VOLUME 4 □ NUMBER 1 JANUARY 1967

IEEE spectrum

features

Departments: *please turn to the next page*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

the cover

A completed section of the new welded-rail railroad line between New York and Washington that will provide high-speed passenger service beginning in October. General problems in moving large masses of people by various means are discussed in the article beginning on page 60.

Spectral lines

The way to a more professional status. From time to time we receive correspondence discussing the professional status of engineering, often reflecting a certain amount of despair or cynicism, and sometimes suggesting changes in Institute policy or activity. These letters are most welcome; they often provoke considerable discussion, and occasionally even action by IEEE. Some of my opinions on these matters are given in the following. They carry no official sanction, but are open for criticism. If they provoke more discussion, so much the better.

One common complaint is that the Institute does not do enough to uphold and improve the professional status of its members; that more should be done to influence legislation and establish practices and particularly to put some controls on the way the title "engineer" is used. The American Medical Association has been cited as an example of what a strong professional society can do in this area.

Should professional societies such as IEEE take a stronger stand in setting standards for engineering practice, as the AMA does for the medical profession? I believe to do so would be to take a backward step, toward trade unionism rather than toward a more professional stature. A distinguishing mark of the professional (in the learned professions) is independence of action, personal (individual) responsibility, and the use of intellectual skills. I cannot see how we could empower the IEEE with political strength without surrendering much of our highly valued independence of action. The Institute does exercise considerable influence in setting technical standards and cooperating with other organizations in many activities on a purely voluntary basis, but to attempt to influence politics, or make mandatory regulations governing our members, would go far beyond our charter. On the contrary, we should encourage a greater individual responsibility and greater intellectual skill to improve our standing in the profession, and thus collectively improve the status of the profession.

The loose usage of the title "engineer" has always been an annoyance, and to some it is a serious matter that the general public often cannot distinguish between a coil winder and a graduate engineer. Actually, however, engineering is not in any unique difficulty with this kind of semantics. Pretenders in all professions have used unearned titles. The only legal protection is in regard to practice, not titles.

Quite aside from the semantics of the situation is the professional nature of the practice of engineering by the legitimate members of our profession. As our correspondents suggest, much of what some of our members do is quite menial and there is no sharp distinction between the professional activities of the graduate engineer and the work of the skilled technician. One purpose of professional engineering licensing in many states is to help separate the sheep from the goats, and it does effectively restrict the private practice of engineering; but it has not generally been accepted by the majority of engineers in industry, whose activities are really determined by their employers.

Should practicing engineers be required to pass licensing examinations, as some of our correspondents seem to think? The majority of electrical engineers have indicated their opinions by the regularity with which they have *not* sought professional licensing. Licensing was instituted primarily as a protection for the public, but few of us ever deal directly with the public. Industry has its own measure of worth and is not much influenced by professional licenses. Personally, I do not see how licensing would help. Many irresponsible people have passed license examinations.

One correspondent suggests that the engineer is less "professional" than the scientist, doctor, or lawyer because he surrenders his freedom of action and submits to industrial standards of work and remuneration; he is "constrained in ways no true professional will tolerate." Certainly, few of today's large engineering projects would prosper without the sacrifice of individual prerogatives to the joint effort. Is this *the* important factor, or does not one take on a more professional stature in being a cooperative part of a large venture, in the "creative application of existing knowledge?"

Another suggestion is that the engineer often submits to the necessity of performing menial tasks that would be beneath the dignity of one in another profession, because industry does not provide enough assistance. This is an interesting point, but I doubt if the writer had very carefully analyzed the activities of the doctor or lawyer. I believe that the greater part of their time is spent on quite routine, mechanical, and often messy tasks, involving information retrieval (including diagnosis) and manual or oratorical skill. The former task—information retrieval—is being taken over by computer and manual and oratorical skills are also becoming machine operations; for example, X rays, laser, diathermy, cauterizers, water-spray drills, and public address systems. The things that make the physician and lawyer different from a skilled technician are intellectual proficency, judgment, and responsibility—and this is also true of the engineer.

This, I think, is the moral: *Professional stature is not determined so much by what one does, but by how one does it.* Public image is quite another matter.

C. C. Cutler

Authors

Electronics and the traffic engineer (page 50)

Wilbur S. Smith (SM), founder and head of Wilbur Smith and Associates, has been a faculty member at Yale University Bureau of Highway Traffic since 1943, serving as associate director until 1957 and, since then, as research associate. He is chairman of the Advisory Committee for the Transportation Center at the University of South Carolina, a member of the Advisory Committee on Urban Transportation Planning for the Polytechnic Institute of Brooklyn, and a member of the Advisory Board of *Ekistics*, the journal for the Athens (Greece) Technological Institute. He serves as chairman of the Board of Directors of Freeman, Fox, Wilbur Smith and Associates, London, and is president and chairman of the board of the Eno Foundation for Highway Traffic Control. A consultant to the Air Force, he is also a member of the Board of Trustees of the World Safety Research Institute, a member of the National Defense Executive Reserve for Public Roads, and director of the International Road Federation. Mr. Smith is a veteran of over 30 years in traffic engineering and transportation research.

Automatic equalization using transversal filters (page 53)

Harry Rudin, Jr., (M) received the bachelor of engineering degree in 1958, the master of engineering degree in 1960, and the doctor of engineering degree in 1964, all from Yale University, New Haven, Conn. His doctoral dissertation, based on research in the area of statistical communication theory, is concerned with various techniques of linear filtering of non-Gaussian random processes.

He was a teaching assistant at Yale University and, from 1961 until 1964, served as an instructor there. He joined Bell Telephone Laboratories, Inc., Holmdel, N.J., in 1964, becoming involved with various problems in the area of data communication. Recently, his efforts have been concentrated in the area of automatic equalization, particularly generalized equalization techniques, and, also, with the application of these techniques to other communication problems, such as network synthesis and echo cancellation.

Dr. Rudin served as a member of the Executive Committee of the IEEE Connecticut Section from 1962 to 1964.

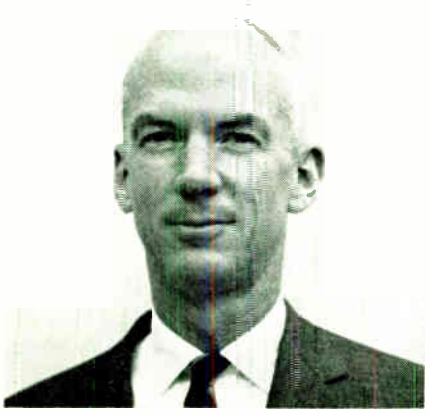
Tomorrow's mass rapid transit—available today! (page 60)

Deane N. Aboudara (M) received the B.S. degree in electrical engineering from California State Polytechnic College in 1950 and, in that year, joined the General Electric Company as a test engineer. In 1953 he became a transportation specialist trainee at the Locomotive and Car Equipment Department, Erie, Pa. Three years later, transferred to San Francisco, Calif., he served as transportation specialist for industrial, railroad, and transit properties. He became application engineer for Automated Material Control Systems at the Industry Control Department, Roanoke, Va., in 1962 and then, in 1964, joined Bay Area Rapid Transit, San Francisco, as electronics and equipment design engineer with responsibility for one of the major engineering groups dealing with design, coordination, and evaluation of activities in the fields of power, automatic controls, fare collection, communications, and equipment. He has held various committee memberships for transportation and general application technical groups.

C. William Woods (M) received the B.S.M.E. degree from Drexel Institute of Technology, Philadelphia, Pa. After serving with the Army, he spent several years in precision manufacturing, becoming interested in automation in 1955. He designed equipment for the automatic assembly of printed circuit boards, worked in microwave equipment design, and later managed the research and development of an advanced infrared, air-to-air guided missile, including design of hot-gas-actuated servomechanisms and electro-optical tracking system and solid-state electronics designed to survive the full range of military environment.

Since joining the Union Switch and Signal Division of WABCO in 1964, he has been responsible for the demonstration of new concepts of automatic control for mass transit, the use of digital computers in railroad operation, and the application of new technologies to equipment designed for the transportation industry.

Raymond S. Silver, director of plans and programs for the Advance Data Systems Division of Litton Industries, is responsible for the planning and implementation of the program for Revenue Control Systems. In this connection he has performed systems analyses and been involved with equipment development for London Transport, the San Francisico Bay Area Rapid Transit District, the Long Island Rail Road, and various other systems throughout the world. In addition, he has developed a construction automation system, and an in-house program for application of digital computer techniques to construction estimating and modular building. Prior to joining ADS, he was with International Business Machines for 13 years, serving as manager of the Time Equipment Division and then as manager of Advanced Systems, Western area, Federal Systems Division. In the latter position he was concerned with data processing systems and applications used in space guidance, missile tracking, and communications.



John C. Beckett (F) was graduated from Stanford University in 1938 and, three years later, received a graduate degree in engineering from the same university. After a tour of duty with the Navy, during which he achieved the rank of commander, he became chief engineer with Wesix Electric Heater Company. He directed the engineering planning for rapid transit in the Bay Area and held the position of vice president of the San Francisco Bay Area Rapid Transit District from its beginning until June 1960. He has served as chairman of the IEEE San Francisco Section, as Section WESCON director, as chairman of the AIEE 1962 Pacific Energy Conversion Conference, as chairman of the AIEE Committee on Electrical Techniques in Medicine and Biology and as president of the Electric Club of San Francisco. Presently, he is Government Relations Manager of the Hewlett-Packard Company, having previously served as general manager of the company's Paeco Division.

Deep-space optical communications **(page 75)**

E. Brookner (M) received the B.E.E. degree from the College of the City of New York and, from Columbia University, the M.S. and Dr.Sc. degrees in electrical engineering. He joined the staff of Columbia University Electronics Research Laboratories in 1953, becoming involved with noise studies and the design of long-range radars. Subsequently, as project engineer at the Federal Scientific Corporation, he was engaged in the analysis and development of coherent radar processors, such as the coherent memory filter. In 1962 he joined the Space and Information Systems Division of the Raytheon Company and studied such topics as decoy design and decoy discrimination. Presently, he is serving as principal engineer and is engaged in the design of experiments for the characterization of millimeter and laser communications channels involving satellite-to-ground and ground-to-ground links. He is also participating in the design of a re-entry wake measurement radar.

M. Kolker (M) received the B.E.E. degree in 1950 and the M.E.E. degree in 1951, both from Rensselaer Polytechnic Institute. Between 1951 and 1955 he served in the Air Force as a research and development project officer responsible for advanced development of bomber defense systems, including fire control, ECM, and missile guidance. From 1956 to 1962, serving General Electric Company, Mitre, and then Cubic Corporation, he was engaged in system engineering analysis and design in the fields of missile guidance, data processing, communications, command, and control. In 1962 he joined Raytheon Company, assuming responsibility for leading systems analysis studies for the application of technology to the problems of command and control, arms control, and disarmament. Recently, he has been responsible for directing advanced studies in developing applications of the laser to space system problems. He is presently working on space laser altimeter development.

R. M. Wilmotte (F), presently a private consulting engineer in Washington, D.C., received the B.A., M.A., and Sc.D. degrees from Cambridge University, England. As a consultant to the broadcasting industry in the field of communication, he built the first broadcasting station directional antenna for the protection of the service area of another station. He also engaged in work on antennas, proximity fuses, fire control, and radar. In 1959 he joined the Advanced Military Systems Group at RCA, Princeton, N.J. Subsequently, he served as project manager for Reley, the first NASA communication satellite. In recent years, he has served as a consultant on reliability management and on electrooptical signal processing. The first to obtain wide-band correlation with electrooptic sonic delay line, Dr. Wilmotte holds over 40 patents and has published 40 papers and articles. He has received the Bureau of Ordnance Development Award.

Computing reliable power spectra (page 83)

Paul I. Richards (A) was graduated from Harvard University. Cambridge, Mass., in 1947. The recipient of a doctoral degree in physics, his thesis dealt with the synthesizing of electrical impedances with sections of transmission line. After graduation he served Brookhaven National Laboratory, Upton, N.Y., for five years and was concerned with theoretical quantum mechanics and experimental mass spectrometry. In 1954 he joined Technical Operations, Inc., Burlington, Mass., where he did research on radiations from nuclear weapons and on machine computations of complex blast waves. He was appointed a Corporate Fellow of Technical Operations, Inc., in 1962 and, since that time, has been engaged in undirected research. Dr. Richards also serves as editor for the Society for Industrial and Applied Mathematics (SIAM). A member of Phi Beta Kappa and Sigma Xi, he has published more than 30 technical papers and two books, a manual of mathematical physics and (with I. T. Richards) a handbook on expository writing.

Planning and operation of a large power pool (page 91)

R. G. Rincliffe (SM), after receiving the B.A. degree in 1921 from Yale University and the M.S. degree in chemical engineering from Massachusetts Institute of Technology in 1923, immediately joined the American Gas Company as an engineering assistant. Subsequently, he served as industrial gas engineer, construction engineer of the coke oven plant at Chester, assistant superintendent of coke ovens and, in 1928, was appointed superintendent of gas production of the Delaware Division of the Philadelphia Electric Company. In 1945 he became vice president of electric operations and, five years later, executive vice president and director. He has served as president and as chairman of the board and, since 1963, has been chairman of the executive committee. In addition, he is president and director of the Philadelphia Electric Power Company, the Susquehanna Electric Company, and the Susquehanna Power Company. He is the recipient of honorary degrees from Pennsylvania Military College, Villanova University, and Saint Joseph's College.

IEEE in Latin America

W. K. MacAdam *IEEE President*

Interest in the IEEE, and IEEE's activities, are growing rapidly in Latin America. This is the predominant impression that I brought back from my recent 25 000-km trip to visit the Institute's Sections and Student Branches in Central and South America. I was accompanied by Francisco Hawley of Mexico, recently appointed Assistant Director of the new Latin American Region 9 of IEEE, and at Rio de Janeiro we were joined by Region 9 Director G. J. Andrews of Buenos Aires.

