VOLUME 4 □ NUMBER 3 MARCH 1967

IEEE spectrum

features

Departments: *please turn to the next page*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE International Convention and Exhibition

departments

the cover

The cover this month is a contemporary adaptation of a poster in the style common at the turn of the century. The old style was chosen purposely to present a sharp contrast to the timeliness and forward-looking scope of the IEEE International Convention and Exhibition.

Spectral lines

The challenge of picture transmission. Television is old art, but it still poses a big challenge to the communications engineer. Most of the concepts basic to present-day picture transmission facsimile, as well as television, were developed 40 years ago by pioneers whose names are already becoming legend. The commercial applications were made some 20 years ago by men who with great insight foresaw the promise, and set the technical base for a great industry.

It would be easy to envy the pioneers who moved in such an open and promising field, with practically unlimited possibilities, with relatively modest competition. It was indeed a time of adventure, but remember the (relatively) primitive electronic tools, the dearth of measuring instruments, the scarcity of technically trained aides. The words "do it yourself" were not yet colloquial, but it was largely a do-it-yourself technology.

As a popular subject for research, television has given way to integrated circuits, quantum electronics, and computer technology, but there remain many challenging problems in the picture transmission area. The problems may not seem very tractable, but good problems seldom are. Indeed, the new technologies hold the key to advances beyond anything seen in the picture systems area for many years.

The last 20 years have witnessed a proliferation of ideas for better picture transmission. The tools for their implementation have been lacking however. Invention has outreached practice to an almost unprecedent degree.

This is evident today in a proliferation of publications. Last fall the Institute put out a call for papers on "redundancy reduction and bandwidth saving" for a special issue of the PROCEEDINGS OF THE IEEE. Of the papers offered, two thirds were directly concerned with picture transmission. Of those that made it into the PROCEEDINGS, the ratio was still two thirds. In addition, recent editions of other journals have carried at least five good articles in this area.

Elsewhere in this issue of SPECTRUM the subject is also given coverage, with one paper describing redundancy reduction techniques, another presenting a serious account of the European color television debate, and a third proving that we have not yet lost our sense of humor!

It is evident that much of the work on redundancy reduction has yet to be built into real systems, and that it will not be practical to do so until the promise of integrated circuits is close to fulfillment.

During the last 15 years or so, many people have struggled with the knowledge that useful pictures are very redundant; that the human viewer cannot begin to utilize the detail present in a picture, and that present methods of picture transmission use orders-of-magnitude more transmission capacity than one might think would be necessary. On one hand, we are told that a "standard" television signal has a capacity of 6×10^7 bits/second, and on the other, that the human visual system is capable of utilizing only about 50 bits per second, a ratio of a million to one. There is more to a picture than "information," but it is also evident that present methods of picture transmission are not well matched to the visual process.

The present series of papers in the PROCEEDINGS demonstrate, mostly by simulation, that savings of communication capacity by factors of from two to ten are likely without too much loss of picture quality, using technology that is just now developing. Is there still another big factor forthcoming?

We are just beginning to understand some of the basic physiological processes of vision. It has been found that physiological processing of the image on the retina is done at a very low level, and that the picture "information" actually reaching the brain is but a small fraction of that available at the retina. Can we learn a lesson from nature, and accomplish the same thing in our electronic systems?

And what about color? Soon most of the world will settle on one or two stable systems for commercial color television. The European systems, PAL and SECAM, are definitely superior to the U.S. system (NTSC). In the United States, European developments are watched with great interest, but unless something quite radical occurs, U.S. color television broadcasting will eventually be second rate.

Closed-circuit television is just getting started, and is not so inhibited by a necessity for standardization. Is it here that the new art will generate pictures with new standards of excellence? Given the channel capacity of a laser beam, the possibilities of redundancy reduction, and new refinements in electron beam and optical scanners, we don't see why not.

What these developments will mean for space exploration and for closed-circuit television systems can only be guessed, but it seems quite certain that they will keep many of us busy for some time to come.

C. C. Cutler

Authors

International standards for color television (page 104)

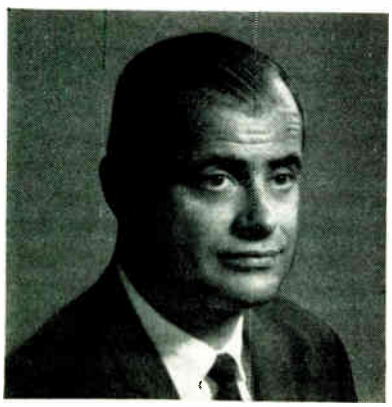
Jack W. Herbstreit (F) received the E.E. degree from the University of Cincinnati in 1939. He joined the FCC in 1940 as a radio inspector and associate engineer in Atlanta and Washington, D.C. During World War II he was consultant on radio-wave propagation to the Chief Signal Officer of the Army. In 1946 he joined the NBS Central Radio Propagation Laboratory. In 1962 he was made deputy director of CRPL, which is now the Institute of Telecommunication Science and Aeronomy. In July 1966 he was named director of the CCIR. He received the IRE Harry Diamond Award in 1959 for his work in radio propagation and the U.S. Department of Commerce Gold Medal Award in 1966 for his contributions in high-precision radio tracking and guidance systems.

H. Pouliquen received the science and engineering diplomas in 1943 and 1946 from l'Ecole Polytechnique, Paris, and the engineering degree from the Ecole Nationale Supérieure des Télécommunications, Paris, in 1948. He joined the Office de Radiodiffusion-Télévision Française, where he was involved in putting into service the first stations of the 819-line television network. Later he transferred to the television department of the ORTF Research Service. He was subsequently appointed engineer in the Specialized Secretariat of the CCIR in Geneva. He has had the opportunity of attending all the Plenary Assemblies of the CCIR since 1953 and all the meetings of the CCIR Study Group XI. He has taken active part in the work of the Study Group as a whole and of working parties set up to discuss television problems.

Universal color television: an electronic fantasia (page 112)

Joseph Roizen (M) is at present the audio/video product manager for Ampex International. He has been with Ampex Corporation for over ten years, where his work has included direct engineering on the first commercial video tape recorders and the first color video tape recorders. Earlier he was with the Television Division of Paramount Pictures, designing color studio equipment and switching system control rooms. He has written more than 50 articles on color television and magnetic recording and holds several patents in these fields. Mr. Roizen attended Sir George Williams College, Montreal, and took graduate courses at the University of California, Los Angeles. He is on the faculties of the University of California, Berkeley, and of Foothill Junior College, Los Altos Hills, Calif., where he teaches courses in electronics, color television, and magnetic recording.

High-power lasers—their performance, limitations, and future (page 115)



F. P. Burns (M) received the B.S.M.E. degree in 1947 from City College of New York and the M.A. in physics and Ph.D. in solid-state physics from Columbia University in 1949 and 1954 respectively. From 1947 to 1954 he taught at City College, where he was assistant professor of mechanical engineering. During the next four years he worked at Bell Telephone Laboratories, where he worked on semiconductor investigations, and Tung-Sol Electric, where he was manager of the Silicon Device Development Section. In 1958 he became manager of industrial device development for RCA and in 1960 he joined Solid State Radiations as manager of the semiconductor laboratory. Since 1962 he has been manager of operations of the Korad Corporation, a subsidiary of Union Carbide Corporation, where he is responsible for product development and engineering of laser systems.

The radio spectrum below 550 kHz (page 121)

Thomas L. Greenwood (SM) attended Washington University, St. Louis, from 1950 to 1951 and the Huntsville branch of the University of Alabama from 1951 to 1953. He has had more than 30 years' experience in electronics and communications, including eight years as a Merchant Marine radio officer; four years as a radio broadcast technician; eight years as chief engineer of radio station WMOB, Mobile, Ala.; and three years in the Communications Department of the Gulf, Mobile and Ohio Railroad. From 1951 to 1960 he supervised activities connected with guided-missile instrumentation development at the U.S. Army Recorder and Electronics Laboratory, Redstone Arsenal, and subsequently was engaged in the development of instrumentation and data-handling systems at the National Aeronautics and Space Administration's George C. Marshall Space Flight Center.

Philosophy of an engineering educator (page 124)

Aaron J. Teller, dean of The Cooper Union School of Engineering and Science, has been active as an educator, research scientist, and industrialist. Since coming to Cooper Union in 1962 he has supervised the establishment of the college's graduate division, offering master's degrees and doctorates in engineering, and has directed a comprehensive revision of the undergraduate curricula. He received the bachelor's degree in chemical engineering from The Cooper Union in 1943, the master's degree from the Polytechnic Institute of Brooklyn in 1949, and the doctorate from the Case Institute in 1951. He began his professional career with the Manhattan Project at Columbia University during World War II and later was on the faculties of Fenn College and the University of Florida. Dr. Teller has a distinguished record of service to private industry and engineering education. His research and consulting engineering work have led to a number of patents.

Data compression by redundancy reduction (page 133)

C. M. Kortman (SM) joined Lockheed Missiles & Space Company in 1956. Since that time he has been involved in a number of advanced development projects in the fields of instrumentation, telemetry, and communications. Since 1962 he has been concerned with the various facets of the data compression activities at LMSC as well as with advanced electronics study and development projects. In his present capacity as manager, advanced techniques, he is responsible for all data compression studies within the company, and for all R & D efforts on ground and vehicle-borne data compression hardware. He received the B.E.E. degree from George Washington University, spent four years in the U.S. Army Air Force and Signal Corps, and also worked for the High Frequency Standards Section of the National Bureau of Standards, Chance Vought Aircraft, and Bendix Aviation.

Foundations of the case for natural-language programming (page 140)

Mark Halpern (M) holds degrees from the City College of New York (B.A., 1951) and Columbia University (M.A., 1955). After pursuing further graduate studies he joined IBM's Programming Research Department and worked on a variety of commercial and scientific compilers; he later transferred to the Programming Systems Department, where he was concerned with the design of compiler-producing systems. In 1961 he moved to the Lockheed Missiles & Space Company's Palo Alto Research Laboratory, where he has led a small group in developing the XPOP "compiler-compiler" system. His principal professional interest is in man–machine communications, particularly in connection with natural-language programming, compilers, and systematization of debugging. He is an editor of the series *Annual Review in Automatic Programming* and a contributing editor of the *Journal of Data Management*.

International standards

Unfortunately, but perhaps not unexpectedly, the recent CCIR Assembly in Oslo failed to agree unanimously on a universal color television system. Hopefully, further efforts will be made to bring about the adoption of a single standard in the future

As outlined in a recent issue of IEEE Spectrum (June 1966, pp. 59-68), it was hoped that last summer's meeting of the CCIR in Oslo would result in the recommendation of a single color television system for adoption throughout Europe and other parts of the world. However, such a recommendation was not forthcoming, chiefly because no one of the systems considered is overwhelmingly superior to the others from the point of view of performance or cost. This article describes the similarities and differences between the various systems proposed and points out the manufacturing alternatives should a multiplicity of standards eventually be adopted.

The most controversial and publicized point on the agenda of the XIth Plenary Assembly of the CCIR (International Radio Consultative Committee), held in Oslo in June-July 1966, was the standardization of color television systems by the countries of the world. This point is particularly important to those countries that do not yet have a regular color television service—that is, almost all countries with the exception of the United States and Japan.

Discussion on this point proved inconclusive; therefore, instead of issuing a recommendation unanimously favoring a single system, the CCIR was able only to issue a report describing the characteristics of the different systems proposed for a basic standard. It is, therefore, at the present time up to the various administrations to make their own choices as to which standard to adopt.

It is always very regrettable when the countries of the world fail to agree on a single standard. This is amply demonstrated by the lack of common standards for such ordinary matters as measuring units (meters vs. feet), power standards (50 Hz vs. 60 Hz, 115 volts vs. 220 volts), and the sizes of nuts and bolts. Without a common standard, the successful development of television will be particularly hampered by difficulties in receiver construction and in the international exchange of programs. The latter will be possible only with the aid of "transcoders," with a resultant loss in quality. In the border areas between the two systems, reception of all available programs will be possible only with multistandard receivers, which will be considerably more costly than receivers designed around a single standard.

In the matter of international standardization, it is often said that a start must be made, but it should come neither too early nor too late. The CCIR has attempted to follow this line of reasoning; as long ago as 1948, the Vth Plenary Assembly of the CCIR adopted Recommendation 29, the philosophy of which is as valid today as it was then. It reads as follows:

"*Recommendation No. 29—Television Standards*

The C.C.I.R.

considering:

(a) that the interchange of television programs between countries is desirable;

(b) that the interchange of such programs should be done in an economical manner;

(c) that the economical interchange of television programs would be facilitated by the adoption of agreed standards for certain characteristics of transmissions;

(d) that technical standards should be coordinated, insofar as possible, to permit such interchange to facilitate the utilization of receiving equipment, and to minimize mutual interference between television services;

(e) that the adoption of such standards will result in the most rapid expansion of the television service, by making more readily available a wider variety of programs and, in addition, giving a reduction in program costs;

(f) that it is desirable that worldwide agreement be obtained on those standards which would permit interchange of programs, both direct and recorded;

(g) that the interchange of programs will be effected by radio relay and cable links for direct programs, and by film for recorded programs. With interchange between different linguistic groups the sound channel characteristics are of secondary importance, and primary attention needs to be placed on the vision signal;

(h) that the question of program interchange is also linked with the desirable technical characteristics necessary to provide:

1. a satisfactory service in the home at reasonable cost;

for color television

Jack W. Herbstreit, H. Pouliquen

International Radio Consultative Committee

2. a reasonable service in the home at minimum cost;

(i) that in consideration of these problems account should be taken of the following factors:

1. the available bandwidth allocated to television is limited;
2. importance is to be attached principally to the cost of receivers rather than that of transmitting equipment;
3. the proposed standards should not preclude in due course, the possibility of reception, by the addition of a suitable frequency converter, of the following:

monochrome pictures on a 'black and white' receiver of the transmissions from a 'color' transmitter, monochrome pictures on a 'color' receiver of transmissions from a 'black and white' transmitter;

(j) that the adoption of transmission standards on as wide a basis as possible will result in the most rapid expansion of the television service, in that it will facilitate the production of receivers at lower cost;

(k) that a factor of prime importance in arriving at world standards is the problem of operating a television service in which the frame repetition rate is not integrally related to the power supply frequency;

(l) that it is inevitable that there will be considerable channel sharing in the existing television bands, and therefore, in view of long distance propagation effects, it is desirable that the standards proposed should be such as to minimize interference between stations.

RECOMMENDS:

that there be undertaken the study of, and publication of Recommendations on the technical factors which would assist in achieving:

(a) Interchange of programs on the widest possible scale,

(b) Coordination of standards to permit the use of a receiver on transmissions differing in a minor degree.

The factors which appear of major importance are:

1. For the interchange of direct programs:
   (a) Frame repetitions rate,
   (b) Frame interlacing,
   (c) Number of lines,
   (d) Aspect ratio;
2. For the interchange of recorded programs:
   (a) The programs should be recorded in such a manner as to make them capable of being reproduced on standard 35 or 16 mm motion picture sound equipment;
   (b) The effects of pattern interference due to transmission of a film on a television system having a different number of lines from that on which the film was recorded;

Among other factors which should be studied to permit interchange of receivers are the following:

(a) Polarity of modulation for vision signal
(b) Distribution of channels in the available spectrum space
(c) Relative frequencies of sound and vision carriers and the positioning of these carriers and associated sidebands within the channel
(d) Type of vision transmission, e.g. double sideband, single sideband, etc.
(e) Type of modulation of sound channel
(f) Form of synchronizing signal
(g) Nonintegral relationship between frame repetition rate and power frequency"

In 1951 the CCIR adopted a study program that highlighted the problem of color television by laying stress on the "point of view of picture quality, program costs and the cost of receivers or converters."

Events took a new turn after the United States adopted the NTSC (National Television System Committee) system and in Brussels in 1955, the CCIR adopted Question 118, which established several criteria, including:

1. Satisfactory picture (color and monochrome) and sound quality
2. Economical use of bandwidth
3. Reliable receivers of reasonable cost
4. Operation of studio, transmitting, and relaying equipment
5. Susceptibility to interference
6. Compatibilities
7. Frequency planning
8. International exchange of programs

9. Scope for development
10. The differences between bands I and III as compared with bands IV and V

At that time there was no call for color television in many countries, which were engaged in installing a black-and-white television network.

In anticipation of the European VHF/UHF Broadcasting Conference, planned for 1961, a meeting was held in October 1959 of an ad hoc group, chaired by Erik Esping, chairman of CCIR Study Group XI (Television), which established some of the essential factors involved in planning standards. In particular, unanimous agreement was reached on a value of 4.43 MHz for the color subcarrier, which had already been proposed at the meeting of Study Group XI in Moscow in 1958.

Before the actual planning conference, which was to meet in Stockholm in June 1961, a preparatory technical meeting was held in Cannes under the auspices of the CCIR in February 1961. Before the meeting, the situation in Europe was one of extreme confusion because of the variety of black-and-white television standards in the broadcasting bands I and III. At this meeting the United Kingdom and France, in a laudable desire for standardization, declared their intention of transmitting color television as well as black and white on 625 lines in bands IV and V. Moreover, agreement was reached on the adoption of 8-MHz channel spacing in those bands.

In the ensuing period, increased interest was shown in the subject and in 1964 a CCIR color television meeting was held in London. This meeting proved to be a turning point, for on that occasion everything was ready for the adoption of a single 625-line color television standard in the European Broadcasting Area. However, extensive studies were still being undertaken in a number of countries on matters concerning color coding for such a standard. The systems proposed at that time were the NTSC, SECAM (séquentiel couleur à mémoire), and PAL (phase alternation line) systems. In fact, a few countries decided that the time was ripe to introduce color and some of them then proposed the adoption of NTSC, being guided in their choice by the experience already gained with NTSC and by the fact that the very slight difference in quality of the pictures in the three systems did not justify the introduction of other systems.

After the 1964 London meeting, studies continued on the different systems. Special mention should be made of the work done by the EBU (European Broadcasting Union) ad hoc Group on Color Television, a group that supplied the CCIR meetings with very carefully prepared data for comparing the different systems.

The next meeting was held in Vienna in April 1965, and it became clear that those countries that had made the greatest efforts in research could wait no longer without jeopardizing the results of their efforts. In fact, it was obvious from the outset that many were ready to take a stand. Since the various ideas on the subject were incompatible, the meeting ended without agreement.

At the end of the Vienna meeting, the more optimistic countries placed their last hopes for a single standard in the Study Group XI meeting to be held during the XIth Plenary Assembly of the CCIR in Oslo in 1966.

Compatible color television systems

At this stage, it would be pointless to analyze, even briefly, the many different color television systems proposed or studied up to about 1950. However, mention should be made of a patent dated January 17, 1938, by Georges Valensi, one-time director of the CCIF (now incorporated in the CCITT), in which the idea of separately transmitting "luminance" (the brightness existing in a monochrome picture) and "chrominance" (color information, that is, hue and saturation) was first advanced. The chrominance information was transmitted in the form of a single parameter characterizing the area of the color diagrams at the figurative point of the area analyzed. This was actually the first example of a compatible system.

For a better understanding of the situation, a brief survey will be made of the technical features of the different systems proposed. We shall start with the NTSC system, which is still, in many respects, a reference system, since it was the first system developed to meet two of the basic requirements of any color system: reciprocal compatibility with black-and-white pictures and transmission in the same channel as monochrome pictures.

Having explained the principles of that system, we shall make some brief comparisons.

NTSC system. After abortive attempts to standardize a field sequential system (that is, a system in which the different pictures corresponding to the three primaries—red, blue, and green—are transmitted one after the other at high speed), advanced studies were resumed in the United States for the development of a system that fulfilled certain standards:

1. Correct reception of color transmission by existing black-and-white receivers (compatibility).
2. Correct reproduction of black-and-white transmissions by color receivers (reverse compatibility).
3. No changes in the frequency plans in the bands allocated to television (in particular, retention in the United States of the 6-MHz channel).

On this basis, the group of scientists forming the NTSC worked out a system where optimum use was made of certain physiological properties of the eye and of coding possibilities, thus achieving a remarkable saving in the radio spectrum.

We shall now deal at a little greater length with some of the basic principles of NTSC, since they are to be found in most systems. We know that the resolving power of the eye for patterns with only color contrast is considerably less than when there is a luminance contrast. Moreover, if we consider a small colored object, the smaller it is the more difficult it is to assess its hue correctly. This means that for the finer details of a color picture, although the luminance should be reproduced with the maximum resolution, a reasonable loss of chrominance resolution is not a disadvantage. Continuing further with this analysis, we notice that the resolving power of the eye for color is lower for some colorimetric variations (green-purple axis) than it is for others (cyan-orange axis); this property has also been used with profit in the coding of the signal.

With respect to transmission possibilities, the NTSC turned to advantage the well-known fact that in the spectrum of a monochrome television transmission, the power is primarily concentrated in the neighborhood of multiples of the line frequency—the logical consequence of the fact that a television signal is mainly a recurring line-frequency signal. Advantage was taken of this property by using the "dips" for transmitting the chro-

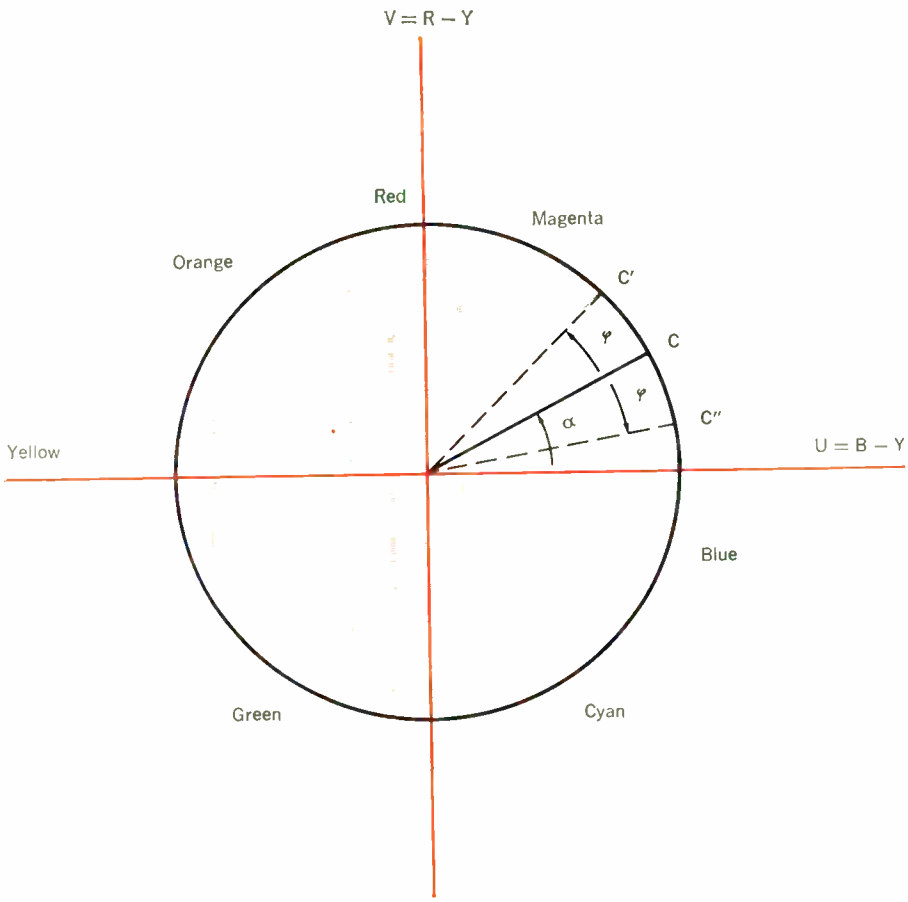


FIGURE 1. Simplified color diagram, NTSC system.

minance information on a subcarrier with a frequency an odd multiple of the half line frequency, $f_H(2n + 1)/2$. On the screen of a monochrome receiver, this subcarrier takes the form of a dotted structure superimposed on the ordinary picture. But, because the frequency of the subcarrier is an odd multiple of the half line frequency, the lighter parts of the dotted structure appearing on a line will be situated exactly under the darker parts of the preceding line. The interlaced structure of the interference pattern thus observed on the picture is much less noticeable than the series of vertical lines that would occur if the subcarrier frequency were a multiple of the line frequency.

To secure a fully compatible color picture, a normal monochrome picture is transmitted in which the luminance signal (transmitted in a bandwidth of 4 MHz) has a voltage Y given by addition of the voltages R, G, B corresponding to each of the three primaries (red, green, and blue) during the scanning of the picture by the television camera.

$$Y = 0.30R + 0.59G + 0.11B$$

The reason for reconstituting Y from R, G, and B is mainly a practical one: a television camera generally contains a device for separating the incoming light ray into three colored rays (red, green, and blue), each of which falls on a scanning tube producing the voltages R, G, and B.

To reproduce a color picture, it is necessary to have three signals R, G, and B at the receiver, or what comes to the same thing: three linear combinations of these signals. One of these combinations is already transmitted by the luminance Y, so it remains to transmit two other combinations of R, G, and B (or of Y, R, G, and B), which are often known as color difference signals.

During the active part of the picture, a chrominance signal consisting of a subcarrier of a frequency of about 3.58 MHz is superimposed on the black-and-white signal. This subcarrier is actually composed of two waves in quadrature, each of which is modulated by color difference signals known as I (in phase) and Q (quadrature).

$$I = -0.27(B - Y) + 0.74(R - Y)$$

$$Q = 0.41(B - Y) + 0.48(R - Y)$$

I and Q are the amplitudes of the two orthogonal quadrature components of the chrominance signal. They correspond approximately, on a color diagram such as the simplified one in Fig. 1, to the cyan-orange and green-purple axes just mentioned. The lower resolving power of the eye for color edges situated in the neighborhood of the green-purple axis has justified a reduction in the transmitted frequency band for the corresponding Q signal (0.5 MHz instead of about 1.3 MHz for I).

The composite chrominance signal is thus written

$$Q \sin(\omega t + 33°) + I \cos(\omega t + 33°)$$

where ω is the angular frequency of the subcarrier.

