VOLUME 4 □ NUMBER 2 FEBRUARY 1967

IEEE spectrum

features

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

the cover

Improved techniques for the generation of random numbers are constantly being devized, as described in an article beginning on page 48. These numbers find application in many deterministic problems as well as in problems involving probability concepts.

VOLUME 4 □ NUMBER 2 IEEE spectrum FEBRUARY 1967

Spectral lines

IEEE Electrolatina. IEEE Electrolatina is the name of the Spanish-language technical periodical that will be started soon in the IEEE Latin American Region. The Institute's Board of Directors authorized the new publication at its November 1966 meeting when it unanimously approved the Publication Board's recommendation to establish a periodical "of, by and for the Latin-American Region of the IEEE."

It had been apparent for some time that the technical publication needs of IEEE members in Latin America were not being met as effectively as they might be. In Latin America, which includes Mexico and the countries of South and Central America and the Caribbean area, the predominant language is Spanish, but for Brazil where Portuguese is spoken. Consequently, the IEEE publications, which are published in English except for an occasional article in another language, are accessible only to those who have a reading knowledge of technical English. Although there are many IEEE members in Latin America who have such knowledge, there are also many who do not.

Another characteristic of this geographic area is that there is not a great deal of interest in some of the highly specialized technical fields covered by IEEE periodicals. There is a strong interest in power generation and distribution and certain aspects of communications, but there is less concern about some of the highly theoretical and scientific topics with which various IEEE periodicals are primarily concerned.

To meet these special needs, the Directors approved the plan for a new periodical to be published on a quarterly basis starting in 1967. The proposal for the publication originated among the members of the Mexico City Section. At first, it was conceived of as a somewhat more localized activity, but gradually it became apparent that in order for a periodical of this type to succeed from both an intellectual and economic standpoint, it would be most desirable for it to cover all of Latin America. Early in 1966, the Board of Directors had designated the Latin American countries as a new Region and so the idea developed of having the publication based on the Region as the geographical unit and on the Regional Committee as the organizational framework.

The plan is that the new publication's Editor, Associate Editors, and Publications Manager will be appointed by and report to the Region 9 Committee, which is chaired by G. J. Andrews of Buenos Aires, Director of Region 9. At the time of this writing, the editorial board is to consist of Erik Wallsten of Mexico as Editor and the following Associate Editors:

Argentina:	Alberto Rincón
Brazil:	R. Costa de Lacerda
Chile:	Reinar Puvogel
Colombia:	Alberto Ospina
Mexico:	Bruno de Vecchi Alfredo Romero F. Daniel Barrios Morales
Peru:	Cesar A. Pera
United States:	A. B. Bereskin
Venezuela:	Hernán Pérez Belisario

To fill the position of Publications Manager, the Region is fortunate in having the services of Ignacio Avilez of Mexico, who is responsible for the publication of several other technical journals.

One of the functions of the Associate Editors will be to collect papers from their areas for publication. It is hoped that potential authors from the United States and Canada and from European countries will direct their papers to Prof. A. B. Bereskin, Department of Electrical Engineering, University of Cincinnati, Cincinnati, Ohio.

It should be noted that this publication will become the first technical publication of the IEEE to be published and circulated on a Regional basis. The IEEE is not a national (or indeed an international) organization, in that it is not organized on a national basis and its technical and scientific interests do not have a national character. In general, its technical activities, including publications, meetings, and recognition of outstanding contributions by bestowing awards, are not related to a particular country or even group of countries. Rather, the IEEE seeks to pursue its objectives, which were so definitively described by Dr. W. G. Shepherd in his article "The uses of a professional society" (IEEE spectrum, Dec. 1966, pp. 35–38), independent of national considerations.

Nevertheless, its non-national character should not preclude the Institute's being sensitive to, and responsive to, the special needs that may exist in particular geographical areas. If these special needs can be met without harm to other activities of the Institute, it is clear that they should be. This certainly appears to be the case for IEEE Electrolatina.

As in any new publications venture, there is concern about whether the timing is right, the contents are well-selected, the composition is appropriate. However, I feel that this publication project is well-conceived and well-planned. We are anticipating that it will be a great success.

F. Karl Willenbrock

Authors

Random-number generation on digital computers (page 48)

R. P. Chambers (M) was born in Toronto, Ontario, Canada, on March 24, 1930. He received the bachelor of science degree in electrical engineering from Newark College of Engineering, Newark, N.J., in 1963. Three years later, he was awarded the master of science degree from the same school.

From 1950 to 1954 he served as a radar mechanic in the United States Air Force. Since 1959 Mr. Chambers has been affiliated with Bell Telephone Laboratories, Inc., Murray Hill, N.J. He has been engaged in work on optical tracking of satellites and on ultrasonic velocity and attenuation measurements in solids and in plastics. He is currently concerned with conducting a computer simulation study of the problem of detecting faint visual signals in noise.

Major recommendations of the U.S. President's Patent Commission (page 57)

B. M. Oliver (F), recently appointed Secretary of IEEE for 1967, received the A.B. degree from Stanford University in 1935 and, a year later, the M.S. degree from the California Institute of Technology, Pasadena. In 1940 he received the Ph.D. degree, magna cum laude, from the latter school. He joined the Bell Telephone Laboratories, Inc., Murray Hill, N.J., in that year, working on the development of automatic tracking radar, television, information theory, and efficient coding systems. He became affiliated with the Hewlett-Packard Company in 1957, serving as director of research and, subsequently, as vice president of research and development.

Dr. Oliver has been awarded over 40 U.S. patents in the field of electronics. He was elected Director-at-Large of the IRE in 1958 and has served as a member of the Board of Directors of WESCON, as IEEE Vice President (1963–64), and as President (1965).

Wave-mechanical uncertainty and speed limitations (page 65)

Panos A. Ligomenides (M) received the diploma in physics with high honors in 1951 and the M.S. degree in technical radio engineering in 1952, both from the University of Athens, Greece. He served in the Radar School of the Greek Royal Navy. Subsequently, he attended Stanford University, where he was awarded the M.S.E.E. degree in 1956 and the Ph.D. degree in electrical engineering and physics in 1958.

He then joined the General Products Division Laboratory, IBM Corporation, Poughkeepsie, N.Y., where he worked on semiconductor and magnetic devices. Subsequently, he worked with the Advanced Technology Group, IBM, San Jose, Calif. In 1964 he joined the engineering faculty of the University of California at Los Angeles. In addition, he presently serves as a consultant to IBM.

Dr. Ligomenides is the author of several papers and is soon to publish a book, *Physical Theory for Solid State Devices.* He is a member of Sigma Xi and the American Physical Society.

EHV transmission in the U.S.S.R. power grid (page 73)

B. P. Lebedev, biography and photo not available at time of publication.
S. S. Rokotian, biography and photo not available at time of publication.

Some aspects of binaural sound (page 80)

Charles J. Hirsch (F) received the B.A. and E.E. degrees from Columbia University in 1923 and 1925, respectively. By 1941 he had served as chief engineer with radio companies in France, Italy, and the United States. In that year he joined Hazeltine Corporation, where he was concerned with the development of color television and of all phases of secondary radar, such as IFF, beacons, and DME. In 1956 he became executive vice president of Hazeltine and, since 1959, he has been with RCA as administrative engineer on the corporate staff of the vice president for research and engineering. He is concerned with the coordination of research and engineering activities of the Home Instrument, Record, and Custom Aviation Divisions with the rest of the corporation.

The author of several articles on air navigation, IFF, DME, analog computers, color television, and stereophonic sound, Mr. Hirsch has been awarded 25 patents in these fields. He serves as honorary secretary of the IEE for the United States.

A new keyboard without keys (page 86)

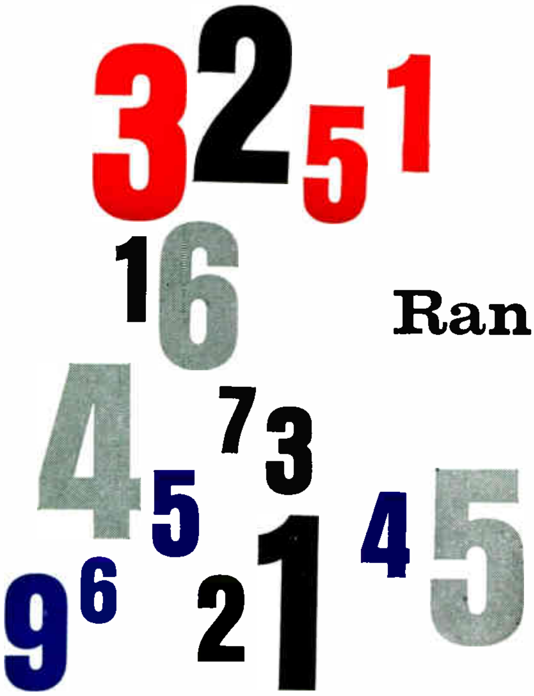
Paul Rosberger, founder and president of Binary Keyboard, Maplewood, N.J., studied composition at Juilliard School of Music between 1958 and 1960. For the next three years he served as store manager of the Gold Seal Bedding Company, a family business. During this time he maintained his interest in music theory and history of music, which, in 1963, led to the establishment of a holding company for the building of prototypes for and development of the binary keyboard. Two years later he assumed responsibility as manager of Quality Control for the Magnus Organ Corporation. However, the holding company was soon reformed for the development of the Binary Touchboard. In addition, the Quarternary Touchboard is presently in the advanced stages of development.

Mr. Rosberger has made tapes of improvisations on the Touchboard, utilizing new types of musical sound effects that can improve performance in areas of electronic music. He has published several papers and recently was invited to speak before the American Society of Music Arrangers.

Dilemmas of engineering education (page 89)

Harvey Brooks, presently dean of Engineering and Applied Physics at Harvard University, received degrees from Yale University, Cambridge University, and Harvard University. In 1942 he joined the Harvard Underwater Sound Laboratory as a special research associate. From 1946 to 1950 he served as a research associate with the General Electric Research Laboratory and as associate laboratory head, Knolls Atomic Power Laboratory. He was a Guggenheim Fellow and a McKay Professor of Applied Physics at Harvard prior to assuming his present position.

The recipient of honorary degrees from both Yale and Harvard, among other schools, Dr. Brooks is a member of the Board of Directors of the American Nuclear Society and of the Board of Syndics of the Harvard University Press. He has served on the Advisory Committees of the Guggenheim and Kettering Foundations, on the Ad Hoc Committee on Reactor Programs and Policies of the Atomic Energy Commission (1958–60), and on the Board of Editors of *Physical Review* (1950–54).



Random-number generation

Because of the usefulness of random numbers in the solution of many different kinds of problems, several techniques have been developed for generating these numbers in a reproducible fashion by means of computers

R. P. Chambers Bell Telephone Laboratories, Inc.

Random sampling methods are valuable not only for providing solutions to problems involving probability but also for solving many problems that are deterministic in nature. Haphazard generation of numbers has several serious disadvantages, since the numbers used in the computation cannot be reproduced and thus rational "debugging" procedures cannot be developed. Over the past 20 years there has been a strong emphasis on arithmetic generators, which are based on recurrence relations involving integers.

Random numbers are useful in many types of computation, such as in problems involving probability and statistics. As a simple example, suppose we wanted to know the probability of drawing a particular type of poker hand. We could, of course, compute this figure analytically. Alternatively, we could take an experimental approach; we could draw a large number of hands, shuffling the cards after each, and count how often our particular poker hand turns up. We could simulate the experiment without cards by using a table of random numbers, with the numbers in the table assigned to the cards in some way so that all cards are equally represented. The probability of the particular poker hand could then be calculated from the results of a large number of draws from the table. For a simple problem such a method has no advantage, but for very involved problems and for those beyond the reach of theoretical methods it can often be quite useful, particularly in connection with high-speed computers. It has been widely used in traffic congestion and telephone exchange studies, war games and other games of strategy, operations research, nuclear reactor design, statistical mechanics, population growth studies, and so on.

A less obvious application of such a random sampling method is in the solution of deterministic problems—that is, problems that do not directly involve probability. Let us, for example, consider the problem of finding the area of an irregular plane figure. Around the figure we could draw a square subdivided by, say, a 10-by-10 grid, and we could estimate the area by counting the small squares. Or we could scatter random points in the large square using coordinates drawn from a table of random numbers and then count the proportion that fell within the figure. This approach can be extended to higher dimensions where its efficiency becomes more competitive. As pointed out by Metropolis and Ulam,[1] the job of examining 10^{20} cells in a 20-dimensional problem is completely unmanageable, whereas a few thousand random samples may give an adequate estimate. Such an estimate will not be certain; it will be subject to random fluctuation and can, in fact, be wildly wrong. But the probable error can be made suitably small if we take enough samples. However, it is characteristic of the random sampling method that to increase the accuracy by a factor of 10 requires 100 times as many samples, and thus an ordinary numerical method will often be more efficient.

If random sampling is to be used to solve a deterministic problem, it is necessary to find a probabilistic analog that has a similar mathematical formulation. For example, in probability theory there is a "drunken walk problem," in which each step forward is as likely as a step backward; its solution is also an approximate

on digital computers

solution to a certain type of partial differential equation. Such analogs have been found for other problems, including the problem of inverting a matrix.[2]

It has become customary to refer to calculations involving random sampling as Monte Carlo methods. Such methods can be traced as far back as 1773 when the French naturalist Buffon estimated the value of π from the number of times a randomly tossed needle intersected parallel lines on a flat surface. Hamming[3] derives the formula involved. McCracken[4] describes his results with this experiment using a line spacing of twice the length of the needle, in which case the probable ratio of total trials to intersections is equal to π. After 1000 trials and a tired arm he had 333 intersections, giving an estimate of 3 for π. The real impetus for the Monte Carlo method came during World War II when Ulam and Von Neumann used it to solve neutron diffusion problems. These problems were too involved for analytical solution and the alternative would have been expensive trial-and-error experimentation.

Methods for generating random numbers

Random numbers can be derived from physical processes such as radioactive decay and thermionic noise. In 1939 Kendall and Babington-Smith published 100 000 random digits read from a spinning disk illuminated by a flash lamp. In 1955 the Rand Corporation[5] published a million digits produced by monitoring a random-frequency pulse source. However, the modern high-speed computer will often consume large quantities of numbers at such rapid rates that reading from storage or tape becomes inadequate. The monitoring of physical devices attached directly to the machine, in addition to being slow, has the serious disadvantage that the numbers generated are not reproducible; therefore, calculations cannot be identically repeated as required in debugging. Consequently, there has been a strong emphasis on arithmetic generators since Von Neumann and Metropolis proposed their mid-square method around 1946.

The arithmetic methods are generally based on some sort of recurrence relation involving integers. Each new number is generated from the previous one, as needed, by what might be considered a scrambling operation, so that the output is "randomly" drawn from the finite population of integers that the machine can produce. An initial value is required to start the recurrence relation. At some point a number that has already occurred will be produced, thus forming a closed-loop sequence, which continuously cycles from that point on. The length of this loop sequence is called the period of the generator and, hopefully, is equal (or nearly equal) to the total integer population of the machine. The period can be greater if a recurrence relation involving more than one previous number is used. The problem is to find a relation that produces a sufficiently random sequence of numbers with a long period and with a minimum of computer time. Computer-generated numbers that manage to pass statistical tests for randomness, even though produced by a completely deterministic process, are customarily called pseudorandom numbers.

mid-square method

In the mid-square method each new number is produced by taking the middle n digits of the square of the previous n-digit number. To illustrate, let $x_0 = 2189$ be the starter.

$$x_0{}^2 = 04\,791\,721 \qquad x_1 = 7917$$

$$x_1{}^2 = 62\,678\,889 \qquad x_2 = 6788$$

$$x_2{}^2 = 46\,076\,944 \qquad x_3 = 0769$$

and so on. But suppose the number $x_n = 3500$ were to occur. Then,

$$x_n{}^2 = 12\,250\,000 \qquad x_{n+1} = 2500$$

$$x_{n+1}{}^2 = 06\,250\,000 \qquad x_{n+2} = 2500$$

Such endless repetition of some number, including zero, is a definite possibility. In any event, repetition of some previous number will eventually occur and the sequence will loop, usually with a period much shorter than the largest machine integer. The method is difficult to analyze and the period can be determined only empirically.

congruential methods

The mid-square method, once widely used, has been largely superseded by congruential methods based on the relation (involving nonnegative integers $< m$)

$$x_{i+1} \equiv ax_i + c \pmod{m} \qquad (0 \leq x_i < m)$$

which means that the expression $ax_i + c$ is to be divided by m, and x_{i+1} set equal to the remainder. The relation reads "x_{i+1} is congruent to $ax_i + c$ modulo m." To illustrate, let m (modulus) = 25, a (the multiplier) = 7, and c = 1, and let x_0 = 1 be the starting value.

$x_1 \equiv 7 \cdot 1 + 1 \pmod{25}$ $x_1 = 22$

$$x_1 \equiv 7 \cdot 3 + 1 \pmod{25} \qquad x_1 = 22$$

$$x_2 \equiv 7 \cdot 22 + 1 \pmod{25} \qquad x_2 = 5$$

$$x_3 \equiv 7 \cdot 5 + 1 \pmod{25} \qquad x_3 = 11$$

multiplicative congruential method

mixed congruential method

and so on. The method was first proposed by Lehmer[6] in 1949 with $c = 0$, in which case it is called the multiplicative congruential method. The mixed congruential method, with $c \neq 0$, appeared in Rotenberg[7] and Coveyou.[8] The congruential methods are fast and can be theoretically analyzed to determine the period. In addition, it is possible to calculate the ith member of the sequence, or conversely, the value of the index i (that is, the position in the sequence) for any specific random number, without having to calculate the intermediate numbers.[9]

Since only the remainder is retained on division by m, the period cannot be greater than m; therefore, m is usually chosen to be one more than the largest machine integer. It is then no longer necessary to divide by m, since the machine cannot register any product larger than $m - 1$ and thus automatically provides the remainder. Also, with this choice of m, normalizing to the unit interval can be accomplished by merely shifting the binary or decimal point, so that a second division by m can be avoided. This is an advantage since it is usually convenient to have random number subroutines provide output values between zero and one, leaving any scaling to the main problem. The constants a and c are chosen to provide speed, a long period, and good statistical behavior. With a suitable choice, surprisingly random, uniformly distributed sequences are obtained. A uniform distribution is a convenient result, since in principle it is possible to derive any other distribution from it.[3, 10]

With a suitable choice of constants in the congruence relations, the full period m can be achieved with the mixed method ($c \neq 0$) and very nearly the full period with the multiplicative method ($c = 0$). Both methods can be speeded up on some machines by choosing the multiplier a so that shift and add instructions can be used in place of the multiplication; within the choice of constants for maximum period it turns out that the mixed method is shorter by one operation.[10] This advantage is lost on some of the newer machines, which have fast multiplication times; for instance, on the IBM 7094 the multiplicative method gives the fastest generators.[11] The mixed generators may be started with any number, x_0; but the multiplicative method requires a nonzero starter and if the maximum period is to be achieved, the starter must be odd.[3] The problem of choosing constants is discussed in Appendix A.

The multiplicative method has generally done well statistically. The mixed generators do not always do as well; in fact some, including most of those with the fastest multipliers, are completely unacceptable.[11] Coveyou[8] mentions trouble in a Monte Carlo application resulting from use of a mixed generator with an unsuitable multiplier. Peach[12] found patterns and periodicities in the sequences from mixed generators. These patterns, he says, constrain the variability of the numbers generated; thus, although good results might be achieved in predicting average behavior in simulation problems, indications of possible extremes of behavior might be missed. Peach colorfully asserts that these patterns and harmonics constitute music rather than noise.

Greenberger[13] reports trouble encountered by Joseph Lach at Yale University with a multiplicative generator supplying coordinate points for random-number plots. Whenever a number smaller than 0.1 was produced by the generator, the next pair of numbers was used as (x, y) coordinates. The result was a plot of scattered points much like television snow, as expected, but with narrow parallel bands devoid of points. A similar problem is discussed in Appendix B.

MacLaren and Marsaglia[14] have used mixed generators which, although passed by standard statistical tests, gave poor results in Monte Carlo calculations, especially when the problems were sensitive to order statistics. They feel that the commonly used tests have little relevance to actual applications and are therefore of limited value. They conducted a series of more stringent tests on multiplicative and mixed generators and on two improved generators which they developed. One of their improved generators uses a stored table of random numbers derived from the Rand Corporation[5] table of random digits. As each number is used, it is replaced by a scrambled version of itself provided by the relation

$$x_i^1 \equiv ax_i + c \pmod{m}$$

so that when the original table is exhausted a new table is available. This avoids the problem of a finite table. Their other method uses two congruential generators, with one shuffling the sequence from the other. Their tests concentrated on n-tuples $(x_1, \ldots, x_n)$. For instance, successive pairs of random numbers were used as coordinates in the unit square and the distribution of points was checked against expected values. In a similar way, triples were used in a unit cube. They also tested for the maximum, the minimum, and the sum of n numbers. They conclude that the mixed methods are not satisfactory and the multiplicative methods are suspect, but that their alternative procedures are satisfactory.

Jansson[15] ran extensive autocorrelation calculations on congruential generators with $m = 2^{35}$. Of those he considered, he found the best to be the mixed generator

$$x_{i+1} \equiv (2^{18} + 1)x_i + 3 \pmod{2^{35}}$$

while the best multiplicative congruential generator was

$$x_{i+1} \equiv (2^{23} + 3)x_i \pmod{2^{35}}$$

He notes that the autocorrelation test is not a sufficient test for randomness, since it is easy to find generators that pass the test even though their randomness is very poor.

Sobol[16] and Shreider[17] mention a perturbation method in which two generators are combined. The first generator delivers m numbers. The second generator then operates on the last delivered number and the result is used

perturbation method

to set the first generator going again. Sobol concludes that such perturbation every m numbers causes an expected increase in the period of the first generator by a factor of $\sqrt{m}$.

In the interest of speed, congruential methods have been tried using addition instead of multiplication, according to the relation

$$x_{i+1} \equiv x_i + x_{i-n} \pmod{m}$$

The result with $n = 1$, $x_0 = 0$, and $x_1 = 1$, known as a Fibonacci sequence, turns out to actually be a multiplicative method with too small a multiplier and it does poorly statistically.[18] A larger n gives better properties[19] but requires indexing, so any speed advantage is reduced.

The decimal digits of π and e have been considered as sources of random numbers. However, generating several digits to produce one random number not only eats up a lot of computer time but also gives results that are rather poor statistically.

One might well object at this point that arithmetic generators, being completely deterministic, cannot produce truly random sequences, and that in view of the many patterns and periodicities reported in the literature, it might be better to fall back on analog methods based on random physical processes. As already mentioned, such methods are relatively slow and are not reproducible. In addition it is no easy matter to keep the associated electronics and other equipment tuned up so that unbiased output is maintained. Hampton[20] describes a hybrid analog–digital noise generator in which analog generation was completely abandoned in favor of a pseudorandom method employing a digital shift register. In a practical sense, "true" randomness is not an essential requirement, as long as the statistical properties of the generator are sufficiently suitable that the right answer to the problem is obtained.

Testing for randomness

Consider an ideal random generator continuously cranking out decimal digits, each having an equal probability of occurrence. This probability is, of course, equal to one tenth and remains constant at all times, so we can expect all digits to occur with equal frequency and with no selective tendency for one digit to follow another. From a gambler's point of view, the next digit can never be predicted. Note that the probability of getting a preselected digit twice in a row is equal to 1/100; and for n in a row the probability is $1/10^n$. Similarly, there is a finite probability of getting any such "nonrandom" pattern we care to imagine. Suppose we wanted to record one million digits from such a generator for use in a table of random numbers. We would not want to pick a string of all zeros even though there is a finite probability of such a string appearing. Such an event is extremely unlikely and we would not consider it typical of the generator's output. The heart of the matter is that since we are picking a finite subset of the generator's infinite output we can't have everything. We choose to reject unlikely sequences in favor of typical ones. But where do we draw the line? For instance, do we allow the same digit to crop up three times in a row but not more? Or four times? Much depends on the problem at hand. For integrating a plane area it is important to get points evenly scattered over the whole plane with a minimum of clustering, whereas for some other problem such a homogeneous population may yield results that are unrealistic. Ideally, the output of the random-number generator would have statistical properties tailored for each new problem. To achieve this situation would be inconvenient, to say the least, and thus the usual goal is a generator with properties adequate for most problems. It is always necessary, however, to be on guard against wrong answers in cases where the generator properties are incompatible with the particular problem at hand.

In choosing an arithmetic generator for a computer we are faced with similar considerations. In this case the digits form integers up to the capacity of the machine and are usually transformed to floating-point numbers on the unit interval. But, again, we are choosing a sequence with a finite length, a length equal to the period of the generator—and again we try to compromise with statistical properties suitable for most problems. In 1938 Kendall and Babington-Smith[21,22] proposed four widely used tests: the frequency, serial, poker, and gap tests. The frequency test counts how often each digit occurs in the sequence. The serial test tallies the frequency of occurrence of all possible combinations of two digits such as 23, 84, 00, etc. The poker test counts how often various poker hands occur. The gap test counts the number of digits that appear between repetitions of a particular digit. The results can then be compared with theoretically expected values by some statistical test such as the chi-square test. See reference 10 (p. 239) or 23 (p. 30) for a discussion of the chi-square test in connection with the frequency and serial tests.

A bewildering variety of other tests has been proposed. One could perform an unlimited number of tests and never get around to using the generator. Von Neumann felt that exhaustive testing is impractical.[24] It is probably better to pick a generator that passes the more straightforward tests, such as the frequency and serial tests, and any tests that are relevant to the particular problem. An important test is a sample computation on the problem at hand for a case where the answer is known.

Conclusions

For many random-number applications, the multiplicative congruential (power residue) method is a good choice with a minimum of hidden pitfalls. On some machines the mixed congruential method may be attractively faster, but its parameters must be chosen with care to assure equivalent statistical behavior. It may often be worthwhile to combine generators for better randomness, as for example in the shuffling scheme used by MacLaren and Marsaglia.[14]

It seems paradoxical that the most convenient random-number generators should turn out to be those that are produced by a completely deterministic process and furthermore by a process that can be theoretically analyzed to determine the period. However, consider the aggravation in trying to conduct a rational debugging procedure when the numbers used in the computation cannot be reproduced—and the feeling of uncertainty involved in using a generator for which the period is not known. If the requirements were for several hundred thousand numbers, it would be a difficult task to check the period empirically. Here lies the real objection to any scheme, however appealing intuitively, to generate random numbers by haphazard methods involving chains of basic machine instructions such as shift, add, rotate, and so on.

Perhaps the thing to emphasize is that, quoting Hull and Dobell,[11] "no finite class of tests can guarantee the general suitability of a finite sequence of numbers. Given a set of tests, there will always exist a sequence of numbers which passes these tests but which is completely unacceptable for some particular application."

One should always be wary of wrong answers due to peculiarities in the generator. It may have patterns that remain undetected in spite of extensive testing. Its output may not be uniform enough for the problem; but, on the other hand, it may be too flat, so that extremes of behavior in a simulated process are not revealed.