Our journey, which took place during the last two weeks of November, included stops in eight cities in seven different countries, where we talked with Institute members and with university faculties and students. The timing was exceedingly important. First, the inclusion of all of Central and South America and the Caribbean in a new Region 9 brings a new community of interest to this area along with a new Regional Committee to help manage its affairs and give it a stronger voice in Institute planning and operation. Second, we are launching a Regional technical journal to be produced by and for the members of Region 9. Even more significant is the fact that this journal will be published in Spanish, except for articles contributed by Brazilian engineers, which may be in Portuguese. Here, we have geared our publications policy to the particular needs of the members of Region 9; this action has no present counterpart in any other Region. These new developments, coupled with the expanded services starting in 1967 for Regions 8, 9, and 10, make it particularly appropriate to undertake discussion at this time with the Sections in an effort to strengthen our bonds of mutal interest.

The first stop on my itinerary was Mexico, where we have a very strong and active Section. Here is where our Latin American journal will be published and where its Editor and Publication Manager are located. However, the editorial contributions will come from all Latin America, with each country having its own Associate Editor. The Mexico Section extended its traditionally warm welcome, and there were constructive discussions with members and the Section Executive Committee, and with the faculty of the University.

From Mexico City we flew to Caracas, Venezuela, for a two-day visit. IEEE members here are enthusiastically planning a new Section and the university faculties are active in forming Student Branches. High on their list of advantages to be gained from an active part in Institute affairs is the new journal, which ultimately is expected to be self-supporting and will be distributed free of charge to all members and students.

Next on our program was a brief stop at Rio, a chat with the past Section Chairman, J. A. Wiltgen, and then we were off to Sao Paulo, Brazil, to present the charter for the Section recently organized by Carlos Lohmann and his associates.

Buenos Aires has had a large, active IEEE Section for many years and its hospitality and interest were characteristic of its growing membership. Totaling 155 at the beginning of 1966, the membership is expected to exceed 300 by March. About half of this will probably be students, which bodes well for the future vitality of the Section. After two days of spring in this beautiful city, we headed west over the Andes to Santiago, Chile, where the local Section is struggling for effective operation within the elongated geography of the country. We traded ideas and suggestions for future progress with the Section members and also held meetings at two universitites where Branches are being formed. Institute members were our hosts on a vist to the NASA Space Tracking Station, and they were able to give us a most encouraging report of the formation of a new Chapter for Aerospace and Electronic Systems.

Our next direction was north, to Lima. Peru, for an hour's stopover with Cesar Pera, communications superintendent of Faucett Airlines, who hopes one day to promote a Peru Section. Another stop was in picturesque Quito, Ecuador. Situated some 25 km from the equator, its 3-km-high elevation gives it 12 months of springlike climate. Oddly enough, Quito is the source of Panama hats. Our final destination on that long day was Bogotá, Colombia, which is situated 2500 meters above sea level on a plateau surrounded by mountains. Here we met with members of the Colombia Section and with the faculty of the University of the Andes, where institution of a Student Branch seems imminent. The Section has more than 100 members but has had difficulty in maintaining an active program, principally because of the job transfer of Section officers. However, the new Regional organization and publications program should provide an incentive for an increase in the Section's activities and membership.

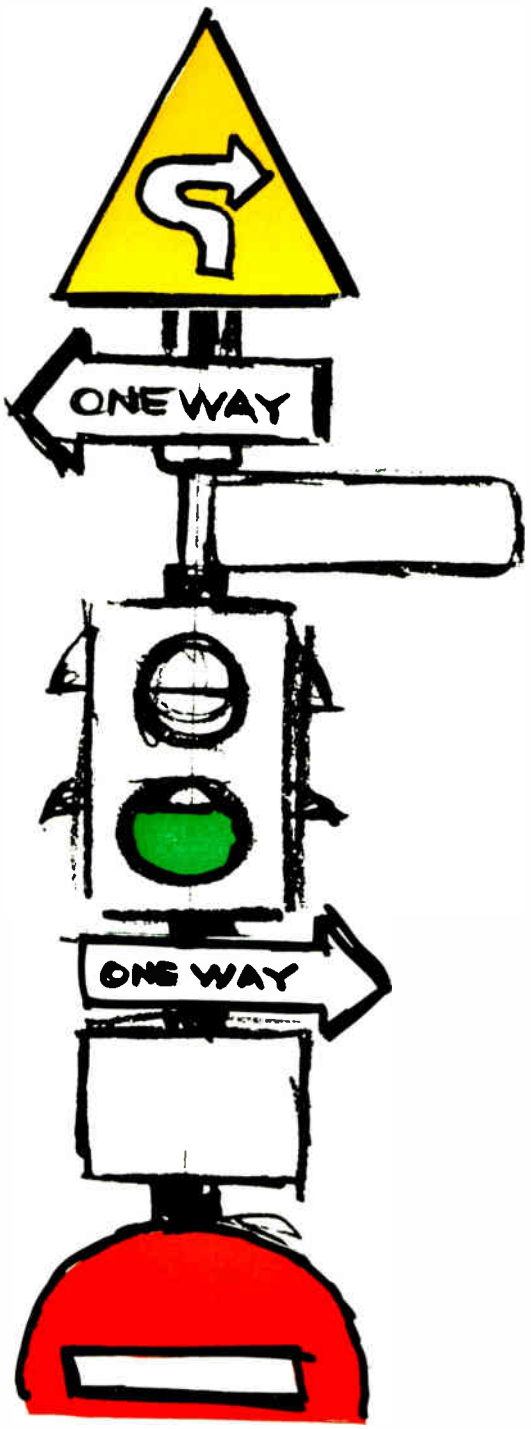
When I arrived back in New York I brought with me a strong feeling of the enthusiasm and support of our Latin American members for their new Region 9 and their new publication. Together with our Regional Directors, I had learned much of the problems and needs of members and students in this area. I believe that this trip will prove an excellent and timely investment in the future of the Institute in Latin America.

Electronics and the traffic engineer

There is an endless opportunity for today's transportation engineer to utilize electronic techniques and devices in finding practical solutions to our increasingly complicated traffic problems

Wilbur S. Smith

Wilbur Smith and Associates



As vehicular traffic increases both in density and complexity, particularly in urban areas, effective control systems to minimize delays become more and more essential. The traffic engineer's problem is twofold: to optimize existing facilities and to plan new ones. Some of the many ways electronic tools are being employed to help solve these problems are described in this article. The challenge is basically one of continued coordination of the several disciplines common to the engineering of transport systems.

Within the Institute of Electrical and Electronics Engineers, there are 31 Groups, ranging from "aerospace and electronic systems" to "vehicular communications." An examination of these fields reveals that a *traffic* engineer or a *transportation* engineer might have a vital interest in the subject matter from approximately ten of these Groups. In his surveys, plans, designs, surveillance, and control of modern urban and rural traffic facilities, the transportation specialist needs to have some knowledge of such subjects as television and radio, communications, automatic control, computers, telemetry, ultrasonics, vehicular communications, and instrumentation and measurements.

Operating as a completely interdependent system, where one component significantly affects another, the control and regulation of traffic has become as scientific as the management of trains or public communications systems.

Modern traffic is complex

The rapid growth in automobile travel has produced demands for road capacity and safety that challenge the very best abilities of transportation specialists. Maximum use must, of course, be made of all existing facilities—those available for both the movement and storage of vehicles. In addition, new facilities must be planned to provide maximum returns in transportation services for the scarce dollars invested. In other words, the traffic

engineer must concern himself with achieving the optimum efficiency in present traffic flows and in sound plans for long-range solutions. There are, nevertheless, many overriding problems: safety, aesthetics, social impacts, and economic forces. Obviously, full use must be made of every possible scientific aid.

The complexity of modern transportation networks and the diverse classes of vehicles traveling on the facilities require that the most sophisticated measurement, analysis, surveillance, and control methods be employed. Several such techniques, now within the electrical and electronic disciplines, are being taxed anew by traffic demands.

The traffic stream appears to be analogous to many other flow problems, such as those in paper mills, refineries, and water supply. Just as the automatic control of these processes requires sensing devices to measure flow rate, temperature, fluid density, and other properties, so must the traffic stream have its sensing devices if automatic control is to be achieved.

Measuring the problem

Traffic engineering journals are replete with advertising and technical articles on sensing devices of many different types. Much of the equipment now used in traffic sensing has, since World War II, gone through an evolution from relatively simple electromechanical devices, to electronic tube and relay devices, to the present solid-state devices now being offered. For almost 20 years, engineers relied on two basic types: the magnetic detector and the pressure-sensitive detector. There are now available: pulsating radar, Doppler-effect radar, Doppler-sonic devices, capacitive loops, inductive loops, and magnetometers. There also have been reports on experimental work in which lasers are used to detect, classify, and measure vehicle speeds. The classification task has been the most difficult to achieve.

In traffic work, the engineer often needs to know more than just the quantity of traffic. Studies of the "quality" of traffic flow are also important, requiring measurements of such characteristics as speed–volume relationships, spacings, and placement of vehicles in the stream. Here again, the traffic specialist is being aided by electronic devices capable of recording vehicle presence, lane occupancy, and density (or vehicles per unit of space). This recording equipment can be connected to on-line computers to help develop and record the traffic flow data, even for instantaneous analysis.

More recently, it is reported that computers can be used to help evaluate aerial photographs—thus electronically scanning a series of exposures and tracing an individual vehicle through a complex series of maneuvers. Techniques like this might soon replace time-consuming and expensive manual processes used in measuring traffic performance.

The computer—an essential tool

The large masses of data produced in traffic studies, and the myriad combinations in which they can be analyzed, make the high-speed, high-capacity computer an essential transportation planning tool. Through use of electronic computers, the traffic engineer produces detailed tabulations of traffic statistics, relates travel demands and characteristics to such basic factors as population and land use, projects the demands for the movement of people and goods by separate transportation modes, and helps to "assign" projected travel volumes to existing and planned transportation networks. Analog and digital computers have thus become a common, if not an essential, tool of the traffic engineer.

The large-scale computer, with rapid data access and large core capacity, has helped to extend transportation planning to a broad geographical scale. The computer also assists in the synthetic mathematical model building process necessary for projection of travel demands. The testing of several system alternatives is facilitated and the optimum solution is made more readily obtainable. Smaller, special-purpose computers also play a vital role in transportation—especially traffic control and operations.

Electronic aids

The most obvious application of electronic aids is in the field of traffic operations. The traffic engineer, especially in an urban area, has a legacy of many poorly designed streets, out of which he must obtain the maximum of traffic capacity. It is in this situation that the accuracy and quickness of electronic devices provide maximum benefit.

Remote "eyeballing" of specific problem areas is possible through strategic placement of video cameras. Surveillance of traffic flows by closed-circuit television or other types of detecting devices permits optimum utilization of crowded urban expressways and special facilities, such as tunnels. These devices can even meter the ramps leading onto the expressway when preset programs determine that additional vehicles will create an unwanted slowdown of the through traffic. Experimental applications of these ramp closings have proved very rewarding.

One unique television setup monitors the operation of a gate associated with the reversible lane of a freeway to prevent its closing on a vehicle. No other sensing device was considered sufficiently safe to allow the introduction of a physical barrier in a freeway traffic stream.

Improvement in sequencing and more responsiveness to competing traffic demands have been achieved through computerized operations of traffic signals. Detection devices are directly connected to automatically operated signs. It is possible to open routes for fire or other emergency vehicles through radio control of local intersection signals, or through other electronic means, which replace central fire station switches.

From such developments come even further improvements, including the "automated highway." Proved experimentally, these new facilities can provide electronic control of motor vehicles, both from within the car and by the use of outside devices. In a manner similar to that employed in the automatic "blocking" of trains, the

new roadways provide driver freedom and increased safety. It is still necessary, however, to fully examine the effect of this automation on drivers and to perfect the control techniques for all possible contingencies.

A limited application of an instrumented roadway is presently under study and, in fact, may be partially installed at this time. Its goals are twofold: first, to study the behavioral patterns of drivers and their response to various stimuli and, second, to develop and evaluate new control techniques. The instrumented roadway will involve facilities already in service (primarily freeways) and should therefore provide an excellent laboratory for the engineering disciplines concerned.

To return to the familiar matter of traffic signals, engineering improvements through extensive use of electronic aids are proving to play a significant part in solving New York City's traffic problem. Traffic Commissioner Henry A. Barnes, faced with the onslaught of drivers wishing to travel conveniently in the city, concluded that the solution lies in electronic control.

As a result, detailed specifications have been developed for the necessary computerized control. The unique solution to this problem resulted in a detection–communication–control system providing 19 computerized master controllers which will supervise a specific area of the city. Further system decentralization is provided by 60 submaster controllers, which, in turn, relay information to local intersection controllers within their area of influence.

Electronic connections are provided between the control areas, and a master surveillance point is maintained by the Department of Traffic, thus avoiding considerable field maintenance expenditure. The keystone to the successful operation of this system is a series of interconnected sensing devices that receive information of traffic volumes, speeds, and spacings, and thus provide the proper division of time for red, yellow, and green phases at individual intersections.

At those intersections where heavy volumes of traffic compete for signal time in both directions, "critical intersection" controllers are being developed that will provide an optimum amount of "green time" to the direction of flow having the most demand. These critical intersection controllers will be responsive, through a series of traffic-sensing devices, to the ever-changing fluctuations in traffic at these key crossing points.

The research and development for the special-purpose computers to handle these system requirements is being done by Sperry-Rand Corporation. Wilbur Smith and Associates, the transportation engineering consultants, in helping to meet the demands of the proposed system, have been engaged in modernizing many signalized intersections throughout the city. The end result should be a vastly improved coordinated traffic signal system utilizing the latest equipment.