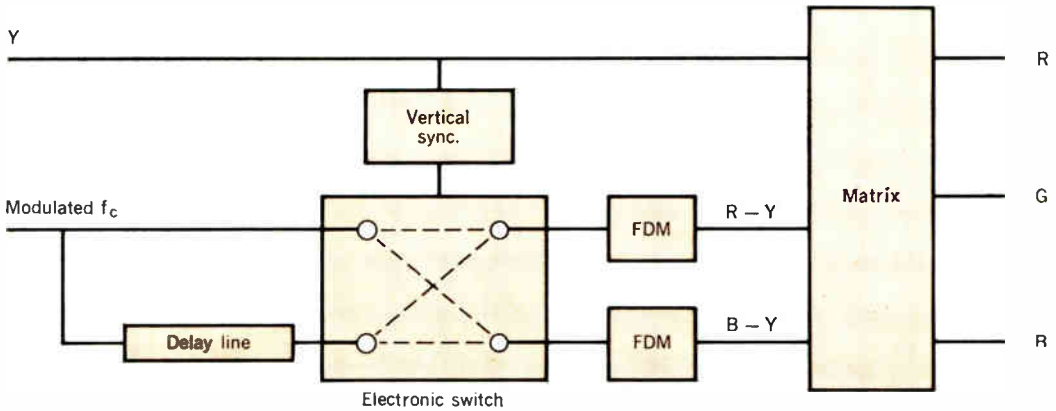


FIGURE 2. Block diagram of SECAM III system.

The signal is in effect an amplitude- and phase-modulated wave; the phase represents the characteristic angle of the color on the diagram in Fig. 1—that is, its hue—and the amplitude represents the degree of saturation.

The two sets of information it contains can be reproduced by "synchronous detection." For example, an electronic multiplication of this signal can be made by cos $(\omega t + 33°)$ and sin $(\omega t + 33°)$, respectively, which results fn a reproduction of I and Q. It is then possible to make a matrix of Y (obtained by normal detection), I, and Q to reproduce the three primaries R, G, and B, which must eventually be fed to the three control grids in a tricolor viewing tube.

This process is possible only if the oscillators used for producing the auxiliary signals in the electronic multiplication are stable in phase. This condition is achieved by having the phases of the oscillators restored to their correct relationship before the beginning of each line by a reference oscillation (burst), at the subcarrier frequency, included in the blanking signal.

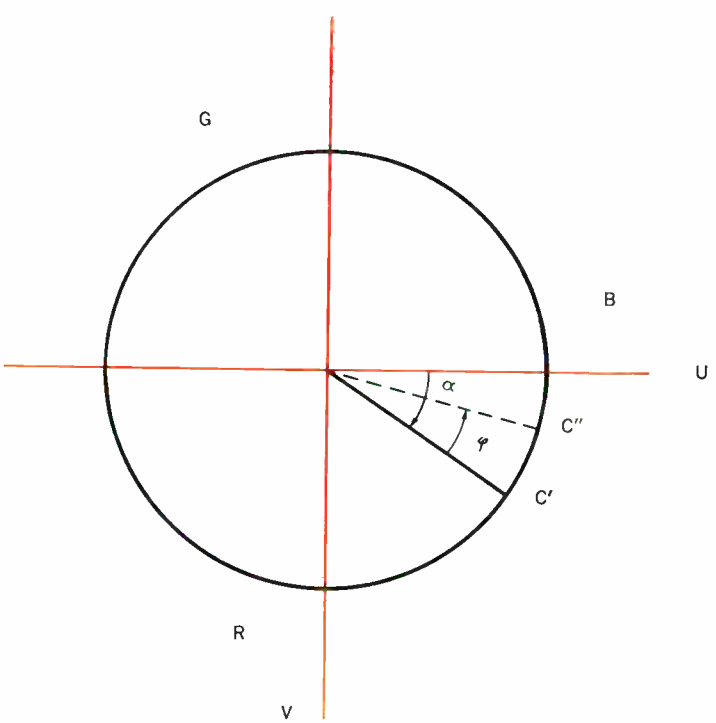
One of the properties of the NTSC system is that because the chrominance alone is transmitted on the subcarrier, any interference or alteration of this subcarrier will not have an appreciable effect on the luminance. The system is said to comply with the "constant luminance principle." In fact, it very nearly fulfilled all the principal conditions that had been set when studies were first started. It marks a decisive stage in the development of color television because when other countries finally come to adopt a system, they will undoubtedly adopt one that is compatible and in which the radio emissions are contained in the same channel as the corresponding black-and-white emissions.

A number of disadvantages in the NTSC system have been pointed out—in particular, difficulties inherent in synchronous demodulation. Actually, its only noteworthy defect is its sensitivity to phase variations, for part of the color information is bound up with the phase modulation of the subcarrier. Considerable unwanted phase variations in transmission often do occur because of multipath propagation, long-distance cable or radio relay transmission, etc., which take the form of color errors at the receiver. This dephasing can be troublesome in that, generally speaking, the errors may not have the same values for the different luminance levels, in this case, the color varies with the luminance (differential phase distortion).

We have gone into a little more detail about the NTSC system because it was the first truly compatible system put into regular service and because all the other systems subsequently studied use most of the basic NTSC principles.

Prior to 1956, a system was devised (TSC system) with two subcarriers, each of which carried part of the color information; for example, $B - Y$ and $R - Y$. The advantage of this system was that it replaces the synchronous demodulation of the NTSC system by two conventional detectors. However, it had shortcomings: irregularities in the amplification of the two subcarriers, whether due

FIGURE 3. Simplified color diagram, PAL system.



to propagation or to the amplifiers, caused color errors; moreover, compatibility was more difficult to obtain because of the complexity of the interference pattern due to the presence of two subcarriers.

Finally, we must mention a more recent version of the NTSC system. In 1963 a proposal was made to add a reference wave with the aim of reducing differential-phase distortion effects. During the recent debates in Oslo this additional reference transmission (ART) system was proposed by the United States delegation as a compromise version of the NTSC system, which was less subject to the shortcomings of that system and was at the same time compatible with it. The proposal, however, did not receive support at that time. With respect, more particularly, to differential phase distortion, mention should be made of the "multiburst" technique, in which the normal color burst is accompanied by two others, one at the gray level and one at the white level, thus providing phase references from which differential phase corrections can be made at the different levels.

SECAM III system. In the SECAM III system the luminance information is transmitted in the usual manner. The chrominance information is also transmitted on a subcarrier f_c, but instead of being modulated by two color signals, it is modulated, for one line, by one color signal $R - Y$ only, for the next line by $B - Y$, then by $R - Y$, and so on. It is necessary to transmit only a line identification signal rather than a burst. The color picture is reproduced at the receiver by means of a delay line, which acts on the subcarrier by retarding it by one line. Thus, for a given line, the direct signal is $R - Y$ and the retarded signal $B - Y$. On the following line the situation is reversed. In this way it merely requires an electronic switch to have two permanent outputs corresponding respectively to $R - Y$ and $B - Y$ (see Fig. 2). This system has required the design of ultrasonic delay lines, working on the subcarrier. It should be noted that no very great precision is required in the delay because the phase of the signal applied to it does not carry information.

In the present version of SECAM III the subcarrier is frequency-modulated. To diminish the visibility of this subcarrier, which is always present, the phase is reversed at every third line and also at each field.

By its very nature, such a system is very little affected by phase variations. Moreover, it is possible to make a direct recording of SECAM III signals on equipment designed for monochrome transmissions. On the other hand, since only one set of chrominance information (instead of two) is transmitted on each line, two successive lines have to be transmitted to re-establish the chrominance information in a single line. This means that in the vertical direction the resolution is reduced by half, which in itself is tantamount to applying, in the vertical direction, the chrominance reduction applied to the horizontal resolution, taking into account the reduced resolving power of the eye for the chrominance.

The SECAM III system was the first to make intensive use of a delay-line device by making practical use of a part of the well-known redundancy of television pictures (a line of a television picture differs very little in general from the preceding line, and in the same way two successive fields are very similar).

The materials problems inherent a few years ago in delay lines have now been resolved: in calculating the cost of a receiver chassis, the prices indicated for these lines are $3 to $5.

The PAL system. In the NTSC system it was seen that coding was done by means of two components modulating two waves in quadrature. In the PAL system, which is in essence a version of the NTSC, U and V are these two components. (They are not exactly the same as components I and Q of the NTSC, but this does not alter the basic principle of the system.)

$$U = 0.493(B - Y) \quad \text{and} \quad V = 0.877(R - Y)$$

A specific color can be represented in a simplified manner on a vectorial color diagram (Fig. 1) by a vector whose angle α denotes the hue; the amplitude gives the color saturation.

In the case of the NTSC system, a color to be transmitted corresponding to a hue C is characterized by the angle α. Any random dephasing (due to propagation, transmission, etc.) equal to φ results in the reproduction of an incorrect color C', corresponding to $(\alpha \pm \varphi)$.

The fundamental characteristic of the PAL system is that at each new line the direction is reversed along the V axis; in other words, it makes a symmetry of the diagram in Fig. 1, which becomes the one shown in Fig. 3. The hue of the color to be transmitted is always characterized by the angle $-\alpha$, but the phase distortion φ, which will affect the corresponding line, will result in the reproduction of an incorrect color C'', corresponding to $(-\alpha + \varphi)$ on the reversed diagram in Fig. 3—that is, $(\alpha - \varphi)$ in Fig. 1.

When the image is reproduced on the screen of the receiver a distortion compensation takes place. The reason is that if the phase error is not too great, the eye sees the average of the two hues reproduced; the impression it receives, except for a slight degree of desaturation, is the hue that would have been reproduced in the absence of distortion. This is called "simple PAL" compensation.

If the phase errors are serious (more than 15° to 20°), the method of compensating by the eye is insufficient. Then, as in the case of SECAM III, a delay line that retards the modulated subcarrier by exactly one line (64 μs) may be used. An electronic switch also ensures the reversal of the signal V at each new line so that this signal will be correctly reproduced. By adding and subtracting the original signal and the outgoing signal of the delay line, two signals at the subcarrier frequency are obtained, amplitude-modulated respectively by U and V (Standard PAL).

We must mention another version of PAL (New PAL), which can eliminate the saturation errors that may occur in Standard PAL. In the PAL system two auxiliary signals are necessary: the burst at the subcarrier frequency and the frequency identification signal $\frac{1}{2} f_H$ for the phase blocking of the electronic switch. The frequency chosen for the subcarrier is

$$f_c = (284 - \tfrac{1}{4}) f_H + \tfrac{1}{2} f_V$$

The NIIR–SECAM IV system. Still another system, the NIIR–SECAM IV, consists of the transmission on two successive lines of (1) a signal amplitude- and phase-modulated with suppressed carrier, resembling the NTSC signal (signal m), and (2) a reference signal r, carrying phase reference information that is used to demodulate the signal m at the receiver. The signal r is amplitude-modulated in the same way as the signal m.

Decoding is accomplished simply by multiplying the two signals m and r. It obviously necessitates a delay line corresponding exactly to one line period and a switch-over device for ensuring the permanent presence of a signal m and a signal r. Since phase distortion is expected to affect both signals in the same manner, no impairment of the picture is to be anticipated.

Although less extensive tests have been made on this system than on the others, it can be said that the quality of the color picture, its compatibility, and its resistance to interference and distortion are comparable to those of the three other systems.

System comparisons

In many cases the differences between the three most predominantly proposed systems—NTSC, SECAM III, and PAL—can be deduced from the basic principles of the system of coding used for the color signals. CCIR Report 406, adopted in Oslo, briefly compares the three systems from what are considered to be the most important technical aspects. In this connection attention is drawn to the important contribution made by the EBU ad hoc Group on Color Television. The numerous comparisons consisted in the main of subjective tests in which the results were worked out on a grading system with regard to quality, impairment, or comparison.

It was agreed that the three systems satisfied the three basic conditions:

1. Color and monochrome systems should be compatible.
2. The signal should be composed of a luminance signal and two signals carrying the color information, insofar as possible in accordance with the constant-luminance principle.
3. The chrominance signal should share the luminance frequency band.

We shall now study the main differences in a little more detail. The *compatibility* of the three systems is satisfactory. Nothing very significant emerged from all the tests made in this sphere, but possibly NTSC has a slight advantage. The quality of the color pictures is satisfactory in the three systems under good transmission and reception conditions. The only differences lie in the horizontal resolution and the alteration of the vertical edges, where SECAM III may be at a slight disadvantage.

With respect to the *receivers*, it should be noted that the presently available NTSC or simple PAL receivers generally require hue and saturation controls, that the Standard PAL receivers require a saturation control, and that neither of these controls is necessary with SECAM III. Estimates of receiver manufacturing costs have been made on the basis of the NTSC-type receiver. An average of the estimates made in five countries shows an increase in price of about 2 percent for SECAM III and 4 percent for the PAL receiver. It must be added that these estimates are very approximate, since they are based mainly on the assumed cost of the delay line.

On the other hand, with regard to *transmission between fixed points*, the SECAM III and PAL systems have the advantage of being less sensitive to defects connected with phase distortion and could be adapted to existing network transmission systems with a minimum of network line improvements and signal correction.

With regard to *studio equipment*, there is little choice between the three systems. In the matter of *tape recording*, however, there are differences. SECAM III signals recorded on a standard video tape recorder cause no problems and provide a good recording, whereas with NTSC considerable extra equipment has to be added. With this equipment the PAL recording is generally of better quality than the NTSC recording.

It is agreed that *transmitters* can easily be altered for the transmission of any of the three systems, except perhaps in the case of the variant of the 625-line system where the subcarrier is relatively near the upper end of the video channel.

Susceptibility to interference and noise varies very little between the three systems. However, under conditions of multipath reception (for example, in mountainous country or in urban areas), the SECAM III and PAL systems generally show a slight advantage. Finally, if the level of interference is such that the picture is naturally already bad, SECAM III is slightly more sensitive to noise.

The *exchange of programs* with different standards is a complex operation and often results in a deterioration in picture quality. It should be pointed out that in all cases when the monochrome standards differ—for example, between a country with a 525-line system and one with a 625-line system—the exchange of color programs will require delicate transcoding. The fact that both countries have the same color standard differing only in frame frequency—both NTSC, for example—will hardly simplify transcoding.

The *additional receiver costs* of SECAM III and PAL receivers over NTSC receivers mentioned previously are on the basis of having a receiver capable of receiving only one standard. If the countries of a region of the world do not agree on a single standard, then multiple standard receivers will be necessary in border areas between two systems. This situation already exists with monochrome in some areas. These multistandard monochrome receivers have been found to cost the television viewers approximately 130 percent as much as a single standard receiver. Multistandard color receivers may probably be expected to have even a greater additional cost.

During the discussions concerning the comparison of systems, priority was often given to future prospects. In this sphere, it can be said that the NTSC system seems more easily adaptable to the single-gun tube. The PAL system is a little less adaptable and the SECAM III system is still less adaptable. Further work is still needed. The NTSC system preserves the vertical resolution, in color, to a greater extent than does the SECAM III system, with the PAL system as intermediate. The PAL and, above all, the SECAM III systems will permit video tape recording more easily.

What are we to make of all these comparisons? That none of the three systems is technically overwhelmingly superior to the other two is, perhaps, a pity. If one of the three showed such an advantage, it is likely that a technical body such as the CCIR would have recommended that system. Unfortunately, this was not the case, and it was obvious at the London meeting in 1964 that other factors were bound to play a part in forming the opinions of many of the delegations taking part in the debates.

One factor, in 1963, was the haste with which some of the European countries wanted to introduce color television service. Another is that some countries had already been engaged in research for many years, often

with large commercial investments. These studies had, with a few exceptions, originally been carried out with NTSC in mind, which would explain why, at the London meeting, some countries proposed that system mainly because there seemed no adequate reason to abandon a system that had stood the test of time. On the other hand, in Vienna in 1965, this argument was dropped by some participants and a preference was shown for the SECAM III and PAL systems, which are less sensitive to phase distortion.

At the beginning of the 1966 Oslo meeting, PAL and SECAM III were still the leading European competitors. However, early in the discussions certain countries that had no strong preference for either system proposed that SECAM IV might be considered; but it was not preferred either by France or by the U.S.S.R. because both countries considered that SECAM III was more nearly ready for commercial production and had already been used for a number of demonstrations. The countries that supported SECAM III (France and the U.S.S.R. in particular) stated that they were prepared, if absolutely necessary, to adopt SECAM IV, provided that the other countries abstained for a limited period from measures—particularly in regard to the manufacture of receivers—that might run counter to the research and development plans to which they would be committed. The delegations of the United Kingdom and the Federal Republic of Germany decided that they could not accept those conditions.

The replies to the various questionnaires distributed to participants in the Oslo meeting showed clearly that no tendency toward a common system could be perceived. As pointed out earlier, the technical, and even cost, differences between the various single systems were marginal, so that the arguments put forward in defense of the different positions were, for the most part, extra-technical. As a result, debates were inclined to be based on considerations that were outside the terms of reference of the CCIR—for, as its name implies, the CCIR is a *consultative body*, whose duties, under the provisions of the International Telecommunication Convention (Montreux, 1965), "shall be to study *technical* and *operating* questions relating specifically to radiocommunications and to issue Recommendations on them."

These terms of reference were referred to by the head of one delegation who appealed to the meeting to maintain the purely technical nature of the CCIR's work and to attempt to reach a decision as to a recommended system of color television on the basis of purely technical considerations, all other considerations being external to the mandate of the CCIR.

If there had been unanimous agreement, the CCIR could have recommended to administrations the use of a single system by issuing a recommendation based on technical considerations alone. However, since there were no overriding technical differences between the various proposed systems and there was no unanimity of opinion among the various delegations, such a recommendation (even on a compromise basis) was not possible.

It is believed that the CCIR has performed its task well in promoting the study of the technical and operating questions concerning television systems, originally set out in its Recommendation 29 in Stockholm in 1948. Perhaps it has done it too well, since so many systems that are satisfactory from the technical viewpoint have been developed that administrations cannot agree on any one of them from technical considerations alone. However, it is hoped that it is still not too late for administrations to again look at CCIR Recommendation 29 and to evaluate the many worthwhile reasons for adopting a single worldwide standard, particularly from the *long term* point of view. For we find ourselves at present faced with the prospect—fortunately not yet a reality—of a multiplicity of color television standards, in part due to the lack of standardization in black and white, but above all due to the inability of the responsible agencies of the various governments to agree on a universal color television system.

Conclusion

It might be useful, in conclusion, to consider, from a purely cost point of view, some consequences of this multiplicity of standards.

In the first place, a certain amount of additional technical equipment, such as transcoders and perhaps, on occasion, slightly more complicated equipment, will have to be used by the broadcasters, thus increasing the expenses of their electronic apparatus. However, this additional cost will be small in relation to the total cost of such material and will become even less significant if one takes into account other expenses at the broadcasting end, such as studios and general administrative outlays, and also the financial resources of broadcasting organizations.

On the other hand, any cost increase due to multiple standards will have to be added to the viewer's only direct investment in color television—that is, his receiver. Nor will this extra cost be limited to viewers living in areas where programs on various standards can be received, for the manufacturer of receivers will be confronted with one of three choices:

1. To manufacture sets for only one standard; this will mean that he will limit his market and hence be obliged to spread his production overhead costs over a limited number of receivers.

2. To manufacture several lines of receivers, each being able to receive one standard. In this case, he will enlarge his potential market, although inadequately providing for markets where programs of more than one standard can be received. It is obvious that the manufacture of several types of receiver will be more costly than that of a single type.

3. To produce multiple-standard receivers, thus perhaps covering all potential markets; however, the receiver itself, as previously pointed out, will be inherently more expensive.

It can be seen that whatever solution is chosen, the manufacturers will be faced with the necessity of producing receivers under less than optimum economic conditions and, as a consequence, their markets will be reduced.

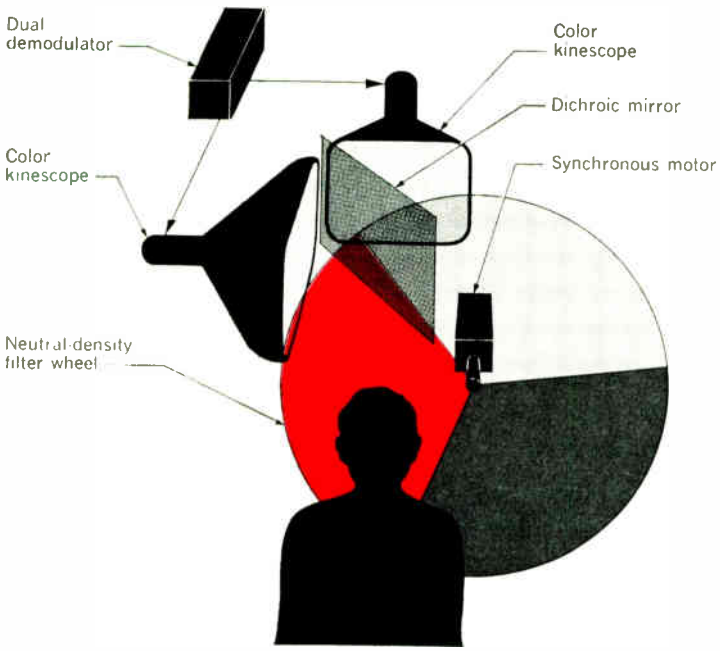
The inevitable conclusion, therefore, is that the multiplicity of standards will cause, all over the world, a higher cost for receivers. Thus, all viewers, wherever they may be and whether or not they live in a multiple-standard area, will have to pay to some extent for the lack of standardization.

Revised text of a paper published in *Telecommunication Journal*, January 1967.

Universal color television: an electronic fantasia

In a zealous attempt to simplify the hitherto complex problems involved in achieving color television standardization, an intrepid American electronics engineer summarizes what has happened to date and offers a foolproof solution

Joseph Roizen *Ampex Corporation*



NUTSEQAMIR receiver.

As can be seen from the preceding article, color television standards—and the repeated failures to arrive at a single standard—have been for years the concern of many experts in the fields of electronics, finance, and politics. One undisciplined representative of the electronics discipline (who, we hasten to add, has written a number of more esoteric articles on color television) has decided that the time is finally ripe to tell the true story.

In my role as an interested bystander, I have been pondering the global color television dilemma now facing the technical experts of the world. Within the scope of my own memory, I can recall the controversy over the three-color rotating filter proposed by one of our largest networks and the subsequent technical battle between the pure Electronic Knights and the dastardly Mechanical Monsters. In the United States the FCC refereed the fight and, after a hurried decision for the MMs, they reversed themselves and ruled for the EKs.

We had glorious living color television and, like most modern things, it got labeled by its initials, to be known henceforth as NTSC.

While NBC's peacock preened his flashy feathers, our European friends started having committee meetings about what kind of tinted television they were going to have. A few hardy members were even dispatched to our shores intent upon assessing the impact and quality of NTSC. They must have disliked what they saw, because they went back to their colleagues with reports so detrimental that NTSC became synonymous with Never Twice the Same Color (oops, Colour).

The gathering snow

The NIH* factor now took over. Suddenly it became imperative to invent a new system that would show those Americans a thing or two. Color systems spewed out of European laboratories faster than early American space failures. Suddenly the technical literature was full of things like FAM, ART, PAL, and BS (Bicolor Sequential, a system developed in Mexico).

The French started the ball rolling with a system that sequenced the color signal on a line-by-line basis, while a one-line delay memorized the last line and added it to the subsequent one.

The Europeans learn fast; they may not like our technology, but they do dig initializing, so the Sequential and Memory system became SECAM. Of course, it had it all over NTSC, except that there were a few niggling problems, so while the boys in the front room were demonstrating to one and all what it could do, the back-room crowd was cleaning up the details. In quick succession, there was a SECAM I, II, III, and so on. Each small deficiency fostered some ingenious cure.

Politics and ocher

It also became a political issue, with government-sponsored efforts at the ministerial level to push SECAM. While French engineers made demonstrations from

*Not Invented Here.

Munich to Moscow, plenipotentiaries followed up with bids of a free system to Argentina and an offer to withdraw from NATO if the Soviet Union would adopt line sequential color TV. It was all done with such zeal that SECAM was soon paraphrased as "Something Essentially Contrary to the American Method." It is my belief that General DeGaulle's visit to Moscow was not for the purpose stated (politics is a devious business) but to get the Russians back on the SECAM track from which they have wavered lately.

The Germans were not about to stand idly by. With Teutonic thoroughness, they re-examined NTSC to see what could be salvaged. After all, we still have a lot of troops over there and they wouldn't want us to get mad at them. NTSC is not without its military adherents, all the way up to brigadier general. There was a false start with a system known as FAM. It retained most of NTSC's good features, but, by this definition, it also didn't eliminate its most obvious ills. Anyway, it came out of a Hamburg laboratory and everyone knows that the really exciting things in West Germany today are taking place in Berlin. Well, it happened: an American dance craze known as the twist hit Berlin (brought in by a Pan-Am stewardess, no doubt) and with a sudden inspirational flash of crystal clarity, a Telefunken scientist combined NTSC with the twist and got PAL. Just twist the phase of every alternate line 180 degrees and the hue errors will cancel themselves out. The PAL system was ready for the next meeting in Vienna and soon it would stand for "Peace At Last."

Tales from the Vienna woods

Plenty of ballyhoo preceded the Vienna meeting of Study Group IX. The *London Times* carried half-page SECAM ads on why two knobs were better than four, carrying the technical fight to the public. After all, they would have to buy the sets when a service was established. They tried to get it on the national ballot, but Harold Wilson turned it down, feeling his victory margin was already too thin.

Weekly news bulletins charted the course of the NTSC color caravan that was touring Europe to eradicate its derogatory reputation and re-establish itself by technical legerdemain. The EIA came out foursquare for the U.S. system and technical treatises made NTSC appear as a virtuous virgin in a pale blue gown (white would bloom) about to be desecrated by the nefarious signal switching interlopers from across the Atlantic. President Johnson, when asked about the color TV problem, said that the law was on the books. The FCC (Federal Color Committee) had ratified it and it was now up to the engineers to go forth and spread the spirit of NTSC.