As with statistical methods in general, one can seldom be certain that the answer is correct. However, some steps can be taken to insure that the answer is not wrong because of the generator. The problem can be run for a case where the answer is known. As a final precaution, the problem can be rerun using a different generator.

Appendix A

Programming the congruential methods[10,18]

The recurrence relation involved is

$$x_{i+1} \equiv ax_i + c \pmod{m}$$

In the multiplicative method, $c = 0$. With $m = 2^b$(for $b > 2$), as is convenient with binary machines, the maximum period is $m/4$ and is achieved with

x_0 odd

$a = 8t \pm 3 \qquad t = 1, 2, 3, \ldots$

Barnett[25] derives the period length for other choices of x_0 and a.

With $m = 10^d$(for $d > 3$), as is convenient with decimal machines, the maximum period is $m/20$ and is achieved with

x_0 not divisible by 2 or 5

$a = 200t \pm r$

$t = 1, 2, 3, \ldots$

$r = 3, 11, 13, 19, 21, 27, 29, 37, 53, 59, 61, 67, 69, 77, 83,$ or 91

In the mixed method $c \neq 0$. With $m = 2^b$ the maximum period is m and is achieved with

x_0 any value

c odd

$a = 4t + 1 \qquad t = 1, 2, 3, \ldots$

With $m = 10^d$ the maximum period is m and is achieved with

x_0 any value

c not divisible by 2 or 5

$a = 20t + 1 \qquad t = 1, 2, 3, \ldots$

It is desirable to pick constants that, in addition to giving a long period, provide good statistical behavior. For the multiplier a, a value near $\sqrt{m}$ is often recommended.[10,18] Coveyou[8] and Greenberger[26,27] derive theoretical estimates of serial correlation for the case of mixed generators with full period m. These estimates are for the entire sequence, so serial correlation in shorter sequences must still be evaluated by empirical testing. The formulas show that multipliers near $\sqrt{m}$ give small serial correlation. To show that this is not a conclusive check, Greenberger[26] cites the combination of $m = 2^{35}$, $a = 2^{18} + 1$, $c = 1$, and $x_0 = 0$, which gives an initial run of several hundred numbers, all in the first half of the unit interval. Jansson[15] extends the work of Coveyou and Greenberger and includes formulas for the multiplicative method.

Allard, Dobell, and Hull[11,28] ran extensive frequency and serial tests, which indicate that x_0 and c are much less important than the multiplier in establishing statistical behavior. They find that very small and very large multipliers can cause poor results. If these are avoided they conclude that the multiplicative method should be consistently reliable. For the mixed method, however, there still remain many unsatisfactory multipliers, including most of the "fast" multipliers—those that can be accomplished with shift-and-add instructions. But they do get particularly good results with the multiplier $10^2 + 1$, which is very fast for the mixed method on decimal machines. However, this multiplier fails some of the tests performed by MacLaren and Marsaglia,[14] including tests involving the maximum and minimum of n numbers and a test that uses successive numbers as coordinate points in the unit square. Greenberger[26] cautions that "no final selection of parameters can be approved until a careful empirical check has been made of the generated sequence." Hull and Dobell[11] note that with $m = 2^{35}$ and $a = 2^7 + 1$, any number in the interval $[0, 2^{-8}]$ is followed by a number that is excluded from half of the unit interval, the excluded regions depending on c. Thus, this generator would not be suitable for any problem in which the numbers following values in $[0, 2^{-8}]$ are required to be uniformly distributed.

The chief virtue of the mixed method is that in situations where speed can be improved by using shift-and-add instructions to accomplish multiplication, it is usually faster, because the shift-and-add multipliers for maximum period are simpler for the mixed method ($2^s + 1$ and $10^s + 1$) than for the multiplicative method ($2^s + 3$ and $10^s + 3$) and thus require fewer additions. However, on the IBM 1620, with its variable word length and two-address instructions, the multiplicative shift-and-add multiplier $10^6 + 11$ cuts generating time from 20 ms to 2 ms, whereas the mixed shift-and-add multiplier $10^2 + 1$ is somewhat slower at 2½ ms.[28] In view of the less consistent statistical behavior of the mixed method, Hull and Dobell[11] feel that the multiplicative method is to be preferred on machines with fast multiplication times.

A word of caution is in order concerning nonrandomness in the least significant digits of numbers from congruential generators. For instance, in binary sequences from multiplicative generators with $m = 2^k$, only the first bit has the maximum period. The other bit periods decrease with bit position until the last bit is always 1.[18,25] Thus random digits, if needed, should be taken from the most significant end of the numbers. However, Hutchinson[29] describes a multiplicative generator with $m = 2^{35} - 31$, which appears to be as random in the least significant bits as in the most significant bits; it has the disadvantage that a multiplication and a division are involved. Greenberger[30] discusses a similar generator coded to avoid multiply-and-divide instructions; he discusses generators based on a number of choices for m with program listings for each.

To illustrate the multiplicative method imagine a binary machine with a word length of six. Let $m = 2^6 = 64$ to avoid division in the congruence relation and to simplify normalizing to the unit interval. For maximum period, $m/4$, the multiplier must be of the form $8t \pm 3$; $t = 1$ gives values near $\sqrt{m}$. Picking $a = 8t + 3 = 11$ gives

$$x_{i+1} \equiv 11x_i \pmod{64}$$

The resulting sequence (picking $x_0 = 1$) is

1, 11, 57, 51, 49, 27, 41, 3, 33,

43, 25, 19, 17, 59, 9, 35, 1, . . .

These values could be normalized to the unit interval by shifting the binary point (or, in Fortran, by actually dividing by 64).

The foregoing sequence is plotted in Fig. 1. The shaded circles plot the pairs (1, 11), (57, 51), . . ., and the open squares plot the pairs (11, 57), (51, 49), ... It is surprising to see that all points lie on families of straight lines as indicated. Straight-line patterns are further discussed in Appendix B, where it will be seen that with generators having sufficiently long periods such patterns are no longer evident.

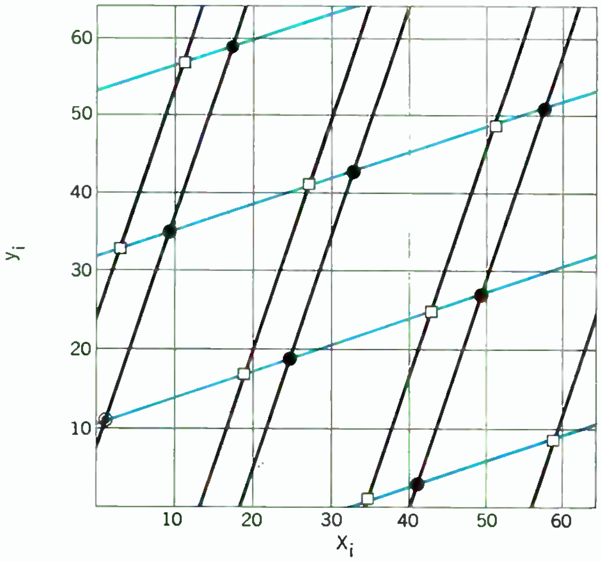
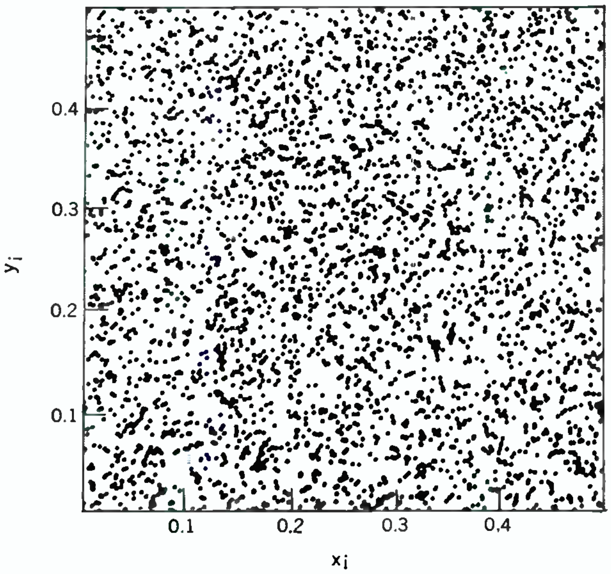


FIGURE 1. "Random" plot: y_i vs. x_i from a multiplicative congruential generator with period $= 2^6$.

FIGURE 2. Random plot: y_i vs. x_i from a multiplicative congruential generator with period $= 2^{33}$.



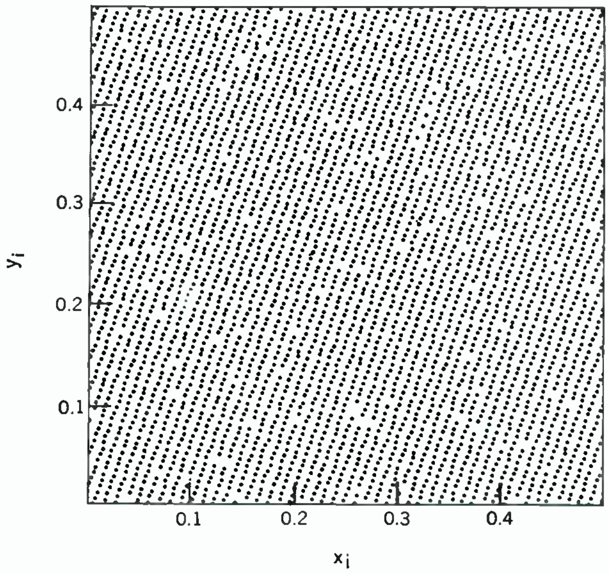
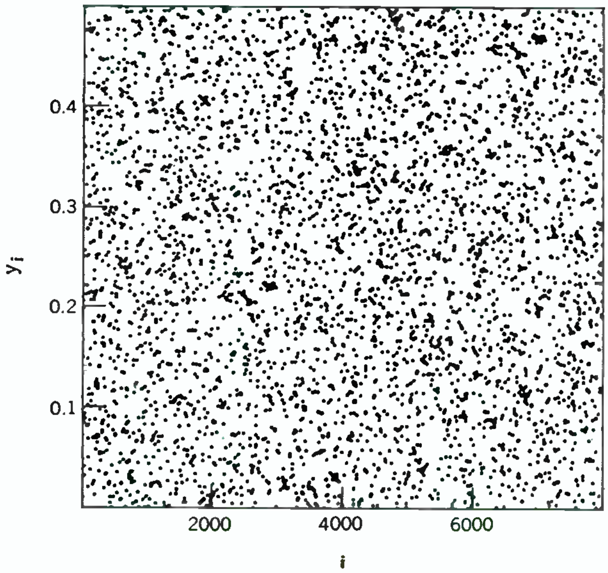


FIGURE 3. Random plot: y_i vs. x_i from a multiplicative congruential generator with period $= 2^{15}$.

FIGURE 4. Random plot: y_i vs. i from the generator that was used in Fig. 3.



Appendix B

Straight-line patterns in random-number plots

Let the sequence of numbers produced by a random-number generator be designated $x_1, y_1, x_2, y_2, \ldots, x_i, y_i$. Figure 2 shows a plot of y_i vs. x_i from a FAP (Fortran assembly program)-coded multiplicative congruential generator with a period of 2^{33}. Such plots were used to simulate photographic background noise.* Plots of y_i vs. i and x_i vs. i were similarly random. A multiplicative generator coded in Fortran II gave unsuitable straight-line patterns as in Fig. 3. Since the maximum integer in Fortran II is 2^{17}, such a generator has a period of $2^{17}/4 = 32\,768$ and thus can produce only 16 384 distinct (x, y) points when called in the above manner. Figures 4 and 5 plot y_i vs. i and x_i vs. i, respectively, and show that the generated sequence is much more random in one dimension. However, fairly prominent straight-line patterns are still evident. Greenberger[13] gives a theoretical discussion of the similar patterns found by Lach that were described earlier.

The interesting thing is that the usual statistical tests do not reveal such two-dimensional straight-line patterns. M. R. Schroeder[31] proposed a Fourier transform test. This test calculates

$$F(n, m) = (\mathrm{RE})^2 + (\mathrm{IM})^2$$

where

$$\mathrm{RE} = \sum_i \cos 2\pi(nx_i + my_i)$$

$$\mathrm{IM} = \sum_i \sin 2\pi(nx_i + my_i)$$

If there is an evenly spaced, parallel straight-line pattern in the sequence, $F(n, m)$ will show a pronounced peak whenever n is a multiple of the number of lines intersecting the x axis and m is the same multiple of the number of lines intersecting the y axis. The test was applied to the sequence that produced Fig. 6. This sequence, unfortunately, was produced by a subroutine of 1959 vintage for which no explanation of the method used is available and which is coded in FAP so that the method is not readily recoverable. However, it will serve to illustrate

* In order to keep these illustrations to a reasonable size, Figs. 2 through 6 and Fig. 9 show only a portion of the unit square.

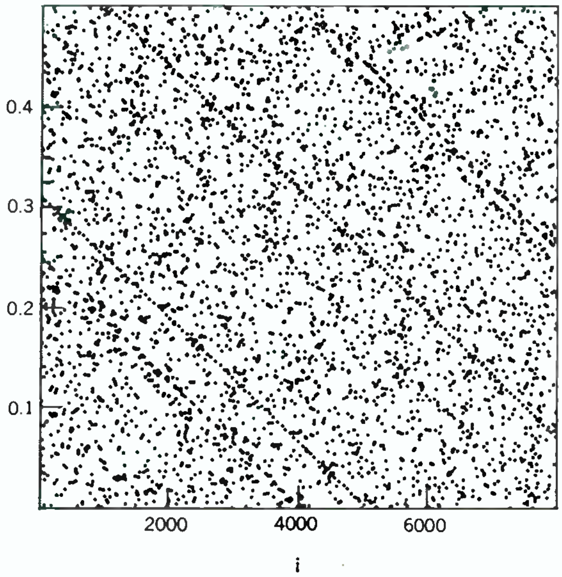


FIGURE 5. Random plot: x_i vs. i from the generator that was used in Fig. 3.

FIGURE 6. Random plot: y_i vs. x_i from a generator of an unknown type.

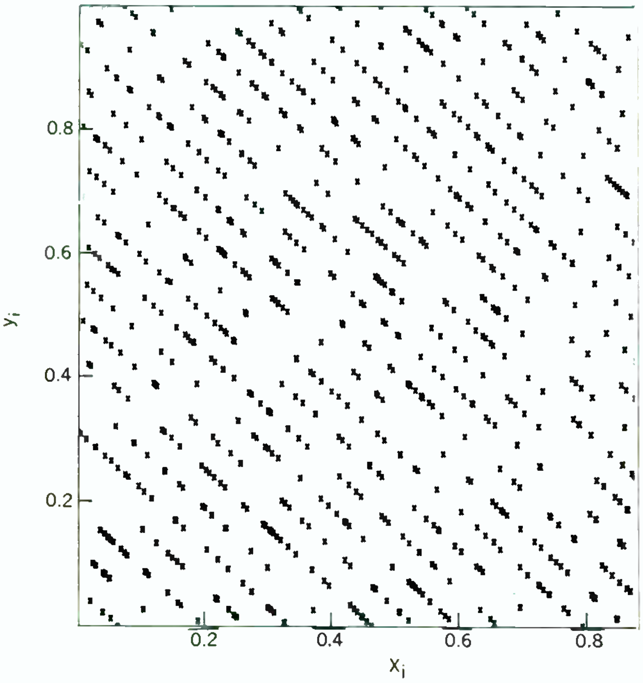
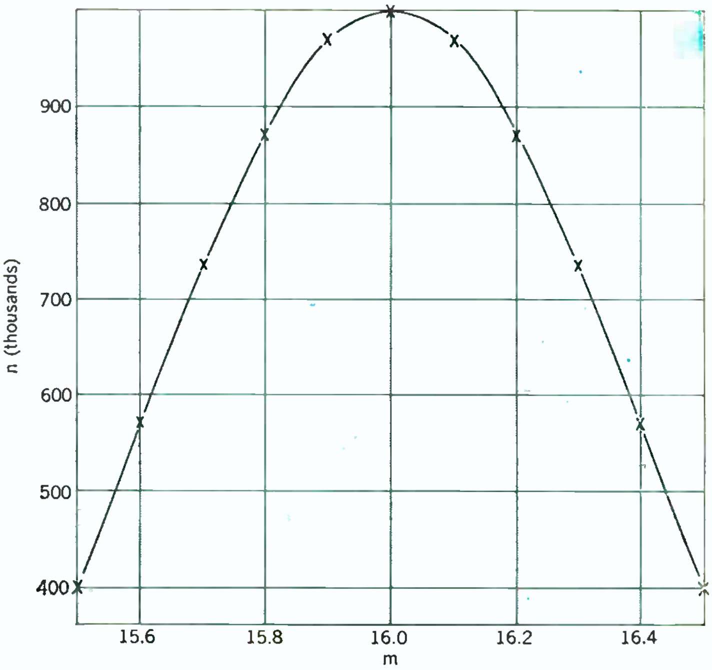


FIGURE 7. Results of the Fourier-transform test on the generator used for Fig. 6; in the neighborhood of a "resonant" value of m and at a "resonant" value of n.



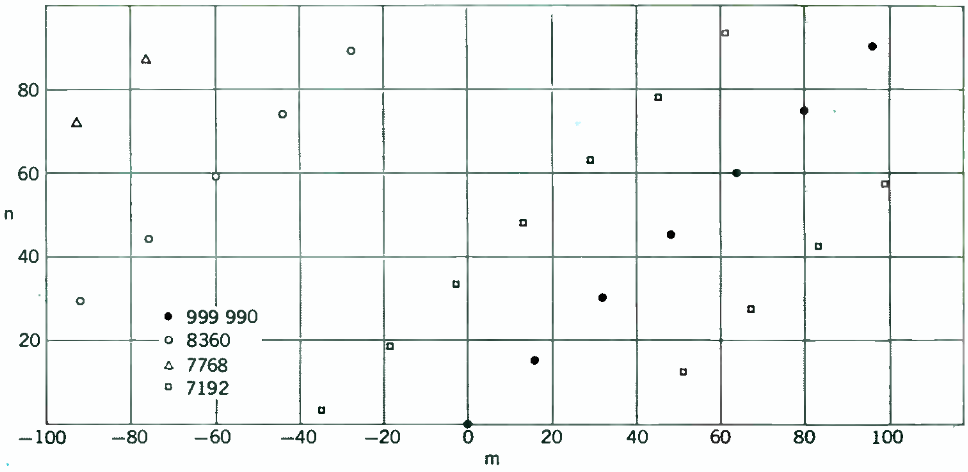
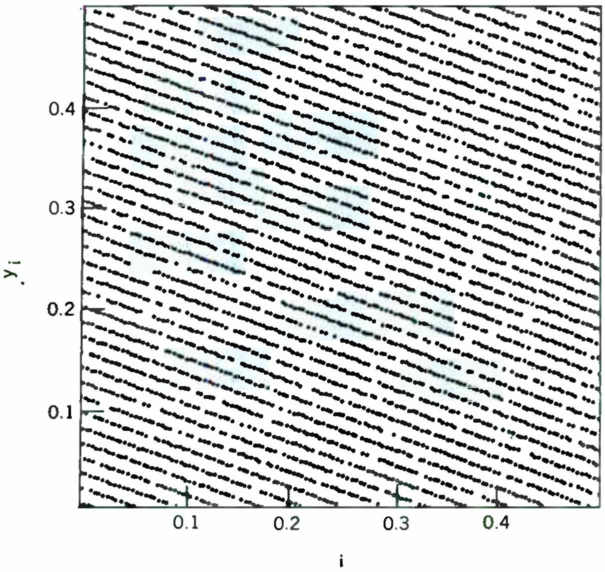


FIGURE 8. Results of the Fourier-transform test on the generator used for Fig. 6; combined plot, for all values of m and n, of test results ⩾ 7192. This parallel-line pattern is extended by results < 7192.

FIGURE 9. Random plot: y_i vs. x_i from the generator used for Fig. 2, except that an improper starter was used.



the Fourier transform test, especially since it had passed a battery of the more usual statistical tests in spite of the straight-line pattern of Fig. 6. This pattern has 15 x-axis intersections and 16 y-axis intersections. $F(n, m)$ was calculated for n ranging from 0 to 100 and for m ranging from −100 to 100, both in steps of one. The results were plotted as $F(n, m)$ vs. m, with a separate plot for each n. $F(n, m)$ was generally less than 9000, but shot to a value of 999 990 at the following points:

$$n = 0, m = 0$$

$$n = 15, m = 16$$

$$\vdots$$

$$n = 90, m = 96$$

Figure 7 shows $F(n, m)$ for $n = 15$, expanded in the neighborhood of $m = 16$. Figure 8 plots values of $F(n, m)$ equal to 7192 or greater.

Perhaps the generator that produced Fig. 2 also contains a straight-line pattern but one whose spacing is so small, due to the long period (over 8.5 billion), that the plot does not resolve it. If so it could be resolved by expanding in a small subspace. However, this process would probably require an impractical amount of computer time, which is perhaps equivalent to saying that such a pattern would be of no practical importance anyway. Straight lines can sometimes result from using an improper starter, so that the period is reduced as in Fig. 9, which was produced by the same generator as Fig. 2 except that the starter was 262 144 = 8^6 instead of one.

This report leans heavily on many of the articles published on the subject, particularly on the Hull and Dobell review paper.[10] I wish to thank Prof. Phyllis Fox of Newark College of Engineering for guidance and encouragement throughout. I also wish to acknowledge many helpful discussions with J. S. Courtney-Pratt and E. Eisner, and to thank W. M. Gentleman and C. S. Sherrerd, of Bell Telephone Laboratories, for their valuable comments.

REFERENCE NOTES

The Hull and Dobell paper[10] is a very useful survey of random-number generators with an exhaustive bibliography. Edmonds,[32] Spenser,[33] and Hammer and Green[34] give shorter accounts. Meyer[35] and the NBS publication[24] contain many valuable source papers. Meyer gives an extensive bibliography.

Hammersley and Handscomb[23] and Shreider[17] are general texts on the Monte Carlo method, but have useful sections on random numbers and include extensive bibliographies. The article by McCracken[4] also gives an interesting account of the Monte Carlo method. Kendall and Babington-Smith,[21] Hamming,[3] and Von Mises[36] present interesting discussions on the concept of randomness.

IBM Manual C20-8011[18] gives a clear exposition of the multiplicative congruential (power residue) method with detailed recipes for binary and decimal machines. Hull and Dobell[10] give a thorough theoretical discussion of the mixed congruential method with a briefer treatment of the multiplicative method. Hamming[3] and Barnett[25] give the theory of the multiplicative method. Peach[12] presents the mixed method theory in terms of high school mathematics rather than number theory. Hull and Dobell[11] give a very good experimental and theoretical comparison of the multiplicative and mixed congruential methods for binary machines. Allard, Dobell, and Hull[28] present a similar treatment for decimal machines. MacLaren and Marsaglia[14] compare test results for the multiplicative and mixed methods.

Muller[37] compares methods available in 1958 for transforming from a uniform to a normal distribution, including an attractively simple method developed by Box and Muller,[38] which is accurate, easy to program, and reasonably fast. Kronmal[39] gives test results with a normal generator based on the method of Box and Muller. Marsaglia and Bray[40] give a variation of the Box and Muller method that not only is as accurate and easy to program but is faster.

Marsaglia, MacLaren, and Bray[41] give a method for transforming to a normal distribution that is accurate and very fast but is moderately difficult to program and uses a fair amount of storage. They recommend it as the basic normal generator in any computer installation. Marsaglia and Bray[40] give another method, which, though somewhat slower, requires very little storage and is much easier to program. Marsaglia[42,43] deals with the general problem of transforming from a uniform distribution to other distributions.

REFERENCES

1. Metropolis, N., and Ulam, S., "The Monte Carlo method," *J. Am. Statist. Assoc.*, vol. 44, pp. 335–341, 1949.

2. Bauer, W. F., "The Monte Carlo method," *J. Soc. Ind. Appl. Math.*, vol. 6, pp. 438–451, 1958.

3. Hamming, R. W., *Numerical Methods for Scientists and Engineers.* New York: McGraw-Hill, 1962.

4. McCracken, D. D., "The Monte Carlo method," *Sci. Am.*, vol. 192, no. 5, pp. 90–96, 1955.

5. Rand Corp., *A Million Random Digits with 100,000 Normal Deviates.* New York: Free Press, 1955.

6. Lehmer, D. H., "Mathematical methods in large-scale computing units," *Proc. 2nd Symp. on Large-Scale Digital Calculating Mach.*, 1949; *Ann. Comput. Lab. Harvard Univ.*, vol. 26, pp. 141–146, 1951.

7. Rotenberg, A., "A new pseudo-random number generator," *J. Assoc. Comput. Mach.*, vol. 7, pp. 75–77, 1960.

8. Coveyou, R. R., "Serial correlation in the generation of pseudo-random numbers," *J. Assoc. Comput. Mach.*, vol. 7, pp. 72–74, 1960.

9. Stockmal, F., "Calculations with pseudo-random numbers," *J. Assoc. Comput. Mach.*, vol. 11, pp. 41–52, 1964.

10. Hull, T. E., and Dobell, A. R., "Random number generators," *SIAM Rev.*, vol. 4, pp. 230–254, 1962.

11. Hull, T. E., and Dobell, A. R., "Mixed congruential random number generators for binary machines, *J. Assoc. Comput. Mach.*, vol. 11, pp. 31–40, 1964.

12. Peach, P., "Bias in pseudo-random numbers," *J. Am. Statist. Assoc.*, vol. 56, pp. 610–618, 1961.

13. Greenberger, M., "Method in randomness," *Commun. Assoc. Comput. Mach.*, vol. 8, pp. 177–179, 1965.

14. MacLaren, M. D., and Marsaglia, G., "Uniform random number generators," *J. Assoc. Comput. Mach.*, vol. 12, pp. 83–89, 1965.

15. Jansson, B., "Autocorrelations between pseudorandom numbers," *BIT*, vol. 4, pp. 6–27, 1964.

16. Sobol, I. M., "On periods of pseudo-random sequences," *Theory Probability Appl. (U.S.S.R.)*, vol. 9, pp. 333–338, 1964.

17. Shreider, Yu. A., ed., *Method of Statistical Testing.* New York: Elsevier, 1964.

18. "Random number generating and testing," Reference Manual C20-8011, IBM Corp., New York, N.Y., 1959.

19. Green, B. F., Jr., Smith, J. E. K., and Klem, L., "Empirical tests of an additive random number generator," *J. Assoc. Comput. Mach.*, vol. 6, pp. 527–537, 1959.

20. Hampton, R. L. T., "A hybrid analog-digital pseudo-random noise generator," *Simulation*, vol. 4, pp. 179–190, 1965.

21. Kendall, M. G., and Babington-Smith, B., "Randomness and random sampling numbers," *J. Roy. Statist. Soc.*, vol. 101, pp. 147–166, 1938.

22. Kendall, M. G., and Babington-Smith, B., "Second paper on random sampling numbers," *J. Roy. Statist. Soc., Suppl. 6*, pp. 51–61, 1939.