Related applications

It is a truism in the traffic environment that much of the fault leading to accidents stems from some instantaneous or inherent human failure. Millions of dollars are therefore being spent on researching the effects of impairment to the human personality and the increase of accident potential because of disability. The use of electrical and mechanical devices to assist in controlling both the individual vehicle and streams of vehicles is an essential function of the modern urban area traffic engineer.

In other uses, toll collections can be handled electronically, providing improved security. Highway illumination improves safety on heavily traveled roadways and intersections. Even in parking operations, electronic gear is optimizing the use of garages and the collection of parking fees.

The use of high-speed analog or digital computers makes possible extensive data-retrieval systems. There is almost no end to the transportation-oriented uses of electronic devices. Agencies now utilize these machines for checking stolen vehicles, for speed enforcement, and for even more rapid checking of driver applicants. Other uses include electronic scales for weighing vehicles in motion and electronic "eyes" for detecting oversize vehicles at tunnel entrances.

Future prospects

The immediate future of the transportation field is vastly more promising than that of the past. Many agencies, public and private, are involving themselves in the development of new concepts, new devices, and new applications.

The U.S. Bureau of Public Roads (BPR) has recently commissioned the Philco Corporation to prepare a report on a "language" that could be a coding technique for route recognition and position description. Such a language would be invaluable in the control of a traffic system embracing the East Coast megalopolis, for example.

So great are the demands on communication facilities that BPR has also commissioned a study to determine the *total requirements* for highway communications. This evaluation would begin when the route is being selected for a highway and continue through such facilities as wire lines and microwave systems that might be required in the operation and policing of the highway. Additional studies are being conducted to determine what information a driver needs and what the saturation point is on visual communications.

The technique of driving involves a continuing series of adaptive reactions to changing road conditions, changing vehicle responses, and changing driver attitudes. Most of the adaptations are simple decisions. However, even simple decisions—easily made—become complex when they are coupled with five or six other such decisions, all of which must be made within seconds.

It is to eliminate the necessity for making these decisions, or to space them over a more reasonable time period, that much of the efforts of engineers concerned with traffic have been directed. Larger or better-located signs, limited-access roadways, and more efficient signal systems are the older tools employed. Variable-message signs, automatic vehicle control, automated highways, and computer-controlled signal systems are a few of the exciting new tools.

Direct driver advisories, utilizing radio and prerecorded messages keyed by roadside transmitters, are perhaps needed. Providing real-time response, computers would be capable of assimilating vast amounts of data and quickly providing the information necessary to the driver.

Automatic equalization using transversal filters

Dispersion, or linear distortion, limits the usefulness of many communication channels. Techniques have been developed that automatically reduce this distortion and thereby permit more efficient communication through the channel

Harry Rudin, Jr. *Bell Telephone Laboratories, Inc.*

The restriction that linear distortion imposes on the flow of information has been well known for some time. Consequently, means have been developed for reducing this distortion through the insertion of compensating linear devices. In the past this reduction has been achieved through the use of conventional lumped-parameter networks, quite static by nature. The use of the transversal filter, however, provides a more dynamic and flexible approach. This device is well suited to automatic and even adaptive operation, and thus it can meet the demands of a channel with time-varying distortion.

A tremendous effort is being made to obtain greater and greater transmission efficiencies from already crowded communications channels. Recent advances in the area of signal design have permitted marked increases in the speed of information transmission. In addition, developments in error control systems permit encoded information to be sent much more reliably. A remaining obstacle to efficient information transmission has been the residual dispersion, or linear distortion, in communication channels, and intensive efforts to solve this problem have been made in the field of automatic equalization.

The general scheme used to achieve automatic equalization is represented in block diagram form in Fig. 1. The nature of the linear distortion introduced by the channel is such that it can be removed by a linear device of sufficient complexity. The linear device used is the automatic equalizer, a control mechanism that adjusts the equalized channel response so that it approximates a desired or reference response.

The problem of signal design deals with the gross or average characteristics of an ensemble of communication channels. In many cases a transmission system designed for the average channel characteristic of such an ensemble is simply not sufficient. This is true when individual channel characteristics deviate significantly from their average characteristic. The problem of equalization deals with the deviations of particular communication channels from the average of the ensemble to which they belong.

Linear distortion

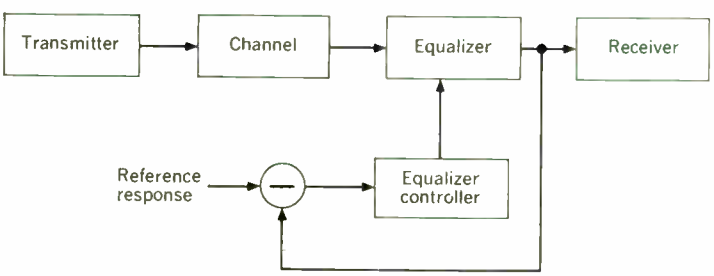
It seems inherent in nature that signals are dispersed—that is, spread out and changed—in transmission. This dispersion manifests itself in two ways: in radio communication it is called multipath and in wire communication it is called linear distortion.*

Figure 2 depicts a multipath situation. The transit time from the transmitter to the receiver varies with the path length so that two signals displaced from each other in time are received although only one was transmitted.[1] This situation is certain to cause difficulties when an effort is made to maximize the information transmission rate in the channel. Multipath transmission can be viewed as communication through a group of parallel, distortion-free channels of different lengths.

Another type of communication, typified by the wire communication channel, may have a single obvious trans-

*It is quite possible that each of the several paths comprising multipath transmission itself contains dispersive distortion.

Fig. 1. Communication system with automatic equalization.



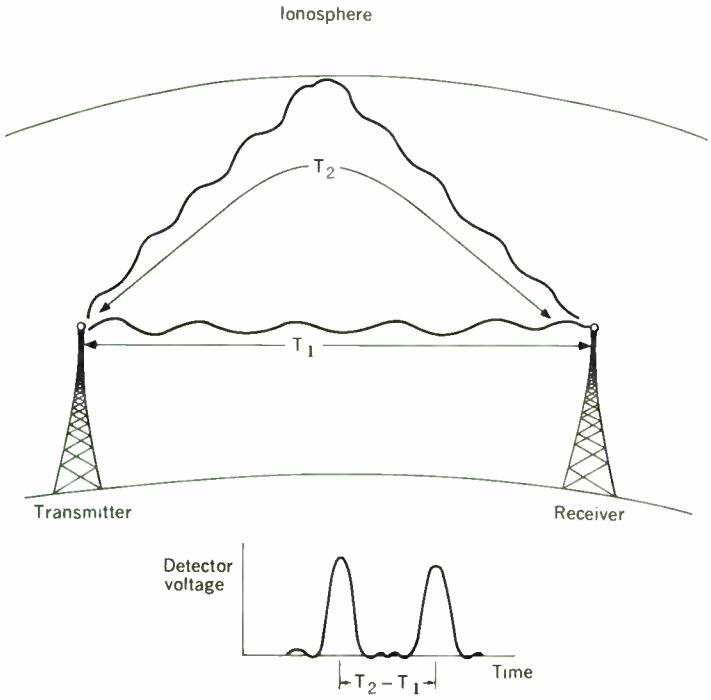
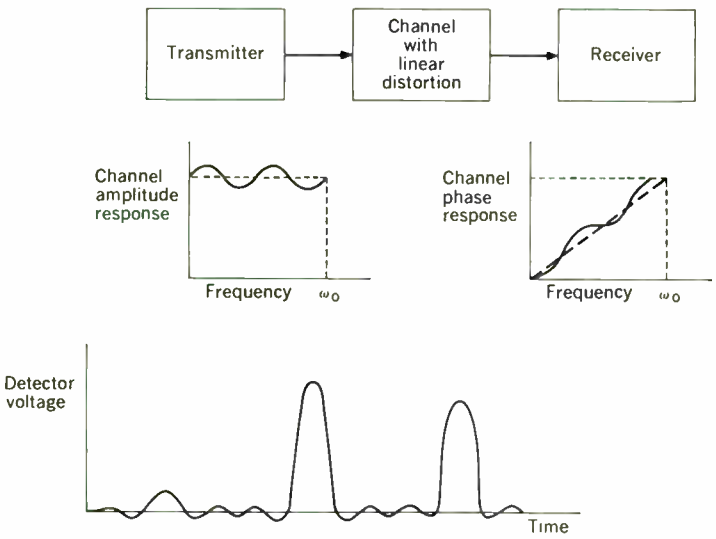


Fig. 2. Typical multipath situation.

Fig. 3. Linear distortion.



mission path but a nonideal transmission frequency characteristic (an ideal transmission frequency characteristic is here defined as one having a flat amplitude frequency response and a linear phase frequency response). Figure 3 displays frequency response curves for such a nonideal channel. When a channel is band-limited (practically speaking, real channels always are), its amplitude and phase frequency response functions can be expressed in terms of Fourier series over the band of interest. A convenient relation between the time and frequency domains can then be given by paired-echo theory.[2]

If, for example, the deviation in the amplitude frequency response can be expressed in terms of a cosinusoid of amplitude $2a$ and period $1/\tau$ (that is, the complete amplitude frequency response is given by

$$A(\omega) = A_0[1 + 2a \cos \omega\tau] \quad (1)$$

where ω is the angular frequency), then it is easy to show that the received response is

$$P(t) = P_0(t) + aP_0(t + \tau) + aP_0(t - \tau) \quad (2)$$

where t is time and $P_0(t)$ is the response that would have been received with no distortion present.*

Similarly, if the deviation in the phase frequency response can be expressed in terms of a sinusoid of amplitude $2k$ and period $1/\tau$ (that is, the phase frequency response

$$\varphi(\omega) = \varphi_0(\omega) - 2k \sin \omega\tau \quad (3)$$

where $\varphi_0(\omega)$ is the original, undistorted phase response), then the received response is

$$P(t) \cong P_0(t) + kP_0(t + \tau) - kP_0(t - \tau) \quad (4)$$

where again $P_0(t)$ is the response that would have been received had there been no distortion. Equation (4) is a good approximation only when k is much smaller than unity. If this restriction does not hold, $P(t)$ becomes more dispersed in time. (A detailed discussion of this may be found in Wheeler[2] or Sunde.[3]) The components of the received signal (those resulting from the several terms required to describe both amplitude and phase) can be added together to produce the total impulse response. Such a superposition of pulses is shown at the bottom of Fig. 3. In general, many terms are needed to describe the channel distortion and so the impulse response of the channel would consist of many, very likely overlapping, pulse waveshapes.

The point here is that the problems resulting from multipath distortion and from linear distortion are equivalent. A device capable of correcting the effects of one is capable of correcting the effects of the other. It is true, however, that multipath is often a time-varying phenomenon. The literature is rather sharply divided by this distinction.

The transversal equalizer

If a portion of the linear distortion is systematic (that is, if it occurs in the entire ensemble of communication channels being considered), this distortion can in principle be compensated for by skillful signal design or fixed network equalization. The remaining portion of the linear distortion is random; it may be the result of natural forces, as in the case of multipath, or it may be manmade, resulting, for example, from manufacturing tolerances in signal-shaping filters. An extremely flexible, easily adjustable network is needed to compensate for the random linear distortion. A device that meets these requirements is the transversal equalizer.[4]

The device is simple in principle, as can be seen from Fig. 4. The distorted signal enters a tapped delay line, and is picked off at various taps (usually equally spaced) on the delay line, delayed in time but unchanged in waveshape. The signal from each tap is passed through an associated variable attenuator; all the attenuator output signals are then summed. The operation of the transversal equalizer can be viewed most easily in the time domain.

The great versatility of the transversal equalizer is apparent if the impulse response of a delay line with infinite

*Note that the cyclical variation is in frequency. Thus one cycle might have a period of, say, 100 Hz.

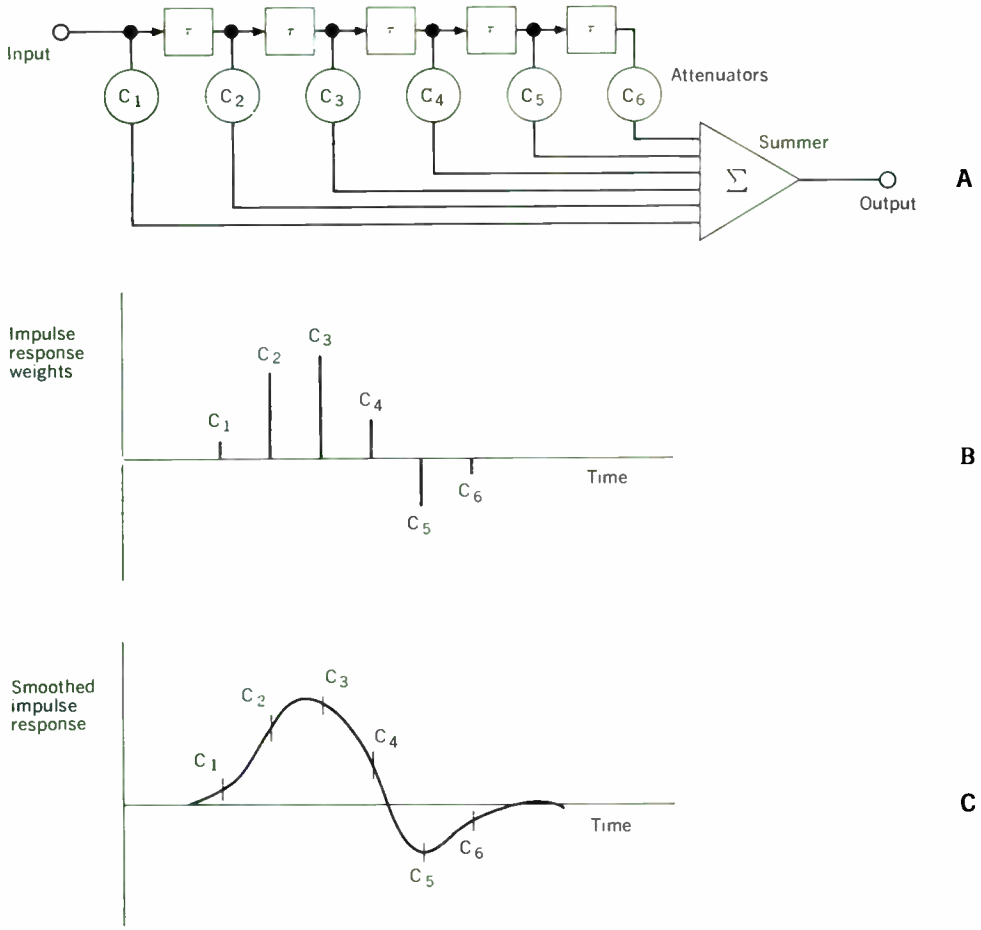


Fig. 4. Ideal transversal equalizer.

bandwidth is examined. Such a case is illustrated in Fig. 4(B), in which the attenuator settings have been picked to generate a specific time function. The attenuators can assume both positive and negative values. As can be seen, the sample values of an arbitrary time function can be specified at the tap-spacing intervals over a time period equal to the length of the delay line. A well-known principle of communication theory is that a waveform band-limited to a frequency of f_0 hertz can be completely specified by samples in the time domain at intervals of $\tau \leq 1/2f_0$ seconds. For a band-limited signal and with the preceding relation satisfied, the smoothed waveform shown in Fig. 4(C) might be obtained.