Only PAL lay low in the manner of a "sleeper" at a horse race, whose backers know it can win when the Munsell chips are down. Polltakers started giving odds on what system was ahead and, with an editorial in *Pravda* praising SECAM, it seemed as though the French would walk away with the contest.

The orders from Camden (or was it Princeton?) were clear; carry the fight into the enemy camp. Gathering all of its adherents, which then still included the BBC, the polychromatic pachyderm, stuffed with its video vitals, wended its way to the inner courtyards of Televidyona Moskva. They should have known better: it was the dead of the Russian winter and NTSC shared the fate of Bonaparte. The retreat was almost as ignominious.

Vienna seemed but a formality. With the Soviet Union's decision came the concurring assent of the Eastern European satellite countries. Even such scientific nations as Mali, Ghana, Tanzania, and Upper Volta declared their support for the French system.

As a last-ditch measure to keep color TV from going down the tubes, NTSC and PAL joined forces, keeping SECAM from a clear-cut majority. This new combination came out as QUAM, although they should have thrown an "L" in there, considering the misgivings of everyone concerned about this pigmented pussyfooting. Nevertheless, the Quamdry was as insoluble as ever.

A breeze from the east

Meanwhile, back in the Urals, the subjects of all this exhortation to jump on the tinctured troika had their own prism to grind. They wanted vivid video for the 50th anniversary of the 1917 revolution. If my addition is right, that makes 1967 the target. It seems almost traitorous to adopt a capitalist system, even in a technical field! Last year in the monthly bulletin of the British Television Society, a new Russian color TV system was described. The publication stated that the revelation was made at a recent wine and cheese party; after reading the technical details of what they had dubbed as SEQUAM, I was convinced that a lot more wine than cheese was consumed at this party.

But lo, and behold, the February 7, 1966, issue of *Electronics*, a magazine not noted for levity, carried a story in its "Electronics Abroad" section detailing a Soviet proposal for a system labeled NIR. What NIR stands for was not stated, but then the Russians always were secretive. I suppose it could mean Not Intentionally Red, or Now Invented in Russia. NIR ends up requiring NTSC-type modulation, PAL switching, and SECAM delay lines. They are leaving nothing to chance! It's no wonder that President DeGaulle went to Moscow.

The solution—or color me brilliant

As I sat contemplating engineer's inhumanity to engineer, I hit upon a solution, one that would transcend all national barriers, let no one lose face or fortune, and combine, out of all of the systems, those elements of which their advocates are so proud. This system, in keeping with established practice, is called NUTSE-QAMIR (National Universal Television Sequentially Encoded, Quadrature and Amplitude Modulated and Intermittently Reversed). To give this system the advantage of a clean new start, I would change the primary colors to vermilion, emerald, and ultramarine, with tertiary colors of heliotrope, turquoise, and ocher. I'll probably have just as much trouble convincing people that vermilion and emerald produce ocher as I now have saying red and green make yellow.

The luminance signal will be a matrix of 30 percent vermilion, 60 percent emerald, and 10 percent ultramarine. I could never see that 59 and 11 percent jazz in NTSC: Who's going to miss one percent of blue, anyway? This will certainly make the mathematics of the system a lot easier.

The subcarrier frequency (always a delicate choice) is selected as 4.432 795 384 265 MHz. The subcarrier is

both alternately phased at 0° and 180° and has an FM swing of ±700 kHz. To avoid the old bugaboo of dot crawl, an offset frequency of 0.5 times General DeGaulle's next odd birthday is added.

The burst signal now consists of a dual burst on an elongated back porch. Most home receivers are overscanned anyway, so a few microseconds lost to more blanking will hardly be noticed. (Luckily, the vertical blanking interval remains the same.) Burst *A* consists of 7 Hz of subcarrier at 0° on line *N* and 180° on line *N* + 1. I haven't yet decided what to do with the rest of the lines, but I'm sure it will work out. Burst *B* has 8 Hz of subcarrier +300 kHz on *N* and −300 kHz on *N* + 1. This gives the FM system a good chance to exercise on a regular basis.

An innovation is suggested for radiation of the NUTSEQAMIR signal. The transmitting antenna is angled at 90° at half altitude. This accommodates both horizontal and vertically polarized systems, bringing the United Kingdom into our TV family.

Naturally, I have given a great deal of thought to the home receiver. So much engineering and ingenuity have gone into past systems that I hated to throw anything away. That left only one path open. The home set will have two color kinescopes and two demodulation systems. One picture tube will display the quadrature signal and the other the sequential. A dichroic mirror will optically integrate the images. The extra brightness can be reduced by a whirling neutral-density filter in front of the mirror. This last touch is added only to appease the CBS holdouts for mechanical color television and may be made optional.

The controls on this receiver are somewhat unique. There are eight operational knobs. The familiar brightness, contrast, hue, and saturation need no explanation. Below these are the BGS, Erh, DeG, and Kos knobs. These do require some clarification. The BGS knob (in case you are wondering, this knob was named after Brigadier General Sarnoff in honor of his contributions to color television) activates the four normal knobs, allowing viewers to chose whatever hue and saturation combination they desire, in true democratic fashion as outlined in The American Bill of TV Rights. (A copy may be had by sending one dollar to the author.)

The Erhard knob deactivates the hue knob, allowing the alternate phase correction to set the color to the proper tint. It has been suggested that a small audio loop recorder repeat every half hour "Achtung das ist eine deutsche entwicklung," but this may be carrying things too far. The DeGaulle knob disconnects both the saturation and hue knobs, the sequenced signals now controlling these elements. If the audio loop recorder is included, a second track could softly play the "Marseillaise" as background music. This leaves us with only the Kosygin knob; it disengages all the normal knobs and puts in preset voltages. The principle is that the government knows best at what levels things should be set (it was established during one of the seven-year plans), and no individual user should fool with them. In addition, the Kos knob increases the red gun output by 30 percent, thus assuring the correct political shade for the program.

The cost of this receiver has been calculated on the same basis that was used for the SECAM and PAL sets. NTSC was taken at 100 percent and the other units were rated at 106 percent and 108 percent. The NUTSEQAMIR receiver would be exactly 314.16 percent of the NTSC base. This would assure a generous margin of profit for manufacturers, giving them the π in the sky that they are always looking for.

I once heard a description of hell as a place where the French are the engineers, the British the cooks, the Germans the police, the Russians the historians, and the Americans the lovers. If one is to take this parody on ethnic characteristics as the gospel, then heaven must surely ascribe more fitting national occupations. Here, the French are the lovers (very few Frenchmen will dispute this), the British the historians (fair play and all that), the Germans certainly qualify for the engineering post, and the Americans could take on the policing duties (witness Korea, Viet Nam, and Santo Domingo). This leaves the Russians to do the cooking, unless we let the French do that too (after all, loving is hardly a full-time occupation). Anyway, there may never be any Russians in heaven, since they don't believe in it as a matter of national policy.

The views expressed in this article are solely those of the author and do not represent those of the Ampex Corporation.

ODE TO THE CCIR COMMITTEE

As a patriotic citizen
NTSC is my lot,
And I must help support it
With everything I've got.

But the Siren Song of SECAM
Lures me from afar,
It's so easy to record it
On a standard VTR.

I may do a sharp reversal
By degrees one-hundred eighty
And decide that PAL is really
The system much more weighty.

And now that NIR is calling
In an alphabet Cyrillic,
Could it be that after all
It's the standard most idyllic?

But if you all suffer jointly
From such indecisive QUA(L)MS
Then remember NUTSEQAMIR
And give thanks for little alms.

It will solve all of your problems,
It will help to smooth the way,
When a single color system
Is the order of the day.

I do not ask for royalties,
I do not seek for fame.
My reward will be sufficient
When receivers look the same.

And when the system's working
To perfection, on the track,
We may find out that Yogi Bear
Is brown instead of black.

High-power lasers—their performance, limitations, and future

The high-power laser has captured the imagination of R&D men, as well as the public, resulting in applications ranging from surgery to the villain's secret weapon in a James Bond movie. The problem is now to find a laser material that can withstand the damage entailed

F. P. Burns *Korad Corporation*

Their many new, imaginative applications have led to the development of lasers with power outputs exceeding the ability of the laser materials to withstand damage for more than a few shots. The apparent necessary remedial steps would be to learn how to increase damage threshold and to determine how to improve laser parameters that will not degrade life. This article discusses how high power is attained in a laser system and it reviews the data on laser performance to establish the principal causes of failure and to determine how to rectify them. An integral part of the analysis is a detailed treatment of the measurement and interpretation of luminance.

The announcement of the first working laser[1] and the successful operation in the *Q*-switched mode[2] opened the gates for a flood of imaginative applications, many of which required the generation of extremely high-intensity output beams. This led to the development of lasers with power outputs that exceeded the ability of laser materials to withstand damage for more than a few shots. The situation clearly was not acceptable. High power was still in vogue, but cost consideration dampened the ardor of the budget makers.

Two courses of action were apparent: a study of the damage mechanisms and their relation to material properties in order to learn how to increase damage threshold, and a reappraisal of application requirements to determine what laser performance specifications are essential to attain desired objectives and the satisfaction of these needs by improving laser parameters that will not degrade life. The materials study was pursued vigorously by several groups and some immediate beneficial results were obtained. The removal of impurities known to be foci of damage occurrence and improvements in material preparation techniques resulted in an increase in damage threshold in both Nd-doped glass and ruby. A review of several of the more important applications reveals that in some cases (e.g., range finding and plasma diagnostics) what is really required is luminance (power per unit area per unit solid angle), because this is the parameter that directly determines the intensity that can be delivered by the laser at a given distance. In other cases, however, such as photography or damage studies, energy delivered in a time considerably longer than that associated with a *Q*-switched pulse is acceptable.

In this article, how one attains high power in a laser system is discussed and the available data on high-power laser performance are reviewed to establish empirically the principal causes of failure, and further, to determine what steps can be taken to retain satisfactory operating life expectancies while meeting system requirements.

Q switching

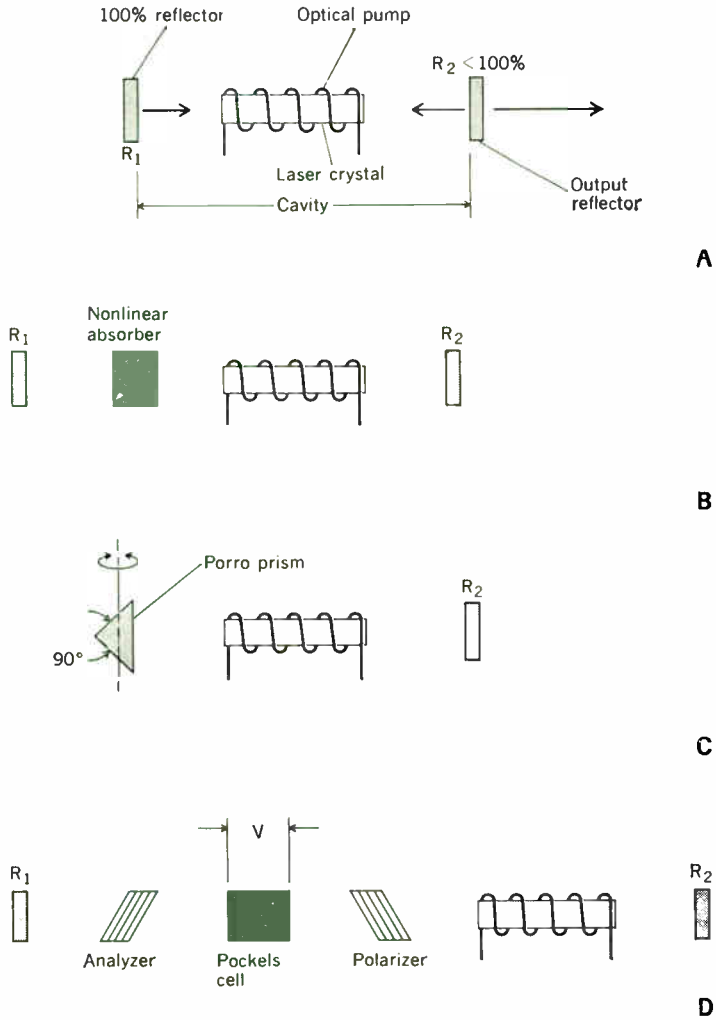
The operating method that is generally used to attain the highest possible output power of a laser is to switch the *Q* of the laser cavity. Figure 1(A) shows a typical laser oscillator; it consists of an optical pump, a laser crystal, and two mirrors, which determine the cavity. The amount of energy that must be absorbed from the optical pump by the laser crystal to initiate oscillation—i.e., the threshold energy—is directly dependent on the losses in the cavity. If these losses are deliberately kept high while the crystal is absorbing energy, large amounts of energy can be stored in the crystal without reaching the condition for oscillation. Suppose that, with the crystal fully pumped, the cavity losses are suddenly decreased. The threshold energy consequently will decrease also and that fraction of stored energy above threshold will be emitted in a short burst of light. The fact that a large amount of energy can be stored and then suddenly released permits the attainment of ultrahigh powers. In conventional operation, energy dribbles out of the rod when threshold is reached and continues to emerge as

long as the crystal is pumped hard enough to be maintained above threshold. In contrast, once a giant pulse is initiated by Q switching, the pulse width is so short that the effect of the flash lamp is negligible.

A laser operating in the conventional mode can deliver up to an order of magnitude more energy than it can when Q-switched. However, whereas a conventional ruby or Nd-doped glass system will typically have an output pulse of several hundred miroseconds, the Q-switched laser pulse width is in the tens of nanoseconds region; thus, the result is a net increase in power by a factor ranging from 10^3 to 10^4. Hundreds of megawatts in single-oscillator and several gigawatts in oscillator–amplifier[3] configurations are readily attainable. The practical limitation is self-damage.

Recently, laser pulse widths of the order of 10^{-12} second were obtained[4] through a method of laser operation known as mode locking. These very short pulses permit higher power intensities without causing excessive damage, but, although there are some interesting applications for such a short pulse, this work is still in the experimental stage and will not be pursued further here.

FIGURE 1. Typical laser cavities. (A) Laser cavity, designed for conventional operation, consisting of optical pump, laser crystal, and two mirrors that determine the laser cavity. (B) Cavity with nonlinear absorber required for passive Q switching. (C) Mechanical Q switching using rotating Porro prism to maintain alignment in one plane. (D) Cavity arrangement using Pockels cell and "pulse-on" method to effect electronic Q switching.



Methods. There are two basic types of Q switching—passive and active. In the former the laser itself effects the necessary change of cavity losses, whereas in the latter an external mechanical or electrical signal produces the required switching. In all cases, the requirements of a fast, substantial increase in the Q of the cavity must be met. If switching is too slow, the energy stored in the laser crystal will dribble out in extended multiple pulses, thus seriously degrading power output. On the other hand, if the change in Q is small, the available energy is also small, resulting in low power output. Having a large amount of stored energy in the laser rod (that is, high gain) tends to decrease pulse width due to the rapid buildup of light intensity in the cavity, all of which further increases the power output of the laser. Thus, to attain high-power output one must pump the laser crystal hard and use a fast-acting, efficient Q switch.

To effect passive Q switching, a material whose optical transmission at the laser frequency is a function of the intensity and duration of exposure to the laser (a nonlinear absorber) is placed between the laser crystal and one of the end reflectors, as shown in Fig. 1(B). The initial transmission is low and thus the material attenuates light directed toward the 100 percent reflector R_1 from the crystal. When the crystal is pumped hard enough to reach threshold, a rapid absorption occurs, creating a condition of high transmission. The sudden decrease of cavity loss lowers the threshold and a giant pulse of light is generated.

Various substances can be utilized for passive Q switching. In liquid cells, cryptocyanine dissolved in alcohol or vanadium thalocyanine dissolved in nitrobenzene for ruby and carbocyanine dye[5] for Nd-doped glass or YAG (yttrium aluminum garnet) have been successfully used. The concentration of the dye—the optical density—in the solvent determines the output power of the laser. The advantages of using liquid cells are their relatively low cost and the narrow spectral line width of the laser output; their disadvantages include dye deterioration, especially on exposure to ultraviolet light and oxygen, lack of control as to when the giant pulse will occur, and the tendency to enhance filamentary laser output (hot spots). Uranyl glass, which requires the use of a separate optical pump to ready it for Q switching by pumping electrons to a given energy level, also works well for ruby.[6] The intensity of the auxiliary optical pump controls the power output of the laser. This method affords easy control of power output, but the glass tends to damage more easily than liquid cells.

Other methods for passive Q switching that have been demonstrated successfully in the laboratory, but are restricted in use, are exploding films, which are good for only one shot at a time, and semiconductor reflectors, in which the increase of reflectivity due to pumping a large number of electrons into the conduction band effects the Q switching. Unfortunately, in the latter case there is a high probability of damage to the semiconductor material.

Mechanical switching. If one of the mirrors of the laser cavity can be made to rotate into precise alignment at

an appropriate speed, single giant pulses can be generated.[7] The rotation of the mirror must be synchronized with the firing of the flash lamp so that the laser crystal is fully pumped when alignment of the mirrors is attained. For example, to produce the highest power in a ruby laser the Q of the cavity must be switched about 600 μs after the lamp is fired.

If the mirror is rotating at 400 Hz and there is an electronic delay of 100 μs between the signal from the transducer on the rotating mirror shaft and the initiation of lamp firing, then the angular setting between mirror alignment and transducer pickup must be about 100°. A Porro prism, as shown in Fig. 1(C), is sometimes used as the rotating mirror to maintain alignment in one plane. The axis of rotation is perpendicular to the line of intersection of the planes containing the 90° angle of the Porro prism. Small angle deviations of the axis of rotation from the laser axis will not affect alignment of the laser cavity. These switches tend to be noisy, especially if the rotor is air driven, and to multiple pulse, but good repeatability and economical operation are in their favor.

Electronic switching. Electronic Q switching permits the laser user to decide precisely when the giant pulse will occur. The principle of operation involves the use of an electric field in an electrooptic material to rotate the plane of polarization of laser light to control the transmission of an analyzer appropriately located in the laser cavity. The Q of the cavity is directly related to the transmission of the analyzer. Electrooptical devices that have been successfully used in this application are the Kerr[8] and Pockels cells. It is also possible to use the Faraday effect[9] to accomplish the same result, employing a magnetic rather than an electric field to activate the device. The optically active material, together with its mount and electrodes, is commonly known as the Q switch.

A typical arrangement in which a Pockels cell is the Q switch is shown in Fig. 1(D). The polarizer and analyzer can be identical components, such as a calcite prism or Brewster plate stack, oriented at right angles to each other. The polarizer assures that the light from the laser crystal is plane-polarized when it reaches the Q switch. The use of a polarizer is essential with lasers that do not emit plane-polarized light, such as Nd-doped glass.

With zero bias on the Pockels cell, the polarized light will pass through the cell and be reflected out of the cavity by the analyzer. This is the off or low Q position. With an appropriate voltage pulse on the Pockels cell the plane of polarization will be rotated 90° as the light passes through the cell; the light then will pass through the analyzer and be reflected back along the axis of the laser by the Porro prism shown. This is the on or high Q position.

It is possible to eliminate the analyzer by biasing the Pockels cell enough to rotate the plane of polarization 45°, to the off position. Removing the bias (i.e., pulse off) would increase the transmission and the Q of the cavity. In this case voltage requirements are half of those for the pulse-on method.

With the electronic Q switch one can cause the giant pulse to appear with a jitter of less than 10 ns.

Power and crystal damage

The critical quantity in causing damage in laser materials is the maximum power intensity (power per unit area) delivered by the laser. Expected life of ruby crystals as a function of power intensity is shown in Fig. 2. The curve is based mainly on the performance of crystals approximately 20 cm long and 1.6 or 1.9 cm in diameter. Data taken with a limited number of 10- by 1.4-cm crystals were consistent with the curve shown. A Kerr cell was the Q switch used under conditions that resulted in pulse widths of approximately 10 ns. A ruby is considered damaged when there is a falloff of 30 percent of output power with the input energy and operating parameters kept constant. Thus, if 10 J/cm² of ruby is required, the corresponding power intensity (one gigawatt) would destroy the ruby in a few shots. This agrees with the data of Avizonis and Farrington,[10] who further show that the threshold energy density for damage to the ruby is increased to 30 J/cm² by increasing the pulse width to 100 ns. They report similar results for Nd-doped glass, as shown in Fig. 3.

The life expectancy for ruby and Nd-doped glass depends strongly on power intensity and to a lesser ex-

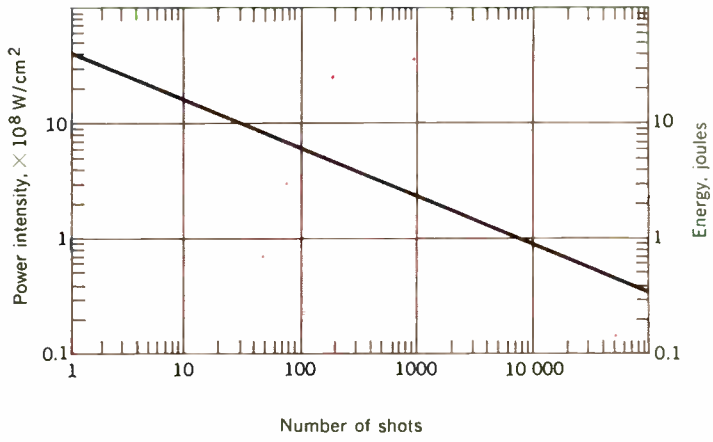
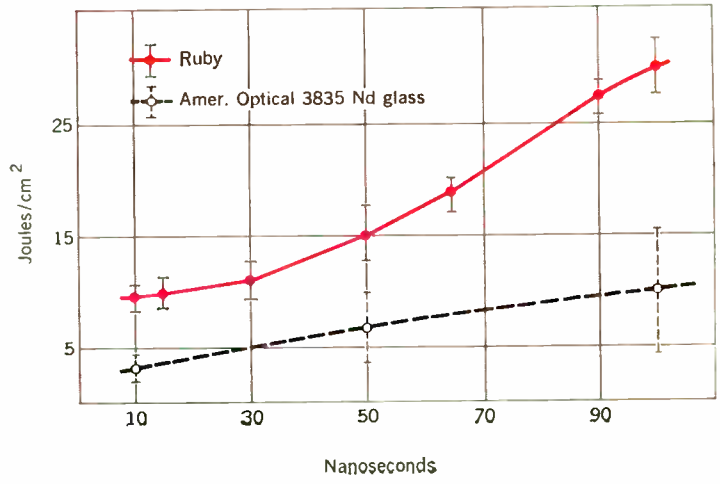


FIGURE 2. Life expectancy of ruby crystals as a function of power intensity. Data points are averages of several crystals of different diameters. Pulse width ≈ 10 ns.

FIGURE 3. Damage threshold in joules/cm² vs. pulse width for ruby and Nd-doped glass.



tent on total energy in the pulse. To extend crystal life, the broadest possible pulse should be used for a given energy whereas, for a given power level, the pulse should be as short as possible. However, it is more meaningful to incorporate power in the specification of luminance. The significance of this term and how it is measured is treated in the following.

Luminance

The concept of brightness has been used for many years in the study of radiation sources. It may be instructive to review briefly its application to diffuse emitters and then consider how to accommodate the unique properties of a Q-switched laser. When a surface is heated sufficiently to emit visible radiation it appears equally bright at all angles of observation and all distances. The measure of brightness or luminance is given by the following relation (see Fig. 4)

$$B = \frac{dE}{dt\, d\Omega\, dS \cos\theta} \tag{1}$$

where B is luminance in power per unit area per unit solid angle, and dE/dt is the amount of power radiated from area dS at angle θ from the normal to the surface S into solid angle $d\Omega$. Lambert's law states that the intensity $dE/d\Omega$ from a diffuse source varies as $\cos\theta$, and thus luminance is independent of angle. As the distance between the eye and the source is increased the decrease in power entering the pupil is compensated by the decrease in size of the image, resulting in constant image brightness with distance.

A laser is markedly different from a diffuse source in that it radiates into a very small solid angle—typically $\sim 10^{-5}$ sr (steradian)—compared with the 2π sr radiation angle of a difiuse source. In addition, the luminance of a laser is a strong function of angle within its cone of radiation. Also, it is essentially monochromatic whereas most diffuse sources have a broad spectrum. We will fix our attention on the reduction of the angle of radiation and the variation of brightness with angle. The laser monochromaticity has important implications but they are not essential in this discussion. Let us now consider the problem of measuring the luminance of a Q-switched laser.[11]

Since the total energy of the laser is included in a cone angle of a few milliradians, the $\cos\theta$ term in Eq. (1) can be set equal to unity. It is reasonable to assume that power intensity and the time duration of the pulse are the same at all points on the radiating surface, permitting (1) to be written as

$$B = \frac{dE}{d\Omega\, At_p}$$

where A is the area of the radiating surface and t_p is the pulse width.

The pulse width is measured with a fast photodiode and oscilloscope. To determine how energy varies with angle, the arrangement shown in Fig. 5 is used. The laser beam is passed through a lens with a focal length of some one to two meters. A white diffuse reflecting block (MgO) is placed in the focal plane of the lens, thus producing the far-field pattern of the laser on the block, the transverse coordinates being proportional to the beam divergence. A multiple-lens camera is used to

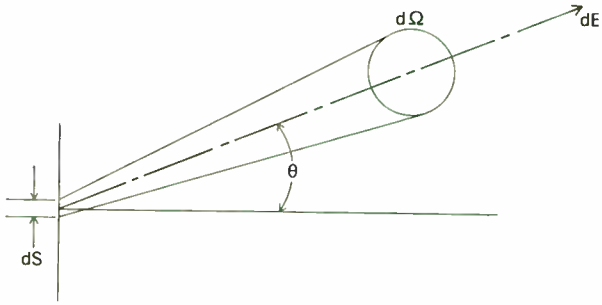
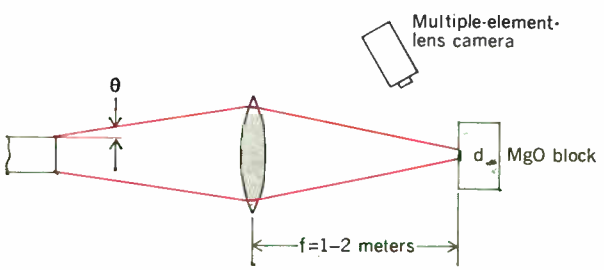


FIGURE 4. Graphic representation of luminance (power per unit area of emitter surface per unit solid angle). Note that solid angle is measured from emitter surface dS.

