, \tag{3}$$

where F_t is the tangential component of F and F_n the normal component, is often introduced.

The grinding power P is given by

$$P = F_t v_s, \tag{4}$$

where v_s is the circumferential velocity of the wheel (the 'wheel speed'). For constant e it can be seen from (4) that a high value of v_s is desirable: for the same power (the same rate of volume removal) smaller forces are necessary, so that less damage and wear may be expected. In practice, there is in fact a tendency to operate grinding wheels at high speeds.

From (2) and (4) we find that the 'reduced' force F_t' (the tangential force per unit width, F_t/b_c) can be

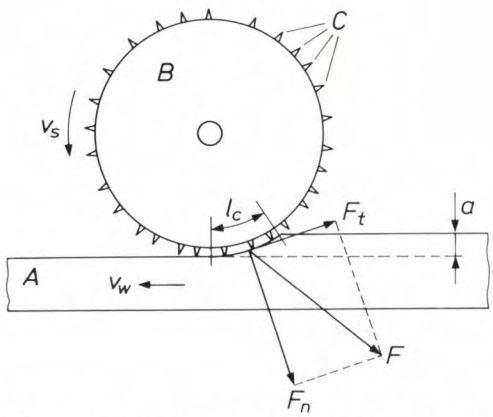


Fig. 2. Diagram representing a surface grinder. The workpiece *A* is passed under the grinding wheel *B*, with circumferential velocity (wheel speed) v_s, at a rate of feed v_w. *C* diamonds, *a* depth of cut, l_c length of the contact zone; the width of this zone (the dimension at right angles to the paper) is b_c. The wheel exerts a force *F* on the workpiece; F_n is the normal component of *F*, and F_t the tangential component.

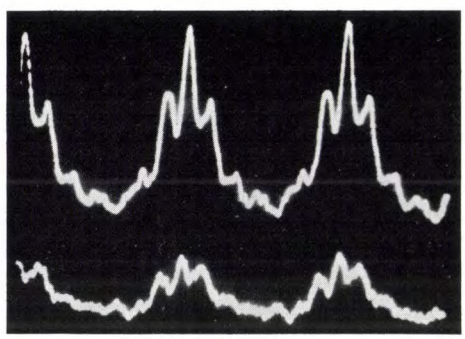


Fig. 3. The normal force F_n (*upper trace*) and the tangential force F_t (*lower trace*) during a little more than two revolutions of the grinding wheel, in an experiment with $v_s = 58$ m/s, $v_w = 60$ mm/s, $a = 0.7$ mm, $h_{eq} = 0.75$ μm. The variations arise because the wheel is not truly circular.

expressed as

$$F_t' = eh_{eq}, \tag{5}$$

where

$$h_{eq} = Z/v_s b_c. \tag{6}$$

The specific energy *e* can therefore be found by measuring the tangential force.

The quantity h_{eq} is useful for comparing different grinding systems: it is the machine-related quantity that determines the reduced force for given *e*. From (6) it follows that for every revolution of the wheel a layer of grindings of thickness h_{eq} would be applied to the circumference of the wheel if the grindings were uniformly distributed and remained densely packed. This can be seen by multiplying the numerator and denominator of the right-hand side of (6) by the time *T* for one revolution; the numerator then gives the volume produced in time *T* and the denominator gives the product of the length $v_s T$ and the width b_c of the layer. The quantity h_{eq} is therefore called the *continuum chip thickness*. The value of h_{eq} is of the order of 0.01 μm for finish grinding and 1 μm for coarse grinding.

Grinding experiments at high wheel speeds

Typical wheel speeds for ordinary grinding are from 20 to 30 m/s. Since high speeds are of interest, we carried out experiments at speeds of about 60 to 100 m/s[1]. The grinding forces were measured with a dynamometer. Strong periodic variations in the forces are unavoidable in such experiments because grinding wheels are not truly circular, see *fig. 3*.

Fig. 4 gives the results of a large number of experiments with Sr Ferroxdure (Sr hexaferrite). The peak values of the reduced forces $\hat{F}_t'$ and $\hat{F}_n'$ are shown in a log-log plot as a function of h_{eq}. Even though there is a large spread in the measured points these results lead to the conclusion that at a given value of h_{eq}, the quantities $\hat{F}_t'$ and $\hat{F}_n'$, and hence the specific energy, do not depend systematically on v_s. We have found this result for other ceramic materials and for most glasses. Some low-melting-point glasses do have a speed dependence at high wheel speeds. This can be attributed to heating effects; we shall not pursue the matter further.

The slope of the line $\hat{F}_t'(h_{eq})$ in fig. 4 does not differ much from 1, so that *e* is also almost independent of h_{eq} (or F_t'). This is *not* a rule, however: *e* usually does depend on h_{eq}. The effect is known as the 'size effect'. A striking example of this is given by MnZn ferrite, in which *e* varies by a factor of 5 over the same range of h_{eq} considered in fig. 4 (see *fig. 5*).

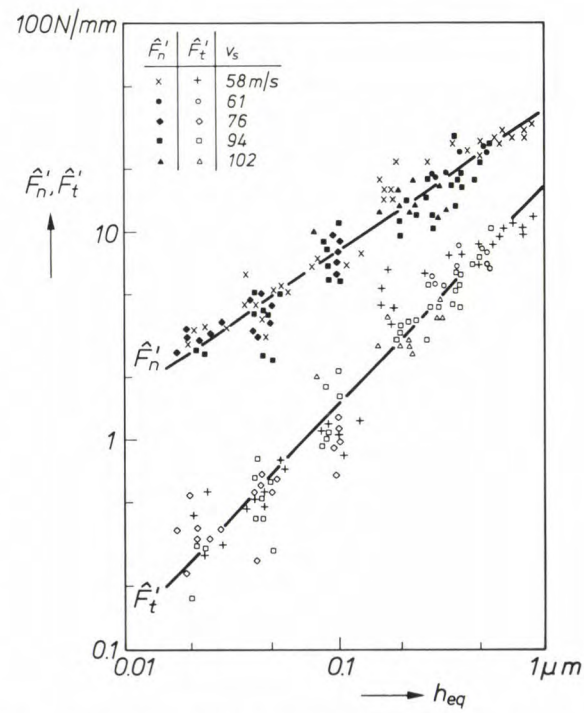


Fig. 4. Peak values of the 'reduced' forces, $\hat{F}_n'$ and $\hat{F}_t'$, as a function of the continuum chip thickness h_{eq}, for speeds of 58 to 102 m/s in grinding Sr Ferroxdure.

[1] A. Broese van Groenou, J. D. B. Veldkamp and D. Snip, J. Physique **38**, C1/285, 1977.

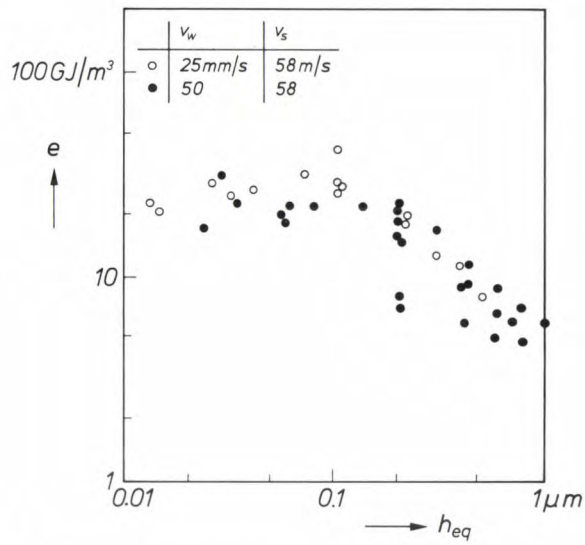


Fig. 5. Specific energy e as a function of the continuum chip thickness h_{eq} for MnZn ferrite. The measured points were obtained at a wheel speed of 58 m/s and at two rates of feed, 25 and 50 mm/s.

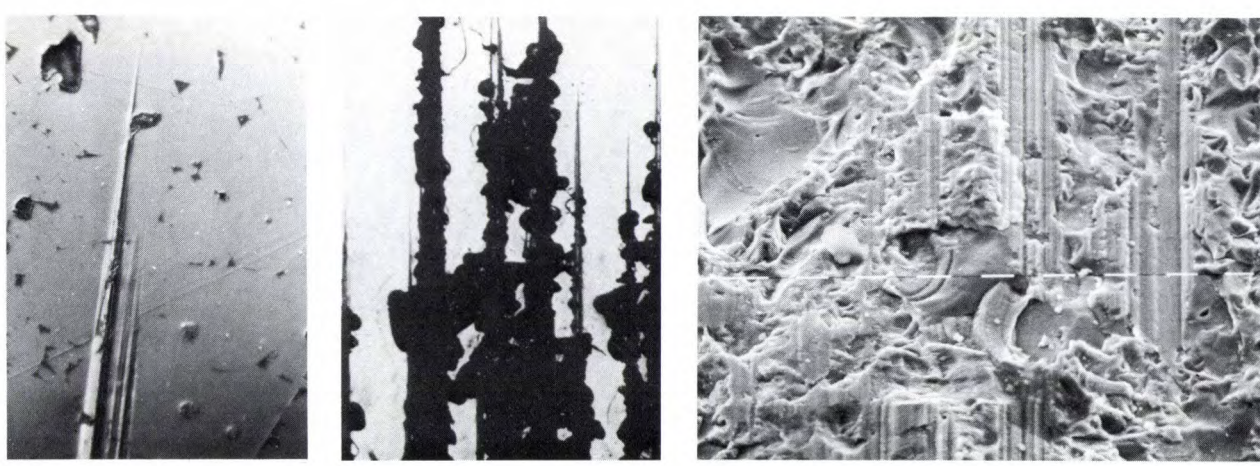
The microphotographs in *fig. 6* give a clue to the explanation of the size effect in MnZn ferrite. They show parts of the zone of first contact on the initially polished surface of a slightly tilted sample that was brought into contact with the wheel. This zone shows concisely what happens on a microscale as the depth of cut gradually increases (increasing h_{eq}). The scratches from the individual diamonds start as well-defined, 'clean' grooves which gradually get deeper. Then material starts to chip away from the edges of the groove and eventually the chipped areas from neighbouring scratches join together. In this final stage the area 'cracked' and loosened by one diamond can be removed by another neighbouring diamond with the expenditure of very little energy, so that the specific energy is relatively low.

These photographs also show clearly that grinding is a multi-point process, and that scratching with a single diamond should be the most suitable experiment for obtaining a better understanding of the microprocesses. The experimental conditions can then be clearly defined, e.g. for the geometry of the diamond and the force on the diamond. The basic experiment is *slow scratching*, since heating effects are then avoided.

The scratching experiment; ploughing

Fig. 7 is a photograph and *fig. 8* a diagram of the apparatus we used for our scratching experiments [2]. The sample S is displaced horizontally underneath the diamond D by the micromanipulator Mi, which is driven by the motor Mo. The diamond is attached to the arm A of a balance. This arm is connected to the frame by a leaf spring LS, so that it can be swung up and down but is stiff horizontally. The balance is brought to equilibrium and displaced vertically until the diamond just touches the sample; the desired load L is then applied. The tangential force on the sample is measured by the strain gauges SG on the bending element BE (fig. 8*b*).

In the experiments to be described here the diamonds were ground to the shape of a square pyramid; the scratches were made with one plane leading as in *fig. 9*. The scratching speed was 1 μm/s. Measurements were made on the materials in *Table I*. The range of hardness values was about 9 times. *Fig. 10* shows two scratches in MnZn ferrite, and *fig. 11* shows the associated tangential forces. When the load is small a clean groove



a *b* *c*

Fig. 6. Scratches on the etched surface of a MnZn ferrite, made by the diamonds on the grinding wheel. The start of the scratches is made visible by setting up the sample so that it is slightly tilted during its brief contact with the wheel. *a*) The start of a scratch: a 'clean' groove. *b*) Scratches with chipped edges after several passes of the same diamonds. *c*) The effects of individual diamonds run together.

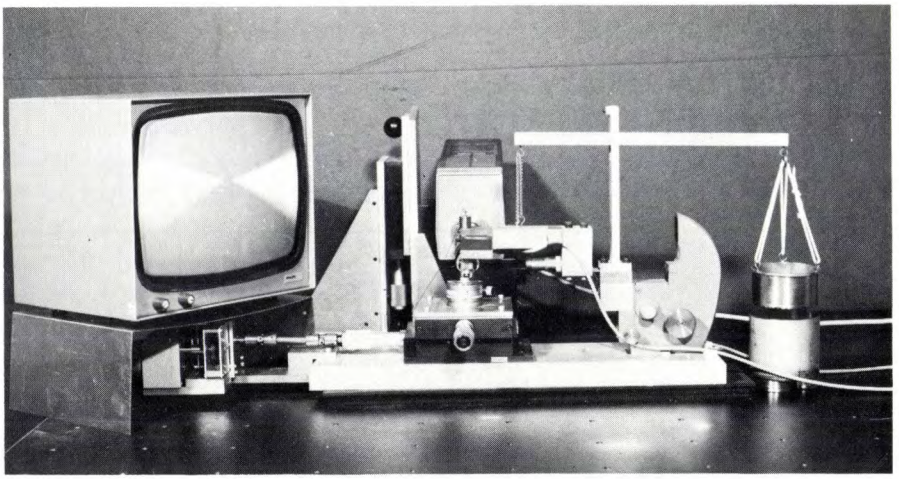


Fig. 7. The scratching apparatus, explained in more detail in fig. 8. The scratching point can be made visible on the monitor on the left.

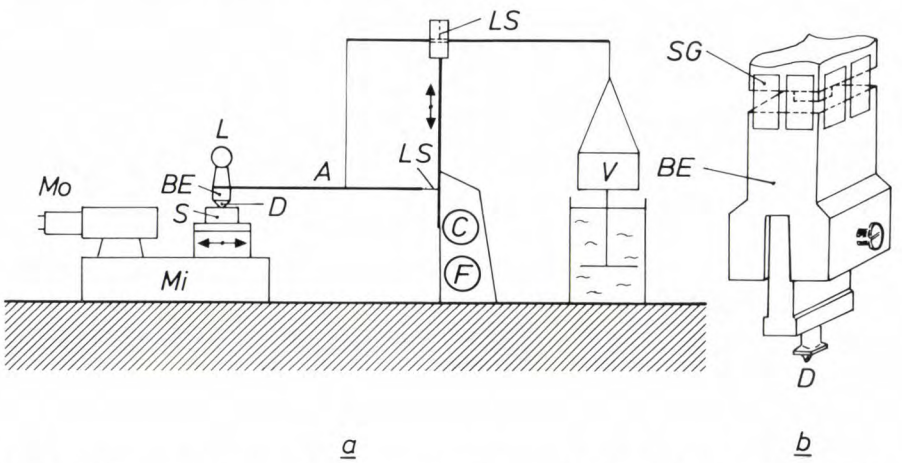


Fig. 8. *a*) Diagram of the scratching apparatus. The sample *S* is displaced underneath the diamond *D* by the motor *Mo* and the micromanipulator *Mi*. The diamond is attached by the bending element *BE* to the balance arm *A*, which can be swung up and down but is stiff horizontally. After the balance has been brought to equilibrium and adjusted vertically the load *L* is applied. *C*, *F* coarse and fine vertical adjustment. *V* counterweight and damper. *LS* leaf springs. *b*) The bending element *BE* with the diamond *D* and strain gauges *SG* for measuring the tangential force. This figure is at true size.

is produced (fig. 10, *lower picture*); with a higher load the groove is wider and there is cracking and chipping (fig. 10, *upper picture*). The large force fluctuations that appear at the higher loads can be connected with the onset of cracking and chipping, but can also be the result of the diamond meeting pores or grain boundaries. In this section we shall only consider the groove.

The formation of grooves in metals has been studied extensively [3]. It is a process of plastic deformation: the diamond ploughs up the material in front of it and pushes it aside. *Fig. 12* shows the plastic nature of the deformation. It may perhaps at first sight seem surprising that such a process can also take place in brittle materials like Ferroxdure and glass. It becomes easier to understand, however, when the pressure

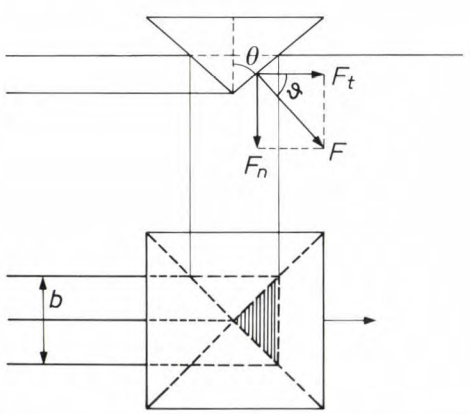


Fig. 9. Geometry of diamond and groove; vertical longitudinal section and plan view. The diamond is a square pyramid (half apex angle θ) that moves with one plane leading. F is the force the diamond exerts on the sample, F_t is the tangential component and F_n the normal component, b the groove width. ('Tangential' and 'normal' relate to the surface of the sample, not to the leading plane of the diamond.) In the ploughing model in its simplest form F is perpendicular to the leading plane.

[2] A. Broese van Groenou, N. Maan and J. D. B. Veldkamp, Philips Res. Repts **30**, 320, 1975.

[3] See for example F. P. Bowden and D. Tabor, Brit. J. appl. Phys. **17**, 1521, 1966.

Table I. The materials investigated. Their hardness H_V is shown on the right in GNm^{-2}.

ZnO	2.6
Lead glass	5.0
Sr hexaferrite (a Ferroxdure)	6.5
MnZn ferrite (a Ferroxcube)	8.0
NiZn ferrite (a Ferroxcube)	9.5
Al_2O_3	23

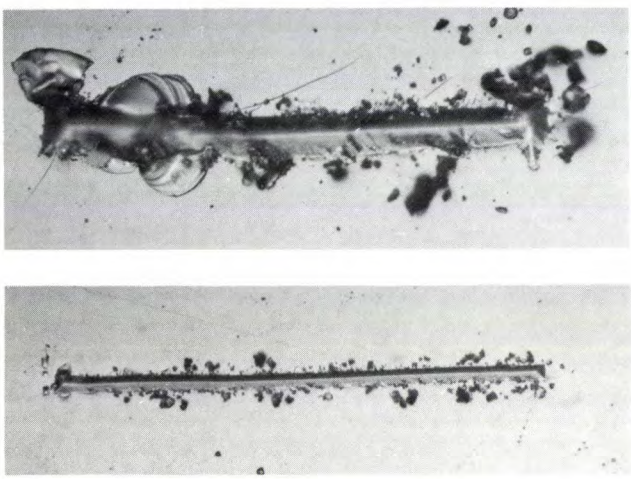


Fig. 10. Slow scratches in a sample of MnZn ferrite made at a load of 0.1 N (*below*) and 0.5 N (*above*). Half apex angle of the diamond: 60°.

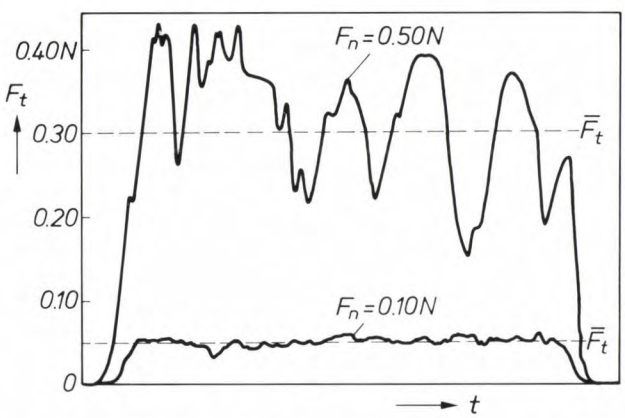


Fig. 11. Tangential forces F_t in making the scratches of fig. 10.

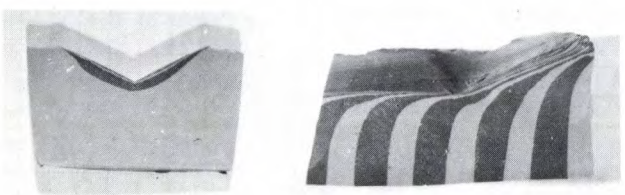


Fig. 12. Cross-section and longitudinal section of a groove made in two-coloured plasticine by a square glass pyramid ($2\theta = 120°$). The plastic deformation is confined to a thin layer below the surface.

beneath the contact area is calculated: it is of the order of 100 000 atmospheres. The plastic nature of the deformation can also be seen from photographs like those in *fig. 13*.

We shall now show that the ploughing model, which was formulated for metals [3], is broadly applicable to groove formation in the materials we have studied. We shall only consider the simplest form of this model here, in which, applied to the situation of fig. 9, *the leading plane takes up all of the pressure and friction plays no part*. This model leads to a few simple rules concerning the force ratio $f\,(= F_t/F_n)$, the groove width b and the '*specific groove energy*' e_g, which can easily be tested.

We shall consider the force ratio first. Under the conditions assumed the total force must be perpendicular to the leading plane of the diamond, which means that

$$f = \cot\theta. \tag{7}$$

The force ratio would thus be completely determined by half the apex angle θ of the diamond, and not dependent on the load or on properties of the material such as hardness or microstructure. This is generally confirmed by the $\bar{F}_t$-F_n diagram of *fig. 14* for two diamonds with different apex angles and for all the materials of Table I; for both apex angles a straight line through the origin gives a fairly good representation of all the measured points. However, f is somewhat smaller than cot θ. This difference, which will not be taken into account from now on, can be attributed to the Coulomb friction, which we have neglected [2].

Now let us consider the groove width. This depends on the forces and the specific groove energy e_g of the material, i.e. the energy required for making the groove divided by the volume of the groove, or expressing it another way, the horizontal force F_t divided by the cross-section A_t of the groove:

$$e_g = F_t/A_t. \tag{8}$$

(e_g can be different from the specific *grinding* energy e because any material that may disappear in the form of chips is not included. We should note here that even when there is considerable chipping the groove width can generally still be determined easily; see for example the upper picture of fig. 10.)

From the geometry of fig. 9:

$$F/A = F_n/A_n = F_t/A_t, \tag{9}$$

where A is the contact area and A_n and A_t are the projections of A on the surface of the sample and on the plane perpendicular to the groove direction. In the

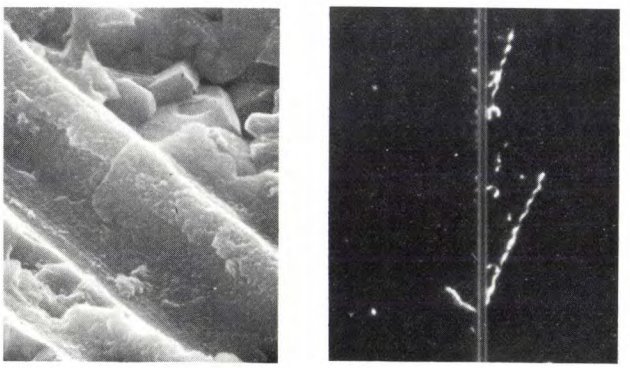
[4] N. Maan and A. Broese van Groenou, Wear **42**, 365, 1977.

ploughing model A_t is identical to the groove cross-section. The central quantity F_n/A_n in (9) is by definition the *scratch hardness* H_s, which is thus equal to e_g in the simple ploughing model, but is often not the same in practice and in less simple models. Since A_n is equal to $\frac{1}{4}b^2$ (see fig. 9), we can use (8) and (9) to find a simple relation linking the specific energy, the groove width and the load:

$$e_g\,(=H_s) = 4F_n/b^2. \qquad (10)$$

The groove width b must therefore be proportional to $F_n^{\frac{1}{2}}$, and independent of the apex angle. The measured results in *fig. 15* illustrate this relation for several materials.

Finally, hard materials would be expected to have a higher specific energy than soft ones. This trend is confirmed in *fig. 16*. In the simplified ploughing model e_g would have to be equal to the pressure F/A at which the material just starts to flow (see fig. 9), and would therefore be proportional to the Vickers hardness H_V. In fig. 16 it can be seen that $e_g \approx 2.5\,H_V$ at a load of 0.1 N — a load per diamond not unusual in grinding.



a *b*

Fig. 13. *a*) Plastic deformation in ground Ferroxdure (magnification 2000×). *b*) Plastic deformation, for a scratch in glass ('shavings' are formed; magnification 900×).

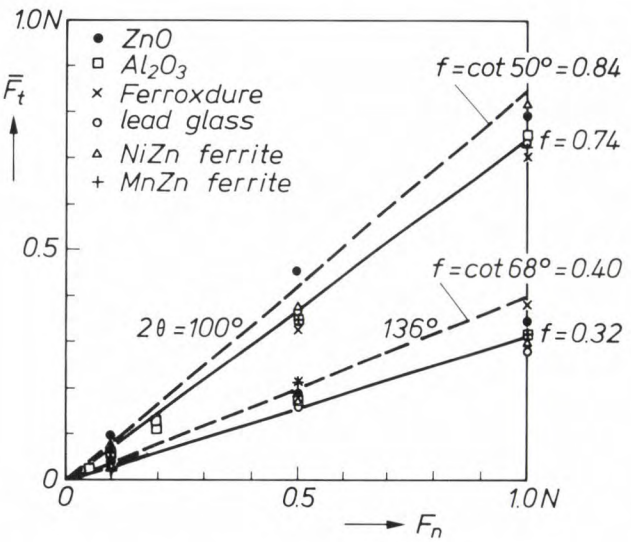


Fig. 14. The average tangential force $\bar{F}_t$ as a function of the load F_n for two diamonds ($2\theta = 100°$ and 136°) and six different materials. In the simple ploughing model (no friction) F_t/F_n would be equal to cot θ (the corresponding values are 0.84 and 0.40, *dashed lines*).

Size effect

As in grinding, the specific energy in scratching is not a true constant of the material: e_g decreases with increasing load, as can be seen from fig. 16. Again, as in grinding e_g is usually considered as a function of the groove depth d (a monotonically increasing function of the load) here. A size effect is then generally found; our results demonstrate this in *fig. 17*. We shall not discuss the explanation of this effect here.

Ploughing and grinding

Let us now consider how realistic it is to interpret 'grinding' as 'ploughing'. To allow the grinding results to be compared with the scratching results we must first convert the continuum chip thickness h_{eq} into an average groove depth d_{av} for the diamonds on the grinding wheel, starting with the assumption that grinding really is ploughing. If C is the number of effective diamonds per unit area of the grinding wheel, b_c the width and l_c the length of the contact zone, then

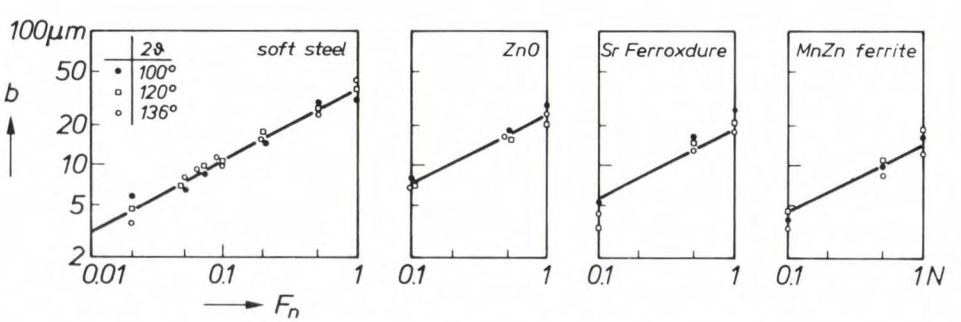


Fig. 15. Groove width b as a function of load F_n for three diamonds in mild steel [4], ZnO, Ferroxdure and MnZn ferrite. The solid lines have a slope of 1 in 2 ($b \propto F_n^{\frac{1}{2}}$). The 'best' straight line for Ferroxdure and MnZn ferrite would be a little steeper (size effect).

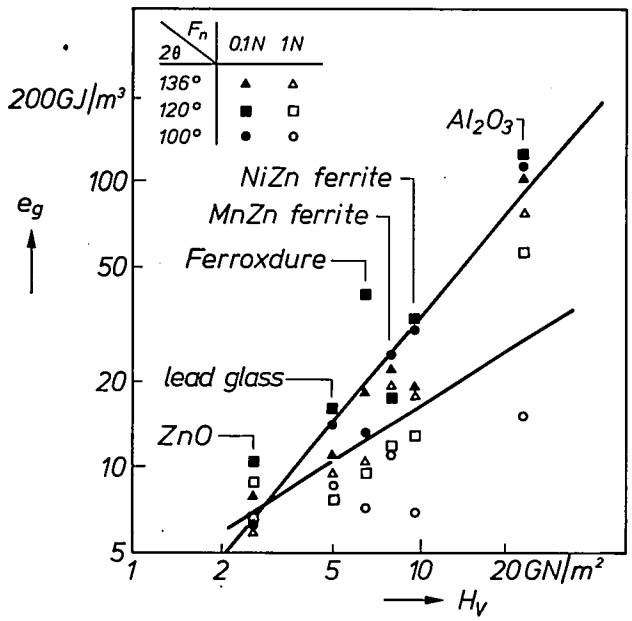


Fig. 16. The specific groove energy e_g for various materials, arranged in order of hardness H_V. In general e_g is higher for harder materials.

is difficult to estimate. Much more closely defined results are obtained from experiments at 5 m/s with a flat grinding wheel (*fig. 18*) [6]. At a given load F_n and wheel speed v_s we measured the tangential force F_t and the rate of volume removal Z (this was obtained from the reduction in height of the sample). Using equations (5) and (6) we calculated e and h_{eq} from the measured quantities. Counting diamonds under the microscope gave C, and $\cot\theta$ was equated to the measured force ratio, so that h_{eq} could be converted into an average groove depth.

The specific grinding energy and the specific groove energy, both as a function of groove depth, can now be compared. The results for Sr Ferroxdure and MnZn ferrite can be seen in *fig. 19*. In view of the great difference in speed (a ratio of $1:10^6$), the agreement between groove and grinding energies for Ferroxdure is remarkable. There is a considerable discrepancy for MnZn ferrite, however: the specific grinding energy is only about a tenth of the specific groove energy.

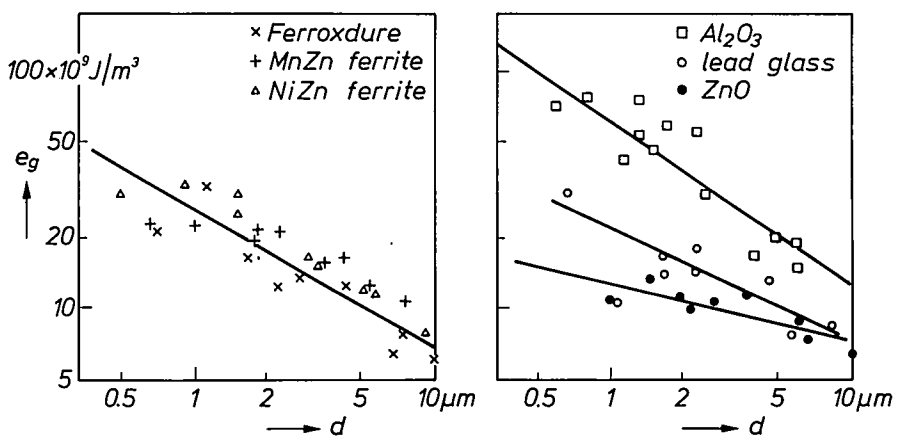


Fig. 17. Size effect for six materials; the specific groove energy e_g depends on the groove depth d.

the number of diamonds in contact with the workpiece at a given moment is equal to $Cb_c l_c$. The rate of volume removal per diamond is $A_t v_s$ (A_t is the groove cross-section) and the total rate of volume removal Z is therefore $Cb_c l_c A_t v_s$. If we also assume that all the diamonds are pyramids of apex angle 2θ, then A_t is equal to $d_{av}^2 \tan\theta$. Combining these relations with eq. (6) we have [5]:

$$d_{av} = \left(h_{eq}\frac{\cot\theta}{Cl_c}\right)^{\frac{1}{2}}. \qquad (11)$$

The high-speed grinding experiments described earlier are not suitable for comparison with scratching experiments: because the wheel is not truly circular (see fig. 3) the number of 'effective' diamonds is rather small and uncertain, so that the average groove depth

MnZn ferrite is much more brittle than Ferroxdure. The results therefore give a more detailed confirmation of the picture of grinding that was outlined before: for

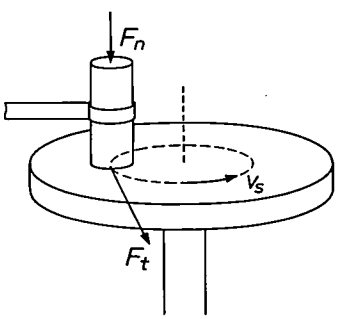


Fig. 18. Grinding with a flat grinding wheel. The sample is fixed in the horizontal direction and pressed vertically against the wheel by a force F_n. F_t tangential force; v_s speed of the wheel at the sample.

'tough' materials grinding is indeed a process of plastic deformation (ploughing), with the specific energy determined by the hardness; in brittle materials, on the other hand, material is removed by chipping as well as by ploughing.

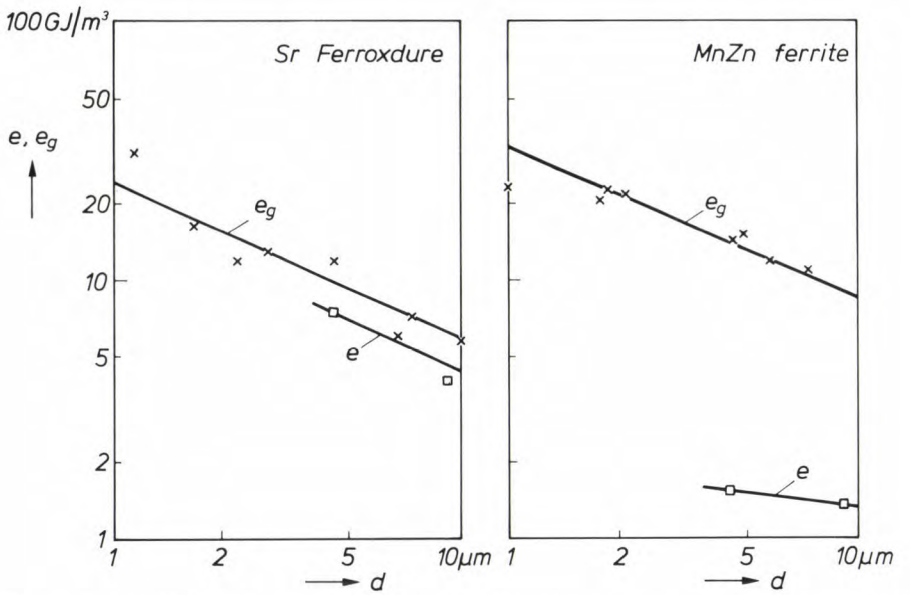


Fig. 19. The specific grinding energy e and the specific groove energy e_g as a function of the groove depth for Ferroxdure and MnZn ferrite.

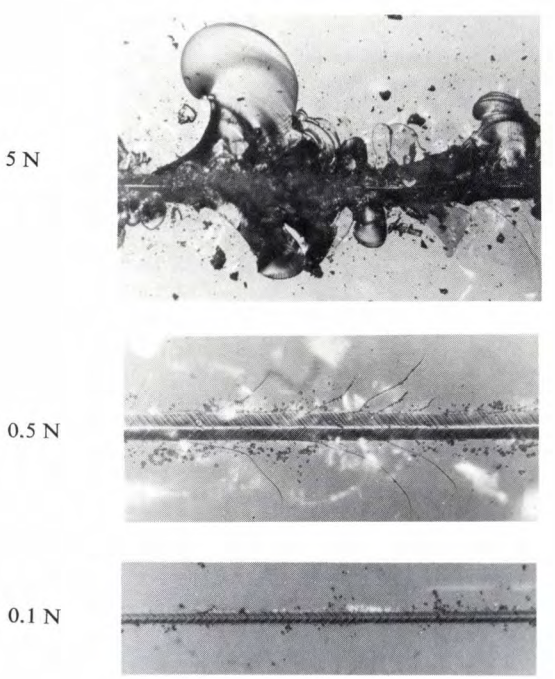


Fig. 20. Effect of the load on the damage in SF58 glass. *From bottom to top:* clean groove ($F_n = 0.1$ N), lateral cracks (0.5 N), chipping (5 N); $v_s = 400$ μm/s.

Cracking and chipping

When the load is increased during scratching a clean groove is in general obtained at first, then cracks start to form and finally there is cracking and chipping (*fig. 20*). Three different sorts of cracks can clearly be distinguished: median cracks, subsurface cracks and lateral cracks [7] (*fig. 21*).

The nature of the damage does not in general only vary with the load, but also with the scratching speed.

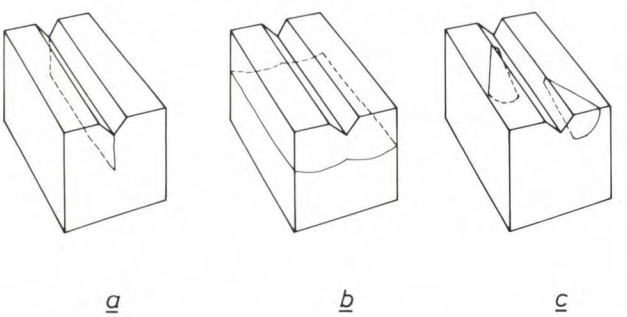


Fig. 21. Types of cracks produced in scratching. *a*) Median crack. *b*) Subsurface crack. *c*) Lateral crack.

This is demonstrated in *fig. 22* for MnZn ferrite and in *fig. 23* for BK7 glass. On increasing the load in glass, median and subsurface cracks are obtained at first, accompanied later by lateral cracks and finally by chipping as well.

We shall give most of our attention to the lateral cracks. These always precede the chipping as the load is increased. They can be directly observed at the

[5] M. C. Shaw, in: New developments in grinding, ed. M. C. Shaw, Carnegie Press, Pittsburgh 1972, p. 220.
[6] J. D. B. Veldkamp and R. J. Klein Wassink, Philips Res. Repts **31**, 153, 1976.
[7] J. D. B. Veldkamp, N. Hattu and V. A. C. Snijders, in: Fracture mechanics of ceramics, vol. 3, ed. R. C. Bradt, D. P. H. Hasselman and F. F. Lange, Plenum Press, New York 1978, p. 273.

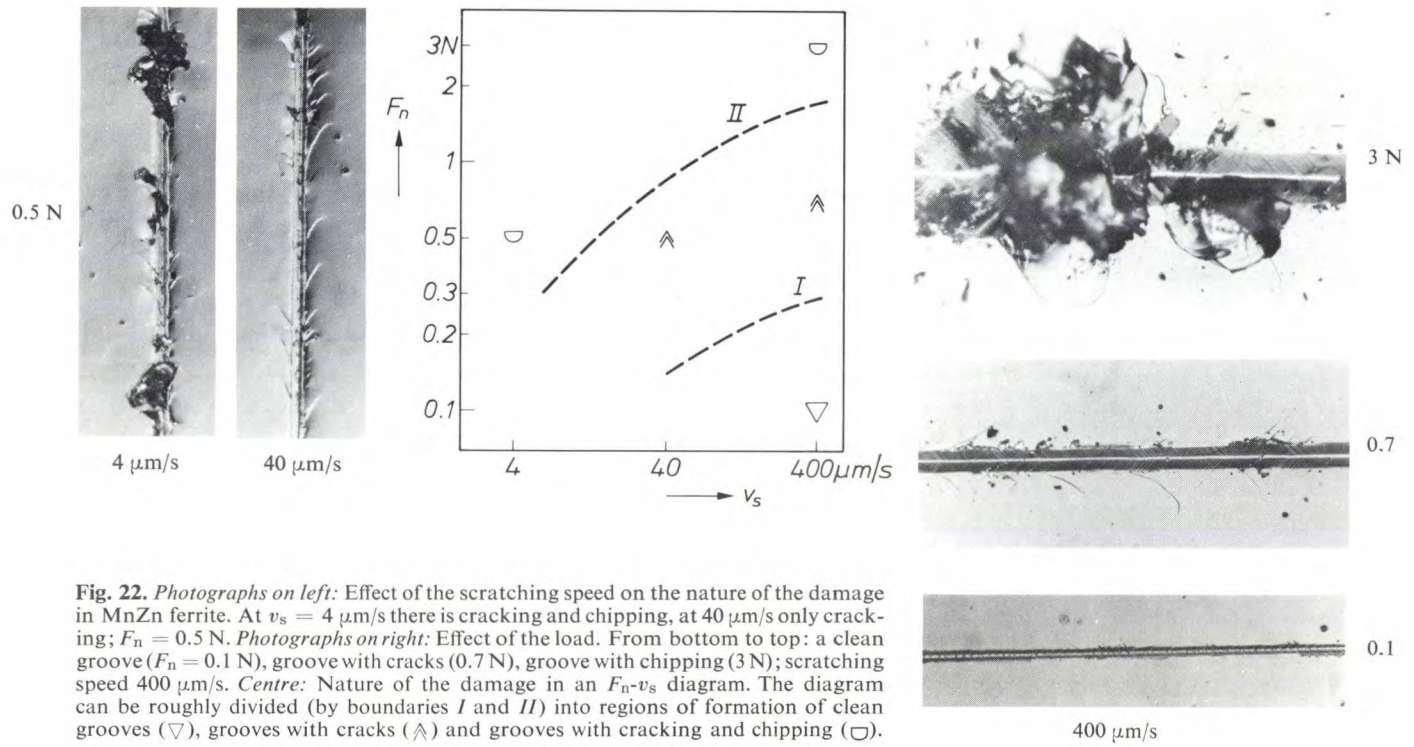


Fig. 22. *Photographs on left:* Effect of the scratching speed on the nature of the damage in MnZn ferrite. At $v_s = 4$ μm/s there is cracking and chipping, at 40 μm/s only cracking; $F_n = 0.5$ N. *Photographs on right:* Effect of the load. From bottom to top: a clean groove ($F_n = 0.1$ N), groove with cracks (0.7 N), groove with chipping (3 N); scratching speed 400 μm/s. *Centre:* Nature of the damage in an F_n-v_s diagram. The diagram can be roughly divided (by boundaries *I* and *II*) into regions of formation of clean grooves (▽), grooves with cracks (⩓) and grooves with cracking and chipping (▭).

surface and can easily be counted and measured. The subsurface cracks are best known in glass, where they are readily visible because of light reflections at the surface of the crack. Median cracks have been investigated in detail some time ago [8].

The boundary *I* in figs 22 and 23 is of particular importance for finish grinding: if the operating point is kept to the right of and below this boundary there will be no damage to the surface. For coarse grinding it is possible to take advantage of chipping by working above boundary *II*. In glass chipping starts in the region between *I* and *II*: the subsurface cracks allow chips to form easily on repeated passes of the scratching particles.

The speed effect expressed in figs 22 and 23 is usually only seen clearly at the lower speeds; at speeds above about 1 mm/s the surface damage is generally only dependent on the load. We shall now discuss the crack formation as a function of the load at these higher speeds, with special attention to the way in which the *minimum force for crack formation* (boundary *I*) depends on the properties of the material.

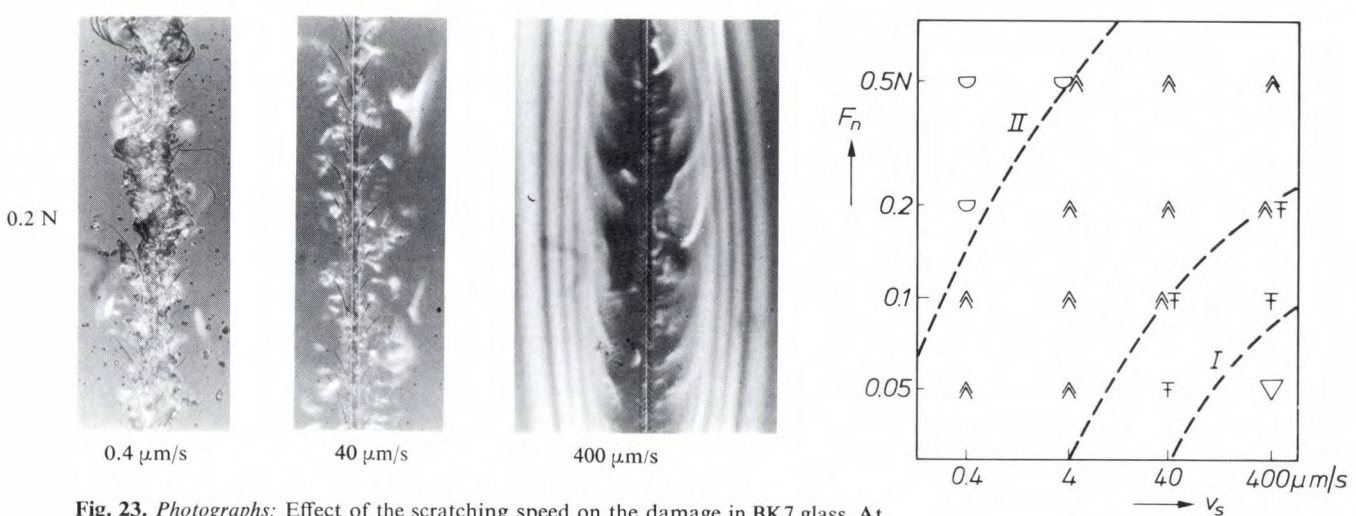


Fig. 23. *Photographs:* Effect of the scratching speed on the damage in BK7 glass. At $v_s = 0.4$ μm/s there is chipping, at $v_s = 40$ μm/s there is lateral cracking and at 400 μm/s only subsurface cracks are formed; $F_n = 0.2$ N. In the F_n-v_s diagram we find from lower right to upper left: clean grooves (▽); grooves with median and subsurface cracks (∓); next there are also lateral cracks (⩓); in the last region there is chipping as well, but there are no more subsurface cracks (▭).

Crack formation as a function of load

A diagram of the model situation for crack formation [9] is shown in *fig. 24*. An infinite sheet of material is loaded by a uniform tensile stress σ in a direction in the plane of the sheet. A planar crack forms across the sheet at right angles to the direction of the stress; the width of the crack is $2c$. The stress in the region of the crack concentrates round the tips of the crack (R). The intensity of the (very inhomogeneous) stress field around either tip of the crack is proportional to σ and to $\sqrt{c}$; it is specified completely by the 'stress-intensity factor' $K = \sigma\sqrt{\pi c}$.

If the stress σ in this situation is increased, K eventually reaches a value K_c, characteristic of the material, at which the material splits at the tips of the crack. We then have an 'explosive' situation; the splitting continues and the tip of the crack propagates in the material at approximately the velocity of sound: the material *fractures*. For simplicity we shall assume that the crack does not propagate at all for $K < K_c$; we can then speak of 'the' cracking condition

$$\sigma\sqrt{\pi c} = K_c. \tag{12}$$

To predict whether the tip of a crack will propagate in a more complicated situation, and if so how far, it is necessary to calculate the stress field and find the values σ and c in fig. 24 should have to give the same stress field around the tips of the crack, since given certain conditions the field around the tip of a crack always takes the same form. The cracking condition again follows from (12) with the values found for σ and c.

The lateral cracks in scratching are related to the 'half-penny' cracks that sometimes form around an indenter in Vickers hardness measurements[10] (*fig. 25*); we shall now consider these briefly by way of introduction. The situation here, for a well-developed crack, is characterized by two quantities: the radius of the crack c and the force F with which the diamond pushes the surfaces of the crack apart. Dimensional considerations now lead us to expect that the stress σ required for (12) is proportional to the characteristic force (F) divided by the characteristic area (c^2). The cracking condition thus takes the form $(F/c^2)\sqrt{c} \propto K_c$, or:

$$F = \alpha K_c c^{\frac{3}{2}}, \tag{13}$$

where α is a constant. This result is in agreement with a more rigorous analysis of the problem [10]. The situation is not explosive: at a given F the crack grows until the condition (13) is satisfied. The linear relationship between F and $c^{\frac{3}{2}}$ has been experimentally confirmed for this type of cracking [10]. At the minimum force for crack formation (F_{min}) cracking just starts; c is then equal to the half-diagonal $\frac{1}{2}b\sqrt{2}$ of the indentation, and this half-diagonal is again related to F_{min} through the *hardness* H of the material:

$$H = F_{min}/b^2. \tag{14}$$

Combining (13) and (14):

$$F_{min} \propto K_c^4/H^3. \tag{15}$$

Let us now return to the lateral cracks associated with scratching (*fig. 26*). The moving diamond is accompanied by a stress field in the material; if the cracking condition is satisfied at a flaw in the material a crack forms. The forward-tilted direction of the lateral cracks (see figs 10, 20-23) corresponds to the

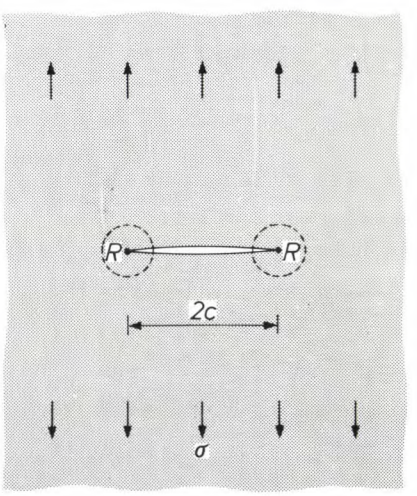


Fig. 24. Model situation for crack formation: an infinite sheet in the plane of the paper is loaded by a uniform tensile stress σ in a direction in this plane. At right angles to this direction there is a planar crack of width $2c$ across the sheet. The stress in the region of the crack concentrates round the tips of the crack (R). The intensity of the stress field at the tip is given by the stress-intensity factor $K = \sigma\sqrt{\pi c}$.

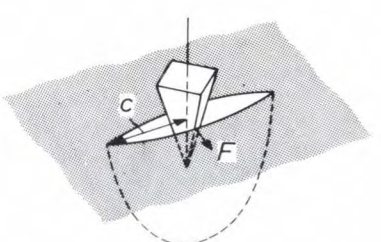


Fig. 25. Diagram of a 'half-penny' crack of radius c caused by a Vickers indenter. The wedge action pushes the surfaces of the crack apart with a force F.

[8] K. Peter, Glastech. Ber. **37**, 333, 1964.
[9] See for example B. R. Lawn and T. R. Wilshaw, Fracture of brittle solids, Cambridge University Press, 1975.
[10] B. R. Lawn and E. R. Fuller, J. Mat. Sci. **10**, 2016, 1975.

calculated stress field below an obliquely loaded point [11]. The horizontal force F_t opens the crack, and is therefore related to F in fig. 25. The widths of the groove and of the cracking and chipping regions for a coarse-grained NiZn ferrite are shown in *fig. 27* as a function of the load. In agreement with eq. (13) the width of the cracking region, $2c$, is approximately proportional to $F_n^{\frac{2}{3}}$.

For scratching as shown in fig. 26 we have set up a crack-formation model, and from this we have obtained a relation between c and F_t [7]:

$$F_t = \alpha K_c c\sqrt{c + \tfrac{1}{2}\beta b}, \qquad (16)$$

where b is the groove width (which is related to F_t by (10) and (3)). If the cracks are large ($c \gg b$), eq. (16) becomes equivalent to eq. (13). The constants α and β are determined by the shape and orientation of the diamond.

We have measured b and c for a square pyramidal diamond with an apex angle of 136° and a leading plane, for a series of loads. Eq. (16) gives a good representation of the measured results for various materials, with $\beta = 12.4$. The constant α was determined by measuring K_c by one of the standard methods for one of the materials (glass) [9]; the result was $\alpha = 2.24$.

It is now possible to derive the minimum force for crack formation from (16), in the same way as above for 'half-penny' cracks from (13). The result is almost the same:

$$F_{t\,\min} = C_s K_c^4 / H_s^3, \qquad (17)$$

where H_s is the scratch hardness (see eq. (10)); C_s is determined by α, β and the force ratio f:

$$C_s = \alpha^4(1 + \beta)^2/f^3.$$

Inserting the values of α and β given above and $f = \cot 68° = 0.40$, we find $C_s = 0.7 \times 10^5$.

Equations (15) and (17) show that a material that is very crack-resistant in tensile tests (high K_c) can nevertheless crack easily in indentation and scratching if it is also very hard.

The high sensitivity of $F_{t\,\min}$ to K_c and H_s explains the very different scratch and cracking behaviour of different materials. *Table II* gives K_c and H_s and the minimum load ($F_{n\,\min} = F_{t\,\min}/0.4$) calculated from them by means of (17) for a number of materials. *Fig. 28* presents this information in a K_c-H_s diagram. In our scratching experiments the load varied from 0.05 to 5 N; the load per diamond in very heavy grinding is estimated at no more than 1 N (solid line in fig. 28). Table II predicts, in agreement with our results, that under these circumstances crack formation cannot be avoided in MgO, MnZn ferrite and television glass, that Ferroxdure and Al_2O_3 must be heavily loaded to

Table II. The minimum load for crack formation $F_{n\,\min}$, calculated from the critical stress-intensity factor K_c and the scratch hardness H_s by using eq. (17), with $C_s = 0.7 \times 10^5$ and eq. (3), $F_n = F_t/f$. The value used for C_s should apply to diamonds with an apex angle of 136° and a leading plane. This value is very sensitive to the force ratio f; this was taken as cot 68° = 0.4. If this should fall to say 0.3, a value often found in practice, C_s and $F_{t\,\min}$ become more than twice as large.

	K_c ($MNm^{-\frac{3}{2}}$)	H_s [*] (GNm^{-2})	$F_{n\,\min}$ (N)
MgO, single crystal {100}	0.8 [9]	17 [13]	0.015
MnZn ferrite	1.2 [6]	22 [7]	0.035
Television glass	1.0 [6]	16 [7]	0.042
Al_2O_3	4.8 [6]	70 [13]	0.27
Ferroxdure	2.5 [6]	24 [13]	0.50
Steel (high C content, hardened)	20 [12]	10 [4]	28000

[*] At a scratching speed of 1 mm/s.

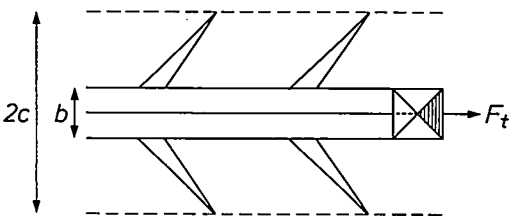


Fig. 26. Diagram of lateral cracks produced in scratching. F_t tangential force, b width of the groove, $2c$ width of the cracking region.

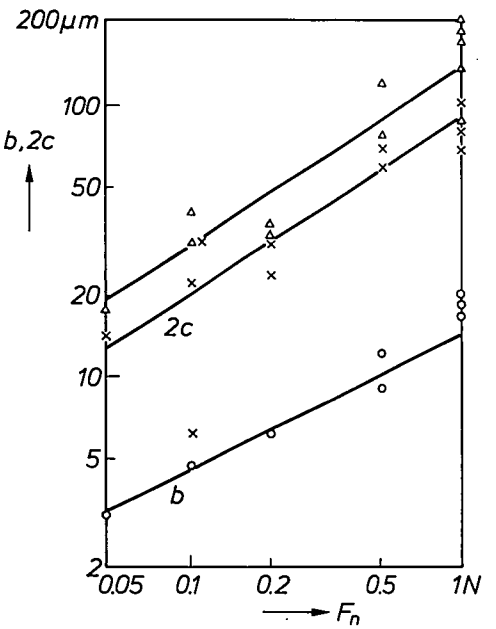


Fig. 27. The width b of the groove (○), the width $2c$ of the cracking region (×) and the width of the chipping region (△) as a function of the load F_n for a coarse-grained NiZn ferrite. The groove width b is approximately proportional to $F_n^{\frac{1}{2}}$, and the width of the cracking region $2c$ to $F_n^{\frac{2}{3}}$.

[11] G. M. Hamilton and L. E. Goodman, Trans. ASME E (J. appl. Mech.) 33, 371, 1966.
[12] P. Kenny and J. D. Campbell, Progr. Mat. Sci. 13, 135, 1967.
[13] Unpublished results of measurements made at these laboratories.

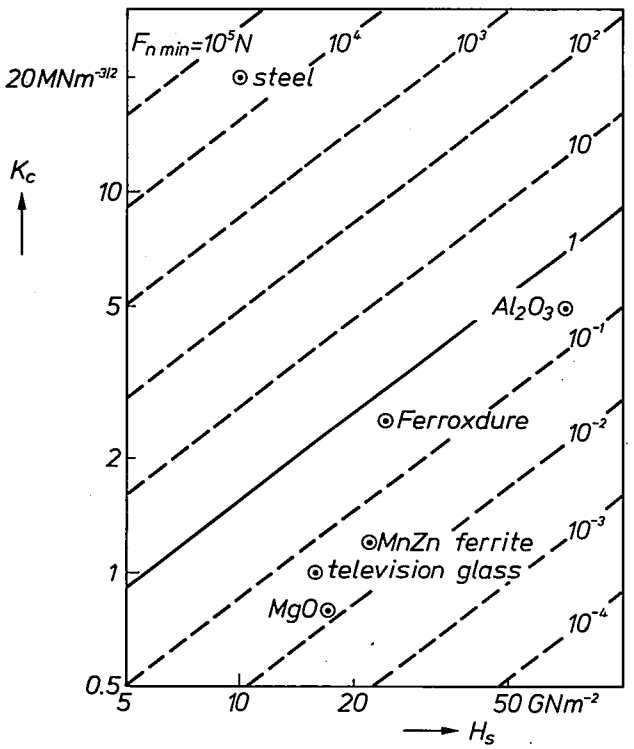


Fig. 28. The K_c-H_s diagram with lines of constant $F_{n\,min}$ ('minimum load for crack formation'), as given by eq. (17) with $C_s = 0.7\times10^5$ and $f = 0.4$. The solid line gives the estimated load per diamond in very heavy grinding ($F_n = 1$ N). The points for different materials represent the values of Table II.

produce cracks and that cracking in steel is impossible. The cracking boundary in Table II for MnZn ferrite and glass agrees in order of magnitude with figs 22 and 23.

As the load increases, not only does the width of the cracking region increase, but the number of cracks per unit length often increases as well. This can largely be explained by the increasing number of flaws in the material ('crack nuclei') at which the cracking condition is satisfied.

The growth in number and magnitude of the lateral cracks eventually leads to chipping; the boundary II in figs 22 and 23 has then been reached. The interaction with other cracks probably comes into play here. Chipping is therefore difficult to analyse quantitatively, and we shall not discuss it further in this article.

Although we assumed for simplicity that a crack does not propagate if K is smaller than K_c, it does in fact usually do so in the situation of fig. 24, but at a velocity v_c that decreases very strongly as K decreases (*fig. 29*). If this relation between K and v_c is taken into account, the equations (13) to (17) remain valid, provided K_c is replaced by K. This is then the value of the stress-intensity factor at the tip of the crack at the associated crack velocity. One of the things explained by this relation is the speed dependence of boundary *I* in figs 22 and 23, if we assume that the crack velocity can be identified with the scratching speed. This assumption appears to be justified for the conditions in our experiments [7].

The knowledge of K as a function of the crack velocity (fig. 29) is essential for calculating the strength and life of objects that are mechanically loaded. The well-known methods [8] for determining $K(v_c)$ are very labour-intensive, however. If the highest accuracies are not required, scratching experiments offer a simple alternative: K can be determined via (16) from a measurement of the width of the cracking region ($2c$) and the groove width (b) at a given load ($F_n = F_t/f$) and the desired velocity. We have been able to determine the K-v_c curve for a number of materials in this way [7].

Conclusions

The objective of the investigation, in brief, was to establish how specific energy and surface damage are related to machine setting and properties of the material. The conclusions are as follows.

The specific energy e is primarily a function of the material: the harder the material, the greater e. The effect of the machine setting on the specific energy can be considerable, however, and here a second material property that shows up clearly in scratching experiments often has an important part to play. This is the minimum force per diamond for crack formation ($F_{t\,min}$). For simplicity we shall also refer to *the* force per diamond in grinding as well as in scratching in order to compare it with $F_{t\,min}$, but in fact this concept is not closely defined, because the diamonds on a wheel do not protrude by the same amount.

If we examine this effect of the machine setting on the specific energy and on the surface damage, we find that in practice the wheel speed v_s has no significant effect. It is therefore always advisable to try to use higher speeds, since the power required for a given rate of volume removal (the product of the wheel speed and the horizontal grinding force) can then be supplied at a smaller force, thus limiting surface damage and the wear of the wheel. This is true if heating effects do not

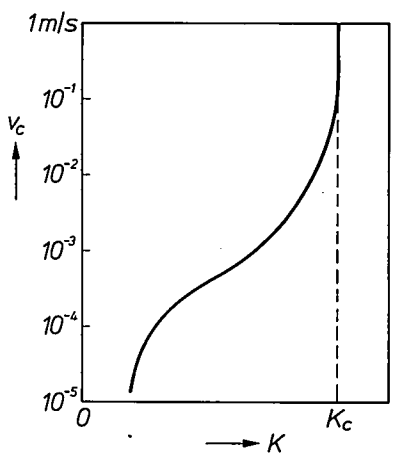


Fig. 29. Diagram showing the relation between the crack velocity v_c and the stress-intensity factor K. The v_c-scale is logarithmic, the K-scale linear. At $K = K_c$ the tip of the crack propagates at approximately the velocity of sound.

come into play. Pronounced speed effects have been observed in scratching, but these are at speeds of microns to millimetres per second, which are of no significance in grinding.

The specific energy and the surface damage are strongly dependent on the grinding forces, however. If the tangential force per diamond, F_t, is smaller than $F_{t\,min}$, grinding is a form of 'ploughing'. The effect of the force on the specific energy is then still relatively small; the surface irregularities remain limited to grooves that are impressions of the diamond points. If F_t is larger than $F_{t\,min}$, however, effects of a different order occur because of cracking and chipping. Much more material is then removed per joule than in ploughing, and the surface damage is much greater.

A high speed and a large force are therefore both desirable for rapid and efficient removal of material; a large force is particularly useful if $F_{t\,min}$ is exceeded. On the other hand, surface damage in the form of cracks can be prevented by making F_t smaller than $F_{t\,min}$; and even then the roughness decreases as F_t becomes smaller. In many cases it will therefore be preferable to grind in two stages: coarse grinding (fast) first, until the correct shape has been obtained; this leaves a damaged layer of thickness equal to the crack depth; then this layer is ground away with a force per diamond small enough (i.e. the removal is slow enough) to give the required quality of finish. If the grinding is done in several stages the operation can be made even more efficient.

The minimum force for crack formation, $F_{t\,min}$, depends very strongly on both the crack resistance and the hardness (see eq. 17), and consequently differs greatly from material to material (see Table II and fig. 28). For example, $F_{t\,min}$ in steel is so large that cracks will never form; Ferroxdure and Al_2O_3 must be very heavily loaded to take advantage of chipping. In materials like MnZn ferrite, glass and MgO, on the other hand, $F_{t\,min}$ is so small that other processes (e.g. lapping, polishing) have to be used to produce a good finish.

Summary. In setting up a grinding machine the 'rate of volume removal' Z and the 'continuum chip thickness' h_{eq} are important quantities. At a given 'specific energy' e for the material, h_{eq} determines the force exerted on the workpiece by the grinding wheel, and hence the surface damage. A high circumferential velocity v_s for the wheel has the advantage that the grinding power P ($= eZ$) required is obtained with a small force. The specific energy (generally high for hard materials) is usually dependent on the machine setting. The influence of v_s, at a given h_{eq}, is very small. The influence of h_{eq} (or the force) on e is also very small in 'tough' materials, but is considerable in brittle materials, because of chipping. The single-point scratching experiment has been used for further investigations of groove formation, cracking and chipping. The groove formation corresponds to the 'ploughing' model; grinding a tough material is 'ploughing'. The minimum force required for crack formation in scratching depends strongly on the properties of the material. This explains the very different scratching and grinding behaviour of the materials investigated.

Real-time orthogonal transformation of colour-television pictures

H. Bacchi and A. Moreau

The amount of information contained in a picture can be very considerable, and the transmission of moving pictures from one place to another corresponds to a very large flow of information. Nature itself has difficulties in handling so much information — the image information picked up by the 150 million or more rods and cones of the human eye cannot be conveyed to the brain by the optical nerve without previous processing, even though the nerve has a million fibres. Nerve cells in the retina provide cross-connections between the receptors, thus reducing the redundancy in the image signal. Something of the kind is also aimed at in the digital transmission of television images. Direct digitizing of each image point results in a bit rate that is too high for the existing digital transmission channels, which were really intended for telephony. It is possible to reduce the bit rate by applying certain geometrical transformations to the image. These are rather complicated operations, which must be performed in real time for television transmission. The authors have built an equipment that will do this.

Introduction

In television, as in many branches of electrical engineering, the use of digital techniques is increasing. They are already in use for image storage and for delaying a television signal to make it synchronous with another television signal. Digital methods are also employed for conversion from one line standard to another.

A digital signal has equally obvious advantages for the long-distance transmission of television signals. Since the signal consists of a train of pulses, the pulse can be regenerated at a repeater station by removing noise superimposed on the signals during transmission. This greatly improves the signal-to-noise ratio, and for this reason digital transmission has for many years been used in long-distance telephony.

The problem in the digital transmission of television pictures is the high rate of information to be transmitted. The straightforward digitizing of a colour-television signal results in a bit rate of 120 megabits per second. This is a very high rate, and it would take up too much of the available capacity if the existing digital telephone-transmission network were also to be used for television. The telephone channels in this network are grouped in an internationally agreed hierarchy; it would be desirable to limit the transmission of television signals to a group of 480 telephone channels, with the associated bit rate of 34 Mbit/s [1]. The bit rate would therefore have to be reduced by about four times, yet without any significant loss of picture quality.

With this aim in view a study was started at LEP to find a more efficient method of transmitting television signals — more efficient than transmitting brightness and colour information for every single image point. Various methods are conceivable. All of them are based on reducing the redundancy in the TV signal. The redundancy is obvious: many neighbouring image points contain the same information, and it is clearly unnecessary to transmit that information for every image point.

The methods of reducing redundancy fall into two main groups, both of which we investigated. One group includes the processes in which the differences between the successive image points, and not their content, are digitized and transmitted (differential pulse-code modulation — DPCM). This process can be carried out with relatively simple equipment and gives good results as long as the bit rate does not have to be reduced by more than three times [2].

H. Bacchi, ingénieur A.M. and E.S.E., and A. Moreau, ingénieur E.S.E., are with Laboratoires d'Electronique et de Physique Appliquée (LEP), Limeil-Brévannes (Val-de-Marne), France.

[1] CCITT Recommendation G. 751 (1977). See also J. Poncin, Le codage numérique de la télévision, Acta Electronica **19**, 215-223, 1976.

[2] J. O. Limb and F. W. Mounts, Digital differential quantizer for television, Bell Syst. tech. J. **48**, 2583-2599, 1969.
A. Habibi, Comparison of *n*th-order DPCM encoder with linear transformations and block quantization techniques, IEEE Trans. **COM-19**, 948-956, 1971.

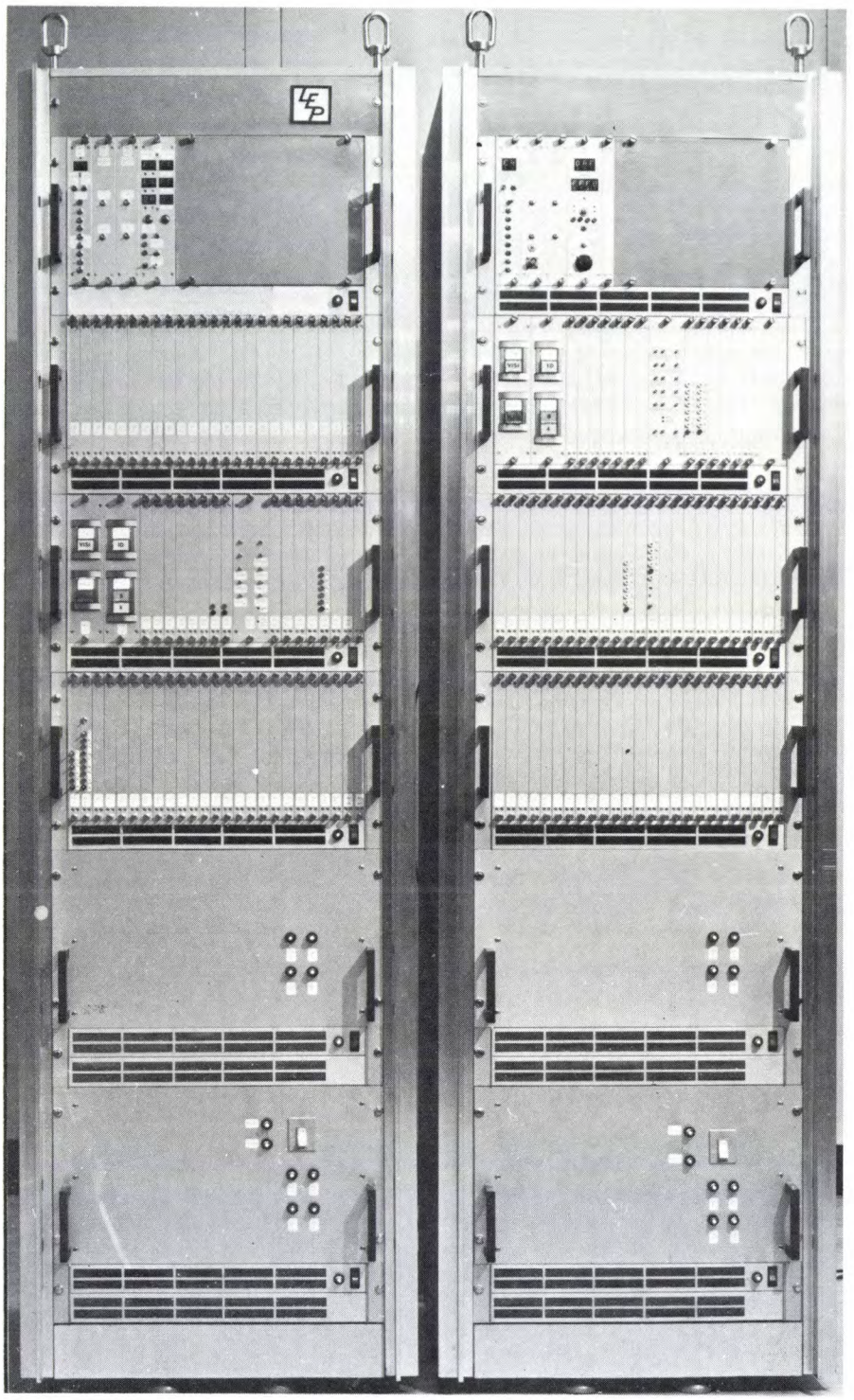


Fig. 1. Two racks containing the equipment for real-time direct and inverse orthogonal transformation of colour-television pictures. One rack contains the equipment for the direct transformation, the other the equipment for the inverse transformation. A choice can be made between the Hadamard and the Haar transformation matrices. The TV pictures are divided into subpictures; the shape and size of the subpictures can be chosen from several possible values. Each subpicture is expressed by the transformation as a sum of basis pictures, each with its own weighting factor. By transmitting only the most-significant weighting factors the bit rate can be greatly reduced.

Since we required a greater reduction in bit rate, we devoted more attention to the second group, the orthogonal transformations [3]. In this group the TV picture is divided into a large number of subpictures, usually square; each subpicture is 'expanded as a series' and treated as a superposition of a number of basic patterns, each with its own weighting factor. It is then sufficient simply to transmit the number of the basic pattern and the associated weighting factor. The subpicture is reconstructed from this at the receiving end in an inverse transformation. The orthogonal transformation of TV pictures requires rather bulky equipment.

There are various ways of dividing a subpicture into 'orthogonal' basic patterns that are not capable of further division. We compared various transformations in computer simulations, and with suitable coding we obtained bit-content reductions of up to ten times. These computer simulations all related, however, to stationary pictures. In real TV transmission the picture changes and there may be signal interference during transmission.

To investigate the usefulness of transformations under practical conditions it is necessary to be able to transform the pictures in real time. In a study carried out for the CCETT [4] we have built equipment that will do this (*fig. 1*). Two transformations were selected for this investigation, both of which can be carried out with logic circuits of reasonable dimensions: these are the Hadamard transformation and the rather similar Haar transformation [5]. Before describing this equipment, we shall first look more closely at the fundamentals of orthogonal picture transformations [6].

Transform picture coding

Redundancy in TV pictures

The television-transmission channels in common use today have a vast capacity for information transfer: they can transmit the brightness (and the colour) of each one of the 400 000 or so picture elements of a TV picture independently of the values of the adjacent elements, for 25 completely different pictures a second. This information capacity is only rarely used to the full, for the following reasons.

In the first place, two consecutive TV pictures are very much the same and, in the absence of noise, some parts of the picture usually do not change at all. The repeated transmission of the unchanged parts of the picture is an example of redundancy in the television signal that could be called 'time redundancy'. However, to economize on transmission channels by only transmitting the differences between successive pictures there would have to be some kind of memory device at the receiving end that could contain an entire TV picture — a rather expensive addition. Although such systems are being studied, we shall not deal with them in this article.

There is another kind of redundancy, however, called 'spatial redundancy', which we will examine here. TV pictures tend to be composed of areas of nearly constant brightness separated by well-defined boundaries [7]. When the pictures are being scanned, the full bandwidth of the television channel is momentarily required for good definition of these boundaries, but between such boundaries only slowly varying brightness levels have to be transmitted. This implies that the full capacity of the channel is not continuously utilized.

It is worth looking for a more efficient method of encoding the picture information than straightforward transmission of the scanned picture — whether or not the transmission is made in analog form or as digitized samples. The new method would be required to spread the information flow more evenly so that the required channel capacity could be determined by the mean rate of information and not by occasional maximum values.

We shall approach this problem by considering the picture structure. Nearly all pictorial configurations contain significant structure. This structure becomes obvious as soon as we divide the picture into an array of subpictures. Many of these subpictures will lie entirely within an area of uniform brightness and will possess no detail at all. Others will be divided by one boundary line into two parts and will thus present a very simple basic pattern. Others again will be divided by two boundary lines, for example a horizontal one and a vertical one, into four parts and thus have a slightly more complicated basic pattern. More elaborate patterns will also occur. Each subpicture can now be treated as the superposition of a varying number of different basic patterns or 'basis pictures', each weighted by a factor that determines its contribution to the total subpicture (see *fig. 2*).

The idea is familiar from Fourier analysis. To take a well-known one-dimensional example, a periodic signal, e.g. an audio signal, may be represented as the sum of a number of Fourier components, the fundamental and harmonics, each of which has its Fourier coefficient as a weighting factor. In principle the Fourier series may be infinite, but higher harmonics tend to be weaker than the lower ones and a good approximation may be obtained by truncating the series after a number of terms. The listener has to decide which approximation is acceptable.

Although in this example the audio signal is a function of time, the same reasoning applies to our TV

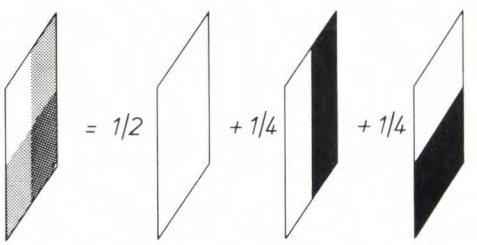


Fig. 2. Each subpicture may be considered as the sum of a number of basis pictures, each with an appropriate weighting factor. The basis pictures are elementary geometrical patterns that increase in complexity with increasing rank number.

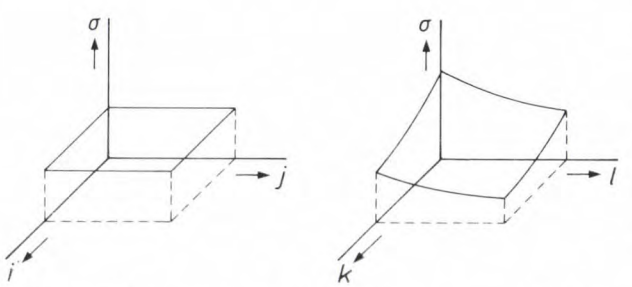


Fig. 3. Unlike the grey levels x_{ij} of the picture elements (*left*), the weighting factors y_{kl} have a variance that decreases rapidly with increasing order (*right*). In encoding for transmission the greater part of the available bits should therefore be devoted to the low-order weighting factors. A reduction of the total bit rate can be achieved in this way.

[3] P. A. Wintz, Transform picture coding, Proc. IEEE **60**, 809-820, 1972 (contains an extensive bibliography).

[4] Centre Commun d'Etudes de Télévision et Télécommunications (Rennes, France), a French government organization responsible for research in this field.

[5] H. Bacchi and H. Tchen, Réalisation d'un transformateur d'images en temps réel, Ann. Télécomm. **30**, 363-373, 1975.

[6] H. Bacchi, Considérations sur le choix des transformations orthogonales et sur leurs perspectives, Acta Electronica **19**, 299-332, 1976.

[7] The treatment is confined here to monochrome pictures. Colour pictures present no essentially new problems; they can be handled by applying the techniques described here simultaneously to a luminance signal and to two chrominance signals, which are of lower resolution.



Fig. 4. Illustrating the small effect on picture quality that can be achieved in spite of a considerable reduction in the bit rate, when appropriate coding is used.

subpictures, where the brightness is a function of position. However, a picture is two-dimensional and therefore requires a two-dimensional analysis. Two-dimensional Fourier analysis is one of the possible solutions, in which the individual terms of the expansion correspond to the basis pictures. Terms of higher order will represent more complex basis pictures. Complexity in this context is synonymous with spatial frequency — the number of black-white transitions per unit length.

To find an efficient way of encoding the TV signal, the relative amplitudes of low and high spatial frequencies have to be determined. This requires an investigation of the statistics of the TV signal. After sampling and quantization, the brightness values of the picture elements are represented by a set of numbers. The picture elements may be regarded as statistical quantities (random variables) but they are not uncorrelated because, as noted above, all natural images and most man-made pictures contain significant structure. Efficient source encoding might be attempted by analysing and classifying this structure. Unfortunately, however, only a limited number of simple structures can be distinctly observed, for instance the areas of constant brightness and sharp transitions mentioned above. A straightforward measurement of statistical properties is of little avail; only first- and second-order



Left: Original picture; 8 bits per picture element. *Above:* Picture after Hadamard transformation and encoding with 0.9 bit per picture element.

statistics, such as mean values and standard deviations, can be accurately measured, and these give only a coarse indication of the picture information.

The constancy of the brightness in some picture areas indicates, in statistical terms, that the correlation between neighbouring picture elements is high. Conversely, at transition boundaries in the picture the correlation is low. However, taken as the average over many pictures the correlation is relatively high (0.7 to 0.95), which indicates that constant-brightness areas are predominant in natural pictures. Obviously, if a TV signal varies slowly for most of the time, it consists mainly of low-frequency components. High-frequency components due to transients will occur rarely or have a low magnitude.

Thus, after a picture has been analysed into basis pictures, the weighting factors associated with the low spatial frequencies will be much greater than those associated with the high spatial frequencies. This is only statistically true, of course. A particular picture may conceivably contain nothing but high frequencies. Nevertheless, statistical analysis of many television pictures shows that after transformation the variance of the weighting factors of the lower-order basis pictures is much greater than that of the higher-order basis pictures (*fig. 3*).

Fewer bits will therefore be sufficient for encoding the weighting factors for the higher-order basis pictures than for the lower-order ones. In addition the eye is less sensitive to errors introduced into high-frequency components than to errors in low-frequency components. This also helps to reduce the number of bits required for encoding the higher-order weighting factors. These can be more coarsely quantized or even discarded. Experiments have shown that the bit-rate reduction that can be achieved in this way is substantial and may even be as high as 8 to 10 times before picture quality is seriously affected (*fig. 4*).

The full sequence of operations to which the TV picture is subjected in orthogonal transformation is summarized in block-diagram form in *fig. 5*. After sampling and analog/digital conversion (A/D) the television picture is divided into subpictures, and the transformation (T) is applied to these subpictures. The resulting coefficients are quantized (Q) and the series of binary numbers representing the quantized coefficients are encoded in the most economical way (C). After transmission, the same operations are carried out in reverse order at the receiving end.

The derivation in the transformation unit T of the coefficients or weighting factors for the basis pictures from the brightness values of the individual picture elements amounts in fact to a coordinate transformation. We shall now discuss how such a coordinate transformation can be performed and consider the transformations we have chosen for use in practice.

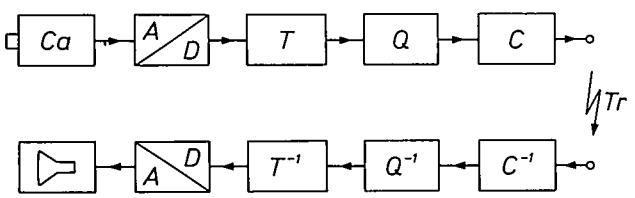


Fig. 5. Block diagram of TV transmission with orthogonal picture transformation. *Ca* camera. *A/D* sampling and analog/digital conversion. *T* division into subpictures and transformation. *Q* quantization of the weighting factors. *C* coding. *Tr* transmission. C^{-1} decoding. Q^{-1} weighting-factor amplitude recovery. T^{-1} inverse transformation and redistribution of picture elements to form picture lines. *D/A* digital/analog conversion.

Orthogonal transformations

So that we can give a mathematical description of the transformation to new coordinates, we represent a subpicture consisting of $N \times N$ picture elements by a matrix

$$X = \begin{pmatrix} x_{11} & \dots & x_{1N} \\ & & \\ x_{N1} & \dots & x_{NN} \end{pmatrix}, \qquad (1)$$

where the elements x_{ij} are a measure of the brightness levels of the individual picture elements.

The basis pictures referred to above may be represented by similar matrices B_{kl}. As indicated in fig. 2, the subpictures X may now be considered equivalent to a series of basis pictures, each with its weighting factor y_{kl}:

$$X = \sum_{k=1}^{N} \sum_{l=1}^{N} y_{kl} B_{kl}. \qquad (2)$$

The information contained in the N^2 picture elements is now expressed by a set of N^2 weighting factors y_{kl}. Instead of the positional coordinates of the picture elements in the subpictures the set of N^2 basis pictures $\{B_{kl}\}$ has been introduced as a new coordinate system.

A condition for the transformation of the subpictures into the new coordinates B_{kl} is that the coordinates should be orthogonal, so that none of them can be expressed as a linear combination of the others. In this sense the terms of a Fourier series are orthogonal.

From eq. (2), which represents a set of N^2 linear equations, no more than N^2 different weighting factors y_{kl} can be determined. This means that the number of basis pictures of progressively increasing complexity is also limited to N^2. This imposes an upper limit on the resolution with which the subpictures can be transformed. The same upper limit, however, has already been imposed by the representation of the subpicture by a finite number of N^2 picture elements. Thus, the transformation process does not in principle imply a loss of resolution.

Equation (2) is equivalent to a double matrix multiplication:

$$X = \Phi Y \Phi^{\mathrm{T}}, \qquad (3)$$

where

$$Y = \begin{pmatrix} y_{11} & \dots & y_{1N} \\ & & \\ y_{N1} & \dots & y_{NN} \end{pmatrix}$$

is a matrix containing the weighting factors y_{kl}, Φ is a particular matrix connected with the set $\{B_{kl}\}$ of basis pictures, and Φ^{T} is the transpose of Φ (rows become columns and columns become rows). A double matrix multiplication is required because the array of picture elements to be transformed is two-dimensional.

In fact, eq. (3) gives the inverse transformation, from which the image X can be reconstructed. If

$$\Phi \Phi^{\mathrm{T}} = \Phi^{\mathrm{T}} \Phi = I, \qquad (4)$$

where I represents the identity matrix, the direct transformation is

$$Y = \Phi^{\mathrm{T}} X \Phi \qquad (5)$$

and the direct and inverse transformations are effected by means of the same matrix Φ. The condition (4) implies that the transpose Φ^{T} is identical to the inverse

matrix Φ^{-1}; in the case of a matrix of real numbers, as here, the matrix is then said to be orthogonal.

To find the relation between the matrix Φ and the basis pictures B_{kl}, we write:

$$Y = \sum_{kl} y_{kl} E_{kl}, \tag{6}$$

where E_{kl} is a matrix of order $N \times N$ in which the element of order kl is unity and all other elements are 0:

$$E_{kl} = \begin{pmatrix} 0 & & 0 \\ & 1 & \\ 0 & & 0 \end{pmatrix}.$$

Instead of (3) we can now write:

$$\begin{aligned} X &= \Phi\left(\sum_{kl} y_{kl} E_{kl}\right)\Phi^{T} \\ &= \sum_{kl} y_{kl}(\Phi E_{kl}\Phi^{T}) \\ &= \sum_{kl} y_{kl}\, \Phi_k \Phi_l^{T}, \end{aligned} \tag{7}$$

where Φ_k denotes the kth column of matrix Φ. It follows that

$$B_{kl} = \Phi_k \Phi_l^{T}, \tag{8}$$

which is the desired relation.

Hadamard and Haar transformations

Several orthogonal transformations exist, each of which has its own transformation matrix Φ and an associated set $\{B_{kl}\}$ of basis pictures. One of them is the Fourier transformation, which we have already mentioned once or twice. In choosing an orthogonal transformation it is necessary to weigh the computing effort required against the performance in terms of bit-rate reduction. For the experimental equipment shortly to be described we chose two transformations for which the associated calculations are much simpler than for other transformations. These are the Hadamard transformation and the Haar transformation. *Fig. 6a* shows the 8×8 Hadamard matrix [8] and fig. 6*b* the 8×8 Haar matrix [9]. Both transformations have the advantage that they can be computed with a fast algorithm that only requires additions and subtractions, unlike the Fourier transformation.

The basis pictures connected with the 8×8 Hadamard matrix are shown in *fig. 7*. They only contain white and black, since the Hadamard matrix only contains the numbers $+1$ and -1. Fig. 7 clearly demonstrates the growing complexity of the basis pictures — or their increasing content of high spatial frequencies — with increasing rank numbers. It is easy to see that subpictures consisting mainly of basis picture B_{88} or one of its neighbours will be extremely unlikely, as discussed earlier in connection with the encoding of the coefficients.

Operation of the equipment

Specifications

In accordance with the specifications issued by CCETT, the equipment built at LEP — which performs the functions of blocks T and T^{-1} in fig. 5 — can execute the Hadamard and Haar transformations on a colour-TV signal in real time. The colour-TV signal consists of a luminance signal sampled at a rate of 12 MHz, with each sample translated into an 8-bit word (256 grey levels), and two 'colour-difference' signals, one for the red and one for the blue component, which are sampled at a rate of 3 MHz and then also translated into 8-bit words.

In the experiments carried out with the equipment, it is necessary to have several options for the size of the

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix}$$

a

$$Hr = \begin{pmatrix} 1 & 1 & \sqrt{2} & 0 & 2 & 0 & 0 & 0 \\ 1 & 1 & \sqrt{2} & 0 & -2 & 0 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & -2 & 0 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & -2 & 0 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & 2 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & -2 \end{pmatrix}$$

b

Fig. 6. *a*) The 8×8 Hadamard matrix H. It contains only the elements $+1$ and -1, which reduces the multiplication of a given matrix by H to simple additions and subtractions of the given matrix elements. *b*) The 8×8 Haar matrix Hr. To obtain orthogonal matrices, as required for the picture transformations, each element of each matrix must still be divided by $\sqrt{8}$.

[8] J. Hadamard, Résolution d'une question relative aux déterminants, Bull. Sci. math., series 2, **17**, part 1, 240-246, 1893.

[9] A. Haar, Zur Theorie der orthogonalen Funktionensysteme, Math. Annalen **69**, 331-371, 1910.

subpictures or 'blocks'. In the equipment built at LEP these can be one-dimensional (N successive picture elements of a TV line) or they can form a square array of $N \times N$ picture elements. In both cases N may be 4, 8 or 16.

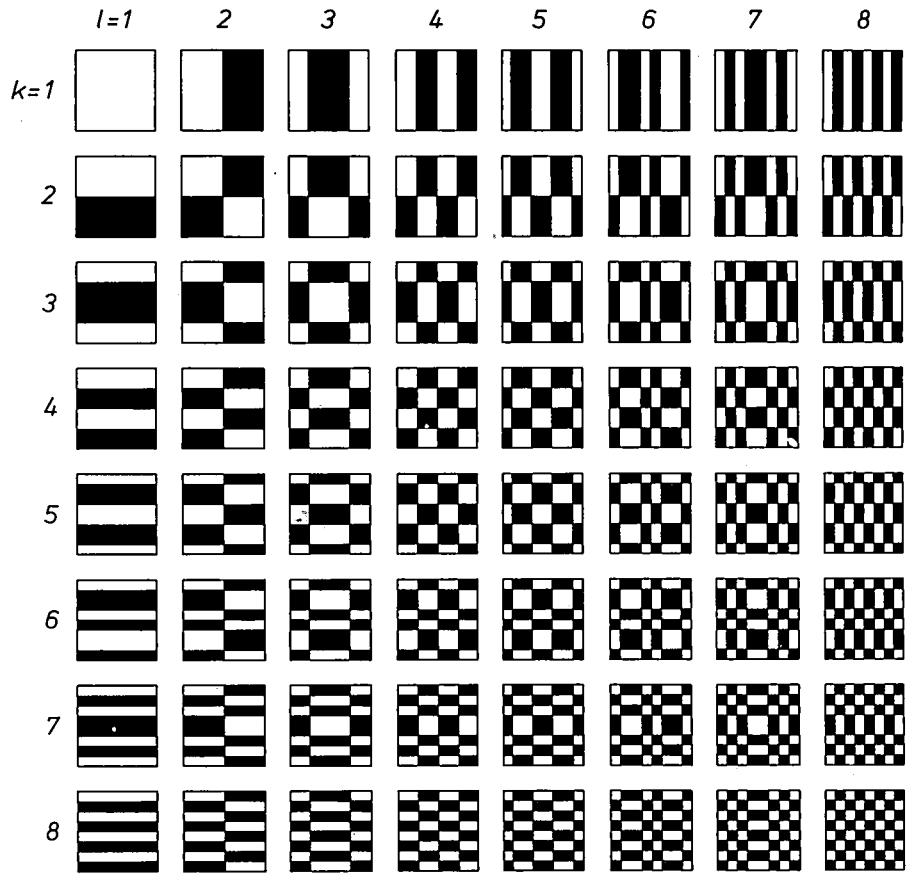


Fig. 7. The basis pictures B_{kl} for the 8×8 Hadamard matrix. As this matrix contains only the extreme values $+1$ and -1, the basis pictures contain only black and white and no intermediate grey tones. The figure illustrates the increasing content of high spatial frequencies with increasing rank numbers k and l.

General description

To perform the required operations on square arrays containing a maximum of 16×16 picture elements it is necessary to have two memory devices, each of which can store 16 complete colour-TV lines. The incoming digitized picture elements are stored line by line in one of these memories, e.g. M_1 in *fig. 8*. Once 16 lines have been stored, M_1 will be read out block by block, while at the same time the 16 following lines are entered into the other memory, M_2.

The blocks are processed in an arithmetic unit that performs the double matrix multiplication by splitting it into two single matrix multiplications:

$$Y = \Phi^{\mathrm{T}} X \Phi$$

is equivalent to

$$Y^{\mathrm{T}} = (X\Phi)^{\mathrm{T}}\Phi. \qquad (9)$$

First $P = X\Phi$ is computed, then P is transposed and finally $Y^{\mathrm{T}} = P^{\mathrm{T}}\Phi$ is computed. This means that multiplication by Φ is performed twice, each time by the same circuit. For transposing P an intermediate memory is available to which P is written row by row and from which it is then read out column by column.

The inverse transformation is accomplished in a similar manner. First $\hat{X} = \Phi(\Phi\hat{Y}^{\mathrm{T}})^{\mathrm{T}}$ is computed (the

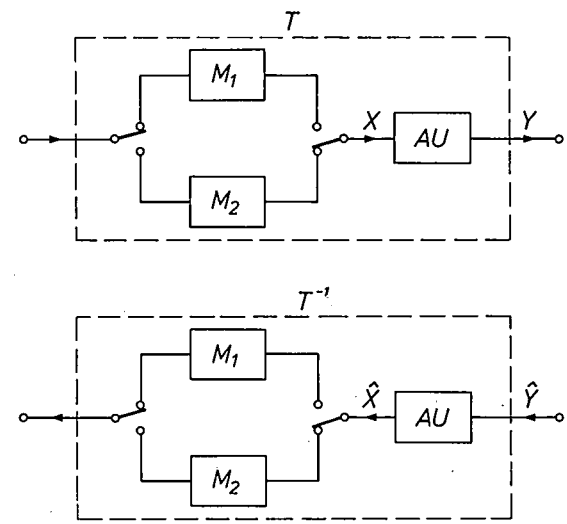


Fig. 8. Block diagrams for the direct-transformation unit T and inverse-transformation unit T^{-1}. Both units contain two memories M_1 and M_2 with a capacity of 16 colour-TV lines each. These memories are used alternately — in the direct-transformation unit T to receive 16 TV lines and then to read out 16×16 blocks, and vice versa in unit T^{-1}. The matrix multiplications are carried out in the arithmetic unit AU.

sign ˆ serves to denote the TV signal after reception). Then two memories, each suitable for 16 TV lines, are used to recover the original line-by-line scanning.

After this general outline, let us now look in more detail at how the Hadamard and Haar transformations are reduced to a set of additions and subtractions and performed by a fast algorithm. This algorithm is very similar in both cases, but each requires its specific preliminary manipulations.

The Hadamard transformation

Fig. 6*a* shows an 8×8 Hadamard matrix. This matrix, which we shall denote by H, is not the matrix Φ_H required for the Hadamard transformation, because it is not orthogonal. To make it orthogonal it has to be divided by $\sqrt{N}$, so that $\Phi_H = H/\sqrt{N}$; in our example this corresponds to dividing by $\sqrt{8}$. The matrix multiplication is performed by H, however, and not by Φ_H, because H contains only $+1$ and -1, so that the matrix multiplication only consists of a series of additions and subtractions. The transformed picture will then be

$$Z = H^T X H = N\Phi_H{}^T X \Phi_H = NY, \qquad (10)$$

which means that all the coefficients are really N times too large. Nevertheless a division by N is omitted because it would introduce rounding-off errors since we are only using integers. It was decided to avoid rounding-off errors consistently, because we only wanted to assess the loss of quality introduced by the encoding and decoding.

The received picture is thus $N\hat{Y}$. The inverse transformation is

$$H(N\hat{Y})\,H^T = N^2\hat{X}, \qquad (11)$$

which will yield the reconstructed picture $\hat{X}$ after a division by N^2. We shall return to this later.

The Haar transformation

Like the matrix H, the 8×8 Haar matrix Hr shown in fig. 6*b* is not orthogonal. Again, the actual transformation matrix is $\Phi_{Hr} = Hr/\sqrt{N}$, and the factor $1/\sqrt{N}$ is treated in the same way as for the Hadamard transformation.

The Haar matrix Hr contains $\sqrt{2}$, or, more generally, powers of $\sqrt{2}$. To avoid multiplications by $\sqrt{2}$, we write $Hr = G\Lambda$, where G contains only $+1$, 0 and -1 and Λ is a diagonal matrix containing powers of $\sqrt{2}$ (see *fig. 9*).

The direct and inverse transformations may be represented by:

$$Y = \frac{1}{N}\Lambda G^T X G \Lambda,$$

$$X = \frac{1}{N} G\Lambda Y \Lambda G^T = \frac{1}{N^2} G\Lambda^2 G^T X G \Lambda^2 G^T. \qquad (12)$$

The matrix Λ^2 only contains powers of 2 and is easily handled by shifting binary numbers; the matrix Λ, on the other hand, which contains powers of $\sqrt{2}$, is difficult to process. For this reason the multiplication by Λ^2 that appears in (12) is performed at the transmitting end, and the matrix $Y_P = \Lambda^2 G^T X G \Lambda^2$ is transmitted. We call the multiplication by Λ^2 'Haar weighting'. The original picture is reconstructed in the receiver by $\hat{X} = (1/N^2) G \hat{Y}_P G^T$.

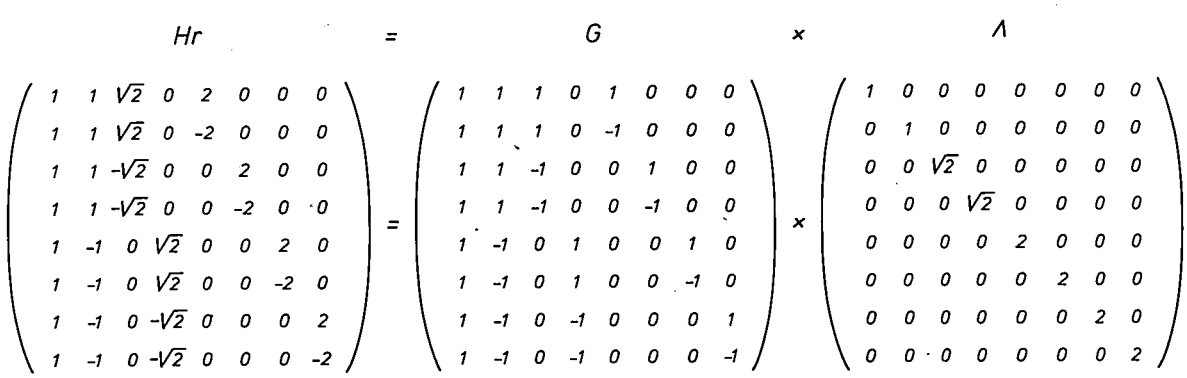


Fig. 9. Analysis of the Haar matrix Hr into a matrix product $G\Lambda$. The elements $\sqrt{2}$, which present difficulties in the binary calculations, only occur in the diagonal matrix Λ and can be eliminated by combining two multiplications by Λ into one multiplication by Λ^2 in which $\sqrt{2}$ does not occur.

The fast algorithm

A matrix product such as $P = XH$ is obtained from a series of vector-by-matrix products: each row of X is multiplied by H to give the corresponding row of P. We have developed a fast algorithm to perform the vector-by-matrix multiplication.

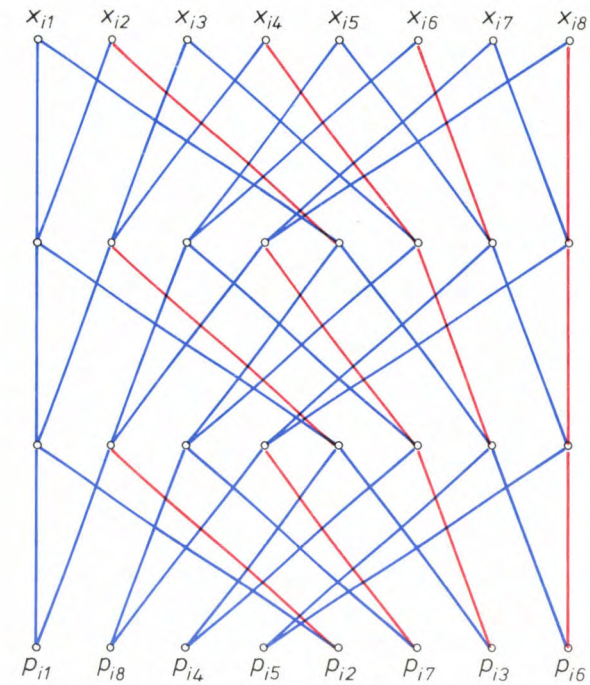


Fig. 10. Fast algorithm for 8×8 Hadamard transformation. Each row $p_{i1} \ldots p_{i8}$ of the matrix product $P = XH$ is obtained in three steps from the corresponding input row $x_{i1} \ldots x_{i8}$ by additions (*blue*) and subtractions (*red*). The graphs for the three steps are identical and the three steps are carried out successively by the same wired logic.

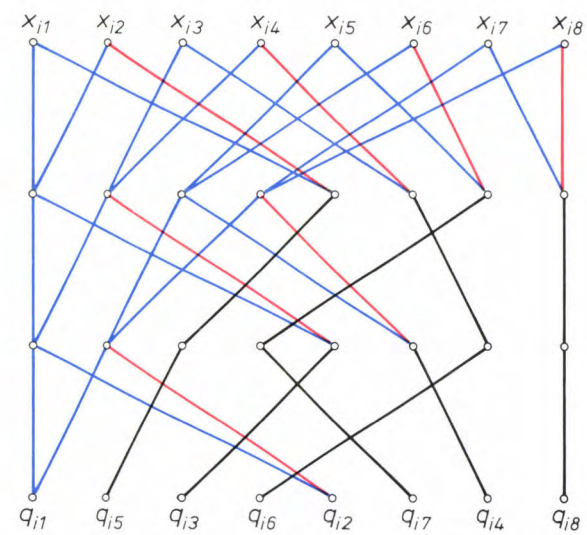
The Hadamard algorithm is shown in *fig. 10*; it performs the multiplication by the matrix H for $N = 8$. The algorithm can be divided into three successive steps, each of which is represented by the same branching system of additions and subtractions. The circuit that performs the elementary step is formed from permanently connected adders and subtractors. This circuit is used three times in succession to perform the complete algorithm. The intermediate outputs are fed back to the input for the next step. For transforming a subpicture consisting of 16×16 picture elements the operator has eight adders and eight subtractors, and four steps are required. The same kind of operator is used for the direct and the inverse transformation.

This is not the case with the Haar algorithm, which is shown in *fig. 11*; this performs the multiplication by the matrix G for $N = 8$. With this algorithm some of the results are obtained before the last step; the arithmetic units that perform the additions and subtractions are then programmed to transfer these results through the final steps unchanged. In the inverse transformation this order is reversed and some of the input data goes through the first steps unchanged.

Apart from these transfer steps which have been included to simplify the wiring and the time relationships, the Haar branching system is very similar to the Hadamard one. The order of inputs q_{i6} and q_{i7} and outputs x_{i6} and x_{i7} has to be reversed in the inverse transformation to obtain maximum similarity.

Rounding-off, word length

The inverse transformation in the receiver is designed to recover the original 8-bit words containing the picture-element data. Because of a transmission error, however, the received data may be so changed that a word with a higher value than 2^8 is produced after the inverse transformation. Normally this would require a ninth bit. If this most-significant bit is dropped because there is only room for eight bits, the result may be a serious error in the reconstructed picture. It is therefore necessary to consider the word length that can be accepted in the various steps of the process.



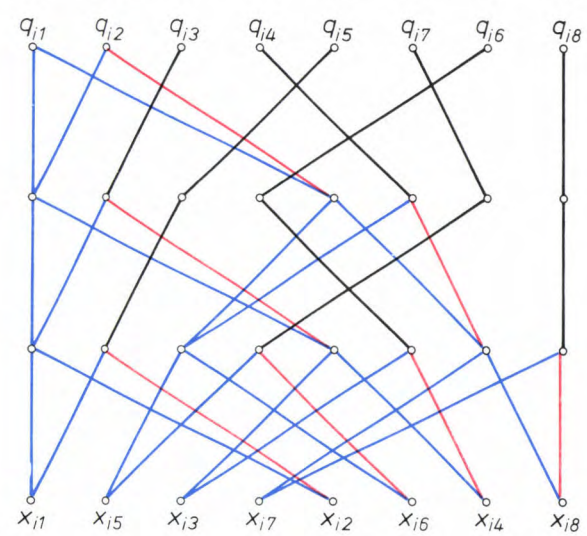


Fig. 11. Fast algorithm for 8×8 Haar transformation; *a*) direct transformation; *b*) inverse transformation. Each row $q_{i1} \ldots q_{i8}$ of the matrix product $Q = XG$ is produced in three steps from the corresponding input row $x_{i1} \ldots x_{i8}$ by additions (*blue*), subtractions (*red*) or by transferring intermediate results unchanged (*black*). The graphs closely resemble those of the Hadamard fast algorithm (fig. 10). To make the similarity as great as possible the inputs q_{i6} and q_{i7} and outputs x_{i6} and x_{i7} have been reversed for the inverse transformation.

Fig. 12 shows how the number of significant bits varies during the computations for a 16×16 transformation. Fig. 12*a* gives the ideal case in which there are no transmission errors. For each matrix multiplication the range of possible numbers can increase by a factor of 16 and the binary words can thus become four bits longer. All these bits are retained in the direct transformation, because the divisions by N are not carried out, to avoid rounding-off errors. The words therefore increase to a length of 16 bits before transmission. The first matrix multiplication in the inverse transformation extends the word length to 20 bits.

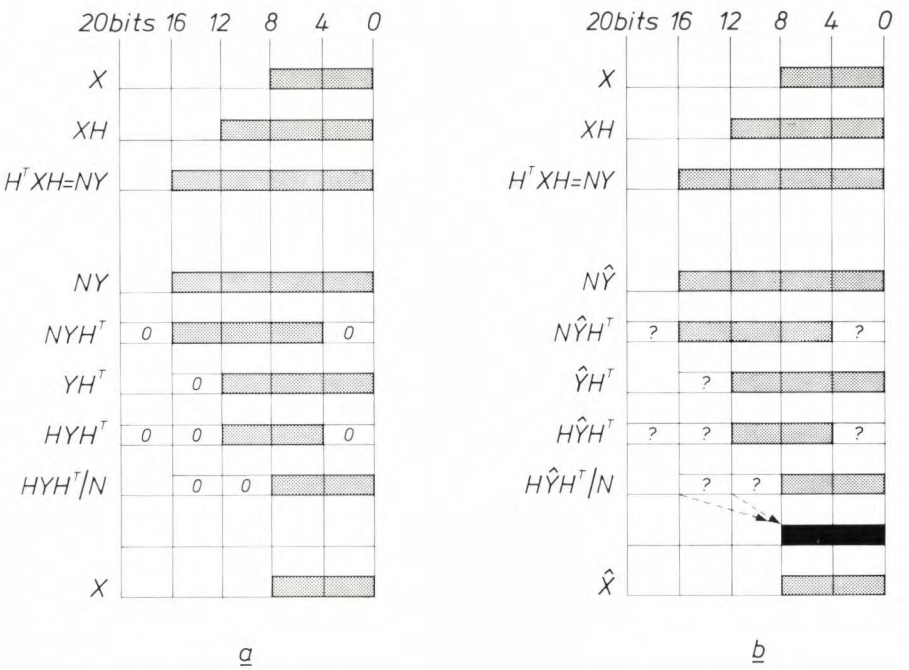


Fig. 12. Number of bits in different stages of the direct and inverse transformations of subpictures consisting of 16×16 picture elements. The Hadamard transformation symbols are used to designate the different steps, but the figure also applies for the Haar transformation. *a*) Ideal case without coding and transmission errors. The extra bits added by the matrix multiplications and retained during transmission to avoid rounding-off errors become zeros in the inverse transformation. Division by a factor $N = 16$ is carried out twice by dropping the four least-significant bits, which are zeros. *b*) Actual situation with possible coding and transmission errors. Because of these errors the bits at the most- and least-significant places may differ from zero in the inverse transformation (see question marks). In the divisions by N the least-significant bits are dropped, but the most-significant bits are subjected to a check. If the check shows that they are not all equal to zero, the word with the highest possible value is chosen for the last eight bits. In this way the error introduced by dropping the eight most-significant bits is minimized.

If the words are not corrupted during transmission, however, the first four bits and the last four bits will be zeros. A division by $N = 16$ can now be performed by merely shifting four places to the right, which eliminates the last four zeros. The same thing happens in the second matrix multiplication, and the remaining bits form the original eight-bit word.

Fig. 12*b* illustrates what happens when transmission errors do occur. The first matrix product of the inverse transformation will not now necessarily only contain zeros at the first four and last four places. The last four bits are nevertheless truncated in the division by N, but the errors introduced will be small. The first four bits are included, however, because they can represent a high value. After the second matrix multiplication the first eight bits are checked in a special circuit. If they are not all equal to zero, the final eight bits, which represent the reconstructed picture element, are replaced by the maximum eight-bit word; in this way the visibility of errors in the reconstructed picture is limited.

Processing time and implementation

The requirement for real-time processing, i.e. that all operations performed on a TV picture must be completed before a new picture is presented, determines the amount of time available for each process step. The logic speeds required are no higher than those that can be obtained from commercially available integrated circuits. Our logic circuits employ transistor-transistor logic (TTL), and the memories are made in

MOSFET technology (MOSFET = metal-oxide-semiconductor field-effect transistor). The equipments for both the direct and the inverse transformations each contain about 2000 integrated circuits.

Line memories

Each of the memories M_1 and M_2 (fig. 8) can store 16 colour-TV lines. In view of the processing of blocks of 16×16 picture elements, each TV line is divided into groups of 16 picture elements. Each group is represented by a single binary word of $16 \times 8 = 128$ bits. In addition to the luminance signal Y [10] two colour-difference signals D_r (for red) and D_b (for blue) must also be stored. As these are sampled at a rate four times lower than the luminance-sampling rate, each line is composed of sequences YYD_rYYD_b, where each character stands for one 128-bit word. In one line there are 39 Y words, 10 D_r words and 10 D_b words, which makes 59 words; in 16 lines there are thus 944 words, each lasting 880 ns. These are stored in a random-access memory (RAM) capable of containing 1024 bits and with a cycle time of 750 ns; 128 of these RAMs are required for each of the two memories M_1 and M_2.

Arithmetic unit with wired logic

There are 16 line periods available, or 16×64 μs = 1024 μs, for computing the transformation of 59 blocks of 16×16 picture elements. This gives a maximum computing time of 17.35 μs for a single block. This includes 8 μs for each of the two matrix multiplications; the rest of the time is required for signal handling. Each matrix multiplication consists of 16 vector-by-matrix products, for each of which 8 μs/16 = 500 ns is available. Each vector-by-matrix product is computed by the fast algorithm in four steps. Every step consists of eight additions and eight subtractions, which are performed simultaneously. The four steps are carried out in succession by the same circuit with wired logic, and the intermediate results are fed back to the inputs via a buffer memory. For every step 125 ns is available. The additions and subtractions, and the transfers for the Haar transformation, are all carried out by standard arithmetic-logic units.

Intermediate matrix-transposing memory

An intermediate memory is necessary for storing the first matrix product P of each transformation and providing it in transposed form P^T for the second matrix multiplication. This memory is composed of an array of TTL RAMs, each of which can contain 16 words of 4 bits. The matrix P is stored in the memory row by row, and is read out column by column. The rows of P are provided by the fast algorithm at intervals of 500 ns. The elements appearing simultaneously, 16 in all, are serialized before entering the memory. A time of 500 ns/16 = 31 ns is available for each element. This corresponds to a word rate of 32 MHz, which is the highest word rate encountered in the transformation circuits. A clock signal of 32 MHz is derived from the sampling rate for the luminance signal (12 MHz), and all the other clock signals are derived from these two frequencies.

Built-in test circuits

To allow easy checking and maintenance, the units for both the direct and inverse transformations are provided with built-in test circuits. These generate special input signals, check the corresponding output signals and can detect single-bit errors. Indicator lamps facilitate the tracing of faulty circuits.

The test circuits can, for example, supply all the basis pictures B_{kl} (for any block size and shape and for both the Hadamard and the Haar transformation) to the input of the direct-transformation unit. After transformation only the coefficient of index kl has a non-zero value. When this output signal is applied to the inverse transformation unit, the corresponding basis picture is recovered, which can then be automatically tested for correctness.

The equipment described here was primarily intended for laboratory investigations. It was therefore designed to be as versatile as possible, with several additional options and a free choice of various parameters.

If the latest digital integrated circuits were used, possibly including charge-coupled devices, the size and power consumption of the equipment could be considerably reduced; only about a tenth of the number of integrated circuits would then be necessary.

Summary. A system is described that performs direct and inverse Hadamard and Haar transformations of colour-television pictures in real time. The equipment is now being used at the CCETT research centre at Rennes, France, to investigate the extent to which the bit rate in the digital transmission of TV signals can be reduced. The TV pictures are subdivided into a large number of subpictures, and each is expressed as a series of basis pictures, multiplied by appropriate weighting factors. It is the weighting factors that are transmitted; many of these have small values and can be set equal to zero without significant loss of picture quality. This allows more efficient signal coding to be used.

The equipment consists of two units, one for the direct transformation and the other for the inverse transformation. The additions and subtractions required in the Hadamard and Haar transformations are carried out in parallel by a fast algorithm. The equipment was built in 1975, using the standard TTL logic circuits and MOSFET memories then available. Each unit contains 2000 integrated circuits. With circuits now available only about a tenth of this number would be required.

[10] We use the standard symbol Y for the luminance signal, although the same symbol has been used earlier for the matrix in the transformed plane.

Measurement of ozone in an aircraft

S. van Heusden and L. G. J. Mans

In the last ten or twenty years civil aviation has largely switched over to the use of jet aircraft. The greater part of a flight in a jet aircraft is made at high altitudes above the Earth's surface. To cut down on fuel consumption it is useful to fly jet aircraft higher than was customary with propeller aircraft, and the altitude at which long-distance flights are made is still tending to increase. On intercontinental flights, for example, an altitude of 10 kilometres is nowadays very common.

An altitude of 10 kilometres corresponds to the height of the tropopause, the zone of transition between the lowest layer of our atmosphere — the troposphere — and the stratosphere. Most weather phenomena occur in the troposphere, and powerful air turbulance can develop; the stratosphere is much more stable. Another striking difference between troposphere and stratosphere is a pronounced vertical temperature gradient, i.e. the temperature in the troposphere decreases with increasing height (to about −60 °C), while above the tropopause it rises again; see *fig. 1a*. One of the reasons for this increase is the presence of ozone (O_3) in the stratosphere, which absorbs part of the short-wave ultraviolet radiation from the Sun.

Fig. 1*b* shows the variation of ozone concentration with altitude. At 50 km the air pressure has dropped to a mere 1 millibar; see fig. 1*c*. At this low pressure there is very little chance of collisions between the particles that take part in the ozone-forming reactions (in which very-short-wave UV radiation plays an important role). As the altitude decreases and the air pressure rises, the probability of collisions becomes greater, and so therefore does the ozone concentration, until at an altitude of 30 km a concentration of about 10 000 ppb is reached (parts per billion, i.e. $1:10^9$). Below this height the intensity of the very-short-wave UV radiation decreases and so does the ozone formation. At the same time there are breakdown reactions in the troposphere and the ozone is also diluted by vertical air currents.

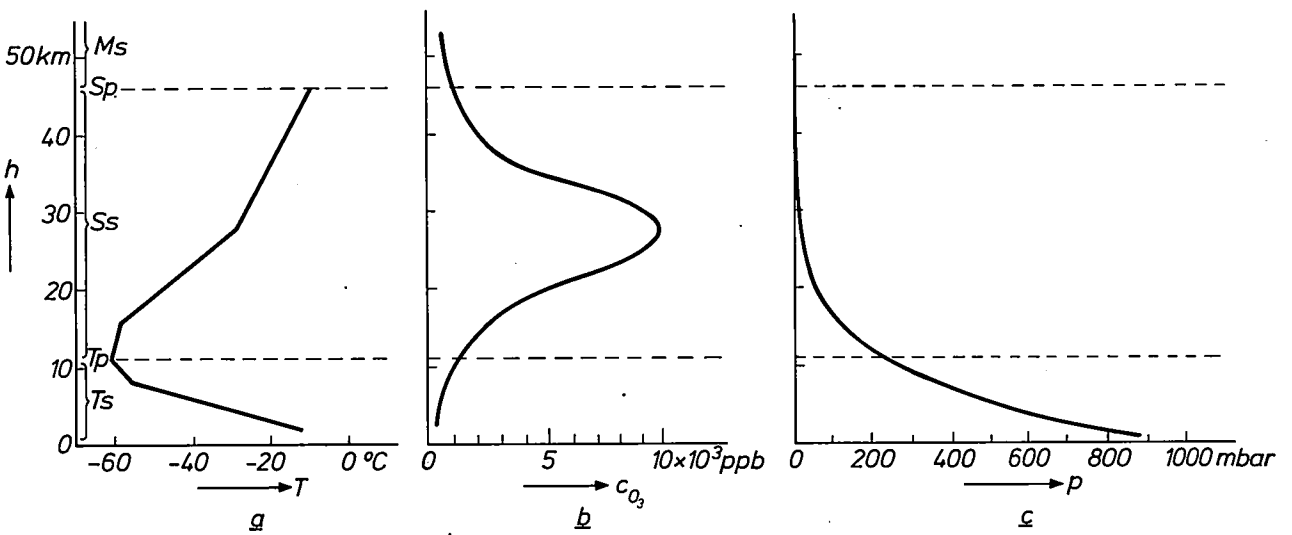


Fig. 1. Temperature of the atmosphere (*a*), ozone concentration (*b*) and air pressure (*c*) as a function of the altitude *h* above the Earth's surface (schematic). *Ts* troposphere. *Tp* tropopause. *Ss* stratosphere. *Sp* stratopause. *Ms* mesosphere.

This does not complete the picture of the Earth's atmosphere. The height of the tropopause — and hence of the region where the ozone concentration begins to rise steeply — is not the same everywhere and at all times, but changes with the seasons and also with the distance from the equator. As a rough approximation, the height of the tropopause is low in the spring and autumn, and decreases with the latitude. Above

Ing. S. van Heusden is with Philips Research Laboratories, Eindhoven; Dr L. G. J. Mans is a physician with the Medical Service of Eindhoven University of Technology.

the poles the tropopause may be as low as 8 km, whereas above the equator it may lie at a height of 16 km. Also to be considered is the still not completely explained phenomenon of the 'ozone finger': in some geographical regions the ozone concentration is much greater than elsewhere at the same altitude and the same latitude.

Although the thick ozone blanket is indispensible to life on Earth — without it all higher forms of life would be exposed to a lethal dose of short-wave UV radiation — this is not to say that living organisms thrive in an environment that contains ozone. Experiments on animals subjected to chronic exposure indicate a striking increase in mortality. Human subjects become aware of the prickling odour of ozone at concentrations as low as 10 to 50 ppb. A concentration slightly higher than 50 ppb may cause a (relatively innocuous) irritation of nose and throat, but a concentration of 2000 ppb can reduce a person's capacity to think and concentrate. As a rule the ozone concentration outside the aircraft at normal flight altitudes is of the order of 1000 ppb.

The ventilation systems of the jet aircraft used in civil aviation today are designed so that the air supply is taken from one of the compressor stages of the jet engine, where the pressure is a few bars and — even more important — the temperature is a few hundreds of degrees Celsius. This method of air intake has the effect of reducing slightly the ozone content in the air, because at higher temperatures ozone molecules are easily broken down into 'normal' oxygen molecules.

Is this reduction sufficient to decrease the ozone content *inside* the aircraft to a medically acceptable level? This question was put to the Medical Department of KLM and the Medical Service of Eindhoven University of Technology. Finally, an approach was made to Philips Research Laboratories, where experience had been gained over a substantial period of time with ozone measurements in automatic air-monitoring systems [1]. It was decided to make a combined study of the problem.

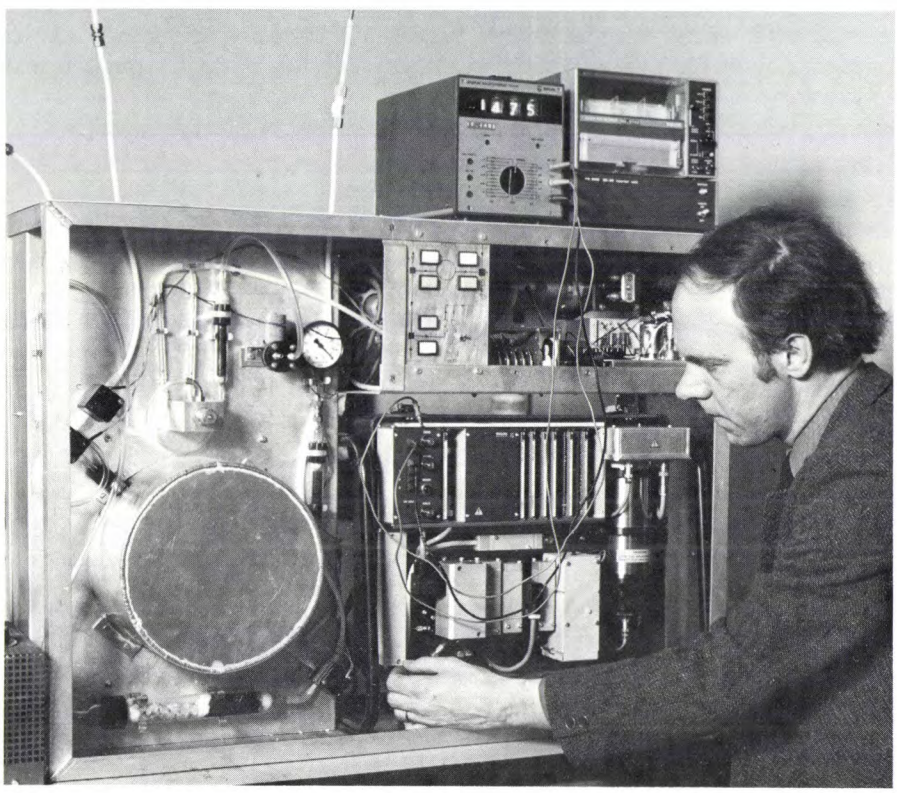


Fig. 2. The ozone-measuring equipment. The Philips PW 9771 ozone meter can be seen at lower right. The power supply and control section are above it. The sampling section is on the left. A pen recorder and a digital voltmeter can be seen at the top of the picture.

So that the measurements could be carried out under practical conditions, it was decided to make them on a scheduled flight from Europe to North America on a DC 10-30 aircraft operated by KLM. This meant that much of the information required for ozone measurements was available right from the start. The outside pressure during such a flight is about 0.2 bar and the outside temperature is about -50 °C; the pressure inside the cabin is 0.8 bar and the temperature 25 °C

(or higher). Other factors taken into account were a low humidity, both inside and outside, and a flying speed of 900 km per hour. The ozone concentrations measured were expected to range from 0 to 1000 ppb.

The methods of measurement available include those based on colorimetry, coulometry, UV absorption spectroscopy and chemiluminescence. We chose to use chemiluminescence, mainly because of its high specificity for ozone. Methods were considered in which ozone reacts with ethylene gas or with the dye rhodamine B, a solid [2]. For safety reasons we preferred to use rhodamine B.

To keep the measuring equipment compact and to ensure good correlation between the inside and outside measurements, we arranged to make both measurements alternately with the same equipment. The heart of the system was a Philips PW 9771 ozone meter, which is used in automatic air-monitoring systems. This meter met our requirements for temperature, humidity, pressure and measuring range. Another advantage in its favour was that it does not require constant attention and, since it is self-calibrating and resets itself to zero, it is capable of continuous measurements for a long period without human intervention. The equipment is shown in *fig. 2*, together with the peripheral equipment. This includes a sampling and processing section, which allowed us to make measurements both inside and outside the aircraft.

The peripheral equipment was designed to enable the samples of the outside air to be brought up to cabin pressure with de-ozonized air. This meant, of course, that allowance had to be made for dilution of the ozone. Another aspect of the preparation was that the equipment had to be modified to operate from the aircraft voltage supply (115 V, 400 Hz). In addition a digital voltmeter and pen recorder enabled us to follow the results of the measurements directly.

In the DC 10-30 the ventilation air is tapped from the eighth compressor stage of the jet engine, where the temperature is 300 °C. The ventilation air is fed to the cabin through the ceiling. The internal samples were taken in the first-class cabin, immediately under the ceiling to avoid disturbing effects from the passengers (walking about, smoking). The measuring equipment was also set up in the first-class section (*fig. 3*). The outside samples were taken near the nose of the aircraft through a flexible tube projecting through the shell in the direction of flight.

Our flight was made on 24th and 25th March 1977 from Amsterdam to Toronto and back. The measurements were carried out in cycles of 13 minutes each. During the first nine minutes of each cycle, seven samples were taken in the cabin and measured. At the same time the outside sampling circuit was flushed through and then filled with an outside sample. This sample was measured during the remaining four minutes of the cycle (the 8th measurement point). Each of the eight measurements was preceded by an automatic zero determination, made with ozone-free air, and by automatic calibration against a known ozone concentration. Since the samples for the 7th and 8th measurement points were taken at virtually the same time, the ozone contents inside and outside the aircraft could readily be compared [3].

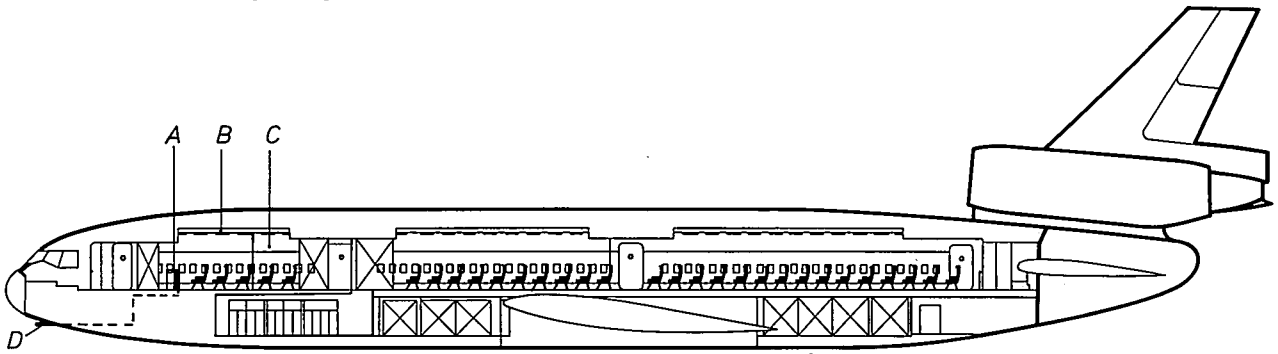


Fig. 3. Longitudinal section of the DC 10-30 jet aircraft in which the measurements were made. *A* location of the measuring equipment. *B* sampling location in the first-class cabin, *C* sampling location in the tourist-class cabin. *D* the place where the samples of outside air were taken.

Fig. 4 shows the results of the measurements during the return flight. Four facts clearly emerge.

1. There is a very definite correlation between the ozone concentrations inside and outside the aircraft.

2. Some 60 or 70% of the ozone present in the atmosphere is admitted to the cabins for passengers and crew.

3. The ozone content depends closely on the altitude relative to the tropopause.

4. The TLV (Threshold Limit Value) [*] of 80-100 ppb laid down in all IATA countries is exceeded during most of the flight.

[1] S. van Heusden, Philips tech. Rev. **34**, 73, 1974.

[2] S. van Heusden and D. J. Kroon, 11e Colloque sur les atmosphères polluées, Paris 1974, p. 32.1.

[3] The measurements are reported in more detail in S. van Heusden and L. G. J. Mans, Avia. Space Environ. Med. **49**, 1056, 1978.

Further measurements in the cabin showed that the ozone concentration at head height — taken as 1.20 m above the floor — was 25 % lower than at ceiling height in the first-class cabin. In the tourist-class cabin the corresponding figure was 40 %. This effect may be due to the presence of passengers (it may also reflect the difference in the numbers occupying seats in each class), smoking, the serving of meals, etc.

No general conclusions can of course be drawn from the results of measurements during a single return flight in one type of aircraft, but it does seem desirable to make a closer study of the ozone content in jet aircraft generally. On the other hand it should be borne in mind that there is little medical knowledge about the consequences of *chronic* human exposure to the ozone concentrations considered here, which may mean that the present TLV figure might be modified in the future.

If in the long run it should prove desirable to introduce measures to reduce the ozone concentration in aircraft, it might be necessary to develop filters with a much higher capacity per unit volume than those now available. The possibility of improvement by leaving the ventilation air for longer or at a higher temperature in a compressor stage could also be studied. In general this will lead to higher fuel consumption. A temporary solution that might be considered is to fly at altitudes at which the ozone-measuring instruments — which every aircraft should then have on board — give a reading that does not exceed a specified ozone concentration. This would of course also make the fuel consumption higher.

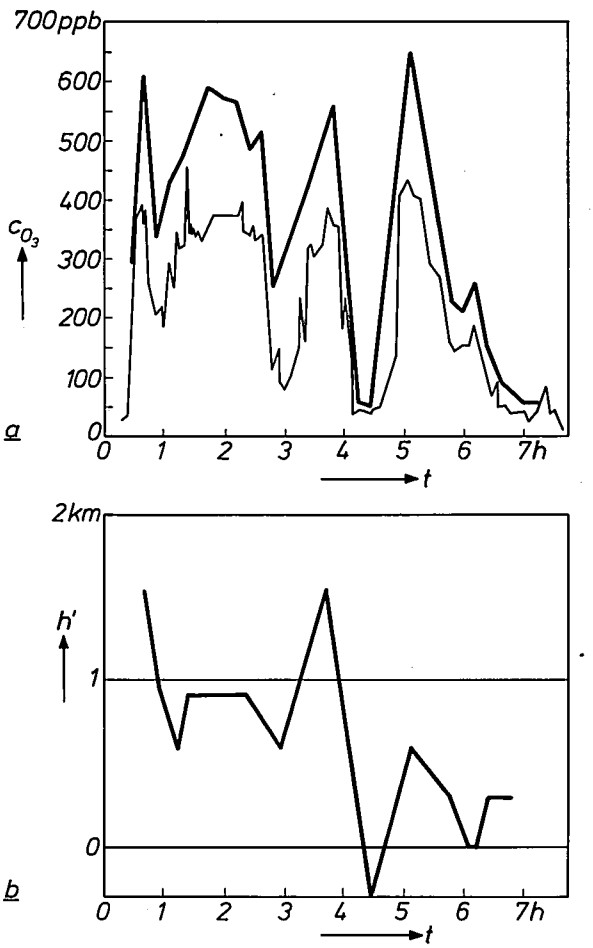


Fig. 4. *a*) Recordings of the ozone concentration during the Toronto-Amsterdam flight. *Heavy line*: concentration outside the aircraft. *Thin line:* the concentration in the first-class cabin, measured at ceiling height. *b*) Altitude *h'* relative to the tropopause. (The altitude relative to the Earth's surface was 9.8 km; the data on the height of the tropopause are taken from the flight plan.)

[*] This should really be TLV-TWA (TWA for Time-Weighted Average). This standard is based on a 40-hour working week and should not be exceeded, unless there is compensation with equivalent concentrations *below* this standard.

Summary. In a joint study KLM, Eindhoven University of Technology and Philips Research Laboratories, Eindhoven, have measured the ozone content in a DC 10-30 jet aircraft during a flight from Amsterdam to Toronto and back. The measurements showed that there is a strong correlation between the ozone content inside and outside the aircraft, and that 60 to 70 % of the ozone contained in the atmosphere is admitted to the passenger and crew compartments. It was also found that the ozone content depends to a great extent on the altitude relative to the tropopause. The measurements reveal that the TLV (Threshold Limit Value) now applicable in the IATA countries was exceeded during most of the flight.

The thermal behaviour of electronic components during soldering

R. J. Klein Wassink

The thermal behaviour of small components like resistors and capacitors during soft soldering is unlikely to reveal many new aspects of a truly scientific nature. Even so, as can be seen from the article below, a theoretical treatment of the subject can be of considerable practical value. The large-scale automation of soldering in modern mass production is of course only efficient if it is reliable and if the heating applied does not noticeably impair the performance of the components. A rough estimate shows that the number of components soft-soldered per year in the entire industrial world is in the region of 10^{12}.

Soldering and mass production

Most soldered joints were formerly made by hand, including those used in the electronics industry. Owing to the enormous growth of production needs, this manual work has now almost entirely been superseded by highly mechanized and automated methods. Nowadays hundreds of components like resistors and capacitors can be soldered simultaneously by a single automatic operation, for example to a circuit wired on a printed board. Recently a great deal of attention has also been paid to soldering in hybrid circuits made by the thick-film technology [1]. This article deals only with soldering on printed boards.

Among the principal methods used dip soldering should be mentioned first. In this method the terminal leads projecting from the components are simultaneously immersed in a stationary solder bath in such a way that the underside of the board just touches the solder. The second method is called wave soldering; here the solder is pumped up like a spout from a fountain, bringing it into contact with the leads on the underside of the printed board (*fig. 1*). Both methods are used for soft soldering, for which the temperature at the point of contact must remain above about 200 °C for a few seconds. To speed up the operation — and for other reasons as well — the parts to be soldered are sometimes preheated. This point will be discussed again at the end of the article.

The soldering method used should be completely reliable, i.e. all the soldered joints required must be good ones. The desired reliability can be achieved by only using components whose characteristics correspond to the conditions and data that are typical of the soldering method. Important in this respect are the choice of materials, the solderability of the various metals, the choice of solder and flux, the dimensions of the components and their leads, and also the dimensions of the printed boards. There are two main types of printed board: single-sided — with conducting tracks on the underside — and double-sided — with tracks on both sides. In the double-sided type the

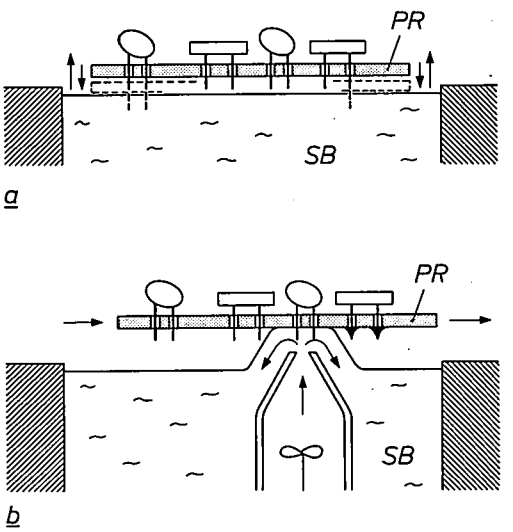


Fig. 1. The principle of dip soldering (*a*) and wave soldering (*b*). *PR* printed board with electronic components, such as capacitors and resistors. Each component has two wire leads, which have to be soldered to conducting tracks on the underside of the board. *SB* solder bath. In dip soldering all the leads are simultaneously immersed in the solder for a few seconds by a vertical movement of *PR*. In wave soldering the leads are soldered consecutively. The solder is forced upwards by a pump. The printed board travels horizontally.

Ir R. J. Klein Wassink, with Philips Philite and Metalware Works, is seconded to the Philips Centre for Technology (CFT), Eindhoven.

[1] W. Funk, Philips tech. Rev. 35, 144, 1975.

soldering takes place in metallized holes that run through the full thickness of the board. This is often the preferred method for professional equipment.

This article deals with the thermal aspect of the reliability of the soldering process. This involves two requirements, which at first sight seem diametrically opposed. The first is that the leads of the components should be raised to a temperature high enough for them to be wetted by the solder. The second requirement is that the temperature to which the capacitor or resistor rises during soldering should *not* be so high as to affect its operating value. Each component has its typical soldering range, i.e. a range of values for the solder-bath temperature and for the soldering time (immersion time) within which both of these requirements can be satisfied. There must be a good overlap between the soldering ranges of components of different types. The simplest means of ensuring this is to adapt the length of the leads.

In the next section an electrical network will be discussed that can give a fairly good simulation of the thermal behaviour of a component during soldering. The method of calculation based on this network allows the required solder-bath temperatures and soldering times to be found easily from a graph. Conversely, the method can also be used to good advantage for designing new components that have to be suitable for a particular soldering method. The general validity of the model will then be examined. Finally, the model is applied to two cases and a number of practical matters are discussed, such as the use of copper wire or iron wire for the leads and the question of preheating.

The model and the temperature-time diagram

The electrical network that serves as the simplest analogue for calculating the thermal behaviour of a component is shown in *fig. 2* [2]. The component consists of a body and a lead, which is inserted through an opening in a single-sided printed board. (The thermal behaviour in the case where the lead is inserted through a metallized hole in a double-sided board requires a more complicated electrical analogue, which will be referred to briefly in the section dealing with the validity of the model.) Dipping in the solder bath corresponds to closing the battery switch in the model. This has the effect of applying a step-function voltage to the network. The resultant increase in the potential at the various points in the network corresponds to the rise in temperature at the corresponding points in the component. The value of the capacitance in the network corresponds to the heat capacity of the body. For simplicity it is assumed that no temperature gradients occur in the body of the component, so that its temperature is everywhere the same. To calculate the behaviour of the potential as a function of time it is sufficient in the case of fig. 2 to solve a simple first-order differential equation. The model used here is therefore generally classified as a 'first-order model'. Solving the differential equation gives $T_w(t)$, the temperature of the lead at the height of the underside of the printed board (after dipping), and also $T_b(t)$, the temperature of the body, both as a function of time. These two temperatures must be known for an assessment of the success of the soldering process. The moment at which dipping starts is taken as the time $t = 0$. The solution found for the temperature $T_b(t)$ is:

$$T_b(t) = T_{pre} + \{1 - \exp(-t/RC)\} \times \times \left[\frac{T_{sol} - T_{amb}}{1 + (R_s + R_w)/R_b} - (T_{pre} - T_{amb})\right]. \quad (1)$$

The resistance R is given by:

$$1/R = 1/R_b + 1/(R_s + R_w),$$

which is the expression for a parallel arrangement of the thermal resistances R_b and $R_s + R_w$. The significance of the various temperature symbols in eq. (1) can be found in fig. 2.

The solution for the temperature $T_w(t)$ is

$$T_w(t) = T_{sol} - \frac{R_s}{R_s + R_w}\{T_{sol} - T_b(t)\}. \quad (2)$$

It thus appears that the temperature of the lead, which is between the (high) solder-bath temperature and the body temperature of the component, follows from a

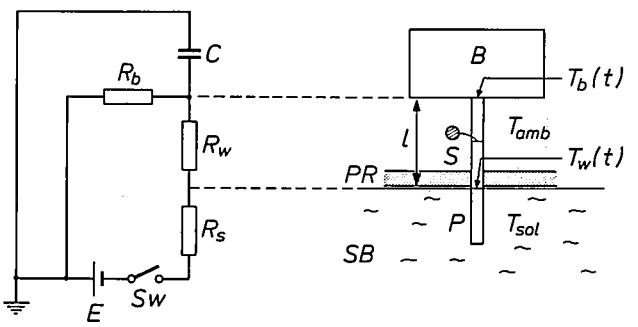


Fig. 2. An electrical network (*left*) for simulating the thermal behaviour of an electronic component (*right*), consisting of a body (*B*) with a lead (*P*) during soldering on a single-sided printed board (*PR*). *SB* solder bath. R_s interface resistance between solder bath and lead. R_w thermal resistance of the lead. R_b interface resistance between body (*B*) and air. *C* heat capacity of *B*. *l* soldering distance ('free' length) of the lead. *S* cross-sectional area of the lead. $T_w(t)$ temperature of the lead as a function of time. $T_b(t)$ body temperature as a function of time. T_{sol} solder-bath temperature. T_{amb} ambient temperature. Closing the switch *Sw* corresponds to the moment of immersion in the solder bath at $t = 0$. If there is preheating, the body temperature (and also the lead temperature) is equal to T_{pre} at $t = 0$.

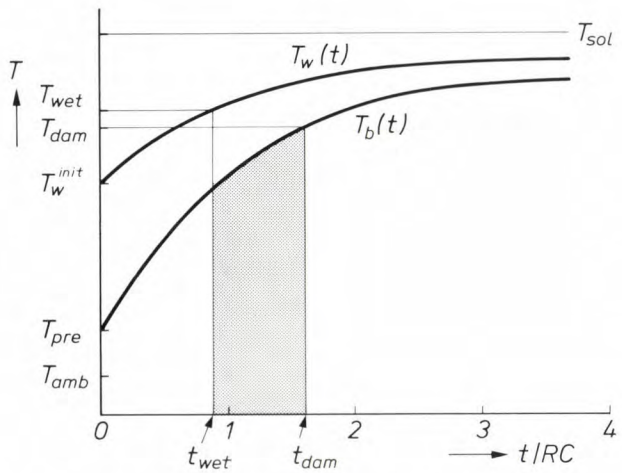


Fig. 3. A calculated temperature-time diagram for the soldering of an electronic component (fig. 2). The numerical values used are given in Table I. The temperatures $T_b(t)$ of the body and $T_w(t)$ of the lead are both plotted as a function of the time that has elapsed since the moment at which the lead is first dipped into the solder bath. T_{pre} temperature of body and lead at the time $t = 0$, as a result of preheating. T_{amb} temperature of ambient air. T_{sol} temperature of solder bath. T_{wet} wetting temperature of the lead. T_{dam} temperature at which the component is damaged. In the model the heat capacity of the lead is neglected so that at the same time as immersion begins its temperature becomes T_w^{init}. RC is the characteristic time constant of the component (see eq. (1)). The limit values t_{wet} and t_{dam} for the soldering time follow from the values of T_{wet} and T_{dam} applicable to the component.

simple 'potential division' in fig. 2. This division involves only the thermal resistances R_s and R_w, and is independent of time.

The temperatures calculated from equations (1) and (2) are plotted in *fig. 3* for a component whose characteristic data are listed in *Table I*. The time in fig. 3 is expressed in units of RC, the characteristic time constant of the component.

When the damage temperature T_{dam} of a component and the wetting temperature T_{wet} of its lead are also known the corresponding time durations t_{dam} and t_{wet} can be derived from fig. 3. The soldering time must be chosen between the limits represented by these quantities.

Fig. 4 shows a temperature-time diagram derived from fig. 3 by giving values to the coordinates such that the curves for $T_b(t)$ and $T_w(t)$ are straight lines. Since a dimensionless quantity (T^*) is plotted instead of the temperature, the diagram also gives a conversion line from which the temperature T corresponding to T^* can be read, where T^* is equal to $(T - T_{pre})/(T_{sol} - T_{pre})$.

The linearized diagram only applies of course to components whose characteristic data (temperatures and thermal resistances) correspond to those given in Table I. Starting with the damage temperature T_{dam} of 145 °C and the wetting temperature T_{wet} of 183 °C, it is first necessary to find the corresponding dimensionless temperatures on the vertical scale. With these and the two temperature lines the limits τ_{dam} and τ_{wet} for the dipping duration are then found on the dimensionless time axis. Each value between these limits, after multiplication by the time constant RC, gives an appropriate soldering time for the component.

The conversion of fig. 3 into a graph containing only straight lines is a simple matter. From eq. (1) it can be directly deduced that $\{T_b(t) - T_{pre}\}$ is proportional to $\{1 - \exp(-t/RC)\}$. The curves in fig. 3 can therefore be transformed into straight lines by taking the scale distances along the t/RC axis from $\{1 - \exp(-t/RC)\}$.

In the case, frequently found in practice, where $R_b \gg R_s + R_w$, the proportionality factor is equal to $(T_{sol} - T_{pre})$, so that T^* then goes from 0 to 1 when t goes from 0 to ∞. If $(R_s + R_w)$ is not negligibly small compared with R_b, then T^* at $\tau = \infty$ is somewhat less than 1. A similar approach applies for $T_w(t)$. The values of T_b^{max} and T_w^{init} can be calculated from the equations given in Table I.

The main thermal quantities in the model

Heat capacity and thermal resistances

The heat capacity C of a component is equal to $V\varrho c$, where V is the volume of the body, ϱ is its density and c its specific heat capacity. The product ϱc and the thermal-conductivity coefficient λ for various materials are given in *Table II*. It can be seen that the product ϱc has about the same value for various materials, but the values of λ differ considerably.

The thermal resistance R_w of the lead can be calculated from the relation for the conduction resistance, which states that R_w is equal to $l/\lambda S$ where l is the soldering distance (or 'free' length) of the lead and S the area of its cross-section (fig. 2). The conduction resist-

Table I. The values of the thermal quantities applicable to the component used by way of example as the basis of fig. 3 and fig. 4. T_b^{max} maximum temperature that the body of the component would have after an infinitely long time. T_w^{init} temperature of the lead immediately after immersion in the solder. The values of these temperatures were calculated from the equations given in the Table, which were derived from equations (1) and (2). The other quantities are defined in the caption to fig. 2.

Temperature (°C)		Thermal resistance (°C/W)	
		R_w	15
T_{pre}	50		
T_{sol}	250	R_s	15
T_{amb}	20	R_b	315
T_b^{max}	230 ←	$\frac{T_b^{max}}{T_{sol}} = \frac{1 + (1 + R_s/R_w)R_wT_{amb}/R_bT_{sol}}{1 + (1 + R_s/R_w)R_w/R_b}$	
T_w^{init}	150 ←	$\frac{T_w^{init}}{T_{sol}} = \frac{1 + R_sT_{pre}/R_wT_{sol}}{1 + R_s/R_w}$	
T_{dam}	145		
T_{wet}	183	The eutectic temperature of tin-lead	

[2] From H. J. Verbeek, DVS-Ber. **40**, 35, 1976.

ance of copper, aluminium and iron leads is calculated per millimetre of length, and with the widely used diameter of 0.6 mm it is equal to 10, 16 and 50 °C/W respectively.

The thermal resistances R_s and R_b are interface resistances, R_s between the solder and the immersed part of the lead and R_b between the body and the ambient air. The values are calculated from the general formula $1/hA$, where h is the appropriate heat-transfer coefficient and A the interface area. There are uncertainties about the thermal resistance R_s, and these will be mentioned in the section on the validity of the model.

Table II. Product of the density ϱ and the specific heat capacity c, with the thermal-conductivity coefficient λ of some metals, alloys and materials for printed boards.

Material	ϱc (J/cm³ °C)	λ (W/m °C)
Iron	3.7	75
Copper	3.5	370
Aluminium	2.4	230
Fernico	3.7	16
Brass	3.3	100
Plastics	about 2	about 0.2
Paperboard	2	0.2
Glass	2	1
Solder (eutectic tin-lead)	1.7	50

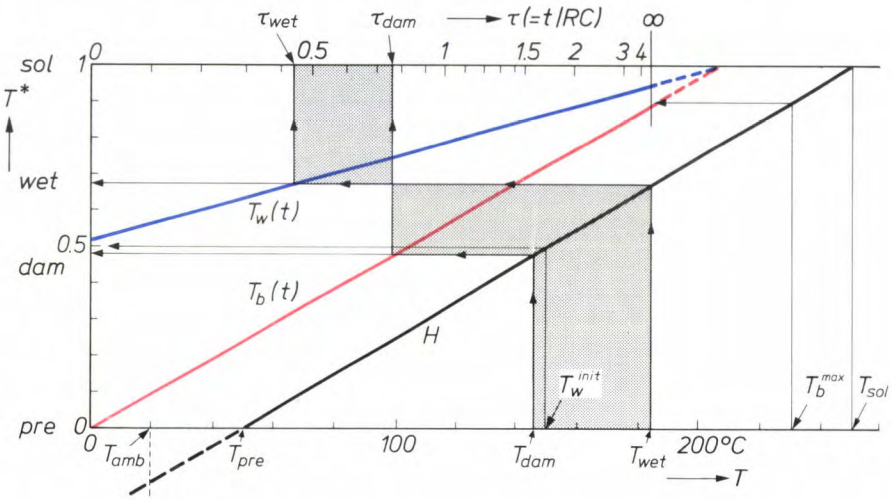


Fig. 4. The temperature-time diagram from fig. 3 with the coordinate axes transformed to make the curves straight lines. H conversion line for transforming the temperature T(°C) into a dimensionless quantity T^*. The dimensionless time axis is at the top; the unit of time is the RC constant of the component. The temperature curves for the body and for the lead are the line $T_b(t)$ and the line $T_w(t)$; these lines are found from the calculated T_b^{max} and T_w^{init}. Taking $T_{dam} = 145$ °C and $T_{wet} = 183$ °C as given data for the component and the solder, the time limits τ_{dam} and τ_{wet} for a properly soldered joint can be read from the graph.

With an interface area of 10 mm² R_s can be assumed to have the order of magnitude of 10 °C/W, with h taken as 10 kW/m²°C.

The value given for h is reasonably accurate for wave soldering. For dip soldering a lower value must be used, about 6 kW/m²°C; the thermal resistance R_s then has a value between 15 and 20°C/W. This means that in dip soldering the quantity R_s/R_w is rather higher, and this is reflected in the longer wetting time required. The increase in R_s affects the temperature of the body mainly through the increase in the time constant RC. If the free length of the leads is not too small, however, the body temperature after wave soldering is found to be only a few per cent higher than after dip soldering (during the same time). Measurements and calculations agree completely on this point.

The interface area for the body is about 100 mm². The heat-transfer coefficient h for the heat transfer from the body to the air is about 10 W/m² °C. The thermal resistance R_b thus amounts to about 10^3 °C/W, a hundred times higher than R_s, so that it is often reasonable to allow for an infinite thermal resistance between the body and the air.

The damage temperature

In the discussion of the model several references have already been made to the damage temperature T_{dam}. In fact T_{dam} is not a fixed quantity, nor is it fixed for each type of component. The damage temperature can best be regarded as the temperature at which certain characteristics of the component change by more than a permissible percentage. The application determines the magnitude of this percentage. The percentage change that actually occurs depends not only on the temperature but also on the time during which the temperature prevails. We shall not go further into this here; little information on the matter is available, for one reason because damage studies are usually limited to a test at one standard combination of time and temperature, e.g. 10 seconds and 260 °C.

The validity of the model

The model described contains a number of simplifications that require some explanation. It is assumed, for instance, that the heat capacity of the lead is negligibly small. This is of course admissible when the heat capacity of the lead is very much lower than that of the body. In general this is in fact the case, although there are exceptions, such as small ceramic capacitors. Because the heat capacity of the lead differs from zero, the initial temperature T_w^{init} of the lead is not produced instantaneously on dipping. There is a delay of the order of 0.1 s. This means that the body temperature starts to rise rather less steeply than indicated by eq. (1). A counteracting effect here is that of the heat capacity of the solder itself round the end of the lead. The presence of the printed board does not significantly affect the thermal behaviour, although in theory the board should increase the thermal inertia. When single-sided boards are dipped the underside with the conducting tracks comes into contact with the surface of the solder bath immediately. Calculations show that the temperature at the underside, which originally had a value of say 20 °C, can then exceed the wetting temperature in a fraction of a second; see *fig. 5*. This time is so short that there is no significant effect on the thermal behaviour, and therefore the model is entirely acceptable in this respect.

For double-sided boards, where the solder penetrates over the whole thickness of the board, the model is still valid, though there are some reservations. In this situation additional heat is transferred by the metallized wall of the hole in the printed board, and the rising solder also brings its own heat with it. The electrical network that could be used as an anologue contains a larger number of resistors and capacitors than the one in fig. 2. The behaviour of the extended network has been calculated from computer programs, but the results for the thermal reliability were almost the same. It is therefore preferable to make the calculations with the original model. Since the upper side of the printed board now has to be wetted as well, the position for calculating the temperature of the lead must be taken at the upper side, and not the underside. This means that the 'potential division' must be modified when eq. (2) is used: the heat resistance R_w in eq. (2) has to be reduced by the correction term $\Delta/\lambda S$, where Δ is the thickness of the board. The dashed curve in fig. 5 represents the calculated time needed to raise the upper side of a board to the wetting temperature, if the presence of the holes is not taken into account. In fact the wetting temperature will be reached much more quickly, because the metallized wall of the holes also conducts heat. The body also heats up more quickly than the model suggests.

The final factor limiting the validity of the model is the variation in the thermal resistance R_s. In the expression $1/hA$ for this resistance neither A nor h are constant. The coefficient h varies because of the marked variation of temperature at the lead/solder interface; in addition the heat transfer is affected by the layers of oxide and flux initially present, and also by the solvent that evaporates from the flux during soldering. The surface area A is not constant owing to the 'climbing' of the solder level; immediately after dipping the meniscus of the solder is curved downwards; at the end, for complete wetting, the meniscus is curved upwards (*fig. 6*). With a lead of 0.6 mm diameter the

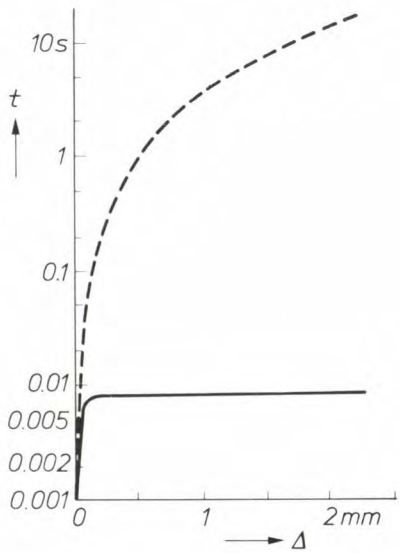


Fig. 5. The solid curve represents the calculated time required for the underside of printed boards to reach a temperature of 183 °C after contact with a solder bath at 240 °C, as a function of the board thickness Δ. The underside of the printed board, which is at ambient temperature (20 °C), comes into contact with the solder bath at the time $t = 0$. The dashed curve gives the corresponding results for the upper side of the printed board. (The heat-transfer coefficient is taken as 10 kW/m² °C. The presence of holes in the boards is not taken into account.)

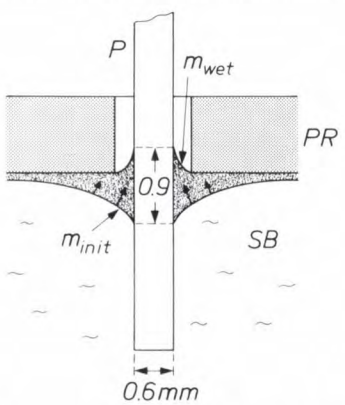


Fig. 6. Solder 'climbing' up a lead during soldering. Immediately after immersion of the lead (*P*) in the solder bath (*SB*) the surface meniscus becomes convex (m_{init}). When the solder meniscus becomes concave (m_{wet}) the lead is fully wetted. *PR* printed board. For clarity the diameter of the hole in the printed board is shown disproportionately large.

difference in solder level can be as much as 0.9 mm. The effect of this is to reduce the free length of the lead, with full wetting, by about 1 mm (maximum), which means that the body is heated up rather more rapidly than the model predicts.

Model calculations on capacitors

As an example the model in fig. 2 was used to calculate the thermal behaviour of two very different capacitors. The data used are listed in *Table III*; the free length of the leads was varied from a few mm to 10 mm. This parameter, which has a considerable effect on the thermal behaviour, usually provides the simplest means of adapting various components to the soldering process used. The results of the calculations are summarized in *fig. 7* and *fig. 8*. Each curve consists in principle of two branches, one more or less horizontal and one more or less vertical. Of the upper four curves in fig. 8 the vertical branch falls completely or partly in the temperature region below 200 °C. In the region on the left of the vertical branch the solder-bath temperature is too low, so that no wetting can occur; in the region above the horizontal branch the soldering time is too long, causing damage to the component. The region between the two branches has the correct combinations of solder-bath temperature and soldering time, so that the wetting occurs without damage to the component. Since the branches are roughly horizontal and vertical, two practical rules for soldering may be formulated:

— if the leads are not wetted in a soldering process, it is preferable to increase the temperature of the solder bath;

— if the body of a component is damaged in a soldering process, it is preferable to shorten the soldering time.

In both cases the other possible change (longer soldering time or lower solder-bath temperature) would have to be relatively much greater to obtain a properly soldered joint.

A study was also made of the extent to which the material of the leads affects the thermal inertia. The materials compared were iron and copper (*fig. 9*). The wetting time for iron leads is about a sixth of that for copper leads, under otherwise identical conditions. The temperature of a body with an iron lead is significantly lower than that of the same body with a copper lead, even at the time corresponding to the wetting of a copper lead.

It should be emphasized here that the better soldering performance of iron leads is not the only factor that determines the choice of material, but the matter will not be discussed further here.

Table III. Data for calculating the thermal behaviour of an electrolytic capacitor (C1) and of a microcapacitor (C2).

Component	C (J/°C)	R_w/l (°C/W mm)	R_s (°C/W)	R_b (°C/W)	T_{dam}	T_{wet} (°C)	$T_{amb} = T_{pre}$
C1	0.28	10	20	∞	85	183[*]	20
C2	0.055	41.5	20	∞	114	183[*]	20

[*] eutectic temperature of solder (tin-lead)

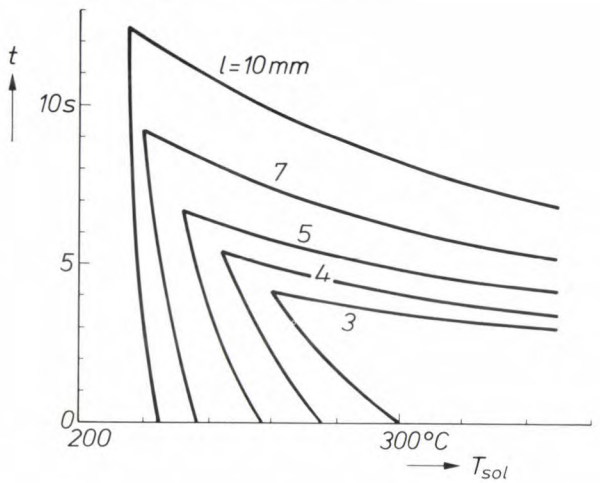


Fig. 7. Calculated thermal-behaviour profiles of the electrolytic capacitor C1 in dip soldering (see also Table III). The soldering time t is plotted vertically, the solder-bath temperature T_{sol} horizontally. Beside the 'horizontal' branch of each profile the free length l is given for the lead to which the curve relates (see fig. 2). Complete wetting of the lead without damage to the capacitor occurs only for combinations of t and T_{sol} on the right of the 'vertical' branch and below the 'horizontal' branch of the appropriate profile. To the left of a vertical branch the wetting is insufficient, and above a horizontal branch the capacitor is damaged.

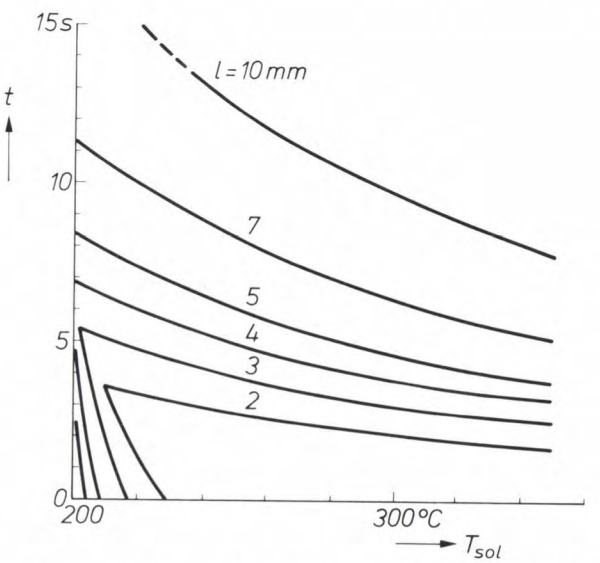


Fig. 8. Calculated thermal-behaviour profiles for the microcapacitor C2 in dip soldering (see also Table III). A fuller explanation is given in the caption to fig. 7. The 'vertical' branches of some of the profiles lie to the left of the vertical axis (T_{sol} = 200 °C) for this capacitor and are therefore not shown. If in addition to a capacitor C2 a capacitor C1 also has to be soldered to a particular board (fig. 7) an acceptable soldering time and solder-bath temperature can easily be found for both capacitors by varying the length l.

Preheating

We shall now return to the question of preheating, which has been the subject of some debate. It is obvious that preheating will make the moment of wetting earlier. Our calculations, which closely resemble those made with the examples given in the previous section, indicate that the time interval between the moment of wetting and the moment at which the body reaches the damage temperature is *independent* of the preheating. Thus, the moment at which damage begins is brought forward to the same extent, so that preheating does not offer any greater margin.

When the conditions of the soldering process are such that the lead is immediately wetted on immersion,

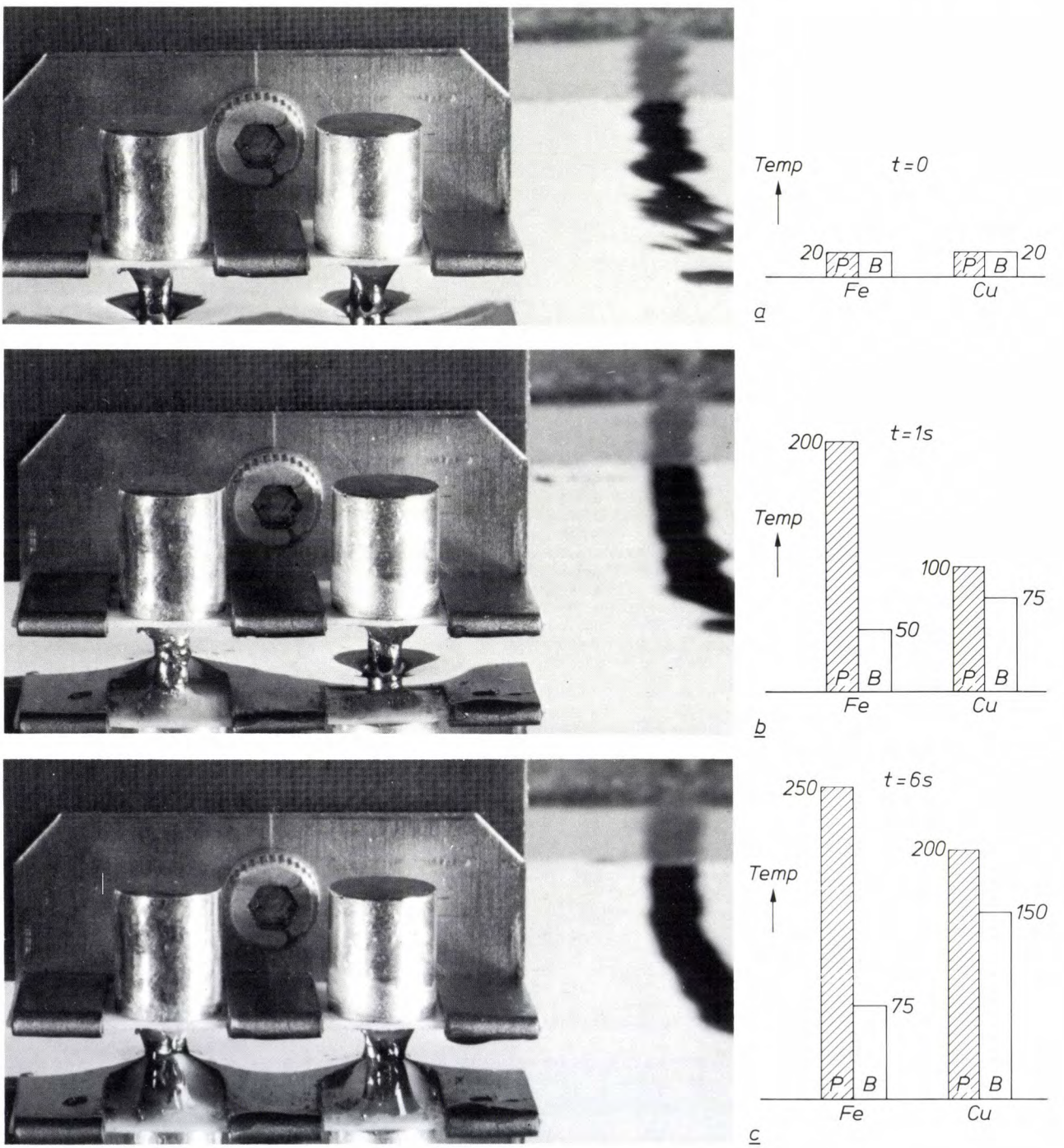


Fig. 9. Comparison of the behaviour of iron leads (*left* in the photographs) and copper leads (*right*). *a*) At $t = 0$ both leads are dipped into the solder. In both cases the meniscus is curved downwards (see fig. 6). *b*) After 1 second the iron lead is completely wetted (meniscus curved upwards). The meniscus for the copper lead is still curved downwards. *c*) After 6 seconds the copper lead is also completely wetted. The histograms give the corresponding lead temperatures T_w and body temperatures T_b in °C. *P* lead. *B* body.

the preheating only makes the damage start earlier. Under these conditions the preheating therefore shortens the available soldering time.

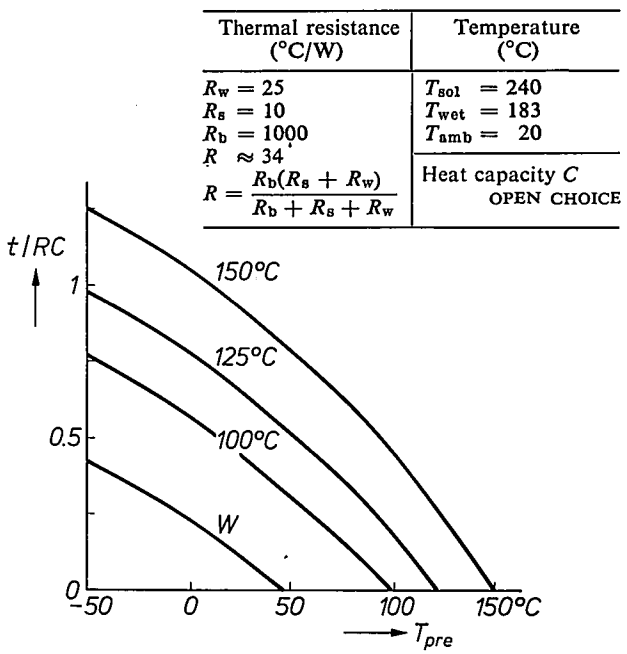


Fig. 10. The effect of preheating on soldering. The calculated (dimensionless) soldering time is plotted vertically, the preheating temperature T_{pre} horizontally. The preheating temperature is the temperature to which the body of the electronic component and its lead are raised before soldering starts. *W* wetting line; this gives the minimum time necessary for wetting the lead to be soldered (fig. 2). The three other curves are 'damage lines'; the value beside them is the temperature at which damage occurs. The acceptable soldering times lie between the damage line and the wetting line *W*. At preheating temperatures below about 50 °C the time between the occurrence of wetting and damage is roughly constant, at higher temperatures it is shorter. The values used for the calculation (fig. 2) are shown in the figure.

In *fig. 10* a number of calculated (dimensionless) soldering times are plotted as a function of the preheating temperature. The characteristic data for the case selected are indicated in the figure. The lowest curve is the 'wetting line', the other curves are the 'damage lines'. The damage temperature is indicated beside each of these curves. The curves in fig. 10 give a further indication that the time interval between wetting and damage is constant, at least at preheating temperatures lower than about 50 °C, since the curves in that range are parallel. Finally, we should note that preheating may be applied for reasons other than those considered here, in particular to give the flux the desired viscosity.