The time-domain impulse response of the equalizer is related through the Fourier transform to the frequency response. The impulse response may be written

$$c(t) = \sum_{n=-N}^{N} c_n \delta(n\tau) \tag{5}$$

where the number of taps is $2N + 1$, c_n is the attenuator setting of the nth tap, and $\delta(n\tau)$ is the Dirac delta function. The Fourier transform can be written

$$C(\omega) = \sum_{n=-N}^{N} c_n e^{jn\omega\tau} \tag{6}$$

It can be seen from (6) that the spectral response is periodic in frequency. Specifically, the amplitude response is even about frequencies $2n\pi/\tau$ and the phase response is odd about the same frequencies.

The foregoing statements are mathematically true, but they must be modified for physical delay lines, which themselves have a band-limited response. If the input signal is properly band-limited, the bandwidth of the delay line need not be a design consideration. A good rule of thumb for lumped-parameter delay lines[3] is that the total delay–bandwidth product is limited to approximately 40. Cost rises very rapidly if larger delay–bandwidth products are desired.

Some insight into the flexibility of the transversal equalizer can be had from the following discussion. The attenuators associated with the tapped delay line can be manipulated in pairs to achieve useful effects (as viewed) in the frequency domain. If the center-tap attenuator on the delay line is set at unity and symmetrically located pairs of attenuators are given equal values (but magnitudes much smaller than unity), it is possible to adjust only the amplitude response, leaving the phase unchanged. For this case, $c_0 = 1$ and

$$c_{-n} = c_n \tag{7}$$

Conversely, if symmetrically located pairs of attenuators are given equal absolute values (again much smaller than unity) but opposite signs, it is possible to adjust only the phase response, leaving the amplitude response flat. For this case, $c_0 = 1$ and

$$-c_{-n} = c_n \tag{8}$$

This property has been recognized by several investiga-

tors, including Wiener and Lee in an early patent,[6] and Sperry and Surenian,[7] who utilized the property in a television channel equalizer.

Another indication of the flexibility of the transversal filter is its use as a mathematical model for time-varying multipath transmission channels by Kailath.[8,9] This same versatility and flexibility was recognized in the first patents of the transversal filter,[6,10-12] the earliest of which was issued in 1935 to Wiener and Lee.[12]

In view of the facility with which a transversal filter can be adjusted, it is not surprising that it should be the center of the search for automatic equalization techniques. One equalizer using digitally controlled attenuators is capable of 2^{108} different settings.[13,14]

Varieties of automatic equalizers

In discussing the various existing automatic equalizers there is a great temptation to categorize them; for example, there are equalizers for synchronous systems and for asynchronous systems and of the preset and adaptive varieties. Perhaps it is best merely to stress the point that all automatic equalizers function in an essentially common manner. Thus, the block diagram of Fig. 1 can represent the operation of any automatic equalizer.

The portion of Fig. 1 with which we are mainly concerned here is shown in greater detail in Fig. 5, which shows the transversal equalizer, consisting of the tapped delay line, its associated attenuators, and the summing amplifier, plus additional control circuitry so that the operation of the transversal equalizer can be made automatic. The input to the transversal equalizer, or equivalently the input to the tapped delay line, is the distorted or received signal. A difference amplifier is used to compare the output of the transversal equalizer—that is, the equalized waveform—with the desired waveform. The result of this comparison is the error signal. This signal is usually given a weighting before it is applied to the control circuitry, which in turn functions in such a way as to minimize the weighted error. The block diagram of the generalized automatic equalizer shown in Fig. 5 will be used to discuss several representative equalizers that have appeared in the recent literature.

A very general kind of equalizer (in the sense that the equalizer is in no way dependent on the modulation scheme used in the channel) has been developed by R. W. Lucky and the writer.[13,14] In this scheme the entire bandwidth of the channel is to be equalized; hence, the spacing between taps must be no greater than the Nyquist interval—that is, the reciprocal of twice the bandwidth. The equalization is established during a training mode during which a properly synchronized desired waveform is compared with the equalized waveform. The difference is the error signal. It is, of course, important that the waveforms used in the training mode have spectral components covering the entire bandwidth, since it is the entire bandwidth that is to be equalized. The attenuators are adjusted to minimize the mean-square value of the error signal.

The technique by which the attenuators are adjusted is an interesting one. The error signal is viewed as including a linear sum of the signals appearing at the various taps. The equalizer strives to minimize the systematic contribution of these signals to the error signal. Under a mean-square error criterion, the measure of such a systematic contribution is cross-correlation. The control circuitry determines the polarity of the cross-correlation between the various tap signals and the error signal and uses this information to change the values of the tap attenuators by a constant increment in the direction that makes the various cross-correlation values more nearly equal to zero. At the conclusion of the training period the control circuitry is disengaged, and the equalized channel is then used for communication.

When many interrelated adjustments are made in a system, the question of stability and convergence to the optimum adjustments naturally arises. This question is a significant one in the field of automatic equalization. In the case of the equalizer just described, stability and convergence can be guaranteed, regardless of the linear distortion in the channel. In the case of other equalizers, the initial distortion in a channel may play an important role in the problem of convergence. The form and degree of dispersion in a channel also affect the number of delay-line taps needed to reduce dispersion to a tolerable level.

The equalizer just discussed is designed in such a manner that the entire channel bandwidth is equalized. Consequently, after the channel has been equalized, it can be used by an arbitrary communication system. If it is known that only certain kinds of signals will be transmitted over a channel, it is often possible to utilize an equalizer of more economical design. This is particularly true in the case of synchronous data transmission, where the receiver is given a clock signal that specifies the regularly occurring instants at which the receiver should sample the output signal of the channel. The information transmitted through the channel is recovered only from the resulting samples. Thus, in this case, it is sufficient—in fact, preferable—to minimize intersymbol interference, the overlap of the information at one sampling instant into that of adjacent sampling instants.

In general, intersymbol interference in a synchronous data system can be controlled by placing taps on the delay line at intervals equal to the reciprocal of the baud or symbol rate. This usually means that a smaller number of taps is needed in the synchronous data case to correct for a specified amount of dispersion. Figure 6 shows waveforms for a binary synchronous data communication system, for the case where a number of "zeros," then a "one" digit, and then more "zero" digits are transmitted. Note that the ideal waveform passes through zero at all but one of the sampling instants. The same statement cannot be made for the received waveform because of the effect of dispersion. Since the detector in the receiver makes decisions at known, regularly occurring instants in time, the signals compared in the difference amplifier of Fig. 5 need be specified only at the sampling instants. Thus, the useful information in the desired waveform can be expressed as a sequence consisting of the desired values of the system's pulse response measured at the sampling instants. For conventional binary communication this sequence would be

$$0, 0, 0, \ldots 0, 1, 0, \ldots 0, 0$$

After the received or distorted waveform has been similarly sampled, the information in the error signal can be expressed as a sequence. It is the components of this sequence that the equalizer strives to minimize in a synchronous data transmission system.

Note that this notion can be extended to equalizers

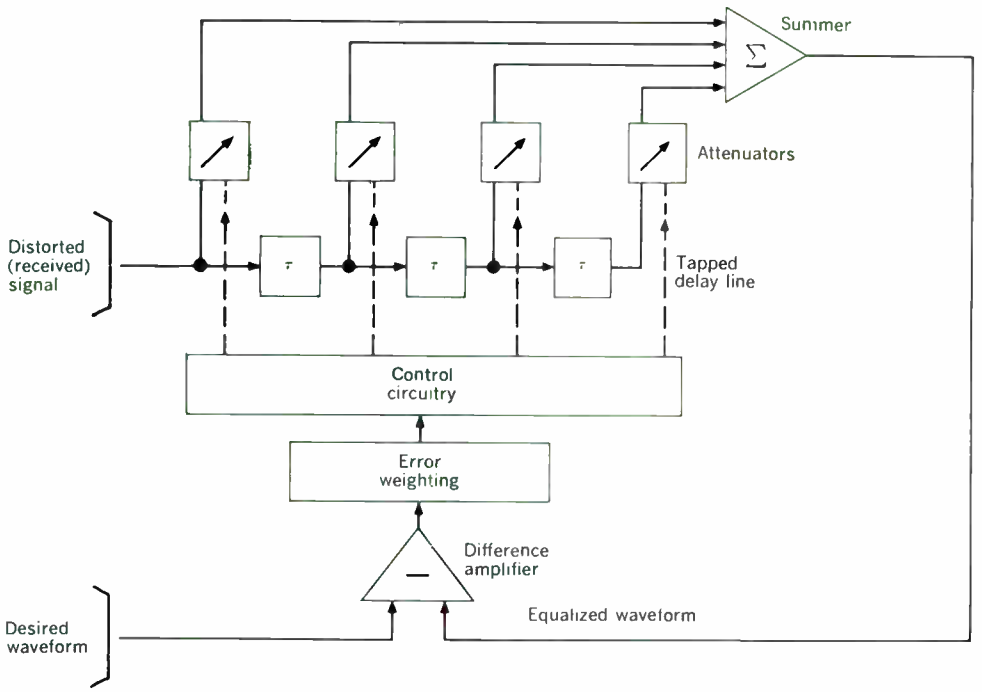
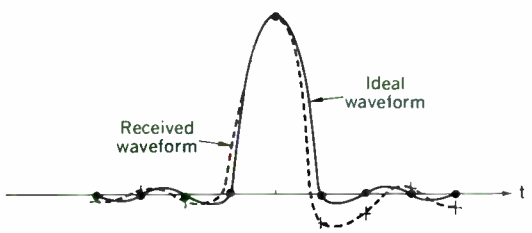


Fig. 5. Generalized automatic equalizer.

Fig. 6. Signals in an equalizer for binary synchronous data communication system.



0, 0, 0, 0, 1, 0, 0, 0, 0	Ideal sequence
e_{-4}, e_{-3}, e_{-2}, e_{-1}, 1, e_1, e_2, e_3, e_4	Received sequence
e_{-4}, e_{-3}, e_{-2}, e_{-1}, 0, e_1, e_2, e_3, e_4	Error sequence

for more general synchronous data communication systems. For example, an equalizer has been successfully constructed for a partial-response data communication system wherein a carefully controlled amount of intersymbol interference is intentionally introduced in the design.[15] For such an equalizer the desired sample sequence might be

$$0, 0, \ldots 0, +1, 0, -1, 0, \ldots 0, 0$$

as in the case of reference 15.

Several representative waveforms from a hypothetical equalizer intended for synchronous binary transmission are shown in Fig. 7. The received signal is shown at the top of the figure; the transmitted sequence is that of Fig. 6: a single binary "one" preceded and followed by binary "zeros." Dots on the waveforms represent the sampling instants and make the intersymbol interference apparent. The equalizer has only two taps, other than the main tap, and these are to be used to reduce the intersymbol interference. The waveform from the first tap is identical to that at the main tap, but delayed in time by a sampling period and halved in amplitude. The waveform from the second tap is also identical to that at the main tap but is delayed by two sampling periods, quartered in amplitude, and inverted in phase. When the three waveforms are added together, the signal at the bottom of the figure results. The intersymbol interference at the two sampling instants immediately following the "one" (corresponding to the two equalizer taps) has been reduced to zero. Note, however, that the equalizer has introduced interference where there originally was none. To find out if a net improvement has been made, some measure of total distortion must be used. A convenient measure is

$$D = \sum_{n=-\infty}^{\infty} |e_n| \tag{9}$$

The values e_n are the samples of the error signal, and it is assumed that both desired and equalized signals are normalized so that both attain unity value in response to a single transmitted "one" digit. Using this measure of distortion, the intersymbol interference has been reduced from 1.0 to 0.5.

An equalizer that functions in this fashion is one described by Schreiner, Funk, and Hopner.[16] Here, as in the example, the equalizer attenuators are to be adjusted to reduce the intersymbol interference to zero within the length of the delay line. This equalizer is of the preset variety; in other words, the channel is equalized prior to the transmission of information.