23. Hammersley, J. M., and Handscomb, D. C., *Monte Carlo Methods.* New York: Wiley, 1964.

24. "Monte Carlo method," Appl. Math. Series No. 12, Nat'l Bur. of Standards, Washington, D.C., 1951.

25. Barnett, V. D., "The behavior of pseudo-random sequences generated on computers by the multiplicative congruential method," *Math. Comput.*, vol. 16, pp. 63–69, 1962.

26. Greenberger, M., "Notes on a new pseudo-random number generator," *J. Assoc. Comput. Mach.*, vol. 8, pp. 163–167, 1961.

27. Greenberger, M., "An a priori determination of serial correlation in computer generated random numbers," *Math. Comput.*, vol. 15, pp. 383–389, 1961.

28. Allard, J. L., Dobell, A. R., and Hull, T. E., "Mixed congruential random number generators for decimal machines," *J. Assoc. Comput. Mach.*, vol. 10, pp. 131–141, 1963.

29. Hutchinson, D. W., "A new uniform pseudorandom number generator," *Commun. Assoc. Comput. Mach.*, vol. 9, pp. 432–433, 1966.

30. Orcutt, G. H., Greenberger, M., Korbel, J., and Rivlin, A. M., *Microanalysis of Socioeconomic Systems.* New York: Harper & Row, 1961, Appendix to Part IV, pp. 356–370.

31. Schroeder, M. R., Private communication.

32. Edmonds, A. R., "The generation of pseudo-random numbers on electronic digital computers," *Comput. J.*, vol. 2, pp. 181–185, 1960.

33. Spenser, G., "Random numbers and their generation," *Comput. Automation*, vol. 4, no. 3, pp. 10–11, 23, 1955.

34. Hammer, C., and Green, L. G., "Generating random numbers," *Instr. Control Syst.*, vol. 37, no. 5, pp. 149–150, 1964.

35. Meyer, H. A., ed., *Symposium on Monte Carlo Methods.* New York: Wiley, 1956.

36. Von Mises, R., *Probability, Statistics and Truth*, 2nd ed. New York: Macmillan, 1957.

37. Muller, M. E., "A comparison of methods for generating normal deviates on digital computers," *J. Assoc. Comput. Mach.*, vol. 6, pp. 376–383, 1959.

38. Box, G. E. P., and Muller, M. E., "A note on the generation of random normal deviates," *Ann. Math. Statist.*, vol. 29, pp. 610–611, 1958.

39. Kronmal, R., "Evaluation of a pseudorandom normal number generator," *J. Assoc. Comput. Mach.*, vol. 11, pp. 357–363, 1964.

40. Marsaglia, G., and Bray, T. A., "A convenient method for generating normal variables," *SIAM Rev.*, vol. 6, pp. 260–264, 1964.

41. Marsaglia, G., MacLaren, M. D., and Bray, T. A., "A fast procedure for generating normal random variables," *Commun. Assoc. Comput. Mach.*, vol. 7, pp. 4–10, 1964.

42. Marsaglia, G., "Expressing a random variable in terms of uniform random variables," *Ann. Math. Statist.*, vol. 32, pp. 894–898, 1961.

43. Marsaglia, G., "Generating discrete random variables in a computer," *Commun. Assoc. Comput. Mach.*, vol. 6, pp. 37–38, 1963.

BIBLIOGRAPHY

Ellis, D. J., and Ryan, P. C., "Tests on the multiplicative congruence method of generating pseudorandom numbers on the NAREC computer," AD-615-313, U.S. Naval Res. Lab., Washington, D.C., 1965.

Franklin, J. N., "Deterministic simulation of random processes," *Math. Comput.*, vol. 17, pp. 28–59, 1963.

Hampton, R. L. T., "Experiments using pseudo-random noise," *Simulation*, vol. 4, pp. 246–254, 1965.

Howe, J. E., "The generation of random numbers from various probability distributions," AD-475-568, U.S. Naval Postgraduate School, Master's thesis, 1965.

Jagerman, D. L., "The autocorrelation and joint distribution functions of the sequences $\{(a/m)j^2,\}$ $\{(a/m)(j + \tau)^2\}$," *Math. Comput.*, vol. 18, pp. 211–232, 1964.

Jansson, B., "Generation of random bivariate normal deviates and computation of related integrals," *BIT*, vol. 4, pp. 205–212, 1964.

MacLaren, M. D., Marsaglia, G., and Bray, T. A., "A fast procedure for generating exponential random variables," *Commun. Assoc. Comput. Mach.*, vol. 7, pp. 298–300, 1964.

Marsaglia, G., "Generating a variable from the tail of the normal distribution," *Technometrics*, vol. 6, pp. 101–102, 1964.

Stoneham, R. G., "A study of 60,000 digits of the transcendental *e*," *Am. Math. Monthly*, vol. 72, pp. 483–500, 1965.

Tausworthe, R. C., "Random numbers generated by linear recurrence modulo two," *Math. Comput.*, vol. 19, pp. 201–209, 1965.

Ushakov, I. A., "A procedure for obtaining random numbers with a uniform law of distribution," *Eng. Cybernetics*, no. 1, pp. 179–183, 1965.

Ushakov, I. A., and Panyukov, B. V., "Generation of random numbers," *Eng. Cybernetics*, no. 4, pp. 194–199, 1965.

Zelen, M., and Severo, N. C., "Methods of generating random numbers and their applications," Sect. 26.8 of "Handbook of mathematical functions," Appl. Math. Series No. 55, Nat'l Bur. of Standards, Washington, D.C., 1964, pp. 949–953.

Major recommendations of the U.S. President's Patent Commission

During the past 130 years there have been no basic changes in the general character of the United States patent system, even though the country has undergone a dramatic transformation into an enormously complex industrial community. A recently released report contains important recommendations that would result in a patent system better able to cope with our exploding technology

B. M. Oliver *Hewlett-Packard Company*

On April 8, 1965, President Johnson ordered the establishment of a Commission to study the U.S. patent system in all of its aspects and to recommend legislation that would improve its service to society. The membership of the Commission was announced on July 23, 1965, and comprised 14 representatives of important areas affected by the patent system: science and engineering, large and small business, universities, patent attorneys, the judiciary, the National Science Foundation and the Departments of Commerce and Defense. Other government agencies were represented by observers. In its report, the Commission made 35 recommendations, some of which, if adopted, will represent the first major changes in our patent system since 1836. In this article one of the Commission members, the IEEE Secretary, presents some of the highlights of the Commission's report.

As inventors or as manufacturers, as applicants, licensees or litigants, most IEEE members in the United States and many in other countries have been affected by the U.S. patent system. Some owe their livelihood to patents; others consider them a damned nuisance. Although individual attitudes about the value of a patent system vary widely, there is no doubt that patents determine to a great extent how our industry operates and the kinds of opportunities afforded the creative engineer. Any major changes that may occur in U.S. patent law are of direct concern to IEEE members in the United States and, because such changes may induce changes in the laws of other countries, they should also be of interest to IEEE members generally.

Quite early in its study the Commission came to the unanimous agreement that a patent system was a necessary part of our technological society. Without patent protection of some sort much of the knowledge that today is disseminated so freely at meetings and in technical society journals and other publications would be retained in secrecy. Countless man-hours would be wasted in reinvention and rediscovery. Since they would soon be copied, many worthwhile inventions would be deemed poor investment risks and would never be produced. The respect for intellectual property, legally enforced by the patent system, would vanish and industry would revert to the guild secrecy of olden times.

By the simple *quid pro quo* of giving the inventor, or his assignee, the right to a limited monopoly in return for making public the knowledge of his invention, a patent system protects (and hence encourages) investment in research and development, assists inventors by making available the latest state of the art, and unites our pluralistic efforts into an open race to advance the utilization of knowledge.

The Commission study revealed several serious problem areas in our present patent system. Today's standards of invention are too low. A great many patents issue on trivial inventions obvious to anyone skilled in the art. Partly as a result of these low standards, the office is flooded with applications. A huge backlog of applications delays examination by over two years. The pressure of this work load forces the Patent Office to reward the examiners for quantity of disposals rather than for the quality of examination. Since it is far easier for the examiner to allow than to reject, it is inevitable that patents of doubtful validity issue. This chokes our courts with suits contesting the validity of issued patents.

The Commission invited and received communications from many organizations and individuals recommending or opposing various changes in our present patent system. On the basis of these communications and its own collective background of experience with the present system, the Commission developed a plan for revision comprising 35 recommendations. The Commission believes that these proposed changes would make our patent system viable and better able to serve its role of promoting invention and progress. Only the primary recommendations of interest to most IEEE members have been included here. Those IEEE members more closely concerned with patents should read the complete Commission report.*

The principal objectives of the reforms recommended by the Commission are:

1. To improve the quality, validity and enforceability of the U.S. patent.
2. To accelerate the disclosure of technological advances.

*The report, entitled "To Promote the Progress of. . . Useful Arts," is available at 65 cents from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.

3. To accelerate the issuance process.
4. To minimize patent litigation and to reduce the cost thereof.
5. To bring U.S. patent practice into closer harmony with that of other countries, as a first step toward an international patent system.
6. To prepare the patent system to cope with further increased rate of invention in the decades ahead.

In the following pages the recommendations are printed in color. Other verbatim excerpts from the report are in quotes.

Recommendation I: Priority of invention

Under present U.S. practice, if priority is disputed the patent is granted to the "first to invent"—that is, to the inventor whose records, supported by independent testimony, show that he was the first to conceive and to reduce the invention to practice, *provided* that he did not publicly disclose the invention through publication or public use or sale more than one year (the so-called "grace period") before filing for a patent. Public use or sale abroad does not prevent a U.S. applicant from patenting the same invention here. The Commission's first recommendation would change these practices, which are unique to the United States.

Prior art shall comprise any information, known to the public, or made available to the public by means of disclosure in tangible form or by use or placing on sale, anywhere in the world, prior to the effective filing date of the application.

A disclosure in a U.S. patent or published complete application shall constitute prior art as of its effective (U.S. or foreign) filing date.

Under this recommendation

1. Patents would be granted to the *first to file*.
2. There would be no "grace period."
3. Prior public knowledge, public use, or sale anywhere in the world would bar a U.S. patent.

Under the present "first to invent" system, when the Patent Office receives two or more applications for the same invention it declares an interference, and priority must be resolved on the basis of evidence showing the earliest date of conception, together with due diligence in reducing the invention to practice. Obviously, witnessed records are important, and a careful inventor or employer will see that these are kept. But often the records are inadequate and can be falsified. What constitutes due diligence? Interferences are expensive and can be wrongly adjudicated.

The first-to-file system eliminates interferences in one stroke. Patent rights would go to the first inventor to file an application, just as in scientific circles credit goes to the first to publish. It is often argued that a first-to-file system would create a race to the Patent Office. Perhaps, but there is the same race to the editor today in scientific publication, and no one is seriously concerned. "The Commission believes it is as equitable to grant a patent to the first to file as to the [first to invent]. . . . circumstances may determine the winner in either case. But the first to file is more apt to be the inventor who first appreciated the importance of the invention and promptly acted to make the invention available to the public."

A first-to-file system would "encourage prompt disclosure of [new] technology, substitute for the delays and expense of interference proceedings a fair and inexpensive [criterion] of priority; and bring U.S. practice into harmony with that prevailing in almost all other industrial nations."

The admission of public knowledge, use, or sale anywhere in the world as prior art seems long overdue in the modern age of jet travel and fast communication. As things now stand, public use or sale in Alaska or Hawaii constitutes prior art against a U.S. applicant, but public use or sale in Canada or a Mexican border town does not! The Commission believes such arbitrary geographical distinctions should be eliminated.

Finally, since publication would bar an inventor who had not yet filed, this recommendation would eliminate the need for "defensive" patenting. A corporation not wishing to exploit the rights to an invention but merely wishing to protect its own rights to use it should, under present law, file for a patent. If no interferences develop, the patent is then granted and simply never exercised. Defensive patenting accounts for a fair fraction of the present patent office load. It is wasted labor.

Recommendation II: The preliminary application

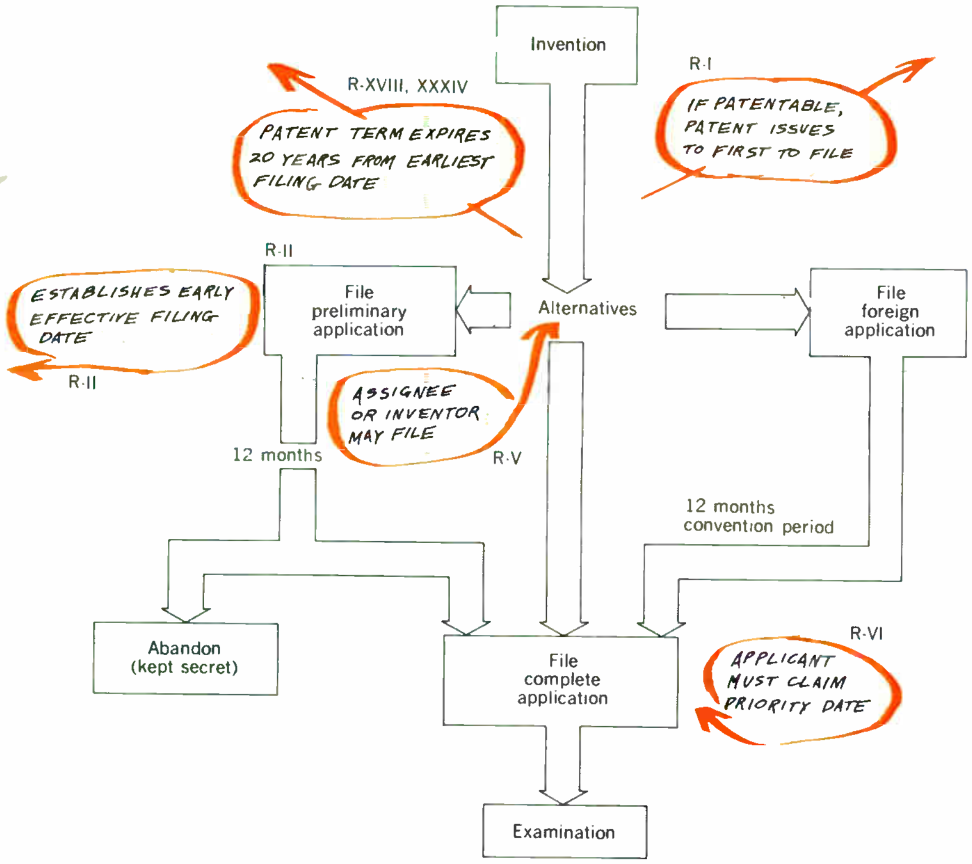
To be fair to the inventor, to allow him time to seek support or to develop and market test his invention, and to allow free discussion of new discoveries and inventions in academic and scientific circles and at technical conferences, the proposed first-to-file system with no grace period needs an "instant" disclosure procedure to establish a filing date in the Patent Office. The procedure should be free of the expense and delay associated with preparation of a formal application. Accordingly, the Commission has proposed the preliminary application.

A preliminary application may be used to secure a filing date for all features of an invention disclosed therein, if the disclosure subsequently appears in a complete application. Requirements as to form shall be minimal and claims need not be included.

One or more preliminary applications may be filed and consolidated into one complete application filed within twelve months of the earliest preliminary application or foreign application relied on.

Under this provision the applicant could file a simple description of the invention that describes in a clear fashion all of the novel features. This could be done without the assistance of an attorney. Additional preliminary applications could be filed to cover aspects of the invention developed subsequently. In effect, such preliminary applications would replace the significant entries an inventor now makes in his notebook, but each would bear the incontestable date stamp of the Office. Within 12 months after filing the first preliminary, the inventor would have to consolidate all the preliminaries into a complete application.

Any novel content in each preliminary would be protected as of the date of its receipt by the Office. Information contained in these preliminaries could then be disclosed to the public without jeopardy to the inventor's rights. The period between filing a preliminary and filing the complete application would in many ways be like the present grace period.



Invention
R-XVIII, XXXIV
PATENT TERM EXPIRES 20 YEARS FROM EARLIEST FILING DATE
R-I
IF PATENTABLE, PATENT ISSUES TO FIRST TO FILE
R-II
ESTABLISHES EARLY EFFECTIVE FILING DATE
R-II
File preliminary application
Alternatives
File foreign application
ASSIGNEE OR INVENTOR MAY FILE
R-V
12 months
12 months convention period
R-VI
APPLICANT MUST CLAIM PRIORITY DATE
Abandon (kept secret)
File complete application
Examination

FILING an application. Appropriate recommendations are keyed by number.

This procedure should create no significant burden in the Patent Office, since each preliminary need merely be stamped and filed. Upon later examination of the complete application, the examiner would have immediately before him proof as to the first-to-file in case of copending applications on the same invention. Preliminary applications not followed up within a year by a complete application would be destroyed.

Recommendation III:
Exhibitions and unauthorized publication

Under a present treaty agreement, inventions may be displayed at certain recognized international exhibitions without loss of patent rights, provided an application is filed within six months thereof. Although there is no real need for this provision given the possibility of filing a preliminary application, the U.S. law must comply until the treaty is modified.

Recommendation III recognizes this situation, and then goes on to state:

> **An unauthorized public disclosure of information derived from the inventor or his assignee shall not constitute prior art against him, if, within six months after said disclosure, a complete application for the invention is filed by the inventor or assignee.**
>
> **Any allegation, that a disclosure should not constitute prior art because it was unauthorized, shall be considered by the Patent Office only if it is verified, sets forth details establishing a *prima facie* case, and is accompanied by proof that notice has been served on the party accused of making the disclosure.**
>
> **If the party accused promptly contests the allegation, the application shall not issue as a patent until the matter is finally judicially determined in favor of the applicant.**

This recommendation protects the inventor against loss of his patent right because of unauthorized disclosure by providing a procedure whereby the effect of such disclosure may be nullified. Should the accused party not contest the allegation, the Patent Office would simply ignore the disclosure as prior art against the inventor. The disclosure would bar other applicants, however. The inventor would not be entitled to the disclosure date for priority purposes since there may have been intervening untainted disclosures that would contribute prior art against him.

Recommendation IV: Patentable subject matter

Under present law, patents are often granted on particular designs, such as on fabrics or in jewelry, or on the styling of an industrial product. Patents may also be obtained on new varieties of asexually reproduced plants, such as tuberous plants that propagate true to type even if heterozygous. It is not clear at present whether computer programs are patentable. In the opinion of the Commission all of these areas should be excluded from patent protection. The report recommends that:

The classes of patentable subject matter shall continue as at present, except:

1. All provisions in the patent statute for design patents shall be deleted, and another form of protection provided.

2. All provisions in the patent statute for plant patents shall be deleted, and another form of protection provided.

3. ["Programs" for data-processing machines] shall not be considered patentable regardless of whether the program is claimed as: (a) an article, (b) a process described in terms of the operations performed by a machine pursuant to a program, or (c) one or more machine configurations established by a program.

The Commission holds "that all patentable subject matter should meet the statutory provisions for novelty, utility and unobviousness and that the above subject matter cannot readily be examined for adherence to these criteria."

The exclusion of computer programs presumably would include built-in programs in special-purpose computers. Thus, while the principle of a particular read-only memory might be patentable, its content would not. Nor could a special-purpose computer so programmed be held to infringe a general-purpose machine programmed by software and vice versa. In view of the enormous number of programs already in the public domain, their rapid proliferation without patent protection, and the subtle ways in which apparently different programs can produce the same end results or almost identical programs produce very different results, the inclusion of programs in the patent system would needlessly produce a nightmarish examination problem.

Recommendation VII: Early publication

Under the present law, patents are published only upon issuance. With the present backlog, and with dilatory practices on the part of the applicant, important inventions can lie hidden in the Office for several years to the detriment of the art and to the later distress of many who have meanwhile been innocently practicing the invention. Recommendation VII would end this problem by requiring all applications to be published within two years unless abandoned by the applicant. An abandoned application could be dedicated to the public. It could also be retained in secrecy as at present.

Publication of a pending application shall occur eighteen to twenty-four months after its earliest effective filing date, or promptly after allowance appeal, whichever comes first.

An applicant, for any reason, may request earlier publication of his pending complete application.

An application shall be "republished" promptly after allowance or appeal subsequent to initial publication, and again upon issuance as a patent, to the extent needed to update the initially published application and give notice of its status.

The only drawback of this recommendation would be a somewhat increased volume of publication, but the advantages seem to far outweigh this cost. In addition, there would be an incentive for the Office to dispose of the case (i.e., allow or reject) prior to the publication deadline and thereby avoid republication.

Recommendation IX: Optional deferred examination

The flood of applications in recent years has caused some countries to consider or adopt a deferred examination procedure. The complete application is merely examined for form and then published. No rights are bestowed until later examination has taken place and a final patent has been granted. The rationale is that only important patents will ever be examined and therefore consume the examiner's time. There is, however, little total cost saving, since more publication takes place. Also, unexamined patents constitute a threat to potential infringers, who must be able to request examination in order to clarify their position. The Commission was almost unanimously opposed to the adoption of a deferred system unless all other measures failed to keep the backlog down, but recognized that such a step might ultimately be necessary. "The Commission clearly favors a high quality immediate examination system if it can be maintained without a constantly increasing backlog. Nevertheless, it is recommended that:

Standby statutory authority should be provided for optional deferred examination.

If an optional deferred system were to be put into effect, the Commission felt certain provisions should be included:

1. The examination shall be deferred at the option of the applicant, evidenced by the election not to accompany the complete application with an examination fee.

Request for examination, accompanied by payment of an examination fee, may be made anytime within five years from the effective filing date of the application.

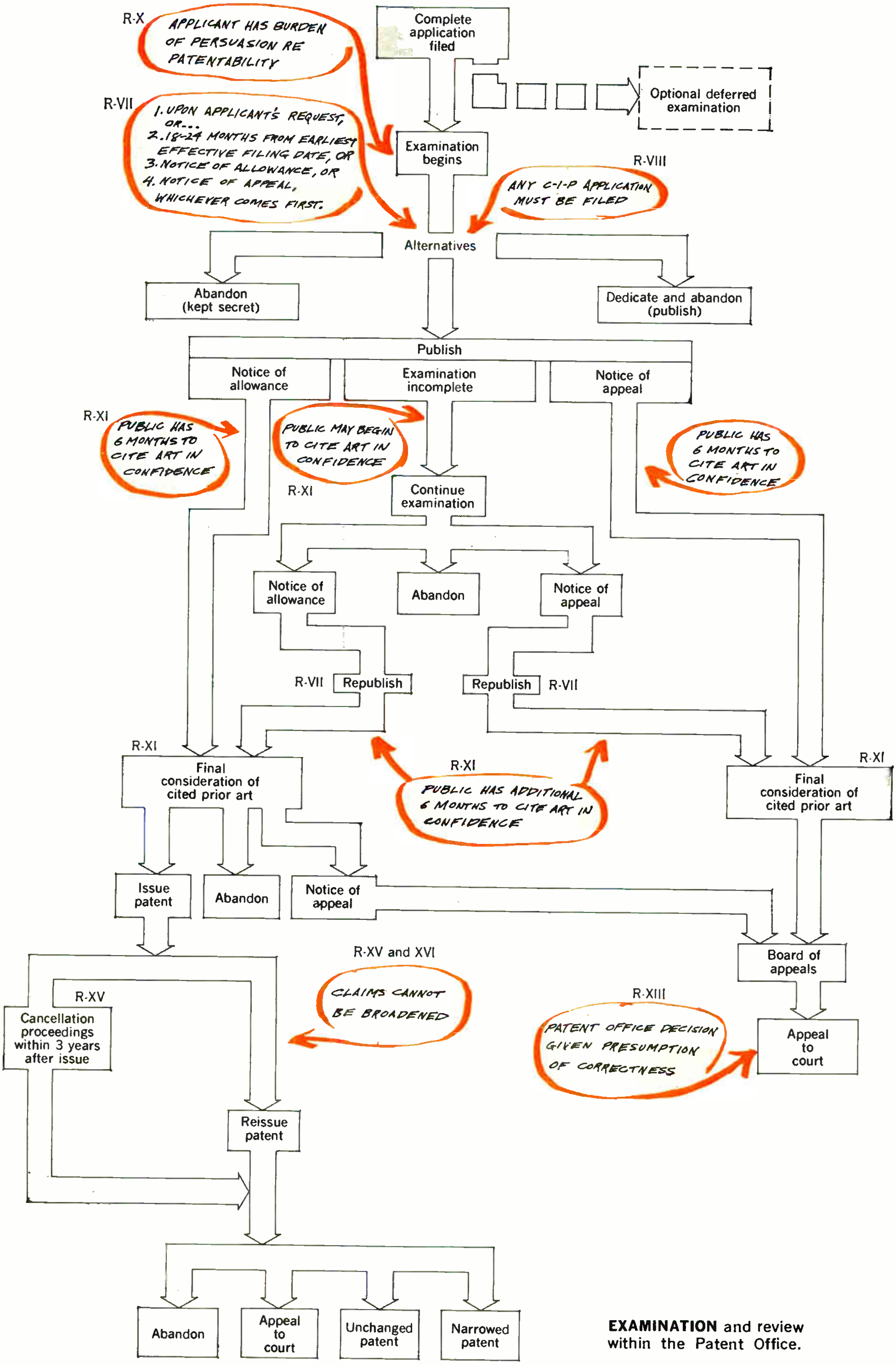
2. A deferred application shall be promptly inspected for formal matters and then published.

3. Any party, without being required to disclose his identity, may provoke an examination upon request and payment of the fee.

4. Unless made special upon the request of any party, an application initially deferred shall be inserted in the queue of applications set for examination in an order based on the date of payment of the examination fee.

5. Examination of pending parent or continuing applications shall not be deferred beyond the time when examination is requested of any of the parent or continuing applications.

No other question took more of the Commission's time than the one of to defer or not to defer, and, if so, how. While perhaps sounding irresolute, the Commission's position was the best that could be reached after a great deal of study.


R-X
APPLICANT HAS BURDEN OF PERSUASION RE PATENTABILITY
Complete application filed
Optional deferred examination
R-VII
1. UPON APPLICANT'S REQUEST, OR...
2. 18-24 MONTHS FROM EARLIEST EFFECTIVE FILING DATE, OR
3. NOTICE OF ALLOWANCE, OR
4. NOTICE OF APPEAL,
WHICHEVER COMES FIRST.
Examination begins
R-VIII
ANY C-I-P APPLICATION MUST BE FILED
Alternatives
Abandon (kept secret)
Dedicate and abandon (publish)
Publish
Notice of allowance
Examination incomplete
Notice of appeal
R-XI
PUBLIC HAS 6 MONTHS TO CITE ART IN CONFIDENCE
PUBLIC MAY BEGIN TO CITE ART IN CONFIDENCE
R-XI
PUBLIC HAS 6 MONTHS TO CITE ART IN CONFIDENCE
Continue examination
Notice of allowance
Abandon
Notice of appeal
R-VII
Republish
Republish
R-VII
R-XI
Final consideration of cited prior art
R-XI
PUBLIC HAS ADDITIONAL 6 MONTHS TO CITE ART IN CONFIDENCE
R-XI
Final consideration of cited prior art
Issue patent
Abandon
Notice of appeal
Board of appeals
R-XV and XVI
CLAIMS CANNOT BE BROADENED
R-XV
Cancellation proceedings within 3 years after issue
R-XIII
PATENT OFFICE DECISION GIVEN PRESUMPTION OF CORRECTNESS
Appeal to court
Reissue patent
Abandon
Appeal to court
Unchanged patent
Narrowed patent

EXAMINATION and review within the Patent Office.