Summary. The article describes the calculated thermal behaviour of electronic components, such as capacitors and resistors, during soft soldering to printed-circuit boards. The components consist of a body, which must remain below a defined damage temperature, and a metal lead to be soldered, which must at least reach the wetting temperature. Each component has a typical range of solder-bath temperatures and soldering times in which both conditions are fulfilled. For different components these ranges can be made to overlap sufficiently by varying the length of the lead. Conversely, the results of the calculations can be a help in designing new components to meet the soldering conditions. The thermal behaviour is simulated by an electrical network that satisfies a first-order differential equation. A dimensionless temperature-time diagram derived from the network simplifies its application to various components. The conditions determining the validity of the model and the main thermal quantities are discussed. The printed boards have little effect on the thermal behaviour. The wetting time for iron leads is about a sixth of that for copper leads. In cases where wetting does not occur 'immediately', preheating has no effect on the time between wetting and damage. Wave soldering gives body temperatures that are a few per cent higher than in dip soldering, with slightly shorter wetting times.

Scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands *E*
Philips Research Laboratories, Redhill, Surrey, England *R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany *H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium *B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. *N*

P. Asbeck: Self-absorption effects on the radiative lifetime in GaAs-GaAlAs double heterostructures.
J. appl. Phys. **48**, 820-822, 1977 (No. 2). *N*

C. I. M. Beenakker: A cavity for microwave-induced plasmas operated in helium and argon at atmospheric pressure.
Spectrochim. Acta **31B**, 483-486, 1976 (No. 8/9). *E*

U. Bergmann, K. Klose & L. Spiess: Focus path of a laser beam deflected by a prismatic polygon mirror: its calculation and optimization.
Appl. Optics **15**, 3084-3088, 1976 (No. 12). *H*

O. H. Blaauw, G. Blaauw-Jansen (both with Botanisch Laboratorium, Utrecht) **& O. Elgersma:** Determination of hit numbers from dose-response curves for phytochrome control of seed germination (Lactuca sativa cv Noran).
Acta bot. neerl. **25**, 341-348, 1976 (No. 5). *E*

B. Bölger, L. Baede & H. M. Gibbs (Bell Telephone Laboratories, Murray Hill, N.J.): Production of 300 W, nanosecond, transform limited optical pulses.
Optics Comm. **19**, 346-349, 1976 (No. 3). *E*

H. Bouma & Ch. P. Legein (Institute for Perception Research, Eindhoven): Foveal and parafoveal recognition of letters and words by dyslexics and by average readers.
Neuropsychologia **15**, 69-80, 1977 (No. 1).

D. J. Breed & R. H. Doremus (Rensselaer Polytechnic Institute, Troy, N.Y.): Hydrogen profiles in water-oxidized silicon.
J. phys. Chem. **80**, 2471-2473, 1976 (No. 22). *E*

R. Brehm & M. Helmig: Het meten van hoge snelheden in een glazen smoor-ejecteur.
Mikroniek **16**, 138-141, 1976 (No. 7/8). *E*

A. Bril & A. W. de Jager-Veenis: Some methods of luminescence efficiency measurements.
J. Res. Nat. Bur. Stand. **80A**, 401-407, 1976 (No. 3). *E*

J. W. Broer: Let's become scientist-proof!
Proc. 23rd Int. Technical Communication Conf., Washington D.C. 1976, pp. 113-118. *E*

J. W. Broer: Impressions of the 1976 STC conference.
Communicator sci. & tech. Inform. No. 29, 2-3, Oct. 1976. *E*

J. Brokken-Zijp: Isomerization and decomposition of alkylaryldiazosulphides by light.
Mol. Photochem. **7**, 399-440, 1976 (No. 4). *E*

S. D. Brotherton, A. Gill & M. J. King: C.C.D.-addressed liquid-crystal displays.
Electronics Letters **13**, 270-272, 1977 (No. 9). *R*

J. Bruinink: Electrochromic display devices.
Nonemissive electrooptic displays, ed. A. R. Kmetz & F. K. von Willisen, pp. 201-221; Plenum Press, New York 1976. *E*

K. H. J. Buschow: Note on the change in magnetic properties of $GdCo_2$ on hydrogen absorption.
J. less-common Met. **51**, 173-175, 1977 (No. 1). *E*

K. H. J. Buschow & F. J. van Steenwijk (Kamerlingh Onnes Laboratorium, Leiden): Magnetic properties and Mössbauer effect of some Eu-Cd compounds.
Physica **85B**, 122-126, 1976 (No. 1). *E*

P. J. Cameron, J. M. Goethals, J. J. Seidel & E. E. Shult: Line graphs, root systems, and elliptic geometry.
J. Algebra **43**, 305-327, 1976 (No. 1). *B*

F. M. A. Carpay, S. Mahajan*, G. Y. Chin* & J. J. Rubin* (* Bell Laboratories, Murray Hill, N.J.): The correlation of textural and microstructural changes in deformed molybdenum single crystals.
Acta metall. **25**, 149-159, 1977 (No. 2). *E*

F. M. A. Carpay & A. L. Stuijts: Characterization of grain growth phenomena during sintering of single-phase ceramics.
Science of Ceramics **8**, 23-38, 1976. *E*

T. A. C. M. Claasen, W. F. G. Mecklenbräuker & J. B. H. Peek: Effects of quantization and overflow in recursive digital filters.
IEEE Trans. **ASSP-24**, 517-529, 1976 (No. 6). *E*

T. A. C. M. Claasen & J. B. H. Peek: A digital receiver for tone detection applications.
IEEE Trans. **COM-24**, 1291-1300, 1976 (No. 12). *E*

T. Conklin (Amperex Electronic Corporation, Slatersville, R.I.) **& E. H. Stupp:** Applications of the pyroelectric vidicon.
Opt. Engng. **15**, 510-515, 1976 (No. 6). *N*

B. H. J. Cornelissen, L. J. M. Esser & L. J. van de Polder: Applications of charge transfer devices in television.
3rd Int. Conf. 'The technology and applications of charge coupled devices' (CCD'76), Edinburgh 1976, pp. 348-363. *E*

J. E. Curran: Vacuum technologies applied to electronic component fabrication.
J. Vac. Sci. Technol. **14**, 108-113, 1977 (No. 1). *R*

J. P. M. Damen, J. A. Pistorius & J. M. Robertson: Calcium gallium germanium garnet as a substrate for magnetic bubble application.
Mat. Res. Bull. **12**, 73-78, 1977 (No. 1). *E*

N. H. Dekkers, H. de Lang & K. D. van der Mast (Philips Development Group Electron Optics, Eindhoven): Field emission STEM on a Philips EM 400 with a new detection system for phase and amplitude contrast.
J. Microsc. Spectrosc. électron. **1**, 511-512, 1976 (No. 3). *E*

M. Delfino, J. P. Dougherty, W. K. Zwicker & M. M. Choy: Solution growth and characterization of L(+) glutamic acid hydrochloride single crystals.
J. Crystal Growth **36**, 267-272, 1976 (No. 2). *N*

P. Delsarte & R. J. McEliece: Zeros of functions in finite Abelian group algebras.
Amer. J. Math. **98**, 197-224, 1976 (No. 1). *B*

P. Delsarte & P. Piret: Semiregular convolutional codes: definition, structure, and examples.
Information and Control **33**, 56-71, 1977 (No. 1). *B*

A. M. van Diepen & Th. J. A. Popma: Mössbauer effect and magnetic properties of an amorphous Fe_2O_3.
J. Physique **37**, C6/755-758, 1976 (Colloque C6). *E*

J. G. Dil & E. M. Brody (University of Rochester, N.Y.): Brillouin scattering from isotropic metals.
Phys. Rev. B **14**, 5218-5227, 1976 (No. 12). *E*

R. J. Dolphin & F. W. Willmott: Band-broadening effects in a column-switching system for HPLC.
J. chromatogr. Sci. **14**, 584-588, 1976 (No. 12). *R*

J. W. F. Dorleijn & A. R. Miedema: The magnetic resistance anisotropy in nickel and iron based alloys.
AIP Conf. Proc. **34**, 50-54, 1976. *E*

J. W. F. Dorleijn & A. R. Miedema: The residual resistivities of dilute iron-based alloys in the two-current model.
J. Physics F **7**, L 23-25, 1977 (No. 1). *E*

H. Dötsch: Dynamics of magnetic domains in microwave fields.
J. Magnetism and magn. Mat. **4**, 180-185, 1977 (No. 1-4). *H*

H. Dötsch: Influence of spin precession on the stability of bubble domains.
Phys. Stat. sol. (a) **39**, 589-594, 1977 (No. 2). *H*

W. F. K. van den Eijnde: Is goede raad nu echt zo duur?
Polytechn. T. pt/aktueel 1976, No. 49, p. 13. *E*

F. L. Engel (Institute for Perception Research, Eindhoven): Visual conspicuity, visual search and fixation tendencies of the eye.
Vision Res. **17**, 95-108, 1977 (No. 1).

J. M. L. Engels & A. J. M. van Kimmenade: The mobility of excess electrons in compressed liquid methane.
Physics Letters **59A**, 43-44, 1976 (No. 1). *E*

U. Enz: A new type of soliton with particle properties.
J. math. Phys. **18**, 347-353, 1977 (No. 3). *E*

L.-E. Eriksson & H. C. van den Elzen: An equalizer structure with reduced sampling time reference sensitivity.
IEEE Trans. **COM-24**, 1337-1343, 1976 (No. 12). *E*

E. Fabre: MIS silicon solar cells.
Appl. Phys. Letters **29**, 607-610, 1976 (No. 9). *L*

J. M. Friedman, H. M. Gibbs, T. N. C. Venkatesan (all with Bell Laboratories, Murray Hill, N.J.), **B. Bölger, D. Polder & M. F. H. Schuurmans:** Absorption notch in the spectrum of a transform limited pulse of small area.
Optics Comm. **20**, 183-187, 1977 (No. 1). *E*

Y. Genin & Y. Kamp: Planar least-square inverse stabilization of 2d recursive digital filters.
Proc. 1976 Eur. Conf. on Circuit theory and design, Genova, pp. 430-435. *B*

C. J. Gerritsma & P. van Zanten: Scanning limitations of twisted nematic display devices.
Nonemissive electrooptic displays, ed. A. R. Kmetz & F. K. von Willisen, pp. 298-302; Plenum Press, New York 1976. *E*

J. A. Geurst: Canonical form of generalised two-fluid equations for helium II.
Physics Letters **59A**, 351-352, 1976 (No. 5). *E*

J.-M. Goethals: The extended Nadler code is unique.
IEEE Trans. **IT-23**, 132-135, 1977 (No. 1). *B*

M. Goscianski & J. Donjon: Matrix display using a twisted nematic liquid crystal dynamically addressed.
Tech. Dig. 1976 Int. Electron Devices Meeting, Washington D.C., pp. 626-628. *L*

R. G. Gossink: Glasfibers voor optische communicatie.
Klei en Keramiek **26**, 230-238, 1976 (No. 12). *E*

G. Groh: Tomosynthesis and coded aperture imaging: new approaches to three-dimensional imaging in diagnostic radiography.
Proc. Roy. Soc. London B **195**, 299-306, 1977 (No. 1119). *H*

G. J. van Gurp: Electromigration in cobalt films.
Thin Solid Films **38**, 295-311, 1976 (No. 3). *E*

J. Hallais: Application de la méthode de croissance aux organo-métalliques à l'épitaxie en phase vapeur des composés III-V.
Proc. Coll. Matériaux et technologie pour la microélectronique, tendances actuelles, Montpellier 1976 (Suppl. Le Vide No. 183), pp. 39-49. *L*

H. W. Hanneman & B. Visser: An experimental TV ghost cancellation circuit with automatic gain control.
3rd Int. Conf. 'The technology and applications of charge coupled devices' (CCD'76), Edinburgh 1976, pp. 371-379. *E*

K. H. Härdtl: Ferroelectric displays.
Nonemissive electrooptic displays, ed. A. R. Kmetz & F. K. von Willisen, pp. 241-259; Plenum Press, New York 1976. *A*

J. Hasker: A new concept for fluorescent lamps.
J. Illum. Engng. Soc. **6**, 29-34, 1976 (No. 1). *E*

J. C. M. Henning & J. H. den Boef: High resolution ESR spectrum of cubic Cr^{3+} in MgO.
Physics Letters **59A**, 241-242, 1976 (No. 3). *E*

D. Hennings: The broadening of the Curie peak by lattice defects in ferroelectric lead titanate — diffuse phase transition.
Science of Ceramics **8**, 203-211, 1976. *A*

D. Hennings: Defect chemistry of La-doped barium titanate.
Reactivity of solids, Proc. 8th Int. Symp., Göteborg 1976, pp. 273-277; 1977. *A*

S. van Heusden & L. P. J. Hoogeveen: Chemiluminescent determination of reactive hydrocarbons.
Z. anal. Chemie **282**, 307-313, 1976 (No. 4). *E*

J. H. C. van Heuven & T. E. Rozzi: Two applications of algebraic invariants to linear transmission networks.
IEEE Trans. **CAS-23**, 641-647, 1976 (No. 11). *E*

B. Hoekstra, F. van Doveren & J. M. Robertson: Magnetostriction constants of epitaxial La,Ga:YIG films measured by microwave resonance.
Appl. Phys. **12**, 261-263, 1977 (No. 3). *E*

B. Hoekstra, J. M. Robertson & W. T. Stacy: The origin of the uniaxial anisotropy in thin films of $(YLaPb)_3(FeGa)_5O_{12}$ and its variation along the growth direction.
Mat. Res. Bull. **12**, 53-64, 1977 (No. 1). *E*

B. Hoekstra, R. P. van Stapele & J. M. Robertson: Spin-wave resonance spectra of inhomogeneous bubble garnet films.
J. appl. Phys. **48**, 382-395, 1977 (No. 1). *E*

P. J. 't Hoen: On a relation between image quality and the modulation transfer function.
7th L. H. Gray Conf.: Medical images, Leeds 1976, pp. 173-183. *E*

L. Hollan: Evolution des techniques d'élaboration de GaAs épitaxial pour les dispositifs hyperfréquences.
Le Vide **31**, 131-137, 1976 (No. 184). *L*

L. Hollan, J. M. Durand & R. Cadoret: Influence of the growth parameters in GaAs vapor phase epitaxy.
J. Electrochem. Soc. **124**, 135-139, 1977 (No. 1). *L*

E. P. Honig & D. den Engelsen: Reflection from stratified anisotropic media: an alternative method.
Optica Acta **24**, 89-95, 1977 (No. 1). *E*

W. H. de Jeu, W. A. P. Claassen & A. M. J. Spruijt: The determination of the elastic constants of nematic liquid crystals.
Mol. Cryst. liq. Cryst. **37**, 269-280, 1976 (No. 1-4). *E*

L. Jeunhomme (CGE, Marcoussis), **A. Cozannet** (CNET, Lannion), **R. Bouillie** (CNET, Lannion) **& J. P. Hazan:** Mesure des caractéristiques de transmission de conducteurs optiques.
Onde électr. **56**, 564-571, 1976 (No. 12 bis). *L*

B. A. Joyce & C. T. Foxon: Growth and doping of semiconductor films by molecular beam epitaxy.
Solid state devices, 1976 (ESSDERC Munich; Inst. Phys. Conf. Ser. No. 32), pp. 17-37; 1977. *R*

G. C. Kenney: Special purpose applications of the optical videodisc system.
IEEE Trans. **CE-22**, 327-338, 1976 (No. 4). *N*

F. M. Klaassen: Physics of and models for I^2L.
Tech. Dig. 1976 Int. Electron Devices Meeting, Washington D.C., pp. 299-303. *E*

F. M. Klaassen: Some considerations on high-speed injection logic.
IEEE J. **SC-12**, 150-154, 1977 (No. 2). *E*

H. A. Klasens (Philips Scientific and Industrial Equipment Division, Eindhoven) **& J. Goossen:** The iodide interference with silver chloride electrodes.
Anal. chim. Acta **88**, 41-46, 1977 (No. 1). *E*

W. L. Konijnendijk: Structure of glasses in the systems $CaO\text{-}Na_2O\text{-}B_2O_3$ and $MgO\text{-}Na_2O\text{-}B_2O_3$ studied by Raman scattering.
Phys. Chem. Glasses **17**, 205-208, 1976 (No. 6). *E*

W. L. Konijnendijk & J. H. J. M. Buster: Raman-scattering measurements on silicate glasses containing AsO_4^{3-} ions.
J. non-cryst. Solids **22**, 379-389, 1976 (No. 2). *E*

W. L. Konijnendijk (Philips Lighting Division, Eindhoven) **& J. M. Stevels** (Eindhoven University of Technology): Viscosity of borosilicate glasses in relation to their structure.
Verres Réfract. **30**, 821-826, 1976 (No. 6).

W. L. Konijnendijk & H. Verweij: Structural aspects of vitreous $PbO.2B_2O_3$ studied by Raman scattering.
J. Amer. Ceramic Soc. **59**, 459-461, 1976 (No. 9/10). *E*

L. Koppens: The decomposition of organometallic precursors for ferrite powders.
Science of Ceramics **8**, 101-109, 1976. *E*

B. M. Kramer, A. C. Derycke (Université de Lille), **A. Farrayre & C. F. Masse:** High-efficiency frequency multiplication with GaAs avalanche diodes.
IEEE Trans. **MTT-24**, 861-863, 1976 (No. 11). *L*

E. Krätzig & H. Kurz: Spectroscopic investigation of photovoltaic effects in doped $LiNbO_3$.
J. Electrochem. Soc. **124**, 131-134, 1977 (No. 1). *H*

J.-P. Krumme, H. Heitmann, D. Mateika & K. Witter: MOPS, a magneto-optic-photoconductor sandwich for optical information storage.
J. appl. Phys. **48**, 366-368, 1977 (No. 1). *H*

M. Lacroix & A. Pirotte: Generalized joins.
SIGMOD Rec. **8**, No. 3, pp. 14-15, Sept. 1976. *B*

P. K. Larsen, G. Margaritondo*, J. E. Rowe*, M. Schluter* & N. V. Smith* (* Bell Laboratories, Murray Hill, N.J.): GaSe valence band structure from angle-resolved photoemission spectroscopy.
Physics Letters **58A**, 423-425, 1976 (No. 6). *E*

P. E. J. Legierse (Philips Philite- en Metaalwarenfabrieken, Eindhoven): Fotochemische metaalbewerking.
Polytechn. T. Werktuigbouw **32**, 27-35, 1977 (No. 1).

J. C. van Lier, G. J. Koel, W. J. van Gestel, L. Postma, J. T. Gerkema, F. W. Gorter & W. F. Druyvesteyn: Combined thin film magnetoresistive read, inductive write heads.
IEEE Trans. **MAG-12**, 716-718, 1976 (No. 6). *E*

J. Lohstroh: Integrated Injection Logic.
Bull. SEV/VSE **68**, 53-59, 1977 (No. 2). *E*

J. Lohstroh: Dynamic behavior of active charge in I^2L transistors calculated with lumped transistor models.
IEEE J. **SC-12**, 176-184, 1977 (No. 2). *E*

F. Mahdjuri: Highly insulated window glazing.
Int. J. Energy Res. **1**, 135-142, 1977 (No. 2). *A*

R. Memming: Electrochemical surface reactions on non-metals.
Nat. Bur. Stand. spec. Publn No. 455, 267-280, 1976. *H*

W. A. van der Meulen & P. J. F. C. Janssen: Automatic versus manual indexing.
Inform. Process. & Man. **13**, 13-21, 1977 (No. 1). *E*

A. R. Miedema: Elektrische weerstand in ferromagnetische metalen.
Chem. Weekbl. Mag. 1976, m 537-540 (okt.); Ned. T. Natuurk. **42**, 193-196, 1976 (No. 16). *E*

A. Mircea, A. Mitonneau & J. Vannimenus (Ecole Normale Supérieure, Paris): Temperature dependence of ionization energies of deep bound states in semiconductors.
J. Physique Lettres **38**, L 41-42, 1977 (No. 1). *L*

R. F. Mitchell: Basics of SAW frequency filter design: a review.
Wave Electronics **2**, 111-132, 1976 (No. 1/2/3). *R*

A. E. Morgan, H. W. Werner & J. M. Gourgout (Cameca S.A., Courbevoie, France): In-depth concentration profiling of garnet epilayers using secondary ion mass spectrometry.
Appl. Phys. **12**, 283-286, 1977 (No. 3). *E*

J. H. Neave & B. A. Joyce: Some comments on electron-beam-induced adsorption.
J. Physics D **10**, 243-248, 1977 (No. 2). *R*

Ngo-Tich-Phuoc, G. M. Martin, C. Belin & E. Fabre: Homogeneity along Cl-compensated THM grown CdTe ingot.
Rev. Phys. appl. **12**, 195-198, 1977 (No. 2). *L*

K. H. Nicholas: Implantation damage in silicon devices.
J. Physics D **10**, 393-407, 1977 (No. 4). *R*

A. G. van Nie: Modulated carriers in nonlinear systems.
Int. J. Circuit Theory & Appl. **5**, 69-79, 1977 (No. 1). *E*

C. van Opdorp, C. Werkhoven & A. T. Vink: A method to determine bulk lifetime and diffusion coefficient of minority carriers; application to *n*-type LPE GaP.
Appl. Phys. Letters **30**, 40-42, 1977 (No. 1). *E*

R. Orlowski, E. Krätzig & H. Kurz: Photorefractive effects in $LiNbO_3$:Fe under external electric fields.
Optics Comm. **20**, 171-174, 1977 (No. 1). *H*

K. R. Peschmann & G. Grosche: Amplification and entailed resolution degradation in high-pressure gas ionography.
Medical Phys. **4**, 202-207, 1977 (No. 3). *A*

R. J. van de Plassche: Dynamic element matching for high-accuracy monolithic D/A converters.
IEEE J. **SC-11**, 795-800, 1976 (No. 6). *E*

A. Posthuma de Boer & A. J. Pennings (State University of Groningen): Polyethylene networks crosslinked in solution: preparation, elastic behavior, and oriented crystallization, I. Crosslinking in solution.
J. Polymer Sci. Pol. Phys. Edn **14**, 187-210, 1976 (No. 2). *E*

H. F. Premsela: A stereo technique for measuring the depth of specimens in transmission electron microscopy.
Micron **7**, 171-173, 1976 (No. 2). *E*

P. Röschmann: Two-magnon-scattering contributions to FMR linebroadening in polycrystalline garnets.
AIP Conf. Proc. **34**, 253-258, 1976. *H*

P. Röschmann & G. Winkler: Relaxation processes in polycrystalline substituted garnets with low ferrimagnetic resonance linewidth.
J. Magnetism and magn. Mat. **4**, 105-115, 1977 (No. 1-4). *H*

W. Schäfer, G. Will (both with Universität Bonn) **& K. H. J. Buschow:** A neutron diffraction investigation on the magnetic properties of ErMg and DyMg.
J. Physics C **9**, L 657-661, 1976 (No. 23). *E*

P. Schagen: Some recent developments in remote sensing.
Nature **266**, 223-228, 1977 (No. 5599). *R*

M. F. H. Schuurmans: Narrowing of level crossing curves in selective reflection.
Z. Physik A **279**, 243-248, 1976 (No. 3). *E*

G. B. Scott & J. L. Page: The absorption spectra of $Y_3Fe_5O_{12}$ and $Y_3Ga_5O_{12}$:Fe^{3+} to 5.5 eV.
Phys. Stat. sol. (b) **79**, 203-213, 1977 (No. 1). *R*

G. B. Scott & J. L. Page: Pb valence in iron garnets.
J. appl. Phys. **48**, 1342-1349, 1977 (No. 3). *R*

R. J. Seymour & M. M. Choy: High-efficiency difference frequency generation in $LiNbO_3$.
Optics Comm. **20**, 101-103, 1977 (No. 1). *N*

J. G. Siekman: Over het boren van diamant met een laserbundel.
Ned. T. Natuurk. **42**, 224-225, 1976 (No. 18). *E*

M. Sintzoff: Résolution d'équations récursives de prédicats.
Colloque sur les arbres en algèbre et programmation, Lille 1976, pp. 239-246. *B*

M. Sintzoff: La conception et l'écriture du logiciel.
Panorama de la nouveauté informatique en France, Congrès AFCET, Paris 1976, pp. 351-357. *B*

J. W. Slotboom: Minority carrier injection into heavily doped silicon.
Solid-State Electronics **20**, 167-170, 1977 (No. 2). *E*

L. A. Æ. Sluyterman & J. Wijdenes: Proton equilibria in the binding of Zn^{2+} and of methylmercuric iodide to papain.
Eur. J. Biochem. **71**, 383-391, 1976 (No. 2). *E*

N. V. Smith*, P. K. Larsen, M. M. Traum* & S. Chiang* (* Bell Laboratories, Murray Hill, N.J.): Miniature analyser and its use in angle-resolved photoemission using synchrotron radiation.
Proc. Int. Symp. on Photoemission, Noordwijk 1976, pp. 119-123. *E*

W. T. Stacy, A. B. Voermans & H. Logmans: Increased domain wall velocities due to an orthorhombic anisotropy in garnet epitaxial films.
Appl. Phys. Letters **29**, 817-819, 1976 (No. 12). *E*

S. Strijbos: Friction between a powder compact and a metal wall.
Science of Ceramics **8**, 415-427, 1976. *E*

E. H. Stupp & A. Milch: Measurement of thickness and diffusion length of thin epitaxial layers of GaP.
J. appl. Phys. **48**, 282-285, 1977 (No. 1). *N*

J.-B. Theeten: Les agrégats, entre l'atome et le solide.
La Recherche **8**, 170-171, 1977 (No. 75). *L*

M. J. J. Theunissen, H. W. Hanneman & H. A. Schmidt (Philips I.C. Development Laboratories, Nijmegen): Design aspects of S-shaped bulk charge-coupled devices.
3rd Int. Conf. 'The technology and applications of charge coupled devices' (CCD'76), Edinburgh 1976, pp. 44-54. *E*

J.-P. Thiran: On unit element structures for wave digital filters.
IEEE Trans. **CAS-24**, 20-28, 1977 (No. 1). *B*

G. E. Thomas: A molecular mechanism for the chemical effect in bombardment-induced light emission.
Radiation Effects **31**, 185-186, 1977 (No. 3). *E*

J. C. Tranchart, A. Farrayre & L. Hollan: Electrochimie du GaAs: étude et applications.
Proc. Coll. Matériaux et technologie pour la microélectronique, tendances actuelles, Montpellier 1976 (Suppl. Le Vide No. 183), pp. 77-94. *L*

H. Vantilborgh: Working set dynamics.
Modelling and performance evaluation of computer systems, Proc. Int. Workshop, Stresa 1976, pp. 377-387; 1977. *B*

H. J. Verbeek (Philips Elcoma Division, Eindhoven): A model to evaluate design and application of components regarding heat transfer during soldering.
DVS-Ber. **40**, 35-39, 1976.

M. L. Verheijke: Some practical procedures in computerized thermal neutron activation analysis with Ge(Li) gamma-ray spectrometry.
J. radioanal. Chem. **35**, 79-93, 1977 (No. 1). *E*

H. Verweij & H. van den Boom: The use of laser Raman spectroscopy in the study of the formation of oxide glasses.
Reactivity of solids, Proc. 8th Int. Symp., Göteborg 1976, pp. 529-534; 1977. *E*

H. Verweij & W. L. Konijnendijk: Structural units in K_2O-PbO-SiO_2 glasses by Raman spectroscopy.
J. Amer. Ceramic Soc. **59**, 517-521, 1976 (No. 11/12). *E*

H. Weiss: Use and abuse of the modulation transfer function.
7th L. H. Gray Conf.: Medical images, Leeds 1976, pp. 161-172. *H*

C. J. Wellekens: Filtres d'amplitude et de phase spécifiées.
Onde électr. **57**, 59-63, 1977 (No. 1). *B*

H. W. Werner: Characterization of ceramics by means of modern thin film and surface analytical techniques.
Science of Ceramics **8**, 55-79, 1976. *E*

R. Wernicke: The kinetic processes of equilibrium restoration in La-doped $BaTiO_3$.
Reactivity of solids, Proc. 8th Int. Symp., Göteborg 1976, pp. 279-284; 1977. *A*

H. P. J. Wijn: Trends in the technology of magnetic devices.
Physics in industry, Proc. int. Conf. Dublin 1976, ed. E. O'Mongain & C. P. O'Toole, pp. 69-73; Pergamon Press, Oxford 1976. *E*

H. W. de Wijn (Rijksuniversiteit Utrecht), **A. M. van Diepen & K. H. J. Buschow:** Samarium in crystal fields, a case of strong *J*-mixing.
Phys. Stat. sol. (b) **76**, 11-30, 1976 (No. 1). *E*

H. J. de Wit, Ch. Wijenberg & C. Crevecoeur: The dielectric breakdown of anodic aluminum oxide.
J. Electrochem. Soc. **123**, 1479-1486, 1976 (No. 10). *E*

H. E. J. Wulms: Base current of I^2L transistors.
IEEE J. SC-12, 143-150, 1977 (No. 2). *E*

Y. T. Yeow*, M. R. Boudry, D. R. Lamb* & S. D. Brotherton* (* University of Southampton): Sources of errors in quasistatic capacitance-voltage determination of the interface state density distribution in the MOS system.
J. Physics D **10**, 83-95, 1977 (No. 1). *R*

W. K. Zwicker, M. Delfino, J. P. Dougherty, A. Sicignano, J. Ladell & J. A. Nicolosi: Formation of secondary phases during crystal growth of $Pb_5Ge_3O_{11}$.
J. electronic Mat. **6**, 125-143, 1977 (No. 2). *N*

J. O. Zwolsman (Philips Elcoma Division, Eindhoven): Influence of the current pulse shape in micro resistance welding.
DVS-Ber. **40**, 79-82, 1976.

Contents of Philips Telecommunication Review **35**, No. 3, 1977:

J. A. Brennan & J. van Duuren: The capacity of SPC Telex Systems (pp. 109-120).
A. de Waard: PRC/VRC-4600 VHF combat area radio equipment (pp. 121-133).
M. Kunst & J. Noordanus: Developments in digital radio link equipment (pp. 134-145).
W. Lulofs: An experimental domestic satellite earth station (pp. 147-153).

Contents of Philips Telecommunication Review **35**, No. 4, 1977:

W. G. Bax: 1 pW per km on a 3600 channel coaxial line (pp. 157-165).
R. P. Wiarda: TV transmission over an 18 MHz system (pp. 166-175).
Ph. Dancot, J. Dreze & L. Witmans: DACOS, a new operating system for DS 714 (pp. 176-185).
R. Azimullah & R. Deraemaeker: Simplex TOR STB 750 for world-wide radio telegraph communication (pp. 186-199).
W. M. Pannell: Private Mobile Radio — the next two decades (pp. 200-206).

Contents of Electronic Applications Bulletin **34**, No. 4, 1977:

The TCA580 integrated gyrator — design and applications (pp. 148-161).
Control circuits for SMPS used in TV receivers (pp. 162-180).
B. F. van der Heyden: The design of stable low-noise IMPATT oscillators for the telecommunications band (pp. 182-194).

Contents of Electronic Applications Bulletin **35**, No. 1, 1977:

An SMPS without mains isolation for TV receivers (pp. 2-23).
D. W. Parker, R. G. Pratt, F. W. Smith & R. Stevens: Acoustic surface-wave bandpass filters (pp. 24-39).
D. Hermans & M. D. Hull: Designing hi-fi speaker systems — part 1 (pp. 40-63).

Contents of Mullard Technical Communications **14**, No. 135, 1977:

A. Ciuciura: Picture centring in the 20AX system (pp. 174-180).
G. van Schaik: An introduction to switched-mode power supplies in tv receivers (pp. 181-195).
Channel electron multipliers (pp. 196-200).
A. Ciuciura: Soft-Flash: a new development in colour picture tube technology (pp. 201-208).

Contents of Mullard Technical Communications **14**, No. 136, 1977:

G. van Schaik: Control circuits for SMPS in tv receivers (pp. 210-226).
K. F. Etherington: Image intensifiers — an introduction (pp. 227-239).
P. G. Noble: A simple hand-held night viewer (pp. 240-244).

Recent United States Patents

Abstracts from patents that describe inventions from the following research laboratories that form part of or cooperate with the Philips group of companies:

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

4 080 719

Method of manufacturing a semiconductor device and device manufactured according to the method

H. J. H. Wilting *E*

A method of manufacturing a metal silicide pattern with respect to which two electrode zones are to be provided in a self-registering manner. According to the invention the pattern is provided in the form of a layer of polycrystalline silicon and, by selective oxidation and masking, only the upper surface of the pattern is exposed to the silicide formation so that passivation problems and short circuit are avoided. The use of silicides which cannot withstand high temperatures is also possible.

---

4 081 793

Device for reading out the charge condition of a phototransistor

J. Lohstroh *E*

A device for reading out the charge condition of a phototransistor by means of sampling pulses, for example read out of a transistor to be selected from a matrix of phototransistors, a voltage which is a measure of said charge condition being taken from the emitter of the phototransistor. In order to avoid the storage effect owing to the time constant which depends on the differential resistance of the base-emitter junction, the emitter circuit of the phototransistor includes a current source which at the instant of sampling is switched on, while furthermore the emitter voltage of the phototransistor is applied to a voltage comparator circuit which supplies a voltage for switching off the current source at the instant that said emitter voltage drops below a specified reference value. In the case of read-out after a dark period some charge is drained owing to the response time of electronic systems, which charge can be compensated for by simple injection means.

---

4 081 855

Current pulse generator

A. W. M. van den Enden *E*
E. de Niet
M. A. Deurwaarder

The invention relates to a device for generating a time-dependent current in a coil, for example, a magnetic coil. Such a magnetic coil can be used, in combination with a second coil of this kind, for generating rotary magnetic fields in a device for transporting magnetic bubbles along a bubble guide structure consisting of discrete elements on a plate of magnetic material. The device comprises two or more current connection terminals, an *LC* oscillator circuit, a switching element connected in series with the coil, and a control unit.

4 083 764

Ion-selective electrode

R. E. van de Leest *E*
N. M. Beekmans
L. Heijne

Ion-selective electrode for selectively measuring the concentration of an ion in a solution of a mixture of ions, consisting of an ion-conducting or mixed ion-conducting - electron-conducting matrix material provided with a thin surface coating of a material which is sparingly soluble and which has one ion in common with the matrix material.

---

4 084 180

Color splitting prism assembly

J. Stoffels *E*
A. A. J. Bluekens
P. J. M. Peters

A beam splitting prism assembly in which at least some of the prisms are separated by an air gap, optionally connected at the edges through the beam splitting layers by means of strips of a thermoplastic material.

---

4 084 185

Record carrier on which information is stored in an optically readable structure

H. de Lang *E*
G. Bouwhuis

A record carrier is described which is provided with an optically readable structure of trackwise arranged areas alternating with intermediate areas. Areas of at least two different types are provided. As a result, a high information density can be achieved. The different types of areas may also be used for centering a read beam relative to a track and for focussing the beam onto the track.

---

4 084 376

Heating system

G. A. A. Asselman *E*
H. H. M. van der Aa
R. J. Meijer

A heating system for supplying thermal energy to a heater of a machine such as a hot gas engine, this system using a heat-transporting medium which flows within a closed space between a heat-receiving portion to a heat-discharging portion. Thermal

energy is supplied from a heat source through a wall in the heat-receiving portion with the transporting medium changing from the liquid phase to the vapor phase, and the absorbed thermal energy is transferred to the heater in the heat-discharging portion by changing from the vapor phase into the liquid phase. The wall parts of the heat-receiving portion, through which thermal energy can be supplied from the heat source to the transporting medium, comprise on their surface contacting this medium a layer of a porous material having a capillary structure whereby liquid-phase medium is distributed on said walls.

---

4 084 471

Circuit arrangement for obtaining a chorus effect

N. V. Franssen E

A circuit arrangement for generating the tones of a tonal scale includes a master oscillator and a number of sets of frequency converters. The output signals of each set of frequency converters form each of the tones of an octave. A frequency shifter is connected between at least one selected set of frequency converters and the master oscillator in order to shift the frequency supplied to the selected set of frequency converters from the master oscillator to cause a desired shift in the frequency of each of the tones of the octave produced by the selected set of frequency converters.

---

4 085 285

Method of manufacturing printed circuit boards

G. J. M. Lippits E
P. J. Janssen
H. A. Debruijn
J. van Ruler

Manufacture of photosensitive material by a dispersion of particles of semiconducting metal oxide having a charge sign opposite to that of the substrate in the dispersion agent, which dispersion is provided on the substrate surface and adhered thereto. In a similar manner walls of a hole for electrical through-connections are made photosensitive as a pretreatment for metal-plating.

---

4 085 430

Thin film magnetic head with a gap formed between a loop shaped core part and a bridging core part

J. T. Gerkema E
F. W. Gorter

A thin-film magnetic head having improved playback properties in which the gap-bounding core parts overlap each other laterally while enclosing a layer having a gap and in which a first core part has the shape of a closed loop and is bridged in the center by a second core part. The parts of the first core part on either side of the bridge are provided with identical coils, while the first core part at the area of the overlap is divided centrally by an extra gap. Upon playing back, the difference signal of the coils then provides an indication as regards the position relative to the track.

---

4 086 431

Compression system

N. V. Franssen E
K. E. Kuijk
J. W. Siebelink

System for the vocoder transmission of speech signal with reduced bandwidth. The speech signal is split into a low-frequency part and a high-frequency part. The high-frequency part is subdivided into a number of sub-bands and for each sub-band a control signal with reduced bandwidth is transmitted. A control signal is representative for the ratio between the energy in the sub-band of the speech signal and the energy in the corresponding sub-band of a reference signal with high-frequency components, which is produced from the low-frequency part of the speech signal.

4 086 512

Camera tube employing siliconchalcogenide target with heterojunction

J. Dieleman E
A. M. E. Hoeberechts

A camera tube target formed by a radiation-receiving silicon layer which on a side to be scanned by the electron beam has a chalcogen-containing layer having intrinsic conductivity which forms a hetero junction with the silicon layer, the chalcogen-containing layer comprising at least one element of the fourth group of the periodic table of elements in an atomic ratio to the chalcogen component lying between 1:1 and 1:2.

---

4 086 559

Electric resistor based on silicon carbide having a negative temperature coefficient

W. F. Knippenberg E
G. Verspui
S. H. Hagen

An electric resistor having a negative temperature coefficient of the electrical resistance whose resistance body consist of *P*-type doped pyrolytic polycrystalline cubic silicon carbide.

---

4 087 671

Device for plasma-MIG welding

W. G. Essers E

A device for plasma-MIG welding comprising a welding torch provided with a non-consumable electrode and a supply of a welding wire, the non-consumable electrode being connected to a first power supply source having a drooping voltage characteristic and the welding wire being connected to a second power supply source having a constant voltage characteristic, the first power supply source being connected in series with the second power supply source, thereby improving the stability of the plasma arc.

---

4 089 588

Nematic liquid crystalline mixture of α-cyanostilbenes and application thereof in image display devices

M. de Zwart E
T. W. Lathouwers

Nematic liquid crystalline mixtures of α-cyanostilbenes with negative electric anisotropy which are nematic liquid crystalline in a temperature range around room temperature. The mixtures contain the compounds 2-(*p*-ethoxyphenyl)-3-(*p*-hexyloxyphenyl)-acrylonitrile and 2-(*p*-butoxyphenyl)-3-(*p*-heptylphenyl)-acrylonitrile or the compounds 2-(*p*-butoxyphenyl)-3-(*p*-pentyloxyphenyl)-acrylonitrile and 2-(*p*-methoxyphenyl)-3-(pentylphenyl)-acrylonitrile and are very suitable for practical application in picture display devices.

---

4 089 799

Luminescent fluoride

J. L. Sommerdijk E
F. M. J. H. Hoex-Strik

Luminescent bivalent europium-activated fluoride according to the formula

$$AB_{1-x-y}Mg_xEu_yF_3,$$

wherein A represents cesium and/or rubidium and B represents calcium and/or strontium and wherein

$0.10 \leqq x \leqq 0.90$
$0.001 \leqq y \leqq 0.20$
$x + y \leqq 0.95.$

4 090 138

FSK transmitter having frequency band-limitation

J. O. Voorman E
J. M. Schmidt

A signal consisting of square pulses is applied to an FSK transmitter and is transformed into a control signal consisting of substantially symmetrical trapezoidal pulses. This control signal is applied on the one hand to a gyrator oscillator modulator to generate an FSK signal and on the other hand to a conversion device which converts the control signal into a direct voltage signal the direct voltage value of which shows a positive pulsatory variation each time during the occurrence of the edges of the trapezoidal pulses of the control signal. Said direct voltage signal is used to subsequently modulate the FSK signal in an amplitude modulator. The amplitude-modulated FSK signal may now be applied directly as a modulation signal to a carrier modulator without the intermediary of a band-pass filter.

---

4 090 218

Method of, and apparatus for, manufacturing a video record

M. C. W. van Buul E
J. A. M. Hofman
L. J. van de Polder
S. L. Tan

When the information of a continuous-motion video record is displayed in the stop-motion mode by repeated display of two fields which form an interlaced picture, motion in the scene gives rise to a frame-frequency flicker phenomenon in the stop-motion picture. By forming suitable video signal combinations in discrete groups which each comprise two associated fields and by recording these combinations in the video record, the flicker phenomenon in a stop-motion picture is eliminated without continuous display being perceptibly affected.

---

4 091 429

Device for the magnetic recording by means of a magnetic strip domain as a recording/playback head

F. A. de Jonge E
J. W. F. Dorleijn
W. F. Druyvesteyn
G. J. Koel

The invention relates to a device for magnetic recording which is provided with a layer of a magnetizable material which is provided on a support and has an easy axis of magnetization which is at least substantially perpendicular to the plane of the layer, which layer is suitable to support magnetic domains having a direction of magnetization which is opposite to the direction of magnetization of the remainder of the layer, the device also comprising a device for producing a magnetization field having a field direction perpendicular to the plane of the layer, said layer being destined to cooperate with a recording medium which is relatively movable with respect to the layer.

---

4 091 459

Store comprising drivable domains

F. A. Kuijpers E
A. W. M van den Enden
A. H. A. M. Roovers

A storage device comprising a main magnetic field generator, a rotary magnetic field generator and a number of plates of magnetic material which are arranged transverse to the main magnetic field and in which magnetic domains can be formed and transported along a domain guide structure consisting of discrete elements. The domain guide structure comprises a switching device incorporating diverging switches which are arranged according to four orientations which correspond to corresponding orientations of discrete elements of said domain guide structure. The domain guide structures comprise diverging switches which may be arranged to be shifted with respect to each other over one or more discrete elements of the domain guide structure in different plates, each plate having a specific rotary magnetic field sequence.

---

4 093 504

Method for producing electrically conductive indium oxide patterns on an insulating support by etching with hydrochloric acid and ferric chloride

J. J. Ponjeé E
H. J. Feil

Etching indium patterns on insulating supports with the use of an etchant-resistant mask, by means of a solution of an acid to which ferric chloride has been added. Passivation of the bath is prevented by adding ferric chloride.

---

4 093 889

Low-pressure mercury vapor discharge lamp

J. Bloem E
A. Bouwknegt

A low-pressure mercury vapor discharge lamp, having a discharge space containing thermally emitting electrodes and a mercury amalgam which is composed of mercury, bismuth, tin and lead. This amalgam causes the mercury vapor pressure to remain stable at the value of 6×10^{-3} torr which is the optimum value for the conversion of electric energy into ultraviolet radiation over a wide temperature range. In addition, the mercury vapor pressure at room temperature is still sufficiently high to ensure rapid starting.

---

4 094 010

Optical multi-channel digital disc storage system

R. Pepperl H
J. Krüger
B. Hill

An optical storage system for the simultaneous or random storage of digital information in a plurality of channels, comprising a rotating storage disc which is scanned by a light beam. The digital information of each channel is stored in points along a spiral track comprising interruptions. A concentrical guide track is associated with each channel, the light beam being focussed onto both tracks of each channel for optical addressing. The light beams of all channels are controlled via deflection systems so that, during the interruptions, they are briefly directed onto the guide tracks while following the spiral tracks. From the radial deviation of the moving disc, measured by way of a detector system, control signals are derived by the deflection systems to compensate for undesired radial motions of the spiralshaped tracks.

---

4 094 013

Optical storage disk system with disk track guide sectors

B. Hill H
R. Pepperl
J. Krüger

A storage disk comprising at least one data track which is composed of spiral-shaped turns or concentrical turns, the data being arranged in the form of data blocks on the turns so that between the data blocks alternately a guide segment or a blank segment is present. The lengths of the unit, measured in angular degrees, consisting of data block and guide segment and the unit consisting of data block and blank segment are substantially the same and are chosen so that the same number of units is accommodated

in each turn. The spirals are sub-divided into sectors which contain only data blocks or only guide segments or blank segments. In each sector containing guide segments the radial distance between these segments is substantially twice the distance between the data blocks in each sector containing data blocks. The invention also relates to a device for writing and/or reading data of a rotating disk as described, comprising a light deflection system which notably includes a slow light deflector and a fast light deflector, the latter deflector scanning the guide segments.

---

4 094 689

Glass compositions

H. M. J. M. van Ass *E*
R. G. Gossink

Glass composition pairs which are suitable for processing into fibers with a gradient refractive index profile for use in telecommunication. They consist of GeO_2, B_2O_3/SiO_2, alkali oxide which in one composition consists of Li_2O and/or Na_2O and in the other composition of Na_2O and/or K_2O and other compatible oxides.

---

4 095 133

Field emission device

A. M. E. Hoeberechts *E*

A field emission device and method of forming same, comprising a substrate on which at least one conical electrode is provided, which substrate, with the exception of the proximity of the tip of the electrode, is covered with a layer of a dielectric material on which a conductive layer is present at least locally, in which in order to form an integrated accelerating electrode the conductive layer extensd in the direction of the punctiform tip of the electrode to beyond the dielectric layer and shows an aperture above the tip so that the conductive layer forms a cap-shaped accelerating electrode surrounding the conical electrode.

---

4 095 647

Heating device

G. A. A. Asselman *E*
J. W. J. M. van der Leegte

A heating device comprising a plurality of heating chambers, each chamber being bounded by the heat-transmission wall of a reservoir in which a heat transport medium is present which completes an evaporation/condensation cycle. The reservoirs are connected, via a common reservoir which also contains heat transport medium, to the same common heat source.

---

4 096 049

Method for producing an ion-selective electrode

R. E. van de Leest *E*
L. Heijne

The production of an ion-selective electrode for selective measuring the concentration of an ion in a solution of a mixture of ions, consisting of silver sulphide as matrix material, which is superficially converted into a hardly soluble substance which has one ion in common with the matrix material.

---

4 096 407

Television camera tube comprising mosaic of conductive regions separated by insulator from semiconductor plate

A. M. E. Hoeberechts *E*

A television camera tube having a semiconductor target which on the side to be scanned by an electron beam comprises a mosaic of electrically conductive regions each determining a picture element and separated from the semiconductor plate by an electrically insulating layer, and a resistive layer extending over the mosaic and the insulating layer which contacts the semiconductor surface via an aperture in the insulating layer present in each picture element and has a RC time between the scanning time of a picture element and the scanning time of the whole target.

---

4 096 461

Magnet system for tunable YIG oscillator and tunable YIG filter

P. Röschmann *H*

A magnet system for compensating the frequency difference between the oscillator frequency f_o of a YIG oscillator and the central filter frequency f_{m1} of a YIG filter to be independent of the oscillator frequency in which two pole-pieces are provided having opposing surfaces, one of which has a raised portion, and one of which has a plate of magnetic material having a saturation magnetization smaller than the smallest field strength desired. The raised surface portion of the pole-piece has a height at which the frequency f_o-f_{m1} is independent of f_o.

---

4 097 120

Liquid crystalline compounds and mixtures

J. van der Veen *E*
T. C. J. M. Hegge

Novel nematic liquid crystalline azo-compounds have been found, which are anisotropic over a wide temperature range which includes room temperature. The compounds are temperature, light and moisture proof. They exhibit no dynamic scattering. The compounds may be employed in E.S.R. and N.M.R. spectroscopy, and, as the case may be mixed with other nematic liquid crystalline compounds, in displays.

---

4 097 747

Device for measuring absorption of radiation in a slice of a body

G. Kowalsky *H*

In third-generation computer tomography apparatus measuring values must be intermediately stored partly for the total scanning time in order to enable processing of all measuring values with a given projection angle. This involves high expenditures for storage means. Moreover, the applied dose is larger than required for obtaining the measuring value per se. The invention consists in that there is provided a shield which moves with the radiator but whose direction does not change in space during the measurement, but does change relative to the radiation beam. As a result, it is achieved that all measuring values with a given projection angle can be obtained within a part of the scanning time, the said part corresponding to the ratio between the opening angle of the radiation beam and the total scanning time. As a result, the cost of intermediate storage as well as the patient dose can be reduced.

---

4 099 079

Secondary-emissive layers

A. G. Knapp *R*

A substrate bearing a secondary-emissive layer which consists of a cermet consisting of a readily evaporable metal, for example Au, Ag, Cu, Ni, Cr, Al or a nickel-chromium alloy, and an alkali metal aluminium fluoride, for example cryolite. The substrate material may be, for example mild steel, or a synthetic plastic material. It is possible to make large dynodes when using mild steel substrates which are much cheaper than silver-magnesium or beryllium-copper and the secondary-emissive layer does not require an activation treatment when incorporated in an electric discharge tube.

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79 No. 6

Galvanic effects in the wet-chemical etching of metal films

J. J. Kelly and G. J. Koel

The method of wet-chemical etching has been widely used for many years for making patterns in metal films. When a metal is etched in contact with another metal, galvanic effects can cause a considerable change in the etch rate. Control of these effects requires the correct choice of etching liquid. The article below describes how knowledge gained from electrochemical research is of value in finding appropriate combinations of metals and etchants.

Introduction

The wet-chemical etching of thin metal films is often an important process stage in the manufacture of integrated circuits. Because of increasing integration of the circuits the requirements that etching methods have to satisfy are becoming more and more demanding. This is particularly so for the accuracy of definition and the reproducibility of the patterns that have to be made in metal films. Etching of metals in solution is an electrochemical process. When a metal film covered with another metal has to be etched it is necessary to take account of the electrical contact between the two metals and between the metals and the etching solution [1]. Such a contact, forming a galvanic element, can significantly affect the etching behaviour [2–4]. The rate at which a particular metal dissolves in solution can for example be increased because of contact with a more noble metal. This can sometimes cause considerable problems when patterns have to be etched in metal films partly covered by a more noble metal. Because of the enhanced etch rate, removal of the metal film is not limited to the unprotected areas, but the film is also partly etched *underneath* the noble metal.

'Undercutting' of this nature occurs when a molybdenum film partly covered with gold is etched [2]. Molybdenum is often used to provide adhesion between another metal, such as gold, and a substrate. In an aqueous solution of cerium ammonium nitrate ($Ce(NH_4)_2(NO_3)_6$) and nitric acid (HNO_3) an unprotected molybdenum film dissolves at a rate of about 0.1 μm per minute. If the film is locally covered with a photoresist, hardly any undercutting can be seen after 15 seconds of etching. A completely different picture is obtained, though, if the molybdenum film is locally covered with gold, which is not dissolved by this etchant. The rate at which the molybdenum then dissolves can be increased a thousand times or more by the galvanic contact with gold. As *fig. 1* shows, this results in considerable undercutting — generally an unwanted effect. After 15 seconds of etching the undercutting may amount to as much as 80 μm. Although this is an extreme example from many cases of under-

Dr J. J. Kelly and Dr G. J. Koel are with Philips Research Laboratories, Eindhoven.

[1] In this article we shall not discuss effects that can arise at other electrical contacts during wet-chemical etching, such as those between a metal and a semiconductor, or between two different types of semiconductor.

[2] J. J. Kelly and C. H. de Minjer, J. Electrochem. Soc. **122**, 931, 1975.

[3] L. T. Romankiw, Proc. Symp. on Etching for pattern definition, Washington D.C. 1976 (Spring Meeting Electrochem. Soc.), p. 161.

[4] D. MacArthur, Proc. Symp. on Etching for pattern definition, Washington D.C. 1976 (Spring Meeting Electrochem. Soc.), p. 76.

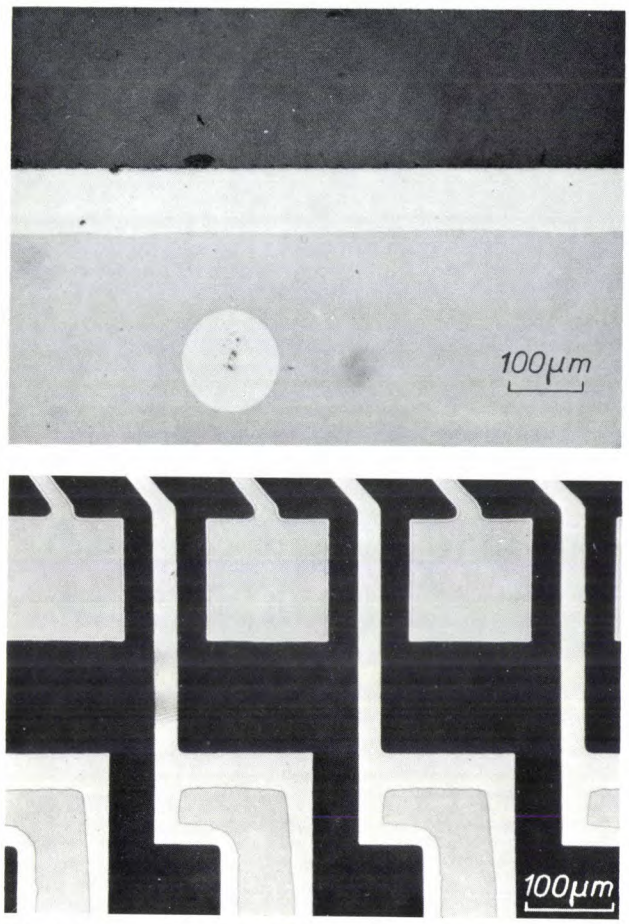


Fig. 1. Photographs showing the undercutting of molybdenum caused by galvanic contact with gold. The upper photograph shows a molybdenum film partly covered with gold that was etched for 15 seconds in a solution of 22 wt% of $Ce(NH_4)_2(NO_3)_6$ in 1% HNO_3; this 'duplex' film was deposited on a glass substrate. To demonstrate the undercutting clearly, the photograph was taken through the glass. In the black area molybdenum was not covered by gold and has been completely removed. The dark-grey area corresponds to unetched molybdenum under gold. The light-grey central area corresponds to molybdenum that has been removed in spite of the gold protecting layer ('undercutting'). The circular light-grey region was produced by undercutting at a pinhole in the gold film. The lower photograph shows the undercutting in part of an iron-garnet magnetic-bubble memory circuit. Here again, in addition to the unprotected molybdenum (*black*), some of the gold-covered molybdenum has been etched (*light grey*). The three square blocks (*dark grey*) show much less undercutting, since the gold area with which they were in contact during etching was relatively small.

cutting [2,3], it does however give an idea of the considerable problems that can arise in etching two-layer metallic films ('duplex' films).

Changes in etching behaviour as a result of galvanic-element formation can also be important in the manufacture of multilayer integrated circuits. When a metal film in which a pattern has been chemically etched has to be uniformly covered with a second layer by evaporation or sputtering, the coverage may not be satisfactory at the steep edges produced by the etching [5,6]. This is one of the problems that arise in the manufacture of thin-film magnetic heads consisting of a nickel-iron alloy (Permalloy), silicon oxide and gold representing the magnetic material, insulation, and electrical conductor respectively. In its simplest form such a magnetic head contains two Permalloy films one above the other with the silicon-oxide and gold films between them; see *fig. 2*. The two Permalloy films are connected through a window in the silicon oxide, with the upper layer of Permalloy 'wrapped around' the gold film. When the upper Permalloy film is deposited the steep etched edges in the gold and silicon-oxide films can cause an uneven coverage. This in turn can result in breaks in the Permalloy film at the edges, as shown in *fig. 3*.

The problem of steep edges in a particular film can be largely avoided by making bevelled or sloped steps. This is done by applying a thin film that dissolves more rapidly in the etchant [5,7,8]. As the etch rate of this top layer is increased, the edges become less steep. If a metal is used as the top layer for bevel-etching another metal, then once again galvanic interaction between two different metals and an etching solution must be considered.

The objective of our investigations was to find answers to the problems caused by galvanic effects in wet-chemical etching. Since the dissolution of a metal in a solution is an electrochemical process, we used an electrochemical approach to determine how the etching behaviour is changed by the formation of a galvanic element. In particular, we examined the influence of the composition of the etching solution on the galvanic effect. The information gained in this way has enabled us to find a number of etchants that signify a considerable improvement for the applications we have mentioned.

In this article we shall first of all look at the electrochemistry of the galvanic-element formation during the etching. Next, we shall show how the undercutting can be explained in electrochemical terms, and how the effect can be prevented by using a suitable etchant. Then we shall consider the influence of the galvanic effect in producing bevelled edges. Here we shall show how the desired bevelling angle can be obtained by an appropriate choice of top metallic layer and etchant.

Electrochemistry of galvanic-element formation during etching

When a metal is chemically dissolved in a solution oxidation and reduction reactions take place. The metal atoms are oxidized while the oxidizing agent of the etchant is reduced at the surface of the metal. This means that there is a transfer of electrons at this surface. The total reaction can be considered as the sum of anodic and cathodic partial processes. The rates of

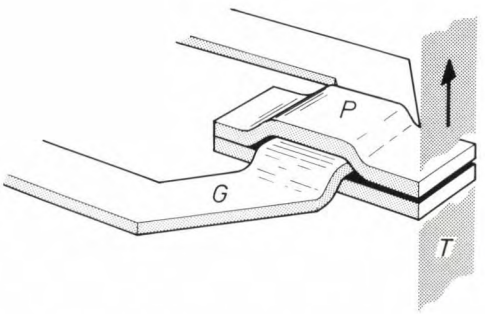


Fig. 2. Diagram of a thin-film magnetic head. *P* magnetic circuit of Permalloy. *G* gold turn. *T* magnetic tape. A silicon-oxide layer separates the gold and the Permalloy. The two Permalloy films are connected through a window in the oxide.

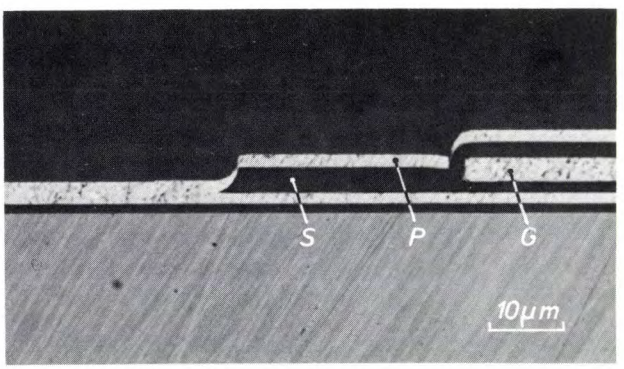


Fig. 3. Photomicrograph of part of the cross-section of a thin-film magnetic head. The breaks in the Permalloy film *P* are due to incomplete coverage of the steep etched edges in the gold conductor *G* and the silicon-oxide layer *S*.

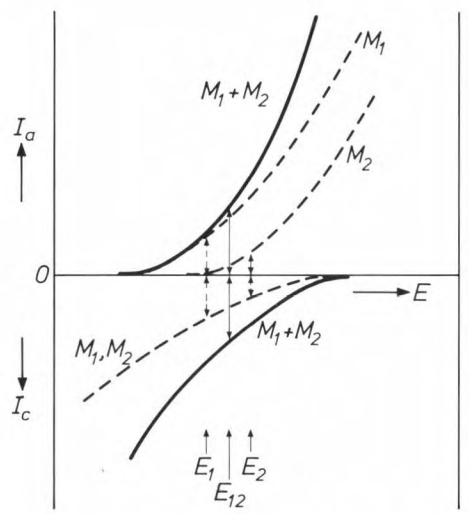


Fig. 4. Partial polarization curves for the dissolution of two different metals M_1 and M_2 in an acidic solution. I_a is the anodic partial current and I_c the cathodic partial current; both are shown as a function of the potential E. During chemical dissolution of the separate metals no external current flows, so that at each mixed potential (E_1 or E_2) I_a is equal to I_c. The etch rate is determined by the value of I_a at the mixed potential. Because of galvanic contact between M_1 and M_2 the mixed potential assumes a new value E_{12}, situated between E_1 and E_2. For the less noble metal this means that there is an increase in the anodic current at the mixed potential, and hence an increase in the etch rate.

both processes depend on the electrode potential of the metal. The effect of this potential on the total reaction can be studied by measuring the cell current as a function of the applied potential, which is varied by means of an external voltage source. According to the 'mixed-potential theory' the total 'polarization curve' thus obtained is equal to the sum of the curves for the anodic and cathodic partial processes [9].

The effect of the formation of a galvanic element on the dissolution of the different metals can be explained by this theory. *Fig. 4* shows the anodic and cathodic partial curves for the dissolution of two arbitrary metals M_1 and M_2 in an acidic solution. To keep the treatment simple it has been assumed that only the following anodic reactions take place:

$$M_1 \rightarrow M_1^{n+} + n\,e\,,$$
$$M_2 \rightarrow M_2^{n+} + n\,e\,.$$

In the cathodic reaction, hydrogen ions are reduced and hydrogen gas is evolved in the reaction

$$2H^+ + 2e \rightarrow H_2\uparrow\,,$$

which can occur at the surface of both metals. The anodic current I_a and the cathodic current I_c both depend on the potential E. We assume that the Tafel equation [9] holds for both reactions:

$$I_a = I_a^0 \exp(b_a E), \qquad (1)$$

$$I_c = I_c^0 \exp(-b_c E), \qquad (2)$$

where I_a^0 and I_c^0 are the currents at $E = 0$ and b_a and b_c are constants. We assume that the cathodic partial curves for the two metals are the same, that the areas of M_1 and M_2 exposed to the solution are equal and that the electrical resistance of the solution is negligibly small. Since no external current flows during chemical dissolution, the anodic partial current must be equal to the cathodic partial current. This determines the value of the mixed potential at which the oxidation and reduction reactions occur. The mixed potentials of the two metals are indicated by E_1 and E_2 in fig. 4. The rate at which a metal dissolves is proportional to the anodic partial current at the mixed potential.

[5] J. J. Kelly and G. J. Koel, J. Electrochem. Soc. **125**, 860, 1978. J. J. Kelly, to be published shortly.
[6] J. L. Vossen, G. L. Schnable and W. Kern, J. Vac. Sci. Technol. **11**, 60, 1974.
[7] T. Yanagawa and I. Takekoshi, IEEE Trans. **ED-17**, 964, 1970.
[8] G. J. Koel and T. C. J. M. Bertens, Extended Abstracts Electrochem. Soc. **74-1** (Spring Meeting, San Francisco 1974), p. 205.
[9] K. J. Vetter, Electrochemical kinetics, theoretical and experimental aspects, Academic Press, New York 1967, pp. 732-747.

When M_1 and M_2 are electrically connected the combined anodic curves and the combined cathodic curves have to be considered. As a result of the contact the mixed potential of the bimetallic system assumes a new value E_{12}, which is higher than E_1 and lower than E_2. From equation (1) it is clear that the contact between M_1 and M_2 has caused an increase in the etch rate of the less noble metal M_1, and a decrease in the etch rate of the more noble metal M_2. On the other hand equation (2) shows that the hydrogen production at M_1 decreases, while that at M_2 increases. In the formation of the galvanic element, M_1 therefore functions as local anode and M_2 as local cathode.

Undercutting

To explain the undercutting shown in fig. 1, we shall consider the electrochemical behaviour of molybdenum in the $Ce(NH_4)_2(NO_3)_6$-HNO_3 solution mentioned earlier. Curve *a* of *fig. 5* is the measured polarization curve for a rotating molybdenum electrode in this solution. In the anodic region the current increases rapidly with increasing potential. In the cathodic region, on the other hand, the current is constant over a wide range of potential; the current density here has a value of about 90 mA/cm². The cathodic current is determined by the rate at which the Ce^{IV} ions are transferred from the solution to the molybdenum surface. As soon as a Ce^{IV} ion arrives at this surface it is immediately reduced in the reaction

$$Ce^{IV} + e \rightarrow Ce^{III}.$$

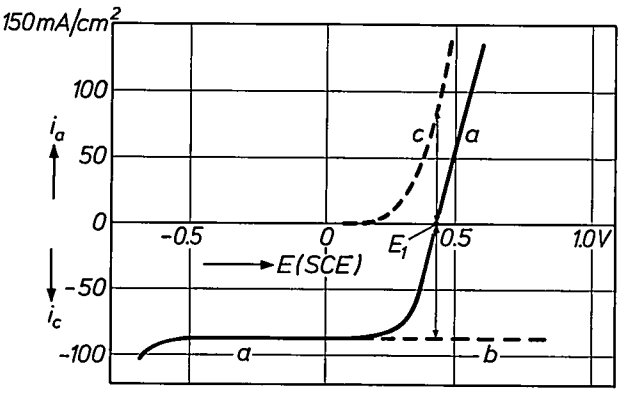


Fig. 5. Polarization curves for a molybdenum electrode rotating at a rate of 500 revolutions per minute in the $Ce(NH_4)_2(NO_3)_6$-HNO_3 solution of fig. 1. The densities i_a of the anodic current and i_c of the cathodic current are plotted against the applied potential *E(SCE)* with respect to a saturated calomel electrode. Curve *a* is the total polarization curve. Curve *b* is the cathodic partial curve, which corresponds to a part of the polarization curve for a platinum electrode rotating at the same rate. The constant cathodic current indicates that the reduction rate of Ce^{IV} ions is determined by mass transfer. The difference between curves *a* and *b* gives curve *c*, the anodic partial curve for the dissolution of molybdenum. It can be seen here that the anodic current increases very rapidly with potential. An increase in the mixed potential E_1, because of galvanic-element formation, can therefore produce a very large increase in the etch rate.

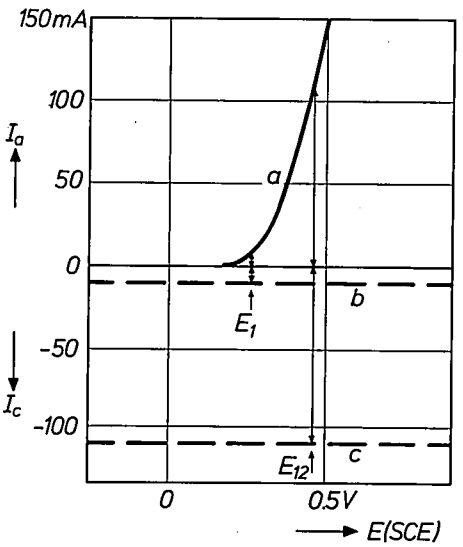


Fig. 6. Polarization curves for a stationary molybdenum electrode in the $Ce(NH_4)_2(NO_3)_6$-HNO_3 solution. Curves *a* and *b* are the anodic and cathodic partial curves for Mo. Electrical contact with gold gives rise to a much larger cathodic current; curve *c* relates to the case where the area of the gold is about ten times that of the molybdenum. Because of this contact the mixed potential E_1 changes to the higher value E_{12}. This considerably increases the anodic current of the molybdenum, with the result that the etch rate is increased.

The same cathodic current, determined by mass transfer, is observed when the molybdenum electrode is replaced by a rotating platinum electrode; part of the polarization curve for Pt is shown in curve *b*. The limiting cathodic current for both metals is proportional to the Ce^{IV} concentration and to the square root of the rate of rotation of the electrode. This is in agreement with the theory that has been developed for electrochemical reactions whose rate is determined by mass transfer at rotating electrodes [10].

The anodic partial curve for molybdenum is obtained by subtracting the cathodic current (curve *b*) from the total current (curve *a*) at each value of the potential. The resultant curve is shown in curve *c* of fig. 5. The anodic partial curve increases very steeply for potential values above 0.2 V, while the cathodic current in this range is practically constant. This implies that the rate of dissolution of molybdenum in this solution is entirely determined by the cathodic current resulting from reduction of Ce^{IV} ions. When the molybdenum electrode is stationary in the solution, transfer of the Ce^{IV} ions to the molybdenum surface is much slower. The cathodic current density is then equal to about 10 mA/cm². This must also be the value of the anodic current density as a result of the oxidation of molybdenum during wet-chemical etching of the metal. The etch rate of molybdenum can be calculated from this value, with the assumption that three electrons are necessary to dissolve a single molybdenum atom. There is good agreement between the calculated etch rate and the measured rate of 0.1 μm per minute.

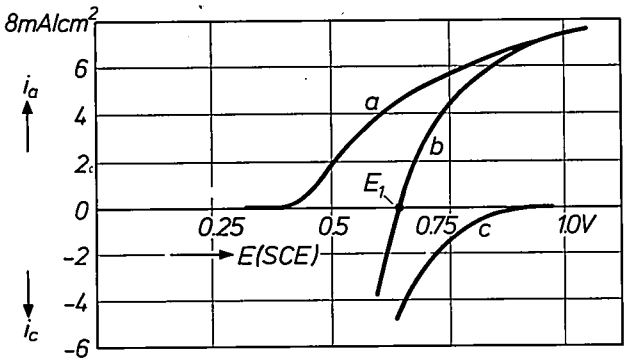


Fig. 7. Polarization curves for a stationary molybdenum electrode in a solution consisting of 70 ml of 85% H_3PO_4, 15 ml of CH_3COOH and 0.8 ml of H_2O. The solution was agitated by a magnetic stirrer during the measurements. Curve *a* is the anodic partial curve. Above about 0.6 V the anodic current density does not increase very rapidly with rising potential. Curve *b* is the total polarization curve for the electrode in the solution after 0.3 ml of the oxidizing agent HNO_3 has been added. The mixed potential E_1 is in the range where the anodic current density does not increase rapidly with potential. Curve *c*, the difference between curves *a* and *b*, gives the cathodic partial current density resulting from the reduction of HNO_3. An increase in the mixed potential because of a galvanic contact does not greatly affect the etch rate, because the anodic current density is not very dependent on potential and the mixed potential cannot rise above 0.95 V.

What happens when the molybdenum is in contact with gold during the wet-chemical etching? Because it is a noble metal the gold itself is not attacked by this etchant. However, a reduction of the Ce^{IV} ions takes place at the surface of the gold, and this causes an increase in the cathodic partial current; the greater the area of gold in contact with the molybdenum, the greater is the increase. In the example of *fig. 6* the area of gold is about ten times that of molybdenum, so that the cathodic current increases from 10 mA to 110 mA. Since anodic and cathodic currents are equal at the mixed potential, the same increase in the anodic current is found, and hence a corresponding increase is observed in the etch rate of the molybdenum. The result of this is an appreciable undercutting, as we saw in fig. 1.

An analogous description can be given for many other cases of accelerated etching, such as the undercutting of an aluminium film that is partially covered with a platinum-gold film, in a solution of sodium hydroxide (NaOH) [2], and the undercutting of an aluminium film in contact with a copper film, in a solution of ferric chloride ($FeCl_3$) and hydrochloric acid (HCl) [3].

Reduction of undercutting

As fig. 4 shows, the increase in the etch rate of a metal as a result of a galvanic contact with a more noble metal is normally caused by a shift of the mixed potential to a higher value. The increase in the etch rate depends on the rate at which the anodic partial current of the metal increases with increasing potential. This suggested that we should try to find etchants in which the anodic partial current of the metal is sufficiently independent of potential so that any change in the mixed potential of the system would not significantly affect the etch rate. Polarization behaviour of this nature is mainly found in solutions used for electrolytic polishing of metals [11]. When the potential is increased in such systems above open-circuit potential a normal increase in the anodic current is observed at first. Above a certain value of the potential, however, hardly any further increase in the anodic current is found. This is because the metal has become coated with a viscous or solid layer of an oxide or a salt [11].

The current density during electrolytic polishing of a metal such as molybdenum is relatively high in most polishing solutions; about 100-200 mA/cm². In general these solutions are not so suitable for the etching of thin molybdenum films, as the etch rate is too high. A solution of phosphoric acid (H_3PO_4) and acetic acid (CH_3COOH) gives polishing-type action even at relatively low current densities (<10 mA/cm²) and can therefore be used to reduce undercutting in thin molybdenum films. Curve *a* of *fig. 7* shows the anodic polarization curve of a molybdenum electrode in this medium. The anodic current increases only slightly with increasing potential above about 0.6 V. To allow this solution to be used as a chemical etchant an oxidizing agent must be added that has a redox potential in the range in which the anodic current does not increase rapidly. Nitric acid is suitable in this case. The total polarization curve for molybdenum in an H_3PO_4-CH_3COOH-HNO_3 solution is shown in curve *b* of fig. 7. The mixed potential E_1 of the metal in this medium is 0.65 V with respect to the saturated calomel electrode and is therefore in the range where the anodic current is not strongly dependent on potential. The difference between curves *a* and *b* gives the partial curve for the cathodic reaction resulting from reduction of HNO_3 (curve *c*).

If the cathode surface area is increased by combining the molybdenum with a noble metal (e.g. gold or platinum), the HNO_3 is also reduced at this second metal. This shifts the mixed potential to a more positive value, but because of the stabilization of the anodic partial current the etch rate of the molybdenum does not change significantly. Also, the mixed

[10] D. P. Gregory and A. C. Riddiford, J. Chem. Soc. 1956, 3756.

[11] P. V. Shchigolev, Electrolytic and chemical polishing of metals, Freund Publishing House, Holon, Israel, 1970.
W. J. McG. Tegart, The electrolytic and chemical polishing of metals in research and industry, 2nd edition, Pergamon Press, London 1959.

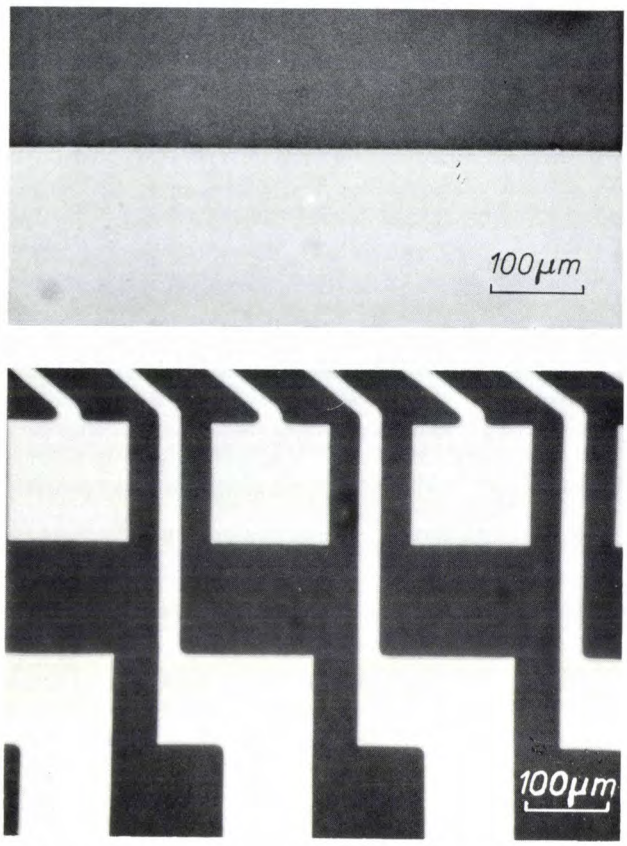


Fig. 8. Photographs showing molybdenum that has been etched in contact with gold without serious undercutting. The upper photograph was taken after etching for 15 seconds in the H_3PO_4-CH_3COOH-HNO_3 solution. The molybdenum not covered with gold has been removed (*black*), while the molybdenum protected by gold has not been seriously undercut (*grey*). The same applies for the lower photograph, which shows part of a circuit that has been treated with the same etchant. This photograph and the lower photograph of fig. 1 were made by K. P. Schmidt of Philips GmbH Forschungslaboratorium, Hamburg.

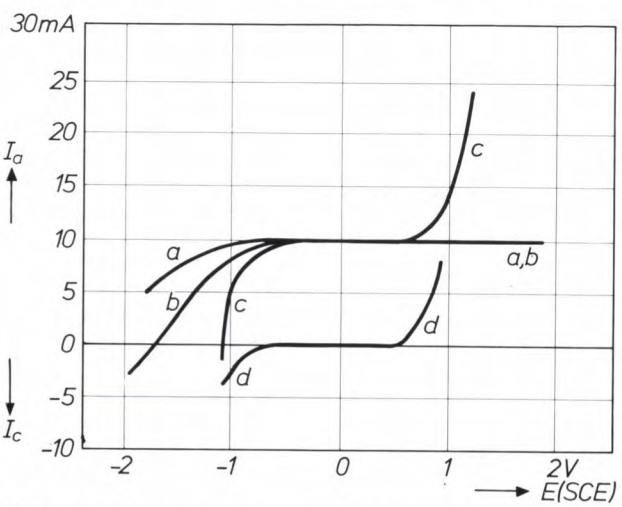


Fig. 9. Polarization curves measured for a solution of 150 g of Na_2CO_3 and 70 g of $Na_3PO_4.12H_2O$ in 1 litre of water. Curve *a* is the anodic partial curve for an aluminium electrode 1 cm² in area rotated at a rate of 550 revolutions per minute. Curve *b* is the total polarization curve for this electrode. Curve *c* is the total polarization curve for the same electrode, but now in contact with a platinum electrode 15 cm² in area. The anodic current of the aluminium is not affected by the aluminium-platinum contact over a wide range of potential. Curve *d* is the polarization curve for the platinum electrode without aluminium.

potential of the molybdenum in this medium cannot rise above 0.95 V, the highest potential at which HNO_3 can be reduced. This again helps to ensure that the increase in the etch rate remains within the permitted bounds. Undercutting of the molybdenum in this solution is therefore much less pronounced than in a $Ce(NH_4)_2(NO_3)_6$-HNO_3 solution; see *fig. 8*.

The undercutting of an aluminium film partly covered with platinum can be reduced in the same way. A suitable etchant in this case is a basic solution of sodium carbonate (Na_2CO_3), sodium phosphate (Na_3PO_4) and potassium ferricyanide ($K_3Fe(CN)_6$). A solution of the first two salts gives an electrolytic polishing action, while the third salt acts as an oxidizing agent. Curve *a* of *fig. 9* shows the anodic partial current corresponding to the dissolution of aluminium in a basic medium in the reaction

$$Al + 4OH^- \rightarrow AlO_2^- + 2H_2O + 3e.$$

The curve was obtained by determining the potential dependence of the dissolution rate of Al; the amount of dissolved metal was determined spectrophotometrically. At potentials greater than −0.8 V a constant anodic current is found. This partial current is identical to the total anodic current, shown in curve *b*; in this range no cathodic reaction takes place. This is however not the case at potentials lower than −0.8 V. Here hydrogen is cathodically evolved in the reaction

$$2H_2O + 2e \rightarrow 2OH^- + H_2\uparrow.$$

Because of this reaction the total current is decreased.

When aluminium is brought into contact with platinum in this solution no change is found in the total current over a wide potential range (−0.8 V to 0.5 V); see curve *c* of fig. 9. Outside this range the contact with platinum does have an effect on the current. The sharper decrease in the current at potentials below −0.8 V is the result of hydrogen evolution at the platinum. The increase in the current at potentials above 0.5 V is due to oxygen evolution at the platinum electrode:

$$4OH^- \rightarrow 2H_2O + O_2\uparrow + 4e.$$

In a solution of Na_2CO_3 and Na_3PO_4 to which $K_3Fe(CN)_6$ has been added aluminium is oxidized in the reaction

$$Al + 4OH^- + 3Fe(CN)_6^{3-} \rightarrow AlO_2^- + 2H_2O + 3Fe(CN)_6^{4-}.$$

In this reaction the trivalent ferricyanide complex is reduced to the divalent ferrocyanide complex. The mixed potential of the reaction in this medium lies in the potential range where the anodic current is constant. We may therefore expect that the rate of dissolution of the aluminium will not change when contact with platinum is made. The experimental confirmation of this is given in *fig. 10*, which shows the number

of moles of aluminium dissolved as a function of the etching time. The measured points for aluminium in contact with platinum are found to lie along the same line as those obtained with aluminium on its own. The current density derived from the slope of this line corresponds to the saturation current density of about 10 mA/cm² observed in the polarization experiments (fig. 9). Since the galvanic contact does not produce any change in the etch rate in this medium, there is hardly any undercutting when an aluminium film partly covered with platinum is etched; see *fig. 11*.

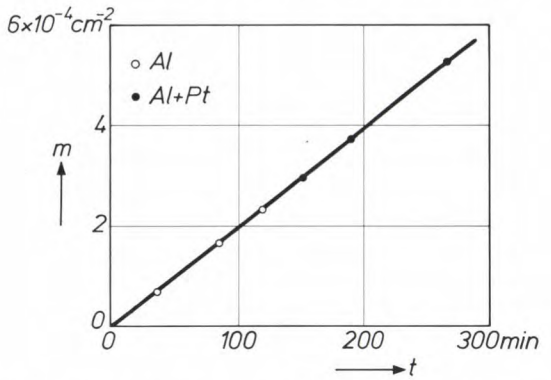


Fig. 10. The number of moles of aluminium m etched as a function of time in the Na_2CO_3-Na_3PO_4 solution of fig. 9 to which 32 g of $K_3Fe(CN)_6$ has been added. The measurements were made at 24 °C with an aluminium electrode 1 cm² in area rotating at 550 revolutions per minute without contact to platinum (*open circles*) and with contact to a platinum strip 15 cm² in area (*points*). The contact in this solution and under these conditions was found to have no effect on the etch rate of the aluminium. In an NaOH solution, on the other hand, the etch rate of the aluminium increased as a result of the contact with platinum.

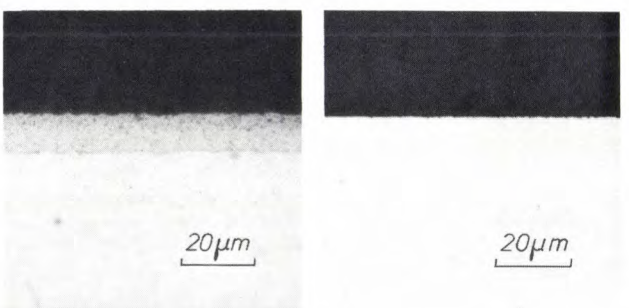


Fig. 11. Photographs made after etching an aluminium film partly covered with platinum and gold. The photograph on the left was taken after etching for three minutes in a 0.25 mol/l NaOH solution. In addition to the non-covered aluminium (*black*), some of the covered aluminium (*dark grey*) has been removed. This undercutting is not found when the platinum-gold film is replaced by photoresist. The photograph on the right was taken after etching an aluminium/platinum-gold film for four minutes in the Na_2CO_3-Na_3PO_4-$K_3Fe(CN)_6$ solution of fig. 10. In this etchant hardly any undercutting of the aluminium is observed, despite the presence of the noble-metal top layer.

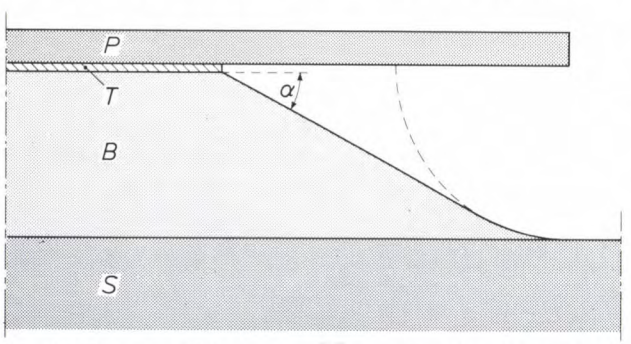


Fig. 12. Diagram of the undercutting profile of a base layer B with a top layer T with a higher etch rate. P photoresist. S substrate. The dashed line indicates the etching profile when the top layer is not present. As the etch rate of the top layer is increased, the bevelling angle α decreases.

Making bevelled edges

Fig. 12 shows diagrammatically how bevelled edges can be made in a film by providing it with a thin top layer that dissolves more rapidly in the etching solution. It can be shown that the bevelling angle α is given by

$$\sin \alpha = r_B/r_T, \qquad (3)$$

where r_B is the etch rate for the base layer and r_T is the etch rate for the top layer. As a rule the top layer is thinner than 100 nm, while the lower layer may have a value of between 2 and 8 μm. When a second metal film is used as top layer for bevel-etching another metal film, the formation of a galvanic element produces a mixed potential that differs from the potentials of the separate metals. This means that the etch rates of both metals may change. The magnitude of this change depends on the polarization behaviour of the two metals in the etchant (see fig. 4) and on the relative areas exposed to the electrolyte. Since the upper layer has a much smaller area than the base layer, the mixed potential is normally determined by the base layer. The etch rate of the base layer does not therefore differ very much from that of the separate metal. On the other hand, the mixed potential of the upper layer can vary because of galvanic contact with the base layer. If the metal of the upper layer is less noble than that of the base layer, this potential assumes a higher value, as we saw for molybdenum in contact with gold (fig. 6).

Since in many cases the etch rate is an exponential function of potential (eq. (1)), the change in potential can cause such a large increase in the etch rate of the top layer that the system becomes impracticable. It is therefore necessary to choose the top layer and the etchant so that an unacceptable acceleration of the etch rate as a result of galvanic effects does not occur.

Obtaining the desired bevelling angle

For making bevelled edges in a metal by using another, less noble, metal as top layer, an etchant can be used in which the anodic partial current of the top layer is almost independent of potential. This ensures

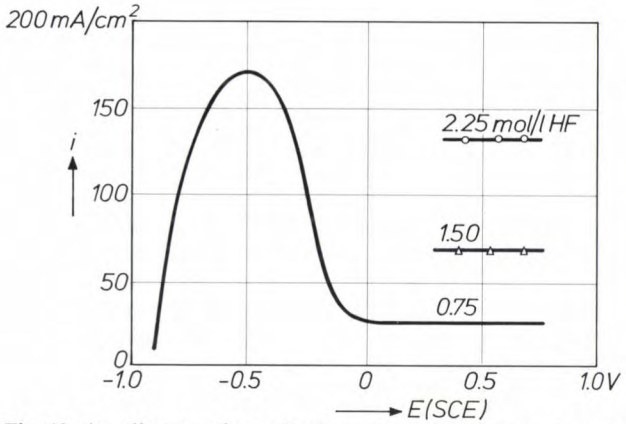


Fig. 13. Anodic curve for a titanium electrode rotating at a rate of 500 revolutions per minute in a solution of 3.9 mol/l H_2SO_4 and 0.75 mol/l HF. The anodic current density is virtually constant over a wide potential range. In this range the value of the anodic current density increases considerably as the HF concentration is increased to 1.50 mol/l or 2.25 mol/l.

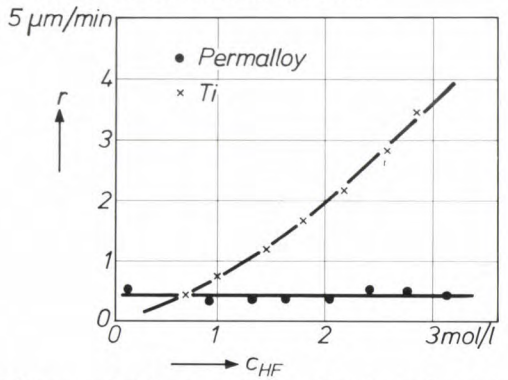


Fig. 14. Measured etch rate r of Permalloy (*points*) and the etch rate of titanium (*crosses*) calculated from the current density in the plateau region of fig. 13, as a function of the HF concentration c_{HF} in a solution of 3.9 mol/l H_2SO_4 and 1.12 mol/l H_2O_2. The etch rate of titanium increases with increasing HF concentration, whereas that of Permalloy is virtually independent of HF concentration.

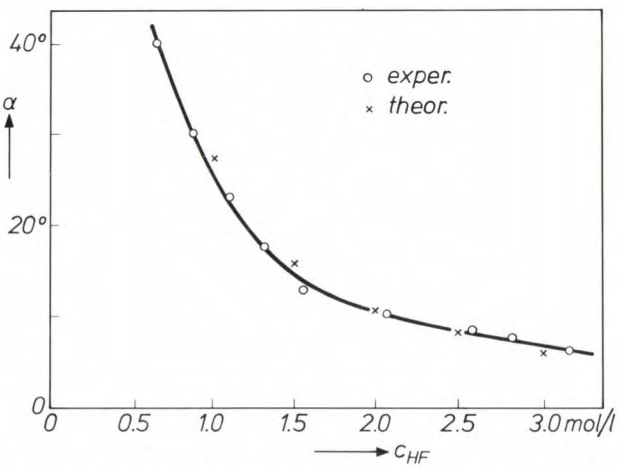


Fig. 15. Bevelling angle α of a Permalloy film obtained with titanium as a top layer, as a function of HF concentration in the H_2SO_4-HF-H_2O_2 solution. The experimental values (*open circles*) agree well with the values calculated (*crosses*) from the etching rates (fig. 14).

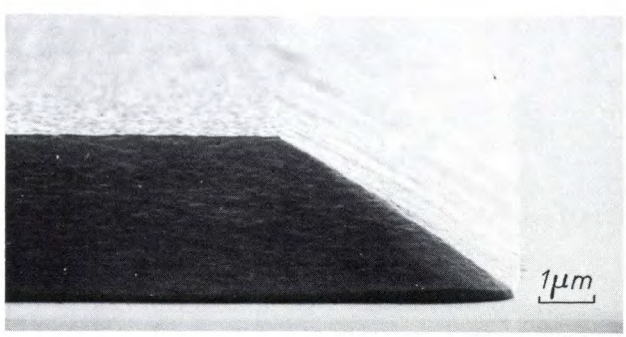
▷

Fig. 16. Photograph of a bevelled slope in a 3 μm thick Permalloy film. The photograph was made with a scanning electron microscope after the photoresist had been removed. The bevelling angle α is about 30°.

that the etch rate of the top layer does not increase significantly because of galvanic contact with the base layer. A reliable etching system is then obtained that will give the desired etching angle α, as determined by eq. (3).

For making bevelled edges in sputtered Permalloy films we have used titanium as the top layer and an etching solution of sulphuric acid (H_2SO_4) and hydrofluoric acid (HF). *Fig. 13* shows the polarization curve for a rotating titanium electrode in an H_2SO_4-HF solution. The anodic current density has a maximum at about −0.5 V. At potentials higher than 0 V the anodic current is virtually independent of the potential. The addition of hydrogen peroxide as an oxidizing agent has hardly any effect on the anodic current in this potential range. This implies that there is virtually no reduction of the peroxide, no doubt because of the poor electronic conductivity of the passive layer on titanium [5,9]. A change in the HF concentration in the solution has a marked effect on the anodic current in this range. As fig. 13 shows, an increase in concentration to 1.50 mol/l or 2.25 mol/l gives a considerable increase in the anodic current, which, however, remains independent of potential.

When an H_2SO_4-HF-H_2O_2 solution is used for etching a thin titanium film on a thicker Permalloy film, the potential of the titanium is the same as that of the Permalloy. This is dependent on the HF concentration, but for all concentrations between 0.2 and 4 mol/l it lies in the range where the anodic current is independent of the potential. The etch rate of titanium is therefore constant in this range. If the HF concentration is increased, the etching rate increases, as a result of the marked increase in the potential-independent anodic current. *Fig. 14* shows the etch rate of titanium calculated from the current in the plateau region of fig. 13 and the measured etch rate of Permalloy as a function of the HF concentration. The etch rate of Permalloy is virtually independent of the HF concentration. The etching angle α can be calculated for various HF concentrations by substituting these etch rates in eq. (3). As *fig. 15* shows, the values of α calculated in this way agree well with the experi-



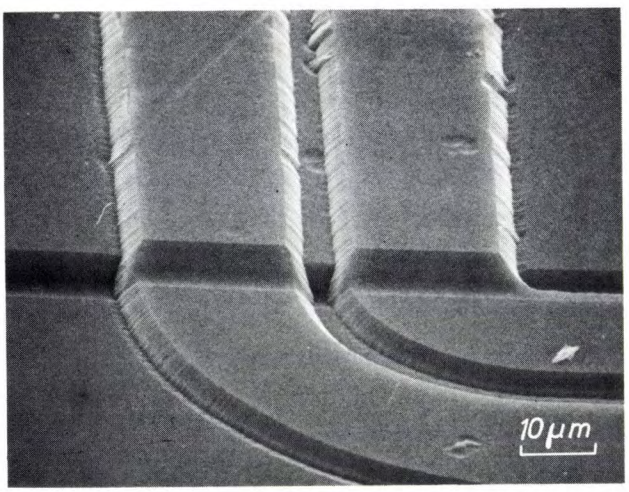


Fig. 17. Scanning-electron-microscope photograph of part of a thin-film magnetic head with multiple turns. The two adjacent gold strips cross the sloping edge of the Permalloy film without any breaks. The gold and the Permalloy are separated by an insulating layer of silicon oxide.

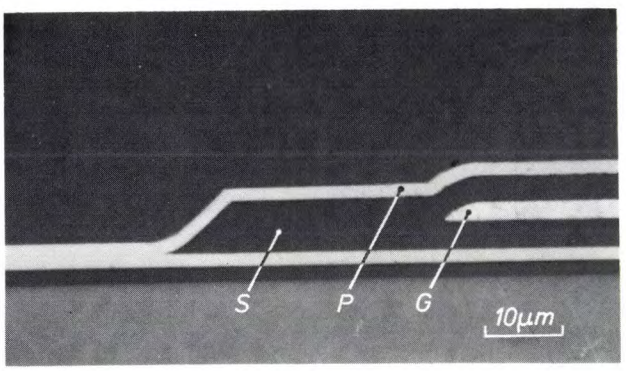


Fig. 18. Photomicrograph of the same part of a thin-film magnetic head as shown in fig. 3. The Permalloy film *P* is now uniformly covered because of the bevelled edges in the gold *G* and silicon oxide *S*.

mentally determined values. This confirms the validity of the electrochemical approach.

By changing the HF concentration in the etching solution the bevelling angle can be varied from 6° to 40°. *Fig. 16* is a scanning-electron-microscope photograph of a 3 μm Permalloy film covered by a 0.05 μm titanium layer that has been etched for 3 minutes in an H_2SO_4-HF-H_2O_2 solution. The bevelling angle is about 30°.

Fig. 17 shows how two adjacent strips of gold film cross the edge of a Permalloy film. The sloping edge of the Permalloy ensures a uniform coverage of gold, and undesirable breaks are avoided.

Breaks in the upper Permalloy layer of a magnetic head (fig. 2) can be prevented in the same way, by first making bevelled edges of a suitable angle in the gold and silicon-oxide layers. *Fig. 18* shows the same part of the magnetic head as fig. 3, but now with the bevelled edges allowing a uniform coverage of Permalloy.

Summary. For many applications in semiconductor technology, patterns are etched in a metal film (e.g. of molybdenum) that is partially covered with a more noble metal (e.g. gold). Galvanic effects arise in this process, and these can sometimes give a considerable increase in the etch rate. This can lead to undesirable undercutting, in which not only the uncovered metal is etched but also a large part of the covered metal. An investigation of the electrochemical properties of such systems has provided information about etchants that should be used to reduce undercutting as much as possible. These solutions are very similar to those used in the electrolytic polishing of metals. Galvanic effects can also play a part in the formation of bevelled edges in a metal film (e.g. Permalloy); such edges are often necessary to permit a subsequent uniform coverage with another material. These edges can be bevelled by first applying a thin film of a less noble metal (e.g. titanium). Electrochemical studies have shown how the desired bevelling angle can be obtained by using an etchant of appropriate composition.

Philips tech. Rev. 38, 158-159, 1978/79, No. 6

Experimental semi-automatic machine for hot splicing glass fibres for optical communication

The application of glass fibres for optical communication [1] requires a completely new technology for joining them together. Experience has shown that in laying or repairing glass-fibre cables the use of connectors at the joints should be avoided as far as possible, largely because they are complicated and introduce losses. It is much better to fuse the ends together directly ('hot splicing'), rather like soldering wires together.

Such splices will often have to be made in a 'hostile' environment, e.g. in a narrow, dusty manhole close to city traffic. Even though the quality of the join must be good, the splicing procedure must be fairly simple, so that the present technicians will not need extensive retraining.

Fig. 1 shows an experimental hot-splicing machine made at Philips Research Laboratories, Eindhoven. The splicing is carried out by heating from an arc discharge. In the machine shown the splicing operation can be easily observed by means of a 'stereogram' projected on a screen at about 60× magnification (fig. 1, upper left). This stereogram is produced by a pair of halogen lamps and various optical components that give images of the splice as seen from two perpendicular directions.

The two glass fibres to be spliced are clamped in two V-shaped grooves. The lamp housing is swung aside to give access to the grooves; see *fig. 2*. For the best results the end faces of the fibres should be reasonably perpendicular to the axis [2].

The first step is to set the two fibre ends *approximately* to a predetermined spacing, with the gap shown near the middle of the screen. Axial displacements (in the z-direction) are made by turning two knobs; the right-hand fibre can be moved transversely (in the x,y-plane) by operating the two levers of a micromanipulator.

When the two fibres have been lined up approximately, the arc discharge can be ignited by pressing the ignition button. First, however, the lamp housing must be swung back into the position shown in fig. 1; this closes a safety contact. Pressing the ignition button moves the front electrode towards the one at the back for a moment. The next operation is automatic, and is set in motion by pressing the starter button. The arc then moves upwards until it is just below the axis of the glass fibres. Any protective plastic coating on the fibres is then burnt away.

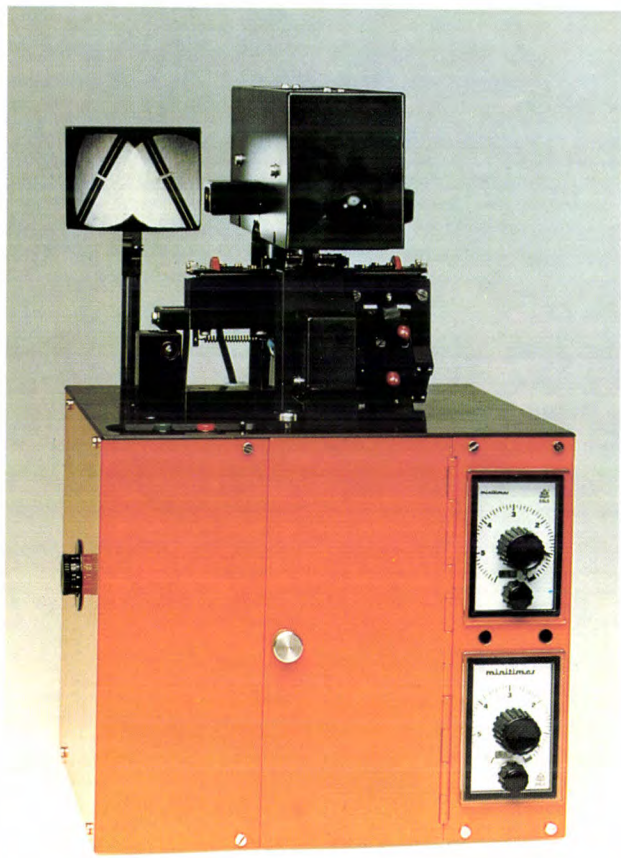


Fig. 1. The machine for hot splicing glass fibres for optical communication. The supports for the fibres (see also fig. 2), the positioning controls and the arc-discharge unit are all mounted on the cabinet. Above these there is a lamp housing, which contains two halogen lamps and optical components that project a 'stereogram' at about 60× magnification on the screen at the left. The lamp housing can be swung aside to allow fibres to be inserted or removed. A safety switch then disconnects the high voltage from the arc electrodes. The positioning unit is operated manually: longitudinal displacements are controlled by two knobs below the lamp housing; radial displacements are controlled by the two small levers of a micromanipulator (the red knobs on the right). The cabinet contains the electrical components and a motor for the movement of the arc electrodes and for the automatic longitudinal displacement of the left-hand fibre during the splicing operation. This is preset by the knob on the left-hand side of the cabinet. The splicing time is preset by the time switch at the upper right. A delay between the preheating and the splicing operation can be introduced if necessary by means of the lower right-hand control. The green button at the left on top of the cabinet starts the automatic operations; the red knob on its right can be used to interrupt the procedure if necessary. The unit is 33 cm high and operates from a 220 V a.c. mains supply. The arc current is 13.5 mA d.c. at a voltage of 2600 V.

[1] See also Philips tech. Rev. **36**, 177-216, 1976.

[2] A method for this has been described by W. J. J. van Hoppe, G. D. Khoe, G. Kuyt and H. F. G. Smulders in Philips tech. Rev. **37**, 89, 1977.



Fig. 2. A closer view of the machine. The arc can be seen at the centre, with glass fibres on either side of it. The fibres are held fast in a V-shaped groove by a spring-loaded clamp (next to the arc). The electrode at the front can be moved towards the one at the back for ignition. There are three height settings for the pair of electrodes. The lowest is the starting position; in the middle position the arc removes the plastic coating, and the actual splice is made in the top position.

The ends of the fibres are then lined up by hand *accurately*, and at the same time the spacing is set to a predetermined value and centred in relation to the electrodes with the aid of marks on the screen; see *fig. 3a*.

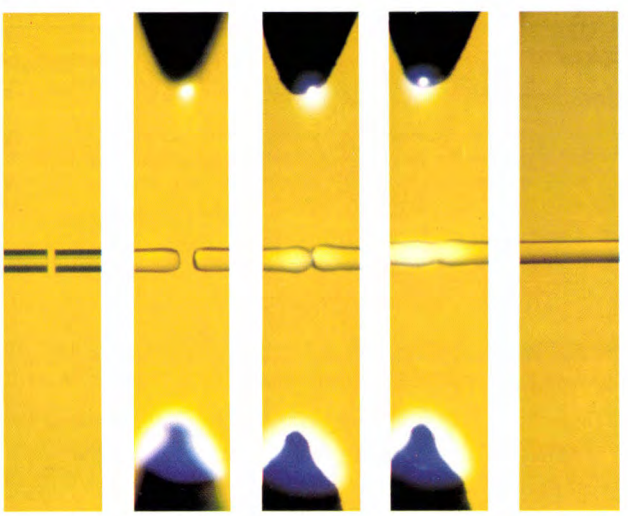


Fig. 3. Two silica fibres to be spliced; their diameter is 100 μm. *a*) After removal of the protective coating and accurate alignment. *b*) After preheating. *c*) The fibres are moved together automatically and *d*) heating is continued. *e*) The splice.

When the starter button is pressed again, the second automatic operation commences. This is the actual splicing operation. The arc moves upwards to the centre-line of the fibres to preheat the end faces (see fig. 3*b*). Next, the left-hand fibre moves a preset distance towards the right-hand one (fig. 3*c*). The join is then heated for 5 to 15 seconds (fig. 3*d*), depending on the setting of a built-in time switch. If the two end faces have been carefully prepared and the preset distance and heating time chosen correctly — these only have to be set once for a particular kind of fibre — a smooth join is produced like the one in fig. 3*e*.

The projection screen is indispensible for obtaining good splices. It permits optical losses to be kept below an average of 0.06 dB and a maximum value of 0.15 dB. Many successful splices in silica fibres have now been made with the new machine.

A. J. J. Franken
G. D. Khoe
J. Renkens
C. J. G. Verwer

A. J. J. Franken, Ir G. D. Khoe, J. Renkens and C. J. G. Verwer are with Philips Research Laboratories, Eindhoven.

Energy production by photoelectrochemical processes

R. Memming

It does not seem very likely that the Sun will be able to make a significant direct contribution in the near future to the production of the electrical energy that society needs. Nevertheless, it is sensible to examine processes that convert solar energy into electrical or chemical energy. Although the production of electrical energy by photoelectric solar cells, for example, could turn out to be rather expensive, these cells could still offer a solution in certain circumstances. Again, cells with a low efficiency might be made so cheaply as to become competitive in certain applications.

Although our articles usually describe research that has been more or less completed and has given results of practical significance, the article below describes research for which no practical applications are as yet in sight. Indeed, we cannot say very much more about this study of electrochemical semiconductor cells than that the fundamental research is virtually complete. Nevertheless, it seems worth while even at this early stage to offer a general picture of the work that has been carried out, showing a realistic view of the present state of affairs, and without venturing to speculate on possible future applications.

Introduction

The increasing demand for energy has greatly stimulated research into new sources of energy during the last ten years or so. Solar energy is of great interest here because it is independent of our energy resources. As we know, Nature makes extensive use of this energy source in photosynthesis, a process in which solar energy is converted into chemical energy. Conversion into electrical energy was first achieved in *P-N* junctions and in Schottky photovoltaic cells. These devices are used in many optoelectronic applications. High costs and other problems arise, however, if these devices are used in large areas for solar energy conversion. Systems of high conversion efficiency have only been made with single crystals. The *P-N* junctions also have to be made by a complicated diffusion technology to ensure that the depth of the junction below the surface is about the same as the penetration depth of light.

During the last few years there has been an ever-increasing interest in photoelectrochemical devices for solar energy conversion. These consist mainly of semiconductor/electrolyte junctions which can easily be made. They can be used for conversion of solar energy either into electrical energy or into chemical energy. In this article the principle of operation of both types of cell will be examined.

A great deal of fundamental research into the properties of semiconductor/electrolyte interfaces has been done since about 1960, and the essential features of most charge-transfer processes are now well understood. This knowledge could be directly applied in investigations of systems that might be suitable for solar cells. Before discussing these systems, let us first look briefly at the fundamental processes associated with a semiconductor electrode in an electrolyte.

Dr R. Memming is with Philips GmbH Forschungslaboratorium Hamburg, Hamburg, West Germany.

[*] Although it is more usual to represent voltages by E in electrochemistry, we shall use V, to avoid confusion with energy.

[1] R. Memming, in: A. J. Bard (ed.), Electroanalytical chemistry, published by Marcel Dekker, New York, in press.

[2] M. Gleria and R. Memming, J. electroanal. Chem. 65, 163, 1975.

Fundamental processes at semiconductor electrodes in an electrolyte

Potential distribution

The potential distribution at a semiconductor electrode in an electrolyte differs from that of a metal electrode in that there is a potential drop not only across the Helmholtz double layer but also across the space-charge layer beneath the semiconductor surface (*fig. 1*). This space-charge layer is produced because the density of the charge carriers in the semiconductor is much smaller than in a metal. The energy bands at the semiconductor surface are therefore curved — upwards if the space charge is positive, downwards if it is negative (*fig. 2*). The Fermi level is assumed to be constant. If the electrode potential V_E relative to a reference electrode is measured, we obtain the expression

$$V_E = V_{sc} + V_H + \text{constant},$$

where V_{sc} is the potential difference across the space-charge layer and V_H the potential difference across the Helmholtz layer; the potential differences to the reference electrode are included in the constant [*].

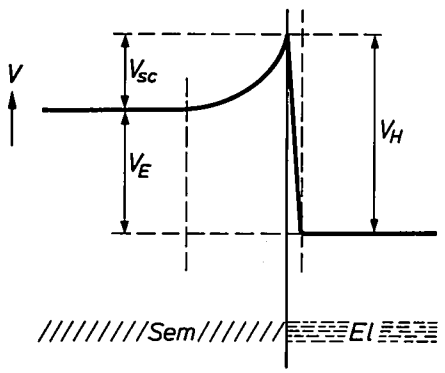


Fig. 1. Potential distribution V at the semiconductor/electrolyte interface. In addition to the potential drop V_H across the Helmholtz double layer in the electrolyte (*El*), as with metal electrodes, there is a much more gradual potential variation V_{sc} in the space-charge layer below the surface of the semiconductor *Sem*.

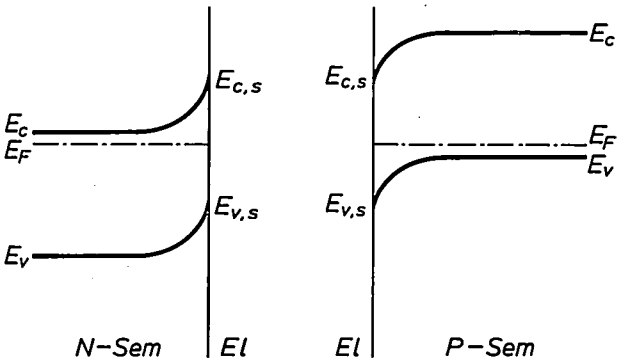


Fig. 3. Position of conduction and valence bands for an *N*-type semiconductor (*left*) and a *P*-type semiconductor (*right*) of the same material. The values of E_c and E_v at the surface ($E_{c,s}$ and $E_{v,s}$) are identical for both cases and depend only on the *p*H of the electrolyte.

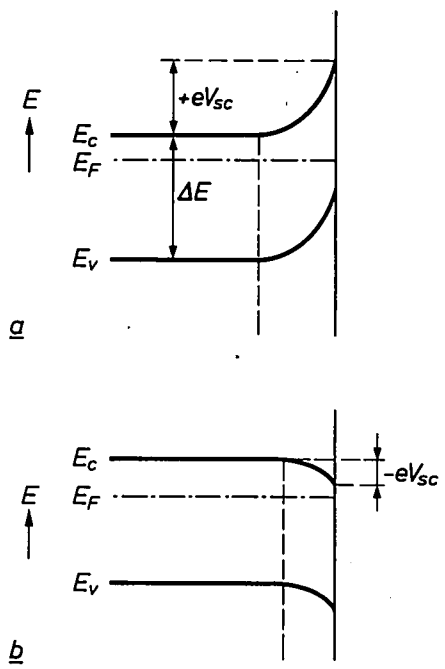


Fig. 2. Energy-level diagram for the semiconductor electrode at and below the surface. E_c bottom of the conduction band. E_v top of the valence band. E_F Fermi level. ΔE band gap. eV_{sc} band-bending. The bands may be bent either upwards or downwards; upward bending signifies a positive space charge and downward bending a negative one.

Quantitative measurements of the space-charge capacitance of many semiconductor electrodes have shown that any variation of an externally applied potential only appears across the space-charge layer, while the potential across the Helmholtz layer remains unchanged [1]:

$$\Delta V_E = \Delta V_{sc}.$$

From this result it follows that there is an electrode potential for which the energy bands in the semiconductor are flat up to the surface. This 'flat-band potential' varies with the doping of the semiconductor crystal. Conversely, at a given electrode potential V_E, the potential drop V_{sc} depends on the doping of the semiconductor. This is also valid if the doping changes from *N*-type to *P*-type. A quantitative analysis of these results has also shown that the position of the energy bands and of the Fermi level at the surface is independent of the doping, as shown schematically in *fig. 3*. Such an analysis shows that the position of the energy bands of semiconductors can be clearly related to the potential of the normal hydrogen electrode (NHE) as a reference [1,2]; see *fig. 4*. This energy scheme is of great importance for selecting a suitable semiconductor electrode for solar cells, as will be discussed below.

Energy levels and the Fermi level in electrolytic systems

Electrolytes that contain a redox system are characterized by the redox potential V_{redox}. Values of this potential are usually given in the conventional scale with the NHE potential as a reference point. In solid-state physics, however, the potential of the vacuum outside the solid is used as a reference point. When this reference point is used, the redox potential corresponds to the energy required for transferring an electron from the redox system in the electrolyte into the vacuum. The electron energy is then defined in the same way as in solids and the redox potential can be considered as the Fermi level $E_{F,el}$ of the electrolyte. The potential difference between the normal hydrogen electrode and vaucum is equal to about -4.5 V [3]. The Fermi level of a redox system is therefore equal to

$$E_{F,el} = -(4.5\ \text{eV} + eV_{redox}).$$

Both energy scales are given in fig. 4. When a semiconductor electrode is in equilibrium with an electrolyte, the Fermi levels at the interface must be equal [1,4]:

$$E_F = E_{F,el}.$$

As in solids, the energy states of a redox system can be occupied by electrons or empty. However, the positions of these energy states can differ considerably owing to the strong interaction between the ions of the redox system and the solvent. Each ion is surrounded by a solvation shell, a shell of solvent molecules. The degree of interaction depends on the size of the ion and its charge. If an electron is transferred from the reduced component of a redox system to the vacuum, there is a rearrangement or reorientation of the solvation shell, and the same is true for the reverse process. The complete cycle is shown in *fig. 5*, where I is the ionization energy, A the electron affinity and λ the reorientation energy. The electron energies of the reduced and oxidized components are therefore not equal. The actual electron transfer between the redox system and an electron donor or acceptor is fast whereas the ensuing

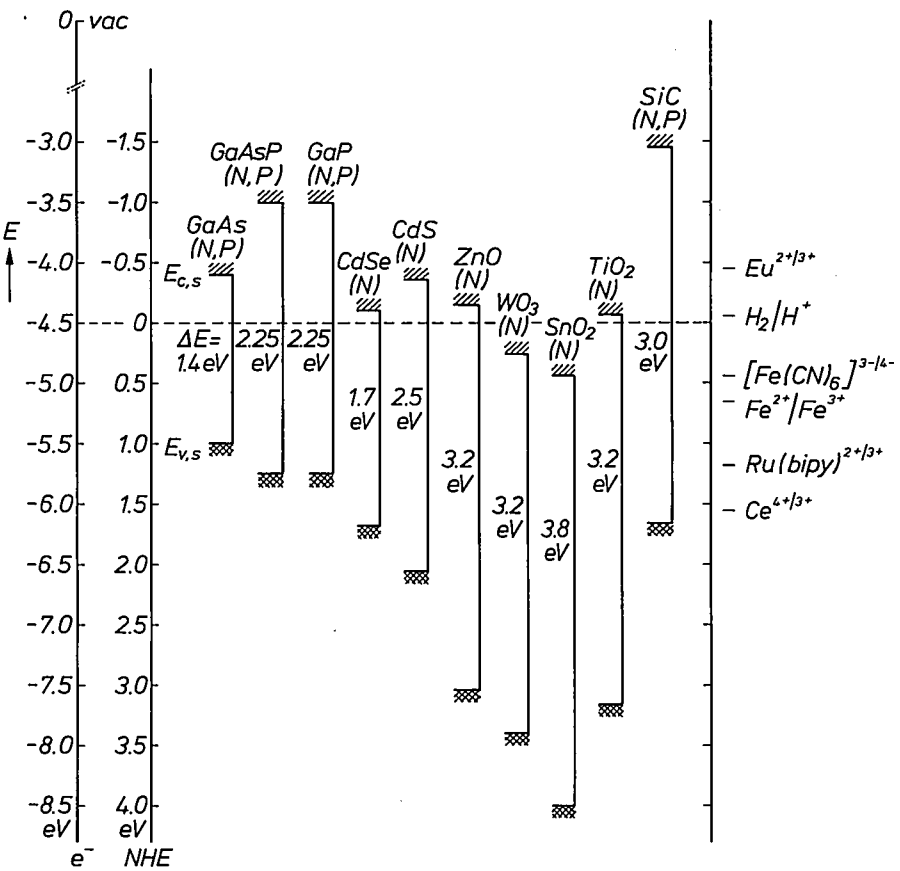


Fig. 4. Position of energy bands at the surface ($E_{c,s}$ and $E_{v,s}$) for various semiconductors at a *p*H of 1 (see fig. 3). The zero of the scale on the far left (e^-) is the energy of an electron in vacuum. The zero of the other scale on the left (*NHE*) is the potential of the normal hydrogen electrode. On the right are shown the standard potentials of several redox couples. *bipy* represents 2,2′ bipyridyl.

reorientation of the solvent molecules is relatively slow (the Frank-Condon principle). The two partial processes together constitute the total electron-transfer process, so that

$$I - \lambda = A + \lambda = E_{\mathrm{F,el}};$$

see *fig. 6a*. The energy levels E^0_{red} and E^0_{ox} of the reduced and the oxidized components of a redox system

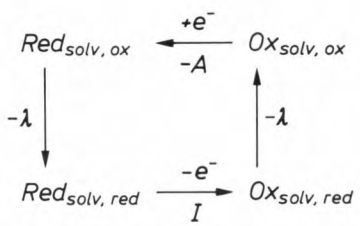


Fig. 5. Diagram of the electron transfer from the reduced component (*Red*) to the vacuum and the reverse process to the oxidized components (*Ox*) of a redox couple. The actual electron transfer, i.e. the ionization of *Red* (supplied energy *I*) and the recapture of the electron by *Ox* (energy $-A$) is followed in both cases by a reorientation of the solvation shell, releasing a reorientation energy λ. This can be formally expressed as $I - \lambda - A - \lambda = 0$.

differ from the Fermi level by λ. These levels are not discrete levels, but are distributed over a certain energy range owing to the fluctuations of the solvation shell. The corresponding distribution function of the density of states can be derived from various theories (see for example notes [1] and [4]), and are given by:

$$D_{\mathrm{red}} = \exp - \frac{(E - E_{\mathrm{F,el}} - \lambda)^2}{4kT\lambda}$$

and

$$D_{\mathrm{ox}} = \exp - \frac{(E - E_{\mathrm{F,el}} + \lambda)^2}{4kT\lambda},$$

where D_{red} corresponds to occupied states (reduced component) and D_{ox} corresponds to the unoccupied states (oxidized component). These functions are shown in fig. 6*b*. The half-width at room temperature is

$$\Delta E_{\frac{1}{2}} = 0.53\lambda^{\frac{1}{2}}\ \mathrm{eV}.$$

Typical values of λ for inorganic redox systems are of the order of 1 eV so that the energy levels of such a system are spread over a considerable energy range. The relative position of energy levels at the interface can now be easily obtained by using the condition that the Fermi levels are equal on both sides of the interface. Examples for two redox systems of different normal potentials are given in *figs 7a* and *b*. Redox systems with a large positive normal potential have low-lying energy levels in this scheme. This can be seen directly from a comparison of the two energy scales in fig. 4.

Charge transfer

The application of an external voltage to an electrode can in principle cause a current to flow. As explained earlier, the relative position of the energy bands at the interface remains unchanged, whereas the potential difference across the space-charge layer is varied, leading to a corresponding change in the band-bending. This is demonstrated for anodic polarization in fig. 7*c* and for cathodic polarization in fig. 7*d*. The Fermi levels are no longer equal but differ by an amount determined by the applied voltage η:

$$E_{\mathrm{F}} - E_{\mathrm{F,el}} = e\eta.$$

Electron transfer, however, can only occur if levels of equal energy exist on both sides of the interface. Various cases are possible. In fig. 7*a* the empty (D_{ox}) and the occupied states (D_{red}) of the redox system partly overlap the conduction band of the semiconductor. In this case an anodic current $i_{\mathrm{c}}{}^+$ is possible through the electron transfer from occupied states D_{red} to the conduction band, in which practically all the states (with a total density N_{c}) are empty. For this current:

$$i_{\mathrm{c}}{}^+ \propto \int_{E_{\mathrm{c}}}^{\infty} N_{\mathrm{c}} D_{\mathrm{red}}\, \mathrm{d}E,$$

an expression which is independent of the applied voltage. The situation is different for a cathodic current $i_{\mathrm{c}}{}^-$, where electrons are transferred from the conduction

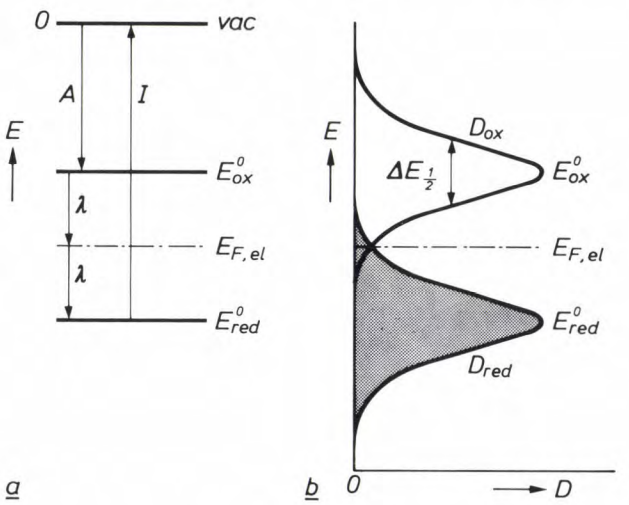


Fig. 6. *a*) Simplified diagram of the energy levels of the components of a redox couple, as derived from the cycle in fig. 5 ($I - \lambda = A + \lambda$). The Fermi level $E_{\mathrm{F,el}}$, corresponding to the redox potential, lies midway between the filled level $E^0{}_{\mathrm{red}}$ and the low $E^0{}_{\mathrm{ox}}$ of the redox system. *b*) In reality the levels are distributed over a certain energy range owing to fluctuations of the solvation shell. *D* distribution functions. $\Delta E_{\frac{1}{2}}$ half-width of *D*.

[3] F. Lohmann, Z. Naturf. **22a**, 843, 1967.
[4] H. Gerischer, in: H. Eyring, D. Henderson and W. Jost (ed.), Physical chemistry, Vol. IX A, Academic Press, New York 1970, p. 463.

band to the empty states D_{ox} of the redox system. In this case we have

$$i_c^- \propto \int_{E_c}^{\infty} n_s D_{ox}\, dE,$$

where n_s is the electron density in the conduction band at the surface. This depends on the electron density n_0 in the bulk of the semiconductor and the potential difference V_{sc} across the space-charge layer. It is given by the Boltzmann equation:

$$n_s = n_0 \exp(-eV_{sc}/kT). \qquad (1)$$

When a voltage is applied the only change is in the potential difference across the space-charge layer; for cathodic polarization therefore

$$n_s = n_0 \exp\{-e(V_{sc}^0 - \eta)/kT\},$$

where V_{sc}^0 is the band-bending at equilibrium and η is shown in *fig. 8a* (solid curve) for the combination of an SnO_2 electrode with the redox system Fe^{2+}/Fe^{3+} [5].

Similar equations are valid for charge exchange via the valence band. This is possible if the energy levels of the redox system partly overlap the valence band, as shown in fig. 7*b* and *d*. The current equations are now given by

$$i_v^- \propto \int_{-\infty}^{E_v} N_v D_{ox}\, dE,$$

and

$$i_v^+ \propto \int_{-\infty}^{E_v} p_s D_{red}\, dE.$$

For cathodic polarization the current i_v^- is determined by the transfer of electrons from the valence band (total density of states N_v) to D_{ox}. This process is independent of the applied voltage. An anodic current

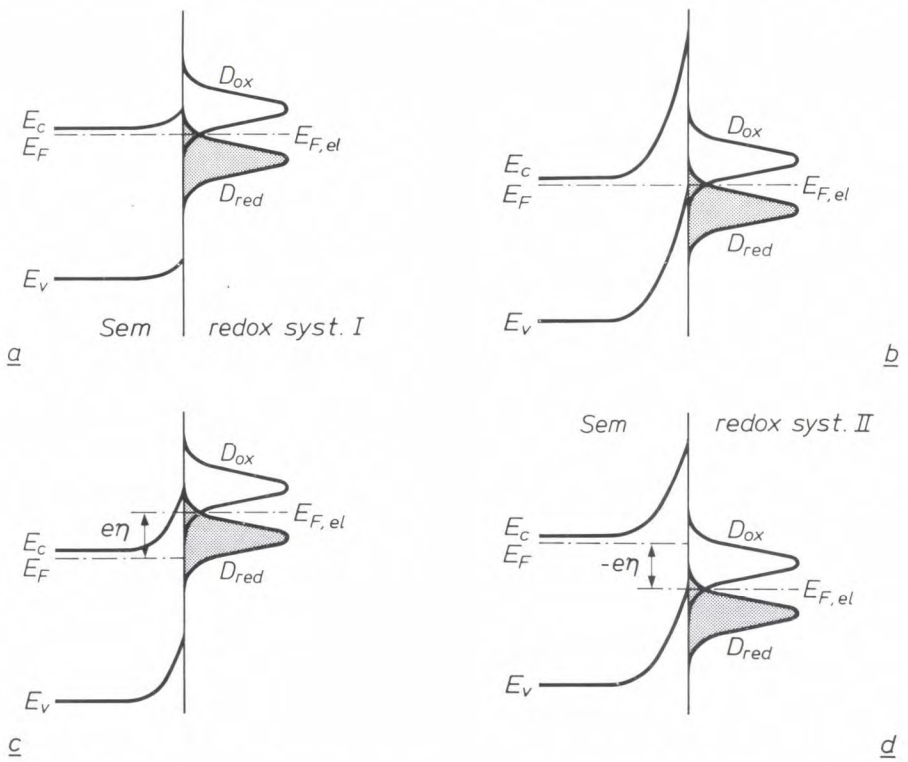


Fig. 7. In equilibrium the Fermi levels of the semiconductor and the electrolyte are at the same height. Since the position of the band edges at the interface is fixed, the band-bending in a given semiconductor is completely determined by the position of the Fermi level at the surface and therefore by the redox potential (*a* and *b*). If the electrode is polarized by an external voltage ('anodically' by making the semiconductor positive on 'cathodically' by a negative voltage), there is a jump $e\eta$ in the Fermi level at the electrode interface (*c* and *d*).

the applied voltage. The cathodic current clearly increases exponentially with η. An example of these two cases of electron transfer via the conduction band is

i_v^+, which is due to electron transfer from D_{red} to the valence band, can only occur if there are unoccupied states (holes) in the valence band at the surface. The

hole density p_s at the surface is given by

$$p_s = p_0 \exp(eV_{sc}/kT) = p_0 \exp\{e(V_{sc}^0 - \eta)/kT\}. \quad (2)$$

An example is shown in fig. 8*b* (solid curve) for the case of *P*-GaP and the same redox system [6].

In this model a potential-dependent current only occurs if the density of the electrons or holes at the surface given by (1) or (2) determines the charge transfer. The doping of the semiconductor is of importance in so far as it determines the densities of the charge carriers in the bulk of the semiconductor (n_0 or p_0). For example, if n_0 is very small, as in *P*-type material, then from eq. (1) n_s must also be very small. If therefore electrons from the conduction band are required for the electrochemical process at *P*-type electrodes, the cathodic current does not depend on the potential, as in *N*-type material, but is limited by diffusion.

Finally, we should note that current flow is also possible if the electrolyte contains no redox system, but only a conducting salt. In this case the cathodic current is usually determined by hydrogen evolution and the anodic current by electrochemical dissolution of the semiconductor electrode (see the dashed curves in fig. 8). The dissolution of the electrode always occurs by electron transfer via the valence band. It can only be observed with *P*-type material since holes in the valence band are required. The reaction leads to severe problems in investigations on solar cells, as will be seen later. Present knowledge indicates that hydrogen evolution only occurs for electron transfer from the conduction band: it has only been observed with cathodic polarization and *N*-type electrodes.

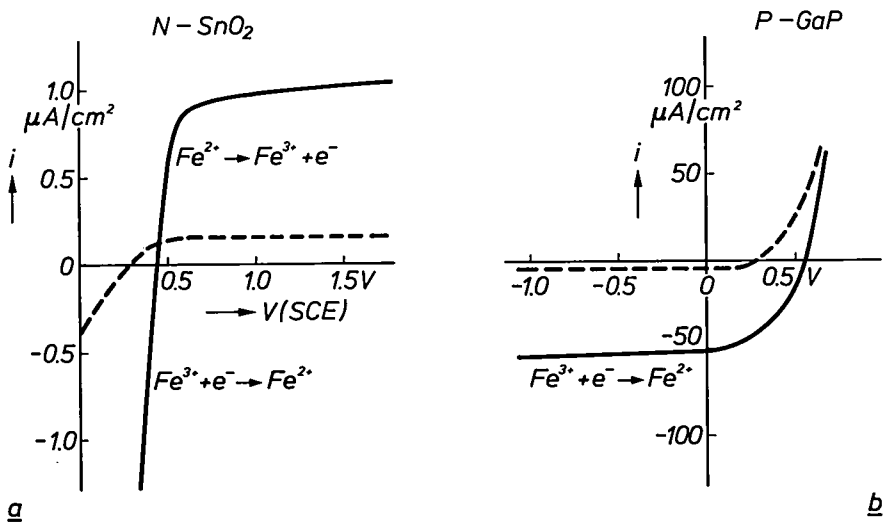


Fig. 8. Dark-current/voltage characteristics for charge transfer between a semiconductor electrode and an electrolyte. The current density i is plotted as a function of the electrode potential V with respect to a standard calomel electrode (SCE). *a*) If the conduction band of a semiconductor (SnO_2 here) overlaps the energy levels of the redox potential of the electrolyte (Fe^{2+}/Fe^{3+} redox couple, see the diagram in fig. 4), charge exchange takes place via the conduction band. The transfer of electrons of ferrous ions to the virtually empty conduction band is independent of the applied voltage. Electron transfer from the conduction band to ferric ions is governed by the concentration of electrons in the conduction band at the interface, and is therefore strongly dependent on the applied voltage (*solid curve*). If the solution contains no redox couple but only a conducting electrolyte, the dashed curve is found. *b*) In the case of charge exchange via the valence band (*P*-GaP and Fe^{2+}/Fe^{3+} here) the transfer of electrons to the electrolyte is independent of the voltage, but the reverse process is governed by the availability of holes at the electrode surface and is therefore strongly dependent on the voltage.

The nature of energy generation in photoelectrochemical cells

The previous sections have shown how the electrons and holes in the semiconductor electrode control the charge-transfer process. It is quite obvious that in cases in which the minority carriers determine the current flow the carrier density can be increased by light excitation. When electron-hole pairs are created by light excitation in the space-charge layer beneath the surface of the electrode the electrons and holes are separated by the electric field in this layer. The effect is illustrated in *fig. 9* for an *N*-type semiconductor with a

[5] R. Memming and F. Möllers, Ber. Bunsen-Ges. phys. Chemie 76, 475, 1972.

[6] R. Memming and G. Schwandt, Electrochim. Acta 13, 1299, 1968.

positive space charge. This leads to a decrease in the electric field under open-circuit conditions and consequently a decrease in the band-bending (fig. 9*a*, dashed lines). The corresponding photovoltage V_{ph} can be measured as a change in the electrode potential from the 'dark value' $V_{E,dark}$ to a more 'cathodic' value $V_{E,ill}$ on illumination:

$$V_{ph} = V_{E,ill} - V_{E,dark}.$$

There is a maximum in V_{ph} at light intensities for which the energy bands are flat up to the surface:

$$V_{ph}^{max} = V_{fb} - V_{E,dark},$$

where V_{fb} is the flat-band potential. If the semiconductor electrode is held at the dark potential $V_{E,dark}$, however, by short-circuiting it, then there is no field decrease on illumination, and the holes are forced to cross the interface to an acceptor molecule in the electrolyte, such as the reduced component of a redox system (fig. 9*b*): a photocurrent is then produced in the external circuit.

Similar effects occur at a *P*-type electrode with a negative space-charge layer. In this case the electrons are driven towards the surface. The signs of the photovoltage and photocurrent are the opposite of those with the *N*-type electrode.

The origin of the photoeffect is therefore to be found in the existence of an electric field within the space-charge region beneath the electrode surface. In this respect the photoeffect is very like that in a *P-N* junction (fig. 9*c*), and there is in fact a definite similarity in the photoelectric behaviour of a semiconductor-electrolyte interface and a *P-N* junction. In both cases the photocurrent is proportional to the light intensity *I*, while for the photovoltage

$$V_{ph} = \frac{kT}{e} \ln\left(1 + \frac{\Delta p}{p_0}\right) \approx \frac{kT}{e} \frac{\Delta p}{p_0} = \frac{kT}{e} \gamma I,$$

where p_0 is the density of holes in the bulk of the semiconductor, Δp the increase in p_0 on illumination and γ a proportionality factor. Similarities can also be found in the current-voltage curves.

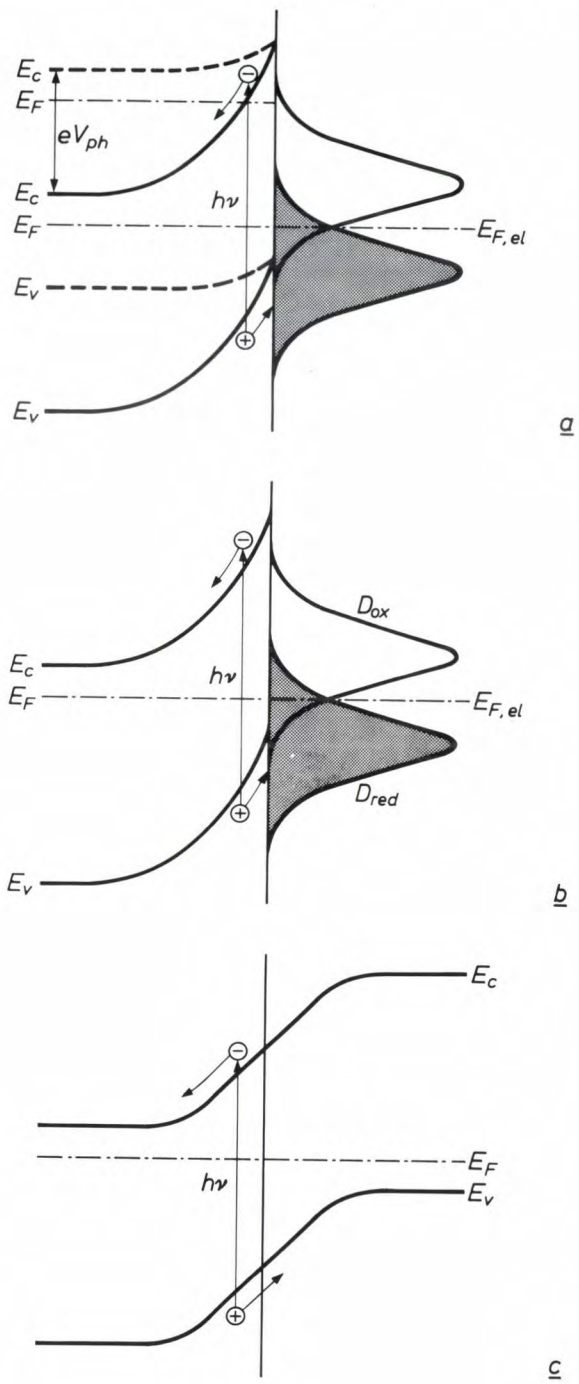


Fig. 9. Absorption of light in the space-charge layer of a semiconductor electrode (*a*,*b*), as in an ordinary *P-N* junction (*c*), leads to the release of (photo)electrons and (photo)holes, which are driven in different directions by the field. If the electrode is illuminated under open-circuit conditions (*a*), the band-bending is completely or partly neutralized and the electrode potential changes with the photovoltage V_{ph}. In a short-circuit (*b*) the charge is transferred across the interface, via the valence band in the case shown here (photocurrent).

Regenerative solar cells

It might seem from the previous section that it is a simple matter to convert solar energy into electricity by putting a semiconductor electrode in an electrolyte with a redox system. Many possible arrangements have to be rejected, however, because of anodic dissolution of the semiconductor. In practice the choice is restricted to 'regenerative cells'.

In selecting a semiconducting material the criterion for the band gap is the same as for cells with a *P-N* junction; the maximum conversion efficiency for solar energy would be expected to occur at a band gap of

[7] M. S. Wrighton, A. B. Ellis, P. T. Wolczanski, D. L. Morse, H. B. Abrahamson and D. S. Ginley, J. Amer. Chem. Soc. **98**, 2774, 1976.

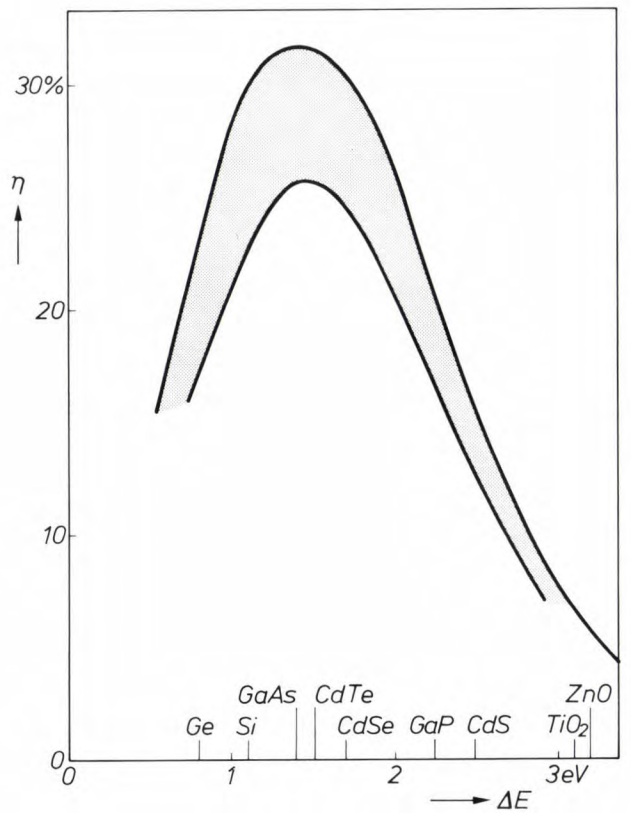


Fig. 10. Theoretical efficiency η for the conversion of sunlight into electrical energy by semiconductor electrodes as a function of the band gap ΔE. The value found for η depends to some extent on the boundary conditions, but lies in the region between the two curves.

about 1.6 eV [7]; see *fig. 10*. Semiconductors with considerably smaller band gaps give a much lower efficiency because the absorption of photons of very high energy leads to an excitation of electrons to higher energy levels in the conduction band. Since such electrons fall back to the lower edges of the conduction band the surplus energy is converted into heat.

In the remainder of this section the practical restrictions of some of the regenerative cells for converting solar energy into electricity will be discussed. First, however, we shall consider how to construct the cell with the greatest energy efficiency.

Efficiency; electrochemical dissolution of the electrode

As mentioned earlier, photocells can be made with *N*-type or *P*-type semiconductors. To obtain cells with high efficiencies semiconductor electrodes have to be combined with a redox system in such a way that, at equilibrium, the highest possible electric field is produced across the space-charge layer. If we consider an *N*-type semiconductor first, then we ought to choose a redox system with a redox potential close to the valence band of the semiconductor. Fig. 4 indicates that the combination of *N*-type CdSe with the redox system Ce^{4+}/Ce^{3+} would be suitable. Since the Fermi level is at the same height everywhere in equilibrium, there is strong band-bending (*fig. 11*). In this case, the semiconductor electrode and the inert metal counter-electrode have the same potential. If the semiconductor electrode is now illuminated, the space-charge field becomes weaker. The maximum photovoltage is obtained when the energy bands are flat. For a heavily doped semiconductor

$$V_{\mathrm{ph}}^{\max} \approx \Delta E/e,$$

where ΔE is the band gap. Such a cell, consisting of a

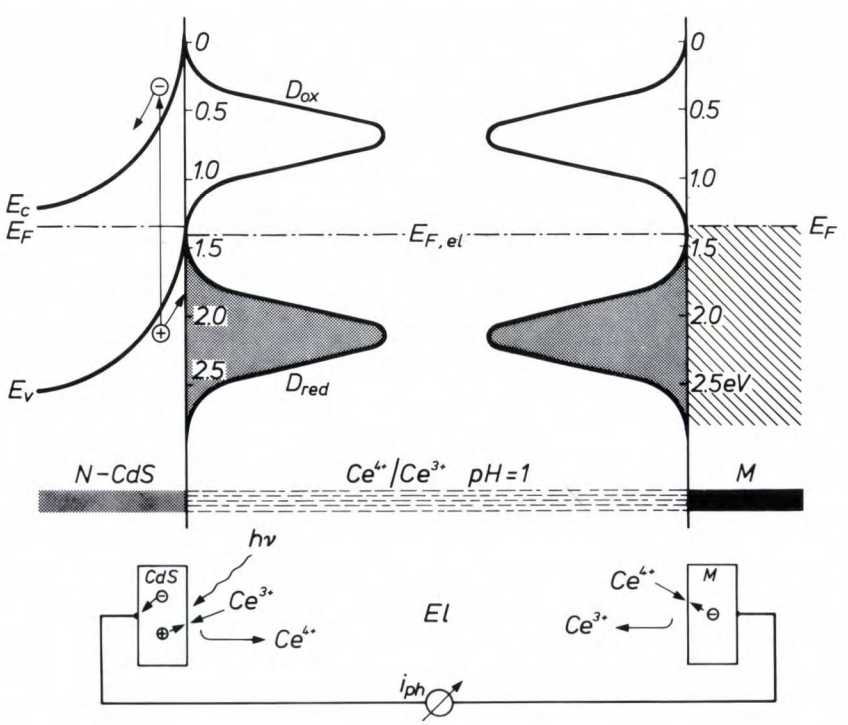


Fig. 11. Energy-level diagram (*above*) and charge-transfer processes (*below*) in the short-circuited system *N*-CdSe/ceri-cero/metal. In the case shown here the band-bending is almost identical with the band gap, since the redox potential coincides with the edge of the valence band, whereas in the bulk of the semiconductor the Fermi level is near the conduction band. Photo-holes are driven to the surface where they oxidize cero ions ($Ce^{3+} \rightarrow Ce^{4+}$). These are regenerated at the metal electrode by reduction of ceri ions. This example is hypothetical, since in this case a non-regenerative process — the semiconductor dissolution — is dominant.

semiconductor, an electrolyte with a redox system and a metal electrode, should in principle operate regeneratively if the semiconductor and metal electrodes are short-circuited; see fig. 11 (*lower diagram*). Positive minority carriers (holes), formed by illumination of the semiconductor, are driven towards the electrode surface and neutralized there in the reaction

$$Ce^{3+} + p^{+} \rightarrow Ce^{4+}.$$

The Ce^{3+} ions taken up here are regenerated at the metal electrode by the reaction

$$Ce^{4+} + e^{-} \rightarrow Ce^{3+}.$$

The electrons required here flow through the external circuit from the semiconductor to the metal electrode.

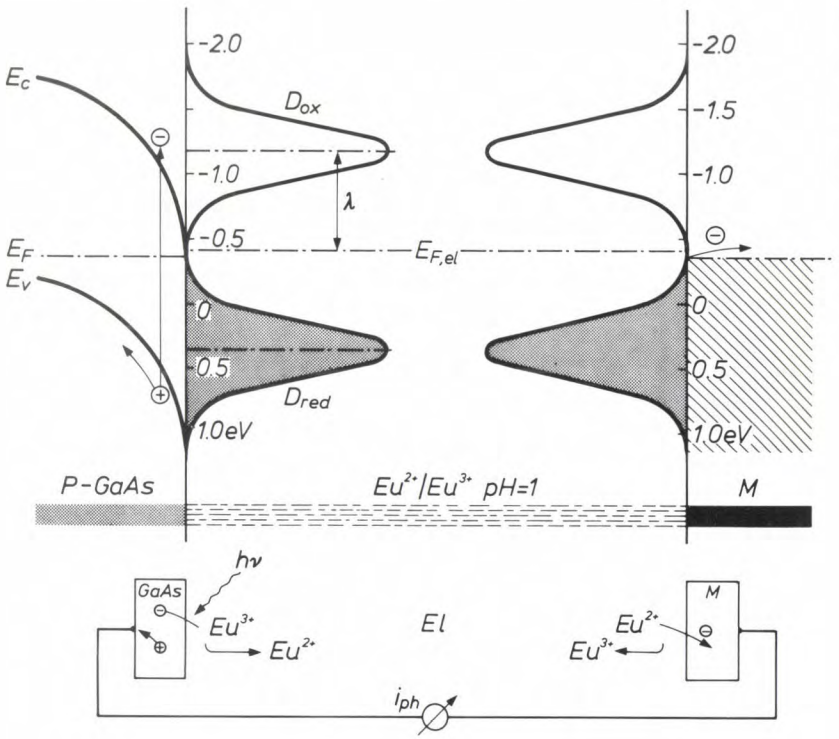


Fig. 12. Energy diagram (*above*) and charge-transfer processes (*below*) for the short-circuited system *P*-GaAs/Eu^{2+}-Eu^{3+}/metal. The redox potential of the europium couple coincides with the conduction-band edge. Photoelectrons created in the space-charge layer are driven to the surface and transferred to Eu^{3+} ions. In addition to this 'regenerative' process, hydrogen generation may also take place by the transfer of photoelectrons to water molecules at the semiconductor interface. This latter process can be neglected in this case.

A similar cell could also be constructed with a *P*-type semiconductor electrode. In this case a redox system should be used with a Fermi level located close to the conduction band of the semiconductor. Fig. 4 indicates that a good combination would be GaAs and Eu^{2+}/Eu^{3+}. The equilibrium condition is shown in *fig. 12*. The charge-transfer processes are now:

$Eu^{3+} + e^{-} \rightarrow Eu^{2+}$ at the semiconductor electrode and
$Eu^{2+} \rightarrow Eu^{3+} + e^{-}$ at the metal electrode.

However, many combinations of a semiconductor electrode and a redox system do not operate regeneratively because other photo-induced reactions with water may be favoured. For an *N*-type electrode the minority charge carriers, the holes, are not always transferred to the redox system: they can also induce reactions such as anodic dissolution or oxygen evolution. At a *P*-type electrode there may be hydrogen evolution instead of the intended redox process. In all these cases a regenerative process is not practical. For instance the combination of fig. 11 (CdSe with Ce^{3+}/Ce^{4+}) does not work at all because the anodic process is entirely due to electrochemical dissolution. While various combinations have been found that do operate regeneratively, in all these cases the factors that determine the efficiency (band-bending and redox potential) seem unfortunately to be less favourable.

Examples of regenerative combinations are *N*-CdS/S^{2-} [7,8], *N*-CdSe/S^{2-} [9], *N*-CdS/$Fe(CN)_6^{4-/3-}$ [10], *N*-GaAs/Se^{2-} [11], *P*-GaAs/$Eu^{2+/3+}$ [12] and *P*-GaP/$Eu^{2+/3+}$ [12]. In all of these cases the photovoltages were found to be much lower than $\Delta E/e$. With the *N*-type electrode this result is obtained because the Fermi level of the redox system is well above the valence band.

These results lead us to the problem of which factors favour the desired redox process and which factors favour undesirable subsidiary reactions. A first approach to this problem via thermodynamic considerations now follows.

Thermodynamic calculations

The most important process in competition with the redox process is the anodic dissolution of the semiconductor electrode. Some information about the potential at which this process starts can be obtained from thermodynamic data by calculating the free energy. If GaP is taken as an example the dissolution of the electrode can be described by the equation [6]:

$$GaP + 6H_2O + 6p^+ \rightarrow Ga(OH)_3 + H_3PO_3 + 6H^+.$$

Six holes are thus required for the dissolution of one GaP molecule. The potential of this electrode reaction can be related to the normal hydrogen potential by combining the equation above with the equation for the hydrogen electrode:

$$3H_2 \rightarrow 6H^+ + 6e^-.$$

Subtraction gives [12]

$$GaP + 6H_2O \rightarrow Ga(OH)_3 + H_3PO_3 + 3H_2.$$

The free energy ΔG of this reaction, which is found by substituting values from the literature [13,14] for the free energy of formation of the reaction components, is -170 hJ/mol (—40.6 kcal/mol). The potential V_{diss} for the anodic dissolution is given by

$$V_{diss} = \Delta G/zF,$$

where z is the number of electrons taking part in the reaction and F is the Faraday constant. At $pH = 0$ this potential is equal to -0.29 V, at $pH = 7$ it is -0.71 V.

The potential at which anodic dissolution of the electrode would be expected is plotted in *fig. 13* for GaP and GaAs and in *fig. 14* for TiO_2 [15]. A summary of these and other values is given in *Table I* [12]. These values, obtained by thermodynamic calculations, indicate that many electrodes start to dissolve at lower anodic potentials than those of many redox processes. Accordingly, in most cases dissolution of the electrode can compete with the redox reaction. In fig. 13 and fig. 14 the energy position of the conduction and valence bands at the surface are also plotted for comparison. It is interesting to note that the energies corresponding to dissolution potentials are near the middle of the band gap. Whether a particular reaction takes place or not cannot be predicted with certainty from thermodynamics alone. For example, it was stated earlier that for electron transfer comparable energy levels must exist on both sides of the interface. Strictly speaking, this condition is only valid for reactions that have a weak interaction with the electrode. The condi-

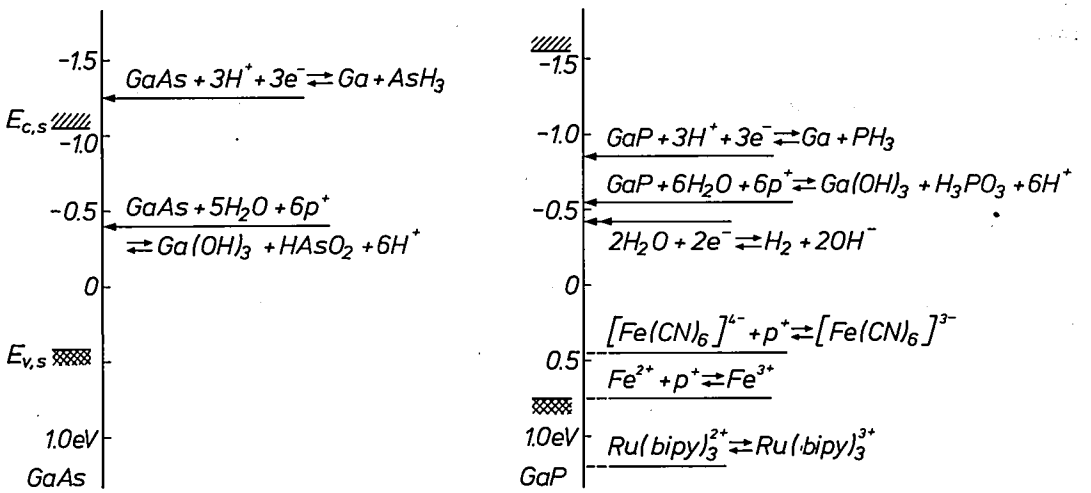


Fig. 13. The energy-level diagrams of GaAs (*left*) and GaP (*right*) and the electrode potentials V_{diss} (*arrows*) at which dissolution of the semiconductor material would be expected. The electrode potentials are calculated from the free energy of formation of the various reaction components for the case of a neutral electrolyte. The zero of the scale is the potential of a normal hydrogen electrode. Also indicated are the potential at which the generation of hydrogen starts (*double arrow*) and the redox potential of some important redox couples.

[8] B. Miller and A. Heller, Nature **262**, 680, 1976.
[9] A. Heller, K. C. Chang and B. Miller, J. Electrochem. Soc. **124**, 697, 1977.
[10] H. Gerischer and J. Gobrecht, Ber. Bunsen-Ges. phys. Chemie **80**, 327, 1976.
[11] A. Heller and K. C. Chang, Proc. Conf. on Semiconductor-liquid interfaces under illumination, Airlie 1977.
[12] R. Memming, J. Electrochem. Soc., in press.
[13] W. M. Latimer, The oxidation states of the elements and their potentials in aqueous solutions, Prentice-Hall, Englewood Cliffs.
[14] C. D. Thurmond, J. Phys. Chem. Solids **26**, 785, 1965.
[15] H. Gerischer, Proc. Congress on Photochemical processes, London 1976.

tion is therefore valid for electron transfer to a redox system but not for anodic dissolution. Although the dissolution energies occur near the centre of the band gap it has been shown experimentally that holes in the valence band are required for this process.

Further investigations are necessary to provide information about the factors that determine whether regenerative redox reactions are suppressed by anodic dissolution.

lysed much more easily by using a rotating ring-disc electrode, consisting of a semiconductor disc and a platinum ring [16]. We shall not go into more detail about this method now; all we need to say here is that it gives the total photocurrent i_{ph} and also the current i_{ox} produced by charge transfer to the redox system. If i_{ph} and i_{ox} are equal, the dissolution of the electrode is negligibly small. *Table II* gives some measured values for i_{ox}/i_{ph}.

Table I. Electrochemical (anodic) dissolution reactions of six semiconductors and the potentials (thermodynamic values) at which the reactions start for $pH = 7$.

Semiconductor (band gap ΔE)	Reaction	V_{diss} ($= \Delta G/zF$)
GaAs (1.4 eV)	$GaAs + 5H_2O + 6p^+ \rightleftarrows Ga(OH)_3 + HAsO_2 + 6H^+$	−0.38 V
GaP (2.25 eV)	$GaP + 6H_2O + 6p^+ \rightleftarrows Ga(OH)_3 + H_3PO_3 + 6H^+$	−0.71 V
CdS (2.5 eV)	$CdS + 2HCl + 2p^+ \rightleftarrows Cd^2 + 2Cl^- + 2H^+ + S$	−0.09 V
TiO_2 (3.0 eV)	$TiO_2 + 4HCl + 4p^+ \rightleftarrows TiCl_4 + O_2 + 4H^+$	+1.4 V
ZnO (3.25 eV)	$2ZnO + 4H_2O + 4p^+ \rightleftarrows 2Zn(OH)_2 + O_2 + 4H^+$	+0.81 V
SnO_2 (3.8 eV)	$SnO_2 + 4HCl + 4p^+ \rightleftarrows SnCl_4 + O_2 + 4H^+$	+1.05 V

Table II. The ratio of the current resulting from charge transfer to the redox system (i_{ox} for anodic processes, i_{red} for cathodic processes) to the total photocurrent i_{ph}, both measured with a ring-disc electrode [16]. If both currents are equal the photocurrent is only determined by charge transfer to the redox system.

Redox system (V_E)	i_{ox}/i_{ph} (*N*-type semiconductor) CdS	GaP	GaAsP	GaAs	i_{red}/i_{ph} (*P*-type semiconductor) GaP	GaAs
$Fe(CN)_6^{4-/3-}$ ($V_E = 0.42$ V)	1.0 (pH = 6-13)	1.0	0.35	0.35	0.04 ($pH = 1$)	0.25 ($pH = 1$)
$Fe^{2+/3+}$(0.73 V)	0.13	0.35	0.1	0.1		
$Ce^{3+/4+}$(1.6 V)	0.08	0.1	0	0		
$Ru(bipy)_3^{2+}$ (1.2 V)	0	0				
Quinone (Q/QH$_2$) (0.5 V)	1.0	1.0	0.2	0.2	1.0 duroquinone[*] (pH = 1-5)	1.0 duroquinone (pH = 1-6)
Fe(II)ox	0.1	0.1				
$Eu^{2+/3+}$(−0.42 V)					1.0 (pH = 1-6)	1.0 (pH = 1-6)
Fe(II)EDTA [*]					0.9	1.0

[*] Duroquinone is tetramethyl-1,4-benzoquinone. EDTA is an abbreviation for ethylenediaminotetra-acetic acid.

Stability of semiconductor surfaces

It is not very easy to determine experimentally the extent to which anodic dissolution competes with the redox process. A current-potential curve (*fig. 15*) as measured for an *N*-type semiconductor does not give any information, since the anodic photocurrent is only determined by the concentration of holes exited by light, and hence by the light intensity. Information could perhaps be obtained by analysing the electrolyte after passing a known current through it. This method, however, is time-consuming and not very accurate. Photoelectrochemical processes can be ana-

The value of i_{ox}/i_{ph} is dependent on the pH, as *fig. 16a* shows for CdS in combination with various redox systems [16] and fig. 16*b* shows for various semiconductors in combination with the ferri/ferrocyanide redox system [12].

Similar experiments have been performed with *P*-type electrodes during cathodic polarization. In this case the redox process has to compete with hydrogen evolution. The values for i_{red}/i_{ph} found with these electrodes are also given in Table II.

The question now arises as to whether it is possible to find semiconductors that are less polar than GaAs,

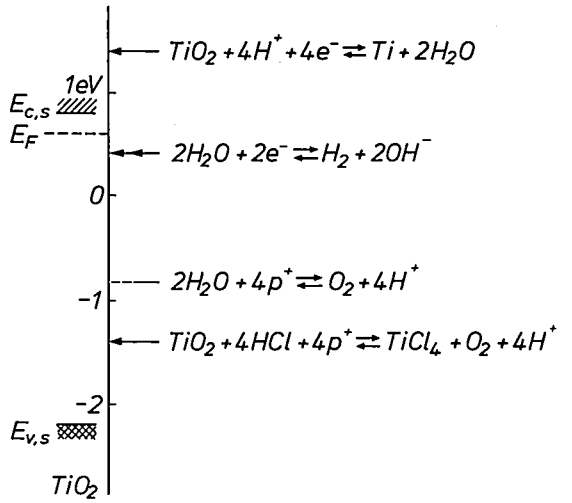


Fig. 14. As fig. 13, but now for a TiO_2 electrode (the pH of the electrolyte is 7).

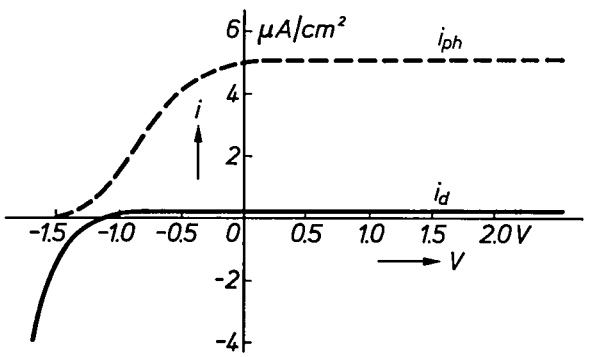


Fig. 15. Current-voltage characteristics for the dark current i_d and the photocurrent i_{ph} for an N-type semiconductor electrode. The photocurrent depends only on the light intensity and gives no information about the nature (regenerative or not) of the electrode reactions.

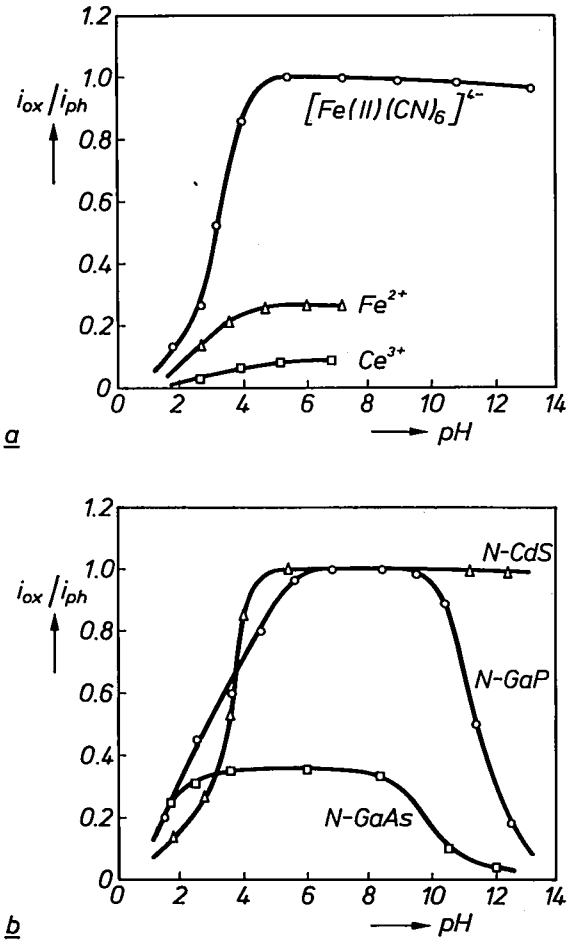


Fig. 16. The contribution of the regenerative electrode process as a fraction of the total photocurrent (varying from 100% for $i_{ox}/i_{ph} = 1$ to 0% for $i_{ox}/i_{ph} = 0$) as a function of the acidity of the electrolyte for N-CdS in combination with various redox couples (*a*) and for various semiconductors in combination with ferri/ferrocyanide (*b*). The contribution of the regenerative process was determined with the aid of a special ring/disc electrode.

CdSe or GaP and hence not so liable to dissolution. This is because photoexcitation of an electron transition from an anion-valence band to the cation-conduction band in polar semiconductors leads to the breaking of a polar bond. The stability should be less critical for transitions between energy bands resulting from d orbitals [17].

Selection rules for semiconductor-redox systems

The experimental results presented in Table II and in fig. 16 show that certain combinations of a semiconductor and a redox system are available in which the anodic dissolution of the semiconductor or the evolution of hydrogen can be avoided. The question now arises as to how successful these combinations are, measured in terms of the energy scheme of the type shown in fig. 11. This problem will be discussed separately for (anodically polarized) N-type and for (cathodically polarized) P-type semiconductors.

N-type semiconductor electrodes

If the ferri/ferrocyanide redox system is taken as an example, then Table II shows that CdS forms a stable electrode under illumination at pH = 7. A quantitative plot of the relative energy distribution at the interface (*fig. 17*) shows however that the valence band of the CdS and the occupied levels of the redox system do not overlap. It is in fact rather surprising that this particular combination is stable. It is equally surprising that GaAs and GaP are not really stable in this combination even though their valence bands do in fact overlap the occupied levels of the redox system. If we now consider CdS in combination with redox systems whose energy levels are much lower (e.g. Fe^{2+}/Fe^{3+} or Ce^{3+}/Ce^{4+}), we find that there is a good agreement between the

[16] R. Memming, Ber. Bunsen-Ges. phys. Chemie, in press.

[17] H. Tributsch (Ber. Bunsen-Ges. phys. Chemie **81**, 361, 1977) has recently started an investigation of the chalcogenides of transition metals, such as MoS_2.

location of the valence band and that of the redox system, but the semiconductor is unstable (in fig. 16*a* $i_{ox}/i_{ph} < 1$). Analysis of this and other combinations leads to the conclusion that a combination is only stable if the common Fermi level at the interface is located somewhere between the conduction and the valence band.

These experimental observations can only be understood in terms of the energy scheme by assuming that there are surface states within the band gap. This interpretation suggests that a redox system should be chosen whose Fermi level is near the middle of the band gap of the semiconductor.

The existence and formation of surface states can be interpreted as follows.

The dissolution of an *N*-type semiconductor requires holes produced by optical excitation. Several charges usually have to be transferred for the dissolution of a single surface molecule, e.g. six for GaP. The transfer of even the first or second hole leads to a considerable weakening of the surface bonds. The electron energy in such a surface group cannot be identical with the energy of an electron in the valence band, but should correspond to a higher energy level above the valence band ('surface state'). These surface states are therefore involved in the dissolution of the semiconductor electrode. They presumably lie close to the thermodynamically calculated potential for the dissolution of an electrode, somewhere in the middle of the band gap. An energy diagram including surface states is given in *fig. 18*.

If the energy levels of a redox system correspond to the energy levels of the surface states, then electrons from the occupied states of a redox system can be transferred to the surface states, and thus regenerate the surface bonds, which have been broken by trapping holes. This transfer seems to offer the best chance of preventing dissolution of the semiconductor.

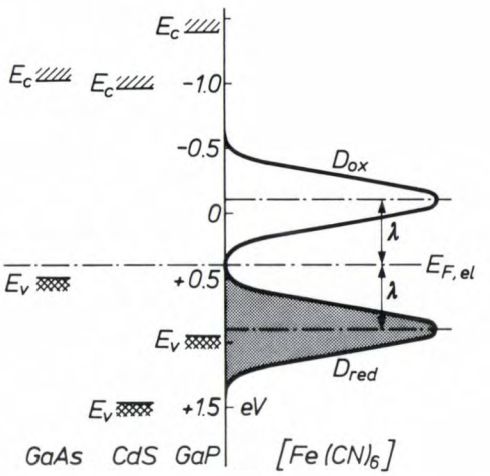


Fig. 17. Energy-level diagram of the interface of various semiconductor electrodes in combination with the ferri/ferrocyanide redox couple. Of the three *N*-type semiconductors shown, only CdS is stable in this combination, although the occupied energy levels of the redox system (*hatched*) lie between the bands of the CdS and do not correspond to the valence band. This indicates the significance of surface states (see fig. 18).

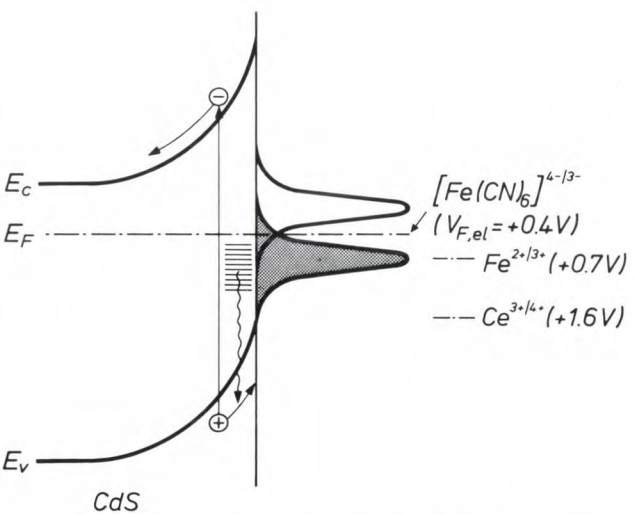


Fig. 18. Stable regenerative combination of *N*-CdS and ferri/ferrocyanide. Photoholes in the valence band are captured at the interface by surface states (*dashed*) that lie in the forbidden zone. The surface states have energies in the same range as the ferrocyanide levels (*grey*), so that the captured holes can be effectively neutralized by electrons originating from the ferrocyanide ions. When *N*-CdS is combined with ceri/cero or with ferri/ferro, then the holes in the surface states are not neutralized, or not so effectively, and the electrode dissolves. Stability is clearly achieved at the expense of less band-bending and a correspondingly lower energy-conversion efficiency.

This interpretation, however, has some rather undesirable consequences for the efficiency of solar cells. If an *N*-type electrode is combined with a redox system whose Fermi level is near the middle of the band gap as in *fig. 19*, the band-bending at equilibrium is much smaller than in the ideal case of figs 11 and 18. Consequently, the highest photovoltage possible now amounts to only $\frac{1}{2}\Delta E/e$, so that the theoretical efficiency will be only one half of that shown in fig. 10. The efficiency values obtained experimentally, 8.9% for *N*-GaAs/Se/Se^{2-} [11] and 7.8% for CdSe/S/S^{2-} [9], are quite close to the new theoretical maximum of 13%.

P-type semiconductor electrodes

If *P*-type semiconductors are to be used for solar cells the redox system must have energy levels that overlap the conduction band (see fig. 12). Transfer of excited electrons from the conduction band to a redox system seems to be much easier, since most redox systems selected on the basis of the normal energy diagram of fig. 4 will operate even with GaAs ($i_{red}/i_{ph} = 1$, see Table II). As an example, fig. 19 shows the energy diagram for the system Eu^{2+}/Eu^{3+}. The reorientation energy λ was estimated at 1 eV, a value similar to that found for the system Fe^{2+}/Fe^{3+}. The unoccupied levels of the Eu^{2+} ion give a good overlap with the edge of the conduction band of several semiconductors. The energy levels of the system H_2/H_3O^+ are not plotted in

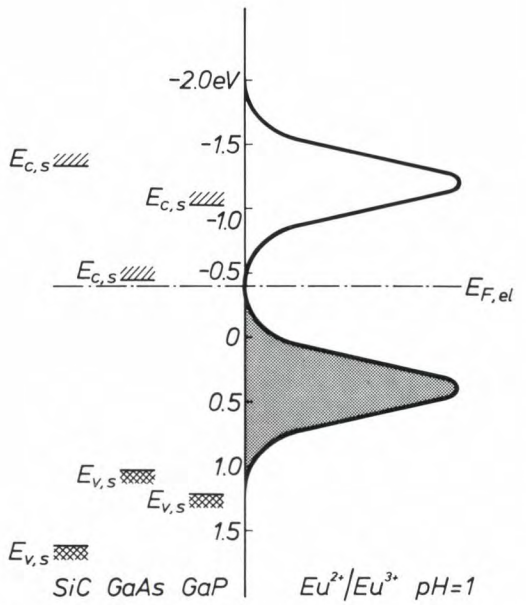


Fig. 19. Energy-level diagram of the combination of various semiconductors and the redox couple Eu^{2+}/Eu^{3+}. Since the redox potential is close to the conduction band, the band-bending is large for these *P*-type semiconductors in this combination (it is a maximum for GaAs). Photoelectrons also transfer from the conduction band to the empty levels of the Eu^{3+}/Eu^{2+} ions. In this apparently desirable situation, however, the majority charge carriers (again holes in the valence band) may cause complications (see fig. 20).