Another equalizer, also intended for synchronous data transmission, has been described by Lucky[17,18] and Becker *et al.*[19] This equalizer functions to minimize the distortion defined in (9). A staircase approximation to the steepest descent technique is used to force the sam-

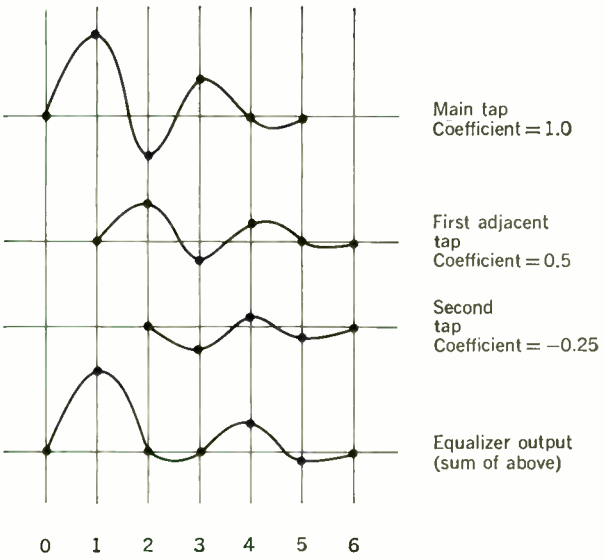


Fig. 7. Operation of an equalizer for synchronous data.

Fig. 8. Unequalized and equalized facsimile transmission.

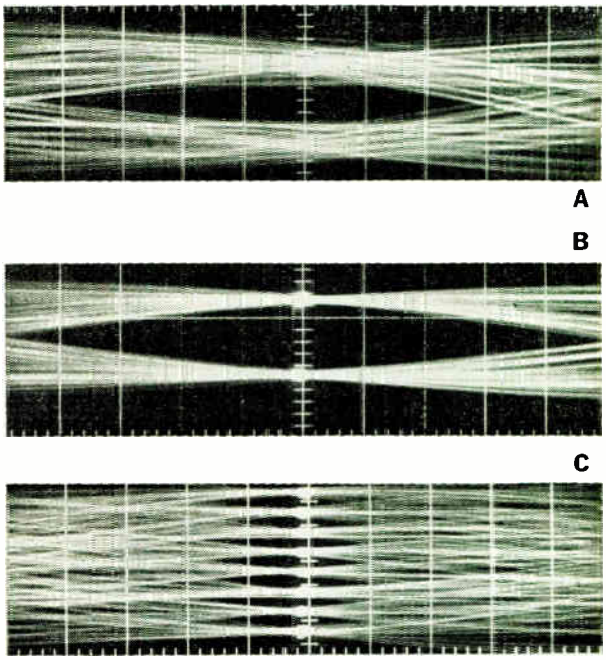


Fig. 9. Eye patterns. A—Unequalized binary eye. B—Equalized binary eye. C—Equalized eight-level eye.

ples of the error sequence to zero within the length of the delay line. A very significant contribution is the proof[17] that this easily implementable procedure is convergent and does the best possible job in minimizing distortion, provided that the initial distortion is less than unity. This equalizer functions either in the preset mode[17] or in the adaptive mode.[18] In the adaptive mode, the channel is equalized while useful information is being transmitted. To accomplish this the equalizer in effect makes an estimate of the channel's impulse response from the data actually received. As the equalization proceeds and the intersymbol interference is reduced, the estimate becomes better and better. Operation in the adaptive mode has the great virtue that should the characteristics of the channel change as a function of time, the equalizer would be able to compensate for this change. The preset equalizer is blind to such a change.

Because communication channels are, in general, linear, the equalizer can be placed either before or after the channel. Normally the equalizer is placed at the receiver because the information needed for equalization is obtained at this end. Equalizers placed at the transmitter, often called predistortion equalizers, have been suggested several times in the literature.[16,20,21] These equalizers are intended to be used in synchronous data transmission.

An interesting hybrid system has been suggested by Gorog,[22] who discusses the design of an equalizer for synchronous data transmission where half of the delay line is of the lumped parameter variety and the other half is digital, the delay being obtained through the use of digital shift registers. A system where the entire delay line is made up of digital shift registers is described by McAuliffe.[23]

Conclusion

Many more equalization strategies exist, but those mentioned here should be sufficient to give the reader a representative view of the field. References to papers discussing other strategies can be found in the bibliography.

A few words are in order about the performance of automatic equalizers. The general-purpose equalizer described earlier[13,14] was tested in conjunction with a facsimile transmission system, with an improvement in performance as shown in Fig. 8. This is an example of equalization used in an asynchronous transmission scheme. Figure 9 shows the improvement in performance for a synchronous data transmission system using the equalizer described in reference 19. The pictures shown are "eye

patterns," obtained by superimposing the channel's response to many different sequences of binary "ones" and "zeros." The width of the eye opening provides a measure of the necessary tolerance with respect to sampling time, and the height of the eye opening indicates tolerance with respect to sampling level or margin against noise. Three "eyes" are shown: the unequalized binary eye, the equalized binary eye, and an equalized eight-level eye. The equalizer makes it possible for the receiver to distinguish eight separate levels with a corresponding increase in the rate of reliable data transmission.

Although much of the groundwork has been laid in the field of automatic equalization, many exciting developments lie ahead as the basic principles are applied to specific communication systems.

The author is grateful to his colleagues at Bell Telephone Laboratories for their encouragement and constructive criticism in the preparation of this article and acknowledges their key roles in making automatic equalizers a practical reality.

REFERENCES

1. Hulst, G. D., "Inverse ionosphere," *IRE Trans. on Communications Systems*, vol. CS-8, pp. 3–9, Mar. 1960.

2. Wheeler, H. A., "The interpretation of amplitude and phase distortion in terms of paired echoes," *Proc. IRE*, vol. 27, pp. 359–385, June 1939.

3. Sunde, E. D., "Theoretical fundamentals of pulse transmission," *Bell System Tech. J.*, vol. 33, pp. 721–788, May 1954; pp. 987–1010, July 1954.

4. Kallman, H. E., "Transversal filters," *Proc. IRE*, vol. 28, pp. 302–310, July 1940.

5. Millman, J., and Taub, H., *Pulse and Digital Circuits*. New York: McGraw-Hill, 1956.

6. Wiener, N., and Lee, Y. W., U.S. Patent 2 124 599, July 1938.

7. Sperry, R. V., and Surenian, D., "A transversal equalizer for television circuits," *Bell System Tech. J.*, vol. 39, pp. 405–422, Mar. 1960.

8. Kailath, T., "Adaptive matched filters," *Symp. Math. Optimization Techniques*, chap. 6, pp. 109–140, 1960.

9. Kailath, T., "Optimum receivers for randomly varying channels," *Symp. Inform. Theory (London)*, vol. 4, pp. 109–121, 1960.

10. Blumlein, A. D., and Kallman, H. E., British Patent 517 516, Feb. 1940.

11. Lee, Y. W., and Wiener, N., U.S. Patent 2 128 257, Aug. 1938.

12. Wiener, N., and Lee, Y. W., U.S. Patent 2 024 900, Dec. 1935.

13. Lucky, R. W., and Rudin, H. R., "Generalized automatic equalization for communication channels," *Proc. IEEE (Letters)*, vol. 54, pp. 439–440, Mar. 1966.

14. Lucky, R. W., and Rudin, H. R., "Generalized automatic equalization for communication channels," *Digest Tech. Papers 1966 IEEE Internat'l Commun. Conf.*, pp. 22–23.

15. Becker, F. K., Kretzmer, E. R., and Sheehan, J. R., "A new signal format for efficient data transmission," *Bell System Tech. J.*, vol. 45, May–June 1966.

16. Schreiner, K. E., Funk, H. L., and Hopner, E., "Automatic distortion correction for efficient pulse transmission," *IBM J. Res. Develop.*, pp. 20–30, Jan. 1965.

17. Lucky, R. W., "Automatic equalization for digital communication," *Bell System Tech. J.*, vol. 44, pp. 547–588, Apr. 1965.

18. Lucky, R. W., "Techniques for adaptive equalization of digital communication," *Bell System Tech. J.*, vol. 45, pp. 255–286, Feb. 1966.

19. Becker, F. K., Holzman, L. N., Lucky, R. W., and Port, E., "Automatic equalization for digital communication," *Proc. IEEE (Correspondence)*, vol. 53, pp. 96–97, Jan. 1965.

20. Lebow, I. L., McHugh, P. G., Parker, A. C., and Rosen, P., "Application of sequential decoding to high-rate data communication on a telephone line," *IEEE Trans. on Communication Theory*, vol. IT-9, pp. 124–126, Apr. 1963.

21. Nyquist, H., "Certain topics in telegraph transmission theory," *AIEE Trans.*, vol. 47, pp. 617–644, Apr. 1928.

22. Gorog, E., "A new approach to time-domain equalization," *IBM J. Res. Develop.*, pp. 228–232, July 1965.

23. McAuliffe, G. K., "ADEM—an adaptively data-equalized high-speed modem," *Proc. 1964 MIL-E-CON*, pp. 332–337.

BIBLIOGRAPHY

Amoroso, F., "Further history of tapped delay-line pulse shaper," *IEEE Trans. on Information Theory (Correspondence)*, vol. IT-11, p. 450, July 1965.

Bellows, B. C., and Graham, R. S., "Experimental transversal equalizer for TD-2 radio relay system," *Bell System Tech. J.*, vol. 36, pp. 1429–1450, Nov. 1957.

Bennett, W. R., and Davey, J. R., *Data Transmission*. New York: McGraw-Hill, 1965.

Bogert, B. P., "Demonstration of delay distortion correction by time-reversal techniques," *IRE Trans. on Communications Systems*, vol. CS-5, pp. 2–7, Dec. 1957.

Boothroyd, W. P., and Creamer, E. M., Jr., "A time division multiplexing system," *AIEE Trans.*, vol. 68, pp. 92–97, 1949.

Comstock, G., "Low-cost terminals triple phone link data rates," *Electron. Design*, vol. 11, p. 8, Jan. 18, 1963.

Coll, D. C., and George, D. A., "A receiver for time-dispersed pulses," *Conf. Record 1965 IEEE Ann. Commun. Conv.*, pp. 753–757.

Coll, D. C., and George, D. A., "The reception of time-dispersed pulses," *Conf. Record 1965 IEEE Ann. Commun. Conv.*, pp. 749–752.

D'Attilio, R., "Boothroyd and intersymbol interference reduction," *IEEE Trans. on Information Theory (Correspondence)*, vol. IT-10, p. 257, July 1964.

DiToro, M. J., "A new method of high-speed adaptive serial communication through any time-variable and dispersive transmission medium," *Conf. Record 1965 IEEE Ann. Commun. Conv.*, pp. 763–767.

DiToro, M. J., "Phase and amplitude distortion in linear networks," *Proc. IRE*, vol. 36, pp. 24–36, Jan. 1948.

DiToro, M. J., Hanulec, J., and Goldberg, B., "Design and performance of a new adaptive serial data modem on a simulated time-variable multipath HF link," *Conf. Record 1965 IEEE Ann. Commun. Conv.*, pp. 769–773.

Enke, H. G., "A method for compensation of existing echoes for any time function," *NTZ Commun. J.*, pp. 33–37, 1965.

Gibson, E. D., "A highly versatile corrector of distortion and impulse noise," *Proc. 1961 Nat'l Electron. Conf.*, pp. 543–556.

Gibson, E. D., "An intersymbol adjustment method of distortion compensation," *Proc. 1961 MIL-E-CON*, pp. 196–207.

Gibson, E. D., "Automatic equalization using time-domain equalizers," *Proc. IEEE (Correspondence)*, vol. 53, p. 1140, Aug. 1965.

Kettel, E., "Ein automatischer Optimisator fur Abgleich des Impulsentzerrers in einer Datenubertragung," *Arch. Elek. Ubertragung*, vol. 18, pp. 271–278, May 1964.

Kisel, V. A., "Phase correcting circuits using delay lines," *Telecommun. Radio Eng.*, vol. 12, pt. I, pp. 33–40, Dec. 1965.

Levy, M., "The impulse response of electrical networks, with special reference to the use of artificial lines in network design," *J. IEE (London)*, vol. 90, pt. III, pp. 153–164, Dec. 1943.

Linke, J. M., "A variable time-equalizer for video-frequency waveform correction," *Proc. IEE (London)*, vol. 99, pp. 427–435, 1952.

Mohn, W. S., Jr., and Stickler, L. L., "Automatic time-domain equalization," *1963 Nat'l Commun. Symp.*, pp. 1–9.

O'Neill, J. F., and Saltzberg, B. R., "An automatic equalizer for coherent quadrature carrier data transmission systems," *Digest Tech. Papers 1966 IEEE Internat'l Commun. Conf.*, pp. 24–25.

Price, R., "Optimum detection of random signals in noise with application to scatter-multipath communication—I," *IRE Trans. on Information Theory*, vol. IT-2, pp. 125–135, Dec. 1956.

Price, R., "The detection of signals perturbed by scatter and noise," *IRE Trans. on Information Theory*, vol. PGIT-4, pp. 163–170, Sept. 1954.

Price, R., and Green, P. E., Jr., "A communication technique for multipath channels," *Proc. IRE*, vol. 46, pp. 555–570, Mar. 1958.