Recommendation X: When in doubt, reject

An invalid patent—one that will not be upheld in court—is worse than no patent. It forces both the patentee and the threatened infringer to waste hundreds of man-hours and thousands of dollars in useless litigation. The following recommendation, while it may seem hard on the applicant, in reality would often protect him and others from his own folly.

The applicant shall have the burden of persuading the patent office that a claim is patentable.

"Until recently, the Patent Office has followed a policy of (a) instructing the examiner to resolve all reasonable doubts in favor of the applicant, and (b) prohibiting the examiner from indicating that he is allowing a claim despite his doubt as to its patentability...."

"Many have long recognized that resolving doubt in favor of the applicant is inconsistent with giving a patent a strong presumption of validity. Little justification exists for giving weight to a Patent Office decision when it resolves doubt in this manner, since it is passing the question of patentability on to the courts instead of exercising its judgment. Inasmuch as the examiner does not indicate when he has applied the rule of doubt, all patents may be questioned in this regard."

A further consequence of this recommendation might be to discourage applications of doubtful patentability once the higher standards required became known to attorneys. This, in itself, could significantly reduce the Patent Office workload. The time an examiner saves at present in allowing rather than rejecting is borrowed time, since his action encourages more low-grade applications.

Recommendation XI: Opposition

Even assuming mechanized searching, which the Patent Office does not now use except to a very limited extent in the chemical field, it is unlikely that the examiner will ever have before him all the prior art that might be cited against an application; e.g., prior art published in an instruction manual for a foreign product or embodied in a discontinued product would very likely be unavailable. The Commission thus recommends that the patent as tentatively allowed by the examiner be published and the public allowed six months to call additional prior art to his attention before issuance.

The Patent Office shall consider all patents or publications, the pertinency of which is explained in writing, cited against an application anytime until six months after the publication which gives notice that the application has been allowed or appealed to the Board of Appeals. If the Patent Office, after the citation period, determines that a claim should not be, or have been, allowed, the applicant shall be notified and given an opportunity *ex parte* both to rebut the determination and to narrow the scope of the claim. The identity of the party citing references shall be maintained in confidence.

Public use proceedings, as at present, may be instituted during the citation period.

In fields where legal staffs actively scan the *Patent Office Gazette* for patents of interest, much additional prior art may be uncovered and potentially invalid patents be rejected or the claims be appropriately narrowed. Again, this provision would protect both the applicant and the public against later expensive litigation. "Citing, or failing to cite prior art during [the opposition] period would not, however, preclude a later challenge on the [same] art."

Recommendation XIII: Review of Patent Office rejections

An applicant whose claims have been rejected by the examiner and by the Patent Office Board of Appeals may take them to the Court of Customs and Patent Appeals or to the District Court of the District of Columbia.

"Currently, the weight given on appeal to a Patent Office decision denying a patent depends upon which court reviews the decision. The Patent Office's decision is presumed correct in the District Court for the District of Columbia and the Court of Appeals for the District of Columbia Circuit, but not in the Court of Customs and Patent Appeals."

The Commission recommends a consistent policy with respect to the review of Patent Office rejections.

A Patent Office decision refusing a claim shall be given a presumption of correctness, and shall not be reversed unless clearly erroneous.

Recommendation XV: Amendment and cancellation

It frequently happens that an accused infringer after a monumental search effort uncovers prior art that clearly anticipates the threatening patent. At the present time he can only try to convince the patentee not to sue. If he fails he must stand the expense of a trial at which the pertinence of the prior art may be less evident to the judge than it would be to an examiner. Although the opposition period may uncover some of this prior art in time to prevent issuance, much will remain hidden. The Commission therefore feels a threatened infringer should be able to seek a Patent Office review in the light of newly discovered prior art, even after issuance.

The Patent Office, upon receipt of a relatively high fee, shall consider prior art of which it is apprised by a third party, when such prior art is cited and its pertinency explained in writing within a three year period after issuance of the patent. If the Patent Office then determines that a claim should not have allowed, the patent owner shall be notified and given an opportunity *ex parte* both to rebut the determination and to narrow the scope of the claim. Failure to seek review, or the affirmance of the Patent Office holding, shall result in cancellation of the claim.

When the validity of a claim is in issue before both the Patent Office and a court, the tribunal where the issue was first presented shall proceed while the other shall suspend consideration, unless the court decides otherwise for good cause.

Anyone unsuccessfully seeking Patent Office cancellation of claims shall be required to pay the patent owner's reasonable cost of defending such claims, including attorney's fees. The commissioner shall require an appropriate deposit or bond for this purpose at the start of the action.

This recommendation was originally proposed without a three-year limitation. The Commission was unwilling to accept the procedure for the full term of the patent until proven workable. It recommended instead that the three-year limitation be reviewed after gaining some experience with cancellation.

Recommendation XVIII: Term of the patent

The Commission unanimously recommends that:

The term of a patent shall expire twenty years after its earliest effective U.S. filing date.

Under the present law the term extends for 17 years from the date of *issue*. It is often to the applicant's advantage to delay issuance and applicants often do just this by what are called "dilatory practices": delaying responses until the last day allowed, filing continuations, amendments, etc. This may allow the art to develop and make the patent more valuable, but it buries for an unconscionably long time the knowledge the applicant should be disclosing to the public.

With the term beginning with the filing date the applicant is motivated to bring about early issuance. Also increased pressure will be placed on the Office by all applicants to secure prompt action. Since the average delay need not exceed three years, the average term of a patent should not be reduced by this recommendation.

Recommendation XXIII: In rem invalidity

A strange situation, incomprehensible if not revolting to the student of natural law, exists today in regard to the validity of patents. Under present law, even though one or more claims of a patent have been held invalid in a suit in one federal judicial circuit, the patentee may sue a different defendant in another circuit for infringement of the same claims.

"As a result, a party may be held liable as an infringer or required to pay royalties in one circuit, while his direct competitor is practicing the same invention without restriction in another circuit. Moreover, the mere possession of a patent, even though held invalid in one or more circuits, serves as a potential threat to persons unwilling or unable to defend a suit on the patent."

To eliminate this situation, the Commission recommends that:

A final federal judicial determination declaring a patent claim invalid shall be *in rem*, and the cancellation of such claim shall be indicated on all patent copies subsequently distributed by the Patent Office.

"Under the proposed recommendation, a claim, once held invalid, would be treated as cancelled from the patent. No one thereafter could be required, on the basis of a royalty agreement previously made part of an infringement judgment, to continue royalty payments on the claim. Furthermore, the proposal would preclude a subsequent suit on a patent claim previously held invalid by a Federal court.

"A patentee, having been afforded the opportunity to exhaust his remedy of appeal from a holding of invalidity, has had his 'day in court' and should not be allowed to harass others on the basis of an invalid claim. . . ."

Recommendation XXV: A 'small claims' patent court procedure

The Federal Rules of Civil Procedure today provide for a liberal discovery procedure during which litigants take depositions and respond to interrogatories of the opposing party. Often this period is protracted and literally thousands of documents are produced. It is the prodigious consumption of time in the pretrial discovery period that makes the cost of litigation so high today. Rather than face a $100 000 trial, many contestants settle out of court with little regard as to the real validity of the patent in question.

An earlier recommendation (XXIV) of the Commission provides for the creation of the office of "Civil Commissioner" in the District Courts to supervise, limit, and prevent abuse of the discovery procedures. This recommendation should have a considerable effect in reducing the cost of important patent cases.

Even with this provision, however, a problem remains. Many a patentee holds patents where the license value is $10 000 or less. Presently, and even with commissioners riding herd on discovery, the cost of a trial would exceed the damages recoverable. Thus many good valid patents are simply defied. The Commission therefore recommends a simplified procedure for these small cases.

A party to a patent case seeking to reduce his litigation costs, with the consent of the adverse party, may submit his case to the court on a stipulation of facts or on affidavits without the usual pretrial discovery. This procedure may be used where no injunctive relief is asked and only limited damages are sought. Incentives shall be provided to consent to this procedure as set forth below.

Damages might be limited to $100 000 or, with the consent of both parties beforehand, this ceiling could be waived. As an inducement to accept this abbreviated procedure, the infringer, if he refuses, might be required to pay all litigation costs including patentee's attorney fees in a regular court procedure if the latter should win.

Recommendation XXVI: A statutory Advisory Council

It was the Commission's unanimous conclusion, after its 18 months of study, that the Patent Office would be better able to serve the nation if its operations and the statutes under which it operates were subject to external review by a representative body on a continuing basis, rather than on a sporadic basis. Accordingly, the Commission makes this strong recommendation:

A statutory Advisory Council, comprised of public members selected to represent the principal areas served by the patent system, and appointed by the Secretary of Commerce, shall be established to advise him, on a continuing basis, of its evaluation of the current health of the patent system, and specifically, of the quality of patents being issued and the effectiveness of any internal patent quality control program then in operation, and whether an optional deferred examination system should be instituted or terminated.

Every fourth year the Council shall report publicly to the President and the Congress on the con-

dition of the patent system together with recommendations for its improvement.

The membership shall consist of not less than twelve nor more than twenty-four. The term of appointment shall be four years, with a maximum tenure of eight years. An executive director, and other support as deemed necessary, shall be provided.

The object here is not just to form another Commission, nor (perish the thought) to endow the present Commission with immortality, but rather simply to let the governed have a greater voice in what governs them. A growing evolving patent system responsive to changing conditions and needs would certainly be better than one belatedly overhauled at infrequent intervals.

Recommendation XXIX: Modernized searching

The Patent Office today uses mechanized searching for prior art only in the chemical field, particularly in one area—steroid chemistry. Significantly, in this area, half of the citations of prior art come from publications other than patents, the highest figure of any area; also significantly, the rejection rate on applications is highest in this area.

Clearly, mechanized searching is the wave of the future. That the Patent Office does not use it more widely today is not entirely the fault of the Office. Only the chemical engineers have developed appropriate key words and descriptions for their subject matter. We electrical engineers have failed to do so.

The Commission feels that the time is ripe for a joint engineering and scientific effort to mechanize prior art information retrieval and indeed *all* scientific and engineering data retrieval. It feels that the Patent Office has a crucial interest in how this is done. Therefore:

A study group comprising members from industry, technical societies and government should be established to make a comprehensive study of the application of new technology to Patent Office operations and to aid in developing and implementing the specific recommendations which follow.

1. The United States, with other interested countries, should strive towards the establishment of a unified system of patent classification which would expedite and improve its retrieval of prior art.

The United States should expand its present reclassification efforts.

2. The Patent Office should be encouraged and given resources to continue and to intensify its efforts toward the goal of a fully mechanized search system.

3. The Patent Office should acquire and store machine-readable scientific and technical information as it becomes available.

The Patent Office should encourage voluntary submission by patent applications of copies of their applications in machine-readable form.

4. The Patent Office should investigate the desirability of obtaining the services of outside technical organizations for specific, short-term classification and mechanized search projects.

Recommendation XXXV: Toward an international patent

Scientists and engineers respect one another's work throughout the world. Yet patent rights must be obtained nation by nation. Often the laws differ widely. One effect of adopting the Commission's recommendations would be to bring U.S. practice closer to the "center of gravity" of the practices of other nations. This would be a first step toward a larger goal.

Presently millions of dollars and hundreds of thousands of man-hours are wasted each year searching and re-searching, examining and re-examining patents that have already been granted in other countries. If we can agree on a common patent law, if we can adopt the same criteria of patentability, if we can use a common method of search with a data base that is international, then we can agree to accept each other's patents on face value. A patent granted in the United States or in any treaty country would be valid in all the other treaty countries.

This is no idle dream, and the Commission's final recommendation supports such a course of action.

The Commission believes that the ultimate goal in the protection of inventions should be the establishment of a universal patent, respected throughout the world, issued in the light of, and inventive over, all of the prior art of the world, and obtained quickly and inexpensively on a single application, but only in return for a genuine contribution to the progress of the useful arts.

Conclusion

The final report has been submitted for executive review to Secretary of Commerce John T. Connor, Office of Science and Technology Director Donald F. Hornig, and Acting Attorney General Ramsey Clark, who will make their recommendations regarding further action to President Johnson. In the event of approval, a bill will be drafted for submittal to Congress. Readers wishing to make comments at this time on the report or any of its recommendations may write to this executive review board.

It would be remiss of me not to conclude by proclaiming my profound respect for this Commission, for its members individually, and for the work it has done. Although selected to have all parishes represented, the membership was far from parochial in its approach. Without exception the members submerged self-interest for the greater good. I can only hope that this attitude will be evident to all who read its final report, and that its recommendations will be received in this light by technical people everywhere and by the legislators who may be asked to convert its recommendations into law.

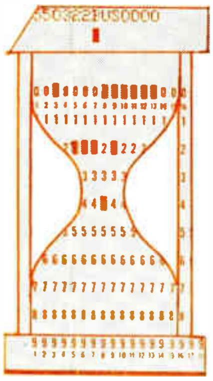
The President's Commission on the Patent System includes the following members: *Public Members*—Harry Huntt Ransom, Simon H. Rifkind (cochairmen), John Bardeen, James W. Birkenstock, Howard W. Clement, Howard K. Nason, Sidney Neuman, Bernard M. Oliver, Horton Guyford Stever, and Charles B. Thornton. *Government Members*—Secretary of Commerce John T. Connor (represented by his designee, Edward J. Brenner); Secretary of Defense Robert S. McNamara (designee John M. Malloy); Small Business Administrator Bernard L. Boutin (designee Eugene J. Davidson); and National Science Foundation Director Leland J. Haworth (designee Charles F. Brown). *Official Observers*—Secretary of State Dean Rusk (designee Eugene M. Braderman) and Office of Science and Technology Director Donald F. Hornig (designee David Z. Beckler).

Wave-mechanical uncertainty and speed limitations

In this era of microtechnology, it is essential to realize that the disturbances caused by the measurement process cannot be merely minimized and then neglected; rather, they become an integral part of the phenomenon being investigated

Panos A. Ligomenides *University of California, Los Angeles*

Modern science is eliminating the old, unrealistic approach to experimental errors. The Heisenberg principle of uncertainty has effected a fundamental change in the scientific attitude toward experimental errors and toward the degree of exactness in stating physical laws. Quantum physics has brought about the realization that a measurement or a mere observation at the quantum level perturbs the observed system in such a way that corrections cannot be made for the "loading" effect of the measuring instrument. There is something more basic and fundamental to our limitations in measurements than the lack of perfection in experimental planning and equipment. This article examines the types of uncertainty in measurements or observations. The interest then focuses on the fundamental importance of the relation between experimental uncertainty in measuring, observing, or in any way identifying a physical quantity, and the energy required to accomplish this in a certain allotted interval of time. This last relation has a direct bearing on the speed limitations of measurements.



An important factor that limits the ultimate speed of digital computers arises in engineering design because of the finite time that it takes for signals to travel from one point in the computer to another. It takes about 3 nanoseconds for light to travel one meter, and therefore at least that much time should be allowed for a computer signal to travel the same distance.

With the advent of very fast switching devices, this "time-of-flight" limitation becomes of engineering importance. If switching devices with gigahertz rates are to be used efficiently in the design of ultrafast computers, the time of flight required for communications between devices must be reduced accordingly. Under such circumstances, transmission times over long wires become significant, and microminiaturization seems to be the obvious answer.

Semiconductor, magnetic, or cryoelectric microtechnology will undoubtedly raise the ultimate upper limit for speed by several orders of magnitude over that of present-day digital computers, even when the latter employ zero-time switching components. It may well be that the time-of-flight restriction will always be the ultimate limitation on the speed of digital computers, even at very high component packing densities. However, there also exists another limitation on the signal (information) processing speed—a limitation that is derived from the field of wave mechanics.

Information processing in computers involves experiments such as the "sensing" of stored binary information. A certain magnitude of a released signal must be measured within a small allotted time interval. Uncertainties in the exact time of occurrence of a signal and in the identification of it, such as by measuring the amount of energy that it carries, may become of enough comparative magnitude to limit the speed of operation.

Depending on the accomplishments of future microtechnologies, it may be that the time-of-flight limitation will always precede the wave-mechanical one by several orders of magnitude. In any event, we consider it profitable to discuss briefly the wave-mechanical situation, because of its various wide implications. With the advent of microsignal techniques and the possible engineering use of new phenomena from quantum physics in information processing, inevitable experimental uncertainties should be studied, understood, and become part of engineering education.

The uncertainty principle

During measurements or observations at the atomic level we can never overlook the disturbance caused by the introduction of the measuring apparatus. The observer and his apparatus become an integral part of the phenomenon under investigation. The viewpoint of classical physics, that the world goes on its deterministic path independent of the observer, is an incorrect one. The observer interferes with the process of the world he observes and his interference always leaves something to be known, because it cannot be accurately measured, predicted, or even post-calculated. This margin of indeterminacy becomes more pronounced during microphysical observations, where the unavoidable interaction leads to a sizable disturbance of the observed phenomenon.

The essence of the foregoing observations is embodied in the statement of the uncertainty principle, which was first formulated by Werner Heisenberg in 1927.[1,2] He stated that an indeterminacy is always found to exist in the values of certain observed or measured mechanical magnitudes associated with an elementary particle. Since only those magnitudes that can be observed or measured, directly or indirectly, have physical significance, this principle is a fundamental feature of wave mechanics.

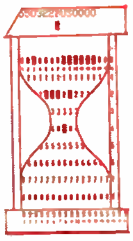
Expressed for position and momentum the principle may be stated as:

$$\begin{array}{c}\text{Uncertainty}\\ \text{of position}\end{array} \times \begin{array}{c}\text{Uncertainty}\\ \text{of momentum}\end{array} = \frac{h}{2\pi} \qquad (1)$$

where h is Planck's constant, an extremely small quantity equal to 6.625×10^{-34} joule-second. A look at this relation shows that, for a particle weighing as much as one milligram, its position and its velocity may be simultaneously determined within one trillionth in units of centimeter and second. Thus, in the measurement of position and velocity of a falling stone, for example, insignificant indeterminacies due to the uncertainty principle are lost in common experimental errors and are undetectable. However, for an electron (mass $= 0.91 \times 10^{-27}$ gram), the uncertainties of simultaneously determining position and velocity may become of the order of 1 cm and 1 cm/s, respectively.

Mathematically, the principle means that if the wave function, which is found as a result of solving the Schrödinger equation, is made more compact by simply compressing along the coordinate axes x, y, z, then the derivatives with respect to these coordinate axes (that is, $\dot{x}$, $\dot{y}$, $\dot{z}$) will be greater and the spreading velocities of the wave will become larger. In other words, smaller wave packets will spread more rapidly.

Further developments in quantum theory have shown that, besides position and momentum, relations exist for other pairs of quantities as well. Such quantities, which cannot be measured in pairs simultaneously to any desired degree of accuracy, are said to be conjugate quantities. Time and energy represent another example:



$$\begin{array}{c}\text{Uncertainty}\\ \text{of time}\end{array} \times \begin{array}{c}\text{Uncertainty}\\ \text{of energy}\end{array} = \frac{h}{2\pi} \qquad (2)$$

Accuracy limitations in observations

During the execution of very fast measurements, as is the case in high-speed electronic computers, the uncertainty principle may be manifested as an ultimate limiting factor on speed. When extremely sensitive measuring techniques and microminiaturization have progressed sufficiently, it is conceivable that wave-mechanical uncertainty will interfere with high-speed measurements on the state of physical systems and signals. At high packaging densities, a $\Delta E \cdot \Delta t$ indeterminacy may impose restrictions on the minimum allotted time for measurements and unbearable demands on power requirements for a given margin of measurement accuracy.

The foregoing proposition will now be demonstrated with an idealized computer measurement problem. Consider the experiment of "sensing" the state of a binary memory cell. A certain amount of energy is released by the cell in the form of a signal, such as an electric current pulse or a single photon, and the task is to "identify" the signal within a certain allotted time interval. Identification may consist of measuring the energy content of the signal.

Consider an ideal box lined with perfect mirrors, which could hold in radiant energy indefinitely (ideal memory cell).* The box is weighed at all times. An ideal spring scale equipped with a pointer recording the weight on a vertical column placed alongside is as good as any. At a chosen instant of time, an ideal shutter, preset like a time bomb, will open to release just one photon. The released photon is the "sense" signal and its energy (frequency) must be measured to provide the desired bit of information. For that, the box is weighed again. The change in mass gives the energy of the emitted photon.

The changing of the mass of the box results in a vertical motion of the box, which will have to be measured. (It may be observed that, in general, most measurements resolve themselves into either distance or time.) Velocity and distance measurements will result in an uncertainty in the height of the box (or of the pointer) above the experimental bench. Also, the uncertainty about its elevation above the earth's surface leads to an uncertainty in the rate of the clock. The clock's rate is affected by the clock's relative position according to the theory of relativity. It can be shown[3] that these uncertainties are consistent with Eqs. (1) and (2).

The complementary indeterminacy that is always present in measurements becomes particularly evident when the level of observation is reduced to subatomic scales. However, when the time allotted for an observation becomes extremely short, comparable to the size of indeterminacy Δt, then speed limitations and excessive power requirements may result. An observation yields a certain amount of information about the observed physical system at the expense of some amount of energy. The availability and the consumption of the required amount of power then may become a serious problem, when the time allotted for the observation becomes extremely short. This point is to be briefly investigated next.

Information theory argues that as long as we have not observed or measured anything about a physical system, any elementary "complexion" of the system is possible. This maximizes the number of possible complexions for the system, and thus its entropy is large and the indeterminacy of its state is maximum. By measurement or observation, the ambiguity about the state of the system is narrowed, the number of possible complexions is reduced, the entropy of the observed system is reduced, and the entropy of the measuring apparatus is increased. This is accomplished by a transformation of a certain amount of energy from an "available" form to an "unavailable" form. The speed of observation is limited by our ability to deliver the amount of energy transformed in the time allotted for the observation. The problem, again, does not become appreciably felt before very high speeds are reached in the measurements. If the power transformed (to heat) during the measurement is reduced, the infor-

*This is a Gedanken, or "thought," experiment, a device often used in theoretical physics. Idealized conditions are admissible as long as there is no contradiction with the accepted laws of physics. The above example is adapted to a computer problem from two historic thought experiments. In 1930, the Sixth International Solvay Congress on Physics took place in Brussels. Niels Bohr, an advocate of the uncertainty principle and the man who developed it into a new philosophy of physics, was present at a meeting when Albert Einstein performed a thought experiment in order to disprove the uncertainty principle. Bohr retaliated with a counterthought experiment to disperse Einstein's arguments.[3]

mation gained by the experiment is also reduced and the uncertainty of the measured value of the physical quantity increases.

Information gained in an experimental observation

An observation made on a physical system provides the observer with a certain amount of information about the system. The amount of information received is limited by any uncertainties above the values of the measured parameters. It follows then from Heisenberg's principle that the amount of information that we are allowed to obtain in a measurement is limited. This generalization was worked out by L. Brillouin,[4] who established the "negentropy principle." Further discussion on the essential ideas and implications on engineering work requires a prior quantitative definition of information.

If P_0 is the number of all possible and equally probable elementary complexions* of a system or of a measurable physical quantity *before* the measurement and P_1 is such a number *after* the measurement, then the amount ΔI of information that is gained by the measurement is defined by the relation[5]:

$$\Delta I = K \ln \frac{P_0}{P_1} \qquad (3)$$

Notice that ΔI is always positive. Also, the accuracy of a measurement (or observation) is high, and thus ΔI is large, if P_1 becomes small. The factor K is a constant depending on units. In accordance with this relation, the "information content" of a measurable system or physical quantity is uniquely determined by the measurement; that is, $P_1 = 1$.

Often in computer problems, the amount of information is measured in units of binary digits (bits). If there are n independent binary variables in an observed or measured physical system, each of them corresponding to an equally probable binary choice (0 or 1), the number of the equally probable complexions (states, outcomes) of the system is

$$P_0 = 2^n$$

If we set

$$K = \frac{1}{\ln 2} = \log_2 e$$

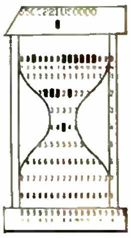
the information content of the system—that is, the amount of all possible information to be gained by an experiment on the system—is

$$\Delta I = K \ln P_0 = \log_2 P_0 = \log_2 P_0 = n \text{ bits} \qquad (4)$$

In terms of such units, we need one bit of information for each binary variable in order to know completely the state of the system.

For example, a system of three binary ferrite cores (part of a digital computer's memory) has a number of possible and equally probable complexions (states) equal to $P_0 = 2^3 = 8$. Its information content is

$$\Delta I = \log_2 P_0 = 3 \text{ bits}$$



If the state of one of the cores is measured and identified, the number of the remaining unknown possible states is four, and the information received by the measurement is

$$\Delta I = \log_2 \frac{8}{4} = 1 \text{ bit}$$

An additive property is inherently implied by the definition of information (3). If $P_0 = P_{01} \cdot P_{02}$ for a combination of two systems, the overall information content is $\Delta I = \Delta I_1 + \Delta I_2$.

Information cost and speed limitations

In many practical experiments the cost of obtained information during the experiment is of prime concern. For such cases the quantity of information received is related to the inevitable increase in the entropy of the measuring apparatus during the experiment. Entropy units are also used to measure information quantities as well as entropy variations. You may have already noticed the similarity between the relation (3), defining the quantity of information, and the statistical definition of entropy by letting $K = k$, where $k = 1.38 \times 10^{-23}$ J/°K is the Boltzmann constant.

$$\Delta I = k \ln \frac{P_0}{P_1}$$

or

$$\Delta I = k \ln 2 \log_2 \frac{P_0}{P_1} \qquad (5)$$

where P_0 and P_1 again represent the numbers of elementary complexions of the physical system before and after the experiment. (This relation indicates that the smallest amount of information which may be gained from an experiment is, in entropy units, equal to $k \ln 2 \approx 10^{-23}$ J/°K. This amount represents one bit in binary digit units.)