FIGURE 5. Schematic of experimental setup for measurement of brightness of Q-switched laser. The MgO block serves as a diffuser. Multiple-lens camera is fixed with a series of attenuators separated by 3 dB.



photograph the image. Each lens is backed by a neutral density filter with known values of transmission, resulting in a series of photographs with a range of relative intensity of about two orders of magnitude. The degree of reaction of the film to the light depends directly on the energy density but is a nonlinear function.

The neutral density filters serve the purpose of calibrating the film density versus energy function. Densitometer traces are taken to determine precisely the change in spot size with increasing transmission; a schematic representation of these traces is shown in Fig. 6. With minimum transmission the spot size x is set at zero. Each value of x can be thought of as representing a ring of diameter x and thickness dx with constant relative energy density $\sigma(x)$ containing energy dE (in normalized units), thus

$$dE = \sigma(x) 2\pi\, dx \tag{2}$$

By use of the relation

$$x = f\theta \tag{3}$$

where f is focal length of focusing lens and θ is the angular divergence of that portion of the beam corresponding to x, the spot sizes can be converted to angular divergence of the laser beam. Relative energy density in the laser beam can now be plotted as a function of beam divergence as shown in Fig. 7. By changing variables, (2) becomes

$$dE = \sigma(\theta) f^2 2\pi\theta\, d\theta$$

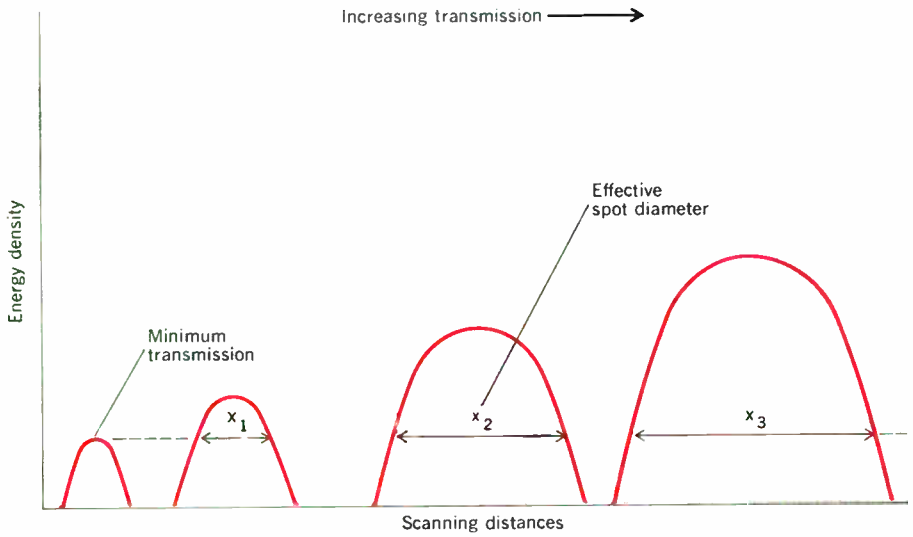


FIGURE 6. Typical densitometer traces of images of focused laser spot on film plate in multiple-lens camera.

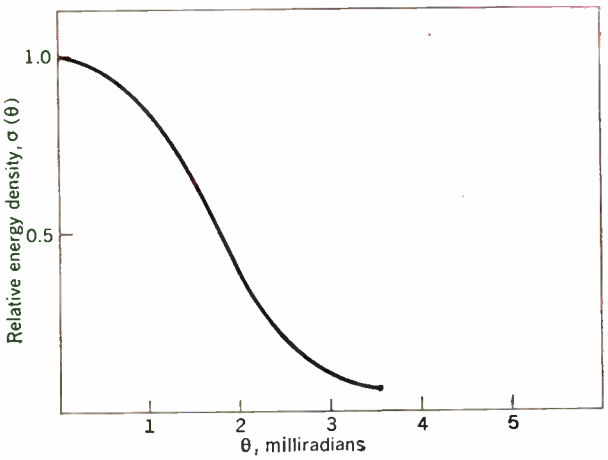


FIGURE 7. Typical relative energy density $\sigma(\theta)$ vs. beam angle. The function $\sigma(\theta)$ is approximated by exp $-(\theta/\theta_0)^2$.

Now $2\pi\theta\, d\theta$ is an element of solid angles $d\Omega$ (assuming $\sin\theta \approx \theta$), emanating from the radiating source containing those light rays that resulted in energy density registering on the photographic plate of value $\sigma(\theta)$; therefore,

$$\frac{dE}{d\Omega} = f^2\sigma(\theta)$$

Since $\sigma(\theta)$ is a relative function, f^2 can be absorbed in a constant C so that

$$B = \frac{dE}{t_p\, d\Omega\, A} = \frac{C\sigma(\theta)}{At_p}$$

Now the integral $t_pA\int B\, d\Omega$ over 2π sr is the total energy emitted from the beam E_t; thus

$$C = \frac{E_t}{\int_{2\pi}\sigma(\theta)\, d\Omega}$$

The value of E_t is easily determined by a calorimeter measurement. The constant C can be determined by performing the indicated numerical integration or analytically using a function that closely fits the curve in Fig. 6.

At $\theta = 0$, $\sigma(\theta) = 1$; therefore,

$$B_0 = \frac{C}{At_p}$$

which is the peak luminance of the laser. The function exp $-(\theta/\theta_0)^2$ matches quite closely most of the curves taken to date. Peak luminance now becomes

$$B_0 = \frac{E_t}{\pi\theta_0^2 At_p} = \frac{P}{\pi\theta_0^2}$$

where P is the peak power intensity and θ_0 is the cone angle of the laser beam, i.e., one half the full beam angle. To the extent that the density function is close to the Gaussian model used here, θ_0 is the angle where power density has fallen to $1/e$ of its peak value. The value of θ, which includes one half of the total energy, is 0.83 θ_0 and the value of θ, where the power is one half maximum, is the same. Each of these angles is referred to as "beam angle" in the literature. Care must be taken to define which of these angles is being considered, as their values vary significantly when the energy density is not Gaussian.

The significance of luminance is more evident when the particular problem of focusing the maximum amount of energy in the smallest spot is considered. If f is the focal length of the lens, d the diameter of the ruby and lens, and θ is the cone angle of the laser output, then S, the radius of the spot, is

$$S = f\theta$$

Practical considerations limit the f stop number (f/d) of the lens to one. Therefore,

$$S = d\theta$$

The intensity in the spot is

$$I = \frac{F}{\pi d^2 \theta^2} = \frac{P}{\pi \theta^2}$$

where F is the total power from the laser. The right-hand side of the equation is the luminance expression obtained previously.

Beam angle reduction

Luminance varies as the inverse square of beam angle and therefore can be significantly increased by relatively small reductions in beam angle. Thus, any scheme that reduces beam angle will increase the performance of the laser without increasing intensity.

One method of improving beam angle is by using an oscillator–amplifier configuration. The output of a particular Korad design is 1.1 GW with a brightness of 2.5×10^{14} W/cm²/sr, which is 2.5×10^{11} times brighter than the sun.

Beam angle is improved over that with a single oscillator in two ways. It has been experimentally determined that the oscillator determines the beam angle out of the amplifier unless the optical quality of the amplifier is poor. The beam angles from small crystals are usually significantly better than those from larger ones. Thus, selecting a good oscillator crystal is relatively easy and far more economical than trying to select an equivalent beam angle with a large crystal. Also, hard pumping of an oscillator crystal degrades beam angle. In the oscillator–amplifier configuration, it is the amplifier that contributes the major share of the energy; this permits the oscillator to work at low pumping levels, which avoids beam angle degradation. Another advantage in having the oscillator work at low levels is that critical components, such as the Pockels cell and reflectors, are less subject to damage.

The output end of the amplifier is usually cut at the Brewster angle to avoid reflection back to the oscillator output reflector, which can cause the amplifier to oscillate, especially when it is fully pumped. No reflective dielectric coatings are used, since these would fail quickly under the effects of a high-intensity laser beam. The 100 percent reflector is a Porro prism cut to within two seconds of arc to preserve the beam angle. A Pockels cell, immersed in a protective fluid, is used as the Q switch.

The system can deliver 300 shots with a brightness of 2.5×10^{14} W/cm²/sr with essentially no damage to the optical components—with the exception of the amplifier crystal, which most likely will show signs of bubble formation at the output end. This performance is practically impossible to obtain with a single oscillator.

The same configuration can be used for Nd-doped glass, except that an additional polarizer stack is needed since the radiation emitted from a glass laser is not polarized. Glass tends to have better beam angles than ruby, so that higher brightness can, in principle, be expected. The Centre de Recherches de la Compagnie Generale d'Electricité has announced a system with an oscillator and two amplifiers, with a brightness of 10^{14} to 10^{15} W/cm²/sr.

The beam angle of even these advanced systems is still more than 20 times greater than the diffraction-limited value (~0.05 mr), so theoretically there is still room for improvement. However, in applications requiring the laser to be used at long ranges, the pointing problem may make further beam reduction undesirable.

Pulse stretching and arrays

To attain a large-energy (200-joule) Q-switched pulse without destroying the ruby, it is necessary to stretch the pulse and decrease power intensity. Pulse widths of more than 100 ns can be obtained by using a long oscillator cavity and actuating the Q switch when the gain is relatively low. Power intensity is kept below damage threshold by an array of output amplifiers. In a large Q-spoiled laser system configuration developed at Korad the output from the system is 200 joules at 125 ns. The beam splitters use frustrated internal reflection as the principle of operation. An isolator between the preamplifier and intermediate amplifier prevent feedback into the preamplifier. A beam shaper converts the circular output from the flat-ended oscillator crystal into an elliptical beam conforming to the Brewster entrance of the preamplifier.

A similar system,[12] designed to deliver several gigawatts, and using an oscillator, preamplifiers, and an array of final amplifiers, has been constructed at the Lawrence Radiation Laboratories. The requirement with regard to this system is for significant energies in extremely short pulses.

The future

Improvements in the optical quality of laser crystals and techniques to decrease beam angles, resulting in brightness of 10^{15} to 10^{16} W/cm²/sr, can be expected in the near future. Damage thresholds probably will not be significantly increased, but imaginative arrays of lasers and larger diameter crystals will permit the generation of higher power lasers (25 to 100 GW) capable of delivering up to 1000 joules in a stretched Q-switched pulse. Electronic feedback circuits will permit pulse stretching without extending the cavity length. When all is considered, lasers have a bright future.

Revised version of a paper presented at the International Communications Conference, Philadelphia, Pa., June 15–17, 1966.

REFERENCES

1. Maiman, T. H., "Stimulated optical radiation in ruby," *Nature*, vol. 187, pp. 493–494, Aug. 1960.

2. Hellwarth, R., *Advances in Quantum Electronics*. New York: Columbia University Press, 1961.

3. Kisliuk, P. P., and Boyle, W. S., "The pulsed ruby maser as a light amplifier," *Proc. IRE*, vol. 49, pp. 1635–1639, Nov. 1961.

4. DeMaria, A. J., Stetser, D. A., and Hynau, H., "Self-mode locking of lasers with saturable absorbers," *Appl. Phys. Letters*, vol. 8, pp. 174–176, Apr. 1966.

5. Soffer, B. H., and Hoskins, R. H., "Generation of giant pulses from a neodymium laser by a reversibly bleachable absorber," *Nature*, vol. 204, p. 276, Oct. 1964.

6. Melamed, N. T., Hirayama, C., and French, P. W., "Laser action in uranyl-sensitized Nd-doped glass," *Appl. Phys. Letters*, vol. 6, pp. 43–45, 1965.

7. Benson, R. C., and Mirarchi, M. R., "The spinning reflector technique for ruby laser pulse control," *IEEE Trans. on Military Electronics*, vol. MIL-8, pp. 13–21, Jan. 1964.

8. McClung, F. J., and Hellwarth, R. W., "Giant optical pulsations from ruby," *J. Appl. Phys.*, vol. 33, pp. 828–829, Mar. 1962.

9. Helfrich, J. L., "Faraday effect as a Q-switch for ruby laser," *J. Appl. Phys.*, vol. 34, pp. 1000–1001, Apr. 1963.

10. Avizonis, P. V., and Farrington, T., "Internal self-damage of ruby and Nd-glass lasers," *Appl. Phys. Letters*, vol. 7, pp. 205–206, Oct. 1965.

11. Winer, I. M., "A self-calibrating technique measuring laser beam intensity distributions," *Appl. Opt.*, vol. 5, pp. 1437–1439, Sept. 1966.

12. O'Neal, W. C., Private communication.

The radio spectrum below 550 kHz

In today's high-frequency world it is surprising how much use is still made of the lower frequencies, for standard broadcasts as well as for such special services as weather data for mariners

Thomas L. Greenwood *Huntsville, Ala.*

The lower radio frequencies are used in worldwide communications to a greater extent than is generally realized. This article discusses the almost-forgotten radio world that exists on frequencies below the United States' standard AM broadcast band. The information is based on personal observations in Huntsville, Ala., of stations in the low-frequency spectrum.

During the years immediately following World War I, low frequencies—or "long waves," as they were called then—were being developed for worldwide "wireless telegraph" communication. This came about as a result of experimental evidence that the attenuation in the transmission medium (earth's atmosphere) is proportional to the frequency. Efforts were therefore made to utilize the lowest practical frequency, which required huge antenna structures and high transmitter powers. It was slowly realized, however, that atmospheric noise (static) was the limiting factor in reception and that a compromise was necessary to achieve reliable communication. A frequency near 50 kHz was found to provide the most favorable signal-to-static ratio where atmosphere was a problem in reception.

Although Marconi's first transatlantic communication in 1901 was accomplished at a frequency of about 200 kHz, further experiments indicated that longer wavelengths (lower frequencies) might afford better results. The Clifden, Ireland–Glace Bay, Nova Scotia, circuit utilized wavelengths of about 4000 meters (75 kHz) and the Marconi Company's station at Caernarvon, Wales, used wavelengths of up to 10 000 meters (30 kHz). The first radio communication between Europe and Australia was achieved in 1918 with a 200-kW spark transmitter on a frequency of about 21 kHz (14 000 meters), and by the 1920s, wavelengths of up to 25 000 meters were being used for long-distance communication. However, plans for high-power (up to 2 MW) low-frequency stations for worldwide communications, with frequency spacings only 100 Hz apart, were never fully realized because of new developments in high frequencies.

Around 1925 it was found that frequencies above 3 MHz had certain advantages for long-distance communications, so the trend since then has been in that direction, although low-frequency stations were still being used after World War II. High-frequency alternators installed in the early 1920s were operated by RCA and the U.S. Air Force until the early 1950s, and the U.S. Navy still operates one in Japan that dates back to 1927. Utilizing a frequency of 17.4 kHz, it can be heard in Huntsville at any hour of the day or night.

The writer recalls that, while serving as Radio Officer in the U.S. Merchant Marine during the early 1930s, he heard news transmitted from WII in New Brunswick, N.J., at a frequency of 18 kHz. The transmitter was a high-frequency alternator with an output to antenna of 200 kW, and it was keyed by a magnetic modulator. Reception was good throughout the southern Atlantic Ocean area.

The advantages of low frequencies are low attenuation, slight fading effects, and a high degree of immunity to solar flare blackouts (so-called "magnetic storms"). During the solar flare in September 1966, high-frequency radio circuits were badly crippled but low-frequency circuits were not affected and continued to handle traffic, which was piling up a backlog for the high-frequency circuits. Also, when high-frequency circuits are saturated with traffic, the low frequencies take care of the overflow.

A new spectrum

During the 1940s, international agreements produced a new radio spectrum structure, which divided the useful frequencies into bands according to their transmission characteristics: ELF, 3 Hz–3 kHz; VLF, 3–30 kHz; LF, 30–300 kHz; MF, 300–3000 kHz (3 MHz); HF, 3–30 MHz; VHF, 30–300 MHz; UHF, 300–3000 MHz (3 GHz); SHF, 3–30 GHz. In the ELF (extremely low-frequency) region, no regular radio communication is carried on at this time. Experimental transmissions have been made on 400 Hz, and transcontinental telegraph communication was achieved with the transmitter in California, using a 100-km section of a 300-kW power line. The reception, in New York State, was limited at times by atmospheric noise and interference from 60-Hz harmonics, but it was demonstrated that operation was practical. A signal-to-noise ratio of 18 dB was achieved during favorable conditions.

The lowest frequency in regular use is 10.2 kHz by the Navy-sponsored Omega long-distance navigation system, which also uses 12.75 and 13.6 kHz for transmitter stations in Trinidad, Hawaii, Panama, and New York. All these stations can be heard in Huntsville. The 10–14-kHz band is reserved for navigation service.

The lowest frequency regularly used for radio communication is 15.4 kHz by NWC in Northwest Cape, Australia. This station is heard in Huntsville, as are others in the VLF range: GBR in Rugby, England, at 16 kHz since 1925, achieving worldwide coverage with 350-kW power; RCC7 in the U.S.S.R., at 16.2 kHz; FUB, in Paris, at 16.8 kHz; NDT, at 17.4 kHz; GQD, Anthorn, England, at 19 kHz; GBZ, Criggion, Wales, at 19.6 kHz; U.S. Navy stations NAA, Cutler, Maine, at 17.8 kHz with 1-MW power; NPM, Hawaii, at 26.1 kHz; NBA, Panama, at 24 kHz; NSS, Annapolis, at 21.4 kHz; NPG, Jim Creek, Wash., at 18.5 kHz. Among the Navy stations occasionally heard are NKA in Asmara, Ethiopia, at 23 kHz, and NJW at 26 kHz. A few stations use power approaching a megawatt; others have 200- to 600-kW power in the antenna.

There are standard frequency transmissions at 20 and 60 kHz from the National Bureau of Standards stations WWVL and WWVB in Boulder, Colo., that can be heard throughout the Western Hemisphere. Other nations in various parts of the world operate standard frequency transmitting stations at frequencies from 15 to 150 kHz. Time signals are often combined with standard frequency transmissions, and can be received hourly, day and night.

Most VLF stations have stabilized frequencies, constant to a few parts at 10^{11} Hz, which are useful in worldwide navigation systems, and for checking inertial navigation systems on ships, submarines, and aircraft. The Navy uses VLF for transmission to submerged submarines because these frequencies have low attenuation in seawater. Reception is possible at a depth of 15 meters.

In the low-frequency band there are numerous telegraph services, both Morse and teleprinter. Some 20 stations, using teleprinter code and frequency shift keying, are heard in Huntsville, but cannot be identified by ear. More than 20 telegraph stations, using Morse code at speeds of less than 25 words per minute, have been identified there. The U.S. Navy operates NAA, NSS, and NPG at frequencies between 60 and 160 kHz; RCA has WCC at 147 kHz; and ITT has ESL at 112 kHz. A few European stations cluster around 50–70 kHz, and an Australian station, VIX in Sydney, is received well during the early morning hours at 44 kHz. Radio navigation aids use frequencies from about 75 kHz to 200 kHz: Decca, 75 to 130 kHz; Loran, 100 kHz; and Consolan, near 200 kHz. Other navigation systems are heard but cannot be identified.

The first transatlantic telephone service, inaugurated in 1927, used the LF band between 50 and 70 kHz; it marked the first use of single sideband in radio. Because of the superiority of HF bands, it is no longer used for that purpose. However, an early successful ship-to-shore radiotelephone transmission was accomplished in 1919 with a high-frequency alternator in the 25–40-kHz band by having the magnetic modulator operated by voice signals instead of on–off keying. The shore station was at New Brunswick, N.J., and a vacuum-tube transmitter of about 3 kW on the *S.S. George Washington* carrying President Wilson to Europe provided the other end of the circuit. The first successful experimental operation of radiotelephone across the ocean occurred four years earlier, in 1915; it used a 50-kHz frequency and 11 kW of power.

Aeronautical stations in Europe are authorized to use low frequency between fixed points, but it is not known whether they do so at present. Before World War II, the writer heard many LF telegraph channels being used by European aeronautical services.

Transatlantic aircraft use the Dectra navigation system, which operates in conjunction with Decca at each end of the ocean passage. These services are at 70–130 kHz, and can be identified by the characteristic sequence of signals used to operate the system's navigation instruments.

The standard low-frequency broadcast band occupies 150–285 kHz, sharing the sub-bands 150–160 kHz and 255–285 kHz with other services. Between 285 and 535 kHz, however, there are broadcast stations located in continental interiors remote from maritime activities, as in Russia, Siberia, and China. These stations are usually limited in power, although the U.S.S.R. has a 50-kW station in Minsk at 400 kHz and Finland has a 10-kW station at 433 kHz at Oulu. Russian broadcasting stations in the low-frequency band are used as beacons for aircraft navigation and have a code letter identification for that purpose. Under favorable conditions some European stations might be heard in the United States—pos-

sibly Radio Luxembourg, which uses 1 MW of power at 233 kHz, or France, which has a half-megawatt station at 164 kHz.

Maritime usage

Low frequencies, 100–150 kHz, are used for certain broadcast services to ships, mostly for weather information and notices to mariners. These frequencies supplement the high frequencies to insure close-in coverage near stations that might be missed due to the "skip-distance" effect. Among the North American stations heard in Huntsville are: CFH, Halifax, 114.5 and 132 kHz; CKN, Vancouver, 110 kHz, and NSS, 89, 121.5, and 161 kHz; and NBA, Panama, 147.5 kHz.

In the years between World War I and World War II most ships used a frequency range of 60 to 160 kHz, but it is no longer very popular. International agreements permit ship usage, on a shared basis, of 14 to 160 kHz. United States commercial ships may use 70–160 kHz, but few do so; U.S. government ships may use 20–90 kHz and 110–160 kHz on a shared basis. In 1930 the writer operated a ship station that had 18 assigned frequencies between 60 and 500 kHz. Formerly, aeronautical services used 333 kHz as a watch and calling frequency, and it is still authorized by international agreement. At present it is used very little, possibly only in Africa and Asia, because interference from land-based aeronautical beacons and broadcasting stations would impair its efficiency. The writer recalls hearing traffic on 333 kHz in South American waters during the early 1930s from the Graf Zeppelin, Dornier DO-X, and other aircraft.

In the early 1920s and 1930s, various fixed services used frequencies in the LF band. Pipeline companies were very early users of radio, especially telegraph in the 150–200-kHz band, but have long since moved to the UHF and microwave bands.

Aeronautical beacons use frequencies between 200 and 400 kHz; the MF (medium-frequency) band, which begins at 300 kHz, includes most aeronautical services. At major airports ground stations of about 1-kW power transmit continuous voice broadcasts of weather forecasts, synopses, and other information of interest to pilots. The information is usually on magnetic tape, updated as necessary. Huntsville Airport does not use this band; Nashville (305 kHz) and Birmingham (224 kHz) are the nearest airports that do. Low-power approach beacons are located near all airports. Those heard in Huntsville are ITS (Redstone South) at 407 kHz, HUA (Redstone North) at 287 kHz, SV (Municipal Inner, at Parkway City) at 220 kHz, and HS (Municipal North) at 237 kHz. Some of these beacons radiate appreciable power on second harmonics of their assigned frequency and can be heard on an automobile receiver when the automobile is near the transmitter. A few of the old A–N radio range stations are still in operation and can be heard in Huntsville. Operation of aeronautical ground stations on medium frequencies is the only practicable way that information can be transmitted to aircraft within a 40-km radius of the station.

The maritime radiotelegraph band is 418–535 kHz, including ships and coast stations. Coast stations use 418 to 500 kHz and ships may use 11 frequencies between 418 and 517 kHz. A few maritime beacons operate between 510 and 535 kHz, but most of them use frequencies between 285 and 325 kHz.

Seagoing ships, as the first users of radio, were able to dominate frequency allocations for a number of years. Efficient equipment was first developed for maritime service using frequencies most suitable for that service. These were determined by the dimensions of an efficient antenna that could be mounted on a ship's mast, and were in the 300–700-kHz range. In early years, transmitter powers ranged up to 10 kW, but most modern ship transmitters do not exceed 1 kW. Ships' antennas for use at medium frequencies are of the grounded quarter-wave (Marconi) type, are nondirectional (which is necessary for ships' use), and are fairly efficient (up to 50 percent).

Coast radiotelegraph stations heard in Huntsville on the 418–500-kHz band are: WCC, Chatham, Mass.; WSL, Sayville, N.Y.; WSC, Tuckerton, N.J.; WMH, Baltimore, Md.; WOE, Lantana, Fla.; WAX, Miami, Fla.; WPD, Tampa, Fla.; WLO, Mobile, Ala.; SNU, New Orleans, La.; KLC, Houston, Tex.; WPA, Port Arthur, Tex.; KPH and KFS, San Francisco, Calif.; KOK, Los Angeles, Calif.; plus Panama, Mexico, Venezuela, and West Indies stations. Transmitter powers of coastal stations range from 1 to 20 kW.

It is probably popularly believed that the maritime services no longer make use of manual Morse code operation. Actually however, because of the particular conditions involved in maritime communications, it is highly improbable that manual Morse will become obsolete in the foreseeable future. Although high-volume traffic, such as between shore and a large passenger liner, is handled by teleprinter equipment, it would not be economical, considering the low average message rate, to install such equipment on the majority of ships.

In the years between the World Wars, 375 kHz was used for direction finding in the maritime service. The U.S. Navy operated shore-based direction-finding stations at all important points on the coasts of the United States. Indeed, all maritime nations had shore-based direction finders. The loop direction finder was the first radio navigation device to be developed, but as few ships were so equipped at that time, they were forced to use on-shore stations for radio-location service. These operated quite well, except during saturation periods in adverse weather. Some nations still operate shore-based direction finders but the system has been discontinued in the United States. The writer recalls that, while serving as Radio Officer on an American merchant ship in 1930, he assisted in guiding the ship from Florida to Philadelphia during a hurricane by using radio bearing from shore-based stations. Because of overcast skies, celestial navigation was impossible for a period of four days.