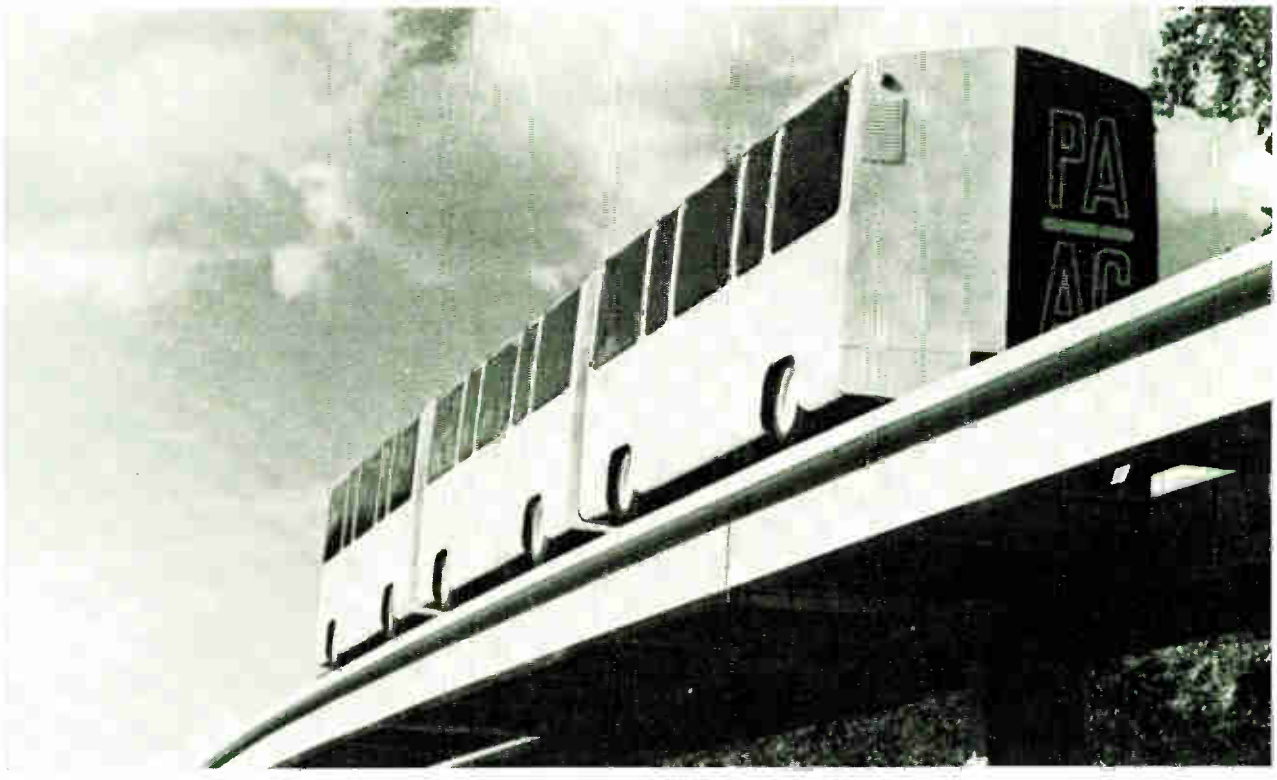
Rappeport, M. A., "Automatic equalization of data transmission facility distortion using transversal equalizers," *IEEE Trans. on Communication Technology*, vol. COM-12, pp. 65–73, Sept. 1964.

Sussman, S. M., "A matched filter communication system for multipath channels," *IRE Trans. on Information Theory*, vol. IT-6, pp. 367–373, June 1960.

Thompson, T. B., and Lyon, J. A., "The design of transversal filters using time-domain methods," *Proc. 1956 Nat'l Electron. Conf.*, pp. 532–539.

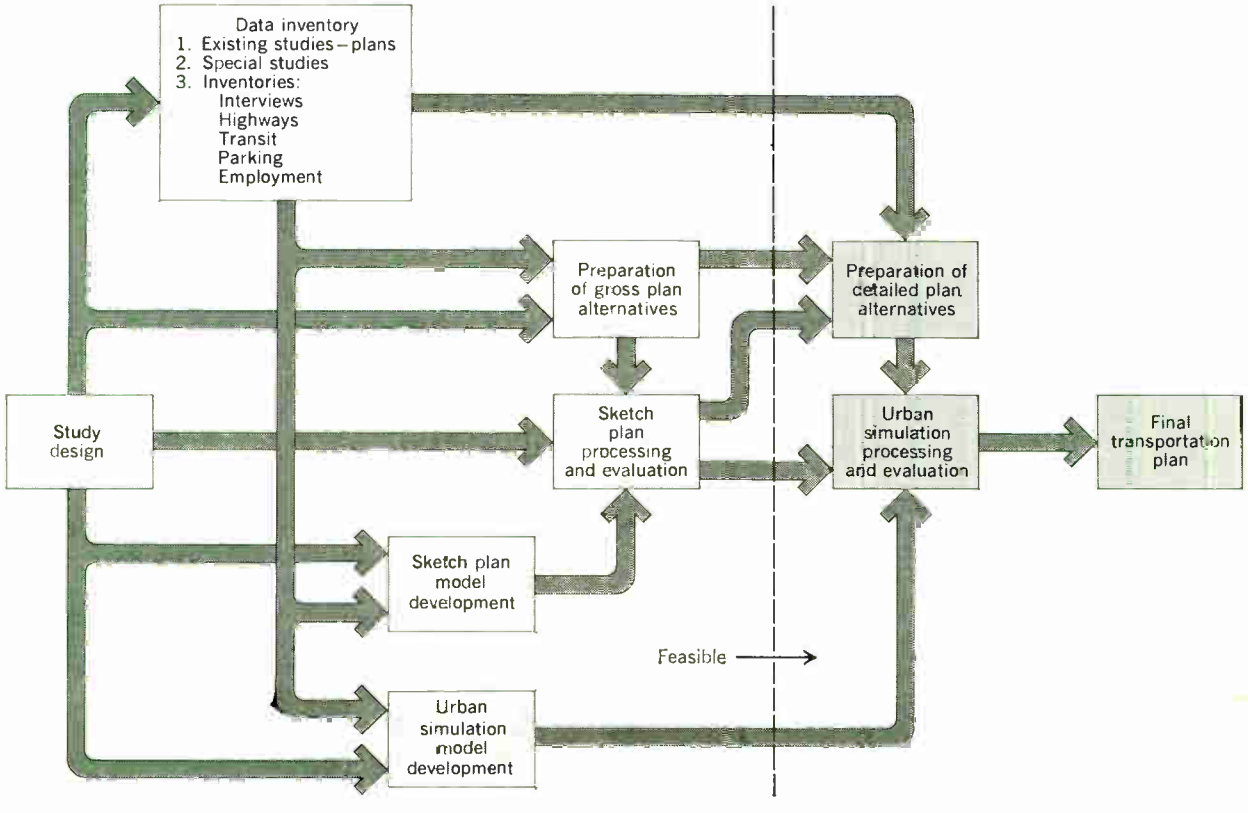
Tuck, W. P., and Wiseman, N. L., U.S. Patent 3 213 196, Oct. 1965.

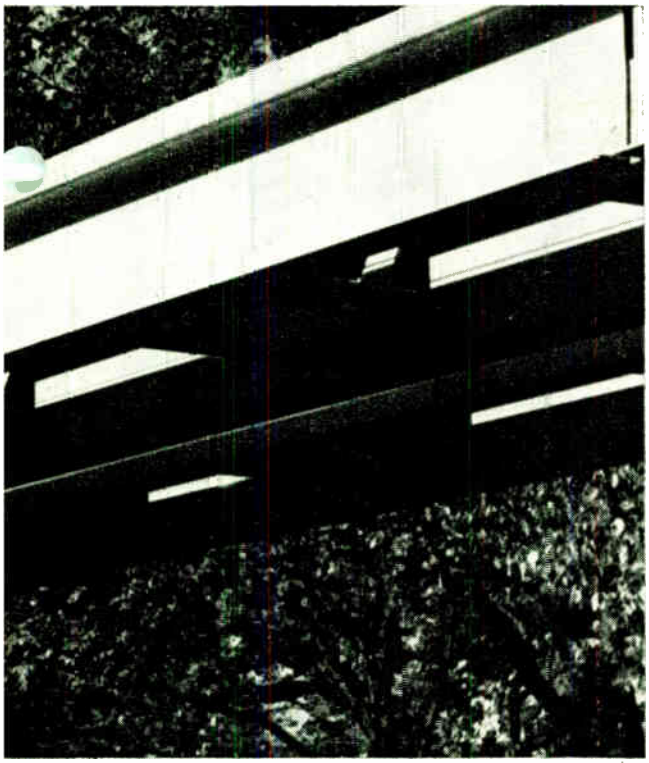
Turin, G. L., "Communication through noisy, random-multipath channels," *1956 IRE Conv. Record*, pt. 4, pp. 154–166.



Transit expressway demonstration project, South Park, Pittsburgh, Pa. This line is specifically designed as a rapid transit system for suburban communities. The system employs lightweight automatic electric vehicles that are operated either as single units or trains on a frequent running schedule.

Fig. 1. Block diagram of the study procedure for the Bay Area (San Francisco) Study Commission. The collected survey data are processed through computers together with suggested transportation models and estimates of future traffic requirements.





Tomorrow's mass rapid transit—available today!

Computer control of high-speed trains and expressway travel, sophisticated multiple-unit propulsion and conductor systems, and new concepts in automatic fare collection are some of the present achievements on metropolitan and interurban lines

Deane N. Aboudara
San Francisco Bay Area Rapid Transit District
C. William Woods
Westinghouse Air Brake Company
Raymond S. Silver *Litton Industries*
John C. Beckett *Hewlett-Packard Company*

Similarity exists between moving large masses of people and the process industry. Both have fixed inputs, fixed outputs, predictable patterns, variables that can be measured, and corrections that can be made. Both entities respond to the principle of closed-loop regulation. From the experience gained in the San Francisco Bay Area Rapid Transit District (BARTD), a planned approach to coordinated and automated mass transportation systems is entirely feasible. Electronic controls are a major factor in providing an economic solution to the problem.

The future of mass transporation, due to changes in demography, looks more favorable today than ever before. Population growth and movement to urban centers have been the key factors in the concerted effort to develop mass transit systems in our major metropolitan regions.

Transportation is one of the services in which modern computer techniques are being applied to acquire a better understanding of the problems and the results to be expected from proposed solutions. Because of the computer, major changes in planning and government organization are taking place, and we can expect some interesting changes in urban transportation.

Systems approach to the problem

A systems approach analysis to achieve a broad balance of transportation modes for people in general is a practical reality today. Mass transportation is concerned with high-capacity service in densely populated areas. The total concept, however, also includes the feeder and distribution services that are essential to a complete system. Thus adequate mass transportation must encompass a coordinated system of rail rapid transit, bus feeder service, privately owned or rented automobiles, specialized water and air transport, and the auxiliary requirements of high-speed expressways, transfer connecting terminals, parking, and the major fixed facilities of bridges and tunnels.

The relative quality of all of these modes of transportation and their components must be maintained in balance to attain a predetermined environmental character. Thanks to the computer, it is possible to project continuously the effects of new industrial and shopping centers, educational institutions—and even new cities—on the transportation patterns of the region. Models can then be developed for testing various proposals for new transportation services, bearing in mind that no single type of service can satisfy all public transportation requirements.

Figure 1 is a flow diagram that outlines the study procedures being used by the San Francisco Bay Area Transportation Study Commission. Data collection is a major part of the study effort, and simulated models reflect the desired objectives. The final plan considers the forces favoring one alternative as opposed to others.

Economic factors are important to the timing of new services and extensions. Certain economic requirements must be satisfied before desired additions can be put into

effect. Again, the computer is the tool that provides predictive data from which a reasonable decision can be made.

Extent of electronic controls

Although the computer is presently the most valuable tool for enhancing the future of mass transportation, other electronic advances are contributing to better standards of coordination of different modes of service.

Electronic controls for the Bay Area Rapid Transit System (BART) will be described in some detail later in this article. Proposals to automate the automobile, however, have been advanced. L. E. Flory of RCA Laboratories[1] urges a planned evolution of the private automobile in urban areas rather than a sudden radical departure from the existing system of transportation. He suggests a system to be developed in three stages:

1. Improved communication with the driver to enable him to make decisions better and faster.
2. An electronic decision-making override and warning to the driver if a dangerous condition occurs that must be handled manually.
3. A completely automated vehicle control either on a continuous basis or as a limited operation.

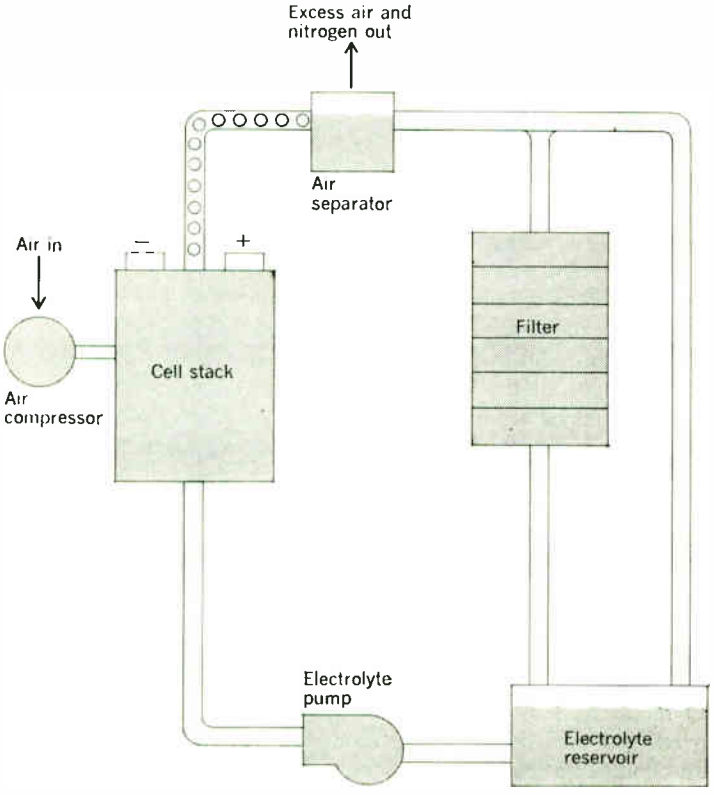
The third stage involves lane control on freeways. It would safely increase the capacity of superhighways by maintaining a controlled space and speed relationship between cars. The driver could relax while the car was under automatic control, but he would be warned in advance of his expected turnoff from the freeway to resume manual control.