The decrease in the entropy of the measured system is given by

$$\Delta S = k \ln \frac{P_0}{P_1} = \Delta I \qquad (6)$$

A physical system with a great number of elementary particles will have an enormous but finite variety of complexions, depending on the positions, velocities, and quantum states of the particles. A computer memory consisting of n binary ferrite cores will have a number $P_0 = 2^n$ of possible states (complexions). The entropy content of such systems, based on statistical considerations, is given by

$$S = k \ln P_0$$

We may recall here that the absolute values of S are irrelevant. Only those changes ΔS in the entropy of a system that occur during physical processes are meaningful. For example, during a crystallization process, the number of possible complexions P_0 of the liquid phase is reduced to that of the crystal P_1, and the system becomes more ordered in the microscopic level by releasing an amount of energy Q. Its entropy is also reduced by an amount given by the relation

*The term "elementary complexion" is used here with a generalized meaning. In the sense of Planck's definition, it represents each discrete quantum configuration relating to a physical system in the subatomic level (in this context it is also used in the definition of entropy). It also means each discrete possible state of a system (physical or mathematic) of binary variables. It may also mean each possible value of a parameter (such as the potential difference between two nodes of an electric circuit or device).

$$\Delta S = k \ln \frac{P_0}{P_1}$$

A positive amount of information is gained about a physical system during an experiment. It is accompanied by an increase in the entropy of the measuring apparatus. This statement sets up a practical relation between entropy and information.

Gain of information ΔI about, and corresponding decrease in the entropy of, an observed system go together. They are also accompanied by an increase ΔS_a in the entropy of the measuring apparatus (environment). The inevitable increase in the entropy of the measuring apparatus is greater than the corresponding amount of gained information ($\Delta S_a > \Delta I$). Thus, each irreversible process of observation or measuring results in a net increase in the net entropy of the universe. This is true for any irreversible process, such as in a measurement on an observation. Accordingly, we may write:

$$\Delta S_a \geqq \Delta I \quad \text{or} \quad \Delta(S_a - I) \geqq 0 \tag{7}$$

where S_a refers to the measuring apparatus. The equal sign holds only under reversible (ideal) conditions.

Relation (7) implies that a finite amount of energy δE (proportional to ΔS_a) must be degraded (changed into heat) in the measuring apparatus during the experiment, as follows:

$$\delta E = T \Delta S_a = T \Delta I \tag{8}$$

where $\Delta I = k \ln P_0/P_1$.

A point of distinction should be made here. If a physical system of n binary variables is measured so that its state is completely known (that is, if the individual state of each of its binary variables has been measured), then the amount of information gained is finite and equal to the information content of the system ($\Delta I = n$ bits). The amount of energy required for the measurement is also finite. If the measurement is made in nonzero time, then the amount of power required is also finite. Similar is the situation with a physical system consisting of individual elementary particles. A finite number of possible complexions (quantum configurations) exists for the system (such as in an atom) and the identification of the exact state of the system (one quantum configuration out of all possible ones) involves a finite gain of information equal again to the information content of the system ($\Delta I = k \ln P_0$).

If the measurement is of a continuous physical quantity, such as a potential difference, with a range of possible amplitude values from 0 to 1 volt, then absolute accuracy in the measurement of its amplitude (zero error) would mean the expenditure of an infinite amount of energy. On the other hand, if an error of, say, 1 millivolt is allowed in the measurement of the amplitude (that is, if the range of 1 millivolt still remains unresolved after the measurement), then a finite amount of energy expenditure is required for determining the amplitude. The power is also finite if the time allotted for the measurement is nonzero. (At least $\delta E = T\Delta I = Tk \ln 1000$ is required, under ideal reversible experimental conditions.)

Cost of experimental accuracy and of high speed

The uncertainties that accompany every physical measurement, caused by imperfect apparatus, human intervention, computational errors, accidents, noise, etc., have been well analyzed in scientific literature. The problem of inherent uncertainties, which are unavoidably present in all physical measurements and observations because of the interference of measuring apparatus with the measured physical system and because of the energy cost of information gained, were also discussed in this article.

The wave-mechanical restrictions on the ultimate speed of an experiment come about because of the time uncertainty Δt, which, at extremely high speeds, may become comparable to the time allotted for the measurement. Unbearably high power demands also develop for a specified degree of measurement accuracy.

Information theory views a physical measurement as an information-seeking experiment. The amount of information gained in the experiment was quantitatively defined in (3) and was related to the entropy changes that occur during the measurement. A measurement is an irreversible process and subsequently results in a larger increase in the entropy of the measuring apparatus, as compared with the magnitude of the decrease in the entropy of the system measured and thus with the amount of the information gained; that is, $\Delta S_a > \Delta I$. (The statistical denominator of these definitions must be well understood.) Thus, the information obtained in a measurement is paid for by an increase in the entropy of the measuring apparatus. The entropy increase in the apparatus is in turn related to an amount of energy δE, which is degraded during the experiment from an "available" form to an "unavailable" form. It should be noted that the smallest possible change in entropy required for our observation (obtaining just one bit of information, a "yes" or "no" to a specific question) is, from (6), equal to $k \ln 2$. Considerations of the amount of power $\delta E/t$ required to be degraded during an experiment in relation to the accuracy (amount of information requested) and speed of the measurement enter at this point.

The implications of the foregoing facts to the designer of modern high-speed electronic computers may soon reach beyond the realm of academic fascination. As the sizes of switching components and their signals are reduced to almost microscopic dimensions, and as the speed of signal manipulation increases to fantastic rates, the task of measuring and identifying the state of a device, a circuit, or a signal, presents us with a very serious problem of reliability and cost of information. A trade-off exists between information gained and power spent in an experiment. Detection and identification of the state of a signal in zero time requires that an infinite amount of power be degraded to a lower, unavailable form of energy. It all seems to point to the conclusion that the efficiency and success of a measurement should be calculated in bits per second per dollar. Also, a complete physical theory must recognize that experimental errors cannot be made as small as one may desire, because they are part of the actual facts of life.

REFERENCES

1. Heisenberg, W., "The actual content of quantum theoretical kinematics and mechanics," *Z. Physik*, vol. 43, pp. 172–198, 1927.

2. Bohm, D., *Quantum Theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1951, pp. 100–101.

3. Gamow, G., "The principle of uncertainty," *Sci. Am.*, vol. 198, p. 54, Jan. 1958.

4. Brillouin, L., *Science and Information Theory*, 2nd ed. New York: Academic Press, 1962.

5. Shannon, C. E., "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, July 1948.

EHV transmission in the U.S.S.R. power grid

In this article, technological information on progress in present-day and future power generation and transmission development in the Soviet Union are freely presented—including a frank discussion of the "bugs" encountered in EHV transmission and some of the difficulties in achieving an All-Union Power Grid

B. P. Lebedev
Committee for the U.S.S.R. Participation in International Power Conferences
S. S. Rokotian
Power Systems Research and Design Institute

In 1965, 507 billion kWh of electricity was produced in the Soviet Union, and the total generating capacity of its power stations reached 114 000 MW. The total length of transmission lines, with voltages of 35 kV and higher, ran to 312 000 km. The capacity of the European power interconnections of the U.S.S.R. exceeded 68 000 MW in March 1966. The control of the power industry is centralized under the Ministry of Power and Electrification of the U.S.S.R. Ninety-four percent of its electric energy is generated at power stations that are part of interconnected systems. Thermal stations predominate; their generating capacity accounts for 77 percent of the total installed generation. Hydroelectric plants represent 22 percent of the total capacity and 18 percent of the total output. About one percent of the total power production is from nuclear plants. Thirty-six percent of the thermal plants are equipped to transmit heat energy as well as electric power.

The most significant trends in the development of the Soviet power industry in recent years are the construction of huge power stations, the marked increase in the size of generating units, and the use of EHV in transmission systems. Typical capacities of the new, large thermal stations are 1800, 2400, and 3600 MW. At present, 12 thermal power plants, with individual generating capacities of 1000 MW and higher, are in operation in the Urals, Donbass, and Siberia. In the 1950s, by contrast, the typical capacity of installed turbogenerator units ranged from 100 to 150 MW.

In 1965, for example, there were 115 generating units in operation with unit capacities of the order of 150–300 MW (including fifteen 300-MW units), for a total capacity of 22 000 MW. Work is presently under way for the installation of the first 500-MW single-shaft (tandem) generator in the Nazarovskaya State district power station, and an 800-MW cross-compound generator at the Slavianskaya State district power plant.

There are also several large hydroelectric power stations, with capacities of 1000 MW and higher, along the principal rivers of European Russia and Siberia: the original Volzhskaya (2300 MW) and the new Volzhskaya (2500 MW) on the Volga; Votkinskaya (1000 MW); and Bratskaya (3800 MW) at Bratsk. Other hydroelectric stations under construction are the Nizhne-Kamskaya (1080 MW), on the Lower Kama; Nurekskaya (2700 MW), on the River Nurek; Krasnoyarskaya (5000 MW); Ust-Ilimskaya (4300 MW), and others. At the Krasnoyarskaya station, 500-MW units are being installed, two of which will be in operation this year.

The main voltage used for transmission trunk lines, for the transfer of power from large generating stations, and for the interconnection of power systems and inter-system ties, is 500-kV alternating current.

During the last three years in the Soviet Union, there has been a steady increase in the generating capacity of power plants of the order of 10 to 11 million kW annually. In 1965, electric energy production was 11 percent greater than that of 1964. By 1970, the rate of commissioning of new generating facilities will increase power production by 15 000 – 18 000 MW annually.

Interconnections in the U.S.S.R.

Electric power stations are interconnected to form a total of 90 district power systems, which control the administrative, economical, and operational aspects of the generating stations that operate within each district.

Interconnected system structure. District power systems, in turn, are integral components of area dispatching centers which form ten regional, interconnected power systems: Center, Lower Volga, Middle Volga, Ural, South, North–West, North of the Caucasus, Transcaucasus, Middle Asia, and Central Siberia. Very soon, the interconnected power systems of North Kazakhstan, Trans-Baikal areas, and the Far East will be formed. Several power systems operate outside the framework of the interconnected systems in the areas of the Extreme North, Kamchatka, and Sakhalin. All area dispatching centers are under the direct control of the Ministry of Power and Electrification of the U.S.S.R.

The development of interconnected power systems in European Russia has led to the creation of more efficient interconnections such as the European Power Grid, which includes the interconnected systems of the Center, North–West, South, Ural, Volga Basin, and Caucasus areas. This power grid has a primary dispatching center that controls the area dispatching centers. The generating capacity of the European Power Grid reached 68 500 MW on January 1, 1966.

Transmission and distribution lines. By the beginning of 1966, the circuit length of EHV transmission lines (330 kV and higher) in the U.S.S.R. was about 15 300 km (see Table I).

During the past five years, approximately 32 000 km of transmission lines, with voltages of 35 kV and higher, were installed annually. Voltages of 35 to 150 kV are used as distribution lines, and 220-kV lines in district power systems are employed as basic system lines. But in large power systems, with a high level of interconnection, the function of the 220-kV lines tends to classify them also as distribution lines.

Transmission lines, with a voltage of 330 kV and higher, are operated for the long-distance transmission of large blocks of electric power and are also used as basic system lines in interconnected power systems. The 330-kV lines are installed in two interconnected systems, the South and North–West, and partially in the Caucasus area. In other interconnected power systems, 500-kV transmission lines are employed.

The first 500-kV transmission line (the first circuit from Volgograd to Moscow), about 1000 km in length, was commisssioned in 1959. In 1963–1964, the double-circuit transmission line (Kuibyshev–Moscow), and the single-circuit line (Kuibyshev–Ural), which had been operated at 400 kV, were converted to 500 kV.

I. Length of transmission lines in the U.S.S.R.

Voltage, kV	Length, thousand km	Percentage of Total Transmission
800*	0.5	2.8
500	8.0	
400	0.2	
330	6.6	2.1
220	34.2	11.0
110–150	129.3	41.5
35	132.5	42.6

* DC transmission line Volgograd–Donbass (in planning stage).

Of a total of twenty-five 500-kV substations in operation, six are step-up substations at power plants, 17 are receiving step-down substations, and two are switching stations. In 1966, 13 substations, and about 2000 km of 500-kV transmission lines, were under construction (see Fig. 1).

Development of 500-kV transmission. During the initial stages of development of the 500-kV transmission lines, individual lines, about 1000 km long, were built with a transmitting capacity up to 1000 MW per circuit. Now the tendency is to develop 500-kV networks that extend over very large areas. Separate sections of such networks, between adjacent substations, are relatively short (300 km and less). Although at the initial stages, 500-kV transmission lines were used mainly for the transmission of power generated at the big hydroelectric stations, these lines are now also used for the exchange of power between interconnected systems, both on a daily and a seasonal basis.

A similar developmental process of network forming took place in relation to the 330-kV lines within the interconnected systems of the South and North–West.

Conductors, current densities, and shunt reactors. Most of the 500-kV EHV lines in the U.S.S.R. have three ACSR conductors per phase, and the cross-sectional area of the aluminum part of each conductor is 500 mm^2. Some lines that carry a lower load have three 400-mm^2 conductors per phase. The conductors in a phase are arranged at the apexes of an equilateral triangle with 400-mm-long sides. The 330-kV transmission lines have two conductors per phase, and the cross section of the aluminum part of each conductor is either 300 or 400 mm^2.

The economic current density in European Russia is within 0.6 to 0.8 A/mm^2 for the aluminum part of the conductor, but in Asiatic Russia, where the cost of power is lower, the economic current density value rises to 0.8–1.0 A/mm^2.

The 500-kV double-circuit transmission lines are designed for intertied schemes, with 500-kV busbars at step-up and receiving substations, and with transverse sectionalizing at intermediate substations and switching stations. Such schemes provide high reliability during outages of separate sections of the transmission system, operational flexibility, and high transmitting capacity in relation to the transient stability requirements. Short, 500-kV radial and single-circuit loop lines are normally backed by high-power double-circuit lines.

Shunt reactors, which are a means of reducing reactive power flows and power losses at low-load operating conditions, are obligatory on 500-kV lines. The specific capacity for a 1000-km-long 500-kV transmission line is 0.7–0.9 kilovar per kilowatt of active power transmitted.

For voltage leveling at terminal and intermediate substations, for the reduction of power losses in the transmission, and to reduce the internal overvoltages, special attention is given to the location of shunt reactors along the line. Up to 70 percent of this equipment is installed on the 500-kV side. Of the total number of shunt reactors, about one third are connected directly (without circuit breakers) at the transmitting end of the line, and two thirds of them are connected through circuit breakers at intermediate substations.

At high-load conditions, a portion of the reactors at intermediate substations is switched off to increase the transmitting capacity of the lines and to reduce power losses.

Shunt reactors are not used at the receiving end of long-distance transmission lines, since the reactive power generated by part of the 500-kV line, adjoining the receiving power system, is used within it, and this reduces the capacity of the synchronous capacitors installed at the receiving substations.

The high transmitting capacity of long-distance 500-kV transmission lines (750 to 1000 MW per circuit for distances up to 1000 km) has been achieved by such measures as the use of bundle conductors, the construction of switching stations (or intermediate substations), the use of high-speed protective equipment and circuit breakers, shunt compensation by the use of shunt reactors, series compensation by the use of capacitors, and the provision of various automatic devices on the lines.

Protective devices for the 500-kV lines

Among the devices of special importance used to improve the stability of long-distance transmission are the "strong action" regulators and counter-disturbance automatic system devices. The strong action field regulation of hydro generators at the stations that supply power over long-distance transmission lines accounts for the increase in the steady-state stability to the limit assumed in accordance with the requirements of constant voltage either at the source of the line or at an electrically closed point. The increase in the steady-state stability limit, achieved in this way, is estimated to be from 8 to 10 percent. This effect was achieved by current regulation and frequency regulation when signals proportional to current derivatives are replaced by signals proportional to the frequency deviation and its derivative.

Considerable effect has also been obtained by the use of strong actuation field regulation of big synchronous capacitors at receiving substations. This has reduced the equivalent reactance of the receiving power system.

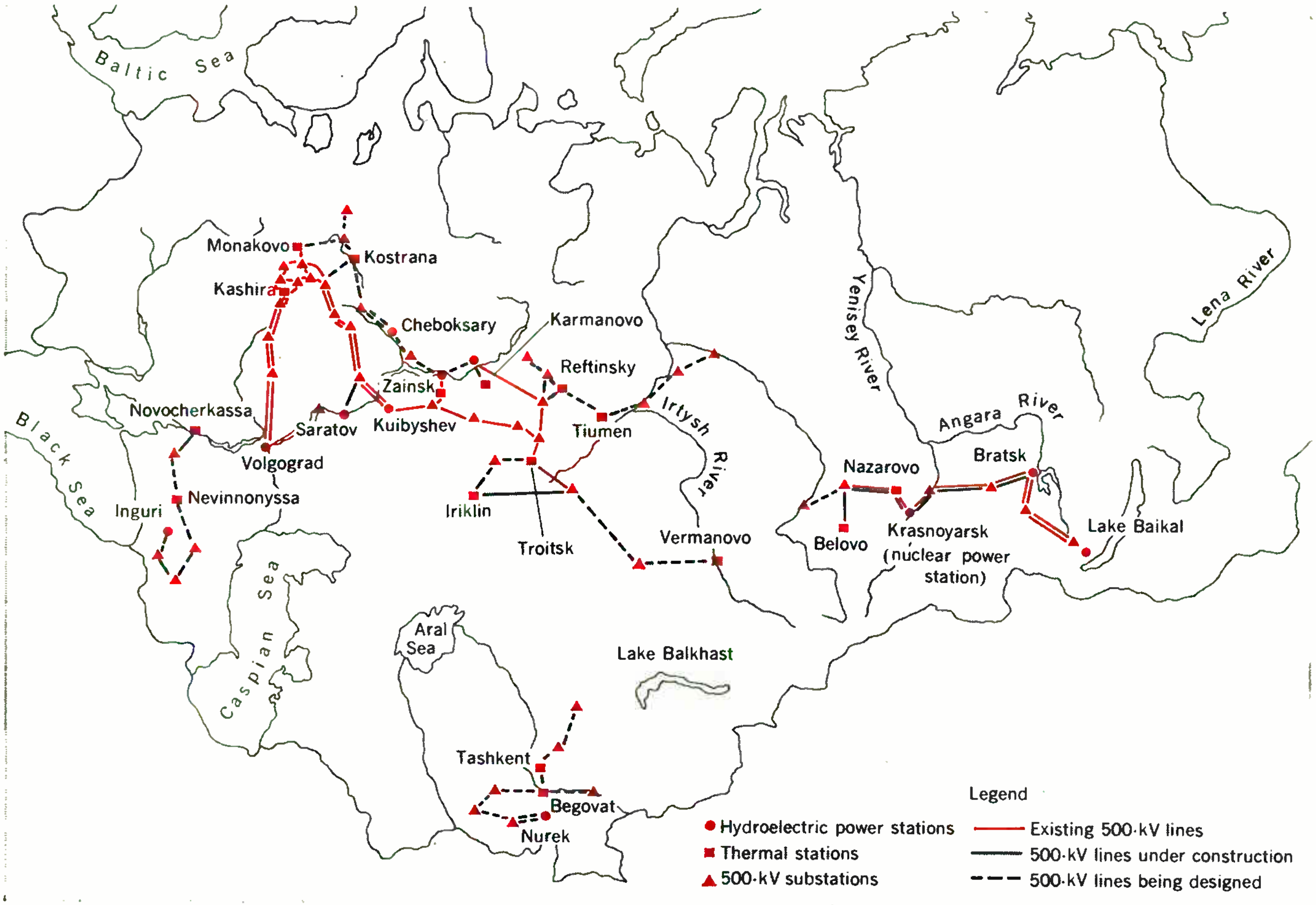
Strong action field regulators and fast electronic exciters have provided considerable improvement in transient stability, rapid damping of synchronous power swings, and the limiting of dangerous voltage variations during disturbances in normal operating conditions.

The introduction of special system automatic devices dates back to the mid-1950s, when the first 400-kV (Kuibyshev–Moscow) line was completed. Ten years of operational experience on this line proved the high efficiency of these devices and it created a widespread demand for this equipment.

Functions of automatic devices. At present, 500-kV transmission lines in the Soviet Union are equipped with automatic devices to

1. Prevent steady-state and transient stability disturbances.

FIGURE 1. Existing and proposed EHV transmission lines, thermal generating stations, hydro plants, and 500-kV substations in European and Central Asiatic Russia.



2. Prevent the development of emergency conditions that involve disturbances in synchronism.
3. Eliminate dangerous power surges into the local and intermediate power system during system faults that disrupt long-distance transmission.
4. Eliminate the hazardous steady voltage rises on the transmission equipment and the overloading of its elements.

To prevent stability disturbances in an unfaulted line, caused by critical overloading, emergency automatic load-shedding equipment is employed that both controls the power transmitted and reduces the load by disconnecting some of the generators at the hydroelectric power station. In the event the hydroelectric station supplies power into transmission lines that run in different directions, or if the hydroplant is connected to a large local power system, automatic system isolation and load-shedding equipment is activated.

When faults occur in one of the circuits of the initial section of a double-circuit transmission line, the automatic load-shedding equipment operates according to the immediate fault indications (i.e., the actuation of protective relays, opening of circuit breakers, and transfer of power to the unfaulted circuit). When faults happen on other sections of a double-circuit transmission line, or a single-circuit line with intermediate taps, automatic devices are normally used with actuating elements at intermediate points of the lines, and the order for the isolation of busbars at hydroelectric stations or for shutting down some of the generators is sent via telecommunication channels.

These automatic devices maintain both transient stability, when phase-to-phase short circuits occur (two phase and two phase to ground), and the required margin of steady-state stability (not less than 8 percent) in after-fault conditions. If stability disturbance still persists, special automatic devices, which respond to the indications of an asynchronous condition, trip out the line and isolate it from the local power system and other transmission lines, thereby limiting the influence of the asynchronism. Operating experience indicates that, in practically all cases, rapid resynchronization (15–20 seconds) is achieved.

Remote control automatic devices are used to prevent dangerous transfers of power that are caused by the disconnection of a line to the intertied low-capacity power system. Other automatic equipment is also employed for the control of active and reactive power transfers, as well as the voltage and frequency on the busbars of the electric power station, to the local power system. In addition, this equipment makes it possible to divert the power of certain generators into the local system when disturbances occur in the parameters just mentioned.

Automatic devices, used to eliminate the danger of lasting voltage rises that are caused by the unilateral opening or interruption of lines, normally control not only the voltage level of the 500-kV busbars, but also the reactive power flows in the 500-kV transmission lines. This equipment operates in steps with different values of time delay for switching on disconnected shunt reactors, for opening unilaterally connected lines, and for disconnecting autotransformers.

Another automatic method is the remote unloading of the Moscow–Kuibyshev transmission line by means of telecommunication when series capacitors are overloaded by the disconnection of a part of the banks of static capacitors that are connected in parallel. And of particular importance are the means for automatic frequency-controlled tripout and automatic reclosing with frequency response. This equipment, used in all power systems of the U.S.S.R., serves to prevent outages that are caused by sudden power losses and the accompanying dangerous drops in frequency.

Estimates of consumers' power requirements. The amount of power used by various consumers, who can be isolated by automatic frequency-control equipment, is estimated for each individual power system. This estimate should exclude the possibility of phenomena such as the frequency avalanche and the voltage avalanche under conditions that involve the disconnection of generating plants or the disruption of individual power systems into isolated segments, and the isolation of areas where load exceeds the generating capacity. Depending upon the power system operating conditions, different modifications of frequency-controlled automatic equipment are used. These include high-speed configurations, with different frequency settings, or with a common frequency setting but different time-delay settings; and special automatic arrangements for eliminating local power deficiencies in separate parts of the power grid.

Operating experience has shown that, with the development of interconnected systems and the establishment of 500-kV networks, the problems of prevention or fast elimination of system disturbances cannot be practically solved without the assistance of automatic equipment.

Importance of transmitting capacity. Of primary importance for the development of a reliable interconnected system is the transmitting capacity of the system comprising the EHV transmission lines. In the Soviet Union, this vital aspect is considered both in the planning stage and during the operation of the network. Within the European Power Grid, the transmitting capacities of 330- and 500-kV transmission lines that intertie different interconnected power systems vary from 35 to 60 to 100 percent of the total capacity of the smaller system.

Construction details of 500-kV transmission lines

For the initial 400-kV lines, steel H-frame self-supporting suspension towers were used with horizontal arrangement of conductors. But in 1965, a new 500-kV tower configuration was designed with an H-frame structure supported by steel guy cables. Back in 1957–1958, in order to conserve structural steel, a new type of suspension tower was designed that employed reinforced concrete tubes. The design of this tower is similar to that of the steel-guyed tower.

Some 'bugs' in the design. Several instances of automatic tripout of the first 500-kV lines occurred. These were caused by ice load on the conductors, which resulted in the "galloping" of the conductors and steel cables that overlapped in the span, and the condition indicated the inadequacy of horizontal and vertical clearances between the conductors and steel wires on the towers. This situation was remedied by increasing the horizontal clearances between the conductors and ground wires from 2.4 to 4.5 meters, and the vertical clearances have been increased up to 11 meters.

During the course of operation of the 500-kV lines, certain improvements were made in the design of line equipment. In particular, the hinging capacity of suspension and strain insulator strings has been increased, and the design of clamps with limited holding strength has also been improved.

To make the fastening of bundle conductors to anchor towers more secure, each phase conductor is separately connected to the crossarm, and the connectors are hinged in both the horizontal and vertical planes.

The operational results for 500-kV lines in the U.S.S.R., obtained from 1956 to 1964, are given in Table II. To make the information more complete, the table also includes data for these lines built for 500 kV but temporarily used for transmission at 220 and 400 kV. Table III gives the physical specifications and costs for typical 330- and 500-kV transmission lines.

Lightning faults and outages. Lightning faults on 500-kV lines in the Soviet Union are very rare because

1. There are only 20–25 stormy days per year along the routes of transmission.
2. The single-circuit H-frame towers do not exceed 30 meters in height.
3. Two ground wires are used along the entire length of transmission, with protective phase angles of 20 to 30 degrees.
4. Tower to ground resistance values are low (normally less than 5 ohms).
5. There are adequate impulse insulation levels of insulator strings.

In addition to lightning, outages of 500-kV lines in Russia have occurred from the galloping phenomenon just described.

Automatic reclosure was introduced on 500-kV lines in 1961–1962.

The 750-kV EHV lines

The capacity of our interconnected systems increases from 150 to 200 percent during a five-year period. Thus the potential transmission capacity of the 330-kV network will be fully utilized in a period of 15 to 20 years, depending on the length of the grid, load densities, and the rate of load growth. And it is estimated that the potential capacity of the 500-kV network will be realized within 20 to 25 years. As these critical time periods elapse, the transmitting capacities of system-forming transmission lines will become insufficient, and the danger of system disturbances will increase. Thus the creation of an up-to-date power grid requires the superimposition of a new transmission network together with a higher transmitting capacity on the existing networks.

At present, planning is under way for a 750-kV network for the interconnected system of the South, where such EHV transmission lines could be used to carry power of the order of 2500–3000 MW from the large hydro stations, or a group of big thermal plants, over distances up to 1000 km. Such transmission schemes could be realized sometime after 1970.