In addition to legitimate signals on frequencies below 550 kHz, many spurious signals arise. Television sweep-frequency harmonics of 15.72 kHz and 60-Hz power system harmonics cause widespread interference. Beat frequencies between nearby broadcasting stations, due to external cross-modulation effects, are much in evidence in Huntsville; 50, 100, 150, 270, 490, and 500 kHz are the most troublesome. The signal at 50 kHz is easily mistaken for the standard frequency transmission from Podebrady, Czechoslovakia. It is a few hertz above the standard 50-kHz signal, and is quite strong in some sections of the Huntsville area. The television sweep frequency harmonics may be used for receiver dial calibration, and so these are not entirely useless.

The philosophy of an engineering educator

The dean of one of the oldest engineering schools in the United States offers his views and recommendations for revising and updating the engineering curricula—and the students' attitudes—to meet the critical needs of our rapidly expanding science and technology in an era of unprecedented change

Aaron J. Teller *The Cooper Union*

We, as a nation, have emerged. The education to maintain and expand our society is available to all. The new challenge to the unique, the intelligent, the eager, and the ambitious is to explore the frontiers of knowledge, discover the magnificence of the order of nature, and then to apply this accrued wisdom to the creation of a new and better life for our whole society. The engineering school must stimulate and challenge its select students to explore and solve our great sociotechnological problems. It must not merely train its young men to maintain the present; it must do what others cannot in using the strength of its unique students and our pioneering heritage to provide this nation with that elite body of men who have always sought the way to a better life. It must supply society with sociotechnological explorer–leaders.

In the last two decades, colleges and universities in the United States have phenomenally expanded their efforts to provide our society with opportunities for higher education. The phrase, "mass education," is used as a common expression that is reflective of this effort.

The expansion of education opportunities was primarily triggered by the emergence of the technological age, imposing upon our society an escalated educational demand that could no longer be satisfied only by elementary and high school education.

Environment for leadership

Traditionally, society has utilized higher education as a prime vehicle to provide not only the practitioners of the many professions, but also to stimulate the development of leadership in professional achievement. Professional activities may be classified in two areas: "reproductive" and "exploratory." The great majority of practitioners in any field are in the reproductive category: those who satisfy the expanded needs of society either by the duplication of existing facilities, or by providing only moderate change.

But to advance society sociologically, technologically, and culturally, the activity of the exploratory or creative practitioner is essential. The paramount concern, therefore, is that, in our quest for educational opportunities for all and the concomitant development of mass education for a heterogeneous school population, *the stimulation to leadership for the superbly capable and motivated individual not be submerged in the massive education program designed for the average.*

Obviously, the educational environment for leadership is greatly enhanced by association with creative teachers and with other gifted students, and by the challenge and stimulation that result from these associations. Further, it must be complemented by the challenge and demand of a curriculum that is designed to stimulate creativity.

Submerging the uniquely capable student in a mass of students of standard capabilities, within a stultified research atmosphere, can produce adverse effects on his educational challenge to achievement. Thus the lack of competitive challenge of classmates, the dimunition of the academic and philosophic demands of the curriculum, and the reduction of contacts with creative educators are an integral part of such an inhibitive environment.

Conventional vs. unique education

In the society of 30 years ago, a college education was relatively uncommon. It was economically difficult to acquire this education, and, in the main, it required a strong motivation coupled with unusual intelligence. The schools were smaller, the student lived in a student environment that was intellectually stimulating, and the senior faculty members were more available to the student than at present.

Today, with the simultaneous massive growth in university and class size, the greater heterogeneity of intellectual capabilities in the student population, and the withdrawal of capable faculty from undergraduate education to "contract research" and graduate school activities, these factors of an earlier day have diminished. Therefore the opportunity for the unique, the motivated, and the highly intelligent student to be challenged intellectually may have actually decreased.

In view of this situation, it is essential that there should be a number of small institutions in the United States that are dedicated to the education and stimulation of intellectually gifted and motivated students, with emphasis on the *exploratory* practices of the various professions. These schools would have the characteristics of

1. A homogeneously elite student body to provide an intellectually competitive environment.
2. A particularly capable, professionally oriented faculty, whose efforts are primarily directed to the stimulation of each individual in the student body.
3. A faculty who is given the opportunity for exploratory research activities with undergraduate and graduate students, but whose educational efforts are not vitiated by massive contract research or teaching duties at lower levels of intellectual achievement.
4. An academic program directed specifically to guiding the student group toward intellectual and creative accomplishments.
5. A sufficient degree of flexibility so that curricula and policies may be varied to achieve the optima in the education process, without having to overcome the inertia of a mass educational effort.

The Cooper Union—tradition and a concept of excellence

The Cooper Union was founded with the objective of excellence, and, over the years, it has acquired the concept of excellence as a tradition. Its School of Engineering has recently undergone a self-evaluation of its effectiveness in the development of practitioners in exploratory engineering and science. It was established that the institutional characteristics of the school generally met the five criteria just mentioned, and thus a sound base existed for the founding of a modern center of excellence in education that would be dedicated to the exploratory aspects of engineering and science.

There existed at The Cooper Union, as in other institutions, however, an unfortunate aspect of tradition that was based upon success. This was the tendency for acceptance of only minimal change in the educational process and it reflected the "fear of failure" that could result from radical change. Thus the philosophy of engineering education was threatened by stagnancy at the time of an increasing need for creative activity.

As a result of the self-evaluation taken during this period, the decision was made for radical change, with the objectives of establishing a modern "center of excellence" that would encompass and promote

1. The development of exploratory practice in the professions of engineering and science.
2. The stimulation of the intellectually gifted individual toward creative achievement.
3. Educational and research activities centered on conceptual knowledge, not information.
4. Exploration in educational content and methods in which rapid feedback and flexibility would minimize the danger of failure.
5. A prototype model that would act as a "pilot plant" in education for other engineering schools.

A prescription for undergraduate education

Professional education must be a total experience. For problem development, there must be an exposure to conceptual thinking and an indoctrination in conceptual knowledge; for implementation, an understanding of the proper use of empirical information. Thus education must combine a pattern of inculcation of an exploratory philosophy with a sense of realism.

The exposure process. Of necessity, exposure to education must be a potpourri that is arranged into an interrelated and integrated system that includes the factual, the psychological, and the philosophical.

Re-evaluation and modification. In the past, the modifications of engineering school curricula have been reflective of evolutionary changes in the practice of engineering; but modification implies that what exists is essentially correct in philosophy—and this assumption is questionable!

It is hoped that Cooper Union's present re-evaluation will be anticipatory, not reflective; that it will recognize the existence of many different objectives in engineering and scientific practice within a given field of endeavor; that the role it must accept will educate, excite, and inspire a limited number of students to the exploratory practice of engineering and science; and that its educational motivation and mechanism will, therefore, be different from those of other institutions.

The Cooper Union was not the only institution to

recognize the inadequacy of the educational process as it existed in a rapidly changing technology, but it is regrettable that many engineering schools, through default in their education processes, stood by while a number of major technological advances were achieved by scientists rather than by engineers.

From empiricism, an evolution. In the last two decades, engineering has evolved from total empiricism to a combination of theory application and empiricism. Since engineering schools did not anticipate this change, we were still educating primarily in empiricism at a time when the realistic exploratory practice demanded the incorporation of theoretical projections.

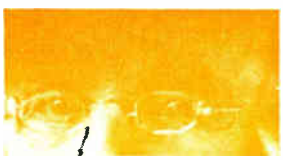
In recognition of an inadequacy, schools reacted rather than evaluated, and developed "core" scientific curricula (that often bear little relationship to the professional course presentations). They also added "theory" in the graduate school effort. This reaction raised the argument that excessive exposure to studies in systems and materials behavior would result in a "scientific" rather than an "engineering" orientation. This is superficially true, since it does inhibit the practitioner from a truly empirical approach and it may limit his effectiveness in achieving the quick and ready solution that is necessary for reproductive practice.

Nevertheless, the engineering philosophy of achieving an effective and an optimum solution to a real problem can be better implemented by the fundamentally oriented engineer. He is not bounded by the limits of empiricism nor the philosophy of "make do." He can approach the problem more realistically from a study of the behavioral aspects of the system and, coupled with empirical knowledge, he can project the most realistic solution.

'Realistic' design of systems and processes

What is necessary in engineering education is the establishment of an educational philosophy that is directed toward the stimulation of achieving realistic design of systems and processes. This necessarily involves the research orientation as well as the economic and social implications. These cannot be dissociated from technological education.

Since the ultimate contribution of the engineer to society is that of the creation of systems and processes, *the design philosophy must pervade the curriculum.* In preparation for exploratory practice, the emphasis of education must be on an understanding of what is known in regard to the behavior of systems and processes, combined with an inculcation of a psychology to use this knowledge or to seek additional knowledge to be used for the extension of the frontiers of application.



Proper recognition must be advanced as to the necessity for the *combination* of theory and empiricism for the solution of real problems, *because the solutions of problems facing engineers in exploratory practice will always require more knowledge and information than that which is presently available.*

The protagonists of "either/or" (empiricism or theory) find themselves in the enigmatic positions in which

1. The empiricist finds his knowledge obsolescent.
2. The theoretician is not capable of solving real problems with this theory alone.

The Cooper Union experiment

At The Cooper Union, it is our intention to experiment—with vigor in philosophy and caution in implementation—within the parameters of the following preconditions and objectives in which

1. The traditional mode of development of an engineering program is questioned from the aspect of sequential development.
2. The philosophy espoused by the instructor is of equal significance with content in course presentation.
3. The application of knowledge in engineering education is of equal import to the knowledge itself, and the experience of utilization of knowledge may have salutary effects on the understanding process.
4. The stimulation to creative endeavor can best be achieved by a creative practitioner in the classroom.
5. The approach to stimulation of a psychology of creativity must be inculcated from the instructor's attitude and the sophisticated problem. (The simulation of creative endeavor by the presentation of "ingenuity problems," requiring little knowledge, can be insulting to the student of high intelligence.)

Implementation—the 'three steps.' The first "cautious" step is to populate the classrooms with instructors who have given evidence of creative achievement. The second step is to redesign the curriculum from the conceptual viewpoint of sequential development and by the introduction of the exploratory attitude in the development and application of knowledge. The third step is to improve faculty-student communications by changing the "mechanism of contact" to provide an increase in the freedom of inquiry and the spectrum of exposure.

Loss of the 'application psychology.' One of the paradoxical aspects in the evolutionary process of engineering education is the expansion of fundamental studies that often appear to be unrelated to application studies. These studies are partially justified by rising new challenges in engineering that require the direct application of fundamentals. Tragically, however, the *application psychology* in the presentation of fundamentals has become a lost art. Thus many engineering students are bored by fundamentals and wade their way through them in the hope of surviving the exposure to reach the promised land of application and empiricism, or some students are pushed irrevocably toward the science philosophy. In both cases, the engineering profession has lost good engineering potential.

Communication and information transmission. Communication is the one area in which the "technique mechanism" can achieve significance. It can aid the competent instructor in stimulating and in eliciting creative capabilities from the student on an individual basis. The classroom technique is stultifying in that the approach must be directed toward the median of class capability and never at the highest level of the individual.

Where information transmission or the establishment of general background is desired, the large lecture procedure should be used—with the most capable and articulate members of the faculty providing this service.

All of these worthwhile objectives can be achieved by a single phenomenon—a talented, imaginative, psychologically oriented and motivated faculty.

Vital role of the faculty

It is axiomatic that the quality of the faculty is the major factor in the development and implementation of any academic program. The definition of quality, however, is quite diffuse and variations of this definition can produce dramatic differences in educational philosophy.

In the milieu of professional education, we have defined two types of faculty orientation or competencies that are necessary for the effective implementation of our objectives: the professional teacher, and the professional engineer. Regardless of orientation, the common characteristics of both must be knowledgeability, excitement for the subject, interest in the stimulation of the individual student, and communication capability.

The professional teacher is one with a broad spectrum of knowledge, but one who has little participation in the creation or application of this knowledge. He must possess a deep understanding of and compassion for the confusion and fears of the student. The faculty member with these attributes is the superb teacher in the "tools" course in which the introduction to the field is presented by means of his broad sweep of knowledge.

Beyond the tool level, the educational process is concerned with the studies of fundamentals in depth and the application of these concepts. To stimulate the student to probe beyond what is known, the teacher himself must be a participant in exploratory work. He cannot teach only on the basis of reflection of past participation, or textbook knowledge, for he would be teaching obsolescent material in a rapidly changing field.

A fallacious argument. A line of argument has been projected that instruction by a professional engineer is not essential in undergraduate education. The fear exists that such education would supplant broadness with depth and that there would be a decreasing correlation of the various facts of the curriculum. The argument is invalid because

1. The professional today is not a narrow empiricist but, rather, one who relies on the effective combination of broad conceptual knowledge and empiricism to achieve realistic solutions.
2. The exposure to study in depth is an essential requisite to exploratory practice.
3. The excitement of achievement by means of an intellectual exercise has a salutary effect on the student's desire to continue this practice.
4. The intellectual quality of the student at The Cooper Union is that, through exposure to specialist capabilities, he will be challenged to "mesh matrices"—the objective necessary for intellectual and creative achievement.

The academic program has strength

One of the major factors in the development of The Cooper Union academic program is the high intellectual quality of the student body that has permitted the level of education to be continually upgraded. Nevertheless, two disturbing questions have naturally developed: Shall we continue to raise the educational requirements for the bachelor's degree? Shall we accept the "normal" level of attainment in undergraduate engineering curricula and reduce the time necessary for the acquisition of the degree?

With the natural reluctance to initiate the type of change indicated by the second alternative, we have progressively increased the intellectual challenge of our program. Within the last three years, the faculty has changed the undergraduate program by increasing the quality of its science and mathematics courses, directing its engineering courses away from the descriptive and toward the conceptual, and introducing the new approach of offering interdisciplinary engineering fundamentals throughout the four-year curriculum. Parenthetically, we believe the last-mentioned mechanism for providing constant exposure to fundamental concepts, with increasing sophistication as the student progresses through the formal course of study, is far superior to a segmented presentation of fundamentals by means of the "core curriculum."

We have expanded the project concept in education, thereby permitting the undergraduate to project his own capabilities in the exploration of engineering.

Finally, the faculty believes that the continuing evolutionary direction of our educational plan toward conceptual content and utilization of knowledge by the student, as an individual, will enhance the effectiveness of the academic program in the objective of providing professional engineering leadership.

...But there are weaknesses

As a result of the concentration on conceptual knowledge, it becomes difficult to convince the student that application does not diminish, but rather enhances, the intellectual challenge. And we must always ask ourselves: Is it fair to the student that, upon completion of the difficult and ambitious program, he receives the same degree as is given elsewhere for completion of less difficult programs? Is it reasonable for our students to find, when entering graduate school, that they have already completed much of what is offered to them?

In its zeal to expose him to the vast horizons and opportunity for intellectual challenge in engineering and science, the faculty has prescribed the academic program to the point where little latitude is given to the student. This situation contradicts the necessity of broadening the spectrum of exposure to the professionally oriented student who has, upon admission to the school, already focused on his objectives.

Proposals for strengthening the program

To overcome the structural weaknesses of our program, there is a plan under consideration to establish a four-year program, including cooperative training by industry, that may culminate in the granting of either a Master of Engineering or a Master of Science degree.

The four-point revised program. It is hoped that this program will provide significant advantages toward the stimulation of the learning and creative processes by

1. Requiring one summer of supervised cooperative training by industry, or in research laboratories, to provide an experience that is related to the intellectual challenge of the application of knowledge.
2. Increasing the latitude of the student in elective courses during the latter years of the program to afford

the broader educational exposure necessary for a professionally oriented student.

3. Adding two summers of sufficiently unique and vigorous "on campus" formal training to provide a total exposure that will match the best to be found in the training for the master's degree.

4. Replacing the formal classroom process, to a large degree, by the seminar—a small group in which co-learning can occur and the individual student can be challenged and helped.

Renaissance of engineering design

Recently, there has been a resurgence of interest in engineering design on the part of educational institutions. This reflects a concern that, in a world burgeoning with new fundamental knowledge, there is a high degree of mediocrity in the practice of engineering design.

The responsibility for this anomaly lies with the educational institutions themselves. It is the natural outgrowth of conflict between the "pure theorists," profound in analysis, and the "practical designers," steeped in accomplishment. Unfortunately, each approach, by itself, is inadequate for today's engineering. It is necessary, therefore, that there be a renaissance in engineering education directed to creative design. Such design must include the conception of the problem as well as the solution, and the "realistic" solution must be favored over the "practical."

The design experience must be projected on three criteria:

1. The establishment of the necessity for solution, either technological or sociological.

2. The understanding of the phenomena involved.

3. The achievement of solutions based on the advancement of fundamental knowledge, the restriction of empiricism, and optimization in the context of technical, social, and economic factors.

Finally, the student...

The writer has already noted that the dramatic changes in the horizons and practices of the engineering profession, and the anticipation of new frontiers, have demanded radical revisions in engineering curricula. In line with this, The Cooper Union has also drastically revised its curricula, with significant increases in mathematics and science content, and by the introduction of a new component of common study in the interdisciplinary fundamentals of engineering.

But underlying all these changes is an assumption—and the validity of the assumption is open to question. The accepted hypothesis is that the essential format of the curriculum is correct; that science and mathematics should precede fundamentals of engineering behavior (and the latter should precede application); and that this sequence of formal courses is best for the student. But for what student? Today's generation may not have the same learning psychology and objectives as the preceding generation!

Viewpoints, past and present. The youth of today has a significantly different viewpoint on life and education from that of a generation ago. Contemporary youth, emerging in an affluent society, has the opportunity to question and to deliberate before making decisions. They have time; we did not. We were forced by economic necessity to make a rapid choice of educational avenues to follow; they are not. We had neither the time nor the social status to question. They have both.

They have been encouraged to question the social structure of society, and, in learning how to question, they do not establish limits on the areas of their questions. Thus, in relation to education, they question choices of profession, objectives of the curriculum, content of courses and their interrelationship. This generation, like ourselves, wants breadth of knowledge, but it also wants to know "why," since it will not—as we did—accept as readily the premise that the educator, as the elder, wiser, and experienced member of society, is necessarily correct in what he does. And it will not submit to his authority as rapidly as we did.

Despite its swaggering attitude on the right to question, youth reflects its immaturity in the fear of making wrong choices in courses, curriculum, and profession.

Freedom vs. responsibility. The present generation faces the magnificent opportunity of having the freedom to make decisions—and the awesome and inevitable consequence of accepting the responsibility for having made them. More than any preceding generation, it cannot escape responsibility by having the forces of society and history make the decisions for it.

These problems can be exacerbating at The Cooper Union because the discussions concerning them are amplified and intensified by the high intelligence level of the students.

A final word on professionalism

The essence of professionalism is the practice of an art or science with the objective of an ultimate benefit to the public welfare. It is the superimposition of the sense of social responsibility in one's work that distinguishes the professional from the technician. This attitude cannot be formally taught; it must be developed.

Professionalism cannot be projected by formal courses in the humanities or the social sciences because it is not based upon information. Information is a necessary component that supports concepts or attitudes, but it is not, in itself, the structure of the attitude.

The concepts of social responsibility are best projected by the involvement in the problems of engineering and science as vital components of the solutions. They are best developed by participation in problem solving that incorporates moral and ethical conditions to the solution.

Thus the seminar technique of mutual involvement of the student and the professor in these problems can and should include the projection, by the instructor, of social effects and responsibility. Since we have already established that the instructor in this activity must be a professional and a practitioner of creativity, he must, of necessity, include these considerations as a part of the problem-solving process.

It must be emphasized that we are not in the business of processing warm bodies. We are in the business of developing leaders and individuals—not "off-the-shelf commodities." We believe this new approach to education is more efficient for the production of the "custom-made product."

Data compression by redundancy reduction

Great strides in improving data communication efficiency are being made through the recent development of several techniques for eliminating large amounts of redundant data from the streams of information being transmitted

C. M. Kortman *Lockheed Aircraft Corporation*

Because of the increasingly pressing problem of spectrum overcrowding in data transmission channels, it is becoming more and more necessary to develop schemes to optimize the use of the available frequencies. The Proceedings of the IEEE is devoting its March issue to the subject of "Redundancy Reduction and Bandwidth Saving." This article, timed to coincide with the Proceedings issue, briefly describes some of the salient features of the various approaches to redundancy reduction.

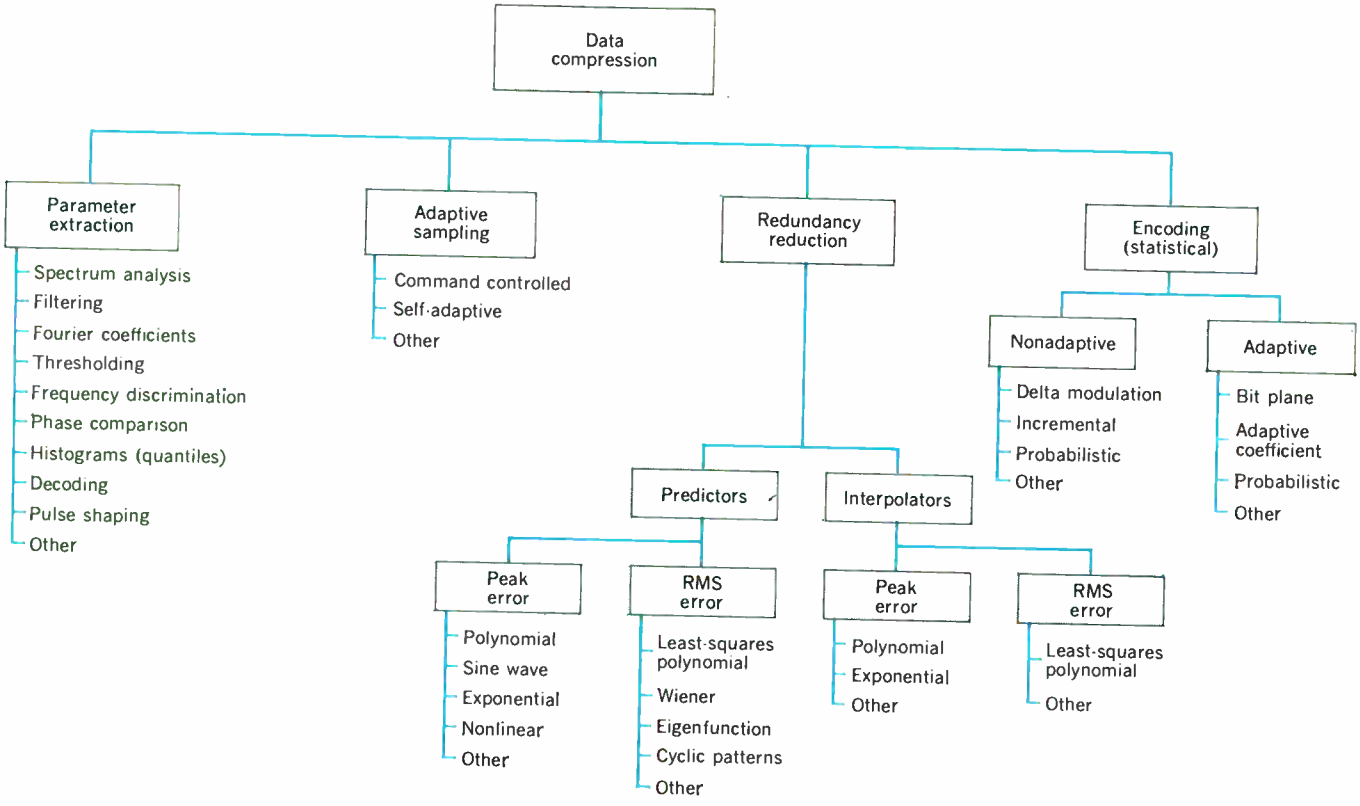
In the nearly 20 years since C. E. Shannon provided a basis for modern information theory, a great deal of study has transpired and much work has been accomplished in applying that theory.[1] However, much of the work has accepted the "information" as it existed and the "channel" was designed to handle the maximum, transient response required rather than to handle the average rate of information flow. This procedure resulted in the transmission of significant amounts of redundant data, simply because there was at that time no practical means for eliminating it.

Relatively recent improvements in techniques—and in devices permitting effective mechanization of these techniques—have provided a means for eliminating a large amount of the redundancy. These techniques have been referred to by various workers as "data compression," "bandwidth compression," "data compaction," "redundancy reduction," etc., and are discussed in considerable detail in the March 1967 issue of the PROCEEDINGS OF THE IEEE, a special issue on "Redundancy Reduction and Bandwidth Saving."

Through the use of these various approaches it is possible to provide a more uniform rate of information flow, with all elements containing meaningful information, so that the application of other facets of information theory is truly useful. In a recent article on interplanetary communication,[2] Dr. L. C. Van Atta states: "Since we are continually faced with a trade-off between power density (watts/cycle) and bandwidth (cycles/sec), we must learn to make maximum use of the bandwidth available at each stage of progress. The data acquired in the vehicle, be they scientific or engineering, video or voice, must be subjected to sophisticated data-reduction and compression techniques in order to assure high efficiency in transmitting the truly essential information."

Figure 1 is a schematic classification of these methods as presented in the previously referenced issue of IEEE PROCEEDINGS.[3] In this grouping, all are broadly classified as "data compression," which is defined as a technique for reducing the bandwidth needed to transmit a given amount of information in a given time or for reducing the time needed to transmit a given amount of information in a given bandwidth. The group titled "parameter extraction" probably includes what Dr. Van Atta referred to as "sophisticated data reduction." Parameter extraction is defined as a technique that reduces the bandwidth required to transmit a given data sample by means of an irreversible information-describing transformation. These transformations are considered irreversible in that, although they provide useful descriptions of the input signal, they so distort the signal that it is impossible to reconstruct the original waveform. Parameter extraction represents the oldest and most widely used form of data compression.