Control philosophy for rail transit. BARTD reached an early decision in favor of automatic train control,[2] in which an attendant will ride each train—at least for a while—and attendants will be located at stations to provide assistance and guidance to passengers who are unfamiliar with the system. Also, attendant personnel are necessary to give passengers a feeling of security and safety. Yet, in the long view, operating personnel on vehicles are a major expense factor and a source of inconsistent operations.

Electronic aids such as public-address units and two-way voice communications facilities at stations and on vehicles can provide added convenience to the rider. Television monitoring of station platforms is also a part of the operation of a modern transit system.

Not all automatic systems, however, are designed to have an attendant on the train. For example, the Westinghouse Electric Corporation has developed a lock-on guided train, with pneumatic tires, called "Transit Expressway" (see title illustration). This system is designed primarily for feeder service to connect with other rapid transit facilities and major air terminals. The advantages of this proposed type of rapid transit include minimal elevated support structures, quiet operation, and lower construction costs. This system, particularly attractive for suburban travel, gives assurance that grade-separated transit in less densely populated areas is economically feasible.

Fig. 2. Simplified schematic flow diagram of a zinc-air battery. Energy is derived from the process by converting zinc to zinc oxide. Oxygen can be recovered while the batteries are being recharged at a central location.



For successful urban transit, an overriding need

The greatest single need for the success of future urban transit is a practical electric automobile. Most of the unsolved problems of a mass transportation system can be answered by the development of such a vehicle. The reader should bear in mind that the private automobile has achieved its pre-eminence in urban transportation because of two significant factors: personal privacy and door-to-door convenience. But the gasoline-powered car has two serious disadvantages: exhaust fumes and noise. The electric car would overcome both objections and would be ideally suited for urban transportation.

Electric propulsion, with a self-contained energy source, is not limited to small vehicles. Studies have shown that it is well-suited for buses and delivery services.[3] Application to rapid transit is also possible by use of batteries to maintain service between station stops and the provision of an electrified right of way only for acceleration out of stations.

Enthusiasm for the electric car has been somewhat dampened by adverse reports on power, weight, and cost factors of batteries and fuel cells.[4] Good performance has been demonstrated in the use of silver-zinc batteries, but these units are too costly for general practical use. And fuel cells continue to be prohibitively expensive despite the predictions for continued technical progress.

The most encouraging development in recent months is the zinc-air battery shown in Fig. 2. This prototype has energy densities similar to the silver-zinc battery and its probable cost would be competitive with that of the lead-acid battery.

Thus the future of the electric automobile is predictably good, and the early doubts are beginning to fade as the result of progress in this revolutionary battery.

Electric car rental service. Time lost at the conventional auto parking facility is a major inconvenience, and the high cost of this service is related to the inefficiency of personal automobile retrieval.

The electric car, with its low maintenance requirement, is ideal for an auto rental subscription service. The subscriber establishes his credit with the service agency in the same manner as one applies for telephone service. He is then privileged to rent a car from any of several storage centers, and he pays for the service at a nominal rate according to usage and type of service. He may use the service and a different car two or three times a day.

This plan would greatly simplify the parking problem, particularly at rapid transit railroad stations. Automatic parking would be fast and efficient, since a user who needs a car would take the first one on the conveyer. Automatic test check-out of each car would be accomplished while the car is on the storage conveyer.

The convenience of the individual self-drive automobile for the short haul to the final destination of an urban trip is a key element in a modern transportation system. The electric car for this purpose should not exceed a capital cost of $1400, including batteries, and it would accommodate two passengers over a range of 160 km on a full charge. The life expectancy of such a car should be from eight to ten years.

Coordinated system as a utility service

A metropolitan area must have all the elements of a mass transportation system to serve all the people in the region. Some elements will be more expensive than others, but the user will be able to make a reasonable choice and can expect good service.

Although the daily commuter is generally viewed as the principal user of mass transportation, there is a continuous turnover of regular commuters, shoppers, and occasional users during the day.

A coordinated system means more than physically integrating elements of the system together at stations and terminal points because, for the user, it means one unified service and not a conglomeration of independent services—each with its own regulations and fare schedules.

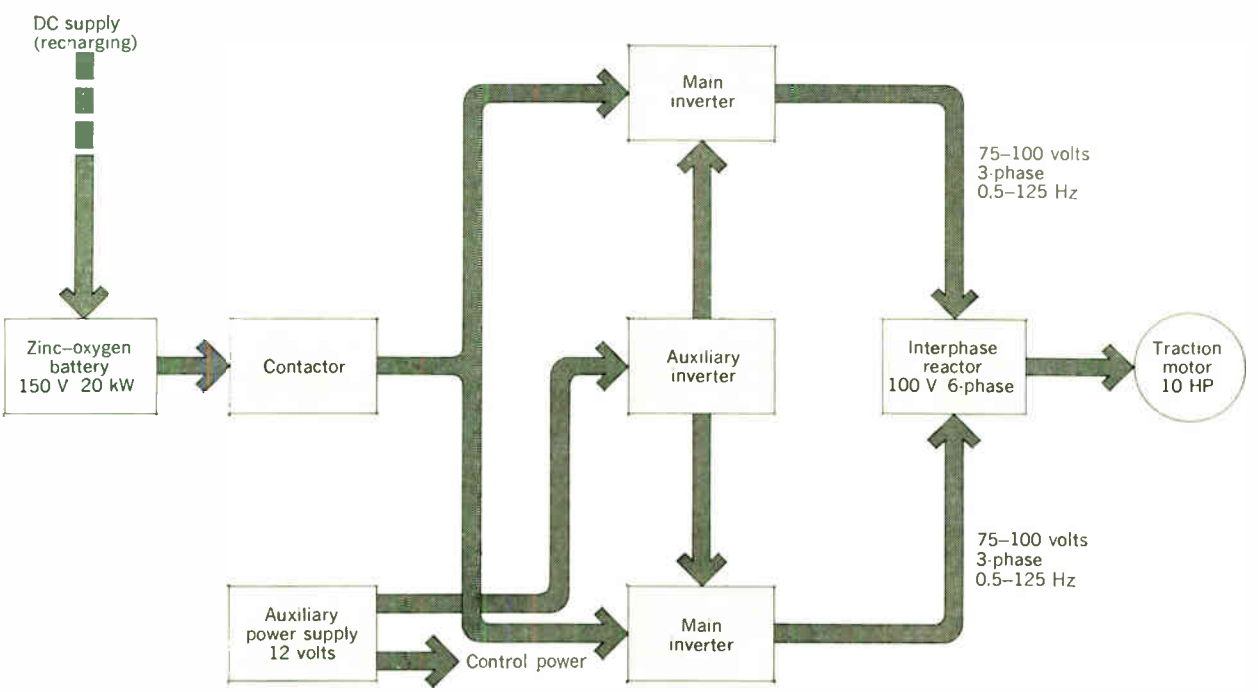
Modern electronic accounting techniques make it possible to have monthly billing for all urban transportation service. Although a subscriber may be required to pay an initial deposit, as is done for many utility services, the policy of paying in advance for transportation would not be necessary. The computer would also separate and credit, in accordance with contract agreement, the amount to be paid to each separately managed component of the overall transportation service. Franchise contracts with bus, car rental companies, and parking facilities can be administered successfully without complicated accounting. In this way, the system becomes an economic unit with technological components.[5]

The variable-frequency motor drive

Progress in the zinc-air battery is only the beginning. Recent results from the BARTD Diablo test track (see "Special Conference Report: National Transportation Symposium, 9th Joint Railroad Conference," IEEE SPECTRUM, pp. 116–121, July 1966) confirm the advantage of variable-frequency motor drives for electric vehicles (Fig. 3) over the conventional dc motor drive.

Variable frequency provides better acceleration characteristics and eliminates jerks during speed changes. Also, regeneration is feasible and this may save about 30 percent of the energy requirement. Energy conservation is particularly attractive in extending the range of a battery-powered vehicle. Whether BARTD will eventually choose

Fig. 3. Block diagram of a possible variable-frequency motor drive appropriate for a battery-operated two-passenger automobile. Power is shown at 20 kW, which compares favorably with the 32 kW that are required for a four-passenger Volkswagen.



the variable-frequency drive for its trains is not known at this time, but the technical progress that has been made on this equipment is directly applicable to the electric automobile or bus.

Electronic controls are closely related to the success of the application of electric drive to automobiles. Variable frequency from 20 to 420 Hz, six phase, as tested on the BARTD track, is highly reliable because of new designs in which solid-state devices are used. Thus the maintenance requirements of the modern electric car, using similar techniques, are expected to be a fraction of those required for the gasoline-powered automobile.

A further introduction to BARTD

BARTD is literally starting from the ground up in engineering the world's most modern rapid transit system. Moving people is a process; it has fixed inputs and outputs, predictable patterns, variables that can be measured, and corrections that can be made, and it can respond to the principle of closed-loop regulation. These are functions that have blood relatives in the steel industry, cement industry, the utilities, and the petroleum industry. In the "process of people" industry, however, absolute fail-safety is mandatory.

In considering the BARTD system, it was evident that the objectives would only be satisfied by extrapolating into the future based upon present-day accomplishments in this fascinating era of rapidly changing technology. To assist in bringing this to fruition, several programs, totaling more than $11 million, were developed and were aided financially by a grant from the U.S. Department of Housing and Urban Development.

Distribution systems

The concept of underground distribution is taking hold in our domestic and industrial areas. Thus we decided to establish the criterion that no overhead distribution or top pantograph pickup would be acceptable.

Heretofore, a standard collector voltage was 600 volts dc. However, by continually asking the question "why?" it became evident that there was no logical reason for blandly accepting 600 volts for BART. In reviewing advanced designs, we found much in the way of accomplishments concerning the application of materials and components, insulation systems, high-speed design, and computer design optimization that gave us justification for *not* settling on the standard value.

It became evident that a 1000-volt dc system was a realistic goal. Technological advancements in solid-state devices for "playing tricks" with alternating current indicated that an investigation should be made to determine the feasibility of an all-ac distribution system.

Our requirement for silent running or noise abatement in the system structure gave us the opportunity to experiment with a three-phase contact system in which the current collection would be accomplished by a three-unit side pantograph. Continuing with this approach, time has allowed us to implement a full-scale shop demonstration and, finally, to conduct actual trial operations at the Diablo track. To date, we have been most encouraged by the performance and by the limited degree of problems.

Propulsion systems

Many techniques used in other disciplines lend themselves nicely to the traction motor concept with steel wheels on a steel rail. For example, chopper circuits will eliminate accelerating resistors and will conserve energy if a dc distribution and propulsion system is chosen. And phase-controlled rectifiers seem quite realistic if the decision is to employ ac distribution and dc propulsion. Also, the adaptability of thyristors is encouraging, and, through our accelerated efforts, we find the use of the ac induction motor a possibility by providing acceptable speed and torque relationships through control by static variable-frequency and variable-voltage means.

The systems mentioned are presently undergoing various phases of testing. But, obviously, there are still some unknowns; for example, the effect of voltage influence from one car with respect to another. There are also certain pulsations that are inherent in the chopper design which can induce specific mechanical phenomena. The phase-controlled rectifier system has demonstrated some intermittent problems in the propulsion and braking modes. These were traced to insufficient voltage creep distance, a predictable situation.

Figure 4 shows a highly simplified schematic diagram of the systems we have discussed, and it also includes systems that were considered in the initial analysis.

Solid-state devices are showing a high degree of reliability in performing the functions of dynamic or mechanical-type electrical components such as relays, regulators, time-delay devices, interlocks, and many other items that are vital to reliability and maintenance.

The essence of automatic train control

BART is pioneering a concept of complete automatic train control. To provide a better understanding of this concept, the following criteria are presented and defined.

Automatic train operation. Operation involves controlling propulsion and braking, which, in conventional systems, is handled by the motorman. In more detail, this function would include

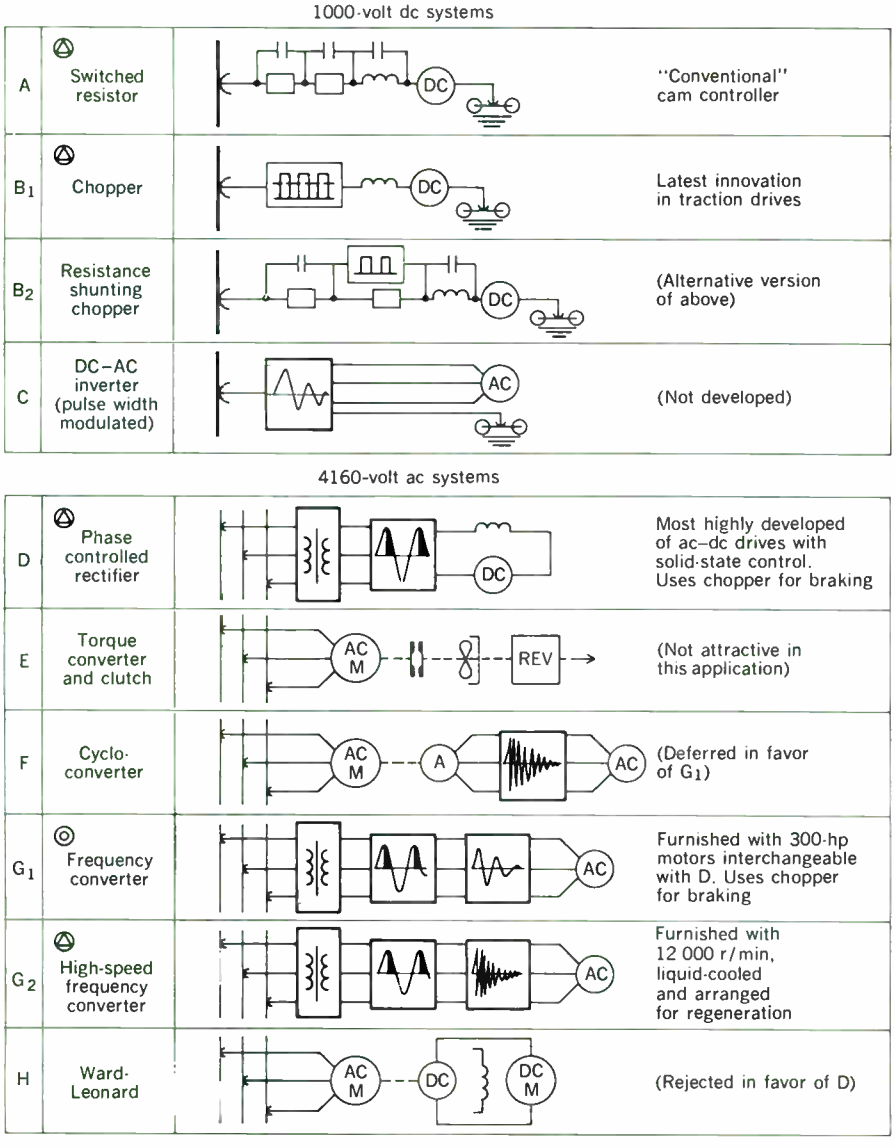
1. Train movement and acceleration from a full stop, speed control while running near a speed limit, and braking to a stop or slower speed.
2. Programmed stops in which the train would be caused to stop at passenger station platforms and other designated positions.
3. Door control and direction reversal.