The experimental line. To facilitate the construction plans for 750-kV lines, a 750-kV experimental and industrial line from Konakovo to Moscow (90 km), with

II. Outage data for 500-kV lines

Years	Total Length of Lines at the End of the Period, km	Number of Outages		
		Non-storm	Lightning	Subtotal
1956–58	2595	11	3	14
1959–60	3910	46	4	50
1961–62	7127	55	21	76
1963–64	8075	60	14	74
Total		172	42	214

III. Specifications and costs for 330- and 500-kV lines

Item	330 kV	500 kV
Effective power transmitted, MW	300–350	900–920
Costs per km of line, thousands of dollars	22–33	33–38.5
Amount of steel used for towers and foundations, metric tons/km	8–20	12–24
ACSR conductors used per km, metric tons	7–11	13–16
Specific capital investments per kW and 100 km of line, dollars	6.6–9.9	3.9–4.4
Costs of transmission of 1 kWh over 100 km, dollars	0.000 22–0.000 33	0.000 17
Specific number of outages per 100 km-years	1.0–1.5	0.6
Design internal overvoltage level (times nominal voltage)	2.7	2.5

IV. Characteristics and costs of 750-kV line

Costs per km of line, thousands of dollars	44–55
Amount of steel used for towers and foundations, metric tons/km	35–40
ACSR conductors used per km, metric tons	25–30
Specific capital investments per kW and 100 km of line, dollars/kW	2.2
Costs of transmission of 1 kWh over 100 km, dollars	0.000 044–0.000 055
Efficiency of transmission, percent	91–92
Design internal overvoltage level (times nominal voltage)	2.1
Type of towers	H-frame, guyed, steel
Tower height, meters	30
Length of crossarms, meters	3.5

a transmitting capacity of 1250 MW, has just been completed. With four conductors per phase, arranged at the corners of a square bundle configuration (60 cm per side), the effective capacity is 2200–2500 MW per circuit. Each of the four conductors has a cross-sectional area of 600 mm². It is estimated that the average annual losses, in the prevailing climatic conditions of the central part of European Russia, will be 20 to 25 kW/km.

Table IV lists some of the pertinent physical and economic characteristics for a typical 750-kV line.

Problems in establishing an interconnected power system

The process of linking the interconnected power systems of the largest economic areas of the country (European Russia, North Kazakhstan, and Central Siberia) with 500-kV lines is expected to be completed in the early 1970s. The transmission capacities of single-circuit 500-kV lines (750 to 1000 MW) would not be sufficient to serve as a basis for the All-Union Power Grid. If the lowest transmitting capacity values of the intersystem power ties (15 percent of the total capacity of the smallest of the interconnected systems) should be taken as a basis—and this is considered to be an unsatisfactory figure—then, by 1975, the minimum transmission capacity of interties between the North Kazakhstan system and the European grid to the west, as well as between North Kazakhstan and Central Siberia, should not be less than 3000 MW each. The transmission capacity of an intertie between the European grid and the Central Siberian system should be of the order of 6000–9000 MW.

Need for higher voltages. In view of the just-mentioned requirements for the interconnection of the European grid, Kazakhstan, and Central Siberia (and later, Middle Asia), it is obvious that transmission lines of more than 500 kV must be used. Therefore, R & D institutions in the U.S.S.R. are studying the possibilities of using 750- and 1000–1100-kV ac transmission lines or 1500–2000-kV dc lines for the creation of the All-Union Power Grid. A combined solution is also considered possible: dc transmission in the east–west direction, and ac transmission in the north–south direction. This alternative solution may be understood by the fact that the main flow of electric energy in Russia will be from east to west.

Fossil-fuel resources. The production of electricity in the U.S.S.R. is planned at 830–850 billion kWh by 1970, and the output for 1975 and 1980 is estimated at about 1450 and 2400 billion kWh, respectively.

About 70 percent of the fuel consumption increase in European Russia will be met by natural gas and oil. But, in this context, an economic problem may arise during the decade of the 1970s, if the fuel balance (approximately 30 percent) of European Russia includes the utilization of cheap—but not transportable—brown coal (cannel coal) from the open-cut mines in Central Siberia and North Kazakhstan, or the development of the more costly deep veins in European Russia. Thus two alternatives were studied from both the technological and economic viewpoints:

1. The construction of big thermal-electric stations in Siberia and North Kazakhstan, burning cheap, brown coal, and the transmission of energy to the Urals and European Russia by very-large-capacity EHV lines.
2. Construction of thermal stations in the Urals and European Russia, coupled with the railway transportation of coals from Siberia and an increase in the extraction of local coal from deep mines.

The comparative results of these two alternatives, by the target date of 1980, are shown in Table V. From this table it is apparent that the advantages of the first alternative would include a smaller capital investment and far less annual operational expenditures.

Long-distance transmission techniques. Conventional ac transmission techniques cannot be applied for very-long-distance transmission and hence compensated, or resonance, transmission should be used. This requires additional equipment, however, such as shunt reactors, banks of static capacitors, synchronous condensers—and a considerable increase in total costs.

High-voltage dc transmission lines—in principle—have no stability limitations, but they must also have lower internal overvoltages. There are good prospects for the construction of long-distance (2000 to 4000 km) dc transmission lines with a transmitting capacity of 10 000 to 20 000 MW per circuit.

A comparison of the design and economic characteristics of dc and ac transmission for two long-distance transmission lines—Krasnoyarsk–Urals (2000 km), and North Kazakhstan–Tambov (2500 km)—shows the economic advantages of EHV dc transmission (see Tables VI and VII, and Fig. 2).

Some design characteristics. The design work to be undertaken for the transmission line from Kazakhstan to Tambov is based on using ±750-kV direct current, with the following design and economic characteristics:

1. Power at the receiving end—5250 MW.

V. Comparative data of alternative power schemes

Alternative	At the Receiving End		Installed Capacity of Thermal-Electric Stations, million kW	Total Capital Invest-ments,* billions of dollars	Annual Opera-tional Costs, millions of dollars	Cost of Energy at the Re-ceiving End, dollars/kWh
	Energy, billion kWh	Power, thousand MW				
1. Transmission of electric energy from eastern areas	225.2	31.7	37.9	5.31	515	0.0023
2. Construction of thermal stations, using local and Asiatic coals	225.2	31.7	33.3	5.58	1027	0.0046

* Fuel extraction, construction of power stations, transmission lines, and rebuilding of railways.

2. Energy at the receiving end—37.5 billion kWh per year.
3. Capital costs, including mining and transportation of fuel, construction of thermal-electric stations in Kazakhstan, dc transmission lines, and 500-kV ac receiving networks—$928 million.
4. Operation and maintenance costs—$87 million per year.
5. Cost of energy at the receiving substation in Tambov—$0.23 per kWh.

The valuable experience obtained during the initial period of operation of the Volgograd–Donbass transmission line (±400 kV dc, 750 MW, 473 km long) is of great importance to the success of the Kazakhstan–Tambov project. The first station of the lower voltage line was commissioned in 1962, and in 1963–1964, the transmitting capacity and voltage attained the design values just mentioned. During this period, the line operated under different conditions—unipolar and bipolar, with a different number of bridges and at varying voltages (100, 200, 300, and 400 kV per pole and up to 800 kV between the poles) and with reversal of power flow.

Results of experience. The experience and the results obtained during the initial period of experimental operation proved that the entire dc transmission system was quite satisfactory, and it did not require any major modifications in design or operational procedures.

The possibility of long-duration operation with current to ground (up to 900 amperes) has been proved. Also, asynchronous operation of the Central and the Southern systems, as well as the regulation of power flows in the dc transmission line—irrespective of the voltage levels and frequency in the power systems being interconnected—has been accomplished in the Volgograd–Donbass transmission line.

The operating experience and studies undertaken on the Volgograd–Donbass line have confirmed that the high-voltage insulation levels of the line and substations are sufficient. Further, the adopted system of protection from internal overvoltages, the scheme of transmission with the successive connection of individual converter units, and the use of protective and automatic devices have also worked out satisfactorily. As far as the performance of carrier-current channels on the dc transmission line are concerned, in reference to communication equipment, no difficulties have been encountered.

The experimental operation of the Volgograd–Donbass dc transmission line, from the time of commissioning up to the present, has confirmed the principal theoretical characteristics and properties of an overhead, high-capacity transmission line, and has proved that the basic scientific, technological, and design solutions were correct. It should be emphasized also that this transmission line, from the outset, has been providing service for industrial loads.

Thus the experience gathered at Volgograd–Donbass has made possible the planning of EHV dc transmission lines of higher capacities—from Siberia and North Kazakhstan to the Urals and European Russia.

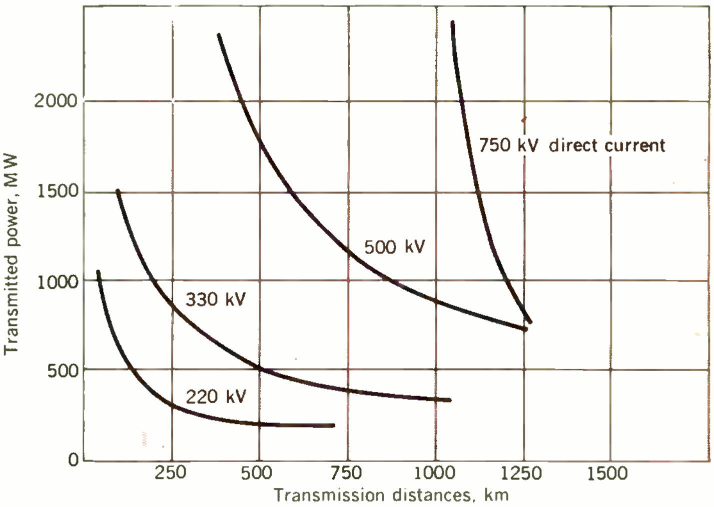


FIGURE 2. Transmitted power plotted against transmission distances for voltages of 220 kV and higher.

VI. Economic cost comparison of alternative schemes for EHV transmission line from Krasnoyarsk to the Urals

Type of Transmission	Capital Investments, millions of dollars	Annual Costs, millions of dollars
1500 kV dc	253	22.9
Alternating current:		
1000-kV resonance transmission	281	28.2
1000-kV compensated transmission	410	38.9

VII. Economic cost comparison of alternative schemes for EHV transmission line from Kazakhstan to Tambov

Type of Transmission	Capital Investments, millions of dollars	Annual Costs, millions of dollars
1500 kV dc	275	24.3
Alternating current:		
1000-kV resonance transmission	334	34.9
1000-kV compensated transmission	517	49.5

This article is based on a paper presented at the American Power Conference held in Chicago, Ill., April 26–28, 1966.

BIBLIOGRAPHY

Collected articles. *500-kV Long-Distance Transmission Lines* (in Russian). Moscow: Energia, 1964.

Akopyan, A. A., *et al.*, "Internal overvoltages in 500-kV long-distance transmission lines and measures for their limiting," CIGRE, 1962 Session, Rept. 405.

Akopyan, A. A., *et al.*, "750-kV experimental and commercial transmission line Konakovo–Moscow," CIGRE, 1964 Session, Rept. 413.

Levitov, V. I., *et al.*, "Operational experience on 500-kV networks in the U.S.S.R.," CIGRE, 1966 Session, Rept. 416.

Venikov, V. A., *et al.*, "Boost regulation in transmission systems in the U.S.S.R.," CIGRE, 1960 Session, Rept. 325.

Some aspects of binaural sound

In binaural listening, the human can hear two normally dissonant sounds without subjectively experiencing dissonance. Among other things, this finding suggests that in the human listener, harmony and melody obey different rules

Charles J. Hirsch *Radio Corporation of America*

Experiments in binaural listening show that the two ears provide almost independent effects on the brain—namely, a sound message confined entirely to one ear does not interfere appreciably with a sound message confined entirely to the other ear. A result of this independence is that harmony, in a restricted sense, between binaurally heard tones does not exist. This fact might be used to create new effects in two-part music. In the most general sense, the experiments described in this article show that any two messages—whether in music, in words, or in symbols—can be separated more easily by the brain when the messages are heard binaurally, a finding that raises certain interesting implications.

Binaural listening has been the subject of many experiments, and psychophysicists are well aware of the fundamentals of hearing and perception.[1] Some of this knowledge is so old, and the experts accept it as so commonplace, that new applications, particularly in music, are sometimes overlooked. This is particularly important today because electronics has made two-channel sound systems (stereo) readily available, and binaural listening with earphones is more and more widely practiced. This article describes some simple experiments, of special interest to the musically inclined, that were rather startling even to sophisticated engineers who participated in them. They are used to construct a new experiment in binaural listening that leads to insights into the physiology and psychology of hearing and suggests a new art form. This article describes this experiment and draws inferences from its results.

This article is not intended to be a scientific treatise but rather a stimulus to further work to explore potentialities of separate channels of sound, one to each ear, for music and entertainment.

To provide a more complete understanding of binaural listening, we should first compare it with normal and with stereophonic listening.

In normal listening, a source of sound is heard by both ears; the proportion heard by each ear depends on the attitude of the listener's head with respect to the source.

In stereophonic listening, which enhances the quality of reproduction of music and speech, two loudspeakers supply the two ears with independent sound messages from two microphones. The loudspeakers are located to form an isosceles triangle with the head of the listener at the apex. Each ear, however, receives some sound largely intended for the other ear, by diffraction around the head and by reflections from the wall.

In binaural listening, each ear receives only the sound intended for it. At present, this practice requires two earphones, which is, admittedly, rather inconvenient. However, the greater enjoyment of hearing stereo recordings by binaural listening has caused most manufacturers to offer comfortable binaural earphones with their stereophonic equipment. For the same reason, the listening to music and to motion picture sound provided by the airlines is entirely by binaural earphones. Binaural listening is the limiting case of stereophonic listening in which the separation between left and right sources is complete.

On the nature of harmony

Since this article deals with harmony in a restricted sense, a few words of clarification are desirable. Harmony is usually defined as the total subjective effect produced when two or more tones are sounded simultaneously. Melody is the subjective effect of tones heard sequentially.

In the case of harmony, the simultaneous tones combine in the ear to create subjective tones whose frequencies are the sums and differences of the frequencies of the original tones. These subjective tones add themselves simultaneously to the original tones to produce the total subjective effect known as harmony. For the purpose of this article we shall say that there is no harmony unless the subjective tones are created.

Harmony includes consonances[2] and dissonances whose effect can be roughly and loosely labeled as pleasant or unpleasant (or unsatisfactory) respectively. Over the centuries musical tastes have evolved and the meaning of consonance and dissonance is no longer divorced from the musical context by musicians. Thus, some intervals that were once considered to be dissonant are now accepted as consonant and some that were consonant now sound "square."

To draw simple conclusions from the experiments described later, and to reach basic understanding, we shall willfully divorce harmony from its musical context and limit our consideration to the harmony of single pairs or chords. The dissonances that we shall consider produce rough grinding sounds that, by themselves, are disagreeable. The composer can use them for his own artistic effects but he should understand their nature.

Binaural listening produces unusual effects that can be used to create new musical sensations amounting to a new art form. As we shall see, there is no, or at least very little, harmony (neither consonance nor dissonance) between tones heard binaurally; harmony exists only between tones that are mixed prior to application to each ear. As will be described, binaural hearing provides new opportunities for the reproduction of existing musical compositions in a novel manner, new opportunities for contrapuntal music, and new opportunities for novel musical effects. Perhaps most important of all, it provides a key to a fuller understanding of the nature of harmony.

Early experiments

The binaural effects derive from the physiological fact that the two ears have *separate and almost-independent* nonlinear effects on the brain. Because of this, the amplitude of a sound heard only by one ear does not appreciably add, or subtract, or intermodulate, or interfere with the amplitude of a sound heard only by the other ear (except for the small coupling, through bone conduction in the skull, usually observed only for loud sounds because of threshold and masking). This important fact was demonstrated by a set of experiments performed about a generation ago, whose results are summarized as follows:[3-6]

Experiment 1: If one note is heard only by one ear and another note of the same frequency is heard only by the other ear (binaural listening), the two notes do not combine to subtract their intensities when they are of opposite polarity.

Experiment 2: If two notes that differ slightly in frequency are mixed and then are applied to one or both ears (respectively called monotic or diotic hearing*), beats are heard at a rate equal to the difference frequency of the primary notes. As this difference frequency increases, the beats are no longer perceived individually and they assume a rough unpleasant quality[3] except when integrally related to the original tones.

If the two notes are heard binaurally, one note in each ear respectively, the beats are almost absent and the roughness disappears; instead the notes are perceived separately or, in some cases, as a compromise pitch.[3,6] When the primary notes are low in frequency and separated by a small frequency difference (say less than 10 Hz), a sound of constant intensity seems to circle back and forth within the head. The sensation of traveling is very weak and to perceive it at all requires much concentration. It disappears as the frequency, or the frequency difference, of the primary tones is increased. This phenomenon has been called a "binaural beat" by some writers and has been the subject of much study.[7] However, it is of a different nature[8] and magnitude than diotic beats. It does not seem to be caused by heterodyning of the original tones.

* Licklider,[1] page 1026, defines *monotic* as the stimulus applied only to one ear, *diotic* as the same stimulus applied to both ears, and *dichotic* as different stimuli applied to the two ears.

To determine if tones heterodyne when applied binaurally, the author repeated experiment 2 (above) with 25 observers, with pure sine waves and also with half-wave rectifiers in series with the earphones to introduce harmonics, at a level of 75 dB above 0.0002 dyne/cm^2. The tests were conducted at 100, 200, 400, and 1000 Hz and with $\Delta f = 5, 10, 20, 40, 50$ Hz, and above. Most of the 25 observers heard no beats at all. Those who heard beats identified them with the circling in the head mentioned earlier and described them as very much weaker than diotic beats. For our purpose such "binaural beats" will be ignored.

It may be that some workers, using very strong primary tones, hear *diotic beats* by bone conduction, which couples the two ears mechanically.

Experiment 3: The pitch of a note of constant frequency depends somewhat on its intensity[5]; as the intensity increases, notes below 500 Hz tend to decrease in pitch whereas those above 3000 Hz tend to increase in pitch.[6] Notes of differing pitch produced by a single frequency do not produce beats.

Experiment 4: With some listeners, a note of constant frequency may produce a slightly different pitch in the two ears. This condition is called diplacusis.[1] Notes having the same frequency but producing a different pitch in the two ears do not produce beats when so combined.

Experiment 5: The sound heard by the left or right ear seems to originate in the left or right side of the head. The listener interprets it subjectively as coming from the corresponding location in front of him. When the two ears receive identical sounds, even from earphones or spaced speakers, the sound seems to originate from a single cramped point source in front of the listener or inside his head. When multiple sounds are heard by the two ears independently, the sounds differing from each other in frequency, amplitude, or phase, they appear to be distributed in a wide area in front of the listener. Each source of sound assumes a location in space that separates it from other sources and helps to identify it.[1]

Novel effects produced by binaural listening

How stereophony makes use of all these effects to increase the reality and richness of reproduction has been described in the literature.[9] However, for our immediate purpose, we shall concentrate on the finding of experiment 2, which shows that "two notes of different frequency do not produce beats or roughness when heard separately by the two ears, although they do produce beats or roughness when both notes are heard by each ear."

Helmholtz suggested that the dissonance produced by two notes having a specific musical interval is caused by low-frequency beats that fatigue the auditory nerves in the manner that "flicker" fatigues the optical nerves.[10]

From this peculiarity, we can deduce that the two tones, say C and C sharp (#), that cause dissonance when heard simultaneously in one ear, do not produce dissonance if one ear hears only C and the other ear hears only C#, because, as shown in experiment 2, the beats are negligible. This is, in fact, found to be the case. When C and C# are combined in one ear, they produce an annoying rough grinding sound, similar to that made by a badly overloaded audio-frequency amplifier. The grinding sound disappears when the tones are heard by separate ears simultaneously. Instead, one hears the two tones individually or, sometimes, a compromise tone.

Conversely, combinations of tones that sound richer when heard together than when heard individually, such as C and G or C and E, do not sound any better, or any worse, than C and C# when heard binaurally. The absence of a beat note, when a musical interval is heard binaurally, eliminates sensations occurring in monaural hearing of the same interval.

A new experiment. To test the result of this observation, the writer asked a cellist to play a musical composition in a given key. It was recorded on one track of a stereo tape. The cellist then retuned his cello by a half or full tone, and recorded the same selection on the second track of the stereo tape, timing himself by a metronome so that corresponding notes on the two tracks were heard simultaneously on stereo playback. The cellist played the following selections in the keys noted:

Selection by J. S. Bach	Left Ear, Track 1	Right Ear, Track 2	Difference in Tone
Sarabande from Suite No. 5	C minor	B minor	1/2
Sarabande from Suite No. 4	E-flat major	E major	1/2
Prelude from Suite No. 1	G major	A major	1.0

When either track is heard alone in both ears, the reproduction is satisfactory. When the two tracks of the stereo tape are mixed, as in the monophonic reproduction of stereo, the music is full of grinding rough sounds and annoying beats. In addition, it seems to be concentrated in a cramped point directly in the center of the head or in front of the listener.

When heard binaurally, the two reproductions are heard without grinding sounds or beats (dissonances in the Helmholtz sense). Each reproduction is free of undesired effects. The two selections are heard separately and simultaneously, each being melodically consistent with itself. Of course, each ear combines the notes it receives to produce beats that may or may not be consonant. The lack of beats is only between notes heard binaurally. In addition, the sound fills the head or seems to be diffused in front of the listener as in stereophony.

The result is, to say the least, unusual. Listeners describe the effect as if there were a wall in the middle of their heads that separates the two sounds. Most listeners try to concentrate on both renditions simultaneously but they report that their attention wanders from one to the other sequentially and becomes confused. Generally, these listeners do not enjoy the result. The wandering of the mind from one selection to the other may be due to small random errors in synchronism between related notes, in the left and right tracks of the tape, which are meant to be heard simultaneously. If a tone is heard by one ear one millisecond or more before it is heard by the other, it will appear to be heard only by the first ear unless the later tone is stronger by 8 dB or more. This is the *precedence effect*.[11] It would be interesting to repeat the experiment by using automatic means to transpose from one key to another.

A few accept the overall effect uncritically and like it, perhaps because of the stereophonic effect. Other listeners are intrigued by the novel effect and try to think of possible applications. For example, as stated above, two notes bearing the frequency relations 1.0 to 1.5, normally the most harmonious of pairs, do not create any harmony when heard binaurally. On the other hand, no discords are heard no matter what their frequency ratio is. If one ear receives the chord c, e, and g, and the other ear receives the chord c#, e#, and g#, the usual discords c# — c, e# — e, and g# — g (where c# — c, etc., means the frequency of c# minus the frequency of c) are not heard and the music may gain in interest from the new sound sensation resulting from the chords c, e, g, and c#, e#, and g# in combination.

Binaural reproduction might be used to create new musical effects and to liberate contrapuntal music from the restraints imposed by the rules of harmony because binaural harmony, in its restricted sense, is practically nonexistent. Harmony applies only to tones that are mixed in each ear. Although, as stated above, most listeners do not enjoy the result of the experiments on Bach, skilled composers may be able to devise new artistic binaural combinations. For example, two sets of instruments might play the various parts of a fugue binaurally in different keys.

I. From Jeans "Science and Music"

Concord associated with small numbers

It is found to be a quite general law that two tones sound well together when the ratio of their frequencies can be expressed by the use of small numbers, and the smaller the numbers the better is the consonance. This will be clear from the following table, in which the intervals are arranged in order of increasing dissonance:

Interval	Frequency Ratio	Largest Number Occurring in Ratio
Unison	1:1	1
Octave	2:1	2
Fifth	3:2	3
Fourth	4:3	4
Major third	5:4	5
Major sixth	5:3	5
Minor third	6:5	6
Minor sixth	8:5	8
Second	9:8	9

Other binaural effects. The increased richness of stereophony is due partly to the spatial distribution of sounds and partly to the fact that pairs of tones of equal frequency and amplitude but largely opposite in phase, which cancel each other when heard monophonically, do not cancel when heard binaurally. Such pairs are very numerous when several instruments play simultaneously, so that fewer tones are heard with monophonic than with stereophonic reproduction.[9] Many listeners prefer the binaural reproduction of stereo by earphones to the more usual reproduction by loudspeakers because the separation between left and right ear is more nearly complete and results in fewer cancellations.

In one experiment a selection was heard in one ear while the same selection was heard in the other ear after an adjustable delay. With very little delay (0.1 second), the effect was that of "quasi-stereo" in that the music appeared to come from a wide angle instead of being concentrated at a point. With more delay (0.5 second), it produced an echolike effect and the wide angle was maintained.

The independence of the ears can be used to increase the rate of flow of information to the brain. It has been known for a long time that radio operators can differentiate between two messages in Morse code as long as each message is confined to one ear. The same principle applies to word messages; for example, the words Ticonderoga and Constantinople remain separately more intelligible when heard binaurally but become confused to something like "conestoga" when combined monaurally.

Physiological basis of harmony

Twenty-five hundred years ago Pythagoras (582–507 B.C.) established the fact that tones heard simultaneously sound best when their frequencies can be expressed as

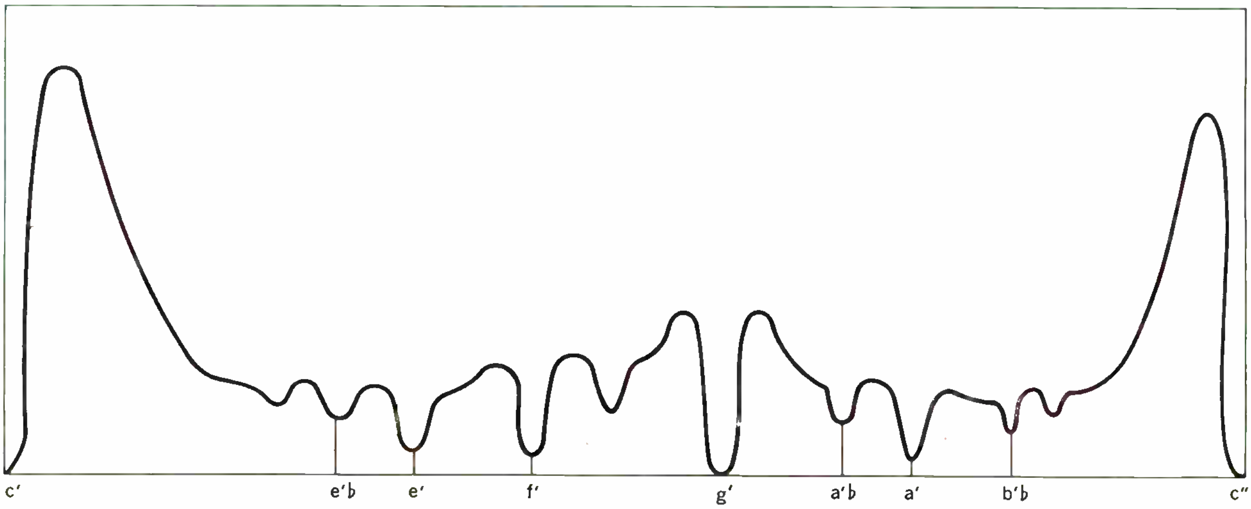


FIGURE 1. Two violin tones sounding together create the degree of dissonance shown here. The lower tone c' sounds continuously while the upper tone moves gradually from c' to c''. This observation by Helmholtz was taken from Jeans "Science and Music."

the ratio of small numbers, such as: 2/1, 3/2, 4/3, 5/4, etc. These are respectively the octave, fifth, fourth, and third notes of the scale of "just intonation." Numerologists, mystics, metaphysicians, and philosophers tried to supply explanations for this relation (e.g., Table I).