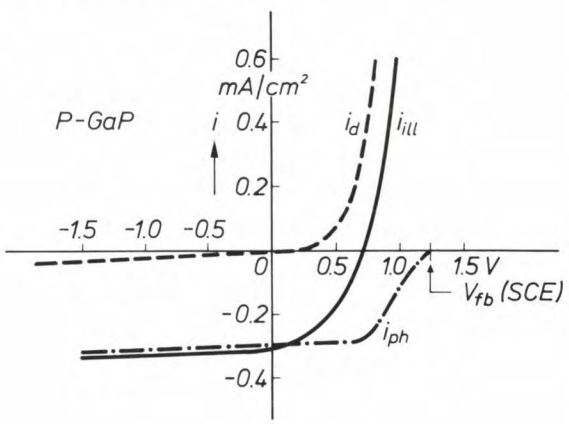


Fig. 20. Current-voltage characteristics for a *P*-GaP electrode with and without illumination. The dark current rises rapidly with increasing positive voltage, because holes (majority carriers) are pulled to the interface and give 'anodic' dissolution of the semiconductor. The same process also occurs during illumination under open-circuit conditions (no net current). In this situation two of the partial currents, the anodic dark current (*dashed curve*) and the cathodic photocurrent (*chain-dotted curve*) are equal, i.e. the partial anodic current gives rise to dissolution.

fig. 19, because their energy distribution may be considerably distorted by strong interaction between water and the crystal surface. Experimental results have shown that the reduction of a number of redox systems can compete efficiently without an accompanying evolution of hydrogen. It is even possible to use redox systems with a Fermi level close to the edge of the conduction band, so that there is considerable band-bending at equilibrium as shown in fig. 12 for GaAs/$Eu^{2+/3+}$. The combination with GaP gives a similar picture.

In practice, however, the situation is less promising than the above might suggest. The maximum photovoltages are in fact found to be much lower than expected: 0.8 V instead of 1.2 V for GaAs and 1.1 V instead of 1.6 V for GaP [12]. These low values can be understood if the current-voltage characteristics are considered, shown for *P*-type GaP in *fig. 20*. The dashed curve shows the dark current, while the solid curve was measured under illumination. The solid curve in the anodic region can be considered as the sum of two components: the dark current and a pure cathodic photocurrent, shown by the chain-dotted line, which ends at the flat-band potential V_{fb}. However, the anodic dissolution commences well below the flat-band potential, and the associated current compensates the cathodic photocurrent. This compensation effect not only decreases the photovoltage and the corresponding conversion efficiency, it also has the result that *P*-GaAs or *P*-GaP electrodes cannot be left open-circuited during illumination, since at zero net current the partial anodic and cathodic currents are equal, i.e. anodic dissolution would occur because of the photocurrent.

Cell characteristics

Apart from the aspects discussed above, an important characteristic of a solar cell is its photocurrent-photovoltage curve. An example obtained with *P*-GaP, Eu^{2+}/Eu^{3+} as a redox system and a gold-amalgam counter-electrode is shown in *fig. 21*.

As mentioned earlier, the highest overall conversion efficiencies have been obtained with systems such as GaAs/Se/Se^{2-} (8.5%) [11] and CdSe/S/S^{2-} (8.1%) [9]. With *N*-type electrodes the choice of redox systems is limited, so that the theoretical efficiency is less than 12%. *P*-type electrodes also have the limitation that they cannot be operated under open-circuit conditions during illumination.

So far only combinations that have been investigated experimentally have been discussed. To conclude, some attention should be given to proposals as yet untried.

A considerable increase in conversion efficiency should be possible from a cell with a common counter-electrode and several semiconductor electrodes with different band gaps. The electrodes should be arranged in series so that the light beam first encounters the semiconductor with the largest band gap. A question that arises is whether each electrode then operates under optimum conditions. No estimates have as yet been made.

Another attractive idea is to combine *N*- and *P*-type electrodes in a single cell, as has been suggested for the direct decomposition of water [18]. The results presented above suggest that a regenerative cell could be constructed with *P*-GaAs and *N*-CdS as electrodes and a solution of Eu^{2+} and hydroquinone as a redox system. The photoreactions at the electrodes are then

$2Eu^{3+} + 2e^- \rightarrow 2Eu^{2+}$ at the GaAs electrode and
$QH_2 + 2p^+ \rightarrow 2Q + 2H^+$ at the CdS electrode.

The regeneration in the electrolyte is expressed by

$$2Eu^{2+} + Q + 2H^+ \rightarrow 2Eu^{3+} + QH_2.$$

At first sight a considerable increase in the efficiency would be expected from the combination of *N*-type and *P*-type electrodes with different band gaps. However, a reasonable characteristic would only be expected if the anodic photocurrent at the *N*-type electrode is equal to the cathodic photocurrent at the *P*-type electrode. The two currents can be balanced by using electrodes of different area, which seems to be rather difficult.

Photoelectrolysis of water

Photoelectrochemical devices with an aqueous electrolyte containing no special redox system have also been studied. In cells such as these H_2O/H_2 and H_2O/O_2 are the active systems. The same basic principles apply as for the redox systems discussed. The performance of such a cell can be derived from the energy-level diagram shown in *fig. 22*. It is based on an *N*-type semiconductor with a certain amount of band-bending at equilibrium. On illumination electron-hole pairs are created; the holes are driven towards the surface and the electrons towards the bulk of the material. The semiconductor electrode is short-circuited to a metal counter-electrode. The holes are transferred at the semiconductor surface to water molecules that correspond to the unoccupied states of the H_2O/O_2 redox couple (only the normal potentials of the couples involved are shown in this energy diagram). The maximum photovoltage is determined by the band-bending. It is however not the intention that the reverse process (the conversion of oxygen into water) should occur at the counter-electrode, but that water should be reduced to hydrogen. This is only possible if the photovoltage at the semiconductor electrode is large enough to raise the Fermi level of the metal above the normal potential of the H_2O/H_2 couple. Electrons can then be transferred from the metal to the water, as shown in fig. 22. The electrode reactions are then

$H_2O + 2p^+ \rightarrow \frac{1}{2}O_2 + 2H^+$ at the semiconductor electrode,
$2H_2O + 2e^- \rightarrow H_2 + 2OH^-$ at the metal electrode.

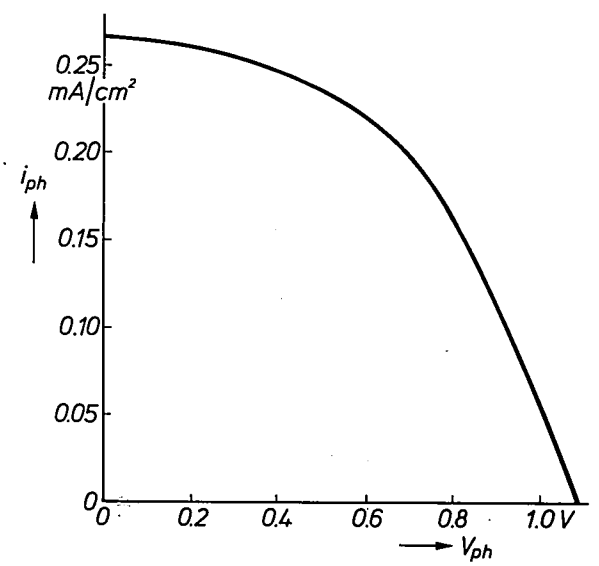


Fig. 21. Photocurrent/photovoltage characteristic for a *P*-GaP electrode in combination with a europium redox couple and a counter-electrode of gold amalgam. The conversion efficiency is at a maximum for the current and voltage that correspond to the maximum surface area of a rectangle *i-V*. Here, however, the electrode has already been severely attacked (see fig. 20).

Clearly, two conditions have to be met. On the one hand the flat-band potential of the semiconductor must be 'more negative' than the potential of the couple H_2O/H_2. The choice must therefore be made from semiconductors whose conduction-band energy is higher than E_{H_2O/H_2} in fig. 4. On the other hand the valence band should be below E_{HO_2/H_2}, since otherwise no transfer of holes from the valence band is possible. Both conditions can be fulfilled simultaneously if

$$E_g = E_c - E_v > E_{H_2O/H_2} - E_{H_2O/O_2} = 1.23 \text{ eV}.$$

A number of the semiconductors shown in fig. 4 do indeed satisfy these conditions, but in most cases anodic dissolution occurs instead of oxygen evolution. The question as to which semiconductors are sufficiently stable can to some extent be answered from the thermodynamic data in Table I and figs 13 and 14. There are in fact very few, such as TiO_2 [19,20], SnO_2 [21], Fe_2O_3 [22,23] and one or two titanates [24]. It has been shown experimentally that only these materials are stable and that oxygen is evolved in the anodic process. SnO_2 and Fe_2O_3 have to be ruled out, however, because their flat-band potential occurs at too 'anodic' a potential. Even TiO_2 is a doubtful case since its flat-band potential is only -0.1 V with respect to the H_2O/H_2 potential, so that a high light intensity is necessary to pass the H_2 potential.

These systems are very attractive because solar energy is converted into chemical energy in the form of hy-

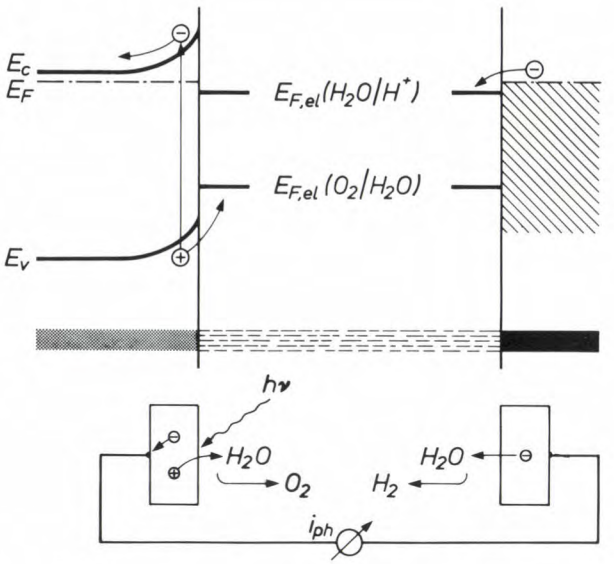


Fig. 22. Diagram illustrating the photoelectrolysis of water. Holes created by illumination are driven to the interface and react with water, generating oxygen. This is only possible if the redox potential of the H_2O/O_2 couple lies above the edge of the valence band. At the metal electrode, which is electrically connected to the semiconductor electrode, electrons are transferred to water and hydrogen generation takes place. The Fermi level, and hence the conduction band of the semiconductor, must therefore lie above the potential of the H_2O/H_2 couple. The potentials of these redox couples differ by 1.23 eV, so that the band gap of the semiconductor must also be at least 1.23 eV. Suitable semiconductors that are also stable (i.e. do not dissolve anodically instead of generating oxygen) all have such a large band gap that they absorb only ultraviolet light, so that the conversion efficiency is small.

drogen, and considerable effort has gone into the study of various systems [25–32]. The main difficulty is again in finding a stable semiconductor for which the anodic process under illumination is the evolution of oxygen and not electrochemical dissolution. Only semiconductors with a large band gap seem to fulfil this condition, and since such semiconductors only absorb ultraviolet light the overall conversion efficiency is very low.

K. Honda [33] has observed the evolution of hydrogen at the metal counter-electrode of a cell with a TiO_2 electrode. On the other hand, W. Gissler [34] was unable to detect any hydrogen evolution in such a cell. He only observed a current in the external circuit if oxygen was present in the electrolyte, and concluded that the cathodic process was only due to the reduction of oxygen.

It is also possible to combine a *P*-type semiconductor with a metal counter-electrode. Excited minority carriers are now used for the hydrogen evolution. If the flat-band potential is more anodic than the H_2O/O_2 potential, then oxygen is evolved at the counter-electrode. These cells have the advantage that corrosion should be avoidable [12]. Hydrogen evolution has indeed been observed in a cell with electrodes of *P*-GaP and platinum. The efficiency, however, was less than 1% [35].

Other cells with two semiconductor electrodes, a *P*-type cathode and an *N*-type anode, have been studied by A. J. Nozik [18]. However, there are still problems with stability, etc., as discussed above. Nozik has also investigated the energy aspect of photoelectrolysis [36]. The efficiency is determined here by factors quite different from those in regenerative cells. Since the photoelectrolysis cells are short-circuited and the output is hydrogen rather than electrical energy, the efficiency is mainly determined by the short-circuit current. As the quantum efficiency for the photocurrent can approach 100%, the efficiency of electrolysis cells can be high, provided that semiconductors of optimum band gap are used.

In principle other reactions can also be used for chemical storage. The decomposition of water has the advantage that the two products can easily be separated.

Sensitization of electrode processes

In the discussion the anodic dissolution of the semiconductor was the greatest problem. There are fewer problems with anodic dissolution for semiconductors with a large band gap. However, because of the large band gap these semiconductors do not absorb visible light, or do so only slightly. The question now arises as to whether it would be possible to sensitize electrode processes for visible light. Two approaches are feasible:
a) the introduction of energy levels into the band gap of the semiconductor by heavy doping;
b) using dye-redox systems, which absorb visible light.
The research has not as yet given any practical results, but many aspects have not yet been explored. The present situation is as follows.

For the first approach, it is well known from photoconductivity measurements that certain semiconductors can be heavily doped so that energy levels are formed near the middle of the band gap. An 'extrinsic' photoconductivity has been observed by excitation of

[18] A. J. Nozik, Appl. Phys. Lett. **29**, 150, 1976.
[19] D. Laser and A. J. Bard, Chem. Phys. Lett. **34**, 605, 1975. D. Laser and A. J. Bard, J. Electrochem. Soc. **123**, 1027, 1976.
[20] F. Möllers, H. J. Tolle and R. Memming, J. Electrochem. Soc. **121**, 1160, 1974.
[21] F. Möllers and R. Memming, Ber. Bunsen-Ges. phys. Chemie **76**, 469, 1972.
[22] K. L. Hardee and A. J. Bard, J. Electrochem. Soc. **124**, 215, 1977.
[23] L. S. R. Yeh and N. Hackerman, J. Electrochem. Soc. **124**, 833, 1977.
[24] A. B. Ellis, S. W. Kaiser and M. S. Wrighton, J. phys. Chem. **80**, 1325, 1976.
[25] K. L. Hardee and A. J. Bard, J. Electrochem. Soc. **123**, 1024, 1976.
[26] S. N. Frank and A. J. Bard, J. Electrochem. Soc., in press.
[27] Y. Nakato, S. Tonomura and H. Tsubomura, Ber. Bunsen-Ges. phys. Chemie **80**, 1289, 1976.
[28] Y. Nakato, K. Abe and H. Tsubomura, Ber. Bunsen-Ges. phys. Chemie **80**, 1002, 1976.
[29] H. Morisaki, M. Hariya and K. Yazawa, Appl. Phys. Lett. **30**, 7, 1977.
[30] J. M. Bolts and M. S. Wrighton, J. phys. Chem. **80**, 2641, 1976.
[31] G. Hodes, D. Cahen and J. Manassen, Nature **260**, 312, 1976.
[32] W. Gissler and R. Memming, J. Electrochem. Soc., in press.
[33] A. Fujishima and K. Honda, Bull. Chem. Soc. Japan **44**, 1148, 1971.
[34] W. Gissler, P. L. Lensi and S. Pizzini, J. appl. Electrochem. **6**, 9, 1976.
[35] M. Tomkiewicz and J. M. Woodall, J. Electrochem. Soc. **124**, 1436, 1977.
[36] A. J. Nozik, J. appl. Phys., in press.

electrons from these levels into the conduction band; see *fig. 23*. Examples are CdS doped with Cu [37,38] and GaP doped with Zn [39]. Electrochemical studies have been made of Cu-doped CdS combined with redox systems with a redox potential such as to ensure an electron transfer from the redox system to the copper levels during excitation (fig. 23). However, no intrinsic photocurrent could be detected [12]. Various experiments were also performed with TiO_2 strongly doped by various elements, also without success [20].

The second approach is illustrated in *fig. 24* by an energy diagram for the interface between a semiconductor and a dye-electrolyte. In this diagram the dye D is characterized by two molecular levels, the lower one occupied by two electrons, the upper one empty. The absorption of light produces an electron transfer from the lower level E_0 via the upper level E_u to the conduction band [1,40]. The electric field in the space-charge layer again forms the driving force that makes the electrons flow via an external circuit to a metal counter-electrode. The dye, which is oxidized at the semiconductor surface, should be reduced at the counter-electrode in a regenerative cell.

Experiments have shown, however, that most dyes are strongly adsorbed on the semiconductor surface and that the oxidized dye molecules are not replaced by new ones, so that the photocurrent vanishes within a relatively short time. This can be avoided by adding a reducing agent X that regenerates the oxidized dye molecules. The X^+ ions formed in this regenerative process can now diffuse towards the metal electrode, where they in turn are regenerated by accepting an electron from the metal. Expressing this formally:

$$\left.\begin{aligned} &D_{ads} + h\nu \rightarrow D^*_{ads} \rightarrow D^+_{ads} + e^- \\ &D^+_{ads} + X \rightarrow D_{ads} + X^+ \end{aligned}\right\} \text{at the semiconductor,}$$

$$X^+ + e^- \rightarrow X \qquad \text{at the metal.}$$

It is possible to make regenerative solar cells in this way. The conversion efficiency, however, will be extremely low because the only excited dye molecules that can take part in the electron transfer are those that reach the electrode within the lifetime of the excited state. For singlet states the lifetime is usually of the order of 10^{-9} s, which implies a diffusion length of 10^{-7} cm, so that only adsorbed molecules take part in the charge-transfer mechanism. Since an adsorbed dye monolayer only absorbs about 1 % of the incident light the conversion efficiency is very low.

A higher efficiency would be expected for dye molecules that exhibit a large intersystem crossing from the excited singlet state to the triplet state [41]. Triplet states have much longer lifetimes, but only in solutions that are free of oxygen. An interesting candidate is $Ru(bipy)_3^{2+}$ [42], a very stable and reversible compound,

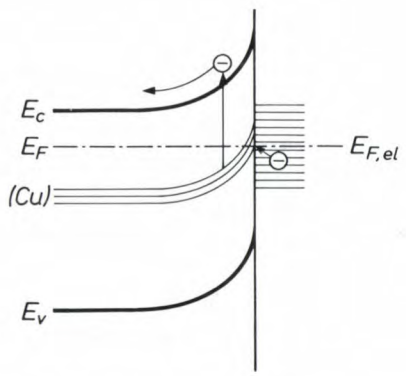


Fig. 23. Hypothetical model of the sensitization of semiconductor electrodes for light whose quantum energy is smaller than the band gap (ΔE), shown for strongly doped CdS as an example. The function of the valence band could be taken over by the narrow doping band in the forbidden zone. Although 'extrinsic' bulk photoconduction is observed in the doped CdS when illuminated by long-wave light, it is found that the doping band is not effective in photoelectrochemical cells.

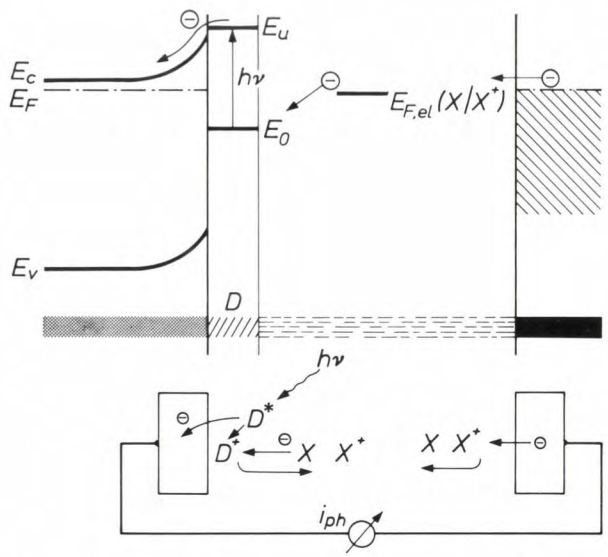


Fig. 24. An alternative approach to the sensitization of the electrode. Here a dye (D) which is adsorbed on the electrode absorbs long-wave light. Absorption of light is associated with excitation of an electron from the ground state E_0 to the excited state E_u, and is followed by electron transfer to the conduction band of the semiconductor. The dye is regenerated by X, the reduced component of the redox couple X/X^+. Unlike the dye, X/X^+ is able to diffuse freely and is in turn regenerated at the counter-electrode. Since only one monolayer of dye molecules, firmly adsorbed to the semiconductor, is operative, and it absorbs only about 1 % of the incident light, the conversion efficiency of such cells is low.

[37] M. Avinor, Thesis, Universiteit van Amsterdam 1959.
[38] H. G. Grimmeiss and R. Memming, J. appl. Phys. **33**, 2217, 1962.
[39] H. G. Grimmeiss, A. Rabenau and H. Koelmans, J. appl. Phys. **32**, 2123, 1961.
[40] H. Gerischer, in: Topics in current chemistry **61**, Springer, Berlin 1976.
[41] R. Memming, Photochem. Photobiol. **16**, 325, 1972.
[42] M. Gleria and R. Memming, Z. phys. Chemie Neue Folge **98**, 303, 1975.
[43] H. Tsubomura, Proc. Conf. on Semiconductor-liquid interfaces under illumination, Airlie 1977.

which has an extremely large intersystem crossing to the triplet state.

A few attempts have also recently been made to increase the photocurrent in cells with dyes by increasing the semiconductor surface area on a microscopic scale [43].

Summary. A survey is presented of the research on photoelectrochemical solar cells, which convert solar energy into electrical or chemical energy. These cells consist of a semiconductor electrode, an electrolyte and a metal counter-electrode. The nature of power generation by light excitation is similar to that in *P-N* junctions: electron-hole pairs are created by the illumination and separated by the electric field across the space-charge layer below the semiconductor surface. The minority carriers moving towards the surface take part in an electrochemical process. A regenerative cell suitable for power generation can be made if a redox system is added to the electrolyte. Its function and construction are described in terms of the energy levels. Problems arise because of side reactions such as the anodic dissolution. The stability problem and the selection of suitable semiconductor material are discussed in detail. In addition, photoelectrolysis cells that can be applied for the production of hydrogen are described. Stable semiconductor electrodes suitable for this kind of energy conversion have large band gaps, which limits the conversion efficiencies to rather low values.

Scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands *E*
Philips Research Laboratories, Redhill, Surrey, England *R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany *H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium *B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. *N*

C. Albrecht & A. K. Jansen: Numerical analysis of non-linear small-signal distortion in *p-n* structures.
IEEE Trans. **ED-24**, 91-98, 1977 (No. 2). *E*

P. Baudet, M. Parisot & R. Veilex: Amplificateur bas bruit à transistors F.E.T. au GaAs.
AGARD Conf. Proc. No. 197, pp. 3.1-3.16, 1977. *L*

J. W. M. Biesterbos, M. Brouha & A. G. Dirks: Thermomagnetic history effects in amorphous Y-Fe evaporated alloys.
Physica **86-88B**, 770-772, 1977 (Part II). *E*

H. van den Boom, A. J. J. van Dijsseldonk & J. C. M. Henning: Determination of exchange parameters of Cr^{3+} pairs in $MgAl_2O_4$ by optical measurements.
J. chem. Phys. **66**, 2368-2377, 1977 (No. 6). *E*

J. R. Brandsma & B. L. A. Waumans: The hardware simulator: a tool for evaluating computer systems.
IEEE Trans. **C-26**, 68-72, 1977 (No. 1). *E*

J. C. Brice & P. A. C. Whiffin: Changes in fluid flow during Czochralski growth.
J. Crystal Growth **38**, 245-248, 1977 (No. 2). *R*

F. G. Brockman: Development of ferrites and other ferromagnetic oxide ceramics.
Amer. Ceramic Soc. Bull. **56**, 216-218, 1977 (No. 2). *N*

R. Bruno, W. Hermann, H. Hörster, R. Kersten, K. Klinkenberg & F. Mahdjuri: Rationelle Energieverwendung und Nutzung der Sonnenenergie in Gebäuden.
Sonnenenergie II (Reihe: Forschung aktuell), Umschau Verlag, Frankfurt am Main 1977, 32 pp. *A*

K. H. J. Buschow: Changes of the magnetic properties upon hydrogen absorption in rare-earth iron compounds.
Physica **86-88B**, 79-80, 1977 (Part I). *E*

J. Cornet: Local order parameters of Te-rich As-Te glasses using X-ray and neutron diffraction and X-ray photoemission spectroscopy.
Proc. Int. Conf. on the Structure of non-crystalline materials, Cambridge 1976, pp. 17-22. *L*

J. J. Daniele, D. A. Cammack & P. M. Asbeck: CW GaAs/GaAlAs DH lasers grown by Peltier-induced LPE.
J. appl. Phys. **48**, 914-916, 1977 (No. 3). *N*

M. Davio & G. Bioul: Interconnection structure of injective counters composed entirely of *JK* flip-flops.
Information and Control **33**, 304-332, 1977 (No. 4). *B*

R. Dessert: Gunn diode oscillator and YIG resonator covering the band 18-26.5 GHz.
Proc. tech. Progr. Int. Microelectronics '76, Brighton 1976, pp. 167-171. *L*

A. M. van Diepen: Mössbauer linewidth of octahedral iron in magnetite.
Physica **86-88B**, 955-956, 1977 (Part II). *E*

A. M. van Diepen & F. K. Lotgering: Electron exchange between Fe^{2+} and Fe^{3+} on octahedral sites in spinels.
Physica **86-88B**, 961-962, 1977 (Part II). *E*

A. G. Dirks, J. W. M. Biesterbos & K. H. J. Buschow: Magnetic compensation temperatures of amorphous rare earth-cobalt alloys.
Physica **86-88B**, 761-763, 1977 (Part II). *E*

R. J. Dolphin & P. J. Pergande: Improved method for the analysis of estrogenic steroids in pregnancy urine by high-performance liquid chromatography.
J. Chromatography **143**, 267-274, 1977 (No. 3). *R*

J. W. F. Dorleijn & A. R. Miedema: The anomalous Hall effect in nickel alloys.
Physica **86-88B**, 537-538, 1977 (Part I). *E*

C. T. Foxon & B. A. Joyce: Interaction kinetics of As_2 and Ga on {100} GaAs surfaces.
Surface Sci. **64**, 293-304, 1977 (No. 1). *R*

G. Frank, R. Olazcuaga & A. Rabenau (Max-Planck-Institut für Festkörperforschung, Stuttgart): Occurrence of corundum-type indium(III) oxide under ambient conditions.
Inorg. Chem. **16**, 1251-1253, 1977 (No. 5). *A*

K. G. Freeman: Experimental direct broadcast reception of 12 GHz television signals from the Canadian Communications Technology Satellite.
Radio and electronic Engr. **47**, 234-236, 1977 (No. 5). *R*

Y. Genin & Y. Kamp: Comments on 'On the stability of the least mean-square inverse process in two-dimensional digital filters'.
IEEE Trans. **ASSP-25**, 92-93, 1977 (No. 1). *B*

P. J. van Gerwen, N. A. M. Verhoeckx, H. A. van Essen & F. A. M. Snijders: Microprocessor implementation of high-speed data modems.
IEEE Trans. **COM-25**, 238-250, 1977 (No. 2). *E*

W. S. M. Geurts van Kessel*, W. M. A. Hax, R. A. Demel* & J. de Gier* (* State University of Utrecht): High performance liquid chromatographic separation and direct ultraviolet detection of phospholipids.
Biochim. biophys. Acta **486**, 524-530, 1977 (No. 3). *E*

A. D. Giles & F. F. Westendorp (Mullard Magnetic Components, Southport, Lancs., U.K.): Simultaneous substitution of cobalt and titanium in linear manganese zinc ferrites.
J. Physique **38**, C1/47-50, 1977 (Colloque C1). *R*

A. D. Giles & F. F. Westendorp (Mullard Magnetic Components, Southport, Lancs., U.K.): The effect of silica on the microstructure of MnZn ferrites.
J. Physique **38**, C1/317-320, 1977 (Colloque C1). *R*

F. E. L. ten Haaf (Philips Scientific and Industrial Equipment Division, Eindhoven) **& M. L. Verheijke:** A simple mathematical model of a liquid scintillation counter.
Liquid Scintillation Counting **4**, 63-73, 1977. *E*

H. B. Haanstra & H. Ihrig: Voltage contrast imaging of PTC-type $BaTiO_3$ ceramics having low and high titanium excess.
Phys. Stat. sol. (a) **39**, K7-10, 1977 (No. 1). *E, A*

P. Hansen: Magnetostriction of Fe^{II} ions in yttrium iron garnet.
J. appl. Phys. **48**, 801-803, 1977 (No. 2). *H*

P. Harrop, R. Dessert & P. Baudet: Performance of some GaAs MESFET mixers.
Proc. 6th Eur. Microwave Conf., Rome 1976, pp. 8-13. *L*

J. C. M. Henning & H. van den Boom: Determination of exchange parameters of Cr^{3+} pairs in $MgAl_2O_4$ by optical and ESR measurements.
Physica **86-88B**, 1027-1028, 1977 (Part II). *E*

I. D. Higgins & R. Davies: An analytical comparison of two simple high-Q Gunn oscillators.
IEEE Trans. **MTT-25**, 412-416, 1977 (No. 5). *R*

B. Hoekstra, J. M. Robertson & R. P. van Stapele: Spin wave resonance in inhomogeneous garnet bubble films.
Physica **86-88B**, 1245-1246, 1977 (Part III). *E*

H. Ihrig & W. Puschert: A systematic experimental and theoretical investigation of the grain-boundary resistivities of n-doped $BaTiO_3$ ceramics.
J. appl. Phys. **48**, 3081-3088, 1977 (No. 7). *A*

B. A. Joyce & C. T. Foxon: Growth and doping kinetics in molecular beam epitaxy.
Proc. 8th Conf. (1976 Int.) on Solid state devices, Tokyo (Jap. J. appl. Phys. **16**, Suppl. 16-1), pp. 17-23; 1977. *R*

J. E. Knowles: Permeability mechanisms in manganese zinc ferrites.
J. Physique **38**, C1/27-30, 1977 (Colloque C1). *R*

A. J. R. de Kock, S. D. Ferris*, L. C. Kimerling* & H. J. Leamy* (* Bell Laboratories, Murray Hill, N.J.): SEM observation of dopant striae in silicon.
J. appl. Phys. **48**, 301-307, 1977 (No. 1). *E*

G. Kowalski: Reconstruction of objects from their projections. A simple reconstruction algorithm, theoretical and simulation studies.
EDV in Medizin und Biologie **8**, 1-8, 1977 (No. 1). *H*

G. Kowalski: Reconstruction of objects from their projections. The influence of measurement errors on the reconstruction.
IEEE Trans. **NS-24**, 850-864, 1977 (No. 1). *H*

B. M. Kramer, A. C. Derycke*, C. F. Masse & P. A. Rolland* (* Université de Lille I): High conversion efficiency frequency multiplier using GaAs avalanche diodes.
Proc. 6th Eur. Microwave Conf., Rome 1976, pp. 679-684. *L*

J.-P. Krumme, H. Heitmann & K. Witter: MOPS, a magneto-optic photoconductor sandwich.
Physica **89B**, 273, 1977. *H*

H. K. Kuiken: Solidification of a liquid on a moving sheet.
Int. J. Heat and Mass Transfer **20**, 309-314, 1977 (No. 4). *E*

H. Kurz, E. Krätzig, W. Keune*, H. Engelmann*, U. Gonser*, B. Dischler & A. Räuber**** (* Universität des Saarlandes, Saarbrücken; ** Institut für Angewandte Festkörperphysik, Fraunhofer-Gesellschaft, Freiburg/Breisgau): Photorefractive centers in $LiNbO_3$, studied by optical, Mössbauer and EPR methods.
Appl. Phys. **12**, 355-368, 1977 (No. 4). *H*

W. Kwestroo, H. C. A. van Gerven & H. A. M. van Hal: The hydrogarnets $Ba_3In_2(OH)_{12}$ and $Ba_3Sc_2(OH)_{12}$.
Mat. Res. Bull. **12**, 161-164, 1977 (No. 2). *E*

W. Kwestroo, H. C. A. van Gerven & C. Langereis: Compounds in the system BaO-In_2O_3.
Mat. Res. Bull. **12**, 157-160, 1977 (No. 2). *E*

D. E. Lacklison, G. B. Scott, A. D. Giles, J. A. Clarke, R. F. Pearson & J. L. Page: The magnetooptic bubble display.
IEEE Trans. **MAG-13**, 973-981, 1977 (No. 3). *R*

P. K. Larsen, M. Schlüter* & N. V. Smith* (* Bell Laboratories, Murray Hill, N.J.): Normal photoemission demonstration of two and three dimensionality of electron states in layer compounds.
Solid State Comm. **21**, 775-778, 1977 (No. 8). *E*

F. H. de Leeuw: Recent developments in the dynamics of magnetic domain walls and bubbles.
Physica **86-88B**, 1320-1326, 1977 (Part III). *E*

N. Maan & A. Broese van Groenou: Low speed scratch experiments on steels.
Wear **42**, 365-390, 1977 (No. 2). *E*

J. Magarshack: Device innovation in communication systems.
Microwave J. **20**, No. 2, p. 50, Feb. 1977. *L*

G. M. Martin, A. Mitonneau & A. Mircea: Electron traps in bulk and epitaxial GaAs crystals.
Electronics Letters **13**, 191-193, 1977 (No. 7). *L*

A. R. Miedema & J. W. F. Dorleijn: On the normal Hall effect in nickel alloys.
J. Physics F **7**, L27-31, 1977 (No. 1). *E*

G. F. Neumark, D. J. DeBitetto, R. N. Bhargava & P. M. Harnack: Auger coefficient of GaP(Zn, O): I. Evaluation from the luminescence decay, II. Evaluation from the temperature dependence of the luminescence efficiency.
Phys. Rev. B **15**, 3147-3155, 3156-3162, 1977 (No. 6). *N*

J. M. van Nieuwland: Eigentrillingen in het geluidveld in niet-rechthoekige galmkamers.
Publ. Ned. Akoest. Gen. No. 40, pp. 59-73, 1977. *E*

J. A. Pals, W. van Haeringen & M. H. van Maaren: Josephson effect between superconductors in possibly different spin-pairing states.
Phys. Rev. B **15**, 2592-2599, 1977 (No. 5). *E*

G. Piétri: Contribution of the channel electron multiplier to the race of vacuum tubes towards picosecond resolution time.
IEEE Trans. **NS-24**, 228-232, 1977 (No. 1). *L*

A. Pirotte: Data base models for natural language users.
Int. Workshop on the Cognitive Viewpoint, Ghent 1977, pp. 165-172. *B*

C. L. Sam & M. M. Choy: Mixing of N_2 laser and dye laser pulses in ADP to generate difference frequency tunable from 0.68 to 1.1 μm.
Appl. Phys. Letters **30**, 199-201, 1977 (No. 4). *N*

J. G. van Santen: Solid state image sensors using the charge transfer principle.
Proc. 8th Conf. (1976 Int.) on Solid state devices, Tokyo (Jap. J. appl. Phys. **16**, Suppl. 16-1), pp. 365-371; 1977. *E*

W. Schäfer, G. Will (both with Universität Bonn) **& K. H. J. Buschow:** Magnetic structure and magnetic properties of TbMg.
J. Magnetism and magn. Mat. **4**, 28-35, 1977 (No. 1-4). *E*

P. J. Severin & H. Lydtin: Chemical vapor deposition and isotope separation.
J. Electrochem. Soc. **124**, 140-146, 1977 (No. 1). *E, A*

J. W. Slotboom: The *pn*-product in silicon.
Solid-State Electronics **20**, 279-283, 1977 (No. 4). *E*

N. V. Smith*, P. K. Larsen & M. M. Traum* (* Bell Laboratories, Murray Hill, N.J.): Miniature plane mirror analyser suitable for angle-resolved photoelectron spectroscopy.
Rev. sci. Instr. **48**, 454-459, 1977 (No. 4). *E*

J. L. Sommerdijk, A. Bril & F. M. J. H. Hoex-Strik: Luminescence of Eu^{2+}-activated Cs(Ca, Mg)F_3 and Rb(Ca, Mg)F_3.
J. Luminescence **15**, 115-118, 1977 (No. 1). *E*

M. J. Sparnaay: Membrane models.
Electrical phenomena at the biological membrane level, Proc. 29th Int. Meeting Soc. Chim. Phys., Orsay 1976, pp. 43-58; 1977. *E*

W. T. Stacy, H. Logmans & A. B. Voermans: Induced orthorhombic anisotropy in (Eu, Lu)$_3Fe_5O_{12}$ epitaxial films.
Physica **86-88B**, 1455-1456, 1977 (Part III). *E*

G. H. A. M. van der Steen (Philips Glass Development Centre, Eindhoven) **& H. van den Boom:** Raman spectroscopic study of hydrogen-containing vitreous silica.
J. non-cryst. Solids **23**, 279-286, 1977 (No. 2). *E*

F. J. van Steenwijk, W. J. Huiskamp, H. Th. LeFever, R. C. Thiel (all with Kamerlingh Onnes Laboratory, Leiden) **& K. H. J. Buschow:** Conduction electron spin polarization in some intermetallic europium compounds.
Physica **86-88B**, 89-90, 1977 (Part I). *E*

E. Stulemeijer: De emittervolger als teruggekoppelde versterker.
Polytechn. T. Elektr. **32**, 79-86, 1977 (No. 2). *E*

D. R. Terrell: Photo and dark conduction of crystal violet-doped (*N*-vinylcarbazole containing) polymers for single layer photothermoplastic devices.
Photogr. Sci. Engng. **21**, 66-72, 1977 (No. 2). *H*

A. Thayse: Static-hazard detection in switching circuits by prime-implicant examination in fuzzy functions.
Electronics Letters **13**, 94-96, 1977 (No. 4). *B*

J. B. Theeten, R. Madar, A. Mircea-Roussel, A. Rocher & G. Laurence: Cadmium telluride layer deposition using the ionized cluster beam technique.
J. Crystal Growth **37**, 317-328, 1977 (No. 3). *L*

T. G. Verbeek (Philips Lighting Division, Eindhoven) **& H. van Tongeren:** Positive column of Na-Ne-Ar ac discharges.
J. appl. Phys. **48**, 577-580, 1977 (No. 2). *E*

L. Vriens: Near-resonant dye laser scattering and two-photon absorption in a sodium neon plasma.
J. appl. Phys. **48**, 653-661, 1977 (No. 2). *E*

L. E. Zegers & C. B. Dekker: A comparison of digital transmission techniques for standard FM mobile radio sets.
IEEE Trans. **COM-25**, 364-368, 1977 (No. 3). *E*

Enclosure of Philips Technical Review
Beilage der Philips Technischen Rundschau
Bijlage van Philips technisch Tijdschrift

PHILIPS

June 1979

Recent United States Patents

Abstracts from patents that describe inventions from the following research laboratories that form part of or cooperate with the Philips group of companies:

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

4 092 208

Method of growing single crystals of rare earth metal iron garnet materials

J. C. Brice *R*
J. M. Robertson

A method of growing monocrystalline bismuth rare earth iron garnet, either as a single crystal or as an epitaxial layer, from a solution containing composing components of the garnet together with a flux. The flux consists essentially of a mixture of Bi_2O_3 and RO_2, wherein R is at least one of the elements Si, Ge, Ti, Sn, Zr, Ce and Te, wherein the system Bi_2O_3—RO_2 includes a eutectic composition having a eutectic temperature which is below the melting temperature of pure Bi_2O_3. By using these Bi_2O_3—RO_2 fluxes, the monocrystalline garnets produced have lower optical absorption coefficients at, for example 5100 Å and 5600 Å than similar garnets grown using lead-containing fluxes. Furthermore higher growth rates and higher growth temperatures are possible when using the Bi_2O_3—RO_2 fluxes rather than lead-containing fluxes.

4 099 122

Transmission system by means of time quantization and trivalent amplitude quantization

M. C. W. van Buul *E*

Indicated as hybrid D-PCM a new transmission system is given which is based on differential pulse code modulation (DPCM) with signal compression at the transmitter end and signal expansion at the receiver end. In the transmitter defined auxiliary information is fed to the DPCM code groups to obtain composite code groups, while in the receiver a local auxiliary signal derived from a reconstruction device is subtracted in a separating device from the received composite signal. With a particularly simple construction hybrid D-PCM is distinguished as compared with DPCM by effective perturbation reduction, in addition to which a pulse can be saved in the transmitted code groups.

4 099 136

Amplifier circuit for high frequency signals, particularly for cable distribution systems, comprising at least a first transistor controlled at its base electrode by a signal source, and a difference amplifier

D. Blom *E*

An amplifier circuit for high freqency signals, particularly for cable distribution systems, comprising at least a first transistor controlled at its base electrode by a signal source, and a difference amplifier in which the distortion produced across the base-emitter junction of the said transistor is eliminated in a simple manner by means of feedforward control.

4 099 840

Multilayer reflector for gas discharge laser

J. van der Wal *E*
W. A. J. Gielens
J. M. M. Pasmans

A multilayer reflector of which the last layer most remote from the substrate and having a high refractive index is vapor-deposited from a mixture containing a metal or oxide of an element from group 4b of the periodic table and a metal oxide from the group of beryllium oxide, magnesium oxide and calcium oxide, is particularly resistant to hydrogen and ultraviolet light and is hence suitable for application in a laser which is to have a long life.

4 100 378

Cross-correlation arrangement

T. A. C. M. Claasen *E*
J. B. H. Peek

Cross-correlation arrangement for determining the presence or absence of a signal $x_2(t)$ of known frequency in an unknown signal $x_1(t)$ if this signal $x_2(t)$ is present for at least a period of time τ_o with an unknown starting instant. An auxiliary signal $a_1(t)$ is added to the signal $x_2(t)$. The sum signal is converted into one-bit code words $s(nT)$ which are multiplied in at least a first correlation channel by one-bit code words $y_{per}(kT)$ which are stored in a storage medium having at least $N/2$ addressable storage locations. The code words $y_{per}(kT)$ are elements of a series $\{y_{per}(kT)\}$ which is formed by the code words which occur in a predetermined time interval having a length NT and which are of a nonperiodical signal of the form sign $[x_2(t) + a_2(t)]$. The product code words obtained are added together in an integrator. This integrator is read each time after a time interval NT and reset. The absolute value of an output code word of the integrator is optionally added to corresponding code words of other correlation channels. The sum code words obtained in this way are integrated in a second integrator whose contents is continuously compared with a detection level. If $N_o = \tau_o/T$, N is smaller than N_o for example a factor of 3. The value of N determines the width of the detection characteristic.

4 100 409

Device for analyzing a surface layer by means of ion scattering

H. H. Brongersma E

A device for analysing a surface layer by means of ion scattering. The device comprises an energy selector having two coaxial cylindrical electrodes. A primary mono-energetic ion beam impinges upon the surface layer and its axis coincides with the axis of the cylindrical electrodes. Back-scattered ions, the paths of which lie on a conical surface having an apical angle of 180° reduced with the scattering angle, are selected for energy and detected.

---

4 100 427

Device for converting solar energy

H. Durand L
G. J. Naaijer

A device is described for converting solar energy into electric power for a load, which device is provided with photocells and electric accumulators. Switching means are provided which at least during starting of the load connect this load to the output terminals of the array of photocells so that during starting the current through the load is determined by the current supplied by the photocells. The device makes efficient use of the available solar radiation, is of simple design and is reliable.

---

4 100 436

Current stabilizing arrangement

R. J. van de Plassche E

A current stabilizing arrangement includes a first and a second current circuit or path in which currents with a mutually fixed ratio are maintained. These currents respectively flow through the series connection of a first semiconductor junction in series with a resistor and a second semiconductor junction. The voltage across the second semiconductor junction is maintained equal to the voltage across said series connection, which results in currents which are linearly dependent on the temperature. In order to add a component with a positive second-order temperature dependence to these currents so as to enable a negative second-order temperature dependence to be compensated for in the case where the arrangement is used as a voltage or current reference source, the current stabilizing arrangement comprises a transistor whose base-emitter junction constitutes said second semiconductor junction, the base circuit of said transistor including a resistor.

Re.29.700
Reissue of 4 004 247

Voltage-current converter

R. J. van de Plassche E

A voltage-current converter having a balanced-voltage input and a balanced-current output. Each of the emitter circuits of the two input transistors includes a three-transistor current-mirror circuit, whose input circuit carries a constant current, whose output circuit carries the balanced-output current and of which the base-emitter junction of the transistor which is included in the input circuit is bridged by a semiconductor junction in the output circuit. The two current mirror circuits are coupled in that a resistor is included between the bases of the transistors which are included in the input circuits of the two current mirror circuits.

---

4 100 493

Transmission system for signal transmission by means of compressed delta modulation

E. C. Dijkmans E
K. E. Kuijk

Transmission system for signal transmission by means of compressed delta modulation provided with leaking integrators in the delta modulator and demodulator, the leaking integrator in the modulator being connected in series with the capacitor of large capacitance between the terminals of a signal source having controlled current sources for reducing the noise and distortion produced by the modulator.

---

4 100 494

Dynamic control of code modulated pulse transmission

H. P. J. Boudewijns E
K. Riemens
L. D. J. Eggermont

A transmission system for the coded transmission of information signals in which the transmitter and receiver comprise a dynamic control circuit, in which a signal generator is included for supplying a dynamic control signal. When a pulse train analyzer which analyzes the transmitted pulse train delivers a series of one-pulses, an exponential variation of the dynamic control signal is obtained by actuating a feedback circuit which applies the dynamic control signal to the input of the signal generator.

---

4 100 580

Facsimile system

H. H. H. Groothuis E

A facsimile system including a picture pick-up arrangement in which the picture information to be transmitted is processed in separate groups of picture elements. The information of a first number of adjacent lines and a second number of adjacent picture elements is applied to a picture element comparison circuit. When only white information occurs in a group of picture elements thus formed this picture element group is represented by a logical 1. When there is black information in the picture element group an instantaneous derived code signal is formed with a logical 0 followed by a number of bits suitable for picture element coding. A considerable bit reduction is the result as compared with a one-bit representation per picture element. A specific choice of a clock pulse frequency results in a simple signal coding.

---

4 100 583

Thin-film magnetic head for reading and writing information

G. J. Koel E
F. W. Gorter
J. T. Gerkema

A thin-film magnetic head for writing information in and reading information from a track of a magnetic recording medium in which the head comprises a magnetically permeable yoke having two limbs between which a writing and reading gap is formed. In a construction in which the limbs of the yoke are opposite to each other, one limb is interrupted by a gap which is bridged by a magneto-resistive reading element. In a construction in which the limbs of the yoke overlap each other at the area of the writing/reading gap, they are bridged by a magneto-resistive reading element in another place.

---

4 100 741

Hot-gas engine

A. P. J. Michels E

A hot-gas engine in which the supply of fuel to the burner device is controlled by means of a control signal which is derived from a differential pressure signal which represents the volume flow of combustion air and which is corrected for variations in temperature and pressure of the ambient air.

4 101 185

Method for producing a low-pressure gas discharge lamp

J. van Esdonk E
J. Hasker
J. C. G. Vervest

A method of producing a low-pressure gas discharge lamp in which a body of a solid material having a thin structure of filaments permeable to the gas discharge such as glass wool is disposed. Coherence is given to the filaments by compressing them whereafter they are sucked into the lamp envelope. Compressing, knitting together and producing the wire-shaped elements is done in a funnel-shaped space which is connected at its constricted side to the tubular discharge space.

---

4 101 344

Method of manufacturing a semiconductor device

E. Kooi E
P. J. W. Jochems
A. T. van Zanten

In LOCOS *N*-channel MOST-IC's underpasses can be obtained below the LOCOS pattern by performing, at the area where the underpasses are to be formed, an As or Sb implantation prior to providing the LOCOS. By using the nitride mask as an implantation mask, the LOCOS and the source/drain zones of the transistors can be provided in a self-registering manner with respect to the underpasses.

---

4 101 779

Apparatus for mechanically supporting a medical device in a plane

W. Schmedemann H

Apparatus comprising a parallelogram which is movable in one plane and which is formed by a first and a second coupling rod and a part of a first and a second supporting rod for a medical device and a counterweight therefor, respectively. The coupling rods are pivotably connected near one end to the first supporting rod and are pivotably connected to the second supporting rod near their other end. The parallelogram is pivotable about a stationary shaft which extends perpendicularly to the plane of movement. This shaft is situated between the two pivots of a coupling rod.

---

4 101 908

Adjustable mount for electrostatic printer stylus within cylindrical printing roller

H.-D. Hinz H
U. Rothgordt
F. Schinke

An electrode stylus which is arranged in an aperture of a rotatable printing roller is secured in a cylindrical insert of the roller by means of an electrode holder. The electrode holder comprises a constriction in the center, so that it is sub-divided into an adjusting portion which supports the electrode and which can be slightly bent, by means of adjusting screws, transversely of the jacket surface of the roller, and a fixing portion which can be fixed in an aperture in the insert by means of a fixing screw. The adjustment of the height of the stylus is effected, after the loosening of the fixing screw, by an arm which can be inserted into the insert and which is radially pivotable therein relative to the roller axis. Microscopic adjustment of the electrode stylus in the tangential as well as the radial direction relative to the roller is thus possible.

---

4 102 559

Optical device for coupling optical fibers to each other and/or a light source

J.-J. Hunzinger L

Two very thick plano-convex lenses with a common axis are positioned with their convex surfaces facing each other, and a radiation source and the input face of a fiber are each disposed on one of these surfaces in the vicinity of the axis. The main feature of the device resides in the considerable thickness of the lenses relative to the radii of curvature of the dioptric faces. As a result, aberrations, in particular spherical aberrations, are reduced while the dimensions of the device are such as to enable easy construction and handling.

---

4 104 185

Latent heat accumulator

J. Schröder A

A heat accumulator in which the heat-energy storage medium consists essentially of a potassium fluoride-water system having a potassium fluoride content between 44 and 48 % by weight.

---

4 106 293

Stirling cycle machine with a control device for supplying working medium

J. H. Abrahams E

A multi-space Stirling cycle reciprocating machine having a control device for supplying working medium to the working spaces. A pressure-controlled annular body in the device allows passage of working medium to the working spaces, exclusively by shape changes, each time at the relevant maximum cycle pressure.

---

4 106 348

Device for examination by means of ultrasonic vibrations

M. J. Auphan L

A device for ultrasonic examination, comprising a flat mosaic of a row of transmission transducers and a number of rows of receive transducers which are symmetrically arranged with respect thereto. The device furthermore comprises means for applying actuation pulses to one transmission transducer at a time, and means for enabling receive transducers in the time interval following each actuation pulse, and for at the same time displaying on an image display device an image of a section of the examined body in a plane perpendicular to the plane of the mosaic and through the line of transmission transducers.

---

4 107 189

Liquid crystalline compounds

J. van der Veen E
T. C. J. M. Hegge

Liquid crystalline 2,3-diphenyl acrylonitrile derivatives, i.e., 2-(*p*-ethoxyphenyl)-3-(*p*-hexyloxyphenyl) acrylonitrile, exhibit dynamic scattering in an electrical field and are useful in electro-optic displays and as solvents in E.S.R. and N.M.R. spectroscopy.

---

4 107 387

Resistance material

A. H. Boonstra E
C. A. H. A. Mutsaers

A resistance material consisting of a mixture of metal oxide and/or metal oxidic compounds and any metals with a binder, which material comprises a metal rhodate of the type $M_3Rh_7O_{15}$, by way of resistance determining component. This material has a small negative TCR which is substantially constant in a large temperature range.

---

4 107 569

Color selection means comprising lens electrodes spaced by grains of insulating material

H. J. Ronde E

In an electrode assembly of at least two electrodes connected together in an insulating manner, said electrodes are kept at a defined distance by means of grains of an electrically insulating

material. Said grains are sunk on two sides in layers of an adhesive material present on the facing surfaces of the electrodes and the sum of the thicknesses of which is smaller than the distance between the relevant electrodes.

---

4 107 724

Surface controlled field effect solid state device

J. E. Ralph *R*

An electronic solid state device comprising an inhomogeneous body in which grains of semiconductor material are present. The contact between adjoining grains is such that in the interface regions there is continuity between the bulk material of the adjoining grains and the dominant current conduction paths in the body lie within the grains via said grain to grain contacts. At the surfaces of the grains adjoining said interface regions rectifying barrier forming means are provided to enable control of the said current paths by surface field effect depletion in the interface regions. Said means can be opposite conductivity type surface layers at or on the grains. In other devices said means are formed by adsorbed gas atoms, ions or molecules. The said control may be employed to yield photoconductive layers with controllable gain and speed of response, said layers being used, for example, in solid state imaging devices.

---

4 109 203

Delta-modulation encoder

L. D. J. Eggermont *E*
E. C. Dijkmans
K. Riemens

In a delta-modulation encoding arrangement having step-size control drift voltages and miss-match between the charge and discharge pulses for the first integrating network in the feedback circuit produce an additional DC voltage on this integrating network. This additional DC voltage is compensated by an additional feedback circuit comprising a second integrating network. However, if the signal to be encoded rapidly decreases from a large to a small amplitude and when after this rapid decrease the information signal only has a small amplitude an additional distortion is introduced for these small signals. To obviate this a modulation arrangement is included in the extra feedback circuit, which modulation arrangement is controlled by a control signal which derives a control signal from the delta modulation output pulses. The energy contents of the delta modulation pulses which are applied through the modulation arrangement to the second integrating network is varied in such a way that this energy contents decreases when the step-size increases and vice versa.

---

4 110 636

Feeding bridge

E. C. Dijkmans *E*

A bridge circuit provided with an isolation transformer having four primary windings which are connected such that the voice currents do and the supply currents do not generate a resulting flux in the core of the transformer and which comprises two transistor circuits. These circuits on the one hand prevent the flow of voice currents via a supply source connected to the bridge and on the other hand balance the bridge circuit so that the influence of longitudinal noise signals which are produced in a transmission line connected to the bridge circuit are suppressed.

---

4 110 777

Charge-coupled device

L. J. M. Esser *E*
M. J. J. Theunissen

The invention relates to a charge-coupled device in which the charge transport takes place in the form of majority charge carriers via the bulk of a surface layer of the first conductivity type which forms a *P-N* junction with a substrate of the second conductivity type. The comparatively thick and high-ohmic surface layer has a comparatively thin low-ohmic buried zone of the first conductivity type which adjoins the said *P-N* junction. The buried zone forms a buffer layer against the depletion zone belonging to the *P-N* junction. Without detrimentally influencing the transport properties, a low-ohmic substrate may be used, which has important advantages in particular with respect to the leakage currents.

---

4 111 002

Cyclic desorption refrigerator and heat pump, respectively

H. H. van Mal *E*
E. T. Ferguson

A cyclic desorption refrigerator and heat pump, respectively, having a plurality of thermodynamic units, each with a first and second chamber, the first chambers comprising the same sorbent and the second chambers comprising different sorbents, the heat of absorption released in a second chamber being used for desorption in an adjacent second chamber.

---

4 111 970

Method of producing chromium (III)-N-acyl-anthranilates

D. J. Zwanenburg *E*

The efficiency of chromium (III)-N-acylanthranilates as control materials for electrophotography is improved when a N-acyl-anthranilic acid is reacted in a methanolic solution, in the presence of sodium methanolate, with a chromium (III) salt, the resulting precipitate is dissolved in toluol and a part of the toluol is vacuum distilled.

---

4 112 345

Stepping motor with energizing circuit

B. H. A. Goddijn *E*

A stepping motor provided with an energizing circuit for simultaneously energizing two phase windings from current sources. For actively damping the rotor movements an active network has been provided which measures the difference between the voltages across the two phase windings and in accordance with a suitably selected function converts it into two currents which are in phase opposition to each other, and which are applied to the two phase windings.

---

4 112 411

Device for echography by means of focussed ultrasonic beams

P. Alais *L*
M. J. Auphan

A device in which focussed ultrasonic beams are used by way of binary distribution of the phase of the signals occurring in a mosaic of transducers. The material of which the transducers are made can be polarized, which enables the prefocussing of the ultrasonic beam to be transmitted or received, and also enables point-wise scanning and scanning by double focussing according to the X-direction and the Y-direction.

---

4 112 563

Color display tube and method of manufacturing same

J. van Esdonk *E*

A method of making a color selection electrode for a color television display tube in which apertured electrodes are placed in contact with metallic coatings on opposite sides of an insulating foil and diffusion bonded thereto by heat and pressure, and thereafter the metallic coatings and insulating material selectively etched to form apertures corresponding to the electrodes therein.

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79 No. 7/8

The PM 2517 automatic digital multimeter

M. P. van Alphen, R. E. J. van de Grift, J. M. Pieper and R. J. van de Plassche

Digital multimeters have been on the market for a long time, and are not in themselves particularly newsworthy. However, the recently introduced Philips PM 2517 multimeter deserves our special attention because it contains a number of interesting and original circuits. Besides an extremely accurate analog/digital converter, these include various other analog circuits, such as an accurate, programmable current generator for resistance measurements and a circuit that determines the r.m.s. value of rectified alternating-current quantities. These circuits, all of them ICs, were designed right from the start for automatic selection of the measurement ranges; the switching required can often be performed in the IC itself. The multimeter is the result of close cooperation between Philips Research Laboratories, the Scientific and Industrial Equipment Division and the Solid-state Special Products Group of the Elcoma Division.

An integrated multimeter

For many years the customary central element in a multimeter has been a moving-coil meter. This is surrounded by various circuits that reduce the measurement of voltages and resistances to the measurement of a current. In addition to this 'analog' multimeter digital multimeters are now available, and these have a number of advantages. The recently introduced Philips PM 2517 automatic digital multimeter is such an instrument; it is portable and gives a reading to four decimal places. *Fig. 1* shows the two versions available, one with light-emitting diodes and the other with a liquid-crystal display.

In the conventional multimeter the force experienced by the moving coil in the magnetic field balances the opposing force of a spring. The magnetic field and the spring constant can be considered here as the reference quantities. In a digital multimeter the reference quantity is necessarily of an electrical nature and not mechanical. In the PM 2517 an accurately constant reference current is used. This is balanced against an input current derived from the test signal. The currents are balanced by applying each to one of the plates of a capacitor; when the currents are equal the voltage across the capacitor is zero. The number of pulses of the reference current required to produce a balance gives the desired information in digital form. The stability and accuracy of the circuit are high; the deviations from linearity are less than $\pm 10^{-5}$ of the full-scale reading.

In moving-coil instruments there is usually a selector switch for setting to the various measurement ranges. As might be expected with the shift from mechanical to electrical functions that is characteristic of digital instruments, this switching is also electronic. The analog/digital converter of the PM 2517 is designed for compatibility with electronic switching; it has two states that differ in sensitivity by a factor of ten, and a choice between them can be made electronically. With this arrangement it is only necessary to switch the voltage-divider resistances once, which can be done with simple electronic circuits.

The digital indication of resistance values is a significant improvement over the hyperbolic resistance scale usually found in moving-coil meters. A voltage is generally used in these meters as the reference quantity, and the meter — which measures current — then indicates the reciprocal of the resistance from $I = V/R$.

Ir M. P. van Alphen and Ir J. M. Pieper are with the Philips Scientific and Industrial Equipment Division at Almelo; R. E. J. van de Grift and Ir R. J. van de Plassche are with Philips Research Laboratories, Eindhoven.

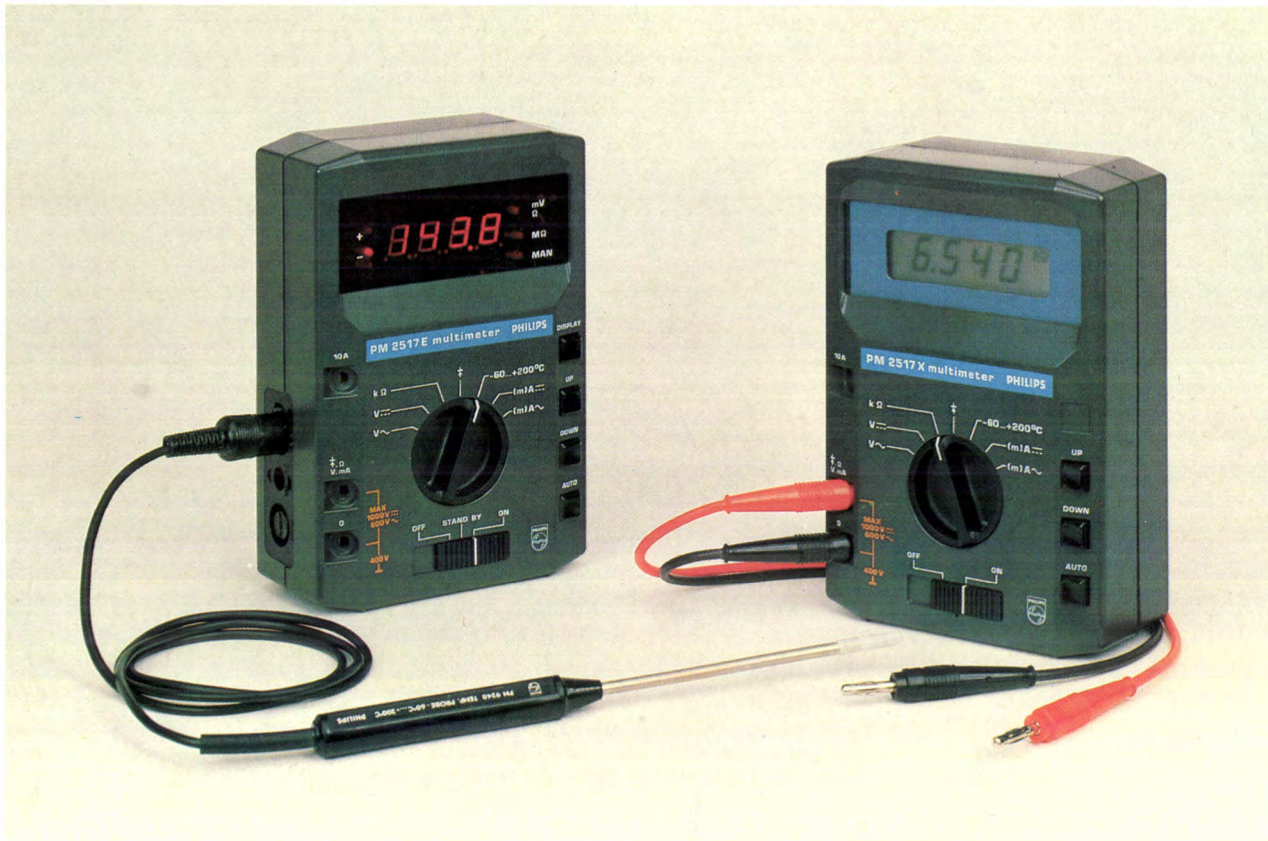


Fig. 1. The PM 2517 automatic digital multimeter. The version on the left (PM 2517E) has a display with light-emitting diodes (LEDs); the one on the right (PM 2517X) has a liquid-crystal display (LCD). A temperature probe (PM 9248) is connected to the instrument on the left; the one on the right has test leads.

The development of a current generator that can be accurately set to various values has enabled us to use a current as reference in the PM 2517, so that the resistance measurement can be reduced to a simple voltage measurement.

The situation is rather similar for the root-mean-square (r.m.s.) value of alternating voltages and currents. Moving-coil meters react to the mean value of the rectified current; they usually have a separate scale, calibrated as the r.m.s. value for sinusoidal alternating currents. This scale reads incorrectly for other waveforms. The PM 2517 includes a circuit that calculates the true r.m.s. value independently of the waveform. This calculation is made by analog methods.

It is only because of the use of integrated circuits that there is room for the many operations that are performed in the multimeter. Six special ICs have been developed for the PM 2517, four of them analog and two digital. A bipolar technology has been used in the four analog ICs; the two digital ICs have been made in LOCMOS (*L*ocal *O*xide separating *C*omplementary *MOS* transistors). The analog ICs in particular contain new circuits, which have been specially designed to meet the requirements for high accuracy. Some unconventional solutions have been found; we shall discuss the more important of these circuits below.

A number of discrete components have been added to the ICs, such as voltage dividers for the various measurement ranges, shunt resistors for current measurement and variable resistors for various adjustments (see *fig. 2*). Discrete components are also included for instrument protection, an important aspect we have not as yet mentioned. When it is switched on an automatic multimeter sets itself to the most sensitive range position: for voltage measurement this is the 1 V range, for current measurement the 100 mA range. If the mains voltage should now be applied, for example, this must not damage the instrument. Many precautions have been taken to prevent such damage. These are so effective that no damage results even if the EHT voltage of a television receiver is accidentally applied. We shall return to instrument protection later.

However, most of the rest of this article will be a description of some of the interesting new circuits in the multimeter. To show clearly the place of these circuits in the complete instrument we shall first explain the general arrangement of the multimeter with the aid of a block diagram.

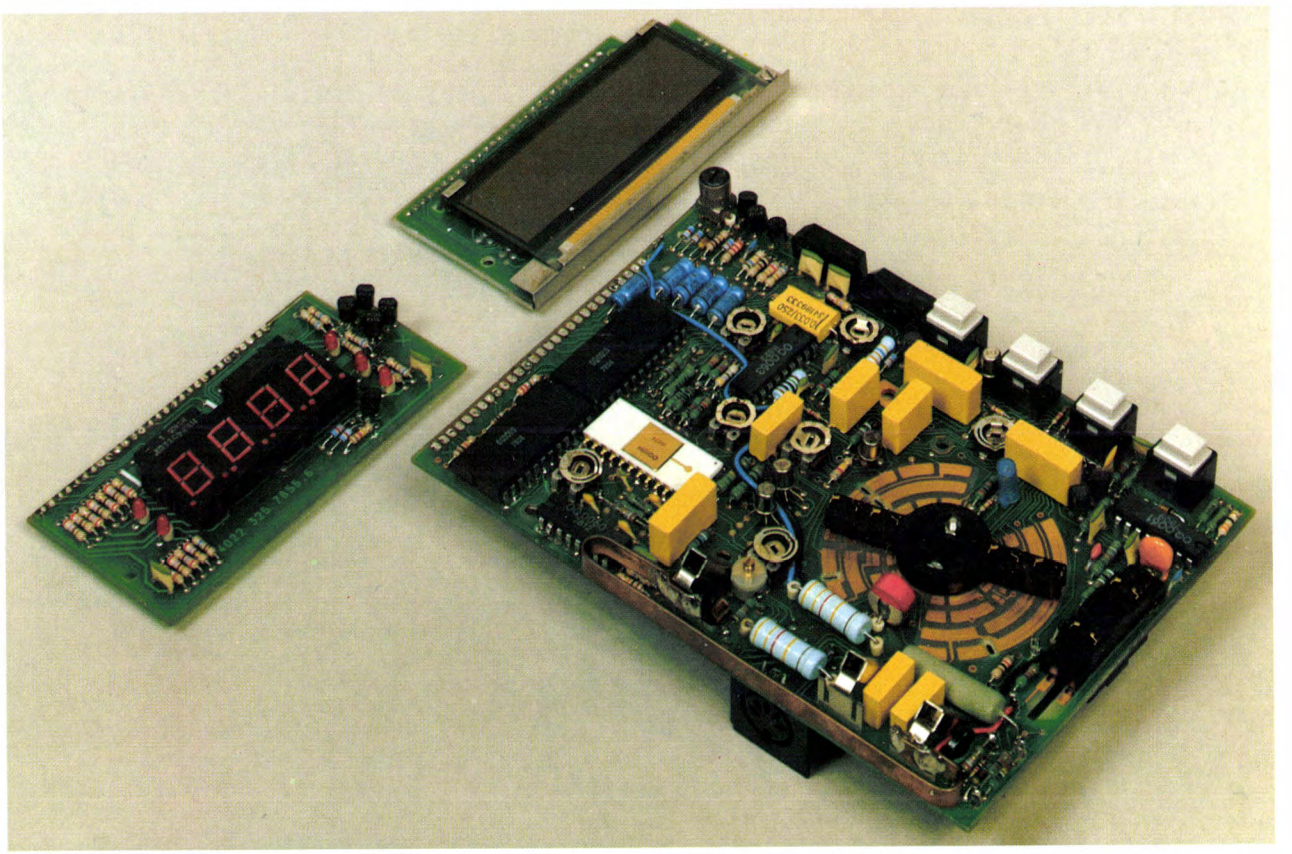


Fig. 2. The interior of the PM 2517. The large panel contains all the measurement circuits. The small panel on the left is included with the LED version (PM 2517E), and the one on the right is included with the LCD version (PM 2517X). Six ICs can be seen on the large panel; all of them were specially developed for the multimeter. Three connector sockets can be seen on the left; the folded metal strip connecting the two outer sockets is a 10 A shunt of 12 mΩ. To the right of this shunt there are two (light blue) resistors, which together form the 9 MΩ series resistance for the voltage measurement; the (red-painted) cylindrical component next to them is a PTC resistor, which provides protection from overload. The segments of the function-selector switch are printed on the panel. The oscillator coil of the d.c. inverter can be seen right at the top.

General description of the PM 2517

A block diagram of the PM 2517 digital multimeter is shown in *fig. 3*. The six integrated circuits in the blue boxes include the measurement and control functions. Between the input terminals and these ICs there is a greatly simplified representation of a number of voltage dividers, shunts, etc., which are in discrete-component form in the PM 2517. There is an optional choice between two displays; these are shown at the right. The upper one is the LED (light-emitting diode) display of the PM 2517E, the one below is the liquid-crystal display (LCD) of the PM 2517X.

The heart of the instrument is the analog/digital converter A/D (in IC *2*) and the voltage/current converter V/I (in IC *1*) that precedes it. Direct voltages are converted here into a current that charges the capacitor C_1, which is also discharged again by digitally counted pulses of accurately known magnitude. The accuracy of the instrument is largely determined here. All the other quantities — alternating voltages, direct and alternating currents, resistances — are first reduced to a direct voltage, and the actual measurement is then made in this part of the instrument.

The digital information provided by the analog/digital converter is accumulated in the control unit *CU* (IC *5*) and processed to form a control signal for the display. If the test signal is too large for the range selected, the control unit gives a command (via the red lines) to reset the voltage dividers etc. and shift the decimal point.

All measurements start at the lowest range. For voltage measurements, where the lowest range goes to 1 V, a voltage divider that reduces the voltage to a tenth is included even in the lowest range. This gives the instrument an input resistance of 10 MΩ, high enough to have an insignificant effect on most of the

circuits on which measurements will be made. The voltage divider also gives some instrument protection.

The range-selection process depends on a combination of two methods. In addition to the 10:1 voltage divider there is another one with a ratio of 1000:1. The step between the two, a factor of 100, is too large, but an intermediate step of 10 times is provided in the voltage/current converter V/I by switching over to a conversion resistance R_{conv} that is 10 times larger; this reduces the sensitivity by a factor of 10. Ranges of 1 V, 10 V, 100 V or 1000 V can thus be selected for voltage measurement. (The maximum readings remain just below these limits because of the four-digit indication: the actual values are 999.9 mV, 9.999 V, 99.99 V and 999.9 V.)

As we said earlier, alternating voltages are reduced to the measurement of direct voltages; this takes place in several steps. After an isolating capacitor C_2 there is again a voltage divider with a ratio of 10:1, but now with a total resistance of 2 MΩ. This is followed by a voltage/current converter V/I (in IC *3*), similar to the one for the direct-voltage measurements. A choice can again be made between two conversion resistances that differ in magnitude by a factor of 10; with a second voltage divider (1000:1) a choice can once more be made from four voltage ranges.

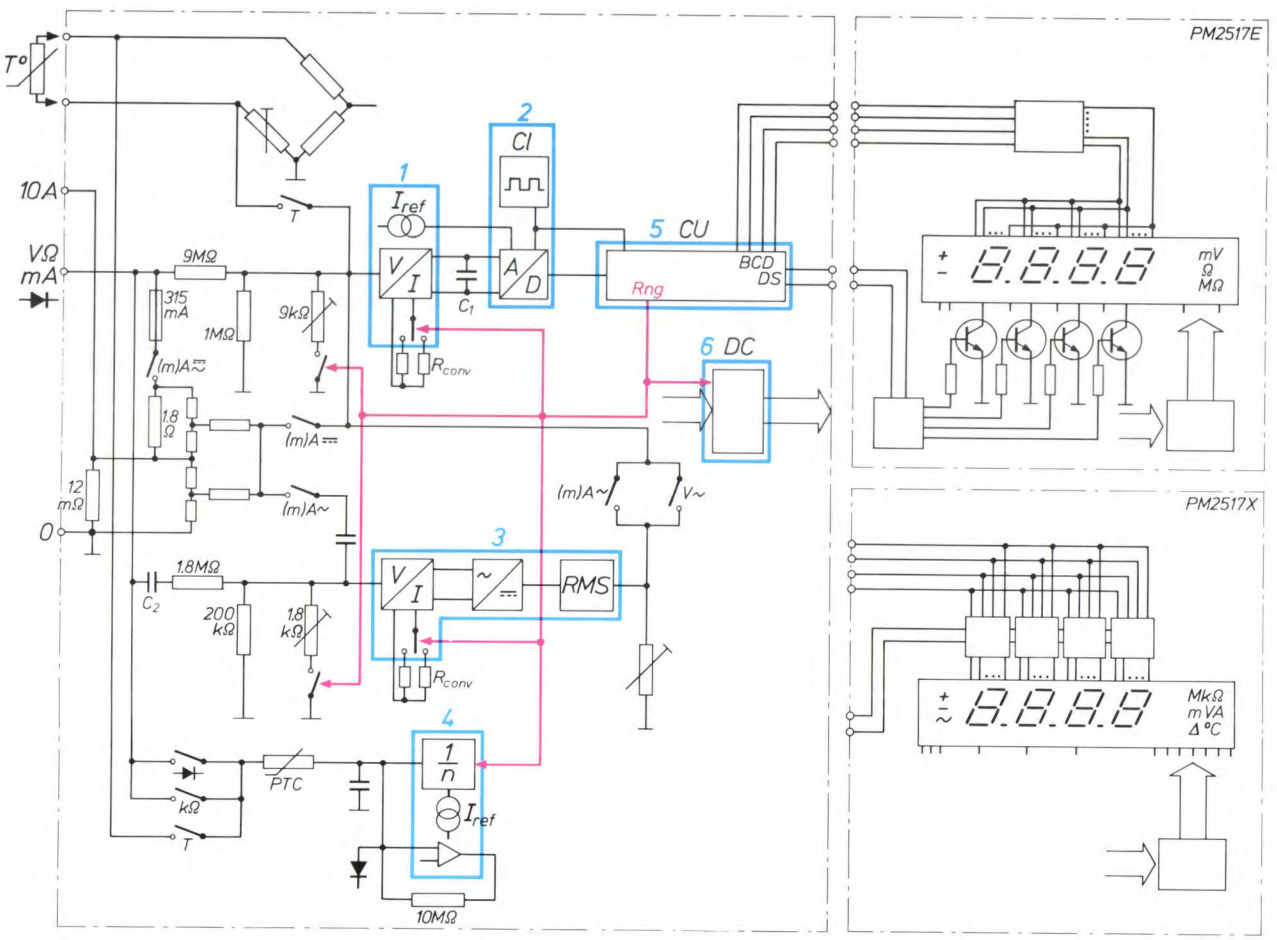


Fig. 3. Simplified circuit diagram of the PM 2517 multimeter. On the left is the large panel with the ICs *1* to *6* (*in the blue boxes*), on the right the two options for the display panels, with LEDs for the PM 2517E and LCD for the PM 2517X. IC *1* contains a voltage/current converter V/I and an accurate reference-current generator (I_{ref}), IC *2* a clock generator *Cl* and an analog/digital converter A/D; the digital output signals go to the control unit *CU* (IC *5*). This supplies binary-coded decimal (*BCD*) digits to the display; a signal on the lines *DS* makes clear which digit is intended. The group of outputs *Rng* (*red*) gives the automatic selection of the correct range. Alternating voltages are processed in IC *3*; after conversion to current and rectification the r.m.s. value is determined (*RMS*). IC *4* contains a current-generator circuit for resistance measurements. IC *6* (*DC*) controls the placing of the decimal point and the indications for polarity and unit of measurement on the display.

The alternating current from the voltage/current converter is rectified in a full-wave rectifier. The r.m.s. value of the resulting pulsating direct current is then calculated in the circuit *RMS*; the calculation does not depend on the waveform. The logarithmic relation between the voltage across a diode and the current through it is used for this calculation; we shall return to the details later. The r.m.s. value of the alternating voltage is eventually represented by the magnitude of a direct current. This passes through a resistor and the voltage across the resistor is directly applied to the input of the direct-voltage measurement circuit.

There are two ranges for the measurement of direct and alternating currents: 100 mA and 10 A. The input socket for the 100 mA range is the same as for voltage measurements. The current flows to the instrument chassis through a shunt resistor. The voltage across the shunt resistor is measured as described earlier. For measurements in the 10 A range there is a separate input socket, which is connected to the chassis by a shunt resistance of 12 mΩ. This is formed by the folded metal strip that can be seen in fig. 2.

The measurement of resistances is reduced to a voltage measurement by passing an accurately known current through the resistance. There are five ranges, from 1 kΩ to 10 MΩ. The current originates from a current-generator circuit that is included, with a current multiplier, in IC *4*, and can be set between 1 mA and 100 nA in five steps of a decade. More will be said about this circuit later. The current of 1 mA gives a voltage of 1 V across a resistance of 1 kΩ; since the resistance to be measured is connected to the same input socket as for voltage measurement, the 'voltage' indication now refers to a resistance value. The symbol Ω appears on the display instead of the symbol mV. If the voltage rises above 1 V, then at a command from the control unit *CU* the current is automatically reduced by a factor of 10 and the decimal point displaced.

The measurement of the forward and reverse voltages of semiconductor diodes is closely related to resistance measurement. In this test a current of 1 mA is supplied from the input socket; the voltage produced across the diode is indicated in mV. If the diode is in the forward direction the voltage will be a few hundred mV; if it is in the reverse direction the voltage will exceed the upper limit of the range and this will be indicated by a special code in the display.

Finally, the PM 2517 can also be used to measure temperatures between −60 °C and +200 °C. This measurement is made with a temperature probe. The probe is connected to a separate input and contains a thermistor that forms one of the arms of a Wheatstone bridge in the instrument. The indication is given in °C. The bridge is supplied by a current of 1 mA from IC *4*.

After this description of the input circuits we shall now say a few words about the control of the display. The four digits that indicate the measured result are provided by the control unit *CU* in binary-coded decimal (*BCD*) form. This information is supplied on 4 parallel lines, which means that only one digit can be transmitted at a time. A code signal on two separate lines (*DS*, digit selection) makes clear which of the four digits is intended.

The control of the segments of the four 7-segment digits is not the same in the liquid-crystal display as in the LED display. The digits in the LED display are illuminated one at a time in rapid succession, largely to limit the number of lines. The anodes of the corresponding segments of the four digits are connected together and receive their control from a logic circuit to which the binary code for the four digits is applied. The cathodes are connected 'per digit', however, and are only energized when the code on the two *DS* lines calls for this.

With the liquid crystals the four digits are shown simultaneously. Since only one digit is transmitted at a time, four circuits — one for each digit — are included to act as a buffer memory and hold the control signals for the seven segments. In both versions there is also a decoder circuit *DC* (IC *6*), which derives the correct decimal point and unit symbol from information about the nature of the test quantity and the range selected.

The power for the PM 2517 is supplied either from four 1.5 V batteries or from a mains adaptor. To eliminate the effects of falling battery voltage, the batteries drive a d.c. inverter. This contains an oscillator, which produces an alternating voltage at a frequency of 200 kHz. The alternating voltage is rectified to give two stabilized direct voltages of ± 8.5 V, which supply all the circuits. The inverter has an efficiency of about 60%.

The analog/digital converter

In the design of an integrated analog/digital converter we had to bear in mind that many analog functions could be accurately and adequately performed in a bipolar technology [1], whereas although the MOS technology was very suitable for digital circuits, there were problems in the design of the accurate analog input circuit and current generator. An optimum analog/digital converter has been obtained by using a bipolar technology for all the analog functions where accuracy is important, and a digital MOS circuit for counting and control. Efforts were made to find a

[1] Th. J. van Kessel and R. J. van de Plassche, Integrated linear basic circuits, Philips tech. Rev. 32, 1-12, 1971.

principle of operation that could readily be applied in an integrated circuit. Such a principle has been found and has been called sigma-delta modulation [2]. The analog signal processing takes place in two bipolar ICs, and will be described in more detail below; the digital shift registers and counter circuits required are included in the digital IC (in LOCMOS technology) shown as the control unit *CU* in fig. 3.

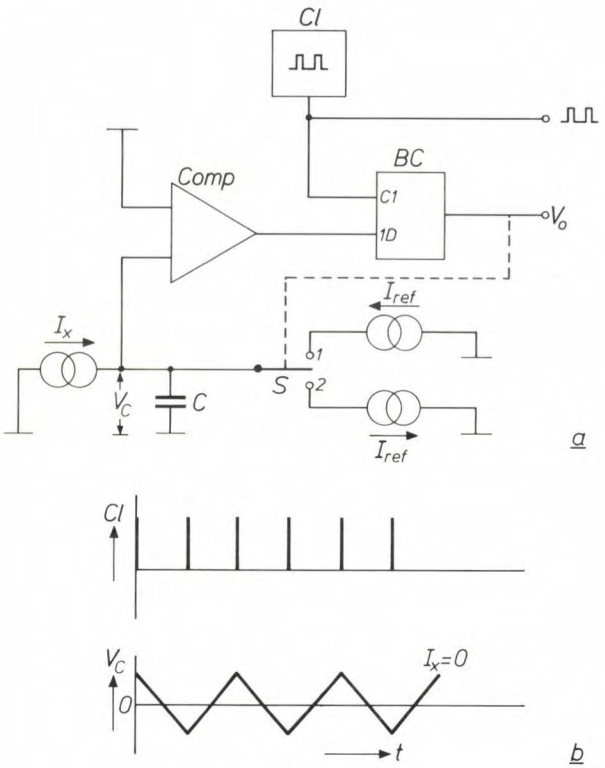


Fig. 4. Principle of the 'sigma-delta modulator'. *a*) Block diagram. The current I_x to be measured charges the capacitor *C*. The circuit is designed so that it keeps the voltage V_C across *C* to zero by carefully regulating the discharge pulses. This regulation is provided by switch *S*, which supplies a charging or discharging reference current as required. *S* is controlled through the intervention of a bistable circuit *BC* by a comparator *Comp*, which monitors the voltage across the capacitor. The control action takes place in synchronism with the clock pulses originating from clock generator *Cl*. The output signal V_o is a digital representation of the magnitude of I_x. *b*) When $I_x = 0$, V_C alternates about the zero.

signal itself, but to its time integral: the voltage V_C across capacitor *C* (*fig. 4a*) is the integral of the current to be measured I_x. We call this 'sigma-delta modulation'.

One advantage of taking an integral is that interference from the mains is greatly reduced. This is because the integration includes many cycles of the mains; the integration time in the PM 2517 is 250 ms.

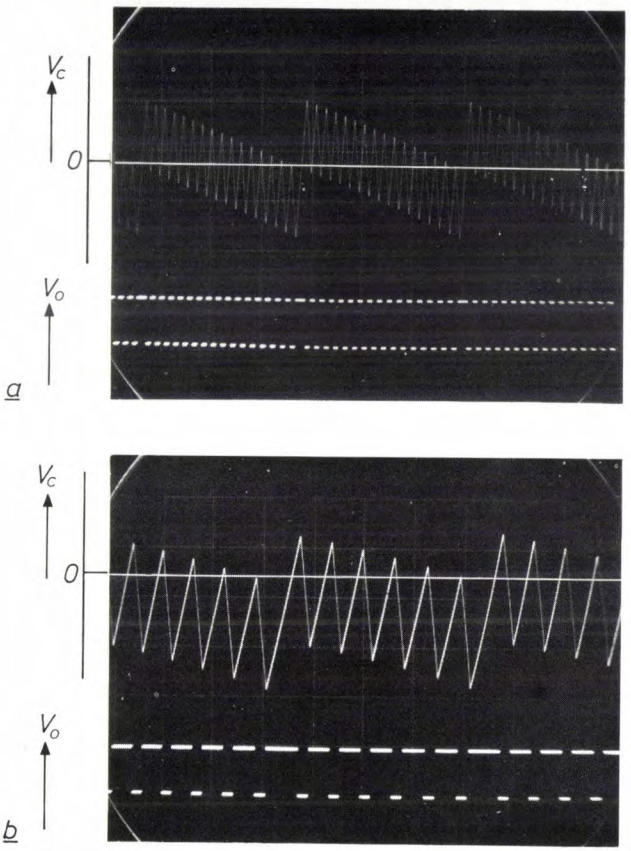


Fig. 5. The capacitor voltage V_C and the output voltage V_o of the sigma-delta modulator as a function of time. *a*) The current I_x to be measured is small with respect to the reference current ($I_x = \frac{1}{33} I_{ref}$). After every 32 clock pulses one extra discharge pulse is necessary to compensate I_x. *b*) $I_x = \frac{7}{19} I_{ref}$. After each charging pulse there are two discharge pulses. This would bring the circuit to equilibrium if I_x was accurately equal to $\frac{1}{3} I_{ref}$. However, I_x is a little larger; after every 18 pulses one extra discharge pulse is required.

Sigma-delta modulation

The basis of the analog/digital conversion is the well-known principle of delta modulation [3]: changes in the signal are followed as closely as possible by steps of the same magnitude, and a binary code indicates whether the step goes up or down. However, in an instrument we are not as interested in the changes in the signal as in its magnitude, which may not change for some time. Delta modulation, which in fact yields the time derivative of the signal, is therefore applied not to the test

The delta modulation is produced with the aid of a circuit that measures V_C and tries to keep it at zero. This is done by charging or discharging *C*, depending on the position of switch *S*, with a reference current I_{ref} at the same time as the test current I_x (fig. 4*a*) is being applied. The switch *S* is operated in synchronism with the clock pulses from the clock generator *Cl* by a bistable circuit *BC*, which puts *S* into position *1* or *2* depending on whether the comparator circuit *Comp* finds a negative or a positive voltage across *C*.

The sequence of control pulses originating from *BC* contains the desired information about the magnitude of I_x in digital form. If $I_x = 0$, *C* is alternately charged and discharged, and the output from *BC* is an alternating train of 'ones' and 'zeros'. The voltage V_C then changes periodically about the zero (fig. 4*b*). If I_x is slightly different from zero, an extra discharge pulse will be necessary now and again to bring V_C back to zero; see *fig. 5a*, where 1 extra discharge pulse appears after every 32 clock pulses; I_x has $\frac{1}{33}$ of the full-scale value here. In fig. 5*b* there is one extra discharge pulse after every 18 clock pulses; here I_x is $\frac{7}{19}$ I_{ref}, so that for each charging pulse there are usually two discharge pulses. The upper limit of the range has been reached when only discharge pulses appear; I_x is then equal to I_{ref}.

The 'ones' and 'zeros' that form the output signal V_o in fig. 4*a* are applied to an up-down counter, which takes a step upwards at a '1' and a step downwards at a '0'; the up-down counter is located in the control unit (*CU* in fig. 3). The virtue of the procedure used is that speed of measurement and accuracy can be traded against one another; the number of numerical values that can be distinguished increases with the time of counting. In the PM 2517 a total of 25 000 pulses are counted; the resolution is then $I_{ref}/25\,000$. The clock rate is slightly higher than 100 kHz, and one 'measurement cycle' lasts about 250 ms, as mentioned above.

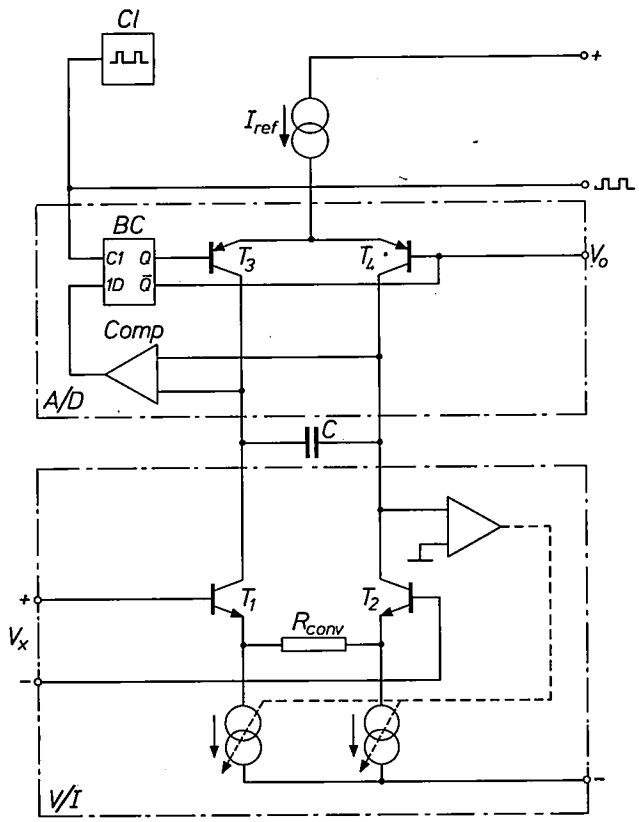


Fig. 6. Simplified circuit of the voltage/current converter *V/I* and the analog/digital converter *A/D*. The function of the switch *S* in fig. 4 is taken over by the transistors T_3, T_4. The current in the conversion resistor R_{conv} is V_x/R_{conv}.

Conversion of voltage to current

The measurement circuit described is based on an input current I_x. This does not mean that it is suitable in this form for current measurements, because the currents to be measured in a circuit will not in general be supplied from a current generator (with an infinitely high output impedance), like I_x in fig. 4*a*. In addition it must also be possible to measure voltages. A voltage/current converter is therefore included in front of the analog/digital converter. *Fig. 6* shows a simplified general picture. The voltage to be measured V_x is applied to the bases of the transistors T_1 and T_2 and appears across the conversion resistor R_{conv}, producing a current V_x/R_{conv} in it. This current is in fact the difference between the collector currents of T_1 and T_2 and charges the capacitor *C*. The switch *S* in fig. 4*a* is replaced here by two transistors T_3 and T_4, which pass the current I_{ref} alternately to one electrode of the capacitor and then the other.

The voltage across the capacitor *C* is kept to zero in this way. The voltage between *C* and earth, however, is not defined, and nor therefore are the collector voltages of the four transistors. This is provided for by a control loop, which includes an operational amplifier that measures the voltage from *C* to earth and adjusts the bias currents of T_1 and T_2 so that the capacitor is approximately at earth potential, i.e. half-way between the positive and negative supplies.

In practice the conversion of voltage to current is much more complicated than fig. 6 would suggest. The output impedance of the emitters of T_1 and T_2 is not low enough in proportion to R_{conv}, so that the full value of V_x does not appear across R_{conv}. To make this output impedance practically zero it is necessary to apply negative feedback. The feedback arrangement is shown in *fig. 7*. The complete circuit consists of four 'current mirrors' (called 'controlled current sources' elsewhere [1]). The combination T_3, T_4, T_5 forms such a current mirror as does T_9, T_{10}, T_{11}. These current mirrors contain the necessary powerful feedback to ensure that independently of the external load and

[2] R. Steele, Delta modulation systems, Pentech Press, London 1975.
R. J. van de Plassche and R. E. J. van de Grift, A five digit A/D converter, 1977 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. tech. Papers, pp. 102-103.
R. J. van de Plassche, A sigma-delta modulator as an A/D converter, IEEE Trans. CAS-25, 510-514, 1978 (No. 7).
[3] J. F. Schouten, F. de Jager and J. A. Greefkes, Delta modulation, a new modulation system for telecommunication, Philips tech. Rev. 13, 237-245, 1951/52.

independently of V_x the same current I_0 always flows in each of the vertical branches. The operating point of the transistors therefore does not change if V_x varies; the source of T_1 therefore follows the voltage fluctuations of the gate and since the base-emitter voltage of T_3 is constant these fluctuations are passed on undiminished to R_{conv}; this is also true for the right-hand side.

Reference-current generator

We now come to the reference-current generator, which is really the corner-stone of the measurement accuracy, since it delivers the current with which the measured quantity is compared. The reference current must be accurate in value and it must also remain constant in practical use if the supply voltage and temperature vary.

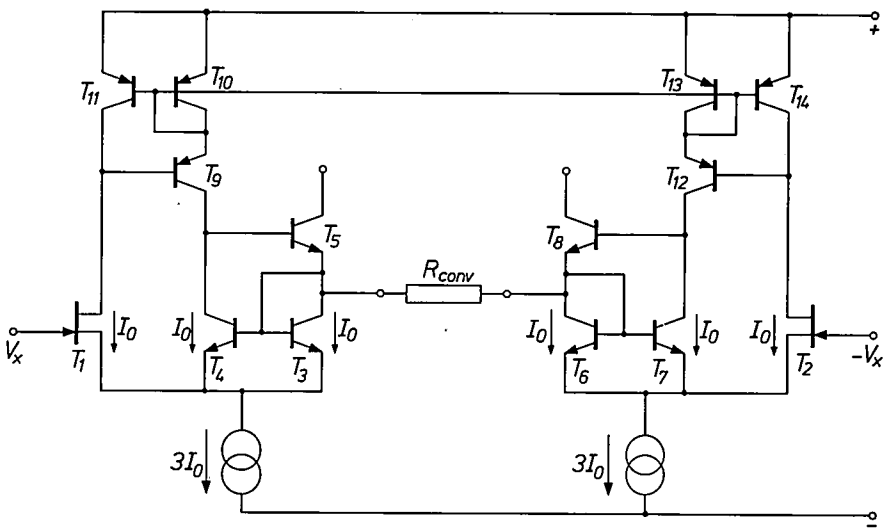


Fig. 7. A more complete circuit diagram of the voltage/current converter. The input transistors T_1, T_2 are junction field-effect transistors, which give a high input resistance. The circuit is built up from 'current mirrors', which ensure that a current I_0 flows through each of the six vertical branches. Since the operating point of the transistors cannot then change, the potential difference between the two input terminals is transferred unchanged to the conversion resistance R_{conv}. The current in this resistor leaves the circuit as a difference between the collector currents of T_5 and T_8.

'Auto-zero' circuit

The PM 2517 indicates either negative or positive voltages; the zero is effectively 'in the centre of the scale'. Stability of the zero and the absence of drift are important requirements for an accurate indication of small direct voltages. These features are ensured by including a circuit that automatically corrects deviations in the zero; we call this an 'auto-zero' circuit. The circuit is shown in *fig. 8*. In the figure any unbalance is represented by the voltage generator V_{offset} on one side in front of the analog/digital converter A/D. The symmetrical voltage input is connected to the A/D converter via a reversing switch. The reversing switch is operated by auto-zero pulses, which reverse it at exactly half the measurement time. This means that the voltage supplied to A/D is $V_x - V_{offset}$ during the first half and $-V_x - V_{offset}$ during the second half; V_{offset} is eliminated by subtracting one from the other. The subtraction is carried out by reversing the output in synchronism with the auto-zero pulses between the direct and inverted output signals from the A/D converter.

These objectives are attained by arranging that the current is defined by a geometrical property of the integrated circuit — the ratio p of the emitter areas of the transistors T_1 and T_2 in *fig. 9*. Imagine first of all that the resistance R_2 is removed from the circuit. We then have two dissimilar transistors that both carry the same current; the current is kept the same by a control system consisting of the two resistors R and the operational amplifier OA. We thus have $I_1 = I_2$; see the straight line in *fig. 10*.

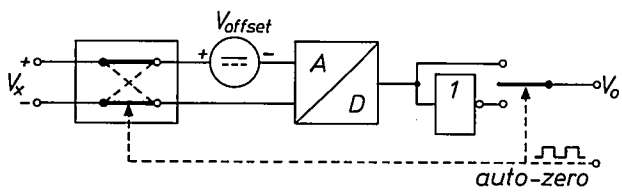


Fig. 8. The 'auto-zero' circuit. Drift and unbalance in the input circuit, represented here by a voltage V_{offset}, are eliminated by periodically reversing the polarity of the input voltage V_x. At the same time the digital output signal is inverted.

There is however another relation between I_1 and I_2. The sum of the base-emitter voltage of transistor T_1 and the voltage drop across R_1 is equal to the base-emitter voltage of T_2. This is true for the whole of the curved line in fig. 10. It is clear from the figure that there is only one point (besides the origin) for which both conditions are satisfied, and the current generator automatically takes up this value.

The relation between the base-emitter voltage V_{BE} and the emitter current I is given by the well-known diode equation

$$I = I_R(\exp\frac{eV_{BE}}{kT} - 1). \quad (1)$$

In addition to the constants e (electronic charge) and k (Boltzmann's constant) the equation contains the absolute temperature T and the reverse current I_R. The reverse current is proportional to the emitter area. For somewhat larger currents (1) can be approximated by

$$I = I_R \exp\frac{eV_{BE}}{kT}. \quad (2)$$

The two conditions shown graphically in fig. 10 are:

$$I_1 = I_2, \quad (3)$$

$$I_1R_1 + \frac{kT}{e}\ln\frac{I_1}{pI_R} = \frac{kT}{e}\ln\frac{I_2}{I_R}. \quad (4)$$

For small currents the term I_1R_1 in (4) can be neglected; the equation then simplifies to $I_1 = pI_2$, which determines the slope at the origin of the curved line in fig. 10. At higher currents, however, the term I_1R_1 predominates over the logarithmic term and the slope of the curve is determined by the magnitude of R_1.

The current that automatically sets itself in this way is not independent of the temperature; it increases with increasing temperature. Compensation can be provided by supplying an additional current that decreases with increasing temperature. This is the current I_3 in fig. 9. The current I_3 is equal to the base-emitter voltage of T_2 divided by R_2. Now for a transistor with the operating point determined as above the base-emitter voltage decreases by about 2 mV per °C, and I_3 therefore decreases in proportion. R_1 and R_2 are chosen so that the two opposing temperature effects cancel one another in the output current I_{ref}; the two resistors are not integrated but mounted externally.

Rectification and determination of the r.m.s. value

To measure alternating voltages it is necessary to rectify the voltage and determine the r.m.s. value. Because it was desired to extend the bandwidth of the instrument to rather high frequencies, these operations could not be performed digitally, since the signal-sampling rate in the PM 2517 is much too low for this. Analog methods therefore have to be used.

The r.m.s. value is not determined by approximating the alternating voltage by a sine function, but the true r.m.s. value is calculated for any waveform by using logarithms. The voltage across a semiconductor diode is proportional to the logarithm of the current in it. The alternating voltage is therefore first put into the form of a rectified alternating current. This is done in the current rectifier.

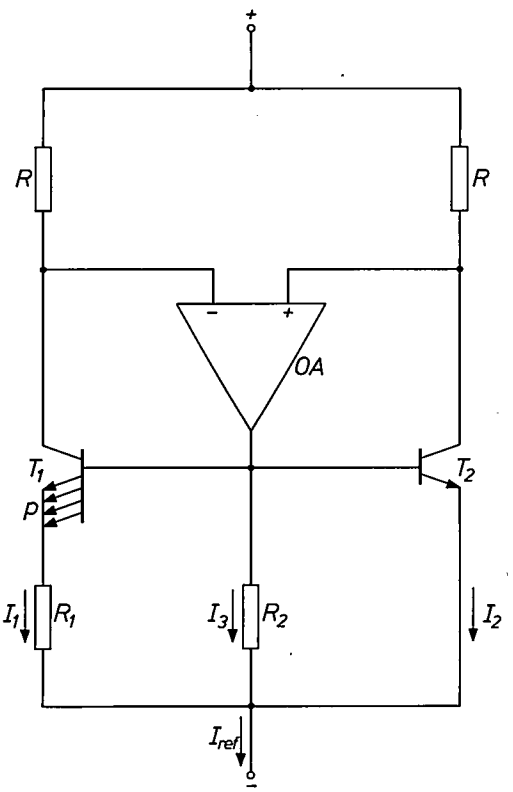


Fig. 9. The reference-current generator. The sum of the currents I_1 and I_2 is entirely determined by the ratio p of the emitter areas of the transistors T_1 and T_2 and is independent of the supply voltage. It does nevertheless increase with temperature. However, I_3 decreases with temperature, so that the correct choice of R_1 and R_2 makes the total current constant even when there are temperature fluctuations.

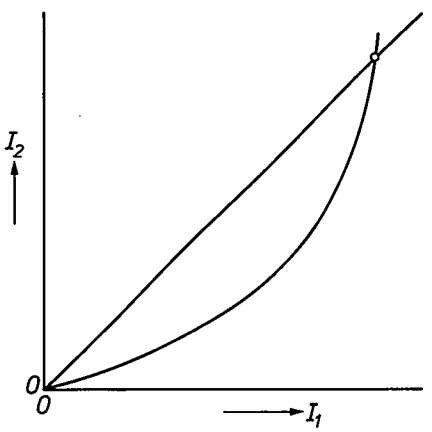


Fig. 10. In the reference-current generator (fig. 9) $I_1 = I_2$ (*straight line*), but at the same time $V_{BE1} + I_1R_1 = V_{BE2}$ (*curved line*). The currents take up the value corresponding to the intersection of the two lines.

The current rectifier

The current rectifier is preceded by a voltage/current converter as shown in fig. 7. The currents supplied by the converter are $I + \Delta I$ and $I - \Delta I$, where ΔI is the alternating-current component to be measured. Both currents are supplied to the current rectifier, shown in simplified form in the circuit of *fig. 11*.

Starting at the bottom, we see that this circuit contains two transistors T_1 and T_2 used as rectifying elements and two special current mirrors T_3, T_5 and T_4, T_6. These current mirrors ensure that equal currents $I + \Delta I$ flow on either side. Since a current of only $I - \Delta I$ is taken off at one side, the difference $2\Delta I$ must find its way to a third output via one of the transistors. In the figure it flows through T_1; when this is conducting T_2 is non-conducting, since the sum of the base-emitter voltages of the two transistors is zero at all times. A half-cycle later the roles of the two transistors are reversed.

A small direct-current unbalance can arise between the two halves of the circuit. To detect this unbalance, the collectors of T_1 and T_2 are not connected together directly, but a capacitor is connected between them, while each collector is connected to the output through a resistor (*fig. 12*). Differences in the direct-current component of the two collector currents cause a direct voltage to appear across the capacitor, and this voltage also appears across the input of the differential amplifier T_7, T_8, which then provides compensation by taking unequal currents at the two inputs from the current rectifier.

Determination of the r.m.s. value

The r.m.s. value of an alternating current is the value of the direct current that gives the same mean power dissipation in a resistor. The arithmetical operations follow from this definition: since the power is proportional to the square of the current, the squaring must be performed first; a mean must then be taken over at least one cycle and finally the square root is taken.

Fig. 13 gives the essential features of the circuit that carries out these operations. The rectifier current I_i is passed through a series circuit of two transistors T_9 and T_{10} connected as diodes; the resulting voltage V_1 across the transistors is proportional to twice the logarithm of the current, and hence to the logarithm of the square of the current.

In transistor T_{11} the logarithm of the current I_0 is subtracted from this; I_0 is the current that expresses the final result of the calculation, and is delivered to T_{11} by the current mirror T_{13}, T_{14}. V_2 is therefore proportional to the logarithm of the ratio I_i^2/I_0.

The reverse conversion from logarithms to ordinary quantities is made in transistor T_{12}, whose collector

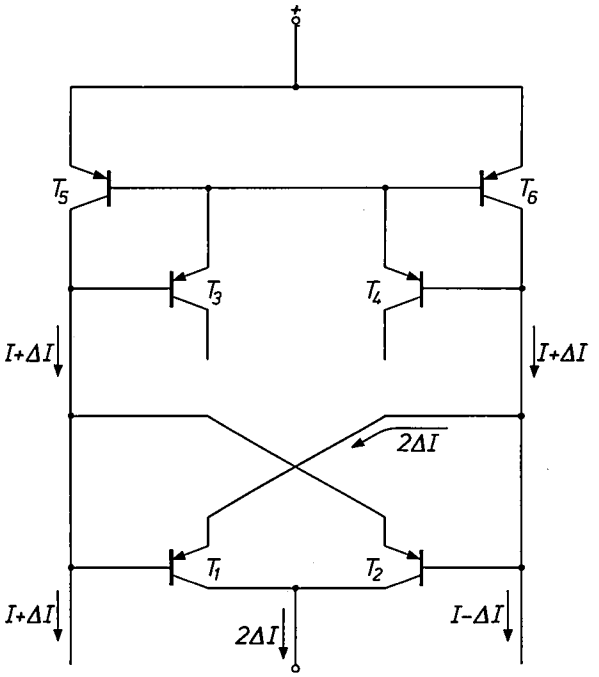


Fig. 11. The current rectifier. Alternating voltages are converted into an alternating current ΔI before the measurement. This current is rectified in the full-wave rectifier formed by T_1 and T_2. The transistors T_3 to T_6 act as a pair of current mirrors that ensure that the currents $I + \Delta I$ on either side are equal. The difference between $I + \Delta I$ and $I - \Delta I$ finds its way to the output; the figure relates to the instant at which this difference current flows via T_1. A half-cycle later it flows through T_2.

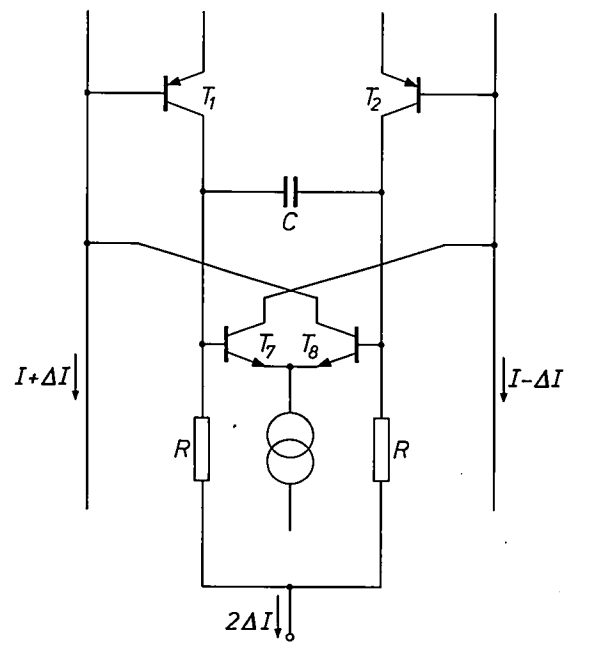


Fig. 12. To correct any asymmetry in the current rectifier (fig. 11) the collectors of T_1 and T_2 are not connected directly, but through the resistors R. When there is asymmetry a direct voltage appears across the capacitor C. This direct voltage is applied to the differential amplifier T_7, T_8, which then provides compensation by taking unequal currents at the input to the current rectifier.

current is I_1^2/I_0. It now only remains to take the mean with the aid of an *RC* filter; a current $I_0 = \langle I_1^2 \rangle / I_0$ is obtained. From this it follows that $I_0 = \sqrt{\langle I_1^2 \rangle}$, where I_0 is the desired r.m.s. value of the current I_1. The total circuit supplies a current of $2I_0$; a correction for the factor of 2 is made later in the measurement circuit.

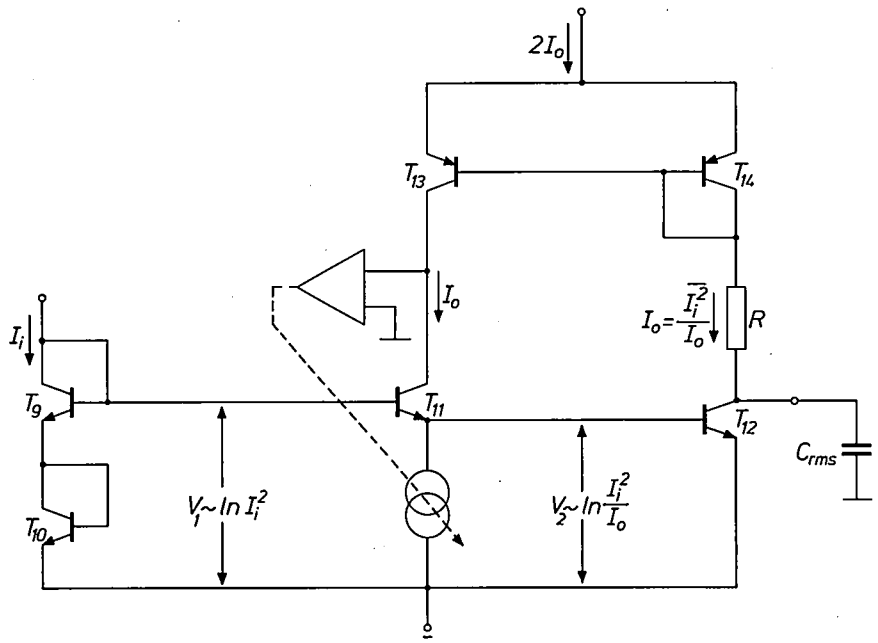


Fig. 13. Analog circuit for calculating the r.m.s. value of a rectified alternating current I_1. In the multimeter this is the current $2\Delta I$ (fig. 12). The direct current I_0 has the same r.m.s. value as I_1.

In fact, of course, we cannot speak of 'the logarithm of the current', as has been done here for simplicity. The logarithmic relation between V_1 and I_1 follows from the earlier equation (2):

$$V_1 = 2\frac{kT}{e}\ln\frac{I_1}{I_R} = \frac{kT}{e}\ln\left(\frac{I_1}{I_R}\right)^2.$$

The logarithm is taken of the ratio of I_1 and the reverse current I_R of the transistors T_9 and T_{10}.

Similarly the expression for V_2 is

$$V_2 = \frac{kT}{e}\left\{\ln\left(\frac{I_1}{I_R}\right)^2 - \ln\frac{I_0}{I_R}\right\} = \frac{kT}{e}\ln\frac{I_1^2}{I_0 I_R}.$$

The programmable direct-current generator for resistance measurements

The direct current that we require for the conversion of a resistance measurement to a voltage measurement has to be accurately set to the correct value. In fact, it has to be accurately set to a number of different values, for the different measurement ranges. The accuracy required is 10^{-3} to 10^{-4}. Now since the resistances in an IC only have an accuracy of a few per cent at best, our requirements cannot be met unless special precautions are taken.

To obtain the desired accuracy we have made use of a principle that has been called 'dynamic element matching' elsewhere [4]. This can be explained with the help of *fig. 14*, where the principle is applied for dividing a current into two equal parts. The current $2I$ from a current generator is divided into two parts I in a divider circuit, but there is an unavoidable small deviation ΔI. The two currents are now connected alternately to the two outputs by a switch S that periodically reverses the connections at a command from the signal from a clock generator *Cl* (fig. 14*a*). Each of the two outputs therefore carries the mean of the currents $I + \Delta I$ and $I - \Delta I$, i.e. I. If the clock signal has a time asymmetry ΔT (fig. 14b), then there is still a difference between the two output currents, but it is now not ΔI but only $(\Delta T/T)\Delta I$, which is an order of magnitude smaller.

The same principle can be applied even if the ratio of the output currents is not 1:1. Let us take a ratio of 10:1 as an example, since this represents the basic operation in the multimeter. Eleven similar current generators are then used, connected to a switching net-

[4] R. J. van de Plassche, Dynamic element matching for high-accuracy monolithic D/A converters, IEEE J. **SC-11**, 795-800, 1976.

work with two outputs. One output carries a single current, while the other carries the sum of ten currents. The switching network is controlled by a shift register with eleven stages. A single 'one' and ten 'zeros' circulate in the shift register. It is always a different current generator that provides the single current. The ratio of the mean values of the output currents is kept accurately at 1:10; there is a ripple on the currents but this is removed by a lowpass filter.

There are two such current multipliers in the PM 2517 (CM_1 and CM_2 in *fig. 15*). The current ratios for each can be adjusted by changing the content of the associated shift registers. The two multipliers are connected by the current mirror *CMi*, which is in fact a current multiplier of ratio 1:1 here. The circuit *CMi* contains a control loop that sets its current generator to the desired value. In the same way the eleven current generators in CM_1 and CM_2 are set to the same value as the reference current. By using different combinations of the transfer ratios of CM_1 and CM_2, five different values of current varying from 100 nA to 1 mA can be derived from the reference current of 10 μA. These currents correspond to resistance ranges from 10 MΩ to 1 kΩ. The complete unit is contained in a single IC, shown in *fig. 16*.

A fraction of the test current does not flow through the resistance to be measured, but through the input resistance (10 MΩ) of the voltage-measurement circuit. This introduces an error, which is largest for the 10 MΩ range. Compensation is necessary, and is obtained by using an operational amplifier that always supplies a current equal to that lost in the input resistance of the voltmeter (*fig. 17*). The voltage V_x across the resistance to be measured is always applied to the input of the operational amplifier. This has negative feedback via a voltage divider consisting of two equal resistances arranged in such a way that it amplifies twice, and gives a voltage $2V_x$ at the output. Across the 10 MΩ resistor, connected between output and input, there is therefore a voltage V_x; the current V_x/10 MΩ produced in this resistance exactly compensates for the current lost in the input resistance of the voltage meter. The operational amplifier is included in the IC, but the voltage divider and the 10 MΩ resistor are external.

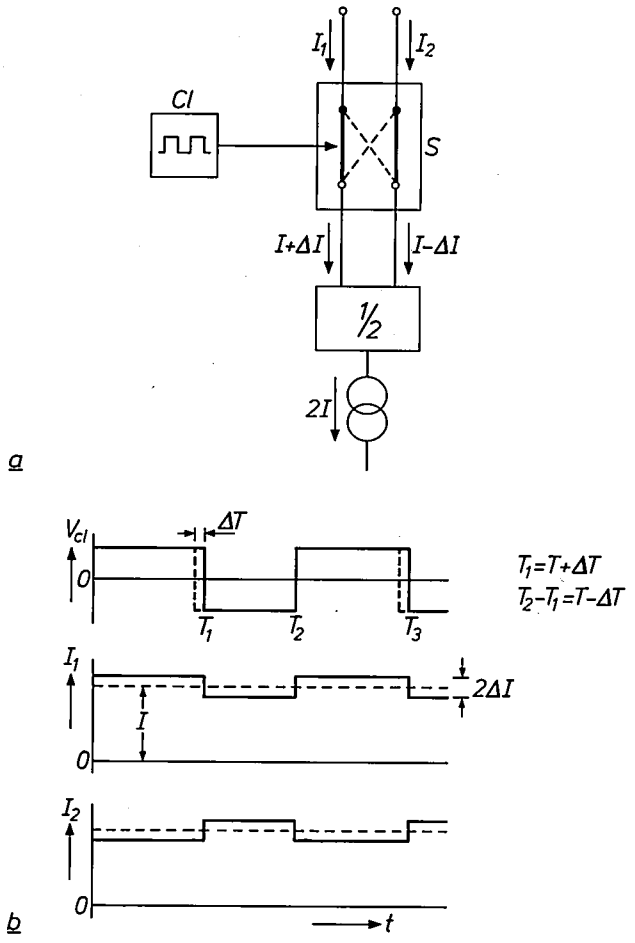


Fig. 14. Accurate division of a current into two equal portions. *a*) The switch *S* periodically reverses the connections of the outputs on the current divider, so that the two output currents I_1 and I_2, averaged over a long time, are accurately equal. Switch *S* is controlled by a clock generator *Cl*. *b*) The clock signal V_{cl} and the output currents I_1 and I_2 as a function of time *t*. A deviation ΔT in the symmetry of V_{cl} gives a deviation between the two mean output currents, but this amounts to $(\Delta T/T)\Delta I$ and is an order of magnitude smaller than ΔT or ΔI itself.

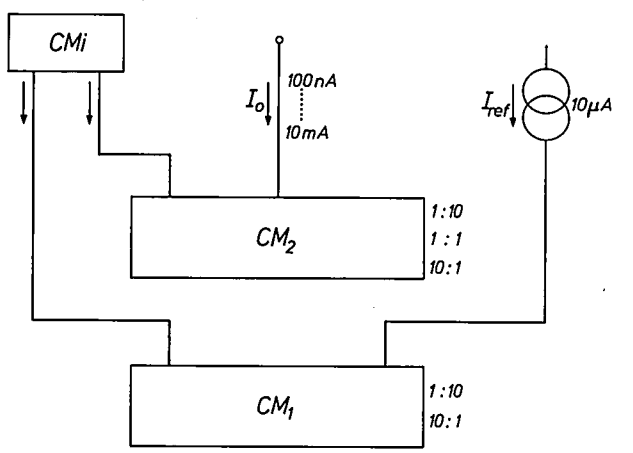


Fig. 15. The programmable direct-current generator for resistance measurements consists of a reference-current generator and two current multipliers CM_1, CM_2 connected by a current mirror *CMi*. The various multiplication factors allow the output current I_o to be set to values between 0.01 I_{ref} and 100 I_{ref}, i.e. from 100 nA to 1 mA.

Instrument protection

A frequently occurring type of overload from which the multimeter must be protected is the accidental application of the mains voltage to the input terminals. Another dangerous hazard is the accidental application of the EHT voltage of a television receiver. The PM 2517 remains undamaged by either of these overloads, even in its most sensitive range.

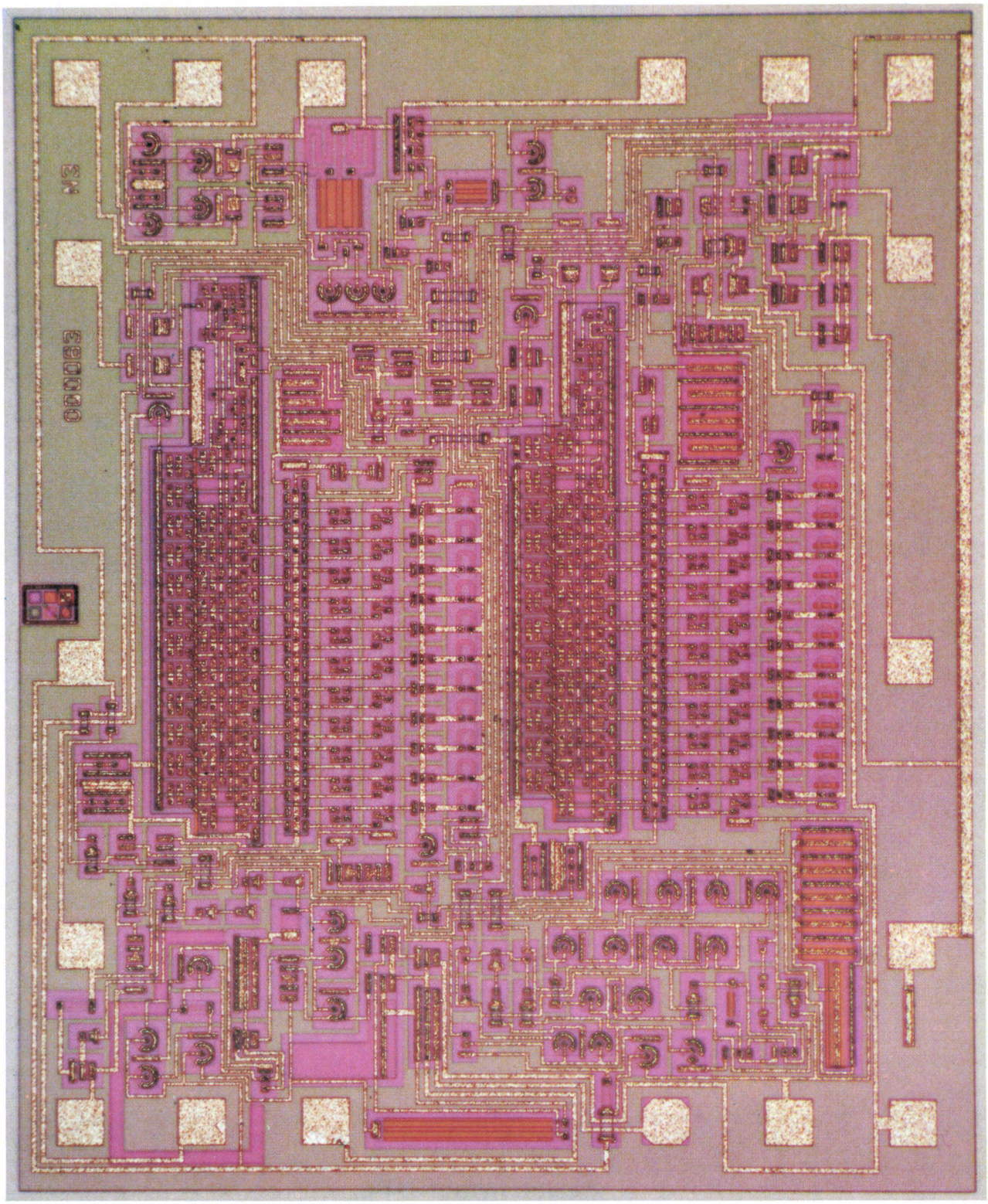


Fig. 16. Photomicrograph of the IC containing the programmable direct-current generator. The dimensions of the IC are 3.2×2.6 mm. The two identical patterns appearing at the centre are the two current multipliers. The closely-packed areas are the shift registers in digital I^2L; on their right can be seen the eleven current generators.

The measurement circuit for direct and alternating voltage is protected from overload by the voltage dividers, which have a high resistance and always give a reduction of ten times even in the most sensitive range. In addition, a voltage limiter that limits the voltage to ± 2.5 V is connected across the input to the integrated circuit; under normal conditions the voltage at the input is no higher than ± 1 V.

To prevent flashover and leakage currents in the input circuit the printed wiring on the panel is arranged to make the leakage paths as long as possible, and the panel is lacquered. The length of the leakage paths was increased by making holes at suitable places in the panel between the segments of the switch, which are printed on the panel.

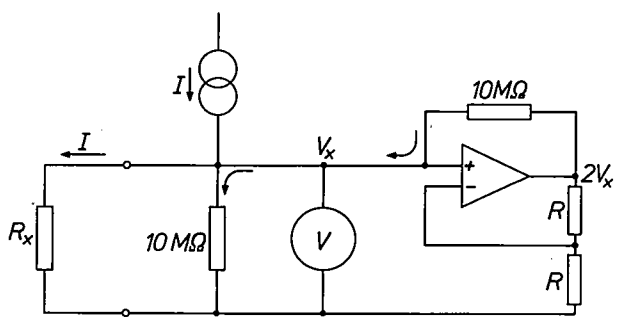


Fig. 17. Compensation of the input current of the voltage-measurement circuit in resistance measurement. By using an operational amplifier with negative feedback a current $V_x/10$ MΩ is generated that is exactly equal to the fraction of the test current I lost in the parallel resistance of 10 MΩ.

In the 'current measurement' position the multimeter is protected by the quick-acting 315 mA fuse shown in fig. 3. The wire-wound 1.8 Ω shunt resistor can withstand a brief overload until the fuse melts. No fuse is provided for the '10 A' input socket; the 12 mΩ shunt resistor connected to it is rugged and can be considered as an almost complete short-circuit to the chassis.

The current-generator circuit for the resistance measurement is connected to the input circuit through a 100 Ω PTC resistor (the red component to the left of the switch in fig. 2). In parallel with the current generator there is a diode circuit that ensures that the voltage across the output of the current generator does not exceed four times the diode voltage. This diode circuit only has to carry a large current for a short time in an emergency, because the PTC resistance heats up and reduces the current.

Summary. The Philips PM 2517 digital multimeter contains six ICs specially developed for the instrument; four of these are bipolar and contain new analog circuits of high accuracy. A 'sigma-delta modulator' functions as an A/D converter. A very stable current generator in the modulator provides a current of constant magnitude, which acts as a reference. For measuring alternating voltages there is a full-wave rectifier, followed by a compact analog circuit for the determination of the r.m.s. value. The test currents for resistance measurements are supplied by an accurate and programmable direct-current generator, adjustable in five steps from 100 nA to 1 mA. The measurement ranges are selected automatically by an internal control unit, and the indication is displayed to four decimals (to 9999) by light-emitting diodes (PM 2517E) or liquid crystals (PM 2517X).

A transducer for medical echography

M. Auphan and G. Dale

A weak point in the current form of medical ultrasonic echography as a diagnostic tool is the lateral resolution. An improvement has been achieved by providing one of the sides of the transducer with five electrodes instead of one, in the form of a central disc and four concentric ring electrodes. Delay lines of the P²CCD type mounted behind the electrodes ensure that the transducer remains focused on a point target on the axis and also that this point travels in step with the transmitted acoustic pulse.

In recent years medical specialists have been able to add ultrasonic echography to the diagnostic tools at their disposal. The heart, several organs in the abdomen, and also the human foetus have already been frequent subjects of echographic examinations; in many hospitals it has now become a routine method of diagnosis [1].

The most important part of an echograph is the ultrasonic probe. The probe contains a piezoelectric transducer, which transmits the ultrasonic energy and also receives the returning echo. In addition to the transducer the probe contains a layer of material whose characteristic acoustic impedance is such as to minimize the transmission losses in the transition from transducer to skin.

The ultrasonic energy is transmitted as a narrow beam of short pulses (about 1 μs) at a repetition rate of a few kHz. The pulses have a high velocity; in human tissue they travel about 1.5 mm in one microsecond. During the passage of the beam through the body echo pulses arise at all points in the radiation field where the characteristic impedance changes abruptly, for example at boundaries of organs or walls of blood vessels. Since the beam is very narrow, the echoes always originate from points on or near the axis of the transducer.

From the echo pulses electronic circuits build up an 'echogram', an image of the area scanned. The echogram, in which the various organs and their possible abnormalities are revealed, can be observed on the screen of an oscilloscope, for example, or on a TV monitor. Depths of about 20 cm in the body are in general easily reached; echoes from greater depths are less easily detected because the ultrasonic energy is then too strongly absorbed.

There are various methods of deriving an image from the echoes [2]. The most usual is the 'B-type' scan. In this case a single image is produced from all the echoes that originate from a flat slice of the body. The A-type scan produces an image only from the echoes received from one position of the probe; see *fig. 1*. *Fig. 2* shows a photograph of a B-type echogram taken with a conventional transducer.

To produce an echogram it is necessary to determine two coordinates: the 'axial coordinate', i.e. the depth at which the echo originates in the body, follows from the time elapsed between the transmission of a pulse and the reception of the echo. This time can be measured accurately by electronic methods, so that the image resolution in the echogram along the axis of the beam may be as good as 0.5 mm or even better. The image resolution in the transverse direction, 'lateral resolution', is almost an order of magnitude worse: in echograms made with a conventional echograph the lateral resolution is not much better than 4 to 8 mm. To allow abnormalities in organs to be diagnosed at an earlier stage it is desirable to have echograms that give a better lateral resolution than 4 mm.

The lateral resolution is poor because there is no focusing, or very little, in the radiation field either during transmission or on reception: the ratio of the

M. Auphan is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France; G. Dale is with Service Central de Radiologie, Hospices Civils, Strasbourg, France.

[1] See for example Proc. Second Eur. Congress on Ultrasonics in Medicine, Munich 1975 (Excerpta Medica, Amsterdam 1975).

[2] A general introduction to the subject is given by P. N. T. Wells in his book Biomedical ultrasonics, Academic Press, London 1977.

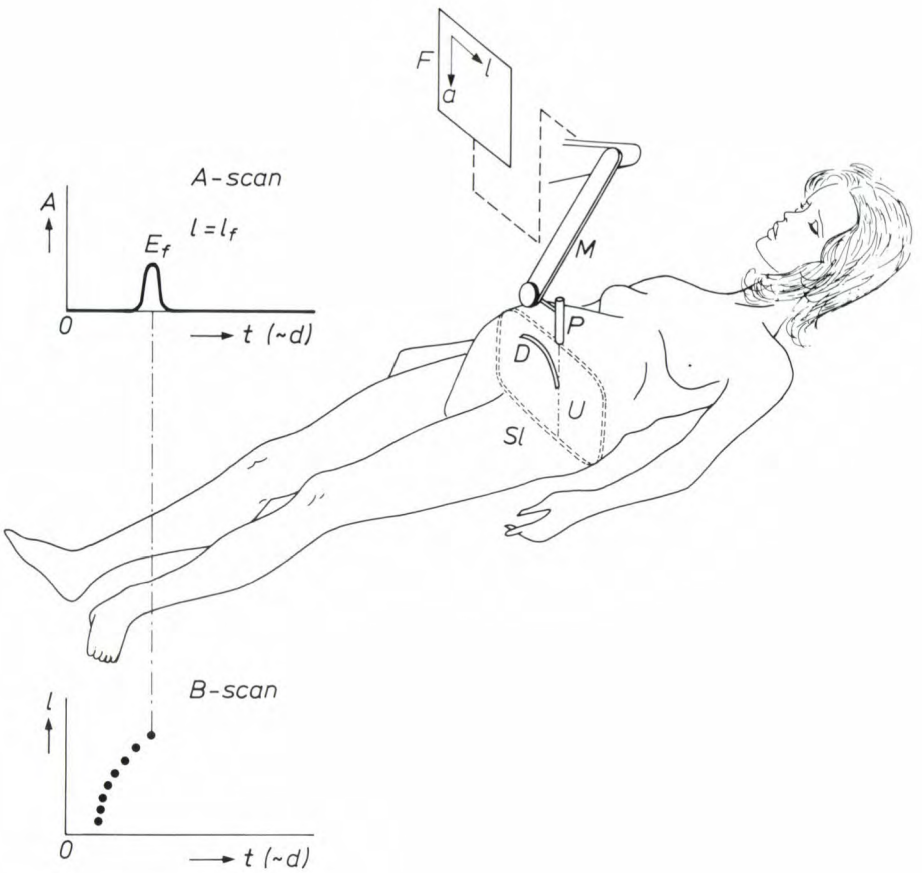


Fig. 1. Medical echography with B-type and A-type scanning. *P* probe, with ultrasonic transducer. This transmits a narrow beam *U* of pulsed ultrasonic energy and receives the echoes E_1, $E_2, \ldots,$ E_f from acoustic scatterers, such as the boundary of an organ (*D*). *M* mechanism for positioning *P* so that as it moves over the patient's skin it scans a flat slice *Sl* of the body. *F* coordinator, which establishes the instantaneous position (*l*,*a*) of the probe. The data required for this can be obtained with sensors located at the pivotal points of *M*.

An A-type echogram consists of the amplitudes of echoes originating on the instantaneous axis of *P*, as a function of the axial depth *d* at which an acoustic scattering centre is situated. This depth is proportional to the pulse travelling time *t*. A B-type echogram gives the amplitudes — in the form of a brightness modulation — as a function of the axial depth *d* and the lateral position *l*. B-type echograms represent a cross-section of the patient's body.

diameter of the transducer to the 'average' wavelength of the pulses — which may be regarded as ultrasonic wavefronts — is no more than 10 to 20, so that refraction effects become serious.

The new transducer and its electronics have been designed to improve the lateral resolution by keeping the transducer focused during the detection of echoes of a transmitted pulse. In addition, the focus travels 'in step' with the transmitted pulse so that any acoustic discontinuity caused by an echo always produces a sharp image. This method is known as 'dynamic wavefront tracking' or 'receiver-focus tracking' [3].

In echograms obtained with the new transducer in a water tank a lateral resolution of 1 to 2.5 mm has been achieved.

Fig. 3 shows the general arrangement of the improved echographic system. The transducer is different from the conventional type: the back of the piezoelectric disc now has five electrodes instead of one, consisting of a central electrode with four concentric rings around it. The central electrode is both transmitter and receiver; the four rings generally act as receivers only. If required, however, they can also be used for focusing the transmitted beam. An experimental model of the new transducer is shown in *fig. 4*.

The rings and the central electrode are connected, via preamplifiers, to individually adjustable delay lines. The system works as follows. The echo returning from a small, point-like discontinuity consists of a wavefront that is approximately spherical and travels towards the detecting electrodes. The delay lines are

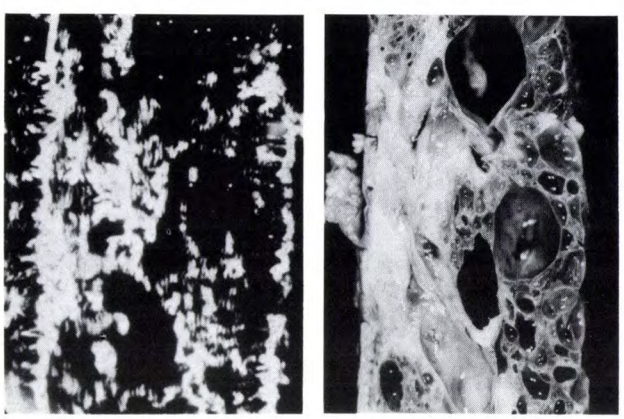


Fig. 2. *Left:* Echogram of a human kidney. The picture shows a longitudinal cross-section, obtained with a conventional echograph on the screen of a television monitor. *Right:* For comparison, a photograph of the same kidney, in the same cross-section, after removal from the body. Numerous cysts are visible. It is not difficult for an experienced observer to discover cysts from the echogram.

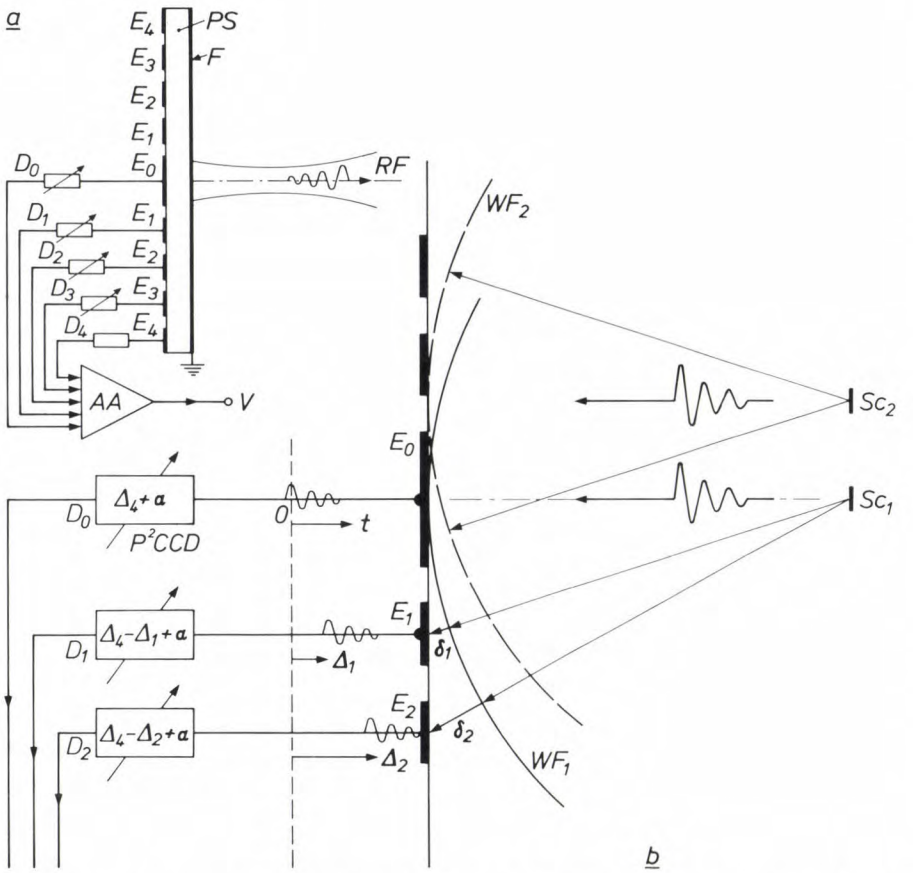


Fig. 3. *a*) Schematic cross-section of the new transducer, with a block diagram of the signal-processing electronics. *PS* thin piezoelectric disc (diameter 30 mm). *F* front electrode. E_0 central electrode for transmission and reception $E_{1,\ldots,4}$ ring electrodes for reception. *RF* the ultrasonic radiation field during transmission. $D_{0,1,\ldots,4}$ delay lines of the P²CCD type. *AA* adding and amplifying circuit. *V* video signal for building up the echogram. A transmitted pulse is a 'packet' of three oscillations, each with a period of $\frac{1}{3}$ μs. *b*) Principle of 'dynamic wavefront detection'. Sc_1, Sc_2 point-like discontinuities that scatter ultrasonic signals. Sc_1 is situated on the main axis of *PS*; Sc_2 is located a short distance away from the main axis. The discontinuities are both at the same depth. WF_1 and WF_2 wavefronts of the back-scattered signals. To obtain the best possible resolution the received signals are delayed in the delay lines $D_{0,1,\ldots,4}$ during periods equal to $\Delta_4+\alpha$, $\Delta_4-D_1+\alpha$, $\Delta_4-\Delta_2+\alpha$, $\Delta_4-\Delta_3+\alpha$, and α respectively. The quantity Δ_1 is the time taken by the pulse to travel the distance δ_1, etc., etc. The period α is the (fixed) time delay in delay line D_4. The five signals from WF_1 reach *AA* simultaneously, those from the wavefront WF_2 do not. During the time the transmitted pulse penetrates further into the body, the delay times (except α) are continuously decreased; this automatic adjustment is synchronous with the increasing depth at which the scattering discontinuities are situated.

adjusted in such a way that the signals produced by a spherical wavefront in the various electrodes activate the adder circuit *simultaneously* (fig. 3*a*). The delay times of the four delay circuits are continuously and automatically varied so as to keep the transducer tuned at all times to the successive reflections of the pulse entering the patient's body. Detection thus takes place with good resolution, irrespective of the depth at which the echo originates.

The delay times relating to a particular depth only produce a strong signal for echoes that come from the axis. Echoes from points of equal depth but not situated on the axis give a weaker signal, as can be seen from fig. 3. The delay lines we use are of the P²CCD type [4]; the delay they give is equal to 128 periods of the clock signal with which they operate. The frequency of the five clock signals can be set to any value between 15 and 30 MHz, giving delays from 8.5 to 4.2 μs. The clock signals are each generated by a separate voltage-controlled oscillator.



Fig. 4. Photograph of an experimental model of the new probe.

The desired modification of the delays is produced by automatically adjusting the control voltages of the clock oscillators for the delay lines D_0, D_1, D_2 and D_3. Oscillator D_4 is set to a fixed frequency, because the wavefront always reaches the outer ring last (E_4, fig. 3*a*). The four variable control voltages are derived from a simple $1/t$ voltage generator, via a bridge circuit at its output. The voltage generator itself is adjustable for selecting the desired value of maximum scan depth. The output signal of the voltage generator was given a $1/t$ time dependence for the following reasons.

The echo received by the central electrode at a time t after transmission of the original pulse comes from a discontinuity at

[3] See also F. L. Thurstone and O. T. von Ramm, Proc. Second World Congress on Ultrasonics in Medicine, Rotterdam 1973, p. 43 (Excerpta Medica, Amsterdam 1974).

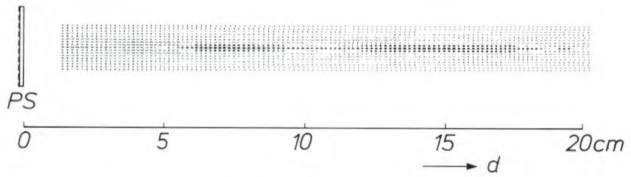


Fig. 5. Calculated radiation pattern of the new transducer (fig. 3). The calculation was based on the assumption that the transducer (*PS*) operates with 'axicon' geometry [6]. The ring electrodes (fig. 3) act as transmitters here, each with an advance that is proportional to the average radius of the electrode. The pressure of the ultrasonic energy in the radiated field is represented by means of a grey gradation; pulse duration 2 μs. At depths (*d*) between 7 and 10 cm and between 13 and 18 cm the ultrasonic energy is concentrated particularly closely around the acoustic axis.

the depth $\frac{1}{2}vt$, where v is the velocity of the ultrasonic pulse in human tissue. The same echo pulse reaches the outer ring electrode (average radius of r_4) after a slightly longer period τ, where:

$$\tau = t\left\{\tfrac{1}{2} + \tfrac{1}{2}\left(1 + (r_4/\tfrac{1}{2}vt)^2\right)^{\frac{1}{2}}\right\}.$$

It is obvious that the greatest delay must be applied to the central electrode signal. The delay time $\Delta(r)$ to be applied to the signal coming from the ring electrode of average radius r can be expressed in the form of a power series in $1/\tau$:

$$\Delta(r) = \left\{\left(\frac{r_4}{v}\right)^2 - \left(\frac{r}{v}\right)^2\right\}\frac{1}{\tau} - \left\{\frac{r^2 r_4{}^2}{v^4} - \left(\frac{r}{v}\right)^4\right\}\frac{1}{\tau^3} + \cdots .$$

This series converges so quickly that $\Delta(r)$ is approximated sufficiently well by its first term, which is proportional to $1/t$.

As we mentioned earlier, the signal can be transmitted from the central electrode alone. There is some pseudo-focusing of the transmitted beam, which is due to the disc shape of the transmitting part of the transducer [5]. The focusing can be improved somewhat, however, by using the ring electrodes as well for transmission. The pulses must then be transmitted from the rings fractionally earlier. If the advance is proportional to the radius of the electrode, the transducer behaves as an 'axicon' [6], i.e. a conical source which concentrates the acoustic energy on its axis. A calculated radiation pattern is shown in *fig. 5*.

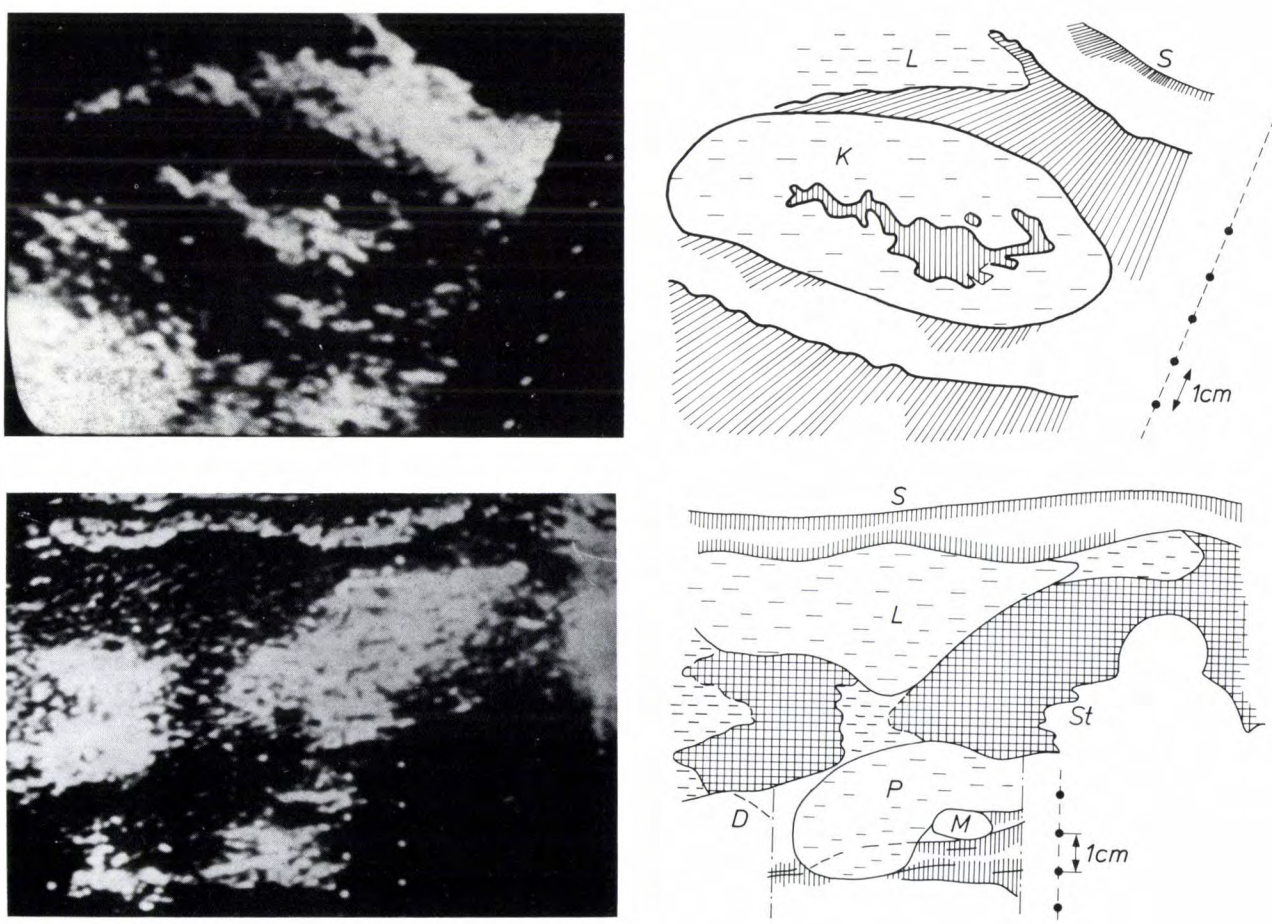


Fig. 6. Two examples of B-type echograms obtained with the new transducer. The echogram above is of a healthy human kidney. The fine detailing indicates a lateral resolution of about 2 mm. *K* kidney. *L* liver. *S* skin. The echogram on the right shows part of the abdominal cavity. It is possible to distinguish the liver (*L*), the pancreas (*P*) and a blood vessel in the mesenterium or vena mesenterica (*M*), as well as shadows of the duodenum (*D*), the stomach (*St*), and the patient's skin (*S*).

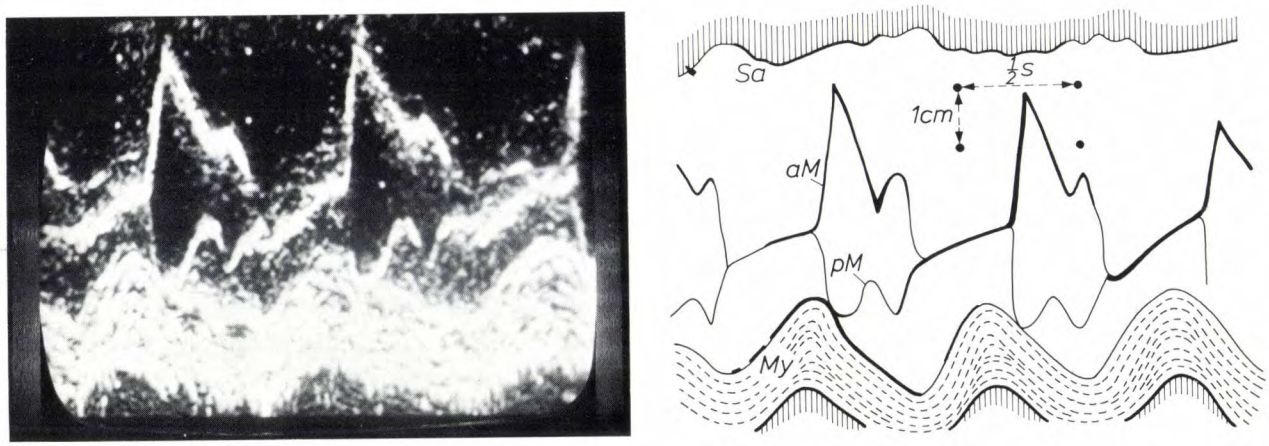


Fig. 7. Echogram of part of a human heart, *in vivo*. Taken in the *M* mode, this echogram allows the movement of various parts of the heart to be followed. This is done by means of an oscilloscope with a slow timebase (horizontal direction) and a long-afterglow screen. *Vertical*: the axial depth. *Sa* septum atriorum. *My* myocardium. *aM* valvula mitralis anterior. *pM* valvula mitralis posterior. This imaging method shows the movements as wavy lines. This echogram was also made with the new transducer.

Our experience with an echograph equipped with the new transducer (fig. 5) shows that echograms giving a lateral resolution of 1 to 2.5 mm are feasible. Drift in the oscillators that control the delay lines appears to limit the resolution in practice. Coupling the transducer to the patient sometimes causes difficulties; a flexible water bag may be required for coupling in certain cases. Two examples of the echograms recorded are given in *fig. 6*, with an explanatory diagram. *Fig. 7* is an example of an investigation of the heart in the 'M mode' (M for motion). In this mode of recording, the horizontal deflection on the screen is proportional to time, and constitutes a slow timebase. The vertical deflection, as in the ordinary echograms, is a measure of depth. Because of the long afterglow of the storage screen, a time plot is obtained showing the motion of various parts of the heart, for example the endocardium, the septum atrorium, and in particular the valvulae with their component parts.

Summary. Description of an improved piezoelectric transducer that acts as a focused receiver of ultrasonic pulses (duration about 1 μs, transmitted at a repetition rate of a few kHz) for medical echography. The transducer (diameter about 3 cm) has four concentric ring electrodes and a central electrode (diameter about 8 mm). The focusing is obtained by subjecting each of the signals from the five electrodes to an appropriate delay in delay lines of the P^2CCD type. While the transmitted pulse travels through the body, oscillators that generate the clock frequencies for the delay lines automatically match the delays to the depth reached. The lateral resolution in the echograms is 1 to 2.5 mm. During transmission the electrodes may in a similar way form a single conical source ('axicon'), which distributes the acoustic energy more homogeneously over the axis. Echograms of organs in the chest and abdomen are shown.

[4] L. J. M. Esser, Peristaltic charge-coupled devices: what is special about the peristaltic mechanism, in: P. G. Jespers, F. van de Wiele and M. H. White (ed.), Solid state imaging, Noordhoff, Leiden 1976, pp. 343-425.
[5] M. Auphan and H. Dormont, Ultrasonics **15**, 159, 1977.
[6] C. B. Burckhardt, H. Hoffmann and P.-A. Grandchamp, J. Acoust. Soc. Amer. **54**, 1628, 1973.

Philips tech. Rev. 38, 200-210, 1978/79, No. 7/8

Plasma etching in IC technology

H. Kalter and E. P. G. T. van de Ven

When an electrical discharge is set up in a gas containing halogens or oxygen, particles are produced that are chemically very reactive. These can be used to etch solids that are converted into volatile compounds by the particles. The method is known as plasma etching, and provides a valuable alternative to wet etching in IC technology. Plasma etching has been increasingly used in IC technology in recent years, largely because the method is highly compatible with the trend towards further miniaturization. The authors describe the three main systems for plasma etching and give results of their investigations on the control of the etching process by variation of the parameters.

Introduction

Until recently the only etchants used for etching small patterns in the production of ICs were liquid ones, such as solutions containing HF. With the continued miniaturization of ICs [1], it is now necessary to etch patterns whose smallest details have dimensions much less than 5 μm. Furthermore, the results must be reproducible. Even after improvement, this is not usually possible with wet etching methods [2]. It can however be done with a number of 'dry' methods: details of 1 μm or smaller can for example be made by 'physical ion etching', a method in which the atoms of a solid are physically removed by bombarding the surface with inert-gas ions [3].

Another dry etching method, plasma etching, is the subject of this article. In this method gas molecules in a gas discharge are fragmented into reactive particles: electrons, ions and neutral radicals, which are allowed to react *chemically* with the surface to be etched. A suitable choice of the reagents ensures that the reaction products are sufficiently volatile for the surface to remain free. In the silicon technology a discharge in tetrafluoromethane (CF_4) is widely used; the products of the discharge include atomic fluorine and the radical CF_3, and also the radicals CF_2 and CF, which are formed by further fragmentation [4]. *Table I* gives a number of equations for fragmentation and recombination reactions that can occur in such a plasma. In etching silicon, silicon dioxide (SiO_2) and silicon nitride (Si_3N_4) in a CF_4 plasma it is mainly atomic fluorine and sometimes the radical CF_3 that gives the conversion into the volatile SiF_4. The equations for the most likely etching reactions are also shown in Table I. As an illustration *fig. 1* shows photographs of a pattern of polycrystalline silicon on silicon dioxide that was obtained after etching with a plasma of CF_4 and oxygen.

In addition to silicon and its compounds, other materials that form volatile fluorides, such as the metals tungsten, titanium, molybdenum and tantalum, can be etched by fluorine-containing plasmas. Chlorine-containing plasmas can also be used. Aluminium can be etched in discharges of CCl_4 or BCl_3 [5], silicon in a discharge of CCl_4, and chromium (for the fabrication of chromium masks) in a discharge of chlorine and oxygen [6]. Volatile chlorides and oxychlorides are

Table I. A number of equations for possible fragmentation and recombination reactions in a CF_4 plasma and for etching silicon, silicon dioxide and silicon nitride with a CF_4 plasma.

Type of reaction	Equation
Fragmentation	$e + CF_4 \rightarrow CF_3^+ + F + 2e$
	$e + CF_4 \rightarrow CF_3 + F + e$
	$e + CF_3 \rightarrow CF_2 + F + e$
Recombination	$CF_3 + F \rightarrow CF_4$
	$CF_3 + CF_3 \rightarrow C_2F_6$
	$F + F \rightarrow F_2$
Etching	$Si + 4F \rightarrow SiF_4$
	$3Si + 4CF_3 \rightarrow 4C + 3SiF_4$
	$SiO_2 + 4F \rightarrow SiF_4 + O_2$
	$3SiO_2 + 4CF_3 \rightarrow 3SiF_4 + 2CO_2 + 2CO$
	$Si_3N_4 + 12F \rightarrow 3SiF_4 + 2N_2$

Drs H. Kalter is with Philips Research Laboratories, Eindhoven; Ir E. P. G. T. van de Ven is with the Philips Electronic Components and Materials Division (Elcoma) in Nijmegen.

formed in these processes. Some other metals such as nickel, iron and cobalt form no sufficiently volatile compound with fluorine or chlorine at the temperatures normally used for plasma etching ($\leqslant 200$ °C) and cannot therefore be etched in a plasma containing fluorine or chlorine. Gold occupies an intermediate position in this respect. Oxygen plasmas are widely used for removing organic substances such as photoresist masks [4].

Plasma etching offers a number of advantages over wet etching. The most important is that it can be used for etching smaller patterns. Detail dimensions can be smaller than 1 μm, while in wet etching the practical limit is about 4 μm. We have to remember here that with modern photolithography images of details of about 2 μm can be produced in routine work, whereas in electron and X-ray lithography the details can be smaller than 1 μm. However, there is little point in using these methods except in combination with an equally refined etching method.

Another advantage of plasma etching is that the process is relatively simple and can be fairly easily automated. The method is also generally more reliable than wet methods, partly because there are no difficulties from capillary and galvanic effects. Finally, there is no need for large quantities of corrosive acids, so that plasma etching is safer and less harmful to the environment than wet etching.

A difficulty with plasma etching is that the method is not yet sufficiently selective for some of the combinations of materials used in IC technology. Under certain conditions the wafers to be etched can also be damaged electrically ('radiation damage') by incident ions, electrons and photons from the discharge.

There are three main systems used for plasma etching, which differ in the shape of the reactor, the electrode connections and the gas pressure. The research described in this article was concerned with the most common of these systems, plasma etching in a barrel reactor.

Studies were made of the variation of the etch rate as a function of gas composition, gas flow rate, pressure and temperature. Special attention was paid to the radiation damage that can affect wafers when photoresist is removed by an oxygen plasma, which appears to be largely caused by the penetration of sodium ions.

Before examining the results, we shall discuss the three main plasma-etching systems and indicate some of their typical applications and limitations.

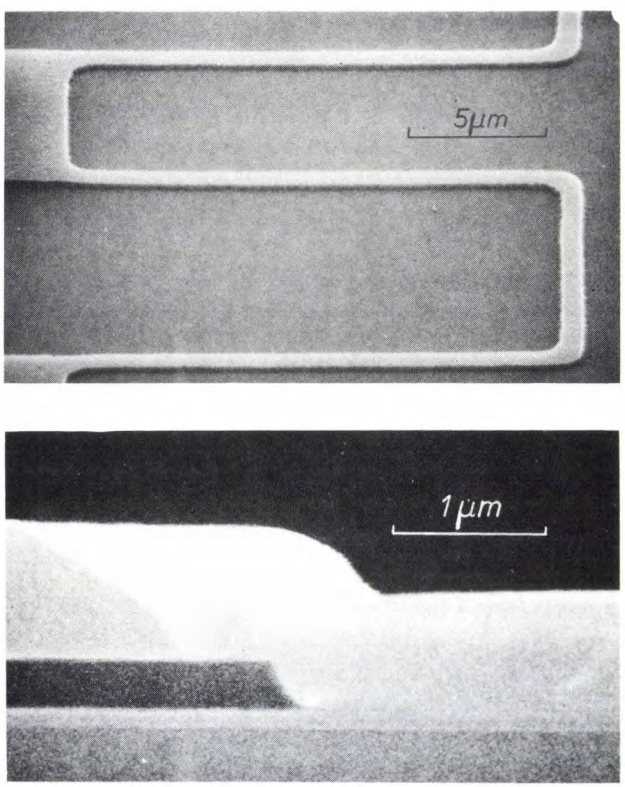


Fig. 1. Photographs made with a scanning electron microscope (SEM) of a pattern of a 0.3 μm polycrystalline silicon layer on a 0.1 μm layer of silicon dioxide after etching with a plasma of CF_4 and oxygen. The upper photograph is a plan view. The lower photograph is a side view; the polycrystalline silicon layer was coated with silicon dioxide beforehand to give better contrast. On the left, from bottom to top, there are the layers of single-crystal silicon, silicon dioxide, polycrystalline silicon and silicon dioxide again.

Systems

Diagrams of the reactors used in the three main plasma-etching systems are shown in *fig. 2*. In system *a* the reactor is tubular. The wafers with the material to be etched are screened from the discharge by a perforated metal cylinder (the 'etch tunnel'). The etching action arises because the neutral radicals formed in the discharge can reach the wafers by diffusion into the tunnel, provided their lifetime is sufficiently long.

[1] See for example the LSI issue of Philips Technical Review: Vol. **37**, 265-356, 1977 (No. 11/12).
[2] See for example R. Kilian and M. Liehr, Philips tech. Rev. **38**, 51, 1978/79, and J. J. Kelly and G. J. Koel, Philips tech. Rev. **38**, 149, 1978/79.
[3] Research on ion etching was the subject of an earlier article in this journal: H. Dimigen and H. Lüthje, Philips tech. Rev. **35**, 199, 1975.
[4] J. R. Hollahan and A. T. Bell (editors), Techniques and applications of plasma chemistry, Wiley, New York 1974.
[5] R. Reichelderfer, D. Vogel and R. L. Bersin, Extended Abstracts Electrochem. Soc. **77-2** (Fall Meeting, Atlanta 1977), p. 414.
T. O. Hordon and R. L. Burke, Proc. Kodak Microelectronics Seminar, Monterey 1977.
[6] C. B. Zarowin and E. I. Alessandrini, Extended Abstracts Electrochem. Soc. **77-2** (Fall Meeting, Atlanta 1977), p. 395.
K. Nishioka, H. Abe and T. Itasaka, *ibid.* p. 406.
H. Abe, K. Nishioka, S. Tamura and A. Nishimoto, Proc. 7th Conf. on Solid state devices, Tokyo 1975 (Jap. J. appl. Phys. **15**, Suppl. 15-1, 1976), p. 25.

Ions are excluded from the etching process by the presence of the tunnel. *Fig. 3* is a diagram of the arrangement for plasma etching in such a reactor. A vacuum pump and a gas-control system provide the desired gas pressure, usually between 30 and 300 Pa. The energy necessary for the discharge is supplied by the radio-frequency generator *RF*. The frequency is usually set to the value of 13.56 MHz, which is a permitted frequency for industrial operations. At such a high frequency there is no need for direct contact between electrodes and plasma. The energy is coupled to the plasma through an impedance-matching network and a capacitive coupling with two electrodes that enclose the reactor chamber.

cause the density of the gas is about one thousandth of its density at 10^5 Pa (1 atmosphere), while the viscosity is about the same. Reactive particles are therefore transported to the wafers mainly by diffusion. In these conditions the diffusion coefficient D and the lifetime τ of the different particles are of importance. The diffusion coefficient is inversely proportional to the gas pressure. For atomic fluorine and oxygen the value of D is about 500 cm²/s at 100 Pa and 150 °C, while the value of τ is about 0.3 s. The distance the particles can travel is about $2\sqrt{D\tau}$ and in this case is equal to about 25 cm. This is larger than the diameter of the tunnel, which implies that the reactive particles can easily cover the distance to the wafers; see *fig. 4*.

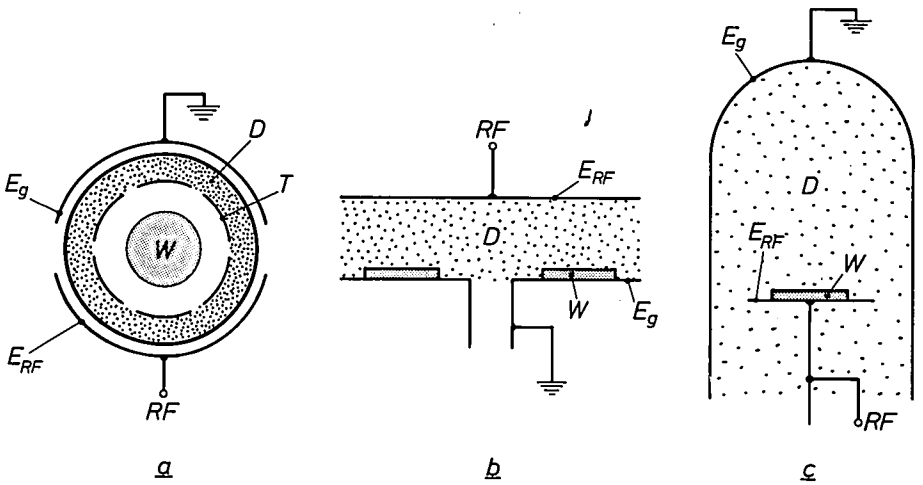


Fig. 2. Diagram of the reactors used in the three main systems for plasma etching.
a) Barrel reactor, containing a perforated aluminium tunnel T with the wafers to be etched W. E_{RF} electrode connected to the radio-frequency generator *RF*. E_g grounded electrode. Since the wafers are not in direct contact with the discharge D, the main contribution to the etching process comes from the long-lived radicals.
b) Planar reactor with the wafers on the grounded electrode. Since the wafers are now in direct contact with the discharge, short-lived radicals now contribute to the etching process. Ions also make a limited contribution.
c) Reactor for reactive ion etching. The wafers are located on the electrode connected to the r.f. generator; this electrode is much smaller than the grounded electrode here. The ion contribution to the etching is generally much larger than in (*b*).

As well as types with capacitive coupling of the r.f. energy there are also types in which the r.f. energy is coupled by a coil. Although these are said to be 'inductively coupled', this is not usually strictly correct. Because of the self-inductance such high voltages are produced across the coil that its elements start to behave like capacitive electrodes. The inductive energy transfer is only dominant at relatively low pressures ($\leqslant$ 1 Pa), since only electrons in circular orbits describe a path long enough for them to collide with molecules and not with the wall.

The reactor chamber is usually made of silica and the etch tunnel of aluminium. The wafers are placed in the tunnel in batches in a carrier. The gas flow goes mainly around the tunnel and is laminar. This is be-

To obtain uniform etching the etching reaction must take place at the same rate at the centre of the wafers as at the outer regions. The concentration of the reactive particles must therefore be as constant as possible between the wafers. This implies that there should be little recombination of the reactive particles during the diffusion between the wafers, and that the particles should not start to attack the surface of the wafers too quickly. If not, there would be non-uniformities from the outer regions to the middle of the wafers as a result of the decreasing concentration of reactive particles (fig. 4). In practice it is found that the reaction probability at each collision between a

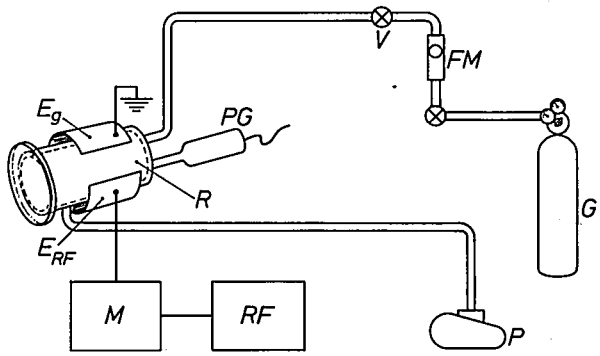


Fig. 3. Diagram of the arrangement for plasma etching in a barrel reactor. The reactor R is enclosed by the grounded electrode E_g and the electrode E_{RF}, which is connected to the generator RF via a matching network M. The reactor is connected to a gas supply G and a vacuum pump P. Gas pressure and the rate of flow can be controlled by the valves V, a pressure gauge PG and a flow meter FM.

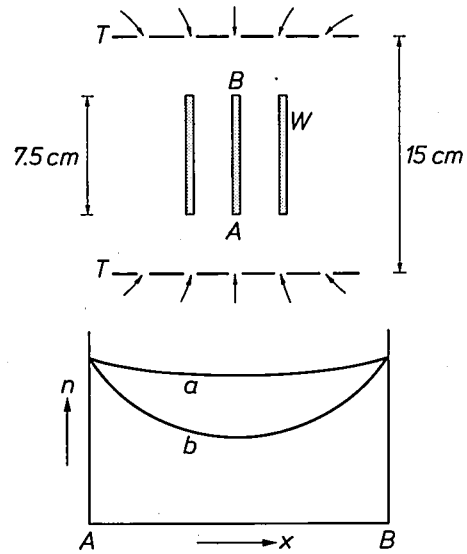


Fig. 4. Diagrammatic representation of plasma etching in a barrel reactor with etch tunnel T (*above*) and the number of reactive particles n, as a function of position x (*below*). Once they have entered the tunnel via the perforations, neutral reactive particles with a sufficiently long lifetime arrive between the wafers W. Because of recombinations and the etching reactions there are always fewer reactive particles near the centre of a wafer than near the edges A and B. In the situation corresponding to curve a the variation of n is very small, so that the wafer is uniformly etched. In the situation corresponding to curve b, on the other hand, the value of n decreases so much from the edges to the centre that the wafer is not uniformly etched.

reactive particle and a wafer must be less than 10^{-3} for uniform etching.