Adaptive sampling is a means by which the sampling rate of a given sensor can be adjusted to correspond to its information rate. The reasoning applied is that the data would not be redundant and therefore not compressible, provided the sampling rate was continuously and perfectly matched to the activity of the data source. Most of the time, present telemetry systems greatly oversample the data. In nearly all cases, the sampling rate is set on the basis of the fastest expected response from the source and not on the basis of the quiescent or normal value. However, to match the sampling rate to the data activity would require an extremely complex machine and to date a practical mechanization has not been developed.



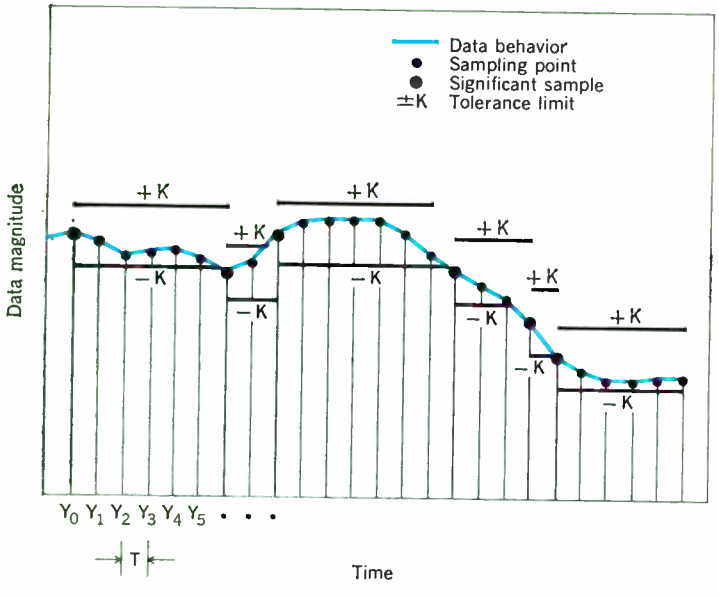


FIGURE 1. Classification of data-compression models.

FIGURE 2. Zero-order floating-aperture predictor.

Redundancy reduction is a technique for eliminating those data samples that can be derived by examination of preceding or succeeding samples, or by comparison with arbitrary reference patterns. The basic difference between adaptive sampling and redundancy reduction is that in adaptive sampling the sampling rate of the original data waveform is varied, whereas in redundancy reduction the waveform is initially sampled at a constant rate and the nonessential samples are eliminated later. However, the final result is essentially the same in that an output is provided only when the data change exceeds a predetermined tolerance.

Encoding involves the transformation of a given message into a corresponding sequence of code words. As in the cases of adaptive sampling and redundancy reduction, an effective coding method requires sequential message words to exhibit a high average correlation. This sequential technique has been the subject of a great deal of study in the field of information theory and will not be discussed further here except to point out that it is generally a reversible process and may be combined with other methods, such as parameter extraction or redundancy reduction, to provide greater overall efficiency of data compression.

Redundancy reduction

Shannon has defined redundancy as "that fraction of a message or datum which is unnecessary and hence repetitive in the sense that if it were missing the message would still be essentially complete, or at least could be completed." Redundancy exists whenever the sampling rate of a multiplexer exceeds the frequency required to describe the input function in accordance with the accuracy requirements of the user.

The choice of reference patterns used to detect redundancy is virtually unlimited. Polynomials, exponentials, and sine waves are good examples of reference patterns by which real data can often be approximated. Arbitrary cyclic patterns, such as the periodic components of an electrocardiogram or a commercial television picture, can be used as references to detect redundant data from a given sensor. The process of redundancy

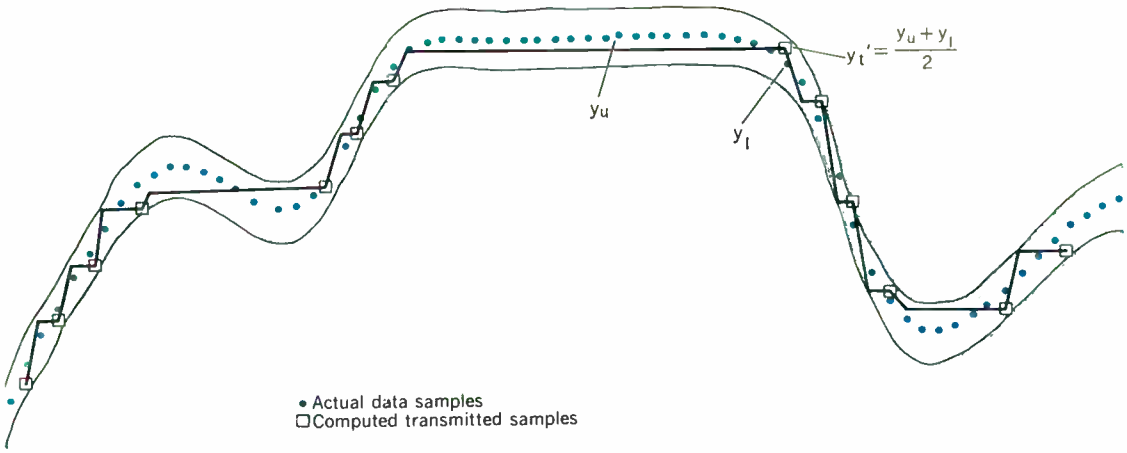


FIGURE 3. Zero-order polynomial interpolator.

reduction can be achieved by means of "prediction" from a priori knowledge of previous samples, or by a posteriori "interpolation" in which future samples are utilized.

For redundancy reduction to achieve reasonable compression efficiencies, it is often necessary to introduce certain errors. These errors are caused by filtering or thresholding (or both) within the redundancy reduction process and do result in slight reductions in the source entropies. However, unlike parameter extraction, adaptive sampling and redundancy reduction are designed so that the original source waveforms can be reconstructed with a guaranteed fidelity. This fidelity can be established to supply the data within the accuracy requirements of the user.

Of the many techniques for redundancy reduction that are possible, at present, the most effective and widely used are the polynomial predictors and interpolators since they give the closest approximations to most real data. Therefore, only the polynomial methods will be included in this discussion.

Predictors

A predictor is an algorithm that estimates the value of each new data sample based on past performance of the data. If the new value falls within the tolerance range about the estimated new value, it is rejected as redundant since it is known that the data value can be reconstructed within the specified tolerance. The zero-order predictor is the simplest. The algorithm predicts that each new data value will be the same as the last transmitted value, within tolerance limits. A version of the zero-order predictor employing a floating aperture, shown in Fig. 2, functions as follows. If y_0 was the last transmitted data sample, a prediction is made that subsequent data samples $y_1, y_2, \ldots y_5$ will be within K percent of y_0. As shown in Fig. 2, these samples are within the tolerance corridor denoted by $y_0 + K$ and $y_0 - K$ and can be discarded as being redundant. Each sample falling on or outside of the corridor must be transmitted as a nonredundant, or significant, sample and is then used as the new reference for subsequent predictions. This use of the new, significant sample as the reference for future predictions results in a tolerance corridor that follows the data and is referred to as a "floating aperture." The term "zero-order polynomial prediction" implies that the redundant portion of a time function will be approximated by a horizontal straight line.

The polynomial prediction philosophy can be extended to include higher orders of polynomial redundancy reduction. Although higher-order polynomial predictors will, at times, provide a greater compression efficiency on highly active data, experience has shown that for much telemetry data the simpler zero-order predictor will provide equal or greater compression efficiency.

Interpolators

A prediction is a guess that can be effective only if the characteristics of the data remain relatively constant from one time interval to the next. If the data are varying continuously in a random manner or if they are perturbed by high-frequency noise, the redundancy reduction efficiency of the predictor will generally be low for reasonable system accuracies. Examination of such data indicates that a greater number of redundant samples could have been eliminated if both future and past data samples had been used. The process of after-the-fact polynomial curve fitting to eliminate redundant data samples is termed interpolation.

Zero-order polynomial interpolator. The functional operation of the zero-order polynomial interpolator algorithm is shown in Fig. 3. The zero-order interpolator is similar to the zero-order predictor in the sense that a horizontal line is used to represent the largest set of consecutive data samples within a prescribed peak-error tolerance. The primary difference pertains to the sample selected to represent the redundant set. It is seen in Fig. 3 that the transmitted sample for the interpolator is determined at the end of the redundant set as contrasted with the first sample in the case of the predictor. Furthermore, the transmitted sample y_t' used for the interpolator is computed as the average between the most positive sample y_u and most negative sample y_l in the set. All samples in the set are within the prescribed peak-error tolerance from the transmitted sample.

First-order polynomial interpolator. As the name implies, the first-order polynomial interpolator algorithm approximates the data with a first-order polynomial curve.

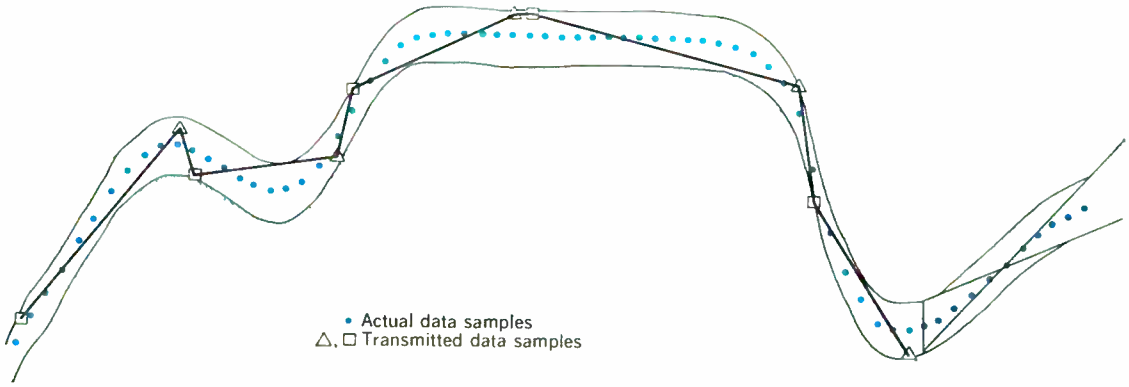


FIGURE 4. First-order polynomial interpolator.

Several methods exist for representing redundant samples by a straight line segment. To achieve the largest compression ratio, it is necessary to select a line segment that is within K percent of as many samples as possible. This optimum first-order algorithm requires freedom of both the starting and end points of the straight line, resulting in four degrees of freedom. The performance of this optimum process is shown in Fig. 4. It is seen that both the starting and end points of each line are values computed in such a way that the length of each line is maximum for the peak-error guarantee. Also, the end point of one line segment can be connected with a straight line to the beginning of the following line segment. Since the four-degrees-of-freedom first-order interpolator is a complex process to implement, it is reasonable to seek less than optimum approximations that can be mechanized more easily. Several such approximations have been developed and are described in the referenced issue of IEEE PROCEEDINGS.

Implementation

Figure 5 shows the essential components required to implement any of the redundancy reduction algorithms that have been discussed. The reference memory stores the data values, tolerance limits, algorithm selection (if needed), and any additional information required to enable the comparator to determine whether each new data value is significant or redundant. If the new value is redundant, the reference information is returned to the reference memory and the next data value is examined. However, if a new value is significant, the reference information is updated and returned to the reference memory. At the same time the significant data value is inserted into the buffer memory and stored until it can be read out.

In practical data compression systems the buffer must be able to permit the acceptance of significant data samples at an irregular (random) rate according to data activity and also be able to transmit the data at a constant, regular (but slower) rate that simplifies the recovery of the data at the receiving end. At this point it is possible to define one measure of data compression efficiency, termed the "element compression ratio." This is the ratio of the number of data values presented at the input to the number of significant data values delivered to the buffer memory during a specific time interval. Although this ratio is a basic measure of the ability of a particular algorithm to remove redundancy, it is not the significant measure of the efficiency of a data compression system to conserve bandwidth or power, or both. The difference arises from the necessity to identify the significant data values with an address, a location, and/or time of occurrence so that it will be possible to reconstruct a time history of each data source within prescribed error tolerances. It is also necessary to provide synchronization information in the output bit stream. Therefore, a second measure of data compression efficiency, the "bandwidth compression ratio," has been formulated. It is the ratio of the number of bits presented at the input to the number of bits delivered at the output of the data compressor. Since it includes all penalties for identification, timing, and synchronization, it is a true measure of overall compression efficiency. This is the compression ratio that is cited in the performance examples presented later.

The operation of the data compressor is controlled by the timing and control logic. A more detailed explanation of data compressor operation is given in reference 3 in the description of an actual machine designed for space vehicle use. However, before performance examples are presented, some additional functions of a practical data compressor should be discussed. The first of these is buffer load or fullness control, the need for which arises from the possibility that, over a period of time, the number of significant data values read into the buffer will exceed the number read out over the same time period. If the buffer is large enough, all the data can be stored and eventually read out, provided that the average input rate drops below the readout rate for a sufficient period. In most practical applications, however, unlimited buffer capacity is not available and thus there is the danger of buffer overflow during periods of high data activity. If the buffer is permitted to overflow, there is uncontrolled loss of data and no way even to approximate what has happened. Therefore, several methods of preventing overflow have been developed.[4] Among these are the deletion of certain low-priority data so that a greater activity in high-priority data can be handled. Another method is to increase the tolerance band on all or a selected number of data sources to reduce the number of data values accepted as significant. All methods result in the degradation of some or all of the data, but in a con-

trolled manner so that the data are still useful.

Still another function that can be realized is effective presampling filtering, accomplished by utilizing the data compressor memory and some additional computation and logic capability. It is considered good design to filter all input signals prior to multiplexing to avoid frequency foldover of unwanted instrumentation noise (commonly called "aliasing errors"). However, most instrumentation systems cannot afford the luxury of presampling filtering because of size, weight, and cost considerations. As an alternate solution, the sampling rate of the multiplexer can be increased sufficiently to pass the system noise and digital filtering applied to average out the unwanted noise components. This solution is particularly attractive when redundancy reduction is included as a part of the telemetry system, because the reference memory can be used to provide the time delay required for filtering.

In the implementation of a link by the use of redundancy reduction there are some additional factors to consider. A very important factor is that of noise. Through years of practical experience, yardsticks of acceptable signal-to-noise ratios have been established for various kinds of data and different methods of transmission. Nearly all of these are heavily influenced by the inherent redundancy in the data. Therefore, if this inherent redundancy is reduced, there is an intuitive feeling that a higher signal-to-noise ratio should be required if the original error rate is to be maintained. Alternatively, some redundancy can be reinserted into the data in the form of coding to maintain the desired error rate without an excessive increase in required signal-to-noise ratio. There are many trade-offs involved in determining optimal parameters for a specific data link and no ready rule of thumb exists at this time.

A final factor to consider is that of "decompressing" the data—that is, at the receiving end of the link, either restoring the data to its original form, or putting it into a form suitable for its end use. Although this article and most of the others written about data compression deal almost entirely with the compression part of the problem, there are some basic differences in the algorithms that affect the decompression or reconstruction process. These differences have been overlooked by many workers in the field and have led to some false conclusions concerning the relative merits of various algorithms.

As defined herein, the predictors are relatively easy to decompress. Since decompression is the inverse of compression, redundancy must be reinserted into the signal. For analog decompression this is easily done by the use of a uniform time base such as that provided by a strip chart recorder. The indicator (pen, stylus, etc.) is set to the initial value (y_0 in Fig. 2, for example) and draws a straight line with the predicted slope (a horizontal line for the zero-order predictor) until a new value is received. This means that the original data had varied beyond the bounds of its tolerance band and, therefore, the reconstructor must shift to the new value and continue until another new value is received. This method will provide a sharp-cornered approximation (staircase for the zero-order predictor) of the original data, which may be smoothed by various methods. Such a record is quite satisfactory for real-time or near-real-time readout of certain data channels. However, the interpolators as defined herein present a very different problem to the receiving-end decompressor. For these algorithms the representative data values and slopes are not known, and thus not transmitted, until the longest possible line segment has been fitted to the data. Therefore, it is impossible to plot a reconstructed replica of the original data without imposing some delay. How much delay becomes a problem.

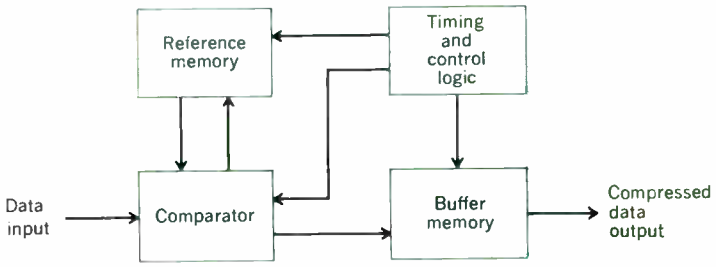
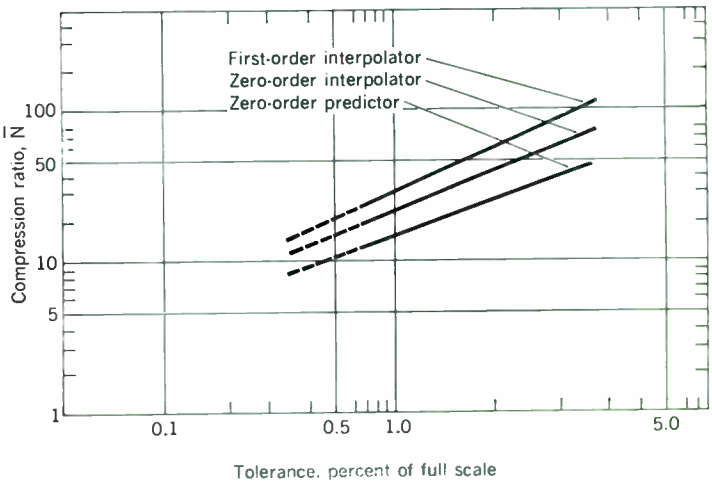


FIGURE 5. Simplified block diagram of redundancy-reduction data compressor.

FIGURE 6. Compression characteristics of Polaris flight data.



Consider using an interpolator algorithm on a dc battery voltage (a poor choice). Since the voltage does not change, the data compressor could wait for days or weeks for a terminal value to use in computing the representative value or values and slope. Meanwhile the decompressor or reconstructor has received nothing to reconstruct and so provides no data output. This problem can be averted in many ways, but the illustration is given to point out this basic difference between predictors and interpolators and to alert potential users of the need to consider this factor in overall system design.

Performance

Data compression in general and redundancy reduction in particular may be applied to a wide variety of data transmission and processing problems. These include telemetry, video, voice, and facsimile. Examples of the performance of these techniques on many types of data are given in the referenced issue of IEEE PROCEEDINGS. The examples that follow were taken from extensive computer simulation studies in which the data com-

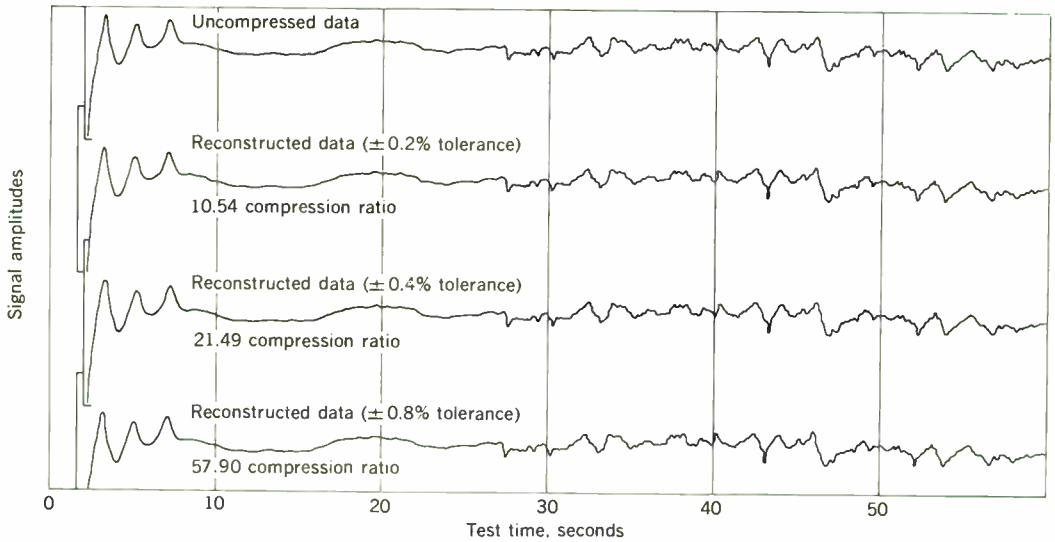


FIGURE 7. Polaris data reconstruction.

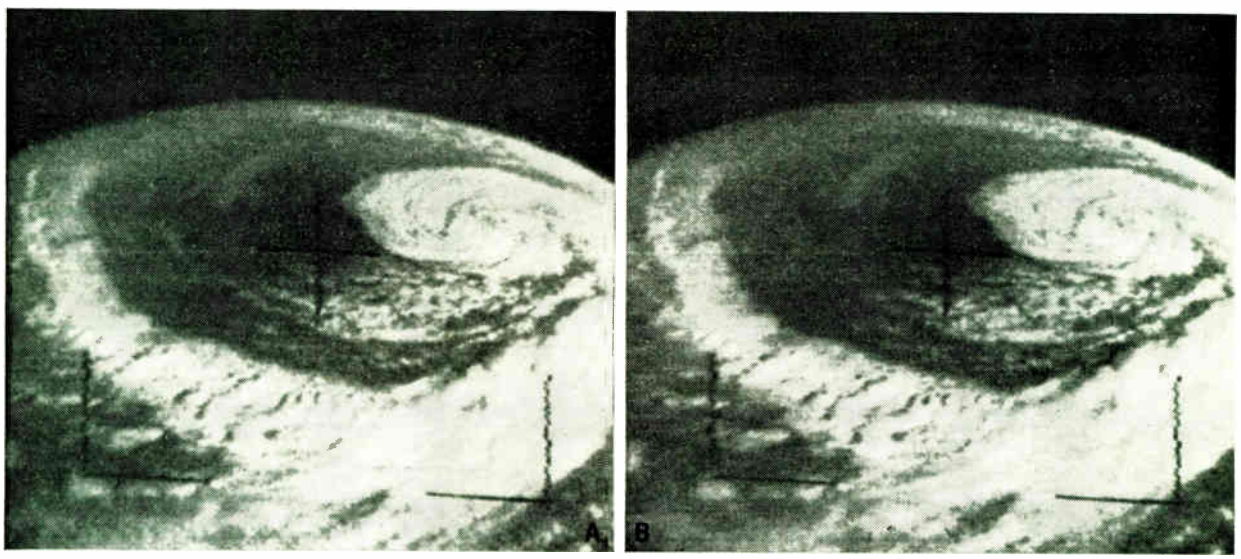
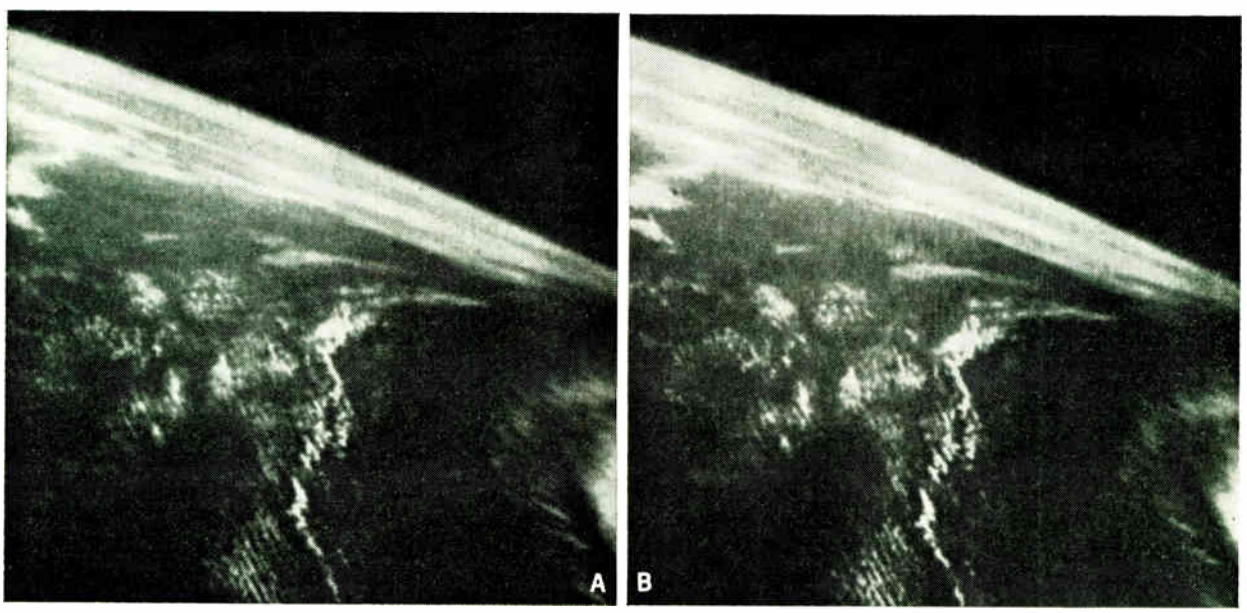


FIGURE 8. TIROS video data. A—Original. B—Compressed, with BWCR ≈ 4.

FIGURE 9. Gemini video data. A—Original. B—Compressed, with BWCR ≈ 6.