Much later, in 1862, Helmholtz[10] theorized that dissonances are caused by low-frequency beats between fundamentals and between harmonics, which produce a sort of aural flicker that tires the aural nerves. Since pairs of tones whose frequency ratio is expressed in small numbers have more of their harmonics exactly in unison, they have fewer harmonics that can beat with each other to create dissonances. Note that the term "harmonic" is used here in the mathematical sense of a multiple of a fundamental frequency. Other expressions for harmonics are "overtones" and "upper partials." The term does *not* imply that the tones are "harmonious." Figure 1 shows the relative dissonance produced by two violins playing different notes. While this theory explains why tones bearing these ratios do not sound badly when heard simultaneously, it does not explain why they sound better than single tones.

As recently as 1937, Sir James Jeans, in his book *Science and Music*,[12] stated: "Innumerable theories are ready to tell us the origin of the annoyance we feel on hearing a discord, but none even attempts to tell us the origin of the pleasure we feel on hearing harmony; indeed, ridiculous though it may seem, this latter remains one of the unsolved problems of music."

Perhaps the following discussion will throw a little light on this mystery. The binaural experiment on Bach, as well as experiments with individual tones, shows that the ears have largely independent effects on the brain, and that the brain does not combine tones, transmitted simultaneously but separately by the two ears, to produce harmony. Harmony, which includes consonance and dissonance (used here in the Helmholtz sense of low-frequency beats), requires that the simultaneous component tones be combined in one ear. It is therefore very likely that harmony is produced by intermodulation, i.e., by the mixture of tones, in a nonlinear channel, producing beat notes. Beat notes are often called subjective tones because they do not exist in the primary (original) physical stimulus. A combination of tones produces a discord, which takes on partly the nature of noise, if the frequencies of the subjective and original tones do not bear a harmonic relation to each other, and approach randomness. This condition is especially true if the primary tones create low-frequency beats, say 4 to 40 Hz (and up), that produce a rough grinding sound, or a type of aural flicker, similar to that produced by a badly overloaded amplifier, which fatigues the aural nerves to the brain. In general, the combination is not unpleasant if the primary and subjective tones bear a harmonic relation to the difference tone of lowest frequency.

The hypothesis is here proposed that: Harmonious pairs and chords sound richer *because the ear creates subjective tones whose pitch is a subharmonic of the tones in the primary pair or chord but is higher than the threshold of tone perception.* Incidentally, Helmholtz states that ". . . notes do not begin to have a definite pitch until about 40 vibrations are performed in a second."[10] The hypothesis does not try to explain why low notes sound richer.

The creation of subjective subharmonic tones is, of course, well known. It is used by organ builders to create low subjective tones from two small organ pipes instead of a single large one. The difference in the frequencies emitted by the two small pipes is made equal to the frequency that would be emitted by the large pipe. It is also used to produce "synthetic bass" from very small loudspeakers.[13,14]

For example, consider a chord consisting of the major triad, a, c#, e, whose frequencies in "just intonation" are exactly a = 220 Hz, c# = 275 Hz, and e = 330 Hz. These notes bear the following frequency ratios to a:

$$\frac{c\#}{a} = \frac{275}{220} = 1.25 \qquad \frac{e}{a} = \frac{330}{220} = 1.50$$

The difference c# − a has a frequency of 275 − 220 = 55 Hz whose frequency ratio to a is 1.25 − 1.00 = 0.25. The difference e − a has a frequency of 330 − 220 = 110 Hz whose frequency ratio to a is 1.5 − 1.0 = 0.5.

Placing the primary and created subjective notes in sequence, as shown in Table II, we see that there are among the new notes created by the first-order nonlinearity: (1) c# − a = e − c# = 55 Hz = AA, which is *exactly* two octaves below the lowest original note, a (it is equal to a/4, c #/5, e/6); and (2) e − a = 110 Hz = A, which is *exactly* one octave below the lowest original note, a (it is equal to a/2, 2c#/5, e/3).

Surely, the combination AA, A, a, sounds deeper

II. Subjective tones created by chord a c# e

Frequency, Hz														
Just intonation	55	110	165	220	275	330	385	440	495	550	605	660	825	990
Nearest in eq. temp.	55	110	165	220	277	330	392	440	444	554	587	659	831	988
Original tones														
Designation				a	c#	e								
Freq. ratio to a				1.00	1.25	1.50								
Subjective tones	e—c#									2c#				
By 1st-order nonlin.	c#—a	e—a						2a	a+c#	a+e	c#+e	2e		
By 2nd-order nonlin.		2a — e	2a — c#	2c# — e		2c# — a	2c# — e	2e — a				3a	3c#	3e
All tones														
Ratio to a	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.75	4.50
Ratio to AA	1	2	3	4	5	6	7	8	9	10	11	12	15	18
Designation	AA	A	E	a	c#	e	g	a′	b′	c′#	d′	e′	g′#	b″

All tones are in "just intonation."
The nearest tone in "equal temperament" is given for identification and designation.
The designation is only approximate for either "just intonation" or "equal temperament." Only a = 220 Hz is defined.

(and richer) than a alone. The three-tone chord a, c#, e, has become the ten-tone chord AA, A, a, c#, e, a', b', c'#, d', e', which is deeper in pitch and completely free of dissonances.

If, in addition, we consider the beats produced by the second-order nonlinearity (which produces the beats $2f_1 - f_2$ and $2f_2 - f_1$, $3f_1$ and $3f_2$) the original chord has been increased to 14 tones: AA, A, E, a, c#, e, g#, a', b', c', d', e', g'#, b''. All these tones are exact harmonics of 55 Hz and, in any combination, cannot create any beats other than 55 Hz or harmonics of 55 Hz. One of these is $2a - c = 165$ Hz = E, which is exactly one octave below e. Of course, these second-order beats are lower in amplitude than the first-order beats. The above example uses the major triad in "just intonation" because it is the ideal combination of tones for illustrating the creation of a perfect consonance with no dissonance. The same reasoning applies to other tone pairs or chords although to a lesser degree; their upper harmonics may not coincide exactly, as in the major triad, and can produce some dissonances. One reason for the fine quality of chamber music may be that its instruments—violin, viola, and cello—often use "just intonation," or the "Pythagorean intonation." That is, they produce combinations of tones with exact sub- and upper harmonics and therefore fewer dissonances than is the case with "equal temperament."

The rules of harmony, in the restricted sense, appear to be imposed by physiological factors such as the nonlinearities of the hearing channel, which imposes strict frequency relations for tones played simultaneously.

On the other hand, melody depends on the short-time memory for a *sequence of tones*, such as the memory of the successive words in a sentence without which there is no meaning. Melody, therefore, seems to be governed by psychological factors such as *acquired* and *impermanent standards* of value and meaning.[15]

It follows from the difference between harmony and melody that the musical intervals used for melody, if a scale is used at all, need not be the same intervals, or frequency relations, that are used for harmony. For example, a pleasing melody might consist of successive single notes or a succession of chords, which might be unpleasant if heard simultaneously. This difference appears to supply a physiological justification for "atonal" or other modern music.

An objection, based on historical data, can be made to the statement that musical intervals for harmony and melody obey different rules. Essentially all early music was monophonic, yet almost all ancient scales used musical intervals closely related to those required by harmony.[10] Perhaps the reason for this strange coincidence lies in the fact that much early music was performed on open-string instruments having appreciable coupling between strings so that plucking one string would cause others to resonate. In addition, the low damping of strings can make successive tones overlap in time, as in the playing of arpeggios today, and create a harmony of sorts. This combination of melody and rudimentary harmony would lead to the modern musical intervals.

Conclusion

Experiments in binaural listening show that the two ears provide almost independent effects on the brain and that a sound message confined entirely to one ear does not mix appreciably or interfere with a sound message confined entirely to the other ear.

A result of this independence is that harmony, in the restricted sense, between binaurally heard tones does not exist. This fact might be used to create new effects in two-part music. Another finding is that two messages in music, words, or symbols can be separated by the brain more easily when heard binaurally.

Conventional harmony is created by the nonlinearity of the ear; the richness of a consonant chord is probably produced because its component notes create subjective tones whose pitch is appreciably lower than the tones in the chord itself, along with the creation of a minimum of dissonances. It seems likely that harmony and melody obey different rules and therefore can be treated independently.

The writer wishes to emphasize that he is presenting deductions from experimental results with suggestions, but no recommendations, as to their use. In the light of modern developments in music, he does not express any judgment on the relative values of consonance and dissonance. He feels that these are means that the artist should use for his own purposes, at his discretion, but that he can benefit from understanding them.

The contributions of the following are gratefully acknowledged: John Phetteplace, formerly of the RCA Record Division, for arranging the technical and artistic details of the binaural Bach recordings; David Moore, cellist, for his patience, skill, and understanding in doing a most unusual job of recording the Bach selections. The author thanks his associates at RCA for suggestions and encouragement.

REFERENCES

1. Licklider, J. C. R., "Basic correlates of the auditory stimulus," in *Handbook of Experimental Psychology*, ed. by S. S. Stevens. New York: Wiley, 1953, pp. 985–1039. This is an excellent review with extensive bibliography.

2. *Harvard Brief Dictionary of Music*. New York: Washington Square Press, 1961.

3. Beatty, R. T., *Hearing in Man and Animals*. London: G. Bell, 1922, pp. 108–115.

4. Stevens, S. S., and Davis, H., *Hearing*. New York: Wiley, 1938, pp. 241–245.

5. Olson, H. F., *Elements of Acoustical Engineering*. Princeton, N. J.: Van Nostrand, 1940, p. 322.

6. Dacos, F., "Sur la notion et l'expression du timbre musical," *Bull. Sci. Assoc. des Ingr. Electriciens Sortis de l'Inst. Electrotech. Montefiore (AIM)*, Liege, Belgium, no. 6, p. 481, Nov.–Dec. 1964.

7. Rutschmann, J., and Rubinstein, L., "Binaural beats and binaural amplitude-modulation tones, successive comparison of loudness fluctuations," *J. Acous. Soc. Am.*, p. 759, 1965.

8. Fletcher, H., *Speech and Hearing in Communication*. Princeton, N.J.: Van Nostrand, 1953, pp. 214–216.

9. Hirsch, C. J., "The non-directional aspect of stereo," *IRE Trans. on Broadcast and Television Receivers*, vol. BTR-7, pp. 36–39, Nov. 1961.

10. Helmholtz, H. L. F., *Sensations of Tone*. Translated from fourth German ed., 1877, by Alexander Ellis. New York: Longmans, fifth ed., 1930, pp. 170–171, 177, 227, 253, 331.

11. Pierce, J. R., and David, E. E., *Man's World of Sound*. New York: Doubleday, 1958, pp. 125–126.

12. Jeans, Sir James, *Science and Music*. New York: Macmillan, 1937; paper ed., Cambridge Univ. Press, 1961.

13. Langford-Smith, F., *Radio Designer's Handbook*. London: Iliffe and Sons, Ltd., fourth ed., 1953, pp. 616, 676.

14. Shepard, Jr., F. H., "Method and means for reproduction of sound frequency vibration," U.S. Patent 2 313 098, Mar. 1943.

15. Meyer, L. B., *Emotion and Meaning in Music*. Chicago: Univ. of Chicago Press, 1956, p. 63ff, p. 230ff.

In addition, an excellent guide is: Seashore, C. E., *Psychology of Music*. New York: McGraw-Hill, 1938.

A new keyboard without keys

A classical exercise or the far-out sounds of space music, either is simple to produce as a result of the application of the art of electronics to the conventional instrument keyboard

Paul Rosberger *Binary Keyboard*

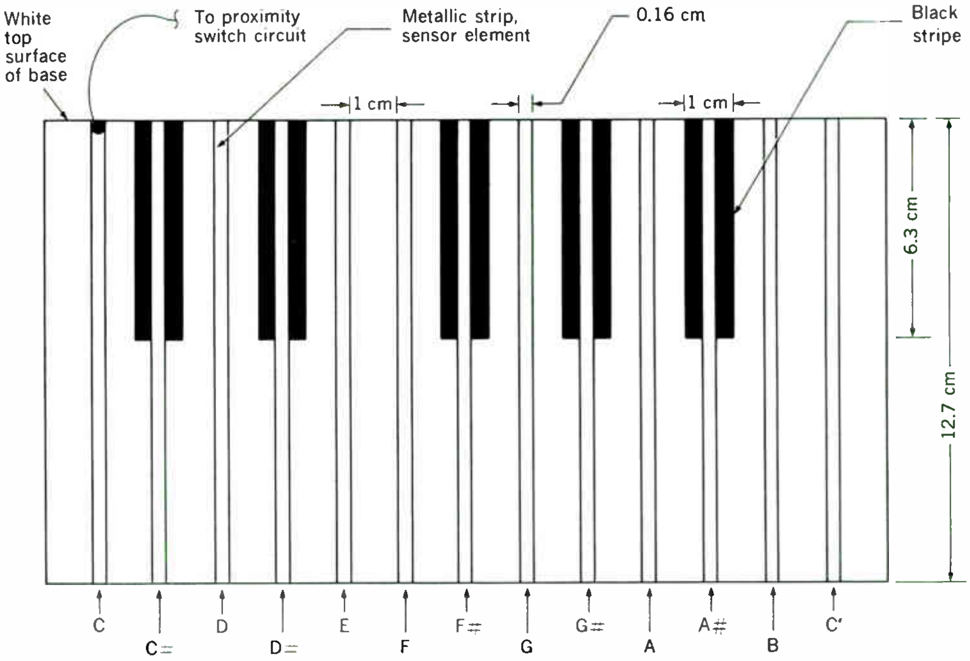

White top surface of base
To proximity switch circuit
Metallic strip, sensor element
0.16 cm
Black stripe
1 cm
1 cm
6.3 cm
12.7 cm
C
C#
D
D#
E
F
F#
G
G#
A
A#
B
C'

FIGURE 1. Details of the new Binary Touchboard.

The advantages of the electronic medium and those of a simple congruent instrument keyboard have been combined to produce the Binary Touchboard described in this article. Basically, it operates by translating a finger touch into a switching signal that activates a note-sounding device. The simplified fingering is especially advantageous in such areas as atonal music, jazz, and what might be characterized as "spatial music."

Much of the credit for the magnificent and abundant sounds of this century belongs to the general evolution of electronics. Sounds can now be produced, imitated, amplified, and recorded with a generally unprecedented improvement in efficiency over mechanical or electrical means. Unfortunately, with all this "coming of age," an area of music that still remains comparatively unexplored is the electronic control of sound, which has been limited to specialized instruments such as the theremin, trautonium, and ondes martenot. And absolutely nothing has been provided electronically to improve the conventional music keyboard—although it has been studied as a mechanical structure with a view to rearranging its irregularly spaced and shaped keys in an even geometrical and spatial division that would benefit the performer.

The proposed keyboards probably never have been accepted either because they are complicated, or, having their own fingering limitations, lack any real improvement. Enthusiasm for the inherent advantages of the electronic medium and those of a simple congruent keyboard inspired the writer to combine these forces. The resulting "keyboard" operates electronically, and since it has no moving parts it is called a Binary Touchboard.*

Some features

As shown in Fig. 1, the Touchboard consists of a rigid dielectric base with a white top surface. On this surface are 0.16-cm-wide parallel metallic strips, which are of minimal thickness (0.010 cm), are reasonably flush with the base, and are secured with a pressure-sensitive adhesive. There is a 1-cm spacing between adjacent strips. Each strip is individually connected to a musical-tone generator and they are arranged to sound the tone generator in a chromatic order. The strips that sound accidental notes have a black stripe centered under them that is about 1 cm wide and extends some 6.35 cm from the back of the Touchboard. This provides a coding that resembles the conventional keyboard.

To sound any desired note, the related metallic strip is merely touched by a finger of the performer. There is no movement or mechanical resistance such as exists with depressible keys.

Briefly, each metallic strip functions as a capacitor sensor element of a proximity switch of the type that is used as an elevator button, automatic door opener, or safety alarm device. Various proximity switch circuits are suitable for the Binary Touchboard. They operate by translating finger touch, through an amplification circuit isolated from the performer, into a strong switching signal. The switching signal can activate any electric or electromechanical note-sounding device. For example, if the device were a solenoid, it would directly operate the switch under a conventional electronic organ key.

The advantages of a single row of congruent sensor elements can be found by comparison with the irregular order of black and white keys on a conventional keyboard. Conservative harmonic structure is fashioned around a key center and utilizes closely related keys from which the composition modulates. For every scale in our duodecimal system there are 12 different key orders to be fingered. This requires remembering 12 different "incomplete images" of the total keyboard for each scale. Unfortunately, the scales vary in their fingering orders and also in difficulty of performance—which compounds the complexity of learning. There is a similar difficulty in locating types of chords and intervals, especially when there is frequent modulation within a passage. Such problems become more severe when contemporary music is encountered.

With the Binary Touchboard all 12 keys of a given scale can be fingered in an identical manner. Also, only one position is required for any given chord or interval and, once familiar, it can be visually relocated by referring to only the tonic or end note. In general, by eliminating the conventional two irregular orders of keys of different lengths and shapes, every note can be sounded without regard to varying key leverage, finger length, or accessibility. A whole tone is always a whole step in distance; a semitone is a half step in distance.

The spatial advantage of the Touchboard can be seen by again referring to Fig. 1. Minimal-width sensor elements with maximum clearance between them provide a total of about 2.1-cm clearance for sounding each note; their area is comparable to that of a natural key on the keyboard were it unobstructed by accidental keys. Since a clearance space is available to either of the sensor elements adjacent to it, this maximum area of the Touchboard has a double function. By having sensor elements no wider than the lost space between natural keys of a conventional keyboard and a Touchboard clearance space that has a binary function, the overall width of the Touchboard becomes one half the width of a keyboard figuratively made up solely of natural keys since the latter would have to be wide enough to permit practical finger clearance and flat positive touch to conventional keys in a congruent single row. The span of intervals is larger than on the conventional keyboard so that a major tenth interval, or greater, is easily reached by any adult.

Where a plurality of manuals is desired, as on an organ, the new Touchboard makes it possible to play on more than one manual at once with a single hand. This is possible because a second manual can be positioned in close relation to the first, similar to the relation of the accidental keys to the naturals on a conventional keyboard. The Binary Touchboard has several other physical advantages. For example, a note can be repeated as fast as the performer can oscillate his finger. There is no mechanical delay, and the degree of finesse and consistency of touch exceeds that possible with any mechanical device. Also, there is no lost motion or random articulation such as is caused by the depression of a mechanical key.

The Touchboard is virtually indestructible. Unlike the wear and tear on conventional keys, excessive operating pressure cannot damage it. However, should damage result from some cause other than operation, the least costly part of the overall unit is the Touchboard surface. It can be frequently and easily replaced, and requires no readjustment of the surface itself or of the overall unit.

Another advantage is that the Touchboard can be tilted at a slight angle so that its back portion is higher than the front, which facilitates finger–thumb crossover motion. The sensor elements can be made touch-sensitive for use in conjunction with a percussion pianoforte instrument. The varying pressure of the performer's

* Patent pending.

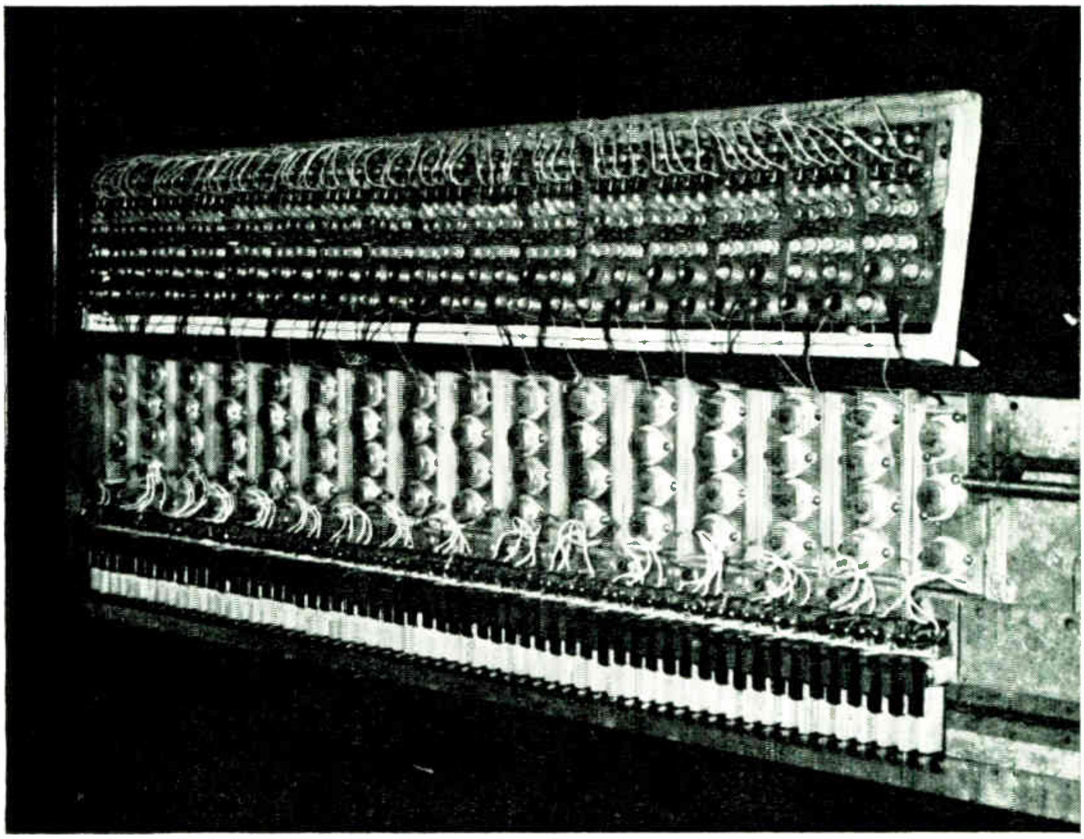


FIGURE 2. A manual of the Touchboard prototype for converting an electronic organ.

finger on the sensor element is in direct proportion to the resulting dynamic level of the responding note.

A new sound

The question of acceptance of the Binary Touchboard is not necessarily one of whether or not it will replace the conventional keyboard. The writer feels that it is not only an improvement, but that its unique features are of independent and equal importance. It creates a novelty that can add new and imaginative dimensions to sound, requiring fresh investigation of the potential of every instrument.

Composers, especially those involved in atonal music or music of complex key modulation, wiil have less difficulty and greater freedom in devising and expanding their concepts. Parallel or semiparallel motion of orchestral compositions can be tried without resorting to fingering in steps. For example, the combined motion of stringed instruments can be reproduced, such as a tremolo of thirds or the glissando and quasi-glissando of a chord. Glissandos of any speed can be produced effortlessly.

The new keyboard is advantageous for chromatic slurs (which are performed by quickly sliding one finger over two or more notes), and provides excellent interpretation of jazz. The amount of fingering can be reduced. For example, a diatonic scale can be played with three fingers since the finger of the note before the half step slides to the next note. The performance of complex harmonies such as cluster chords is also simplified since no effort is required to play a multitude of adjacent notes.

Novel sound effects are possible; these might be termed "spatial music." An example would be two fingers ascending the Touchboard while the distance between them is regularly widened or narrowed, or the fingers might be "drummed" as they are moved. Cluster chords can be glissandoed, and the glissando motion can follow any cyclic pattern and need not be linear or unidirectional. Cluster chords can be sounded for short durations, providing a percussive sound that may be varied in pitch and density.

Because most spatial sound effects are pleasing, a person who is new or inept at a keyboard can perform musically acceptable phrases. Such music can be of psychotherapeutic benefit for patients by providing an expressive musical medium that requires no initial formal education.

New research areas are opened up by the new keyboard. Experimenting with any number of divisions of the octave merely requires recoding a universal Touchboard and connection to desired frequencies. For example, it is now practical acoustically to test and develop performance of Joseph Yasser's 19-tone system. A 19-tone octave can easily be reached.

Figure 2 shows one manual of the latest Touchboard prototype for converting an electronic organ. The new equipment is easily fitted into the space provided by removal of the conventional keys. No physical modification of any kind is necessary.

Particular credit is due to Howard Levi, circuit designer, for successful completion of the prototype.

Dilemmas of engineering education

The values and attitudes true of the specialist are not always compatible with the values and needs of the professional, and therein lies a fundamental dilemma of engineering education today

Harvey Brooks *Harvard University*

Modern technology has posed a special dilemma for today's professional, who is confronted by rapid change in both the body of knowledge he must use and the needs and expectations of the society that he must serve. The resulting problem for professional education—to ascertain the proper balance between science and art in training for a profession—is considered in this article.

Medicine, engineering, business management, and education—these are the learned professions that depend primarily on scientific knowledge. I include business management and education among these because I believe that they face the same problems relative to the maturing social sciences that medicine and engineering have faced with respect to the biological and physical sciences respectively. The problem arises essentially out of the fact that, to an increasing degree, in each of these professions science has overtaken art, although to a varying degree in different parts of each profession.

A "learned" profession as considered here is one that rests on a basis of systematic theoretical and empirical knowledge, on some mixture of scholarly disciplines or "sciences." But the professional is much more than a man with knowledge, he is the middleman or intermediary between a body of knowledge and society. To the professional belongs the responsibility of using both existing and new knowledge to provide services that society wants and needs. This is an art because it demands action as well as thought, and action must always be taken on the basis of incomplete knowledge.

The dilemma of the professional today lies in the fact that both ends of the gap he is expected to bridge with his profession are changing so rapidly: the body of knowledge that he must use and the expectations of the society that he must serve. Both these changes have their origin in the same common factor—technological change. Technology has created a race between opportunities and expectations that was not foreseen by the 19th century science fiction writers. Our problem cannot be usefully phrased in terms of too much technology. Rather it is whether we can generate technological change fast enough to meet the expectations and demands that technology itself has generated. And the four professions—medicine, engineering, business management, and education—must bear the brunt of responsibility for generating and managing this change. This places on the professional a requirement of adaptability that is unprecedented.

The resulting problem for professional education concerns the relationship between science and art in the training of the professional. What is the proper balance in training between learning the techniques of the profession, i.e., how to apply existing knowledge, and the body of knowledge itself? What are the expectations of society going to be during the professional's working life, and how must he prepare himself to acquire the relevant scientific knowledge? Not only does much of this knowledge not yet exist, but the tasks he will be called on to perform for society may not yet be defined. Even in medicine, where the cure of disease seems a well-defined task, I suspect the reorganization and rationalization that the next 20 years will bring to the delivery of medical care is going to necessitate a radical redefinition of the task of the doctor as an individual professional, if not of medicine collectively as a profession. In business management the proliferating roles of enterprise in our society and the internationalization and "socialization" (in the sense of performing an increasing variety of social service functions) equally call for redefinition of the businessman's role.