Aluminium is used for the metal etch tunnel in the barrel reactor, since it becomes coated with a protective layer of aluminium fluoride in a fluorine-containing plasma. A problem here is that this layer becomes partly hydrolysed in air, so that a porous film of aluminium oxide and aluminium fluoride gradually forms on the tunnel. A thick porous layer assists various undesirable recombination processes, and these lead to irreproducible results. It is therefore important to clean the etch tunnel regularly, e.g. with a solution of HF or H_3PO_4.

In system *b* the reactor consists of two flat electrodes forming parallel capacitor plates at a spacing of 1 to 5 cm. The wafers for etching are placed on the lower electrode and are in direct contact with the discharge. The gas pressure has a value between 10 and 100 Pa. The lower electrode is grounded; the upper one is connected to the r.f. generator. In system *c* the electrical connections are reversed. The r.f. electrode with the wafers for etching is small in relation to the grounded electrode and the gas pressure is often lower than in *b* (1-10 Pa).

During the discharge the electrodes assume a negative potential with respect to the plasma [7]. This is because the electrons have a much higher mobility than the ions, and are therefore able to reach the electrodes more easily. The magnitude of the negative potential depends on the discharge parameters and the geometry. The contribution of the ions to the etching process depends on the magnitude of the negative potential of the electrode with the wafers and on the gas pressure. In system *b*, where the wafers are located on the grounded electrode and the pressure is fairly high, the ions receive relatively little kinetic energy, so that the ion contribution to the etching process is limited. In system *c*, where the negative potential of the electrode with the wafers is relatively high, the ions can make a considerable contribution to the etching process, particularly when the gas pressure is relatively low ($\simeq$ 1 Pa). This arrangement for plasma etching is therefore sometimes called 'reactive ion etching' [8]. In *b* and *c* there is a contribution from neutral radicals as well as from the ions. Because of the free contact between the wafers and the discharge this contribution is not limited to the long-lived radicals, but those with relatively short lifetimes also take part in the process. It seems likely that at the same time the reactions of the neutral radicals are catalysed by electrons and ions.

The same kind of arrangement as in fig. 3 can be used for systems *b* and *c*, but with the different shapes for the reactors and different connections and electrode geometry. The r.f. frequency can vary very considerably, from 30 kHz to 30 MHz. In both systems the relative concentration of the various kinds of reactive particles, and hence the etching process, can be greatly affected by making variations in the electrode geometry, the gas composition and the pressure.

[7] J. L. Vossen and J. J. O'Neill, Jr., RCA Rev. **29**, 149, 1968.
[8] J. A. Bondur, J. Vac. Sci. Technol. **13**, 1023, 1976.

The etching profile

The profile obtained with plasma etching depends strongly on the relative contributions from the various kinds of reactive particles and therefore on the system; see *fig. 5*. Let us look first at etching in the barrel reactor. At a pressure of about 100 Pa the mean free path of the reactive particles in the reactor is of the order of 50 μm. This distance is much greater than the width of the etch grooves and the thickness of the layer to be etched. Since for each collision the probability of a reaction with a wafer is certainly no greater than 10^{-2}, the particles are incident from arbitrary directions and easily reach the sides of the etch grooves. The etching process is therefore isotropic and there is undercutting. The lateral dimensions of the pattern are thus to some extent determined by the thickness of the layer and the uniformity of the etching process. *Fig. 6* shows an example of the undercutting that occurs in isotropic plasma etching. Some of the polycrystalline silicon layer has been etched away even though it was covered with photoresist. When the dimensions become smaller than about 2 μm undercutting of this nature can reduce the accuracy too much and is no longer permissible.

In this respect plasma etching in a planar reactor can give better results because of the contribution from short-lived radicals and from ions. The short-lived radicals have a greater reaction probability on collision with a wafer than the long-lived ones. This means that they have a smaller probability of arriving at the sides of the etch grooves after collisions. As the relative contribution from short-lived radicals increases in relation to that from long-lived radicals, etching in the depth direction (anisotropic) increases and the undercutting decreases. The same is true, only more so, for the relative contribution from ions, since these are incident at right angles to the wafer and do not therefore cause undercutting. The situation is similar when the ions catalyse the reactions of neutral radicals, as probably happens in etching aluminium and polycrystalline silicon with a CCl_4 plasma or silicon dioxide with a CHF_3 plasma. *Fig. 7* gives an example of plasma etching in a planar reactor with hardly any undercutting.

To conclude, reactive ion etching is mainly anisotropic because of the relatively high ion contribution. In this respect this system is rather like the physical ion etching with accelerated inert-gas ions that we mentioned earlier [9]. In comparison with that system, however, reactive ion etching has the advantage that the etching action is mainly chemical, so that a lower ion energy is sufficient. This means that there are no problems with 'ditches', which are formed at the resist edges in physical ion etching (*fig. 8*). Another advantage over physical ion etching is that the reaction products are volatile, so that there is no undesired redeposition of material that has been removed. The diagram of *fig. 9* gives a schematic representation of three different redeposition processes that adversely affect the final result in physical ion etching.

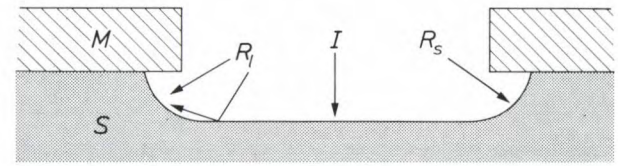


Fig. 5. Illustrating the effect of the different types of reactive particles on the etching profile. *M* photoresist masking. *S* specimen to be etched. Long-lived neutral radicals R_l can cause undercutting if they reach the sides of the etch groove directly or via collisions with the surface. Short-lived neutral radicals R_s give less undercutting since they can only arrive at the sides directly, not via collisions. Ions *I* are incident at right angles and do not give any undercutting.

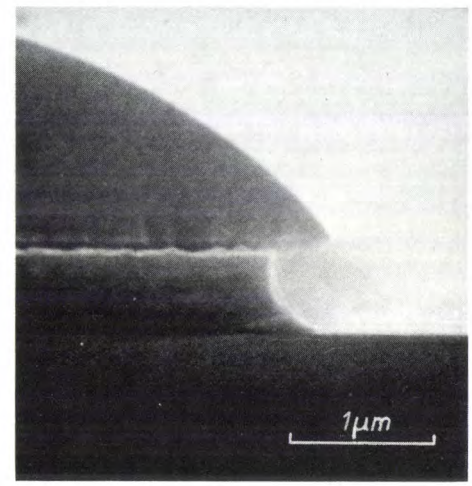


Fig. 6. SEM photograph taken after plasma-etching a polycrystalline silicon layer in a barrel reactor. Some of the silicon beneath the photoresist has also been etched away.

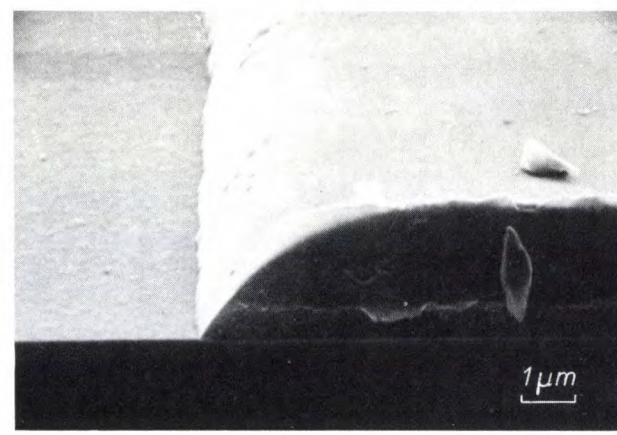


Fig. 7. SEM photograph taken after plasma-etching a polycrystalline silicon layer in a planar reactor. The silicon beneath the photoresist is almost unaffected by the etching.

Applications and limitations

Because of the differences in relative contributions from the reactive particles the typical applications and the limitations of the three systems are rather different. Plasma etching in a barrel reactor is widely used for etching silicon nitride and for removing photoresists. This system is unsuitable for dimensions smaller than 2 μm because of the isotropic etching profile. Another limitation is that it is not very easy to etch silicon dioxide on a layer of silicon without removing silicon as well. This happens because the etch rate of the oxide is usually much less than that of the silicon. It is also not possible to etch aluminium in a barrel reactor, since the aluminium is coated with a protective skin of oxide.

Aluminium can however be etched successfully with a CCl_4 plasma in a planar reactor and by reactive ion etching. The oxide skin is then removed by the bombardment of the charged particles or by reactions with the short-lived radical CCl_3. Plasma etching in a planar reactor and reactive ion etching are also suitable for etching polycrystalline silicon or silicon dioxide on silicon. Since the etching can be anisotropic the detail dimensions can be very small ($\leqslant 1$ μm). A difficulty with anisotropic etching in a planar reactor is that the etch rates of different materials are not very different in the gases commonly used in this system, so that the etching is not particularly selective. Disadvantages of reactive ion etching and, to a lesser extent, of plasma etching in a planar reactor are the erosion of photoresists and the radiation damage caused by ions and electrons. The radiation damage in reactive ion etching can be reduced by using a higher gas pressure and larger electrodes. Measures such as these can make the system comparable with plasma etching in a planar reactor.

Table II gives a summary of special features, the materials most commonly etched and limitations for the three main plasma-etching systems.

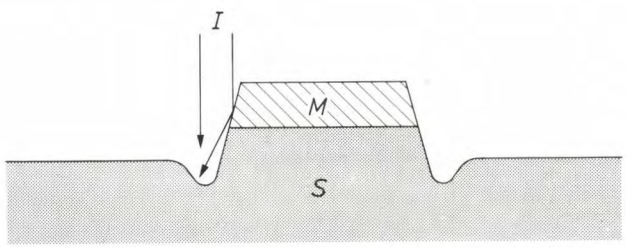


Fig. 8. Illustrating the formation of ditches in 'physical ion etching'. *M* photoresist masking. *S* specimen to be etched. In this form of etching there is some etching of the mask as well. The mask is etched most rapidly in the neighbourhood of the transitions between the upper surface of the mask and the sides. The slope of the sides of the mask with respect to the surface to be etched then becomes smaller, and ions are no longer incident parallel to the sides. Some of them collide with the sides of the mask before they arrive at the etch surface. There is consequently an increase in the number of incident ions close to the edges. This gives a local increase in the etch rate, so that ditches are formed.

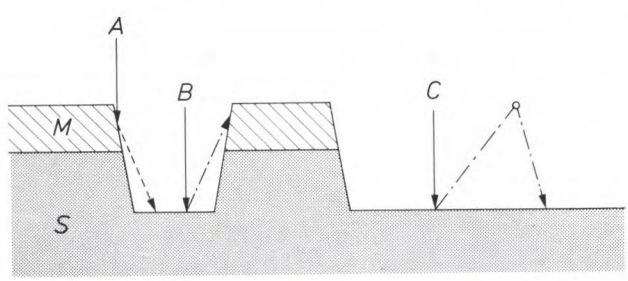


Fig. 9. Diagrammatic representation of some undesirable redeposition processes that can arise in physical ion etching. In process *A* particles from the photoresist masking *M* are scattered on to the surface of the specimen *S*; this delays the etching process. In process *B* particles from the specimen are sputtered on to the sides of the mask, which can cause the formation of sharp edges after the resist is removed. In process *C* particles that were originally sputtered from the specimen arrive back at the etch surface after colliding with gas atoms. This also delays the etching process. The effect increases in magnitude with increasing gas pressure.

The etch rate in a barrel reactor

As in other etching methods, the etch rate in plasma can be varied by changing the conditions. In our investigations of plasma etching in a barrel reactor we have found that the etch rate in this system is strongly dependent on the composition and flow rate of the gas, and on the temperature.

Gas composition

We have studied the effect of the gas composition on the etching of silicon and its compounds by CF_4 plasmas. The etch rate with a pure CF_4 plasma is relatively low. Adding oxygen increases the etch rate considerably, since more atomic fluorine is formed. *Fig. 10* shows the etch rate as a function of the oxygen content of the plasma for two kinds of silicon dioxide and for silicon nitride. The etch rate increases exponentially at first with the oxygen content and reaches a maximum at 2.5 to 3%. A further increase to 15% can be seen to have very little effect on the etch rate.

At low oxygen contents the effect of the gas composition on the etch rate depends on the material to be etched. The selectivity of the etching process can therefore be changed by changing to another oxygen content. This is illustrated in *fig. 11*, where the ratio R_N/R_O of the etch rate of silicon nitride to that of silicon dioxide is plotted as a function of the oxygen content.

A marked effect on the selectivity has also been found in etching silicon and silicon dioxide with a

[9] See the article by Dimigen and Lüthje quoted in [3].

plasma of CF_4 and oxygen. With a plasma of CF_4 with 80% of oxygen these materials are etched at about the same rate, whereas with a plasma of CF_4 with 8% of oxygen the etch rate for silicon is many times larger than that for silicon dioxide [10].

Gas flow rate

The effect of the gas flow rate for etching silicon dioxide and silicon nitride with a plasma of CF_4 with 4% of oxygen is shown in *fig. 12*. The etch rate R is found to be inversely proportional to the gas flow rate q [11]. A relation of this kind between $1/R$ and q would also be expected on theoretical grounds. In a fluorine-containing plasma the value of R is determined by the concentration c_F of atomic fluorine and the reaction-rate constant α for the chemical reaction that produces the etching:

$$R = \alpha c_F. \qquad (1)$$

The fluorine concentration is determined by the lifetime τ in the reactor and the generation rate g:

$$c_F = \tau g. \qquad (2)$$

The generation rate is assumed to be constant, since the r.f. power used in our experiments was always the same. The lifetime depends on the rate r of the heterogeneous recombination reactions at the reactor walls, the gas flow q and the rate of the homogeneous recombination reactions in the gas, which is a function of the pressure p. The reciprocal of the lifetime τ is given by:

$$1/\tau = c_1 r + c_2 q + c_3 f(p), \qquad (3)$$

where c_1, c_2 and c_3 are constants. On combining the

Table II. Some special features, the most widely etched materials and some limitations, for the three main systems for plasma etching.

	Plasma etching in a barrel reactor with etch tunnel	Plasma etching in a planar reactor	Reactive ion etching
Location of the wafers	outside discharge	inside discharge	inside discharge
Gas pressure	30-300 Pa	10-100 Pa	1-10 Pa
Reactive particles	long-lived radicals	long- and short-lived radicals	long- and short-lived radicals
		ions (small contribution)	ions (considerable contribution)
Etching profile	isotropic	isotropic or anisotropic	anisotropic
Materials most widely etched	silicon nitride	aluminium	aluminium
	polycrystalline silicon	polycrystalline silicon	polycrystalline silicon
	photoresist	SiO_2 on silicon	SiO_2 on silicon
Limitations	undercutting	undercutting sometimes	erosion of photoresist
	will not etch SiO_2 on silicon	some erosion of photoresist	radiation damage
	will not etch aluminium	some radiation damage	

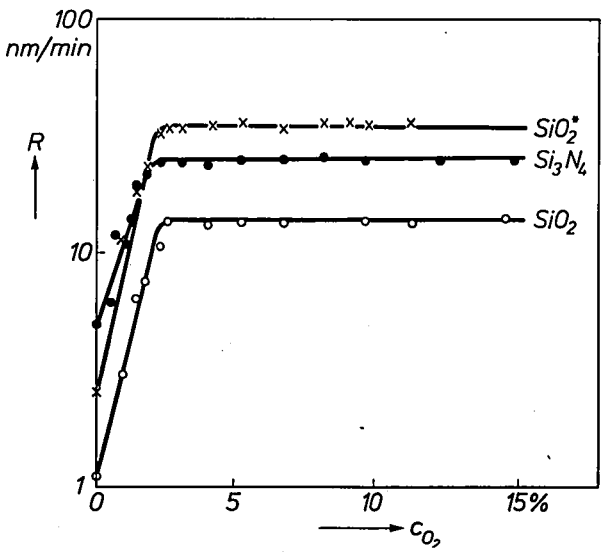


Fig. 10. The etch rate R plotted on a logarithmic scale against the oxygen content c_{O_2} for the etching of three silicon compounds in a barrel reactor with a plasma of CF_4 and oxygen. The label SiO_2* refers to silicon dioxide that has not been prepared by thermal oxidation of silicon, but by a gas-phase reaction between silane (SiH_4) and oxygen at a relatively low temperature (400-450 °C). The experiments were performed at a gas pressure of 133 Pa (1 torr), an r.f. power of 100 W, a temperature of 100 °C and a gas flow rate of 300 cm³/minute. Increasing c_{O_2} from 0 to 2 or 3% gives a large increase in R. At still higher values of c_{O_2} the value of R remains constant.

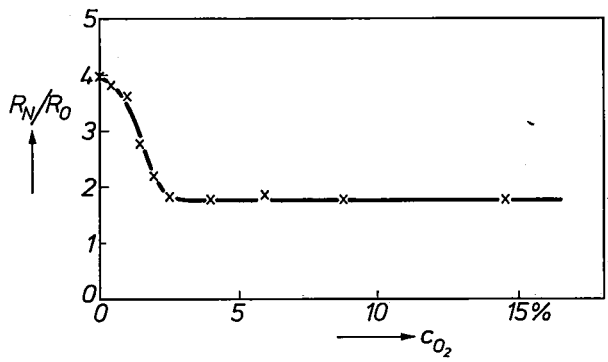


Fig. 11. The ratio R_N/R_O calculated from fig. 10 of the etch rate of silicon nitride and silicon dioxide, as a function of the oxygen content c_{O_2}. When c_{O_2} is increased to about 2% the value of R_N/R_O decreases and so therefore does the selectivity for both materials. At higher oxygen contents R_N/R_O is virtually constant.

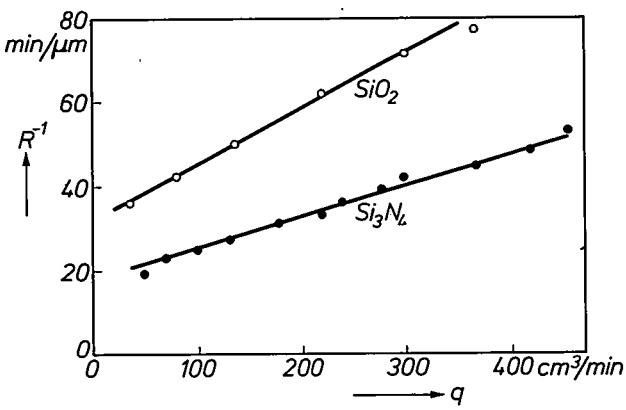


Fig. 12. Reciprocal of the etch rate R, plotted against the gas flow rate q, for silicon dioxide and silicon nitride etched in a barrel reactor with a plasma of CF_4 containing 4% of oxygen at a gas pressure of 133 Pa, an r.f. power of 100 W and a temperature of 100 °C. The relation between $1/R$ and q is linear for both materials.

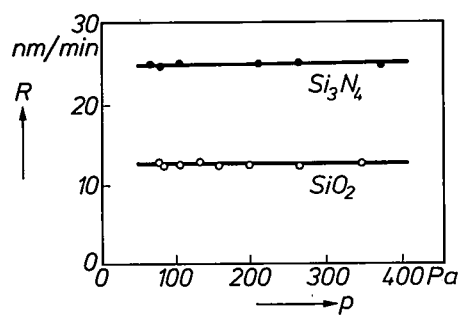


Fig. 13. Etch rate R as a function of the gas pressure p for silicon dioxide and silicon nitride etched in a barrel reactor with a plasma of CF_4 containing 4% of oxygen at an r.f. power of 100 W, a gas flow rate of 300 cm³/minute and a temperature of 100 °C. In the pressure range investigated the etch rate was found to be virtually independent of the pressure for both materials.

three equations above we obtain the following expression for the reciprocal of the etch rate:

$$1/R = c_1 r/\alpha g + c_2 q/\alpha g + c_3 f(p)/\alpha g. \quad (4)$$

In the pressure range that we studied the etch rate of silicon dioxide and silicon nitride in the CF_4 plasma with 4% of oxygen is almost independent of the pressure; see *fig. 13*. The homogeneous recombination reactions do not appear to play any significant part here. This implies that the constant c_3 in fig. 4 can be equated to zero, so that

$$1/R = c_1 r/\alpha g + c_2 q/\alpha g. \quad (5)$$

The parameters in this equation, apart from the reaction constant α, are independent of the material to be etched. The value that can be derived from the straight lines of fig. 12 for $c_1 r/c_2$, for example, is therefore found to be the same, 250 cm³/minute, for the two materials considered.

Equation (5) also indicates that a change in the gas flow rate does not affect the selectivity of the etching process. The ratio R_N/R_O of the etch rates of silicon nitride and silicon dioxide is equal to the ratio of their reaction constants for all values of q. The value of R_N/R_O calculated from fig. 12 is 1.77 ± 0.01.

As equation (5) shows, the etch rate reaches its maximum value when the gas flow rate is so low that the second term is negligible compared with the first. However, a maximum for the etch rate by no means guarantees optimum results from the etching process. Since the value of $c_1 r$ is determined by the interaction between the reactive particles and the reactor walls, it is strongly dependent on impurities and on the nature and roughness of the wall. These effects make it difficult to obtain reproducible results at low flow rates.

Temperature

The etch rate in plasma etching also varies strongly with the temperature. The magnitude of the temperature dependence is a function of the material to be etched and the gas composition. In all cases we have found that the effect of the temperature on the etch rate can be described by Arrhenius's equation for reaction velocity:

$$R = c \exp(-E_a/kT), \quad (6)$$

where c is a constant and E_a is the activation energy of the chemical reaction. The measured etch rates for four materials are plotted on a logarithmic scale in *fig. 14* as a function of $1/T$, for etching with a plasma of CF_4 with 4% of oxygen. *Table III* shows the values of the activation energy determined from the slope of the straight lines.

Differences in the activation energies of the various materials offer a means of varying the selectivity of the

[10] G. Bell and B. Hasler, Extended Abstracts Electrochem. Soc. 77-2 (Fall Meeting, Atlanta 1977), p. 411.

[11] For a planar reactor, however, the etch rate at constant pressure is almost independent of the gas flow rate. See C. J. Mogab, J. Electrochem. Soc. 124, 1262, 1977.

etching process by changing the temperature. As an example we show in *fig. 15* how the ratio R_N/R_O of the etch rate of silicon nitride to that of silicon dioxide increases with increasing temperature.

The activation energy is determined not only by the material, but also by the nature of the active particles. The etch rate is plotted on a logarithmic scale as a function of $1/T$ in *fig. 16* for the etching of silicon nitride by a CF_4 plasma with no oxygen, and by a plasma with an oxygen content of 8%. In spite of the large difference in etch rate there is no difference in slope, so that the activation energy is the same: 0.17 eV. We can therefore assume that the etching particles are the same in both plasmas. This is in complete agreement with the effect we discussed earlier of adding oxygen when etching with a CF_4 plasma.

The examples discussed here show that the temperature is an important factor in plasma etching. Good temperature control is therefore necessary for uniform and reproducible etching. When more than one wafer is being etched at the same time the temperatures of the wafers should be as far as possible kept the same. It must also be possible to etch a number of successive batches of wafers at an almost constant temperature. Good temperature control can be attained in a barrel reactor because of the use of the aluminium etch tunnel and by starting the etching process by preheating to a predetermined temperature in an inactive nitrogen plasma.

The experience summarized above provides a sound basis for highly reproducible plasma etching of silicon and its compounds. The desired etch rate for each material can be obtained by an appropriate choice of gas composition, gas flow rate and temperature. If necessary the selectivity for the different materials can be altered by varying the gas composition or the temperature.

Removal of photoresist masks

The plasma method can also be used for removing masks of organic photoresists [12]. This is done by allowing an oxygen plasma to act upon the photoresist in the barrel reactor; the resist is then converted into CO, CO_2 and water. The method is an attractive alternative to the conventional wet stripping in IC manufacture with reagents such as fuming nitric acid. Its advantages fall into two main groups. The first includes the reduction in environmental pollution, the absence of the need for certain safety precautions and the associated reduction in the process costs. The second group includes the increased compatibility with modern processing, the virtual absence of unwanted

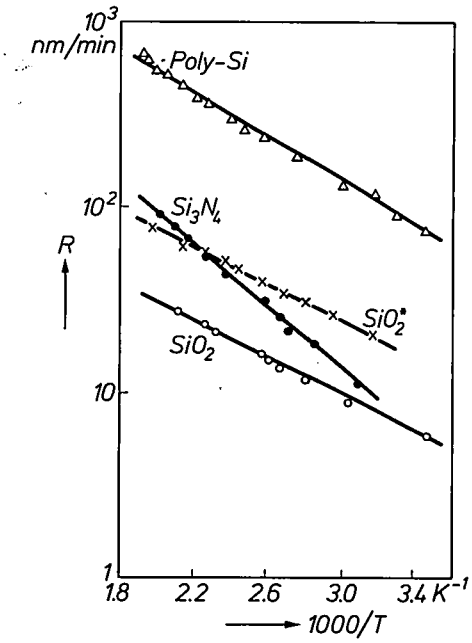


Fig. 14. Etch rate R, plotted on a logarithmic scale as a function of the reciprocal of the absolute temperature T for four silicon-containing materials etched in a barrel reactor with a plasma of CF_4 containing 4% of oxygen at a gas pressure of 133 Pa, an r.f. power of 100 W and a flow rate of 300 cm³/minute. In each case the plot is a straight line.

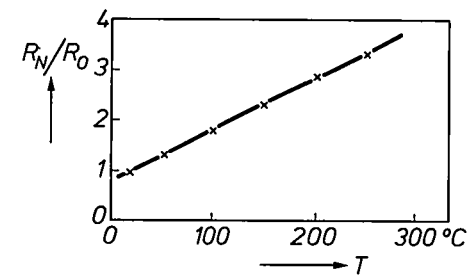


Fig. 15. The ratio R_N/R_O calculated from the results of fig. 14 of the etch rates of silicon nitride and silicon dioxide, as a function of the temperature T. In the temperature range investigated the value of R_N/R_O increases nearly linearly with temperature.

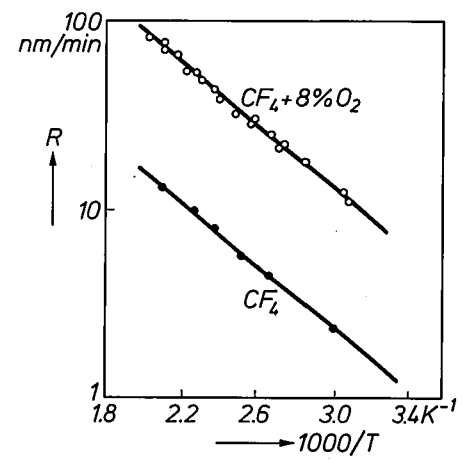


Fig. 16. Etch rate R, plotted on a logarithmic scale as a function of the reciprocal of the absolute temperature T for silicon nitride etched in a barrel reactor with a plasma of CF_4 containing no oxygen and 8% of oxygen. The gas pressure was 133 Pa, the r.f. power 100 W and the flow rate 300 cm³/minute. The plot is again a straight line in each case. The etch rates are very different, but the slopes are almost equal.

Table III. Activation energy E_a for the etching of several materials with a plasma of CF_4 with 4% of oxygen in a barrel reactor with an etch tunnel. The values of E_a were derived from the slopes of the straight lines in fig. 14.

Material	E_a (eV)
Polycrystalline silicon	0.12
SiO_2 (from thermal oxidation)	0.10
SiO_2* (from gas phase)	0.10
Silicon nitride	0.17

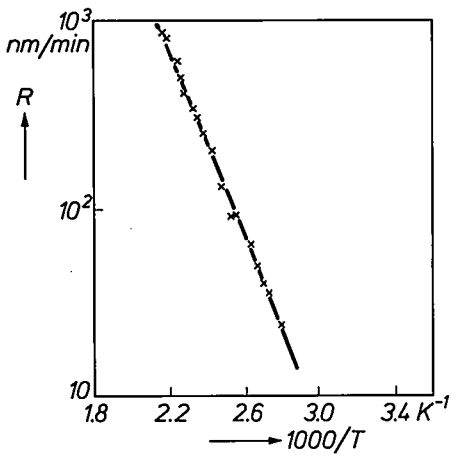


Fig. 17. Rate of removal R of a photoresist (Waycoat III) in a barrel reactor by an oxygen plasma, plotted on a logarithmic scale as a function of the reciprocal of the temperature T. As in the plasma etching of silicon-containing materials the plot is a straight line.

effects on the aluminium interconnections, and the reduction of the problems that can arise after ion implantation with photoresist masking.

In the investigation described here our main interest was in the temperature sensitivity of the process and the radiation damage in the processed wafers.

In a barrel reactor with no tunnel the wafers are quickly heated up to a temperature of 200 °C or more because of the direct contact with the plasma. With the tunnel the heating up takes much longer, so that the effect of the temperature on the etch rate is much more noticeable [13]. *Fig. 17* illustrates how the removal of a photoresist from a silicon wafer depends on the temperature. Just as in the plasma etching of silicon and its compounds (fig. 14) we again have a straight line, corresponding to Arrhenius's equation. The activation energy is found to be much higher in this case; it is about 0.5 eV.

For reasonably rapid removal of photoresists the temperature of the wafers should be at least 120 °C. The wafers can of course be heated up by omitting the tunnel. However, this has the disadvantage that ion and electron bombardment can cause radiation damage to the outer edges of the wafers before the resist has been completely removed from the centre. The wafers for sensitive circuits are first preheated in a nitrogen atmosphere at a relatively high r.f. power (about 500 to 800 W), with the plasma *inside* the tunnel. The photoresist, which is almost unaffected by this, protects the wafers from damage due to ion and electron bombardment. Next, the photoresist is removed in an oxygen discharge at a relatively low r.f. power (e.g. 100 W), with the plasma *outside* the tunnel.

Increasing the gas pressure from about 100 Pa to about 400 Pa has hardly any effect on the rate of removal of the photoresist. Changing the gas flow has a considerable effect, however. If the flow rate is not too low the rate of removal is inversely proportional to it. The results obtained are completely analogous to those obtained in etching silicon and its compounds (figs 12 and 13).

Radiation damage

When photoresists are removed by an oxygen plasma there may be some radiation damage to the wafers by ions, electrons and ultraviolet radiation. This can lead to instabilities in the finished circuits, as in an MOS transistor [14]. The seriousness of the damage is found to depend on the number of sodium ions that penetrate into the silicon-dioxide layer from the photoresist during the discharge [15]. It has been found from radiochemical analysis that the sodium content of the various types of photoresist is relatively high. Photoresists also contain a number of other elements that cause less damage and usually occur in considerably lower concentrations; see *Table IV*.

Table IV. Concentrations (wt ppm) of a number of elements in three types of photoresist. The concentrations were determined by radiochemical analysis after excitation by a neutron flux of 2.2×10^{11} cm^{-2} s^{-1} for three hours.

Resist \ Element	Na	K	Cu	Zn	As	Br	Ag
Waycoat III	2.8	<0.2	0.40	5.6	0.03	0.17	0.4
Shipley AZ1350H	7.8	0.54	0.32		0.013	0.14	<0.4
Shipley AZ111	13	0.65	0.30	11	0.02	0.15	<0.4

The sodium-ion transport can be traced with the radioactive isotope ^{24}Na. When the photoresist is removed without making use of the etch tunnel it is found that more than half of the sodium is located in the silicon-dioxide layer afterwards, and can no longer be removed by rinsing with water or concentrated nitric acid or by etching in a 1% HF solution; see

[12] This is often referred to as plasma stripping of photoresists.
[13] E. P. G. T. van de Ven and H. Kalter, Extended Abstracts Electrochem. Soc. **76-1** (Spring Meeting, Washington D.C. 1976), p. 332.
[14] J. A. Appels, H. Kalter and E. Kooi, Philips tech. Rev. **31**, 225, 1970.
[15] H. Kalter and E. P. G. T. van de Ven, Extended Abstracts Electrochem. Soc. **76-1** (Spring Meeting, Washington D.C. 1976), p. 335.

Table V. Residue of ^{24}Na on and in SiO_2 and SiO_2 coated with 30 nm of phosphosilicate glass (PSG), after removal of the photoresist (Waycoat III) in an oxygen plasma, both with and without tunnel. The figures quoted are percentages of the amount of ^{24}Na (about 10 ppm) initially used for doping the resist.

Material	Percentage of ^{24}Na	
	without tunnel	with tunnel
SiO_2		
without rinsing	100	100
after rinsing in H_2O or HNO_3	55	2
after etching for 30 seconds in 1% HF	52	1
SiO_2-PSG		
without rinsing	100	100
after rinsing in H_2O or HNO_3	45	20
after etching for 30 seconds in 1% HF	2	1

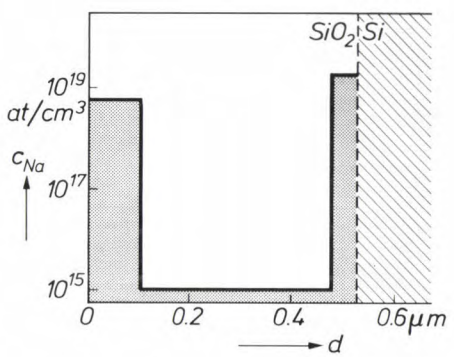


Fig. 18. Schematic diagram of the variation in the concentration c_{Na} of the radioactive isotope ^{24}Na in a silicon-dioxide layer as a function of the depth below the surface. The ^{24}Na originated in the photoresist (Waycoat III), which was first doped with the ^{24}Na, then removed from the oxide layer by an oxygen plasma in a barrel reactor with no etch tunnel. A considerable number of sodium ions are concentrated close to the boundary (*dashed line*) with the silicon layer beneath the oxide.

[16] R. A. Kushner, D. V. McCaughan, V. T. Murphy and J. A. Heilig, Phys. Rev. B **10**, 2632, 1974.

[17] This glass is generally used in IC technology to give protection from external contamination.

Table V. The diagram of *fig. 18* shows how the sodium has become distributed in the oxide layer. A significantly large number of the sodium ions are found to have travelled right through the oxide layer during the discharge, finally arriving close to the boundary with the silicon layer beneath the oxide. The driving force for this migration is a potential gradient in the oxide layer resulting from various interactions between the plasma and the wafer [16]. In an MOS transistor such a migration causes instabilities because of the occurrence of charge variations in the oxide layer [14].

When the etch tunnel is used the situation is much better. After the surface of the silicon-dioxide layer has been cleaned with a 1% HF solution, it is found that only 1% of the original amount of ^{24}Na is left in the oxide layer.

A considerable improvement can also be attained by coating the silicon-dioxide layer beforehand with a thin layer of phosphosilicate glass (PSG), which has the property of combining chemically with sodium ions [17]. This makes it almost impossible for the sodium ions to penetrate into the oxide layer. Even when the etch tunnel is not used most of the sodium ions do not travel further than the upper part of the PSG upper layer, as can be shown by etching away a part of this layer with a 1% HF solution (Table V).

Summary. In plasma etching solid materials are converted into volatile compounds by chemical reactions with reactive particles that are formed in a gas discharge. This method of etching is now more and more widely used in IC technology because of the trend towards further miniaturization. The extent to which the various kinds of reactive particles contribute to the etching process mainly depends on the particular system used for the plasma etching, which therefore has a considerable influence on the results. In the investigations described here, a study was made with one particular system — plasma etching in a barrel reactor — to find out which parameters were important in etching silicon-containing materials with CF_4 plasmas and photoresist masks with oxygen plasmas. The etch rate was found to be dependent on the material, the gas composition and the temperature, strongly dependent on the gas flow, but almost independent of the gas pressure. The electrical damage that often occurs to wafers when photoresist is removed is found to be the result of the migration of sodium ions originating in the photoresist. This damage can be prevented by choosing the conditions appropriately.

Garnet films for fast magnetic bubbles

D. J. Breed, F. H. de Leeuw, W. T. Stacy and A. B. Voermans

In bubble memories magnetic bubbles move about in a thin film of iron garnet on a non-magnetic garnet substrate. The maximum velocity of the bubbles is a particularly important quantity. An understanding of the dynamics of bubbles and Bloch walls, together with the experimentally discovered and surprisingly strong effect of an 'in-plane field' on the velocity of a Bloch wall, led to the idea that it ought to be possible to increase the maximum velocity considerably above the value that was usually found. This idea could be put into practice because of the control over the garnet system provided by many years of 'molecular engineering'. The many chemical substitutions that are possible make this system particularly adaptable, so that bubbles can usually be 'made to measure'. A subtle application of this approach at Philips Research Laboratories has resulted in films in which the maximum velocity is more than an order of magnitude greater than before.

Introduction

A bubble memory consists of a ferrimagnetic iron-garnet film grown epitaxially on a single-crystal substrate of a non-magnetic garnet, usually gadolinium gallium garnet, $Gd_3Ga_5O_{12}$ ('GGG') [1]. The principal conditions that will allow magnetic 'bubbles' to exist in such a film are a preferred direction of magnetization perpendicular to the plane of the film and a bias magnetic field of the correct strength, also perpendicular to the film. Under these conditions the film is one large, homogeneous magnetic domain that can contain small, mobile cylindrical domains of opposite magnetization — the bubbles (*fig. 1*). A memory device consists of shift registers for the bubbles; a bubble at a register position represents the binary '1' and the absence of a bubble the '0'. Small bubbles are preferable for a high bit density; in practice their diameter is between 2 and 7 μm [2].

In the present-day bubble memories the movement in the registers is produced by poles in Permalloy strips on the film, which are made to change in polarity by a rotating field in the plane of the film (*fig. 2*) [3]. With this 'field control' the information-processing rate is determined by the frequency of the rotating field. The technical problems of making rotating fields with the required characteristics limit the frequency to about 0.5 MHz, and hence the bit rate to 0.5 Mbit/s. Higher bit rates can be achieved by combining a number of parallel, field-controlled 'memory loops' with a single major loop that is 'current-controlled' (controlled by current pulses in patterns of conductors on the film). In this case the bit rate is limited by the fact that the movement of a bubble cannot be speeded up indefinitely: *the bubble velocity is tied to a maximum.* This is about 20 m/s in modern bubble memories,

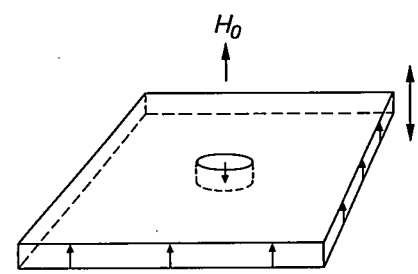


Fig. 1. Magnetic bubble in a thin ferromagnetic film. The double-headed arrow indicates the preferred axis of magnetization. Bubbles can only exist if there is a constant, homogeneous magnetic field H_0 of appropriate strength (the bias field).

Dr D. J. Breed, Dr F. H. de Leeuw, Dr W. T. Stacy and A. B. Voermans are with Philips Research Laboratories, Eindhoven.

[1] See for example J. A. Pistorius, J. M. Robertson and W. T. Stacy, Philips tech. Rev. 35, 1, 1975.

[2] A general treatment of magnetic bubbles is given in A. H. Bobeck and E. Della Torre, Magnetic bubbles, North-Holland, Amsterdam 1975, and in Hsu Chang (ed.), Magnetic bubble technology: integrated-circuit magnetics for digital storage and processing, IEEE Press, New York 1975.

[3] See W. F. Druyvesteyn, F. A. Kuijpers, A. G. H. Verhulst and C. H. M. Witmer, Philips tech. Rev. 36, 149, 1976.

which in practice would allow bit rates of about 1 MHz. In this article we shall discuss films in which we have measured bubble velocities that are more than an order of magnitude larger, about 500 m/s, which opens the way to much higher bit rates. In this introduction we shall mention some of the main stages in the development that led to this result; we shall deal with them in more detail later.

'Bubble dynamics' is based on the dynamics of domain walls. The simplest domain wall is flat (see *fig. 3*). Such a wall can be made to move by a magnetic field. In the situation shown in fig. 3, for example, the upward-directed field H_d favours the region of upward-directed magnetization and thus propels the wall towards the right. When the wall velocity (v) is plotted as a function of the 'drive field', the theories on the movement of domain walls unanimously predict a critical field-strength at which the velocity sharply decreases (*fig. 4*). At this field-strength there is a transition from a laminar to a turbulent movement [4]. In laminar movement the magnetization pattern in the wall is simple and also constant to an observer travelling with it. During the turbulent movement disordered disturbances occur, and the associated loss of energy is responsible for the lower velocity.

In experiments with the usual garnet films the laminar region is often obscured, because the critical field does not exceed the coercivity (the smallest field capable of detaching walls from defects in the material). The measured maximum velocities of about 20 m/s therefore broadly correspond to the theoretical value for the turbulent region [5]. To the best of our knowledge, the existence of a critical field-strength was first clearly observed in our experiments with an 'in-plane field', a constant field in the plane of the film [6] (*fig. 5*). Very briefly, the effect of such a field is to shift the critical drive field and the maximum velocity to higher values at which they can be observed. This also implies the possibility of much greater (laminar) velocities (see fig. 5).

A constant in-plane field, however, makes it impossible to control bubbles by Permalloy strips and a rotating field. As we shall see later, the in-plane field has a desirable effect because it introduces an 'in-plane preferred direction'. From this it is a natural step to think of a *built-in in-plane anisotropy*, i.e. of films possessing a magnetic *orthorhombic anisotropy*, in which the magnetization, in addition to the strong (primary) preferred axis perpendicular to the plane of the film, has another (secondary) preferred axis in the plane of the film. This is *not* an obstacle to the usual manipulation of bubbles. It was in films of this type that we measured the high bubble velocities of about 500 m/s mentioned above.

We should note here that very high wall velocities were observed some years ago in thin single-crystal wafers of an entirely different group of materials, the orthoferrites (e.g. $YFeO_3$) [7]. At the time these high velocities were also brought into relation with the magnetic orthorhombic anisotropy [8] that exists naturally in these materials because of their crystalline structure. The orthoferrites, however, can only contain bubbles larger than 25 μm, and therefore they are not of interest for memory applications.

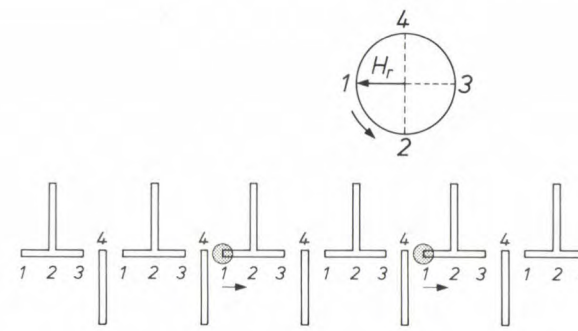


Fig. 2. Example of a pattern of Permalloy strips on a film for manipulating the bubbles. The rotating field H_r gives rise successively to the poles *1*, *2*, *3*, *4*, which carry with them any bubbles present. Considerably modified versions of this system are now used.

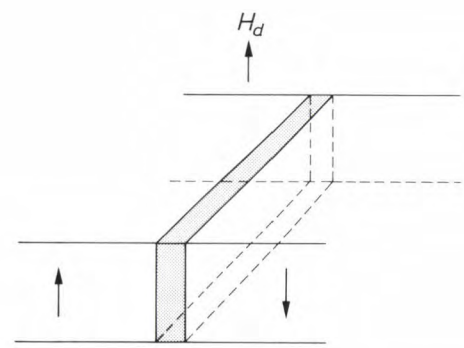


Fig. 3. Flat domain wall between two domains of opposite magnetization. The drive field H_d makes the wall move towards the right.

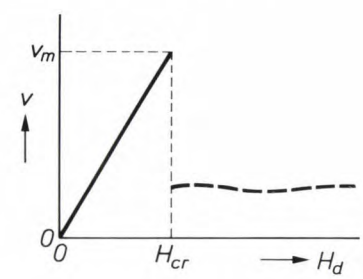


Fig. 4. Theoretical expectation for the wall velocity v. With increasing drive field H_d a 'laminar-turbulent' transition takes place at a critical field-strength H_{cr}; at this transition the velocity drops abruptly from its maximum v_m to much lower values.

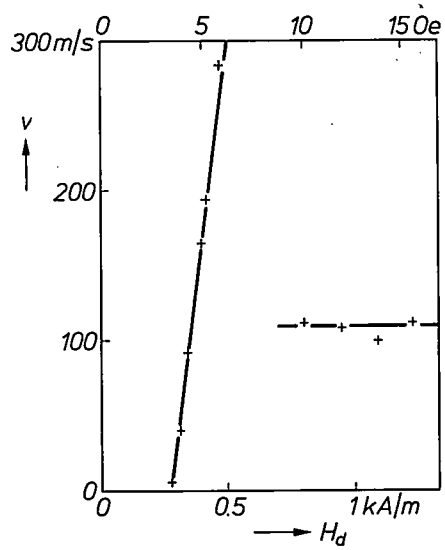


Fig. 5. Laminar-turbulent transition, observed in a film with a composition $Y_{2.9}La_{0.1}Fe_{3.8}Ga_{1.2}O_{1.2}$ in an in-plane field of 32 kA/m (400 Oe) [6].

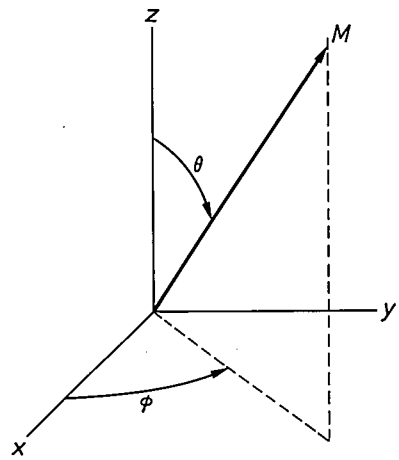


Fig. 6. The coordinate system used in the description of a Bloch wall. The z-axis is the preferred axis of the saturation magnetization M.

Iron garnets are cubic by nature. The anisotropy of the epitaxial layer, which must be at least uniaxial for bubbles and is obtained by means of a suitable choice of the composition and growth conditions [1], is of two types: 'growth anisotropy' and 'stress anisotropy'. The growth anisotropy is of great practical importance but its origin is not yet entirely understood. The stress anisotropy arises from a difference in the lattice constants of substrate and epitaxial layer. Since this can easily lead to cracks or other defects in the film, efforts to avoid it are often made. The films discussed in this article, however, owe their desired properties partly to the effect of stress anisotropy. The film is under compressive stress in its own plane. Owing to the magnetostrictive properties of the material, this stress produces the required primary preferred axis perpendicular to the film as well as the desired secondary preferred axis in the plane of the film.

Before discussing the films with orthorhombic symmetry that we have made, we shall discuss, with the aid of a simple model of the Bloch-wall movement, the basis of the beneficial influence of an in-plane preferred direction, and briefly summarize the methods used for measuring wall velocities.

Model of the laminar movement of a Bloch wall; effects of an in-plane field and of in-plane anisotropy

Let us consider an infinite medium with *uniaxial anisotropy*. In the coordinate system in *fig. 6* the easy direction of magnetization is that of the z-axis. The direction of the saturation magnetization M is given at every point by the polar angle θ and the azimuth ϕ.

Fig. 7 gives the structure of the *ideal flat Bloch wall* for the situation where the magnetization is directed upwards on the left and downwards on the right. The magnetization gradually rotates from $+z$ to $-z$ on passing through the wall along the y-axis, and remains parallel to the x,z-plane ($\phi = 0$). Each arrow in fig. 7 represents the magnetization at every point of a plane parallel to the x,z-plane. The structure is a compromise between the exchange interaction that tries to make neighbouring atomic magnets ('spins') parallel, and the anisotropy, which favours the $\pm z$-axis. It is the state of minimum energy under the given boundary conditions '$\theta = 0$ for $y = -\infty$' and '$\theta = \pi$ for $y = +\infty$'. The mathematical result $\theta = \theta(y)$ of the minimization is given in *fig. 8*. The quantity Δ defined in this figure is a measure of the wall thickness and is given by [9]:

$$\Delta = \sqrt{A/K_u}, \qquad (1)$$

where A is the exchange constant and K_u the anisotropy constant, i.e. the constants in the exchange-energy density w_{ex} (for $\phi = 0$) and the anisotropy-energy density w_{an}:

$$w_{ex} = A\theta'^2 \quad (\theta' = d\theta/dy), \qquad (2)$$

$$w_{an} = K_u \sin^2 \theta. \qquad (3)$$

The thickness of the wall therefore increases with the strength of the 'smoothing action' of the exchange interaction compared with the preference for the $\pm z$-axis.

The movement of such a Bloch wall can be traced back to the precessions of spins in magnetic fields

[4] See E. Schlömann, AIP Conf. Proc. 5, 160, 1972.
[5] J. C. Slonczewski, J. appl. Phys. 44, 1759, 1973.
[6] F. H. de Leeuw, IEEE Trans. MAG-9, 614, 1973.
[7] F. C. Rossol, Phys. Rev. Lett. 24, 1021, 1970.
[8] A. A. Thiele, Phys. Rev. B 7, 391, 1973.
E. Schlömann, J. appl. Phys. 47, 1142, 1976.
[9] See S. Chikazumi, Physics of magnetism, Wiley, New York 1964.

(see *fig. 9*). Let us imagine that in fig. 7 there is a drive field H_d in the direction of the z-axis. After a short time, owing to the precession in this field, all the spins acquire the same azimuth $\phi \neq 0$. Since the y-component (M_y) of the magnetization is now no longer zero, each thin layer parallel to the x,z-plane has become a magnetic double layer, which has no outward

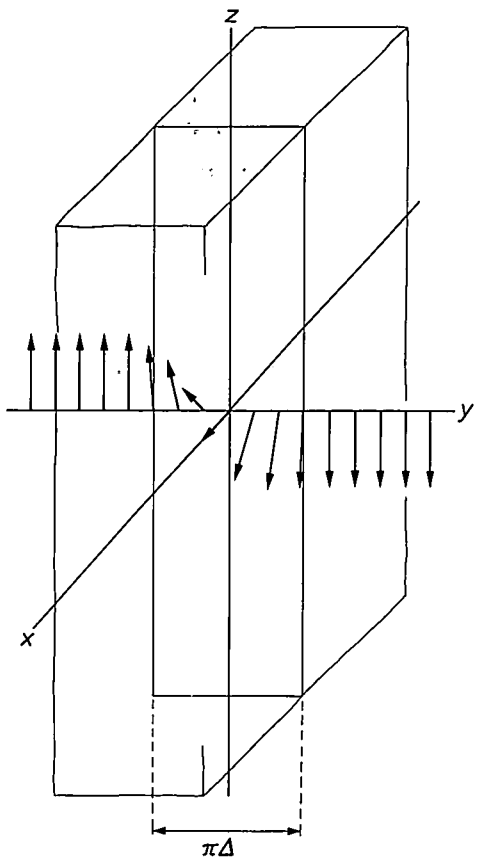


Fig. 7. Ideal planar Bloch wall. Each arrow represents the direction of magnetization at every point of a plane parallel to the plane of the wall (the x,z-plane); all the arrows are parallel to this plane.

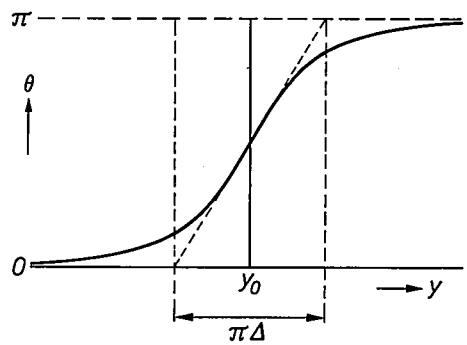


Fig. 8. The polar angle θ of the magnetization as a function of y for the ideal Bloch wall (fig. 7); θ is given by the equation

$$\sin\theta = 1/\cosh\{(y-y_0)/\Delta\}.$$

The quantity Δ in this equation is a measure of the wall thickness. As much as 74% of the change in θ takes place within the indicated interval $\pi\Delta$, 'the' wall thickness. The quantity Δ is often loosely referred to as the wall thickness.

effect, but operates internally as a 'demagnetizing field' $\boldsymbol{H}_m$ directed towards the left and given by:

$$H_m = M_y = M\sin\theta\sin\phi \text{ [10]}. \quad (4)$$

The Bloch wall moves because the spins successively complete a half-precession around this field. The movement is thus inseparably connected with a demagnetizing field, and hence with a value of the azimuth ϕ that differs from zero.

A drive-field pulse sets the wall in motion *gradually*, of course: the two precessions together form a single flowing movement. Besides, it is more realistic to envisage a medium in which precessions and wall movements are accompanied by 'friction losses'. Since an increasing number of spins become aligned in the direction of the drive field, this field supplies energy to the system. When this energy is set equal to the friction losses, the following relation is found:

$$v = \mu H_d, \quad \mu = \mu_0\gamma\Delta/\alpha, \quad (5)$$

where μ is the wall mobility, Δ is the wall thickness (1), γ is the gyromagnetic ratio, α is Gilbert's damping constant (see fig. 9) and μ_0 is the permeability of free space.

Closer analysis shows that the relatively very strong exchange interaction ensures that variations in ϕ in a moving wall are negligible over small distances; such variations would increase the exchange-energy density. This implies that, to an approximation, the spins rotate in a plane 'ϕ = constant', that is to say around the component $\boldsymbol{H}_\perp$ of $\boldsymbol{H}_m$ which is perpendicular to this plane (*fig. 10*):

$$H_\perp = M\sin\theta\sin\phi\cos\phi = \tfrac{1}{2}M\sin\theta\sin 2\phi. \quad (6)$$

The exchange-energy density is $A(d\psi/dy)^2$, where $d\psi$ is the angle between two spins at a distance dy. If ϕ also depends on y, we find:

$$w_{ex} = A(\theta'^2 + \sin^2\theta\,\phi'^2), \qquad (\phi' = d\phi/dy). \quad (7)$$

When $\phi \neq 0$, the energy density also has a demagnetization term:

$$\tfrac{1}{2}\mu_0 M_y H_m = \tfrac{1}{2}\mu_0 M^2\sin^2\theta\sin^2\phi. \quad (8)$$

The total energy density is therefore given by:

$$w = A\theta'^2 + A\sin^2\theta\,\phi'^2 + (K_u + \tfrac{1}{2}\mu_0 M^2\sin^2\phi)\sin^2\theta. \quad (9)$$

For $\phi = 0$ and $\phi' = 0$ we obtain from this, by minimization of the total energy, the structure in fig. 7 and fig. 8 for a stationary wall. Since the exchange interaction is a very strong effect, the energy becomes much greater when $\phi' \neq 0$. Very powerful means are therefore necessary to bring this about. If we exclude these, then ϕ is constant. On the other hand, with quite 'modest' means (a drive field) it is possible to obtain $\phi \neq 0$ and hence set a wall in motion, because the demagnetization makes only a modest contribution to w.

The structure of a moving wall with a given ϕ can be derived from that of the stationary one by substituting $K_u + \frac{1}{2}\mu_0 M^2\sin^2\phi$

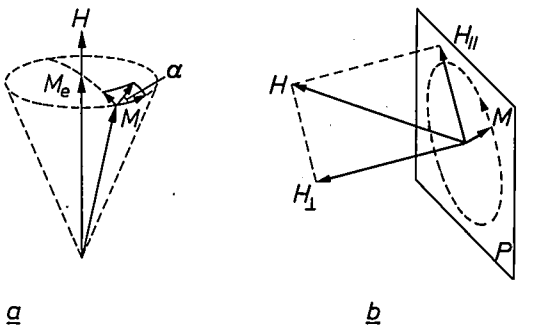


Fig. 9. *a*) Larmor precession. In a magnetic ion the magnetic moment M is coupled to an angular momentum L ($= -M/\gamma$); this makes the ion behave like a top ('spin'). The change in the angular momentum per second, dL/dt, is equal to the couple acting on the top, $\mu_0 M \times H$. It can easily be shown from this that in the magnetic field H the magnetic moment M describes a precession with an angular velocity

$$\omega = \mu_0 \gamma H.$$

The gyromagnetic ratio γ depends on the electron configuration of the ion. If the system is damped by its environment, then M spirals around H to reach its equilibrium state M_e. The velocity of the vector point then makes an angle α with the precession circle ($\alpha \ll 1$); α is referred to as Gilbert's damping constant [10].

b) If M is held in a plane (P) by other effects, then in an arbitrarily directed magnetic field H it can only describe the precession around the component $H_\perp$, perpendicular to P, with an angular velocity $\omega = \mu_0 \gamma H_\perp$; the precession around $H_{||}$ has been 'quenched'.

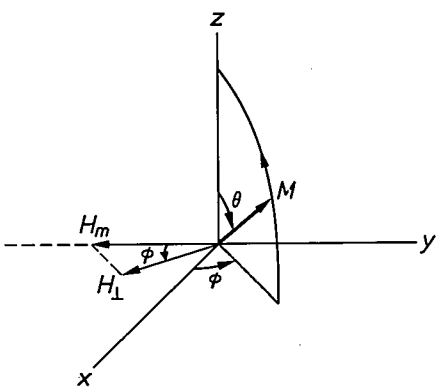


Fig. 10. In a moving wall, M precesses in a plane 'ϕ = constant' around the component $H_\perp$ of the demagnetizing field H_m perpendicular to this plane.

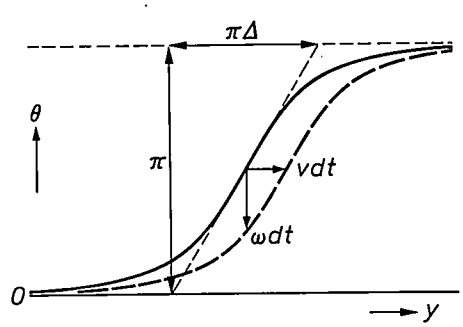


Fig. 11. The 'wall pattern' $\theta = \theta(y)$ at a given instant (*solid curve*) and at a time dt later (*dashed*). The wall has been displaced to the right because θ has decreased for the spins in the wall (they have rotated upwards in fig. 7). Evidently $v/\omega = \pi\Delta/\pi = \Delta$, where v is the wall velocity and ω the angular velocity of the spin at the centre of the wall.

for K_u, as appears from a comparison of (9) with (2) and (3). The thickness Δ_m of the moving wall is thus given by:

$$\Delta_m = \sqrt{A/(K_u + \tfrac{1}{2}\mu_0 M^2 \sin^2\phi)} = \Delta/\sqrt{1 + (\tfrac{1}{2}\mu_0 M^2/K_u)\sin^2\phi}, \quad (10)$$

where Δ is given by (1). The movement thus gives rise to *wall contraction*. This effect is usually a minor one ($\frac{1}{2}\mu_0 M^2 < K_u$). In the case of in-plane anisotropy, however, wall contraction can have a very marked effect (see p. 217).

Fig. 11 gives a plot of $\theta(y)$ at two instants separated by a time interval dt. The wall pattern shifts because θ changes for every spin in the wall. If we consider the centre of the wall ($\theta = \pi/2$, $\sin\theta = 1$), it follows from fig. 11 that:

$$v = \omega\Delta, \quad (11)$$

where ω is the angular velocity of precession at the centre of the wall for $H_\perp$ (see fig. 9 and eq. (6)):

$$\omega = \mu_0 \gamma H_\perp = \tfrac{1}{2}\mu_0 \gamma M \sin 2\phi,$$

so that:

$$v = \tfrac{1}{2}\mu_0 \gamma M \Delta \sin 2\phi. \quad (12)$$

The azimuth ϕ takes up a value at a given drive field such that the magnetic couples and the damping couples are in equilibrium. Calculation of this equilibrium via (12) gives (5). It follows from (12) that the velocity has a *maximum value*,

$$v_p = \tfrac{1}{2}\mu_0 \gamma M \Delta, \quad (13)$$

known as the Walker limit [11], which is reached when $\phi = \pi/4$. This is an indication of a 'laminar-turbulent' transition: velocities greater than v_p do not fit into this description of a laminar movement. If the drive fields are too strong, 'something else' must happen. A discussion of this does not come within the scope of this article.

The velocity limitation (13) can be avoided by applying a field H_1 in the x,y-plane. (Anticipating the later choice of the x,y-plane as the plane of the film, we will refer to it as an in-plane field.) If this field is much stronger than the demagnetizing field, it determines, for a given ϕ, the precession velocity and the wall velocity, which are therefore much greater. This situa-

[10] The equations in this article are valid when the international system of units (SI) is used; μ_0 is the permeability of free space ($= 4\pi \times 10^{-7}$ H/m), which should of course be distinguished from the Bloch-wall mobility μ. The equations valid in the Gaussian system can be obtained by substituting $4\pi M$ for M, $1/4\pi$ for μ_0 and $4\pi\gamma$ for γ. The values of quantities are expressed in both SI units and Gaussian units — the Gaussian units in brackets.

[11] L. R. Walker, quoted by J. F. Dillon Jr in: G. T. Rado and H. Suhl (ed.), Magnetism, Vol. III, p. 450, Academic Press, New York 1963.
N. L. Schryer and L. R. Walker, J. appl. Phys. **45**, 5406, 1974.

tion for the spins at the centre of the wall is illustrated in *fig. 12*. The precession velocity is given by

$$\omega = \mu_0 \gamma H_\perp = \mu_0 \gamma H_1 \sin \phi_1, \qquad (14)$$

where ϕ_1 is the angle between H_1 and the plane of rotation of the spins; here again ϕ and hence ϕ_1 is determined by the drive field. The maximum velocity is thus $v_m = \mu_0 \gamma H_1 \Delta$. In a more detailed model [6] the outcome is:

$$v_m = \tfrac{1}{2}\pi \mu_0 \gamma H_1 \Delta = (\pi H_1/M) v_p. \qquad (15)$$

In general terms, an in-plane anisotropy has the same effect as an in-plane field: it also introduces a preferred direction in the x,y-plane. The total anisotropy is then orthorhombic with the easy direction of magnetization along the z-axis. The anisotropy energy density w_{an} now takes the form:

$$w_{an} = (K_u + K_i \sin^2 \phi_1) \sin^2 \theta, \qquad (16)$$

$$K_u > 0, \qquad (16a)$$

$$K_u + K_i > 0, \qquad (16b)$$

where ϕ_1 is the azimuth measured from the in-plane axis x_1 (see *fig. 13*). The conditions (16a) and (16b) ensure that the z-axis is the preferred axis. The secondary axis is x_1 if K_i is positive and y_1 if K_i is negative. Even in the simple cases where the secondary axis is situated in the plane of the Bloch wall or is perpendicular to it, the theory [8] yields complicated expressions for the maximum velocity. If, however, the demagnetizing field has no effect ($|K_i| \gg \frac{1}{2}\mu_0 M^2$) and $|K_i|$ is much smaller or much larger than K_u, the result is simple again. After introduction of the dimensionless constants

$$Q_1 = 2K_u'/\mu_0 M^2, \quad Q_2 = 2|K_i|/\mu_0 M^2, \qquad (17)$$

where K_u' is the smaller of the two values K_u and $K_u + K_i$, we find:

$$v_m \approx 2\gamma \sqrt{\tfrac{1}{2}\mu_0 A Q_2} \quad (Q_2 \gg 1 \text{ and } Q_2 \gg Q_1), \qquad (18)$$

$$v_m \approx \gamma \sqrt{\tfrac{1}{2}\mu_0 A Q_2 (Q_2/Q_1)} \quad (Q_1 \gg Q_2 \gg 1). \qquad (19)$$

The situation corresponding to (18) is preferable to that corresponding to (19) when it is desired to obtain a high maximum velocity: for a given Q_2 ($\gg 1$) equation (18) always gives a higher value than equation (19). *Q_2 must therefore preferably be chosen much larger than Q_1.*

It is possible to derive relations rather similar to (18) and (19) by way of the following simple approach. As a result of the exchange interaction the spin rotation is again limited to a plane 'ϕ is constant'. At the centre of the wall ($\sin \theta = 1$) the change per unit time of the angular momentum, $-|\dot{M}|/\gamma$, is equal to the

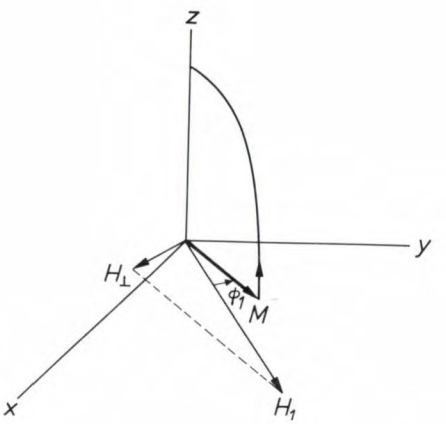


Fig. 12. Precession of the spin at the centre of the wall in the case of a strong in-plane field $\boldsymbol{H}_1$. The angular velocity is determined by $\boldsymbol{H}_\perp$, the component perpendicular to the plane of rotation 'ϕ = constant'.

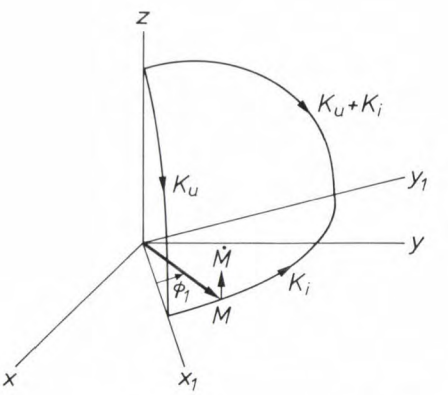


Fig. 13. The orthorhombic axes z, x_1, y_1. K_u, K_i and $K_u + K_i$ are the energy densities that have to be supplied to rotate the magnetization from z to x_1, from x_1 to y_1 or from z to y_1, respectively.

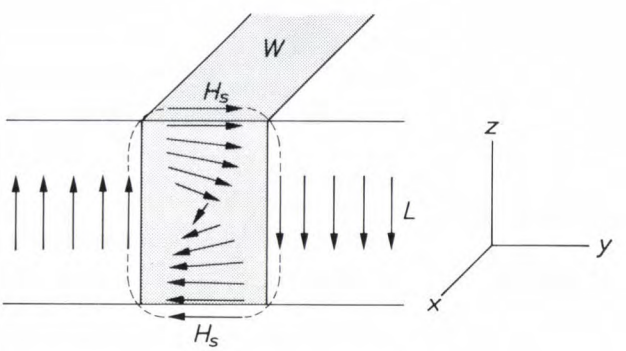


Fig. 14. Bloch wall W in a thin layer L. The magnetization to the left and right of the wall gives a stray field H_s that orients the magnetization in the wall at the surface of the layer along the y-axis. At the centre of an ideal wall the magnetization is directed along the x-axis (fig. 7); it is thus rotated at the surface of the layer.

couple $-\delta w_{an}/\delta\phi_1 = -K_1 \sin 2\phi_1$. The angular velocity ω is equal to $|\dot{M}|/M$. With $v = \omega\Delta_m$ we thus find for the maximum velocity ($\sin 2\phi_1 = 1$):

$$v_m = \Delta_m \gamma K_1/M = Q_2(\Delta_m/\Delta)v_p, \qquad (20)$$

where Δ_m is the thickness of the moving wall, which in this case, at the maximum velocity ($\phi_1 = \pi/4$), is given by:

$$\Delta_m = \Delta/\sqrt{1 + (Q_2/Q_1)\sin^2\phi_1} = \Delta/\sqrt{1 + Q_2/2Q_1}. \qquad (21)$$

In the extreme cases envisaged here this yields (18) and (19), apart from a factor $\sqrt{2}$ in the first case.

The movement of bubbles

The simplest bubble is a cylindrical domain with a closed Bloch wall; see *fig. 15a* and *b*. The wall is in equilibrium because the force on the wall F_H, due to the bias field H_0, which is everywhere directed inwards, is in equilibrium with two other forces, F_m and F_w. These forces can be derived from the magnetization energy of the bubble in its environment and the wall energy. The equilibrium determines the radius of the bubble [2]. If the bubble is in a field with a gradient (fig. 15*c*), then the force on one side predominates and the bubble starts to move. The Walker limit (13) should apply to the bubble in fig. 15*a*, *b*, *c*.

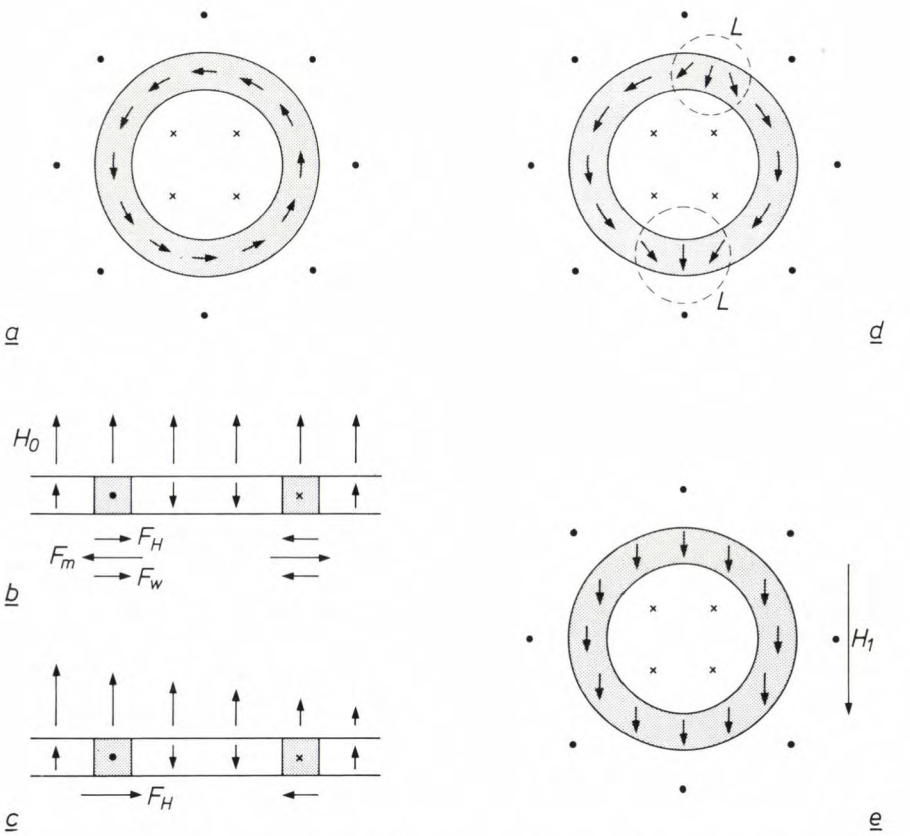


Fig. 15. *a*) The simplest bubble, seen from above. Inside the bubble the magnetization points into the paper (×), outside the bubble it points outwards (•). In the wall the magnetization is shown only at the centre of the wall. *b*) Cross-section of the bubble. H_0 bias field. F_H, F_m, F_w are radial forces acting on the wall resulting from the bias field, the demagnetizing energy and the wall energy, respectively. The balance between these forces determines the bubble radius. *c*) In a field with a gradient the bubble experiences a net force in one direction. *d*) A bubble with two Bloch lines L (also referred to as Néel segments), where the magnetization at the centre of the wall reverses direction. *e*) In a strong in-plane field H_1 the magnetization in the middle of the wall is everywhere in the same direction.

The theory outlined above for an infinite medium is directly applicable to *thin films*, provided that the in-plane field or the in-plane anisotropy is strong enough. If there is *no* in-plane field, the theory for thin films (Slonczewski's theory [5]) is very complicated owing to the stray field at the surface of the film, which causes a rotation of the magnetization at the centre of the wall (*fig. 14*). A strong in-plane field or a strong in-plane anisotropy straightens the magnetization in the wall, so that this complication does not occur and the theory for an infinite medium remains valid.

Apart from the complication in fig. 14, the structure of a bubble is often more intricate than that illustrated in fig. 15*a*. In fig. 7 the magnetization rotates clockwise upon passing through the wall. In terms of energy, however, there is no difference from a wall rotating anticlockwise. Both cases may easily occur in one bubble; there are then gradual transition regions,

referred to as Bloch lines (fig. 15*d*). This can complicate the behaviour quite considerably. In addition to 'normal bubbles' it has been shown experimentally that there are also 'hard bubbles', whose behaviour is very different; for one thing they are very much less mobile. This is due to the presence of a large number of Bloch lines.

Hard bubbles are unlikely to occur in a strong in-plane field (fig. 15*e*). Since the angle between the in-plane field and the wall has no effect on v_m, the same maximum bubble velocity is to be expected in this case as for a straight wall, provided that the in-plane field is strong enough. Both expectations have been confirmed by experiment: hard bubbles do not occur, and the velocity does not level off even at 400 m/s (*fig. 16*). Hard bubbles are also not found in the case of a strong in-plane anisotropy.

Wall-velocity measurements

The measurements that yielded the results in fig. 5 were carried out with the arrangement shown in *fig. 17a*. The garnet substrate with the film is located in the 'quadrupole field' (fig. 17*b*) produced by the four poles of two magnets. The film is situated in the central horizontal plane, where there is a vertical field H with a gradient (fig. 17*c*). At the centre, where H changes polarity, there is a *straight Bloch wall* ($y = y_1$).

At a time t_1 the field-strength is suddenly increased by a constant amount H_p by means of a small coil on the substrate. The equilibrium position of the wall is now situated at $y = y_2$, and the wall starts to move from y_1 to y_2. In an appropriate experimental arrangement (fig. 17*a*) one domain is more transparent than the other, owing to the Faraday effect, and the movement of the wall can thus be recorded with a photocell. The rate at which the photocell signal varies after $t = t_1$ is a direct measure of the wall velocity.

Situations with an in-plane field are obtained in this arrangement by vertical displacement of the film; it can be seen in fig. 17*b* that the field in the film then

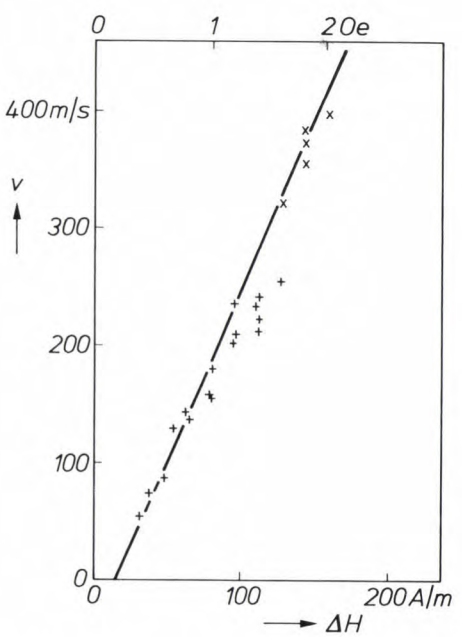


Fig. 16. Bubble velocity v as a function of the difference ΔH in field-strength across the bubble (see fig. 15c), for an in-plane field of 14.7 kA/m (185 Oe) (+) and an in-plane field of 22.7 kA/m (285 Oe) (×) [12].

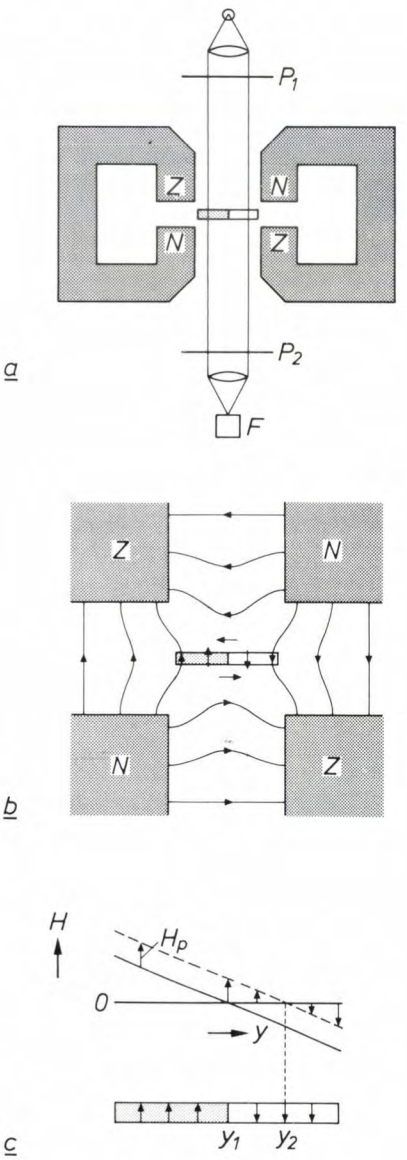


Fig. 17. *a*) Arrangement for measuring the velocity of a straight wall in a thin film. If there is only one wall in the film, and the position of the polarizers P_1, P_2 is correct, one domain is more transparent than the other (Faraday effect), so that the signal from the photocell F is a measure of the location of the wall. *b*) The quadrupole field in which the film is situated. In this field a thin film situated in the horizontal central plane has one straight wall at the centre. *c*) Field-strength H as a function of the horizontal coordinate y for $t < t_1$ (*solid line*) and for $t > t_1$ (*dashed*). From the time $t = t_1$ the straight wall moves from y_1 to y_2.

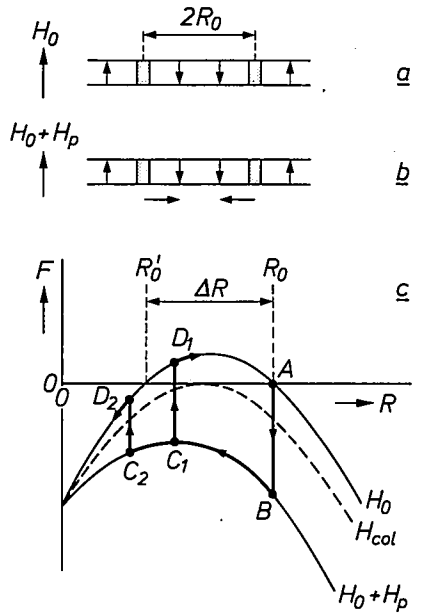


Fig. 18. The bubble-collapse technique. *a*) Bubble of radius R_0 in stable equilibrium in the bias field H_0. *b*) During an additional field pulse H_p the wall is subject to an inward-directed force. *c*) The radial force F as a function of the radius R of the bubble, for various fields; forces directed outwards are taken as positive, forces directed inwards are taken as negative. R_0, R_0' stable and unstable equilibrium values of R for $H = H_0$. ABC_1D_1 magnetic cycle for a short pulse; ABC_2D_2 for a longer pulse. In the first case the bubble returns to A after the pulse, in the second case it collapses. In the border-line case (pulse duration τ) the velocity is about $\Delta R/\tau$ ($\Delta R = R_0 - R_0'$). H_{col} is the collapse field.

acquires a horizontal component. In fact, it was during our experiments that we discovered the strong influence of an in-plane field, since the results were highly sensitive to vertical displacement of the film.

In our work on orthorhombic layers we made a great deal of use of the *bubble-collapse technique* [13]; see *fig. 18*. A film with bubbles is subjected to rectangular field pulses in the direction of the bias field H_0, of strength H_p. The radial force F on the wall (the sum of the forces F_H, F_m and F_w in fig. 15*b*) is shown in fig. 18*c* as a function of the bubble radius R for the fields H_0 and $H_0 + H_p$. A positive value of F corresponds to an outward force. In the bias field H_0 the force is zero — the bubble is in equilibrium — for two values of R, R_0 and R_0'; the equilibrium is *stable* only for $R = R_0$. When H is slowly increased, the bubble becomes smaller until the collapse field H_{col} is reached. With stronger fields the force is directed inwards for all values of R; the bubble collapses. During a field pulse the bubble passes through a series of states, for example ABC_1D_1 or ABC_2D_2, depending on the pulse duration. If $R > R_0'$ for the end point (e.g. D_1) then the bubble returns to A; if on the other hand $R < R_0'$, the pulse collapses. The pulse duration at a given pulse height H_p is now gradually increased and the pulse duration τ noted at which the bubbles (observed via the Faraday effect) start to collapse. In this case the wall has a mean velocity $\Delta R/\tau$ ($\Delta R = R_0 - R_0'$; R_0' can be calculated from the data for the bubble film); H_p is the drive field at the start of the pulse. A theoretical analysis [14] shows that, to obtain the wall velocity as a function of the drive field, all that is necessary is to plot $\Delta R/\tau$ as a function of H_p; both functions are identical as long as the movement of the bubble wall is determined by equation (5).

Finally, the movement of bubbles in a field with a gradient can also be directly observed. Short pulses of current are passed in the same direction along two parallel conductors on the film. An accurately known gradient field then exists between the conductors, and the pulse duration necessary to cause a bubble to cover a certain distance is then measured [15]. We used this method to obtain the results shown in fig. 16.

The orthorhombic stress anisotropy

The properties of a bubble film depend on the chemical composition, the orientation of the substrate surface and the conditions under which the film was grown. The composition can be considered to be derived from $Y_3Fe_5O_{12}$ ('YIG', yttrium iron garnet) by completely or partly replacing the yttrium by rare-earth ions and partly replacing the iron by another metal ion. The film is obtained by immersing a polished piece of substrate crystal (GGG) in a melt strongly oversaturated with the required oxides. When certain precautions are taken, the iron garnet grows epitaxially as a thin film on the surface of the substrate ('liquid-phase epitaxy', LPE [1]). The temperature and the degree of supercooling of the melt can be varied in this process. Growth anisotropy is obtained, as mentioned earlier, by choosing the appropriate growth conditions, and stress anisotropy is the result of a difference in lattice constants between film and substrate.

If we are to be certain of the orientation of the preferred axes relative to the film, only the crystallographic {111}, {100} and {110} planes of the cubic garnet crystal (*fig. 19*) can be used for the plane of the film (the surface of the substrate). In all other cases the preferred axes are temperature-dependent and sensitive to slight variation in composition and growth conditions [16]. In practice the {111} layers are often used

[12] F. H. de Leeuw and J. M. Robertson, AIP Conf. Proc. **24**, 601, 1975.
[13] A. H. Bobeck, I. Danylchuk, J. P. Remeika, L. G. Van Uitert and E. M. Walters, Proc. Int. Conf. on Ferrites, Kyoto 1970, p. 361.
[14] J. W. F. Dorleijn and W. F. Druyvesteyn, Appl. Phys. **1**, 167, 1973.
[15] G. P. Vella-Coleiro and W. J. Tabor, Appl. Phys. Lett. **21**, 7, 1972.
[16] R. D. Pierce, AIP Conf. Proc. **5**, 91, 1972.

because the anisotropy is then uniaxial with the axis perpendicular to the plane of the film. For our purpose, {111} layers and {100} layers were both ruled out from the start: the symmetry does not allow orthorhombic anisotropy to be produced in these layers. This does not apply to {110} layers, and we did indeed succeed in producing such layers with orthorhombic anisotropy. The relation between the x_1,y_1,z-coordinate system, in which the orthorhombic anisotropy is described by equation (16), and the cubic crystal axes is given in *fig. 20*.

Now it is known that in iron garnets containing rare-earth ions the preferred axis produced in a {110} layer by growth anisotropy is seldom perpendicular to the plane of the layer, as is required for bubble films [17]. We therefore opted for the addition of a strong stress anisotropy.

The difference in lattice constants between film and substrate causes an elastic stress in the plane of the film. Owing to the coupling between the elastic and the magnetic states of the material, this difference leads to a magnetic anisotropy. The same coupling is also responsible for the magnetostriction of the material; the anisotropy constants K_u and K_i are therefore directly related to the magnetostriction constants λ_{100} and λ_{111} of the material [9]:

$$K_u = -\tfrac{3}{4}\sigma(\lambda_{100} + \lambda_{111}), \qquad (22a)$$

$$K_i = +\tfrac{3}{4}\sigma(\lambda_{100} - \lambda_{111}), \qquad (22b)$$

where σ is the tensile stress, assumed to be isotropic,

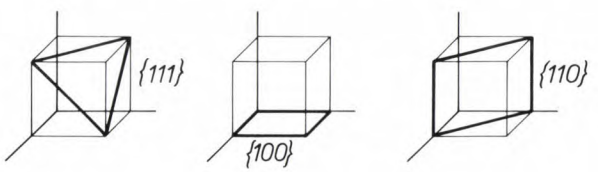


Fig. 19. A {111} plane, a {100} plane and a {110} plane in a cubic crystal structure. These are the only planes suitable for the surface layer in garnet films.

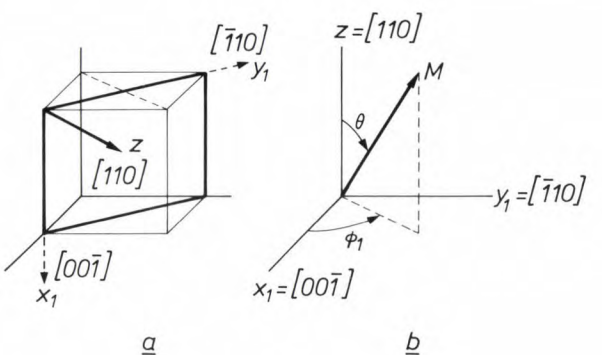


Fig. 20. The orthorhombic axes z, x_1, y_1, *a*) in the cubic crystal structure, *b*) in the position previously used.

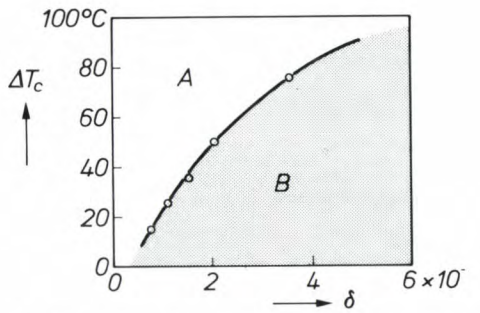


Fig. 21. In a film produced by liquid-phase epitaxy stress relief is a process that introduces impermissible defects in the layer. Its occurrence depends to a great extent on the mismatch and the degree of undercooling ΔT in the LPE process. The undercooling necessary to produce a film free from defects increases with the mismatch. The figure gives the critical degree of undercooling, ΔT_c, as a function of the mismatch parameter δ ($= (a_l - a_s)/a_s$, where a_l is the lattice constant of the stress-free layer, and a_s the lattice constant of the substrate). *A* is the region of defect-free films, *B* the region of films with a rough surface, due to stress relief. The measured points were found with $(Tb,Lu)_3Fe_5O_{12}$ and $(Eu,Lu)_3Fe_5O_{12}$ films on a {111} GGG substrate. The saturation temperature of the melt was about 940 °C.

in the plane of the film. A strong in-plane anisotropy thus requires a large difference in the magnetostriction constants.

To obtain a stress anisotropy large enough for our purpose it is necessary to have a much greater stress in the film than usual. For this reason we have to rely on *compressive stress*, i.e. a negative value of σ; large tensile stresses cannot be used since they would fracture the film. Furthermore, the z-axis in fig. 13 has to be the primary preferred axis, so that K_u and $K_u + K_i$ must both be positive. It follows from this, and (22),

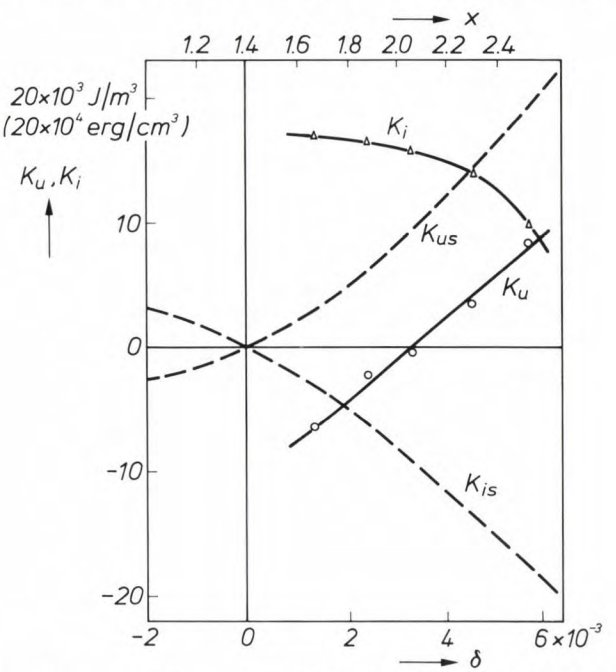


Fig. 22. Anisotropy constants K_u and K_i as a function of the mismatch δ for {110} $Eu_xLu_{3-x}Fe_5O_{12}$ films. K_{us} and K_{is} are the contributions to K_u and K_i predicted by equation (22). The Eu concentration x is derived from δ.

that there are two conditions for the magnetostriction constants:

$$\lambda_{100} + \lambda_{111} > 0, \quad (23a)$$

$$\lambda_{111} > 0. \quad (23b)$$

Films with europium

In bubble-garnet technology special physical properties are usually produced by means of rare-earth ion substitutions. *Table I* shows clearly the extent to which such substitutions are useful for our purpose. This table lists the lattice constants and the magnetostriction constants of the iron garnets $M_3Fe_5O_{12}$, where M represents a rare-earth ion or yttrium. As can be seen, the conditions (23) limit the choice to Eu or Tb. Because of the large damping constant — which would make very high drive fields necessary — Tb has to be ruled out, so that Eu remains.

Table I. The lattice constant a_0, the magnetostriction constants λ_{100}, λ_{111}, $\lambda_{100} + \lambda_{111}$ and the damping constant $\alpha M/\gamma$ of the iron garnets $M_3Fe_5O_{12}$, where M represents a rare-earth ion or yttrium [18]. In practice, it is found that α, Gilbert's constant used in fig. 9 and in equation (5), is not only proportional to the concentration of the damping ions but also inversely proportional to the total magnetization M. The product αM is therefore more suitable than α for comparing the damping contributions from the various kinds of ion; in practice the quantity $\alpha M/\gamma$ is used for this, where γ is the gyromagnetic ratio. A value is also quoted for the lattice constant of $Gd_3Ga_3O_{12}$ (GGG), which is used as the substrate for the garnet films.

M	Atomic number	a_0 (nm)	λ_{100} (10^{-6})	λ_{111} (10^{-6})	$\lambda_{100} + \lambda_{111}$ (10^{-6})	$\alpha M/\gamma$ (10^{-8} Js/m³rad or 10^{-7} Oe²s/rad)
Y	39	1.2376	−1.0	−3.6	−4.6	0.006
La	57	767				0.07 [*]
Sm	62	529	159	−183	−24	25
Eu	63	498	86	9.7	96	2.2
Gd	64	471	4.0	−5.1	−1.1	0.19
Tb	65	436	67	560	637	142
Dy	66	405	−254	−145	−399	52
Ho	67	375	−80	−56.3	−136	58
Er	68	347	10.7	−19.4	−8.7	8.5
Tm	69	323	25	−31.2	−6.2	1.3
Yb	70	302	18.3	−14.4	+3.9	2.2
Lu	71	283	−1.0	−3.6	−4.6	
$Gd_3Ga_5O_{12}$		1.2383				

When the yttrium in the starting material YIG, which has a somewhat smaller lattice constant than GGG, is gradually replaced by the much larger europium, the film gradually comes under increasing pressure. The stress that such a film can take, however, is limited; if the mismatch is too great, stress relief occurs via a complicated dislocation process [19]; defects arise that make the film unusable. The permissible mismatch depends on the growth conditions; one of the results from a special investigation of this dependence is the important practical result shown in *fig. 21* [20]. To keep the permissible Eu content as large as possible for a given maximum mismatch, we have combined the Eu with the smallest ion of the rare-earth series, Lu; this introduces no complications since Lu is magnetically inactive.

We have checked the expected influence of the Eu substitution on the anisotropy experimentally by measuring the constants K_u en K_1 in a number of films with the composition thus obtained [20]:

$$Eu_xLu_{3-x}Fe_5O_{12}. \quad (I)$$

The method used, in which the couple on a specimen in a magnetic field is measured by a torsion magnetometer, will not be dealt with here. The results of the measurements are given in *fig. 22*, which shows a plot of K_u and K_1 as a function of the 'mismatch parameter' δ, i.e. the relative difference in lattice constants ($\delta = (a_l - a_s)/a_s$, where a_l is the lattice constant of the layer and a_s that of the substrate). We measured δ for each film by X-ray diffraction (see *fig. 23*). The Eu content x was derived from δ, which was derived from a_l.

These results can probably be interpreted as follows. Of the two components from which the anisotropy is in general built up the growth anisotropy does not primarily depend on δ, whereas the stress anisotropy does depend completely on the mismatch. The stress-

[17] R. Wolfe, R. C. LeCraw, S. L. Blank and R. D. Pierce, Appl. Phys. Lett. **29**, 815, 1976.
[18] G. P. Vella-Coleiro, AIP Conf. Proc. **10**, 424, 1973.
[19] D. C. Miller and R. Caruso, J. Crystal Growth **27**, 274, 1974.
[20] W. T. Stacy, A. B. Voermans and H. Logmans, J. appl. Phys. **48**, 4634, 1977.
[*] Result of measurements at these laboratories.

anisotropy values expected from equation (22) are shown as dashed lines in fig. 22. The experimental curves are in fact reasonably parallel to the dashed ones; the fairly constant difference is then the growth anisotropy.

It follows from fig. 22 that in composition (I) films a δ of 3.5×10^{-3} is necessary to obtain a primary preferred axis perpendicular to the plane of the film ($K_u > 0$) i.e. more than ten times greater than the usual values of mismatch.

in-plane anisotropy ($Q_2 = 0$); the wall velocity is about 5 m/s, independent of the drive field. This value is broadly in agreement with the value that Slonczewski's theory [5] would predict for the 'turbulent' wall movement in this film. In film *B* the in-plane anisotropy dominates the demagnetization ($Q_2 \gg 1$); the velocity is an approximately linear function of the drive field ('laminar' movement) and shows no tendency to level off. The largest measured value is about 100 times greater than that in film *A*.

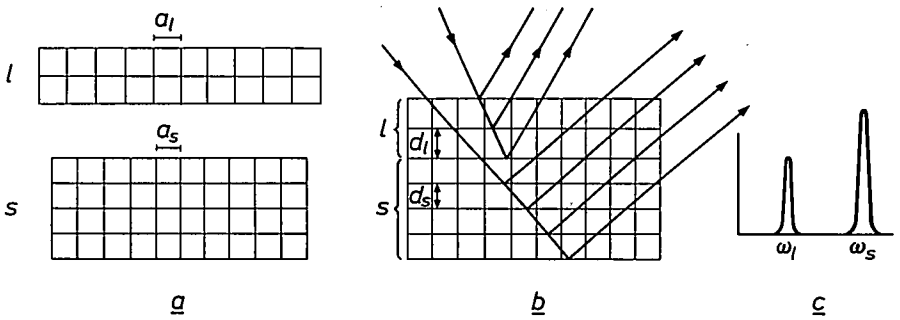


Fig. 23. Determining the difference $a_1 - a_s$ by X-ray diffraction. *a*) Substrate *s* and stress-free layer *l*. The layer *on* the substrate (*b*) is under compression ($a_1 > a_s$). Consequently the distance between the horizontal lattice planes d_l is greater than a_1, and therefore certainly greater than d_s, the distance between the horizontal planes of the substrate. The angle of incidence ω at which an X-ray peak appears is thus not the same for the layer as for the substrate (*c*). The angular difference gives $d_l - d_s$, and this in turn gives $a_l - a_s$ from the relation $(a_l - a_s)/a_s = \{(1 - \nu)/(1 + \nu)\}(d_l - d_s)/d_s$, where ν is Poisson's ratio.

In the calculation of the dashed curves from equation (22), σ was calculated from the relation $\sigma = E(a_s - a_l)/(1 - \nu)a_s$, where E is Young's modulus of elasticy and ν is Poisson's ratio. The λ-values were obtained by interpolation of the values for the pure Lu iron garnet and the pure Eu iron garnet.

As in all bubble garnets, the composition must also be magnetically diluted to obtain bubbles of the desired size (2-7 μm); the Fe is therefore partly replaced by Al (or Ga). The composition of our bubble garnet layers is therefore of the type

$$(Eu,Lu)_3(Fe,Al)_5O_{12}. \qquad (Ia)$$

One of our first proofs that the in-plane anisotropy also has the desired effect on the velocity of a Bloch wall can be seen in *fig. 24* [21]. Here the mean wall velocity, measured by means of the bubble-collapse method, is plotted as a function of the pulse field for a {110} layer and a {111} layer with approximately the same composition:

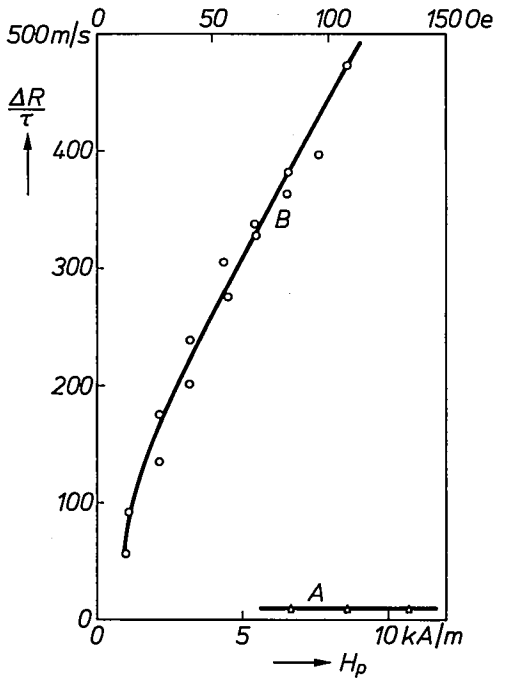
$$Eu_{2.7}Lu_{0.3}Fe_{4.3}Al_{0.7}O_{12}. \qquad (Ib)$$

Film *A*, as would be expected for a {111} film, has no

Films with manganese

As we have seen, europium is the only rare-earth ion that can be considered for rare-earth substitution to obtain the desired magnetostriction. It does, however, cause damping. From the slope of curve *B* in fig. 24 it follows that the mobility has a value of about 0.05 m²/As (400 cm/sOe), which is comparable with the mobility of bubbles in similar films without in-plane anisotropy. Although much higher velocities are therefore possible in the {110} Eu film than in the films normally used, proportionally stronger drive fields are necessary to achieve them. Although the damping contribution from Eu compares favourably with most other ions in Table I, it is still twelve times larger than that of Gd, for example. We wondered whether it might perhaps be possible to introduce the required magnetostriction by means of Fe substitutions, which are used elsewhere only for magnetic dilution. The rare-earth substitutions would then only be necessary to achieve

[21] W. T. Stacy, A. B. Voermans and H. Logmans, Appl. Phys. Lett. 29, 817, 1976.
[22] D. J. Breed, W. T. Stacy, A. B. Voermans, H. Logmans and A. M. J. van der Heijden, IEEE Trans. MAG-13, 1087, 1977.



◁ **Fig. 24.** Wall velocities ($\Delta R/\tau$) as a function of the drive field (H_p) for a {111} film (*A*) and a {110} film (*B*) of approximately the same composition $Eu_{2.7}Lu_{0.3}Fe_{3.4}Al_{0.7}O_{12}$. The results were obtained with the bubble-collapse technique (see fig. 18). The most important quantities are:

Film	Orientation	δ	M	$4\pi M$	Q_1	Q_2
A	{111}	3.5×10^{-3}	21.6 kA/m	(271 Gs)	27	0
B	{110}	3.0×10^{-3}	13.5 kA/m	(170 Gs)	12	48

the desired mismatch, and could be carried out with low-damping ions such as Gd, Y, La and Lu.

A search of the literature indicated that the Mn^{3+} ion is the only transition-metal ion that makes a large contribution to the magnetostriction and at the same time causes little damping at room temperature. We therefore investigated films of the type [22]

$$(Gd,Lu)_3(Fe,Mn,Al)_5O_{12}. \qquad (II)$$

Manganese films have another important practical advantage over europium films. In europium the anisotropy is directly connected with the mismatch; both are determined by the Eu concentration. The radius of the Mn^{3+} ion, however, is the same as that of the Fe^{3+} ion, so that the Mn substitution has no effect on the mismatch, which is determined by the Gd/Lu ratio. Mismatch and anisotropy can therefore be controlled independently in Mn layers.

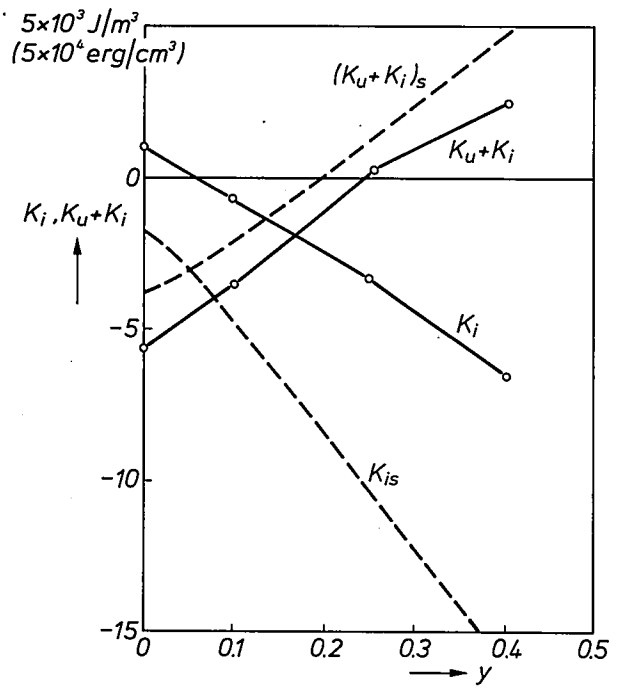
To determine the effect of the anisotropy, we again measured the constants K_u and K_i in magnetically non-dilute material; *fig. 25* gives the result for films with the composition

$$Gd_{2.1}Lu_{0.9}Fe_{5-y}Mn_yO_{12}. \qquad (IIa)$$

Fig. 25 also shows the curves that would be expected from equation (22) for the stress anisotropy. The experimental curves here are again seen to be approximately parallel to the theoretical ones, and the vertical displacement can again be attributed to growth anisotropy.

Fig. 26 gives the wall velocity as a function of the drive field, measured by the bubble-collapse method, both for a europium film and for a manganese film. The linear relationship points to laminar movement. We were unable to measure velocities higher than 300 m/s in these samples with our equipment, but the straight lines show no levelling off.

The mobility derived from the slope of the straight line is about four times greater in the Mn layer than in the Eu layer; the mobilities are 0.19 and 0.044 m^2/As respectively (1500 and 350 cm/sOe). We verified that the Mn ion does in fact introduce little damping from ferromagnetic resonance measurements on films of $(Gd,Lu)_3Fe_5O_{12}$ and $(Gd,Lu)_3(Fe,Mn)_5O_{12}$: the linewidths were the same.



◁ **Fig. 25.** The anisotropy constants K_1 and $K_u + K_1$ as a function of the Mn concentration y in {110}-$Gd_{2.1}Lu_{0.9}Fe_{5-y}Mn_yO_{12}$ films. In the range of practical interest K_1 is negative. This means that $K_u + K_1$ must be greater than 0 to obtain a primary preferred axis perpendicular to the plane of the film; y must therefore be greater than about 0.25. The $[\bar{1}10]$ axis (the y_1-axis in fig. 20) is then the secondary preferred axis. The Mn concentration y has no effect on the mismatch, and the anisotropy constants can therefore be varied at a constant mismatch by varying y. In the film used here, δ was equal to 3.4×10^{-3}. K_{1s} and $(K_u + K_1)_s$ are again the contributions predicted by equation (22).

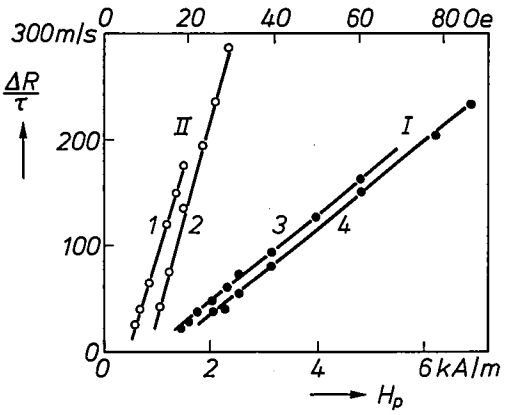


Fig. 26. Wall velocities ($\Delta R/\tau$) as a function of the drive field (H_p) for a europium film (*I*) and for a manganese film (*II*). The straight lines *1*, *2*, *3* and *4* were obtained at different values of the external field H_0. The slope indicates a mobility of 0.044 m²/As (350 cm/sOe) for the Eu film and 0.19 m²/As (1500 cm/sOe) for the Mn film.

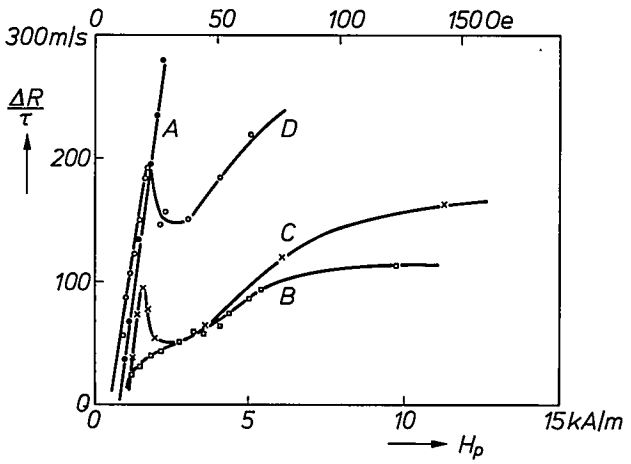


Fig. 27. Wall velocity as a function of the drive field for four films of the type $(Gd,Lu)_3(Fe,Al)_{5-y}Mn_yO_{12}$ with different y and different mismatch δ.

	y	δ	Q_1	Q_2
A	0.40	2.5×10^{-3}	25	41
B	0.25	1.4	11.5	9.2
C	0.30	1.6	13.0	11.6
D	0.35	1.8	17.9	12.6

We made use of the adaptability of the manganese system to make visible the levelling off or the 'laminar-turbulent' transition — which could not be done with the films mentioned above — by reducing both the Mn concentration and the mismatch, and hence the in-plane anisotropy [23]. *Fig. 27* gives the results of bubble-collapse measurements; the caption gives the specifications of the films used. Film *A*, with the greatest mismatch and the highest Mn concentration, has a large Q_2 and gives no levelling off. Film *B*, with the smallest mismatch and the lowest Mn concentration, has the smallest Q_2 and is the first to level off. The other two films are intermediate cases, with a distinct 'Walker peak'.

Further research

To conclude we should briefly mention some results we have obtained from further research. In the first place the smoothing action of the in-plane anisotropy is not so complete as that of an in-plane field (fig. 15*e*); although films with strong orthorhombic anisotropy do not contain any hard bubbles, they do appear to have bubbles of several types.

In the second place, the possibilities for the mobility have certainly not yet been exhausted with the films discussed above. In films with the composition

$$(Gd,Y)_3(Fe,Mn,Ga)_5O_{12}$$

we found a mobility three times as high: 0.63 m²/As (5000 cm/sOe). Magnetic dilution with Ga was found to favour low coercivities. We used a film of this type to make an experimental bubble circuit that works at a frequency of 4 MHz. This is not much compared with the potentialities of the film, but it is more than has ever been achieved before. The frequencies reported so far have never been higher than 1 MHz.

[23] D. J. Breed, A. M. J. van der Heijden, H. Logmans and A. B. Voermans, J. appl. Phys. 49, 939, 1978.

Summary. The highest frequency at which a magnetic bubble memory can work depends on the maximum velocity of the bubble. In the iron-garnet films generally used for making bubble memories the maximum velocity is about 20 m/s. This article discusses garnet films in which velocities of more than 500 m/s are possible because of an orthorhombic magnetic anisotropy of the film. For bubbles to exist the magnetization must have a preferred axis perpendicular to the plane of the film. The beneficial effect of an in-plane anisotropy as well is related to that of an in-plane field. Both increase the maximum angular velocity of the spin precession that is the basis of the 'laminar' movement of a Bloch wall. The movement becomes turbulent and the velocity decreases sharply when the drive field becomes too high. A strong magnetic in-plane anisotropy is obtained in {110} layers with the composition $(Lu,Eu)_3(Fe,Al)_5O_{12}$ and $(Gd,Lu)_3(Fe,Mn,Al)_5O_{12}$ on a substrate of $Gd_3Ga_5O_{12}$. In the plane of the film these layers are under a strong compressive stress, which is caused by the mismatch between the film and the substrate, and, because of the contribution made by Eu or Mn to the magnetostriction, gives the desired anisotropy. The Mn films have the highest value for the wall mobility (0.19 m²/As = 1500 cm/sOe).

Scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

J. P. André, A. Gallais & J. Hallais: GaAs-GaAlAs heterostructures grown by the metal-alkyl process.
Gallium arsenide and related compounds (Edinburgh), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33a), pp. 1-8; 1977. *L*

P. Ashburn, C. Bull*, K. H. Nicholas & G. R. Booker* (* University of Oxford): Effects of dislocations in silicon transistors with implanted bases.
Solid-State Electronics **20**, 731-740, 1977 (No. 9). *R*

G. A. A. Asselman, C. L. Spigt & R. J. Meijer: Design considerations on a thermal energy storage Stirling engine automobile.
Soc. Automotive Engrs. (SAE) Preprint 770080, 12 pp., 1977. *E*

W. J. Bartels & W. Nijman: Asymmetry of misfit dislocations in heteroepitaxial layers on (001) GaAs substrates.
J. Crystal Growth **37**, 204-214, 1977 (No. 3). *E*

F. Berz: A simplified theory of the *p-i-n* diode.
Solid-State Electronics **20**, 709-714, 1977 (No. 8). *R*

P. A. Blakey: Computer simulation of charge generation in trapatts.
Electronics Letters **13**, 434-435, 1977 (No. 15). *R*

P. Blood: Ion implantation of semiconductor devices.
Physics Bull. **28**, 416-418, 1977 (Sept.). *R*

M. C. Boissy, D. Diguet (both with RTC La Radiotechnique-Compelec, Caen), **J. Hallais & C. Schemali:** An efficient GaInAs photodiode for near-infrared detection.
Gallium arsenide and related compounds (Edinburgh), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33a), pp. 427-436; 1977. *L*

J. van der Borst, F. J. A. den Broeder & T. Scheffers (Philips Philite and Metalware Manufacturing Division, Eindhoven): The effect of metalloid composition on some magnetic and mechanical properties of $Co_{75}Si_{25-x}B_x$ glasses.
J. appl. Phys. **48**, 1724-1726, 1977 (No. 4). *E*

C. H. J. van den Brekel: Morphological stability analysis of growth from the vapour.
Philips J. Res. **33**, 20-30, 1978 (No. 1/2). *E*

J. C. Brice: The growth of insulating crystals.
Repts. Prog. Phys. **40**, 567-602, 1977 (No. 5). *R*

A. L. J. Burgmans: The Hanle effect for the $3\ ^2S_{1/2} \leftrightarrow 3\ ^2P_{1/2}$ transition of sodium.
Physics Letters **60A**, 453-455, 1977 (No. 5). *E*

K. H. J. Buschow, M. Brouha, J. W. M. Biesterbos & A. G. Dirks: Crystalline and amorphous rare-earth transition metal alloys.
Physica **91B**, 261-270, 1977. *E*

F. M. A. Carpay: Discontinuous grain growth and pore drag.
J. Amer. Ceramic Soc. **60**, 82-83, 1977 (No. 1/2). *E*

T. A. C. M. Claasen & W. F. G. Mecklenbräuker: On the transposition of linear time-varying discrete-time networks and its application to multirate digital systems.
Philips J. Res. **33**, 78-102, 1978 (No. 1/2). *E*

P. J. Courtois: Decomposability: queueing and computer system applications.
Book, XIII + 201 pp.; Academic Press, New York 1977. *B*

J. R. Dale & R. C. Oldfield: Mechanical stresses likely to be encountered in the manufacture and use of plastically encapsulated devices.
Microelectronics and Reliability **16**, 255-258, 1977 (No. 3). *R*

J. J. Daniele: Experiments showing the absence of electromigration of As and Al in Peltier LPE of GaAs and $Ga_{1-x}Al_xAs$.
J. Electrochem. Soc. **124**, 1143-1144, 1977 (No. 7). *N*

M. Davio & A. Thayse: Optimization of multivalued decision algorithms.
Philips J. Res. **33**, 31-65, 1978 (No. 1/2). *B*

M. Delfino, G. M. Loiacono, B. J. Fitzpatrick & W. A. Smith: Deuterium isotope effect in L(+)-glutamic acid hydrochloride.
J. solid State Chem. **20**, 391-395, 1977 (No. 4). *N*

J.-P. Deschamps & A. Thayse: The module structure of discrete functions.
Proc. 7th Int. Symp. on Multiple-valued logic, Charlotte, N.C., 1977, pp. 14-19. *B*

A. M. van Diepen & K. H. J. Buschow: Hydrogen absorption in $CeFe_2$ and $ThFe_3$.
Solid State Comm. **22**, 113-115, 1977 (No. 2). *E*

H. Durand: Cell and module test procedures seen from the manufacturer and the user point of view.
Proc. Workshop Terrestrial photovoltaic measurements-II, Baton Rouge 1976, pp. 301-307. *L*

H. Durand: Production, consommation et économies d'énergie: rôle joué par l'industrie et les produits électroniques.
Coll. Int. Electronique + 5, Paris 1977, pp. 65-70. *L*

W. Eckenbach: Regeneration of sensitizing dyes in ZnO binder layers.
J. photogr. Sci. **25**, 141-146, 1977 (No. 4). *A*

J. M. L. Engels & A. J. M. van Kimmenade: The motion of excess electrons in liquid mixtures of methane and argon.
Chem. Phys. Letters **48**, 451-454, 1977 (No. 3). *E*

E. Fabre, J. Michel & Y. Baudet: Photocurrent analysis in MIS silicon solar cells.
12th IEEE Photovoltaic Specialists Conf., Baton Rouge 1976, pp. 904-906. *L*

M. Goscianski: Optical characteristics of twisted nematic liquid crystals: application to the improvement of the scanning capability in matrix displays.
J. appl. Phys. **48**, 1426-1431, 1977 (No. 4). *L*

J. Hallais, A. Mircea-Roussel, J. P. Farges & G. Poiblaud (RTC La Radiotechnique-Compelec, Caen): Properties and thermal conversion of semi-insulating GaAs.
Gallium arsenide and related compounds (St Louis), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33b), pp. 220-227; 1977. *L*

P. Hansen & R. Krishnan (C.N.R.S., Bellevue, France): Anisotropy and magnetostriction of 4d and 5d transition-metal ions in garnets and spinel ferrites.
J. Physique **38**, C1/147-155, 1977 (Colloque C1). *H*

L. Heyne & D. den Engelsen: The speed of response of solid electrolyte galvanic cells for gas sensing.
J. Electrochem. Soc. **124**, 727-735, 1977 (No. 5). *E*

A. Huijser, J. van Laar & T. L. van Rooy: Electronic surface properties of uhv-cleaved III-V compounds.
Surface Sci. **62**, 472-486, 1977 (No. 2). *E*

L. Jacomme: Parallel beam analysis of a localized index gradient multimode fiber.
Optics Comm. **21**, 177-181, 1977 (No. 1). *L*

W. H. de Jeu & J. A. de Poorter: X-ray diffraction of the smectic phases of N-(p-n-heptyloxybenzylidene)-p'-n-pentylaniline.
Physics Letters **61A**, 114-116, 1977 (No. 2). *E*

G. D. Khoe & G. Kuyt: Cutting optical fibres with a hot wire.
Electronics Letters **13**, 147-148, 1977 (No. 5). *E*

U. Killat: Coupled wave theory of hologram gratings with arbitrary attenuation.
Optics Comm. **21**, 110-111, 1977 (No. 1). *H*

K. Kosai & B. M. Singer: Use of the WKB and variational methods to calculate the final stages of charge transfer in a CCD.
IEEE Trans. **ED-24**, 705-709, 1977 (No. 6). *N*

D. Küppers, J. Koenings & H. Wilson: Application of the plasma-activated chemical vapour deposition (PCVD) process to the preparation of fluorine doped fibres.
NTG-Fachber. **59**, 12-14, 1977. *A*

P. K. Larsen, S. Chiang* & N. V. Smith* (* Bell Laboratories, Murray Hill, N.J.): Determination of the valence band structure of InSe by angle-resolved photoemission using synchrotron radiation.
Phys. Rev. B **15**, 3200-3210, 1977 (No. 6). *E*

H. J. Leamy & A. G. Dirks: The microstructure of amorphous rare-earth/transition-metal thin films.
J. Physics D **10**, L 95-98, 1977 (No. 8). *E*

F. K. Lotgering & A. M. van Diepen: Electron exchange between Fe^{2+} and Fe^{3+} ions on octahedral sites in spinels studied by means of paramagnetic Mössbauer spectra and susceptibility measurements.
J. Phys. Chem. Solids **38**, 565-572, 1977 (No. 6). *E*

J. H. A. Melis: Design aspects and potential characteristics of a demand responsive bus system.
Coll. Int. Electronique + 5, Paris 1977, pp. 129-137. *E*

D. Meyer-Ebrecht: Imaging technology: a European survey.
IEEE Spectrum **14**, No. 5, 46-50, May 1977. *H*

J. Michel: Black and thin silicon solar cells.
12th IEEE Photovoltaic Specialists Conf., Baton Rouge 1976, pp. 603-605. *L*

A. Mitonneau, G. M. Martin & A. Mircea: Investigation of minority deep levels by a new optical method.
Gallium arsenide and related compounds (Edinburgh), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33a), pp. 73-83; 1977. *L*

J. M. van Nieuwland, N. Hazewindus & A. Baan (Philips Cyclotron Dept, Eindhoven): The central region of the SIN Injector Cyclotron.
Nucl. Instr. Meth. **142**, 339-351, 1977 (No. 3). *E*

L. P. A. S. van Noorden (Institute for Perception Research, Eindhoven): Minimum differences of level and frequency for perceptual fission of tone sequences ABAB.
J. Acoust. Soc. Amer. **61**, 1041-1045, 1977 (No. 4).

T. G. J. van Oirschot, W. J. Bartels & C. Langereis: Lattice matching and crystal perfection in $Ga_yIn_{1-y}As_xP_{1-x}$-InP heterostructures.
Gallium arsenide and related compounds (St Louis), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33b), pp. 105-109; 1977. *E*

C. van Opdorp, A. T. Vink & C. Werkhoven: Minority carrier recombination at surfaces, dislocations and microdefects: evaluation of parameters from near band edge luminescence.
Gallium arsenide and related compounds (St Louis), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33b), pp. 317-330; 1977. *E*

K. Pasedach: Einsatz blockzirkulanter Matrizen zur Rekonstruktion einer Funktion aus ihren Linienintegralen.
Z. angew. Math. Mech. **57**, T 296-297, 1977 (No. 5). *H*

J. G. J. Peelen: Alumina: sintering and optical properties.
Thesis, Eindhoven 1977. *E*

W. J. J. Rey: M-estimators in robust regression: a case study.
Recent developments in statistics, Proc. Eur. meeting of statisticians, Grenoble 1976, pp. 591-594; 1977. *B*

U. Rothgordt, H.-D. Hinz & K. Witter: Triboelectric charging of thin foils and its application to a new recording principle.
IEEE Trans. **IA-13**, 223-226, 1977 (No. 3). *H*

G. Rouweler & B. J. de Maagt (Philips Lighting Division, Eindhoven): The reaction of bromine and oxygen with a tungsten surface, Part I. Measurement of the rate of reaction in a gas-flow experiment.
Philips J. Res. **33**, 1-19, 1978 (No. 1/2).

C. Schiller: Lattice deformation of epitaxial pseudobinary III-V compounds determined by x-ray diffraction.
Gallium arsenide and related compounds (Edinburgh), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33a), pp. 25-34; 1977. *L*

M. F. H. Schuurmans: Velocity tuned two- and three-photon processes in a three-level system.
Physics Letters **61A**, 445-448, 1977 (No. 7). *E*

J. M. Shannon: Thermionic-field emission through silicon Schottky barriers at room temperature.
Solid-State Electronics **20**, 869-872, 1977 (No. 10). *R*

F. J. van Steenwijk (Kamerlingh Onnes Laboratorium, Leiden) **& K. H. J. Buschow:** Isomer shifts in divalent europium intermetallic compounds.
Physica **85B**, 327-330, 1977 (No. 2). *E*

M. Tasto: Reconstruction of random objects from noisy projections.
Computer Graph. & Image Proc. **6**, 103-122, 1977 (No. 2). *H*

D. R. Terrell & H. Katiofsky: Melt viscosity characteristics of methacrylate-styrene copolymers.
J. appl. Polymer Sci. **21**, 1311-1322, 1977 (No. 5). *H*

D. L. A. Tjaden: First-order correction to 'weak-guidance' approximation in fibre optics theory.
Philips J. Res. **33**, 103-112, 1978 (No. 1/2). *E*

C. van Trigt: Worst-case analysis of algorithms.
Philips J. Res. **33**, 66-77, 1978 (No. 1/2). *E*

R. Veilex: High speed GaAs circuits.
Solid state circuits, 1976 (ESSCIRC Toulouse), pp. 117-141; 1977. *L*

A. G. H. Verhulst, T. W. Bril & L. Postma: A highly sensitive magnetic bubble detector in an essentially single-mask overlay.
Appl. Phys. Letters **30**, 548-550, 1977 (No. 10). *E*

J. C. Verplanke: Accurate determination of shallow Zn-diffusion profiles in GaP and $GaAs_{0.1}P_{0.9}$ using anodic oxidation and a double radioactive tracer technique.
J. Electrochem. Soc. **124**, 469-471, 1977 (No. 3). *E*

J. C. Verplanke & R. P. Tijburg: The analysis of anodic oxide films on GaAs and GaP by means of radioactive tracer techniques.
J. Electrochem. Soc. **124**, 802-804, 1977 (No. 5). *E*

J. O. Voorman: The gyrator as a monolithic circuit in electronic systems.
Thesis, Nijmegen 1977. *E*

Q. H. F. Vrehen & H. M. J. Hikspoors: Two-photon induced stimulated Raman scattering in cesium vapor.
Optics Comm. **21**, 127-131, 1977 (No. 1). *E*

Q. H. F. Vrehen, H. M. J. Hikspoors & H. M. Gibbs (Bell Laboratories, Murray Hill, N.J.): Quantum beats in superfluorescence in atomic cesium.
Phys. Rev. Letters **38**, 764-767, 1977 (No. 14). *E*

J. A. Weaver: Experiments in visually controlled flexible automation.
Radio and electronic Engr. **47**, 385-392, 1977 (No. 8/9). *R*

K. Weiss: Roton second sound in the relaxation-time approximation.
Phys. Rev. B **15**, 4227-4233, 1977 (No. 9). *E*

C. J. Wellekens & A. N. Godard: Simultaneous flat approximations of the ideal low-pass attenuation and delay for recursive digital, distributed, and lumped filters.
IEEE Trans. **CAS-24**, 221-230, 1977 (No. 5). *B*

C. Werkhoven, C. van Opdorp & A. T. Vink: Non-radiative recombination in n-type LPE GaP.
Gallium arsenide and related compounds (Edinburgh), 1976 (Proc. 6th Int. Symp.; Inst. Phys. Conf. Ser. No. 33a), pp. 317-325; 1977. *E*

J. P. Woerdman & M. F. H. Schuurmans: Doppler-free three-photon cross-resonances in Na_2 vapor.
Optics Comm. **21**, 243-246, 1977 (No. 2). *E*

J. Wolter & R. E. Horstman: Quantitative measurement of the phonon energy transmission coefficient for a silicon-HeII boundary.
Physics Letters **61A**, 238-240, 1977 (No. 4). *E*

Contents of Philips Telecommunication Review **36**, No. 1, 1978:

W. F. Hawes: Special issue on mobile radio (pp. 1-2).
F. Grimm: Future development of mobile radio (pp. 2-4).
S. W. Watkinson: A dual phase lock loop receiver (pp. 5-12).
M. Arrowsmith: A new FM portable equipment: the P 5000 series (pp. 13-18).
R. J. Howes: M206 mobile radiotelephone (pp. 19-24).
R. E. Grove: MASCOT 1000 multi-access systems control terminal (pp. 25-30).
J. O. Voorman & A. Wijker: PAGYR: a new component for selective calling (pp. 31-40).
A. M. Morriën & W. Oude Vrielink: Digital paging encoders/decoders (pp. 41-52).
D. Muilwijk: Spectrum efficiency in mobile radio communication (pp. 53-60).
J. Chastagnier: Ground-to-train communication (pp. 61-68).
J. Mezan de Malartic & D. Leclerc: VHF radiotelephone with LSI synthesizer (pp. 69-76).
H. J. Wermter: MATS B2, a new Mobile Automatic Telephone System (pp. 77-84).
F. P. van Enk & G. Devery: P-TMC mobile/fixed equipment (pp. 85-91).
J. Kuchel: Rural radio telephones with solar power supply (pp. 92-96).
I. Swinson: UHF Citizens Band radio equipment FM-320 (pp. 97-99).
R. Baird: The SI-10 000 Status and Identification System (pp. 100-103).

Contents of Philips Telecommunication Review **36**, No. 2, 1978:

A. Rijbroek & H. W. de Haan: PCM system 8TR 623 featuring per-channel encoding (pp. 109-115).
P. W. L. van Iterson & J. C. Leguyt: A domestic satellite communication system (pp. 116-132).
P. W. L. van Iterson: Encoding and modulating in Domestic Satellite networks (pp. 133-138).
J. S. Visser: Twoports with reactive short-circuit and open-circuit impedances (pp. 139-146).

Contents of Electronic Applications Bulletin **35**, No. 2, 1978:

G. van Aller: A comparison of first and second-generation image intensifiers (pp. 66-75).
D. Hermans & M. D. Hull: Designing hi-fi speaker systems — part 2 (pp. 76-101).
R. Lauer: Accelerometers using piezoelectric ceramic sensors (pp. 105-116).

Contents of Electronic Applications Bulletin **35**, No. 3, 1978:

T. S. Christian: SI units in permanent magnet calculations (pp. 118-135).
Liquid crystal displays — part 1 (pp. 136-145).
J. Jongsma: Minimum-loss transformer windings for ultrasonic frequencies. Part 1: Background and theory (pp. 146-163).

Contents of Mullard Technical Communications **14**, No. 137, 1978:

G. van Schaik: An SMPS without mains isolation for tv receivers (pp. 246-264).
A. G. van Nie: Low-level properties of microwave crystal detectors (pp. 265-278).
L. E. Jansson: Mains-operated switched-mode power supply for an oscilloscope (pp. 279-288).

Contents of Mullard Technical Communications **14**, No. 138, 1978:

P. J. Jackson: IEC/IEEE data transmission bus interfaces (pp. 290-298).
P. J. Jackson: Power strobing of bipolar read-only memories (pp. 299-305).
20AX Hi-Bri ctv picture tubes (pp. 306-308).
S. R. Longley: Design and application information for broadband solid-state noise sources (pp. 309-316).

Volume 38, 1978/79, No. 7/8 pages 181-228 Published 12th October 1979

Supplement to Philips Technical Review
Beilage der Philips Technischen Rundschau
Bijlage van Philips technisch Tijdschrift

PHILIPS

August 1979

Recent United States Patents

Abstracts from patents that describe inventions from the following research laboratories that form part of or cooperate with the Philips group of companies:

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

4 059 781

Shadow mask each aperture of which is defined by a quadrupolar lens

W. M. van Alphen *E*
J. Verweel

A post-focusing type color cathode ray tube comprising a quadrupole lens formed in each aperture of the color selection mask.

---

4 059 840

Television camera and pick-up tube having stripes for improved resolution and linearity

L. J. van de Polder *E*
S. L. Tan

A black-white or color television camera has a pick-up tube having a signal electrode comprising separate strips. Before a line scanning operation to be performed by an electron beam takes place, the signal electrode strips are coupled to a reference potential. After the line scan, the strips which are coupled separately to parallel inputs of a switching circuit which is provided with switches coupled to the parallel inputs, are coupled for supplying video information. This results in a separation between electron beam scanning and picture signal supply by the pick-up tube, which has the advantage that the beam scanning does not affect the picture quality in the direction of line scan. Furthermore, the construction with strips yields an improved signal-to-noise ratio. The switching circuit may, possibly together with a shift register, be integrated in a semiconductor body, which is provided near the signal electrode in the pick-up tube or outside the pick-tube.

---

4 060 248

Playing apparatus, in particular a video player for playing a record which is made of a transparent material and provided with a central hole

H. G. Lakerveld *E*
G. E. van Rosmalen
K. Bulthuis
J. P. Sinjou

A playing apparatus, more specifically a video player for playing round records with a central hole, which are made of a transparent material, which records at one side have a transparent surface and at their other side are provided with information tracks which are optically readable with the aid of a read beam, which tracks are covered by a light-reflecting layer. In order to enable both relatively thick rigid records and relatively thin flexible records to be played on one playing apparatus, the invention provides a transparent adapter disc, which together with a flexible record can be positioned on the playing apparatus and driven in a rotary fashion. In order to prevent dust particles from being trapped between the flexible record and the transparent adapter disc, the adapter disc may be provided with a central recess with a number of air admission apertures at the circumference thereof, so that during operation a thin film of air is formed between the flexible record and the adapter disc.

---

4 061 832

Positive-working electron resists

E. D. Roberts *R*

A positive-working electron resist which comprises a film-forming co-polymer B-D which is cross-linkable by carboxylic acid anhydride groups. The co-polymer B-D is formed from an unsaturated organic compound B and an unsaturated carboxylic acid chloride D. The cross-links are formed by heating a film of the resist in a moist atmosphere, some of the carboxylic acid chloride groups being hydrolysed so as to form carboxylic acid groups, at least some of these carboxylic acid groups reacting with carboxylic acid groups of different molecules of the co-polymer B-D so as to form carboxylic acid anhydride bridge links between these different molecules. The cross-linked material is insoluble in solvents in which the monomers B and D and the copolymer B-D are normally soluble. Electron irradiation of the cross-linked material breaks the cross-links and restores the readily soluble straight chain forms, so that during development of an irradiated resist pattern, a solvent is used only to dissolve soluble material in the areas which have been irradiated.

---

4 061 920

X-ray installation comprising an image intensifier/image pick-up tube system and an automatic X-ray exposure device

R. Möllendorf *H*
P. Lux

The invention relates to an X-ray installation which comprises a television display device and an automatic exposure device and in which dominant selection is no longer required. The television image produced during fluoroscopy is sub-divided into a large number of sub-fields, a storage position being assigned to each



sub-field. In each storage position the part of the video-signal is stored which corresponds to the assigned sub-field. The sub-fields are then determined which contain the regions of the maximum and the minimum brightness. An effective, organ-programmed intermediate value is derived therefrom. Moreover, the mean value of the video signal is obtained in the part of the television image assigned to the sole central measuring field of the automatic exposure device. The mean value is compared with the intermediate value, and the switch-off dose or switch-off voltage is corrected in dependence of the difference thus found.

---

4 062 014

Natural binary to logarithmic binary conversion device with means for range selection

U. Rothgordt *H*
B. Ehlers
E. Bunge
H. Piotrowski

In practice it is often desirable to represent only a given, selectable dynamic range of a measuring signal, preferably in logarithmic form. This is achieved in accordance with the invention in that from the measured value in the form of a binary number only the positions corresponding to the desired dynamic range are selected for forming the base number, the weight of the most significant bits having the value "1" being converted into a binary number. The mantissa is derived from the following positions, i.e. independent of the selected dynamic range, so that even when the most significant bit is in the last position, a mantissa comprising the full number of positions is formed. The magnitude of the dynamic range is chosen by omitting the correspondingly more significant positions of the base number formed for the maximum range. The position of the dynamic range can be shifted in that the inputs of the circuits for forming the base number and the mantissa are shifted with respect to the lines supplying the individual positions of the measured value.

---

4 062 107

Method of manufacturing infra-red detector

M. V. Blackman *R*
M. D. Jenner

A method of manufacturing an infra-red detector in which a printed form of lead-out contact pattern is applied to the surface of each infra-red sensitive element. A detector element of infra-red sensitive material is provided having at one major side at least one active surface area defined between spaced contact layers which extend over oppositely located curved edges of the element at said side. The element is adhered via the opposite major side to an insulating substrate having a contact pattern provided thereon. Electrically conductive material is deposited to forming interconnections between the contact layers on the element and adjacently situated end portions of lead-out conductors of the contact pattern.

---

4 063 074

Device for measuring radiation absorption or radiation emission distributions in a plane through a body

W. Wagner *H*

For measuring the absorption from measuring values whereby the radiation attenuation in a plane of a body is measured, a large number of calculations must be performed. When use is made of a digital computer, a large quantity of storage locations and a long calculating time are required. According to the invention, use is made of a charge storage tube wherein the correction values calculated in a preprocessing unit on the basis of the measuring values are superimposed, so that a substantial reduction of the calculation time and a reduction of the number of storage locations are realized.

4 063 287

Tracking mirror device for a video disc player

G. E. van Rosmalen *E*

A tracking mirror device, in particular for a video disc player, in which for following an information track on a video disc with a beam of radiation, use is made of a mirror which oscillates at a high frequency, which causes the beam of radiation, which serves for scanning the information track, to oscillate transversely to the information track at a high frequency and a small amplitude, while the filtered out high-frequency signal, as a function of its amplitude and its phase relationship with the high frequency voltage which is applied to the oscillating mirror, yields a control signal for following the information track. In accordance with the invention, use is made of an oscillating mirror which is driven electromagnetically and which is connected to oscillation compensation elements by means of springs which elements oscillate in phase opposition to the mirror. Thus, an oscillator mirror is obtained which can readily be excited to natural resonance and thus requires little power. On the mirror a piezo-electric miniature acceleration transducer can be mounted, which can provide a feedback signal so that the oscillating mirror can be included in a self-oscillating circuit.

---

4 063 914

Method of producing an optical fiber

E. Roeder *A*
E. Steinbeck

A method of and a device for a simultaneous extrusion of multi-component co-axial rods with very uniform, centered cores. A block of cladding material has core material co-axially contained in the geometric relationship desired for the ultimate product, and extrusion is by pressure opposite to the direction of material flow.

---

4 064 393

Device for measuring radiation absorption in a section of a body

K. Pasedach *H*
W. Wagner

A device for examining a body section by means of X-radiation including a plurality of detectors arranged to intercept radiation passing through the body in straight lines or strips and a calculating circuit which interpolates, on the basis of absorption values measured, absorption values which can be expected along sets of mutually extending strips which cross each other at different angles. The absorption in the section of the body can then be calculated in known manner on the basis of these values.

---

4 064 505

Signal transmission circuit comprising a coder and a decoder

N. V. Franssen *E*

A coder and a decoder having delay elements of the same delay time arranged one after the other in which the signal transmission ratio code of the signal paths from the input of the circuit to the output is chosen in accordance with Bessel coefficients so as to prevent disturbing echoes.

---

4 065 674

Rotary-anode X-ray tube

H. G. Lakerveld *E*
J. A. Rietdijk

The stator of the drive motor of a rotary-anode X-ray tube is composed of a plurality of radially displaceable sector elements. As a result, the anode can be accelerated and decelerated while using a narrow airgap, and the airgap can be widened for bridging the high voltage to be applied at a later stage. The invention can be realized in automated form by using a blocking device for the high voltage and a tachometer.

4 065 735

Electrical filters including coupled resonators

J. S. Palfreeman R
M. Redwood
F. W. Smith
R. F. Mitchell

A surface acoustic wave band-pass filter comprising a piezoelectric substrate having on one surface first and second SAW resonators tuned to the same frequency and coupled together. An input transducer is coupled to the first resonator and an output transducer is coupled to the second resonator. Each resonator is preferably formed by two spaced apart aligned grating reflectors and the two resonators are preferably coupled by an array of discrete parallel metallic strips extending between the two cavity resonators.

---

4 065 808

Network computer system

H. Schomberg H
F. Heubach

A network computer system in which the individual processors are arranged in the form of a matrix, with each processor connected with its directly neighboring processors. The individual processors are checker-board-like assigned to one of two groups, and each group is connected, by its own instruction line, to a control unit for connection to the control computer. The system is useful in performing iterative calculations in which the processors of the one group calculate new values on the basis of initial data previously calculated by the directly adjacent respective processors of the other group.

---

4 066 923

Color selection lens electrodes connected by diffusion bonds

J. van Esdonk E

In a color display tube of the post-focusing type the color selection means comprise a first and a second system of lens electrodes. A lens electrode belonging to the first system is connected by means of a diffusion bond to one of the metallized sides of an electrically insulating member which is metallized on two sides. A lens electrode belonging to the second system is formed either by the other metallized side of the insulating member or is secured to said other metallized side of the insulating member also by means of a diffusion bond. The color selection means are manufactured by starting from an insulating foil which is metallized on two sides and which is connected on the one side by a diffusion connection to each lens electrode belonging to the first system and which is provided on the other side with each electrode belonging to the second system, the desired apertures in the color selection means being obtained by locally etching away at least the metallized foil by means of a selective etching process.

---

4 070 585

Starter motor lockout system

D. J. Parkyn R

An electronic lockout system for the starter motor of an internal combustion engine includes means for operating the starter motor only if the engine flywheel is stationary. The system also includes means for automatically de-energizing the starter motor when the engine achieves a predetermined rotational speed and independently of the continued operation of the starter switch.

---

4 071 763

Electroradiographic device

K. Peschmann A

Xonics chambers can normally be used only with comparatively low electrode voltages in the Townsend-plateau region of the discharge where no charge carrier multiplication occurs. The sensitivity is then comparatively low. The sensitivity can be increased by increasing the electrode voltage, but this gives rise to a high electrode voltage for a reasonable charge carrier multiplication. According to the invention a gas is added whose ionization energy is smaller than the energy of the lowest metastable levels of the rare gasin the chamber. The charge carrier multiplication, and hence the sensitivity is thus improved without the electrode voltage being substantially increased. For example, when 0.2% by volume trimethylamine is added, an electrode voltage of 13 kV already suffices.

---

4 071 765

Electron microscope

K. J. van Oostrum E
G. G. P. van Gorkom

In an electron beam apparatus corrections are made for vibrations and field disturbances which cause a lateral beam shift. A detector is arranged in an intermediate image near the beam path and a signal derived therefrom is used for controlling a beam deflection device. When a second beam deflection device is added, the undesired deflection in a scanning electron microscope can be distinguished from the deflection caused by the scanning.

Re. 29 530

Magneto-optical information storage device using photoconductive control element

J.-P. Krumme H
B. Hill

The invention relates to a device for magneto-optic memories controlled by light and/or heat in an external magnetic field, in which the magneto-optic material of the memory has a photoconductive layer which can be activated by the control beam and which can be controlled by means of electrodes provided thereon via a current or voltage source.

---

4 072 915

Acoustic surface wave devices

R. F. Mitchell R

An electrode structure within the cavity of an acoustic surface wave resonator, e.g. an interdigital transducer or a multistrip coupler, has two sections separated by an odd integral multiple of a quarter wavelength of the standing wave in the cavity and located with respect to the center of the cavity so that all the electrodes are midway between a node and an adjacent antinode of the standing wave. Distortion of the resonator response due to reflection by the electrodes of the transducer or coupler is suppressed by this arrangement.

---

4 073 432

Circuit arrangement for determining the average value of a frequency

G. Schröder H

A known arrangement for continuously averaging a frequency comprises a counter whose count-up input receives a measuring signal and whose count-down input receives the delayed measuring signal, by which the averaging period is determined. However, spurious signals produce a cumulative indication error. In order to avoid this, an additional counter is included which each time counts the measuring signal from zero onwards and after the predetermined time interval transfers its count to the measuring counter and starts again at zero. As a result, the measuring counter is always corrected after the predetermined time interval so that errors cannot accumulate. The predetermined time interval can be obtained by means of a further counter. For delaying the measuring signal, a RAM can be used which is controlled by the further counter.

4 073 567

Pivoting mirror device

H. G. Lakerveld E
G. E. van Rosmalen

A pivoting mirror device for pivoting a pivoting mirror about two mutually perpendicular pivoting axes under the influence of electric control signals, in particular for a video player which reads information from a video disk with the aid of a light beam using optical means, the mirror being journalled so as to be pivotable to all sides and the device comprising electro-dynamic drive means which include permanent magnets and control coils for controlling the pivoting movements. The mirror is supported near its center, while permanent magnetic positioning means which co-operate with permanent magnetic means which are connected to the mirror continually subject the bearing to pressure. In order to obtain a high efficiency the permanent magnetic drive means are magnetized in such a way that the magnetic field lines, at least at the location of the control coils, extend substantially parallel to the reflecting surface of the mirror, while moreover at least a part of the turns of the control coils extend parallel to the reflecting surface and transversely to said field lines.

---

4 073 574

Optical projector

J. A. Clarke R
H. Howden

An aspheric element, placed near a principle plane of an wide-angle, wide-aperture image forming system and having in its aspheric function a term proportional to the third power of the distance from the optic axis improves off-axis revolution.

---

4 074 149

Peak detecting with constant fractional offset

G. J. Naaijer L

A peak detector for electrical signal peaks including a capacitor which is charged, via a rectifying element, to a voltage corresponding to the peak value. At the end of the peak a discharging element connected parallel to the capacitor is activated for a predetermined period of time, with the result that the capacitor is discharged over a fixed fraction of its peak voltage. At the end of the predetermined period of time, the discharging element is reactivated, with the result that a substantially constant voltage remains as a reference level.

---

4 074 172

Discharge lamp operating circuit

G. A. Wesselink E
H. Roelofs
G. J. Ponsen

The invention relates to a device for operating a low-pressure sodium vapor discharge lamp without a stabilizing ballast. The lamp is provided with a discharge tube having first and second main electrodes opposing a third main electrode. An automatically operating change-over switch occasionally switches the lamp current to either the first or second electrodes. A reservoir of liquid sodium near the first electrode affects sodium vapor pressure in the discharge tube by the switching action.

---

4 074 199

Vestigial-sideband transmission system for synchronous data signals

F. de Jager E
J. J. Martony

A vestigial-sideband transmission system for synchronous data signals. The transmitter is provided with a filtering and modulating circuit for generating a vestigial-sideband channel signal which at the location of its carrier frequency and at the location of a frequency which is spaced from this carrier frequency by a distance equal to one half of the clock frequency of the data signals is a double-sideband modulated signal. In the receiver the reference carrier for coherent demodulation and the reference clock signal for regeneration can be simply recovered from the transmitted vestigial-sideband channel signal itself.

---

4 074 261

Keyboard contact and signal line arrangement

H. van der Gaag E

The invention relates to a keyboard comprising at least three contact points per key. One of a group of signal connections is always alternately connected to a group of keys. Each key is then alternately connected to one signal connection of said group. At least three of such groups are present so as to obtain a comparatively low value for the ratio of the number of signal connections to the number of keys.

---

4 074 302

Bulk channel charge coupled semiconductor devices

R. J. Brewer R

A semiconductor device includes a semiconductor body having a semiconductor layer of one conductivity type, means for enabling the semiconductor layer to be locally fully depleted without avalanche multiplication in order to allow the movement of majority charge carriers representative of information via the interior of the layer to one or more storage sites in the layer, means for enabling the reading of information present at least in one storage site comprising an insulated gate field effect transistor structure having its channel region present at the surface of the semiconductor layer above said one storage site and between source and drain regions of the opposite conductivity type, and means for enabling the temporary confinement of the quantity of majority charge carriers in said one storage site to the interior of the layer without mixing of the stored charge carriers in said one storage site and the mobile minority charge carriers in the transistor channel in order that the current flow in the transistor channel when the transistor gate electrode is maintained at any given potential is dependent upon the quantity of charge in the said one storage site. Various semiconductor devices, particularly of the type referred to as bulk channel charge coupled devices, are disclosed, including a charge coupled imaging device, a device having a relatively high charge handling capacity, a photo-detector, a tapped delay line, a random access memory and an image display device.

---

4 074 312

Automatic dithering optical focusing system using an oscillating mirror in the path of the radiation beam

G. E. van Rosmalen E

An automatic optical focusing system suitable for an apparatus for reading an information track on a record carrier with the aid of a read spot which is focused in an imaginary focusing plane and which is formed by a radiation beam which is produced by a radiation source, more in particular suitable for a video disc player which employs optical reading, and serving for automatically keeping the read spot focused at the plane in which the information is disposed (information plane) during playing of a record carrier, comprising an optical focusing device, an oscillating optical element which cooperates with the radiation beam and which can effect high-frequency oscillations so that a high frequency oscillation of small amplitude about the coarse position of the focusing plane is obtained, an electromagnetic drive system for the high-frequency drive of the oscillating optical element, a radiation-sensitive detector which responds to the radiation beam which has been modulated by the information track, and a control circuit for producing a control signal which depends on the output signal of the detector, which control signal is applied to the control coil of the optical focusing device. The oscillating optical element is constituted by an oscillating mirror which reflects the radiation beam which mirror effects high-frequency bending oscillations of which at least a component has the direction of the reflected part of the radiation beam. Preferably, the oscillating mirror is round and effects high frequency bending oscillations between a convex and a concave shape.

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79 No. 9

PHLIQA 1, a question-answering system for data-base consultation in natural English

PHLIQA Project Group

I. Organization and performance

II. The artificial languages and translation operations

The common aim of all research in the broad field of artificial intelligence is to raise the level of 'intelligence' at which machines can perform their tasks. The article below deals with a somewhat distinct sector of this field, in which efforts are made to enable computers to 'understand' a natural language. This language is usually English. If communication between man and machine were simpler, more people would be able to benefit from the large quantities of information stored in computer memories. Until now users have had no alternative but to learn a programming language. It now begins to look as if the computer can be made to adapt to man. The PHLIQA 1 (Philips Question-Answering) system is a first step in this direction; it enables a computer to understand and to answer questions formulated in natural English. Part I of the article describes the organization of the system and describes its capabilities. Part II will deal with the syntax and semantics of the artificial languages used, i.e. their grammar and the rules of meaning, and discusses the English grammar used and the translation operations [*].

I. Organization and performance

From card index to data base

There is a growing tendency to store large files of data in computer memories. Whatever advantages such 'data bases' may have over the traditional card-index system, they have one major drawback in common: their inaccessibility to people who have not learned to program. Even for programmers, it is only specialists in the management of data bases who are able to use them with ease, and then only if they know the organization of the data base. There is a large group of potential users who only want to interrogate a data base occasionally, and it would be too much to expect them to learn a special programming language. In the present situation it is in fact more difficult for this group to obtain the information they want than it used to be in the days when the information was contained in a card index.

The feasibility of using a natural language, usually English, as a means of communicating with computer

The PHLIQA Project Group consists of W. J. H. J. Bronnenberg, Ir S. P. J. Landsbergen, Ir R. J. H. Scha, Ir W. J. Schoenmakers, and Drs E. P. C. van Utteren of Philips Research Laboratories, Eindhoven. Until 1976 the project group also included Drs H. C. Bunt, now with the Institute for Perception Research (IPO), Eindhoven, and Ir P. Medema, now with the Philips Information Systems and Automation Department (ISA), Eindhoven.

[*] Part II will appear in the next issue.

systems has been the subject of investigations since the early sixties [1]. In our laboratories and elsewhere promising results have already been achieved, especially for applications where the subject of the dialogue is restricted, which is the case for data bases [2]. Programs have been made that enable a computer to answer a simply formulated question relating to data in a data base and expressed in ordinary, if somewhat limited, English. Our program, called PHLIQA 1, is the subject of the present article.

Of course, such a program has an underlying formalism, a function described by mathematical rules, which, given a question and a data base, defines the answer to that question. Part of this function, for example, is a 'formal' grammar of the English language.

The program itself is the actual algorithmic description of the function in a particular programming language. Part of this is a parsing program, which allocates to a sentence a structure that follows from the rules of the formal grammar.

Finally, there is of course the computer, which, during the execution of the program, behaves to all intents and purposes as a machine that 'understands' English — in the case of PHLIQA 1 it is a Philips P1400 computer.

This entire system will be referred to as the PHLIQA 1 question-answering system.

PHLIQA 1 was developed in conjunction with its own data base, specially built for the purpose, but one of the requirements was that it should be organized in such a way as to provide confidence in the possibility of equipping any other data base with a question-answering system. For this reason the parts of PHLIQA 1 that depend on the chosen data base and those that do not are kept carefully separated, and an effort has been made to derive the parts dependent on a data base in a systematic way from the structure of the data base and its subject domain.

The data base designed by way of example for PHLIQA 1 contains fictitious data. It is also small enough to fit completely into the fast memory of the computer, thus avoiding a number of problems of data-base management. The organization of the data base, however, does comply with the relevant standards [3]. No attempt has been made to organize the data base so as to make it suitable in advance for a question-answering system. On the contrary, some features that cause difficulties in such a system, and which are also found in 'real' data bases, have been deliberately introduced. These arise in particular from the fact that the designers of a data base envisage an organization in which the storage and retrieval of data is performed as efficiently as possible. This often leads to an 'unnatural' representation of the subject domain in the data base.

In the next section of this first part we shall deal with the general outline of PHLIQA 1. We shall discuss the types of question handled and their answers, and touch briefly on the important concept of "representation of meaning". We shall then deal with the answering operation as a series of successive translation operations, followed by a value calculation made with the aid of the data base. In this context attention is paid to the data base chosen as an example.

In part II the syntax and semantics of the artificial languages used will be outlined. The English grammar used and the principal translation mechanisms are elucidated by simple examples. Finally the operation of PHLIQA 1 is illustrated by an example that shows how the whole succession of translation operations yields an answer to a particular question.

PHLIQA 1 in outline

For the user the PHLIQA 1 system is a machine capable of answering questions posed in English about a specific subject domain. The (fictitious) information about that subject domain — in our case computers installed at European establishments — is stored in a computer memory. It is not necessary to know how the data base is organized to use the system.

The PHLIQA 1 program, as stated above, is a translation program. In outline what happens is the following. The user types in his question on a keyboard with a visual display (*fig. 1*) on which both the typed-in question and the reactions from the computer appear. The system analyses the question and translates it, via a number of intermediate steps, into an internal logic language. The resultant expression is rather like an algebraic expression. The system calculates the value of the expression with the aid of the data from the data base. The result is the answer to the question.

PHLIQA 1 is an experimental precursor of a system that is eventually intended to offer greater scope. It has allowed attention to be concentrated on the vital part of the problem, i.e. the development of formalisms such as the grammar mentioned above and programs

[1] A more general picture of computer linguistics, including the work on question-answering systems, can be found in: A. Zampolli (ed.), Linguistic structures processing, North-Holland, Amsterdam 1977.

[2] For recent work on communication in natural languages with computer systems, see F. J. Damerau, Automated language processing, in: M. E. Williams (ed.), A. Rev. Inform. Sci. Technol. **11**, 107-161, 1976, and S. R. Petrick, On natural language based computer systems, IBM J. Res. Develop. **20**, 314-325, 1976.

[3] The CODASYL standards, which are given in a report published by the Data Base Task Group of the Conference on Data Systems Languages, ACM, New York, April 1971.

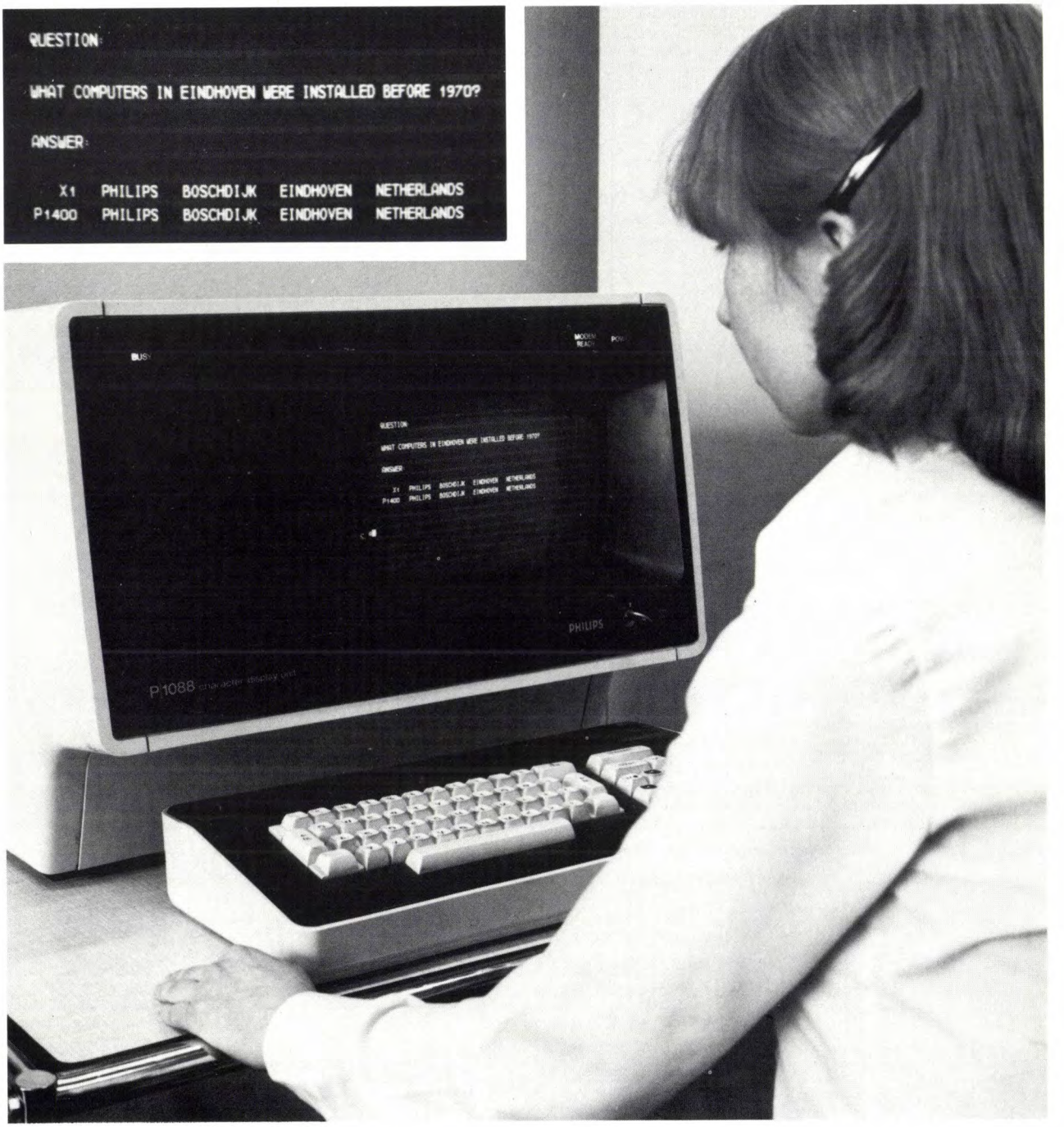


Fig. 1. The user of the PHLIQA 1 question-answering system simply has to type in his question in natural English. The answer 'computed' by the system appears on the display screen, which continues to show the question.

based on them that would be capable of providing the answer to a question in a systematic and reliable way. The emphasis here is on the correctness of the answers, and for the time being not so much on the convenience of use of the system. The user can ask a question and receives a fairly short answer to it or a brief indication of the reason why the question cannot be answered. He can then ask another question. If a question has more than one meaning, all the answers that relate to the various meanings can be obtained; the present system does not however indicate what those meanings are.

Table I gives some idea of the capabilities of the present system. It lists a number of questions with the answers given by PHLIQA 1 after interrogating the data base.

The English grammar used is subject to certain

▷

Table I. Ten examples of questions, in natural English, which can be put to the PHLIQA 1 question-answering system. The questions and the answers given appear on a display screen (fig. 1), which is connected to a Philips P1400 computer. The answers are brief, e.g. "YES" or "NO", a number, or a list with names and addresses. As can be seen from the answers, the data base contains fictitious data. In its present stage the system cannot synthesize complete sentences for answers. It only processes questions that consist of one sentence, and it takes no account of context. A 'follow-up question', e.g. "And how many in Germany?" after question 1 cannot therefore be answered. Question 6 is an ambiguous question. The two possible answers are given here.

1	QUESTION: HOW MANY COMPUTERS ARE THERE IN THE NETHERLANDS? ANSWER: 9
2	QUESTION: WHAT IS THE NUMBER OF IBM COMPUTERS IN GERMANY? ANSWER: 1
3	QUESTION: WHAT COMPUTERS IN EINDHOVEN WERE INSTALLED BEFORE 1970? ANSWER: X1 PHILIPS BOSCHDIJK EINDHOVEN NETHERLANDS P1400 PHILIPS BOSCHDIJK EINDHOVEN NETHERLANDS
4	QUESTION: IN WHAT MONTH IN 1972 DID SHELL BUY A COMPUTER FROM IBM? ANSWER: JANUARY 1972
5	QUESTION: DOES EACH COMPUTER IN EINDHOVEN HAVE A CPU MADE BY PHILIPS? ANSWER: NO
6	QUESTION: WHAT COMPANIES HAVE A CONFIGURATION WITH A CPU THAT COSTS MORE THAN 100 000 DOLLARS? FIRST ANSWER: AKZO SECOND ANSWER: AKZO SHELL SIEMENS UNILEVER
7	QUESTION: WHAT IS THE PRICE OF THE MOST EXPENSIVE CONFIGURATION OF UNILEVER? ANSWER: 538 000 DOLLARS
8	QUESTION: HOW MANY BITS OF CORE MEMORY DOES AKZO'S ARNHEM COMPUTER HAVE? ANSWER: 131 072 BITS
9	QUESTION: WHICH OF THE COMPANIES THAT POSSESS MORE THAN 2 CONFIGURATIONS BOUGHT A CPU BEFORE MAY '75? ANSWER: AKZO PHILIPS SHELL SIEMENS UNILEVER
10	QUESTION: HOW EXPENSIVE IS EACH OF AKZO'S CONFIGURATIONS? ANSWER: PDP15 AKZO YSSELLAAN ARNHEM NETHERLANDS 191 000 DOLLARS SI2002 AKZO GOLDSCHMIDT-STRASSE STUTTGART GERMANY 196 000 DOLLARS GAMMA30 AKZO AV. MONTAIGNE LYONS FRANCE 973 000 DOLLARS ICL4120 AKZO TEMPLE PLACE LONDON UNITED KINGDOM 483 000 DOLLARS

limitations. Furthermore, although PHLIQA 1 is capable of analysing English sentences, it cannot perform a synthesis. Nor can the user refer back (for example with pronouns) to a previous question or a previous answer. A real dialogue with the system is therefore not possible.

It will not be easy to remove these limitations in future versions, but even so the associated problems may be regarded as secondary to the problem of answering single questions correctly.

The questions and their answers

PHLIQA 1 can handle questions of two kinds: Yes/No questions and 'WH' questions (the abbreviation WH stands for words such as "what", "which", "when", "how", etc.). An example of a Yes/No question is "Did Shell buy a computer in 1970?" Examples of WH questions are "What computer did Shell buy in 1970?" and "How many computers did Shell buy in 1970?". Both kinds of question relate to a description of an object; an unambiguous identification of that object, or at least as unambiguous as possible, can be regarded as an answer to the particular question. In the case of Yes/No questions the term "object" refers to a truth value (TRUE or FALSE). In the example given of a Yes/No question this is the truth value of "Shell bought a computer in 1970". The unambiguous description of this in the answer can be "YES" or "NO". In WH questions the identified object can for example be a set of computers ("the computers that Shell bought in 1970"); the answer then consists of an enumeration of those computers, characterized for example by the model of their CPU (central processing unit) and the address of the establishment where they are installed.

The representation of meaning

In determining the answer to a question, PHLIQA 1 performs two main tasks. The first is to produce an explicit representation of the meaning of the question ('understanding' the question), and the second is to use this representation to answer the question with the aid of the information in the data base. An explicit rep-

resentation of meaning is an expression in an artificial language. For every expression in such an artificial language the rule is that as soon as values are assigned to the 'primitive terms' in that expression, the value of the entire expression is established. The assignment of values to the primitive terms of an artificial language is referred to as an *interpretation* of that language.

every assignment of a value to the primitive terms of the expression leads to a value for the entire expression that is a correct answer to the question, for the state of the world as reflected by the value assignment (*fig. 2*).

In the choice of an artificial language for the representation of meaning it is important to determine which primitive terms this artificial language should contain.

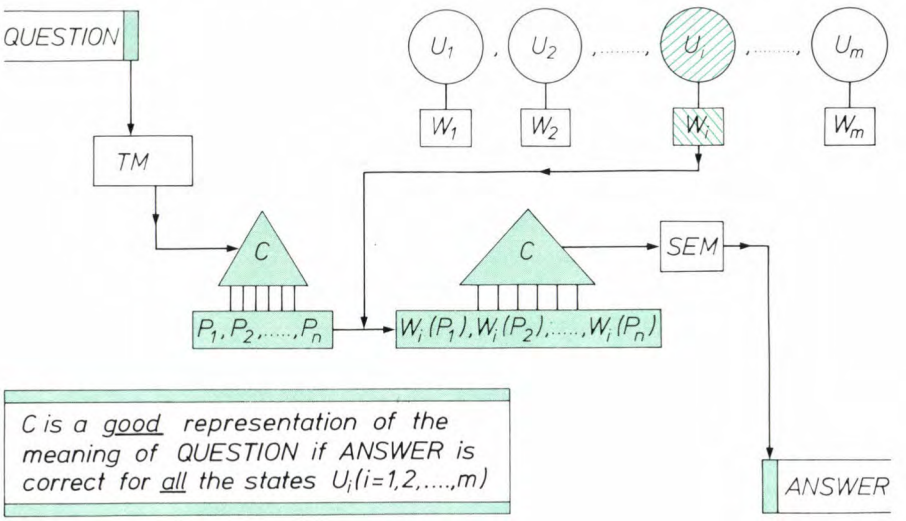


Fig. 2. Representation of the meaning of an interrogative sentence. The translation mechanism *TM* transforms the sentence QUESTION into the expression *C* in an artificial language. This expression contains the primitive terms $P_1, P_2, \ldots, P_n$. The possible states of the world are $U_1, U_2, \ldots, U_m$. Associated with the state U_i is the value assignment W_i, which assigns the values $W_i(P_1), W_i(P_2), \ldots, W_i(P_n)$ to the primitive terms $P_1, P_2, \ldots, P_n$. For each value assignment W_i the value of *C* is established by the semantic rules *SEM* of the artificial language, giving ANSWER.

An expression in an artificial language may to some extent be likened to an algebraic expression, such as $a \times (b + c)$, where *a*, *b* and *c* are the primitive terms. The language of algebra has rules for multiplication and addition that establish the value of such an expression; in fact they do this for every set of values of *a*, *b* and *c*, that is to say for every interpretation of the language. The rules for the calculation constitute the semantics of the language. In the artificial languages of PHLIQA 1 the primitive terms not only stand for numbers but also for entities such as countries or computers. In addition to the ordinary arithmetical rules the semantics of these languages also comprises logic operations and operations with sets.

Each interpretation of an artificial language reflects a possible 'state of the world'. Suppose, for example, that the language contains the primitive term COUNTRIES, which is considered to stand for all the countries of the world, then the assignment of a value to this term that is in accordance with the *present* state of the world will include the country Surinam. If, however, reference is to be made to the state of the world in 1970, then Surinam does not form part of the value of COUNTRIES. For PHLIQA 1, the 'world' is confined to the subject domain of the data base.

An expression in an artificial language is a correct representation of the meaning of a given question if

The three artificial languages

A special feature of the PHLIQA 1 system is that it uses *three* internal artificial languages, with the same syntax and the same semantics but different primitive terms. In response to a given question a representation of its meaning is produced in each language, and the answer is then calculated (*fig. 3*). The three languages are EFL (*E*nglish-oriented *F*ormal *L*anguage), WML (*W*orld-*M*odel *L*anguage) and DBL (*D*ata-*B*ase *L*anguage).

During the translation operation English→EFL the system takes the *logic* aspects into account. The *referential* semantic aspects, that is to say the aspects relating to the world, are handled by the next two translation operations, EFL→WML and WML→DBL. What it amounts to is that the two translation operations establish the relation between the English words in the question and the primitive terms in the data base.

The semantic analysis of a question is thus made in steps. This modular approach is of advantage for the reliability of the system, and during the design of PHLIQA 1 it was therefore possible to divide the problem of making the system into a number of much sim-

pler problems, which could be solved independently of each other. Furthermore, the sequence of the translation operations allows PHLIQA 1 to respond more effectively to unanswerable questions than systems that have only one translation operation (English → Data-Base Language).

The primitive terms of the three artificial languages were chosen as follows:

— The primitive terms of EFL correspond to the referential aspects of English words. For example EFL contains the term COMPUTERS, corresponding to the words "computer" and "computers", and also the term BUY, corresponding to the verb "to buy" in all its conjugated forms. The primitive terms of EFL, like words in natural English, can be ambiguous.

— The primitive terms of WML correspond to the 'elementary' concepts of the subject domain. They are not ambiguous. Although the subject domain is largely determined by the data base, the terms of WML will not in general coincide with the primitive terms of the data base. This is because the data-base primitive terms were chosen more for the efficient storage and retrieval of data, whereas the WML terms are designed to 'model' the subject domain in a more natural manner than in the data base. Cities, years and CPUs are examples of objects that belong to the subject domain of PHLIQA 1 and therefore have corresponding WML terms (CITIES, YEARS, CPUS), although there are no corresponding primitives in the data base.

— The primitive terms of DBL correspond to the data-base primitives. For DBL the values of the primitive terms are stored in the data base. Since the semantic rules (the rules of meaning) are known for DBL, just as they are for the other languages, the value of each DBL expression can be calculated from the values of the terms.

During the translation operation English→EFL the data is processed that comes from the grammatical structure of the sentence, from the 'function words' such as "the", "each", "and", "than", and from the 'formal aspects' of the other words, such as the distinction between the singular and plural of nouns. In this phase the referential semantic aspects of the English words are *not* analysed; the semantic analysis brought about by the translation operation English→EFL is therefore independent of the subject domain.

During the translation operation EFL→WML the terms of EFL are translated into WML expressions, and in this way a relation is established between English words and the primitive concepts of the subject domain. The translation operation EFL→WML is of course dependent on the subject domain, but is independent of the storage structure of the data base. If the translation operation EFL→WML is successful, this implies that the question makes sense in the context of the subject domain.

During the translation operation WML→DBL the storage structure of the data base is taken into account. Some examples of the problems associated will be given after we have discussed the organization of the data base.