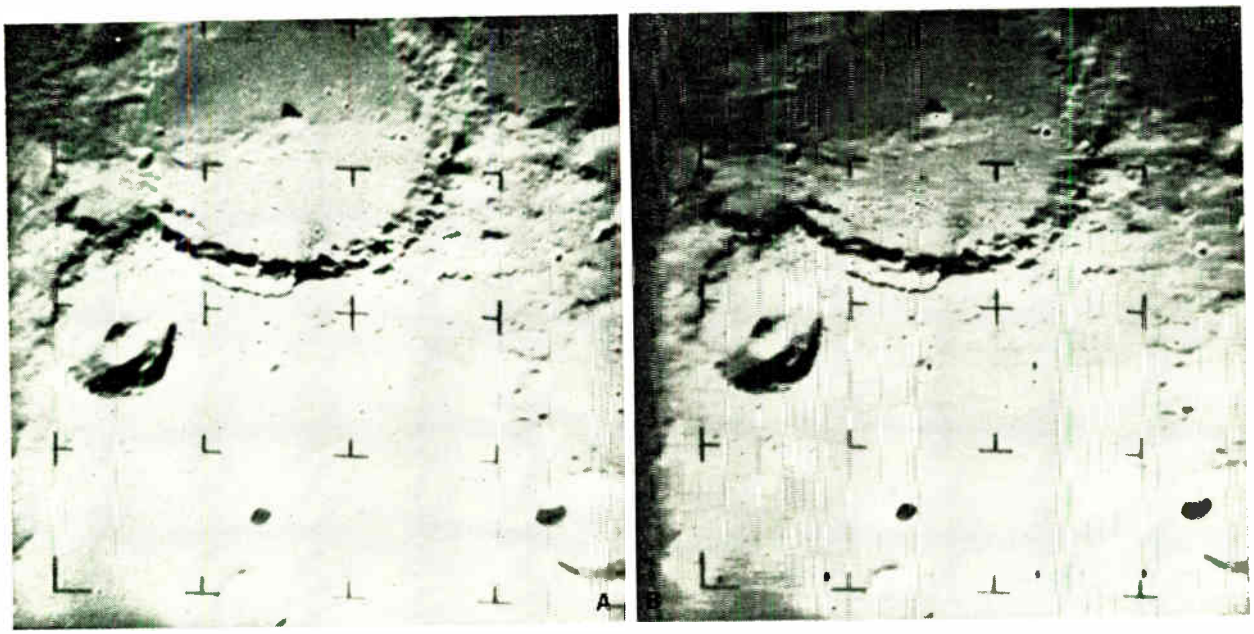


FIGURE 10. Ranger video data. A—Original. B—Compressed, with BWCR ≈ 5.

pression and reconstruction were performed by a general-purpose computer.

Telemetry data. Data obtained from a Polaris flight were analyzed to determine the performance characteristics of several redundancy-reduction algorithms. Figure 6 shows the relationship between compression ratio $\bar{N}$ and tolerance K for the zero-order predictor, zero-order interpolator, and disjoined first-order interpolator. These results have shown that the data from a complete FM/FM telemetry system occupying 80 kHz of bandwidth could have been transmitted over a 3-kHz voice-grade telephone line in real time had data compression been employed. Figure 7 illustrates the fidelity of a servo error signal for several compression ratios, following compression and reconstruction. It is evident that for even large tolerances and compression ratios the usual fidelity of measurement data can be maintained.

TIROS television data. Based upon the success in compressing sampled telemetry data, the more efficient redundancy reduction algorithms were applied to TIROS television data. Figure 8 shows a reproduction of an original picture combined with an example of reconstructed compressed data. Excellent fidelity is achieved for bandwidth compression ratio (BWCR) values as high as 4 to 1.

The original TIROS picture contained moderate levels of recorder and transmission noise. Because redundancy reduction cannot distinguish between information and noise, a penalty in terms of compression efficiency results when data transmitted by analog techniques are subsequently compressed. To obtain a greater feel for the penalty imposed by such noise, some high-resolution photographs were compressed and reconstructed using similar redundancy reduction algorithms. Figure 9 shows the original and a compressed example of a photograph taken on a Gemini mission. A somewhat higher bandwidth compression ratio was achieved. Figure 10 is a similar example of a Ranger photograph of the lunar surface. Even though overall bandwidth compression ratios of 4 to 6 are impressive, new improvements in techniques promise even higher ratios. It is appropriate to note here that the maximum benefit accrues when the compression is accomplished as near to the data source as possible.

Applications

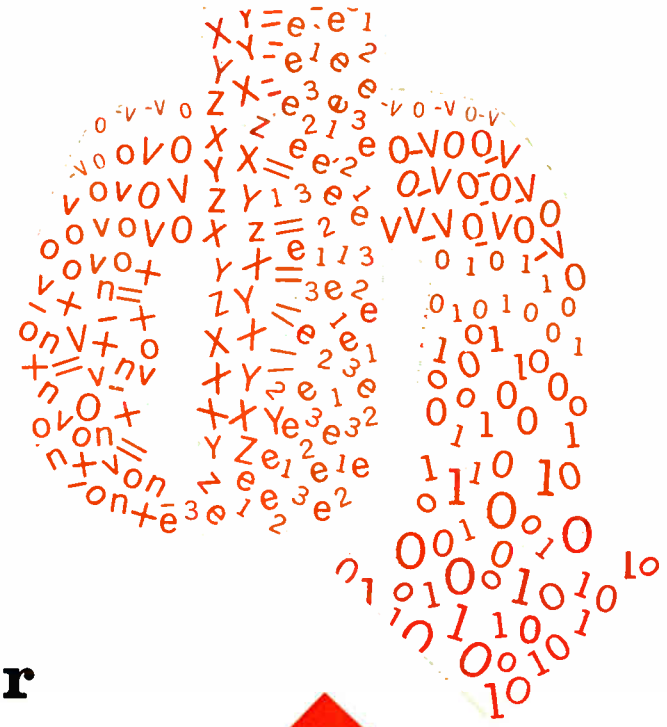
The applications of data compression are virtually unlimited and equipment has been designed for both vehicle and ground use. A detailed description of a vehicle-borne data compressor and of several ground applications are given in reference 3. Additional applications are discussed in other papers included in the IEEE PROCEEDINGS previously referenced and therefore will not be repeated in this article.

Conclusions

Data compression by redundancy reduction provides the communications engineer with a much-needed tool for decreasing the redundancy in this data, thereby enabling him to transmit and process only true information—that is, changes from some norm or reference. Many techniques are now available and much effort is being expended in determining the best ones to use for specific types of data. All of these methods represent practical applications of information theory and may be combined with other methods to improve the overall channel efficiency.

REFERENCES

1. Gilbert, E. N., "Information theory after 18 years," *Science*, vol. 152, pp. 320–326, Apr. 15, 1966.
2. Van Atta, L. C., "Interplanetary communication," *Internat'l Sci. Tech.*, pp. 50–61, Nov. 1966.
3. Kortman, C. M., "Redundancy reduction—a practical method of data compression," *Proc. IEEE*, vol. 55, Mar. 1967.
4. Medlin, J. E., "The prevention of transmission buffer overflow in telemetry data compressors," submitted for publication.



Foundations of the case for natural-language programming

Misconceptions and conflict have long impeded useful discussion on the question of the suitability of natural language for programming. This article sharpens the issues and argues the case for natural language

Mark Halpern

Lockheed Palo Alto Research Laboratory

There are more and deeper issues involved in the notational question in computer programming than are covered in the usual easy antitheses between natural sloppiness and formal precision. This article argues that it is far from clear that a formalism patterned on mathematical notation is the answer to any burning problem in practical programming. In arguing the case for natural language, it is often as necessary to take issue with those favoring the cause as with those opposed to it, for the natural-language "party," also, has been guilty of perpetuating invalid ideas. Thus, this article attempts to lay the groundwork for more useful exchanges between the formalist and natural-language schools.

> Few things worth saying can at once be said clearly; and whereof one cannot yet speak clearly, thereof one must practise speaking.*
>
> —*Wittgenstein renovated*

A running debate, mostly subterranean, has long been going on over the suitability of natural language for use as a programming language. From time to time the debate surfaces in the form of sharp exchanges at technical conferences and strong letters to the journals, but these casual encounters have been insufficient even to make clear to the general reader what the issues are, let alone to resolve them. The absence of open and lively debate between those who favor and those who oppose natural-language programming has left the problem to be dealt with by each language designer as best he can, without benefit of others' experience and ideas. Two opposite but equally undesirable ways of handling the conflict are in common use: one is a mutual turning of backs by the two parties, as may be seen in the increasingly wide gulf between research and practice in the design of programming languages; the other is a tendency toward superficial, makeshift compromise, with the usual result of satisfying no one. The possibility that the issues underlying the controversy are too fundamental to allow any useful exchange between the two parties cannot be dismissed, but it seems worth some effort to find out; the result should be at least a clearer idea of what we are disagreeing about.

The root issue may be put roughly this way: one school believes that a programming language need not and should not have its form dictated by the fact that it is in some sense addressed to a machine, but should be very close to the language its intended users ordinarily employ in their work, apart from the computer. The other school believes that the fact that a programming language is addressed to a machine is the inescapably decisive force in determining its shape, and that such a language will almost certainly be quite different from those in use between man and man. The first or "natural-language" school is often considered to be advocating the use of plain English as a programming language. This characterization, true so far as it goes, is a simplification of their position that is liable to serious misinterpretation; it may easily be taken to mean, for example, advocacy of languages such as Cobol, which it does not. The second or "calculus" school (as we shall call it) sees programming languages as needing a massive infusion of the rigor, precision, and economy exhibited by mathematical notation. Its members often suggest that a language with these qualities would amply repay its users for their trouble even if it were not implemented on a computer, because it would give them for the first time a proper representation of their procedures (see box below).*

There is no quarrel between the two schools over how to call for numerical computation; they agree that the algebraic notation commonly offered by compilers is right for that purpose. Their agreement on this point, however, represents no new insight on either side, but only a happy coincidence of prejudices: in mathematical (and symbolic-logical) notation we have the one language that is both a natural language, in the sense defined above, and a calculus. The dispute over almost all other computer applications continues. In this dispute the writer makes no pretense to neutrality; this article, as its title indicates, is an attempt to lay the foundation for a case in favor of natural-language programming. In doing so, however, we will have almost as much occasion to take issue with others favoring that cause as with those opposing it, for we of the natural-language party have been remiss in developing and presenting our ideas, and have been guilty of perpetuating a number of errors.

The problem—what it is and what it isn't

It is of the utmost importance to identify correctly the problem that natural-language programming proposes to deal with. A number of the proposal's critics have pointed out that by far the greater part of the effort that

* Note: Digressions from the author's ongoing argument are set off in boxes.

> Occasionally there are signs that pedagogical and moralistic considerations enter into the calculists' thinking, as when H. Zemanek says "we do not want the easy application that a natural language offers because the user would then not reflect enough on what he instructs the machine to do."[1] There is a hint here of satisfaction at the discipline imposed by the computer on its users; some suggestion that easy, natural-language access to it would be a tragic waste of the opportunity it offers to straighten out the confused, woolly thinking of nonmathematicians. In other contexts, of course, the boast of the mathematician is that his notation spares him from thinking about what he is doing, allowing him to develop his argument by purely formal manipulation, to avoid the trap of assigning a meaning—a physical interpretation—to intermediate results, and to venture into realms of abstraction where "meaning" in this sense has no meaning.

* The epigraph to this article is a play on a dictum by the Austrian-born philosopher Ludwig Wittgenstein that runs, in translation, "Everything that can be said can be said clearly. Whereof one cannot speak, thereof one must be silent." In this original form it was adopted as the motto of the formal, printed presentation of the Algol language, and well expresses the viewpoint of the school called the "calculists" in the present article. It seemed particularly appropriate, therefore, that the opposing viewpoint be summed up in a variant form of the same dictum.

The suggestion has been made that the programming problem might be sidestepped by systems that would take the initiative in the man–machine dialogue, presenting to the user at each step a number of alternatives from which he would select one. This technique has its uses, but these seem to be limited to cases where the number of logical alternatives is nowhere greater than about six. Its area of practical application, therefore, is not in the programming of new procedures, but the parameterization of existing, general procedures such as the report-program generator. For broader applications, the CRT screen would have to show the equivalent of a conventional programming manual at each step.

Natural-language programming will come into its own, of course, only when supported by a system permitting vocal communication with the computer. So long as a keyboard stands between man and machine, unpremeditated and informal programming is blocked, and the potential advantage of natural language unrealizable. Happily, projects designed to achieve man–machine voice communication are under way at a number of laboratories, and are sufficiently promising of practical results to encourage the parallel pursuit of a natural-language programming capability.[6,7]

goes into programming lies in the analysis of the problem and the design of the algorithm, making the question of what notation the algorithm is finally expressed in relatively unimportant.* This point is not always valid, we suggest, even for professional programmers—in many routine applications, the filling up of the coding form may take more time and effort than the design of the algorithm, which may in fact be done in the programmer's mind as fast as he codes. But even if the critics' point were always valid, it is irrelevant to the problem for which natural-language programming gives promise of being a remedy. The problem we have in mind is that very many computer customers are perfectly capable of performing the hard part of programming—the analysis of their problem and the design of an algorithm to deal with it—but are frustrated as potential programmers by their unfamiliarity with the easy part—the exact form of the currently acceptable programming languages. We can agree with the critics that this is for professional programmers often the trivial part of their total task; all the more intolerable, then, that it should be for so many thousands of highly trained (but nonprogramming) professionals a practically insurmountable barrier between themselves and the machine. These people find it as a rule impossible to practice their own demanding disciplines and also master the myriad details of the programming systems nominally available to them; as a result, they are dependent on professional programmers. These latter are growing ever scarcer relative to both available computing power and the number of potential customers for it. The problem of the programming-language barrier, then, may be a trivial one to the professional programmer, but its practical importance to many others would be hard to exaggerate.

* See, for example, the discussion appended to reference 2.

English, active and passive

"Natural-language" programming, it cannot be too strongly emphasized, means programming in whatever terminology is standard for the application in question; this is not necessarily the ordinary English vocabulary. (As we noted earlier, the natural language for numerical computation is algebraic notation.) These pages are largely devoted to the particular natural-programming language that is ordinary English because doing so puts the issues involved in the sharpest relief.

One reason why it can be dangerously misleading to talk about "English-language programming" is that the speaker usually has in mind *active* English, whereas the listener understands him to mean *passive* English. The difference between them is critical, and confusion on this point to fatal to any possibility of understanding. Passive English is a language that *reads* sufficiently like English so that almost anyone with some idea of the application involved can understand a procedure described in it, but only those trained in its use (and, usually, with ready access to a manual) can *write* it. The Cobol language is an example of passive English. Its "natural" appearance to the reader of a listing gives little hint of the difficulties in writing it, and the misguided reader who supposed from its appearance that he could use the language without training and reference to a manual would soon discover that it had all the trickiness of any professional programmer's language.

An active-English programming language would permit the free use of a subset of natural English for the *writing* of programs, the subset being open-ended and limited only by those constraints inherent in the application, not by any arbitrary decisions of the system's designer. This means that users would be offered not a fixed number of stereotyped statements to be used exactly as given in the manual except for the replacement of dummy operand names by actual ones, but a stock of words that may be used in any reasonable, straightforward way, and altered or expanded as necessary. No example of this facility can be cited among operational programming systems.*

The distinction is a critical one because the two kinds of English, similar though their appearance on a listing may be, are as far apart in suitability for the role of programming language as can be imagined. For passive English there is simply nothing good to be said. Combining the wordiness and noisiness of a natural language with the rigidity and arbitrariness typical of programming languages, it exhibits the worst features of both and the virtues of neither. (The readability of a Cobol listing is valuable, but has nothing to do with the use of English as a source language; a processor can easily be made to produce such documentation regardless of the input notation.)

English can be justified as a programming language only if it is *active*. (We shall mean active English when we use the term "English" in what follows.) When it can be used

* The writer and a group of associates are engaged in an effort to realize such a system,[3] and a few projects reported in the literature seem to share our approach, in spirit at least.[4,5]

actively, its potential importance is immense: it could make programming just enough easier to let many who are now dependent on the services of a professional programmer get at the machine directly. As we noted earlier, the growth of the computer's availability to greater numbers of users is now and for the foreseeable future communication-limited, and every promising attack on this constraint should be energetically pressed. The number of people able and willing to make programming their life's work is probably already near its limit; we can no more expect a professional programmer to be available to everyone who needs the computer than we can provide a chauffeur to everyone who needs a car. And even if programmers were plentiful, there would always be many for whom no amount of professional help would be as satisfactory as direct contact (see box on page 142).

Translation versus comprehension

The other great distinction that must be made, just as important as the one between active and passive language, is that between *translation* and *comprehension* of natural language by computer. Many of the skeptics about natural-language programming are mistakenly supposing that the processing difficulties it would involve are as formidable as those that have made fully automatic translation of, for example, Russian to English so elusive. But there is a radical difference between translation—the conversion of a text from one natural language to another—and comprehension—the execution of a procedure described in a natural language. Translation systems treat natural-language statements as data; comprehension systems treat them as commands. In the former the user talks to another human *through* the computer, which is programmed to perform on Russian-language input a transformation under which all information-bearing features of the text are preserved with only their representation undergoing change. The difficulties in doing this are too notorious to need any discussion here.[8]

In a comprehension system the user talks *to* a computer in order to generate coding or to parameterize existing coding; the required processing of source statements, as compared with that demanded by translation, is simple and straightforward. There are two saving graces in comprehension: one is that the user painlessly foregoes practically all the natural-language features that offer great difficulty. Users of such systems do not care to chat idly with the machine; they want to give data and order that certain procedures be performed on that data. (The procedure specifications may be couched as questions or requests, but are commands nonetheless.) The syntactic complexity of the statements they will want to make is, accordingly, unlikely to be great. Similar considerations suggest that the vocabulary to be dealt with should be within the powers of a realizable system. Users will address such a system within the framework of some particular application, and this is a powerful force for elimination of ambiguity, since it is not often that terms are ambiguous within the vocabulary of a single discipline (see box on page 144).

The other saving grace is that the general form of the target language—machine language—is a fixed one, independent of that of source-language statements, and known in advance to the processor. The comprehension process is therefore a restricted many-to-one transformation, not an indefinitely-many-to-indefinitely-many mapping such as translation. The man–machine communication channel offered the user by a comprehension system may be thought of as an ear trumpet worn by the computer, with a wide but well-defined mouth toward the user, and a rather fine-meshed filter somewhere downstream of that mouth.

The practical promise of such a system lies in the indications that the necessary constraints, although powerful enough to spare the comprehension system the most formidable of the problems facing a translation system,* are in no way arbitrary; they should leave the user feeling unimpeded, since they prevent him from doing only what he is little tempted to do in any case. The grotesque sentences that are regularly exhibited in papers on mechanical translation to show how profoundly ambiguous English can be, featuring labyrinthine syntax and words with five meanings (each of which suggests an entirely different interpretation of the whole), have no bearing on the present problem. None of these features is at all likely to appear in a statement composed by a serious worker trying to enlist the computer's help in solving a problem; if once in a great while they do, the system will be doing him a favor by rejecting it.

Further, such a system offers the possibility of a programming language that includes all the semantic and syntactic resources the users of an application-oriented procedure would naturally employ in invoking it. Such a language would be largely built by the users themselves, the processor being designed to facilitate the admission of new functions and notation at any time. The user of such a system would begin by studying not a manual of a programming language, but a comparatively few pages outlining what the computer must be told about the location and format of data, the options it offers in output media and format, the functions already available in the system, and the way in which further functions and notation may be introduced. He would then describe the procedure he desired in terms natural to himself.

The system's ability to comprehend immediately (i.e., compile correctly from) his description could be guaranteed only if he restricted himself to those constructions and that vocabulary already introduced to it; although these conditions will usually be met without conscious effort by any user who has had a few earlier contacts with the system, this high probability of immediate success is not what the system finally depends on. Its real strength is its permanent readiness to be introduced to new notation, even by ordinary users who are by no means professional programmers. The substitution of linguistic open-endedness and cumulative improvability for a fixed notation (whether passive English or a formal calculus) should go far to relieve many users of daily dependence on professional programmers, and in doing so generate a new wave of applications.

Objections and counterproposals

With this understanding of what "natural-language programming" means, we are prepared to examine the controversy over it. The richest source in print of argu-

* Substantially the same distinction is made by D. M. MacKay.[9] He is chiefly interested in making clear the inadequacy of what is here called comprehension for completely general communication with the computer, but agrees with us that the "nonlinguistic" treatment of linguistic tokens we propose is sufficient "to enable the computer to accept data and answer questions in verbal form."

ments against it seems to be the writings of Prof. E. W. Dijkstra (of the Mathematics Centre, Amsterdam), who has expressed his opposition vigorously, lucidly, and at substantial length. We will draw heavily for diabolical advocacy on two publications of his;[10,11] the first of these is a letter that seizes the occasion of the appearance of a report on the MIRFAC[12] system to attack the entire natural-language concept.

The heart of the case Dijkstra makes against natural-language systems is that there is a sharp and ineradicable contrast between human and mechanical response to instructions:

> If we instruct an "intelligent" person to do something for us, we can permit ourselves all kinds of sloppiness, inaccuracy, incompleteness, contradiction, etc., appealing to his understanding and common sense. He is not expected to perform literally the nonsense he is ordered to do; he is expected to do what we intended to order him to do. A human servant is therefore useful by virtue of his "disobedience." This may be of some convenience for the master who dislikes to express himself clearly; the price paid is the nonnegligible risk that the servant performs, on his own account, something completely unintended.
>
> If, however, we instruct a machine to do something we should be aware of the fact that for the first time in the history of mankind, we have a servant at our disposal who really does what he has been told to do. In man–computer communication there is not only a need to be unusually precise and unambiguous, there is—at last—also a point in being so, at least if we wish to obtain the full benefits of the powerful obedient mechanical servant. Efforts aimed to conceal this new need for preciseness—for the supposed benefit of the user—will in fact be harmful; at the same time they will conceal the equally new possibilities in automatic computing of having intricate processes under complete control.

The picture Dijkstra draws of man addressing machine is in fact quite unrealistic, and the oversight he commits explains much of the misunderstanding between him and those whose views he is attacking. (For that matter, the

Although the novelty in the phrase "natural-language programming" lies in the first term, we must not forget the second. The application to which we are proposing to put the language is programming—that is, the calling on functions that a computer has been wired or coded to perform on demand. The only advantage sought in such use of natural language is a significantly greater ease of calling upon those generators, subroutines, and macros that have already been stored in the machine; it will make no more sense to address to such a system a natural-language statement that does not refer to one of these routines than to write a check without funds behind it. Any attempt to realize a system of this description, then, must begin with a body of programmed functions for application to a well-defined class of problems, and a body of potential users of them. For each function, the system is taught the various statement forms by which these users will attempt to call for it. (Note that we do *not* proceed by contemplating arbitrary English statements and asking what operational interpretation should be placed on them if encountered; we begin with known operational routines, and ask how they might be called upon.) There is little chance that all the needed forms will be anticipated at the outset, but the inadequacy of the first-generation vocabulary and syntax is expected and allowed for in our plans. The key here is that new statement forms can be taught to the system as fast as experience shows them desirable, and taught to it by the users themselves in the course of dealing with it. Defining new notation to the system does not constitute programming in the sense that defining new functions would; it does not, in particular, require any knowledge of the machine itself as programming in the usual sense always does. The procedure whereby this teaching of notation is accomplished (explained elsewhere[3] in detail) does not require that each trivially different statement be individually introduced. All that need be defined are the separate words and symbols (other than "noise" words) and such variants on a standard syntactic pattern as may be required. The notation that emerges from this empirical evolution will presumably converge asymptotically toward perfection (where "perfection" means a state in which any statement employed in good faith by a user is immediately comprehended).

This process lends itself to an experiment designed to record the rate of this presumed convergence, and thereby introduces some measure of objectivity into the expression of at least one kind of programming-language merit. The experiment calls for a series of attempts by several successive groups of nonprogramming users to get the system to perform the functions of which it is capable, the users knowing only what these functions are and nothing more. Each such group of users would make a fixed number of attempts at communication, the number of their successes would be recorded, and then the system would be taught all those forms the group had unsuccessfully tried. Each group would presumably find the system's comprehension better than its predecessor did, and the statistics collected should eventually enable us to express, with stated confidence level, the probability with which future users may expect to find themselves immediately understood; the statistics should also indicate the payoff to be expected from each further increment of effort at teaching the system additional notation.

latitude he thinks permissible in giving orders to humans would be astounding to an army officer, doctor, or anyone with experience in directing men in the performance of complex procedures.)

The nature of machines and their response to instruction is not properly at issue, because programmers practically never address machines directly; they address programming systems, which act as buffers between them and the less amiable characteristics of computers. These systems already permit many users to disregard for almost all purposes the particular model of computer on which their programs are to run, and there is every reason to think we can develop them much further in the direction of liberating the programmer from concern with the purely mechanical, should we decide to. Any appeal to the characteristics of naked machinery, then, begs the question; the *effective* machine, the machine as it is known to almost all programmers, is very nearly whatever we want it to be—and what we should want is a philosophical, not a technical, question.

Redundancy

One of the ways in which the bare computer's demand for precision and explicitness can be mollified is through one of the very characteristics of natural language that calculists most abhor: redundancy. In many other technologies, the word "redundancy" is an honorable one, standing for a means to reliability. Electronic circuits, mechanical structures, communication channels, and hydraulic lines are made doubly and triply redundant, and users are rewarded by diminishing failure rates; why should this technique have been so unsuccessful when applied to programming? The answer, of course, is that it has not been unsuccessful, merely untried. We have long been playing a game that might be called "But Don't Tell the Computer!" the point of which apparently is to see how little information we can give the computer and still get some output. If we tell the same thing twice to one of the programming systems designed to play this game, or tell it something it does not demand to be told, it not only fails to avail itself of the possibly vital information being offered, but may well gag, stop compiling, and disgorge a core dump at us. What is loosely called redundancy in programs is taken simply as noise by today's systems, when they accept it at all; no system in common use today is capable of profiting from real redundancy. Programming systems that know how to use redundancy would reward programmers with the same resistance to minor failure other redundancy-exploiting systems exhibit, using the extra information offered them at one place to compensate for minor omissions and inconsistencies elsewhere in the program. As a trivial example for those to whom this notion is utterly foreign, consider a program unit that contains two statements implying that the value of X is A, and one that implies its value is B; the conventional system either fails to notice the contradiction and proceeds to compile a faulty program, or notices it and merely throws the problem back at the programmer. A system capable of using redundancy would use majority-decision logic to conclude that X's value should probably be A (and would, of course, notify the programmer of its assumption).