However, I feel the problem of the engineer is an order of magnitude more difficult. The term "engineer" comprises a much greater diversity of activities and skills than is true of the other intellectual professions. These range from research and development to rather repetitive and routine engineering practice, though the latter is likely to be rapidly automated out of existence. At one extreme is an individual who is traditionally called an engineer but is really a scientist, for example, a man doing basic research in structural mechanics or aerodynamics. He is called an engineer, and finds his home in an engineering department partly because the physicists have repudiated his discipline—as is also true of the applied mathematician who has been expelled from the mathematical fraternity. At the other extreme is the systems manager who coordinates and plans the activities of hundreds of specialists toward a common goal.

Within the last 20 years the engineering profession has had to absorb many new technologies—nuclear power, transistors, microwave ferrites, the maser, the laser and nonlinear optics, the computer, microelectronics, cryogenics, ultravacuum technology, high-pressure technology. Just over the horizon we can see holography, man–machine communication, magnetohydrodynamics, optical communications, and probably many other technologies that are still just physics or laboratory instrumentation. Entire new areas of engineering application are also opening up—the technology of education, the technology of the delivery of medical care, the technology of urban planning, new transportation technologies, a new information technology evolving toward what one might term a social or collective brain, new technologies of environmental management.

Limitless horizons and finite capacity

What does all this mean for engineering education? What kind of faculty? What kind of research? What kind of curriculum and courses? I raise these questions in that order because I think this is the order of their significance and importance. In fact, as somewhat of an outsider, I would say that the main fault of engineering education is the excessive preoccupation with curriculum, as though the student were a computer into which just so much data must be programmed in a given time. In my view, the heart of the problem lies in the character and orientation of the engineering faculty. In the long run the courses and curriculum, and the knowledge and motivations of the students, are bound to reflect the research interests, the consulting experience, and the values of the faculty.

In this connection it might be useful to consider what is happening in some of the other professions or professional schools. In medical schools we have seen a tremendous growth of basic science departments and in the numbers of Ph.D.'s as basic scientists on medical faculties; in fact, in many schools there are more Ph.D.'s than M.D.'s. In business schools we find a growing number of Ph.D. economists, sociologists, and, to a lesser extent, psychologists and applied mathematicians—although the last group should grow rapidly with the invasion of data processing and quantitative methods into management. In engineering, we find increasing numbers of Ph.D.'s in physics and mathematics and, to a lesser extent, in biology and chemistry. Engineering as yet has few economists but probably should have more. By and large, engineering faculties have been somewhat more resistant to admixture with basic scientists than have the faculties of other professional schools.

Although I feel this trend to be both inevitable and good with regard to balance, it does present problems, and therein lies the fundamental dilemma of professional education. The specialist in a discipline brings with him to the professional school a set of values and attitudes that are not always compatible with the values and needs of the professional, especially the professional who plans to devote himself to service, which the great majority must do. The problem becomes more acute if the intellectual stature of the disciplinary specialist within a professional school is sensibly above that of the faculty representing the profession itself.

However, this tension of values is not entirely undesirable. Increasingly the task, even for the practitioners in the learned professions, is to apply the same intellectual standards to practice as to research. Too often sloppy thinking or technique is excused in the professional that would be inexcusable in the basic scientist. On the other hand, the basic scientist tends to carry with him one value that is fatal to the professional—a willingness or tendency to exclude from his purview all aspects of a problem not relevant to his own discipline. Thus the professional school instruction constantly faces the threat of becoming like a group of blind men describing an elephant.

The tension I am describing is not helped by the tendency of the engineering profession to exaggerate the difference between science and engineering. Too often the engineer, in describing what a scientist does, tends to forget that the experimental scientist, in setting up his equipment, often adopts the attitudes and approach of an engineer. Indeed, this is quite obvious in fields such as high-energy physics, where specialization has produced a new class of machine designers who behave as engineers but consider themselves physicists. I have observed that much of the motivation for building things that used to attract students to engineering now tends to attract them to the basic sciences, where they often have an equally challenging opportunity to use their skills. The point I am trying to make is that the simple picture of engineers interested in design and building and scientists interested only in theory is a caricature that confuses everybody.

Ideally, I like to think of the educational process as one in which each student is exposed to many different specialists, so that ultimately the general knowledge of one generation is formed from the combinations of specialized knowledge of the previous generation. This should be true for the scientist as well as the professional. Each specialist should sift his own discipline for those aspects having the greatest generality and applicability to pass them on as part of the general background of the student; thus in a sense each generation of students stands on the shoulders of the preceding generation. The student in each discipline should end up knowing more about neighboring disciplines than his professor in the same discipline—the budding physicist should know more mathematics than his physics professor, and the engineer more physics than his engineering professor. Implicit in this is the assumption of a sort of hierarchy of generality that places mathematics above physics, and physics above engineering. It is important, however, that this hierarchy not be misunderstood. Not all mathematics is more general than physics, nor is all physics more general than engineering, and this is where the dilemma arises. The disciplinary specialist must be sufficiently in tune with the professional environment in which he teaches that he himself forms a bridge between his discipline and the profession just as the professional forms a bridge between science and society. Thus the physicist or the mathematician in an engineering school should be an individual who can talk as an equal to his professional colleagues in mathematics and physics, but who can also talk without condescension to his engineering colleagues. I consider this "bridging" capacity to characterize the somewhat misnamed "applied scientist."

A large part of the faculty of a modern engineering school must inevitably consist of people trained in and working in various relevant disciplines rather than of professional engineers. These disciplines may be traditional disciplines of either pure science or engineering science; the distinction is an artificial one. One quality that is unique about the engineering faculty member is, however, that he must face two directions—toward his discipline and toward engineering. This is partly achieved simply by belonging to a single coherent enterprise. If the various disciplines within this enterprise become closed enclaves, impervious to all interests and concerns except their chosen science, the point of the enterprise is lost. In many ways the smaller schools have an advantage in this respect because they find it easier to maintain the mixture of disciplines which is essential to a coherent educational enterprise, and which is also characteristic of the best industrial laboratories.

The engineer

Where does this leave the engineer in the engineering school? Who is he and what is he? Where does he come from, and how is he trained? In this respect the medical profession has a happier situation, at least in theory, than the engineering profession. The teaching hospital provides a natural meeting place for the clinician and the basic scientist, for the student and the practitioner, for which there is no effective counterpart in the engineering school. The situation is further confused in engineering by the existence of a class of people who are really scientists by any sensible definition but who call themselves engineers. Of course, just as a basic scientist may behave as an engineer when he is designing and building his apparatus, so an engineering scientist may behave as an engineer when he consults on a design problem. The labels "engineer" and "scientist," even in the broader sense that I have used them, are not mutually exclusive. But it would seem that some of the engineering faculty should be involved in what would be the equivalent of patient care in a hospital as a primary responsibility if the overall educational effort is to be balanced; in other words, some of the faculty should be opportunity-oriented and some should be problem-oriented. The opportunity-oriented people are, roughly speaking, the scientists, and the problem-oriented people are the engineers. This is the problem that engineering education has neither solved practically nor resolved conceptually. Indeed, conversation with some of my medical friends suggests that even the doctors have not solved the problem as well in practice as in theory. Unfortunately, I have no easy solutions. The basic problem is that, contrary to the situation in the pure sciences, the real professional environment of engineering lies in industry or government. The Massachusetts Institute of Technology has found that organizations such as the Lincoln and Instrumentation Laboratories have some of the educational values of teaching hospitals, but even here little effort has been made to use these centers in a systematic way for educational purposes. They do however serve as a training ground for faculty to some extent. Various devices have been used by other schools to bring faculty and students closer to real engineering industry. All of this involves a substantial commitment of time, and the question arises as to whether the industrial experience might not be better achieved in series rather than in parallel as far as the student is concerned.

Another possible solution is the adjunct professor, who spends most of his time in industry or in consulting, but it is difficult to keep such individuals intimately involved with the rest of the educational enterprise. To be really effective, adjunct professorships would probably have to be nearly equal in number to the full-time staff positions. Further, the adjunct professors ought to be true engineers, engaged in design or systems management, not engineering scientists engaged in professorial-type activities—but unfortunately such individuals are seldom able to make the necessary time commitments.

One solution that has been little tried is the use of a part of the engineering activity to develop or provide services for the university itself. With the increasing importance of physical technology in education and medicine there is new opportunity for engineering involvement. There may be similar opportunities in the design of large research instrumentation, and even in aspects of engineering services for the university's building program. The difficulty is that this cannot be done with full-time faculties, both because of schedule conflicts and because of the tradition that faculty members should not act in a supporting capacity to other faculty members, no matter how challenging the assignment. At Harvard we have been operating an electronics design center on a modest scale that is staffed by part-time academic appointees. Computing centers also have this aspect, and it is certainly true that the computer is creating an opportunity for providing a more realistic, but manageable, problem-oriented environment within the context of the university.

Still another mechanism which has been proposed is the creation of cooperative research projects involving partnership between industry and universities. This is not a new idea. The Mark I calculator, the immediate progenitor of the modern digital computer, was built at Harvard in the 1940s through a true intellectual partnership between the university and IBM. More recently, the Advanced Research Projects Agency of the Defense Department has supported three "coupling" programs involving partnerships between an industry or government laboratory and a university in some phase of materials technology. It is too early to say whether this type of collaboration is effective or even possible, either educationally or technologically. Its cost is such that only considerable technological success could justify its further proliferation. Considered solely as education it is probably too expensive, and this is the basic difficulty with many of the proposals that are made for improving the real environment of the engineering school.

Conclusions

I am afraid that I am better at posing dilemmas than at proposing solutions. The only solutions I have to offer are flexibility, experimentation, and feedback—flexibility in curriculum and organization, experimentation with many different mechanisms, and much more effort at thoughtful evaluation of the result. It seems to me, however, that intellectual standards must be primary. Granted that engineering faculties will be increasingly populated with physicists, mathematicians, chemists, and life scientists, just as medical faculties are becoming populated with biologists and chemists, they must not be allowed to become simply the refuge of people who cannot make the grade in their own disciplines. Here, I think, the basic disciplines within the university setting have a responsibility they have tended to neglect—the responsibility for taking the dilemmas of engineering education seriously and helping to provide a hospitable intellectual and moral climate for the basic scientist who chooses to cast his professional lot with engineering and applied science.

Essentially full text of the banquet address at the New England Region Meeting of the ASEE Conference, Yale University, New Haven, Conn., October 21, 1966.

Special Conference Report

Pattern Recognition 1966 IEEE Workshop

George Nagy *International Business Machines Corporation*

Self-Organizing, Bionic, Heuristically Programmed, Pattern Recognizing, Learning, Neuronal, Cybernetic, Goal-Seeking, Problem-Solving, Microprogrammed, Multiprogrammed, Multi-Input, Redundant, Adaptive, Self-Repairing, Self-Teaching, Time-Sharing, Self-Reproducing, Cluster-Seeking, On-Line, Trainable, Stochastic, Kilomegacycle, Optimal, Artificially Intelligent, Synnoetic Computing Machines—these terms comprised one speaker's list of key words necessary to describe the range of topics discussed at a recent "happening" (the chairman's characterization) instigated by the Pattern Recognition Subcommittee of the IEEE Computer Group.

Some 52 pattern classifiers in search of recognition, divided evenly between private industry (26) and other categories—universities (15), government agencies (5), and nonprofit laboratories (6)—attended sessions of the Workshop on October 24–26, 1966, at the El Conquistador and Dorado Hilton Hotels, Puerto Rico. The invitations had been mailed on the basis of recommendations by the members of the subcommittee as well as in response to inquiries resulting from an announcement in the September 1966 issue of IEEE's *Computer Group News*.

Fully a third of the formal presentations dealt with some form of character recognition, showing that this endeavor still remains the most active single area in pattern recognition. The characteristics of several commercial print readers as well as specific applications, such as the reading of social security forms, zip codes, driver's license applications, editorial copy for typesetting, and military allotment forms, were reviewed both from the users' and the manufacturers' points of view. The economic aspects of processing rejects and substitution errors were also discussed.

Reported development work on systems aimed at handprinted characters seemed to favor the on-line approach with immediate display of the interpreted character, which allows the user to make on-the-spot corrections and to adjust his style to suit the recognition logic. The one off-line project described relies on the context inherent in a programming language such as Fortran to keep the error rate within acceptable limits.

Further contributions in character recognition consisted of new algorithms designed to improve maximum-likelihood decisions based on features by taking into account the interfeature statistical dependences and the Markovian properties of natural language.

Other applications-oriented presentations covered holographic techniques for fingerprint recognition, polynomial decision boundaries for electrocardiograms, automated photometric blood cell analysis, adaptive networks for sonar phased antenna arrays and for aerial photoreconnaissance, a sequential decision model for blackjack, graphic input for computers, the superposition of flight paths on contour maps, and the analysis of three-dimensional projections. Among these, the electrocardiogram analysis seems closest to practical applicability. Several of the other projects, notably the work on fingerprints, blood cells, sonar, and graphic input, also make use of realistic data sets.

The outline of a general-purpose pattern recognition and manipulation system was also presented. Some of the subroutines used for pattern segmentation and description are already operational.

Several of the theoretical papers described "unsupervised learning" schemes. A Fourier series expansion was applied to the decomposition of multivariate normal distributions with a finite number of samples, and a signal identification problem (additive noise, number of messages unknown) was solved by means of a correlation integral equation. Markovian statistics were drawn upon

George Nagy is a research staff member, IBM Watson Research Center, Yorktown Heights, N.Y.

to investigate the convergence properties of an error-correcting training rule on a partially mislabeled sample set.

Other statistical contributions reported improved estimation methods of a posteriori probabilities from additional samples, and new bounds on the risk on the nearest neighbor decision rule with an infinite number of samples.

The capabilities and limitations of two-layer threshold nets (simple perceptrons) were clarified in a paper linking recognizable geometric properties with the maximum number of connections to a unit of the first layer (the order of the net).

An interesting review paper described the difficulties encountered in various approaches to "clustering," examined the relationships among the procedures in use, offered standard distributions for the evaluation of clustering algorithms, and listed actual and potential applications. The use of several varieties of two-dimensional projections of n-dimensional spaces of interest was suggested as an aid in the study of multimodal distributions.

Of interest to psychologists and engineers alike was a "grass fire" (self-propagating) global pictorial processing scheme designed to extract fairly abstract properties, including connectivity. Even longer-range strategy for pattern recognition was represented by a report on studies of human perceptual phenomena.

Philosophic advice from one quarter suggested that efforts should be made to find "impotence conditions" in pattern recognition similar to those deduced in other disciplines.

"Ask not what a pattern will do for you," urged an advocate of the definition of suitable features through synthesis rather than analysis.

Since most of the papers presented are now awaiting publication, and since preprints of many are available from the authors, there will be no further attempt here to summarize each individual contribution. Rather, we shall now endeavor to review some of the issues that generated discussion, argument, controversy, and excitement.

The matter of debate ranged from metaphysics (pattern recognition as a problem in pattern recognition) to practical detail (how to find the gap in C). The discussion here will follow a descending order of abstraction rather than a chronological sequence.

Pattern recognition and the scientific method

Thesis: Pattern recognition is the central problem of psychology and philosophy. The general aim is to build up a hierarchy of computer-recognizable patterns, somewhat parallel to human concepts, in terms of which any complex may be analyzed. *Ad astra per aspera.*

Antithesis: Why use computers to perform "intellectual tasks?" At best the computer is a handy tool for arithmetic, though even there its usefulness is often overrated; a day or two with pencil and paper sometimes yields more insight than months of simulation. Nothing is more fatuous than the current trend of correlating everything with everything else, in the vain hope of understanding obscure causal relationships.

Thesis: Computerized clustering techniques may prove to be powerful aids in structuring the universe, and have already suggested improved classification schemes to mathematical taxonomists in several branches of biology.

Antithesis: The severest limitation encountered by computer programs aspiring to intelligent activity consists of the very restricted scope of knowledge available to them. It is difficult to conceive of a clustering algorithm, no matter how sophisticated, capable of arriving at the fundamental distinction between vertebrates and invertebrates without the programmer's having been aware of the importance of backbones as a feature for consideration at the outset.

Synthesis: Higher-order languages, leading to the description, analysis, and synthesis of arbitrary patterns by computers, must continue to be developed. In the meanwhile, computers in pattern-recognition circles must earn their rent by performing a variety of less glamorous tasks, such as selecting "features" for character readers, verifying assumptions and evaluating statistical parameters for specific data sets, working quickly through complicated logic trees and combinatorial calculations, facilitating man–machine interaction in engineering problems and information retrieval, and testing hypotheses in perceptual psychology.

Context

Thesis: The easy way to take advantage of context in natural languages is to use Markovian properties in terms of bigram and trigram frequencies.

Antithesis: Chomsky has shown that natural languages are not Markov processes. Indeed, Markov processes seldom occur in nature. Why not use syntactical rules combined with dictionary look-up techniques?

Thesis: Markov processes, in addition to providing an analyzable model to any required degree of approximation, do occur, as in thermal noise. A more complicated model would not be tractable even with the use of high-speed computers.

Antithesis: The correction of errors in graphic input should occur at a level much higher than that of the symbols constituting the communication link between man and machine. In graphic input as used for electron circuit design, for example, the machine should analyze the system represented in terms of Maxwell's equations rather than look for wiggly lines without connections to both ends.

Thesis: There are many levels of "well-formedness." Begin at the lowest usable level, that of the symbols, instead of trying to teach the machine the complexities of the real world.

Antithesis: In many instances the lowest level of abstraction is not the most economical in terms of the information to be transmitted to the computer. In Samuel's checker-playing program, for example, the moves in the training games used to improve the program's strategy are evaluated by the weighting algorithm of the program itself in order to detect keypunching errors.

Synthesis: The introduction of context, at some level, is necessary in all but the simplest problems. The systems point of view engendered by contextual considerations may occasionally reveal that the task in hand is not, after all, best suited for pattern-recognition techniques.

Parallel vs. serial processes

Thesis: The flexibility of stored-program digital computers more than offsets any advantages in speed that may be obtained from special-purpose hardware.

Antithesis: Analog methods, as in holography, can successfully attack correlation problems of a magnitude not even conceivable in terms of present-day general-purpose equipment.

Thesis: In complicated problems the sample size is often too small to allow training by example; in such cases, we must resort to training by description, which is essentially a serial process requiring an adequate formal language to convey the necessary heuristics.

Antithesis: In complicated problems the elements of solution are usually insufficiently well understood to allow the trainer to specify them explicitly to the machine; thus, we must resort to adaptive nets and error-correcting algorithms, which are well suited to classification tasks on poorly structured patterns.

Synthesis: The distinction between "serial" and "parallel," while superficially obvious, is difficult to define rigorously. Speaking loosely, it seems likely that systems designed for complicated pattern-recognition functions in real-world environments will be parallel at the front end, and serial, or heuristic, at the back, with suitable feedback paths in between. Some biological and psychological evidence appears to favor this view.

Features and templates

Thesis: For multifont character recognition, where several hundred distinct shapes are involved, templates are impractical.

Antithesis: The extraction of features necessarily degrades the information content of the images; hence, templates are inherently superior.

Thesis: Templates are too sensitive to noise when the distinguishing feature between two classes is only a small fraction of the total area of the characters.

Antithesis: There are more ways than one to skin a cat. To be successful, template matching must be preceded by adequate preprocessing techniques, including size normalization, line thinning, random-dot elimination, and unskewing.

Definitions wanted

For performance specifications for commercial print readers, should a given set of character shapes be referred to as a single "font" even when produced by different printing mechanisms?

Should the word "feature" be reserved for a distinct geometric subset of a pattern, or could it also be used to designate more complicated attributes such as connectivity properties? Another suggested definition would use "feature" for any quantity resulting from any data processing between the original image and the final decision.

The opinion was expressed that some attempt should be made to differentiate among various processes on the basis of the underlying mathematical formulation. The word "learning" appeared too pretentious for simple distance-minimizing algorithms; these, in turn, were deemed a cut above "tracking" algorithms, operating on slowly varying systems, which avoided the problem of local traps by remaining close to the real minimum at all times.

Overview

To a casual listener from another field of endeavor, the claims of pattern recognition to the status of a cohesive discipline, as substantiated by this Workshop, might have sounded a trifle exaggerated. Despite the definite community of interest among the participants, evidenced by animated technical discussions in pool and surf, at breakfast and over drinks, in rain forest and casino, as well as in the meeting hall, it is clear that heuristic methods have little to do with adaptive nets, or either of them with commercial print readers. To be sure, psychologists willingly discuss the effects of retinal stabilization with students of asymptotically efficient estimators, and biologists interested in chromosome counts seek the advice of practitioners of holography, but this intercourse might seem to be more in the nature of a spirited conversation among intelligent people of reasonably broad interests than an interchange of technical information designed to advance specific research goals.

This diversity of interests is particularly apparent to anyone attempting to organize a university course, graduate or undergraduate, covering the general area of pattern recognition. A quick survey of existing courses reveals little agreement as to the basic topics to be included in such a course. Thus, perhaps fortunately, the next generation of pattern recognizers is likely to be as heterogenous a group as the present set.

Future concourse

From one point of view, the rather chaotic state of the art only provides additional impetus for workshops where the much heralded cross-fertilization can take place. Then, as more and more recondite problems are tackled, truly interdisciplinary solutions will be developed, with a fusion of the many interesting techniques now being perfected in dark corners. This, at best, is the pious hope; at worst, workers in different corners will discover that they are weaving the same web.

Several participants at the Workshop raised the question as to what is proper material for presentation at such a meeting. Should completed work be reviewed even when already published elsewhere, with the object of stimulating discussion? How numerous and how long should the formal presentations be and how much detail on a given project is of interest to the whole group? Would panel discussions on set topics be helpful? Does the overwhelming response to the call for ten-minute papers, with no advance abstracts required, indicate that this formula will be as successful as in other disciplines? The IEEE Subcommittee on Pattern Recognition welcomes all suggestions.

on, or fired, by means of a sharp pulse of electric current applied to a control electrode, or "gate," and not by a beam of light. It thus has two limitations that the new device was specifically conceived and developed, by scientists at the Westinghouse Research Laboratories, to overcome: danger of damage to the device from the initial surge of current that turns it on, and difficulty in firing a series of units all at the same instant in time. The new technique eliminates the gating electrode, thus changing the three-terminal device into a two-terminal LASS; isolates the switch electrically from its firing source and simplifies the firing circuitry; allows faster-rising surges of current at turn-on of the switch without danger to it; and fires a stack of switches at the same instant by piping to each a simultaneous pulse of light.

The light used in firing LASS comes from a solid-state laser of gallium arsenide. This laser radiates in the infrared region, which means that the switch cannot be turned on accidently by sunlight or ordinary artificial light. The infrared radiation is coupled to the switch through a flexible fiber optics light pipe about half the diameter of a lead pencil. One GaAs diode can fire about 100 switches.

The working element in LASS is a four-layer silicon wafer that is smaller and thinner than a dime. The structure is such that a large area of the top layer (cathode) is flooded directly with infrared light, which is of the proper frequency to penetrate the silicon material and make it electrically conducting.

University of Illinois offers computer science program

The Department of Computer Science of the University of Illinois has announced a graduate program in computer science leading to the degrees of master of science and doctor of philosophy in computer science. Research areas include but are not restricted to digital computer arithmetic, switching and automata theory, circuit design, computer organization, computer applications in physics, software systems and languages, numerical analysis, pattern recognition, and information retrieval.

A limited number of assistantships and fellowships are available. Details of the program may be obtained from Prof. John R. Pasta, Head, Department of Computer Science, University of Illinois, Urbana, Ill. 61801.

Technical correspondence

Relativity and electricity

In a recent article[1] and a thought-provoking textbook,[2] R. S. Elliott has suggested basing the electromagnetic theory on the Cavendish–Coulomb electrostatic inverse square experimental law and on Einstein's special theory of relativity, as a generalization of an earlier idea by Page.[3] Besides being nonconventional (which I admire!) and not in accordance with historical scientific progress (with the earliest law and the latest theory as foundations), it seems to have a basic philosophical flaw. The special theory of relativity by Einstein[4] is based on two postulates:

1. The laws of physics (including electrodynamics and optics) are covariant in all inertial frames of reference.
2. Light always propagates in empty space with velocity c.

One uses light pulses in order to derive the Lorentz transformations in the special theory of relativity. The second postulate definitely implies the existence of light waves (which propagate in vacuum) and, therefore, the existence of the wave equations and Maxwell equations that describe its behavior mathematically. It seems reasonable that one should *not* base the derivation of the electromagnetic theory and Maxwell equations on the special theory of relativity, because they are required as a basis for the second of its postulates mentioned above (if not also the first one).

Another important aspect of the problem should be pointed out. One could develop a "gravitational theory" based on the Hooke–Newton[5,6] gravitational inverse square experimental law and on Einstein's special theory of relativity, along the same lines as done by Elliott[1,2] for the electromagnetic theory. A "gravitational magnetostatic" field,* "gravitational Maxwell equations," and "gravitational waves" could be derived along the same lines.

A development of the "gravitational theory" suggested above, though based on almost identical foundations as the ones suggested by Elliott,[1,2] is in direct contradiction to the general theory of relativity[7-14] (the theory of gravitation) developed by Einstein, with its tensor-analysis mathematics and non-Euclidean space ideas. (The general theory of relativity does show, however, that weak gravitational fields in vacuum propagate as waves with the velocity of light.) It seems that Elliott should accept a priori the experimental postulates of Biot-Savart and Faraday as evidence for the existence of the electromagnetic theory developed by him; by the same token, the lack of those experimental laws or their equivalent in the "gravitational theory," developed on almost identical foundations, should form the basis for its rejection, besides its contradiction to Einstein's general theory of relativity. A unified theory of gravitation and electrodynamics seems to be the only general solution.

In this context, it might be mentioned that the electrostatic potential $\phi(R)$ due to a point charge q, could be looked upon as having the general functional form $f(R)$ with the Laurent's expansion:

$$\frac{1}{q}\,\phi(R) = f(R) = \sum_{-\infty}^{+\infty} c_n R^n \qquad (1)$$

where c_n are universal constants. All the terms in (1) except $c_1 = {}^1/_{4\pi\epsilon}$ could be neglected for the usual distances $A <<< R <<< G$ where A = atomic dimensions and G = galaxial dimensions. This seems to put the inverse-square Cavendish–Coulomb experimental law of electrostatics on more rational grounds.

H. Unz
University of Kansas
Lawrence, Kans.

*Of course the "gravitational magnetostatic" field cannot be detected as long as there is no compensating[2] (negative) mass in existence.

1. Elliott, R. S., "Relativity and electricity," *IEEE Spectrum*, vol. 3, pp. 140 152, Mar. 1966.

2. Elliott, R. S., *Electromagnetics.* New York: McGraw-Hill, 1966.

3. Page, L., "A derivation of the fundamental relations of electrodynamics from those of electrostatics," *Am. J. Sci.*, vol. 34, pp. 57–68, 1912.

4. Einstein, A., "On the electrodynamics of moving bodies," *Ann. Phys.*, vol. 17, pp. 891–921, 1905; see also Lorentz, H. A., Einstein, A., Minkowski, H., and Weyl, H., *The Principle of Relativity.* New York: Dover, 1923.

5. Mason, S. F., *A History of the Sciences.* New York: Collier, 1962.