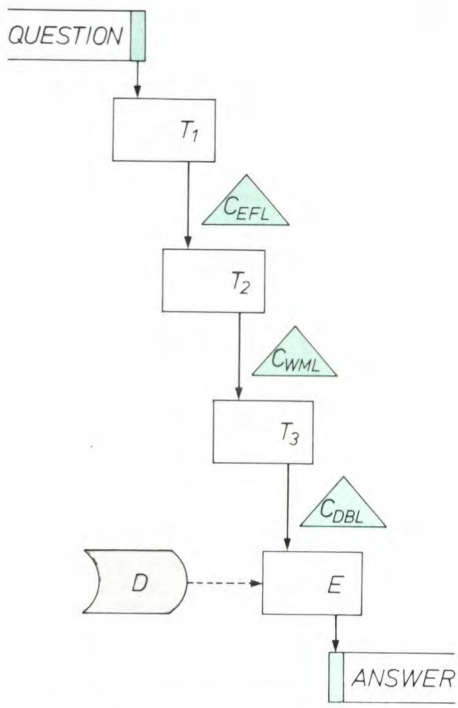


Fig. 3. The PHLIQA 1 question-answering system. The simplified breakdown into modules shows the translators T_1, T_2 and T_3, the evaluator E and the data base D. The user formulates the question in natural English; this is then translated into the artificial languages EFL (English-oriented Formal Language), WML (World-Model Language) and DBL (Data-Base Language); E then makes use of D to formulate ANSWER. C_X is the translation of QUESTION into the artificial language X.

Unanswerable questions

To understand the organization of PHLIQA 1 it is also useful to consider the various types of question that do not lead to an answer. One of the advantages of the three-stage analysis in PHLIQA 1 is that the system is able to indicate why it cannot answer such questions.

— There are questions that contain words that are not to be found in the lexicon used for the translation operation English→EFL. In this case the system identifies these words; the user can then type in other words in their place or rephrase the whole question.

— There are questions that cannot be translated into EFL because the system regards them as ungrammatical. Two examples of such questions are "The of computers are what?" and "How many computers is there in Holland?" The system then answers that it cannot parse the sentence.

— There are questions that can be translated into EFL but not into WML. Such questions are syntactically correct; they may make sense in a particular subject domain, but are meaningless in the subject domain of PHLIQA 1. An example is the question "What is the price of Germany?" The concepts of "price" and "country" are both known at the level of WML, but the WML function "price" is not applicable to the elements of the WML set "countries", which means that countries do not have a price in the subject domain of PHLIQA 1. In such cases the system responds with an indication of why the translation operation into WML has failed.
— There are questions that reach the level of WML, and may thus make sense in the limited subject domain but cannot be translated into DBL because DBL does not contain the appropriate constants. What this really means is that the factual information is not contained in the data base. An example is the question "When were AKZO's card readers installed?" The data base contains the installation dates of the complete computer configurations, that is to say the CPU with the peripherals connected to it at that time. The installation dates of peripherals connected later are not stored in the data base, and therefore it is not known when the individual card readers were installed. Although the system admits such questions, it cannot answer them because it does not have the facts. In such cases the response indicates why the translation operation into DBL has failed.
— Finally there is a class of questions that do reach the DBL level but whose value cannot be computed because the question asked makes incorrect presuppositions about the actual state of the world. An example is the question "How expensive is the computer owned by Shell?" This implies the presupposition that Shell possesses just one computer. If this presupposition is incorrect, the system responds in a manner that indicates what the incorrect presupposition is. (Since the program treats each question separately, as an isolated question, it takes no account of the possibility that the part of the sentence "the computer owned by Shell" may be a reference to a computer mentioned in a previous question.)

These cases are shown in the diagram of *fig. 4*.

Ambiguous questions

Besides questions for which no translation is produced, there are also questions for which more than one translation is possible in one of the three translation operations. In that case we are confronted with *ambiguous* questions, that is to say questions with two or more meanings. Ambiguities of two kinds may be distinguished:

— Syntactic ambiguities. These occur in the translation operation English→EFL when a question can be parsed in more than one way. An example is the question "What companies have a configuration with a CPU that costs more than 100 000 dollars?" from Table I. Here the subordinate clause "that costs more than 100 000 dollars" may relate to "a CPU" or to "a configuration with a CPU".
— Semantic ambiguities. These occur during the translation operation EFL→WML when an EFL term (which corresponds to an English word) can be translated into WML in more than one way. Thus, for the verb "to have" a choice has to be made between the interpretation "contain as a part" (as in "Does AKZO's computer have two card readers?") and the interpretation "to possess" (as in "Does AKZO have two card readers?"). An ambiguity occurring during a translation operation does not always lead to more than one possible answer to the question, since the result of a translation operation may turn out 'later on' in the system to be untranslatable.

When an ambiguity arises, the system starts to examine one of the possibilities and it continues to do so until this possibility either comes to a dead end or leads to an answer. After the answer has been given, the user is asked if he wants a search to be made for other meanings. In this case, and also in the case where the possibility examined comes to a dead end somewhere along the line, the system so to speak retraces its steps to the point where the ambiguity first arose. Once that point has been found, the system works on a new possibility, and the procedure is thus repeated until the questioner indicates that he is no longer interested in other meanings or until the system discovers that no other meanings exist. This procedure is illustrated schematically in *fig. 5* for the example "What company has a computer with two card readers made by Siemens?" This question is syntactically ambiguous, because the subordinate clause "made by Siemens" could relate to "a computer" (the meaning *1*) or to "two card readers" (the meaning *2*). The verb "has" contains the semantic ambiguity mentioned above. Other ambiguities that are possible will be disregarded here for the sake of clarity. The system works through the meanings in the sequence of the numbers (fig. 5). The meaning *1.1* is discarded because a company cannot have a computer as a component. The system then tries *1.2* and finds that it makes sense. During the value calculation, however, *1.2* is found to contain a false assumption, which is that exactly one company has a computer made by Siemens with two card readers. PHLIQA 1 then returns to the last branch point, in the program that performs the translation operation English→EFL, and proceeds further with meaning *2*.

It discards *2.1* for the same reason as it did *1.1*. Meaning *2.2* leads to an answer, because there is in the data base exactly one company that has a computer equipped with two card readers made by Siemens.

In the case of ambiguity the system treats the meaning with which it gets farthest as the most likely. Only the response corresponding to this result is communicated to the questioner. In the example in fig. 5, PHLIQA 1 gives the answer corresponding to meaning *2.2* and does not mention the other paths pursued that led to a dead end.

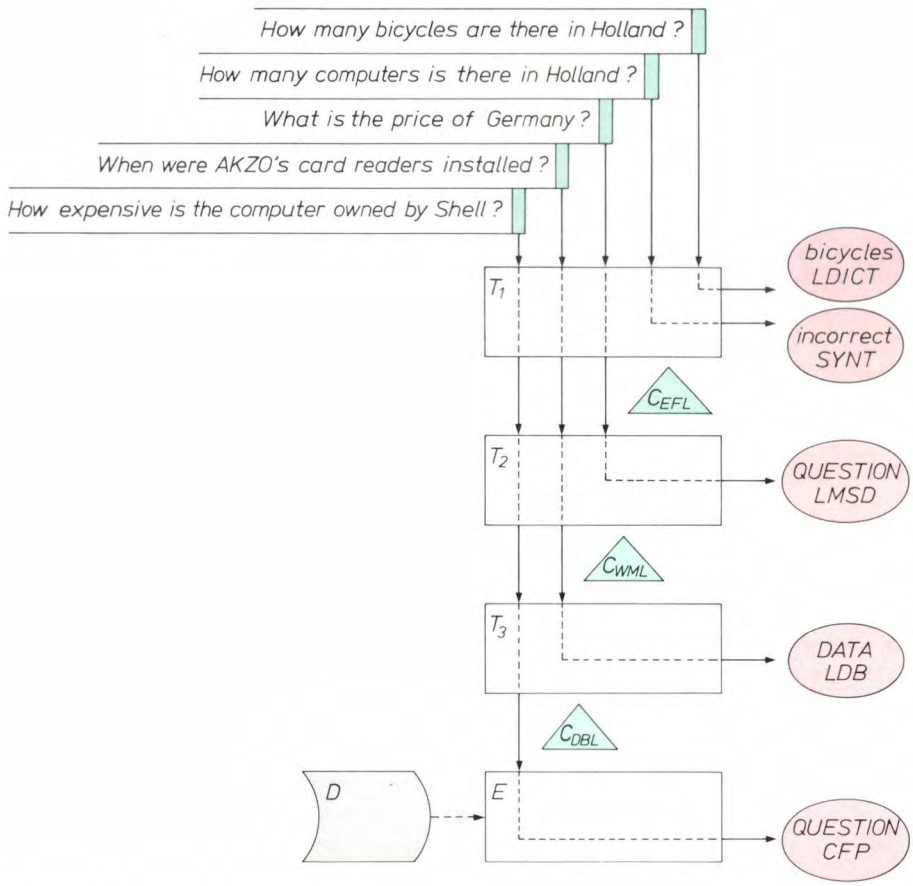


Fig. 4. The diagram illustrates the step-by-step nature of the semantic analysis to which PHLIQA 1 subjects a question. As examples, five different *un*answerable questions are given; this is a useful way of showing in what steps of the semantic analysis the system is able to indicate the unanswerability of the question. *LDICT* means not in lexicon. *SYNT* syntax. *LMSD* no meaning in subject *domain*. *CFP* contains false presupposition. *LDB* not in data base. The significance of the other symbols is given in fig. 3.

As noted above, in the development of PHLIQA 1 the emphasis was placed on the analysis and not the synthesis of English sentences. Full benefit will not be derived from the potentialities of the system until it also becomes possible to synthesize sentences, for only then will the system be able to indicate, in a manner that is also clear to the uninitiated, why a question has not been answered or what meaning the system has chosen.

The organization of the data base

The organization of the data base chosen as an example for PHLIQA 1 is shown schematically in *fig. 6*.

The subject domain, computer installations in Europe and their users, includes objects of various kinds: concrete ones, such as CPUs, and abstract ones, such as CPU models. In dealing with the artificial languages WML and DBL it was stated that not all objects of the subject domain are necessarily explicitly represented in the data base. Where they do need to be explicitly represented, it is done by means of 'records', comparable with the cards of a card-index system. For each object there is a separate record. Objects of the same kind, e.g. countries, are represented by records of the same type (such records are comparable with the cards in the same drawer). In fig. 6 the boxes represent record types of this kind, the name being given in capital letters, e.g. "COUNTRY", "CONFIGURATION", etc. Each record stores data on the objects it represents. All records of the same type have the same internal

structure, that is to say they contain the same kind of data. This data has the form of 'attribute values'. Thus, for records of the type "CONFIGURATION", the value of the attribute "Date-Installed" is mentioned. This attribute value gives in coded form the month and the year when the computer configuration was installed for the present user (e.g. March 1969, coded as 6903). Fig. 6 gives the attributes for each record type.

lows that the single-headed and the double-headed arrow must always be present in combination.

In this way information is stored in the data base about computers in use with European companies. Records of the type "CONFIGURATION" represent the computers. In addition there are record types for the location ("SITE") where the computer is installed and about the company ("CORPORATION") to which the es-

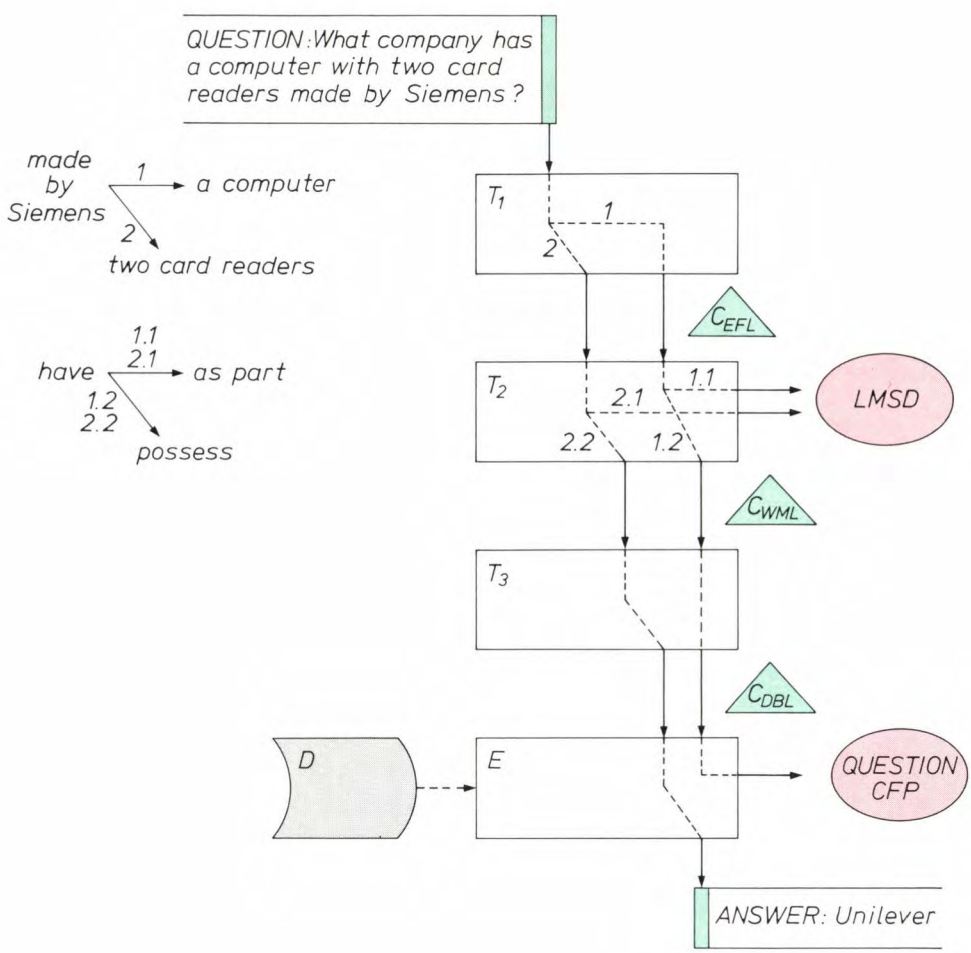


Fig. 5. An example of the processing of a question that contains both a grammatical ambiguity and an ambiguous term. The verb form "has" means both "to have as a part" and "to possess". The system works through the various possible meanings in the sequence of the numbers indicated. *LMSD* no meaning in the subject domain. *CFP* contains a false presupposition.

There is a second way of storing information about the objects represented by the records, and this is by making use of relations with other objects. These relations are indicated in fig. 6 by arrows. A single-headed arrow from record type "A" to record type "B" means that a functional dependence exists between the two types, such that each record of type "A" has one record of type "B" associated with it. The double-headed arrow from record type "B" to record type "A" means that each record of type "B" has a set of records of type "A" associated with it. From the CODASYL standards that have generally been adopted [3] it follows that the single-headed and the double-headed arrow must always be present in combination.

tablishment belongs. For the model of the CPU of the computer there is the record type "CPU MODEL". There is also data about the peripherals ("PERIPHERAL-GROUP", etc.).

The primitive terms of DBL can be directly derived from the organization scheme of the data base. They are the names of the record types, the attributes, and the functional relations indicated by the arrows.

The users of a conventional system with a data base must know the exact meaning of the various record types and attributes, more exactly than is possible from names alone. The users of PHLIQA 1, however, do *not*

need to know the data-base terms and their meanings. The PHLIQA 1 system establishes for them the relation between English words and the data-base terms. Just how complicated that relation can be is evident if we compare the data-base terminology with the forefore have to count the records of this type that represent those computers in the Netherlands. Records of the type "CONFIGURATION" do not, however, contain direct information about the country where the computer is situated. PHLIQA 1 has to deduce that this is

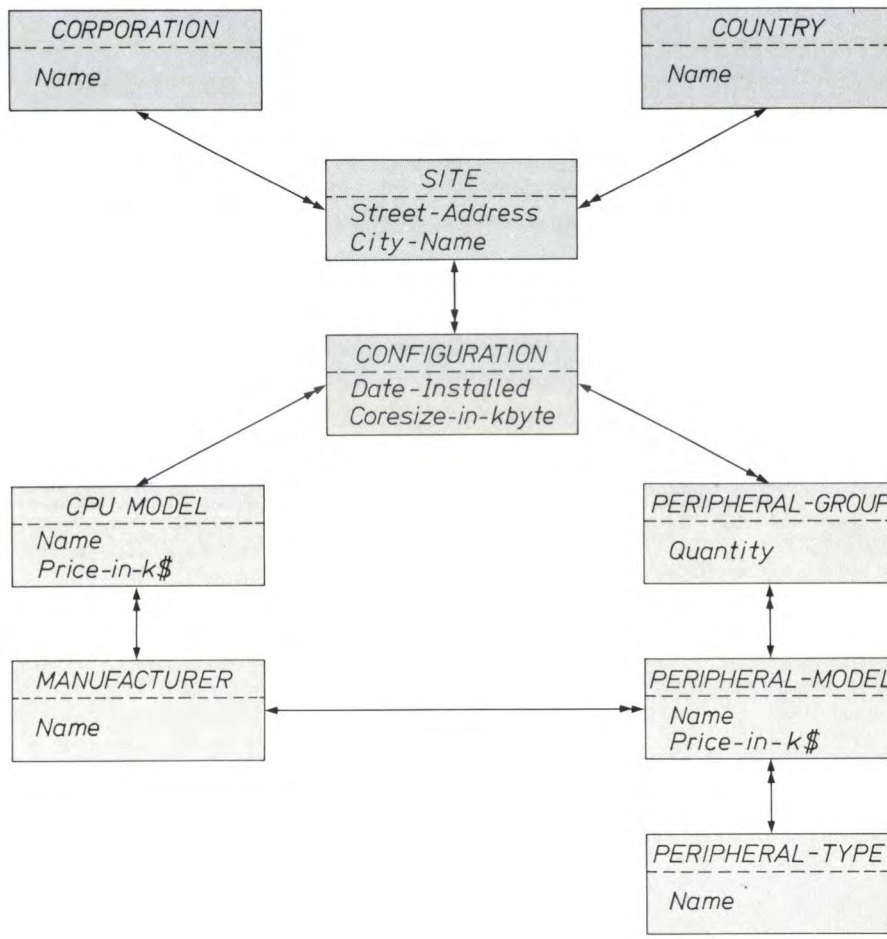


Fig. 6. Organization of the data base (*D* in fig. 3) that was chosen as an example for the development of the PHLIQA 1 question-answering system. The subject domain relates to European computer installations and their users. The boxes in the diagram represent 'record types'; the upper part contains their name, in capitals. Below the name the 'attributes' are mentioned. For each object the data base contains a separate record with the attribute values applicable to that object. The relations existing between the various objects are represented in the diagram by the single-headed and double-headed arrows. The data base is organized in accordance with the CODASYL standards [3].

mulations of the questions given by way of example in Table I. The question "How many computers are there in the Netherlands?" is one of the simplest questions one can imagine. Even so, this question presents problems when we try to answer it from the data in the data base. In the first place the number to which the question relates is not explicitly present in the data base. We assume that records of the type "CONFIGURATION" correspond to computers; the system will therethe country of the establishment where the computer is being used.

Other difficulties are encountered with words like "month", "CPU" and "companies" in the example questions. Neither months nor CPUs are represented in the data base, but the codes mentioned above for months and year together are in fact present, and there are also records for CPU models. "Companies" can be represented in the data base in two different

ways: computer users as "CORPORATION", computer manufacturers as "MANUFACTURER". A company that is both is represented twice.

The distinction discussed above between the three levels of the representation of meaning is the first step towards a systematic treatment of problems of this type. In the second part of this article we shall deal in greater depth with the methods developed for this.

Summary. The experimental computer program PHLIQA 1 enables a computer of medium capacity (e.g. the Philips P1400) to answer certain questions put to it in natural English. The answers are taken from a data base, organized in accordance with the CODASYL standards, which contains fictitious data about European computer installations and their users. The question, typed in by means of a keyboard with a visual display, undergoes a series of translation operations. The first, English→EFL (English-oriented Formal Language), determines the logical relations between the various parts of the sentence. The second, EFL → WML (World-Model Language), reformulates the question in terms of the subject domain and checks whether the question makes sense in that domain. The third translation operation, WML→DBL (Data-Base Language), provides the link between the contents of the question and the chosen data base. The computer calculates the value of the expression in DBL and produces an answer on the display screen ("YES"; "NO"; a number; a list of names and addresses). If ambiguity exists the computer can systematically trace different meanings; if the questions are unanswerable the computer indicates at what language level the answering procedure breaks down and why. The organization of the data base and the syntax of the artificial languages used are described. All the artificial languages have the same grammar; they differ, however, in their primitive terms. The language expressions are referred to as constructions; the extent to which they are 'well-formed' is determined by means of type rules. The syntactic structure of the constructions corresponds to their semantic structure. For translating into EFL the computer uses a lexicon, reduction rules and semantic composition rules. To translate from EFL into WML and DBL the computer uses referential and formal transformations. The various parsing and translating mechanisms are explained with a number of examples. An example illustrates how the whole sequence of translation operations provides, in the form of a three-stage analysis followed by a value calculation, an answer to a particular question. The article is in two parts. Part I gives a general introduction to the organization and external behaviour of the system, and part II deals more with technical particulars, the syntax and the semantics of the artificial languages and the mechanisms of the translation operations.

Two microchannel-plate photomultipliers with subnanosecond characteristics

J. P. Boutot and J. C. Delmotte

This article describes two new photomultiplier tubes, types PM 137 and PM HR 305, that have recently been developed at the Laboratoires d'Electronique et de Physique Appliquée for time measurements on phenomena of ultra-short duration. The chief applications of the tubes are in plasma physics and high-energy nuclear physics.

The main difference between the characteristics of the two tubes stems from a trade-off between gain and pulse response. The PM HR 305 tube has the better pulse response; in the best case the full width at half maximum height (FWHM) of an output pulse can be as small as about 0.20 ns. The FWHM value of the other tube is about five times greater, but the pulse-amplitude gain of the tube is more than a thousand times higher (*Table I*). A pulse response with an FWHM value of about 1 ns, as in the PM 137, is still reasonably good, however. The high signal gain is associated with an exceptionally low spread in the average time that electrons need to cover the distance from photocathode to anode in the tube: for signals originating from a single photoelectron the standard deviation of this time is less than 0.10 ns. This statistical spread, or time dispersion, is usually referred to as the time resolution of a tube. It is an important characteristic, because it determines the precision with which the time of occurrence of a 'primary' event can be found, for example a light flash in a crystal connected to the photomultiplier tube as a detector. Another and obviously different case is finding the intensity distribution of a light flash; here a response with a small FWHM is almost a necessity. (The output pulse of a photomultiplier tube only corresponds exactly with the shape of the input signal in the theoretical case in which the FWHM is zero.)

The high gain and the good time resolution make the PM 137 particularly suitable for situations where very short time intervals have to be measured under conditions of low light level, as in fluorescence spectroscopy and in high-energy nuclear physics. An example of such a time measurement is the determination of the ultra-short time that elapses between the creation of two nuclear particles in a cascade decay, which often occurs in high-energy experiments. The unknown delay can be found from the time difference between the instants the particles hit a scintillator connected to the photomultiplier — provided, of course, that the difference in their times of flight is taken into account.

The applications of the first tube, the PM HR 305, usually relate to analysis of the shape of an individual light-emission profile, which can be performed in real time. Typical applications are found in laser fusion experiments, where the photomultiplier is a useful instrument not only for radiation in the visible part of the spectrum but also for X-rays (detected with the aid of a plastic scintillator).

Both new tubes employ microchannel plates for the electron-multiplication process. It is well known that replacing the conventional multiplier section by a microchannel plate is a good means of obtaining a photomultiplier design with better characteristics for time measurements [1]. Besides changing the characteristics, the incorporation of the microchannel plate also affects the shape of the tube considerably, mainly because of the small thickness (0.4 mm to a few mm) of the microchannel plate as compared with a conventional dynode structure, which may have a length of several centimetres (*fig. 1*).

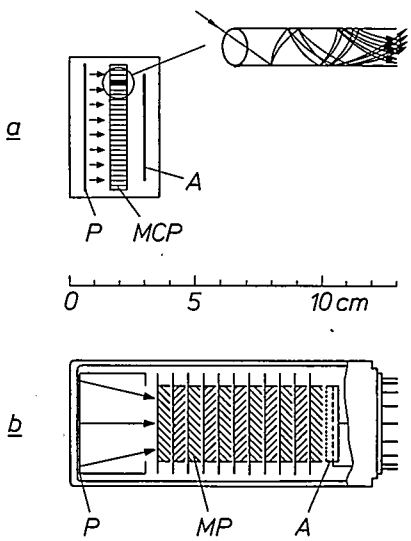


Fig. 1. Longitudinal cross-section of a microchannel-plate photomultiplier (*a*) and a conventional photomultiplier (*b*). *P* photocathode. *A* anode. *MCP* microchannel plate. The multiplier section *MP* in the conventional tube consists of dynodes with the Venetian-blind geometry. The two tubes differ considerably in length.

Dr J. P. Boutot and J. C. Delmotte are with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

Table I. Constructional data and performance figures for the photomultiplier tubes PM 137 and PM HR 305. SER single-electron response. G gain. σ standard deviation.

Feature	PM 137	PM HR 305
	Constructional data	
Dimensions*		
diameter (mm)	55	50
length (mm)	50	80
Terminals		
multiplier section	pins on base	rings
anode	pins on base	coax. connector (tapered)
Photocathode		
geometry	end window (flat)	end window (flat)
aperture (steradians)	2π	$< 2\pi$
useful diameter (mm)	25	13
spectral type	S**	S**
Multiplier section		
microchannel plate	two, in cascade	one
thickness (mm)	3.2 and 0.5	0.4
channels	curved straight	straight
diameter (μm)	40 12.5	7.3
number	2.5×10^5 2.5×10^6	2×10^6
	Performance	
Gain		
average	5×10^6	10^3
maximum	10^7	5×10^3
variation ($=\sigma/\bar{G})^2$, the relative variance of the max. gain	$\leqslant 5\times10^{-2}$	about 1
Time measurements		
transit time (ns)	2.5 - 3	< 1
pulse response FWHM (ns), at max. gain	about 1.3	0.20 - 0.25
time resolution, standard deviation in average transit time (ns) for SER	< 0.1	not measurable
Linearity limits (with full cathode illumination)		
d.c. mode (μA)	1	0.2
pulse mode		
charge/pulse (C)	about 3×10^{-9}	about 10^{-9}
peak current (A) (FWHM = 1 ns)	3	1
Magnetic-field effect		
gain variation $\Delta\bar{G}/\bar{G}$ (%) for transverse magnetic field	$\geqslant 50$	$\geqslant 50$***
magnetic induction (tesla)	> 0.03	> 0.1***

* Overall, unpotted.
** Photosensitive layer can be any type S.
*** Estimated.

The main contribution to the improvement comes from the shortening of the signal transit time in the tubes. The signal is a pulse-like electron avalanche in which the individual electrons that form the same signal have slightly different transit times. The transmission of the avalanche through the tubes is governed by rather complex probability laws. The time distribution of the electrons that form the output signal resembles to some extent a Gaussian distribution, the familiar bell-shaped curve. A shorter transit time is desirable because the *spread* in the transit time is also smaller; this improves both time resolution and pulse response. As can be seen from Table I, the signal transit times in the new photomultipliers are about 2.5 ns in the

[1] J. P. Boutot and G. Piétri, IEEE Trans. **ED-17**, 493, 1970. See also J. P. Boutot, Acta Electronica **15**, 271, 1972.

PM 137 and less than 1 ns in the PM HR 305. In conventional photomultiplier tubes the transit time is of the order of 30 ns.

To compare the new photomultiplier tubes with conventional ones correctly, it is sufficient to look at the time spread caused by the multiplier section only. This is done by considering the 'single-electron' response (SER), which relates to the arrival-time distribution of the electrons at the output end that arises from a single electron at the input end of the multiplier section. As a result of most primary events the photocathode of a photomultiplier will emit several electrons instead of one, which then travel to the input of the multiplier and hence introduce an extra time spread in the output signal. An SER is rather rare; good examples are the background pulses (dark-current pulses) that arise as thermal noise from the photocathode. *Fig. 2* shows an SER as measured for the PM 137. This response was obtained at average gain, with an oscilloscope of bandwidth 0.5 GHz. This bandwidth, although large, still introduces some broadening of the photomultiplier output pulse. Taking this extra broadening into account, we found that the FWHM of the photomultiplier pulse itself in this case amounted to about 1.3 ns. *Fig. 3* shows a pulse response measured for the PM HR 305; this response was obtained at maximum gain, with an ultra-fast oscilloscope (bandwidth 4 GHz). The FWHM found was about 0.25 ns. The photomultiplier pulse itself had an estimated width of 0.20 to 0.23 ns. In this case the original light flash came from a laser; the wavelength was 538 nm and the FWHM value less than 0.05 ns [2]. The pulse broadening from 0.05 to 0.20 ns is mainly caused by the electron-multiplication processes in the microchannel plate.

The design of the tubes is explained in *fig. 4*. Both tubes have an input stage, a microchannel-plate section, and the anode (or collector) stage. In the high-gain tube the microchannel-plate section is composed of two microchannel plates in cascade. The first plate

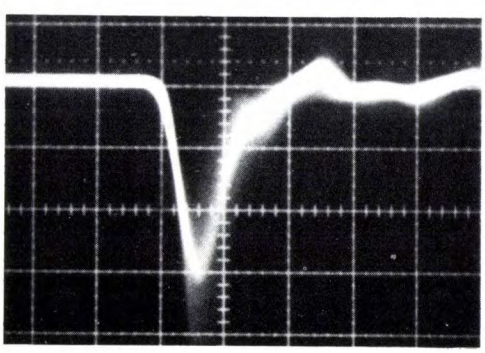


Fig. 2. Oscillogram of the single-electron pulse response (SER) of the PM 137 photomultiplier tube, used at average gain (5×10^6). The oscilloscope had a bandwidth of 500 MHz. If the broadening effects were contributed only by the PM 137, the estimated FWHM value of the output pulse would be 1.3 ns.

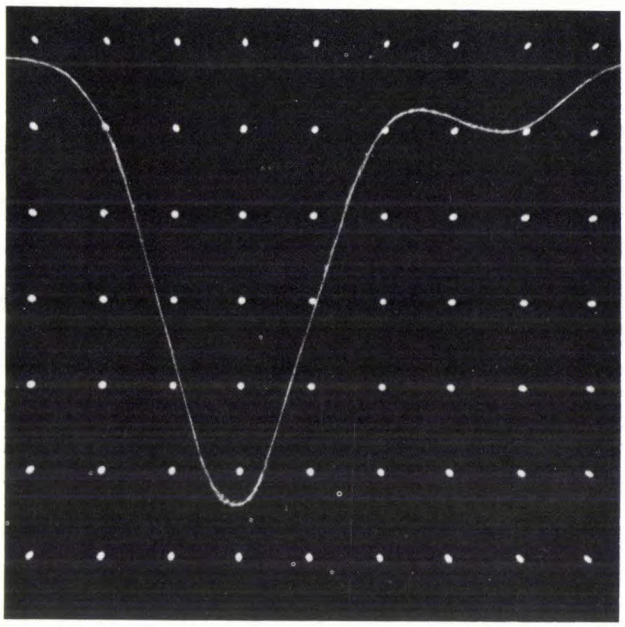


Fig. 3. Oscillogram of an output pulse from the PM HR 305 photomultiplier tube, used at maximum gain (5×10^3). The oscilloscope had a bandwidth of 4000 MHz. The FWHM value, as contributed by the PM HR 305, is estimated at about 0.23 ns. The input signal was an ultra-short laser light pulse (FWHM less than 0.05 ns, wavelength 538 nm).

(thickness 3.2 mm) has curved channels, a feature that overcomes ionic feedback and also increases the gain [3]. (In straight channels there is a high probability that the positive ions created inside and travelling back to the photocathode will generate spurious pulses and cause a slow deterioration in the photocathode sensitivity, as a result of sputtering.) The second microchannel plate, which is 0.5 mm thick, has about ten times as many channels. The electron avalanche that arrives as an output signal from a channel at the first plate is therefore divided between about 10 channels of the second plate. The channel diameters are 40 μm for the first plate and 12.6 μm for the second. A careful investigation of the operation of such a cascaded microchannel-plate arrangement has shown that the relative importance of gain, statistical gain fluctuations, noise and counting-rate performance has to be taken into account [4]. Gains exceeding 10^7 can be obtained by means of this cascading. The investigation indicated that the geometry illustrated in fig. 4 was the best compromise.

In the ultra-fast tube, the PM HR 305, the microchannel plate is a single unit whose thickness, about 0.4 mm, is much smaller than the value for the cascaded unit in the other tube. The channel diameter is 7.3 μm.

In both tubes the distance between the photocathode and the input face of the multiplier section is short (2 mm and 5 mm). The accelerating voltage applied to this input stage is chosen somewhere between

200 and 500 V; the exact value enables the user to obtain some desirable compromise between detection efficiency, noise, and transit-time fluctuations. In the input stage the transit-time fluctuations are virtually negligible, on account of 'proximity focusing'. This focusing is brought about by the simple expedient of making the distance between the (flat) photocathode and the multiplier input face small with respect to their cascaded microchannel plates. The tube gain can be adjusted from 1×10^6 to about 6×10^6 by increasing the voltage across the first microchannel plate from about 1200 V to 1600 V, with a fixed voltage (800 V) across the second plate (*fig. 5*). The statistical gain fluctuations are low enough to allow discrimination between the output signals corresponding to the emission of one photoelectron and of two photoelectrons.

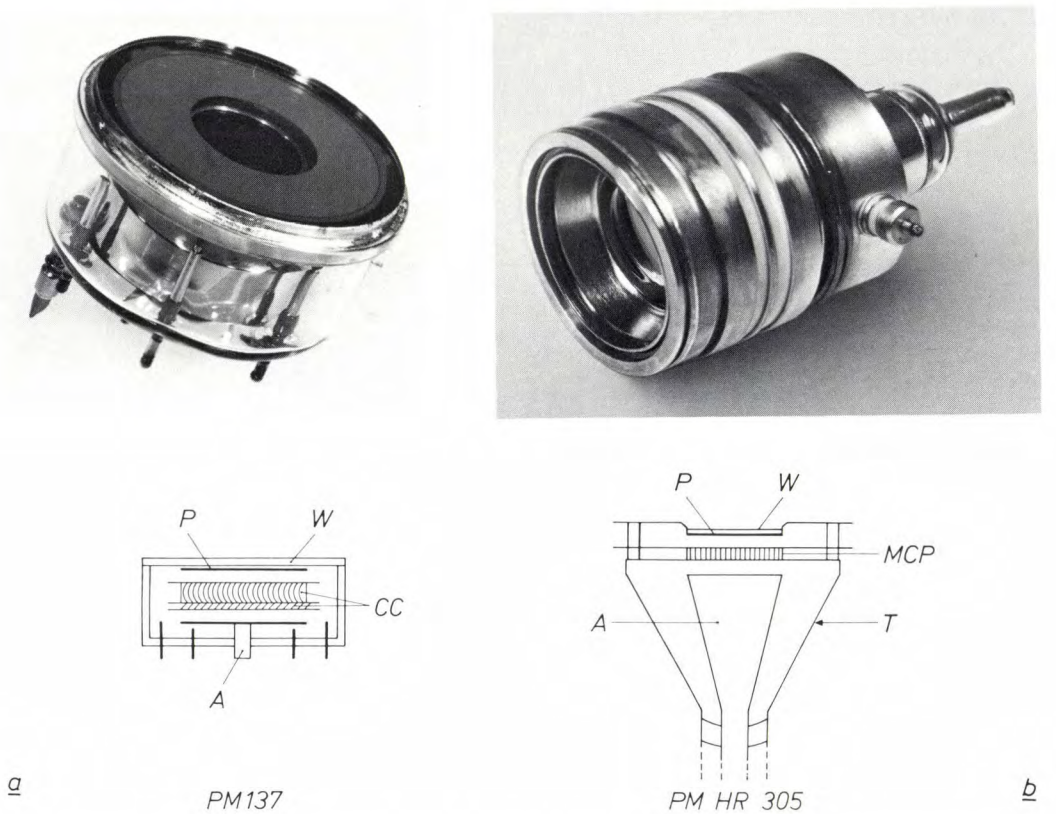


Fig. 4. The design of the photomultiplier tubes PM 137 (*a*) and PM HR 305 (*b*). *W* window. *P* photocathode. *A* anode. *CC* microchannel-plate multiplier (cascade of two). *MCP* microchannel-plate multiplier (single). *T* tapered output section. Further constructional data and performance figures are given in Table I.

diameters, and the electric field-strength high. The proximity focusing makes the operation of the tubes insensitive even to strong magnetic fields; measurements have shown that this sensitivity is at least a thousand times smaller than with conventional tubes, in which electrostatic focusing methods are employed [5]. This is a particularly important advantage in nuclear physics, where stray magnetic fields are often unavoidable.

The output stages of the tubes reflect the difference in characteristics. In the high-gain tube the anode is a simple plate located 3 mm from the multiplier output face. Perfect matching to the output line, a 50 Ω coaxial cable, would be superfluous in this case, on account of the relatively large FWHM value expected from the

The electrical resistance of the cascaded microchannel plates is about 2×10^8 Ω. Under conditions of maximum gain, therefore, the direct current — also termed 'standing current' or 'strip current' — in the multiplier is 7.5 μA. *Fig. 6* shows that the average output current of the tube increases linearly with the primary event rate, at least until the output current amounts to about 10% of the strip current. This was

[2] J. P. Gex, R. Sauneuf, J. P. Boutot and J. C. Delmotte. Proc. 13th Int. Congress on high-speed photography and photonics, Tokyo 1978.
[3] J. P. Boutot, G. Eschard, R. Polaert and V. Duchenois, Adv. in Electronics and Electron Phys. **40A**, 103, 1976.
[4] M. Audier, J. C. Delmotte and J. P. Boutot, Rev. Phys. appl. **13**, 188, 1978.
[5] C. C. Lo, P. Lecomte and B. Leskovar, IEEE Trans. **NS-24**, 302, 1977.

established by irradiating a scintillator crystal connected to the tube with gamma rays whose quantum energy was such that each scintillation (primary event) caused an average of 170 electrons to reach the multiplier input face. Under these conditions the linearity is maintained up to about 10^4 scintillations per second. If there are more than 10^4 primary events per second, the strip current can no longer deliver the electrons needed at the channel walls at the proper rate — and this causes a loss of gain and linearity. Linearity is a very useful feature, particularly when the tube is used to measure light-intensity levels.

The linearity of microchannel plates is much better than that of conventional multiplier structures. In the photomultiplier tubes with microchannel plates electric field-strengths up to 10 kV/cm are permissible, which is much higher than in conventional multipliers; this means that under otherwise identical conditions the loss of linearity caused by the build-up of space charge is distinctly lower. Since the tube with the higher gain can deliver charge pulses of up to 3×10^{-9} C in the linear range, peak current pulses of more than 1 A can occur in that range (for pulses with a width of 1 ns).

In the tube with the better pulse response the anode is of a much more complicated design than in the tube just described. The surface area of the anode and its distance from the output face of the microchannel plate have been calculated to minimize the effects of transit-time dispersion and electrical capacitance upon the pulse response of the tube. The coaxial output structure is tapered, to provide a gradual match to the characteristic impedance of the coaxial cable and hence to decrease pulse-broadening reflections. The impedance match to the cable is further improved by a special bypass capacitor between the output end of the microchannel plate and the earthed outer conductor of the coaxial cable.

As noted earlier, the overall dimensions of photomultipliers with microchannel plates can be considerably smaller than those of photomultipliers with conventional multiplier sections. The dimensions of the PM 137 tube do not exceed 5 cm. This tube has a simple glass envelope, with the connector pins in the base. The photocathode is made by means of the transfer technique [6]; a remote sealing method is then used to connect the end window to the body of the tube, using a hot indium alloy as gasket material. The PM HR 305 tube has a length of only 8 cm, its maximum dimension. The structure of this tube is rather unconventional. The body consists of a stack of metal rings (the electrodes) and ceramic rings (the insulators), hard-soldered together. The advantage of the rings is of course that their shape easily matches the required

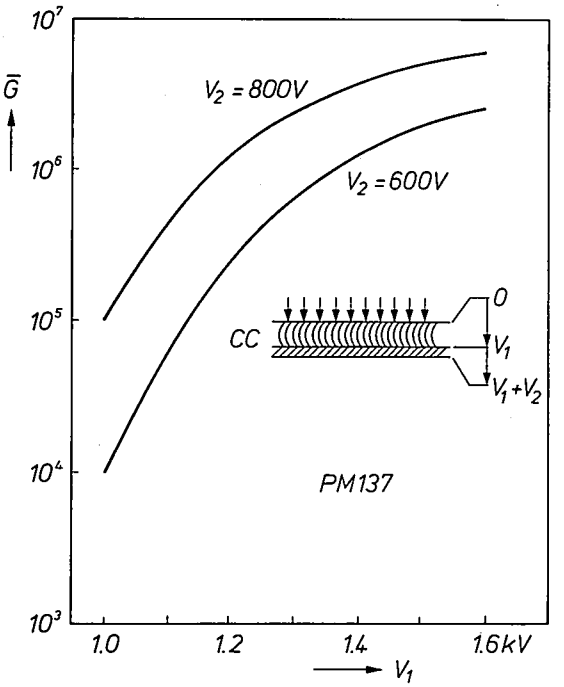


Fig. 5. The average d.c. gain $\bar{G}$ for the PM 137 photomultiplier tube plotted as a function of the voltage V_1 across the first microchannel plate, with the voltage across the second microchannel plate of the multiplier set at 800 V and at 600 V. *CC* microchannel-plate multiplier (cascade of two).

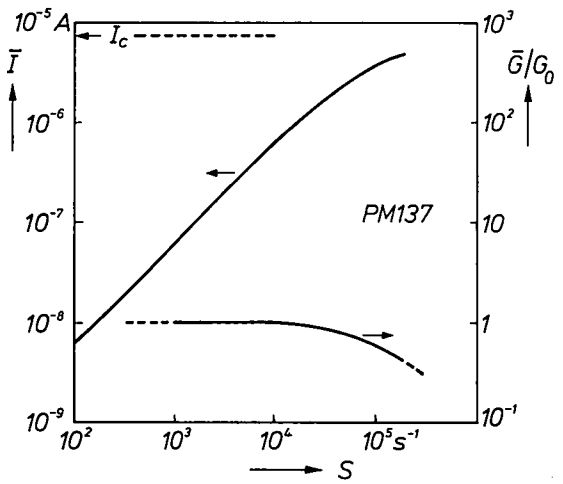


Fig. 6. The average output current $\bar{I}$ of the PM 137 photomultiplier tube as a function of the primary event rate S. Each primary event causes a 'cloud' of 170 electrons, which impinge on the multiplier input face. Loss of linearity becomes noticeable at $S > 10^4\ s^{-1}$. The lower curve is the relative average gain. At $S = 10^4\ s^{-1}$ the loss of gain amounts to 2%. $G_0 = 2.3 \times 10^6$. I_c strip current, the direct current in the multiplier at maximum gain.

coaxial output structure. The techniques adopted to make the photocathode and to finally seal the end window are very similar to the ones just described for the other tube. At the end of the tapered output section a bead-shaped glass-to-metal seal provides an effective

[6] R. Legoux, Philips tech. Rev. 30, 234, 1969.

vacuum seal (fig. 4) and gives a good electrical match.

The problems of spurious pulses and photocathode decay due to sputtering — which arise from the difficulty of suppressing ion avalanches in the input stage — are reduced in the PM HR 305 tube by coating the microchannel plate with a thin layer of Al_2O_3. This coating reduces the rate of ion feedback to such an extent that the total output charge from the PM HR 305 tube (calculated cumulatively during the life of the tube) can be taken to be of the order of 0.1 C.

Summary. The construction and performance of two microchannel-plate photomultiplier tubes, PM 137 and PM HR 305, are described. The PM 137 contains two microchannel plates in cascade, and produces a high gain (max. $10^7\times$); the PM HR 305 contains a single plate, and gives a fast pulse response (full width at half maximum height 0.20 ns). The time resolution of the PM 137 is better than 0.10 ns. The sensitivity to external magnetic fields is low. Ion feedback in the channels is reduced by using a plate with curved channels in the PM 137 and by coating the plates in the PM HR 305 with a thin layer of Al_2O_3. The applications of the tubes are in high-energy nuclear physics and in plasma physics.

Apparatus based on Philips PW 1100 diffractometer for crystal-structure research at high pressures

W. Denner and Heinz Schulz

In many areas of science technology it is desirable to have information on the behaviour of matter, whether gaseous, liquid or solid, under extreme conditions, for example at high or low temperatures or at high pressures. Typical instances are found in the conditions that occur in the processing of certain materials, in the shock waves from an explosion, and also below the Earth's crust. Among the areas of science where the effects of high pressure are of special interest are solid-state physics, mineralogy, geology and chemistry.

Important information on the behaviour of the solid state under extreme conditions is obtained from changes in the parameters of crystal structure. Routine measurements have long been used for determining the effects of temperature on the structure of crystals, but until recently this did not apply to the measurement of effects at high pressure. Routine measurements at high pressure have only really become a practical proposition since it became possible to make sufficiently small high-pressure cells, in which diamond anvils are used.

In the article below the authors describe an apparatus that can be used for measuring X-ray reflection intensities from crystals subjected to high pressure in cells of this type. The apparatus is based on the Philips PW 1100 single-crystal diffractometer. The necessary modifications to the software were made in cooperation with Dr E. Keulen, Ir H. Vossers and Ing. J. K. Huften of the Philips Scientific & Industrial Equipment Division at Almelo.

Introduction

High hydrostatic pressures can be generated in various ways. One of the simplest — known for some time as the 'diamond-anvil high-pressure technique' — is based on a principle developed by P. W. Bridgman [1]. In this technique the head faces of two diamond anvils exert pressure on a very thin steel disc which has a circular hole in it. The cylindrical cavity formed by the hole is the pressure chamber; it contains the crystal under investigation, a reference crystal for the pressure measurement and a liquid that transmits the pressure to the crystals.

It has recently become possible to make high-pressure cells of this type with such small dimensions that they can easily be built into commercially available measuring equipment. This has led to investigations in which a large number of physical phenomena, including absorption, luminescence, radioactivity, Raman scattering and electrical conduction, have been measured as a function of pressure [2]. These investigations made it clear that the results of the measurements cannot be unambiguously interpreted and evaluated in quantitative terms without a knowledge of the changes that the crystal structure undergoes when subjected to pressure. Although routine methods have long been in use for systematic determination of the effects of temperature on the structure of a crystal, measuring the effects of pressure on crystal structure has until recently been a problem.

For investigating these pressure effects it is desirable to carry out X-ray investigations of single crystals

Dr W. Denner and Prof. Dr Heinz Schulz are with the Max-Planck-Institut für Festkörperforschung, Stuttgart, West Germany.

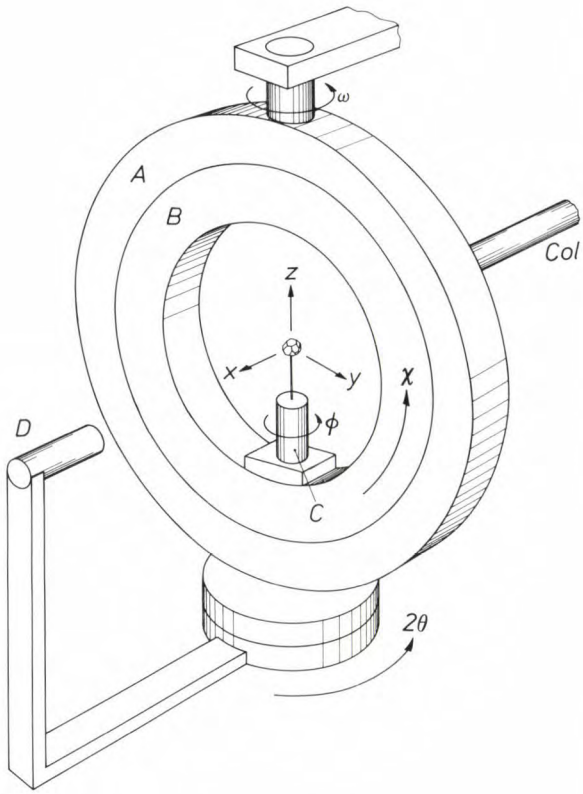


Fig. 1. Diagram of the Eulerian cradle of the Philips PW 1100 four-circle diffractometer. Ring *A* rotates about the (vertical) ω-axis. Ring *B* rotates inside ring *A*, the χ movement. The crystal holder *C* turns about the ϕ-axis and the detector *D* about the 2θ-axis, which coincides with the ω-axis. *Col* collimator. All the components are at their zero settings.

subjected to high hydrostatic pressure. We have therefore modified a type of high-pressure cell now in use [3] so that it can be built into the Philips PW 1100 single-crystal diffractometer (*figs 1* and *2*). A detailed description of this four-circle diffractometer has been given in an earlier article [4] published in this journal, which will be referred to here as I. Since all high-pressure cells suitable for this purpose are based on the same principle, the modifications of the PW 1100 that we discuss here are equally applicable to high-pressure cells of other types.

[1] P. W. Bridgman, Proc. Amer. Acad. Arts Sci. **79**, 127, 1951. See also M. Brouha and A. G. Rijnbeek, Philips tech. Rev. **36**, 245, 1976.

[2] W. K. Hensley, W. A. Bassett and J. R. Huizenga, Science **181**, 1164, 1973.
L. S. Whatley and A. Van Valkenburg, High pressure optics, in: Advances in high pressure research **1**, 327-371, 1966.
B. A. Weinstein and G. J. Piermarini, Phys. Lett. **48A**, 14, 1974.
D. U. Gubser and A. W. Webb, Phys. Rev. Lett. **35**, 104, 1975.
G. J. Piermarini and S. Block, Rev. sci. Instr. **46**, 973, 1975.
S. Block and G. J. Piermarini, Physics Today **29**, Sept. 1976, p. 44.

[3] The high-pressure cell we used is described in: R. Keller and W. B. Holzapfel, Rev. sci. Instr. **48**, 517, 1977.
Information on other high-pressure cells will be found in: D. Schiferl, J. C. Jamieson and J. E. Lenko, Rev. sci. Instr. **49**, 359, 1978.
C. E. Weir, G. J. Piermarini and S. Block, Rev. sci. Instr. **40**, 1133, 1969.
L. Merrill and W. A. Bassett, Rev. sci. Instr. **45**, 290, 1974.
R. Fourme, J. appl. Cryst. **1**, 23, 1968.

[4] J. Hornstra and H. Vossers, Philips tech. Rev. **33**, 61, 1973.

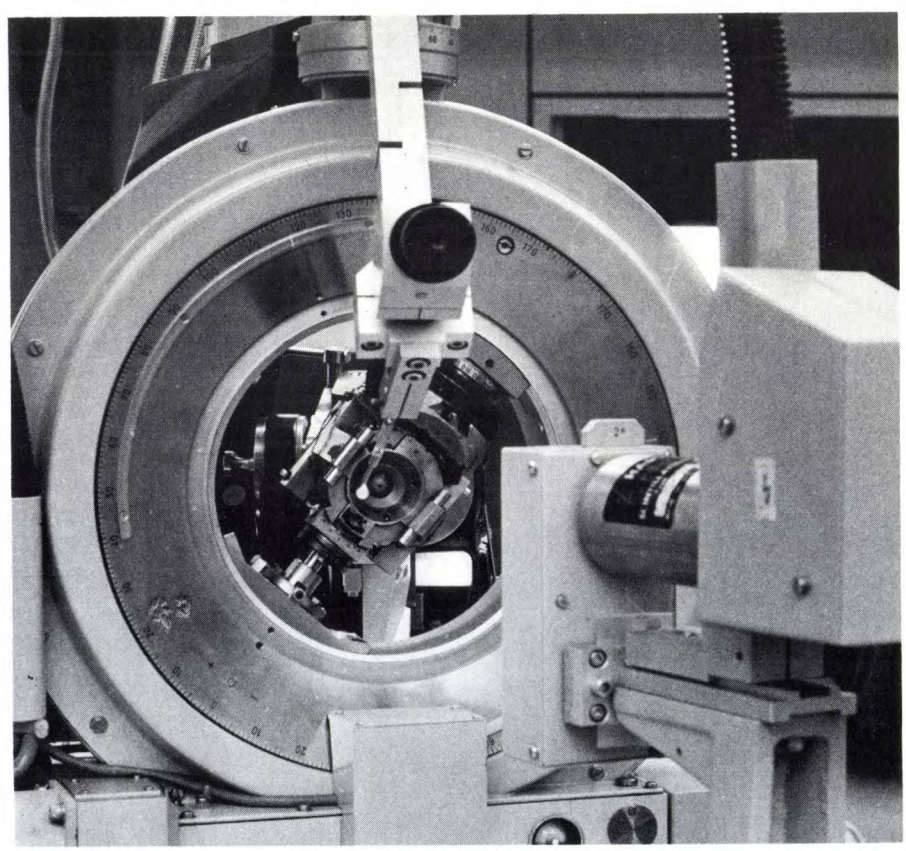


Fig. 2. The Eulerian cradle of the PW 1100 with built-in high-pressure cell in an arbitrary position. The detector can be seen on the right.

The diamond-anvil high-pressure cell

A cross-section of the high-pressure cell used in our apparatus is shown schematically in *fig. 3*. The cross-section is taken perpendicular to the ϕ-axis and in the plane of the incident X-ray beam (fig. 1). The two diamond anvils are embedded in beryllium (the hatched parts in fig. 3). As well as possessing high mechanical strength, beryllium has a relatively small absorption coefficient for X-rays. The diamond anvils through which the X-ray beam penetrates into the high-pressure cell and leaves it again are also sufficiently transparent to both the incident and the diffracted X-ray beam. The other parts of the high-pressure cell serve to generate the pressure, and are made of Ultrafort steel; these parts do not transmit the X-rays.

The size of the pressure chamber determines the size of the crystal sample. In our case the pressure chamber has a diameter of 180 μm and a height of 150 μm at a pressure of 1 bar. With rising pressure the diameter changes only very slightly, but the height becomes very much smaller. With the types of steel we have used so far (Inconel or Hastalloy) the pressure chamber at the frequently applied pressure of about 60 kbar has only been about 70-80 μm high. This means that the crystal sample must not be larger than 70 μm, at least in one direction. The other dimensions of the crystal are limited by the fact that the pressure chamber must also have room for the reference crystal.

Media that were very successful for transmitting the pressure are a mixture of methanol and ethanol (in the ratio of 4 to 1) and a mixture of pentane and isopentane (in the ratio of 1 to 1) [5]. With the methanol/ethanol mixture a hydrostatic pressure up to about 100 kbar can be reached, and with the other a pressure of about 65 kbar.

The basic modifications

Mechanical modification of the high-pressure cell to suit the PW 1100

In the mechanical modification of the diamond-anvil high-pressure cell to suit the PW 1100 four-circle diffractometer, great care was taken to ensure that the crystal sample, which has to be positioned exactly at the common centre-point of the four circles, remains in position even during lengthy measurements and without unduly loading the bearings of the apparatus. Care was also taken to see that the freedom of movement of the Eulerian cradle would not be restricted [6].

The mechanical modification of the PW 1100 also led to changes in the procedure used for aligning the crystal sample before the automatic intensity measurements. In the conventional diffractometer the alignment is performed manually, by translation of the crystal in three directions perpendicular to each other. That is to say, the position is found for the crystal at which no displacement is observed when the crystal is rotated about the different axes of the apparatus.

To make this procedure possible the high-pressure cell has to be mounted on a goniometer head, which is used to displace the cell — weighing about 0.6 kg — in three directions at right angles to each other (x, y and z in *fig. 4*). The translation in the z-direction presented particular problems. Because of the dimensions of our high-pressure cell there was not enough room in the z-direction to allow us to use a commercially available goniometer head. We therefore used an arrangement of our own to incorporate the high-pressure cell in such a way as to permit a translation in the z-direction, and we used the corresponding part of a slightly modified STOE goniometer head for the translations in the x- and y-directions.

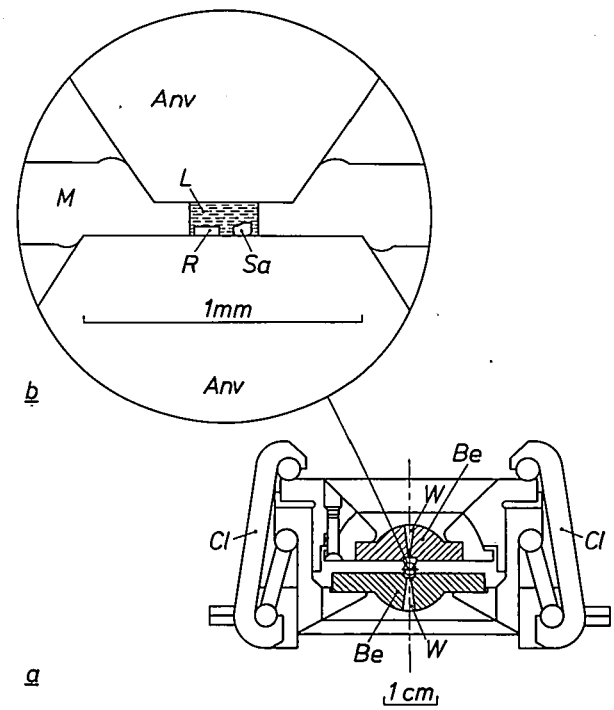


Fig. 3. Section through the high-pressure cell (*a*) and through the central part of the cell (*b*). *Anv* diamond anvils. *Be* beryllium components with windows *W*. (One of the windows can also be seen in fig. 2.) *Cl* clamp that produces the pressure. *M* metal seal. *L* liquid that transmits the pressure to the crystals. *Sa* crystal sample. *R* reference crystal for the pressure determination.

Since the crystals may have to remain in the X-ray beam in any position of the apparatus for a long time during the measurements, and since the high-pressure cell together with the new goniometer head weighs nearly 1 kg, we took steps to avoid overloading the bearings of the Eulerian cradle. These included providing the high-pressure cell with *two* bearings, i.e. the

second bearing at $\chi = 180°$ (fig. 4) in addition to the existing bearing at $\chi = 0°$. The position of the axis of this second bearing can be adjusted by means of two sets of screws (S_1 and S_2 in fig. 4) so that this axis exactly coincides with the ϕ-axis [6]; this is necessary because even slight frictional forces can cause some irregularity in the ϕ movement.

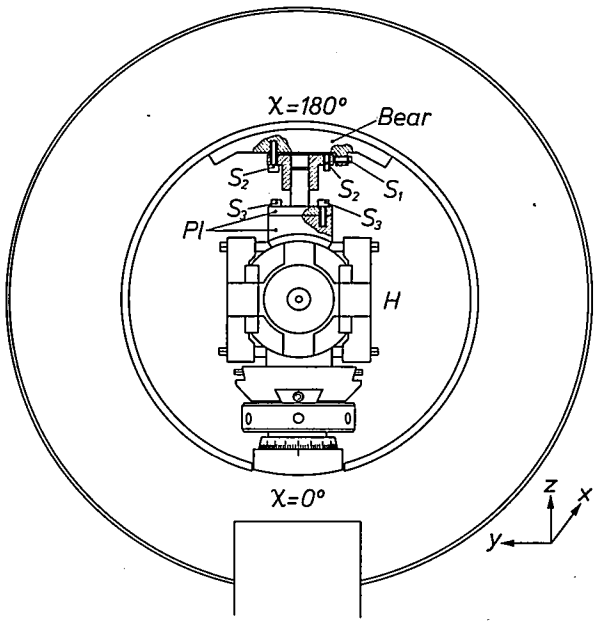


Fig. 4. Cross-section of the χ ring with built-in high-pressure cell *H*. *Bear* the second bearing. *Pl* plane-parallel plates, enabling the high-pressure cell to be displaced during alignment. S_1, S_2 and S_3 screws. S_1 and S_2 are for adjusting the needle axis of this bearing with respect to the ϕ-axis, and screws S_3 are for locking the high-pressure cell in position after alignment.

For the second bearing we chose a needle bearing, partly to minimize the friction that is always present even with optimum adjustment. To prevent this bearing from making it more difficult to align the crystal, the high-pressure cell and the bearing are kept apart by means of two plane-parallel plates. During alignment, especially during the translation in the x- and y-directions, these plates can be displaced parallel to each other, and then locked firmly together by means of the two screws S_3 (fig. 4).

This second bearing not only guarantees the necessary stability but also relieves the other bearings, in particular those of the χ and ϕ rings (fig. 1). To relieve the load on the stepping motor for the χ ring, the weight of the bearing at $\chi = 0°$ is made approximately equal to that of the bearing at $\chi = 180°$. To minimize the extent to which the high-pressure cell restricts the freedom of movement of the diffractometer, the collimator had to be made 2 cm shorter, and the primary-beam stop is located further from the centre of the apparatus. Fig. 2 shows a photograph of the Eulerian cradle with the built-in high-pressure cell.

Modification of the alignment procedure

For accurate alignment of the crystal contained in the high-pressure cell it must be possible to observe it. As shown in fig. 3, openings ('windows') are therefore provided in the beryllium parts. With a suitable microscope, the crystals in the high-pressure cell can be observed through these windows and through the transparent diamond anvils. The alignment is carried out with the aid of a microscope that gives a magnification of 200 × and has a working distance of 3-4 cm. It is fixed to the Eulerian cradle in such a way that its optical axis lies in the x,y-plane of the apparatus and so that an image of the crossed wires in the eyepiece can be produced at the centre of the Eulerian cradle. For the translation parallel to the optical axis the microscope can be finally adjusted to an accuracy of a few microns. By using this microscope the crystal sample can be aligned just as accurately as in a conventional diffractometer.

Alignment in the plane perpendicular to the optical axis is performed in the usual way, as described earlier, with the aid of the crossed wires in the eyepiece. This cannot be done in the direction parallel to the optical axis, and the alignment is then performed by sharply focusing the image of the crystal; it has to be done in this way because the crystal disappears from the picture after a certain rotation since the steel parts of the high-pressure cell come between the microscope and the crystal. In this case the crystal is properly aligned when its image is sharp again after a rotation of 180°. It is also necessary in this procedure to make allowance for the difference in optical path lengths resulting from the difference in the size of the two diamond anvils [7]. The displacements of the crystal in the direction of the optical axis can be made by fine adjustment of the goniometer head.

Fig. 5 is a photograph of the Eulerian cradle with both the high-pressure cell and the alignment microscope. After alignment the microscope can easily be removed from the apparatus and just as easily be reproducibly repositioned if any subsequent alignment is needed. (The microscope has to be removed if intensity measurements are to be carried out.)

[5] G. J. Piermarini, S. Block and J. D. Barnett, J. appl. Phys. **44**, 5377, 1973.
[6] W. Denner, Strukturanalyse an Einkristallen unter hydrostatischem Druck, dissertation, Karlsruhe 1977.
W. Denner, W. Dieterich, Heinz Schulz, R. Keller and W. B. Holzapfel, Rev. sci. Instr. **49**, 775, 1978.
[7] R. Keller, Einkristall-Röntgenbeugung an Tellur unter hohen Drücken bis 40 kbar, dissertation, Stuttgart 1975.

Effect of the high-pressure cell on the measurement of X-ray intensities

Measurement range in reciprocal space

Although our apparatus is designed in such a way that the high-pressure cell does not interfere with the movements of the four rings, the cell does limit the measurement range in reciprocal space (*fig. 6*). Fig. 6*a* shows the high-pressure cell in a cross-section perpendicular to the ϕ-axis, with the apparatus in the zero setting; in fig. 6*b* the cell has been rotated by about 40° around the ϕ-axis. In the latter position the steel parts of the high-pressure cell obstruct the primary X-ray beam and the X-radiation is completely absorbed. In certain angular regions during rotation about the ω-axis the high-pressure cell also makes it impossible to measure the reflection intensities [8].

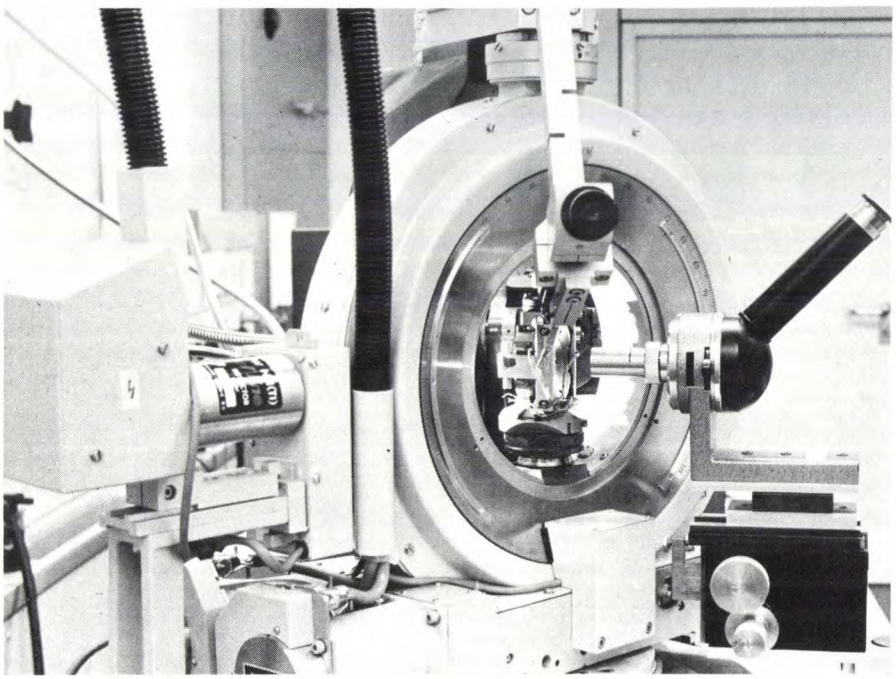


Fig. 5. Eulerian cradle with the aligning microscope in a setting required for alignment.

In the standard measurement program of the PW 1100 a vector of reciprocal space is first rotated about the ϕ-axis to bring it into the plane of the χ ring, then about the χ-axis to bring it into the horizontal plane and finally about the ω-axis until the lattice plane corresponding to the vector reaches the reflecting condition (fig. 1). This setting, at which the azimuth angle ψ is zero, is called the bisector setting or the symmetrical setting [4]. Because of the limitation in the measurement range with the ϕ and the ω rotation, this standard program can only be used for measuring a part of the reciprocal space. The measurement range can be considerably enlarged if it is possible to search not only at the azimuth angle $\psi = 0$ but also at other azimuth angles. In an azimuthal rotation, with varying ψ, one lattice plane always remains in the reflection setting, and the crystal with its holder always rotates at the same angle about the vector of reciprocal space corresponding to that lattice plane. The PW 1100 has a program which can produce this motion by means of a rotation about the ω-, the χ- and the ϕ-axes simultaneously [4].

Systematic errors in the reflection intensity caused by absorption, shadowing and overlapping effects

Since for azimuthal rotation — i.e. varying ψ for the same reflection — the path of the X-ray beam is different at the different ψ-values, the reflection intensity also varies in a different way. And for every other reflection the intensity also varies in a different way. There are two reasons for these intensity variations: the first is the absorption in the crystal sample, and the other is the screening effect of the steel that surrounds the pressure chamber (the 'seal'). Because the path travelled is different each time, the absorption in the crystal varies, and the shadowing effect of the seal can

[8] W. Denner, Heinz Schulz and H. d'Amour, J. appl. Cryst. **11**, 260, 1978.

[9] W. Denner, H. d'Amour, Heinz Schulz and W. Stoeger, J. appl. Cryst. **10**, 177, 1977.

mean either that the primary beam does not cover the entire crystal or that only part of the diffracted beam is detected (*fig. 7*). The shadowing effect does not only differ with different reflections or other ψ-values for the same reflection, but also differs when the reflection and the ψ-value are the same but the pressure is different. This is because a change in pressure varies the height of the seal, while there can also be slight changes in the diameter of the pressure chamber and in the orientation of the crystal.

Both of these factors, absorption and shadowing effect, are difficult to treat analytically. They can best be corrected by experimental means.

To cover the largest possible measurement range in reciprocal space it is necessary to find the best value of ψ for each reciprocal lattice vector separately, i.e. the ψ-value at which the shadowing effect and the absorption are smallest. The way in which this optimum value is determined is described later.

Even if each reflection is measured separately at its optimum ψ-value, the measured intensity can still give a systematic deviation, caused by overlapping effects. These arise because the X-ray beam, in passing through the high-pressure cell, encounters a total of four single crystals: the two diamond anvils, the reference crystal and the crystal sample. The reciprocal lattices of these four crystals overlap, and it is therefore possible that points of the reciprocal lattice of two or more different crystals will lie so close together that they are detected simultaneously. In every intensity measurement it is therefore necessary to check whether the value of the reflection intensity has been distorted by overlapping [8].

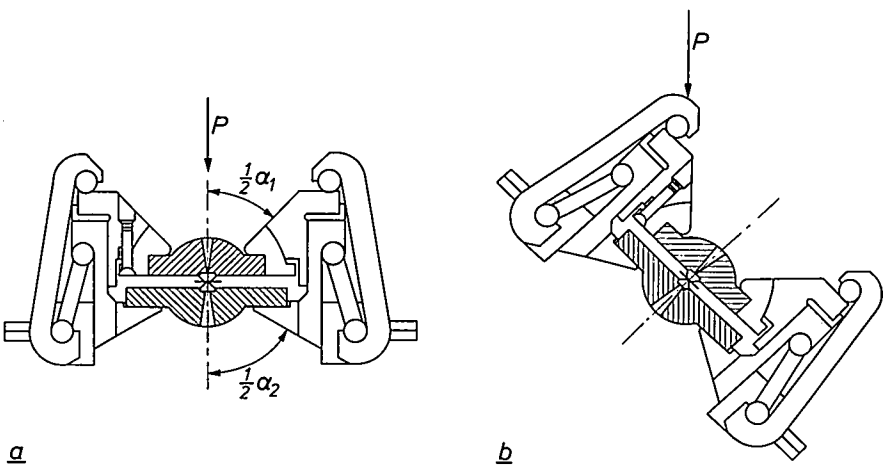


Fig. 6. a) In the zero setting of the diffractometer, i.e. at $\omega = 2\theta = \chi = \phi = 0$ the high-pressure cell does not obstruct the primary X-ray beam *P*. *b*) In the position $\omega = 2\theta = \chi = 0$ and $\phi \approx 40°$ primary-beam obstruction does occur. α_1 and α_2 are the aperture angles of the high-pressure cell.

Correction of systematic errors

Overlapping

Once the orientation matrices of all the crystals contained in the high-pressure cell have been determined [4], the values of ω, 2θ, χ and ϕ at which reflection occurs can be calculated for all points of the reciprocal lattice of these crystals. Overlapping occurs if a reciprocal lattice vector h_1 of the crystal specimen is found at a reflection position (ω_1, $2\theta_1$, χ_1 and ϕ_1) that is 'insufficiently discriminated' from the reflection position (ω_2, $2\theta_2$, χ_2, ϕ_2) of another reciprocal lattice vector h_2 of another crystal [9]. Overlapping reflections are not measured.

Shadowing effect and absorption

To determine a correction factor $A'_n(\omega, 2\theta, \chi, \phi)$ for the shadowing and absorption effects of the high-pressure cell we use reflections that are symmetrically

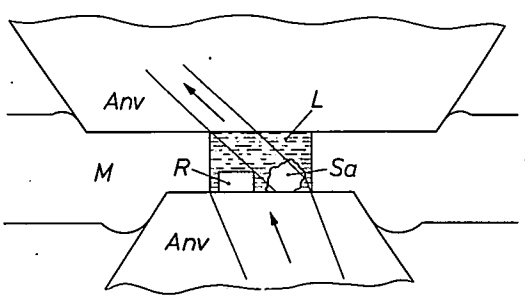


Fig. 7. Illustrating the absorption and shadowing effects. The diagram shows the pressure chamber and the path of the X-ray beam passing through it. The path is different for every ψ-value that can be chosen for a reflection, and also for any other reflection. The meaning of the symbols is given in fig. 3*b*.

▷

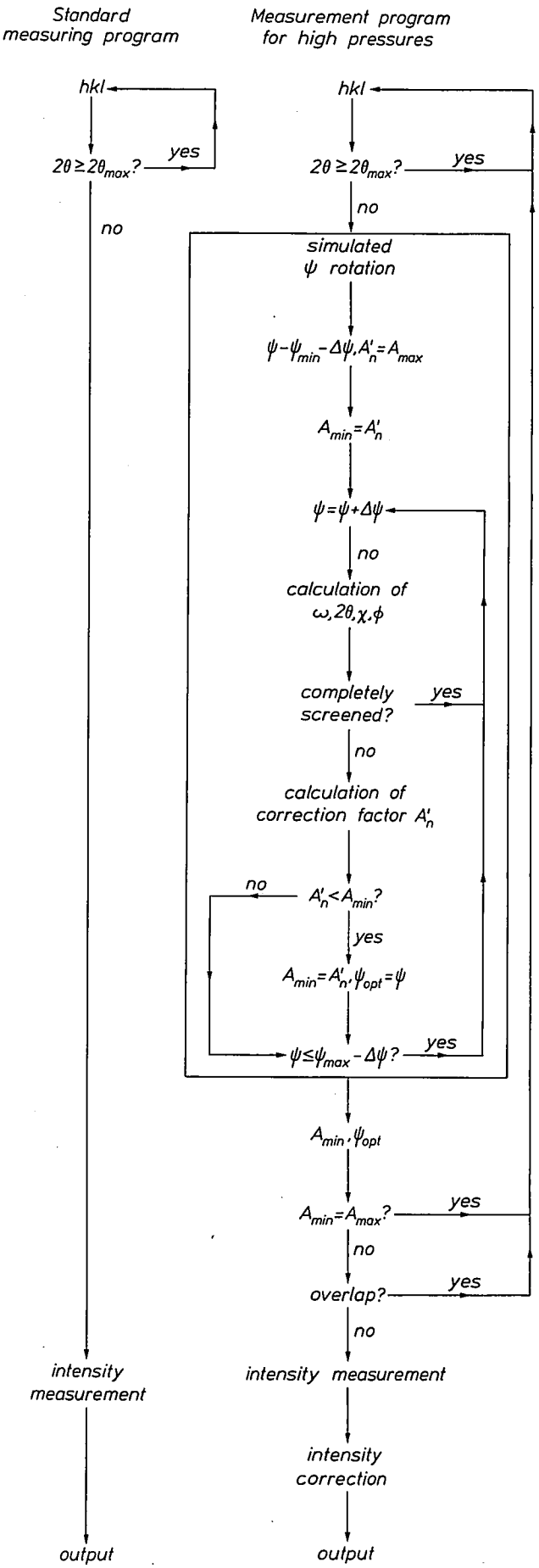
Fig. 8. The standard measuring program of the PW 1100 (*left*) and the program we have designed for intensity measurements at high pressure.

related to each other [10]. The aim is to include as many such reflections in the determination as possible, and also to extend the measurement to as many ψ-values as possible.

A modified program for intensity measurements

To enable the reflection intensities to be measured completely automatically at high hydrostatic pressure and in a wide measurement range in reciprocal space, the measurement program described in I has been modified and supplemented in such a way that the shadowing and absorption effects are corrected during the measurements. This new program for measurements at high pressure yields intensity data that are just as reliable as those used for normal structure research. *Fig. 8.* shows the difference between the standard measurement program of the PW 1100 and the new program for measurements under pressure [8].

The main difference between these two programs is that in the high-pressure program a ψ rotation is first simulated in the control computer before the measurement of a reflection intensity is started. At the beginning of this simulation ψ is set at the lower limit and an arbitrary, high value A_{max} is assigned to the correction factors A'_n $(\omega, 2\theta, \chi, \phi)$ and A_{min}. In an 'internal loop' of the simulation every ψ-value is checked to see whether the high-pressure cell transmits the X-ray beam at the corresponding combination of ω, 2θ, χ and ϕ-values (i.e. whether or not the outer steel parts of the cell obstruct the X-ray beam). For the ψ-values at which the X-ray beam is transmitted, the corresponding correction factor A'_n $(\omega, 2\theta, \chi, \phi)$ is calculated and a check is made to find whether the value found is smaller than the value A_{min} stored in the computer memory. If this is the case, then the value of A'_n $(\omega, 2\theta, \chi, \phi)$ is stored in the memory as a new value of A_{min}, together with the corresponding ψ-value as an optimum value of ψ (ψ_{opt}). If the upper limit has not yet been exceeded at the next value of ψ, another search is made via the internal loop. In this way we obtain from the simulation the optimum ψ-value of each reflection at which the effect of the high-pressure cell on the reflection intensity is at a minimum. If A_{min} still has the value A_{max} after the simulation, there is no ψ-value at which the X-ray beam can pass through the high-pressure cell, owing to the shadowing effect of the seal. After the simulation, all the reflections that are not absorbed in this way by the high-pressure cell are checked to see whether they overlap at their optimum ψ-value. As mentioned above, such reflections are not measured.



The final intensity measurement is made on all other reflections at their optimum ψ-value, ψ_{opt}, and the intensity is multiplied by A_{min}. The results are then recorded on magnetic tape.

Making the measurements

Pressure calibration

The pressure in a diamond-anvil high-pressure cell is determined in most laboratories by means of a small ruby chip. The pressure is derived from the shift of the R_1 luminescence line of ruby with the aid of a spectrophotometer. This method of pressure determination is fast and also fairly simple, but a disadvantage of using it in crystal-structure research with a diffractometer is that the high-pressure cell has to be removed from the apparatus for every pressure measurement, and then put back, re-aligned and so on.

If, however, a reference crystal is used whose changes under pressure are known, e.g. an NaCl, KCl or CsCl crystal [11], the pressure can then be derived directly from the lattice constants determined with the diffractometer. Since with the PW 1100 the lattice constants have already been found when determining the orientation matrix, which is necessary anyway for the correction of the overlap effects, pressure calibration by this method does not involve extra work. The accuracy of the pressure measurement depends on the error in the determination of the lattice constant. With the PW 1100 this is about 0.2% and when an NaCl or a CsCl crystal is used, this value corresponds to an inaccuracy of approximately 0.5 kbar in the measured pressure.

Determination of the orientation matrices

The orientation matrices of the four single crystals contained in the high-pressure cell can be determined by the 'peak-hunting' subroutine described in article I [4]. Special attention needs to be paid here to the limitations upon rotation about the ϕ-axis. To carry out this subroutine the reflections originating from different crystals have to be kept distinct from each other. If at least three linearly independent reflections are found for each crystal, the orientation matrices can be calculated as described, for example, in article I [12]. When this has been done, the high-pressure measuring program can be started.

Since the effects due to the high-pressure cell are corrected during the actual measurements, the data obtained in this way can be used directly for the standard programs, as we have shown elsewhere [13].

In conclusion, we wish to express our thanks for the support we have received in part of our work from the Deutsche Forschungsgemeinschaft.

Summary. Description of an apparatus for crystal-structure research at high pressures (up to 100 kbar), based on the Philips PW 1100 single-crystal diffractometer. A diamond-anvil high-pressure cell, modified to suit the PW 1100, is mounted in the χ ring on two sets of bearings to guarantee the necessary stability. The crystal is aligned with the aid of a microscope. The measurement range is not limited to $\psi = 0°$, as it is in the standard program for the PW 1100. To cover the widest possible range of reciprocal space, every reflection is measured separately at an optimum ψ-value. Corrections for shadowing and absorption effects are made during the intensity measurements. Reflections overlapped by reflections from the diamond anvils or from the reference crystal are not measured. The data obtained can be used in standard programs for crystal-structure determination.

[10] H. D. Flack, Acta cryst. **A30**, 569, 1974 and **A33**, 890, 1977.
[11] D. L. Decker, J. appl. Phys. **42**, 3239, 1971.
[12] A detailed description is given in: W. R. Busing and H. A. Levy, Acta cryst. **22**, 457, 1967.
[13] W. Denner, Heinz Schulz and H. d'Amour, Naturwiss. **65**, 257, 1978.
W. Denner, Heinz Schulz and H. d'Amour, Acta cryst. **A35**, 360, 1979.
H. d'Amour, W. Denner and Heinz Schulz, Acta cryst. **B35**, 550, 1979.

Scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

G. van Aller & W. Kühl (Philips Elcoma Division, Eindhoven): Low-light-level imaging.
Acta Electronica **20**, 231-239, 1977 (No. 3).

R. N. Bates: Design of microstrip spur-line band-stop filters.
IEE J. Microwaves Opt. Acoust. **1**, 209-214, 1977 (No. 6). *R*

E. Bauböck: Mask design for integrated circuits — an application for computer graphics.
Proc. 3rd Eur. Electro-Optics Conf., Geneva 1976 (SPIE Vol. 99), pp. 310-316; 1977. *H*

C. I. M. Beenakker: Evaluation of a microwave-induced plasma in helium at atmospheric pressure as an element-selective detector for gas chromatography.
Spectrochim. Acta **32B**, 173-187, 1977 (No. 3/4). *E*

V. Belevitch: Capacitances of circuits in a tight screened pair or quad.
Philips J. Res. **33**, 139-148, 1978 (No. 3/4). *B*

J. Bloem, A. Bouwknegt (Philips Lighting Division, Roosendaal) **& G. A. Wesselink:** Some new mercury alloys for use in fluorescent lamps.
J. Illum. Engng. Soc. **6**, 141-147, 1977 (No. 3). *E*

J. van den Boomgaard: Composieten.
Klei en Keramiek **27**, 42-55, 1977 (No. 3). *E*

J. van den Boomgaard: Gas bubble formation ahead of a solidification front.
Philips J. Res. **33**, 149-185, 1978 (No. 3/4). *E*

P. W. J. M. Boumans: Is optical emission spectrometry being tailored to the requirements and budgets of a large variety of users?
Proc. Anal. Div. Chem. Soc. **14**, 143-151, 1977 (No. 6). *E*

A. Broese van Groenou: Wat gebeurt er bij slijpen van keramiek?
Klei en Keramiek **27**, 125-140, 1977 (No. 7/8). *E*

P. J. Courtois & J. Georges: On starvation prevention.
RAIRO Informatique **11**, 127-141, 1977 (No. 2). *B*

M. Cratipus (S.F.P.), **R. Astor & Y. Sannier** (T.R.T.): Equipement de liaison hertzienne pour reportages mobiles de télévision.
Radiodiffusion-Télévision **11**, No. 47, pp. 9-14, 1977. *L*

A. L. Dalisa: Electrophoretic display technology.
IEEE Trans. **ED-24**, 827-834, 1977 (No. 7). *N*

M. Delfino, J. C. Jacco, P. S. Gentile* & D. D. Bray* (* Fordham University, Bronx, N.Y.): Solid-state study of hydrazonium tartrate and deuterated hydrazonium tartrate enantiomers.
J. solid State Chem. **21**, 243-251, 1977 (No. 3). *N*

P. Dolizy, O. De Luca & M. Deloron: Photoémission des antimoniures multialcalins de type (Sb, Na, K, Cs).
Acta Electronica **20**, 265-296, 1977 (No. 3). *L*

M. Fouassier, J.-C. Rosier & J. Dietz: Résistance à l'éblouissement des tubes intensificateurs d'images à galette de microcanaux et à double focalisation de proximité.
Acta Electronica **20**, 369-378, 1977 (No. 4). *L*

R. French: Error performance in mobile radio data transmission in the urban environment.
NTG-Fachber. **61**, 83-88, 1977. *R*

K. Gawron & J. Schröder: Properties of some salt hydrates for latent heat storage.
Int. J. Energy Res. **1**, 351-363, 1977 (No. 4). *A*

Y. V. Genin & Y. G. Kamp: Two-dimensional stability and orthogonal polynomials on the hypercircle.
Proc. IEEE **65**, 873-881, 1977 (No. 6). *B*

H. C. de Graaff, J. W. Slotboom & A. Schmitz: The emitter efficiency of bipolar transistors. Theory and experiments.
Solid-State Electronics **20**, 515-521, 1977 (No. 6). *E*

G. J. van Gurp: Electromigration and Hall effect in cobalt films.
J. Phys. Chem. Solids **38**, 627-633, 1977 (No. 6). *E*

B. F. van der Heyden (Philips Elcoma Division, Eindhoven): Design of stable, very low noise, cavity-stabilized IMPATT oscillators for *C* band.
IEEE Trans. **MTT-25**, 318-323, 1977 (No. 4).

A. Huizing, H. A. M. van Hal, W. Kwestroo, C. Langereis & P. C. van Loosdregt: Hydrates of manganese (II) oxalate.
Mat. Res. Bull. **12**, 605-611, 1977 (No. 6). *E*

G. Jacob & D. Bois: Efficient injection mechanism for electroluminescence in GaN.
Appl. Phys. Letters **30**, 412-414, 1977 (No. 8). *L*

D. Kasperkovitz & D. Grenier (R.T.C. La Radiotechnique-Compelec, Caen): Travelling-wave dividers: a new concept for frequency division.
Microelectronics and Reliability **16**, 127-134, 1977 (No. 2). *E*

U. Killat: Spatial frequency distribution of thermoplastic frost deformation.
Proc. 3rd Eur. Electro-Optics Conf., Geneva 1976 (SPIE Vol. 99), pp. 144-147; 1977. *H*

W. L. Konijnendijk & J. H. J. M. Buster: Raman-scattering measurements of silicate glasses containing sulphate.
J. non-cryst. Solids **23**, 401-418, 1977 (No. 3). *E*

G. Kowalski: Fast 3-D scanning systems using a limited tilting angle.
Appl. Optics **16**, 1686-1690, 1977 (No. 6). *H*

J. G. M. de Lau & A. Broese van Groenou: High-frequency properties of Ni-Zn-Co ferrites in relation to iron content and microstructure.
J. Physique **38**, C1/17-22, 1977 (Colloque C1). *E*

G. Laurence, B. A. Joyce, C. T. Foxon, A. P. Janssen*, G. S. Samuel* & J. A. Venables* (* University of Sussex, U.K.): Adsorption-desorption studies of Zn on GaAs.
Surface Sci. **68**, 190-203, 1977. *L, R*

R. E. van de Leest: Solid-state ion-selective electrodes for metal ions.
Analyst **102**, 509-514, 1977 (No. 1216). *E*

B. J. Mulder: 'Integration' of a 1000 W hydrogen discharge lamp and a vacuum monochromator.
J. Physics E **10**, 490-492, 1977 (No. 5). *E*

Th. G. J. van Oirschot, J. Hallais & J.-P. André: Les matériaux III-V pour photocathodes fonctionnant en transmission.
Acta Electronica **20**, 323-332, 1977 (No. 4). *E, L*

J. A. Pals: Microwave-enhanced critical currents in superconducting Al strips with local injection of electrons.
Physics Letters **61A**, 275-278, 1977 (No. 4). *E*

C. Piaget, J.-C. Richard, M. Monnier & P. Guittard: Les photocathodes à affinité électronique apparente négative.
Acta Electronica **20**, 333-352, 1977 (No. 4). *L*

G. Piétri: Introduction *(to issues on Low-light-level imaging)*.
Acta Electronica **20**, 225-230, 1977 (No. 3). *(In English and in French.)* *L*

A. Pirotte & P. Wodon: A comprehensive formal query language for a relational data base: FQL.
RAIRO Informatique **11**, 165-183, 1977 (No. 2). *B*

R. Polaert & J. Rodière: Tube intensificateur à galette de microcanaux pour cinématographie rapide.
Acta Electronica **20**, 379-385, 1977 (No. 4). *L*

H. Rau: The chromium-sulphur system between 873 K and 1364 K.
J. less-common Met. **55**, 205-211, 1977 (No. 2). *A*

J.-C. Richard, D. L. Lamport, E. Roaux & C. Vanneste (Université de Nice): Performances des systèmes de vision nocturne passive utilisant des tubes intensificateurs d'images. Influence de la réponse spectrale de la photocathode.
Acta Electronica **20**, 353-368, 1977 (No. 4). *L, R*

J. M. S. Schofield, J. Smith, D. G. Simmons & G. F. Weston: A d.c. plasma display panel with inherent storage.
Proc. 3rd Eur. Electro-Optics Conf., Geneva 1976 (SPIE Vol. 99), pp. 215-221; 1977. *R*

L. A. Æ. Sluyterman, J. Wijdenes & G. Voorn: Dimerization of papain induced by mercuric chloride and a bifunctional organic mercurial.
Eur. J. Biochem. **77**, 107-111, 1977 (No. 1). *E*

J. L. Sommerdijk, P. Vries & A. Bril: Line emission of $LiBaAlF_6:Eu^{2+}$.
Philips J. Res. **33**, 117-123, 1978 (No. 3/4). *E*

A. L. N. Stevels: Thermoluminescence of UV irradiated CsI:Na.
Philips J. Res. **33**, 133-138, 1978 (No. 3/4). *E*

B. Strocka, P. Holst & W. Tolksdorf: An empirical formula for the calculation of lattice constants of oxide garnets based on substituted yttrium- and gadolinium-iron garnets.
Philips J. Res. **33**, 186-202, 1978 (No. 3/4). *H*

A. W. Veenis & A. Bril: On the vacuum-ultraviolet excitation spectra and quantum efficiencies of luminescent powders, mainly rare-earth activated borates.
Philips J. Res. **33**, 124-132, 1978 (No. 3/4). *E*

J. Visser, B. Symersky & A. J. M. Geraerts: A versatile cryopump for industrial vacuum systems.
Vacuum **27**, 175-180, 1977 (No. 3). *E*

H. W. Werner: Applications of secondary ion mass spectrometry (SIMS).
Mikrochim. Acta, Suppl. VII, 63-83, 1977. *E*

Contents of Philips Telecommunication Review **36**, No. 3, 1978:

L. P. van den Toorn & M. J. Jordaan: COMSYS, a private packet switching network (pp. 149-164).
A. J. de Kleijn & R. P. Wiarda: TV field trial on a 60 MHz system (pp. 165-175).
H. L. Bakker & R. P. Wiarda: Allocation of six TV channels in the 60 MHz system (pp. 176-183).
D. Hek & J. van der Vegte: Design-400 goes coaxial (pp. 185-193).

Contents of Philips Telecommunication Review **36**, No. 4, 1978:

R. Bodart & H. C. van Rossum: Code and speed dependent TDM system 3TR 1600 for telegraph and data transmission (pp. 197-209).
C. W. M. Barrow, J. M. L. Hruschka & D. Schuitemaker: Introducing PRX in the world (pp. 210-216).
G. J. Kamerbeek, B. H. Radstake & S. Dijkhuizen: Modernizing the Saudi Arabian telephone network (pp. 217-223).
J. de Boer & C. Hooijkamp: The required load capacity of FDM multi-channel amplifiers (pp. 225-242).

Contents of Electronic Applications Bulletin **35**, No. 4, 1979:

Liquid crystal displays — part 2 (pp. 172-187).
A mains-isolated switched-mode power supply for TV receivers (pp. 188-210).
J. Jongsma: Minimum-loss transformer windings for ultrasonic frequencies. Part 2: Design methods and aids (pp. 211-226).

Contents of Mullard Technical Communications **14**, No. 139, 1978:

A. P. A. van Beekvelt, F. A. van Heukelom & G. Wolf: A switched-mode power supply with mains isolation for tv receivers (pp. 319-342).
'Plumbicon' camera tubes (pp. 343-345).
P. G. Noble: The safe operation of power transistors (pp. 346-375).
NORBIT 2 industrial control modules (p. 376).

Contents of Mullard Technical Communications **14**, No. 140, 1978:

A. Ciuciura: Systematic evaluation of the video drive requirements for colour tv picture tubes (pp. 379-391).
H. M. Farmer: A design for the display logic of an inexpensive VDU (pp. 392-401).
P. G. Giles, P. J. Hart & I. Thomas: Radio clock for reception and display of MSF time-code transmission (pp. 402-423).
L. E. Jansson: A sample-and-hold balancing and protection circuit for push-pull switched-mode power supplies (pp. 424-438).
Vision-i.f. ICs TDA2540 and TDA2541 (pp. 439-441).
Liquid crystal displays (pp. 442-444).

Contents of Electronic Components and Applications **1**, No. 1, 1978:

S. J. Op het Veld: Microprocessor-controlled video games (pp. 4-28).
J. J. A. Geboers, D. J. G. Jansen & J. C. Kaire: Two-tone telephone dialling (pp. 29-41).
UHF broadband amplifier for transposer bands IV and V (pp. 43-44).
P. G. Noble: Understanding rectifier diode data (pp. 45-55).
A. Franken: Improved performance 25 mm diameter 'Plumbicon' tube (pp. 57-59).
A.M./F.M. receiver IC (pp. 60-62).

PHILIPS TECHNICAL REVIEW

VOLUME 38, 1978/79 No. 10

The atom as a metallurgical building block

A. R. Miedema

For many metals the available experimental data relating to the latent heat of vaporization, surface tension, energy of dissociation of the diatomic molecule, heat of formation of monovacancies and other quantities of great importance in the practice of metallurgy are either incomplete or unreliable. The author proposes a simple model and from it he establishes relationships between different quantities that enable him to predict values that are not available. In this model the binding energy of the atoms in molecules or in the solid is interpreted as surface energy of the atoms that is released where they make contact. The usefulness of the model for making predictions is demonstrated by examples relating to diffusion rates, surface segregation and monolayers of mercury.

A metal atom: the smallest particle of metal

In the sciences of metallurgy and metal physics, atoms are sometimes pictured as the building blocks that determine the properties of the material. When we speak of crystal structures and crystal lattices we always tend to think of stacks of atoms with clearly defined dimensions. Many properties of metals, however, are described by the energy-band model, in which a solid is considered to be a collection of regularly stacked, positively charged ions surrounded by a sea of electrons. Such an approach is based on the collective behaviour of the valence electrons and the concept of the 'atom' has lost much of its significance.

In this article it will be shown that the atomic approach is extremely useful for making predictions, particularly for metals. Even the somewhat naive definition of the atom which says that 'an atom is the smallest particle of a substance that still contains all the properties of the substance' appears to be a good starting point. What follows may show that sticking literally to this definition can lead to far-reaching and sometimes surprising conclusions [1].

If a block of metal is sawn in two the energy of the metal is increased because the surface area of the metal is now larger. The increase in energy is equal to the product of the surface tension γ, i.e. the surface energy per unit area, and the increase in the area of the metal/vacuum interface, i.e. twice the new cross-section. If a number of small blocks are split off from a large block of metal (*fig. 1a*), the change in the surface area of the large block can be neglected; the increase in surface area is equal to the combined surface area of the small blocks.

When the atom is defined in the simple way stated above this approach leads to the conclusion that the *molar latent heat of vaporization* of a metal, ΔH^{v}, must be equal to $N_0\gamma$ times the surface area of a single atom; N_0 is Avogadro's number. To an atom in the form of a cube we must assign an edge $(V_{\mathrm{m}}/N_0)^{\frac{1}{3}}$, where V_{m} is the molar volume of the solid. The surface area is therefore $6(V_{\mathrm{m}}/N_0)^{\frac{2}{3}}$. For ΔH^{v} we thus have

$$\Delta H^{\mathrm{v}} = a_1 \gamma V_{\mathrm{m}}^{\frac{2}{3}}, \qquad (1)$$

where $a_1 = 6N_0^{\frac{1}{3}} = 5.1 \times 10^8$.

Dr A. R. Miedema is with Philips Research Laboratories, Eindhoven.

[1] Background information about the many experimental data used in this article is given in:
a) A. R. Miedema and R. Boom, Z. Metallk. **69**, 183, 1978 (liquid).
b) A. R. Miedema, Z. Metallk. **69**, 287, 1978 (solid).
c) A. R. Miedema, Z. Metallk. **69**, 455, 1978 (surface segregation).
d) A. R. Miedema and F. J. A. den Broeder, Z. Metallk. **70**, 14, 1979 (interface energy).
e) A. R. Miedema, Z. Metallk. **70**, 345, 1979 (vacancies).

Fig. 2 shows that, with four clear exceptions, a large group of metals obey this relation. In this figure the experimental values of ΔH^V are plotted against $\gamma V_m^{\frac{2}{3}}$ for the solid state at zero temperature. The slope of the straight line gives 5.2×10^8 for the proportionality constant a_1. The agreement with the calculated value is striking, but, as we shall see later, to some extent accidental. We shall return shortly to the four exceptions: Mg, Zn, Cd and Hg.

Solid-state reactions are inconceivable without diffusion, and for diffusion it is essential to have vacancies. The *enthalpy of formation of a monovacancy* in a pure metal, ΔH_{vac}, is therefore an important quantity in practice. A monovacancy is a cavity in the metal equal in size to a single atom (fig. 1*b*). If we attribute the (molar) enthalpy of formation to the surface energy of N_0 cavities then we find:

$$\Delta H_{vac} = a_2 \gamma V_m^{\frac{2}{3}}, \qquad (2)$$

where a_2 should be equal to a_1.

Fig. 3 shows that condition (2) is also satisfied, except for tungsten. The value of the proportionality constant a_2 is considerably smaller than that of a_1. This may be explained by the fact that the walls of the cavity are not completely free surfaces; there is some interaction between opposite walls. We shall return to this point later. In fig. 3 the transition metals as a group behave somewhat differently from the non-transition metals. The slopes of the solid straight lines give values of:

$a_2 = 1.17 \times 10^8$ for the non-transition metals,

$a_2 = 1.42 \times 10^8$ for the transition metals.

A third conclusion arrived at by considering a metal atom A as a 'fragment of metal with surface energy' relates to the dissociation energy D_{AA} of the molecule A_2 (fig. 1*c*). On dissociation, the interface between the two atoms changes into two interfaces with vacuum so that the surface energy increases. If we assume that the area of contact is the same fraction of the total atomic surface area for all metals, then D_{AA} must also be proportional to $\gamma V_m^{\frac{2}{3}}$:

$$D_{AA} = a_3 \gamma V_m^{\frac{2}{3}}. \qquad (3)$$

Fig. 4 shows that for the metals in fig. 2 for which an experimental value of D_{AA} is known [2], there is indeed a correlation between D_{AA} and $\gamma V_m^{\frac{2}{3}}$, which is described reasonably well by (3). Once again, the elements Mg, Zn, Cd and Hg are clear exceptions. The large spread of the other points about the solid line can probably be largely explained by the fact that there is a considerable uncertainty in some of the experimental values of D_{AA}. The straight line gives $a_3 = 2.6 \times 10^8$.

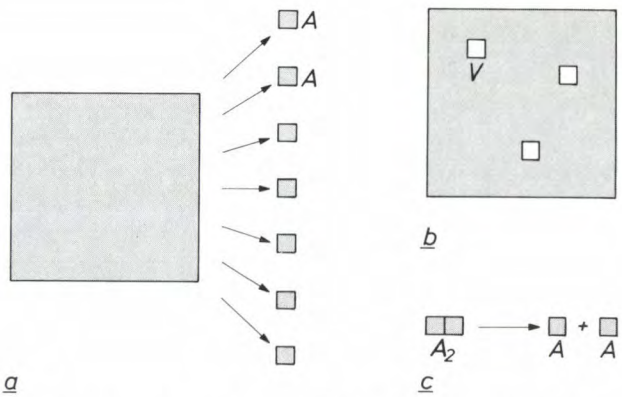


Fig. 1. The atom as a building block. *a*) A large number of small blocks of metal the size of an atom (A) split off from a large block of metal (vaporization). The energy required, i.e. the latent heat of vaporization, is equal to the combined surface energy of the atoms. The change in the surface area of the large block can be neglected. *b*) The energy of a monovacancy (V) can be taken to be the surface energy of a cavity of the same size as an atom. *c*) When the molecule A_2 dissociates into two A atoms the total surface energy increases; the dissociation energy is taken to be the added surface energy.

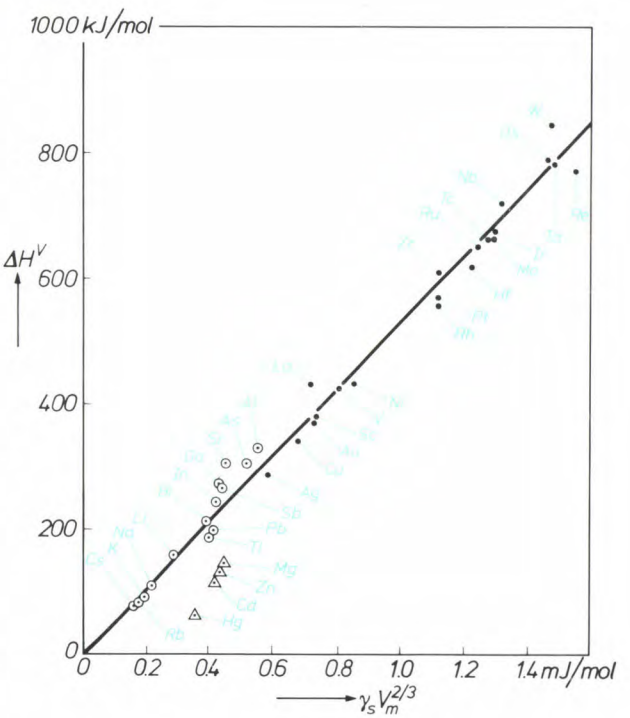


Fig. 2. The latent heat of vaporization ΔH^V plotted against the product $\gamma_s V_m^{\frac{2}{3}}$ for a large number of metals; γ_s is the surface tension and V_m the molar volume of the solid, both values reduced to zero temperature. The straight line represents relation (1) with $a_1 = 5.2 \times 10^8$. ○ non-transition metals. • transition metals, △ the divalent metals Mg, Zn, Cd and Hg. In this article Cu, Ag and Au are included in the transition metals.

We have seen how by considering a metal atom as a fragment of metal it is possible to predict simple relationships that are in fact found to exist for a large number of metals. In such cases it is the exceptions

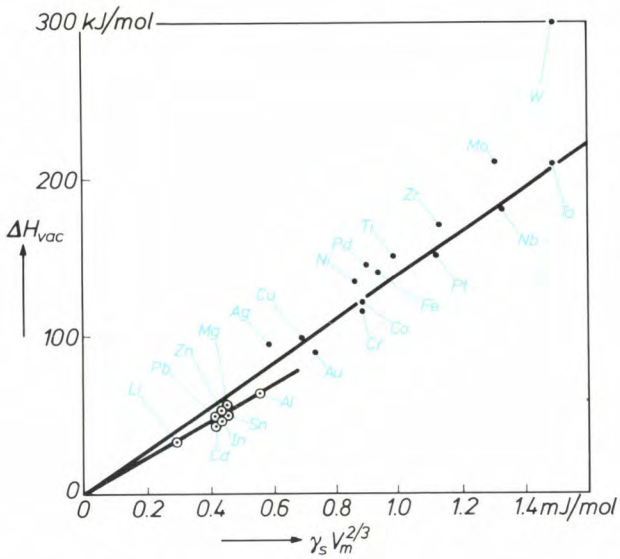


Fig. 3. The molar enthalpy of formation of monovacancies, ΔH_{vac}, plotted against $\gamma_s V_m^{\frac{2}{3}}$ for a number of metals. The straight lines represent equation (2) with $a_2 = 1.17 \times 10^8$ (non-transition metals, ○) and $a_2 = 1.42 \times 10^8$ (transition metals, •).

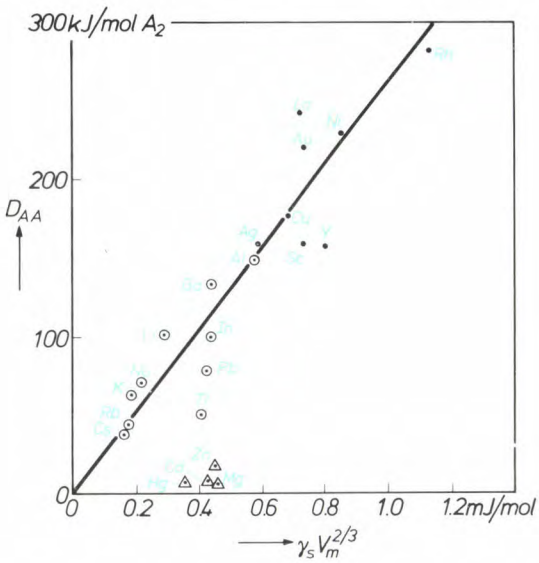


Fig. 4. The dissociation energy D_{AA} of the molecule A_2 as a function of $\gamma_s V_m^{\frac{2}{3}}$ for the metals for which a value of D_{AA} is known. The symbols have the same significance as in fig. 2. The straight line represents equation (3) with $a_3 = 2.6 \times 10^8$.

that become interesting; attention will be given to these in the next section.

An explanation will then be given of the derivation of the quantity γ_s in figs 2, 3 and 4, i.e. *the surface tension of the solid at zero temperature* — not generally a very well known quantity — from the much better known surface tension of the liquid metal at the melting point, γ_{l,T_f}.

The quantity γ_s features prominently in this article. It will be shown that it is closely related to the electron density n_{WS} at the interface between two neighbouring atoms in the crystal. This approach also enables the values that have been found for a_1, a_2 and a_3 to be interpreted in greater detail.

In metallurgy the energy quantities γ_s and ΔH_{vac} discussed above have considerable practical importance, but the experimental data available are far from complete and are sometimes very unreliable. Conversion of the surface tension of the liquid to that of the solid at zero temperature, together with the above relationships, enables a list of these quantities and of D_{AA} to be made for all metallic elements, even for those elements for which no experimental data are available (Table I, p. 262).

The final section will deal with some examples of a less general nature that will clearly demonstrate how useful the atomic model is for making predictions.

Ordinary and special metals

In concise terms 'metals' can be defined as elements that can exist in the solid or the liquid state at room temperature and are good electrical conductors. In the periodic system of the elements in *fig. 5*, from which the seventh group has been omitted, the elements we have examined are indicated by colours. The elements Be, B and C are borderline cases: while they are in many respects metallic they have been left out because the necessary experimental data (in particular γ_{l,T_f}) are not available. On the other hand Si and Ge, which are semiconductors at room temperature, are counted as metals here because they are good conductors above their melting point. In fig. 5 the metals have been subdivided into a number of groups: V_1 to V_5, i.e. metals with valencies of 1, 2, 3, 4 and 5; T_3, T_4 and T_5, the transition metals in which, generally speaking, the 3d, 4d and 5d shells are filled beneath the 4s, 5s and 6s shells respectively; and R, the rare-earth metals in which the 4f shell is filled beneath the 6s shell.

The divergent behaviour in fig. 2 of Mg, Zn, Cd and Hg (the right-hand branch of V_2 in fig. 5) can readily be associated with their atomic electron configuration. This is characterized by an s^2 outer shell ($3s^2$, $4s^2$, $5s^2$ and $6s^2$) with a closed shell ($2p^6$, $3d^{10}$, $4d^{10}$ and $5d^{10}$) inside it. Such a closed, inert-gas-like configuration would be expected to have a high stability. This is confirmed by *fig. 6*: the first ionization potential is particularly high compared with that of the neighbours in the same row of the periodic system. Moreover, the next element (Al, Ga, In, Tl) might be expected to be easily converted to such a configuration and thus easily

[2] A more detailed examination and a summary of the experimental values of the dissociation energy is given in A. R. Miedema and K. A. Gingerich, J. Physics B **12**, 2081, 1979.

ionized; this is also clearly shown to be the case in fig. 6. There is, however, no question of an s^2 configuration and an extra stability *in the solid*: the valence electrons have become conduction electrons of a mixed s, p and d nature; the atoms are 'metallically bonded'. The boiling point (T_b) (Trouton's rule: $\Delta H^V/T_b = 100$ J/mol K).

All this gives us the following picture of 'ordinary' and 'special' metals. In both cases the atoms are 'metallically bonded' in the interior of a piece of solid mater-

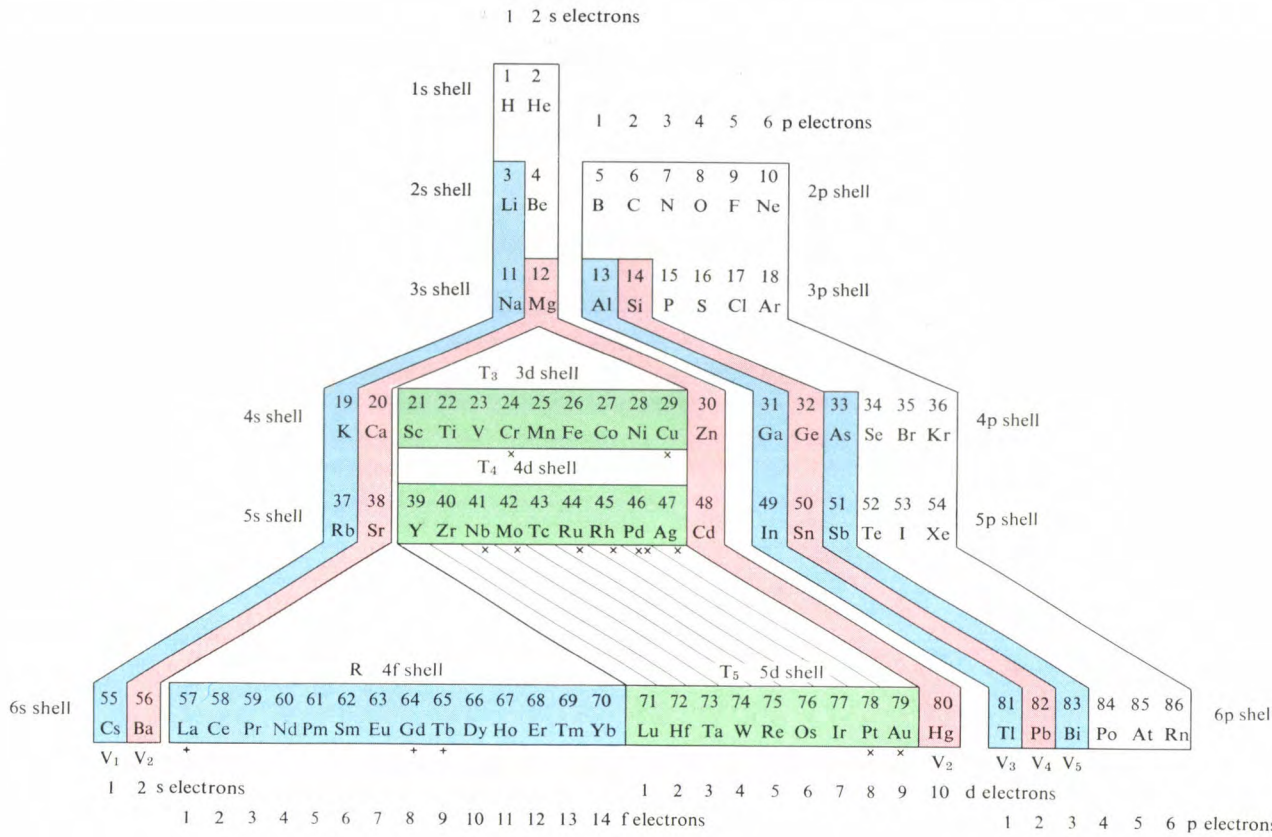


Fig. 5. The metals considered in this article (*coloured*) in the periodic system of the elements; the seventh group has been left out. Si and Ge are included in the metals, B and C are not. The metal Be has not been considered. The metals fall into groups: the metals with valencies 1, 2, ... (V_1 to V_5), the 3d, 4d and 5d transition metals (T_3, T_4, T_5) and the rare earths (R). The atomic electron configuration of Mg, Zn, Cd and Hg — a full s outer shell (s^2) with only full shells inside it — makes the atomic state of these elements particularly stable, like that of inert gases; this explains the unusual behaviour of these elements in figs 2 and 4.

The atomic electron configurations can be easily read off from the system. Atom S (No. 16), for example, has all the shells up to 3s (i.e. 1s, 2s, 2p and 3s) completely filled and four electrons in the 3p shell ($3s^23p^4$); Zr (No. 40) has all the shells filled up to 5s and two electrons in the 4d shell, ($4d^25s^2$). 'Errors in nature' are indicated by crosses:

× one s electron is replaced by one d electron,
×× two s electrons are replaced by two d electrons,
+ one f electron is replaced by one d electron.

Thus the configuration of Pt (No. 78), for example, is not ... $4f^{14}5d^86s^2$ but ... $4f^{14}5d^96s$.

surface energy will therefore be 'ordinary' but the latent heat of vaporization will be relatively small because of the lower energy of the atomic state. This explains the divergent behaviour in fig. 2. Quantitatively the deviations in fig. 2 (about 100 kJ/mol) are in fact in fair agreement with those in fig. 6 (about 1.5 eV per atom, which corresponds to about 140 kJ/mol). The low energy of the atomic state also explains the low boiling point of Hg, Cd and Zn; a low latent heat of vaporization is always associated with a low

ial; the absence of such bonds at the surface is identical with surface energy. The free atoms of ordinary metals can still be considered as being largely metallic; the complete absence of bonds can be interpreted as surface energy. The free atoms of special metals, however, are like those of an inert gas; there is no longer a relationship with the surface energy of the metal.

The divergence in the energy of dissociation of the 'dimers' Mg_2, Zn_2, Cd_2 and Hg_2 (fig. 4) can also be attributed to the inert-gas nature of the atoms, as we

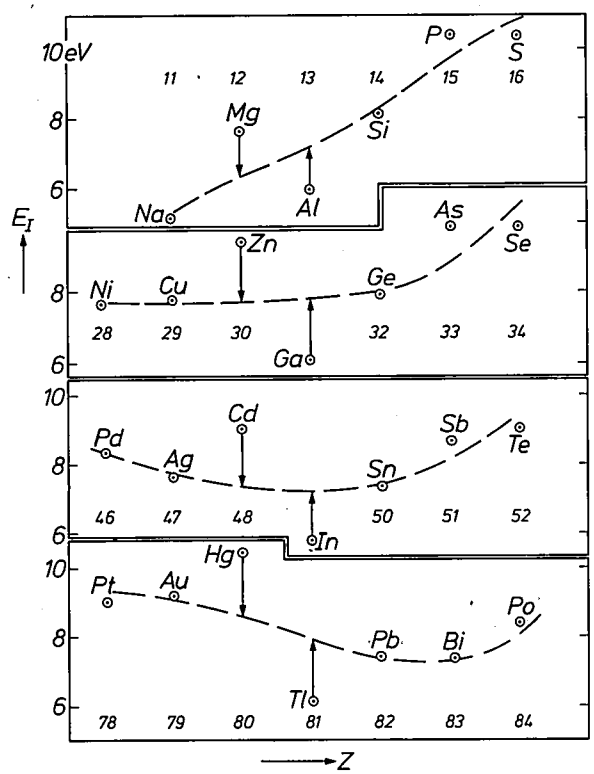


Fig. 6. The first ionization potential E_I as a function of the atomic number in the region around Mg, Zn, Cd and Hg. The particular stability of the atomic electron configuration of these elements means that E_I is relatively high for these elements. E_I is relatively low for the following elements since they change to this configuration on ionization.

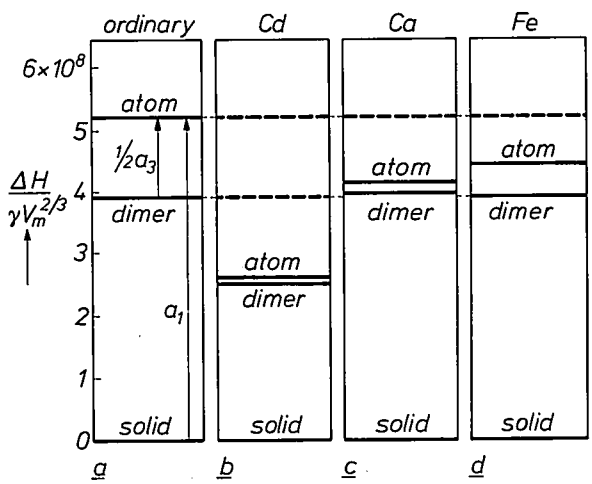


Fig. 7. The energy ΔH per gram atom at $T = 0$, divided by $\gamma_s V_m^{\frac{2}{3}}$, for free atoms and dimers as compared with the solid. *a*) 'Ordinary' metals. *b*) Cd as a representative of the group Mg, Cd, Zn and Hg. *c*) Ca, which also represents Sr and Ba. *d*) Fe, which also represents Ti, V, Cr, Mn, Co and Pd. In (*b*), atom and dimer are both specially stable, in (*c*) and (*d*) the dimer is approximately ordinary, but the free atom is more stable than the ordinary one.

shall see from *fig. 7*. Here the energy per mole of atoms at $T = 0$, divided by $\gamma_s V_m^{\frac{2}{3}}$, has been plotted for free atoms and dimers, with the energy of the solid taken as the zero. For ordinary metals (fig. 7*a*), as represented by the straight lines in fig. 2 and 4, the free-atom level is at a_1 ($= 5.2\times10^8$) whereas the difference between atom and dimer levels is $\frac{1}{2}a_3$ ($= 1.3\times10^8$). The levels in fig. 7*b* for Cd are also representative of Mg, Zn and Hg. The level of the inert-gas-like free atom is much lower than in the ordinary case. It is in fact much lower than the level of the dimer that consists of two atoms with a metallic bond, so that such a dimer will not be formed. In the Cd_2 dimer, which has a somewhat lower energy than two free atoms, the atoms cannot therefore have an ordinary metallic bond. Just as with inert gases, the two atoms with their full-shell configurations can only enter into a weak association by mutual polarization (Van der Waals forces). There is in fact a resemblance to inert-gas molecules such as Xe_2 or Kr_2 in the optical spectra (vibration levels).

It should be said here that in figs 2 and 4 the situation has been deliberately simplified a little. In addition to the ordinary metals four clear exceptions have been included, but a number of less clear-cut exceptions have been left out. These are the alkaline-earth metals Ca, Sr and Ba (the left-hand branch of V_2 in fig. 5), the 3d elements Ti, V, Cr, Mn, Fe and Co (from T_3) and the 4d element Pd (from T_4).

It is obvious that Ca, Sr and Ba ought to be exceptions: they are also divalent and have an s^2 outer shell with a closed shell inside it (a p^6 shell). These metals are represented by Ca in fig. 7*c*. As would be expected, the free-atom level is again lower than in the ordinary case but not as low as in the case of Cd. The dissociation energy of the dimer is so small that the associated level is almost at the ordinary height. This means that we can no longer tell from energy considerations whether Ca_2, Sr_2 and Ba_2 are 'inert-gas-like' or 'metallic' dimers, and there are no other experimental data available that give a definite answer to this. The other special metals are represented in fig. 7*d* by Fe. As a result of magnetic effects, which are much more important in the free atom than in the solid, the atomic level here is also lower than 'ordinary' (but not as low as in Ca); the level of the dimer is again more or less 'ordinary'.

Fig. 7 would seem to indicate that many 'special' metals (those of figs 7*c* and *d*) are 'ordinary' in their relationship between solid and dimer; it is only the free atom that is different.

The surface tension of the solid at zero temperature

For pure metals in the crystalline state it is usual to assign to each lattice point an atomic cell (a 'Wigner-Seitz' cell) whose shape is such that all the cells together fill the entire space. In our model free 'metallic' atoms are separate Wigner-Seitz cells. Only in the simplest

Table I. Table of values for characteristic quantities of the metallic elements. T_f is the melting point, γ_{l,T_f} the surface tension of the liquid at the melting point, γ_s the surface tension of the solid at zero temperature, ΔH_{vac} the enthalpy of formation of monovacancies and D_{AA} the dissociation energy of dimers. The values of T_f are taken from the book by R. Hultgren *et al.* [3], those of γ_{l,T_f} from the Handbook of Chemistry and Physics [4] apart from some values obtained later [1a]. The values for γ_s and ΔH_{vac} were derived as described in the article (pages 262/263 and 265/266); the values of D_{AA} are experimental [2] except for those in brackets, which were obtained in the way described on page 265. The elements are subdivided into the groups T_3, . . ., V_1, . . . of fig. 5. The rare earths are represented by La.

	Element	T_f (K)	γ_{l,T_f} (J/m²)	γ_s (J/m²)	ΔH_{vac} (kJ/mol)	D_{AA} (kJ/mol A_2)	$\frac{\Delta H_{vac}}{RT_f}$
T_3	Sc 21	1812	(0.87)	1.20	104	158	6.9
	Ti 22	1943	1.50	2.05	140	125	8.7
	V 23	2175	1.90	2.60	151	238	8.4
	Cr 24	2130	1.70	2.40	126	151	7.1
	Mn 25	1517	1.10	1.60	86	42	6.8
	Fe 26	1809	1.83	2.55	134	100	8.9
	Co 27	1768	1.83	2.55	127	167	8.6
	Ni 28	1726	1.75	2.45	122	229	8.5
	Cu 29	1357	1.31	1.85	97	196	8.6
T_4	Y 39	1799	(0.80)	1.10	114	156	7.6
	Zr 40	2125	1.43	1.95	160	(294)	9.1
	Nb 41	2740	1.96	2.70	187	(344)	8.2
	Mo 42	2890	2.13	2.95	185	(337)	7.7
	Tc 43	2473	2.35	3.05	182	(333)	8.9
	Ru 44	2523	2.25	3.05	177	(325)	8.4
	Rh 45	2233	1.97	2.75	160	281	8.6
	Pd 46	1825	1.48	2.10	128	104	8.4
	Ag 47	1234	0.91	1.25	84	158	8.1
$R+T_5$	La 57	1193	0.72	0.90	102	241	10.3
	Hf 72	2500	1.60	2.20	175	(320)	8.4
	Ta 73	3287	2.18	3.05	212	(389)	7.8
	W 74	3680	2.34	3.30	210	(386)	6.9
	Re 75	3453	2.65	3.65	223	(408)	7.8
	Os 76	3300	2.50	3.50	209	(382)	7.6
	Ir 77	2716	2.25	3.10	185	(338)	8.2
	Pt 78	2042	1.86	2.55	159	(292)	9.4
	Au 79	1336	1.13	1.55	103	220	9.3
V_1	Li 3	454	0.41	0.53	34	101	9.0
	Na 11	371	0.20	0.26	25	71	8.1
	K 19	336	0.11	0.15	22	53	7.9
	Rb 37	313	0.085	0.12	20	45	7.7
	Cs 55	302	0.07	0.095	19	38	7.6
V_2	Mg 12	922	0.57	0.79	54	5	7.0
	Zn 30	693	0.77	1.02	53	18.5	9.2
	Cd 48	594	0.59	0.78	50	8.5	10.1
	Hg 80	234	0.485	0.61	41	7.5	21
	Ca 20	1112	0.35	0.49	50	11	5.4
	Sr 38	1041	0.295	0.43	51		5.9
	Ba 56	1002	0.255	0.37	49		5.9
V_3	Al 13	933	0.865	1.20	65	144	8.4
	Ga 31	303	0.715	0.83	50	133	20
	In 49	430	0.56	0.69	51	100	14.3
	Tl 81	577	0.465	0.61	47	50	9.8
V_4	Si[*]14	1685	0.80	1.29	63	310	4.5
	Ge[*]32	1210	0.64	1.03	55	272	5.5
	Sn 50	505	0.57	0.71	53	185	12.6
	Pb 82	601	0.46	0.61	49	79	9.8
V_5	As[*]33	(886)	(0.59)	(1.0)	61	378	8.3
	Sb[*]51	904	0.39	(0.68)	53	299	7.1
	Bi[*]83	595	0.38	(0.55)	46	196	10.2

[*] The values for γ_s and ΔH_{vac} apply to these metals in a normal close-packed crystal structure.

case are they cubes; generally speaking they are polyhedra somewhere between cubes and spheres. The latent heat of vaporization is the product of the combined surface area of the atoms and the 'atomic surface tension'. If we wish to relate this atomic surface tension to *the* surface tension of a large volume of the metal, then we have to take the surface tension of the solid at zero temperature, because this is an inherent property that is independent of temperature effects. This quantity, γ_s, plays an important part in many other problems, such as the interface energy between solid and molten material or the adhesion between different metals, and it is therefore of more general interest to have a list of γ_s values for all metals. As has already been stated, direct experimental data here are far from complete and often inaccurate. A complete and what is thought to be a reasonably reliable list is given in *Table I.* The γ_s values in this table have been derived from γ_{l,T_f}, the surface tension of the liquid at melting point, for which there *is* a reliable and almost complete list of experimental values. The derivation, which will now be explained in greater detail, consists of a temperature correction and a roughness correction.

The surface tension is an atomic energy per unit area. It is therefore dependent on the temperature because the number of surface atoms per unit area varies during expansion. This number is inversely proportional to $V_m^{\frac{2}{3}}$. We can correct for this effect by considering the energy for a fixed number of surface atoms, $\gamma V_m^{\frac{2}{3}}$. This quantity, however, is still dependent on the temperature, because an atom at the surface has more degrees of freedom than other atoms and thus represents extra *entropy*. In the first instance we would expect this effect to be the same for all metals. If we therefore write

$$(\gamma V_m^{\frac{2}{3}})_T = (\gamma V_m^{\frac{2}{3}})_0 + bT, \qquad (4)$$

where b is a measure of the surface entropy per atom, then b would have to be the same for all metals. Differentiating (4) with respect to temperature:

$$b/\gamma V_m^{\frac{2}{3}} = (1/\gamma)\mathrm{d}\gamma/\mathrm{d}T + \tfrac{2}{3}\alpha, \qquad (5)$$

where $\alpha(= (1/V)\mathrm{d}V/\mathrm{d}T)$ is the (volume) expansion

coefficient. Where experimental data are available for γ and $d\gamma/dT$, equation (5), together with the known values for α, does provide a reasonably constant value for b, of about -0.5×10^{-7} J/K. By putting this value in equation (4) the value of $\gamma V_m^{\frac{2}{3}}$ for the liquid state at $T = 0$ has been derived from the value at the melting point for all the metals listed in Table I.

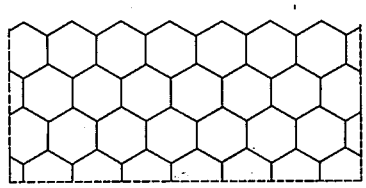


Fig. 8. Two-dimensional model of a crystal with hexagonal Wigner-Seitz cells (atoms). The model is fairly realistic in the sense that it gives values both for the roughness factor (1.15) and for the ratio of the area of the outer faces to the total surface area of a surface atom ($\frac{1}{3}$) that are approximately the same as those that follow from the three-dimensional Wigner-Seitz models of several crystal structures. The roughness factor is the ratio of the 'true' surface area of the crystal to the macroscopically measured area.

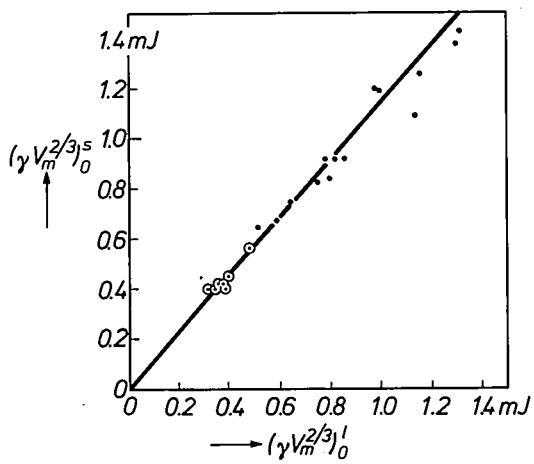


Fig. 9. Comparison of the product $\gamma V_m^{\frac{2}{3}}$ at $T = 0$ for the solid (s) and the liquid (l) states for those elements for which experimental values of γ_s are also known. The gradient of the straight line (1.13) is the mean ratio of the roughness factors.

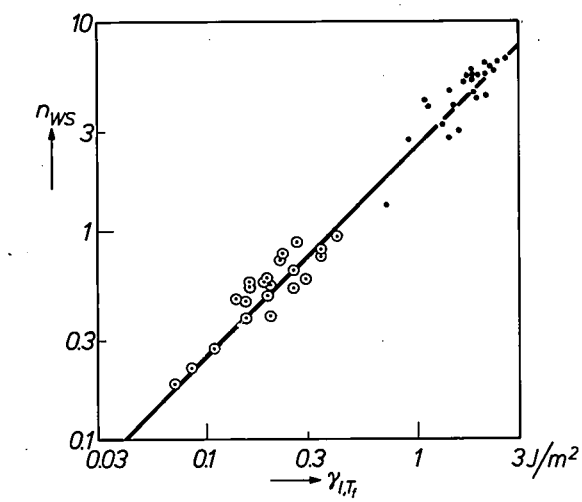


Fig. 10. Relationship between the electron density at the boundary of an atomic cell, n_{WS}, and the surface tension of the liquid at the melting point, γ_{l,T_r}. ○ non-transition metals. • transition metals [5].

On an atomic scale the metal/vacuum interface of a crystal is rougher than that of the liquid. The surface area is larger than the macroscopically measured surface area; the surface energy per macroscopically measured unit of surface area, i.e. the surface tension, is therefore also proportionately greater. The associated 'roughness factor' is dependent on the shape of the atom. *Fig. 8* shows a two-dimensional crystal with hexagonal cells, a shape that is intermediate between the two extremes of cube and sphere. In this case the roughness factor would be 1.15. For other shapes factors that differ from this by 5 to 10% are found. For a crystal, a value of $\gamma V_m^{\frac{2}{3}}$ about 1.15 times larger than that for a liquid would therefore be expected. This is confirmed by *fig. 9*. Here the values of $\gamma V_m^{\frac{2}{3}}$ at $T = 0$ for the solid and the liquid have been plotted against one another for those elements for which a reliable value is known for the solid. The straight line gives a roughness factor of 1.13. This value has been used to derive the values of γ_s given in Table I. We should also note that the above argument suggests that the surface tensions for different crystal planes of the same crystal could differ by 5 to 10%. Such differences have in fact been observed.

Surface tension and electron distribution

In liquid or solid metal the charge distributions of neighbouring atoms overlap. The electron density at the boundary of a Wigner-Seitz cell, n_{WS}, is a measure of this overlap [5]. If a new surface is formed by making a section between two atomic layers, then if the charge distribution were to remain unchanged there would be a 'jump' of n_{WS} in the electron density at the surface. Although such a jump is not sustained — the distribution is smoothed out — it is nevertheless clear that the surface tension and the initial jump are interrelated. *Fig. 10* shows that this is most definitely the case; this figure shows a log-log plot of n_{WS} as a function of γ_{l,T_r}. The straight line in fig. 10 has a slope of 1. The quantity n_{WS}, therefore, is in fact approximately proportional to γ_{l,T_r}.

[3] R. Hultgren, P. D. Desai, D. T. Hawkins, M. Gleiser and K. K. Kelley, Selected values of the thermodynamic properties of binary alloys, Amer. Soc. for Metals, Metals Park, Ohio, 1973.

[4] G. Lang, in Handbook of Chemistry and Physics, CRC Press, Cleveland, 57th edition (1977), pp. F 25, B 257.

[5] The quantity n_{WS} has been determined empirically for transition metals by means of thermodynamic data; for the other metals it has been estimated theoretically; see A. R. Miedema, R. Boom and F. R. de Boer, J. less-common Met. **41**, 283, 1975, and **46**, 67, 1976.

Closer analysis shows that the spread of the points about the straight line in fig. 10 is not random: a log-log plot of the ratio $\gamma_{l,T_f}/n_{WS}$ against the number of valence electrons z gives a pattern approximated by two lines with a slope of $-\frac{1}{3}$, one for the transition metals and one for the non-transition metals. We therefore have the approximate relation

$$\gamma_{l,T_f}/n_{WS} = Pz^{-\frac{1}{3}}, \quad (6)$$

where P clearly differs for transition metals and non-transition metals. In *fig. 11* $\gamma_{l,T_f}/n_{WS}$ is plotted as a function of $z^{-\frac{1}{3}}$ with linear scales. In figs 10 and 11 the directly measured γ_{l,T_f} has been used, although $\gamma(T = 0)$ would have been preferable. The difference is small, however.

The explanation for equation (6) must lie in the stronger spatial modulation of the electron distribution for higher values of z. This is because the screening of positive charges in an electron gas increases with the mean electron density $\bar{n}$ (the 'screening length' l_s is proportional to $\bar{n}^{-\frac{1}{6}}$). In *fig. 12* a comparison is given of two cases with equal values of n_{WS} but different values of z (and hence different values of $\bar{n}$). If we consider the electron density to be a superposition of atomic electron distributions, then we can see in fig. 12 that for small values of z the overlap is much greater than for large values of z. Because of this, the separation of neighbouring atoms will be more difficult even though the values of n_{WS} are equal, and γ will therefore be larger in accordance with equation (6).

The values of a_1 and a_3; relaxation

As stated earlier, the striking agreement between the experimental value of a_1 and the estimated value on page 257 is a chance occurrence. This becomes clear if we examine a_1 more closely.

In our model the molar latent heat of vaporization is given by the equation

$$\Delta H^V = N_0 \gamma_a S_a, \quad (7)$$

where γ_a is the (hypothetical) atomic surface tension and S_a the surface area of the atom. If we take equation (1) as defining a_1 then from equation (7) we find:

$$a_1 = N_0(\gamma_a/\gamma_s)\, S_a/V_m^{\frac{2}{3}}. \quad (8)$$

For a given atomic volume V_a, the area S_a is dependent on the shape of the atom:

$$S_a = f_s V_a^{\frac{2}{3}}.$$

It can easily be verified that the 'shape factor' f_s is equal to 6 for a cube, and $(36\pi)^{\frac{1}{3}}$ i.e. about 4.8 for a sphere. It is assumed that the average ($f_s = 5.4$) applies for most atoms. With $V_a = V_m/N_0$ we find

$$a_1 = f_s N_0^{\frac{1}{3}}(\gamma_a/\gamma_s).$$

Identifying γ_s with γ_a *(fig. 13a)* gives the value of 5.1×10^8, mentioned earlier, for cubes, but a value of 4.5×10^8 for polyhedra with $f_s = 5.4$. If the roughness of the solid surface is taken into account the deviation from the experimental value (5.2×10^8) becomes even greater; with $\gamma_s/\gamma_a = 1.15$ (see fig. 13*b*) we find $a_1 = 3.9 \times 10^8$.

The explanation for the deviation must lie in the fact that the surface tension at an external face of a surface atom, γ_{sa}, and the surface tension of a free atom, γ_a, are not equal (fig. 13*c*). A charge redistribution has taken place in the surface atom; the charge has moved from the energetically unfavourable, free atomic planes to the energetically more favourable internal atomic planes so that γ_{sa} is smaller than γ_a. This effect is known as relaxation of the surface, and we speak of the transfer of 'bonds' of a surface atom from the neighbours that have been removed by the

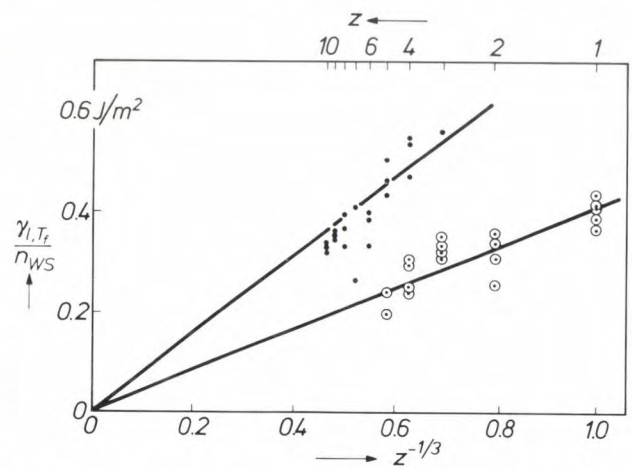


Fig. 11. The ratio $\gamma_{l,T_f}/n_{WS}$ as a function of $z^{-\frac{1}{3}}$ where z is the number of valence electrons. For the non-transition metals (○) z varies from 1 to 5 and for the transition metals (●) it varies from 3 to 10. Valence electrons are the electrons outside a completely closed shell system; for the 3d metals, for example, they are the two 4s electrons and the electrons in the 3d shell.

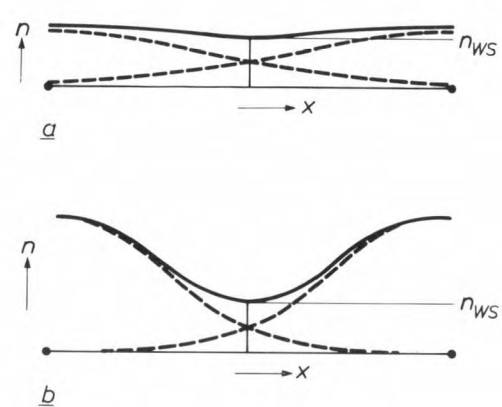


Fig. 12. Electron density n as a function of position x. The dots indicate the centres of the atoms. *a*) Few valence electrons, practically uniform charge distribution. *b*) Many valence electrons, the charge distribution is strongly modulated in space. When the density is the same at the boundary between the two atoms (n_{WS}) the overlap of the two atomic electron distributions (*dashed*) is much more marked in case *a* than in case *b*.

splitting to the neighbours that remain. For the situation in fig. 13*c* we find

$$a_1 = f_s N_0^{\frac{1}{3}}(\gamma_a/\gamma_{sa})(\gamma_{sa}/\gamma_s) = 3.9 \times 10^8\,(\gamma_a/\gamma_{sa}).$$

This would mean that for an 'ordinary' metal $\gamma_{sa}/\gamma_a = 3.9/5.2 = 0.75$. It can also be said that the outer surface has been effectively reduced by a factor of 0.75. The area of the external faces of a surface atom can be estimated from geometrical models (e.g. the one in fig. 8) at about a third of its total surface area. With the relaxation correction an effective value of about $\frac{1}{4}$ ($= 0.75 \times \frac{1}{3}$) is found for this ratio.

Relaxation probably occurs in diatomic molecules as well. If two cubic atoms are placed with lateral faces against each other or two hexagonal (two-dimensional) cells with edges against each other, then the area of the contact face is $\frac{1}{6}$ of the surface area of an atom. The dissociation energy, calculated per atom, should therefore be $\frac{1}{6}$ of the 'total' energy per atom (i.e. the latent heat of vaporization) and $\frac{1}{2}a_3$ should be equal to $\frac{1}{6}a_1$. For 'ordinary' atoms (see fig. 7), however, $\frac{1}{2}a_3 \approx \frac{1}{4}a_1$. Apparently the effective area of the contact face is $\frac{1}{4}$ of the surface area of an atom. This can be interpreted as a transfer of charge towards the contact face (*fig. 14a*) or as an effective increase in area of the contact face (fig. 14*b*).

The dissociation energy of the elements for which no reliable experimental value is available was calculated from equation (3), with the values of γ_s found earlier and taking $a_3 = 2.6 \times 10^8$. The result is given in Table I.

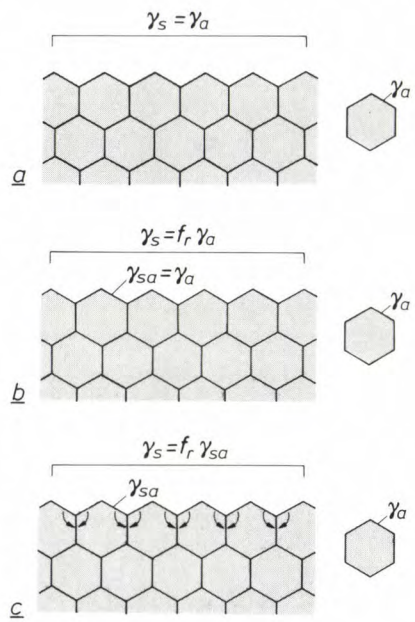


Fig. 13. Surface roughness and relaxation. γ_s is the macroscopically measured surface tension, γ_a the hypothetical surface tension of a free atom, γ_{sa} that of an external face of a surface atom. *a*) If the roughness and relaxation corrections are neglected, γ_s is equal to γ_a. *b*) The roughness correction alone gives $\gamma_s = f_r\gamma_a$ ($f_r = 1.15$, see fig. 8). *c*) γ_{sa} is smaller than γ_a because of charge transfer from the energetically unfavourable external faces to the internal faces ('relaxation', 'bond transfer'). Analysis of the results gives $\gamma_{sa} \approx 0.75\,\gamma_a$. The 'effective' area of the external faces of a surface atom is therefore not $\frac{1}{3}$ but only $0.75 \times \frac{1}{3}$, i.e. a quarter, of the total surface area of the atom.

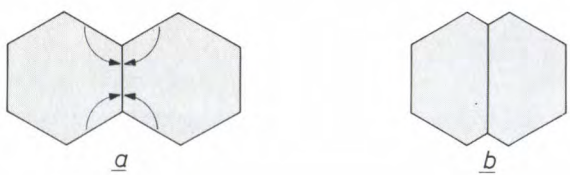


Fig. 14. In dimers the 'effective' area of the contact face between the two atoms is not $\frac{1}{6}$ but about $\frac{1}{4}$ of the total surface area of an atom (because of relaxation). This may be interpreted as charge transfer (*a*) or as deformation (*b*).

The value of a_2

In the simplest form of our model the surface area of a monovacancy is equal to the surface area of a free atom, and the surface tensions are also equal. This would make $a_2 = a_1$. In view of the discussion above, relaxation (charge transfer or deformation and shift of the neighbouring atoms, resulting in a smaller cavity surface area, *fig. 15*) might explain a difference of the order of 20%. However, a_2 is in fact three times as small as a_1. As has already been stated, this can be explained by the interaction that still exists between the walls of the cavity. The electron density at the atom surface, which is n_{WS} in a solid, does not have to match that in vacuum ($n = 0$) (*fig. 16a*), but only the density (n_v) in the middle of the cavity, which is not zero (fig. 16*b*). An estimate of n_v — obtained by superposition of the charge distributions in free atoms located at the lattice points around the cavity — gives much the same result for many non-transition atoms: $n_v \approx 0.2\,n_{WS}$. The situation in figs 15 and 16*b* is comparable with that of two different metals that are in surface contact with one another (fig. 16*c*). As shown in the caption to *fig. 17*, it seems reasonable in this case to assume that matching the two electron densities n_A and n_B requires an energy proportional to $(n_A^{\frac{1}{2}} - n_B^{\frac{1}{2}})^2$. If this result is applied to fig. 16*b*, then the matching energy required is proportional to $(n_{WS}^{\frac{1}{2}} - n_v^{\frac{1}{2}})^2$. The surface energies of a cavity and a free atom should therefore be in the ratio

$$a_2/a_1 = (n_{WS}^{\frac{1}{2}} - n_v^{\frac{1}{2}})^2/n_{WS};$$

when $n_v = 0.2\,n_{WS}$ this gives $a_2/a_1 = 0.3$. The differences that still remain between theory and reality — of the order of 10 or 20% — are probably due to relaxation.

The values of ΔH_{vac} in Table I have been obtained by using equation (2) with $a_2 = 1.42 \times 10^8$ for the

transition metals and $a_2 = 1.17 \times 10^8$ for other metals (see fig. 3). In this way a full list of values has also been obtained for ΔH_{vac}, which are probably more reliable in many cases than the limited direct experimental data available.

The enthalpy of formation for monovacancies, ΔH_{vac}, exhibits a striking relationship with the melting point T_f. Table I gives the ratio $\Delta H_{vac}/RT_f$. If we leave out extreme cases — Hg, Ga, In and Sn on the one hand, and Ca, Ba, Sr, Si and Ge on the other — the ratio is more or less constant: it varies from 6.8 to 10.3. A constant value of $\Delta H_{vac}/RT_f$ means that the concentration of monovacancies (c) has a fixed value at a temperature just below the melting point ($c = c_0 \exp(-\Delta H_{vac}/RT)$). It looks as if a metal melts at the temperature at which the concentration of monovacancies has reached a value that is about the same for most metals.

Three special metallurgical problems

In this last section three examples will be discussed that illustrate the usefulness of the model for making predictions.

The first example concerns the *diffusion rates* of the components in an alloy. These diffusion rates are extremely important for the homogenization of an alloy by heat treatment after the actual manufacturing process. Let us consider the compounds $LaNi_5$ and La_3Ni as typical examples. The diffusion rate is determined to a large extent by the concentration of vacancies. Now the heat of formation of an La vacancy in La metal is almost the same as that of an Ni vacancy in nickel (see Table I); the rates of 'self-diffusion' are also approximately equal. It might therefore be expected that the concentration of vacancies at La and Ni sites would also be the same in the two compounds. Nothing is further from the truth, however, because the 'atomic surface areas' ($V_m^{\frac{2}{3}}$) of La and Ni differ by a factor of more than two: 8 cm² for La and 3.5 cm² for Ni.

In $LaNi_5$ an La atom is entirely surrounded by Ni atoms; an La vacancy is therefore effectively a cavity in nickel metal; it is however much greater than an Ni vacancy in nickel. From the above we would expect to find that

$$(\Delta H_{vac})_{La \text{ in } LaNi_5} = (V_{La}/V_{Ni})^{\frac{2}{3}} (\Delta H_{vac})_{Ni \text{ in } Ni} = 2.3\, (\Delta H_{vac})_{Ni \text{ in } Ni}. \quad (10)$$

With $(\Delta H_{vac})_{Ni \text{ in } Ni}/RT \approx 8$ (see Table I) the factor 2.3 in the exponent produces a ratio of $e^{-2.3\times 8}/e^{-8} = 10^{-5}$ in the concentrations. In fact, equation (10) still underestimates the energy of an La vacancy because the electron density n_v at the centre will be much smaller than for an Ni vacancy in Ni. It is very rare therefore for La vacancies to occur in $LaNi_5$, even close to the melting point. In heat treatments of alloys of this kind — which include for example $SmCo_5$, the well-known permanent-magnet material — it can be assumed that only the small atoms diffuse.

In La_3Ni, on the other hand, an Ni atom is completely surrounded by La atoms. An Ni vacancy can be regarded as a cavity in La metal, but of very small dimensions. The energy of such a cavity is particularly small, so that the Ni atoms in La_3Ni will have a particularly high diffusion mobility.

The second example concerns *surface segregation* in alloys. If a metal A is dissolved in a metal B, the concentration of A in the outermost atomic layer (c_1^A) usually differs considerably from the concentration inside the metal (c_∞^A). There are two forces at work to produce the segregation, which may act in the same or in opposite directions. In the first place, the component with the smallest surface tension is driven towards the surface. Secondly, if the heat of solution is positive, the minority component A is driven out because A atoms in the surface layer are not as completely surrounded by B as in the interior. In an earlier article an expression for c_1^A/c_∞^A was derived indirectly [6]. It will now be shown that this expression can also be obtained directly from the model under discussion, clearly revealing how important surface tension is for the surface segregation. The result will also be corrected for relaxation effects, which make the effective area of the external faces of a surface atom equal to $\frac{1}{4}$ instead of $\frac{1}{3}$ of the total surface area of the atom. This is indicated symbolically in *fig. 18* by representing the atoms as squares.

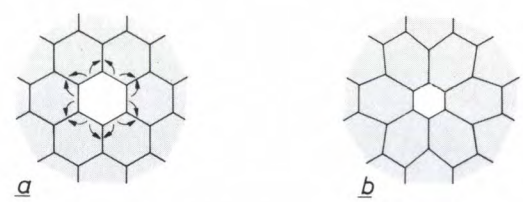


Fig. 15. Two forms of relaxation in a vacancy. *a*) Charge transfer. *b*) Transfer and deformation of the neighbouring atoms resulting in a cavity that is smaller than the free atom. These effects could explain a difference of about 20% between a_2 and a_1 but the difference is much greater than this ($a_2 \approx \frac{1}{3} a_1$).

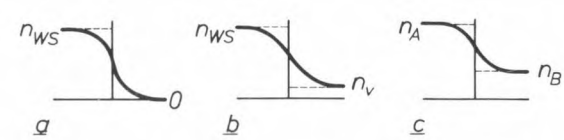


Fig. 16. Matching of the electron density (*a*) between metal and vacuum ($n_{WS} \rightarrow 0$), (*b*) between metal and the centre of a vacancy ($n_{WS} \rightarrow n_v$), (*c*) between two different metals ($n_A \rightarrow n_B$).

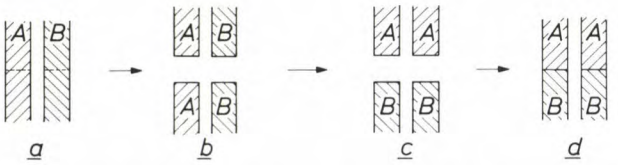


Fig. 17. Interface energy between two different materials. For materials that have Van der Waals interaction it is known that an interface like the one in situation *d* represents an energy $(\gamma_A^{\frac{1}{2}} - \gamma_B^{\frac{1}{2}})^2$ per unit interface area if the pieces of solid A and B (*a*) are taken as a reference of energy zero. This follows from the proportionality of the interaction energy of two pieces of material to the polarizabilities (*P*) of the two pieces. The splitting of the materials ($a \rightarrow b$) expends an energy $K(P_A^2 + P_B^2)$; the interchange ($b \rightarrow c$) expends nothing, and the combination ($c \rightarrow d$) yields an energy $2KP_AP_B$. Situation *d* therefore represents an energy $K(P_A^2 + P_B^2 - 2P_AP_B) = K(P_A - P_B)^2$. For KP^2, the energy of split material (for example A or B in *b*), we can of course also write $2\gamma S$, hence $P = (2\gamma S/K)^{\frac{1}{2}}$, where *S* is the area of the horizontal section. The energy of situation *d* is therefore $2S(\gamma_A^{\frac{1}{2}} - \gamma_B^{\frac{1}{2}})^2$. Because the surface tension in materials with Van der Waals interaction is proportional to a surface electron density and because this is also approximately true for metals (fig. 10) it is reasonable to assume that the interface energy for two metals is proportional to $(n_A^{\frac{1}{2}} - n_B^{\frac{1}{2}})^2$.

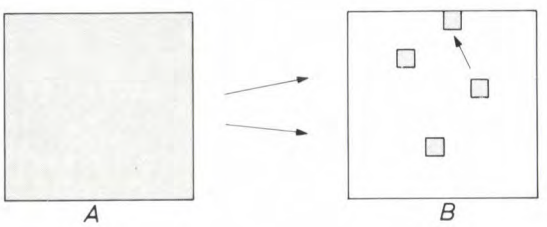


Fig. 18. Solution of a metal A in a metal B. The heat of solution is the interface energy A-B of N_0 atoms of A in the interior of B. If N_0 atoms of A move from the interior to the surface the total energy is reduced by a quarter of the heat of solution and by the difference in surface energy between N_0 atoms of B and N_0 atoms of A.

The energy of one mole of A atoms surrounded by B instead of A is the heat of solution $\Delta H_{sol}{}^A$. In our model this is the interface energy of the total atomic interface A-B. If these atoms are brought to the surface then in the first place a quarter of this interface energy, $\frac{1}{4}\Delta H_{sol}{}^A$, is liberated. Secondly, each A atom receives some additional surface energy, i.e. a quarter of the latent heat of vaporization per mole: $\frac{1}{4}a_1\gamma_A V_A{}^{\frac{2}{3}}$. Finally, the surface energy $\frac{1}{4}a_1\gamma_B V_A{}^{\frac{2}{3}}$ is liberated by the B atoms that have had to make room for these atoms over the effective surface $\frac{1}{4}a_1 V_A{}^{\frac{2}{3}}$. The difference in energy between a state with one mole of A atoms in the interior and a state with one mole of A atoms at the surface is therefore

$$\tfrac{1}{4}\Delta H_{sol}{}^A + \tfrac{1}{4}a_1(\gamma_B - \gamma_A)V_A{}^{\frac{2}{3}}.$$

Hence:

$$c_1{}^A/c_\infty{}^A = \exp\left[\Delta H_{sol}{}^A + a_1(\gamma_B - \gamma_A)V_A{}^{\frac{2}{3}}\right]/4RT. \quad (11)$$

Except for the factor 4 in the denominator of the exponent, this agrees with the result obtained earlier. Equation (11) is in excellent agreement with experimental data [1c].

If the atoms are very different in size the A atoms will not fit into the B lattice properly, and this might have the effect of forcing them to the surface. This has not been taken into account in equation (11); in most cases it is of minor significance [1c].

The following values apply for a solution of copper in iron at an after-treatment temperature of 1000 K: $\Delta H_{sol}{}^A = 56$ kJ/mol [7], $\gamma_B - \gamma_A = 0.7$ J/m^2 (see Table I), $V_A{}^{\frac{2}{3}} \approx 3.7$ cm^2/mol and $4\,RT = 33$ kJ/mol. This gives 1.7 for the first term of the exponent and 4.1 for the second. The result therefore is $c_1{}^A/c_\infty{}^A = e^{5\cdot8} \approx 330$. This example is typical in the sense that the second term in (11) is usually the more important, and considerable surface segregation can be expected even with quite moderate differences in γ. The surface segregation of copper in transition metals can be put to good use: even a small dose of copper gives a material that is more resistant to corrosion.

In our last example let us examine somewhat more closely the 'special' metal *mercury*. On pages 259-260 the conclusion was drawn that Hg atoms occur in two extreme forms: inert-gas-like atoms with an approximately s^2 configuration and metallic atoms with a mixed s,p configuration. This probably explains some of the remarkable properties of free atoms, monolayers and small particles of mercury [8]. The inert-gas nature of the free atoms, for example, directly explains the extremely low 'sticking coefficient' of Hg when vacuum-evaporated on to non-metallic substrates at normal temperatures; the atoms appear to bounce back like billiard balls. The sticking coefficients of Cd and Zn are also very low.

Experiments with monolayers of mercury on glass (vacuum-evaporated at low temperature) indicate that there are two kinds of adsorbed mercury if the population exceeds that of about a $\frac{1}{4}$ monolayer of metallic mercury. They differ in their heat of adsorption for example and can be converted from one to the other by heat treatment. In view of this, let us try to find out which type of mercury our model would predict for a monolayer. The monolayer is assumed to be in free space. In a surface layer one quarter of the atomic surface area is effectively external surface; in a monolayer — a two-sided surface layer — this would therefore be a half. Now the surface area of a free mono-

[6] A. R. Miedema, Philips tech. Rev. **36**, 217, 1976.
[7] Tables of the heat of solution of a metal in any other metal are given in A. R. Miedema, F. R. de Boer, R. Boom and J. W. F. Dorleijn, Calphad **1**, 353, 1977.
[8] A. R. Miedema and J. W. F. Dorleijn, shortly to appear in Phil. Mag.

layer may be reduced somewhat by deformation of the atomic cells, which is not possible with a surface layer. We have estimated the surface area of the outside at 40% of the atomic surface area.

To make an estimate for an inert-gas-like monolayer, let us assume for this kind of atom as well that the

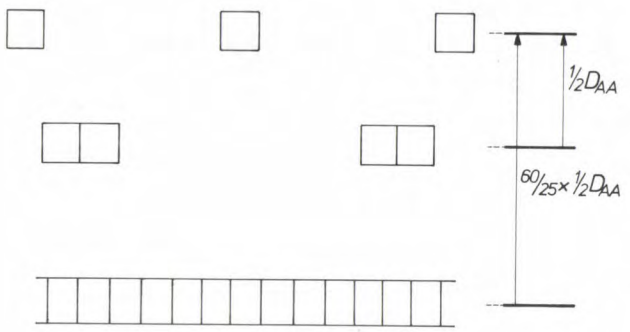


Fig. 19. Calculation of the binding energy of a monolayer of inert-gas-like mercury atoms.

binding energy is the surface energy that is released where atoms make contact. (The corresponding surface tension is of course much smaller than that for metallic atoms.) In combining to form dimers (*fig. 19*) they lose 25% of their outer surface; the energy then falls by $\frac{1}{2}D_{AA}$, i.e. 3.75 kJ/g.at. In combining to form a monolayer they would lose 60% of their external surface area; the energy would then fall by (60/25) × 3.75 kJ/g.at. or 9 kJ/g.at. This is the dissociation energy of an inert-gas-like monolayer.

Fig. 20 illustrates the estimate for a metallic layer. If Hg were an 'ordinary' metal (i.e. metallic in all of its states of aggregation), the energy of free atoms (the latent heat of vaporization) would be $a_1(\gamma V_m^{\frac{2}{3}})_{Hg}$ or 183 kJ/g.at. The energy of a monolayer with respect to the solid is 0.4 times as great: 73 kJ/g.at. The actual latent heat of vaporization ΔH^V, however, is only 60 kJ/g.at. The dissociation energy of the metallic monolayer is therefore negative (−13 kJ/g.at.).

From these energy considerations it would follow that monolayers of the inert-gas type can exist, but metallic layers cannot; however, the energy differences are small. For layers applied to a substrate the chance of metallic layers occurring is relatively greater than for the inert-gas type, because the decrease in energy due to the adhesion is greater for the metallic type. Different results can therefore be expected for different substrates. It should also be pointed out that the inert-gas-like atoms are much larger and that a complete monolayer of such atoms corresponds approximately to a $\frac{1}{3}$ monolayer of metallic atoms. It therefore seems reasonable to assume that inert-gas-like atoms are deposited up to this level of population and that the two forms occur simultaneously beyond it.

It would appear that the 'coordination number' (the number of surrounding neighbours) determines the nature of the atom: if it has many neighbours then it is metallic, if it has few or no neighbours, then it is inert-gas-like. Now for small particles the surface layer is noticeably curved and therefore the coordination number is smaller than for a planar surface layer. Thus, small particles of mercury could well have a metallic core with an inert-gas-like outer layer. This might explain some of the optical properties of small particles, such as the gradual change in the optical constants in layers of about 0.5 nm thick [8].

In spite of the resemblance between mercury and the other divalent metals, mercury is unique in the ambivalent nature of its monolayers. In cadmium, for example, the latent heat of vaporization (112 kJ/g.at) is greater than $0.4a_1\gamma V_m^{\frac{2}{3}}$ (89 kJ/g.at) so that even a free monolayer will be metallic.

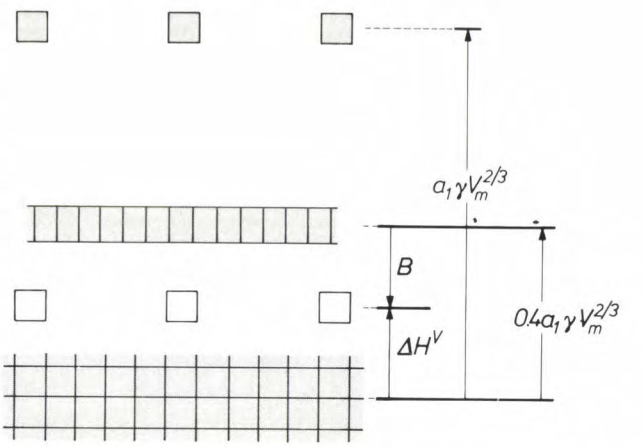


Fig. 20. Calculation of the binding energy (B) of a monolayer of metallic mercury atoms. *White:* inert-gas-like atoms, *grey:* metallic atoms.

Summary. The article describes a model of a metal, in which the binding energy of atoms in dimers and in the solid is interpreted as a reduction in surface energy. The model predicts relationships for the latent heat of vaporization (ΔH^V), the dissociation energy of molecules (D_{AA}) and the heat of formation of monovacancies (ΔH_{vac}) with the surface tension (γ_s) of the solid at zero temperature. A large group of metals are found to comply with these relationships. Exceptions are the metals Mg, Zn, Cd and Hg, because of a particularly stable (inert-gas-like) electron configuration of the free atom. The quantity γ_s is derived from the surface tension of the liquid at the melting point, which is known for all metals, by making corrections for temperature and roughness. The surface tension is approximately proportional to the electron density at the boundary of an atomic cell in the solid. Relaxation (transfer of bonds to interior atomic faces) and the interaction between the walls of a cavity help to explain the values found for the coefficients in the above relations. The investigation has made it possible to compile a complete list of values for γ_s, ΔH_{vac} and D_{AA}. Finally, the model is used for making predictions of diffusion rates, surface segregation and monolayers.

PHLIQA 1, a question-answering system for data-base consultation in natural English

PHLIQA Project Group

II. The artificial languages and translation operations

In part I of this article [1] a description was given of the basic structure of PHLIQA 1. It was shown that a question put to it in natural English is not directly translated by the system, but subjected to three successive translation operations: from natural English to EFL (*E*nglish-oriented *F*ormal *L*anguage), from EFL to WML (*W*orld-*M*odel *L*anguage), and from WML to DBL (*D*ata-*B*ase *L*anguage). The value of the expression obtained in DBL is then computed with the aid of the data contained in the data base. During the translation operation English→EFL the question is parsed and its logical structure is established in the form of an expression in EFL. The translation operations from EFL via WML to DBL establish the relationship between words in natural English and primitive terms in the data base. This is done by translating the primitive terms of EFL, which correspond to natural English words, into expressions in DBL that have the same meaning.

In this second part we shall examine the nature of the artificial languages and deal in more detail with the translation operations [2]. It should first be noted that between EFL and WML and between WML and DBL the PHLIQA 1 system contains other, intermediate languages. These serve to divide up the translation process into small steps, which is important for clarity and ease of approach to the system in the design stage. It would be going too far to deal with these other languages here, though we shall touch upon them in passing.

In the section that now follows we shall discuss the rules of grammar and the semantic rules of the artificial languages. First of all we show how the expressions of these 'construction languages' — the constructions — are built up. We then discuss the assignment of 'types' to the constructions. The types make it possible to judge whether the constructions are 'well-formed'. Next we deal with the introduction of variables as primitive terms. The last part of the section is devoted to the semantic rules, which make it possible to assign a value to any construction in the language concerned, once the values of the primitive terms are given.

In the final section we consider the principal mechanisms of the translation operations and illustrate them with some examples. The first translation operation, English→EFL, is obviously different in character from the two other translation operations owing to the fact that English is a natural language, whereas in the other cases both the languages are construction languages.

The artificial languages

Constructions

The expressions of the artificial languages used in PHLIQA 1 are the constructions referred to above. An example of a construction is given in *fig. 1*. It is a branching diagram that resembles a tree (upside down).

Compared with ordinary sentences, i.e. rows of letters and characters, constructions have the advantage that their syntactic structure is completely visible. This facilitates their treatment in a computer. The syntactic structure of a construction corresponds to its semantic structure, as will be explained below.

The PHLIQA Project Group consists of W. J. H. J. Bronnenberg, Ir S. P. J. Landsbergen, Ir R. J. H. Scha, Ir W. J. Schoenmakers and Drs E. P. C. van Utteren of Philips Research Laboratories, Eindhoven. Until 1976 the Project Group also included Drs H. C. Bunt, now with the Institute for Perception Research (IPO), Eindhoven, and Ir P. Medema, now with the Philips Information Systems and Automation Department (ISA), Eindhoven.

[1] PHLIQA Project Group, PHLIQA 1, a question-answering system for data-base consultation in natural English, I. Organization and performance, Philips tech. Rev. **38**, 229-239, 1978/79 (No. 9).

[2] A more detailed treatment of these subjects is given in: W. J. H. J. Bronnenberg, H. C. Bunt, S. P. J. Landsbergen, R. J. H. Scha, W. J. Schoenmakers and E. P. C. van Utteren, The question answering system PHLIQA 1, to appear in: L. Bolc (ed.), Natural language question answering systems (Natural language communication with computers, part II), Hanser-Verlag, Munich/Vienna, and Macmillan, London, 1979.

The construction in fig. 1*a* consists of two *primitive terms*, i.e. the language constants COMPUTERS and DUTCH, and what we have called a *branching*. A branching consists of a branching category (in this case "selection") and a number of selectors associated with this branching category ("head" and "mod" (= modifier), in the case of "selection"). The branching category determines the type of branching involved. The selectors point to what is termed a subconstruction. The names of the selectors are chosen so as to characterize the roles played by the subconstructions within the total construction. In fig. 1*a* the subconstructions are the primitive terms COMPUTERS and DUTCH. The subconstructions may have branchings themselves. An example is given in fig. 1*b*. To make complicated branching diagrams clearer in this article, we shall occasionally substitute green triangles containing a piece of text in natural English for certain subconstructions (fig. 1*c*). For simplicity the expression 'branching with the branching category a' will be abbreviated to 'a-branching'. A construction with an a-branching at the top will be referred to as an 'a-construction'. The construction in fig. 1*b* is thus an existential-quantification-construction containing a selection-construction as subconstruction.

Each branching has a corresponding semantic rule (further particulars are given in the last part of this section) that lays down how the *value* of a construction that is built up with the aid of that branching from subconstructions follows from the values of the subconstructions. In this way it is known what the value of every construction is, provided that the values of the primitive terms are given. The case in fig. 1*a* can again be taken as an example. We take the value of the language constant COMPUTERS to be The Set Of All Computers In The World and the value of the language constant DUTCH to be a function that yields the value TRUE when applied to an object in the Netherlands and FALSE for all other objects. The semantic rule for the selection-branching is: take from the set 'under' the selector "head" the subset consisting of the elements with the property that application of the function 'under' the selector "mod" yields the value TRUE. This subset is the value of the selection-construction. In our example that value is The Set Of All Computers In The Netherlands.

The second branching in fig. 1*b* is of the category "existential quantification". This category has the selectors "for some" and "holds". The subconstruction to which the selector "for some" points must relate to a set, and the subconstruction under the selector "holds" must relate to a predicate. The value of the existential-quantification-construction is TRUE if there is an element in the set for which the predicate is true, otherwise it is FALSE. The value of the complete construction in fig. 1*b* is thus TRUE if there are expensive computers in the Netherlands.

In addition to the branching categories "existential quantification" and "selection", which we have discussed at some length, we shall briefly mention here another four, which will be used later in examples (fig. 1*d*).