Dijkstra's attitude toward redundancy is a mixed, even a troubled, one. He is here, as elsewhere, more thoughtful than most of the calculus school in seeing some positive value in it, but differs sharply from our position because he judges it solely as a device for optimizing the object program's use of space or time, rather than as a means of improving man–machine communication. Judged by the criterion he employs here, redundancy is clearly a tricky thing, perhaps better avoided; if it can sometimes improve the object program, it also complicates the processor and consumes translation time. He accordingly concedes it some value, but warns:

> . . . if the redundant information is to be a vital part of the language, the defining machine *must* take note of it, i.e., it must detect whether the rest of the program is in accordance with it and this makes the defining machine considerably more complicated.

and later adds:

> But we can hardly speak of "good use of a computer" when the translator spends a considerable amount of time and trouble in trying to come to discoveries that the programmer could have told it as well!

His final position is that redundancy is permissible, but is to be kept optional—optional not only in that programmers need not use it, but in the deeper sense that processors need not use it if offered, and should not require it to produce good object programs.

Given his premises, Dijkstra's conclusions are unavoidable—but those premises are questionable, and Dijkstra himself, as will be shown, questions them by implication elsewhere in the same paper. The rule suggested by the second of the remarks, quoted in the foregoing, that the processor should be spared anything the programmer can do for it, is absurd if taken literally. Unless it is checked by some superior principle, it cuts the ground from under all software. But Dijkstra does not, of course, intend this. On an earlier page he rebukes those who by "good use of a machine" mean simply use that is economical of time and space, saying:

> I have a suspicion, however, that in forming their judgment they restrict themselves to these two criteria, not because they are so much more important than other possible criteria, but because they are so much easier to apply on account of their quantitative nature . . . . there is sufficient reason to call for some attention to the more imponderable aspects of the quality of a program or of a programming system.

He concludes this section of his paper by saying, in words with which the present writer is thoroughly in accord: "in the last instance, a machine serves one of its highest purposes when its activities significantly contribute to our comfort."

This is the superior principle that provides justification for the use to which we would put redundancy, as it does for the existence of software in general.

What calculus, and why?

Dijkstra aligns himself with Prof. John McCarthy in calling Cobol "a step up a blind alley on account of its orientation towards English, which is not well suited to the formal description of procedures." It should be noted

in passing that we agree with McCarthy and Dijkstra that the Cobol language is unsatisfactory, but for reasons diametrically opposed to theirs. What we see as wrong with it is precisely its lack of "orientation toward English"; the Cobol language, as a passive subset of English, is as unacceptable from our point of view as from theirs. Dijkstra says of such languages that "giving a plausible semantic interpretation to a text which one assumes to be correct and meaningful, is one thing; writing down such a text in accordance with all the syntactical rules and expressing exactly what one wishes to say, may be quite a different matter!"[11] His diagnosis of the trouble with passive systems is astute, but the cure (insofar as language can offer one) is to make the program easier to write, not harder to read.

The calculus school would undoubtedly regard active English as even less suitable for the "formal description of procedures" than the passive type. What they strikingly fail to say is what language *would* be suitable for that purpose, and what programmers would be capable of using it. It seems clear that whatever the calculus school has in mind for programmers, it is something far more rigorous and succinct than they now use, something much closer in spirit to mathematical notation—McCarthy's use and Dijkstra's citation of the ominous phrase "formal description" is probably sufficient indication. The belief of the calculists that mathematical notation constitutes a distinct language that makes substantive error harder to commit (or easier to find) is simply unfounded, however. The calculists have overlooked or forgotten the ancestry of their notation. Both historically and logically, mathematical notation is an encoding of a subset of natural language, and is not an independent language (see Box I). The elementary operations and operands of mathematics can be defined only in one of the natural languages; more elaborate operations and operands are defined in terms of the elementary ones—and, where necessary, further natural-language definitions. If English is incorrigibly imprecise, then all mathematical notation is congenitally infected with that imprecision, for English is the ground from which it sprang and to which it still returns at frequent intervals for support.[13]

Mathematical notation is, in fact, nothing more than a shorthand (see Box II) that facilitates the very compact graphic expression of natural-language statements belonging to a certain special universe of discourse.

In principle, all mathematical discourse could just as well be carried on in the ordinary vocabulary of which that notation is simply a condensed representation. In practice, the notation is indispensable because it permits the expression in graspable form of propositions that, expressed verbally, would be too tenuous, too rarefied, and too linear to be apprehended as coherent wholes.

But such propositions so expressed do not in practice form the fabric of mathematical arguments. The notation as it is actually used in mathematical papers is more a medium for the presentation of intermediate results, while the burden of exposition is invariably carried by a prose narrative typically starting "Let L_1 be a . . ." and ending ". . . which completes the proof." The equations and other symbolic expressions embedded in the prose serve to record and isolate for possible inspection the important milestones along the way; their study is seldom necessary for comprehension of the argument, and unless the reader suspects an error in formal manipulation he will not dwell

Box I

A *language* is the direct symbolic representation of an original and independent anatomization of experience into objects of interest. As such, it almost certainly will not map perfectly into any other language—as French, for example, cannot be mapped element for element into English. A *code* is a derivative representation, artificially created through the application of some transformation to a language or subset of one. This derived representation has a completely determinate relationship to the language from which it derives, and neither adds to nor subtracts from it any information. Its practical advantages may be great, but whether these lie in economy, security, or convenience, they do not spring from the code's supposed superiority as a representation of reality, but from its intrinsic properties, such as brevity, secrecy, or mnemonic power.

Box II

Many mathematicians claim for their notation a considerable heuristic power, crediting it with suggesting to them relationships and developments they might otherwise have missed. Their notation clearly has such power for practitioners steeped in its use, but it is hardly unique or even highly unusual in this. Poets, for example, have testified that they often find a line or word they have just written suggesting a successor that would not have occurred to them but for some feature of their "notation"—rhyme, alliteration, pun, etc. The point is that any notation, indeed any tool, whose user is saturated in it comes to be so much a part of his nature that it seems sometimes to work of its own accord. While there remain better and poorer notations, then, the distinction between them does not hinge on this common property of seeming to come alive in an experienced user's hand.

Not only is it useless as a differentiator between good and bad notations, but its value as a means of arriving at significant results in mathematics has been questioned at the highest level. Gauss, for example, had occasion to reprove his contemporary, Waring, for inordinate emphasis on the heuristic value of notation. Waring had said of certain theorems that they were very hard to prove because of the "absence of a notation to express prime numbers." Gauss, himself the inventor of the standard sign for the congruence relationship, replied sharply that mathematical proofs depend on *notions*, not *notations*.[14]

on them. Supporting this observation is the fact that a non-Russian-speaking mathematician will generally be unable to read a paper in a Russian mathematical journal. That mathematicians need to learn foreign languages or employ translation services just as much as any other scientist shows that little of the meaning of even the most formal mathematical discourse is carried by the internationally accepted notation of the discipline.

Insofar as the call for a programming calculus is a demand that any programming insights we manage to achieve should be, as far as possible, reflected in our programming tools, it is unexceptionable. But we must beware of assuming that mathematical notation is analogous to programming notation when it is only homologous to it. If we are to compare things of like function rather than things merely of like form, the parallel between them breaks down, since they play different roles in their respective worlds. Any analogy between programming and mathematics, then, is hazardous; to the extent that mathematical notation *has* an analog in programming, it would be the "comment" facility offered by most programming systems.

Consider, for example, just these two differences:

1. Mathematical objects are imaginary and arbitrary; programming objects are, in general, real and given. A mathematical object is created by fiat, and has exactly and only those qualities that are given it explicitly (and those logically implied by them). The objects of ultimate interest in programming, other than the special subset that is pure-mathematical, pre-exist in nature, and can neither be altered nor exhaustively described.

2. Mathematical goals are flexible and opportunistic; programming goals are fixed and predetermined. If a mathematician trying to prove an intuitively obvious theorem finds it not merely impossible to prove true, but in fact demonstrably untrue, he is far from disappointed; he has a much more important result than he was trying for. A programmer unable to reach his original goal is simply a programmer defeated.

These differences are so fundamental and far-reaching that any argument founded on the resemblance between a proof and a program should be held suspect. The argument is particularly likely to be specious if it assumes that resemblance implies an identity of goals, methods, or problems in the procedures that underlie the two.

In sum, the qualities of precision and rigor that make mathematics so deeply satisfying to some temperaments (among them the writer's), and the yearning for which underlies the calculists' position, are rooted neither in the methods nor the notation of mathematics, but in its subject matter. In general human affairs, even flawless reasoning does not guarantee correct conclusions; the imperfection of the data with which we must work taints every step like an intellectual original sin. The only objects about which perfect reasoning guarantees perfect conclusions are those whose nature is known absolutely because they have been invented, not discovered; and only those disciplines similarly prepared to limit their universe of discourse to objects created by stipulation—imaginary objects—can enjoy the same certainty of result. To attempt to secure that rigor by aping mathematical symbology, jargon, and other surface aspects is to practice homeopathic magic, with its promise that eating our enemy's heart will endow us with his courage.

These remarks, it should be needless to say, are not offered as a definitive treatment of the many problems—linguistic, psychological, metamathematical—that they touch upon. They are intended only to suggest that there are more and deeper issues involved in the notational question than are covered in the usual easy antithesis between natural sloppiness and formal precision, and that it is far from clear that a formalism patterned on mathematical notation is the answer to any burning problem in practical programming.

Consequences for debugging

Each of the two schools of thought claims that the adoption of its proposed type of language would make for a significant advance in debugging, whether by making errors harder to commit or easier to find. Gawlik, in the paper that incited Dijkstra to the writing of the letter I have been quoting, claimed for MIRFAC that programs written in it could be checked for correctness by anyone who understood the problem, even if he knew nothing of programming:

> MIRFAC has been developed to satisfy the basic criterion that its problem statements should be intelligible to nonprogrammers, with the double aim that the user should not be required to learn any language that he does not already know and that the problem statement can be checked for correctness by somebody who understands the problem but who may know nothing of programming.[12]

Dijkstra, as we have seen (in the preceding section), rejects this contention, but says in outlining his own ideas on language design:

> In particular I would require of a programming language that it should facilitate the work of the programmer as much as possible, especially in the most difficult aspects of his task, such as creating confidence in the correctness of his program.[11]

Both parties, then, beneath their conflict over notation, agree that a good language would go far in helping programmers with debugging, and both, I think, are wrong.

One can make two kinds of error in writing a program: formal and substantive. Formal errors result from infractions of rules for using the language; in another familiar nomenclature, they are known as syntactic errors. Substantive errors are those that prevent the procedure embodied in the program from solving the problem—either because it does not really say what the programmer intended, or because what the programmer intended is wrong. It is clear that the debugging advantages claimed for their language types by Gawlik and Dijkstra alike must be limited practically exclusively to the formal errors. Neither claims that a language of the type he proposes would make it impossible to mistake the problem or misstate its solution. Even the weaker claim that they would make the detection of these substantive errors significantly easier cannot be allowed; it is a common experience among programmers to desk-check their programs for hours, only to find that they have repeatedly passed as correct an error they will later call "obvious," recommitting at each iteration the mental lapse that gave rise to the error originally. The suggestion that the programmer should have a colleague check his program is unrealistic. It amounts to nothing less than a demand that each problem be programmed twice, since an independent

checker is useful only if he is, at least mentally, programming the problem in parallel as he reads the original.

But the formal or syntactic errors that both parties in this controversy promise to eliminate or minimize are by far the less important class of errors in a compiler-language program, and many processors already offer detection of all such errors in the first compilation of the program, so that even if the promised advantage should materialize, it will not be worth giving up much to get. Dijkstra's own words support our position. In a later passage from the paper previously cited, he recognizes the two kinds of error described here, and grants that only a response from the party addressed (human or mechanical) can reveal substantive error, but he then unaccountably dismisses this genus from consideration as if it were either negligible or irremediable:

> . . . we badly need in speaking the feedback, known as "conversation." (Testing a program is in a certain sense conversation with a machine, but for other purposes. We have to test our programs in order to guard ourselves against mistakes, which is something else than imperfect knowledge of the machine. If a program error shows up, one has learnt nothing new about the machine—as in real conversation—one just says to oneself, "Stupid!")[11]

The self-directed cry of "Stupid!" is a familiar one to programmers; it is traditionally uttered after long searching of dumps and listings for a clue as to why a syntactically perfect program ran wild or gave wrong answers. And it is in finding these substantive errors, and not those that any compiler will pinpoint on the first run, that the programmer desperately needs help. Since no one has shown how a wise choice of source language offers any help to speak of in dealing with substantive error, we conclude that debugging is not an important consideration in the design of a programming language, or in choosing one from among others on the same logical level.

This is not to say that nothing can be done to help programmers with their debugging problem; I have elsewhere described what a properly designed programming *system* can do in that regard.[15] For the problem of ensuring that the procedure to be executed is formally correct, without loose ends or inconsistencies, the most attractive solution so far is the rigorous tabulation of the decision rules implicit in the procedure. With such a tabulation, the detection of this important subclass of errors can be reduced to the mechanical checking of a table for completeness and consistency, and such a representation of the procedure can even be used directly as computer input.[16] (Whether it is desirable to use it so is doubtful, since the tabular arrangement throws a strong light on formal error at the price of obscuring the practical meaning of the program, which is better conveyed by a narrative form. It may well turn out that such decision-rule matrices, like Cobol-language statements, are not really wanted as a source language, but as a by-product of compilation—in short, as output rather than input.)

Specimen confusions

Much of what little literature exists on the topic of natural-language programming, whether pro or con, suffers from failure to take account of elementary distinctions of the kind insisted on here. Two examples, chosen not for their egregiousness but merely for simplicity and brevity, will suffice as illustration. Their author* makes the familiar contrast between sloppy English and a precise calculus, then gives what he takes to be supporting evidence:

> The written and spoken English of the average adult is imprecise, often redundant and incomplete. Compared to mathematics and symbolic logic it is a poor vehicle for the expression of thoughts in precise, logical form. In one experiment, for example, the subjects were shown the single logical premise: "What can you say about B if you know that all A are B?" The subjects tended to continue by concluding that all B are A. In other studies it has been demonstrated that verbal habits operate as a substitute for thought and often lead to errors in logic. Evidence of this type might lead one to reject a Near-English language as a medium for man–computer problem solving.[17]

Several kinds of error are entwined in this brief passage. The experiment, if correctly reported, was ill-designed. We are asked to accept its results as showing that symbolic-logical notation is better than English for conveying precise notions; what it does show is merely that some people do not understand the idea of set membership. Evidence that any greater number of them would have understood the relation between A and B if it has been expressed as "$A \subset B$" is not given; common sense suggests that no such evidence exists. Certainly all who recognize the expression "$A \subset B$" would know that the relation is not symmetrical—but just as certainly all of them, plus some who are not familiar with such notation, would understand this from the English version. If the experiment had been run properly, with a control group, more of those given the problem in English would have gotten it right than those given it in technical notation. More striking yet, it is far from clear how the rather complex question that Van Cott calls a "single logical premise" could be stated in technical notation. What is the symbol for "What can you say about . . .?" It would seem that not only might fewer people have given the right answer if English had been ruled out, but that the problem could not even have been put to them without its use. Insofar as this anecdote suggests anything, it would tend to show that English has some powers that even a symbolic logician might be unwilling to forego.

Van Cott offers further evidence against English:

> The rationale for the development of Near-English user languages was based on the assumption that the human could not adapt to a new, more formal language easily, rapidly or with any degree of reliability.
>
> Anecdotal evidence suggests that this assumption may be false. For example, people rapidly adapted to the use of the telegraph as a means of communication—reducing the redundancy and ambiguity characteristics of normal English in order to reduce message lengths, save money and avoid misunderstanding.[17]

* Dr. H. P. Van Cott, Associate Director, Institue for Research in Organizational Behavior, Washington, D.C.

IEEE publications

scanning the issues
advance abstracts
translated journals
special publications

Scanning the issues

Honesty, Gentlemen! Which foot is a man to put forward when he is going to lecture engineers about "Ethics as a Prerequisite to Proposal Success"? Here is how Thomas E. Altgilbers (who is a proposals administrator) does it:

"Engineers in general are admirable people, honest and upright; they are respected members of society, and rightly so. They love their wives and children, they go to church, they obey the law as well as anyone else, and they litter only in the immediate vicinity of their desks. However, in the proposal environment, they undergo a remarkable Jekyll-to-Hyde transition. As soon as the proposal cycle starts, whether the moon is full or not, the primordial beast awakes and takes over the planning function. Now proposal evaluators are a perspicacious lot, while the beast is merely pseudo-clever; the result is a clearly identified poverty syndrome.

"As with any malady, the patient (the proposal engineer) must be motivated for a cure; the motivation is provided by recognizing the cause-and-effect relationship between proposal ethics, bookings, billings, and salary. The cure lies in identifying the symptoms and treating the affected areas. These symptoms, which appear throughout the proposal, are classified into three groups: (1) misrepresentation, including overstatements, unsupported claims, and dirty lies; (2) concealment of fact through circumlocution, ambiguity, and confusing organization of material; and (3) omission of key information either requested or not requested."

All of these venal activities, Altgilbers' experience suggests, are almost universal characteristics of proposals. Furthermore, he propounds, there is a direct correlation between proposal success and proposal ethics, a fact that is independent of the proposal scope or the type of customer. For those "innocents" who do not know whereof Altgilbers speaks, and for those Jekyll's who do, there is a delightful few pages of reading ahead. His conclusion could not be more straightforward: "You write a proposal because you think you can win; you think you can win because you have more to offer than the competition. If you are wrong, it's better to find it out early; so just present the facts and let the buyer decide. Be accurate, clear, truthful, and well paid. Of course you will present yourself in the most favorable light, but do it without subterfuge. Put your best foot forward, making sure that it is in fact your own foot, and that it isn't clamped between mandible and maxilla."

If we all really followed Altgilbers' advice, our technological world and our so-called information revolution (which is partly a garbage revolution in disguise) might become so much more straightforward that we'd have a fifty–fifty chance of coping with it. Abandon easy virtue, gentlemen, and take the harder line. All those who are with Altgilbers, say "Aye!" (Thomas E. Altgilbers, "Proposal Perfidy, A Poverty Syndrome," *IEEE Trans. on Aerospace and Electronic Systems*, January 1967.)

Reading Much Lately? When the human animal is faced with an emotional or nervous overload, one of his responses may be to flee the scene; another may be to freeze up and play dead. Faced with an overload of information not only in the mass media but in his own specialty as well, while there dangles over his head the vague threat of "obsolescence," an engineer might well withdraw. At the very time he supposedly should be reading more than ever, he may in fact be reading less than ever. Is this the case? Whatever the case (and the problems implicit in today's technological information overload are by no means simple or superficial), it should be of some interest to see what is turned up in a survey of the reading habits of engineers. One such survey appears in the current issue of the IEEE TRANSACTIONS ON EDUCATION.

Although the inquiry made by J. M. Lufkin and E. H. Miller is narrow in one respect—in that they have tapped less than 2000 engineers, and these in the aerospace and avionics industry—and although the results they seek are qualitative (not usually regarded as a virtue in engineering circles), their analysis is thoughtful and lucid, and for this reason alone carries more weight than many more statistical accounts.

There is no point here in going into the form and specific questions of the Lufkin/Miller survey. What is most interesting is their interpretation. They say: The most interesting single finding of this survey is that the people who have been singled out for excellence, whether by promotion, or by publication, or by special recognition for creativity, all read a great deal more than the average. It is also interesting, and somewhat disturbing, to find that about half of the engineers in this industry, however much or little time they may spend on technical reading, are making little use of the periodicals that offer the most for continuing education.

Considering the fact that there *is* some kind of correlation between reading habits and achievement, the authors then ask if this fact can be used to encourage the others (the nonreaders) to read more. Why, in fact, if some read so much, do others read so little? Are the pressures of industrial working conditions and necessary emphasis on *results* so pervasive as to discourage reading? In answering this question, the authors say, we must remember that our supervisors, who are known to be working under more than average pressure, apparently read a good deal more than the average in spite of this.

For those who "have no time" to dig out the other interesting material from this relatively short, "relatively casual" survey, the motto is: read more! (J. M.

It is not clear who the "people" are who adapted to the telegraph. Are they the professional telegraphers who pounded the key all day long, or the customers who passed their hand-printed messages over the counter to be transmitted? It is just possible that Van Cott means the former, since it does not appear that the public had to do any adapting at all (unless to improve their penmanship), but it hardly seems warranted to make any broad statement about "people" if it is only the comparative handful of highly practiced operators who are the sample group. If he is referring to the telegraphers, we have here another example of the confusion of *language* and *code* that we pointed out in our discussion of mathematical notation. All a telegrapher does is to encode a natural-language message, not translate it into another language; the language remains English, but in a different physical representation, and no conclusions about human adaptability to a "new, more formal language" are warranted.

However, it seems far more likely that Van Cott is referring not to the telegraphers, but to the users and their resort to "telegraphese," meaning "language characterized by terseness and elliptical expressions such as are common in telegrams." If so, one set of objections is replaced by another, yet more devastating. Before drawing Van Cott's or any conclusion from this observation, the assertion that "people rapidly adapted" to this new language must be examined. Of course, customers paying by the word often tried to minimize the number of words in their messages, even at the price of taking more care in their composition. The disproof of their adaptation, however, is given by their immediate reversion to customary habits of speech and writing as soon as this pressure was removed. If this "new language" had any merits other than that of saving money under the rate structure fixed by the telegraph companies, the general public evidently failed to see them. But let us waive this objection; the most curious thing about this second count in Van Cott's indictment of English is that, like his first, it demonstrates the value of English if it demonstrates anything. If telegraphese is an example of "a new, more formal language," then we have some unlooked-for evidence that a subset of English makes a good vehicle for the economical conveyance of clear ideas.

Conclusion

There has been no attempt in this article to make a complete case for natural-language programming, but only to clear away the greatest of the misconceptions and confusions that have long impeded useful discussion of the subject. The writer's original intention of disposing of these incidentally in the course of illustrating the positive advantages of natural language had to be dropped as it became apparent how many and deep-seated these misconceptions were, and how much analysis was needed to show them as such. The more positive side of the argument has had to be deferred, but should be published at an early date.

Chief among the points we sought to make here are:

1. Natural-language programming is an attempt to put nonprogrammers in direct touch with the computer, not to spare the advanced professional programmer what may be to him an insignificant part of his total job.

2. A natural programming language is one that can be written freely, not just read freely.

3. The task to be performed by the processor for such a language is qualitatively different from that of translating one natural language to another.

4. The redundancy of natural language is one of its greatest potential advantages, not a prohibitive drawback.

5. Finally, the possibility of user-guided natural-language programming offers a promise of bridging the man–machine communication gap that is today's greatest obstacle to wider enjoyment of the services of the computer.

Valuable criticism of early drafts of this paper came from Daniel L. Drew and Allen Reiter of Lockheed Missiles & Space Company, Edward Theil of the Department of Mathematics, University of California at Davis, and Christopher Shaw of System Development Corporation. Although these friendly critics have much improved the presentation of our argument, it must not be assumed that any of them subscribe to it.

This article is a revised version of a paper presented at the 1966 Fall Joint Computer Conference, San Francisco, Calif., Nov. 7–10.

REFERENCES

1. Zemanek, H., "Semiotics and programming languages," *Commun. Assoc. Comput. Mach.*, vol. 9, pp. 139–143, Mar. 1966.

2. Sammet, J. E., "The use of English as a programming language," *Commun. Assoc. Comput. Mach.*, vol. 9, pp. 228–230, Mar. 1966.

3. Halpern, M. I., "A manual of the XPOP programming system," Lockheed Missiles and Space Co., Palo Alto Research Lab., Oct. 1965.

4. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Commun. Assoc. Comput. Mach.*, vol. 9, pp. 36–45, Jan. 1966.

5. Yershóv, A. P., "One view of man–machine interaction," *J. Assoc. Comput. Mach.*, vol. 12, pp. 315–325, July 1965.

6. Bhimani, B. V., *et al.*, "An approach to speech synthesis and recognition on a digital computer," *Proc. of the 21st National Conference of the Association for Computing Machinery*. Washington, D.C.: Thompson Book Co., 1966, pp. 275–296.

7. Lindgren, Nilo, "Machine recognition of human language—Part I," *IEEE Spectrum*, vol. 2, pp. 114–136, Mar. 1965.

8. Bar-Hillel, Y., *Language and Information*. Reading, Mass.: Addison-Wesley, 1964, pp. 153–184.

9. MacKay, D. M., "Linguistic and non-longuistic understanding of linguistic tokens," Memorandum RM-3892-PR, Rand Corp., Mar. 1964.

10. Dijkstra, E. W., "Some comments on the aims of MIRFAC," *Commun. Assoc. Comput. Mach.*, vol. 7, p. 190, Mar. 1964.

11. Dijkstra, E. "On the design of machine independent programming languages," in *Annual Review in Automatic Programming*, vol. III, ed. by R. Goodman. New York: Pergamon Press, 1963, pp. 27–42; p. 31; p. 30; p. 33.

12. Gawlik, H. J., "MIRFAC: A compiler based on standard mathematical notation and plain English," *Commun. Assoc. Comput. Mach.*, vol. 6, pp. 545–548, Sept. 1963.

13. Cajori, Florian, *A History of Mathematical Notations*. Chicago: Open Court, 1928.

14. Reid, Constance, *From Zero to Infinity*. New York: Crowell Apollo Editions, 3d ed., 1966, p. 129.

15. Halpern, M. I., "Computer programming: the debugging epoch opens," *Comput. Automation*, pp. 28–31, Nov. 1965.

16. Boerdam, Wim, *et al.*, "DETAB/65 Language," Working Group 2, SIGPLAN, Los Angeles Chapter, Assoc. Comput. Mach., June 1964.

17. Van Cott, H. P., "Flexible machine language for commander-computer chats may be key to flexible C&C," *Armed Forces Management*, vol. 11, p. 95, July 1965.