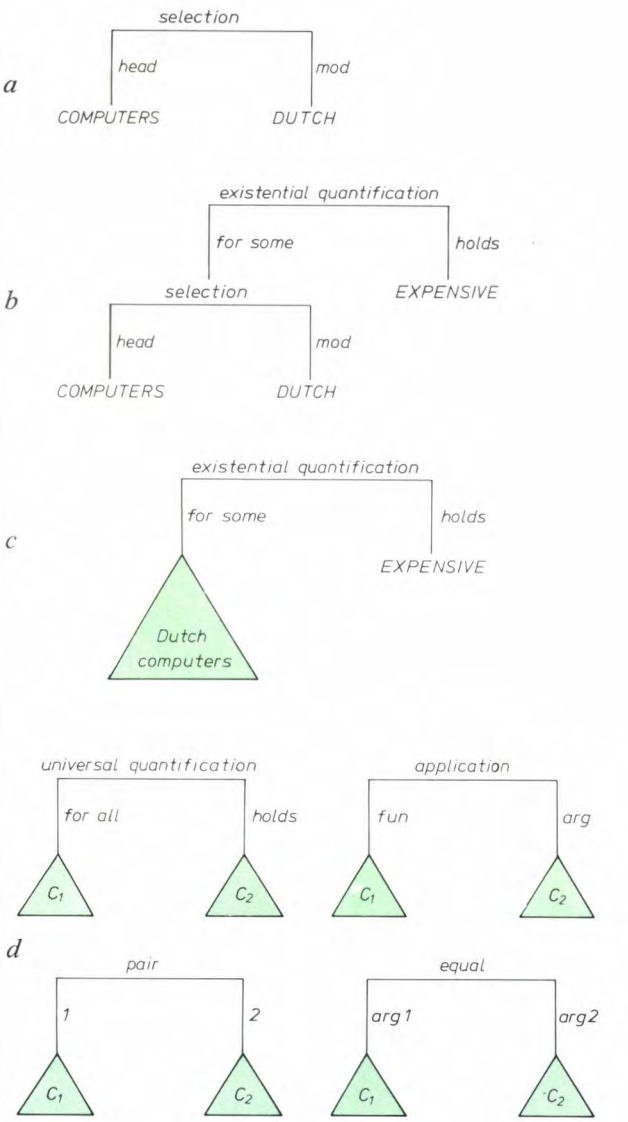


Fig. 1. *a*) Example of a construction. This represents the logical structure of the phrase: "Dutch computers". The construction contains one branching, with two selectors, and two primitive terms – located at the end points. The 'branching category' is shown above the horizontal line of the branching. The selectors are indicated beside the vertical lines. *b*) Example of a construction with two branchings. This represents the logical structure of the sentence "Some Dutch computers are expensive". The existential-quantification-branching represents a well-known concept from logic: the existential quantifier. The construction in (*a*) occurs here as a subconstruction. *c*) An abbreviated form of the construction in (*b*). The triangle with the text in natural English replaces the selection-branching, which is a subconstruction of the complete construction. *d*) The branching categories "universal quantification", "application", "pair" and "equal". These are used in the examples of translations in EFL, WML and DBL (figs 9, 10, 11).

The first is the branching category "universal quantification". This has two selectors, "for all" and "holds". The selector "for all" indicates — by virtue of the type rules, which we shall presently discuss — a subconstruction that has a *set* as its value; the selector "holds" points to a subconstruction that corresponds to a predicate. The value of the construction is TRUE only if the predicate under "holds" is *true* for all elements of the set that is the value of the subconstruction under "for all".

The second is the branching category "application". To this belong the selectors "fun" (= function) and "arg" (= argument). The value of the application-construction is the result of applying the function that is the value of the subconstruction under "fun" to the value of the subconstruction under "arg".

The third is the branching category "pair", with the selectors "1" and "2". These refer to two subconstructions, which thus become the first and the second member of a pair.

The fourth branching category is the category "equal", with the selectors "arg 1" and "arg 2". The value of the equal-construction is TRUE if the values of the subconstructions under "arg 1" and "arg 2" are equal, and otherwise FALSE.

The grammar

To define a construction language — a set of constructions — we need a grammar of that language, that is to say a specification of branchings and primitive terms, and rules telling us how constructions are built up from the branchings and the primitive terms.

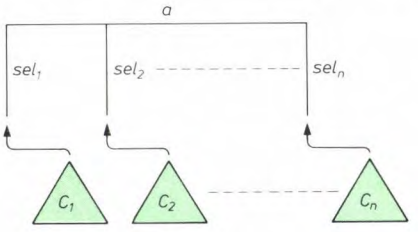


Fig. 2. The composition of a construction. The constructions $C_1, C_2, \ldots, C_n$ are connected to the branching with branching category a, to which the selectors $sel_1, sel_2, \ldots, sel_n$ belong. The rules for forming the construction are part of the grammar of the appropriate construction language. The construction C obtained can also be written:

$$a\ (sel_1 : C_1, sel_2 : C_2, \ldots, sel_n : C_n).$$

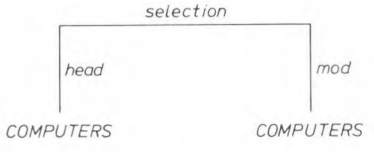


Fig. 3. Example of a meaningless construction. The selector "head" should refer to a subconstruction that has the value of a set, whereas the selector "mod" should refer to a subconstruction that has the value of a predicate. The primitive term COMPUTERS cannot refer to a set and to a predicate at the same time. The construction is therefore meaningless.

To begin with we must therefore have a set of branchings, each consisting of a branching category and the associated selectors. There must also be a set of primitive terms. The branchings can be regarded as a generalization of the language constructions like those that occur in logical languages, for example in proposition logic and predicate logic [3]. All artificial languages of PHLIQA 1 have the same branchings. As already mentioned in part I of this article, these languages differ in their sets of primitive terms.

The rules for forming a construction consist of two parts. The first part states that the primitive terms themselves must be counted among the constructions. The second part establishes that a construction is formed by taking a branching and connecting an arbitrary group of constructions $C_1, \ldots, C_n$ to its group of selectors $sel_1, \ldots, sel_n$ (*fig. 2*).

The subconstructions $C_1, \ldots, C_n$ of the newly formed construction may be primitive terms or constructions that in their turn also contain branchings. When they are written out in full, however, all branching diagrams end with primitive terms.

Any arbitrary combination of branchings with primitive terms at the ends is a construction, according to the rules. This would imply that a construction language might be built that could also contain meaningless constructions. To exclude such expressions, the grammar must in addition comprise rules indicating which constructions belong to the language and which do not. In the case of PHLIQA 1 this has been done by introducing *type rules*, which assign a *type* to every construction. The system accepts the constructions with a 'permissible type' as belonging to the language concerned and the others as not belonging to it. An example of a construction with an unacceptable type is given in *fig. 3*.

Types

To ensure that all grammatically correct expressions of a construction language also have a meaning, we can extend the grammar with the rules mentioned earlier for the types [4].

The type rules are introduced as follows. First of all, a type language is defined, in much the same way as a construction language. The expressions of the type language are the *types*. Next, every expression of the construction language is assigned a type by means of

[3] See for example R. Rogers, Mathematical logic and formalized theories, North-Holland, Amsterdam 1971.

[4] More detailed information on the use of types in defining a language is given in: R. J. H. Scha, Semantic types in PHLIQA 1, Proc. 6th Int. Conf. on Computational Linguistics, Ottawa 1976.

a 'type function' (T). There is a special type, empty, reserved for all constructions regarded as ungrammatical.

Only constructions of a type not equal to empty are regarded as belonging to the language. The construction language thus obtained is referred to as a 'typed construction language'.

A function that has a type of the form ($\beta \rightarrow$ truthvalue), where β is an arbitrary type, is called a predicate. The type ($\beta \rightarrow$ truthvalue) is therefore called 'predicate type'. Such a predicate indicates a property that objects of the type β can have. If the function, applied to a particular object, yields the value TRUE, the object has that property, otherwise it does not.

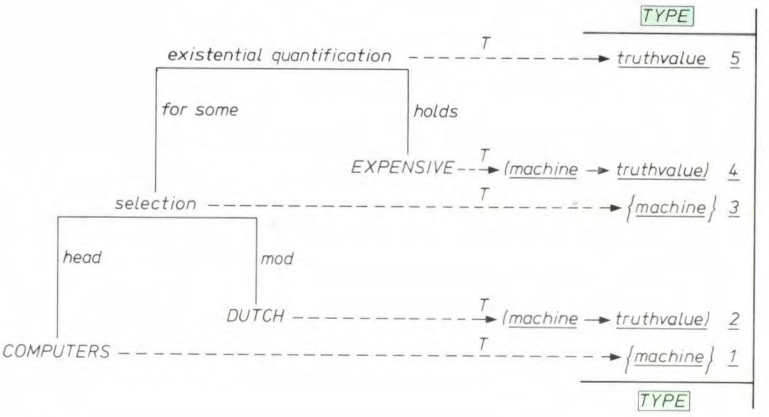


Fig. 4. Example of the application of the type function T for the assignment of a type. The types to be assigned to the EFL constructions in fig. 1*b* are given, each with a number, in the right-hand column. Since machine from *1* also occurs as an argument type in *2*, T assigns to the selection-subconstruction a type that also contains machine. If *1* and *2* were to differ in this respect, *3* would be the type empty. The subconstructions under "mod" and "head" would not then match each other in terms of type, making the selection-branching into a syntactically incorrect construction. Much the same applies to the assignment of the type *5*, where T proceeds from the types *3* and *4*.

A type language is obtained by first of all specifying a set of 'atomic types', and secondly establishing a number of type rules by means of which compound types can be built up from the atomic types. The various artificial languages of the PHLIQA 1 system have various sets of atomic types. The rules for making the compound types are the same for all languages. Atomic types include country, machine, company and truthvalue. The semantic aspects of the type language will be discussed later; we can say here informally that an atomic type characterizes a particular *kind* of object in the subject domain.

Two examples will serve to show how compound types are built up by means of the type rules. The first example relates to the type rule that builds up from an arbitrary, e.g. atomic, type α the compound type $\{\alpha\}$, the type of a set of objects that themselves each have type α; α is called the 'element type' of this set. An actual example is {country}, the (compound) type of a set of countries. The second example is the formation of the type of a function. If α_1 and α_2 are arbitrary types, then $(\alpha_1 \rightarrow \alpha_2)$ is the type of a function formed from them that can be applied to any object of the type α_1 and which yields as a function value an object of the type α_2. Here α_1 is referred to as the argument type of the function type and α_2 as the value type. An actual example of a function type is (machine → truthvalue), a function that is applicable to any object of the type machine and which then yields a function value which is either TRUE or FALSE, i.e. an object of the type truthvalue.

The type function T establishes a relation between the construction language and the type language by assigning a type to each construction. This is done by first of all establishing what T does with the primitive terms of the associated construction language. For the language constant HOLLAND, for example, we have:

$$T(\text{HOLLAND}) = \text{country}.$$

Then, for each branching it is established how the type of a construction that consists of that branching, with associated subconstructions, follows from the types of those subconstructions. For unwanted combinations of subconstructions the type of the construction is empty. In this way the system can compute the type of every construction.

To illustrate how the type function T works for constructions with branchings, we return to the construction in fig. 1*a*. Here the set type {machine} has been assigned to the language constant COMPUTERS and the type (machine → truthvalue) has been assigned to the language constants EXPENSIVE and DUTCH. The application of T to a selection-branching implies that a check is made first of all to see whether the type of the subconstruction under the selector "head" is a set type, i.e. of the form $\{\alpha\}$, and whether the type of the subconstruction under the selector "mod" is a predicate type, in other words of the form ($\beta \rightarrow$ truthvalue). Next a check is made to see whether the element type α of the set 'is contained in' the argument type β of the predicate. (The relation 'is contained in' between

types will not be discussed further in this article; it is sufficient here to know that α is certainly 'contained in' β if $\alpha = \beta$.)

In this case the selection-construction receives the same type as the subconstruction under the selector "head", here {machine}; in the other cases the construction receives the type empty. In the branching

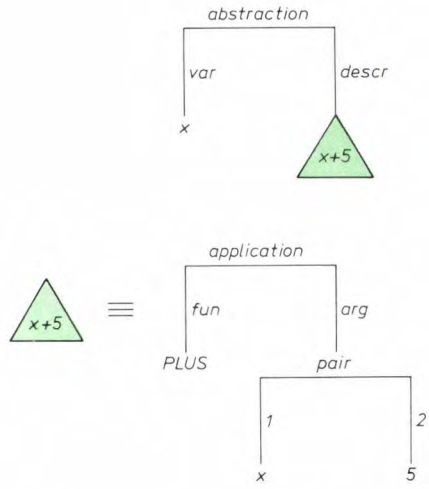


Fig. 5. *a*) A construction with the variable x as primitive term. This example is the representation of the function $f: x \rightarrow x + 5$ in the languages used in the PHLIQA 1 system. *b*) The subconstruction $x + 5$ of (*a*) written out in full. The actual addition can be made in the formal language by using an application-branching with the primitive term PLUS and the pair-branching, a subconstruction that turns the two primitive terms x and 5 into the first and second arguments of the function PLUS.

category "existential quantification" T imposes the same kind of requirement on the agreement between the types of the subconstructions under the selectors "holds" and "for some". If this requirement is satisfied, the type truthvalue is assigned to the existential-quantification-construction. The construction in fig. 1 is again shown in *fig. 4*, but now with the types included. It will be evident from the foregoing that the application of T in the case of fig. 3 will lead to the type empty being assigned to the meaningless expression.

Variables

The expressive scope of the typed construction languages has been enlarged by admitting *variables* as primitive terms in addition to the kinds of language constants already mentioned. The PHLIQA 1 system is therefore also capable of representing functions that do *not* already occur as primitive terms. It does this by means of a branching category called "abstraction". This has two selectors, "var" (= variable) and "descr" (= descriptor). The subconstruction under "var" must be a variable. The subconstruction under "descr" may be any arbitrary construction in which the associated variable may occur. The value of an abstraction-construction is a function. With each value that the variable x can assume, this function associates the value of the subconstruction — for the same x — under "descr".

An example is the mathematical function $f: x \rightarrow x + 5$, where the variable x has a value range consisting of the set of integers. This function is given by the construction in *fig. 5*.

Like the other language constants, variables have a type. Thus, the variable x in fig. 5 has the type integer, which indicates that the value range comprises all integers. The subconstruction for $x + 5$ also has the type integer. From these two pieces of information the type function T assigns to the construction in fig. 5 the type (integer → integer), a second example of a function type.

Semantics

In the above discussion of the branchings we touched briefly on the associated semantic rules. The complete set of semantic rules is called the semantics of the related construction language. The semantic rules assign to each construction of the language a particular value, given an assignment of values to the primitive terms. Such an assignment of values to the primitive terms is called an *interpretation* of the language.

In a typed construction language the interpretation proceeds in two phases. The first yields an interpretation of the type language, referred to as the *type interpretation*; on the basis of this, the second phase provides the interpretation of the construction language itself, called the *language interpretation*.

A type interpretation is an assignment of a *value domain* to each atomic type. A value domain is a set of objects from the subject domain that are characterized by the associated type. The value domain of the atomic type country, for example, could be the set of all countries. The value domain chosen in the PHLIQA 1 system is the set of the countries that are members of the EEC. Some atomic types, called formal types, have the *same* value domain for *each* interpretation; for example, the type truthvalue always has the set {true, false} as its value domain. The non-formal atomic types are called referential types, since they refer to 'real' objects from the subject domain.

The semantic rules of the type language indicate how the value domain of a compound type follows from the value domains of the atomic types contained in it. Two examples will help to make this clear. The value domain of the compound type {country}, our first example, is defined by these rules as the set whose elements are the *sub*sets of the value domain of the atomic type country. The value domain of the com-

pound type (machine → truthvalue), in a similar way to that of the first example, is the set of all (predicate) functions that are applicable to all elements in the value domain of the type machine and which, when applied, yield an element in the value domain of the type truthvalue. In this way the value domains of all compound types can be derived from the value domains of the atomic types.

A language interpretation is an assignment of a value to each primitive term of the construction language. This value assignment must be in accordance with the type of the primitive term concerned and the type interpretation; this means that the assigned value of a primitive term C should be an element of the value domain that has been assigned to the type of C:

$$\text{Value } (C) \in \text{Value domain } (T(C)),$$

where T is the type function. If, for example, the language constant HOLLAND has the type country, and the type interpretation assigns to country the value domain consisting of all countries, then the value of HOLLAND must be one of these countries. Some primitive terms, called formal constants, have the same value for *each* interpretation. An example is the constant 5, whose value is always the number 5. The non-formal constants are called referential constants.

Every branching of the construction language has a semantic rule, which specifies how the value of the construction consisting of that branching together with the subconstructions present is derived from the values of those subconstructions. The foregoing sentence implies that the semantic structure of a construction corresponds to the syntactic structure — which is determined by the branchings. The semantic rules for a few examples have already been touched upon in our discussion of the branchings in the first part of this section.

The semantic rules and the definition of the type function for the branchings are matched in such a way that for each construction t of the language we can write:

$$\text{Value } (t) \in \text{Value domain } (T(t)).$$

To understand this properly it is important to make a sharp distinction between the interpretation of a language and the semantic rules of that language. The interpretation does not belong to the language itself, but establishes the relation between the primitive terms of the language and what they denote. The semantic rules of the branchings, on the other hand, do form part of the language definition. They are mathematical rules that indicate, by operations on sets or on separate elements, how the value of a construction follows from the values of the subconstructions and ultimately from the values of the primitive terms.

A semantic rule that belongs to a particular branching operates in fact as a function that maps the values of the subconstructions under the selectors of the branching on to the value of the *total* construction. The way in which the types in the grammar and in the semantics of the construction language function guarantees that the semantic rules will in fact be applicable to *all possible* values of the subconstructions (for all possible interpretations of the language). The relation

$$\text{Value } (t) \in \text{Value domain } (T(t))$$

mentioned above thus ensures that the total construction has a well-defined meaning.

As an example, let us look once again at fig. 4. The language constant COMPUTERS in this construction has the type {machine} and therefore receives as its value for each interpretation a subset of the value domain of machine. The language constant DUTCH has the type (machine → truthvalue) and its value is thus a function that is applicable, again for each interpretation, to *all* elements of the value domain of machine. It follows from the two foregoing sentences, to put it rather informally, that DUTCH is applicable to all elements of the set COMPUTERS, for *every* interpretation. This conclusion implies that the selection construction with DUTCH and COMPUTERS as subconstructions will therefore have a value for every interpretation. The type rules for the selection-branching thus guarantee that the selection-construction built up from DUTCH and COMPUTERS will have a well-defined meaning.

The language DBL is the only construction language of PHLIQA 1 for which an interpretation is given *directly*, that is to say by the data base. For example, the data base assigns to the language constant $\text{COUNTRIES}_{\text{DBL}}$ the set of records of the record type "COUNTRY".

The languages EFL and WML can only be interpreted in an *indirect* manner, that is to say by means of the translation operations of EFL via WML to DBL. The translation rules establish the relation between the values of the primitive terms in EFL and WML, as the case may be, and the values of constructions in DBL. We shall return to this point in the next section, where we deal with the translation operations.

There is one respect in which EFL differs from the other artificial languages: the primitive terms of EFL can have more than one value. An interpretation of EFL assigns to each primitive term a *set* of values, and not a unique value. The reason for this is that EFL terms correspond to English words, and a word in natural English can have more than one meaning. The

term COMPUTERS$_{EFL}$, which corresponds to the English word "computers", is an obvious example. The PHLIQA 1 system assigns two values to COMPUTERS$_{EFL}$, the one in the meaning of complete computer configurations (including peripherals) and the other in the meaning of central processing units (= CPUs).

In the next section we shall deal with the principal aspects of the various translation operations in PHLIQA 1. The consequences of the multiple meaning of primitive terms of EFL will be discussed there in more detail.

The translation operations

From English into EFL

The operation of translating from natural English into EFL is performed by a "parser" program. This program parses the question, that is to say introduces a syntactic analysis of the question, and with its aid synthesizes the corresponding EFL construction, the first representation of the meaning of the question.

Table I. An example of information from the lexicon used by the "parser" program in the PHLIQA 1 system in translation from English into EFL. The system needs this information to parse the question "Does each computer have a CPU?". The right-hand column contains the EFL language constants that correspond to the referential aspects of certain English words (nouns and verbs in this case). The syntactic categories for each of the English words and a few other syntactic features are shown under the heading "syntactic information".

English	Syntactic information		EFL
	Category	Feature	primitive term
does	aux[*]	{"Stem": do "Verb form": sing 3	—
each	det[*]	{"Quant"[*]: universal "Exp[*] number": singular	—
computer	noun	"Number": singular	COMPUTERS$_{EFL}$
have	verb	"Verb form": infinitive	HAVE$_{EFL}$
a	det	{"Quant": existential "Exp number": singular	—
CPU	noun	"Number": singular	CPUS$_{EFL}$

[*] aux = auxiliary, det = determiner, Quant = quantification, Exp = expected.

The "parser" program is based on an English grammar consisting of a *lexicon* and a set of *reduction rules*. With the aid of the lexicon a syntactic category and a set of syntactic features are assigned to each word. In the case of referential words, such as nouns and verbs, the lexicon also gives the corresponding EFL term. The reduction rules describe how a string of successive parts of a sentence can be reduced to one compound part when the syntactic categories and the syntactic features of these sentence elements satisfy certain conditions. Associated with every reduction rule is a composition rule, which can compose the EFL construction corresponding to the new sentence element from the EFL constructions that belong to the constituent sentence elements.

The reduction process begins at the original string of words with their syntactic categories, syntactic features and EFL terms; the process attempts to reduce this string of words to a single sentence element of the syntactic category 'sentence'. The result of this reduction process can be represented in the form of a tree diagram, displaying the syntactic structure of the sentence (compare the example given below in fig. 6*a* and *b*). With each completed reduction step the appropriate composition rule forms the EFL construction that represents the EFL translation of the sentence element produced by the reduction. Once the string has been reduced to 'sentence' the result of the composition rules carried out is the EFL translation of the complete sentence. A string of words that cannot be reduced to 'sentence' is treated by the system as ungrammatical.

A sentence is said to be syntactically ambiguous if different syntactical structures can be assigned to it. An example of such a sentence is "What companies have a computer with a CPU that costs more than 100 000 dollars?" Here the clause "that costs ..." could refer to "a CPU" or to "a computer with a CPU".

In general, and in this example as well, different syntactic structures lead to different EFL constructions. The "parser" program has the task of finding all possible syntactic structures of the question and of making the corresponding representations of their meanings.

We shall now illustrate the translation operation English → EFL by a highly simplified example. The simplification relates mainly to the grammar used here. The English sentence to be translated reads: "Does each computer have a CPU?".

We shall not go into the algorithmic aspect of the matter; we will, however, use the example to show how the "parser" program applies the rules.

Table I shows the part of the lexicon relevant to this question. We see here that the word "does" has the syntactic category 'auxiliary'. Also, this word has two syntactic features. These have the form of a pair consisting of an attribute followed by a value:

"Stem" : do
"Verb form" : sing 3,

which indicates that the stem "do" belongs to "does" and that the verb form is the third person singular. "Does" is a function word, which has no separate EFL terms corresponding to it.

The reduction rules we use in this example are:

determiner + noun → noun phrase, (R1)

verb + noun phrase → verb phrase, (R2)

auxiliary + noun phrase + verb phrase → sentence, (R3)

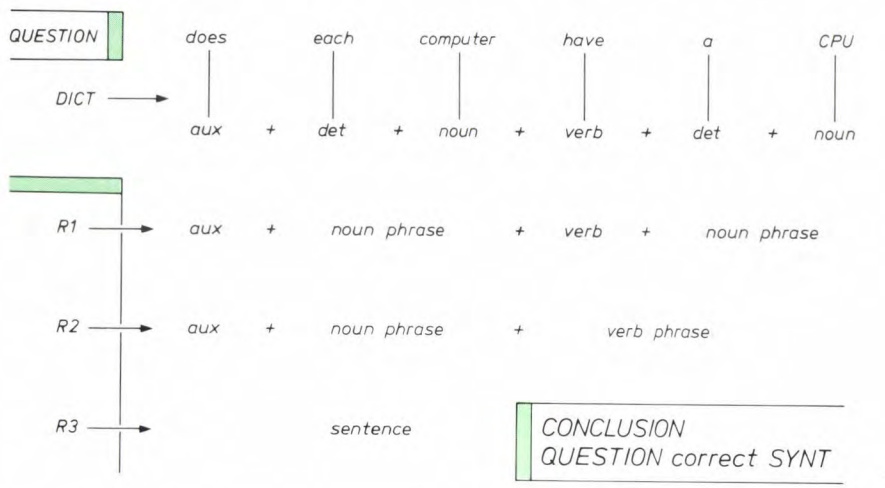
where a 'determiner' is an article, a demonstrative pronoun, etc.

ement. Not until the conditions for the syntactic features have been fulfilled can the reduction be carried out.

When parsing begins, the sentence elements are the individual words. In our example the syntactic categories of these are:

auxiliary, determiner, noun, verb, determiner, noun. (1)

In this sequence 'determiner' occurs twice followed by 'noun'. Rule R1 is applicable to this, provided the



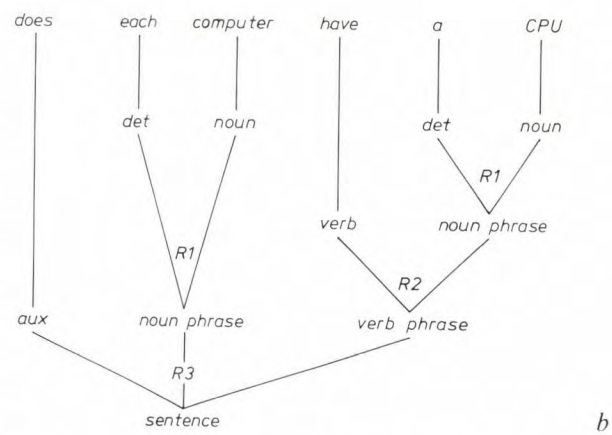


Fig. 6. *a*) Example of the reduction process during the translation operation English → EFL. The process provides the syntactic structure of the sentence QUESTION expressed in natural English. The English words are replaced by means of the lexicon DICT by the syntactic categories (Table I). The reduction rules R1, R2 and R3 reduce the string of syntactic categories to the category 'sentence'. On the basis of this result, QUESTION is recognized by the system as a grammatically correct sentence. *aux* auxiliary. *det* determiner. *b*) The tree diagram giving the syntactic structure of the reduced sentence in (*a*). The 'nodes' resulting from the application of R1, R2 and R3 are labelled as appropriate.

A reduction rule indicates that if in a sentence we have a succession of elements with syntactic categories like those to the left of the arrow (e.g. 'determiner' and 'noun' in rule R1), these successive elements can be reduced to a single sentence element with the syntactic category to the right of the arrow ('noun phrase' in R1). To each reduction rule certain conditions are attached which the syntactic features of the sentence elements must satisfy, and also an instruction for determining the syntactic features of the newly formed sentence elconditions for the syntactic features have been satisfied. In this case the condition is that the value of the attribute "Exp(ected) number" of the determiner shall be equal to the value of the attribute "Number" of the noun. In the example the value of both these attributes is singular, and therefore the condition is fulfilled. If the sentence to be parsed contained 'each computers' instead of 'each computer', the reduction rule R1 would not be applicable and the sentence would finally be treated as ungrammatical.

The reduction rule R1 can be applied to (1) twice. The successive sentence elements determiner + noun are joined to form a single sentence element of the category 'noun phrase'. Both noun phrases are given the syntactic feature "Number" : singular, originating from their nouns; the first also receives the feature "Quant": universal, and the second the feature "Quant": existential, originating from their determiners.

In this way (1) has been reduced to:

auxiliary, noun phrase, verb, noun phrase. (2)

Here we see the sequence verb + noun phrase, to which the reduction rule R2 is applicable. In what follows we shall take no further account of the syntactic features.

Rule R2 now reduces the string of sentence elements (2) to:

auxiliary, noun phrase, verb phrase. (3)

Finally, application of rule R3 leads to:

sentence. (4)

The sentence "Does each computer have a CPU?" is thus treated by this grammar as grammatically correct. The reduction process has assigned to the sentence a syntactic structure (*fig. 6a*) that can be represented in the form of a tree diagram (fig. 6*b*). The branchings of the tree correspond to the reduction rules applied.

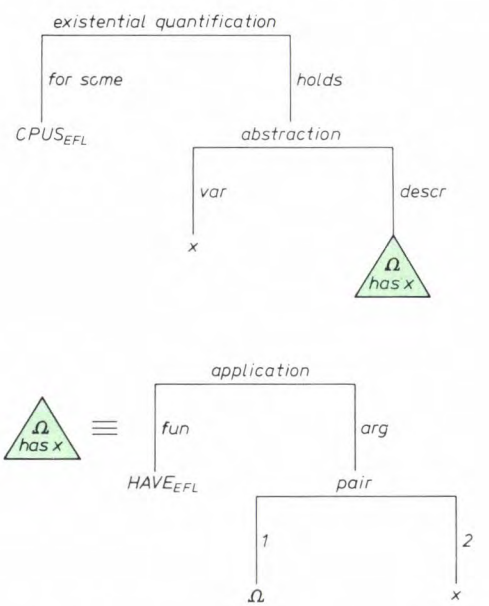


Fig. 7. *a*) Translation of the phrase ". . . have a CPU", which is part of the sentence QUESTION (fig. 6), into EFL. The construction shown results from the operation of the composition rules associated with the reduction rules R1 and R2. *b*) The subconstruction for which the triangle in (*a*) is used as an abbreviated notation. The primitive term Ω relates to a subject, not yet named, associated with the verb in ". . . have a CPU".

As noted, each reduction rule has a composition rule associated with it, which composes the EFL construction corresponding to the newly formed sentence element. This composition begins with the EFL terms associated with the referential words. In our example these are the language constants COMPUTERS$_{EFL}$ and CPUS$_{EFL}$, which both have the type of a set, and the language constant HAVE$_{EFL}$, which has the type of a two-place predicate. The word 'two-place' indicates that the predicate in this case is applicable to a *pair* of arguments (which corresponds to the fact that the verb "to have" has a subject and a direct object).

Let us now look first at the part of the sentence ". . . have a CPU". The reduction rules R1 and R2 assign to this the syntactic category 'verb phrase'. Application of the composition rules leads to the EFL construction given in *fig. 7*. The primitive term Ω contained in this relates to the subject of ". . . have a CPU", which has not yet been named.

The reduction rule R1 has led here to the noun phrase "a CPU" (fig. 6). The primitive term corresponding to the English word CPU is CPUS$_{EFL}$ (Table I). The composition rule associated with R1 merely passes on this primitive term, so that CPUS$_{EFL}$ is also the EFL construction that corresponds to the newly formed noun phrase.

In the next step, R2 leads to a verb phrase. The composition rule associated with R2 constructs an existential-quantification-branching and places CPUS$_{EFL}$ as a subconstruction under the selector "for some" of this branching. The existential-quantification-branching is chosen because the syntactic feature "Quant": existential belongs to the noun phrase. This feature of the noun phrase of the sentence comes from the determiner "a" which occurs in it (see Table I).

Under the selector "holds" of the existential-quantification-branching a function is constructed by means of an abstraction-branching, with a variable (x) and an application-branching. In the application-branching the two-place predicate HAVE$_{EFL}$ is applied to the argument pair Ω and x. The argument x corresponds, via the abstraction-branching and the existential-quantification-branching, to CPUS$_{EFL}$ (which itself corresponds to the direct object "a CPU"). The argument Ω stands for the as yet unspecified variable that corresponds to the subject. The construction shown in fig. 7*a* may be paraphrased as follows: 'The set of CPUs contains at least one (in the figure: "existential quantification") element x for which it can be said ("abstraction") that the expression "Ω has x" is true of an unknown element Ω'. The expression "Ω has x" is shown in more detail in fig. 7*b* as 'the result of the application of the two-place predicate HAVE$_{EFL}$ to the pair ("pair") of elements (Ω,x)'.

Finally, as soon as the reduction rule R3 becomes operative, the remaining parts of the sentence — the auxiliary verb "Does" and the noun phrase "each computer" — also become involved in the process. The associated composition rule, for reasons similar to those applied in the case CPUS$_{EFL}$, places the sub-construction COMPUTERS$_{EFL}$ under a universal-quantification-branching (*fig. 8*). The primitive term Ω from fig. 7, as yet unknown, can now be filled in (in the figure it is the variable y). The construction in fig. 8 is the expression in EFL of the whole of the original question. We can express this construction in words as follows: 'For each computer y it holds that there is at least one CPU x such that "y has x" is true'.

The EFL construction in fig. 8 has the type truthvalue. Our example thus obeys the general rule that Yes/No questions lead to EFL constructions of this type. WH questions ("what", "which", "who", etc.) lead to EFL constructions that have the type string. This is because WH questions are interpreted as questions about a *description* of objects, not as questions about the objects themselves. What that description turns out to be depends on the kind of object asked about and on the information present. If the question is about companies, the description consists of a list of names and addresses (a string of symbols); if the question is about CPUs, the description is a list of models. The system decides on this during one of the translation operations following EFL.

For the remainder of this article it is important to bear in mind that a complete EFL construction — a construction that corresponds to the *complete* question — always has a formal type: truthvalue, string or integer (the type integer in the case of a WH question about a quantity).

From EFL into WML and from WML into DBL

The translation operations from EFL, via intermediate languages, into DBL need only operate on the referential language constants and on the variables. This is connected with the fact that all the artificial languages in the PHLIQA 1 system, as mentioned earlier, have the same branchings and also the same formal language constants. In the translation operation English → EFL the system has already taken all the decisions concerning the logical aspects of meaning. The other translation operations are concerned with the referential aspects of meaning, in particular with the question of the relations that the EFL language constants, such as COMPUTERS$_{EFL}$, CPUS$_{EFL}$ and HAVE$_{EFL}$ in fig. 8, have with the language constants of WML and DBL.

Every operation involved in the translation from a source language into a target language — the first source language is EFL, the final target language is DBL — is carried out by a *referential convertor* [5]. The number of referential convertors in the system thus depends directly on the number of intermediate languages introduced. The application of such a convertor to a construction t of a source language implies that the language constants and the variables of t are replaced by constructions and variables of the target language, respectively.

A referential convertor operates with two sets of transformations: *constant transformations*, which transform the constants of the source language into constructions of the target language and *variable transformations*, which transform the variables of the source language into variables of the target language. Both sets of transformations have to fulfil certain type requirements. These requirements follow from a third set of transformations, called *type transformations*.

The *type transformations* show how the types of the appropriate source language are related to the types of the target language. They have the form

$$\alpha_i \Rightarrow \phi_i,$$

where α_i is an atomic type of the source language and ϕ_i is an atomic or compound type of the target language.

The *constant transformations* have the form

$$C_j \Rightarrow e_j,$$

where C_j is a language constant of the source language and e_j is a construction of the target language. A constant transformation works 'in accordance with' the type transformation. This means that the type of the construction e_j 'is contained in' the *transformed* type of the associated language constant C_j. (As stated, the relation 'is contained in' between types will not be dealt with in this article; it is sufficient here to know that a type α is certainly 'contained in' a type β if $\alpha = \beta$.) The transformed type of C_j is the type that occurs when the atomic types α_i found in the type of C_j are replaced by the types ϕ_i that correspond to them in accordance with the type transformations.

The *variable transformations* have the form

$$x_k \Rightarrow y_k,$$

where x_k is a variable of the source language and y_k is a variable of the target language. The type of this

[5] S. P. J. Landsbergen, Establishing the 'soundness' of a question-answering system, to appear in Proc. 7th Int. Conf. on Computational Linguistics, Bergen (Norway) 1978.

variable of the target language is the *transformed* (as described above) type of the associated variable x_k.

It can be shown that the application of a referential convertor to a syntactically correct construction of a source language always leads to a syntactically correct construction of the target language. ('Syntactically correct' means that the type of the construction is unequal to empty.) The system thus does not produce any surprises in this respect. The proof, which we shall not give here, is based on the type requirements for the constant transformations and the variable transformations and also on certain conditions that the type function T, operating between construction language and type language, fulfils for the various branchings.

The semantic aspects of the referential convertors

In addition to the syntactic aspects discussed, the translation operations also have semantic aspects, of course. In our system we distinguish between two classes of referential convertors with regard to semantics. The first consists of the synonymic referential convertors, or *synonymy convertors*, where the value of the constructions is preserved during the translation operation, and the second consists of the identifying referential convertors, or *identification convertors*, where there is no equality but a one-to-one relation between the values.

In the case of the synonymy convertors the constant transformations $C_j \Rightarrow e_j$ mentioned above are *value-preserving*. More precisely, a constant transformation is only correct for those pairs of interpretations of the source language and target language where C_j and e_j have the same value.

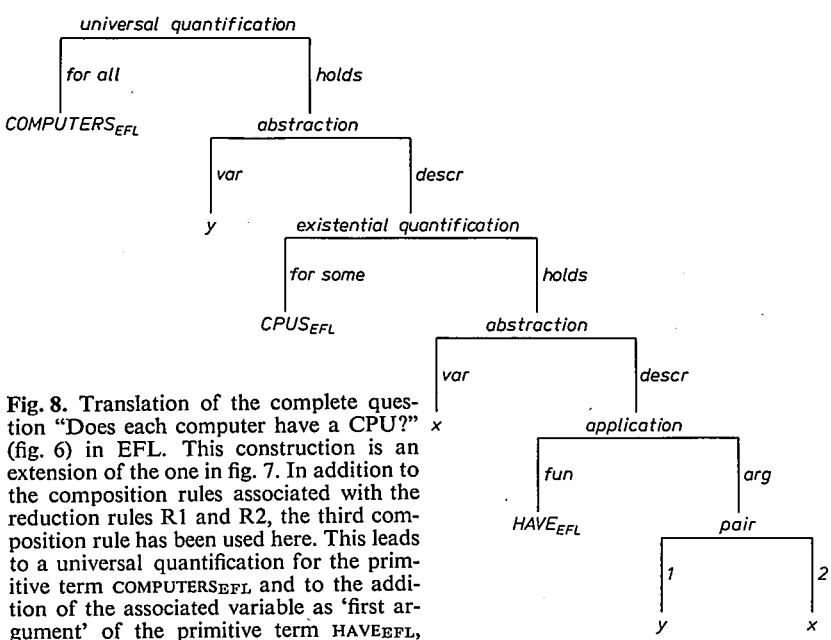


Fig. 8. Translation of the complete question "Does each computer have a CPU?" (fig. 6) in EFL. This construction is an extension of the one in fig. 7. In addition to the composition rules associated with the reduction rules R1 and R2, the third composition rule has been used here. This leads to a universal quantification for the primitive term COMPUTERS$_{EFL}$ and to the addition of the associated variable as 'first argument' of the primitive term HAVE$_{EFL}$, which has the type of a two-place predicate.

In designing a system like PHLIQA 1 it is necessary to determine whether each constant C_j of the source language has the same meaning as the corresponding construction e_j of the target language. This is done by establishing whether C_j and e_j do in fact have the *same* value for every possible state of the subject domain. In PHLIQA 1 this meant that the meaning of English words had to be brought into relation with the meaning of the primitive terms contained in the data base, since EFL is the first source language and DBL the final target language. The correctness of the established relations is easy to check in this case for the designers — and in principle also for others — because these relations are established by simple rules that transform only the *primitive terms* of a source language and not the larger constructions.

It can be shown that every *complete* construction of the source language (i.e. the representation of the meaning of a complete question) is transformed by the convertor into a target-language construction that has the same value — this only applies of course to interpretations of the source language and the target language for which the constant transformations are value-preserving. The fact that the synonymy convertors, and also the other convertors not yet dealt with, possess this feature gives the system its reliability. The term 'reliability' means here that correct constructions are correctly translated as soon as primitive terms are correctly translated.

There are certain kinds of semantic relationships between the artificial languages of PHLIQA 1 for which the synonymy convertors are not sufficient. For this reason *identification convertors* were introduced. These are required for situations where it *is* possible to 'speak' about objects of a particular type in the source language, whereas this is *not* possible in the target language. This does not imply that the constructions of the source language are untranslatable in such situations: the questioner is not interested in the objects themselves as much as in their *properties* that the system is capable of presenting in language form. In many cases such *properties* can be 'spoken' about in the target language, thus establishing a semantic relationship with the source language.

The operation and the usefulness of the identification convertors for semantic relationships of this kind will be illustrated by the following example. WML contains the atomic type cpu, which relates to the type of individual, physical CPUs; on the other hand DBL does not contain the type cpu, nor a type synonymous

with it. Language constants in WML that relate to CPUs, whose type thus contains the atomic type cpu, cannot therefore be translated into DBL with the synonymy convertors. This is because no DBL constructions can be made with a value equal to that of the WML language constants referred to. Nevertheless, questions about CPUs in general can be answered, because the information about a CPU is to be found in the record of the computer configuration of which this CPU is a part, or in the record of the model of the CPU (of record type "CPU MODEL", see fig. 6, part I) that is linked with the record of the computer configuration (with the record type "CONFIGURATION"). Identification transformations can now translate all language constants relating to CPUs into constructions about *configurations*, with the addition of a mark that distinguishes them from constructions that are 'really' about configurations. In this way CPUs are thus *indirectly* identified, by means of the computer configuration to which they belong.

In these identification transformations the value of the translated primitive terms is not preserved. There is, however, an 'isomorphism' between the values of these terms and the values of their translations in the target language.

Isomorphism — a mathematical concept — refers to similarity in form, a structural identity, of two sets; it amounts to stating that the relations between the elements of the one set agree with the relations between the corresponding elements of the other set, although the elements themselves do not have to be equal to each other.

On the subject of isomorphism two further remarks should be made. Firstly, it can be shown that if isomorphism exists between the values of the *primitive terms* and their translations it will also exist between the values of complete *constructions* and their translations. Secondly, for the formal types, e.g. truthvalue, integer, string, isomorphism amounts to *identity*. As we have seen in the section discussing the translation operation English→EFL, the complete constructions in EFL always receive a formal type. In the subsequent translation operations the complete constructions preserve their formal type. From the foregoing it follows that, like a synonymy convertor, an identification convertor is value-preserving for *complete* constructions.

The instantiation convertor

In the discussion on semantics at the end of the section dealing with the artificial languages we pointed out that primitive terms of EFL are sometimes ambiguous ('can have more than one value'). In translation operations from EFL into the next, unambiguous construction language this has the consequence that each primitive term is translated into a *set* of expressions in that language, called *instantiations* of the EFL term. For dealing with such ambiguity the system possesses a special convertor, called the *instantiation convertor*. This generates all possible combinations of instantiations of the different primitive terms in the original EFL construction, and then finds which combinations are well-formed in regard to type. Only these well-formed instantiations are passed by the instantiation convertor to the next convertor.

Instead of a single expression the instantiation convertor thus yields a set of expressions, and in this respect it differs fundamentally from the convertors already discussed. The set can be empty in which case the question is meaningless (see fig. 4, part I).

During the instantiation the types of the EFL terms are limited to those of WML, to which the actual — limited — subject domain corresponds. For example, CPUs have a price in that subject domain whereas cities do not. In EFL, however, both "expensive CPUs" and "expensive cities" are correct. During the instantiation the system concludes that the combination "expensive CPUs" is permissible but not "expensive cities".

The formal convertors

The convertors so far described constitute in principle a powerful mechanism for all translation operations from EFL into DBL, ensuring both syntactic and semantic correctness. This does not necessarily mean, however, that these convertors process a submitted question in the *simplest* way. Owing to their 'locality' (they operate on only one primitive term at a time) the referential convertors often generate unnecessarily awkward constructions. To simplify such expressions the system uses *formal convertors*. These consist of a set of formal transformations that are not local but operate on larger parts of an expression — within the same language. A formal transformation reduces a construction to a simpler one that is logically equivalent to it, that is to say, it has the same value as the original construction for all interpretations of the language.

The application of a formal transformation can usefully be compared with the simplification of an algebraic expression, e.g. $\sqrt{a^2} \Rightarrow a$, since the constructions with which the PHLIQA 1 system works have much in common with algebraic expressions.

The formal transformations may occasionally have an advantageous side-effect in the system, since they may transform constructions with an untranslatable primitive term into a logically equivalent construction (in the same language) in which this term does not occur. This has further increased the translation capacities of PHLIQA 1, since it makes translation possible in cases where the identification transformations and the synonymy transformations alone are unable to produce a translation.

The evaluator

After a question has been translated into DBL by the convertors discussed above, the resultant representation of its meaning in DBL is passed to the evaluator (see fig. 3, part I). This computes the value of the construction and produces it as the answer to the question. The value of a construction has so far been defined in mathematical terms: an interpretation assigns values to the language constants, and the semantic rules of the branchings show how the value of the construction follows from the values of the language constants contained in it.

To calculate the value of a construction in practice the evaluation program must make use of actual *representations* of values. In PHLIQA 1 these representations are again expressions of a construction language, called the *value language*. The primitive terms of this value language are unique names of records and formal constants (the truth values TRUE and FALSE, numbers and strings of symbols such as P-H-I-L-I-P-S). The value language contains only a few branchings, one of which permits a set to be represented by enumerating its elements.

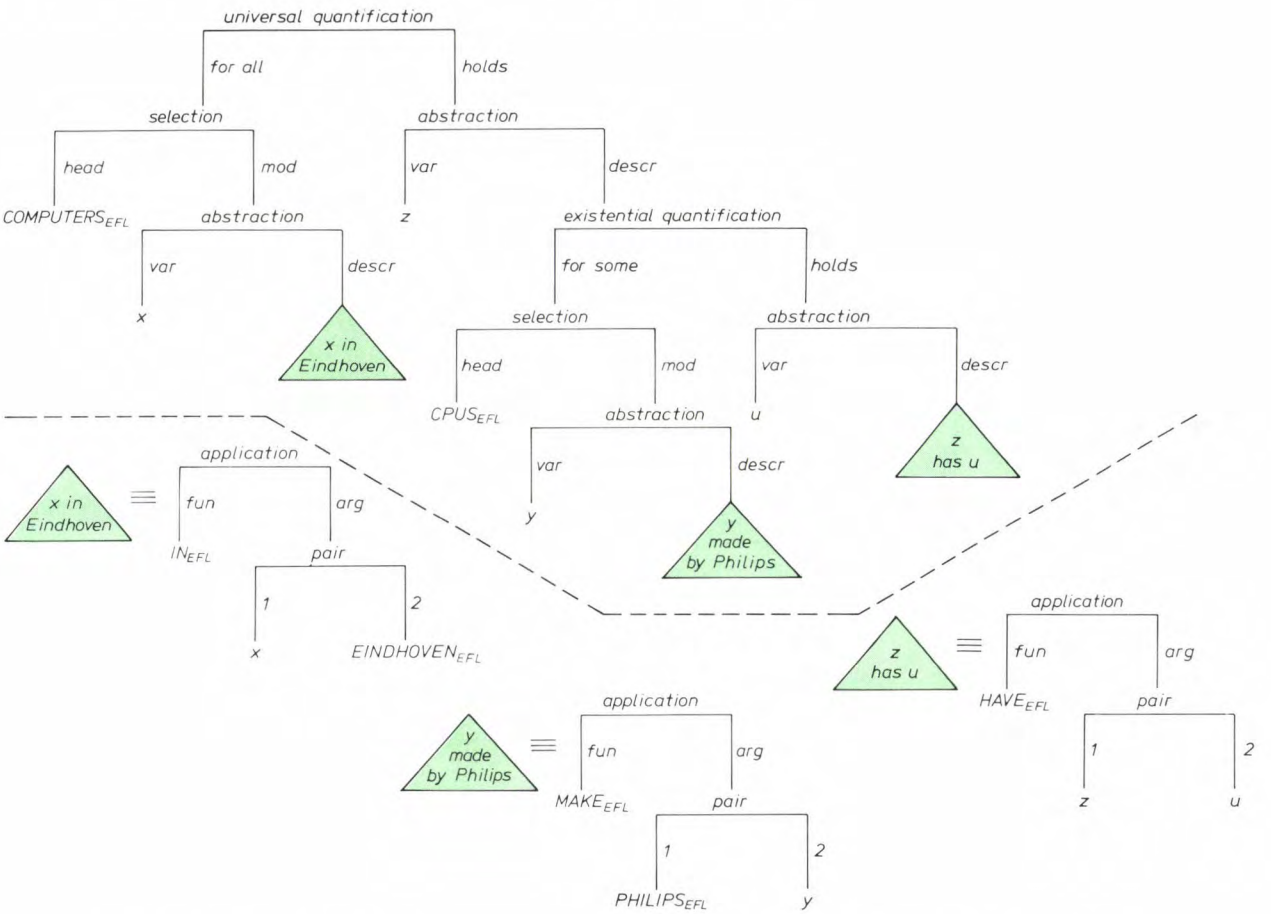


Fig. 9. Translation of the question chosen as an example: "Does each computer in Eindhoven have a CPU made by Philips?" in the English-oriented Formal Language (EFL). The complete construction is shown above the dashed line, and the subconstructions for which the triangles are used as an abbreviated notation are shown below. The construction may be paraphrased as follows: 'Is it correct for every computer z located in Eindhoven that at least one CPU u made by Philips exists such that "z has u" is true?' The resultant branching diagram is the representation of the meaning (in EFL) of the question.

Assigning values to the language constants of a construction, which is what the evaluator does with the aid of the data base, amounts in the terminology of the value language to applying a synonymy convertor. A transformation from this synonymy convertor replaces a language constant (of DBL) by an expression in the value language that represents its value unambiguously. Thus, for example, the language constant CONFIGURATIONS$_{DBL}$ is replaced by a list of names of records with the record type "CONFIGURATION".

In the evaluator the semantic rules of the branchings take the form of formal transformations that convert a branching whose subconstructions are expressions in

the value language into a new expression in the value language. This ultimately results in an expression in the value language that represents the value of the complete DBL construction.

The evaluator can thus be regarded as a synonymy convertor followed by a formal convertor. In fact there is a third component, called the *answer formulator*, which presents the expression obtained in the value language in a readable form to the questioner. The answer formulator presents the value TRUE as "YES" in an external format and the value FALSE as "NO", and gives lists of items such as names and addresses in clearly tabulated form.

An example of the translation operations EFL→WML and WML→DBL

Having described the formal means available for the translation operations, we shall conclude by giving a rather simplified example to show how the system can use these mechanisms in the translation process.

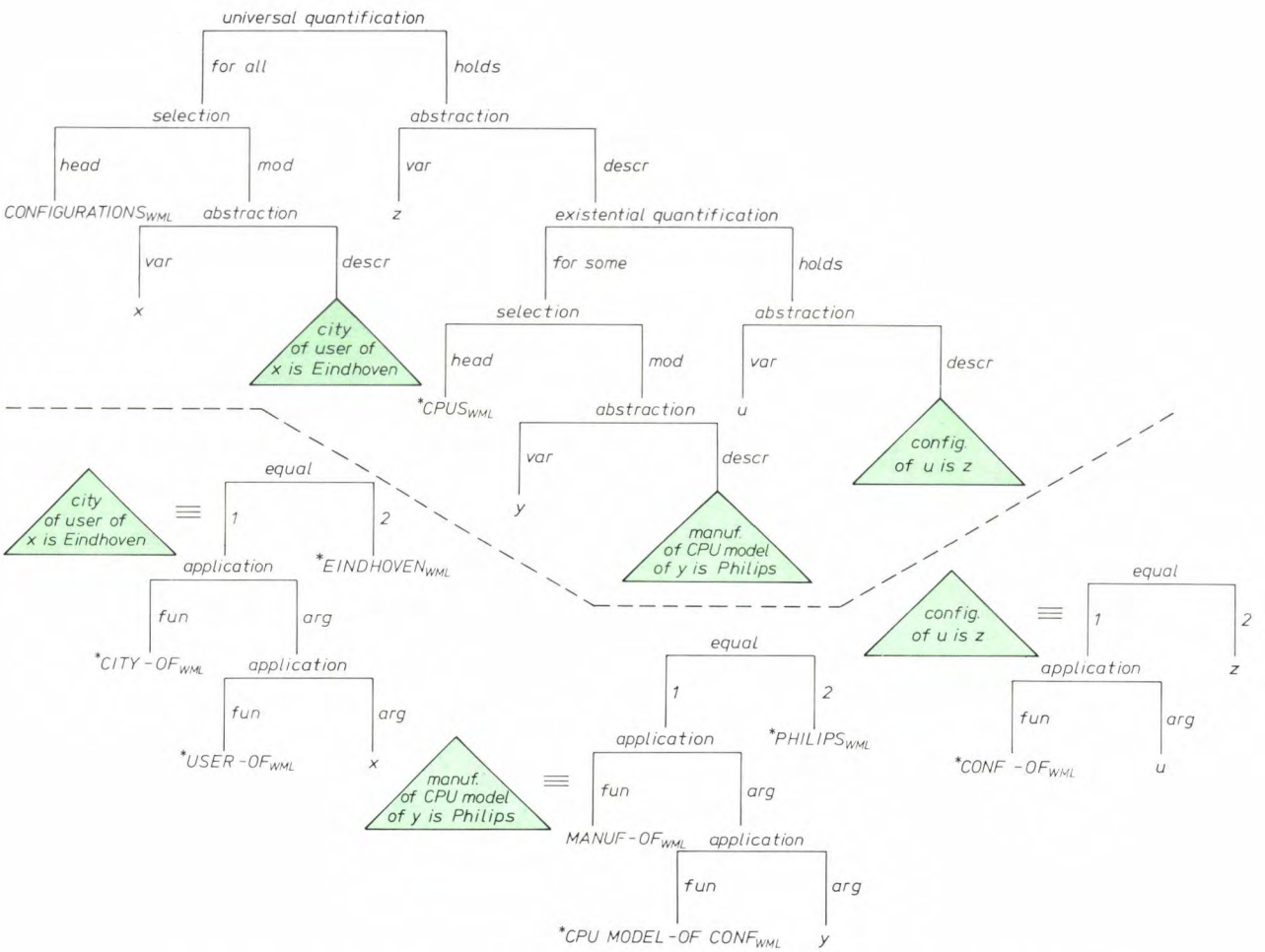


Fig. 10. Translation of the question "Does each computer in Eindhoven have a CPU made by Philips?" in the World-Model Language (WML). The complete construction is shown above the dashed line, and the subconstructions for which the triangles are used as an abbreviated notation are shown below. This branching diagram, obtained from the construction in fig. 9 by the translation operation EFL→WML, is also a representation of the meaning, but now in WML, of the original question. An asterisk in front of a primitive term indicates that it has to be translated into DBL by means of an identification transformation. Synonymy transformations operate on the other primitive terms.

The example is the question "Does each computer in Eindhoven have a CPU made by Philips?". *Fig. 9* shows the EFL construction for this question. Apart from the additions "in Eindhoven" and "made by Philips", which are displayed in an abbreviated form, as triangles, this expression has the same structure as that in fig. 8, the example of the translation operation

English → EFL. As in our treatment of the "parser" program, the procedure shown is simplified.

The principal mechanisms for translating in this example are:

— The instantiation convertor. The primitive terms COMPUTERS$_{EFL}$ and HAVE$_{EFL}$ are ambiguous, since the word "computers" can refer either to CPUs or to complete computer configurations. The meaning of the verb "to have" can include "to possess" and "to have as a part". Only one of the possible combinations of the instantiations in our example leads to a construction of a permissible type, and this is the combination where COMPUTERS$_{EFL}$ is instantiated to the primitive term CONFIGURATIONS$_{WML}$, and where HAVE$_{EFL}$ is instantiated to the primitive term HAVE-AS-PART (a term of an intermediate language between EFL and WML).

— A synonymy convertor. The language constant HAVE-AS-PART, which has the type of a two-place predicate, does not correspond to one WML constant, but is translated into a subconstruction, called C_{WML} here, in which this predicate is written out for various possible combinations of argument types. In C_{WML} we find, for example, the language constant COUNTRY-OF$_{WML}$ with the type (city → country), a function that for each city gives the country where it is situated, and we also find the language constant CONF-OF$_{WML}$, with the type (cpu → conf), a function that for each CPU gives the configuration to which it belongs.

— A formal convertor. This reduces, for instance, the construction C_{WML} written in full to the part that is relevant in this context, i.e. the function CONF-OF$_{WML}$.

The operation of these three kinds of convertors leads to the WML construction shown in *fig. 10.*

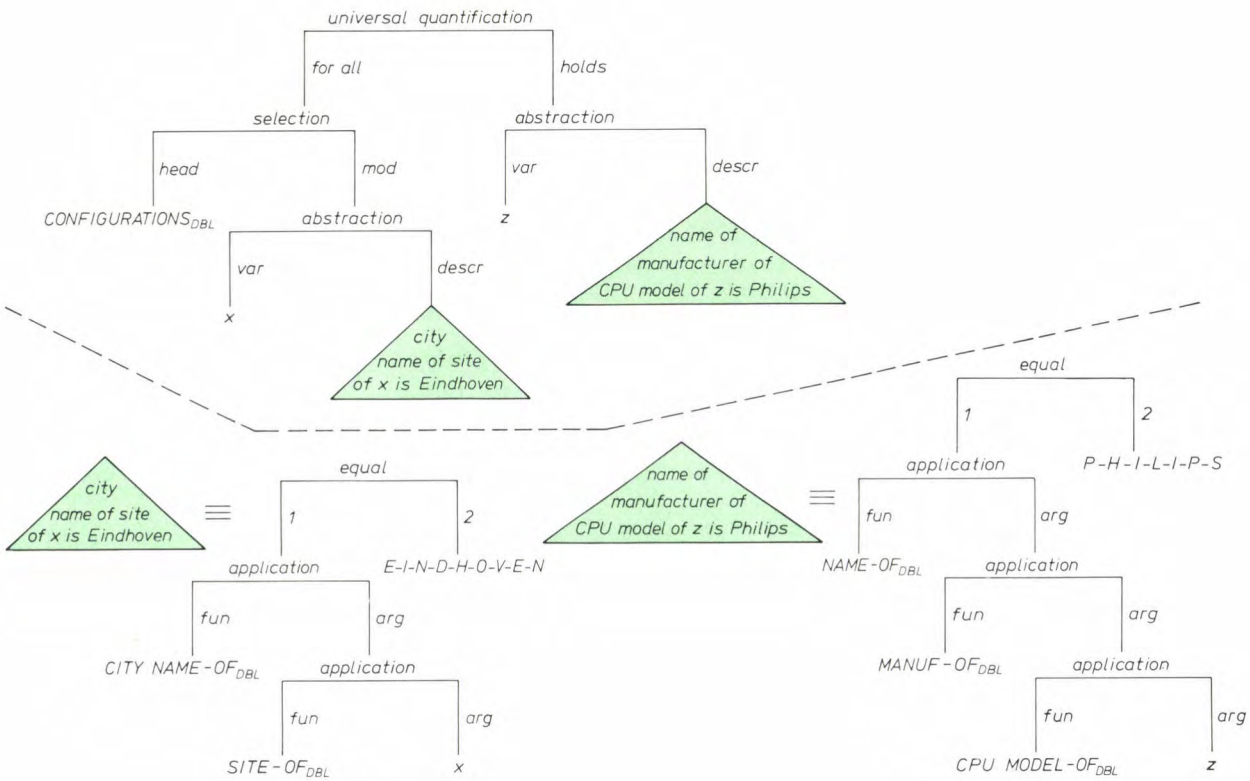


Fig. 11. Translation of the question "Does each computer in Eindhoven have a CPU made by Philips?" in the Data-Base Language (DBL). The complete construction is shown above the dashed line, and the subconstructions for which the triangles are used as an abbreviated notation are shown below. This branching diagram, obtained from the construction in fig. 10 by the translation operation WML→DBL, is the third representation of the meaning of the original question. The system can assign to this translation the value TRUE or FALSE, on the basis of the data contained in the data base. The answer "YES" or "NO" then appears on the screen.

The application of the translation operation WML→DBL to the construction in fig. 10 leads to the construction given in *fig. 11*. The most striking difference compared with fig. 10 is the absence of the existential quantification for CPUs. This is achieved by successively using an identification convertor, which reformulates primitive terms relating to CPUs in terms of configurations, and a formal convertor, which simplifies the expression. Fig. 11 incidentally shows only one of the many DBL constructions that are possible for the given question and are logically equivalent. Just which construction is generated depends on the formal transformations available.

In the DBL construction shown in fig. 11 the primitive term CONFIGURATIONS$_{DBL}$ stands for the set of records with the record type "CONFIGURATION" (the computer configurations). The primitive term SITE-OF$_{DBL}$ is a function that, when applied ("application" in the figure) to a given record of the record type "CONFIGURATION", gives a record of the record type "SITE" (of the establishments where the configuration is used). The function CITY NAME-OF$_{DBL}$ gives for a given record of the record type "SITE" the name of the city where the establishment is situated. The evaluator carries out these function applications and investigates whether the city thus found for each configuration (the variable x in the figure) is "equal" to the string E-I-N-D-H-O-V-E-N, which is found under the equal-branching as a primitive term (that has the type string). (A primitive term of the type string is a formal language constant, and its value is therefore the same for each interpretation. The value *is* the string itself.)

Only records of the record type "CONFIGURATION" that fulfil this requirement are included in the value of the selection-branching (the configurations in Eindhoven). The part of the construction on the right in fig. 11 is treated similarly by the system.

In this way the evaluator computes the value of the construction. Depending on the data, the value is TRUE or FALSE. The answer formulator displays this result on the screen as "YES" or "NO".

Summary. The experimental computer program PHLIQA 1 enables a computer of medium capacity (e.g. the Philips P1400) to answer certain questions put to it in natural English. The answers are taken from a data base, organized in accordance with the CODASYL standards, which contains fictitious data about European computer installations and their users. The question, typed in by means of a keyboard with a visual display, undergoes a series of translation operations. The first, English→EFL (English-oriented Formal Language), determines the logical relations between the various parts of the sentence. The second, EFL→WML (World-Model Language), reformulates the question in terms of the subject domain and checks whether the question makes sense in that domain. The third translation operation, WML→DBL (Data-Base Language), provides the link between the contents of the question and the chosen data base. The computer calculates the value of the expression in DBL and produces an answer on the display screen ("YES"; "NO"; a number; a list of names and addresses). If ambiguity exists the computer can systematically trace different meanings; if the questions are unanswerable the computer indicates at what language level the answering procedure breaks down and why. The organization of the data base and the syntax of the artificial languages used are described. All the artificial languages have the same grammar; they differ, however, in their primitive terms. The language expressions are referred to as constructions; the extent to which they are 'well-formed' is determined by means of type rules. The syntactic structure of the constructions corresponds to their semantic structure. For translating into EFL the computer uses a lexicon, reduction rules and semantic composition rules. To translate from EFL into WML and DBL the computer uses referential and formal transformations. The various parsing and translating mechanisms are explained with a number of examples. An example illustrates how the whole sequence of translation operations provides, in the form of a three-stage analysis followed by a value calculation, an answer to a particular question. The article is in two parts. Part I gives a general introduction to the organization and external behaviour of the system, and part II deals more with technical particulars, the syntax and the semantics of the artificial languages and the mechanisms of the translation operations.

Scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands *E*
Philips Research Laboratories, Redhill, Surrey, England *R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany *H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium *B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. *N*

G. Aaron: Codage bidimensionnel d'images monochromes pour la télécopie.
Acta Electronica **21**, 39-53, 1978 (No. 1). *L*

M. Auphan & H. Dormont: Pulsed acoustic radiation of plane damped transducers.
Ultrasonics **15**, 159-168, 1977 (No. 4). *L*

R. W. Bach, H. A. Friedrichs (both with Rhein.-Westf. T.H., Aachen) **& H. Rau:** p-V-T relations for HCl-H_2O mixtures up to 500 °C and 1500 bars.
High Temp. - high Press. **9**, 305-312, 1977 (No. 3). *A*

G. Bakker: Ervaringen met een hydrodynamisch precisie-glijlager.
Mikroniek **17**, 133-136, 1977 (No. 6). *E*

H. Bantelmann, H. Kalis & A. Montag: MIMUS — a flexible multiprocessor system.
Proc. IMACS (AICA)-GI-Symp. on Parallel computers — parallel mathematics, Munich 1977, pp. 173-177. *H*

U. Bergmann, A. Glättli, K. Klose, L. Spiess & I.-D. Stromberger: Microfilm recording without after-treatment.
J. photogr. Sci. **25**, 159-163, 1977 (No. 4). *H*

H. J. Bex: Crossmodulation in ring modulators.
Proc. IEE **124**, 1155-1159, 1977 (No. 12). *A*

P. A. Blakey: Simple, efficient, self-starting TRAPATT oscillators.
Proc. 7th Eur. Microwave Conf., Copenhagen 1977, pp. 492-496. *R*

S. Boronkay, P. Gustin (MBLE, Bruxelles) **& D. Rossier:** Analyse de documents pour la télécopie.
Acta Electronica **21**, 55-70, 1978 (No. 1). *L*

G. Bosch (Philips Video Division, Eindhoven): Anomalous current distributions in power transistors.
Solid-State Electronics **20**, 635-640, 1977 (No. 7).

M. J. G. Braken: Aangepaste werkstukconstructies, lastoortsvormen en gereedschap voor TIG-lassen.
Mikroniek **17**, 138-142, 1977 (No. 6). *E*

P. C. Brandon & G. N. van Boekel-Mol: β-bromo-β-nitrostyrene as a facile and photosystem I-specific electron acceptor.
FEBS Letters **80**, 201-204, 1977 (No. 1). *E*

J. C. Brice: Trends in liquid phase epitaxy.
1976 Crystal growth and materials, ed. E. Kaldis & H. J. Scheel, pp. 571-604; North-Holland, Amsterdam 1977. *R*

J.-J. Brissot: Recent achievements in low-cost silicon solar cells.
1976 Crystal growth and materials, ed. E. Kaldis & H. J. Scheel, pp. 795-822; North-Holland, Amsterdam 1977. *L*

A. Broese van Groenou: Wat gebeurt er bij slijpen van glas en keramiek?
Mikroniek **17**, 121-128, 1977 (No. 6). *E*

A. Broese van Groenou: The high speed size effect in grinding: the role of heat generation.
Wear **44**, 203-211, 1977 (No. 2). *E*

A. Broese van Groenou, J. D. B. Veldkamp & D. Snip (Philips Centre for Technology, Eindhoven): Scratching and grinding parameters of various ferrites.
J. Physique **38**, C1/285-289, 1977 (Colloque C1). *E*

L. K. Brundle, M. D. Coleman & S. R. Longley: Apertured ferrite slab subsidiary resonance power limiters.
Proc. 7th Eur. Microwave Conf., Copenhagen 1977, pp. 647-651. *R*

E. Bruninx: The coprecipitation of Cu, Ag, Au; Ge, Sn, Pb; Mn and Ca with ferric hydroxide.
Philips J. Res. **33**, 264-280, 1978 (No. 5/6). *E*

R. Bruno, W. Hermann, H. Hörster, R. Kersten & F. Mahdjuri: Analysis and optimization of solar hot water systems.
Int. J. Energy Res. **1**, 329-340, 1977 (No. 4). *A*

E. Bunge: Automatic speaker recognition by computers.
Proc. IEEE Computer Soc. Conf. on Pattern recognition and image processing, Troy N.Y. 1977, pp. 88-93. *H*

D. J. Burnett & H. R. Sethi: Packet switching at Philips Research Laboratories.
Computer Networks **1**, 341-348, 1977 (No. 6). *R*

K. H. J. Buschow: Note on the magnetic properties of Y-Mn compounds and their hydrides.
Solid State Comm. **21**, 1031-1033, 1977 (No. 11). *E*

K. H. J. Buschow, M. Brouha, H. J. van Daal & A. R. Miedema: Valency changes of Ce, Yb and U in intermetallic compounds with Ni and Cu.
Valence instabilities and related narrow-band phenomena, ed. R. D. Parks, pp. 125-136; Plenum Press, New York 1977. *E*

K. H. J. Buschow & R. C. Sherwood (Bell Laboratories, Murray Hill, N.J.): Hydrogen absorption and magnetic properties of rare earth intermetallics.
IEEE Trans. **MAG-13**, 1571-1573, 1977 (No. 5). *E*

S. R. Chinn (M.I.T., Lexington Mass.) **& W. K. Zwicker:** Flash-lamp-excited NdP_5O_{14} laser.
Appl. Phys. Letters **31**, 178-181, 1977 (No. 3). *N*

T. A. C. M. Claasen, W. F. G. Mecklenbräuker & J. B. H. Peek: Een overzicht van niet-lineaire effecten in rekursieve digitale filters.
T. Ned. Elektronica- en Radiogen. **42**, 105-109, 1977 (No. 4). *E*

T. A. C. M. Claasen & J. B. H. Peek: Een eenvoudige digitale korrelatie-ontvanger voor toonsignalen.
T. Ned. Elektronica- en Radiogen. **42**, 93-96, 1977 (No. 4). *E*

D. J. Coe: The lateral diffusion of boron in polycrystalline silicon and its influence on the fabrication of sub-micron MOSTs.
Solid-State Electronics **20**, 985-992, 1977 (No. 12). *R*

D. J. Coe, H. E. Brockman & K. H. Nicholas: A comparison of simple and numerical two-dimensional models for the threshold voltage of short channel MOSTs.
Solid-State Electronics **20**, 993-998, 1977 (No. 12). *R*

M. Davio & J.-P. Deschamps: Synthesis of some partially symmetric Boolean functions by ternary algebra.
Electronics Letters **13**, 473-474, 1977 (No. 16). *B*

P. A. Devijver: Entropies of degree β and lower bounds for the average error rate.
Information and Control **34**, 222-226, 1977 (No. 3). *B*

A. M. van Diepen & F. J. A. den Broeder: Effect of thermal stress on the easy axis of amorphous $Fe_{40}Ni_{40}P_{14}B_6$.
J. appl. Phys. **48**, 3165-3166, 1977 (No. 7). *E*

C. Dierieck: Some remarks on *H*-sets in linear approximation theory.
J. Approx. Theory **21**, 188-204, 1977 (No. 2). *B*

J. Donjon: Large screen projection using Titus tubes.
Television (J. Roy. Telev. Soc.) **16**, No. 9, pp. VIII-IX, May/June 1977. *L*

E. Dormann (T. H. Darmstadt) **& K. H. J. Buschow:** Does nickel contribute to magnetic ordering in $GdNi_2$?
Physica **86-88B**, 75-76, 1977 (Part I). *E*

F. Dumont, D. Rossier, P. Bonhomme*, A. Beorchia* & B. Meunier* (* Faculté des Sciences de Reims): Phototitus: image intensifier and converter for an electron microscope.
J. Physics E **10**, 520-524, 1977 (No. 5). *L*

H. C. van den Elzen: Echo cancelling in tweedraads full-duplex datatransmissie met bipolaire codering.
T. Ned. Elektronica- en Radiogen. **42**, 97-103, 1977 (No. 4). *E*

K. Gawron & J. Schröder: Eigenschaften einiger Salzhydrate zur latenten Speicherung thermischer Energie.
VDI-Ber. No. 288, 79-86, 1977. *A*

J. J. Goedbloed & M. T. Vlaardingerbroek (Philips Elcoma Division, Eindhoven): Theory of noise and transfer properties of IMPATT diode amplifiers.
IEEE Trans. **MTT-25**, 324-332, 1977 (No. 4). *E*

J.-M. Goethals: Single-channel error-correcting convolutional codes.
Philips J. Res. **33**, 248-253, 1978 (No. 5/6). *B*

J. Graf, A. W. Woodhead & C. G. Sluijter: The microchannel plate (film).
Proc. Soc. Photo-Opt. Instr. Engrs. (SPIE) **97**, 223-227, 1977. *L, R, E*

P. C. M. Gubbens, A. M. van der Kraan (both with Interuniversitair Reactor Instituut, Delft) **& K. H. J. Buschow:** Mössbauer effect study of $Tm_2Fe_{17-x}Co_x$ and related compounds.
Physica **86-88B**, 199-200, 1977 (Part I). *E*

G. Harding: Optimum spread functions in linear tomography.
Phys. Med. Biol. **22**, 747-759, 1977 (No. 4). *H*

A. van Herk: Side fringing fields and write and read crosstalk of narrow magnetic recording heads.
IEEE Trans. **MAG-13**, 1021-1028, 1977 (No. 4). *E*

B. Hill: Optische Speichersysteme.
NTG-Fachber. **58**, 307-327, 1977. *H*

B. Hill & K. P. Schmidt: Fast switchable magneto-optic memory-display components.
Philips J. Res. **33**, 211-225, 1978 (No. 5/6). *H*

B. Hoekstra, E. M. Gyorgy*, G. Zydzik* & P. J. Flanders** (* Bell Laboratories, Murray Hill, N.J.; ** University of Pennsylvania, Philadelphia): Magnetostriction measurements with a recording rotating field magnetostrictometer.
Rev. sci. Instr. **48**, 1253-1255, 1977 (No. 10). *E*

R. E. Horstman & J. Wolter: Reflection of high-frequency phonons at a solid-liquid helium interface: the role of phonon focusing.
Physics Letters **62A**, 279-281, 1977 (No. 4). *E*

F. Hottier, J. B. Theeten, A. Masson* & J. L. Domange* (* ENSCP, Paris): Comparative LEED and RHEED examination of stepped surfaces; application to Cu(111) and GaAs(001) vicinal surfaces.
Surface Sci. **65**, 563-577, 1977 (No. 2). *L*

B. B. van Iperen & J. J. Goedbloed: Intrinsic response time and related quantities describing semiconductor avalanches.
Electronics Letters **13**, 448-449, 1977 (No. 15). *E*

L. Jacomme: Laser excitation of cladded multimode fibres — Comment.
Opt. Quant. Electronics **9**, 360, 1977 (No. 4). *L*

L. Jacomme: Approximations in the evaluation of the impulse response of nearly parabolic fibres.
Opt. Quant. Electronics **9**, 197-202, 1977 (No. 3). *L*

L. Jacomme & D. Rossier: Parallel-beam impulse response of multimode fibres: numerical results.
Opt. Quant. Electronics **9**, 203-208, 1977 (No. 3). *L*

W. H. de Jeu & J. van der Veen: Molecular structure and nematic liquid crystalline behaviour.
Mol. Cryst. liq. Cryst. **40**, 1-17, 1977 (No. 1-4). *E*

T. Karstens, K. Kobs, R. Memming & F. Schröppel: Photoisomerization mechanism of thioindigo dyes. Thioindigo in nonpolar solvents.
Chem. Phys. Letters **48**, 540-544, 1977 (No. 3). *H*

J. L. W. Kessels: An alternative to event queues for synchronization in monitors.
Comm. ACM **20**, 500-503, 1977 (No. 7). *E*

G. D. Khoe & G. Kuyt: A Luneberg lens for the efficient coupling of a laser diode and a graded-index fibre.
NTG-Fachber. **59**, 176-178, 1977. *E*

A. J. R. de Kock: Effect of growth conditions on semiconductor crystal quality.
1976 Crystal growth and materials, ed. E. Kaldis & H. J. Scheel, pp. 661-703; North-Holland, Amsterdam 1977. *E*

A. M. van der Kraan (Interuniversitair Reactor Instituut, Delft) **& K. H. J. Buschow:** Magnetic properties of rare earth compounds of the type RFe_4Al_8.
Physica **86-88B**, 93-94, 1977 (Part I). *E*

E. Krätzig: Materialien für optische Datenspeicher.
NTG-Fachber. **58**, 295-306, 1977. *H*

J. van Laar, A. Huijser & T. L. van Rooy: Electronic surface properties of Ga and In containing III-V compounds.
J. Vac. Sci. Technol. **14**, 894-898, 1977 (No. 4). *E*

A. J. Lambell (MEL, Crawley, Sussex) **& R. P. Vincent:** Precision approach monitoring and guidance using interferometer techniques.
Proc. 7th Eur. Microwave Conf., Copenhagen 1977, pp. 509-513. *R*

S. P. J. Landsbergen: Wat is computer-linguistiek?
Wetenschap & taal, ed. B. T. Tervoort, pp. 47-65; Coutinho, Muiderberg 1977. *E*

J. Langerhorst: Vervaardiging van lagerboring voor precisie-lager met behulp van doelbewuste temperatuurvariatie.
Mikroniek **17**, 137, 1977 (No. 6). *E*

A. J. Linssen & H. L. Peek: Effects of trichloroethane oxidation of silicon wafers on SiO_2 and Si properties.
Philips J. Res. **33**, 281-290, 1978 (No. 5/6). *E*

J. H. van Loenhout, J. J. C. Buelens & M. J. G. Braken: Micro-plasmalassen in de praktijk.
Lastechniek **43**, 141-145, 1977 (No. 7). *E*

M. G. Luijpen, P. C. M. Gubbens, A. M. van der Kraan (all with Interuniversitair Reactor Instituut, Delft) **& K. H. J. Buschow:** Mössbauer effect study of $Y(Fe, Co)_2$.
Physica **86-88B**, 141-142, 1977 (Part I). *E*

P. Lux: A novel set of closed orthogonal functions for picture coding.
Arch. Elektronik u. Übertragungst. (AEÜ) **31**, 267-274, 1977 (No. 7/8). *H*

J. Magarshack: Impact of active microwave solid state devices on modern communication systems.
Eurocon '77, Proc. Conf. on Communications, Venezia 1977, Vol. I, paper 2.13.2, 6 pp. *L*

A. E. Morgan & H. W. Werner: Test of a quantitative approach to secondary ion mass spectrometry on glass and silicate standards.
Anal. Chem. **49**, 927-931, 1977 (No. 7). *E*

A. E. Morgan & H. W. Werner: Quantitative SIMS studies with a uranium matrix.
Surface Sci. **65**, 687-699, 1977 (No. 2). *E*

A. E. Morgan & H. W. Werner: Semiquantitative analysis by secondary ion mass spectrometry.
J. Microsc. Spectrosc. électron. **2**, 285-290, 1977 (No. 3). *H, E*

B. J. Mulder: Vacuum sealing of lithium fluoride windows with lead fluoride.
J. Physics E **10**, 591-593, 1977 (No. 6). *E*

G. F. Neumark: Comment on 'Comparison of experimental and theoretical carrier concentrations in heavily doped *n*-type silicon'.
J. appl. Phys. **48**, 3618-3619, 1977 (No. 8). *N*

A. G. van Nie: On fields and currents in microstrips.
J. appl. Sci. Engng. A **2**, 205-212, 1977 (No. 3). *E*

J. M. van Nieuwland & C. Weber: Eigenmodes in non-rectangular reverberation rooms.
9th Int. Congress on Acoustics, Madrid 1977, paper C2, 6 pp. *E*

J. Paredaens & R. Vyncke (Computing Centre V.U.B.-U.L.B., Brussels): A class of measures on formal languages.
Acta Informatica **9**, 73-86, 1977 (No. 1). *B*

M. Parisot & J. Magarshack: Microwave circuit design by means of a simple simulator with an on-line minicomputer.
1977 IEEE MTT-S Int. Microwave Symp., San Diego, pp. 137-139. *L*

R. E. Pearson: Novel microstrip bandstop filter.
Electronics Letters **13**, 561-562, 1977 (No. 19). *R*

W. S. Quon: Multilevel voltage select (MLVS): a novel technique to *X-Y* address an electrophoretic image display.
IEEE Trans. **ED-24**, 1121-1123, 1977 (No. 8). *N*

B. V. Rejda*, J. G. J. Peelen & K. de Groot* (* Free University of Amsterdam): Tri-calcium phosphate as a bone substitute.
J. Bioengng. **1**, 93-97, 1977 (No. 2). *E*

P. J. Roksnoer, J. M. P. L. Huijbregts, W. M. van de Wijgert & A. J. R. de Kock: Growth of dislocation-free gallium-phosphide crystals from a stoichiometric melt.
J. Crystal Growth **40**, 6-12, 1977 (No. 1). *E*

P. Röschmann & H. Dötsch: Properties of magnetostatic modes in ferrimagnetic spheroids.
Phys. Stat. sol. (b) **82**, 11-57, 1977 (No. 1). *H*

D. Rossier: Digital telecopy.
Acta Electronica **21**, 9-19, 1978 (No. 1). *(Also in French.)* *L*

U. Rothgordt: Document printing.
Acta Electronica **21**, 71-82, 1978 (No. 1). *H*

U. Rothgordt, G. Aaron & G. Renelt: One-dimensional coding of black and white facsimile pictures.
Acta Electronica **21**, 21-37, 1978 (No. 1). *H, L*

T. E. Rozzi: Interacting step discontinuities in planar dielectric waveguides.
Proc. 7th Eur. Microwave Conf., Copenhagen 1977, pp. 158-162. *E*

R. Schayès & P. Gustin (MBLE, Bruxelles): Réalisation expérimentale d'un télécopieur du groupe III.
Acta Electronica **21**, 83-100, 1978 (No. 1).

H. Schomberg: A peripheral array computer and its applications.
Proc. IMACS (AICA)-GI-Symp. on Parallel computers — parallel mathematics, Munich 1977, pp. 183-186. *H*

M. F. H. Schuurmans: Power broadening of Doppler-free two- and three-photon resonances.
Physics Letters **63A**, 25-27, 1977 (No. 1). *E*

B. Singer: Theory and performance characteristics of pyroelectric imaging tubes.
Adv. in Image Pickup and Display **3**, 1-82, 1977. *N*

J. W. Slotboom: Analysis of bipolar transistors.
Thesis, Eindhoven 1977. *E*

J. W. Slotboom & H. C. de Graaff: Bandgap narrowing in silicon bipolar transistors.
IEEE Trans. **ED-24**, 1123-1125, 1977 (No. 8). *E*

W. A. M. Snijders: Microprocessor-implementatie van datamodems.
T. Ned. Elektronica- en Radiogen. **42**, 87-92, 1977 (No. 4). *E*

J. L. Spoormaker: Reliability prediction of hairpin-type springs.
Proc. 1977 Annu. Reliability and Maintainability Symp., Philadelphia, pp. 142-147. *E*

J. P. Stagg: Drift mobilities of Na^+ and K^+ ions in SiO_2 films.
Appl. Phys. Letters **31**, 532-533, 1977 (No. 8). *R*

J. B. Theeten, L. Hollan & R. Cadoret (U.E.R. Sciences, Aubière): Growth mechanisms in CVD of GaAs.
1976 Crystal growth and materials, ed. E. Kaldis & H. J. Scheel, pp. 195-236; North-Holland, Amsterdam 1977. *L*

D. L. A. Tjaden: Birefringence in single-mode optical fibres due to core ellipticity.
Philips J. Res. **33**, 254-263, 1978 (No. 5/6). *E*

W. Tolksdorf: New experimental developments in flux growth.
1976 Crystal growth and materials, ed. E. Kaldis & H. J. Scheel, pp. 639-659; North-Holland, Amsterdam 1977. *H*

C. H. F. Velzel: On the imaging properties of holographic gratings.
J. Opt. Soc. Amer. **67**, 1021-1027, 1977 (No. 8). *E*

J. F. Verwey & A. Heringa: Avalanche injection and near avalanche injection of charge carriers into SiO_2.
IEEE Trans. **ED-24**, 519-523, 1977 (No. 5). *E*

J. F. Verwey, A. Heringa, R. de Werdt & W. v. d. Hofstad: Drift of the breakdown voltage in *p-n* junctions in silicon (walk-out).
Solid-State Electronics **20**, 689-695, 1977 (No. 8). *E*

Q. H. F. Vrehen: Single-pulse superfluorescence in cesium.
Cooperative effects in matter and radiation, ed. C. M. Bowden, D. W. Howgate & H. R. Robl, pp. 79-100; Plenum Press, New York 1977. *E*

P. J. de Waard: Stripe-geometry d.h. lasers with linear output/current characteristics.
Electronics Letters **13**, 400-401, 1977 (No. 14). *E*

J. S. C. Wessels: Fragmentation.
Photosynthesis I (Encyclopedia of plant physiology, new series, Vol. 5), ed. A. Trebst & M. Avron, pp. 563-573; Springer, Berlin 1977. *E*

J. C. Williams: Cross-fed printed aerials.
Proc. 7th Eur. Microwave Conf., Copenhagen 1977, pp. 292-296. *R*

J. C. Williams: Doppler direction sense module.
Proc. 7th Eur. Microwave Conf., Copenhagen 1977, pp. 504-508. *R*

J. P. Woerdman: Polarization dependence of near-resonant two-photon absorption of Na_2 vapor.
Chem. Phys. Letters **50**, 41-44, 1977 (No. 1). *E*

A. W. Woodhead & R. Ward: The channel electron multiplier and its use in image intensifiers.
Radio and electronic Engr. **47**, 545-553, 1977 (No. 12). *R*

L. E. Zegers & N. A. M. Verhoeckx: Digital signal processing and LSI in modems for data transmission.
Philips J. Res. **33**, 226-247, 1978 (No. 5/6). *E*

Supplement to Philips Technical Review
Beilage der Philips Technischen Rundschau
Bijlage van Philips technisch Tijdschrift

PHILIPS

December 1979

Recent United States Patents

Abstracts from patents that describe inventions from the following research laboratories that form part of or cooperate with the Philips group of companies:

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

4 074 314

Apparatus for optically reading a record carrier and correcting focus error

C. H. F. Velzel *E*
P. F. Greve

An apparatus is described for reading a record carrier on which information is stored in an optically readable reflecting structure. By projecting two radiation spots one before and one behind the plane of the information structure, it is possible to ascertain whether a read beam is properly focused on the information structure. An optical wedge-field lens combination extends the range of operation of the focus correction device.

---

4 075 484

Device for measuring radiation absorption or emission

D. Meyer-Ebrecht *H*
G. Kowalski

A device for correcting measuring values of a group of radiation detectors for measuring the intensity or distribution of radiation, comprising an additional radiation source for direct irradiation of the detectors. The radiation of the said additional radiation source can be switched on each time when the radiation to be measured is not incident on the radiation detectors. The device furthermore comprises a calculating device which forms for each radiation detector the quotient of the measuring value (M1(x)) and the reference measuring value (M2(x)) of the radiation detectors when exposed to the radiation of the additional radiation source.

---

4 075 490

Device for measuring radiation absorption in a layer of a body

G. Kowalski *H*

A device for measuring the absorption of radiation in a layer of a body, utilizing a radiator emitting a fanned radiation beam which passes through the body and which is measured by a number of adjacently arranged detectors. The radiator/detector system is rotated with respect to the body during the measurement. Integral measuring values of mutually intersecting sets of parallel strips are used for calculating the absorption. To this end, the device comprises means for sampling signals supplied by the detectors at different instants and/or means for effecting the sampling after different signal delay times.

4 075 495

X-ray luminescent screen

A. L. N. Stevels *E*
A. D. M. Schrama-de Pauw

Luminescent screen, in particular X-ray image intensifier screen, provided with a luminescent alkaline earth metal fluoro-halide which is activated by Eu^{2+}, contains samarium and has a composition defined by the formula $Me_{1-p-q}Eu_pSm_qFX$, where Me represents barium and/or strontium and X represents chlorine, bromine and/or iodine. Up to 20 mol % of the Ba and/or Sr can be replaced by Ca. Furthermore $0.001 \leqslant p \leqslant 0.20$; $0.001 \leqslant q \leqslant 0.10$; $0.02 \leqslant q/p \leqslant 0.50$.

---

4 075 665

Method and apparatus for increasing the recording capacity of a disc

J. P. Borne *L*
G. J. Naaijer

A method of recording and reproducing audio information on a disc shaped record carrier. During recording the audio information is divided into consecutive information segments of a duration τ_t. Each information segment is subjected to time compression before recording. The duration τ_t and the time compression are selected so that optimum use is made of the recording capacity of the record carrier without information loss. During reproduction the information segments are subjected to a complementary time expansion and recombined in their original sequence.

---

4 075 844

Hot-gas reciprocating engine having controlled coupling of a combustion air fan

J. W. Schiferli *E*

A hot-gas reciprocating engine whose shaft is coupled to the combustion air fan via successively a variable transmission, constructed as a slip coupling whose degree of slip is inversely proportional to the temperature of the heater, and a fixed transmission.

4 077 112

Method of manufacturing charge transfer device

M. J. J. Theunissen E
R. P. Kramer
H. L. Peek

The invention relates to a charge transfer device (C.T.D.) with polycrystalline silicon electrodes which are provided on a nitride layer. The nitride layer has apertures between the polyelectrodes. Electrodes of a second metallization layer, for example, of aluminium, are provided via said apertures. The charge storage capacities per surface unit (and with equal voltages) can be made equal by subjecting the device for a short period of time to an oxidation treatment prior to providing the Al electrodes so that the oxide layer in the apertures can become thicker than below the Si electrodes.

---

4 078 177

Layer-wise reproduction of three-dimensional objects by simultaneous super-position of encoded images

U. Tiemens H

A method of layer-wise reproduction of three-dimensional images by way of a number of simultaneously recorded, encoded superposition images of different image planes. The object is irradiated from different directions. Simultaneous-superposition images of the object are recorded in encoded form, by means of a large number of m sources, from different directions in a number of n planes which are situated one below the other. The decoding of these images for the layer-wise reproduction of the three-dimensional object is effected so that for each individual layer each of the n simultaneous-superposition images, whose scale has been adapted relative to each other, is simultaneously or consecutively shifted and summed on a single recording medium a number of times which corresponds to the number of m sources during the recording of the object.

---

4 078 208

Linear amplifier circuit with integrated current injector

C. M. Hart E
A. Slob

A new integrated circuit in which bias currents are supplied by means of a current injector, a multi-layer structure in which current is supplied, by means of injection and collection of charge carriers via rectifying junctions, to zones to be biased of circuit elements of the circuit, preferably in the form of charge carriers which are collected by the zones to be biased themselves from one of the layers of the current injector. By means of said current injector circuit arrangements can be realized without load resistors being necessary, while the wiring pattern may be very simple and the packing density of the circuit elements may be very high. In addition a simple method of manufacturing with comparatively few operations can in many cases be used in particular upon application of transistors having a structure which is inverted relative to the conventional structure.

Re. 29 570

Television system having aperture correction

H. Breimer E
S. L. Tan

In a color television camera system of the type having camera tubes for the production of red, green and blue color signals, contour signals derived from only one of the color signals are added to all of the color signals.

4 078 301

Method of manufacturing a deflection coil for a cathode ray tube

M. A. Renders E
W. W. J. Degger

A deflection coil consisting of a number of elementary coils formed by folding flat coils along four folding lines such that at the area of the folding lines each turn crosses all other turns. Very accurate deflection coils can thus be manufactured in a simple manner.

---

4 079 257

Automatic control system for a scintillation camera

M. Jatteau L
J. Pauvert

A scintillation camera has photomultipliers and a number of auxiliary radioisotope sources to calibrate the photomultipliers. The calibration circuits can be either analogue or digital.

---

4 079 284

Mounting piezoelectric elements

D. G. J. Fanshawe R

A method of mounting an elongate piezoelectric element on a laminated printed wiring board by forcing said element into the board at an angle normal to the surface of the board and at a point where the element can contact a conductor line on the board. Piezoelectric elements made of ferroelectric ceramic materials are suitable especially when mounted on a printed circuit board composed of copper-clad phenol formaldehyde laminated plastic.

---

4 079 355

Method for the transmission of binary information by means of a frequency-modulated signal and a circuit for performing that method

H. van der Gaag E

A transmission of a binary signal which transmission is immune to interference is obtained by transmitting the bits of the binary signal by means of a frequency-modulated signal in which a cycles of a first frequency f_1 are followed by b cycles of a second frequency f_2 for a binary zero value and b cycles of the first frequency f_1 are followed by a cycles of the second frequency f_2 for a binary one value.

---

4 079 460

Device operating with the displacement of magnetic domain walls

F. H. de Leeuw E

A device operating with the displacement of magnetic domain walls and comprising a layer of a magnetizable material having an easy axis of magnetization perpendicular to the plane of the layer, in which layer magnetic domain walls can be produced, maintained and moved, as well as a device to cause a magnetic propulsive field to act upon magnetic domains present in the layer so as to move them between previously determined positions. In order to obtain maximum speeds, the propulsive field acting upon a domain wall during the movement should at any instant have a value which is below the value for which a drop occurs in the function which characterizes the magnetizable material and which represents the relation betweeen the propulsive field and the speed.

4 080 044

Device for coupling optical fibers

R. R. Gousseau *L*

A device for coupling optical fibers each of which is secured in a capillary tube by means of a cap such that the end of the fiber is situated in one plane with the end face of the cap. The two capillary tubes are arranged one in the prolongation of the other in the boring of a deformable core which can be deformed such that radial and axial forces are exerted on the capillary tubes, with the result that the latter are simultaneously aligned and secured.

---

4 080 169

Determination method for mercury in organic compounds

J. G. Kloosterboer *E*
A. M. Stalpers-Kiemeneij

Photolysis of organic mercury compounds in mercury-containing samples to determine mercury therein by means of irradiation with a light source emitting predominantly below 260 nm.

---

4 080 269

Method of producing coatings having a high absorption in the range of the solar spectrum

H. Scholz *A*
H. Jungk

Method of producing a black coating highly absorbant in the solar spectrum range by electrodeposition from an aqueous solution containing a soluble cobalt compound and a large amount of alkali thiocyanate.

---

4 082 138

Heat regeneration

A. R. Miedema *E*
K. H. J. Buschow
J. Mulder

A filling mass for low-temperature regenerators comprising one or more of the compounds of the formula $A_{1-x}B_xRh_{1-y}X_y$, wherein A = Sm, Gd, Tb, Dy or a mixture thereof, wherein B = Ho, Er, Tm, Yb or a mixture thereof, wherein X = Cu, Zn, Ru, Pd, Ag, Re, Os, Ir, Pt, Au or a mixture thereof, and wherein $0 \leqq x \leqq 1$ and $0 \leqq y < 0.2$.

---

4 082 556

Light-sensitive polymeric compounds

D. N. Hunter *R*

Light sensitive cyanuric azide polycondensates find use in production of positive and negative photographic relief images.

---

4 113 515

Semiconductor manufacturing method using buried nitride formed by a nitridation treatment in the presence of active nitrogen

E. Kooi *E*
J. G. van Lierop

The invention relates to a method of manufacturing a semiconductor device in which a surface of a silicon semiconductor region covered at least partly with a silicon oxide-containing layer is subjected to a nitridation treatment forming a buried zone of a nitrogen-containing material between the silicon oxide layer and the silicon region, which zone is used in a further phase of the manufacture or in the manufactured semiconductor device.

4 115 036

Pump for pumping liquid in a pulse-free flow

D. Paterson *R*

A peristaltic pump comprises two concentric cylinders at least one of which includes a plurality of piezoelectric elements which are successively electrically energized to produce moving seals in the region between the cylinders.

---

4 116 528

Optical viewing device of the periscope type with rotating means for focussing

J. A. Clarke *R*

A head-worn viewing device for alleviating night blindness includes an image intensifier tube between periscopic reflectors with an objective lens in front of the upper reflector and an eyepiece lens behind the lower reflector. The objective and upper reflector are mounted for rotation together about a pivot so that, for short viewing distances, a continuous adjustment is provided which will refocus the objective at the front plane of the intensifier tube and compensate for the parallax of the periscopic reflectors.

---

4 116 535

Picture display cell with memory effect

J. J. P. Ponjeé *E*
J. Bruinink

The invention relates to an image display cell in which a 4,4'-bipyridinium compound is used as a reduceable redox material. According to the invention said redox material is dissolved in a cell medium which contains a mixture of water and an electrochemically inert organic solvent. A suitable organic solvent is, for example, dimethyl sulfoxide. As a result crystallization of the solution takes place only at temperatures far below 0° C so that the picture display cell can also be used at lower temperatures. The 4,4'-bipyridinium compound may be so chosen that the resulting picture can be erased directly electrochemically without the use of an auxiliary couple.

---

4 117 277

Arrangement for simultaneous two-way data transmission over two-wire circuits

H. C. van den Elzen *E*
P. J. van Gerwen
W. A. M. Snijders

An arrangement for simultaneous two-way data transmission of data signals with a given symbol frequency over two-wire circuits of a type found in telephone networks or of a comparable type. The arrangement comprises an echo canceler with an adaptive filter for producing from transmit channel signals approximated echo signals which are differentially combined with receive channel signals for forming substantially echo-free residual signals. A special type of code conversion is employed in the transmit channel, whereby p-level data symbols are converted into modified p-level data symbols which are thereafter converted into $(2p\text{-}1)$-level data symbols. The echo canceler comprises a digital adaptive filter to which the modified p-level data symbols and the residual signals with a sampling frequency equal to the symbol frequency are supplied. These measures result in a transmission signal having favorable spectral properties, a simple inverse code conversion and an echo canceler which combines simplicity of implementation with favorable convergence properties.

---

4 117 328

Infrared radiation sensitive image pick-up device

M. J. Auphan *L*

An infra-red imaging device consisting of two sheets of interconnected discs of piezoelectric material, at least one of which has a ferroelectric transition temperature. A metal foil is provided

between the sheets and mutually parallel, conductive strips are provided on both outer surfaces of the assembly thus formed, the direction of the strips on the one surface differing from that of the strips on the other surface. A generator for periodic signals which can be sequentially connected to each strip on one of the surfaces of the first sheet generates therein, via a piezoelectric effect, ultrasonic vibrations which are intercepted by the other sheet which converts these vibrations into electric signals. These signals are amplified and carry the image to be examined.

---

4 117 384

Tachogenerator processing circuits and motor speed control systems including such circuits

D. R. Armstrong *R*

A circuit for processing the a.c. output of a tachogenerator whose frequency is proportional to the speed of a rotor of the tachogenerator, said circuit including voltage level detection means responsive to said a.c. output, key pulse producing means responsive to at least one output of the detection means to produce a key pulse whose duration is inversely proportional to the rotational speed, first gating means responsive to an output of the detection means and an output of the key pulse producing means to produce a reset pulse after a predetermined period in the interval between each key pulse, and a voltage generator responsive to each reset pulse to provide a predetermined output level and responsive to each key pulse to ramp that output level at a predetermined rate for the duration of that key pulse, whereby the output level of the voltage generator during each said predetermined period is a predetermined function of the rotational speed.

---

4 117 509

Color television system for recording line sequential signals with 90° phase shift

E. de Boer *E*

A system for recording a color television signal. The color information is recorded as an amplitude modulation. This color information consists of a first and a second color-difference signal, which two color-difference signals are transmitted in a line-sequentially alternating fashion exhibiting a 90° phase-shift during consecutive line periods.

---

4 117 515

Television camera having signal electrode strips

M. G. Carasso *E*
L. J. van de Polder
S. L. Tan

A black-white or color television camera comprising a pick-up tube, provided with a signal electrode composed of separated strips. Prior to a line scan to be performed by an electron beam, the signal electrode strips are connected to a reference potential. After a scan the strips are through-connected successively through associated switches for supplying information. The switch outputs are alternatingly, in at least two groups through-connected and connected to at least two output leads. Always two stitches, which are connected to two adjacent strips are closed simultaneously. In this way it is avoided that a capacitive voltage distribution across the strips and switch inputs influence the picture signal obtained in a disturbing manner.

---

4 117 695

Thermodynamic method and device for carrying out the method

C. M. Hargreaves *E*

A thermocentrifugal pumping device in which a gaseous medium is supplied to a rotor element rotating about an axis, is first conducted inside the rotor element mainly in a radial direction away from the axis of rotation through a compression duct in which the medium is compressed by centrifugal action and in which heat of compression is withdrawn from the medum, in which the medium inside the rotor elements is then mainly conducted in the radial direction towards the axis of rotation through an expansion duct in which the medium expands against the centrifugal action before leaving the rotor element and in which so much thermal energy is supplied to the medium that the medium temperature in the expansion duct is always higher than the medium temperature in the compression duct so that flow of medium is produced in the rotor element in the direction from the compression duct to the expansion duct.

---

4 118 099

Image coding and decoding using coherent codes

H. Weiss *H*
E. Klotz
U. Tiemens

A method of imaging a three-dimensional object with the aid of coding and decoding using coherent point codes, whereby the object is recorded in two steps, each with a plurality of radiation sources of different orientation on the same recording material. Negatives of the two images are produced, after which the positive of one image is combined with the negative of the other image, the combined images being subsequently multiplied by means of point holograms formed with the aid of a reference wave and a multiplicity of light sources whose co-ordinates correspond to those of the point images of the two source arrays recorded by a pinhole camera, the two multiplied images being superposed in one plane, the coded images being decoded by changing their scale with the aid of two synchronously controlled zoom lenses whereby the coordinates of the point image function of an image of an object layer in the coded images are adapted to the points stored in the two point holograms.

---

4 118 734

Optical videodisc with variable width tracks

G. Bouwhuis *E*
J. J. M. Braat
B. A. J. Jacobs

A disc-shaped record carrier is described on which information is recorded in an optical structure of trackwise arranged areas and intermediate areas. By adapting the average dimensions of the areas to the average spatial frequencies on the record carrier a read signal can be obtained of sufficient modulation depth and minimal distortion. This is accomplished by increasing the track width with decreasing radius of the record carrier.

---

4 119 483

Method of structuring thin layers

H. Hübsch *H*
U. Convertini
H. Dimigen
H. Lüthje

Am ethod of structuring oxide layers, nitride layers or magnetic layers in such manner that a photolacquer mask is manufactured on a substrate and the layer to be structured is provided by means of cathode sputtering both on the photolacquer mask and on the surfaces of the substrate not covered with lacquer. The substrate is then treated with a solvent attacking the lacquer mask; the mask swells up and the parts of the layer to be structured present thereon are chipped off. In order to stimulate this latter process, a layer is provided below the photolacquer mask on the substrate relative to which photolacquer has a small adhesive capacity and, after providing the layer to be structured, an increase in volume of the lacquer mask is produced by thermal treatment. After the complete removal of the photolacquer mask and the parts of the layer to be structured present thereon, the structured thin layer remains on the substrate as a negative of the pattern of the photolacquer mask.

4 119 744

Method of manufacturing semiconductor devices in which a layer of semiconductor material is provided on a substrate

J.-J. L. E. Brissot L
C. Belouet
R. Martres

A method of manufacturing semiconductor devices comprising a layer of semiconductor material, comprising the steps of providing a strip-shaped solid substrate having a main surface and providing a solid support having a substantially horizontal surface and comprising a substantially floating liquid mass of such semiconductor material on the substantially horizontal surface, the main surface being wettable by the liquid semiconductor material. The liquid mass is contacted with the main surface of said substrate, onto which substrate the layer of semiconductor material is to be provided, the substrate is moved in its longitudinal direction along and in contact with said liquid mass so that a liquid layer of said semiconductor material is formed on said strip and taken along with it, and the liquid layer is substantially progressively solidified.

---

4 120 080

Method of manufacturing grid electrodes for electron tubes

J. W. A. Krol A
B. Lersmacher
H. Seifert

Disclosed is a method of manufacturing a pyrolytic graphite grid electrode in which after formation of the grid, the electrode is annealed at a temperature of 700° to 1200° C in a reactive atmosphere to remove carbon particles adhering to the electrode surface.

---

4 121 131

Color television display tube and method of manufacturing same

J. van Esdonk E
P. F. A. Haans

In an assembly of at least two electrodes which are connected together in an insulating manner, said electrodes are kept at a given distance from each other by at least one member of an electrically insulating material situated between the electrodes which comprises a core which determines the distance between the electrodes and a jacket which directly adheres to the electrode material by heating. The material of the core has a higher melting point than the material of the jacket so that the core during effecting a connection between the jacket and the electrode material maintains its shape.

---

4 121 255

Television camera apparatus

P. P. M. Schampers E
M. G. Carasso
F. T. Klostermann

The target of a television camera tube comprises photoconductive material in the form of a regularly interrupted structure so that both the photoconductive material and parts of the signal electrode are accessible to the electron beam. The potential difference between the signal electrode and the cathode is adjusted so that during scanning, the signal electrode accepts a proportion of the beam current dependent on the local potential of the photoconductor surface, and the photoconductor is stabilized only during fly-back. The camera tube thus has an adjustable inherent amplification without increased inertia.

4 121 256

Column electrode drive for gas discharge display device

J. H. J. Lorteije E
G. Warrink

A saving in column memory elements while preserving a stable display with a gas discharge television display device is obtained if a plurality of bits of the digital video information which is entered in the line trace period is passed from the line memory circuit to the relevant column electrode in the line blanking period while bypassing a column memory circuit. The display elements are excited only once during a line period at the beginning of the line period.

---

4 121 296

Digital signal processing arrangement

W. A. M. Snijders E
N. A. M. Verhoeckx
P. J. van Gerwen
H. A. van Essen

Digital signal processing arrangement such as a digital filter for computing output signal samples which are each equal to the sum of products of a given plurality of pairs (for example N) of input signal samples. One of the input signal samples of the pair is, for example, formed each time by a sample $x(n)$ of a digital signal to be filtered, the other input signal sample of the pair by a filter coefficient $z(n)$, wherein either the sample $x(n)$ or the filter coefficient $z(n)$ can assume only a limited number of values 0, $\pm A_1, \pm A_2, \ldots \pm A_M$, M being much smaller than N. In the case $x(n)$ can assume only this limited number of values, $z(n)$ is first multiplied, owing to the $x(n)$ belonging to the pair, by the sign of $x(n)$ for generating a product of the form $z(n)$. sgn $[x(n)]$. Independent of the magnitude of $n(n)$ this product is applied to an accumulator which corresponds with this magnitude of $x(n)$. N products have been applied to one of the M accumulators the partial sums stored therein are multiplied by the magnitude $|x(n)|$ corresponding therewith and the products obtained are added together.

---

4 121 933

X-ray electrophotographic recording material and method of manufacturing same

H. Dannert A
H.-J. Hirsch
E. Klein
K.-H. Panstruga

The sensitivity of photoconductive layers of tetragonal lead monoxide in a binder is increased when the tetragonal lead monoxide produced by prior art process after being subjected to a further treatment has a grain size of 1 to 50 μm and is dispersed in the binder without mechanical force and is sedimented on a layer carrier without mechanical force.

---

4 122 328

Device and welding torch for plasma-MIG-welding

W. G. Essers E
G. Jelmorini
G. W. Tichelaar

A compact plasma-MIG welding torch in which the contact tube and the nozzle are directly connected together electrically, thereby enabling welding with a very short extension of the welding wire.

---

4 122 484

Display device for three-dimensional television

S. L. Tan E

A three-dimensional, black-white or color television display device which is compatible for two-dimensional television. Located in front of the screen of a direct-view display tube is a

removable strip or dot filter having alternatingly light-reflecting and light-dispersing strips or dots and light-transmissive, polarizing or light-selecting strips or areas. The display device is provided with a projection display tube which gives with three-dimensional display an image on the strip or dot filter via a polarizer or light-selecting filter. By means of spectacles having polarized glasses or light-selecting glasses an observer can observe a three-dimensional image.

---

4 122 505

Magneto-resistive reading head with suppression of thermal noise

K. E. Kuijk *E*

A magnetic reading head having a magneto-resistive element of the type in which the measuring current is forced to flow through the element at an angle with the easy axis of magnetization. For suppressing thermal noise (resistance variations under the influence of temperature fluctuations as a result of the contact of the element with a moving record carrier), the element is subjected to an auxiliary field parallel to the easy axis of magnetization, the direction of which is reversed at a frequency f_s. The measuring current alternates at the same frequency in phase with the reversals of the auxiliary field. The output signal of the element is applied through a low-pass filter which cuts off signals of a frequency f_s to a detection device.

---

4 122 542

Memory array

F. Camerik *E*
C. M. Hart
A. Slob

An integrated circuit memory array having a plurality of memory cells including two cross-coupled transistors of one conductivity type, load transistors of the other conductivity type, and a bit line, connected to the base region of one of the cross-coupled transistors through a bit line transistor. The array features a common node, directly interconnecting all of the base regions of the load transistors and the emitter regions of the cross-coupled transistors, for each of the memory cells; and a row selection line connected to the emitter regions of the load transistors in an associated row of memory cells.

---

4 122 831

Solar collector comprising an elongate absorber in an evacuated transparant tube

F. Mahdjuri *A*

A solar collector comprising an elongate absorber centrally arranged in an evacuated, transparent, circular-cylindrical tube provided with a reflector in the form of a $\frac{1}{3}$-circular cylinder in contact with the absorber.

---

4 123 147

Electrically controlled adjusting device for a light beam with digital and analog deflection control

B. Hill *H*
R. Pepperl
J. Krüger

A light deflector used for tracking a spiral information track alternately and periodically interrupted by data blocks and guide segments arranged within sectors of a rotating record carrier. The light deflector is controlled during the first turn of the information track by a slow analog light deflector having a small stroke and is controlled during the remainder of the recording by a digital light deflector control device having a relatively large stroke.

4 123 652

Apparatus for reading a radiation-reflecting record carrier

G. Bouwhuis *E*

An apparatus is described for reading a record carrier with an optical information structure, in which apparatus errors in the focussing of a read beam on the information plane are detected with an astigmatic focussing-error detection system. By including a beam splitter in the path of a beam which has been reflected by the record carrier and by including an astigmatic focussing-error detection system in each of the paths of the two sub-beams thus obtained, the astigmatism of the one detection system being opposed to that of the other detection system, a focussing-error signal can be derived which is independent of optical faults in the read apparatus.

---

4 123 672

Circuit arrangement for frequency division of high-frequency pulses

W. G. Kasperkovitz *E*

A circuit arrangement for the frequency division of high frequency pulses in which a cyclic sequence of transistors are connected together via tapped resistances. In integrated circuit technology the tapped resistances may be the parasitic resistance of a semiconductor layer which serves as a common collector for the transistors.

---

4 123 695

Pattern recognition system

J. A. G. Hale *R*
A. R. Turner-Smith

An automatic machine including a two-dimensional positioning table, a device for measuring the dispositions attained by the table in each of said dimensions and a tool holder having an operative center in respect to said two dimensions. Furthermore means are provided for imaging a graticule means to an operative plane of the table so that the center of said image is substantially aligned with the operative center of the tool holder. A visual input device is positioned to view a part of said operative plane including said graticule image and comprises pattern recognition apparatus for:

(a) identifying a predetermined mark on a flat surface of an object carried by said table, said flat surface lying in said operative plane,
(b) identifying said graticule image, and
(c) measuring the separation of said mark from said graticule image center.

---

4 123 783

Magnetic bubble display systems

R. F. Pearson *R*
D. E. Lacklison
G. B. Scott
J. S. Palfreeman

A display arrangement including a source of polarized radiation; a magnetic material capable of supporting magnetic bubble domains in the path of the polarized beam, for producing a modulated beam; a device for representing an image encoded signal by means of the magnetic bubble domains; and a display device in the path of the modulated polarized beam for providing a visual display of the encoded image.

---

4 124 933

Methods of manufacturing semiconductor devices

K. H. Nicholas *R*

A method of manufacturing a semiconductor device in which a masking layer is formed on part of the surface of a deposited layer of relatively high resistivity polycrystalline semiconductor

material present on an insulating layer provided at a surface of a semiconductor body or body part and a relatively low resistivity conductive region having a substantially uniform narrow line width is defined in the polycrystalline layer by effecting a diffusion process to laterally diffuse a doping element into a portion of the polycrystalline layer underlying an edge portion of the masking layer without diffusing the doping element through the insulating layer into the semiconductor body or body part.

---

4 125 401

Method of making copies of information tracks on carriers

E. J. Spiertz E
C. F. W. Flinsenberg
L. K. H. van Beek

Copying by photographic means of plates having coded audio and video information of a very fine structure. Use is made of an aromatic diazosulphide as light-sensitive system, wherein, after exposure, a metal nuclei image is produced by contact with metal salt, which metal nuclei image is at or below the surface and which is intensified by means of physical development to an externally reflective and, as such, readable image.

---

4 125 418

Utilization of a substrate alignment marker in epitaxial deposition processes

D. J. Vinton R

An alignment marker on a substrate surface is covered with polycrystalline semiconductor material during the growth of an epitaxial layer on the monocrystalline substrate. This polycrystalline material is then removed with a selective etchant to re-expose the marker for use in defining an area for processing at the epitaxial layer surface. Permits accurate alignment between buried layers and regions formed from the epitaxial layer surface. Permits provision of the marker on the substrate when it is undesirable to provide the marker on the epitaxial layer surface. Particularly advantageous for electron image projection exposure of electron-sensitive resists.

---

4 125 425

Method of manufacturing flat tapes of crystalline silicon from a silicon melt by drawing a seed crystal of silicon from the melt flowing down the faces of a knife shaped heated element

J.-J. L. E. Brissot L

A method of manufacturing bodies from meltable crystalline material in which a continuous tape of said material manufactured by causing melted material to flow along at least one surface of a heated element, which surface is wetted by the melted material, and causing the tape to grow by drawing away a seed crystal near the lower side of the surface. Application in particular to the manufacture of semiconductor bodies, for example of silicon.

---

4 125 803

Current distribution arrangement for realizing a plurality of currents having a specific very accurately defined ratio relative to each other

R. J. van de Plassche E

A precision current distribution arrangement includes a multiple current source which supplies q approximately identical currents. These currents are separately applied to the q input terimnals of a coupling circuit with p output terminals. With the aid of, for example, shift registers these q currents are transferred to the p outputs in a cyclically permuting fashion, in such a way that at each output during each time interval of the cycle always the same number of currents is transferred, so that per cycle each of the input currents is transferred an equal number of times to a specific output terminal. By subjecting the currents at these output terminals to a low-pass filter action a number of currents is obtained whose ratio relative to each other is very accurately defined.

---

4 125 866

Non-recursive discrete filter

H. A. van Essen E
P. J. van Gerwen

A non-recursive discrete filter for simultaneously realizing for example two band-pass filter characteristics starting from a given lowpass transfer characteristic,where in one bandpass filter has an in-phase transfer characteristic and the other filter the quadrature transfer characteristic of this in-phase filter. The central frequency of these filters is f_0 and the output sampling frequency is $f_s = 8f_0$. This discrete filter comprises multiplying means in which input signal samples $x(n)$ are modified by an associated filter coefficient $a(i)$ for generating modified input signal samples $z(n,i) = a(i)x(n-i)$. The coefficients $a(i)$ are the same for both transfer characteristics and are given by:

$$a(i) = h_1(i) \mid \cos(\pi i/4) \mid \quad \text{for } \cos(\pi i/4) \neq 0$$
$$a(i) = h_1(i) \mid \sin(\pi i/4) \mid \quad \text{for } \cos(\pi i/4) = 0,$$

wherein $h_1(i)$ represents the discrete impulse response of the low-pass filter. To realize the in-phase transfer characteristic the samples $z(n,i)$ are thereafter multiplied by a factor $\mathrm{sgn}[\cos(\pi i/4)]$ and the products thus obtained are accumulated. This results in the output signal sample

$$y_1(n) = \sum_{i=0}^{N-1} z(n,i) \cdot \mathrm{sgn}\,[\cos(\pi i/4)].$$

To realize the quadrature transfer characteristic the samples $z(n,i)$ are thereafter multiplied by a factor $\mathrm{sgn}[\sin(\pi i/4)]$ and the products thus obtained are accumulated. This results in the output signal sample

$$y_2(n) = \sum_{i=0}^{N-1} z(n \cdot i) \cdot \mathrm{sgn}\,[\sin(\pi i/4)].$$

---

4 126 899

Junction field effect transistor random access memory

J. Lohstroh E
J. J. M. Koomen
R. H. W. Salters
C. M. Hart

A random access memory (RAM) in which each memory cell includes a JFET having two gate electrodes selectable by means of a single word line and a single bit line. The JFETs have a common electrode formed from the substrate of a semiconductor body common to each of the memory cells, which serves as one of the main electrodes of each of the JFETs.

---

4 126 900

Random access junction field-effect floating gate transistor memory

J. J. M. Koomen E
J. Lohstroh
R. H. W. Salters
A. T. van Zanten

JFET memory structures, in particular for RAMs with non-destructive reading-out of the charge state of a floating gate electrode in which the primary selection is realized by means of capacitive coupling with the floating gate electrode. The secondary selection takes place on one of the main electrodes of the JFET structures in which the other main electrode can be connected to the supply. By means of a second common gate electrode the pinch-off voltage of the channels can be adjusted so that the channels are non-conductive in the non-selected condition and a good detection of the information state is obtained in the selected condition.

4 126 995

Hot-gas engine with protected heat reservoir

G. A. A. Asselman E
H. Fokker
R. J. Meijer

A hot-gas engine in which the transfer of heat from the heat-source to the meltable material of a heat storage reservoir is effected exclusively indirectly via the working medium in order to prevent overheating, fast corrosion and cracking of the reservoir walls.

---

4 127 789

Light-pervious, heat-reflecting filter and electric lamps having such a filter

H. Köstlin A
R. Jost
H. Auding

The life of light-pervious, thermal radiation-reflecting filters of tin-doped indium oxide is drastically elongated when used in oxidizing or reducing medium by coating the indium oxide layer with pyrolytic or hydrolytic silicon dioxide.

---

4 127 864

Semiconductor device

P. J. W. Jochems E

A semiconductor device having a bipolar transistor of the lateral type, preferably a *PNP* transistor which is provided in a homogeneously doped semiconductor layer and which may be provided both in an *N*-type and in a *P*-type semiconductor layer and of which the base comprises a highly doped contact region and an associated substantially non-depleted active base region, while the emitter zone is situated substantially entirely within the active base region. Herewith, high frequency complementary transistors can be formed in a single epitaxial layer. The invention furthermore comprises a suitable method of manufacturing said transistor in which use is made of underetching.

---

4 128 781

X-ray tube

P. Flisikowsky H
H. Peemöller

An X-ray source having a cathode rotatable relative to the anode and a motor for rotating the cathode to produce a corresponding rotation of the X-ray beam emitted by the source.

---

4 128 842

Electrostatic printing device with monocrystalline stylus

U. Rothgordt H
H.-D. Hinz
G. Verspui

In order to increase recording quality the electrode pin of a device for electrostatically printing characters consists of a whisker which is mounted to be electrically conductive in a metal holder. A whisker is by definition a monocrystalline member.

---

4 130 758

Device for determining gaseous components

S. van Heusden E
L. P. J. Hoogeveen

Device for determining the concentration of vaporous components in a gas current, wherein the gas current reacts in a heated reaction room with ozon and the intensity of the chemiluminescent radiation emitted herewith is measured by a photoelectric cell.

4 130 806

Filter and demodulation arrangement

P. J. van Gerwen E
R. A. van Doorn
W. A. M. Snijders
H. A. van Essen

A digital filter and demodulation arrangement for passband signals, whose corresponding baseband signal has a bandwidth limited to a given maximum frequency. The passband signals are filtered according to two bandpass characteristics which, apart from their asymmetrical distortion relative to their central frequency, are versions from one another shifted 90° in phase. The filtered passband signals are demodulated with an in-phase carrier and a quadrature carrier and the demodulated signals are combined to a baseband signal. In the digital filtering process the sampling frequency is reduced from a value higher than twice the highest frequency in the passband signals to a value which is not higher than twice said maximum frequency in the baseband signal and in the digital demodulation and combination processes the reduced sampling frequency is also used so that a considerable reduction in the internal processing speed is obtained.

---

4 130 835

Electrical waveform analysis

L. H. Guildford R

A circuit for detecting changes in the amplitude of a video signal features a sampler for sampling the signal. The sampled signal is then held and delayed by a selected amount. The delayed and undelayed signals are then compared and when the difference exceeds a threshold, an output signal is supplied.

---

4 131 524

Manufacture of semiconductor devices

A. C. M. Gieles E

A method of manufacturing a semiconductor device comprising the steps of providing a plate-shaped semiconductor body, removing by spark erosion a first portion of the semiconductor body, and then using a selective etching process to remove only a second portion of the semiconductor body.

---

4 131 525

Method of manufacturing a body having a gold pattern and body manufactured according to the method

R. P. Tijburg E
T. van Dongen

The invention relates to a method of manufacturing a body having a gold pattern in which an etchant-resistant and electrically insulating masking layer is provided locally on the surface of a gold layer present on the body and parts of the surface of the gold layer not covered by the masking layer are subjected to a chemical etching treatment in an etching bath in which the gold pattern is formed. According to the invention, the gold layer during the chemical etching treatment is given a potential with respect to an electrode in the etching bath and this potential is adjusted to a value such that the etch rate is maintained at a rate no higher than the rate of the chemical etching treatment.

---

4 131 764

Arrangement for converting discrete signals into a discrete single-sideband frequency division multiplex signal and vice versa

T. A. C. M. Claasen E
F. G. Mecklenbräuker

A multiplex system has an inverse Fourier transformer that is unequal to its forward one, therefore an optimum distribution of elements is possible.

PHILIPS TECHNICAL REVIEW

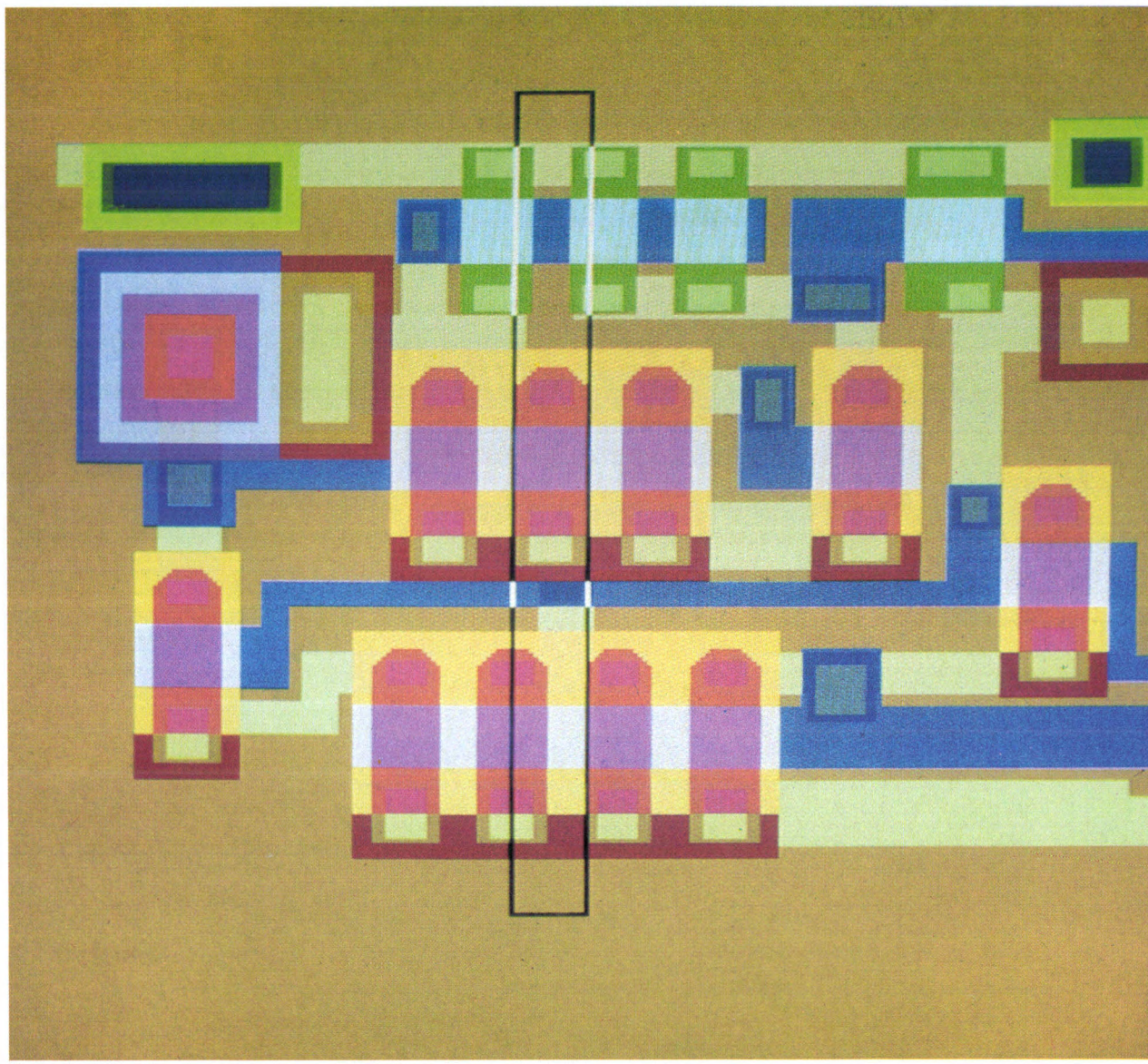
VOLUME 38, 1979/80 No. 11/12

Processing pictorial information

Although the cameras and film we have today will give us pictures of extremely high resolution, it is not always so easy to bring out particular details or structures. We also have industrial robots at our command, capable of carrying out all manner of complicated functions, but they are unable to pick up a component lying in a random orientation and position on a conveyor belt, because they cannot 'look at' things as people can. On the other hand, people are not such reliable observers as one might wish. They easily tire and become less attentive when inspecting large numbers of industrial components or when looking at large numbers of medical X-ray pictures, and they are unable to keep their criteria for judgement constant for hours on end. And there are some visual inspection tasks that are beyond the range of human capability.

Largely through the development of the computer (LSI) and of the appropriate software, we can now submit pictorial information to all kinds of complex operations that go some way to meet the above requirements. Of course, the pictorial information first has to be reduced to digital form. In particular cases we can even go a step further and use the computer to build up a picture from data already tailored to suit the purpose and which the human eye would be unable to assimilate.

Developments in this field are proceeding apace. Robots that can look at things have already been produced on an experimental scale, and there is even work on systems that can learn a little for themselves. This issue of our journal gives a selection from results achieved in recent years and of work in this field now in progress in the laboratories of the Philips Research Organization. The introductory article gives a brief review of the applications and methods of picture processing. As a challenge and a source of inspiration — but not as an example to be imitated — it gives an outline of the organization of the human or animal visual system, in which incoming information is also subjected to elaborate processing before it is communicated to the brain.



The computer design of IC masks can be checked with an instrument developed at Philips Research Laboratories (see p. 326) that displays the masks together but in different colours on a screen. The electronic marker allows faults in the layout and alignment to be located with a resolution of 1 μm.

The challenge of picture processing

G. Groh

Introduction

About one-and-a-half thousand million years ago a biological evolution began that led to the origin and the development of the visual system. Some micro-organisms became capable of distinguishing between light and darkness and learned to react accordingly. This accomplishment greatly improved their chances of survival. An extremely primitive kind of visual scene analysis had proved its usefulness. The rest of the development is well known. Arrays of sensors with increased processing power appeared and finally — at a rather late stage — the eye with its image-forming lens was 'invented'.

Technological evolution followed a different course. Image formation by means of mirrors and lenses and picture recording by photography were the starting points. Later some methods of processing were introduced to compensate to some extent for the often poor quality of the products of the early systems. Picture restoration by manually retouching photographs is the best-known example. Other methods aimed at an improved presentation of pictorial information by manipulating the image. Edge enhancement, which gives an apparently sharper or clearer picture, is one example. The latest stage has been the advent of more 'intelligent' procedures, such as coding for data compression, feature extraction and classification, pattern recognition and finally automated picture evaluation, all strongly stimulated by the increasing capabilities of digital computers.

Thus, the technical discipline of picture processing has three different origins. The first is optics and photography, with its fairly long tradition. The second is television, whose refined analog electronic facilities allow the implementation of many methods of picture manipulation in real time. The third and most recent one is digital data processing, which enables artificial intelligence to be used in picture processing.

These different schools of technological development emanating from different scientific disciplines have not yet been successfully merged together. Compared with the biological competitor, technical picture processing is still in its infancy. That is the challenge!

I shall return to the biological world in the last section of this introduction. For the time being I shall sketch briefly some of the major applications and methods of modern picture processing. Obviously, this cannot be a complete description of the present state of the art [1]. The emphasis will be placed on those areas that are related in some way to the topics dealt with in the other articles in this issue.

First, I shall look at some of the methods of picture processing from the viewpoint of applications. This will demonstrate the fairly long tradition of picture processing and will also indicate the motivation for the considerable research effort being put into this field. Secondly, I shall describe some of the basic principles underlying the methods used. Thirdly, I shall discuss some of the techniques used to perform these operations. Each of the different views will confront us with open questions and reveal some seemingly intractable problems. These are challenges for future research.

Applications of picture processing

There are many areas of human activity where picture processing is useful or necessary. Instead of trying to present a comprehensive but perhaps somewhat trivial list, I shall focus on two major areas of application that are both important in themselves and representative of many others: medical diagnostics and industry.

In present-day medical diagnostics wide use is made of X-ray pictures, or radiographs. These are 'shadowgrams', produced by exposing the patient to irradiation from a point-like X-ray source and recording the projected absorption pattern ('radiation relief') on a photographic film. The absorption differences vary over more than three orders of magnitude. The radiograph therefore represents a grey-tone picture with an enormous dynamic range in optical density, which can only be reasonably covered by means of special double-

Dr G. Groh is a Deputy Director of Philips GmbH Forschungslaboratorium Hamburg, Hamburg, West Germany.

[1] Extensive information can be found in several books, such as:
W. K. Pratt, Digital image processing, Wiley, New York 1978;
A. Rosenfeld and A. C. Kak, Digital picture processing, Academic Press, New York 1976;
K. S. Fu (ed.), Digital pattern recognition, Springer, Berlin 1976;
R. O. Duda and P. E. Hart, Pattern classification and scene analysis, Wiley, New York 1973.
The most recent information will be found for example in the Proceedings of the Fourth International Joint Conference on Pattern Recognition, Kyoto 1978.

coated films. Closed-circuit television, coupled to X-ray image intensifiers, which is increasingly used for certain types of examination, only shows up to about 50 grey levels, however. It is primarily the *display* equipment that restricts the use of advanced electronic systems to applications with less exacting requirements.

This physical limitation, however, can be easily overcome by means of a simple type of processing called '*dynamic windowing*'. Only the part of the intensity range of interest to the human observer is selected and displayed with the full grey scale of the equipment. Two control knobs allow him to scan the full range and to define the 'width' of the 'window'. One application in which this technique is now used is in displays for computerized tomography [2].

Image formation in radiography is often limited by photon noise. The appearance of the noisy picture can be improved by applying spatial *smoothing* techniques. This corresponds to a slight defocusing of the image, which gives an averaging over the statistical fluctuations in the neighbourhood of each individual pixel.

The spatial resolution of radiographs is limited by the size of the focal spot of the X-ray tube. The apparent sharpness of the picture can be improved by edge enhancement. This method was introduced as long ago as 1930. It was then called 'harmonization' and made use of photographic processes to superimpose the original picture and its slightly defocused negative [3].

When we consider the great density range of radiographs, it is obvious that coarse, irregular structures of low contrast may remain indiscernible. A solution to this problem is contrast enhancement by altering the distribution of grey levels in the picture.

Examples of applications of these techniques are given in the articles by A. Hoyer and M. Schlindwein [4], A. Hoyer and W. Spiesberger [5] and L. H. Guildford [6] in this issue. Whereas the first two papers deal with general digital data-processing systems, the third discusses a new approach for real-time equipment, which is important in such applications as infrared imaging and television at low light levels.

Before discussing further the increasing refinement and complexity now found in the applications of picture processing, I should like to draw attention to a relatively new development, which is growing in importance but is sometimes overlooked in this context: *image formation by picture processing*.

In medical diagnostics a prominent example is computerized tomography, in which a rotating linear array of X-ray detectors is used to measure the absorption values in about 100 000 locations and directions of a cross-sectional 'slice' of a patient's body. The 'one-dimensional shadowgrams' thus obtained are processed so as to generate a two-dimensional picture of the local absorption coefficient in that slice. This is not the place to go into the rather complicated mathematics involved [2]. It will be useful, however, to look briefly at the ideas underlying this approach.

Radiology deals with the imaging of three-dimensional objects by means of X-radiation. Physics does not offer any advanced image-forming devices for X-radiation. All we have are highly primitive shadowgrams, showing a superposition of all the layers at different depths in the patient. Without *a priori* knowledge, location in depth is not possible. Owing to the superposition of so many structures, even the considerable density range of an X-ray film is not sufficient for recording and resolving the tiny absorption differences in soft tissues, something that is of primary importance in early cancer diagnosis.

The solution offered by computerized tomography is a two-step process of image formation. First, a well-defined set of data is generated, in accordance with the physical measuring possibilities and boundary conditions. Then the desired final image is formed by post-processing. As described earlier, potential limitations of the display unit are overcome by picture manipulation.

Other medical examples of image formation by picture processing are tomosynthesis [7] and coded-aperture imaging. A very recent version of the latter is described in the article by E. Klotz *et al.* on flashing tomosynthesis [8]. In this case imaging of the two-dimensional object is performed by simultaneously 'flashing' an array of X-ray tubes, irregularly distributed in a plane above the patient. The resulting superposition of images, called the coded picture, is not directly comprehensible but contains information on the depth. Only after appropriate processing can pictures of slices at different depths — the tomograms — be constructed.

Two-step image formation using picture processing is also gaining acceptance in other medical applications, such as ultasonic imaging [9] and spin imaging [10]. The latter utilizes nuclear magnetic resonance for characterizing tissues. Similar principles have already established a certain tradition in nuclear medicine, for example in generating patterns of the distribution of isotopes in the human body. A much older and well-known example in another field is 'synthetic-aperture radar' [11], which again uses picture processing to form a radar image with vastly increased spatial resolution.

The next step on the ladder of increasing refinement is automated recognition of patterns in a picture. The most prominent sample for many people will be machine reading of printed or even handwritten docu-

ments by means of optical character recognition [12]. Other applications are found in medical diagnostics [13]. More stringent demands, however, are nowadays found in industrial applications. Examples are the inspection of printed-circuit boards, mask inspection in the production of integrated circuits, and also non-destructive testing of materials and semifinished products by means of X-radiation [14], with a view to detecting flaws such as pores in welding seams, cracks, porosities and cavities in castings, or irregularities in the fabric structure of rubber tyres. In general, quality control by automated inspection of pictures — which are generated by infrared radiation, visible radiation, X-rays or ultrasound — is gaining increasing importance now that many industries manufacturing 'critical' parts for motor cars, aircraft, nuclear power plants, etc. are insisting on an extremely high reliability for their products.

The hope is that automated inspection by means of computerized pattern recognition will prove to be more reliable and more objective than a wearisome visual inspection by people. As yet, however, no general solution has been found.

The severest requirements of all come from visual scene analysis. This is an important function in robotics, which is mainly used in assembling products, sorting goods, controlling vehicles, etc. In these cases an 'intelligent' interpretation of a three-dimensional scene under different kinds of illumination will ultimately be required. This means that the machine has to recognize the shape, location and orientation of a two-dimensional object. It has to distinguish real parts of an object from shadows. It must not be misled by printed patterns on the surface of the object. The spatial relationship of parts that overlap in one projection have to be recognized and interpreted. And all these processing procedures have to work in real time on sets of grey-tone pictures showing different perspectives.

This applies to general-purpose machines for automated assembly work. Fortunately, there are other applications where the objects are flat (essentially two-dimensional) and where uniform illumination from behind can be used to generate black-and-white pictures. These can be unambiguously described by features of their contours, which considerably reduces the amount of data needed and makes real-time processing possible. The possibilities and potential of visual scene analysis in this area of applications are described in the article by P. Saraga and J. A. Weaver [15], who have successully made an experimental visually controled pick-and-place machine.

The examples given in the foregoing represent only a very small fraction of the full range of existing and potential applications of picture processing. The potentialities are so promising that they are motivating a vast research effort all over the world. In Japan, for example, a huge research programme on 'Pattern Information Processing Systems' (PIPS) is being sponsored by MITI's Agency of Industrial Science and Technology [16]. This project was started in July 1971 and estimates put its total costs at about 25 thousand million yen (about 100 million dollars) by 1980. Major participants are the leading Japanese companies like Mitsubishi, Toshiba, Hitachi and Fujitsu. In the United States of America a government-sponsored research program on 'Image Understanding' is in progress, which has an approved annual budget of three million dollars in the years 1972 to 1981 (this does not include military applications).

Some principles

Various different methods and techniques are used today for picture processing, depending on the different disciplines in which the picture originated. This

[2] G. Kowalski and W. Wagner, Generation of pictures by X-ray scanners, Optica Acta **24**, 327-348, 1977.

[3] Gottfried Spiegler and Kalman Juris, Ein neues Kopierverfahren zur Herstellung ideal harmonischer Kopien nach kontrastreichen Negativen, Fortschr. Geb. Röntgenstr. **42**, 509-518, 1930.
See also E. Zieler and K. Westerkowsky, The Ampliscope, an experimental apparatus for 'harmonizing' X-ray images, Philips tech. Rev. **24**, 285-293, 1962/63.

[4] A. Hoyer and M. Schlindwein, this issue, p. 298.
A related method has recently been applied to optical character recognition: W. G. Baig, An edge-enhancing double-focus camera for image processing, Philips tech. Rev. **37**, 180-187, 1977.

[5] A. Hoyer and W. Spiesberger, this issue, p. 347.

[6] L. H. Guildford, this issue, p. 310.

[7] See for example D. Meyer-Ebrecht and H. Weiss, Tomosynthesis — 3-D X-ray imaging by means of holography or electronics, Optica Acta **24**, 293-303, 1977.

[8] E. Klotz, R. Linde, U. Tiemens and H. Weiss, this issue, p. 338.

[9] See for example H. Schomberg, An improved approach to reconstructive ultrasound tomography, J. Physics D **11**, L 181-185, 1978.

[10] See for example P. Mansfield, A. A. Maudsley and T. Baines, Fast scan proton density imaging by NMR, J. Physics E **9**, 271-278, 1976.

[11] L. J. Cutrona, E. N. Leith, L. J. Porcello and W. E. Vivian, On the application of coherent optical processing techniques to synthetic-aperture radar, Proc. IEEE **54**, 1026-1032, 1966.

[12] See for example P. Saraga, J. A. Weaver and D. J. Woollons, Optical character recognition, Philips tech. Rev. **28**, 197-202, 1967, and M. Beun, A flexible method for automatic reading of handwritten numerals, Philips tech. Rev. **33**, 89-101 and 130-137, 1973.

[13] See for example W. Spiesberger and M. Tasto, The automatic measurement of medical X-ray photographs, Philips tech. Rev. **35**, 170-180, 1975, and the article by Hoyer and Spiesberger in this issue [5].

[14] See for example D. Meyer-Ebrecht and W. Spiesberger, Neue Verfahren in der medizinischen Röntgendiagnostik — eine Herausforderung für die industrielle Röntgenprüftechnik?, Materialprüfung **19**, 409-415, 1977.

[15] P. Saraga and J. A. Weaver, this issue, p. 329.

[16] H. Nishino, PIPS (Pattern Information Processing System) Project — background and outline, Proc. 4th Int. Joint Conf. on Pattern Recognition, Kyoto 1978, pp. 1152-1161.

does not matter and can even have its advantages, as long as the application-oriented approach is adopted and the best combination selected in each case. Picture processing requires an interdisciplinary attitude. Neither optics nor television electronics nor digital data processing alone can offer general solutions, as is clearly demonstrated by the articles in this issue. Nevertheless, there does exist a set of fundamental operations, which are used again and again. I shall try to illustrate this with a few examples that all relate to two-dimensional grey-tone images.

Mathematically, two-dimensional grey-tone images can be described by a brightness or intensity function I, whose value varies with the location (x,y) in the image plane:

$$I = f(x,y). \qquad (1)$$

The black-and-white ('binary') image is a special case in which I takes one of two possible values at each point (x,y).

Some simple primitive operations are negation (i.e. black changes to white and vice versa) of the images and addition of several images. Others are translation, rotation, scaling and distortion or stretching. Their usefulness and potential is demonstrated in the article by Klotz *et al.* [1] and — indirectly — in the article by Saraga and Weaver [8,15]. Whereas in scene analysis these operations are mainly used for searching and comparison and have to be supplemented by more refined operations, in tomosynthesis only translation, scaling and addition are necessary for performing the decoding process.

Integration over or addition of a sequence of noisy pictures of the same object taken at different times has the effect of reducing noise. Edge enhancement can be obtained by spatial differentiation. This can be approximated by adding the negative of the blurred image to the original one, as we have seen earlier.

Here a new operation comes in: blurring or defocusing, which can be described as weighted spatial integration over the neighbourhood of each pixel. This is easy to understand by considering the simplest of all images, i.e. a very small white spot on a black background. Defocusing will make this spot larger and a smoother transition will take place from black to white. The new intensity profile is called the *point-spread function* of the system that causes the blurring. A more interesting picture obviously consists of many pixels; its blurred version is the superposition of all their point-spread functions.

Of course, a pixel is a mathematical fiction that does not exist in reality. Every real image-forming system has a limited resolution, which is determined by the width of the point-spread function. For this reason it is permissible, without losing too much information, to scan a picture with a small spot line by line, as in television, to convert it into a time-sequential signal, or to describe the picture as a two-dimensional matrix of discrete individual pixels, to enable it to be processed by a digital computer. However, the scanning or pixel sampling has to be fine enough to prevent the raster of scanning lines or pixels from showing up.

There can be no doubt that *scanning* and (two-dimensional) *spatial sampling* are the most important basic operations for converting an image into a time-sequential signal or a list of computer data and vice versa. More details on conversion by scanning will be found in the article by Guildford [6].

The digital computer needs an additional operation. The original analog intensity values have to be quantized, i.e. converted into a finite number of grey levels. Again the quantizing operation has to be fine enough to avoid 'artefacts'.

Further processing of signals or sets of data generated in accordance with these basic conversion operations is easily performed. In particular, nonlinear distortions of the grey scale, as in thresholding methods (for generating binary images) and dynamic windowing methods (to stretch and clip the grey scale) are simply a question of using conventional electronic circuits or of rearranging data.

In most pictures the spatial context of the pixels is of decisive importance. This is taken into consideration by the local operator, a small window or 'keyhole' of say 5×5 pixels, which can be deliberately moved over the whole picture. The operations described earlier, or others like spatial gradient operations, are carried out only within the area of this window. Examples of this technique are given in the article by Hoyer and Spiesberger [5].

An important characteristic of a picture is the distribution of the probability of occurrence of the grey levels, the grey-level histogram. The manipulation of the grey-level distribution yields impressive results. Examples will be found in several articles in this issue [4–6].

Finally, *spatial filtering* should be mentioned. The basic operation is the two-dimensional Fourier integral of the image function I:

$$\mathfrak{F}(I) = \iint f(x,y)\, e^{-2\pi i(xv+y\mu)}\, dx dy.$$

The new variables v and μ are called spatial frequencies and correspond exactly to the well-known signal frequency in the time domain.

Spatial filtering means manipulation of the spatial-frequency content of the image. For example, suppression of the high frequencies (lowpass filtering) reduces the sharpness of the picture and causes de-

focusing. Suppression of the low frequencies (highpass filtering) results in a sharper picture.

Although far from complete, this list of basic image operations will serve to illustrate the potentialities. They can be carried out in many different ways.

The same or similar procedures as have been outlined above are used for feature extraction. The general principle is firstly to enhance special features (e.g. edges, contours, local maxima, spatial frequencies) in the picture itself or in the spatial-frequency domain, and secondly to apply a threshold or masking operation to suppress the rest of the image.

In most cases the extracted features themselves form a two-dimensional pattern, i.e. an image which contains less information, but usually still too much.

In the next step of data reduction some of these features are numerically described, resulting in data sets that are called 'features'. The nature of the feature of interest depends to a great extent of course on the particular application. In medicine the doctor may only be interested in the shape of the heart contour or in the area it encloses. If all the other information about location or orientation is irrelevant, one ends up with a small set of data or even with a figure [13]. In mammography [5] all that is required for further investigation is the location of bright spots in the mammogram that betray the possible presence of microcalcifications. In non-destructive testing of the rubber tyres of motor cars, deviations from the periodic structure in the embedded fabric have to be detected and localized. In visual scene analysis [15] an enclosed contour may have to be characterized further to see whether it is perhaps identical with the stored model of the object. If it is, only its orientation and the location of its centre of gravity may be of further interest.

These few examples should provide sufficient illustration. The number of applications requiring modified approaches is overwhelming. This demonstrates the main problem and the difficulties of more advanced picture processing, namely the characterization and extraction of useful features from images. Some general principles are known, but there is as yet no complete theory from which solutions for all conceivable problems could be derived, at least in principle. The classification process is better understood; it does at least have some underlying theory.

In nature, biological evolution is *not* based on deduction. Species of plants and animals are to be considered as self-organizing systems. Mutations that lead to a better chance of survival will remain, others will fade out. This holds in particular for the 'intelligent', i.e. image-understanding part of the visual system. During its evolution numerous modifications of the 'hardware' have been tried and selected for their 'usefulness' in the struggle for survival. Even during the life of the individual creature, the system and especially its 'software' is further trained and improved.

In the world of technology there is nothing comparable. The wishes and dreams are as old as computer science. As early as 1953 C. E. Shannon [17] formulated a set of challenges, from which I can quote in this context:

'Can we organize machines into a hierarchy of levels, as the brain appears to be organized, with the learning of the machine gradually progressing up through the hierarchy? . . . Can manipulative and sensory devices functionally comparable to the hand and eye be developed and coordinated with computers? . . . Can more satisfactory theories of learning be formulated? . . . How can a computer memory be organized to learn and remember by association, in a manner similar to the human brain? . . .'

These questions could have been asked today, a quarter of a century later.

Nevertheless, considerable progress has been made in artificial intelligence. The article by Saraga and Weaver in this issue gives one example [15]. The articles by T. J. B. Swanenburg [18] and E. H. J. Persoon [19] describe some recent experiments on self-organizing memories and recognition systems. Although the class of objects in the pictures is rather restricted, the feasibility of the procedures is clearly demonstrated and there is every expectation that they can be extended to more general pictures.

Practical realization

By far the most important progress has been made in the technology, i.e. in the hardware. In 1949 W. S. McCulloch [20] made the picturesque remark that a computer with as many tubes as man has neurons (10^{10}) would require the Empire State Building to house it, the Niagara Falls to power it and the Niagara River to cool it (the human brain dissipates some 25 watts!). Today we are starting with Very Large Scale Integration (VLSI) which promises by 1985 to achieve integration of 10^5 to 10^6 gates on a thin silicon chip of one square centimetre! If closely packed, 10^{10} neuron-like gates would require the volume of a shoebox. The design and construction of an intelligent visual system like the one that forms part of the brain has become technically feasible, at least in terms of volume.

[17] C. E. Shannon, Computers and automata, Proc. I.R.E. **41**, 1234-1241, 1953.
[18] T. J. B. Swanenburg, this issue, p. 364.
[19] E. H. J. Persoon, this issue, p. 356.
[20] W. S. McCulloch, The brain as a computing machine, Electr. Engng. **68**, 492-497, 1949.

The real bottlenecks, however, still exist. We do not know how to organize such an enormous number of logical functions, that is to say the architecture of such a system and most of the more advanced algorithms are as yet unknown. We even have difficulties in feeding the data flow from a high-resolution television camera into a computer in real time [21].

One of the reasons is that both our television and our computer systems are organized time-sequentially. This is poorly matched to pictorial information, which is essentially parallel. For example, the standard television picture of about 300 000 pixels ($= 500 \times 600$) each consisting of 64 permissible grey levels (six bits) contains as many as 1.8 million bits of information. The elementary operation of just adding two pictures by sequentially passing these pixels through an adder circuit, requiring only a hundred nanoseconds per operation, will take in all 30 milliseconds, which corresponds roughly to the normal frame rate. More refined operations will require much more time. The only solution to this problem is parallel transmission and processing of pictorial data.

There is one approach which is by its very nature parallel, i.e. optics. Many linear operations, like spatial filtering, can be easily carried out in real time on large amounts of data. The only limitation in this respect is the velocity of light. Klotz *et al.* demonstrate in their article in this issue [8] the simplicity and elegance of optical decoding. Another very successful approach is used in synthetic-aperture radar [11]. Numerous system proposals for advanced applications, using coherent light and holography for pattern recognition, have also been discussed in the literature [22]. They all have the advantages of parallelism, but fall short in three important aspects:

— the class of image operations that can be carried out is too restricted; they are mainly linear operations;
— these operations cannot be changed quickly under program control, i.e. they resemble hard-wired filter functions;
— optical processors are not yet compatible with electronic or digital systems, which could give the desired flexibility and make other operations possible, for example Boolean operations.

This situation would soon change with the appearance of new interface components for input, output and filter modification. Then, of course, the analog or digital electronic circuits will still have to be structured in parallel. Otherwise, most of the time will be lost in parallel-to-sequential conversion of the signals and vice versa. The absence of such interfaces and the incompatibility of the subsystems are the main reasons for the present bottleneck. We can make picture-processing systems in one of the three basic technologies, but can seldom come up with combined systems that utilize the specific advantages from each system.

By definition the general-purpose digital computer is very flexible and useful for a wide range of operations. It is therefore a most excellent tool for investigating new algorithms, methods and applications, and for testing the procedures by simulation. For practical use it is usually much too slow, too difficult to handle, too clumsy and too expensive.

The trend is towards dedicated digital processors, which offer just the right degree of programmability and range of operations required by the special applications. Growing use is being made of the parallel architectures that have become economically feasible in modern microprocessor technology [23]. For input and output these dedicated processors are interfaced with analog electronic components, such as camera tubes and cathode-ray-tube displays.

The challenge

The continual challenge, however, is the biological visual system. It shows us what is in principle feasible. Perhaps the study of its functioning may stimulate our creativity and lead to novel, unconventional approaches. I shall therefore briefly summarize some of its more remarkable features.

In the human eye, picture processing starts at the retina, which consists of 1.25×10^8 receptors (94% rods for low light levels, black-and-white, and 6% cones for colour and daylight). They feed only 10^6 nerve fibres, which transmit the information to the brain. This considerable data reduction takes place mainly in the periphery, where up to several thousand receptors (rods) are somehow connected so as to jointly transmit their preprocessed information along only one nerve fibre. The principles of this data compression still remain a closed book. The greatest attention is paid to the two thousand cones in the rod-free fovea centralis. Each cone has a nerve fibre of its own. Their information is projected on to a disproportionately large area of the visual brain without substantially changing the spatial arrangement of the pictorial information.

In the main area of visual reception around the fovea centralis (corresponding to a viewing angle of about two degrees) some 6×10^4 pixels are resolved,

[21] See for example H. Bacchi and A. Moreau, Real-time orthogonal transformation of colour-television pictures, Philips tech. Rev. **38**, 119-130, 1978/79 (No. 4/5).
[22] G. Groh, Holographie, Verlag Berliner Union, Stuttgart 1973.
[23] M. Tasto, Parallel array processors for digital image processing, Optica Acta **24**, 391-406, 1977.

each capable of identifying about 250 grey levels (i.e. eight bits). The maximum information rate is 10 Hz at an illuminance of one microlux, but goes up to 60 Hz at one lux. This corresponds to a total flow to the brain of 6×10^4 (pixels) $\times 8$ (bits) $\times 10$ (Hz) or 5×10^6 bits per second. These five million bits per second are then further processed and reduced. The rate of perception of the visual system is less than 300 bits/s! As we know, for example from experiments on reading, the maximum total processing rate measured in terms of the conscious output, i.e. what finally gets through to our consciousness, is only about 50 bits/s. This means that the main part of the visual information (not even taking the periphery of the retina into account) is reduced by a factor of more than 10^5 without any loss of the relevant 'useful' information. That is the miracle and the challenge!

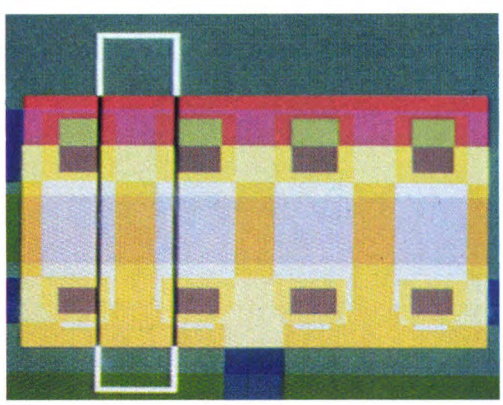
Some methods that we use in technical systems are used by the visual system as well. The best known is Mach's effect, which corresponds to an edge enhancement. But in addition the eye has a remarkable capacity for adaptation, even to systematic distortions or imperfections of the imaging system. For example, after a training phase of a few weeks it can even compensate for a 180° rotation of the images it receives (i.e. an upside-down world) or correct for severe chromatic aberrations.

Another interesting feature, whose effect on picture processing is not understood, is the irreducible tremor of the eye. It consists of irregular small movements of 30 seconds of arc, on which are superimposed slow drifts and saccadic movements of up to five minutes of arc. All we know is that without these irregular movements only moving objects would be observed, while stationary parts would gradually become invisible.

Is this just to compensate for a shortcoming of the light receptors, which apparently cannot detect light of continuously equal intensity, or is an additional preprocessing phase involved, which somehow takes the environment of every pixel into account statistically?

Technologically we process pictures with the aid of a computer, which is interfaced to a television-like receiving system. Most procedures are time-sequential. We are only at the very beginning of the introduction of parallel processing methods. Biologically, we have the fully parallel organized brain, part of which is coupled to the retina of the eye directly and in parallel. There is no real interface, since the 'format' of the signals does not change. 'Computations' for picture processing begin right at the detector, take place in some way or another even during transmission, and are finally completed in the brain. This multilevel processing is facilitated by a uniform structure of almost identical components — neurons. Embryologically and functionally the retina is part of the brain.

It would certainly be too naive to try to copy the biological system, whose functioning we do not even understand. On the other hand the differences between the most successful biological system and our rather mediocre technological approaches are very fundamental and most striking.



Digital image enhancement

A. Hoyer and M. Schlindwein

Computers and pictorial data

Processing pictorial data by digital computers has become a serious competitor of processing by the conventional optical means, which are typical of an analog method. The power and versatility of computer techniques seem to permit much greater flexibility in the manipulation of images than could ever have come from purely optical methods. The large storage capacity together with the high speed of present-day computers makes all kinds of mathematical or algorithmic operations on images feasible.

A major objective of all such operations is image enhancement, the subject of this article. The images can be improved in their general visual appearance — a rather subjective property — or in relation to preparation for the extraction of closely defined specific information, a task that can be performed by dedicated machines (or by people). Good examples of such tasks are to be found in modern X-ray diagnostics [1].

At Philips Forschungslaboratorium Hamburg work is under way on the design of a system for real-time processing and manipulation of images, a system that should have a very wide field of applications. A well-documented set of strategies for image enhancement can form the nucleus of such a general system. The programs for image enhancement and manipulation described in this article are the result of the first attempts to produce such a general package of strategies.

In the present context the term 'pictorial data' is limited to 'natural' black-and-white images, usually of square format. Various shades of grey can be present, but no colour. Such images are represented mathematically by a spatial brightness function $g(x,y)$ where x and y are the planar coordinates of a point in the (flat) image. The function value $g(x,y)$, which cannot be negative, is the brightness value or grey level of the point. The grey value is taken from a given 'grey scale', as the range of brightness values is called. For computer processing the continuous function $g(x,y)$ is replaced by its discrete counterpart $g(x_i,y_j)$, where i and j can be any integer from 1 to N. The $N \times N$ brightness values, which represent the original image, are stored in the computer in the form of an array of N^2 numbers (the digitized picture); these numbers correspond to the individual picture elements (or 'pixels'). Each pixel is a small square, whose area is equal to the total area of the original image divided by N^2. From now on the simpler notation $g(i,j)$ will be used for the discrete brightness function. The 'coordinates' (i,j) of the pixels are of course also stored in the computer.

The complete image-processing procedure, beginning with the original image and ending with the enhanced version, is outlined in *fig. 1*. The input/output equipment used in the laboratory at Hamburg is shown in *fig. 2*. Some of the details of this versatile and important part of the hardware are listed in *Table I*. A minicomputer (Philips P880) is used as the central processor; it also controls a number of the peripheral tasks, for example various setting-up and adjustment operations for the scanners. When 'digitizing' a picture the input equipment can work with a spatial resolution of up to 1600 pixels per mm^2. This high resolution limit imposes severe demands on the stability and

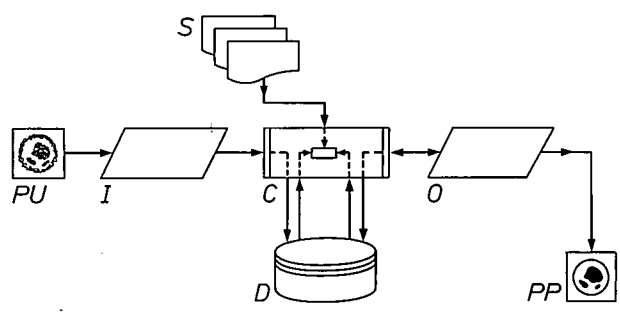


Fig. 1. Block diagram for image processing. *PU* original image. *PP* output picture. *I* input equipment ('digitizer'). *C* central processor. *S* image-processing software. *D* magnetic-disc store. *O* output equipment (display and hard copy on film). The software package consists mainly of image-enhancement procedures and a few peripheral programs, e.g. standardizing a number of the functions of *I*. Details of *I* and *O* are given in figs 2 and 3 and Table I.

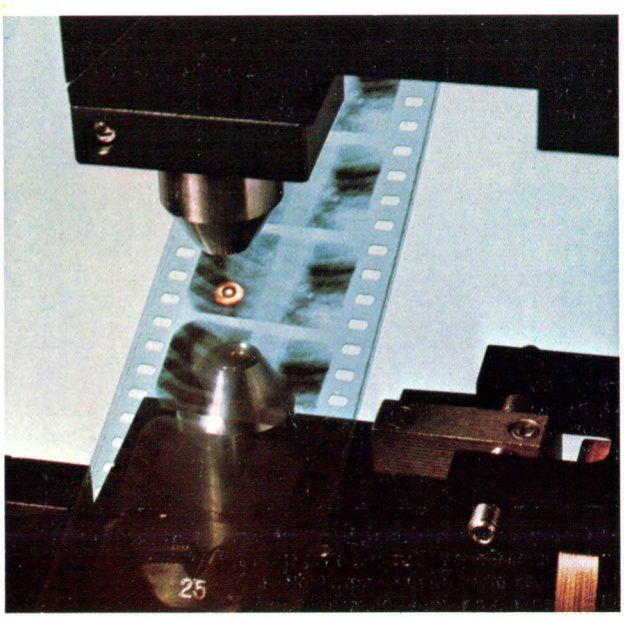
Dipl.-Phys. A. Hoyer, formerly with Philips GmbH Forschungslaboratorium Hamburg, is now with Zentralbereich Information, Philips GmbH, Hamburg; Dr M. Schlindwein is with Philips GmbH Forschungslaboratorium Hamburg, Hamburg, West Germany.

[1] W. Spiesberger and M. Tasto, The automatic measurement of medical X-ray photographs, Philips tech. Rev. 35, 170-180, 1975.



Fig. 2. This photograph shows part of the input and output equipment (*I* and *O*, fig. 1, see also Table I). The operator is handling a picture intended for the drum photoscanner, which processes the larger formats (up to 350×430 mm). The cover of the scanner has been moved back to show the drum on which the picture is mounted. The smaller formats (up to 57×57 mm) can be digitized with the flying-spot scanner, the equipment near the centre of this photograph. The tall rack contains power supplies, interfaces for peripheral units and various control circuits. The small monitor screen at the top of the rack indicates the total range of brightness values present in the unprocessed image for the flying-spot scanner; by modifying the illuminance of the scanning beam this total range can be adjusted, so that it just matches the range of the detector (a photomultiplier). The large screen (177×177 mm) on the right belongs to the output display unit. The output equipment in the foreground (only partly visible) provides film copies of the output image.

accuracy of the scanner, especially for larger pictures. We use a drum scanner, which gives a good performance in this respect; its head is shown in *fig. 3*. For images of dimensions 180 × 240 mm the scanning time ranges from 8 to 160 minutes, depending on the resolution. (In our experimental work most pictures are smaller than the format mentioned here.) The software is able to handle image arrays of up to four million (2048 × 2048) pixels. In the cases we shall deal with here the grey scale has an 'information depth' of 8 bits, that is to say it consists of 256 (= 2^8) different grey levels. The grey level '0' represents black, and '255' white.



▷

Fig. 3. The scanning head of the drum photoscanner (fig. 2). By adjusting the diameter of the light spot a resolution between 25 and 100 μm can be selected. The scanning motion is achieved by a translational movement of the head combined with a rotation of the drum on which the picture is mounted.

The equipment described here has mainly been used for processing and manipulating — on an experimental basis — medical data and also data from the non-destructive testing of materials. Important projects for application have been automatic measurements on chest radiographs and cine-angiograms of the left heart ventricle [1], and also — recently — the automatic detection of early-stage tumours in X-ray pictures of the female breast [2].

The treatment of the subject here is limited to a number of image-enhancement techniques that we have developed or tested; they have all been applied in the projects just mentioned. Of course, the whole subject of image enhancement comprises much more [3], but we do at least know that our modest collection of strategies can offer real help in applications of the types considered. In the next section we deal with 'histogram' transformation, a method that modifies the distribution of the grey levels (the grey-level histogram) of the pixels, to obtain an image with more contrast than the original. Another method of image enhancement to be discussed leads to 'edge enhancement', which sharpens the contours or edges present in an image. Expanding the brightness function of an image into a Fourier series, followed by manipulating the expansion coefficients and generating the new image from the modified brightness function, is briefly treated as a third method of image enhancement. Cleaning-up images to remove noise, for example film-grain noise and blurring, is discussed in the final section under the heading 'spatial filtering'.

Contrast manipulation

A common way to improve images is to change the shape of the grey-level distribution. Each pixel is assigned a new grey-level value, which depends on the original value but is independent of the position of the pixel in the image. The latter condition is not always imposed. We have adopted it here for reasons of simplicity; it implies that the ways of assigning new grey values are restricted to single-valued distribution

Table I. The image features and the operating characteristics relating to the input and the output equipment in the image-processing system of fig. 1. A linear resolution of 25 μm corresponds to a resolution of 1600 picture elements (pixels) per square millimetre. NA means 'not applicable'.

Unit	Image		Technical details			
	Type	Dimensions (mm)	Resolution (μm)	Number of grey levels	Scan time (μs/pixel)	Generation time (μs/pixel)
	Input equipment					
Drum scanner Optronics P1000 [a]	Transparent black-and-white, positive or negative	Variable: up to 350×430	25, 50, 100	256	71 [c]	NA
Flying-spot scanner Dicomed 57 [b]	Transparent black-and-white, positive or negative	Variable: up to 57×57	25, 50, 100, 200	64, 256	35-1300 [d]	NA
	Output equipment					
Display Dicomed 36 [b]	Screen	Variable: up to 177×177	86	64	NA	57
Hard copy, Optronics Photowrite P1500 [a]	Transparent black-and-white, positive or negative	Variable: up to 550×430	25, 50, 100	256	NA	50

[a] Optronics International, Inc., Chelmsford (Mass.), U.S.A.
[b] Dicomed Corp., Minneapolis (Minn.), U.S.A.
[c] For maximum image length (350 mm).
[d] Dependent on signal/electronic noise.

transformations [4]. The total number of pixels remains unaffected, of course. Such a manipulation is frequently made to enlarge the total range of grey levels; for example, X-ray pictures and landscape photographs that were taken under poor illumination can be enhanced strikingly by this 'stretching' of their grey scale (which increases contrast). In changing the grey-level histogram three separate elements must be taken into account: the distribution present in the original picture, the distribution present in the new picture and the 'transfer characteristic', i.e. the function that produces the change of grey levels. Assuming that the first element, the original distribution, is known, then either the transfer characteristic or the final distribution can be specified — and the computer can then calculate the remaining element from this. Examples of both cases will be discussed.

A specified transfer characteristic

Fig. 4 illustrates the increase of contrast that may result from linearly transforming grey levels, an example of the case where the transfer characteristic has been specified. The image is a radiograph of a cast-iron stub axle. The X-ray picture was taken as a non-destructive test to check for air inclusions. From the original picture it is difficult to be sure whether inclusions are present or not. The enhanced image proves beyond any doubt that inclusions are in fact present in the stub axle.

The transformation procedure in this case is simple. The computer assigns to each pixel of the original image

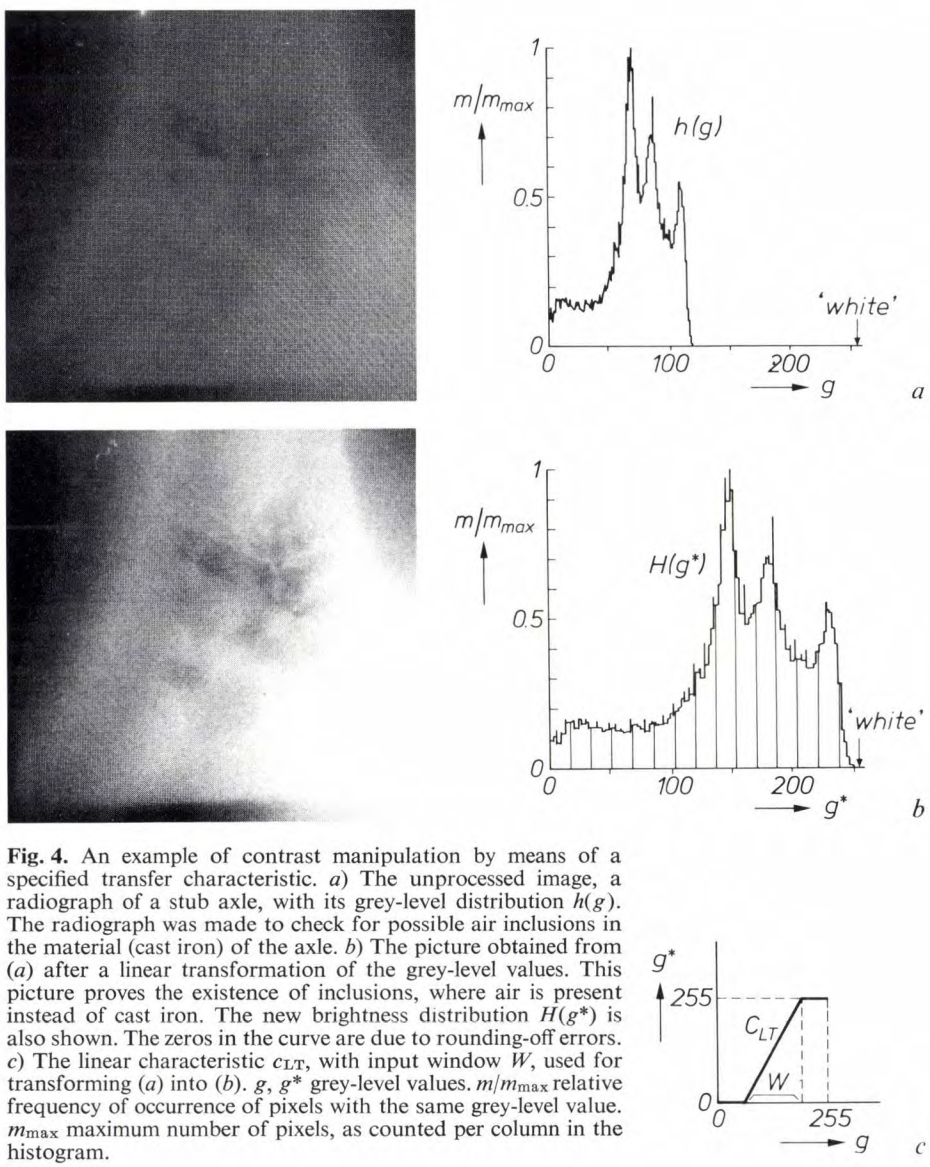


Fig. 4. An example of contrast manipulation by means of a specified transfer characteristic. *a*) The unprocessed image, a radiograph of a stub axle, with its grey-level distribution $h(g)$. The radiograph was made to check for possible air inclusions in the material (cast iron) of the axle. *b*) The picture obtained from (*a*) after a linear transformation of the grey-level values. This picture proves the existence of inclusions, where air is present instead of cast iron. The new brightness distribution $H(g^*)$ is also shown. The zeros in the curve are due to rounding-off errors. *c*) The linear characteristic c_{LT}, with input window W, used for transforming (*a*) into (*b*). g, g^* grey-level values. m/m_{max} relative frequency of occurrence of pixels with the same grey-level value. m_{max} maximum number of pixels, as counted per column in the histogram.

[2] A. Hoyer and W. Spiesberger, this issue, p. 347.
[3] See, for example, A. Rosenfeld and A. C. Kak, Digital picture processing, Academic Press, New York 1976, and also W. K. Pratt, Digital image processing, Wiley, New York 1978.
[4] R. Hummel, Computer Graph. Image Process. **6**, 184, 1977.

a new grey level, which it calculates by using the linear relationship

$$g^* = ag + b,$$

where g is the original grey level of the pixel and g^* is its new grey level; a and b are constants with a specified value. These constants determine the 'stretching' of the original grey scale. The set of pixels with the new grey-level values represents the new version of the image, which the output equipment can either display on a screen or reproduce as 'hard copy' (i.e. on a film).

The linear transfer characteristic in fig. 4 does not cover the total input range of 256 grey levels; that is to say, in transforming the grey levels the computer uses an upper and a lower limiting input value. A small fraction of the pixels may have grey levels that are outside this 'window'; for those pixels the new grey levels are set at the values 255 or 0. Since this clipping mechanism may destroy part of the information contained in the original image, the transformation described must be used with some care.

Specified grey-level histograms

For preventive medical mass screening, it it necessary to have standardized pictures, i.e. pictures with identical grey-level distributions; otherwise reliable and consistent machine processing of the images would become impossible. In this situation the transfer characteristic is the unknown quantity, which depends upon both the original and the desired histogram. In many cases the characteristic will be far from linear. We have written a computer program that makes it possible to find the required transfer characteristic for almost any pair of grey-level distributions. The desired histogram is drawn, with a light pen, on the screen of a graphic display; using the data from the original histogram the computer then calculates the required transfer characteristic.

In our experience a group of five specified brightness distributions are usually sufficient in practice for image enhancement (*Table II*). The corresponding transfer characteristic, also listed in Table II, is a function that contains an integral over the original distribution as the unknown parameter. The upper bound of the integral is the grey-level value to be transformed — the independent variable. The computer must determine this integral before it can assign to the pixels their new grey-level values — which together form the specified histogram. The transformation described by the first of the five distributions in Table II is known as *equalization*, which is a widely used and well established strategy for image enhancement.

For the case of equalization in Table II we show how the required transfer characteristic $T(g)$ can be calculated from the original grey-level distribution $h(g)$ and the new distribution $H(g^*)$. We consider the group of pixels whose original grey-level value lies between some lower limit g_{min} and the value g. The total number n of pixels in this group is a function of g and is given by

$$n(g) = \int_{g_{min}}^{g} h(g')\,dg'.$$

The function $n(g)$ increases monotonically with g. In the new histogram $H(g^*)$ this particular group of pixels have been assigned the grey-level values that range from g^*_{min} to g^*, where

Table II. Five grey-level distributions $H(g^*)$ that can be been specified for increasing the picture contrast. $h(g)$ represents the original distribution of grey levels g. The transfer characteristic for assigning the new grey levels to the pixels can be calculated from $h(g)$ and $H(g^*)$. All these transformations have been incorporated in the program package developed for image enhancement. *GLD* grey-level distribution. α is a parameter that can be varied to obtain the best visual appearance, for example.

GLD (original, known)	Transformation type	*GLD* (new, specified) $1/K = g^*_{max} - g^*_{min}$	Required transfer characteristic $g^* = T(g) \quad n(g) = \int_{g_{min}}^{g} h(g')dg'$
$h(g)$	equalization	$H(g^*) = K$	$g^* = g^*_{min} + n(g)/K$
$h(g)$	hyperbolization	$H(g^*) = \frac{1}{g^* + \alpha} \frac{1}{\ln(g^*_{max} + \alpha) - \ln(g^*_{min} + \alpha)}$	$g^* = -\alpha + (g^*_{max} + \alpha)^{n(g)} (g^*_{min} + \alpha)^{1-n(g)}$
$h(g)$	exponential function	$H(g^*) = \frac{\alpha}{1 - \exp(-\alpha/K)} \exp\{-\alpha(g^* - g^*_{min})\}$	$g^* = g^*_{min} - \alpha^{-1} \ln[1 - n(g)\{1 - \exp(-\alpha/K)\}]$
$h(g)$	Rayleigh	$H(g^*) = \frac{1}{\alpha^2} \frac{g^* - g^*_{min}}{1 - \exp\{-1/(2\alpha^2K^2)\}} \exp\{-(g^* - g^*_{min})^2/2\alpha^2\}$	$g^* = g^*_{min} + [-2\alpha^2 \ln\{1 - n(g)[1 - \exp(-1/(2\alpha^2K^2))]\}]^{1/2}$
$h(g)$	power function	$H(g^*) = (\alpha + 1) K^{\alpha+1} (g^*_{max} - g^*)^{\alpha}$	$g^* = g^*_{max} - [\{1 - n(g)\}/K^{\alpha+1}]^{1/(\alpha+1)}$

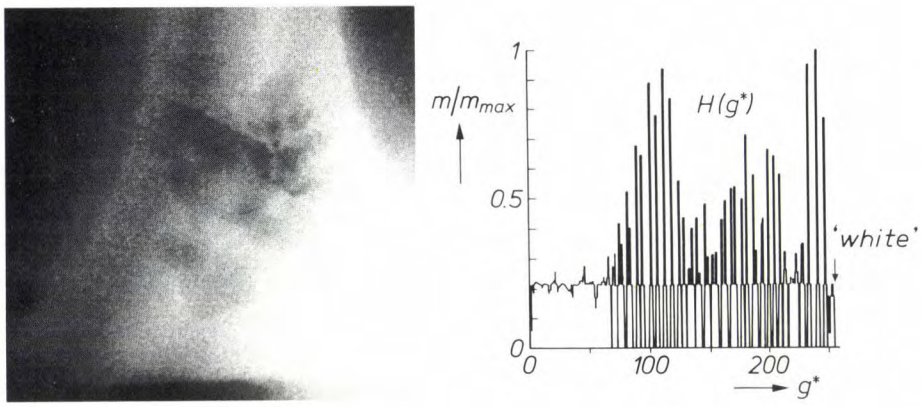


Fig. 5. The radiograph of the stub axle of fig. 4*a*, as it appears after a transformation of the equalization type (Table II) has been applied to the grey-level distribution. The new grey-level distribution $H(g^*)$ is shown on the right. The discrete nature of the original and the new distribution — they are histograms — introduces rounding-off errors, which explain the scattering of the new distribution values around the specified equalization curve.

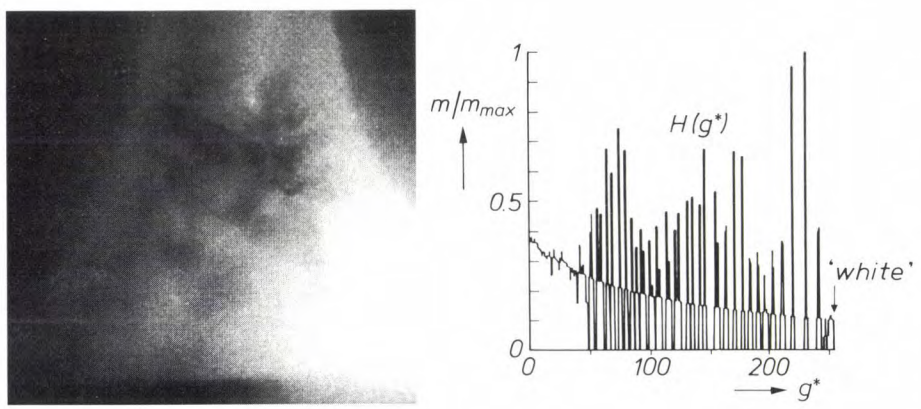


Fig. 6. The radiograph of the stub axle in fig. 4*a* as it looks after applying a transformation of the hyperbolization type (Table II) to the grey-level distribution. As in fig. 5, rounding-off errors are responsible for the scattering of the values in the new grey-level distribution $H(g^*)$ around the specified curve. The boundary of the inclusions is more clearly visible than in fig. 5.

$g^*_{\min} = T(g_{\min})$ and $g^* = T(g)$. In the assignment the *number* of pixels is not affected, so we have

$$n(g) = \int_{g^*\min}^{g^*} H(g^{*\prime})\,\mathrm{d}g^{*\prime}.$$

Substituting the specified distribution for $H(g^*)$ in the last equation, we obtain

$$n(g) = \int_{g^*\min}^{g^*} K\,\mathrm{d}g^{*\prime}.$$

Computing the integral on the right-hand side and substituting the expression for $n(g)$ we find

$$K(g^* - g^*_{\min}) = \int_{g\min}^{g} h(g')\,\mathrm{d}g'.$$

an equation that can be written as

$$g^* = \frac{1}{K}\int_{g\min}^{g} h(g')\,\mathrm{d}g' + g^*_{\min}.$$

which is the required expression

$$g^* = T(g).$$

To ensure that the transfer characteristics of the transformations in Table II are well-defined, we have set $n(g_{\max})$ equal to 1, i.e. we have normalized the histograms.

It has been proposed by W. Frei [5] that instead of equalization the somewhat more elaborate method of hyperbolization can achieve the goal more effectively. *Fig. 5* shows the improved X-ray image of the stub axle (fig. 4) after histogram equalization and *fig. 6* shows it after hyperbolization.

The fundamental assumption in both equalization and hyperbolization is that the amount of information coming from an image is maximized by making all grey levels equally probable. Hyperbolization differs from equalization in that it compensates for the logarithmic sensitivity response of the human eye; it equalizes the grey-level distribution *subjectively*, that is to say 'behind' the eye.

The example given (figs 4, 5 and 6) shows that hyperbolization can give a better result. In particular, comparison of the brighter parts of figs 5 and 6 shows immediately that hyperbolization gives a much more clearly defined boundary for the inclusions.

We have also applied hyperbolization in a 'zonal' form of image processing (*fig. 7*). The word 'zonal' refers to the fact that only a part (or zone) of the original picture is subjected to the enhancement procedure. The technique employed is 'interactive', that is to say the observer, viewing the complete original picture on the input screen of the equipment, draws the closed boundary of the zone of interest with a light pen. The computer then generates the new picture,

[5] W. Frei, Computer Graph. Image Process. **6**, 286, 1977.

ignoring all parts of the image outside the boundary. Some medical and technical users prefer this relatively slow way of working, because it gives them a more direct and flexible control of the image-processing activities; interactive zonal processing allows 'closer contact' with the material.

Let us return for a moment to the simplest of the methods given in Table II, equalization. Although this method is useful in many cases, images are occasionally encountered in which equalization does not produce any useful results at all. If, for example, an image contains certain parts or segments that appear in the original histogram as well-separated peaks, then equalization would destroy this information completely. In such cases threshold techniques make it possible to display only one part of each new image by simply omitting all pixels whose grey levels lie outside certain limits. Of course, to adapt a threshold technique to some kind of machine processing on a large scale, the various peaks are required to have standardized positions, which again calls for a prescribed grey-level histogram.

In general, the specification of both a grey-level histogram and of a transfer characteristic requires careful consideration of the image and of the particular objectives.

As already mentioned, in all the situations described in this section the grey-level values of the pixels are changed *independently of the positions* of the pixels. This feature does not apply in the 'edge-enhancement' procedure that we shall now discuss. Here the grey-level changes of the pixels depend on their direct environment.

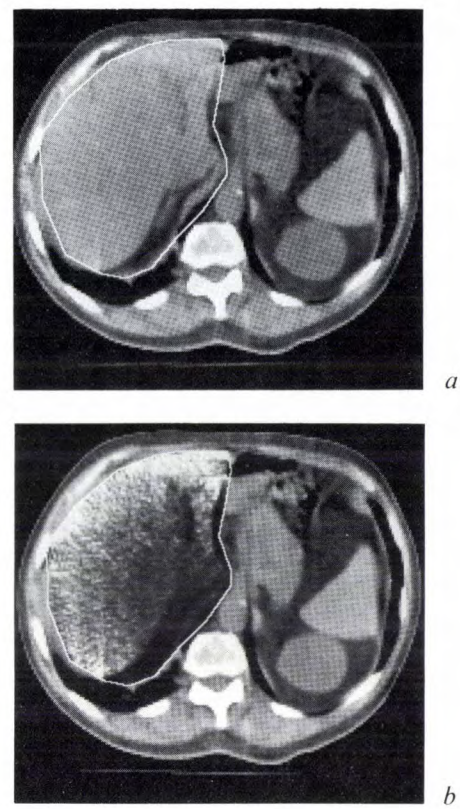


Fig. 7. An example of interactive 'zonal' image processing. *a*) The original picture, a horizontal cross-section of the abdomen recorded and reconstructed with a Philips computer tomograph (256×256 pixels, with a grey scale of 256 levels). The zone of interest is the liver, which has been indicated by drawing the boundary with a light pen. *b*) The picture after histogram hyperbolization — within the region of interest only. An important groove-shaped area on the lower surface of the liver, the porta hepatis, is clearly visible in the upper right-hand corner.

Edge enhancement

For analysing image content it is frequently necessary — and sometimes sufficient — to detect the edges or contours in the images. 'Edge enhancement', as the expression suggests, makes the edges more conspicuous. Apart from 'sharpening' the information about the edges, the method gives an increase in efficiency since less data has to be transmitted or stored if it is sufficient to base further image analysis on the edges alone.

An edge reveals its presence by steep jumps in the grey-level value when two neighbouring pixels — one inside, one outside the edge — are compared. The method is based on the calculation of differences in grey level for pairs of neighbouring pixels. These differences, when inserted into a suitable mathematical expression (for calculation of a 'gradient'), determine the grey-level values for the enhanced image that only shows the contour.

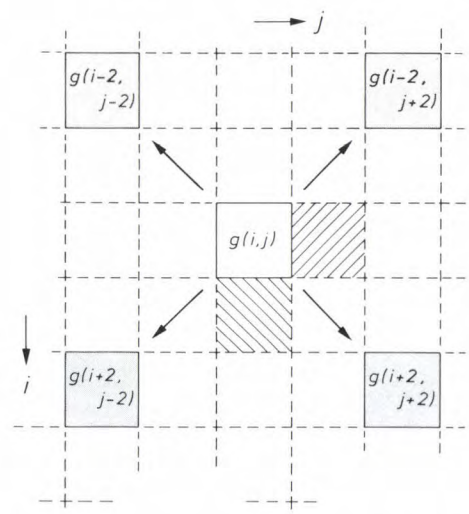
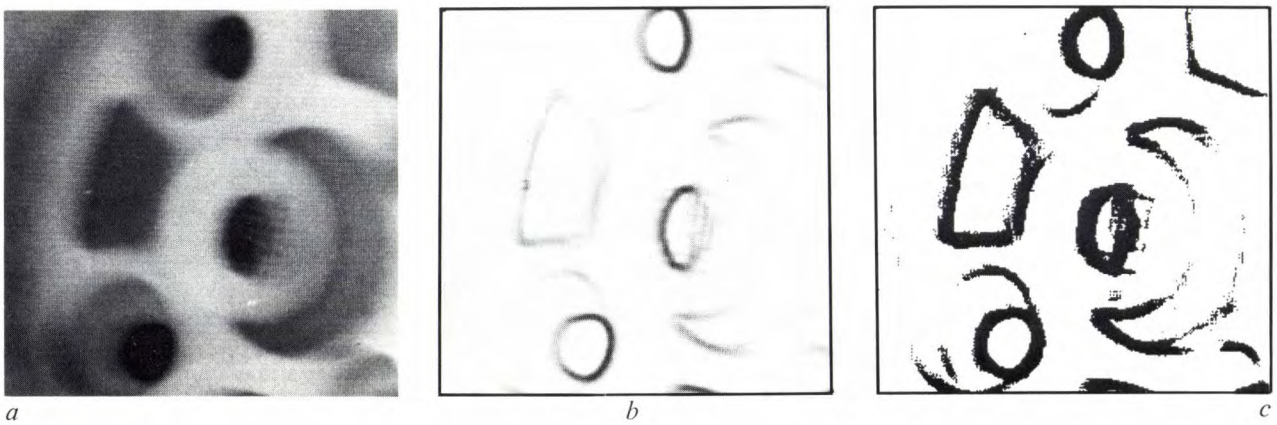


Fig. 8. In edge enhancement the new grey-level value $g^*(i,j)$ is calculated from the four original values that apply to the (grey) corner pixels of a local environment consisting of 5×5 pixels. This symmetric arrangement leads to less noise enhancement than the other one: the centre pixel itself and the two hatched ones — an arrangement that corresponds more closely to the gradient of the brightness function.



a *b* *c*

Fig. 9. An example of edge enhancement by means of the (spatial) first derivative of the brightness function. *a*) The unprocessed image, a radiograph of a car wheel. *b*) The same picture after edge enhancement. *c*) The picture of fig. 9*b* after a threshold procedure to reduce noise. The procedure assigns to all pixels either 'white' ($g = 255$) or 'black' ($g = 0$). Some of the noise pixels, apparently occurring on the 'wrong' side of the threshold, remain visible.

We have found from our experiments that the most useful gradient expression — which could be used, for example, in the study of mammograms — is:

$$g^*(i,j) = [\{g(i-2,j-2) - g(i+2,j+2)\}^2 + \{g(i-2,j+2) - g(i+2,j-2)\}^2]^{\frac{1}{2}}.$$

Here g^* is the transformed grey-level value. The environment of a pixel that is used for calculating the grey-level differences consists of a square of 25 pixels (*fig. 8*). The expression given for g^* most closely meets the foremost requirement that even weak edges should become clearly visible, and that edge broadening and the presence of noise should not spoil the result. (Noise in images reveals itself mostly as isolated pixels or small clusters of pixels whose grey-level value differs considerably from the values of the surrounding pixels. This may lead to the 'detection' of spurious edges.) The calculation procedure is a discrete form of a spatial differentiation (hence the term 'gradient') of the brightness distribution. At the edges a gradient possesses an extreme value, which explains the improved perceptibility.

Fig. 9 shows the effect of edge enhancement on a picture that represents an X-ray image of a car wheel. The edges or contours in fig. 9*b* are clearer than in fig. 9*a*; fig. 9*c* is better still, because of reduced noise. Fig. 9*c* was obtained by subjecting the pixels of fig. 9*b* to a thresholding operation. All the pixels whose grey-level value lies beyond the threshold are in this way made white, whereas all the other pixels are turned black. Thresholding is successful provided the noise pixels give smaller gradients than the edge pixels. This condition has apparently not been met for all the noise pixels in fig. 9*b*, because a few of them are still present in fig. 9*c*. To deal with such cases we have developed a final clean-up procedure, in which all the pixels that originally have less than two black (nearest) neighbours are made white.

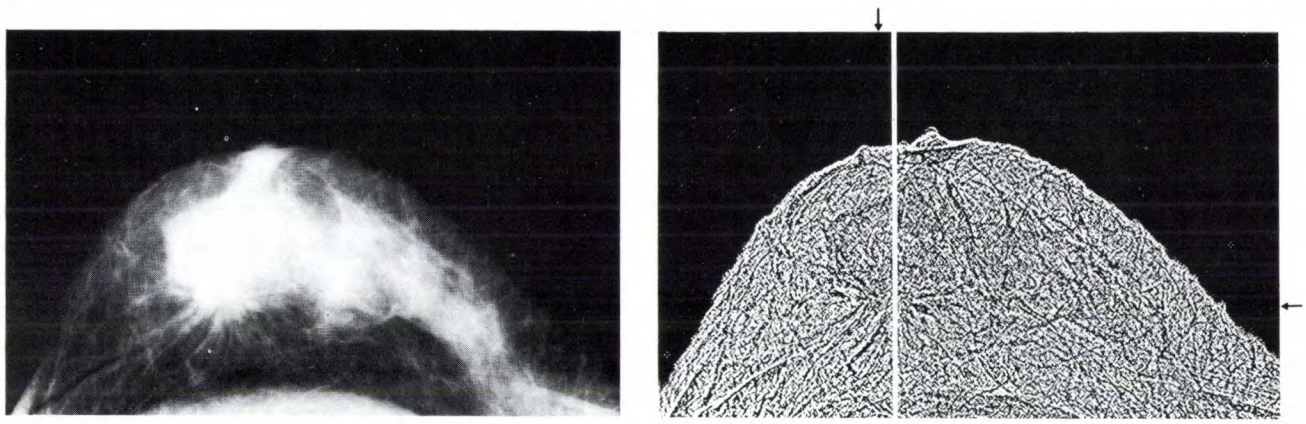
The existence of an extreme value at the edges implies that the second derivative goes through zero there. This fact can serve as the basis for another method of image enhancement. The method employs a discrete version of the Laplacian operator (the second derivative) in two dimensions. In terms of programming the method is more complicated than the method based upon the gradient, because the second derivative takes positive *and negative* values near edges. The second derivative cannot therefore be used directly as a new grey-level value. We have investigated a number of strategies that circumvent this difficulty. The resulting method of edge enhancement would, for example, assist the early diagnosis of a malignant tumour by making it more clearly visible in a mammogram as illustrated in *fig. 10*.

Fourier-domain processing

Another method of image enhancement can be based upon the discrete Fourier expansion of the brightness function $g(i,j)$. For this purpose we have included in the computer software a program with which the Fourier (expansion) coefficients can be calculated from the inverse Fourier transformation:

$$G(u,v) = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} g(m,n) \exp\left\{\frac{-2\pi i}{M}(um + vn)\right\}.$$

Here $g(m,n)$ are the $M \times M$ grey-level values of the original image; $G(u,v)$ with $u,v = 0, 1, 2, \ldots, M-1$ are the $M \times M$ Fourier coefficients. The set of these coefficients, each treated as a transformed grey-level value, forms the Fourier image. *Fig. 11* shows an example. Although it is not immediately apparent, the Fourier image contains all the pictorial data from the original image. The best proof of this is the fact that the transformation is fully reversible.



a *b*

Fig. 10. An example of edge enhancement in a radiograph of the female breast (a 'mammogram') using the second derivative, the Laplacian. *a*) Unprocessed picture, composed of 1024×512 pixels with 256 grey-level values. The lower left of the main white area looks 'suspect', but the evidence of an abnormality is not strong. *b*) The processed picture. The arrows indicate the location of a star-like structure, which greatly enhances the suspicion that a tumour (scirrhus) is present. The vertical white boundary is an artefact (the image has been processed in two parts).

The geometric centre of the Fourier image (fig. 11*b*) falls within the pixel for which $u = v = 0$, called the zero-frequency point. The word 'frequency' here comes from the fact that m and n in the expression for the inverse Fourier transformation are to be seen as spatial *coordinates*, so that u and v become spatial *frequencies*. In scanning an image it is found that the discrete gradients of the grey-level values determine the spatial frequencies present. Low spatial frequencies are related to smooth regions of nearly uniform greyness, whereas high frequencies refer to edges or points where the grey level abruptly changes. Straight lines in the Fourier image that pass through the geometric centre correspond to straight edges in the original picture.

The coefficients $G(u,v)$ are generally complex numbers, with a modulus (or absolute value) and an argument. Both the argument, which is a kind of phase angle, and the modulus can be used as grey-level values to define a Fourier image. In fig. 11 the logarithm of the modulus has been used as the grey-level value.

To build a 'normal' image from a Fourier image, the computer calculates the set of grey-level values $g(m,n)$ from the Fourier transform:

$$g(m,n) = \frac{1}{M} \sum_{u=0}^{M-1} \sum_{v=0}^{M-1} G(u,v) \exp\left\{ \frac{2\pi i}{M} (mu + nv) \right\};$$

this expression is the inverse of the first one. The technique of Fourier-domain processing reduces in fact to a manipulation of the expansion coefficients $G(u,v)$, followed by a recalculation of the 'normal' picture from the transformation given above. In this way certain degradations present in the original can be removed. The problem is to find the mathematical description that models the image degradation in the Fourier image; as soon as the mathematical description is available the Fourier coefficients can be suitably modified to 'filter out' the degradation. *Fig. 12* illustrates a simple case of such a manipulation. The only change in this case is that all coefficients for which the spatial-frequency expression $(u^2 + v^2)^{\frac{1}{2}}$ exceeds a threshold value v_{thr} are set equal to zero. This procedure works as a lowpass filter, and as such can be used for suppressing picture noise. The example in fig. 12 shows that this noise 'clean-up' is accompanied by a certain amount of defocusing in the image.

If the procedure is reversed, so that all Fourier coefficients with $(u^2 + v^2)^{\frac{1}{2}}$ *smaller* than v_{thr} become zero, pictures are obtained in which the edges are enhanced (and, to some extent, the noise too). These pictures resemble the pictures based on gradient calculations, as shown in the example of fig. 9*b*.

Spatial filtering

From the preceding sections it will be evident that picture noise is a problem for which there is no universal remedy. We have found that thresholding, the clean-up procedure and Fourier-domain processing, as described, are all procedures that can cope with noise to some extent, but they have their shortcomings. Situations are encountered in which these procedures are inadequate or even fail outright. The complete failures that we have experienced could usually be attributed to unwanted side-effects of the noise suppression procedure on the good parts of the picture.

In an effort to solve these difficult cases of noise removal we have studied some methods of 'spatial filtering' and added them to the software package of our image-processing system. They are based on an increasingly refined method of averaging the grey-level

[6] M. Schlindwein, Iterative three-dimensional reconstruction from twin-cone beam projections, IEEE Trans. **NS-25**, 1135-1143, 1978.

values of small groups of neighbouring pixels. The methods work well in pictures where the noise consists of an ensemble of isolated pixels or small clusters of pixels. The positions of the noise pixels may be distributed randomly over the picture, producing what is aptly described as 'salt-and-pepper' noise. Known causes of this are errors during image transmission, the grey-level values in a square of $(2k + 1) \times (2k + 1)$ pixels with (i,j) at the centre:

$$\overline{g_k(i,j)} = \frac{1}{(2k+1)^2} \sum_{m=-k}^{k} \sum_{n=-k}^{k} g(i+m, j+n).$$

Figs 13a and *b* show results obtained with this procedure; the original picture is a simulation of a computer-

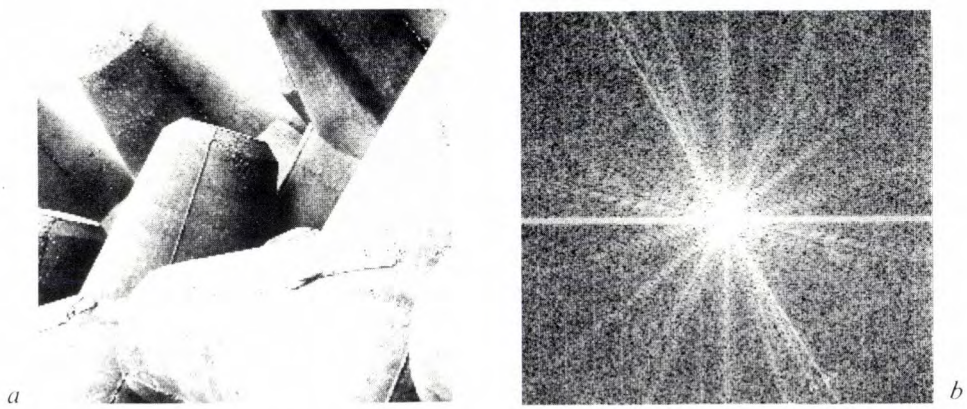


Fig. 11. An example of a Fourier transformation. *a*) The unprocessed (digitized) picture represents concrete blocks for beach conservation on the German island of Sylt. *b*) The image obtained by subjecting (*a*) to a Fourier transformation. The centre point has the coordinates $u = v = 0$ (frequency zero, see text). A straight edge in (*a*) representing a transition from $g = 0$ to $g = 255$, corresponds to a straight white line in (*b*) in the perpendicular direction. The brightness function $G(u,v)$ possesses a constant modulus on such a line, independent of the spatial-frequency coordinates u and v.

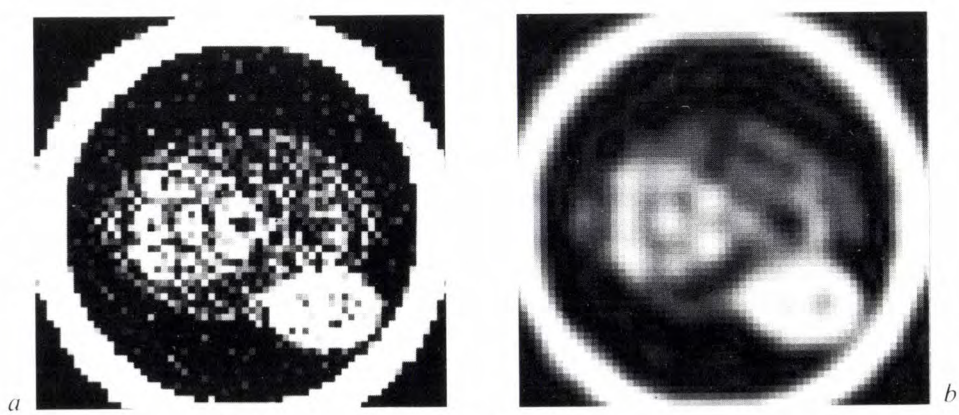


Fig. 12. An example of noise clean-up by means of Fourier-domain processing, in a (simulated) horizontal cross-section of the abdomen. *a*) The unprocessed (digitized) picture. The appearance of two major organs, the white patch close to the boundary and the one slightly off-centre, is seriously marred by noise pixels. *b*) The picture after manipulating the Fourier expansion coefficients and recalculating the 'normal' picture. Most of the noise is suppressed. In this case the manipulation of the Fourier coefficients has acted as a lowpass filter, making the coefficients of all spatial frequencies above a certain threshold value equal to zero. The 'normal' picture (*b*) arises from back-transforming the Fourier-transform image (not shown) of (*a*).

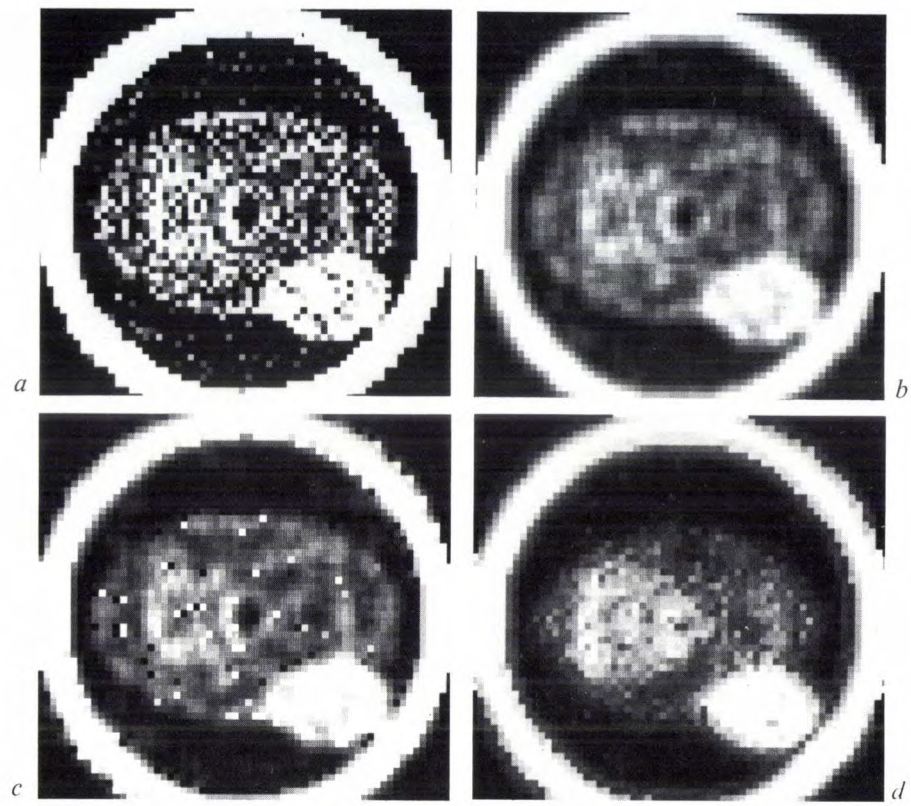
noisy components of the electronic system (e.g. sensors and image intensifiers), and the grain structure of the film. The simplest method of spatial filtering is *conventional averaging*. In this case the grey-level value $g(i,j)$ of the pixel (i,j) is replaced by the average $\overline{g_k(i,j)}$ of

produced tomogram [6]. In the procedure described *all* pixels of the original image are subjected to the averaging, irrespective of local pictorial content. Sharp edges, for example, tend to suffer from the averaging procedure; to some extent they can lose their sharpness.

The second method, called *adaptive averaging*, was introduced to avoid this side effect which blurs or smoothes the edges. Fig. 13*c* shows the result when adaptive averaging is applied to the original picture (fig. 13*a*). This method enables the computer to decide for each pixel whether the averaging calculation described above is to be performed or not. To make the decision, the computer calculates the 'difference operator' $d(i,j)$:

$$d(i,j) = |g(i,j) - g(i+1,j+1)| + |g(i,j+1) - g(i+1,j)|.$$

If $d(i,j)$ exceeds a certain threshold value, no average is taken for that particular pixel, because the value of $d(i,j)$ is then interpreted as an indication of the presence of an edge. As fig. 13*c* shows, the effect of this type of spatial filtering can be remarkably good. Some noise is still present at isolated pixels, however; apparently there has been no averaging here, because these pixels have (wrongly) been interpreted as belonging to edges.



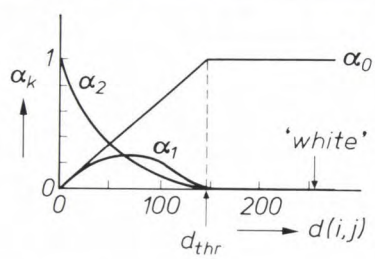


Fig. 13. An example of noise clean-up in a (simulated) cross-section of the abdomen, by means of three spatial filtering methods. *a*) The unprocessed (digitized) picture. The appearance of two major organs, the white patch close to the boundary and a region just off-centre, is badly affected by noise pixels. *b*) The picture after conventional averaging (over squares of 3 × 3 pixels). Sharp edges, e.g. various boundaries, become blurred, a side-effect of this procedure. *c*) The picture after adaptive averaging, a procedure that avoids the edge blurring adequately. Isolated noise pixels are still present. *d*) The picture after continuous-adaptive averaging, a nonlinear method of 'averaging averages' (see text). The isolated noise pixels in (*c*) have been suppressed. The graph gives the coefficients α_0, α_1, α_2 of the various averages chosen in the continuous-adaptive averaging procedure that has resulted in (*d*). The difference operator $d(i,j)$ is plotted horizontally (see text). d_{thr} is the value above which the 'averaging of averages' is omitted; only the original grey-level value $g(i,j)$ has been taken into account here.

We have overcome this spurious interpretation by making the decision function less sharp (more 'fuzzy'). This means that the decisions are no longer of the discontinuous yes/no type.

The third method of spatial filtering we shall discuss here is '*continuous-adaptive*' *averaging*. Fig. 13*d* shows

the result when this method is applied to the original picture (fig. 13*a*). To calculate the new grey-level values, this type of spatial filtering works with a rather complicated expression, a (short) series of averages;

$$g^*(i,j) = \sum_{k=0}^{2} \alpha_k \overline{g_k(i,j)}.$$

The average $\overline{g_0(i,j)}$ is simply the original grey-level value $g(i,j)$. The remaining two averages are calculated as indicated for the first method of spatial filtering, over squares of 3×3 and 5×5 pixels. These squares *both* have the pixel (i,j) as their centre points. The program for the calculations makes it possible to extend the series by adding the averages over still larger squares; whether this is a useful facility, however, depends upon the relative importance of any improvement in noise removal in comparison with the longer processing time required.

The coefficients α_k of the series are functions of the difference operator $d(i,j)$ such that the sum of all coefficients is always equal to 1. In this way the various averages are themselves averaged, which counteracts the crude yes/no behaviour of $d(i,j)$ as encountered in the second method of spatial filtering. Fig. 13 also shows the coefficients used in the processing that results in fig. 13*d*.

Unlike the first type of spatial filtering, both the adaptive types are nonlinear in the original grey-level values, since there is an additional dependence upon the grey-level values — via the difference operator.

Outlook

The various strategies of image enhancement described in this article are meant to be part of a complete system for real-time picture processing, on which work is in progress at PFH Hamburg. Such a system will give the user two more advantages of great practical value:

— efficient storage; storing pictures in digitized form requires far less space than any other method of storage, such as keeping radiographs in ordinary filing cabinets;

— easy accessibility; subjecting a stored picture to a data-processing operation becomes a simple matter of using the appropriate computer program.

The larger hospitals, for example, could benefit considerably from such a system; instead of the cumbersome files of conventional pictorial data — adding half a million (!) new radiographs every year is no exception nowadays — an electronic 'picture bank' promises to become the basis for the work of medical specialists. Finally we should note that for diagnostic purposes it is by no means necessary to make a radiograph first and then to digitize it. Pictures can be recorded directly in digitized form and the enhanced version displayed in real time on the screen.

Summary. A software package has been developed for digital image enhancement that comprises contrast manipulation (histogram transformation), edge enhancement (gradient and Laplacian), and Fourier-domain processing, all with optional interactive zonal processing. Noise suppression by spatial filtering is included. 'Continuous-adaptive' averaging, one of the methods investigated for spatial filtering, avoids spurious edge-detection. Applications to pictures from medical X-ray diagnosis and non-destructive materials testing demonstrate the enhancement of visual appearance and prove at the same time that standardized picture preparation for automatic feature extraction is possible. Input and output of film transparencies up to 350×430 mm (black and white, positive and negative, 256 grey levels, resolution up to 1600 pixels/mm^2). The hardware consists of a flying-spot scanner, a drum photoscanner, a 'photowrite' hard-copy unit and a storage-tube display, all controlled by a Philips P880 minicomputer.

Philips tech. Rev. 38, 310-325, 1978/79, No. 11/12

The Dot Scan CCTV, a flexible system for real-time image-processing experiments

L. H. Guildford

Introduction

Image processing is most frequently performed 'off-line', the processing operation being well separated in space and time from the acquisition of the image. Weather maps, for instance, are produced by satellites scanning the Earth's atmosphere and transmitting the video information to ground stations, where the original picture is reconstructed at a relatively slow rate. To extract all the relevant information from the received images, they are processed frame by frame in elaborate ways involving large computer systems. In medical radiography, exposure times and the number of exposures per patient must be kept to a minimum, and here again elaborate processing is required — later and elsewhere — to extract all information from the image. Generally speaking, where the acquisition of the image in some way presents difficulties and there is no time penalty, the processing will normally be carried out off-line.

Quite a different situation arises if an infrared imager is used, for instance to study the habits of nocturnal mammals, or for observing military vehicles or inspecting for 'hot spots' in industrial processes. In these instances only hot objects are observed and the unfamiliar picture may be made much more intelligible by some form of processing. This operation must obviously occur in real time to be of any use. The same argument holds for the processing of X-ray images used by doctors for observing the function of organs in a patient infused with contrast fluid. Similarly, if a car is equipped with radar for use on an airfield under conditions of very poor visibility, then again any picture processing needs to be done in real time. Image processing in real time could also be very useful in other fields of imaging, such as television at low light levels, ultrasound imagers, electron-microscopy and facsimile.

In all such cases, real-time electronic processing of the video signal is the best method to use. The designer of processing 'operators' for these purposes then meets the following problem: in general it is not advisable to build a system from the paper design without preliminary experiments. This is particularly so in the case of imaging systems, as it is very difficult to visualize from the design what subjective impression the processed images will have on the observer. It is even more difficult to pass on the impression to a potential user. On the other hand, preliminary experiments that *will* demonstrate the effect of processing the image are often impractical as they involve building an almost complete imaging system. This is because the nature of the scanning system from which the video signal is derived may vary enormously from one field of application to another. It may for instance be an electron beam scanning the image of the scene on a target electrode, as in a vidicon tube, or it may be a single detector element scanning the scene itself by means of a system of rotating or reciprocating mirrors or prisms, as with many infrared imagers. As a consequence, the array of picture elements ('pixels'), the order in which they are scanned, i.e. the 'scan format', and the scanning rate all differ widely (*Table I*). Now, the implementation of a processing operator is closely associated with the scan format. Consider a processor that is required to enhance vertical edges in the scene. This may be achieved by subtracting from the original picture an identical picture of reduced amplitude, shifted horizontally by a small amount. For a system with sequential line scanning this would be a simple operation of subtracting from the incoming signal an attenuated video signal delayed by a few pixel periods. For a radial scan, however, there is no such simple relation between 'horizontal neighbours', and a much more complicated operation would be required. In general, a processing operator cannot always be subjectively tested by means of a standard television camera and display monitor; there is a wide range of 'professional' imaging systems that of necessity incorporate unusual scan formats.

With this problem in mind we have built our 'Dot Scan Closed Circuit Television' system. This imaging system can simulate all kinds of imaging systems and image-processing operators, as it is equipped with a highly programmable scan format. The basic block diagram is shown in *fig. 1*. Various processing opera-

L. H. Guildford, M.I.E.R.E., is with Philips Research Laboratories (PRL), Redhill, Surrey, England.

Imaging system	Pixel period / Frame period	Scanning technique	Scan format
Commercial TV		electronic	rectilinear, interlaced
Experimental TV		electronic	rectilinear, multiple interlace
Facsimile		mechanical (rotating drum) mechanical with electronic line scan	rectilinear
Radar		mechanical electronic (phase arrays)	radial
Infrared		electronic (pyroelectric vidicon) mechanical (mirrors, prisms, ...) mechanical with electronically scanned detector arrays	rectilinear rectilinear, sinusoidal, ...
Ultrasound		electronic (phase arrays)	radial
Scanning electron microscopy		electronic	rectilinear

10^{-8} 10^{-6} 10^{-4} 10^{-2} 1 10^{2} s

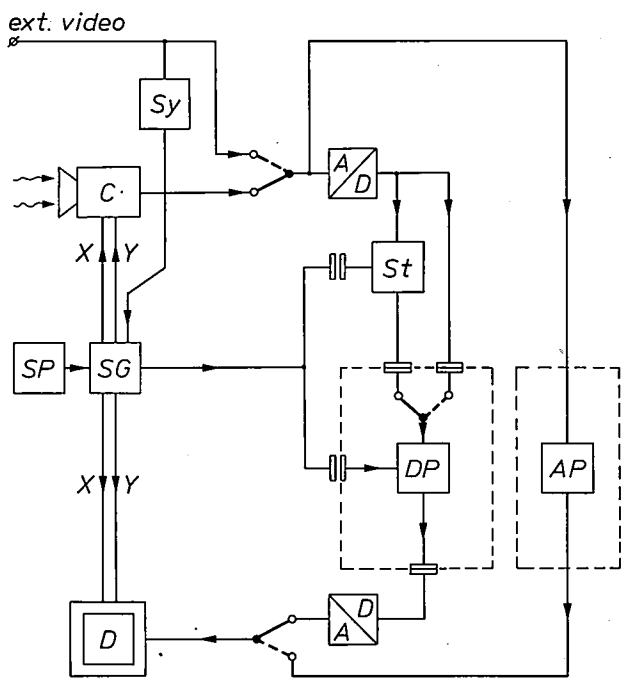


Fig. 1. Block diagram of the Dot Scan CCTV system. Between camera *C* and display *D* digital or analog processing operators (*DP*, *AP*) can be connected. Scan formats of camera and display are freely programmable. The deflection voltages (*X*, *Y*) of the scanning electron beams are produced by the scan program *SP* and the digital scan generator *SG*. The synchronization circuit *Sy* provides synchronization of *SG* if an external video signal is used. The store *St* has available, for processing in real time, the signal values belonging to a group of pixels surrounding the pixel that is currently being displayed. *A/D* and *D/A* are converters from analog to digital and vice versa.

tors (*DP*, *AP*) can be incorporated into the system between the camera (*C*) and the display (*D*). The sequences of the *X* and *Y* deflection voltages for the electron beam in the vidicon — which has electrostatic deflection — and the display tube are obtained from a central, programmable, digital scan generator (*SG*). In place of the camera an external video source can be connected to the system. The scanning program is then adapted and synchronized to this external video source.

In the context of this article the term 'real time' means that the action on the display is continuous and occurs without any significant delay in relation to the scanned scene or incoming video signal. This implies that the computation made by the processing operator must be completed within each frame period, and preferably within a single pixel period. The computation invariably involves the signal values of a group of neighbouring pixels, and this information must be available in real time. For this purpose a store (*St*) has been incorporated in the system. The store and the processing operator are synchronized to the scanning generator. *Fig. 2* shows the simulator in use.

The store in the present system has a capacity for the video information of eight bits per pixel on an area of 4096 (= 64 × 64) pixels. In the near future a larger store will be incorporated, with a capacity up to 512 × 512 10-bit pixels. This will allow full-frame processing to be undertaken. Another extension of the system will make it possible to acquire and display an

image using standard cameras and display tubes, i.e. with a format of two interlaced fields per frame, and still examine the effects of unusual scan formats and processing operators. For this purpose two additional full-frame stores will be added for the conversion from standard format to the format under investigation and scan formats or processing operators. It will enable them to obtain an impression of the effects of various types of processing operators prior to building the proposed system. Alternatively the system may be used directly as an image processor for laboratory experiments. For instance, if a biological process such as

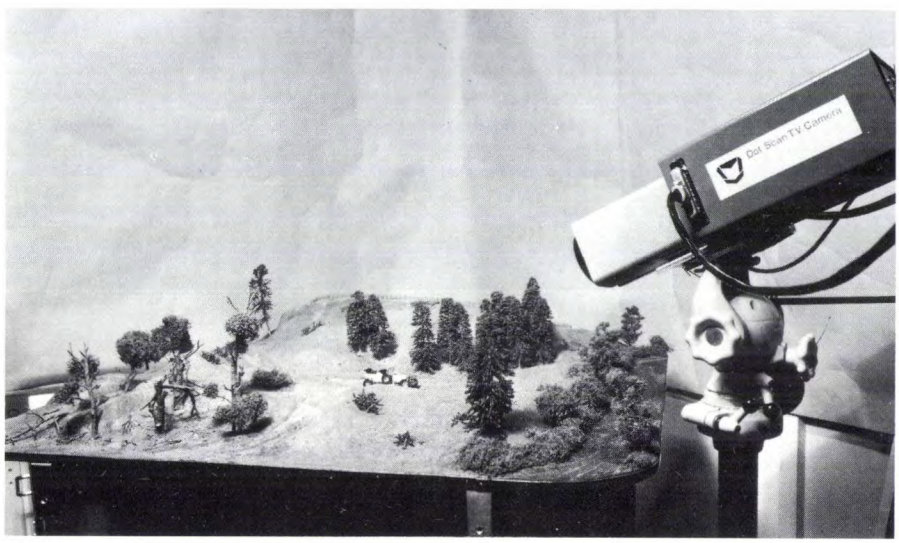




Fig. 2. The Dot Scan CCTV system. *Above:* the camera, viewing the diorama that is used in many experiments where repeatable lighting conditions are required. Conditions of poor visibility are simulated by the use of carbon-dioxide vapour. *Below:* the upper of the two displays on the left has electrostatic beam deflection and is therefore better suited for demonstrating novel scan formats. The lower one has a higher resolution (800 lines) and produces better contrast, but cannot follow highly random scans. On top of the two displays is an oscilloscope showing the grey-level histogram of part of the scene. The video signals are controlled and the scan signals are generated in the rack at the centre. The rack on the right houses the store and the processing operators.

vice versa. Standard videotape can then be used both as an input medium and as a means for recording the output of the processor.

The system is primarily intended to provide an experimental service for our colleagues who are planning to build imaging systems incorporating unusual cell division has been recorded on videotape and it is felt that edge enhancement would clarify the process, the tape could be processed on the Dot Scan CCTV.

In the following sections the hardware will be examined in more detail and three examples of scanning and processing techniques will be discussed: tele-

vision systems with a slow scan rate, systems with edge enhancement and systems with contrast enhancement. These examples will provide an opportunity of illustrating the special features of the Dot Scan CCTV system and for discussing the way in which the equipment is used.

horizontal bar moving four times per second from top to bottom. Secondly, objects in continuous motion are displayed as jumping their position four times a second. One solution to these problems is to break up the 'coherent' picture-replenishment pattern and replace it by a 'two-dimensional multiple interlace' scan format

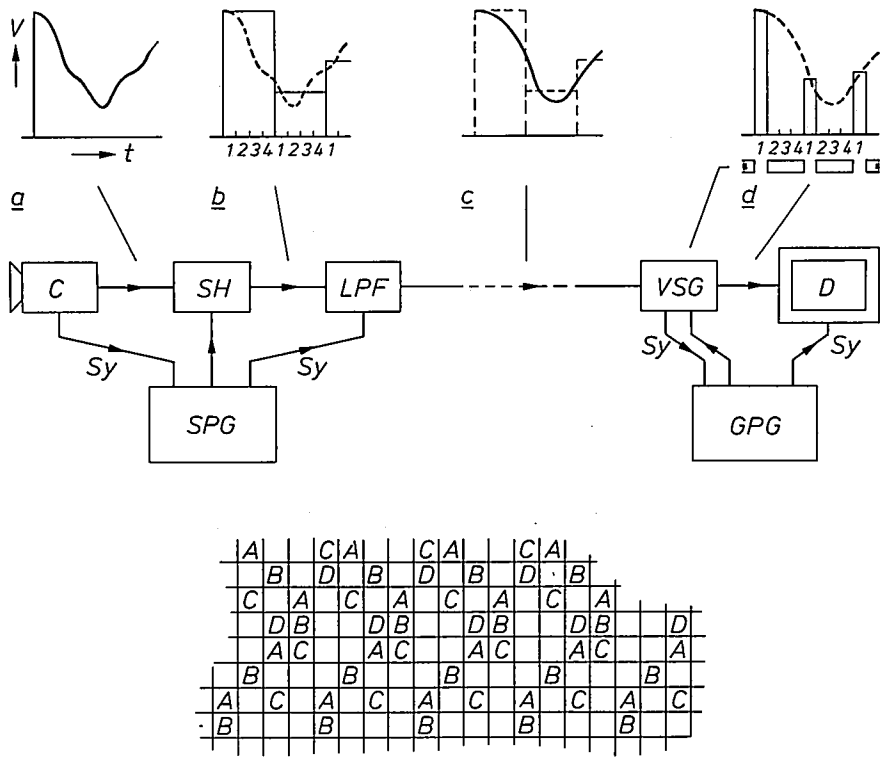


Fig. 3. 'Sampledot', a method of breaking up the coherent picture-replenishment pattern at low frame rates in order to produce acceptable pictures for transmission via a channel of reduced bandwidth. The figure applies to the reduction of the bandwidth by a factor of 4. *Above:* in the sample-and-hold unit *SH* the video signal (*a*) from a conventional camera *C* is sampled every fourth pixel and held for four pixel periods (*b*). Passage of the signal over the transmission path (the lowpass filter *LPF*) results in the waveform (*c*), and the video-sampling gate *VSG* reproduces from *c* the original signal value of the sampled pixels (*d*). By a correct offset of the sampling time for each field, all pixel values are reproduced in eight fields. The units *SPG* and *GPG* are the sampling and gating pattern generators. *D* display, *Sy* synchronization. *Below:* pseudo-random sampling, obtained by starting the sampling in each line at a pseudo-random moment.
A: first TV field, odd lines,
B: second TV field, even lines,
C: third TV field, odd lines,
D: fourth TV field, even lines, etc.

Television systems with a slow scan rate

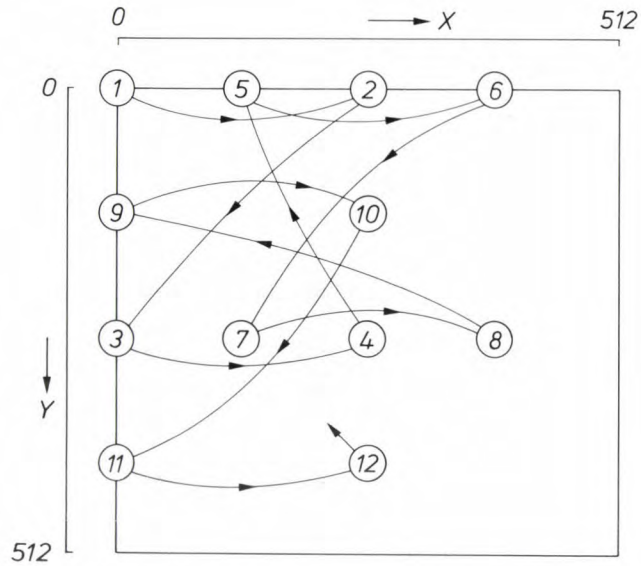
When a television system is operated over a video transmission channel of less than normal bandwidth, the transmitted information rate will also be lower than normal. For a system with sequential line scanning the straightforward answer would be to reduce the scanning rate from the usual 25 frames per second to, let us say, four frames per second. This entails two marked inconveniences. In the first place one does not see a completely flicker-free picture, but one with a

at the display [1]. The point at which the phosphor is activated does not then follow sequential lines but jumps within the display area. The price to be paid is mainly a loss of resolution of moving objects. The usefulness of such scanning techniques is determined by the subjective impression of the images they produce. We have used our Dot Scan CCTV to investigate various techniques in this respect.

[1] P. M. G. Toulon, U.S. Patent No. 2479880, Aug. 1949.

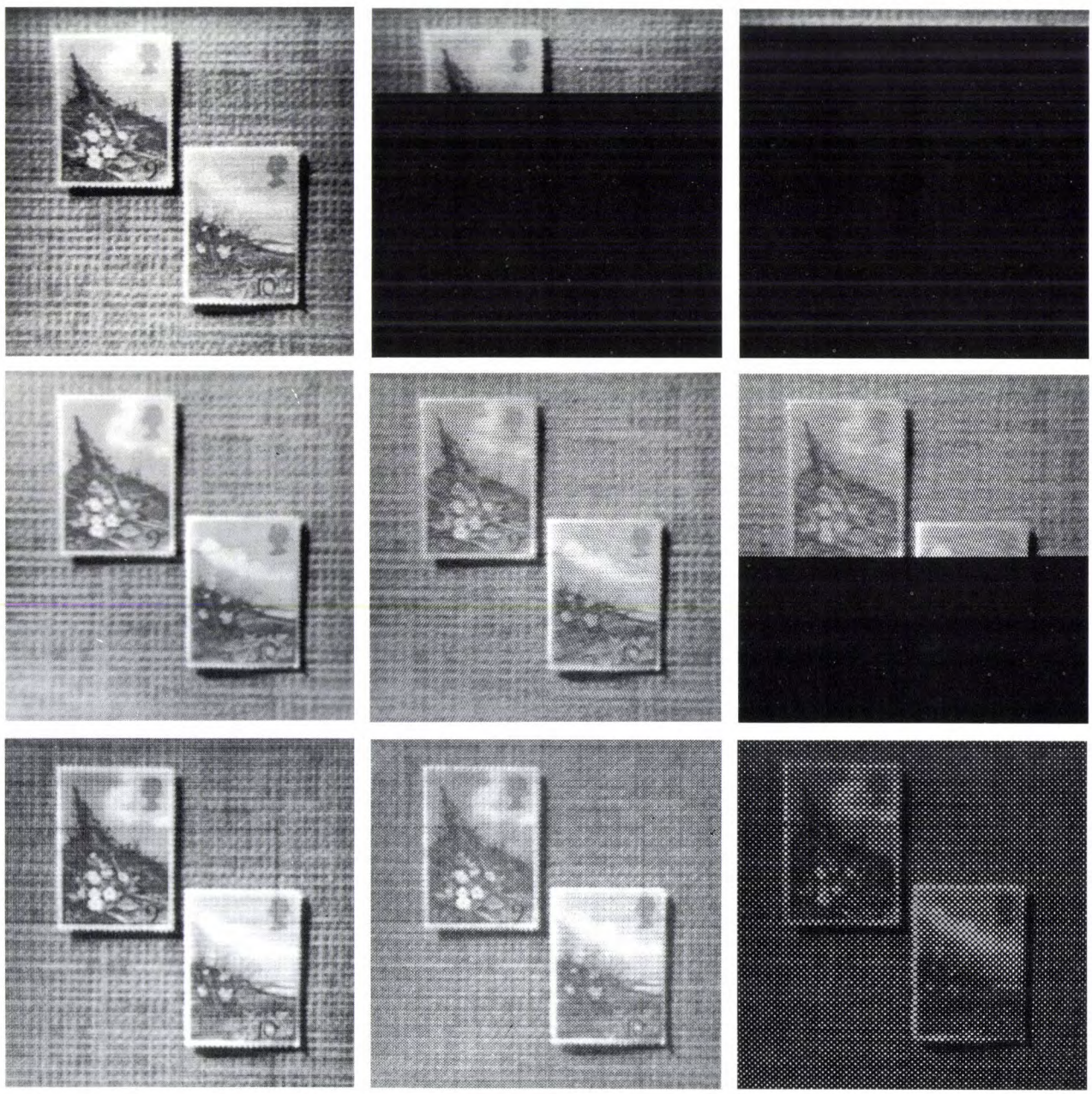
▷

Fig. 4. A Pseudo-Random Scan pattern produced with the Dot Scan CCTV. Successively scanned pixels are at least half the scan width apart, so that the coherence of the picture replenishment is very effectively broken up. When normal phosphors (P4) are used, acceptable pictures are obtained at frame rates down to 0.5 per second. The manner in which this pattern is generated is shown in fig. 6.



▽

Fig. 5. Photographs of a scene on the display with exposure times of a full frame period (*left*) and of $\frac{1}{4}$ and $\frac{1}{16}$ of a frame period (*centre, right*). *From top to bottom:* sequential line scan, 'Sampledot' and the Pseudo-Random Scan of fig. 4. Events taking place in a quarter of the frame period would be easily missed with sequential line scanning. Sampledot and the Pseudo-Random Scans, however, pick up such an event, though at reduced resolution. The $\frac{1}{16}$th frame period would only pick up events by chance, except when using the Pseudo-Random Scan. When this scan is used moving objects tend to become elongated in their direction of motion, as their image is being continuously updated.



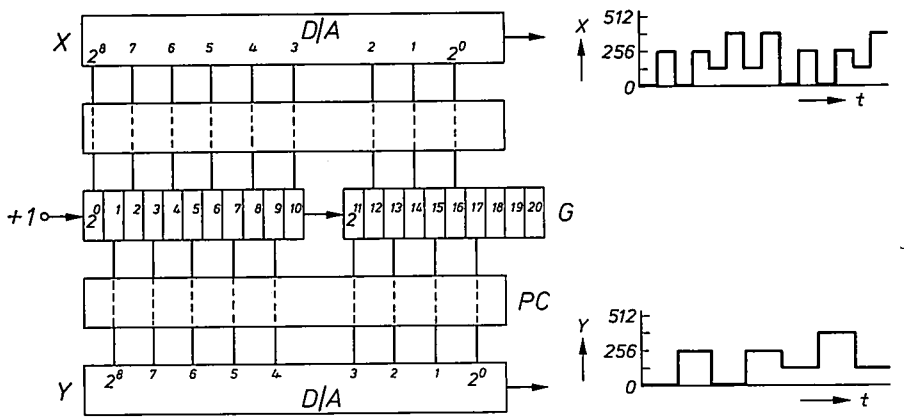


Fig. 6. A circuit that generates the scan shown in fig. 4. The scan generator *G* operates in this case as an 18-bit counter, connected by patch cards *PC* to the digital inputs of the deflection systems. The low-order bits of the counter are connected to the high-order bits of the deflectors, resulting in a large scan jump for each count. *Right:* The coordinates *X* and *Y* of the pixel at which the scan has arrived, as a function of time.

The first of these is known as 'Sampledot' [2]. It can be applied to a standard television system, using a standard television camera and a standard display, as the multiple interlaced format at the display is not achieved by unconventional scanning but by manipulating the video signal between camera and display. *Fig. 3* demonstrates the method for achieving a reduction in bandwidth by a factor of 4. The scene is scanned in the normal manner, but only every fourth pixel along each line is sampled and transmitted (fig. 3, *above*). By correctly off-setting the time at which sampling commences in each television field, each pixel in the scene is sampled once in eight fields per frame, instead of the usual twice per frame. The sampling pattern can be given a degree of randomness by changing the sampling offset in every line (fig. 3, *below*). We have simulated Sampledot on the Dot Scan CCTV by programming the scan generator to produce, at both the camera and the display, a scanning format that reproduces the Sampledot format. Subjectively acceptable pictures have been obtained at scanning rates down to about three frames per second, with scenes that included moving objects.

Better pictures at the same scanning rate and acceptable pictures down to half a frame per second may be obtained by means of a 'Pseudo-Random Scan' [3], as illustrated in *fig. 4*, which introduces a still larger degree of randomness. Its realization is described below (see fig. 6).

Although of course it is not possible to demonstrate on paper the subjective effect of reducing the scanning rates with these techniques, *fig. 5* does give some idea of the effect experienced. A scene on the display was photographed with exposure times of 1, $\frac{1}{4}$ and $\frac{1}{16}$ of a frame period. Thus, in the last two cases only a proportion of the pixels in each frame are scanned. For sequential line scanning this results in 'slicing' the picture, and for the Pseudo-Random Scan it has the effect of reducing overall detail. Sampledot is an intermediate case.

In contrast to Sampledot, the Pseudo-Random Scan of fig. 4 is not so easily achieved. Using conventional cameras and displays, for example, at least two random-access memories would be needed, one at the transmitter and one at the receiver. Against this, however, the system does offer some specific advantages. For instance, it is particularly suitable for keeping a transmission path confidential by 'scrambling' the pixels. Furthermore, this system reduces the effects of coherent interference picked up in the transmission path; the dark banding which this causes when conventional scanning is used is pseudo-randomly distributed over the displayed picture.

The scan generator

Fig. 6 shows how the scan in fig. 4 is achieved in the 'Dot Scan CCTV'. The generator has been programmed to function as an 18-bit counter, adding one count per pixel period. After 2^{18} ($= 512 \times 512$) counts it starts anew at zero. The 18 parallel data bits from the counter are directly connected to the digital inputs of the *X* and *Y* deflection systems controlling the electron beams in both the camera and the display tube: the 'even' bits to the *X* input, the 'odd' bits to the *Y* input, both in reverse order of significance. Thus a change in the first bit of the counter, which occurs every pixel period, produces a half full-scale deflection in *X*, and a change in the second bit, occurring every two periods, produces a half full-scale deflection in *Y*. These are in turn modified by smaller deflections generated by the higher-order bits. The resulting scan pattern has 'maximum interlace'; the electron beam successively jumps from one pixel to a more distant one and only returns much later to a neighbouring pixel.

This is a very simple program; it involves only criss-cross connections between a binary counter and the digital inputs of the deflection systems. Other programs consist of sets of instructions to the generator. Sequential line scanning, for example, has the following set of instructions: 'add 1 to *X* every clock period

[2] R. F. Stone, IEEE Trans. **BC-22**, 21, 1976.
[3] S. Deutsch, IEEE Trans. **COM-21**, 65, 1973.

until X equals 512; then reset X to zero and add 1 to Y; repeat until Y equals 512; then reset Y to zero; repeat'. The scan, or other activities such as variations in scan density and reversed scan, may be limited to certain areas of the field by special commands called 'scan-limit flags'.

The camera, the display and associated electronics

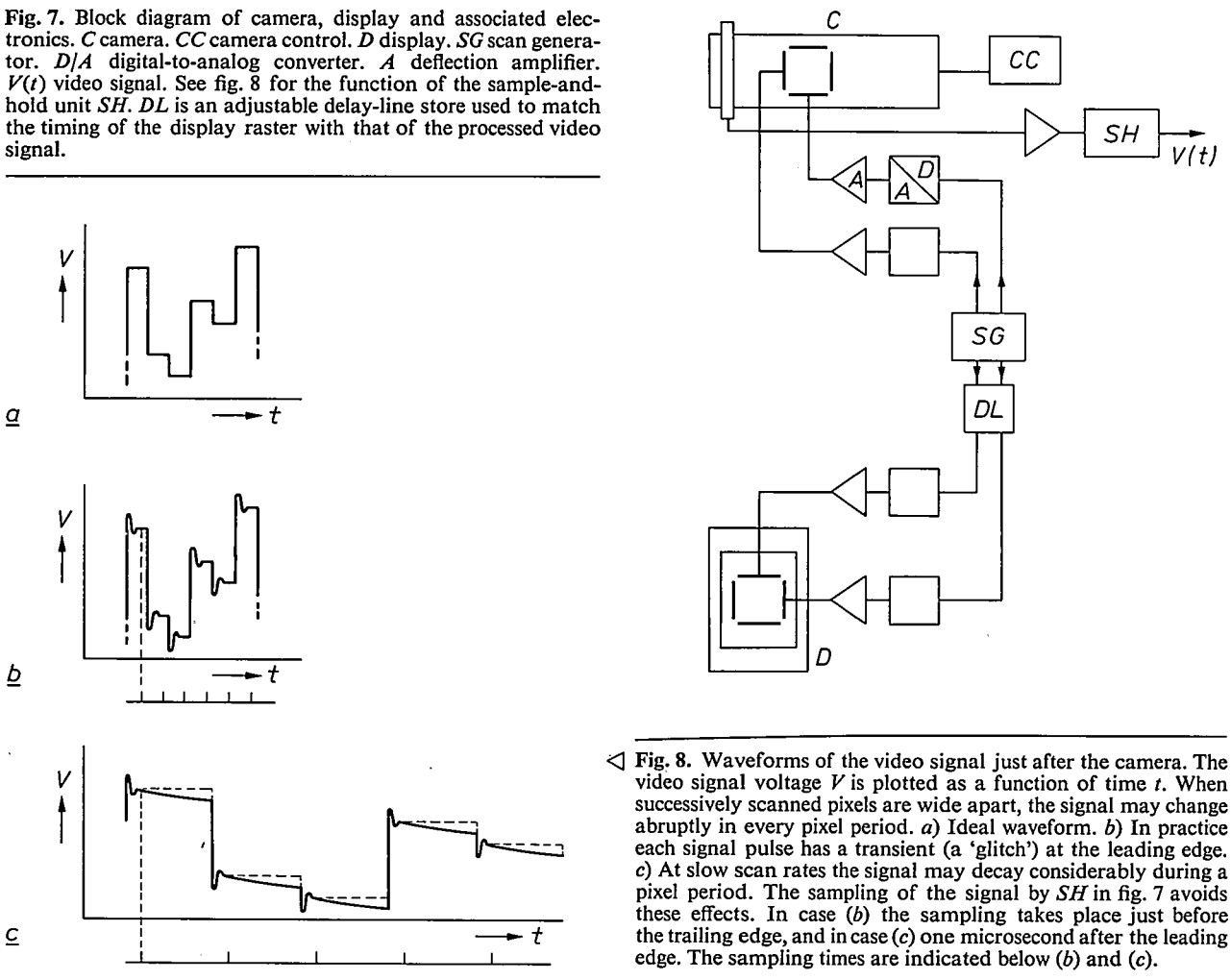
Deflections of the electron beam such as those of fig. 4, produced at normal scanning rates, imply very fast signal variations (high slew rates) and these impose high requirements on the camera, the display and the associated electronics (*fig. 7*). To begin with, the vidicon tube used in the camera must possess an electrostatic beam-deflection system. In our camera we use an RCA 4514. The digital signals from the scan generator are converted into analog signals of about 1 V peak-to-peak (D/A in fig. 7) and applied to deflection amplifiers (A) of high quality and very wide bandwidth (from 0 to 34 MHz; output voltage 40 V peak-to-peak). With this combination of camera and amplifiers it is possible to deal with random scans covering the full field of view, as illustrated in fig. 4, at rates up to about 12 frames per second. The more common sequential line scans can be dealt with up to 25 frames per second. If necessary very much higher frame rates over limited areas may be produced.

Both random scans and very slow scans require special precautions regarding the video signal. If the vidicon beam jumps from light to dark positions in the field of view, the 'ideal' waveform has vertical edges (*fig. 8a*). In practice, after each transition a transient (a 'glitch') is superimposed upon the video waveform owing to capacitive coupling between the deflection system and the target electrode (fig. 8*b*). This is aggravated because the impedance of the internal connection to the mesh electrode of the vidicon becomes relatively high at the frequencies involved. Furthermore, slow scans imply long pixel periods; during these periods the decay of the signal may become appreciable (fig. 8*c*).

To overcome these effects a sample-and-hold unit is incorporated (*SH* in fig. 7). In the first instance the video signal is sampled towards the end of the pixel period, and in the second instance it is sampled a microsecond after the leading edge.

▷

Fig. 7. Block diagram of camera, display and associated electronics. *C* camera. *CC* camera control. *D* display. *SG* scan generator. *D/A* digital-to-analog converter. *A* deflection amplifier. *V(t)* video signal. See fig. 8 for the function of the sample-and-hold unit *SH*. *DL* is an adjustable delay-line store used to match the timing of the display raster with that of the processed video signal.



◁ **Fig. 8.** Waveforms of the video signal just after the camera. The video signal voltage V is plotted as a function of time t. When successively scanned pixels are wide apart, the signal may change abruptly in every pixel period. *a*) Ideal waveform. *b*) In practice each signal pulse has a transient (a 'glitch') at the leading edge. *c*) At slow scan rates the signal may decay considerably during a pixel period. The sampling of the signal by *SH* in fig. 7 avoids these effects. In case (*b*) the sampling takes place just before the trailing edge, and in case (*c*) one microsecond after the leading edge. The sampling times are indicated below (*b*) and (*c*).

Scan information reaching the *display* is similarly processed by digital-to-analog converters and wideband deflection amplifiers. In addition, the scan information passes through adjustable delay lines (*DL*) that compensate for the delay of the video signal en route from the camera to the display.

The image is normally presented on two displays; see fig. 2. The first, an *XYZ* oscilloscope, has electrostatic deflection and is particularly suitable for random scans; the bandwidth in X, Y and Z is 20 MHz. Its resolution, however, is rather limited (400 lines) and the image contrast is rather poor. We therefore use this display mainly for demonstrating novel scan formats. The second, a display of our own design, is based on a magnetically deflected triode cathode-ray tube with a resolution as high as 800 lines. This display, which also gives higher contrast, is better suited for use in experiments concerned with resolution and grey-level manipulation. The large signal bandwidth of the magnetic deflection system (0-600 kHz) allows us to use it in experiments involving modestly randomized scans, such as the Sampledot format, though only at low frame rates, as well as in experiments with normal sequential line scanning over a wide range of aspect ratios and frame rates.

Two-dimensional edge enhancement

Our second example, two-dimensional edge enhancement [4], will serve to illustrate the purpose and organization of the line store shown in fig. 1. For one dimension its principle is shown in *fig. 9*, where it is applied to a slow transition between two grey levels (fig. 9*a*, solid curve A). A purposely blurred reproduction of the original (the 'mask', upper dashed curve $\bar{A}$ in fig. 9*a*) is partly or wholly subtracted from the original and the result (fig. 9*b*, *c* and *d*) is amplified until the overall contrast is restored (fig. 9*e*, *f*, *g*). When the mask is fully subtracted (fig. 9*d*, *g*), all intermediate grey levels are removed from the 'background' (left and right in the diagrams of fig. 9) and the original edge is represented by a black-and-white outline. The processing operator functions as a spatial filter; the lower spatial frequencies are filtered out and the higher spatial frequencies, in effect contrasts over small areas, are reproduced and enhanced, as they are either not present or are of lower amplitude in the mask. A similar principle, known as 'unsharp masking', in which a picture is combined with a slightly defocused mask has been used by photographers for more than forty years [5].

Fig. 10 shows some results obtained with the Dot Scan CCTV. The picture is blurred by taking the average of the real-time values of the video signal $a_1, a_2, \ldots a_9$ in a 3×3 matrix of pixels around the pixel to be processed (see *fig. 11*); a_5 is the signal value that would be displayed in the unprocessed picture. The value a_5' of the processed pixel should be:

$$a_5' = g(a_5 - \alpha\bar{a}),$$

where $\bar{a}$ is the average of $a_1, a_2, \ldots a_9$. Thus a real-time matrix of video signal values $a_1, a_2, \ldots a_9$ is required for each display pixel, and the purpose of the store in fig. 1 is to provide this data in real time. Once such a matrix is available, the implementation of the processing involves straightforward analog electronic techniques, as illustrated by the diagram in *fig. 12*.

The store

Fig. 13 shows the organization of the store of the Dot Scan CCTV. It can provide, at pixel rate, the real-time signal values belonging to a local area of 8×8

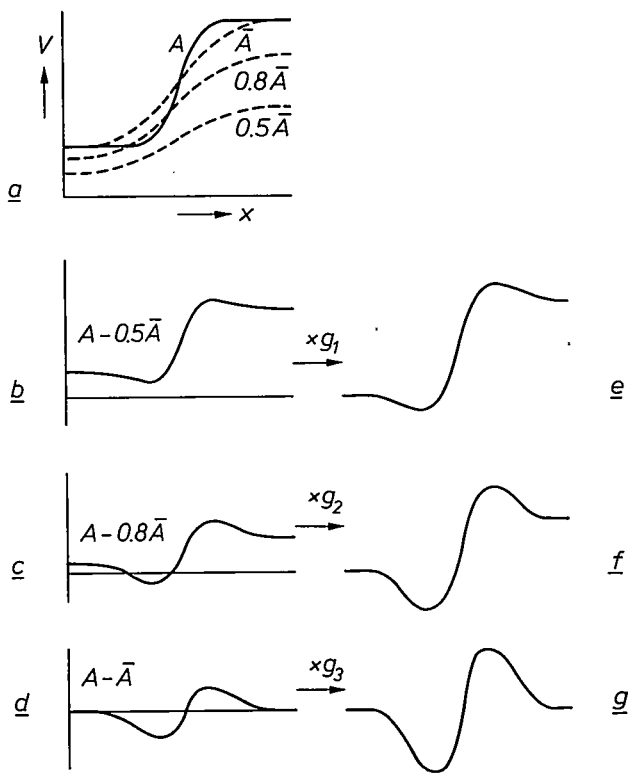


Fig. 9. Principle of edge enhancement for one dimension (x). *a*) A is a transition in space between two signal values V (grey levels). Local averaging of the signal produces the unsharp 'mask' $\bar{A}$. *b,c,d*) When $\bar{A}$ is partly or wholly subtracted from A, local variations in A are enhanced. *e,f,g*) The signal amplitude is increased until the maximum contrast (the peak-to-peak value of V) is restored to its original value. The result (*e*) is 'near ideal', (*f*) is over-emphasized. In the extreme case (*g*) differences between 'flat' areas in the scene (left and right in (*a*) are completely suppressed, leaving only the edges as black-and-white lines.

[4] A. Rosenfeld and A. C. Kak, Digital picture processing, chapter 6.4; Academic Press, New York 1976.
J. O. Drewery, Proc. IEE 123, 299, 1976.

[5] W. F. Schreiber, Pattern Recog. 2, 117, 1970.
L. Levi and M. Mossel, in: Applications of holography and optical data processing, Proc. Int. Conf., Jerusalem 1976, p. 73.

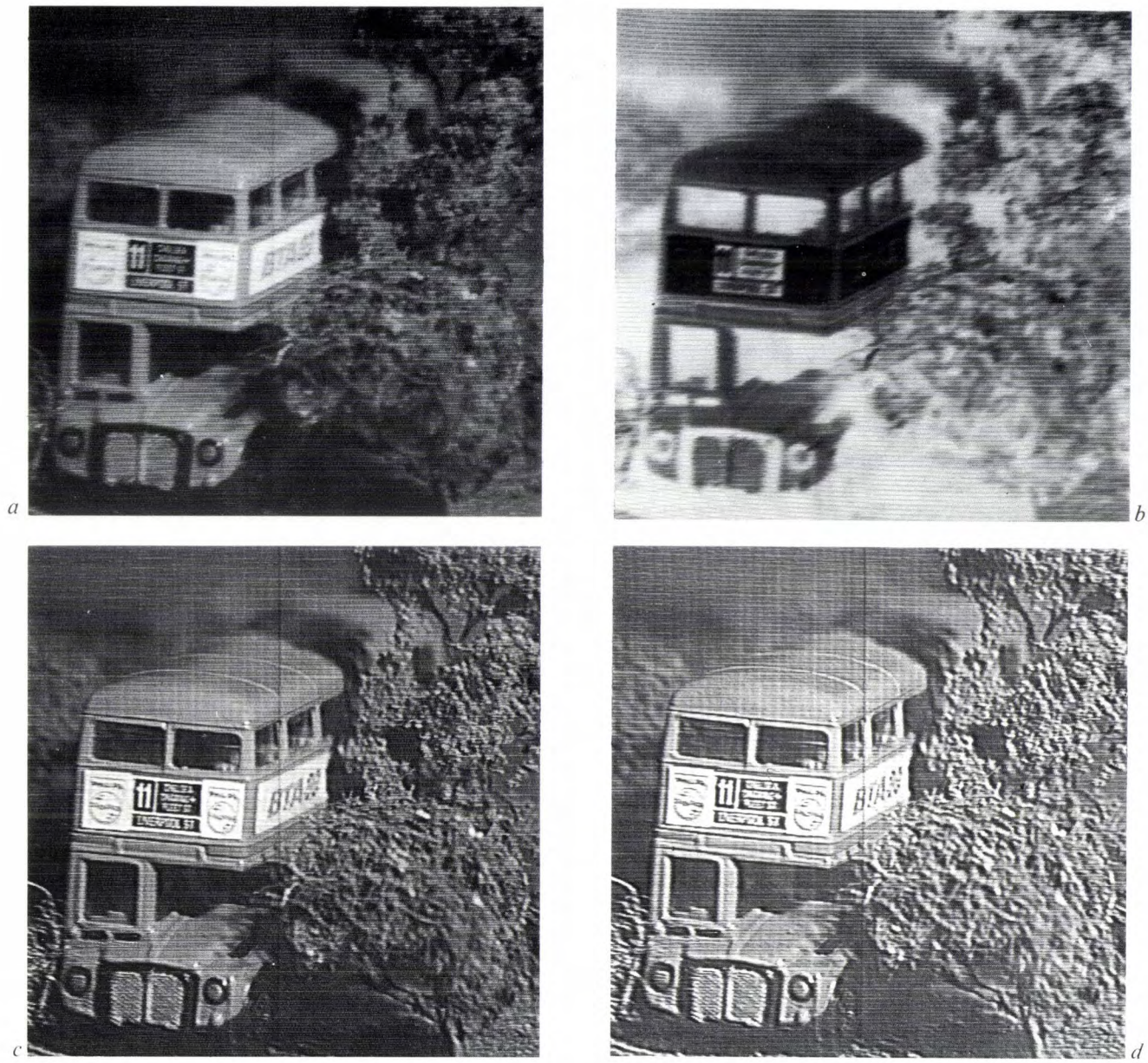


Fig. 10. Two-dimensional edge enhancement with the Dot Scan CCTV. *a*) Original, corresponding to A in fig. 9*a*. *b*) The mask; $V = -\bar{A}$. *c*) Near-ideal edge enhancement; $V = 2(A - 0.5\,\bar{A})$. *d*) Over-compensation; $V = 4.35\,(A - 0.77\,\bar{A})$.

pixels. It consists of a 'main store' (MS) of 8×512 shift-register cells and a 'scratch pad' (SP) of eight rows of eight cells, where each cell can contain an eight-bit pixel word. At a particular moment during a frame the scan will arrive at point p in fig. 13 (*below*). The main store then contains the eight-bit words P, ... Q representing the video signal of all pixels between p and q, counted along the lines backwards from p. On reaching the end of a main-store register, the words are fed into a row of the scratch pad, as well as into the next register. At the instant depicted in fig. 13 the scratch pad thus contains the 64 words A ... B, C ... D,

a_1	a_2	a_3
a_4	a_5	a_6
a_7	a_8	a_9

Fig. 11. Matrix of 3×3 pixels around the pixel (*grey area*) where the display scan has arrived; a_1, a_2, ... a_9 are the real-time signal values of this matrix. Their average $\bar{a}$ yields (except for the sign) the instantaneous value of the mask ((*b*) in fig. 10); a_5 is the value displayed in the unprocessed picture ((*a*) in fig. 10).

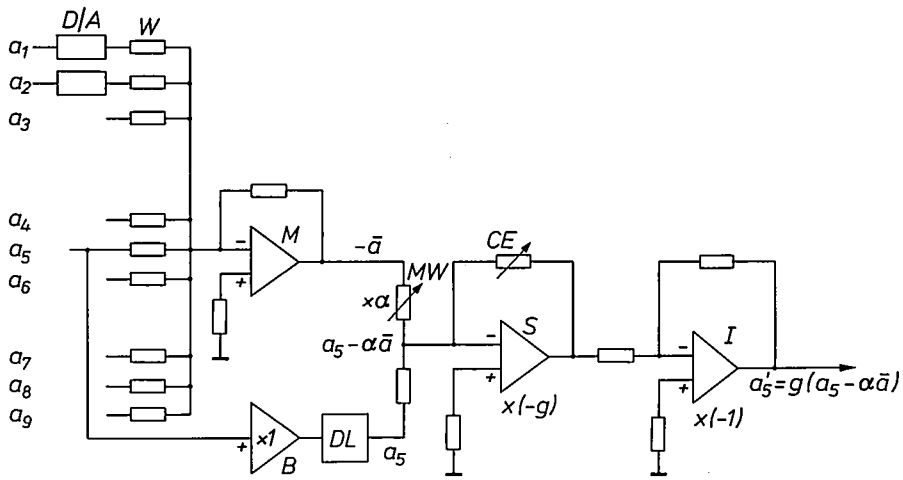
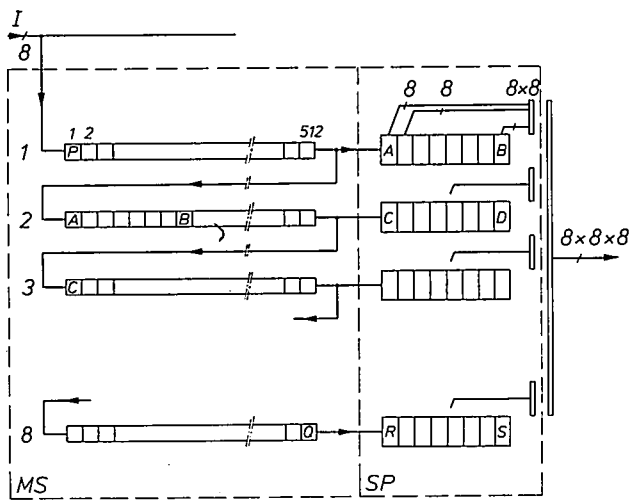


Fig. 12. Processing operator that performs the edge-enhancement operation $a_5' = g(a_5 - \alpha\bar{a})$. The digital signals $a_1, \ldots a_9$ are converted into analog signals (D/A), weighted (W) and combined in the masked amplifier M. After mask weighting (MW) the mask signal $-\alpha\bar{a}$ is obtained. At this point the signal value a_5 of the centre pixel is added, after it has passed the buffer amplifier B and the equalizing delay line DL. The signal is normalized to 1 V black-to-white by the contrast equalizer CE of the amplifier S, and inverted by the inversion buffer I, finally resulting in the required output signal.



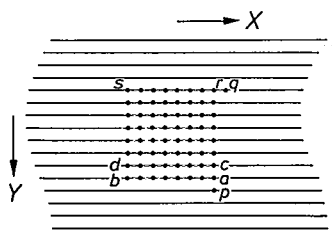


Fig. 13. *Above:* the line store of the Dot Scan CCTV consisting of the main store MS (8 lines of 512 cells) and the scratch pad SP (8 rows of 8 cells). *Below:* scan lines and pixels. When the scan arrives at p, the video word P (8 bits) belonging to p enters the first (8-bit) cell of MS. At this instant MS contains the words $P, \ldots Q$ that belong to the 8×512 pixels between p and q. At the end of each line of 512 cells the words are transferred to the next line and to a row of the scratch pad. The latter contains the values $A \ldots B$, $C \ldots D$, ..., $R \ldots S$ that belong to the pixel matrix $a \ldots b$, $c \ldots d$, ..., $r \ldots s$. This 'real-time matrix' of 64 pixels traverses the whole picture at the scan rate. All cells in SP may be tapped, so that $8\times8\times8$ bits are instantly available for processing.

..... , $R \ldots S$, belonging to the matrix of pixels $a \ldots b$, $c \ldots d$, ... , $r \ldots s$ in fig. 13 (*below*). This array of 8×8 pixels traverses the whole picture at the scan rate. The cells of the scratch pad are tapped and the words contained are available for processing. The processor making use of the scratch pad has, of course, also to be capable of operating at pixel rate.

Contrast enhancement

The final example, contrast enhancement or, more generally, grey-level manipulation [6], is a more complex operator. As we are committed to real-time processing with a modest amount of hardware, the processing must in this case be restricted to a fraction of the total number of pixels. *Fig. 14b* shows the result of the operator applied to a 'keyhole' in the frame that contains the image of a car in shadow (fig. 14*a*). The basic principle of the operator will be discussed first, and then the manner in which it has been realized in our system.

Very little detail of the car can be distinguished in the shadowed area in fig. 14*a*. This is not because of lack of information in the video signal — as is evident from fig. 14*b*, which originates from an identical signal — but to a loss of information between the signal and the combination of display and viewer. In fact, there is always some loss of information ('data reduction') between signal and display. If the video signal of a high-quality television system is digitized, usually 256 levels (eight bits) are required to preserve all the information; in infrared imagers a dynamic range of more than 1000 discernable levels may be encountered in the video signal. Such information is necessarily reduced when displayed, because under ideal conditions cathode-ray-tube displays barely resolve 64 grey levels to the eye and more generally less than 32. An exag-

[6] A. Rosenfeld and A. C. Kak, Digital picture processing, chapter 6.2; Academic Press, New York 1976.
R. Hummel, Computer Graph. Image Process. 6, 184, 1977.
R. P. Kruger, E. L. Hall, S. J. Dwyer and G. S. Lodwick, Int. J. bio-med. Computing 2, 215, 1971.

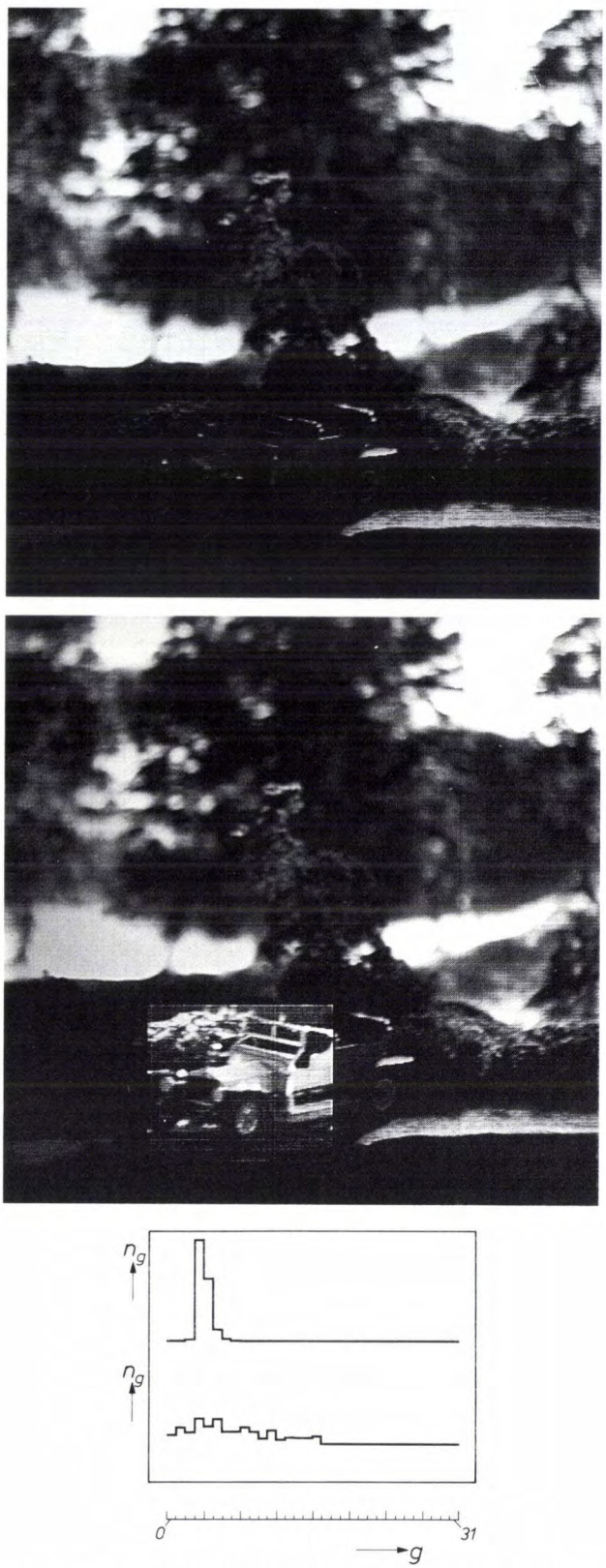


Fig. 14. Contrast enhancement. *a*) Picture on the display screen of a scene showing a car in shadow; unprocessed video signal. *b*) Contrast enhancement applied to the 'keyhole' containing the car image. The video signals originating from the camera are identical in (*a*) and (*b*). *c*) Histogram of grey levels in the keyhole in (*a*) and (*b*). In (*a*) only four grey levels are used, in (*b*) the information in the keyhole is divided over a much larger number of levels. How this is done is explained in more detail in figs 16 and 17.

gerated example of data reduction is shown in *fig. 15*.

The contrast on the display may be enhanced (i.e. expanded or compressed) by a redistribution of the input signal levels over the available grey levels in the display. The principle is illustrated in *fig. 16* for a simplified case of 16 possible input levels and four possible display levels. Fig. 16*a* shows a rather grey original scene (*below*) and its histogram of input levels (*above*); only levels midway between black (*0*) and white (*15*) are present. For a 'linear coding' of the input levels to the displayed levels (fig. 16*b*), only the levels *1* and *2* of the four possible grey levels *0*, *1*, *2*, *3* are present on the display. The contrast, however, can be expanded so that black (*0*) and white (*3*) are also used. This entails a 'recoding' of the information. A well-known approach is to recode in such a way as to approximately equalize the histogram of the display levels (fig. 16*c*).

Fig. 17 shows a block diagram of the processing operator that we have used for histogram equalization [7]. For simplicity it is assumed in fig. 17 that there are 16 inputs and four display levels in a keyhole of 6 × 6 pixels, as in fig. 16. The number of data bits indicated along the interconnections refer to this case; the numbers between brackets refer to the actual circuit (256 input levels, 32 display levels and a keyhole of 64 × 64 pixels).

The processor is timed by commands F from the scan generator. These include the system clock to maintain synchronization, and the 'keyhole flags' which 'flag in' the keyhole period T_k (see fig. 17*b*). By their logical combination the keyhole flags indicate when the scan is *not* in the area T_q or T_k, and is therefore in T_n, hence permitting the processor to perform its recoding operation. This normally takes a period (T_c) of 1.5 line-scan periods. During the non-keyhole period ($T_n + T_q$) the multiplexing switches M_1 and M_2 are in the dashed line positions and the video signal is fed unprocessed to the display. Within the keyhole period (T_k) the multiplexers are in the 'solid-line' positions. Each word of the digitized video signal then addresses a random-access memory (the '*Code RAM*' in fig. 17). If this RAM is encoded as shown and the video word at a particular moment has the value 11, for example, then the grey-level value 2 is presented on the display. This latter value is also used to address another store (the '*Count RAM*' in fig. 17) and the data, a count of 12, stored at that address is incremented by 1 by the adder S. The count RAM thus counts the number of times each of the display levels has been presented during each keyhole period. In this way a histogram count of the displayed grey levels in the keyhole is

[7] P. K. Bailey, private communication.

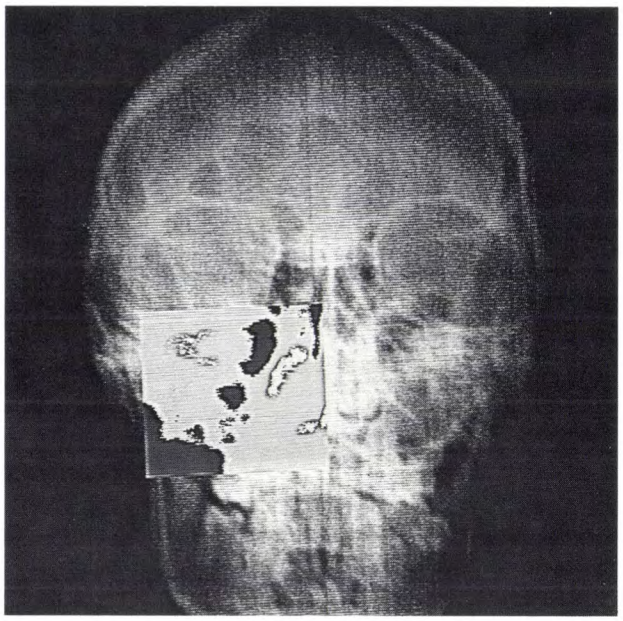


Fig. 15. Data reduction in a keyhole of 128×128 pixels to only four grey levels. Compare the information in the keyhole with the unprocessed information in the corresponding area of the other cheek.

recoded at the end of a keyhole period. The histogram is displayed on a cathode-ray tube (D_2 in fig. 17) for photographic recording and analysis. The count RAM also provides the information required for recoding the contents of the code RAM, which proceeds as follows.

During the recoding period T_c (approximately 30 μs or 1.5 times the line-scan period) the multiplexers M_1, M_2 and M_3 are in their dashed line positions. The recode module *RM* reads the data in the count RAM and compares it with preset values for an ideal histogram count, in this instance an equalized count. Consider, for example, two successive display-level counts in the histogram, where the first is too small and the second too large. Under these conditions, the module *RM* rewrites the data in the code RAM so as to code more input levels into the first display level and fewer into the second. To increase the slew rate of the processor, large errors are translated into large changes in code and small errors into small changes. Hunting between successive field scans is prevented by thresh-

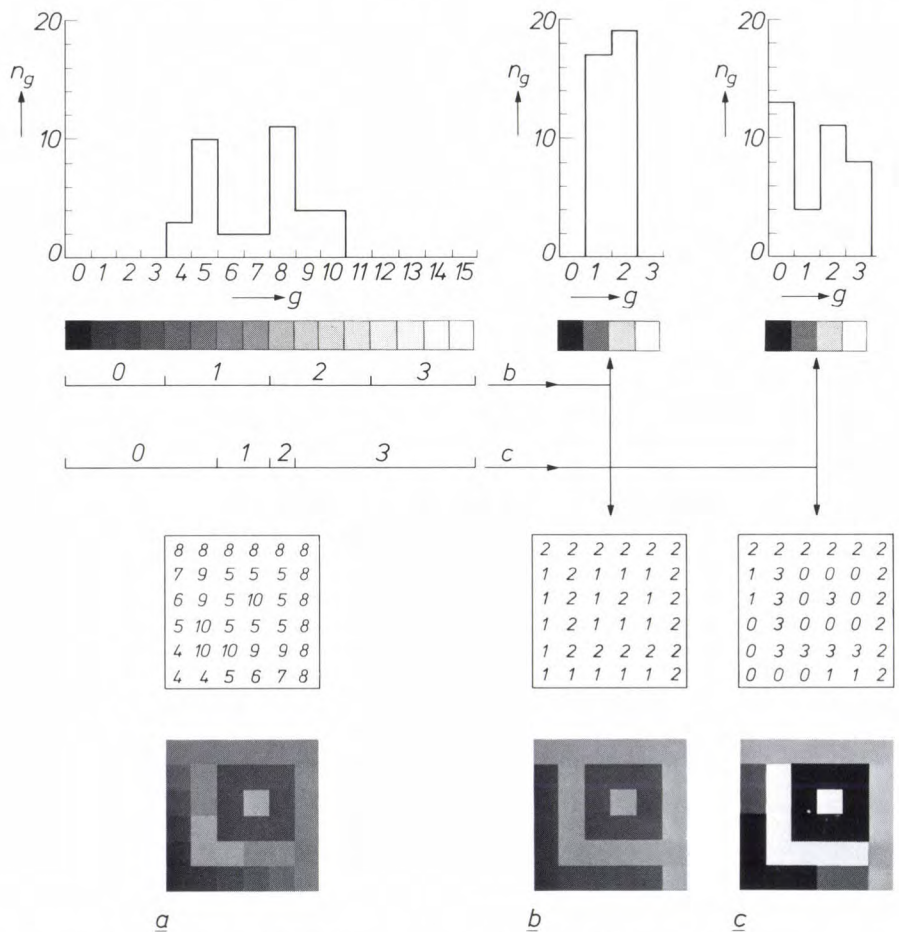


Fig. 16. Contrast enhancement, shown schematically for a situation with 16 input levels, four display levels and a matrix of 6×6 pixels. *a) Below:* the original picture using 16 input levels. *Above:* histogram of grey levels of this picture. In the matrix above the picture the number of the grey level used in each pixel is given in terms of the scale under the histogram; counting the number of fours, fives, ... in this matrix results in the histogram. *b)* Display with linear coding of the 16 input levels to the four display levels; only levels *1* and *2* have been used. *c)* Result of recoding the input levels to the four display levels, so that the histogram is approximately equalized; levels *0* and *3* are then also present in the display.

olding the count error, i.e. no longer correcting for errors below a particular threshold. The processor thus functions in an iterative manner and usually takes some ten field scans to settle when switched from linear coding to histogram-equalized coding. Thereafter the video signal is normally tracked within a delay of one to two field scans, even with moving objects present. Recoding begins from black (*0*) and proceeds to peak white (*31*); in operation one can see the histogram being modified from left to right during successive field scans.

Fig. 18 illustrates the subjective effects that may be achieved under rigidly controlled conditions. The background has been produced electronically by replacing the digital video signal by data from the X output of the scan generator. This produces a pattern of 256 grey levels that are spatially locked to the scan and whose sequence may be altered by changing the cross-connections. This makes it possible to carry out a variety of experiments concerning keyhole size, number of resolvable grey levels and spatial resolution.

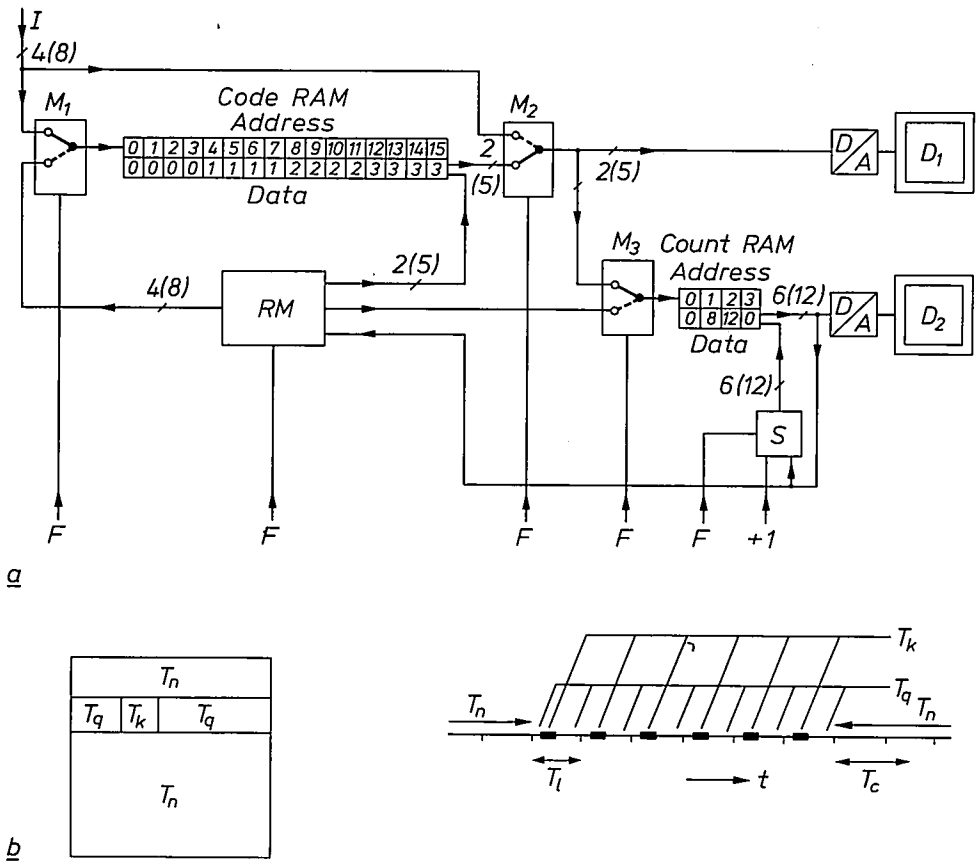


Fig. 17. *a)* Block diagram of the processing operator for histogram equalization. *Code RAM*, *Count RAM* random access memories for encoding the video signal and providing the histogram of display levels; *RM* recoding module; *I* digital video signal input; M_1, M_2, M_3 multiplexers; *D/A* digital-to-analog converters; D_1 display, D_2 oscilloscope. *F* 'flag' commands from the scan generator. *b)* The various periods involved in this operation: T_l line period; T_k keyhole period; T_n, T_q non-keyhole periods; T_c period equal to about 1.5 line-scan periods in which the recoding operation takes place. During T_k (multiplexers in the positions drawn) the digitized video signal addresses the code RAM from which the recoded data passes to the display, and the count RAM builds up the histogram of displayed grey levels by incrementing, by means of the adder *S*, the count data of a level by 1 each time it is presented. The histogram is displayed on D_2. During $T_n + T_q$ (multiplexers in the dashed line positions) the video signal is fed direct (unprocessed) to D_1 and recoding takes place, the histogram data in the Count RAM being used by *RM* to recode the Code RAM. This operation only takes up the period T_c. The numbers at the cross-connections indicate the numbers of data bits for the schematic case of fig. 16; the numbers between brackets refer to the actual circuit.

The keyhole

The size and position of the keyhole are determined by the scan generator. The keyhole is always rectangular, but its size, aspect ratio and position can be adapted to the area of interest. Its size is limited only by the number of pixels that can be handled by the operator in use. When used with the store in fig. 13, for instance, its size is limited to 64×64 ($= 4096 = 2^{12}$) pixels.

When used with the grey-level manipulation operator discussed above, it is the capacity of the count RAM and the speed and capacity of the recode module that restrict the keyhole area. The present upper limit is 128×128 ($= 2^{14}$) pixels.

The keyhole performs two functions. Firstly, it defines an area to which processing can be limited, should the processing operator be so complicated that real-time processing of the whole scene ($512 \times 512 = 2^{18}$ pixels) is impractical. Secondly, an operator may be more effective if restricted to the area of interest. For example, in the case of the histogram equalization just discussed, the best result is obtained when a keyhole is used that just encompasses the object of interest, because in that case the recoding is solely determined by

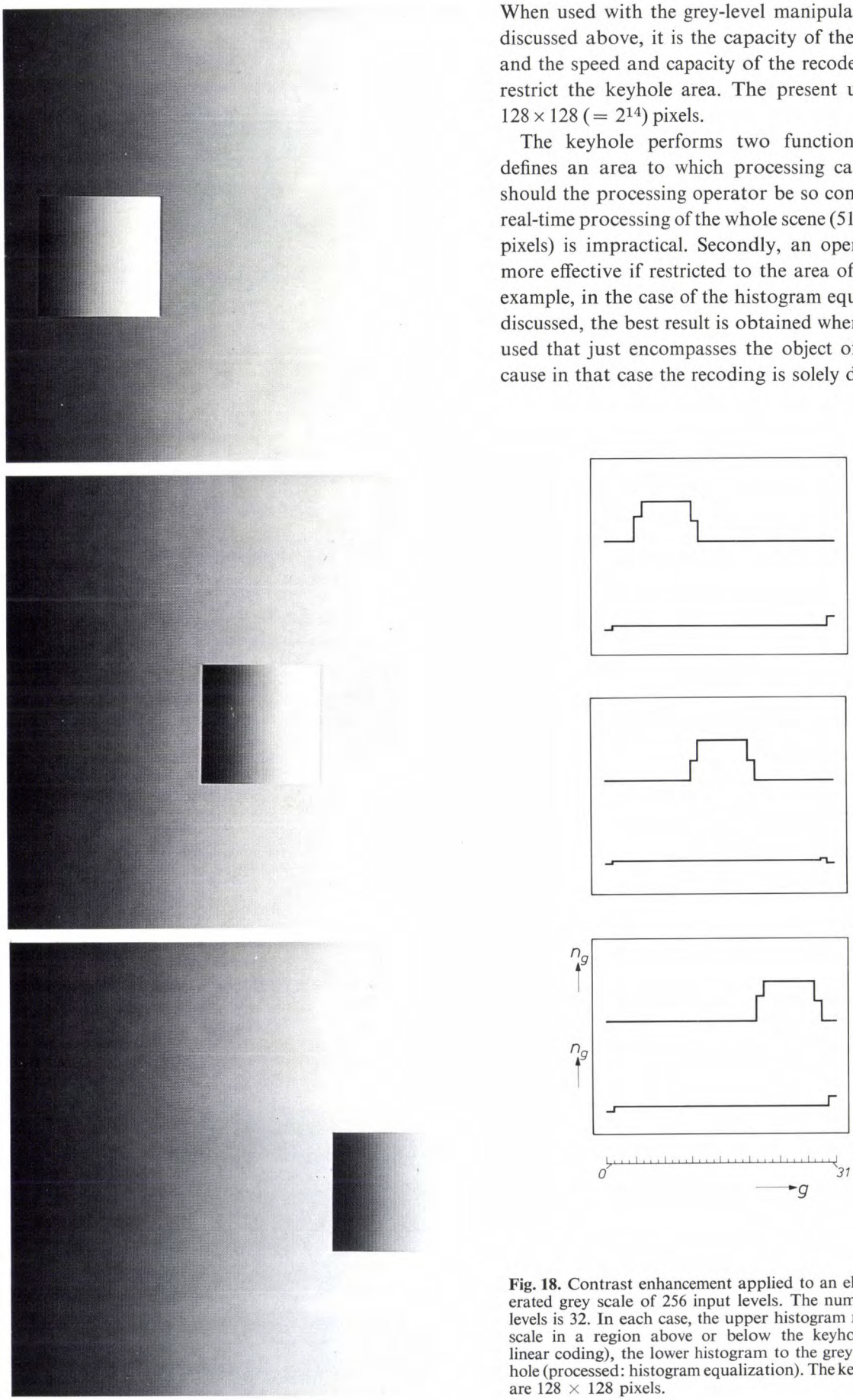


Fig. 18. Contrast enhancement applied to an electronically generated grey scale of 256 input levels. The number of displayed levels is 32. In each case, the upper histogram refers to the grey scale in a region above or below the keyhole (unprocessed: linear coding), the lower histogram to the grey scale *in* the keyhole (processed: histogram equalization). The keyhole dimensions are 128 × 128 pixels.



the grey levels of specific interest. This would still apply even if the count RAM and recoding module could handle 2^{18} pixels. The effect of varying the size of the keyhole is shown in *fig. 19*.

If all parts of the scene are of equal interest, then grey-level manipulation can also be applied to the whole scene using our particular processor, but the local enhancement is then generally less effective. An example is shown in *fig. 20*. In this case the code in the code RAM is produced by the count RAM and the recode module operating from samples taken at every eighth pixel along every fourth line. These pixels are sampled by means of the switch M_3. In this case, too, *all* incoming video information passes through the code RAM. The first four lines at the start of each frame are reserved for recoding.

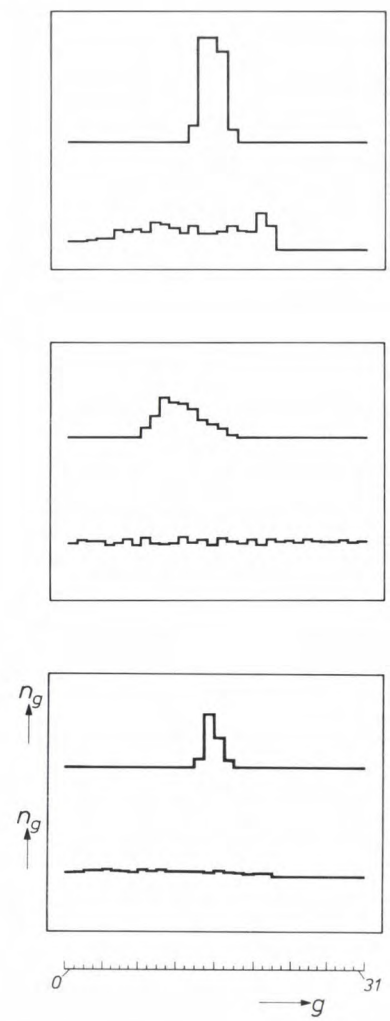


Fig. 19. Illustrating how the size of the keyhole affects the contrast-enhancement operation. The keyhole dimensions are 128×128, 128×64 and 64×64 pixels. The quality of enhancement improves as the processed area approaches that of the object of interest.



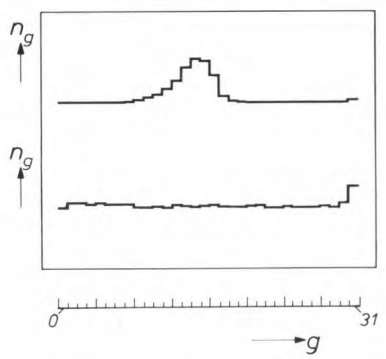


Fig. 20. Contrast enhancement applied to the whole picture. In this case recoding takes place on the basis of each eighth pixel of every fourth line. The upper histogram again refers to the unprocessed picture, the lower one to the processed picture.



Summary. The application of image processing in infrared, X-ray, radar, ultrasound and other imaging systems may often only be useful if it occurs in real time. Preliminary experiments associated with the design of such a system are often difficult to set up because of the very special scanning technique and format of the system in question. The Dot Scan CCTV system is able to simulate many systems in this respect as it has a freely programmable scan format for camera and display. Various processing operators can be incorporated in the system and their effects on the displayed image can thus be studied. Some special features of the system — such as the possibility of generating Pseudo-Random Scans; the requirements to be met by camera, display and associated electronics for coping with such scans; a line store that produces real-time signal values of groups of pixels for processing — are discussed with reference to three scanning and processing techniques: multiple interlaced scans, two-dimensional edge enhancement and grey-level manipulation.

Philips tech. Rev. 38, 326-328, 1978/79, No. 11/12

A universal instrument for digital picture processing

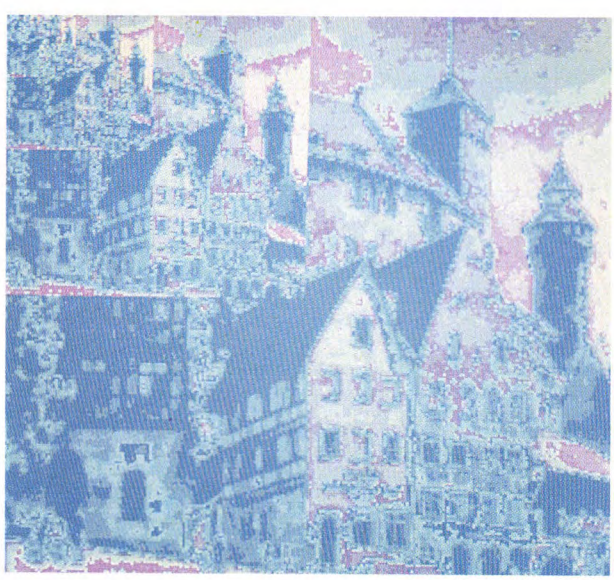


The title photograph shows a computerized instrument developed at Philips Research Laboratories for our own general experiments in pictorial information analysis. Referred to as *P*icture *S*tore and *D*isplay (PSD) it is a universal instrument that stores, processes and displays digitized images taken with a 625-line television camera. An instrument of this type is now being used in three laboratories of the Philips Research Organization, for applications such as checking masks, pattern recognition and studying human vision. The equipment consists of a core memory — in which the grey-level values (on a scale of 0-255) of more than a quarter of a million pixels can be stored — a picture-acquisition unit, an eight-channel television display output (with one colour channel) and a Philips P857 minicomputer as the central processor and control unit. Picture acquisition ('inputting') and display take place at the normal television frame rate. The user can choose a grey scale with fewer than 256 levels for image display, right down to two levels if necessary. The colour-display output is particularly well suited for pseudo-colour manipulations, i.e. making grey-level

Fig. 1. *a*) Digitized urban scene, unprocessed. 256 × 256 pixels. Pseudo-colouring of 16 grey levels. *Top left:* versions reduced in size ($\frac{1}{2}\times$, $\frac{1}{4}\times$, $\frac{1}{8}\times$, $\frac{1}{16}\times$).
b) The same scene after picture-noise reduction. Pseudo-colouring of 256 grey levels. There is some edge blurring as a side-effect.
c) The same scene after *directionally selective* edge detection (differentiation at 45°).
d) As *c*, but with directional selection at 45° and 135°.

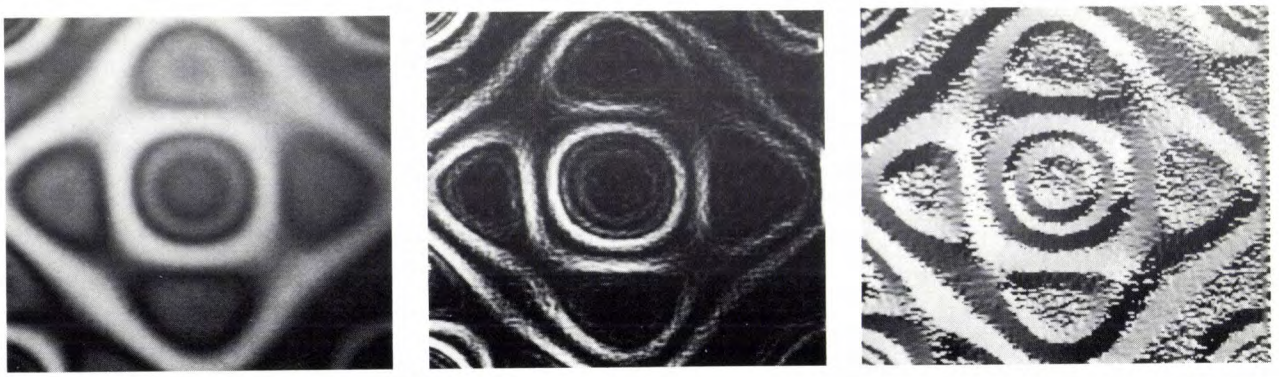
▽

a



b





a *b* *c*

△

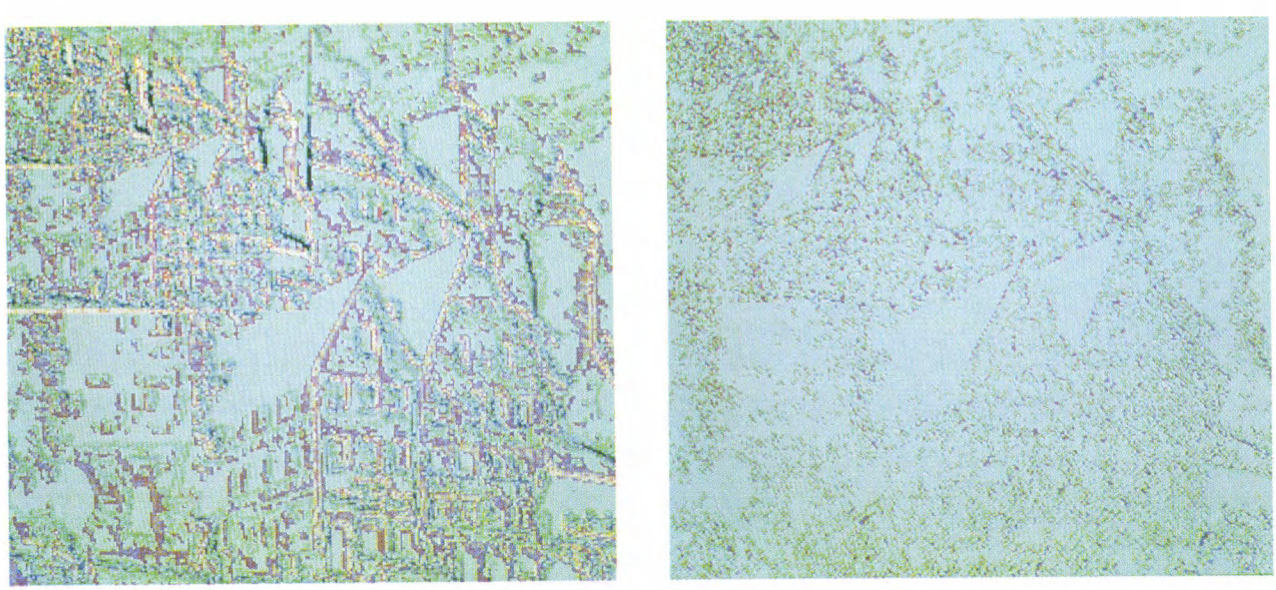
Fig. 2. *a*) Digitized hologram of vibrating metal plate, unprocessed. 256×256 pixels, 256 grey levels.
b) The same hologram after edge detection independent of direction. The brightness is a measure of the *sharpness* of the edges (maximum white).
c) As (*b*). The brightness is now a measure of the local *direction* of the edges. 256 grey levels correspond to a directional range of 360°.

differences more visible by converting them into colour differences. The number of pixels in the output displays is a minimum of 4×4 (highest magnification) and a maximum of 512 × 512. Zonal image processing is possible within a 'window' that can be adjusted in both position and size. With some sacrifice of resolution or the number of grey levels, *series* of pictures, for example an original with three processed versions, can be displayed simultaneously (and stored, of course).

To demonstrate some of the possibilities for image processing and manipulation, noise suppression and edge detection and enhancement, three series of pictures are shown here that were made with the PSD instrument. The software is based on spatial-filtering methods, where the new grey-level values are (weighted) averages over squares of $n \times n$ (n = 3, 5, 7, 9, etc.) pixels. Experience has shown that pseudo-colouring, as used here, generally gives the best presentation of the desired effects.

In the test picture shown on p. 290 for the computer design of a set of IC masks another type of colouring was used, with one colour per mask. Here colouring has the advantage that the

c *d*



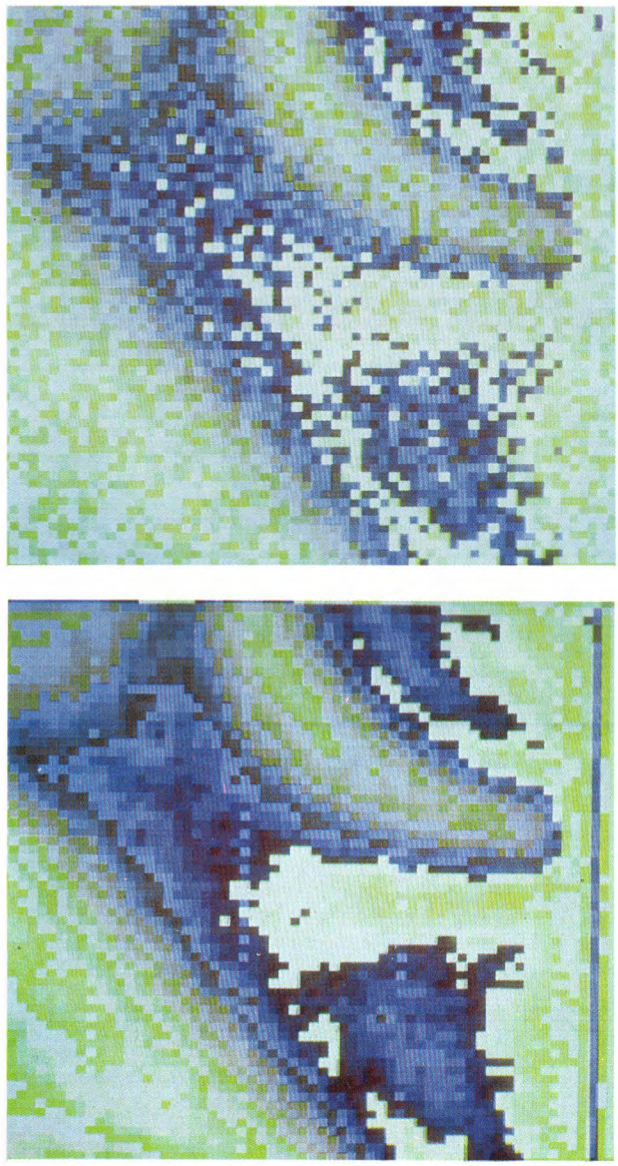


Fig. 3. *Top:* Magnified detail of the digitized hologram (fig. 2*a*, *top right*), unprocessed. 64 × 64 pixels. Pseudo-colouring of 256 grey levels.
Bottom: The same hologram after picture-noise reduction.

visual inspection relates to paths and areas rather than lines. (The conventional test picture, a line drawing, has the disadvantage that there may be more than one interpretation.) Also, the determination of depth is much simpler; with lines it is sometimes difficult to tell which line is 'on top', but with coloured paths it is easy. Overlapping by the upper mask can be made 'transparent'; this also helps to make the estimation of depth easier.

G. J. Scholl

G. J. Scholl is with Philips Research Laboratories, Eindhoven.
An important contribution to the design and building of the PSD instrument was made by Dr C. E. Peterson, formerly with Philips Research Laboratories, Eindhoven. The CIRCUITMASK *program (see Philips tech. Rev.* **37**, 278, 1977), *for designing and checking IC masks, has been extended and modified for use with the PSD equipment by Dr E. H. J. Persoon and Ing. J. H. C. Verbaken, of Philips Research Laboratories, Eindhoven.*

An experiment in flexible automation

P. Saraga and J. A. Weaver

Introduction

The value of the products of the electronic industry lies not so much in the raw materials employed as in the skill and know-how built into them. This industry is not only concerned with the fabrication of the bits and pieces of electronics — both passive, like resistors or capacitors, and active, like semiconductors — but also with the assembly of such parts into subassemblies and finished goods like tape recorders and TV sets.

Much of the detailed assembly of such products is still done by people. Whilst the cost of the basic raw materials, which is in any case usually rather a small proportion of the final product value, is rising, the cost of the labour needed to assemble the products is rising much faster.

In an attempt to keep the end price of its products down in an era when costs are rising, the industry must automate its design and production activities. In any case, it is the simple and repetitive assembly operations that are most easily automated. As a result of automation human labour can be released for more interesting work.

Flexible automation

The classical style of automation, known as 'hard automation', has been used for many years. In this, special-purpose machines are built specifically to produce a particular component or subassembly. Such automation is expensive and therefore implies very large quantity production (mass production) to justify the costs. A characteristic of such hard automation is the necessity to establish and preserve extreme accuracy in the position and dimensional tolerances of the workpieces, and in the control of the machine. This is one reason for the expense. So long as this precision is maintained in both machine and workpieces, hard automation can be very successful. However, components which are broken, out of position or out of tolerance cause severe difficulties.

'Flexible' automation is a new concept, in which the 'intelligence' of computers is allied with the use of sensors — which may be 'visual' or 'tactile' — together with novel methods of mechanical handling.

This flexible automation can do more complex fabrication and assembly tasks than hard automation, and it can easily be changed from task to task. The basic system, which will be quite expensive, can be used for several consecutive tasks, possibly on unrelated products, and this may allow automation to be economically and effectively applied to short production runs. Most of the system components can be used over and over again.

The major part of the control of flexible automation resides in computer programs rather than in the mechanical design of the automation machinery. The characteristics and performance of the system are determined primarily by this software in the controlling computer. Thus, to use such systems an equivalent of the tool setter on an automatic lathe is required, except that the job specification will be wider: he must be capable of 'setting' the software to achieve the required task goal as well as setting the mechanics.

Because software effectively defines the functional capability of the system, it is possible to use the same machinery for operations on a range of different, but similar parts — just by changing the programs or program parameters. In some circumstances it may even be possible to work on different tasks with little or no change-over time simply by switching to other software. In practice this is only possible where no tool changes are required (e.g. the manipulator grippers or special jigs).

In all software writing, modularity and ease of use is important, and in software for this kind of automation it is crucial. In particular it is essential that the software is structured in such a way that changes in function can be easily made at shop-floor level. For this purpose, a user-friendly, high-level language programming system called INDA [1] has been written, which is structured to facilitate easy program development.

P. Saraga, M.A., M. Phil., and J. A. Weaver, M.A., are with Philips Research Laboratories (PRL), Redhill, Surrey, England.

[1] W. T. Park and D. J. Burnett, An interactive incremental compiler for more productive programming of computer-controlled industrial robots and flexible automation systems, 9th Int. Symp. on Industrial robots, Washington D.C. 1979, pp. 281-296.

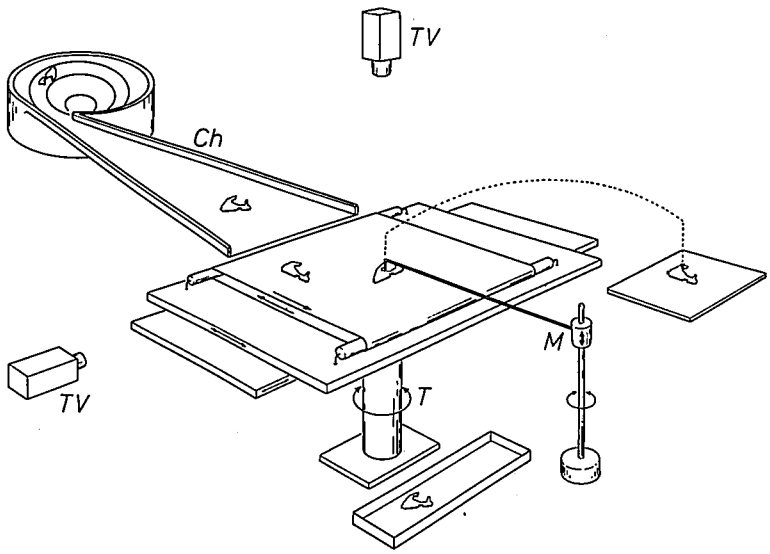


Fig. 2. Experimental visually controlled pick-and-place machine built at PRL. *Ch* chute. *T* table, consisting of two horizontal orthogonal slides pivoted about a central axis. A conveyor belt is mounted on the upper slide of the table. Translation and rotation of the table brings the object into a fixed orientation at the axis of rotation. *M* manipulator which picks up objects that have reached the fixed position described and performs a simple 'goal-post' action to transfer them to a fixed receiving station. *TV* television cameras which provide the computer with the information required for the translations and the rotation.

The set-up time for new tasks should be very much reduced, particularly if one uses the concept of 'teaching by showing', in which a human operator 'shows' the system the task to be performed. Depending upon the intelligence of the system, the human teacher may have to give more or less detail in his training demonstration. For relatively unintelligent systems great detail will be needed with full specification of part and operation to be performed. The exact sequence of operations will have to be defined, together with a specification of paths to be followed by the handling devices, or at the least a number of set-points along the paths. The system may then interpolate between these set-points. Most of the currently available handling systems (robots) can be programmed in this way by 'leading' the machine through the required operation — under human control — and probably at slow speed. The system stores information about the path and sequence of operations, and at the completion of the teaching sequence can play back the operation — often at much greater speed.

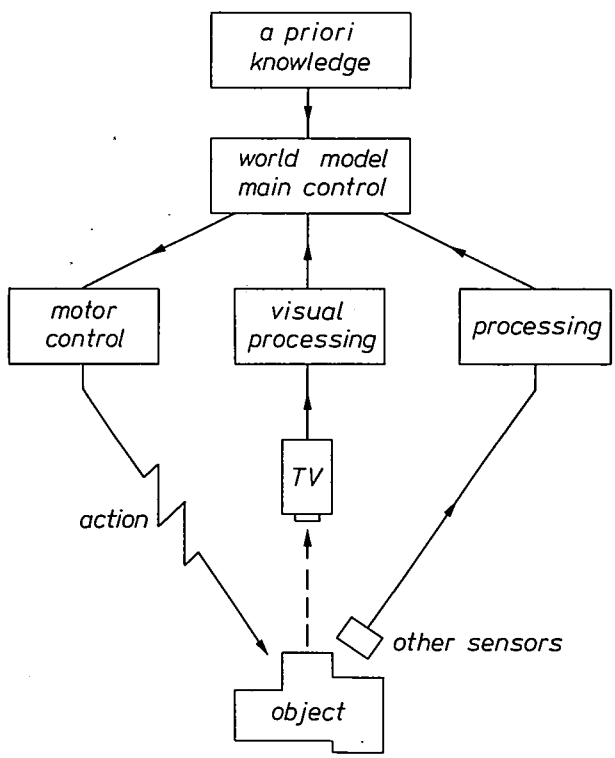


Fig. 1. Conceptual diagrams of a machine controlled by sensors.

It is now becoming feasible to define the requirements of a task to an intelligent machine in much more general terms — the basic function and perhaps the final end points, together with demonstrations of 'good' (i.e. acceptable) and 'bad' piece parts and positions — and to leave the machine itself to determine the best strategy in detail for the whole operation. The action of the machine may be modified by details of circumstance, observed via sensors, so that the course of each operation can be varied while the machine is on its way to the final objective.

Requirements for a sensorily controlled system

A system that is to do a task in a way that is modified by the circumstances in the real environment in which the task is actually done must be able to do four things:

— It must observe the environment.
— It must process the sensor data to extract from it only significant information about the task in hand.
— It must form a plan of action by comparing the requirements of the task with the situation as observed.
— It must issue the detailed command signals to the effectors — the actual mechanical manipulators that will carry out the commands.

Any machine controlled by sensors can be described in terms of the general structure shown in *fig. 1*. The machine obtains images of the real world via its sensors. The main control interprets these images (after processing and selection from the stream of data) in terms of its own conception of this real world stored in a 'world model', which is constructed from sensed data and *a priori* information that defines both the task and the environment. It then instructs the motor control to perform the desired action.

As well as using sensed information to determine a course of action, the sensor-controlled machine can use its sensors to monitor the progress of the action itself. The machine can therefore compensate for errors in the motor control or the mechanisms by observing the results of the action and giving correction instructions. This suggests that machines equipped with visual feedback could attain a greater precision than the accuracy of the machine parts would otherwise allow, as well as allowing the machine itself to correct

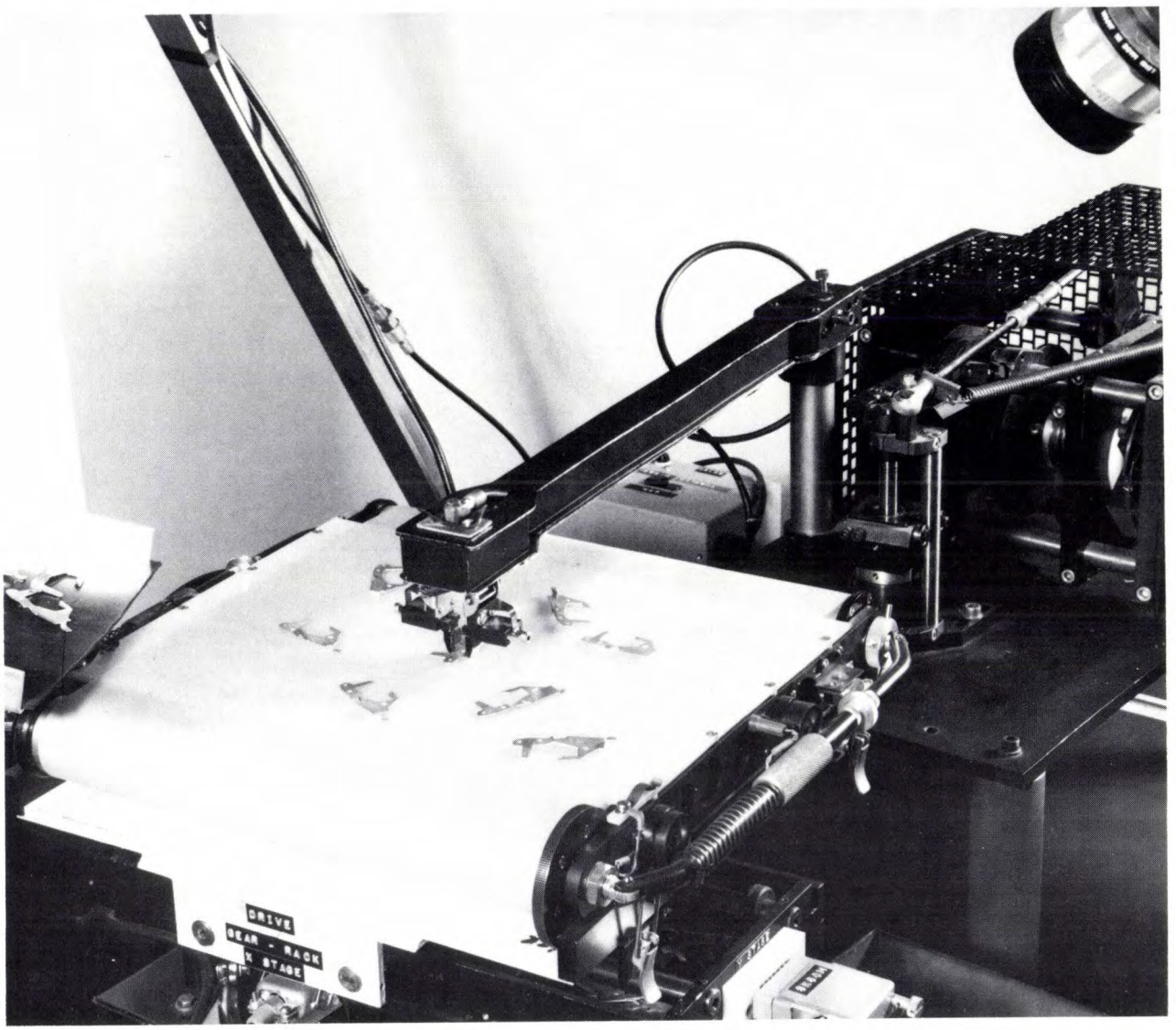


Fig. 3. Photograph of part of the pick-and-place machine. The chute is on the left. One of the two cameras can be seen at top right. The manipulator is in the process of picking up an object. On top of the table is a translucent conveyor belt. Lights within this belt provide back lighting for use with highly reflective parts.

for positional errors in component placing. Thus it may no longer be necessary to specify such high accuracies either for component positioning or for the machine. These high accuracies are currently sometimes necessary more to allow automatic assembly than for any functional need in the component itself.

Vision is certainly not the only sense required. The operation of screwing a bolt into a tapped hole highlights the need for other senses. First the bolt and hole have each to be individually located and brought together. This can be done using visual sensors for control. Once the bolt is positioned at the start of the tapped hole, vision is of little use — either for man or machine — as it may be physically impossible to see enough of the situation in which the bolt is being inserted. In any case it may not be possible to derive sufficient precision from the analysis of the visual image to control the action of screwing the bolt home. For this purpose it may be essential to use tactile sensing — force feedback — together with a degree of controlled compliance in the 'wrist' of the gripper or manipulator.

An experimental visually controlled pick-and-place machine

Pick-and-place operations

Pick-and-place operations are basic to countless everyday tasks performed by each one of us. The operation involves selecting the required object, determining its position and orientation, picking it up and placing it into another specified state. There are many repetitive pick-and-place tasks in industry currently being performed by people who could do more rewarding and profitable jobs.

If it were always possible to keep hold of components from the point of manufacture to final assembly, then these tasks could be solved by hard automation. However, the efficient use of production machinery such as presses, and the need for components to pass through processes such as deburring, often mean that the components are stored in a loose jumbled state. There is therefore a need for intelligent, sensorily controlled pick-and-place machines that can sense their environment and are capable of reacting to changes in that environment. An experimental machine of this type, which has been built at PRL, will now be described.

Structure of the machine

This experimental machine, whose construction is shown schematically in *fig. 2*, can deal with two-dimensional and simple three-dimensional parts. It is controlled by a Philips P 857 computer, and the programs are written almost entirely in RTL/2, a flexible high-level language that is similar to ALGOL 68. This machine performs the seven operations of delivery, separation, positioning, inspection, orientation, pick-up and placement of the component.

The heart of the mechanical part of the machine is the table T. This consists of two slides at right angles mounted on a base that can be rotated. Any object can thus be put into a fixed orientation at the axis of rotation. On top of the table is a conveyor belt. The commands to control the table are derived by the computer from the information it receives from two television cameras TV. A manipulator M picks up the objects that have reached the fixed position just described, and then performs a simple 'goal-post' action to transfer them to one of two output stations, each of which can be positioned at the put-down point.

The parts are supplied by a system consisting of a bowl feeder and a linear chute. No special attachments are fitted to the bowl feeder, which can therefore feed many different kinds of parts without adjustment. The parts arrive on the table in random order and position, of course. *Fig. 3* is a photograph of the system.

The motors driving the mechanical peripherals are connected to the P 857 computer via a communication system called 'Dataline', which has been developed in the laboratory. The system consists of a closed loop of coaxial cable linking up one or more computers of the P 800 family with a number of peripheral devices, each of which can be inserted into the loop at any desired point.

World model

The world model for the visually controlled machine shown in fig. 2 contains two types of information, a set of invariants and a machine-state vector. The invariants include the assumptions about the real world and the task to be performed, as specified by the program designer. As an example, one of the assumptions is that a linear relationship exists between the orthogonal movements x_{tb} and y_{tb} of the table in the real world and the corresponding displacements x_{TV} and y_{TV} in the television picture. This linear relationship is expressed by:

$$\begin{aligned} x_{TV} &= S_x(x_{tb}\cos\theta - y_{tb}\sin\theta), \\ y_{TV} &= S_y(y_{tb}\cos\theta + x_{tb}\sin\theta), \end{aligned} \qquad (1)$$

where θ is the angle between the two frames of reference. The actual values of the scale factors S_x and S_y are elements of the machine-state vector, since they include those model parameters that might be changed when new information is received from the sensors. The main elements of the machine-state vector for the pick-and-place machine are listed in *Table I*.

Program strategy and operation

The flexibility of a visually controlled machine is illustrated not only by its ability to visually identify and locate widely different types of objects, but also by two other features. Firstly, it can be simply trained to perform new tasks without reprogramming. Secondly, it can use its sensors to calibrate itself. For example, the machine must know the relationship between the position of an object in the TV image and its position on the table in order to decide how far to move the table to bring the object to its pick-up position.

The use of the machine is thus divided into the three distinct phases of initialization/calibration, training and operation.

Initialization

In the initialization process the machine is put in a specified state and then allowed to calibrate itself. Calibration involves measuring the relationship between the various frames of reference within the machine such as the TV and the table. The matrices required to define these relationships are obtained by making a few key measurements. For example, the position of the centre of rotation of the table in the TV frame is determined by locating an object in the TV picture (at x_1,y_1, as shown in *fig. 4*) and then rotating the table through 180° and relocating the object at x_2,y_2. The centre of rotation is at the point $\frac{1}{2}(x_1 + x_2),\frac{1}{2}(y_1 + y_2)$.

The scale between the TV frame and the table and their angular relationship are defined by equations (1). The values of S_x, S_y and θ can be calculated by locating an object at one position, moving the table by known distances in x and y and relocating the object in the TV field.

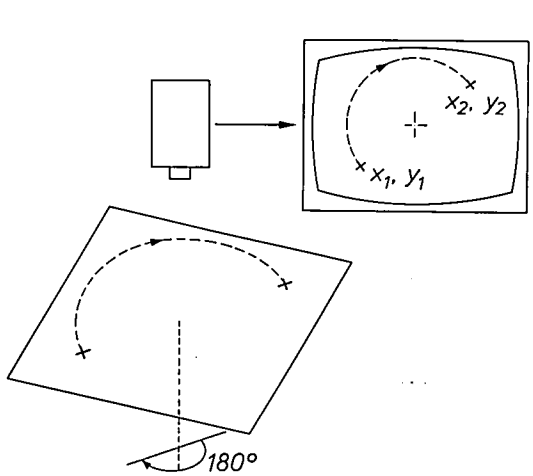


Fig. 4. Determination of the centre of rotation in the TV frame. During the calibration of the machine, the position of the centre of rotation of the table in the TV frame is determined by locating an object in the TV picture at x_1,y_1 and then rotating the table through 180° and relocating the object at x_2,y_2. The centre of rotation is at $\frac{1}{2}(x_1 + x_2),\frac{1}{2}(y_1 + y_2)$.

The training phase

In the training phase the machine must learn the task and acquire the information which is necessary for the task to be performed automatically during the operating phase. The information that must be transferred includes a description of the object, its correct state for being picked up, and a record of the conditions under which the pictures were taken and processed. This constitutes a 'model' of the object. Since the training is performed off-line (i.e. the machine is not in the production line during the training phase), the time spent on training is not critical but the application of the resultant model in the operating phase must be very rapid.

Table I. The contents of the world model of the experimental visually controlled pick-and-place machine.

Invariants	Machine-state vector
Task to be performed	Table location and orientation
Options within the program	Scales: TV scene/real world
Signal-handling strategies, e.g. thresholding	Setting of zoom lens
Machine structure	Histogram of grey levels
Useful life and confidence machine-state vector	Position of arm

After initialization the machine is 'shown' examples of the correct object on the table in the correct position and orientation for pick-up. The object can be placed in this correct state by loading the pick-up arm with the object at the put-down station, and then driving the arm in reverse and making it place the object on the table. The machine then rotates the table to several positions and stores the resultant rotated views of the object.

The machine is also shown false objects that it might encounter during the operation phase. The scenes from both TV cameras are processed and subjected to a set of modelling programs to select models which can distinguish the views of the correct object from those of false ones and which can define the pick-up orientation and attitude of the correct object.

In the operating phase the machine may have to deal with mixtures of different objects. This training procedure must therefore be repeated for every different object that the machine has to handle. In the case of objects having more than one stable state which cannot be transformed into one another by translation or rotation of the table, each stable state is treated as a separate object. The flat object shown in *fig. 5* has two stable states.

Modelling procedures

The models which are currently available are all based on the outline of the object as extracted from the TV pictures stored in the computer. The interface between the cameras and the computer samples the video signal and converts it to a digital signal with 5 bits per pixel (picture element) which is stored in the computer. A thresholding algorithm then produces a binary image of the scene and all closed edges are extracted [2]. The threshold level can be determined automatically using a grey-level histogram or alternatively it can be determined interactively during the modelling phase to ensure that the correct edges are traced. As a result of this 'global' scene processing the modelling program has all internal edges and the outline of the object available. The centre of gravity *CG* of the area within the outline is used to define the position of the object (see *fig. 6*) since this is independent of the orientation of the object. Features of the outline and their position round the edge of the object can be used to derive both identity and orientation models. These features include local maxima and minima in the distance d from *CG* to the edge, and sharp changes in the edge direction (i.e. corners). If for example an object has a unique maximum in d not possessed by any false object then the existence of the maximum is an identity model and the direction of the edge point (at the maximum) from *CG* defines the orientation of the object. Identity and orientation models are derived from vertical views while attitude models use side views.

The available models include:

Model A — basic model

This program uses the object outline to determine the area A, the outline length l, the centre of the area *CG*, and d_{max}, which is the distance from the furthest point on the edge to *CG*. This model is used solely to characterize the position of the object, and to distinguish correct from false objects using A, l and d_{max}.

Model B — intersections of a circle

In this case the program finds the intersections of a circle, centred at *CG*, with the edge of the object (*fig. 7*) [3]. The relative spacing (round the circle) of the intersections can be used to characterize the identity of the object and their absolute position to characterize the orientation. The modelling program chooses 'good' radii for potential circles according to the following criteria:

— the number of intersections should be reasonable (2 to 10);

— the intersections should be stable despite small changes in radius (i.e. they should not move, appear or disappear);

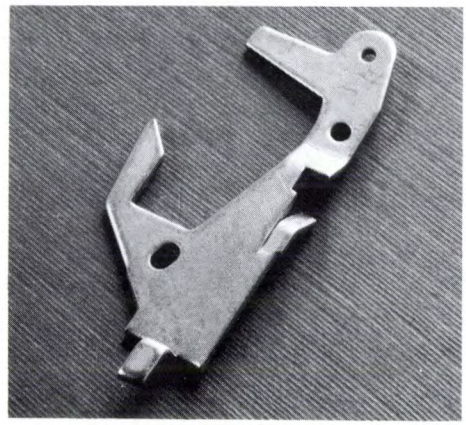


Fig. 5. Flat object which was used during the pick-and-place experiments. It is part of the mechanism of a cassette player.

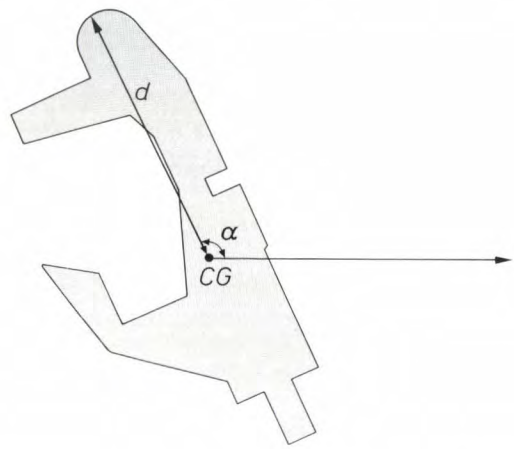


Fig. 6. The centre of gravity *CG* of the area within the outline is used to define the position of the object. Local maxima and minima in the distance d from *CG* to the edge and their angular position α relative to a reference direction (*arrow*) are features that can be used to derive both the identity and the orientation of an object.

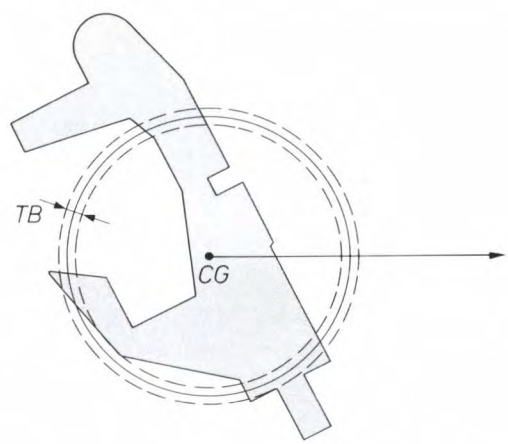


Fig. 7. Circle model. The intersections of a circle, centred at *CG*, with the outside are found. Their relative spacing around the circle characterizes the identity of the object, their absolute positions its orientation. *TB* tolerance band; an intersection is only recorded after the edge has passed this band.

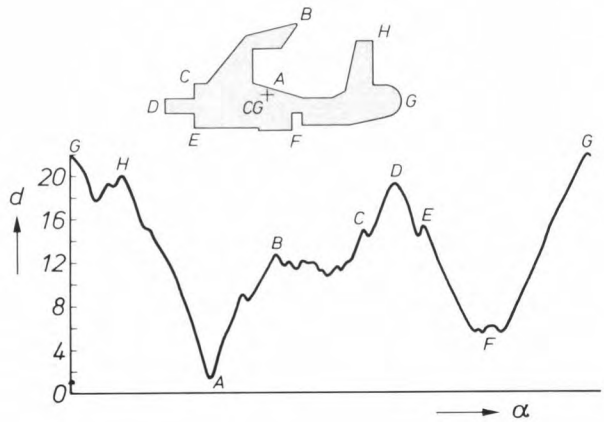


Fig. 8. Recorded distance d of each point of the outline to CG.

— additional degrees of apparent rotational symmetry should not be introduced.

Uncertain intersections are ignored if they do not cause the edge to leave the tolerance band TB on either side of the circle (see fig. 7). A large number of edge points within the band is another indication of a 'poor' radius.

In this model the direction of d_{max} from CG is also used as a feature. The comparison of circle models with the standard is performed with a matching algorithm which allows for small variations caused by changes in lighting, deviations from camera linearity, quantization and other errors.

The matching algorithm finds the best match between a list of the angular positions of the features of an object and a similar list for a standard object stored in the memory. The list for the standard object is twice as long as for the unknown object, containing maximum and minimum limits for the angular positions of each feature.

For an object to be recognized it must contain the same features as the standard list in the same sequential order, although not necessarily with the same feature at the top of the list.

The standard list is designed to have a 'unique' feature or combination of features at the top. The first stage of matching is to find this unique feature in the 'unknown' list. Then the average angular displacement γ between the corresponding features in the two lists is computed. If the angular positions of the 'unknown' features, minus γ, lie within the limits of the standard list, the object is recognized.

Model C — distance maxima and minima

The program follows the outline of the object, starting at the point furthest from CG, and records the maxima and minima in the distance d (see *fig. 8*). The modelling program has an adjustable parameter defining the minimum change Δd which must occur on either side of a singularity before it is recorded as a maximum or minimum. As with model B, the program chooses a Δd which produces a reasonable number of stable maxima and minima.

Although both the angular position and the distance from CG of each maximum and minimum are recorded, it is usually only necessary to use the angular information which is matched with other objects using the same elastic procedure as under model B.

Not necessarily all of the maxima and minima have to come into play. There are different strategies of combining maxima and minima, and very often a single feature, e.g. the first maximum after the minimum, will be sufficient.

Model D — straight edge

The application of this model is limited to the determination of object attitudes from the horizontal camera. The edge information is used to detect whether there is a horizontal flat edge at the top or bottom of the object. This model will for example differentiate between a standard integrated circuit lying on its back and one standing on its legs.

Model E — corners

The outline is examined for sharp changes in the direction of the edge. The change in direction is determined by calculating the angle between two vectors on either side of the edge points being examined (see *fig. 9*). Again the program chooses the values for adjustable parameters, such as the number of edge points spanned by the corner-defining vectors, and the permissible range of angles within which the feature must lie to be recorded.

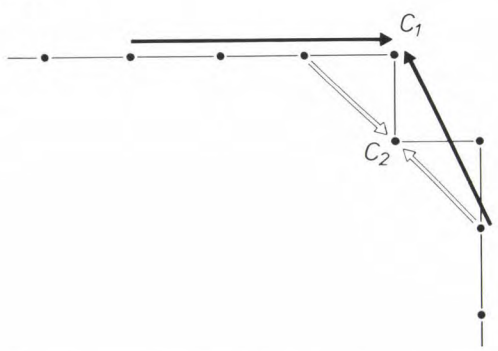


Fig. 9. Finding corners. The angle between two vectors on either side of a prospective corner point is calculated. The number of edge points spanned by the vectors is an adjustable parameter, and so is the permissible range of angle for a corner to be recorded. The point C_1 is recognized as a corner point, irrespective of the length of the vectors. The point C_2 will not be recorded as a corner point if vectors are used spanning three edge points, but it will if they are shorter (two edge points) or longer (four or more).

[2] P. Saraga and P. R. Wavish, Edge tracing in binary arrays, Proc. Conf. on Machine perception of patterns and pictures, Teddington 1972 (Inst. Phys. Conf. Ser. No. 13), pp. 294-302.

[3] A. Pugh, W. B. Heginbotham and P. W. Kitchin, Visual feedback applied to programmable assembly machines, 2nd Int. Symp. on Industrial robots, Chicago 1972.

Operation

In the operating phase the results of applying the appropriate models are compared with the values obtained during modelling. These values are available from the store. Assuming that some time has elapsed since the training phase, the machine must first be initialized to bring it into the state used during training. At the start of the load cycle the table moves to an extreme position, where objects are delivered from the bowl feeder on to the belt, which rotates to spread the objects along its length.

By using pictures from the vertical camera, an object lying in any position and orientation on the table can be located and brought to the axis of rotation by the x and y movements of the table. Its orientation can be determined by comparison with a stored model. The table can then be rotated to bring the object to a predefined position and orientation. If necessary, the object attitude can be determined by using the picture from the side camera. The manipulator can then pick up the object and place it into one of two output stations. For the flat objects, as shown in fig. 5, one output station can be used to receive objects one way up, the other any objects the other way up.

When all objects in the acceptable attitude have been picked from the table a new load cycle is started. As the belt moves forward during the loading operation all objects remaining from the previous cycle fall off into a hopper, ready to be recycled.

Running under an operating system that allows 32 kilowords of memory storage for the pick-and-place program, the machine currently takes about five seconds to locate and pick up an object, of which at least three seconds are taken up by the mechanical movements of the arm and table.

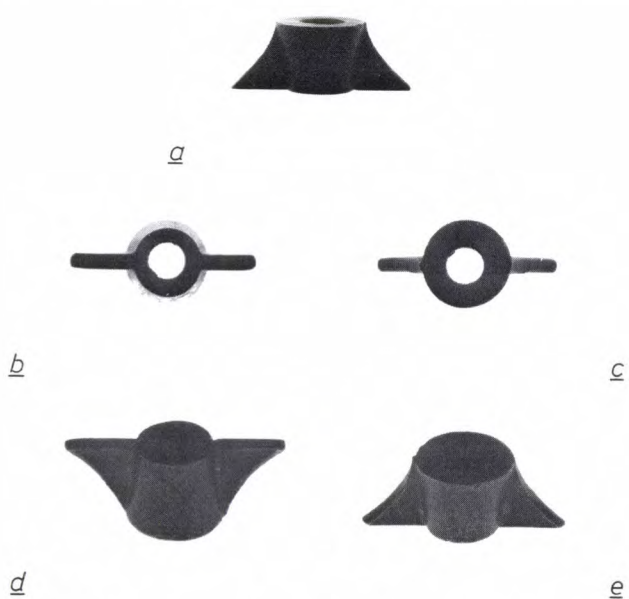


Fig. 10. Wing nut, as seen by the vertical TV camera (*a*, *b*, c) and by the side TV camera (*d*, *e*). The wing nut has three stable positions: lying on its side (*a*), standing upright (*b*) and standing upside down (*c*); the vertical camera cannot distinguish between (*b*) and (*c*), but the side camera can.

Extensions

Three-dimensional objects

The system as described so far can handle three-dimensional objects, provided that each stable state presents a unique shape to the vertical camera. This is not true for all objects. For example, the wing nut shown in *fig. 10* has three stable states, but two of these (fig. 10*b* and *c*) cannot be distinguished in the pictures from the overhead camera, once these are reduced to outlines. These two states are treated as identical, and the object is rotated and moved to the pick-up position. At this point the two states present quite different views to the side camera (fig. 10*d* and *e*) and can easily be distinguished using an attitude model, for instance based on corner features.

Assembly of two-dimensional objects

As well as the pick-and-place operation, simple two-dimensional assembly has also been achieved. By an extension of the training described above, the machine is 'shown' a spatial arrangement of several different objects, such as the assembled jigsaw puzzle and it learns the identity, position and orientation of each object. If in a subsequent operating phase, these objects are lying randomly on the table the machine can reposition them into the learned arrangement by bringing each object in turn to its pick-up position, picking it up, moving the table, and then replacing the object on the table; see *fig. 11*.

Conclusions and future outlook

In its present form, the machine is clearly too slow to be acceptable in a factory. It is important to draw a distinction, however, between the present machine and a likely implementation in a factory; indeed it is not the intention of the research activity to attempt to make actual factory-automation systems. The objective is rather to generate knowledge about techniques and 'tools' which are applicable to advanced factory automation. It is hoped that the work described here demonstrates the practicality of the exciting concept of production methods that can deal effectively with locally varying conditions and which can also be used for a number of different tasks in quick succession

[4] This manipulator has been constructed at Philips Research Laboratories/Centre for Technology, Eindhoven.

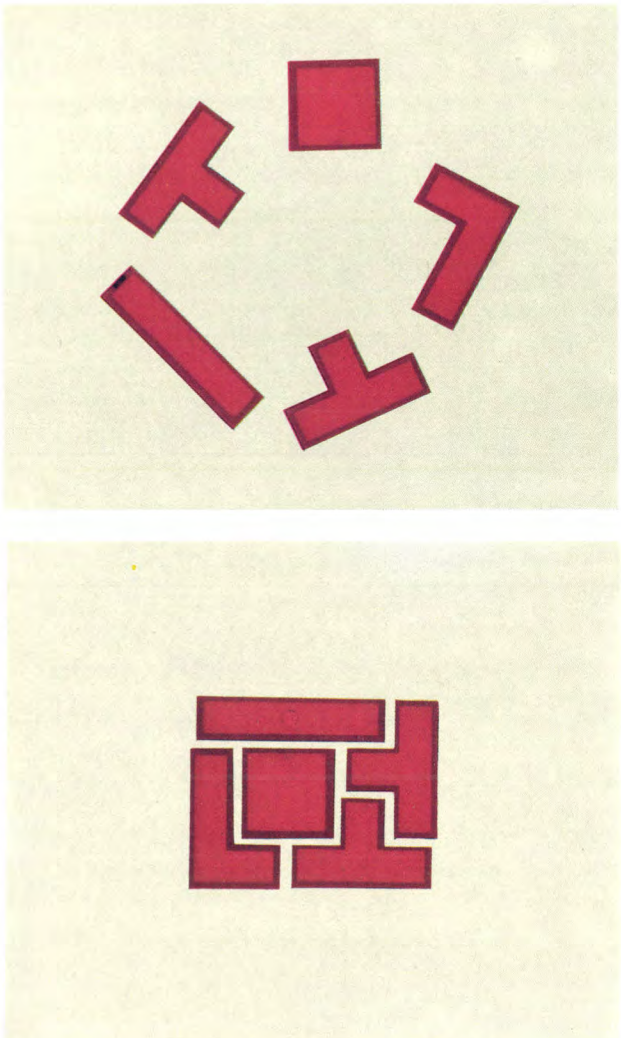


Fig. 11. Jigsaw puzzle before and after assembly by the pick-and-place machine.

without a complete redesign of the production machinery.

As well as the approach described above, others are under study. Instead of using a movable table and a relatively simple manipulator, a more complex manipulator having all the required degrees of freedom and working above a fixed table could be used. Experiments are now being done with such a hydraulic 'arm-like' manipulator which has four degrees of freedom and is equipped with force feedback [4].

In order to achieve the speed required for a factory system, it seems necessary to distribute the processing power among the constituent parts. The mechanics and the picture processing may, for example, be controlled by separate microprocessor systems.

There still are several areas of difficulty in which current knowledge is barely adequate for the task. We do not yet know enough about how to program multiple processor systems, nor yet how to connect them together effectively. Moreover, software is needed that is natural and easy to use in the workshop, yet effective in this multiprocessor environment. More efficient methods for the real-time processing of sensory data will have to be worked out, as well as cost-effective path-control algorithms for the handling devices — both to be implemented in cheap microprocessors.

A more effective programmable handling device, suited to the needs of our industry, may also be needed. However, in spite of present shortcomings of available technology, the techniques of computer science, artificial intelligence and pattern recognition will combine to make a significant impact within the next few years on the production methods applied to a wide range of electronic products — from integrated circuits to washing machines.

Summary. Rising labour costs in the electronic industry are a powerful stimulus to seeking a higher degree of automation which will also relieve human labour from routine assembly tasks. To be cost-effective also for small production runs, handling devices are sought which can be reprogrammed to handle a new task, and which are tolerant of positional errors in the system ('flexible automation'). Visual and tactile sensory feedback greatly enhances flexibility. As an experiment, a visually controlled pick-and-place machine has been built which can recognize objects lying randomly on a table, pick them up and put them down at a specified point in a specified orientation. The machine, whose main parts are a gripper arm, a movable table and two TV cameras, is controlled by a P 857 minicomputer.

Flashing tomosynthesis

E. Klotz, R. Linde, U. Tiemens and H. Weiss

When a radiograph is made the shadows cast by the different layers of the object are superimposed on each other. The conventional X-ray technique is therefore essentially two-dimensional; the distribution of the features along the third dimension cannot be reconstituted from the picture.

Several techniques have been developed for retaining the information about the third dimension. They all rely on a reconstruction of a particular layer of the object from a number of pictures taken from different angles. The geometrical manipulation required can be performed optically, as in the system to be described here, but it can also be performed electronically — after scanning the picture — e.g. by a computer [1]. In the system to be described an array of X-ray tubes is flashed simultaneously. The resulting shadow images are all recorded together on the same film. The image of the desired layer is synthesized by means of a simple and straightforward optical system. We call the procedure 'flashing tomosynthesis'. It is now undergoing a field test in the university hospital at Würzburg, West Germany. Examples of pictures taken by means of flashing tomosynthesis will be discussed in this article.

The system has distinct advantages over the system of 'tomography' that has been in use for a number of years. In tomography the X-ray source and the film are moved simultaneously along circular paths in such a way that the projection of one layer of the object on the film remains stationary, whereas the projections of the other layers move and are therefore blurred. Since this system depends on an integration over time, it does not allow pictures of moving objects to be taken — e.g. a pulsating heart or fast contrast media injected into blood vessels. The search for the appropriate diagnostic plane may require a repeated exposure of the patient (sometimes up to sixty times), resulting in a high radiation dose. In addition the mechanical equipment required is large, yet must nevertheless be accurate to a tenth of a millimetre.

With flashing tomosynthesis long exposures are not necessary since all the processing can be carried out afterwards on photographic images [2]. In this post-processing or 'decoding' the details of any desired layer can be extracted from the multiple image recorded on the film. This multiple image consists of projections of the object from different angles (*fig. 1*). The decoding consists in projecting the multiple image through an array of lenses similar to the array of X-ray sources. This allows the individual projections to be displaced so that the desired detail in the different projections coincides and is thus enhanced. Details from other layers do not coincide and form a 'noise background'.

The process described here is an example of what is known as 'coded' or 'synthetic-aperture' imaging. The name derives from the fact that the optical aperture of the system is composed of a number of elements in a special way. We shall have more to say about this in the next section, where we shall also discuss the signal-to-noise ratio of the decoded picture, which is of vital importance for the practical value of the system. In the sections that follow we shall describe a prototype apparatus constructed in our laboratories, and also advanced coding and decoding methods for optimizing the signal-to-noise ratio.

Synthetic-aperture imaging

In recent years the technique of synthetic-aperture imaging has been extensively developed for a number of diverse applications. It is adopted wherever antennas or lenses of appropriate size do not exist, or where the radiation is of a kind that cannot be focused at all. Radar and radioastronomy are examples of the successful use of synthetic apertures. Its recent application to γ-ray imaging has attracted considerable attention.

When speaking of γ-rays we tend to think of gamma-ray astronomy, isotope imaging and medical X-ray diagnosis [3]. In γ-ray astronomy the synthetic aperture is used to obtain a higher signal-to-noise ratio [4]. In the other applications a more important objective is to obtain a tomographic effect, i.e. to image a single layer of a three-dimensional object.

Synthetic apertures can be roughly divided into two types: continuous ones such as Fresnel zone plates [5], and discrete ones, which are point distributions [6]. Point distributions have the advantage that they can be produced in practice either as a 'passive aperture' consisting of an array of holes, or as an 'active aperture', e.g. consisting of an array of X-ray tubes, as in the system to be described here.

Ing. grad. E. Klotz, Ing. grad. R. Linde, Ing. grad. U. Tiemens and Dr H. Weiss are with Philips GmbH Forschungslaboratorium Hamburg, Hamburg, West Germany.

Synthetic-aperture imaging in flashing tomosynthesis

In flashing tomosynthesis the first step is the formation of a multiple, 'coded' image. This is illustrated in *fig. 2*, where the object is assumed to consist of a square and a circle situated in two different planes O_i. The object is illuminated by three point sources distributed in space in a point distribution P. A multiple image O' of the object is obtained, consisting of the three projections of the object from different angles.

The second stage of the process is to decode the image. In our system this consists in projecting the multiple image through an array of lenses L, which is similar to the point distribution P (fig. 1). With proper adjustment images O'' can be obtained as shown in *fig. 3*. In fig. 3*a* the three projections of the multiple image are adjusted so that the three images of the square coincide and the square is therefore enhanced, whereas the other images together form the noise background. By readjusting the projections, three images of the circle can be made to coincide (fig. 3*b*), and the circle will then stand out from the background.

Ideally, this background should be zero. Obviously, this cannot be achieved — although the actual perturbation from the background depends on the geometry of the point array. What are termed non-redundant point distributions (NRPD) [7] are the best patterns

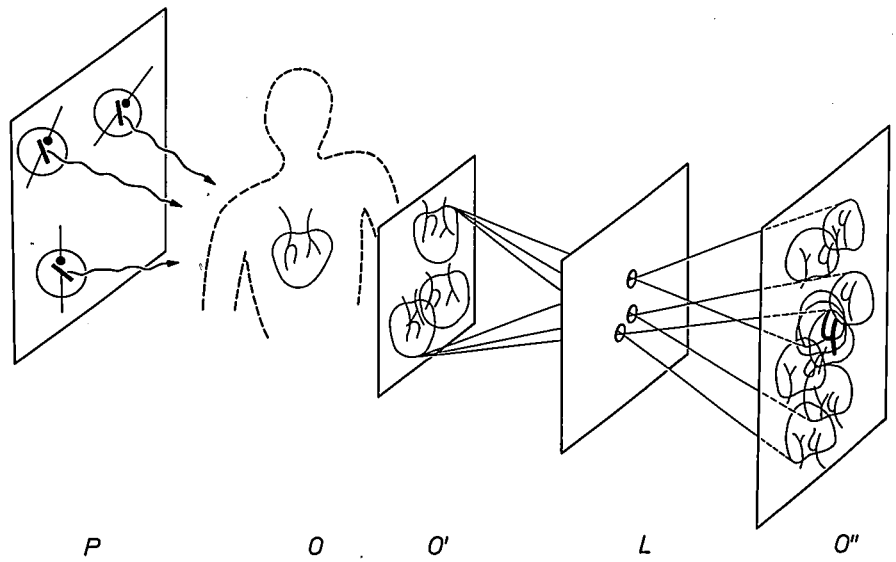


Fig. 1. Simplified diagram of flashing tomosynthesis. An array P of X-ray tubes is flashed simultaneously to give a 'coded' image O' of the object O. The coded image O' (which is recorded on X-ray film) is projected through a lens array L geometrically similar to the X-ray tube array to give a decoded image O'' of the desired layer of the object. The desired image is accompanied by artefacts due to overlapping in the coded image and second maxima in the function describing the autocorrelation of the point array in P.

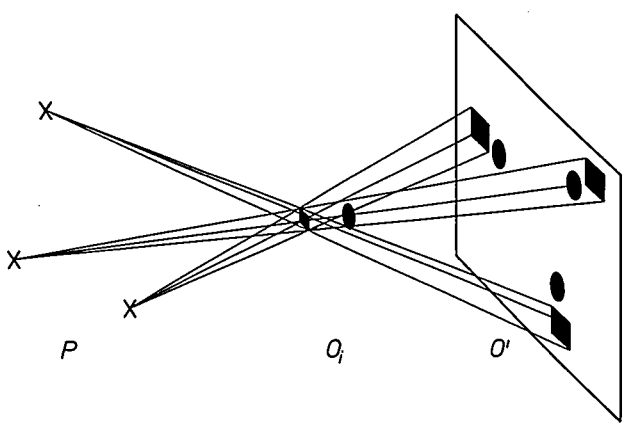


Fig. 2. The encoding. An object is assumed to consist of a square and a circle in two different layers O_i. Illumination by three sources arranged in a point distribution P results in a coded image O' composed of three different projections of the object.

[1] H. Weiss, E. Klotz and R. Linde, Deconvolution systems for coded aperture images of three-dimensional X-ray objects, Optics and Laser Technol. 7, 117-120, 1975.

[2] The short exposure time used makes it possible to record a moving film showing the movements in one particular layer, e.g. of the pulsating heart (cinetomosynthesis). See H. Weiss, E. Klotz and R. Linde, Flashing tomosynthesis: three-dimensional X-ray imaging, Acta Electronica 22, 41-50, 1979 (No. 1).

[3] E. Klotz and H. Weiss, Three-dimensional coded aperture imaging using nonredundant point distributions, Optics Comm. 11, 368-372, 1974.

[4] G. Groh, G. S. Hayat and G. W. Stroke, X-ray and γ-ray imaging with multiple-pinhole cameras using a posteriori image synthesis, Appl. Optics 11, 931-933, 1972.

[5] W. L. Rogers, K. S. Han, L. W. Jones and W. H. Beierwaltes, Application of a Fresnel zone plate to gamma-ray imaging, J. nucl. Med. 13, 612-615, 1972.

[6] H. Weiss, E. Klotz, R. Linde, G. Rabe and U. Tiemens, Coded aperture imaging with X-rays (flashing tomosynthesis), Optica Acta 24, 305-325, 1977.

that can be used here. When n sources are arranged in an NRPD a main image is obtained of intensity n and $n(n-1)$ side images of intensity 1. Fig. 3 illustrates the case for $n = 3$.

The side images are one of the two factors that determine the signal-to-noise ratio of the reconstructed layer. If all the side images were to interfere with the main image the signal-to-noise ratio would be $n/n(n-1)$, i.e. less than 1, and no useful decoded image would result. Whether this will be the case or not depends on the three-dimensional geometry of the object under investigation; many objects, however, give pictures that do not overlap to such an extent that all the side images interfere with the main image. The artefacts created by neighbouring layers form the second factor, and these cannot be predicted either. A general calculation of the signal-to-noise ratio cannot therefore be given.



Fig. 3. The decoding. Three projections of the coded image O' of fig. 2 are superimposed after being displaced in such a way that either three squares coincide (*a*) or three circles (*b*). The crosses indicate the centres of the three projections. The non-coinciding picture components form a 'noise' background.

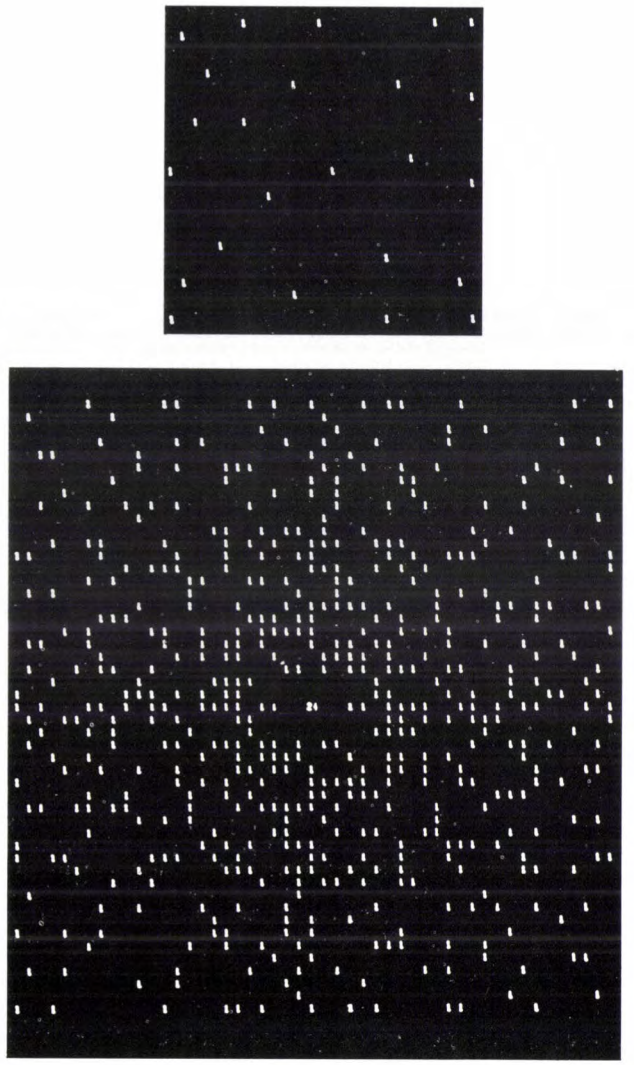


Fig. 4. *a*) Non-redundant point distribution (NRPD) of 24 points. *b*) Its correlation function. The peak in the centre reaches a height of 24, the background side peaks all have a height of 1.

Mathematical treatment

The multiple image O' can be described as the convolution product of the object O and the function P that describes the geometry of the array of X-ray sources:

$$O' = O * P. \qquad (1)$$

The asterisk denotes the convolution. O, O' and P are functions of two variables that define the position in the different planes. This position will be designed here by a two-dimensional positional vector $\boldsymbol{x}$.

The function P is called the point-spread function. If the number of layers is finite, as in fig. 2, and equal

to m, eq. (1) may be expressed by a sum:

$$O' = \sum_{i=1}^{m} O_i * P_i. \qquad (2)$$

The point-spread function P_i for every layer O_i of the object is similar to the distribution function for the array of sources, but multiplied by a scale factor λ_i, which is determined by the central projection geometry. If there are n sources, P_i is given by:

$$P_i = \sum_{l=1}^{n} \delta(\boldsymbol{x} - \lambda_i \boldsymbol{x}_l), \qquad (3)$$

where δ has the value 1 for an argument equal to zero and the value 0 for any other argument (Dirac function); $\boldsymbol{x}_l$ represents the positions of the focal spots of the X-ray tubes in the source-array plane.

The projection of the multiple image O' through the array of lenses is equivalent to the mathematical operation of correlation. If the lens array coincides with the point-spread function P_k that will produce an image of layer O_k, then (using the symbol $\circledast$ for correlation):

$$\begin{aligned} O'' &= O' \circledast P_k \\ &= O_k * P_k \circledast P_k + \sum_{i \neq k} O_i * P_i \circledast P_k. \end{aligned} \qquad (4)$$

This equation presents all the problems of synthetic-aperture imaging in a nutshell. For a perfect reconstruction of the layer O_k the autocorrelation of P_k should yield a δ-function:

$$P_k \circledast P_k = \delta, \qquad (5)$$

and all the cross-correlations should cancel:

$$P_i \circledast P_k = 0, \quad i \neq k \qquad (6)$$

These requirements can only be fulfilled approximately, of course. For an NRPD of n points the δ-function is approximated by a central peak of intensity n surrounded by $n(n-1)$ side peaks of intensity 1. An example of a 24-point NRPD is shown in *fig. 4a*, and its autocorrelation function in fig. 4*b*. The NRPD was calculated using an iterative computer program incorporating the requirement that all possible $n(n-1)$ pairs of coordinate differences between the n points should not occur more than once.

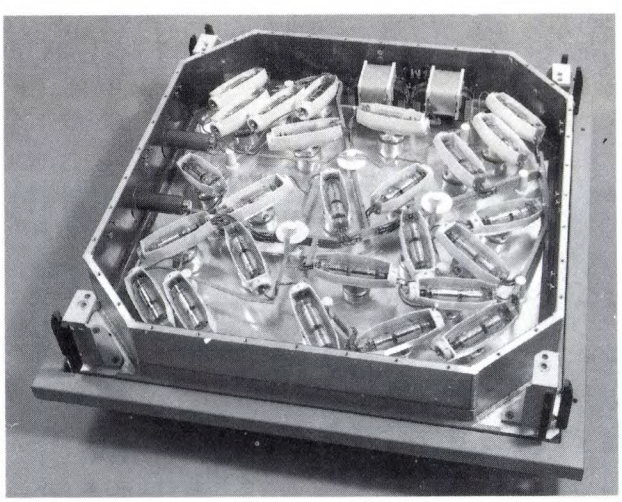


Fig. 5. An array of 25 small X-ray tubes whose focal spots are arranged in a non-redundant point distribution.

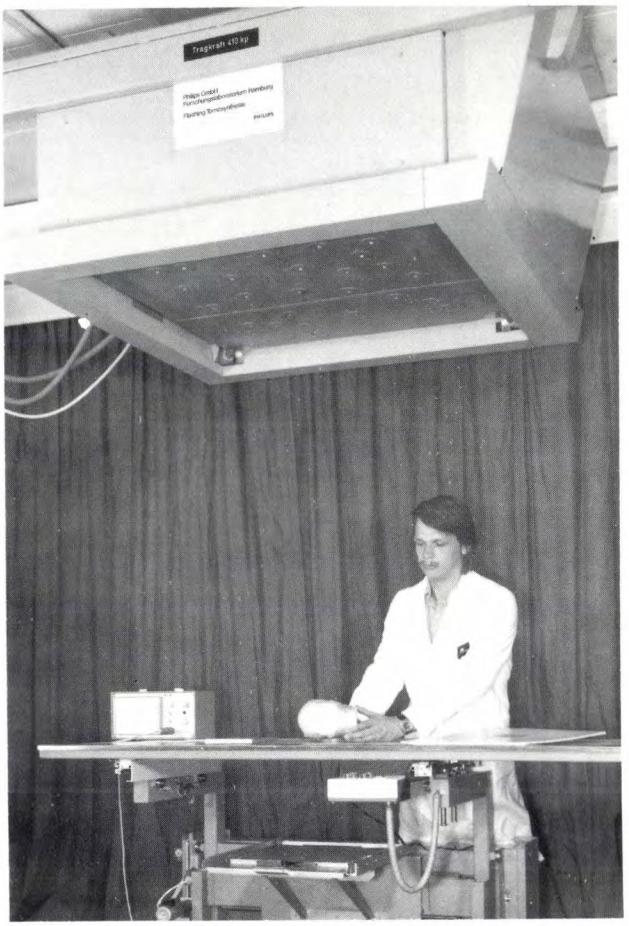


Fig. 6. The complete flashing-tomosynthesis X-ray equipment. *From top to bottom:* exit apertures of X-ray tube array, object table, film cassette.

A prototype flashing-tomosynthesis system

A prototype flashing-tomosynthesis system has been built in our laboratories. *Fig. 5* shows the array of X-ray sources. It consists of 25 small X-ray tubes whose focal spots are arranged in a non-redundant point distribution. The tubes operate in parallel with voltages up to 100 kV and with currents of up to 20 mA each, thus covering all diagnostic applications. The total radiation dose is of the same order of magnitude as for an ordinary radiograph.

The complete equipment needed for taking the radiographs is shown in *fig. 6*. In the overhead fixture

[7] M. J. E. Golay, Point arrays having compact, nonredundant autocorrelations, J. Opt. Soc. Amer. **61**, 272-273, 1971.

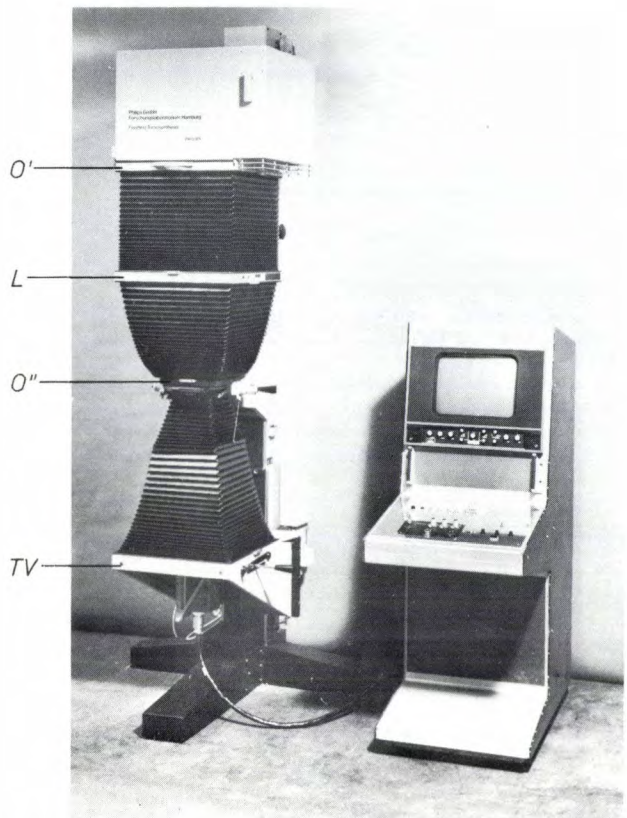
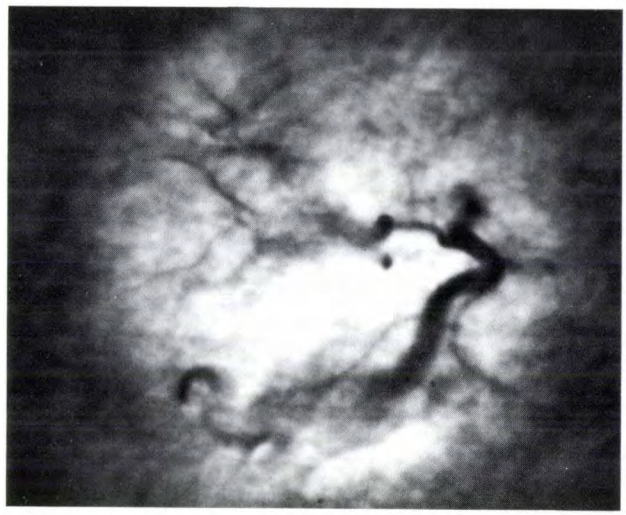
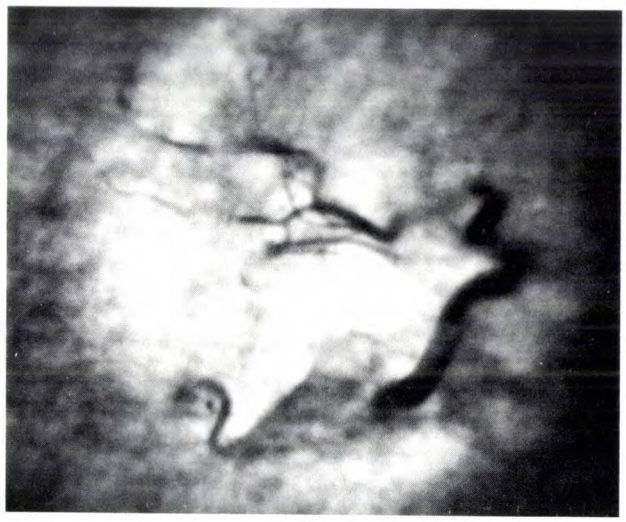


Fig. 7. Flashing-tomosynthesis decoding equipment. The column on the left contains, from top to bottom, a conventional light box illuminating the coded image *O'*, a lens array *L*, a screen displaying the reconstructed image *O''*, and a television camera *TV* connected to the monitor console on the right. By moving the screen up and down the doctor can 'travel through' the object and study the different layers on the TV monitor. A photographic cassette can be inserted at the screen to record a tomogram.



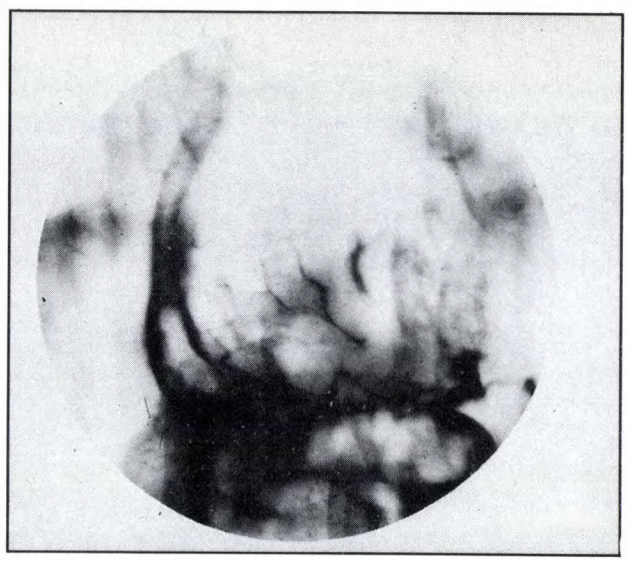
b



c



a



d

Fig. 8. Images of a slice of a model of a human head (a 'phantom') whose vessels are filled with contrast medium. The images were obtained with the equipment shown in the preceding figures. *a*) Multiple coded image. *b*,*c*) Reconstructed images of two different layers. Object-field diameter 90 mm. *d*) Tomogram of the bony structure of the forehead. Object-field diameter 120 mm.

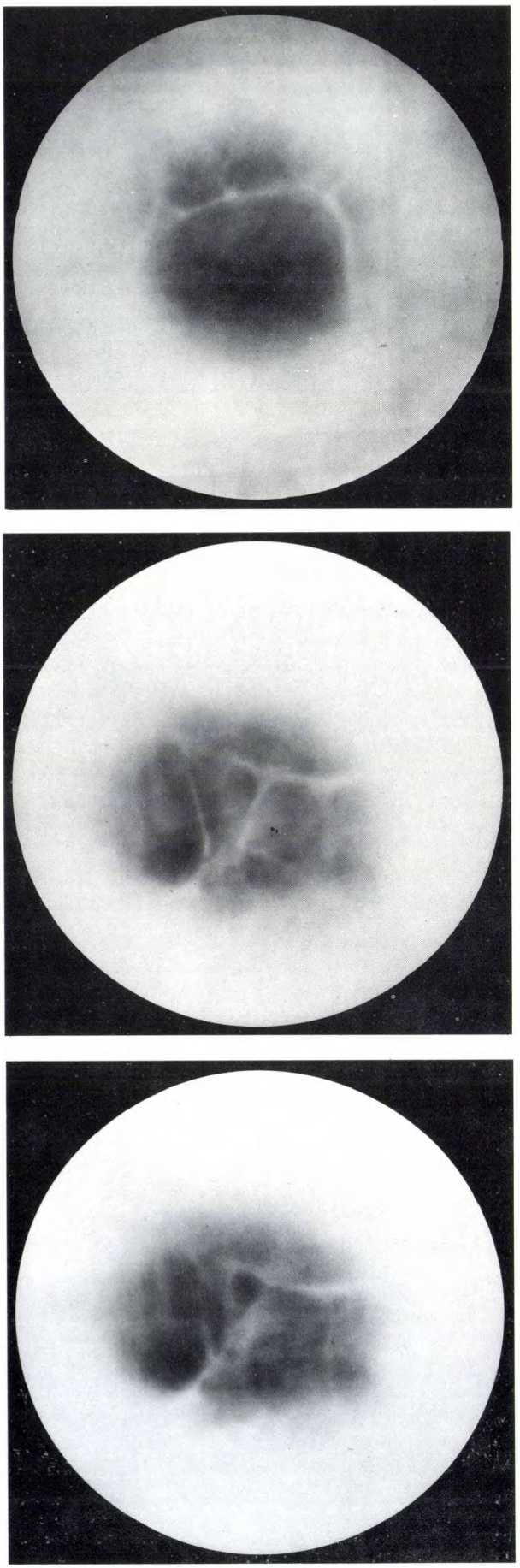


Fig. 9. Three tomograms of a human skull in the region of the eye socket (the orbit), showing the hole for the optic nerve, and other details. The object-field diameter has been reduced to 60 mm, which contributes greatly to the absence of artefacts.

the exit apertures for the X-ray beams can be seen. The table and cassettes are conventional models.

The equipment used for decoding the radiographs is shown in *fig. 7*; this is the equipment installed in the university hospital at Würzburg. The multiple image O' (see fig. 1) is placed in front of a conventional light box. In the plane L twenty-five lenses are arranged to form the same NRPD as the X-ray tubes; a reconstructed three-dimensional image is obtained at O'', where successive cross-sections of the object are shown when a screen is passed through the reconstruction volume. The tomograms can be directly recorded on film or displayed on a monitor via a television camera.

Results obtained with flashing tomosynthesis are presented in figs 8 to 11. *Fig. 8a* shows the multiple image O' obtained by imaging a slice of a model of a human head (a 'phantom'), with its vessels filled with contrast medium. Three reconstructed tomograms are shown in figs 8*b*, 8*c* and 8*d*. Although the paths of the vessels (figs 8*b* and 8*c*) or the bony structure of the forehead (fig. 8*d*) can be clearly seen, artefacts are also observed that are due to the side peaks of the autocorrelation function of the NRPD.

The artefacts are weaker if there is less overlapping in the multiple image O'. The amount of overlapping depends on the object-field diameter chosen. In figs 8*b* and 8*c* the object-field diameter is 90 mm, in fig. 8*d* it is 120 mm. A diameter of 60 mm yields the excellent tomograms in figs 9 to 11. Figs 9 and 10 are images of part of the human skull. *Fig. 9* shows tomograms made in the region of the eye socket (the orbit) and shows the hole for the optic nerve — a very difficult situation in tomography. *Fig. 10* presents two cross-sections derived from one lateral picture of the human skull; the upper tomogram shows the cochlea at the near side, the lower tomogram shows the other side of the head. Together they cover a depth of about 15 cm.

Fig. 11 shows a few early results obtained in the university hospital at Würzburg by Prof. M. Nadjmi[8]. The three tomograms were derived from a multiple image of the human neck taken laterally. The quality of the pictures is comparable with that of ordinary tomograms. To obtain pictures of similar quality from larger object fields it is necessary either to reduce overlapping by increasing the film size or to resort to advanced coding and decoding algorithms. Such algorithms may well provide the answer for the future.

[8] M. Nadjmi, H. Weiss, E. Klotz and R. Linde, Flashing tomosynthesis: first clinical experience, to be published in Medicamundi.

Advanced coding and decoding algorithms

There are two ways of trying to improve upon the coding and decoding of tomographic images. The first is to search for point distributions with a better autocorrelation function, i.e. smaller side peaks; the second is to decode by more effective means than simple autocorrelation.

The first way seems to be illogical since NRPDs by definition possess optimum autocorrelation functions. This only applies, however, if all of the points are assigned the same value, e.g. +1, against a background of zeros. This was the case with the point distributions dealt with so far — which are called 'incoherent codes'.

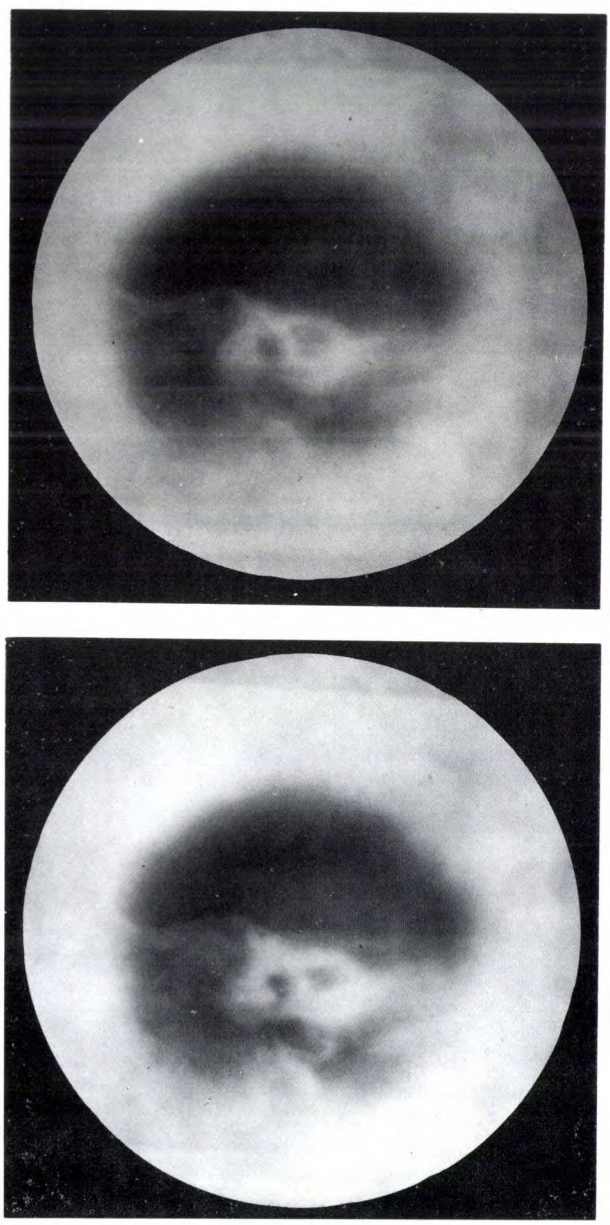


Fig. 10. Two more tomograms of the human skull, derived from a single coded image taken from the side. They represent the cochlea at the near side (*top*) and at the far side (*bottom*) and together span a depth of about 15 cm.

'Coherent codes' also exist, however. They have positive and negative values, which means in practice that it is necessary to take two or more multiple images and to subtract them one from another later [6]. The coherent codes have substantially better, i.e. more δ-like, autocorrelation functions.

Coherent codes

As an example of a coherent code we show the point distribution (*fig. 12a*) and the autocorrelation function of a code for which $n = 16$ (fig. 12*b*). The sum of the intensities of the side peaks is 48, which guarantees a signal-to-noise ratio better by a factor of 5 than that of an incoherent code of order 16, which has 240 side peaks (16×15), all of an intensity +1.

Fig. 13 shows a few successive steps in the decoding process. Fig. 13*a* is a multiple image of a ring. It consists of the superposition of two images, one the coded image resulting from the ten '+1' points of the distribution, and the other from the six '−1' points. The latter has been negated to form a negative image and added to the first. Correlating is also carried out separately for the 'positive' and the 'negative' points. The left-hand image in fig. 13*b* shows the result of correlating the coded image with the positive part of the array, the right-hand image that of correlating with the negative part. Subtracting the second from the first gives the end result in which the background noise is noticeably reduced (fig. 13*c*).

The optimization of two-dimensional coherent codes is still a subject for research. They have the disadvantage that at least two coded images have to be taken instead of one in the incoherent case. If more than two coded images are admitted, a further improvement could be obtained by using two complementary coherent codes [9]. The sum of the autocorrelation functions of these complementary codes does not have any side peaks at all.

Inverse filtering

If incoherent codes are chosen there are still ways of improving the performance. A better signal-to-noise ratio than is possible with simple autocorrelation can be obtained, for example, by decoding with a method called inverse filtering.

The idea of inverse filtering is based on the important property of Fourier transforms that a convolution product in the image domain is equivalent to an ordinary product in the Fourier domain. Image decoding can be described in the image domain by a convolution with a function Q:

$$O'' = O * P * Q \qquad (7)$$

Now the function Q has to be found. Ideally, $P * Q$

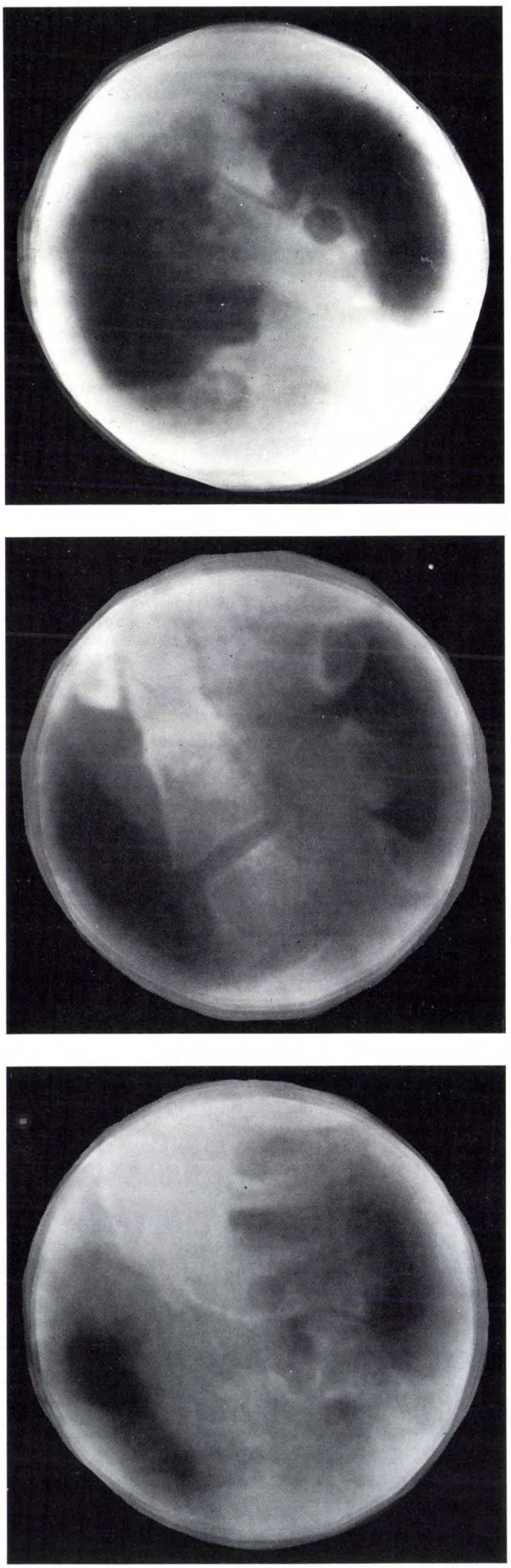


Fig. 11. First results from the trial of the prototype equipment in the university hospital at Würzburg. Three tomograms of the human neck are shown, reconstructed from one encoded image taken laterally.

should be equal to δ. After a two-dimensional Fourier transformation of both sides, the equation will be:

$$\mathfrak{F}\{O''\} = \mathfrak{F}\{O\}.\,\mathfrak{F}\{P\}.\,\mathfrak{F}\{Q\}, \qquad (8)$$

and perfect decoding is seen to be achieved if

$$\mathfrak{F}\{Q\} = 1/\mathfrak{F}\{P\}.$$

Although a mathematical solution has thus been found, putting it to practical use is another matter. The function $1/\mathfrak{F}\{P\}$ becomes infinite at many poles and has to be approximated. In principle it can be approximated adequately by means of a hologram [10], which permits an optical approach to the algorithm. It is not easy, however, to record the large differences in the value of the function $1/\mathfrak{F}\{P\}$ linearly on photographic film. Nevertheless there is hope of success now that a method has been found that considerably simplifies the holographic procedure [11].

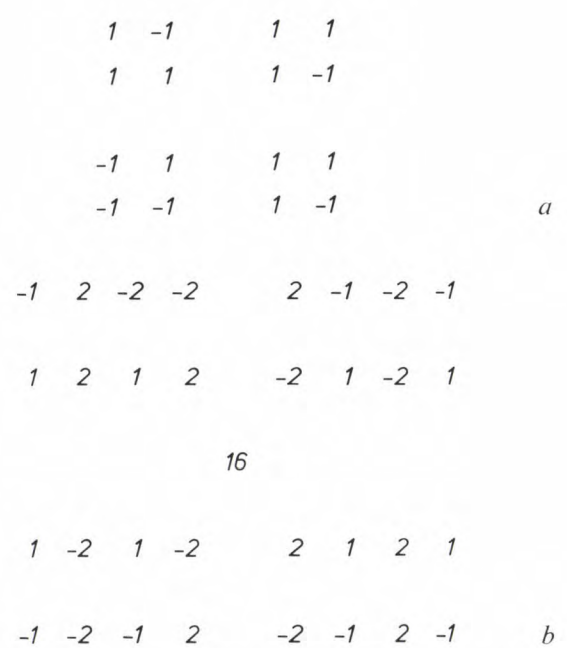


Fig. 12. *a*) Sixteen-point distribution showing positive and negative values, an example of a 'coherent code'. The coherent code is produced in practice by adding positive and negative images. *b*) Autocorrelation function of the code. The sum of the side peaks is only 48, which is far less than in the case of a non-redundant point distribution of 16 points, where it is 240.

[9] N. Ohyama, T. Honda and J. Tsujiuchi, An advanced coded imaging without side lobes, Optics Comm. **27**, 339-344, 1978. K. Pasedach and H. Weiss, W. German Patent No. 28301862.
[10] G. W. Stroke and R. G. Zech, A posteriori image-correcting 'deconvolution' by holographic Fourier-transform division, Phys. Lett. **25A**, 89-90, 1967.
[11] W. J. Dallas, R. Linde and H. Weiss, Simplified production of spatial inverse filters, Optics Lett. **3**, 247-249, 1978.

An image taken of the head slice as before, but now decoded by inverse filtering, is shown in *fig. 14*. A comparison with figs 8*b* and 8*c* clearly shows the reduction of artefacts.

Other approaches to an improvement of the signal-to-noise ratio can be made by starting directly from eq. 6 and attempting to find a deconvolution code Q that satisfies the requirement $P * Q = \delta$. Such codes have indeed been found; we mention here the 'uniformly redundant code' and the 'compensated code', which operate within a restricted region of interest [12].

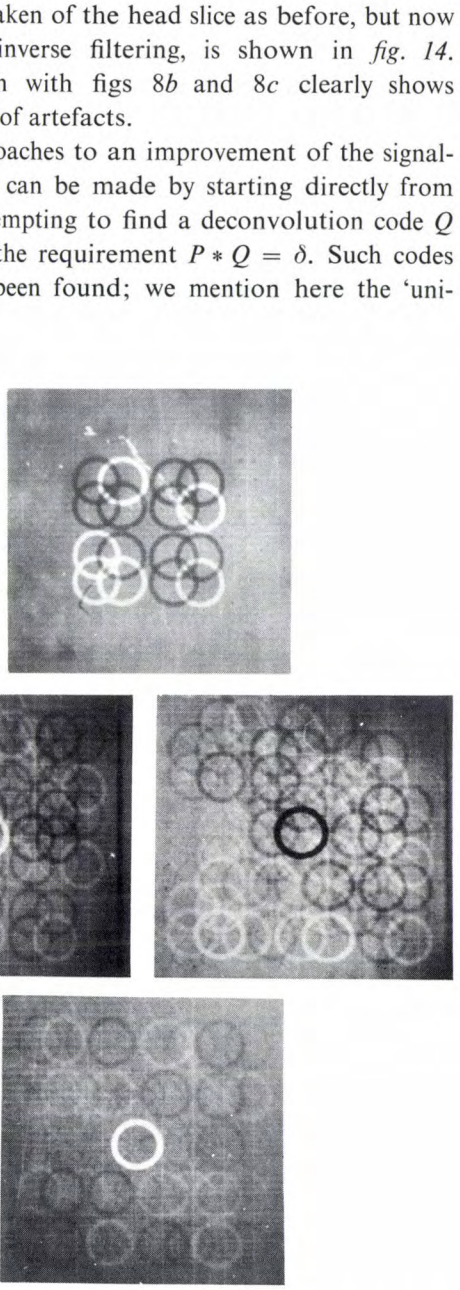


Fig. 13. Decoding the image of a ring encoded by the coherent code shown in fig. 12. *a*) Coded image. *b*) Correlation of the coded image with the positive part of the array (*left*) and with the negative part (*right*). *c*) Result obtained by taking the difference of the two images in (*b*). The background is considerably attenuated.



Fig. 14. Tomogram of head slice obtained by inverse filtering. The absence of artefacts is noticeable when the tomogram is compared with those in figs 8*b* and 8*c*.

In conclusion we can say that, in principle, adequate mathematical solutions of the synthetic-aperture problem have been found, but that working out the algorithms in practice still presents difficulties. Optical methods, although very fast, are rather inflexible; digital methods, which can be very accurate, take up too much time for data input, calculations and data output. If this limitation can be removed, digital processing could become a most powerful tool in synthetic-aperture imaging.

The work described here has received financial support under the medical and technical programme of the Ministry of Research and Technology of the Federal Republic of Germany within the project 'Laser in Medicine and Biology' of the Gesellschaft für Strahlungsforschung. The authors are responsible for the contents of the article.

[12] E. E. Fenimore and T. M. Cannon, Coded aperture imaging with uniformly redundant arrays, Appl. Optics **17**, 337-347, 1978.
W. J. Dallas, Artifact-free region-of-interest reconstruction from coded-aperture recordings, Optics Comm. **30**, 155-158, 1979 (No. 2).

Summary. A 'flashing-tomosynthesis' system has been built with which radiographs of a single layer of an object are produced by simultaneously flashing twenty-five X-ray tubes irradiating the object from different angles. The multiple 'coded' image thus obtained is subsequently decoded by projecting it through an array of lenses similar to the array of X-ray tubes. The coded image is optically correlated in this way with the source array, which is scaled to suit the particular layer of the object to be displayed. For optimum correlation the X-ray tubes are arranged in a 'non-redundant point distribution'. The system is now under trial in the Würzburg university hospital. The object-field diameter can be varied between 60 mm and 120 mm. Future improvements of the signal-to-noise ratio can be expected from more refined decoding techniques.

Computerized mammogram processing

A. Hoyer and W. Spiesberger

Introduction

Breast cancer is a serious disease that can only be treated successfully if it is diagnosed at an early stage. A mass-screening technique capable of routinely detecting breast cancer at a very early stage would therefore be very widely appreciated. The most effective way of detecting breast cancer as yet in mass-screening is that of self-palpation. Obviously, only the larger growths can be discovered in this way; incipient growths are too small.

To improve early diagnosis of breast cancer the idea of combining mammography — i.e. the recording of the radiographic absorption pattern of the breast, see *fig. 1* — with other clinical methods is being considered. Several research groups have studied the reliability of mammography and of clinical methods of examination [1], and have found that a combination of techniques, applied in a systematic diagnostic procedure, could reduce breast-cancer mortality, and that the contribution of mammography in such a procedure is indispensable. The recording of mammograms is an accepted radiological procedure in current medical practice. However, a reliable diagnosis is not always possible, even with skilful and experienced radiologists, since the inspection of a large number of mammograms in a mass-screening examination is tiring, and it is difficult to apply explicit diagnostic principles consistently.

To improve diagnostics, various computer techniques have been proposed that are capable of giving results that are more objective and above all more reliable [2]. In this article we shall describe a procedure for mammogram processing that has been devised at Philips Forschungslaboratorium Hamburg, as part of our general programme of research on medical image processing [3].

Mass-screening examinations for breast cancer are concerned not only with large cancerous structures — such as solid, scirrhus and medullary carcinomas — but also with microstructures, which occur at a very early stage. It is the detection of these microstructures that is of such vital importance. The equipment must therefore be capable of detecting fine as well as gross structures.

For the detection of gross cancer structures we have developed an algorithm that is based on textural features. The principles of measuring these features will be discussed in the next section.

To detect microscopic cancer structures, which are hardly visible in mammograms, indirect indications of cancer have to be used. An important pointer of this type is the presence of microcalcifications: These can be found in the mammogram by the analysis of characteristic structural elements. The procedure that we have developed for this purpose will be described in the final section of this article.

All analysis techniques have been tried out in our experimental image-processing system, which is briefly described in the article by A. Hoyer and M. Schlindwein [4].

Dipl.-Phys. A. Hoyer, formerly with Philips GmbH Forschungslaboratorium Hamburg, is now with Zentralbereich Information, Philips GmbH, Hamburg; Dipl.-Phys. W. Spiesberger was formerly with Philips GmbH Forschungslaboratorium Hamburg, Hamburg, West Germany.

[1] P. Strax, L. Venet, S. Shapiro and S. Gross, Mammography and clinical examination in mass screening for cancer of the breast, Cancer **20**, 2184-2188, 1967.
P. Strax, L. Venet and S. Shapiro, Value of mammography in reduction of mortality from breast cancer in mass screening, Amer. J. Roentgenol. **117**, 686-689, 1973.

[2] L. V. Ackerman, A. N. Mucciardi, E. E. Gose and F. S. Alcorn, Classification of benign and malignant breast tumors on the basis of 36 radiographic properties, Cancer **31**, 342-352, 1973.
L. V. Ackerman and E. E. Gose, Breast lesion classification by computer and xeroradiograph, Cancer **30**, 1025-1035, 1972.
W. G. Wee, M. Moskowitz, N. Chang, Y. Ting and S. Pemmeraju, Evaluation of mammographic calcifications using a computer program, Radiology **116**, 717-720, 1975.
F. Winsberg, M. Elkin, J. Macy Jr., V. Bordaz and W. Weymouth, Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis, Radiology **89**, 211-215, 1967.
R. L. A. Kirsch, G. Larsen and J. J. Thomas, Digital computerized mammographic image enhancement: a preliminary report, Conf. Proc. Biosigma 78, Paris 1978, pp. 251-256.

[3] See for example W. Spiesberger and M. Tasto, The automatic measurement of medical X-ray photographs, Philips tech. Rev. **35**, 170-180, 1975.

[4] A. Hoyer and M. Schlindwein, this issue, p. 298.

The analysis of structural abnormalities

The various kinds of microscopic cancerous growths show up in the mammogram in different ways. A solid cancer appears as a light circular area whereas a scirrhus cancer usually appears in the form of star-shaped light regions, with or without a solid centre.

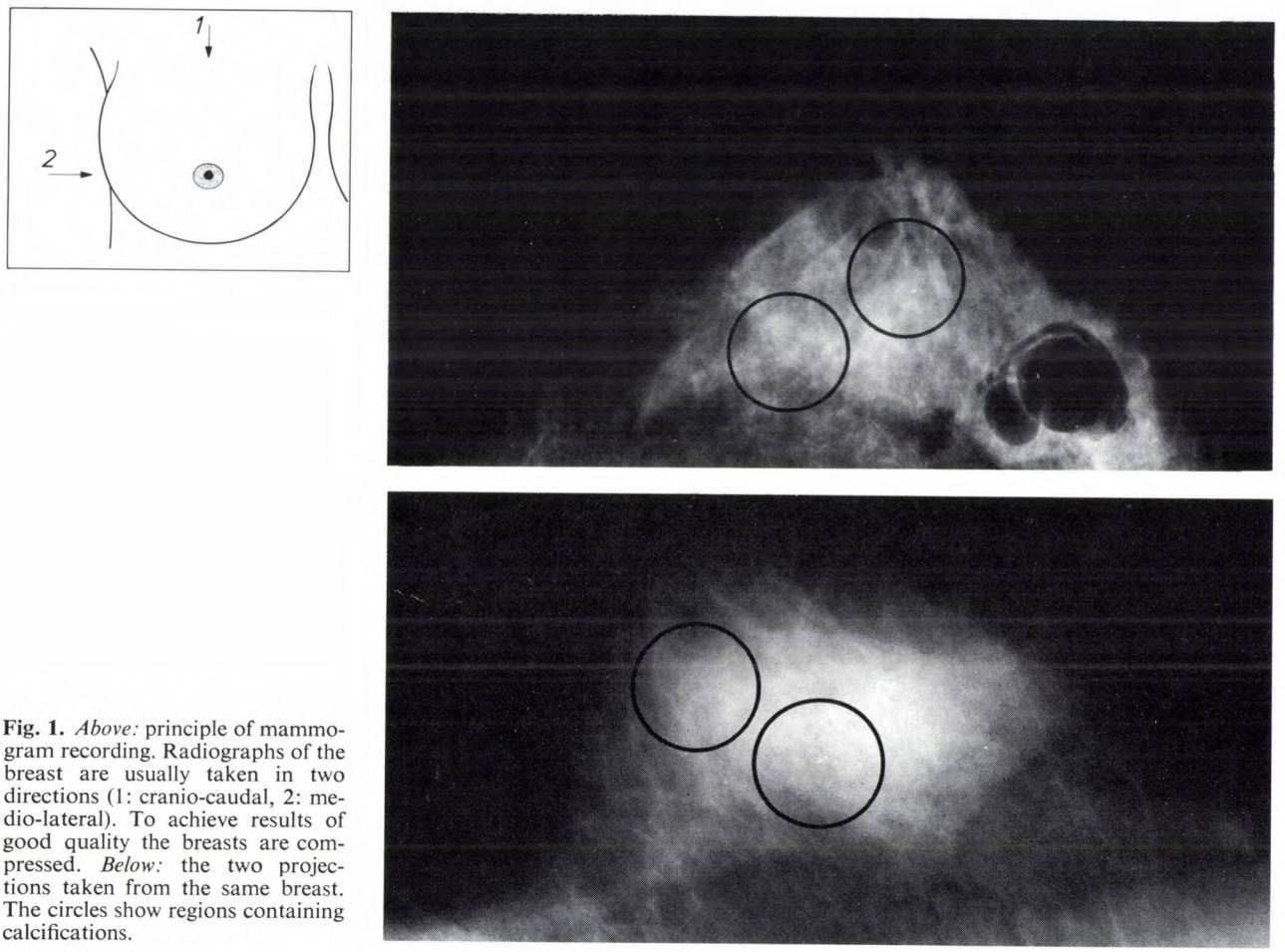


Fig. 1. *Above:* principle of mammogram recording. Radiographs of the breast are usually taken in two directions (1: cranio-caudal, 2: medio-lateral). To achieve results of good quality the breasts are compressed. *Below:* the two projections taken from the same breast. The circles show regions containing calcifications.

The structures vary widely from one mammogram to another, because the structure of the breast varies from woman to woman, and each mammogram is influenced by the exposure conditions. A radiologist, however, is able to recognize similarities in the features of these structures, and so to distinguish cancerous from normal tissue. He does this by making use of information about 'texture'. To give an example: two areas of the image may have the same average density, yet one area may contain a large number of small grey-level 'steps' whereas the other may contain a small number of large grey-level steps. We can quantify this information by determining the distribution of grey levels in the radiograph. This technique is called texture analysis and is well known in the analysis of aerial and satellite photographs [5]. The image is divided into segments and textural features are assigned to each segment. Regions of deviant texture can be found with the aid of pattern-recognition techniques. In mammography, however, the segmentation of the picture is more complicated than in aerial photographs.

The best way to find abnormalities is to compare each part of one of the breasts with the corresponding part of the other one, because normally these parts give the same texture in the radiograph. However, because the breasts are compressed when the mammograms are recorded, corresponding parts do not have symmetrical coordinates and cannot therefore be compared in a simple manner. We have therefore developed a special segmentation procedure. This procedure, which is related to that of C. Kimme *et al.* [6], who have worked on xerograms, consists of three steps. First, the breast contours are calculated from the image data. Secondly, the regions of the breast are segmented into strips; the number of strips is the same for both breasts, so that the width of the strips can be different. Thirdly,

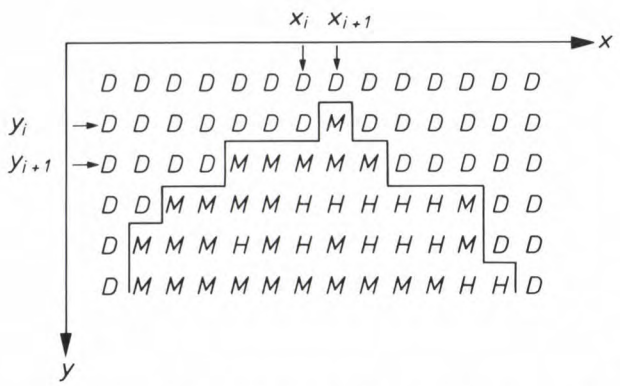


Fig. 2. Schematic presentation of the digitized mammogram. *D* background pixels (*dark*), *M* and *H* medium and high grey levels of breast-tissue pixels.

the strips thus defined are subdivided into rectangles of about the same size as the smallest radiographically visible cancer (1 cm). This procedure ensures that rectangles with the same position represent corresponding regions in both breasts.

of *CON* is at least 10 times as large as a threshold value *T*. This threshold value can be derived from measurements of the background in the mammogram (*D* in fig. 2).

The coordinates in the contour points can be found by scanning the mammogram line by line and retaining one preceding line in each scan; only two lines then have to be stored. To avoid ambiguity, only the first detected contour point of each line is stored in the memory. For the segmentation we take the vertical from the nipple to the breast wall as reference. The position of the nipple can be simply determined by finding the maximum of the breast-contour curve, and the breast wall is approximated by the edge of the image. Starting from the reference line, each half of the breast is subdivided into *N* strips; see *fig. 3*. These strips are then subdivided into rectangles, the number of rectangles depending on the position of the strip. If the number of strips (*N*) is equal to 10, then one mammogram is divided into 108 segments, as can be seen in *fig. 4*.

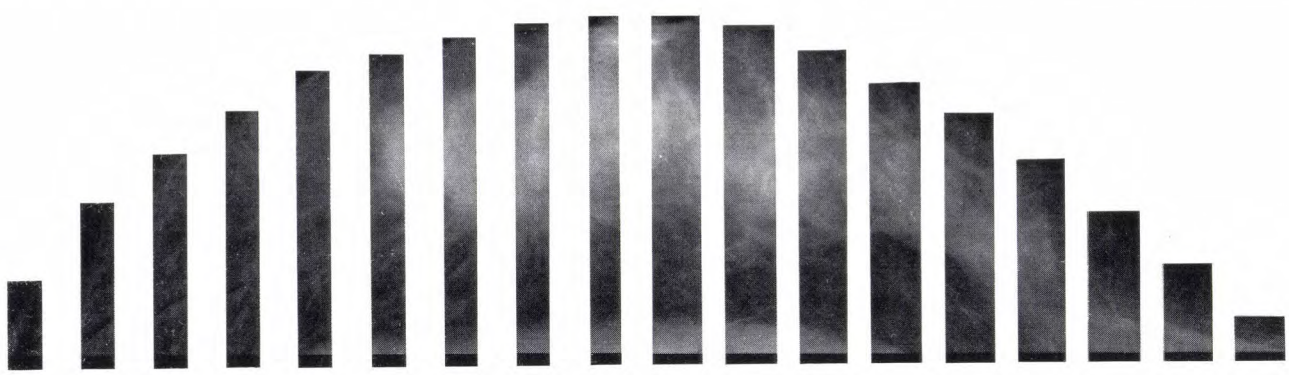


Fig. 3. After the determination of the breast contour each half of the breast is divided into a number of strips, beginning at the nipple. For further processing the strips are subdivided into rectangles, which can be compared with corresponding regions of the mammogram of the other breast (fig. 4). The part of the picture outside the contour is not used.

Determining breast contour and segments

In principle the breast contour can be found by means of a 'gradient method' [7]. A drawback of such methods is that they require a great deal of computer time and memory capacity. We have therefore developed a simpler method that completely answers our requirements. It is based on the large differences in grey level that are to be expected in the region of the breast contour both in the vertical and the horizontal direction. In particular the product of these differences will have a high value at points on the contour. With the notation of *fig. 2* we define

$$CON = \left|g(x_i,y_i) - g(x_{i+1}y_i)\right| \times \left|g(x_i,y_i) - g(x_i,y_{i+1})\right|.$$

We assume that (x_i,y_i) is a contour point when the value

Characteristic features of breast cancer

A comparison of corresponding regions in the mammograms of the two breasts should now show whether they have a similar texture or not. Various methods of describing texture are known from aerial-picture processing. To find out whether the features calculated with these methods are also relevant to our problem, we made use of the two best-known tech-

[5] R. M. Haralick, K. Shanmugam and I. Dinstein, Textural features for image classification, IEEE Trans. **SMC-3**, 610-621, 1973.
[6] C. Kimme, B. J. O'Loughlin and J. Sklansky, Automatic detection of suspicious abnormalities in breast radiographs, Proc. Conf. on Computer graphics, pattern recognition, and data structure, Los Angeles 1975, pp. 84-88.
[7] A. Rosenfeld and A. C. Kak, Digital picture processing, Academic Press, New York 1976.

niques and calculated textural features. The two methods are briefly described below.

The first is known as the run-length method [8]. A run is the number of adjacent pixels with the same grey level, 'adjacent' relating to all four image directions from each point, as illustrated in *fig. 5*. Thus, for each picture four run-length distributions are obtained. It is clear that a large number of long runs describes a very different texture from that represented by mainly short runs. In this way it is possible to derive textural features from the run-length data. A list of possible features is given in *Table I*.

In the second method characteristic features are derived from the grey-level co-occurrence matrix [5]. This matrix gives the distribution of pixel pairs with the same grey levels. Just as for the run-length distribution, the number of these pairs can be determined in four directions, so that four matrices are obtained from which textural features can be extracted. Statistical quantities such as *mean variance* and *higher-order* moments can be computed from the co-occurrence matrices. These quantities are called 'second-order statistical features' as opposed to first-order features, which are directly derived from the grey-level distribution. *Table II* gives a selection of features derived from the co-occurrence matrix as described in the article quoted above [5].

As a preliminary test of the usefulness of the textural features described, corresponding regions of both breasts of the same patient were scanned and quantized into eight grey levels; one of the breasts contained a scirrhus cancer. Since the textural features in this specific problem may not be orientation-dependent, we worked with the mean of the features computed from the matrices for the four directions. We soon found

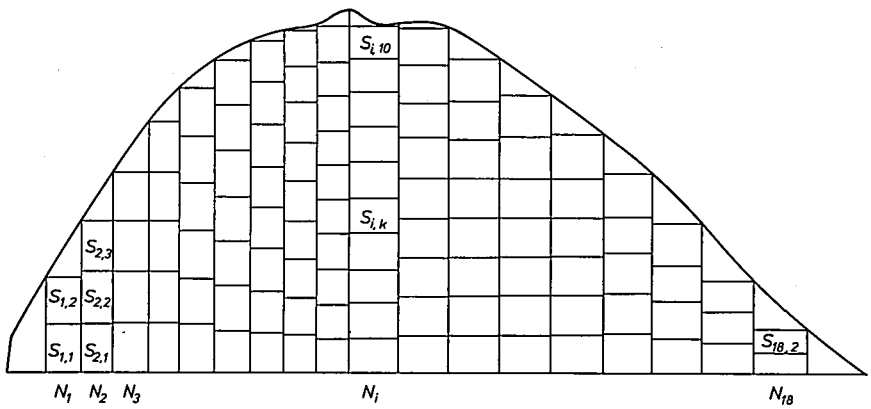


Fig. 4. Principle of the segmentation of a mammogram for texture analysis. The number of subregions per strip depends on the strip number. Storage of the number N_i of the strip and of the number S_i of the subregion gives access to corresponding regions in the mammograms of the other breast.

Table I. Some of the features that characterize the run-length distribution of a mammogram.

Definition	Feature
$RF1 = \sum_{i,j} \frac{p(i,j)}{j^2}/A$	short-runs emphasis
$RF2 = \sum_{i,j} j^2 p(i,j)/A$	long-runs emphasis
$RF3 = \sum_i (\sum_j (p(i,j))^2/A$	grey-level nonuniformity
$RF4 = \sum_j (\sum_i p(i,j))^2/A$	run-length nonuniformity
$RF5 = \sum_{i,j} p(i,j)/N^2$	run percentage

Notation:
$p(i,j)$ element of run-length matrix
$A = \sum_{i,j} p(i,j)$
N number of lines

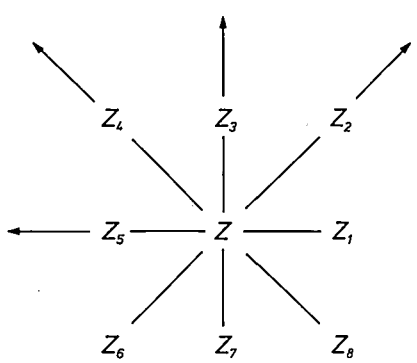


Fig. 5. The four directions (see arrows) for calculating grey-level distributions.

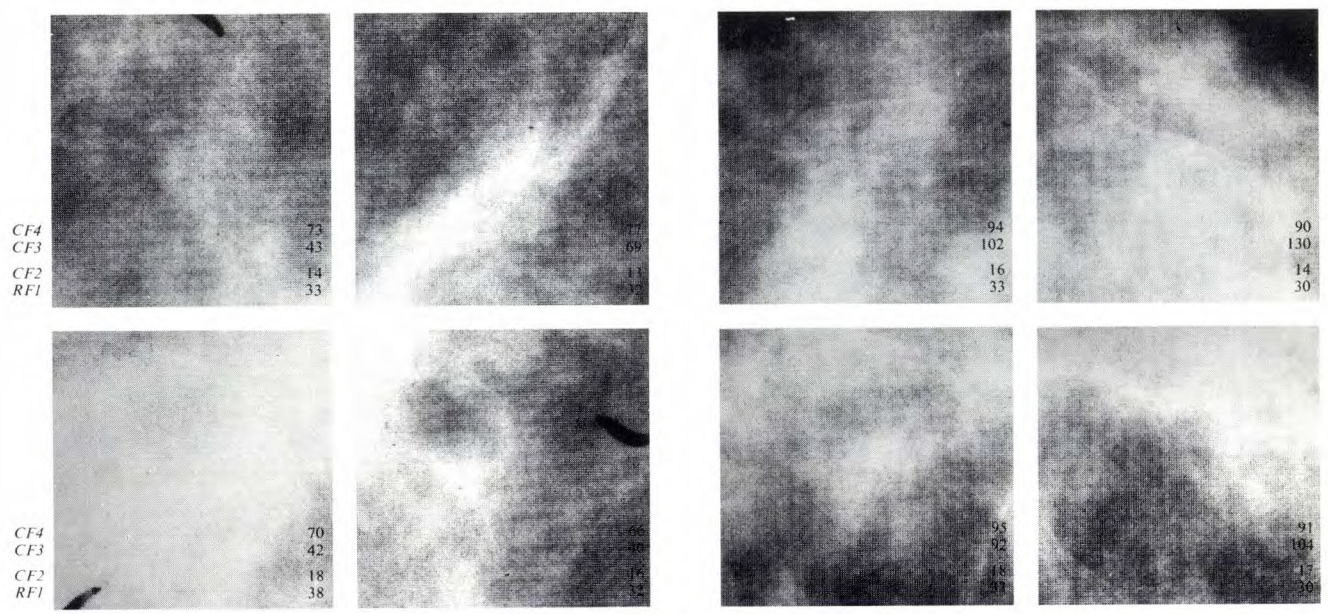


Fig. 6. Examples of structures in mammograms with values of selected textural features in corresponding areas. *Left:* cancerous tissue. *Right:* normal tissue. The features CF_4 and CF_3 (see Tables I and II) show significant differences and are therefore useful for this particular diagnostic problem, whereas CF_2 and RF_1 are virtually insensitive to the difference in texture.

that there was not much difference between the textural features of the two mammograms. The differences were much smaller than those we found with two test plates, which had a much more regular structure than is normally found in mammograms.

It thus appears that there are some features that, although they are sensitive to large differences in texture, are not useful for our purposes. There are other features, however, that do have sufficiently large differences. *Fig. 6* shows some examples of pairs of mammograms investigated and lists a few feature values. The usefulness of the features found in this preliminary investigation for detecting growths of the scirrhus cancer will now have to be proved with a large number of cases. For other types of cancer similar groups of features will have to be found and evaluated.

Detection of calcifications

The presence of microcalcifications can be a significant indication of the presence of an incipient cancerous growth. In 40 to 60% of all cases where microcalcifications were found by mammography, a carcinoma *in situ* or an 'occult' (i.e. not clinically detectable) cancer was detected after additional histological investigations. Since the calcifications are very small (with diameters ranging from 0.1 to 1.0 mm), conventional film mammograms have to be examined extremely carefully. Doctors generally do this with a magnifying glass. However, since it is very tiring work, and also because of the difficulties inherent in mass screening, it seems appropriate to use a computer with an algorithm that automatically detects calcifications.

Table II. Second-order statistical features.

Definition	Name
$CF1 = \sum_{i,j} (p'(i,j))^2$	angular second moment
$CF2 = \sum_{m=0}^{Ng^2} n^2 \sum_{i,j} p'(i,j)$	contrast
$CF3 = \dfrac{\sum_{i,j} ij\, p'(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$	correlation
$CF4 = -\sum_{i,j} p'(i,j) \log p'(i,j)$	entropy
$CF5 = \sum_{i,j} (1-\mu)^2 p'(i,j)$	variance

Notation:

$p(i,j)$ co-occurrence-matrix element
$A = \sum_{i,j} p(i,j)$
$p'(i,j) = p(i,j)/A$
Ng number of rows/columns of the matrix
μ_x, μ_y averages over columns and rows
σ_x, σ_y variances of columns and rows

[8] M. M. Galloway, Texture analysis using gray level run lengths, Computer Graph. Image Process. **4**, 172-179, 1975.

For this to be done, the mammograms must be digitized with sampling steps of 25 μm, or at the most 50 μm, to give about 65×10^6 pixels (or in the other case 17 million pixels).

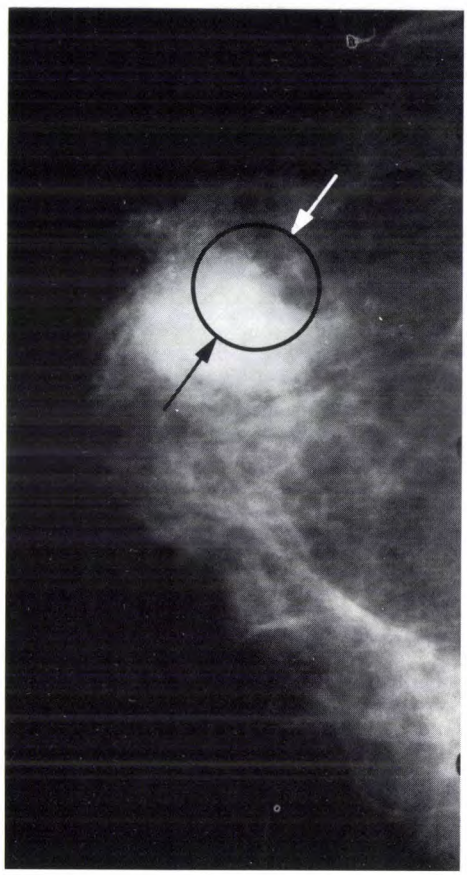


Fig. 7. Mammogram showing traces of calcifications. These appear as small bright spots (between the arrows). Clusters of calcifications, as in this case, are a serious indication of early breast cancer.

The segmentation problem that now confronts us is to 'cut out' small regions of interest from the mammograms and to use their image data for computing characteristic features. Unlike the 'static' segmentation technique that we used for the recognition of structural abnormalities, in which the segmentation was completed before calculations were started, we now apply 'dynamic' segmentation. A small window, measuring about 0.5 by 0.5 mm, which corresponds to 19×19 pixels at a spacing of 25 μm, is moved over the picture in such a way that each pixel in turn becomes the centre of the segment. The image information contained in each segment is collected and processed.

Since calcifications absorb X-rays much more strongly than the surrounding tissue, they appear as small bright spots in the image (*fig. 7*). This permits a simple preselection of areas of interest. If the central pixel of the window contains a local brightness maximum, we regard this area as a possible calcification. After this preselection only a few segments remain for which a more exact analysis must be performed, so that the amount of data for processing has been considerably reduced (*fig. 8*).

After the preselection it is necessary to verify whether there are really calcifications in the selected regions or whether the bright spot is due to some other cause. To do this we calculate certain mathematical features — not comparable with those that doctors employ in examinations — which we use in a *decision algorithm* for pattern recognition; see *fig. 9*. In our opinion this algorithm was preferable to the procedure of computing all the features first and then deciding whether or not a calcification is present. The structure of the algorithm reduces the number of regions with possible

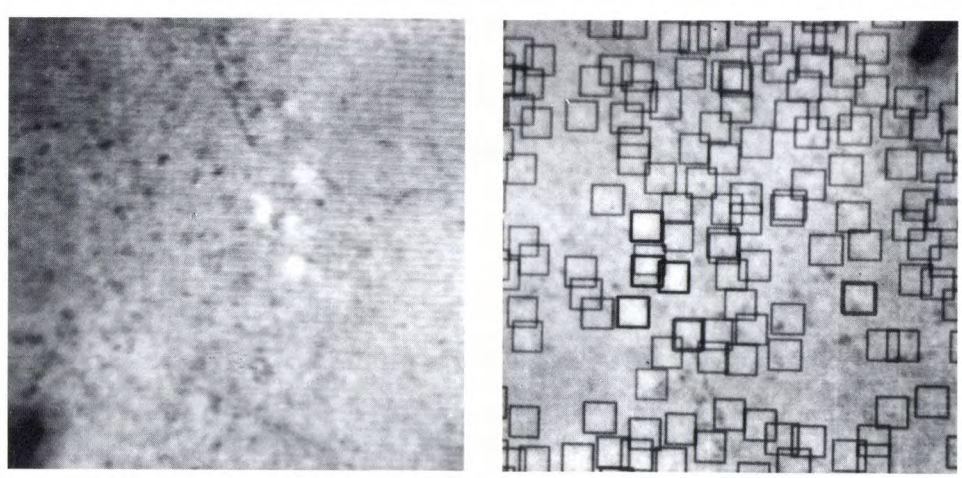


Fig. 8. The effect of preselection on the reduction of the amount of data in searching for calcifications. *Left:* unprocessed subarea of a mammogram. *Right:* processed subarea. The boxes enclose the regions that have been selected as possible sites of calcifications (local brightness peak in the centre). Only these regions are investigated further.

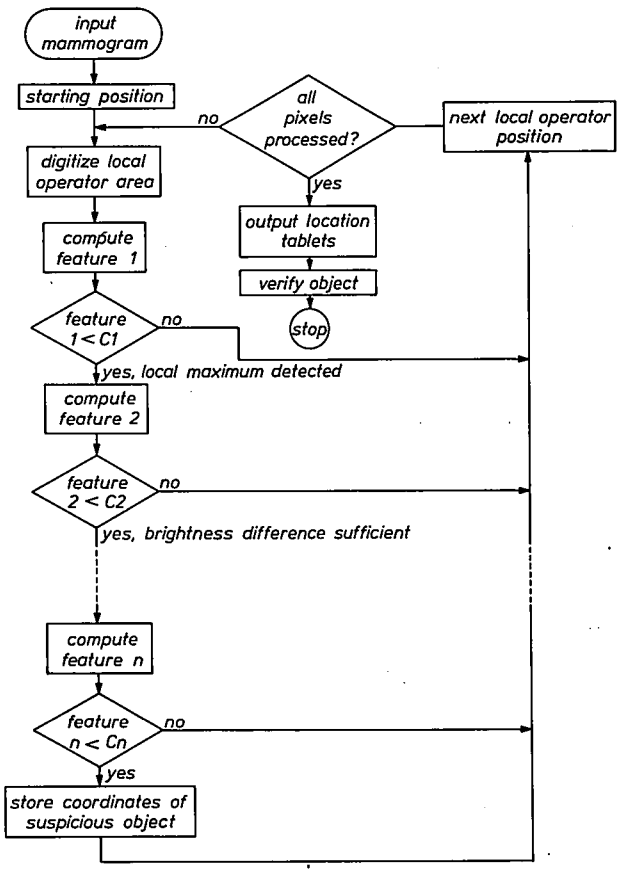


Fig. 9. Flow chart of the algorithm for calcification detection. The process has a tree structure; the different features are calculated sequentially. Only the regions for which all the previous features have given a positive result are processed further.

calcifications step by step, thus diminishing the computer time and memory capacity required.

By way of illustration we shall describe two of the mathematical features that are used as criteria. The first is the difference in brightness between the mean of the central area of the window — 3 × 3 pixels — and the mean of the boundary pixels of the window. If the

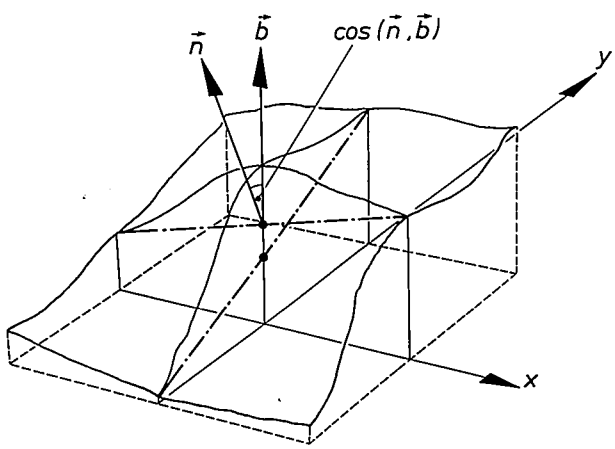


Fig. 10. The slope cos (n,b) of a plane that best fits the background can be used as a measure of the background fluctuations.

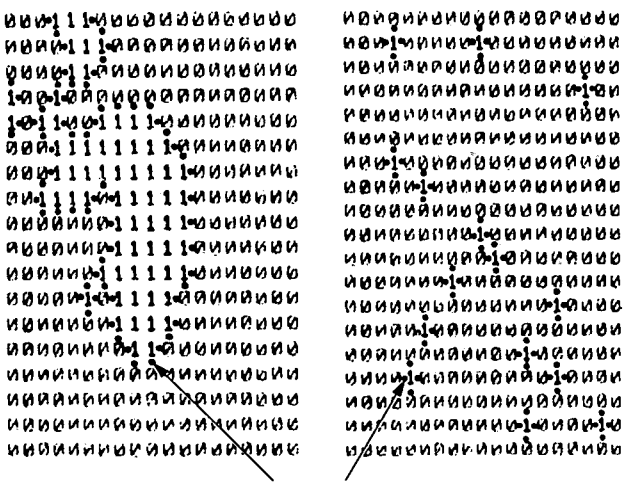


Fig. 11. Principle for calculating the measure of compactness $C = T/N$ for a compact (*left*) and a diffuse (*right*) object. T number of transitions (see the arrows) from '1' to '0'. N number of elements of value '1' (object). The regions inside the window have been converted into a binary code by establishing a threshold at a grey level midway between the average at the centre of the window and the average of the boundary points.

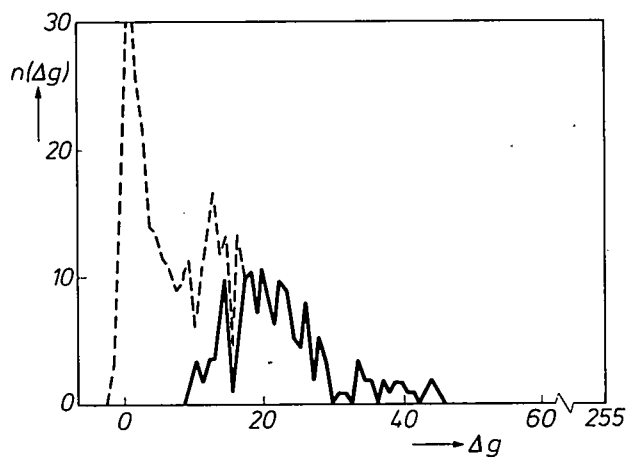


Fig. 12. Histogram of brightness differences Δg between the centre and the boundary of a window for all possible window positions (*dashed line*) and for calcifications alone (*solid line*). The large differences in the positions of the peaks show that object and background can be separated by establishing a threshold.

digitizing process has followed a logarithmic grey-level characteristic, this value gives the difference in optical density between a calcification and its environment. Acceptance or rejection of the segment as a pointer to the presence of calcification depends on whether the brightness exceeds a threshold value, which is based on the background noise. As a measure of the background we use both the standard deviation σ, calculated from the window boundary points, and the slope of a plane that best fits the background; see *fig. 10*.

The second decision criterion we use is the compactness of the bright object in the window segment.

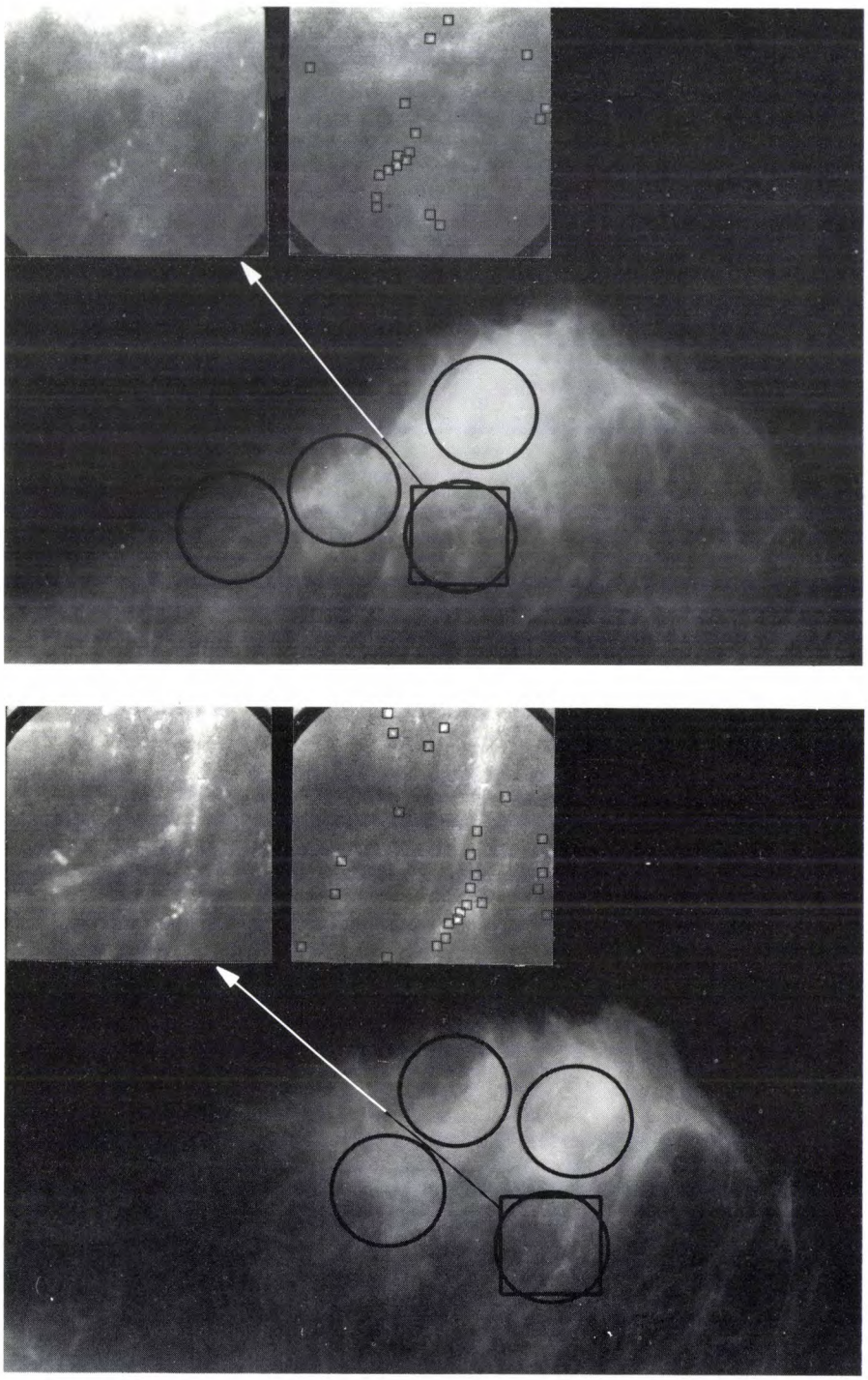


Fig. 13. Results of using the algorithm for the detection of calcifications in subareas of two mammograms of the same breast. *a*) cranio-caudal, *b*) medio-lateral. By way of illustration the subareas are shown enlarged before and after processing. The boxes enclose the location of the calcifications detected.

Calcifications are very compact, whereas noise appears as a more diffuse shape. To calculate a measure of compactness we convert the brightness data of the segment into binary values by establishing a threshold at a grey level midway between the central mean and the boundary mean (*fig. 11*). Elements of the object are assigned a '1' and background elements a '0'. The ratio $C\ (= T/N)$ of the number of transitions from '1' to '0' (T) to the number of object elements (N) is called the compactness. As can be calculated from fig. 11, diffuse objects are typified by values of C in the region of 4, whereas compact objects have C values between 0.5 and 1 [9].

The decision values for each feature are found by trial and error. In a learning process doctors were included as referees to judge the influence of combinations of features and decision values. To illustrate this

[9] M. Tasto, Automatische Mammographieauswertung: Erkennung von Mikroverkalkungen, Biomed. Technik **20**, Supplementary volume, pp. 273-274, 1975.

we show the effectiveness of the brightness difference in two histograms in *fig. 12*. The dashed line represents the brightness differences for central and boundary points in all windows (background and calcifications), and the solid line corresponds to the brightness differences in the windows of the calcifications alone. In the latter case the brightness differences are greater than the decision value, which was put at 2σ (σ is the standard deviation of the boundary points). As can be seen, there is a considerable difference between the main peaks, so that the brightness difference from this decision value is a useful quantity for the characterization of calcifications.

An example of the result of a search for calcifications with the aid of our detection algorithm is shown in *fig. 13*. It can be seen that most of the white spots are found. The method may be further improved by adding other features, e.g. length and orientation of the main axis or shape descriptors. This is important to minimize the number of false positive results, such as errors and scratches on the film, while at the same time increasing the number of calcifications found.

Recognition rate

The method described here for detecting calcifications was verified by processing 105 mammogram subareas of 1.63 cm^2, each consisting of 262 144 pixels. Each area was later examined by doctors, and the computer results were compared with their medical interpretation. The recognition rate turned out to be 68 % true positives (calcifications), 1 % false positives (objects that were not calcifications) and 31 % false negatives (calcifications that were not detected.)

In evaluating these results it must be borne in mind that we are talking about the detection of a single object. However, calcifications are only a strong indication of cancer if they occur in clusters, e.g. between five and ten in an area of one square centimetre. On the other hand, most undetected calcifications were in the mammograms containing a very large number of objects and from which it was obvious that an incipient cancer was present. Looked at objectively the results are promising. But to make the method suitable for use in mass screening it will be necessary to find ways of reducing the number of false negatives.

Since the results described above were first obtained, discussions in the medical world about the hazards of ionizing radation have introduced some uncertainty into the idea of mass screening for cancer by mammography. Before continuing our research we shall have to wait until the hazards of mammographic examinations are properly understood.

The work described in the foregoing was sponsored by the Federal German Ministry of Science and Technology, under contract No. DVM 135. The authors are responsible for the contents of this publication.

Summary. To support the diagnosis of breast cancer, computer programs have been developed that permit automatic detection of abnormalities in mammograms (breast radiographs). These programs can be important in mass screening for cancer detection, where very large numbers of mammograms have to be examined. Two different processing methods are used, one for large and one for very small structures. The detection of large cancerous structures, such as scirrhus cancer, is based on texture measurements, while the detection algorithm for microcalfications, which are the most important indications of early breast cancer, uses specific shape features. The calcification-detection procedure has been statistically tested with the cooperation of specialized radiologists. The results are promising and demonstrate the feasibility of automatic mammogram processing with a view to the detection of calcifications.

Philips tech. Rev. 38, 356-363, 1978/79, No. 11/12

A system that can learn to recognize two-dimensional shapes

E. H. J. Persoon

In spite of the great progress made in visual information processing during the last twenty years, there are still many 'tasks' in this field that present no difficulties to man but which a machine is as yet unable to carry out satisfactorily. One of these is the recognition of objects by their shape in two dimensions [1]. The objects are assumed to be flat or nearly flat, resting on a flat support, arranged at random, and possibly touching or even partly overlapping. A recognition task of this kind is important for the automation of assembly work. After recognizing components scattered indiscriminately on a conveyor belt, the machine would be able to pick them up and position them. In the inspection of assembly-line products, or for identifying and counting objects, it is also important to be able to recognize their shapes.

This article describes a recognition system that we have developed, which has two striking features. The first is that the system first 'learns' the shape to be recognized. This feature makes the system very flexible: within certain limits, it can be used for the automatic recognition of any arbitrary shape. Secondly, the system can still recognize an object when other objects touch it or even partly overlap it.

To make it clear where the difficulties lie in designing a system capable of recognizing objects that may be partly overlapped, let us briefly consider two methods of recognition with the aid of *fig. 1*. We shall assume that the rectangle in fig. 1*a* is stored in some way in the memory of the system; the system is asked whether the rectangle is present in the scenes (*b*) and (*c*).

In the first method the rectangle is translated and rotated over the scenes to see how well it 'fits'. In both (*b*) and (*c*) it is found that there is one position in which the rectangle fits very well, which 'implies' that it is present. With this method it is possible, in principle, to recognize partly covered objects; its implementation on a computer, however, takes too much time. Let us assume that the pictures are black and white and consist of 100 × 100 picture elements ('pixels'); each pixel can thus have the value 1 or 0 (i.e. be either black or white). A 'good fit' means that a certain function f of the position of the rectangle has a sharp minimum; f is, for example, the number of pixels that differ in value in the patterns compared. Now in the first place, calculating a new position, that is to say performing the transformations 'translation' and 'rotation' for every point of the rectangle or of its boundary, is in itself extremely cumbersome. Secondly, the number of possible positions is very large: broadly speaking there are 100 orientations for every one of 100 × 100 locations. An initial refinement might consist in giving priority to transformations for which f becomes smaller, but even this will often be unsuccessful because the rectangle arrives at a *relative* minimum of f and not at the *absolute* minimum desired.

The second method that should be mentioned here is much more effective, but it can only be used if the objects in the scene do not touch one another. In this method the coordinates of the pixels of each object

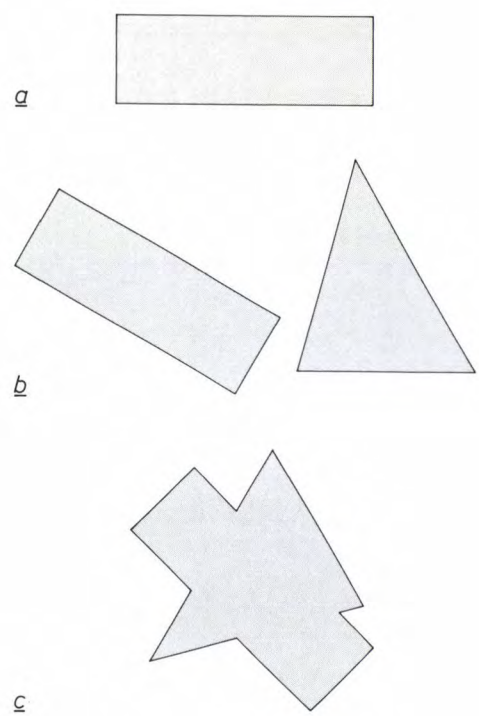


Fig. 1. Example of a recognition problem. The aim is to design systems capable of recognizing the rectangle (*a*) in the images (*b*) and (*c*).

Dr E. H. J. Persoon is with Philips Research Laboratories, Eindhoven.

in the scene are determined, for example by scanning the boundary [2], and from these certain shape features are calculated, such as the greatest distance from the centre of gravity to the boundary, or the 'circularity' (i.e. the ratio of the area to the square of the circumference — a ratio that is a maximum for a circle). The object to be identified is 'recognized' when sufficient agreement is found between its features and those of an object in the scene. If objects touch one another, this method cannot be used: there is no criterion for deciding where one object stops and another begins.

In our system an object in scenes like those of fig. 1*c* is primarily recognized by distinct 'local' features of the uncovered parts, e.g. the right angles *1*, *2*, *3* and *4* in *fig. 2*. In the following such local features will be called *shape elements*. It is obvious, however, that even when the system has established that the displayed scene has a number of such elements in common with the given object, this would still be only a weak criterion for recognition if at the same time the relative locations and orientations of the recognized elements were not taken into account. In our system a shape is therefore characterized in *two phases:* in the first phase a search is made for the shape elements, and in the second phase the relative positions of the elements are characterized. The same two phases recur in the analysis of a scene. First it is established whether the given shape elements are contained in the scene, and next whether they have the appropriate relative positions. This procedure makes it possible to characterize the shape of an object with relatively few elements. Moreover, the same elements often recur in the recognition of more than one shape, so that the memory capacity required does not increase in proportion to the number of shapes to be recognized.

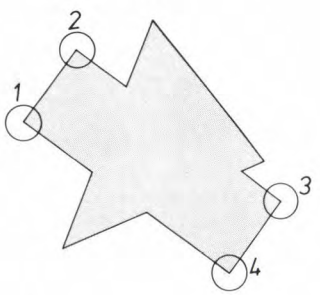


Fig. 2. The right angles *1*, *2*, *3* and *4* are salient 'local' features of the rectangle that can be used for the recognition, even though the rectangle is partly overlapped by a triangle.

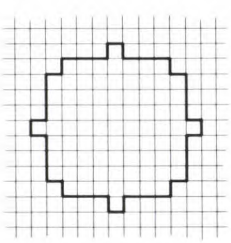


Fig. 3. The window used in searching for shape elements: it is a good approximation to a circle with a diameter of eleven pixels.

Searching for the shape elements and recognizing them

The shape elements we use are patterns 'cut out' from the image by the window of *fig. 3*. This window is a good approximation to a circle with a diameter of eleven pixels on the orthogonal grid. The choice of diameter is a compromise: if the window is made too small, it is no longer possible to discriminate between shapes, such as straight edges, sharp corners, right angles, and so on; if it is too large, the number of possible shape elements is too large. In conjunction with the measures described below, the window of fig. 3 is found to provide the desired degree of discrimination.

It is evident that not every pattern in this window can be used as a shape element: the window has 81 pixels and there are therefore 2^{81} possible patterns. The number of possible shape elements is drastically limited by only accepting a window pattern as a 'shape element' if *the centre of gravity of the boundary points of the complete picture as selected by the window is located in the centre of the window.*

To find such shape elements the screen is scanned from left to right and from top to bottom by the window. As soon as one or more black points appear in the window, the centre of gravity of the selected boundary points is calculated. The window is then displaced in such a way that this point arrives at the centre. Since this has the effect of changing the pattern in the window, the boundary after this step is not usually completely centred, but it is centred after one or two more centre-of-gravity calculations and translations. In all this the *boundary points* of a picture are taken to be the black and the white points that have a white and a black point respectively as a neighbour in the horizontal or vertical directions (see *fig. 4*).

[1] M. Yachida and S. Tsuji, A versatile machine vision system for complex industrial parts, IEEE Trans. **C-26**, 882-894, 1977.
W. A. Perkins, A model-based vision system for industrial parts, IEEE Trans. **C-27**, 126-143, 1978.
B. Neumann, Interpretation of imperfect object contours for identification and tracking, Proc. 4th Int. Joint Conf. on Pattern Recognition, Kyoto 1978, pp. 691-693.
[2] E. H. J. Persoon and K. S. Fu, Shape discrimination using Fourier descriptors, IEEE Trans. **SMC-7**, 170-179, 1977.
P. Saraga and J. A. Weaver, An experiment in flexible automation, this issue, pp. 329-337.

The coordinates x_c, y_c of the centre of gravity are calculated as follows from the matrix $D(x,y)$ of the boundary in the window (see fig. 4):

$$x_c = \frac{\Sigma D(x,y)\,x}{\Sigma D(x,y)}, \qquad y_c = \frac{\Sigma D(x,y)\,y}{\Sigma D(x,y)},$$

where the summations (Σ) extend over all window values of x and y. In fig. 4 we have $x_c = 0$, $y_c = 2\frac{1}{2}$. A boundary section is said to be centred when the distance from the centre of gravity to the centre of the window is smaller than or equal to $\frac{1}{2}$.

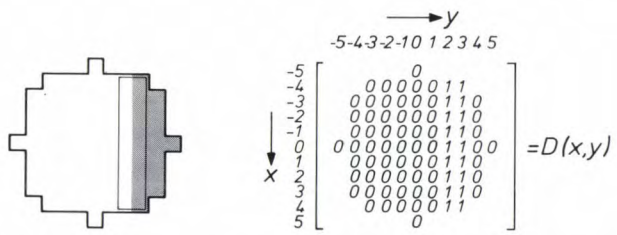


Fig. 4. Boundary points (*framed*) in a simple window pattern and the corresponding boundary matrix $D(x,y)$.

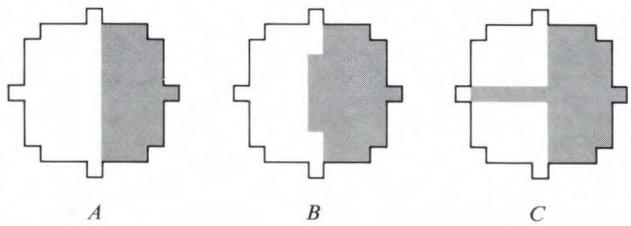


Fig. 5. Three window patterns; the conventional 'distance measure' d (the number of pixels that differ in value) between A and B is equal to that between A and C ($d = 5$). Measured by this criterion A and C are identical if A and B are taken as identical ('elements are identical if $d \leqslant 5$'); conversely, A and B are different if A and C are taken to be different ('elements are different if $d \geqslant 5$'). In the system used here, A and B are taken to be identical because the difference only relates to boundary points. This does not apply to A and C, which are thus 'different'.

Shape elements must be compared with each other to determine whether they are 'identical' or 'non-identical'. In comparing two pictures P and Q it is usual to count the pixels that have a different value in P from the value in Q. The number counted, $d(P,Q)$, is the conventional 'distance measure' between P and Q. The two pictures are regarded as non-identical if $d(P,Q)$ is larger than a given threshold value. This criterion does not however do justice to the following important boundary effect. In the binary pictures that we process the pixels are white or black depending on whether the video signal exceeds a given threshold or not. At a boundary the signal is close to the threshold. Small variations in the threshold, noise in the signal or variations in the sampling position in relation to the picture can easily cause the value of the boundary points to change. The patterns A and B in *fig. 5* may therefore easily originate from the same piece of a given scene. This is very unlikely for patterns A and C. Nevertheless, the conventional distance measure between A and B is equal to that between A and C: $d(A,B) = d(A,C) = 5$. We therefore use a different distance measure, $d'(P,Q)$. In this quantity, as in $d(P,Q)$, pixels are not counted if they have the same value in P and Q; however, pixels with different values are also *not counted if they are boundary points in both pictures.* Our criterion for the identity of two pictures P and Q is now: $d'(P,Q) = 0$. This is a more general criterion than the corresponding conventional one ($d = 0$): it allows differences in the values of boundary points. According to the new criterion, A and B in fig. 5 are identical, but A and C are non-identical: $d'(A,B) = 0$, $d'(A,C) = 4$.

Summarizing, what we understand by two *identical shape elements* are two window patterns in which the centre of gravity of boundary points is centred in the window and for which d' is equal to zero. During the learning process the shape elements found are stored in the memory if they differ from the elements previously stored. During the recognition process a check is made to see which elements in the memory are also present in the input scene. This, in brief, constitutes the first phase of learning and the first phase of recognition.

To make a further saving in memory capacity, a few simple computer operations are used for the identifica-

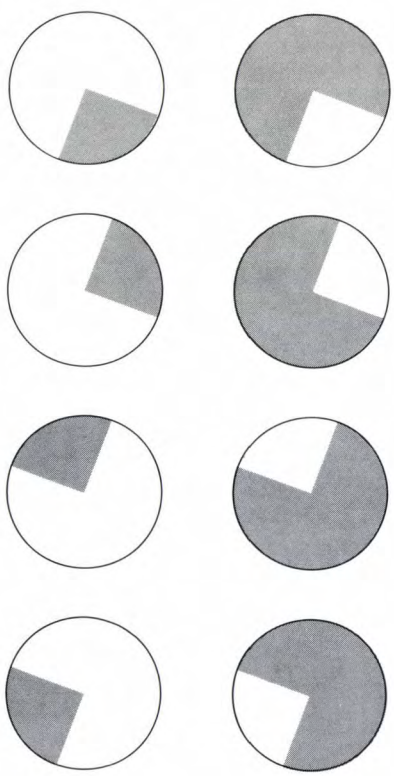
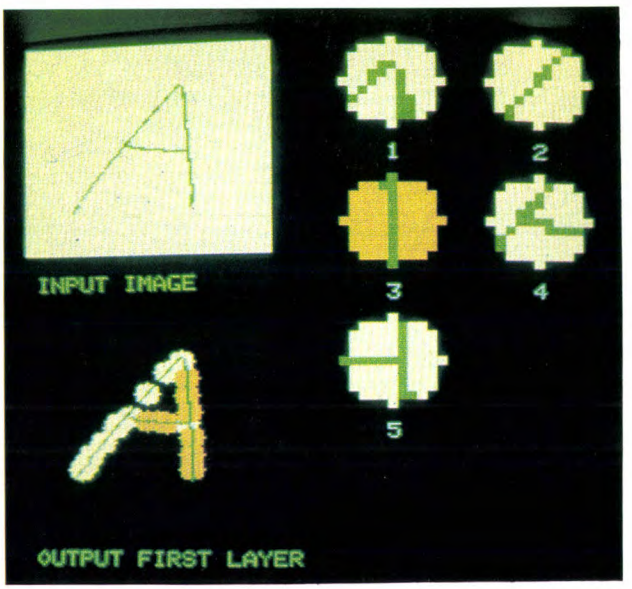
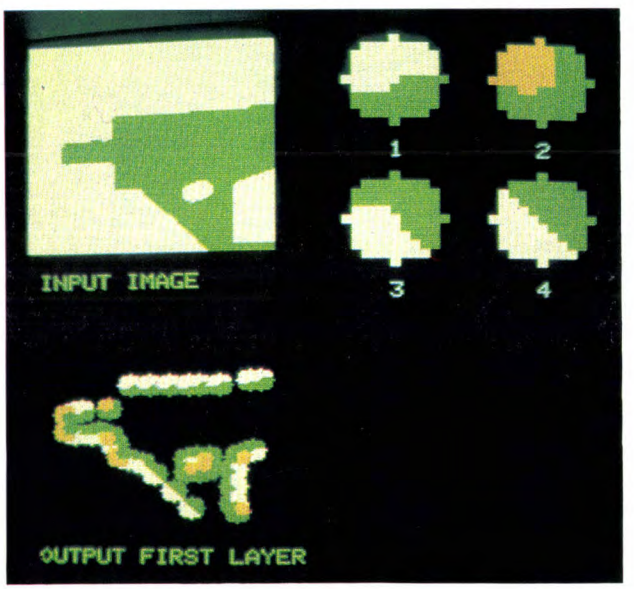


Fig. 6. A shape element (*upper left*) and the elements derived from it. The window is shown as a circle. Once the element at the upper left is stored in the memory, new elements are compared not only with this one but effectively with the other seven as well. This reduces the memory capacity required by a factor of almost eight. The centring of the boundary has not been taken into account in this figure.





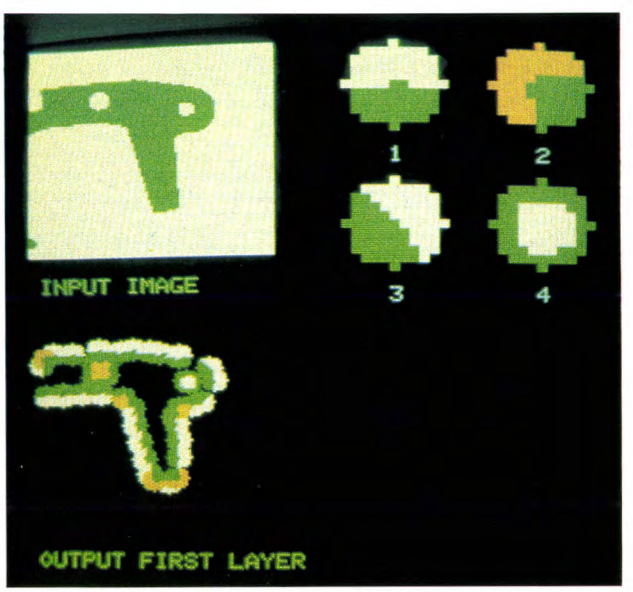


Fig. 7. Three analyses of a picture into shape elements. In each photograph the elements at the upper right were found in the picture at the upper left. The elements (11 pixels in diameter) are shown greatly magnified with respect to the picture (100 × 100 pixels). The display at the lower left shows the reconstruction of the picture with the elements found. A different background colour makes it easy to find element *3* in the first photograph and element *2* in the two others, seen directly, rotated through 90°, 180° or 270°, or negated or both.

tion of the elements. These are rotation of the elements through 90°, 180° and 270° and 'negation' of the pixel values (0 becomes 1 and 1 becomes 0). In the comparison these operations are applied to the stored element; a new element is therefore in fact not only compared with the stored elements themselves but also with the elements derived from them by these operations. For example, once the element at the upper left of *fig. 6* has been stored, it is also used for identifying the other seven in this figure. The measure reduces the memory capacity required by a factor of about 8. (In fig. 6 and also in figs 9 and 13 the window has been drawn for convenience as a circle.)

As can be seen from *fig. 7*, a surprisingly small number of elements are sufficient under these conditions, even for fairly complicated shapes. In each of the photographs, the binary picture that was displayed for learning on a screen of 100 × 100 pixels is shown at the upper left. The shape elements that were found are shown, greatly magnified, at the upper right; there are only five in the first and only four in the other two photographs. The picture at the lower left was obtained by displaying each element — at the original scale — at the locations and in the orientations and negation states in which it was found for the first time or recognized later. In the first photograph element *3* has been given a different colour, so that it can easily be identified, in both the vertical and the horizontal strokes. In the other two photographs element *2*, picked out in the same way, can be seen to appear directly, rotated or negated or both.

The elements found are not in general usable for recognizing the object if it has been rotated through an arbitrary angle in the scene, or 'turned over' to give the mirror image. For this the elements themselves would have to be rotated and mirrored. We decided against such complicated operations. Instead, in the first phase of the learning process we 'show' the system the object, right way up and turned over, in a large number of orientations in one quadrant. *Fig. 8* shows the shape elements of an object that were obtained in this way. Despite the large amount of input material, the number of elements remains fairly small, because of the limitation in the discrimination; usually only five orientations per quadrant can be distinguished upon rotation

of an element. Our procedure of not counting value differences of boundary points contributes significantly to the reduction in the number of elements. In a recognition attempt the number of elements that can be identified with the elements of fig. 8 is again almost eight times the number of elements in fig. 8.

The relative positions of the shape elements

Before considering the second phase of the learning and recognition processes, we should note that an object, as displayed in a scene, should be recognized by its shape *regardless of size*. One reason why this is necessary is to allow for the differences that can easily occur in optical adjustments when viewing the learning scene and the scene to be analysed. This has not been taken into account until now because it is of minor significance in the recognition of shape elements. For recognizing entire objects however, it has to be considered right from the beginning.

The second phase will be discussed by referring to the example given in *fig. 9*. The shape to be learnt is shown in fig. 9*a*, and the scene to be analysed in fig. 9*b*. As a result of the first phase of the learning process the elements *1*, *2*, *3*, *4* and *5* in fig. 9*c* have been stored in the memory. As a result of the first phase of the recognition process, element No. *2* has been identified at the points *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *K* in fig. 9*b*, either directly or in one of the rotated states *2a*, *2b* or *2c*.

We shall first look at the second phase of recognition and come back later to the second phase of learning because this involves recognition as well. Thus, we shall assume for the moment that the second phase of the learning process has been completed. The result of this is a number of 'learning lists' in which the positions of the shape elements of the rectangle have been recorded for each of the five orientations in fig. 9*d*. As we shall presently see, these do not have to be exactly the orientations at which the elements *1*, *2*, *3*, *4* and *5* have been learnt in the first phase. One such learning list, L_2, contains the shape elements of the rectangle in orientation *2* in fig. 9*d* and their positions:

L_2	shape element	...	2	2a	2b	2c	...
	position	...	P	Q	R	S	...

Besides the shape elements mentioned, L_2 contains straight-boundary elements that occur at a large number of positions if the rectangle is large. The memory contains in addition lists L_1, L_3, L_4, L_5 for the other orientations. If the object did not possess mirror symmetry, there would be five more lists for the mirror-image shape.

For recognition it must be possible to cover the rectangle in fig. 9*b* with the one in fig. 9*a*, and to see whether this can be done it is a straightforward matter to start from the orientation that has shape elements in common with the scene, i.e. orientation *2* in fig. 9*d*. However, to obtain coverage the rectangle will not only have to be translated and changed in size, it will usually also have to be rotated through a small angle. For although there is no distinction between the shape elements of the two rectangles, it is still possible, even after the best translation and change in size, for the positions to differ perceptibly because the angular resolution for complete objects is much sharper than for elements (1 in 50, i.e. about 1° for shapes of 50 pixels in length). However, because the rotation remains restricted to a small angle, the general calculation for rotations,

$$x' = x \cos \phi - y \sin \phi,$$
$$y' = x \sin \phi + y \cos \phi,$$

can be replaced by the simple transformation

$$x' = x - \phi y,$$
$$y' = y + \phi x.$$

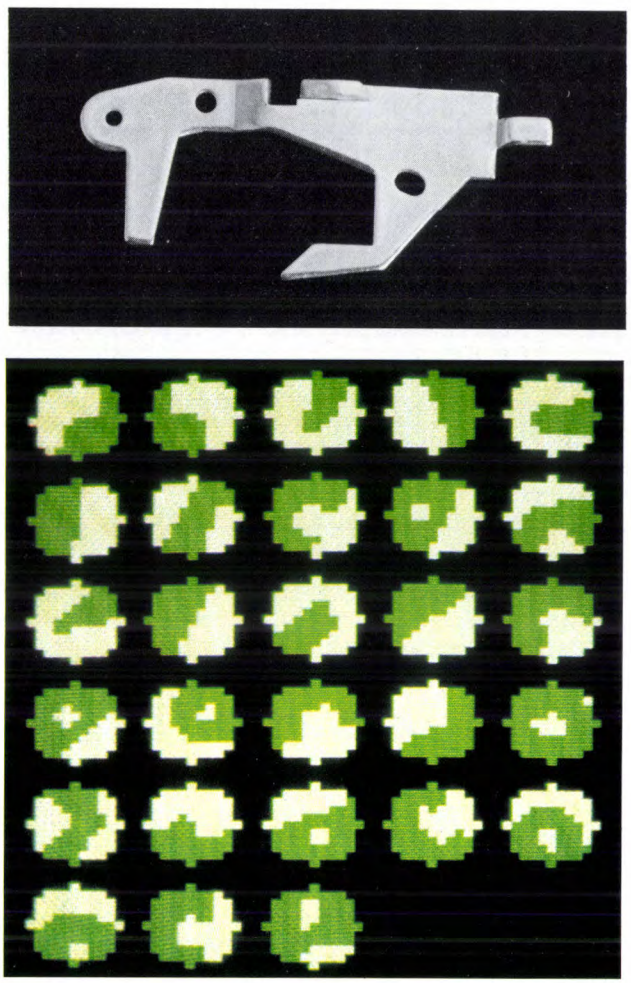


Fig. 8. *Above:* an object that has been presented for learning, both right way up and turned over, in a large number of orientations. *Below:* the shape elements found as a result.

The attempt at recognition in the example given now proceeds further as follows. In the analysis of the scene in fig. 9*b* the recognized shape elements are recorded in a 'scene list' L_s at the positions where they were found:

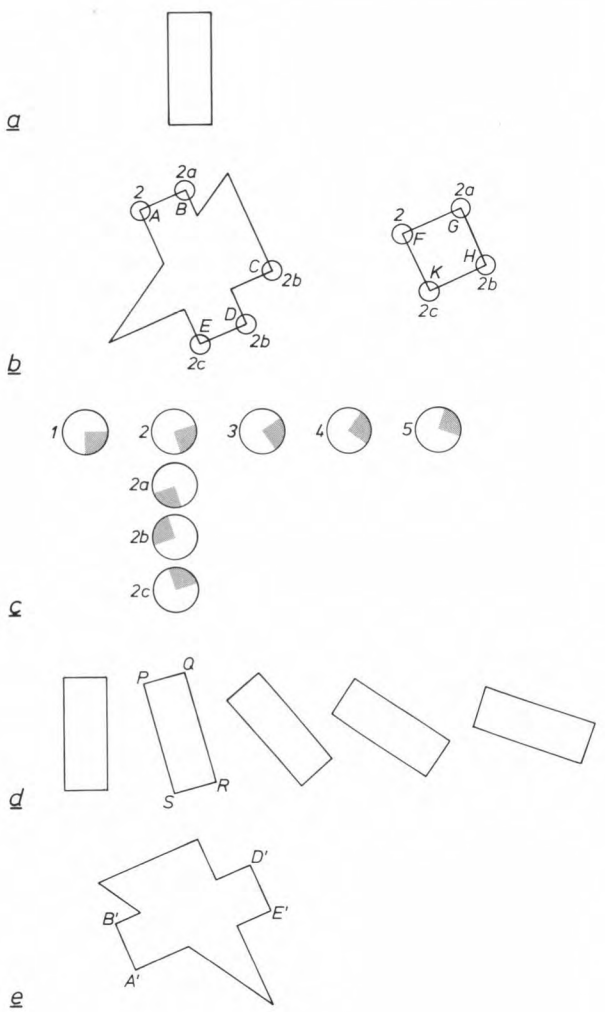


Fig. 9. The two phases of the learning and recognition processes, demonstrated by a simple example. *a*) The rectangle searched for. *b*) The scene to be analysed. *c*) Some of the shape elements of the rectangle that are stored in the memory during the first phase of the learning process. With the aid of No. *2* the elements *2a*, *2b* and *2c* can also be identified. In the scene (*b*) this has happened at *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H* and *K* (the first recognition phase). *d*) Orientations of the rectangle for which, in the second phase of the learning process, the elements found before are recorded with their positions in 'learning lists'. One such list is L_2, which contains the items '*2* at *P*', '*2a* at *Q*', . . In the second phase of the recognition process a 'scene list' L_s is made of the scene in (*b*) : . . , '2 at *A* and *F*', '*2a* at *B* and *G*', . . . The system establishes that L_s and L_2 have (at least) three elements in common (*2*, *2a* and *2b*), and searches for a translation, a rotation through a small angle and a change in scale that will transform the positions of these elements of L_2 into those of L_s. This does not succeed for the elements at *A*, *B* and *C* or at *A*, *B* and *H*, but it does for those at *A*, *B* and *D*. If the transformation found has the effect of matching at least half of all the elements of L_2 (including in this example a large number of straight boundary elements) with elements of L_s, the rectangle has been 'recognized'. *e*) A scene list is effectively also confronted with the learning lists rotated through 90°, 180° and 270°. In this way the rectangle *A'B'D'E* is also recognized through the rectangle in orientation *2* of (*d*).

L_s	position	. . .	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*K*	. . .
	element	. . .	*2*	*2a*	*2b*	*2b*	*2c*	*2*	*2a*	*2b*	*2c*	. . .

The system looks for elements that are common to L_s and the learning lists. In our example L_s and L_2 appear to have a number of elements in common. From these elements the system chooses three at particular locations in the scene, e.g. *2* at *A*, *2a* at *B* and *2b* at *C*. It checks to see whether the corresponding elements of L_2 can be made to cover them by a translation, a rotation and a change in size of the rectangle. This does not succeed for the elements of L_s at *A*, *B* and *C*; a trial with *A*, *B* and *H* also fails, but a trial with *A*, *B* and *D* does succeed. Our criterion for 'coverage' of two positions with the coordinates x,y and x',y' is that $|x - x'|$ and $|y - y'|$ should both be smaller than 2. After this success, the transformation thus found is applied to all the elements of L_2. If the percentage of these elements that then covers one of the elements of L_s is greater than 50%, the system gives the signal 'recognized'. The elements of L_s that are covered in this way in the example are the right-angle elements at *A*, *B*, *D*, *E* and a large number of straight-boundary elements.

For determining the required transformation one group of three elements from L_2 is more suitable than the other. The right-angle elements, for example, are more suitable than straight-boundary elements because they are more characteristic. This is not the place, however, to go further into the choice of a suitable group of three.

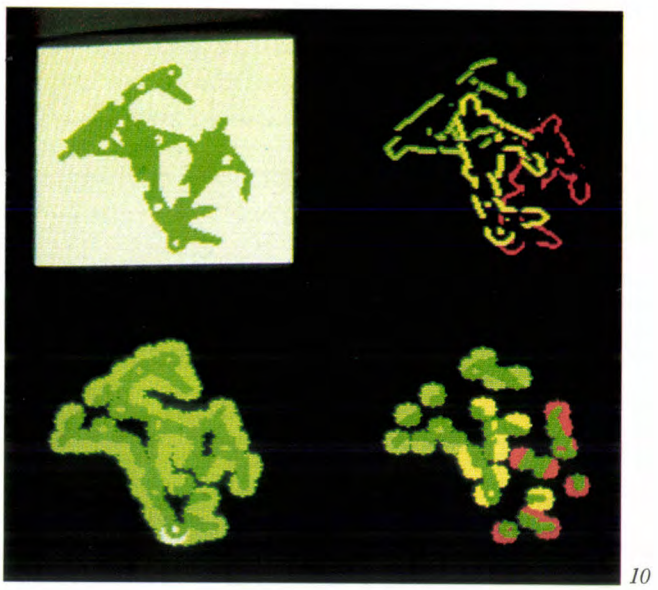
It remains to describe how the learning lists are obtained. The object to be learnt is at first shown in one orientation. Shape elements are searched for and the elements found are compared with the set found in the first phase; the elements that are recognized are recorded with the appropriate coordinates. This yields the first list (L_1). To save memory capacity and computer time the list is restricted to a reasonable number, e.g. 15 elements, evenly distributed around the periphery. Next, the object is rotated through about 2°, and a check is made to see whether it can be recognized from list L_1. If it can, the object is then rotated further. For the first orientation in which the system fails to recognize the object, a new list is made. Usually this happens after about 20°, because the shape elements themselves are no longer recognized. In one quadrant five lists are usually obtained in this way, and another five for the mirror image.

As regards orientation the rectangles in fig. 9*d* lie in the same quadrant as the rectangle of fig. 9*b*. For the recognition of shapes in the other quadrants a scene is effectively confronted not only with the ten lists mentioned, but also if necessary with the 'lists rotated

through 90°, 180° and 270°', which are obtained from the original ones by a short calculation during the confrontation. In this way the rectangle in fig. 9*e* would also be recognized through the rectangle in the orientation *2* in fig. 9*d*. This measure not only saves memory capacity but also shortens the learning time of the second phase.

Fig. 10 gives a more realistic example of a recognition problem. The system has learnt the shape of fig. 8 and is instructed to analyse the picture shown at the upper left. In the first learning phase the shape elements of fig. 8 are stored. In the second phase a number of these elements and a number of the elements derived from them are recorded with the appropriate coordinates in ten learning lists. *Fig. 11* shows a representation in each quadrant of one of the lists; the shapes consist of 15 of the 28 elements of fig. 8. The picture at the lower left in fig. 10 is a representation of the scene list L_S, where the elements of the scene that were identified with elements in the memory are reproduced at the location of identification. At the lower right are the elements of L_s that, after the required transformations of a list based on three elements had been found and applied, were also found to correspond in terms of position to an element of a learning list. Each background colour corresponds to one list. Finally, a visual representation is given at the upper right of the three lists that have been found, after the transformations. It is clear that the system has recognized the three objects in the original scene.

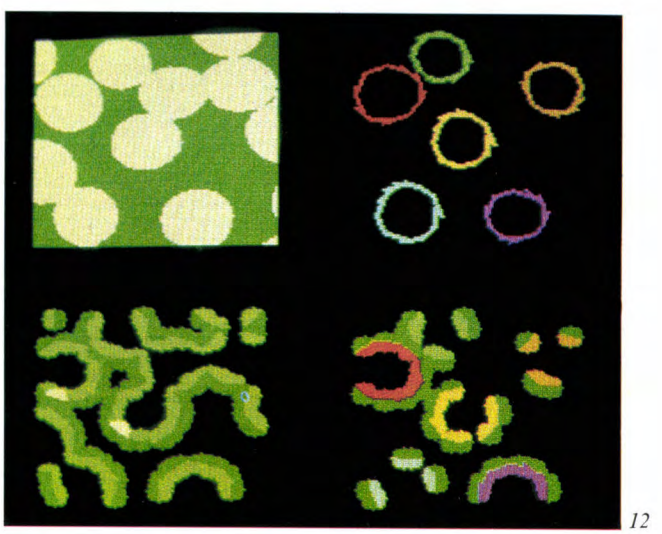
For simple shapes much less computer time may be required than for complex ones. A disc, for example, need only be shown in one orientation and thus yields only one learning list. *Fig. 12*, which is built up in the same way as fig. 10, gives an analysis of a scene containing small plastic balls; the analysis is based on this one learning list.



10



11



12

Fig. 10. Searching for the object of fig. 8 in a scene. *Upper left:* the scene to be analysed. *Lower left:* a visual representation of the scene list L_S: the elements of fig. 8 that have been recognized in the scene, at the location of the recognition. *Lower right:* the elements of L_S that have been covered by the elements of a learning list by means of a transformation. Each background colour corresponds to one learning list. *Upper right:* visual representation of the recognized lists, after the transformations.

Fig. 11. Visual representation in four quadrants of one learning list (15 elements) of the object in fig. 8.

Fig. 12. Searching for discs and counting them. The arrangement of the figure is the same as in fig. 10. Because of the simple shape of the object searched for, there is only one learning list.

Final comments; further research

In the foregoing only a broad outline of the method of operation of the system has been given; many details have been omitted. For example, the elements found in the first phase of learning are tested for the possibility of retrieval before they are stored in the memory. If the element in *fig. 13a* is found 'by chance' in the search process, then the probability of retrieval is small, because centring leads to a different element when it is approached somewhat differently (fig. 13*b* and *c*). With such checks the first phase yields a series of elements that forms suitable starting material for the second phase. Another detail should just be mentioned. To prevent the system from marginally failing to recognize stored elements when it is analysing a scene, the window is centred in turn on each of the eight pixels surrounding each pixel at which an element was found but not recognized, until recognition takes place or all eight have been examined [3].

Further research is being directed primarily at shortening the processing time. Hardware suitable for specific tasks must be developed which can carry out several processing operations in parallel. This has already met with success for the first phase. We are also investigating the feasibility of applying the method discussed in the foregoing to pictures that are not just black and white but also contain grey levels.

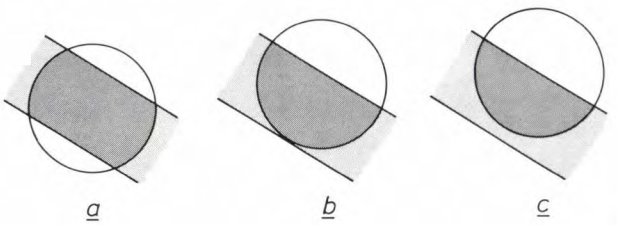


Fig. 13. Retrievability of shape elements. If the shape element (*a*) has been found 'by chance' during the learning process, it is hardly ever found again later because the centring of the boundary leads to a different element (*c*) if the same piece of scene is approached somewhat differently (*b*). Shape elements that are as difficult to retrieve as the one in this example are not stored in the memory.

[3] For further details see E. H. J. Persoon, Principles for self-organisation and their application to picture recognition, Proc. 4th Int. Joint Conf. on Pattern Recognition, Kyoto 1978, pp. 379-383.

Summary. A system has been designed for the automatic recognition of objects by their shapes in two dimensions. The process of 'learning' an object and the process of recognizing it in a displayed scene are both carried out in two phases. In the first phase a search is made for 'shape elements', i.e. patterns in a small window (with a diameter of 11 pixels) such that the centre of gravity of the selected boundary points is situated in the centre of the window. Two shape elements for which only some of the boundary points differ in value are regarded as 'identical'. In the learning process the (flat) object is shown to the system both right way up and turned over in a large number of orientations. The small dimensions of the window limit the discrimination between shape elements and so restrict their number, which is not very large even with complex shapes and in spite of the large amount of input material. In the second learning phase the system establishes the relative positions of the elements in a maximum of ten 'learning lists', five of them for five orientations of the object in each quadrant and the same number for the mirror image. In the recognition process a 'scene list' is made of shape elements that have been recognized in the first phase, with their positions. If about half or more of the elements of one learning list can be made to cover the corresponding elements of the scene list by transformation, the object has been recognized. The method can also be used for recognizing objects even if other objects touch or partly overlap them.

Philips tech. Rev. 38, 364-371, 1978/79, No. 11/12

Self-organizing systems

T. J. B. Swanenburg

Introduction

Modern information-processing equipment is extremely successful at carrying out tasks that are extensive, yet simple in principle. An example is classification of the data in a data bank in accordance with certain characteristics. For such tasks the programmer can provide a complete and unambiguous instruction (an algorithm) that ensures that the task will be carried out correctly. In pattern recognition — a fascinating part of the broad field of visual information processing — the situation is completely different. The automatic recognition of characters from a single simple fount is quite feasible. However, the automatic recognition of hand-written characters, or of other rather less sharply defined characters or patterns, easy though it may be for people, is in fact such a complicated task that it cannot usually be broken down into a set of algorithms or logical operations that can be carried out by an electronic system[1]. Also, patterns can often be so complex that, even in the cases where it does seem possible to specify such operations, not all the possibilities can in fact be envisaged, and errors are made as a result.

The same problems are encountered whenever highly complex information has to be processed automatically. In this article an attempt will be made to free us from having to specify every step in the information processing in advance. This means that we have to look for a small number of simple algorithms that imply 'self-organization', which means that the system develops an internal configuration relevant to the environment in which it has to operate. It cannot be taken for granted that such algorithms exist. However, the study of biological systems does provide the hope that they do. A living organism clearly seems to be 'self-organizing': it grows into an environmentally adapted system that is much more complex than the 'algorithms' that govern its growth — in this case the hereditary information stored in DNA molecules[2,3].

In this article the principles of self-organization will be discussed first; the essential difference between 'self-organizing' and 'adaptive' systems will also be indicated. A simple self-organizing pattern-recognition system and a number of computer experiments made to test the principles used will then be described. Possible applications of self-organization in other areas of information processing will be mentioned in the final section.

Dr T. J. B. Swanenburg is with Philips Research Laboratories, Eindhoven.

Self-organization

In designing a system it is usual to begin with a clearly defined objective. Also, so that the system can be improved once it has been built, it is necessary to be able to establish its quality, and this requires some sort of measure for the deviation from the objective. In designing a system for processing highly complex information difficulties arise with these two points right from the start. It does not yet seem possible to give an adequate description of a machine that would be able to recognize three-dimensional objects, for example, or minor tissue anomalies in X-ray pictures of the thorax. Even if such a machine is built, it is very difficult to determine its quality. All that can be done is to try to evaluate the results on the basis of more or less fortuitous examples, and then modify the system in accordance with the conclusions. But since the objective cannot be properly translated into algorithms, there is no certainty that these modifications will in fact turn out to be improvements in later examples.

In some cases a clear definition of the objective may be possible but it proves difficult to analyse it in terms of algorithms. The answer may then be an *adaptive system*. A number of parameters that affect the results are given the opportunity to adapt themselves to the objective. In fact the system then has to look for one or more maxima in a space whose dimensionality is given by the number of parameters. The volume of this space, and hence in general the time necessary for finding these maxima, will increase exponentially with the number of parameters. The direct conclusion is that this approach is not feasible with extremely complex systems, since convergence would take far too long[4]. What is more, the system can never compare the quality of its earlier forms of organization with its present form, because it has no memory: in searching for optimum parameter values it does not retain the earlier values. So this path leads to a dead end.

Self-organizing systems seem to offer better prospects. It has already been noted that living organisms are clearly 'self-organizing'. Let us now take a closer look at this concept by considering two other examples of self-organization from Nature.

The first example concerns the evolution of a biological species. In the first stage of the evolutionary process, as it is now understood, a particular species never occurs in only one form; small differences between individuals reflect small differences in genetic information. This genetic information is the *memory* of the species; it is assumed that the information common to all individuals is the oldest and in the past offered considerable chances of survival. The small differences are due to *mutations*, probably random, which give the species the opportunity to evolve towards a form of organization that offers even better chances of survival. Most of the mutations do not imply an improvement; the individual has a smaller chance of survival. If the species has enough individuals, however, this mechanism will generally lead to a further evolution of the species, owing to the rare mutations that do imply an improvement[2].

The second example concerns the operation of the human brain in the storage and processing of information received through the senses. It is unlikely that all the information received can be stored; but even if it could, the knowledge gathered would be pointless, for the probability that a situation once recorded will occur again in exactly the same form is zero. The only conclusion is that the brain forms *concepts* (abstractions, generalizations, '*Gestalten*') from the incoming flow of information, and that these concepts make it possible to 'recognize' later situations that are similar but not exactly the same. The expansion of the stored information — the formation of concepts of increasing complexity in an increasingly more complex network of neurons — closely resembles the evolutionary process outlined above. It is evident, particularly for that part of our knowledge that we are consciously aware of, that we are constantly engaged in forming the simplest possible picture of the phenomena around us. By forming more complex concepts we try to arrive as quickly as possible at an interpretation of what we see or hear. These complex concepts are probably brought about by means of more or less arbitrary combinations of a number of simpler ones, and by a conscious or unconscious selection on the basis of their suitability for giving a simpler description of the world around us.

Both examples might be characterized by 'algorithms' primarily involving the *preservation of information* (virtually exact reproduction; memory) and *fortuitous mutations* (the formation of new concepts). If we wanted to point to an 'objective', for example maximizing the chance of survival or finding an adequate description of our environment — *no such objective is explicitly present in the algorithms of the system.* What ultimately happens or is formed comes about through the interaction of the system with its environment. Species that survive in the Arctic Ocean and those that survive in a tropical forest are bound to be different, i.e. to have different sets of specific algorithms. Again, working with horses results in a different 'adequate' store of knowledge from that gained in working with engines.

These examples suggest the following characterization of a self-organizing system. To begin with, there must be sufficient latitude to allow the formation of *complex structures.* For the evolution of the species this latitude is found in the world of organic chemistry, with its unlimited possibilities; for the brain it is found in the neurons, which are capable in principle of forming networks of virtually unlimited complexity. On the one hand a *memory* must ensure that structures once formed will be preserved for some time; on the other hand it must be possible for new structures to emerge from old ones through *mutations.* These cannot be other than random, because there is no explicitly defined objective. The structures that are finally formed are determined by the algorithms of the system and by *interaction with the environment.* Expressed in somewhat anthropomorphic terms: 'the system is constantly trying out new structures in its environment'. Its memory enables it to try out a number of possibilities side by side; it possesses structures that have so far been the best, plus a number of variations on them, some of which may prove to be better.

If the latitude — for example the total number of structural elements or the memory capacity — is limited, the less-successful variants have to be discarded to maintain the possibility of continuously forming new variants (*selection*). It should be noted that this also leads to faster evolution, since the old information is no longer fully available for trying out new forms. On the other hand it also means that the best form of organization is not always found: the decay of information leads to the exclusion of a number of forms[2,5].

[1] R. O. Duda and P. E. Hart, Pattern classification and scene analysis, Wiley, New York 1973.
[2] M. Eigen, Self-organization of matter and the evolution of biological macromolecules, Naturwiss. **58**, 465-523, 1971.
[3] G. J. Chaitin, Information-theoretic computational complexity, IEEE Trans. **IT-20**, 10-15, 1974.
[4] M. Minsky and S. Papert, Perceptrons, M.I.T. Press, Cambridge (Mass.) 1969.
[5] M. Eigen and P. Schuster, The hypercycle, a principle of natural self-organization, Naturwiss. **64**, 541-565, 1977, and **65**, 7-41 and 341-369, 1978.
M. Eigen and R. Winkler-Oswatitsch, Das Spiel, Piper-Verlag, München 1975.

The main points of difference compared with an adaptive system are the memory and the lack of an objective explicitly defined in the algorithms.

A self-organizing system for pattern recognition

A simple pattern-recognition system designed in accordance with the principles explained above will now be discussed. The system is designed for the recognition of binary patterns on a lattice of 8 × 8 picture elements (pixels). At first sight such patterns may perhaps look very simple. In fact, however, there are 2^{64} of them (about 10^{19}). Here again, the probability of an exact correspondence between two random patterns is therefore virtually zero.

Let us consider for a moment the concept of 'recognition'. Usually it is associated with the identification of previously named objects; the concept as it is used here, however, has a somewhat different connotation. Most readers will 'recognize' something in the patterns shown in *fig. 1a*: they will see a vertical black line against a white background. This is because 'vertical straight line' is a clear-cut concept in the human mind. The points that do not correspond to the description are rejected as irrelevant, as noise. Recognition in this sense is the ability to interpret, with the aid of concepts stored in the memory, an input signal never before perceived.

This is the kind of recognition which our system is capable of performing. The idea will be explained by reference to the class of patterns to which the pictures in fig. 1*a* belong: the patterns produced by the generation of a white field with a vertical black bar superimposed at an arbitrary position, with the pixel value inverted at four arbitrary locations (i.e. noise). The number of patterns in this class is about 5×10^6. The system is presented with arbitrary patterns from this class. After it has seen a small proportion of the total number of possible patterns, it has 'discovered' the structure: it has formed in its memory the concept of 'white field' and eight concepts corresponding to 'vertical black bar', even though this configuration has seldom if ever occurred in a pure form. (The eight vertical bars are eight *different* concepts for the system.) The recognition or interpretation of a new pattern in this class now takes place as shown in *fig. 2*: it is reconstructed with the aid of the concept 'white field', one of the eight concepts 'vertical bar' and four noise pixels. This is what will be meant in the rest of this article by the term 'recognition': the reconstruction of an input pattern with the aid of existing patterns in the memory.

The three patterns shown in fig. 1*b* belong to a series that contains just as much information (structure)

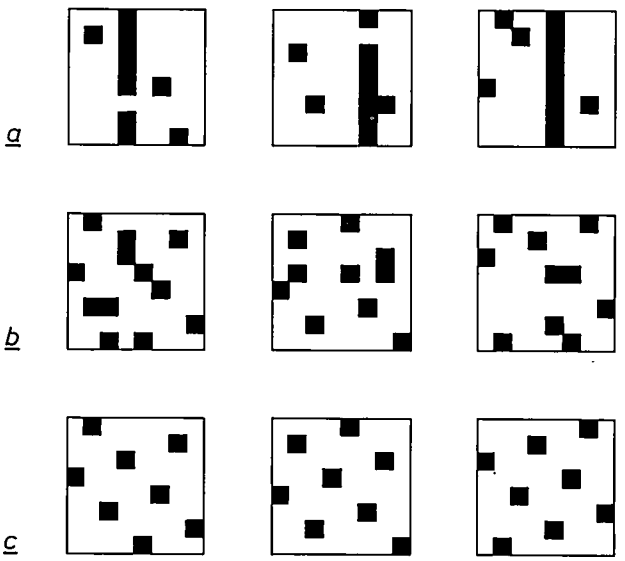


Fig. 1. Binary patterns of 8 × 8 picture elements (pixels) with a particular structure and some noise. The structure in the three patterns (*a*) — a vertical black bar against a white background — is easily guessed, because the concept 'vertical black bar' is well established in the human memory. The structure (*c*) of the three patterns (*b*) does not become clear until the noise is removed. The 'signal-to-noise ratio' is the same in (*a*) as it is in (*b*); in both cases there are eight basic patterns in which four arbitrary pixels have changed value.

as those in fig. 1*a*. Most readers, however, will not recognize the structure without first having seen it in its pure form (see fig. 1*c*): it is not so clear-cut as those in fig. 1*a*. The system, however, has no more difficulty with this series than with the first series, because in both cases it starts with a clean memory and the information content of both is the same. This is one of the important aspects of the system: given a sufficient number of input signals it discovers structure or information that is far from evident.

Let us return for a moment to the reconstruction of an input signal (fig. 2). The concepts and noise points are written on a series of memory matrices. Their pixels have *three* possible states: white, black and 'unoccupied'. During the reconstruction each successive memory pattern writes over the previous patterns, but the unoccupied pixels have no effect in this process, so that the previous patterns are partly retained in the reconstructed pattern. Thus, in fig. 2 we see the black bar *2* and the black pixels *3*, *4* and *6* in the white field *1*, and the white pixel *5* in the black bar *2*.

In general terms the system works as follows. The memory structure is illustrated in *fig. 3*: there are 128 permanent elementary patterns stored in the memory, consisting of one white or one black pixel and a number of memory locations that are initially unoccupied. As soon as the first input pattern is presented there is 'recognition' in the sense described above: the pattern is reconstructed with the aid of 64

of the 128 elementary memory patterns. In addition, some of the elementary patterns that have been used are arbitrarily combined into new (two-pixel) patterns and stored in unoccupied matrices. Later input signals are recognized with the aid of the memory patterns present at that moment, and with as few of these as possible. A memory pattern is given a rating that is proportional to the frequency with which it occurs in the input signals. New memory patterns arise as a result of the combination of two patterns that were used for the last recognition. Such mutations, for which *fig. 4* gives two examples, take place on a statistical basis, with a high probability for combinations of patterns with a high rating. In general the patterns used continue to exist, and the new combination is stored in the matrix with the lowest rating; any information that happens to be stored there is lost. In this way the rating of the successful patterns increases

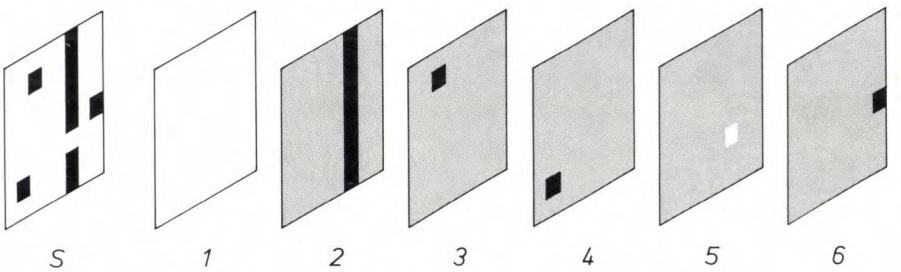


Fig. 2. Recognition. Once the system has learnt the structure 'white field/black bar' by 'looking at' a number of patterns of this structure, it reconstructs the input signal *S* with the aid of the 'concepts' *1* and *2* and the noise pixels *3*, *4*, *5* and *6* in its memory. This is what is meant by 'recognition' in this article. The pixels in the memory matrices have three possible states: black, white or 'unoccupied' (*grey*). Each successive memory pattern fills empty pixels in the output matrix with its 'information' (the black or white pixels) and writes over the information already present.

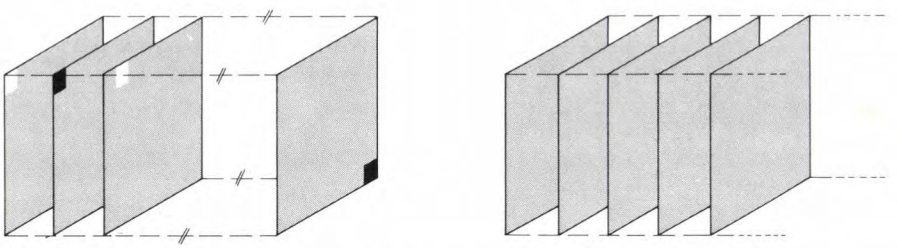


Fig. 3. Organization of the system for pattern recognition. The system contains 128 elementary memory locations, each containing one black or one white pixel, and a large number of other locations that are initially empty and into which concepts are written during the learning process.

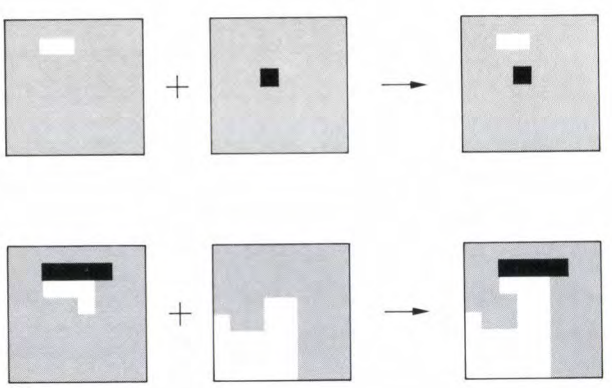


Fig. 4. Two examples of a mutation. At more or less arbitrary moments of time, new concepts are formed in the memory by the combination of two existing concepts. The two concepts remain in the memory.

at the expense of the less successful patterns, which eventually disappear. Finally, in the case of the class of patterns mentioned above, one matrix of the memory contains a white field and eight contain a vertical black bar. The optimum representation has then been reached: the system has discovered the structure in the input patterns.

There is therefore no distinction between a learning and a recognition phase; the system 'recognizes' right from the start, though not very efficiently. The merit of the system is not just that it can recognize, but that it learns to recognize more and more efficiently. In the case mentioned above it requires 64 patterns at first for the reconstruction, but only six in the final phase, i.e. two of the nine concepts formed and four

elementary patterns. The 'final phase', incidentally, is a concept unknown to the system: it goes on trying to find a better organization.

To explain the operation of the system in more detail, the algorithms on which it is based will now be discussed.

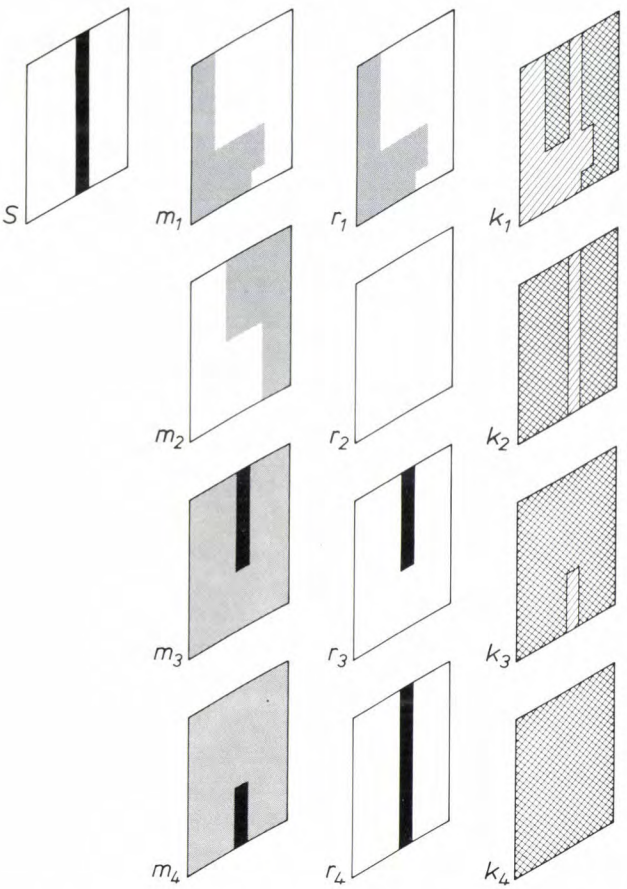


Fig. 5. Illustrating the recognition algorithm. The input signal S is reconstructed in steps in the output matrix (r_1, r_2, ...) with the memory patterns m_1, m_2, ... The pattern in the memory that gives the best match is found to be m_1, and this is written into the output matrix as the first step in the reconstruction (r_1). In k_1 it can be seen which part of S has been matched after this step (*double hatched*) and which part has not been matched (*single hatched*). The memory pattern that best matches the uncovered part appears to be m_2. This pattern is written over r_1 in the output matrix (r_2); now only the bar remains uncovered (k_2). The next best memory patterns are m_3 and m_4. After these four steps the input signal has been completely reconstructed.

Algorithms

The algorithm for *recognition* reads as follows (see *fig. 5*). Search the memory for the pattern that most closely matches the input signal. If there are several patterns that match equally well, choose the one with the smallest number of pixels. If there are more than one such pattern, choose any one of them. Read out the pattern thus found in the output matrix and compare it with the input signal. Repeat the operation for the pixels that have not yet been matched. Each successive memory pattern writes over the information already written in the output matrix at that moment.

It should be noted that this algorithm does not always lead to the reconstruction with the smallest number of memory patterns. If this were required, the system would have to try out all the possible combinations of memory patterns. This would be prohibitive for large memories, since the time required would increase exponentially with the number of memory patterns. The algorithm discussed works rapidly — the time required is only proportional to the number of memory patterns — and usually it works well. We therefore decided to let speed prevail over the certainty that the most efficient reconstruction will always be found.

The frequency of creation of new memory patterns (i.e. the frequency of mutations) is an important quantity: if it is too low, the process of evolution is too slow; if it is too high, valuable information may be lost. There must therefore be an optimum for the mutation probability. Without exactly defining what is meant here by 'optimum', the result of our research can be given here; this is the algorithm for *rating and mutation* that represents this optimum. It reads as follows: note after each recognition the rating w_i of each pattern, i.e. the total number of times it is used for recognition. Divide the patterns into 'plus' and 'minus' patterns, i.e. the patterns for which w_i is greater or less than the average rating ($\overline{w}$). Decide 'by casting lots' after each recognition whether the first plus pattern is to be used for a mutation, and give it a probability $(w_i - \overline{w})/w_i$. If the result is positive, combine it with the next plus pattern to form a new

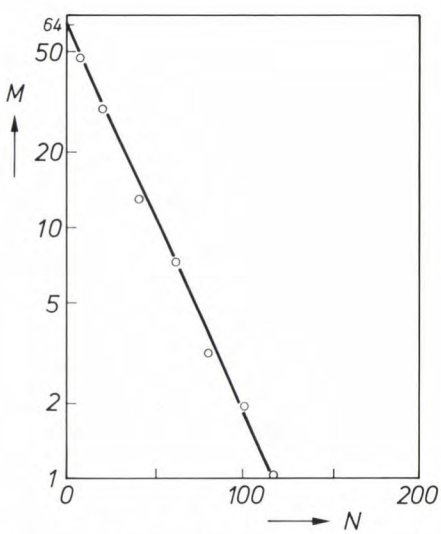


Fig. 6. Experimental learning behaviour of the system, showing the learning of a single pattern without noise. M is the number of patterns in the memory that are used in the reconstruction after the input signal, which always remains exactly the same, has been presented N times. The points were found in a simulation of the system on a minicomputer.

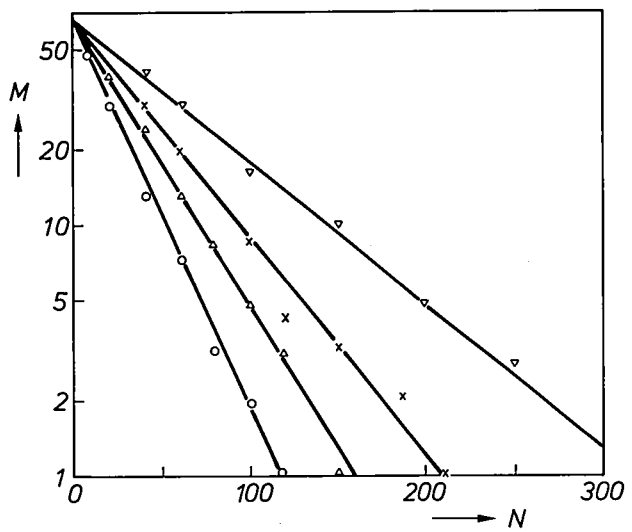


Fig. 7. Learning behaviour when a white field with noise is presented. M is the number of memory patterns used in the recognition of the pattern *without noise*, after the pattern *with noise* had been presented N times.
○ noise level 0 (0 noise pixels)
△ „ „ 1/16 (4 „ „)
× „ „ 1/8 (8 „ „)
▽ „ „ 1/4 (16 „ „)

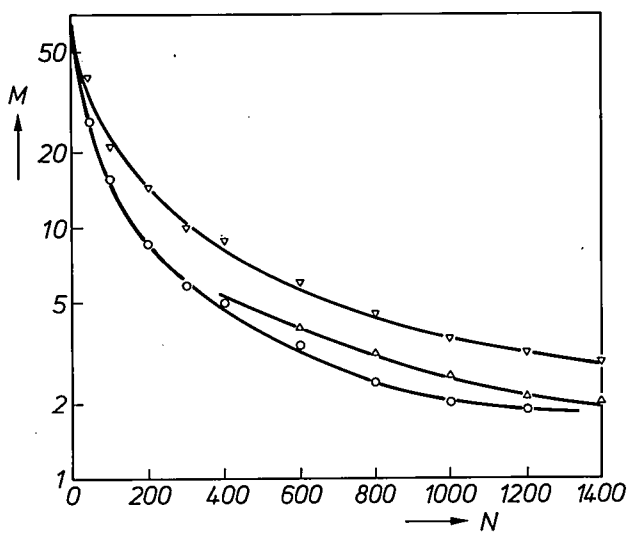


Fig. 8. Learning behaviour when the system is confronted with a white field with a vertical black bar at an arbitrary position and noise (see fig. 1*a*). M and N are as in fig. 7.
○ noise level 0 (0 noise pixels)
△ „ „ 1/32 (2 „ „)
▽ „ „ 1/8 (8 „ „)

pattern. The patterns used are retained, the new one is stored in the matrix with the lowest rating. Whatever the result, repeat the procedure with the following plus patterns.

This algorithm embodies the statistical, non-deterministic character of the system; in particular the number of mutations after each recognition can vary from zero to half the number of plus patterns present. It also implies the disappearance of the less successful patterns.

Experiments

To test the principles of the system discussed here, we simulated it on a Philips minicomputer (P855M) and performed a large number of experiments with it. A few of these experiments will be discussed here; the results are displayed in *figs 6, 7* and *8*. In these figures M is the number of memory patterns required for the reconstruction of an input pattern *without noise*, after the system has seen N input patterns, possibly including noise. (To reconstruct a pattern including noise the number of memory patterns required is therefore M plus the number of noise pixels.) The figures thus give the learning behaviour of the system; a steep curve implies 'rapid learning'. The figures differ in the nature of the input signals.

Fig. 6 relates to the learning of a single pattern. The same picture is presented, without noise, time after time to the system. As can be seen, the pattern is recognized with the aid of a single concept after the system has seen the pattern about 120 times.

The exponential relation between M and N expressed in fig. 6 can be explained from a highly simplified model of the system, in which, after each recognition, the probability of a mutation per memory pattern is set at a fixed number α. Assume that the system has M_0 pixels (64 in our case); this is the number of patterns required for the first recognition. The number of mutations that follow this recognition averages αM_0 in the model. Now in the present case (always the same pattern, no noise) each memory pattern formed from a combination of two used ones is again usable; each mutation therefore reduces the number of memory patterns required by one. The number of memory patterns required after the first recognition is thus $M_0 - \alpha M_0$, or $M_0(1-\alpha)$. After the second recognition it is $M_0(1-\alpha)^2$, after the third $M_0(1-\alpha)^3$, and so on. In general

$$M = M_0(1-\alpha)^N. \qquad (1)$$

We shall come back to the importance of this exponential relationship later.

Fig. 7 shows the learning behaviour for input signals that consist of a white field with a certain amount of noise. There are two reasons why the learning process becomes slower as the noise increases. In the first place, not every combination is now successful, and the probability of success decreases with increasing noise. In the second place, the rating for good combinations increases more slowly; it takes longer for the system to 'know' which combinations are good ones and which are bad ones.

Finally, it can be seen in fig. 8 how rapidly the system learns the patterns consisting of a 'white field with a vertical bar' (fig. 1*a*) at different noise levels.

The 'learning time' here is longer because the system has to learn not one concept but nine (a white field and eight vertical bars). Also, the learning curves are no longer exponential. This is because each curve is in fact the sum of two exponential curves with different 'time constants': the white field is a 'more important' concept and is therefore learnt more rapidly than the black bars. In all cases the system finally succeeds in reconstructing each pattern from two concepts and noise.

in more complex systems: the number of times (N) that a pattern has to be shown before the system 'knows' it ($M = 1$) increases only logarithmically with the complexity (the number of pixels M_0) of the system. This is very different from adaptive systems where, as we saw, the time required for adaptation would be expected to increase exponentially with the complexity (the number of parameters).

It should be noted that the learning time can of course be shortened by providing the system with

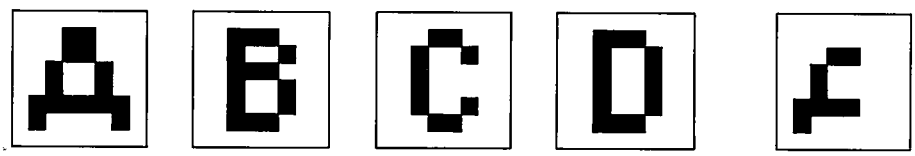


Fig. 9. 'Association'. Once the system has learnt the patterns (*a*), (*b*), (*c*) and (*d*), the 'hint' (*e*) is sufficient for it to recognize (*a*) from among them. It reconstructs (*e*) with the aid of the concept (*a*) and seven noise pixels.

Conclusions; further possibilities for self-organization

From the experiments discussed and from many others it is evident that 'self-organization' for systems is an effective principle. When the system has seen only a relatively small number of patterns with a particular structure, it has nearly always formed the best memory concepts for recognizing the structure. It should be noted that the concepts formed are 'ideal' although the system has probably never seen the structure in its ideal form, but always with errors. In this respect it might perhaps serve as an — admittedly extremely simple — model for the formation of concepts in the human brain. This might apply in other respects as well. If for example the system has learnt the patterns in fig. 9*a*, *b*, *c* and *d* and it is offered the pattern in fig. 9*e*, then this hint is all it needs: it interprets the pattern of fig. 9*e* as fig. 9*a* and a number of errors (by 'association').

Our investigation has led to two important quantitative results. The first is the *optimum mutation probability* (page 368). When this probability is used, the system evolves 'as rapidly as possible' towards a memory content capable of interpreting the presented material 'as well as possible'.

Secondly, a very important conclusion can be drawn from equation (1). Substituting β for the positive number $-\ln(1-\alpha)$, then from equation (1):

$$N = (\ln M_0 - \ln M)/\beta.$$

It follows from this relation that there are good prospects for the use of the principles described above *a priori* knowledge of the structure of the input signals, e.g. the concept 'white field' in the situation of the experiments of fig. 8. The system then only needs to find the eight bars; the self-organization begins at a higher level and therefore takes less time.

Conversely the time required for recognition of a pattern can be shortened if less information is sufficient. Once the system has learnt the patterns in *fig. 10* and similar patterns, the first step in the recognition is the same for all three: the memory pattern that fits best is a white field. It may be that this information is in itself sufficient, for example, if we only want to know whether the input signal is mainly white or black. In the next step the first two patterns are still identical, but the third one is different, and in the last step the first two are also different. This hierarchy in the

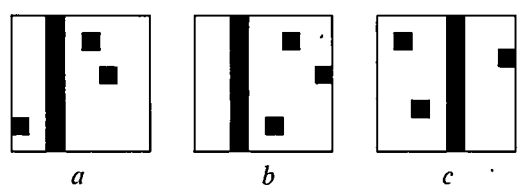


Fig. 10. Hierarchy in the recognition process. If the system has learnt the concept 'white field' and eight concepts 'vertical black bar', the three patterns (*a*), (*b*) and (*c*) are 'identical' at first sight: the first memory pattern is a white field. In the second instance, (*a*) and (*b*) are still identical (bar No. 3) but differ from (*c*) (bar No. 5). The difference between (*a*) and (*b*) only becomes apparent after further analysis. This hierarchy in the recognition process can be used for shortening the recognition time if only rough information is needed.

recognition process seems to be a very convenient feature; the system can stop comparing input patterns with stored patterns as soon as the information required has been obtained.

The system is capable of extension in many different ways. The first extension that comes to mind is the transformation of patterns stored in the memory. In the case 'white field with vertical bar' it would be possible to manage with two concepts instead of nine if, in the attempt at reconstruction, the vertical bar in the memory could be displaced horizontally. The system then has more latitude for making generalizations, leading to a more efficient use of its memory capacity and hence to the ability to store structures of greater complexity. In a further stage translations in arbitrary directions and rotations could be introduced.

It might even be possible to envisage entirely different forms of information processing. To take a random example, it might in principle be possible to make a system that could learn to 'read' when it was provided with appropriate peripherals. The system discussed associates certain black or white pixels that frequently occur together into a single memory concept. Reading is the same: it is the association of characters and sounds that go together.

Finally, let us return to the element of uncertainty in the recognition procedure. To make sure that the system would always find the best (most efficient) description of the input signals presented, it would first be necessary to define what the 'best' description is, and this would require a knowledge of all the input signals. But in fact the very thing we expect the system to do is to learn useful concepts from a small number of samples. Secondly, the system does not try out all the possibilities when presented with certain patterns — this would take far too long — but confines itself to the further evolution of the successful combinations. This uncertainty is the price to pay for achieving a reasonable degree of self-organization in a relatively short time. There is nothing new in this. All human beings are continuously confronted with this uncertainty, as they wonder whether their interpretation of the world around them is right or not[6].

[6] T. S. Kuhn, The structure of scientific revolutions, 2nd edition, Univ. of Chicago Press, Chicago 1970.

Summary. Pattern recognition is an example of information processing of such complexity that it is difficult to carry out automatically. This is mainly because the recognition process is difficult to analyse into precisely defined, simple algorithms. In this article an attempt is made to find a solution in the form of self-organization. By analogy with the evolution of a biological species and with the formation of concepts in the brain, a self-organizing system is characterized here mainly by a *memory* and by *random mutations*. The mutations make it possible to eventually adapt the contents of the memory to the environment. In the case of pattern recognition 'the environment' is the pictorial material presented. Computer simulation experiments with a system designed for the recognition of black-and-white patterns consisting of 8 × 8 pixels demonstrate the effectiveness of self-organization. The system has determined the structure in a certain class of patterns after having only seen a small proportion of them: the patterns accumulated in the memory enable the system to reconstruct each new input pattern in the most efficient manner.

Volume 38, 1978/79, No. 11/12 pages 289-372 Published 22nd April 1980

Supplement to Philips Technical Review
Beilage der Philips Technischen Rundschau
Bijlage van Philips technisch Tijdschrift

PHILIPS

January 1980

Recent United States Patents

Abstracts from patents that describe inventions from the following research laboratories that form part of or cooperate with the Philips group of companies:

Philips Research Laboratories, Eindhoven, The Netherlands	*E*
Philips Research Laboratories, Redhill, Surrey, England	*R*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France	*L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany	*A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
Philips Research Laboratory Brussels, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium	*B*
Philips Laboratories, N.A.P.C., 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A.	*N*

4 131 794

Fluorescent X-ray spectrometer

E. Bruninx *E*

A location-sensitive detector which can be read on-line is included in a fluorescent X-ray spectrometer comprising an analyzing crystal. Thus, the analysis result can be directly displayed and recorded without a film density measurement being required. For the detector use can be made of, for example, a gas-filled counter comprising a properly defined collector wire, a solid-state detector consisting of a row of detection elements, or a high-pressure camera which comprises electrodes which are constructed to be selective.

---

4 131 950

Charge transfer device

J. G. van Santen *E*

A charge transfer device in which series-parallel or parallel-series conversion of information which is present in the form of stored charge packets can take place and in which in a semiconductor body a row of charge storage sites is present in which beside said row several substantially parallel charge transfer registers are provided. The charge storage sites and the registers are interconnected by controllable charge transfer paths in the transverse direction so that charge packets can be distributed from the row between the registers or conversely charge packets distributed between the registers can be transferred to the row of storage sites.

---

4 132 896

Method of forming layered images of objects from superposition images of different image planes

E. Klotz *H*
U. Tiemens

A method for forming a layered image of a three-dimensional object which is irradiated from different directions by a plurality of radiation sources. *n* coded superposition images, in different image planes, of a single movement phase are sequentially recorded on a mechanically moved detector face. Decoding of the *n* superposition images is realized by means of a coordinate matrix whose coordinates correspond to the geometry of the radiation source distribution, so that these images are successively electronically superposed and summed, for a number of times which equals the number of radiation sources. The same layer of the object is decoded *n* times and the *n* decoded layers are summed to form a layered image.

4 132 959

Gas discharge laser having an asymmetrical coupling-out mirror

G. Bouwhuis *E*
J. van der Wal

A gas discharge laser comprises a multilayer coupling-out mirror including a plurality of alternate dielectric layers of high (*H*) and low (*L*) indices of refraction. A radiation-attenuating film element, preferably a metal film, is sandwiched between a pair of *L* dielectric layers located near the mirror substrate to provide an asymmetrical characteristic whereby the laser becomes relatively insensitive to light reflected towards it, which in turn reduces any tendency toward fluctuation in the intensity of the laser beam. The device is especially useful in systems for reading out information from a video record carrier wherein the reflected light is modulated and if returned to the laser would produce output power fluctuations therein.

---

4 134 014

Spectroscopy

J. H. Neave *R*
M. R. Boudry

A method of performing Auger electron spectroscopic surface analysis, e.g. of silicon, in which the secondary electrons which leave the target material are analyzed without the use of an electron optic device. An a.c. signal modulated retard potential is applied in ramp form to the target and a detection circuit connected to the target measures the component of current to earth from the target at the second harmonic of the a.c. modulation signal. The output of the detection circuit is plotted as a function of the retard potential to produce a spectrum of the first derivative of the secondary electron energy with respect to the retard potential so as to enhance the display of Auger transitions. Alternative realizations are described of a circuit arrangement including a four-port network for coupling the target, a.c. modulation signal source, retard potential source, and detector circuit.

---

4 134 123

High voltage Schottky barrier diode

J. M. Shannon *R*

An improved high voltage Schottky barrier diode includes a semiconductor layer having two adjacent sublayers of the same type conductivity but different doping concentrations. A plurality of isolated discrete regions of a second type conductivity op-

posite to that of the first are provided along the boundary region between the sublayers and beneath the Schottky junction. The invention results in an improved high voltage Schottky diode in which the reverse characteristics are substantially enhanced.

---

4 134 134

Apparatus for picture processing

P. Lux H

A Hadamard domain encoder wherein adding and subtracting operations on digitized intensities of 2×2 picture points yield secondary results. Likewise processing of groups of 2×2 secondary results yields further intermediate and/or final results. The complement of unprocessed secondary and final results fully characterizes the original picture but is better suited for data storage and/or transfer because the energy content of the picture is substantially restricted to a limited region. Redundancy in the stored/transmitted data is thus reduced.

---

4 134 388

Solar collector comprising a U-shaped absorber

R. Kersten A
E. Kuhl

A solar collector comprising a U-shaped absorber sealed in an evacuated cover tube provided with a semi-cylindrical reflector on its inner surface, the U-shaped absorber being arranged between the cover tube axis and the reflector, with the central plane of the U-shaped absorber including the center lines of its legs and coinciding with the symmetry plane of the reflector.

---

4 134 391

Solar collector comprising an elongate Dewar vessel

F. Mahdjuri A
H. Hörster
R. Kersten

A solar collector comprising an elongate solid cylindrical body extending lengthwise into an elongate cylindrical Dewar-type vessel in a manner to form annular-shaped supply and outlet ducts for a heat-transport medium, the sum of the cross-sectional areas of the supply and outlet ducts being less than one-quarter of the inner cross-sectional area of the vessel.

---

4 135 107

Multi-phase alternating current machine with stepped stator

E. M. H. Kamerbeek E
H. G. Lakerveld

A multiphase electric machine whose stator winding is disposed in an air gap between a smooth stator wall and the rotor. The coils of the stator winding are wound stepwise in such a way that per coil the field variation in the direction of movement of the rotor is substantially sinusoidal and the winding has a constant thickness.

---

4 135 111

Shadow mask having magnetic quadrupoles around each mask aperture

J. Verweel E

A cathode ray tube for displaying colored pictures of the post focusing type. A magnetic quadrupole lens is formed in each aperture of the color selection means. The defocusing direction of quadrupole lenses is parallel to the phosphor strips of the display screen.

4 135 206

Objective mount for video disc player

W. J. Kleuters E
G. E. van Rosmalen
M. P. M. Bierhoff
K. A. Immink

An optical scanning device which scans a rotary video disc with the aid of a light beam comprises an objective which performs focusing movements, for focusing a light spot on a recording surface of the video disc, as well as tilting tracking movements and time-error correction movements. Near its lower end the objective is suspended in an impregnated corrugated loudspeaker diaphragm of fabric fibre which enables the various movements of the objective to be made without giving rise to annoying resonant effects. The tracking movements and the time-error correction movements are obtained with the aid of coils which are connected to the objective near the upper end and which are disposed in air gaps of associated permanent magnetic stators.

---

4 135 207

Apparatus for reading an optical radiation-reflecting record carrier including a narrow focus control beam

P. F. Greve E
W. G. Opheij

An apparatus is described for reading an optically encoded radiation-reflecting information carrier, using an opto-electronic focusing error detection system. The path of the read beam which is directed towards the information carrier includes a radiation-deflecting element producing a narrow focusing beam that co-operates with two focusing detectors to produce an error signal.

---

4 136 006

Electrode for electrochemical machining

G. Verspui E

An electrode for the electrochemical machining of electrically conductive workpieces, which electrode has a layer of polycrystalline silicon carbide via an insulator layer which is at most 0.5 μm thick.

---

4 136 342

Resolving angular ambiguity

R. N. Alcock R
D. A. Lucas
R. P. Vincent

Angular ambiguity, particularly in measurements made in radio interferometers, is resolved by comparing an ambiguous value for a desired angle with an unambiguous value for an angle which is ideally equal to the desired angle but which may differ from it by an error angle less than π. For example, ambiguity in the phase difference between a pair of widely-spaced antennas may be resolved by reference to the sum of two unambiguous phase differences between each of those antennas and an intermediate antenna. The effect on the ambiguity-resolution process of certain sources of phase errors in practical systems can be eliminated, leaving a large tolerance to remaining errors.

---

4 136 398

Digital filter having coefficient number generator

L. D. J. Eggermont E

A digital filter for generating digital output signals from information signals in accordance with a predetermined transfer characteristic includes a coefficient number generator for supplying digital numbers which indicate the values of coefficients which represent the transfer characteristic. The coefficient number generator includes a device for storing increments of successive coefficient values in numerical form and a decoder connected thereto for recovering the coefficient numbers from the numbers stored in the storage device.

4 136 531

^{3}He-^{4}He dilution refrigerator

F. A. Staas E
A. P. Severijns

A ^{3}He-^{4}He dilution refrigerator is provided with two interconnected mixing chambers arranged at different levels. One end of a superleak opens into concentrated ^{3}He contained in the upper mixing chamber, while the other end opens into a portion of the apparatus containing dilute ^{3}He for the supply of superfluid ^{4}He, via the superleak, to the upper mixing chamber under the influence of the osmotic differential pressure prevailing across the superleak.

---

4 136 733

Heating device

G. A. A. Asselman E
D. B. Green
A. P. J. Castelijns
P. A. Naastepad
J. W. de Ruiter

An isothermal heating device comprises an annular tubular body providing a heating chamber for objects. Between the inner wall and the outer wall of this tubular body there is provided a plurality of separate ducts which are situated in a ring-shape about the heating chamber and which extend parallel to the tubular body axis. These ducts contain an evaporable heat transport medium and are in communication with each other preferably at one end of the tubular body. Such interconnection is in turn connected with a gas buffer reservoir containing an inert control gas.

---

4 137 107

Method of manufacturing a semiconductor device utilizing selective masking, deposition and etching

W. Nijman E
P. J. de Waard

The invention relates to a method in which a system of layers with a contact layer of gallium arsenide is formed epitaxially. In a second epitaxy treatment a layer of gallium aluminum arsenide is formed selectively. In order not to form the latter layer on the contact layer of gallium arsenide, the latter is shielded from gallium aluminum arsenide by means of a masking layer having a composition which differs from that of the layer to be provided selectively, so that the masking layer can afterwards be removed selectively.

---

4 137 122

Method of manufacturing a semiconductor device

G. A. Acket E

The invention relates to a method of manufacturing a semiconductor device in which a compound consisting of at least two semiconductor materials and having an energy gap which is smaller than that of a substrate is deposited epitaxially on the substrate. According to the invention, the composition of the mixture is determined during the deposition by means of measurement of the thermal emission.

---

4 137 458

Electron image projection masks

H. N. G. King R
J. P. Scott

An electron image projection mask consisting of an optical mask comprising a transparent substrate, bearing an opaque mask pattern disposed on the substrate, a transparent coating extending over the mask pattern and the areas of the substrate exposed through the apertures in the opaque mask pattern, and a metal image extending over the transparent coating. The metal image has a pattern corresponding to the apertures in the opaque mask pattern, the optical density of the metal image being such as to reduce the average light transmission through the apertures in the mask to a value between 25 and 80% of the average light transmission of the apertures in the absence of the metal image. Also, a method of producing the electron image projection mask.

---

4 137 465

Multi-stage integrated injection logic circuit

C. M. Hart E

A multi-stage I^2L circuit includes a first switching transistor in a higher stage coupled as a bistable trigger to a second transistor in a lower stage. Additional circuitry is supplied to increase the switching speed of the I^2L circuit by draining stored charge from the base of the second transistor. Regenerative feedback is employed to enhance the effect.

---

4 137 477

Electrodes, for example grid-like electrodes for use in electron tubes, and a method for manufacturing same

J. W. A. Krol A
B. Lersmacher
H. Lydtin
H. Seifert

Disclosed are grid electrodes of glassy carbon which have good mechanical and electrical properties and can be manufactured in a simple manner with a great dimensional stability. The grid electrodes may be made by two alternative methods. In one method a body of cured synthetic resin is formed or processed into the shape of the desired electrode, after which it is carbonized. In the second method a preshaped body of cured synthetic resin which is in the unprocessed state, is carbonized after which the resulting glassy carbon body is further processed to the desired electrode, preferably by laser beam cutting.

---

4 137 536

Electrostatic printing device with air cushion guiding

H.-D. Hinz H
U. Rothgordt

An electrostatic printing device comprising a rotatable printing roller with stylus electrodes. In order to keep the distance of the images or of the recorded sheets small and the images themselves clear and pure, notably at the edges, the stylus electrodes are moved transversely of the longitudinal direction of the tape-like record carrier. The foil tape is at the same time continuously transported. Thus, the translatory movement of the printing roller supporting the stylus electrodes is eliminated.

---

4 138 530

Magnetic structures

D. J. Breed E
A. B. Voermans
H. Logmans

A magnetic structure having a high magnetic bubble mobility which has the property that magnetic bubbles can be transported in it at very high velocities while using comparatively weak driving fields, comprising a substrate having a (110) oriented deposition surface on which a layer of rare earth-iron garnet with a substitution of Mn^{3+} in iron lattice sites has been grown in compression in such a manner that the layer of rare earth-iron garnet has an orthorhombic anisotropy.

4 139 250

Gas discharge display panel and method of manufacturing the same

J. H. Jacobs E
J. van Esdonk
J. P. T. Franssen

A vacuum-tight seal is produced between plate-shaped elements one of which is glass, of a gas discharge display device while maintaining such a pressure on the surface of the elements that they are brought in permanent contact with a spacing member which separates them. Furthermore, at least the glass face plate is preferably heated to such a high temperature that a possible elastic deformation of said element resulting from a pressure exerted on the surface thereof is converted into a plastic deformation.

---

4 139 795

Television camera tube

M. G. Carasso E
P. P. M. Schampers

A photoconductive layer which is vapor-deposited on a carrier in a television camera tube is provided with a structure of regions having at least one dimension which corresponds in the order of magnitude, measured transversely of the layer, to the transverse dimension of an image point in the layer. These regions are separated by partitions which consist of the same material as the further conductive layer, but for which the structure is such that they exhibit a comparatively low transverse conduction.

---

4 141 033

Device for automatically controlling the asymmetry of a signal

E. de Boer E

When a color television signal is recorded on a video disc and said signal is read a complication arises owing to the asymmetry which is produced in the signal by the various process steps. This gives rise to spurious components whose influence should be minimized by a suitably selected coding system. The invention proposes a device with which the influence of this asymmetry is automatically reduced. This device comprises two separator stages for supplying two sub-signals which respectively represent the rising and falling edges of the information signal. The phase relationship between these two sub-signals is controlled with the aid of a phase shifter. The control signal required for this is produced with the aid of a detector which responds to an asymmetry component in the information signal. The two corrected sub-signals are subsequently recombined to the information signal.

---

4 141 051

Variable dynamic range magneto-resistive head

K. E. Kuijk E
F. W. Gorter
J. A. L. Potgiesser

A magnetic reading head having a magneto-resistive element of the non-magnetically biased type, in which the current is forced to flow through the element at an angle with the easy axis of magnetization. For adjusting the dynamic range of the element, the element is subjected to a variable strength auxiliary field which is parallel to the easy axis of magnetization.

---

4 141 603

Rotation-insensitive spiral groove bearing

G. Remmers E
L. P. M. Tielemans

A bearing having two parts which are rotatable relative to each other, and two facing bearing surfaces on the respective parts cooperating with each other. One or both surfaces are provided with three patterns of shallow, parallel lubricant pumping grooves, the two outer patterns having grooves which have a pumping effect in the same direction, and the inner pattern having grooves which have a pump effect in the opposite direction.

---

4 142 090

Method of and device for plasma MIG welding

W. G. Essers E
G. Jelmorini
G. W. Tichelaar

Plasma MIG welding in which a plasma flow is generated between a non-consumable electrode and a workpiece, and a consumable electrode is fed through such plasma flow toward the workpiece to establish a MIG arc between the same. The current intensity in the plasma flow is increased by applying additional current thereto from an annular auxiliary non-consumable electrode surrounding the plasma flow and positioned between the primary non-consumable electrode and the workpiece.

---

4 142 218

Magnetoresistive head

F. W. Gorter E

A magnetic reading device having two electrically series-arranged magnetoresistive elements, in particular for interference compensation and positioning control. A pattern of equipotential strips is provided on each element to adjust the quiescent angle α between the direction of magnetisation and the direction of current passage of one element between 30° and 60° and between 210° and 240°, respectively, and that of the other element at 360° $-\ \alpha$ and 180° $-\ \alpha$, respectively, the device has for various applications an automatic compensation of the signals across the elements so that the use of difference amplifiers is superfluous. In addition, due to this adjustment, the reproduction characteristic has a linear variation.

---

4 142 509

Solar collector comprising an evacuated cover

W. Hermann A
H. Hörster
J. Schröder
F. Mahdjuri

A solar collector including at least one evacuated transparent cover tube provided with means for supplying hydrogen to and extracting hydrogen from the cover tube.

---

4 142 987

Luminescent material dispersion

D. J. Zwanenburg E
H.-G. Junginger

A proper adherance of luminescent material particles to the surface of a photoconductor is obtained when vinyl pyridinealkyl methacrylates polymers of 1-6, 6-14 and 14-20 carbon atoms are used as non-ionic macromolecular compounds.

---

4 143 365

Device for the acquisition and storage of an electrical signal

J. C. Cayzac L
C. J. P. F. le Can
R. Brun
J. Devillers
E. Joinnet

A device for the acquisition and storage of an electrical signal. The signal is divided into adjoining elementary time intervals of specific duration. During each of them the amplitude minimum and amplitude maximum of the signal are established, which values are first stored in analog form and subsequently in digital form. The device may be utilized in oscilloscopes, in particular those of the type comprising a matrix display panel.