Train protection. The function of protection has to do with providing fail-safe operation of the train itself with relation to other trains. There are eight subdivisions of this function:

1. Train detection.
2. Train separation (headway).
3. Route interlocking.
4. Operating speed restriction.
5. Speed reference and actual speed detection.
6. Absolute stop assurance.
7. Right-of-way hazard detection.
8. Attendant emergency stop.

Line supervision. The line-supervision function encompasses the regulation of the complete system by individually controlling trains and their interrelation to assure adherence to schedules and routing. For line supervision, the criteria are

1. Dispatch of trains into revenue service.
2. Assignment of train identification numbers.
3. Assignment of train routes.
4. Control of train routes.

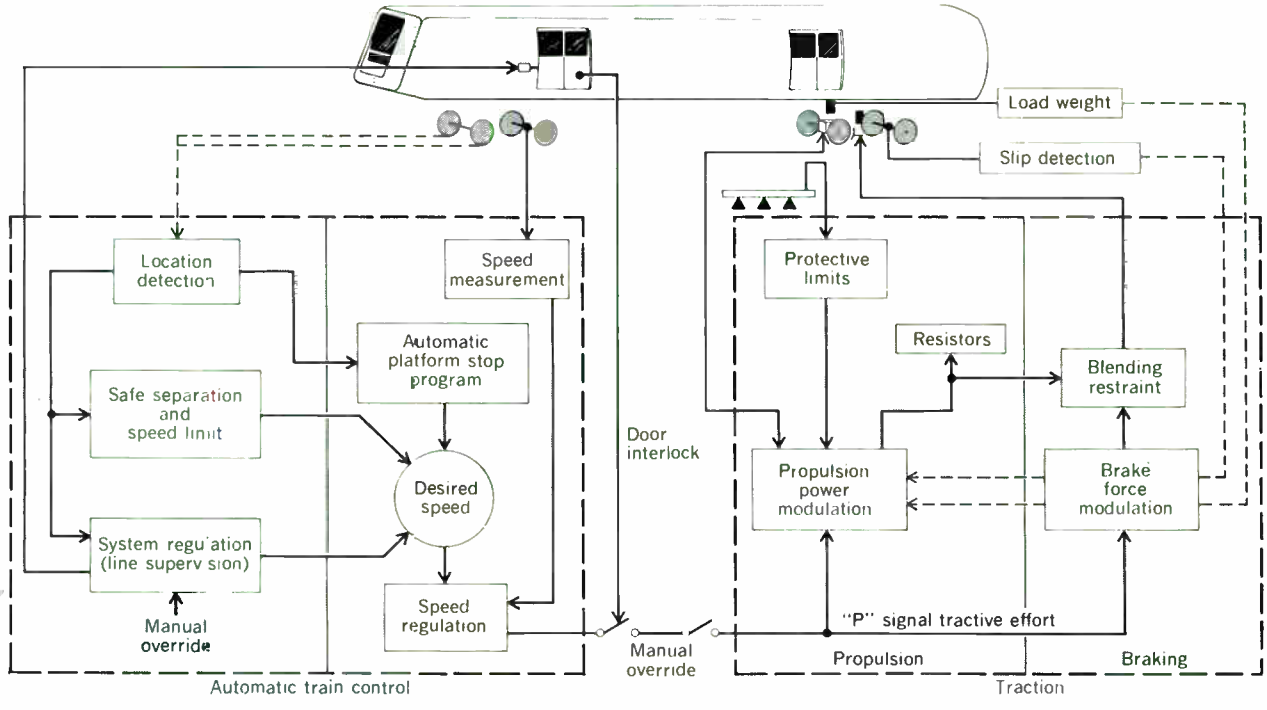


Represented on laboratory cars

Shop-tested for project

Fig. 4. Schematic diagrams of various propulsion concepts suggested for the Bay Area Rapid Transit District (BARTD).

Fig. 5. Diagram showing functional relationship of automatic train control and traction systems.



5. Monitoring and control of train departures from stations and their headways.
6. Removal of trains from revenue service.
7. Alarm operating discrepancies.
8. Initiation of remedial action.
9. Log system operation.
10. Display system operating status.

With this very brief exposure as to what constitutes automatic train control (ATC), Fig. 5 has been prepared to illustrate the functional relationship of ATC and the traction system.

Objectives of a transit system

A transit system must be so designed that, regardless of the means of operation or supervision provided, it must be able to continue to move people under all circumstances—even though some of the equipment may not be in full service. The ability to keep trains running must exist at the local level at all times. This philosophy demands the application of as much equipment as possible for safety on board the trains.

Train safety and train control—from traditional to future

Many existing transit systems throughout the world operate with no more than visual safety protection. It was recognized early, however, that increased efficiency in the utilization of track could be obtained by the installation of wayside block signals for safety and restriction of speed where required. In this case, safety is provided by dividing the system into blocks or zones of sufficient length to provide a safe stopping distance at the authorized speed for each zone. Also, interlockings are provided where more than one route is available. These protect the selected route against the establishment of any other routes through the interlocked area. The traditional supervision of transit systems has been by train order that works against a fixed timetable and utilizes human supervisors at key points. These supervisors are linked together and with other fixed points by telephonic communication. Thus it can be seen that only limited dynamic information is available and optimization of performance of the system is not possible.

As the demands on transit systems increased, the problems with wayside indicators became evident. These problems, which include lack of continuous signal indication to the operator and lack of immediate information when signal conditions change, led the Union Switch & Signal Division of the Westinghouse Air Brake Company to develop a "cab signal"—a coded signal transmitted through the running rails by means of distinct geographical circuits, or track circuits. Information transmitted through the track circuits is picked up inductively through coils that are mounted above the rails ahead of the leading wheels of the first car on the train. A visual indicator in the operator's cab displays either a zero speed permitted (because of lack of signal received), or increasingly permissive speeds as decoded from a variety of transmitted code rates. This system also provides rear-end protection because of the geographical track division into blocks. The transmitted signal in each block is shunted by the wheels of the train, thereby indicating block occupancy. This cuts off the permissive speed signal in the following block to provide zero indication for a safe stopping distance behind a train.

Noise reduction. To reduce noise, shock vibration, and wear, welded ribbon rails have been developed for both railroad and transit systems. Coincident with this development is the logical application of audio frequencies for track circuits. Frequencies in the range of 2 to 6 kHz are used for this purpose. The characteristics of these audio frequencies, the relatively high impedance of steel rails to such frequencies, etc., are such that they present an effective shunt at the block limit points to attenuate the imposed frequency. This permits the use of a few frequencies, repeated at intervals, without the requirement for insulated joints to arrest the signal propagation. Because of crosstalk between parallel tracks, multiple frequency assignments are made so that sufficient separation exists on adjacent sections of track.

Automatic operation—ATO in detail

With cab signal indications on board the vehicle giving rear-end protection and speed commands, the addition

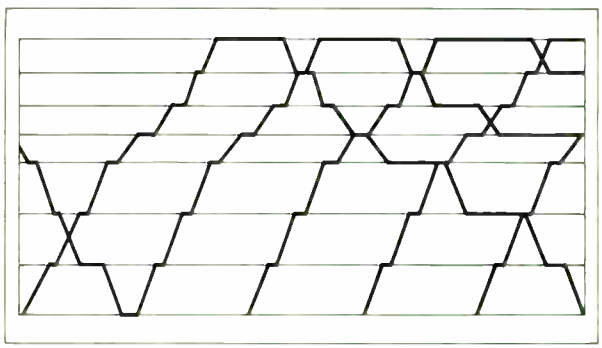
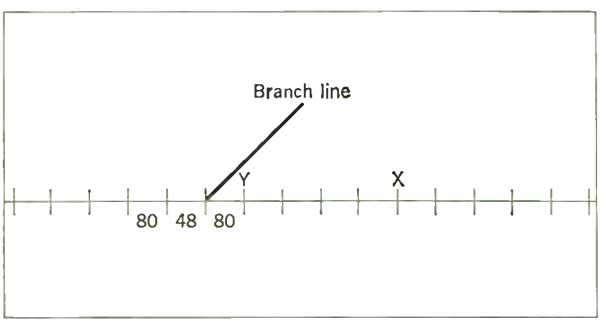


Fig. 6. Portion of a typical train graph in which the horizontal lines represent intermediate stations and the upper- and lower-limit lines are turnaround points.

Fig. 7. Schedule-keeping graph example of a 48-km run with stations 3.2 km apart. All speed limits are 128 km/h, except where noted. Uniform dwell between stations = 20 seconds.



Time, seconds

Percent	At 128 km/h	At 80 km/h	At 48 km/h
100	135	157	245
98	138	160	250
96	140	163	255
94	143	166	260
92	146	170	265

of automatic overspeed protection provides the system with complete ATC. This permits manual operation of the train with an audible warning whenever the speed limit in that zone is exceeded. If this warning is not heeded within a set time, an automatic speed reduction is put into effect.

Systems now being installed by Westinghouse Air Brake for the Delaware River Port Authority's new high-speed rail line from Philadelphia to southern New Jersey, and the "Expo Express" for the Montreal World's Fair in 1967, provide this system as one of the modes of operation. The system, known as automatic train operation (ATO), moves one step further by the elimination of manual operation. In this case, the speed limit signals are decoded in the same manner as before and are used actually to control the speed of the train.

The 'P' wire. Parsons Brinckerhoff-Tudor-Bechtel, the joint engineering firm for BARTD, has pioneered an interface between ATO equipment and the propulsion and braking system called the "P" wire. The "P" wire signal, delivered by the ATO equipment, is in analog form from zero to one ampere, and it indicates the desired positive or negative tractive effort. Coast, or neutral position, is at 0.5 ampere. Increases between 0.5 and 1.0 ampere are proportional to the positive torque required. Decreases between 0.5 and zero are proportional to the negative torque, or braking action. Zero signal, caused either by failure of equipment or a specific request, equals full service brake application. The addition of automatic station-stopping equipment, which will be described next, completes the ATO requirement.

Supervision of the system

For the operation of any transit system, it is necessary to formulate a set of operating rules that indicate the procedures and priorities to be followed. One of the significant operating rules concerns scheduling—either on a timetable or a continuously variable basis—dependent upon demand conditions. Situations favoring timetable operation are infrequent scheduling and the requirement to meet feeder lines for interconnection. Situations favoring continuous variable scheduling include the optimum use of a single section that is fed by multiple branches during rush hours. Other special varying load conditions would involve sporting events, Christmas shopping, and emergency situations.

Train graphs are drawn to plan and visually demonstrate the running and dwell (stopping) times of trains on a transit system. Also indicated are the turnaround times at the end of the line and the schedule density at any time. These are time–distance curves, with time plotted on the X axis and distance on the Y axis. Figure 6 shows a portion of a typical train graph in which the horizontal lines represent intermediate stations, and the upper- and lower-limit lines are turnaround points. Schematic representation of train running time is used, and the dwell time is plotted along the horizontal lines that represent each station. Turnaround time allowance is made at each end of the line. This has been the traditional method of making up for time lost during each previous run.

The train graphs show that, although dwell time can make minor adjustments in performance, the only major tool available to make dynamic changes in schedule performance is to run at some percentage of maximum speed to maintain normal scheduling and to provide incremental increases between that and 100 percent performance. Without using this tool, lost schedule time can only be recovered by reducing dwell at the ends of the route.

Figure 7 shows a typical simplified example in which, if the schedule is based upon 92 percent of maximum performance, a delay of up to 43 seconds at point X can be corrected by the time a train reaches point Y, by using only the correction to 100 percent performance.

Without reducing the dwell time of station stops to a point of uselessness, only about 12 seconds of additional makeup time could be obtained. Thus the significance of using run time control is evident. It should be noted that correction for early arrival at a station can be made by increasing the dwell time so that the train departs at the correct schedule time.

Referring again to Fig. 7, we see that each 2 percent change in performance yields a schedule difference of 2.7 seconds in 3.2 km, thereby indicating the value of incremental changes. Note also that, in our hypothetical system, the average speed from one end of the run to the other equals 66 km/h at 92 percent of maximum performance.

The electronic computer in supervision

The schedule optimization of a system with branch lines is a simple matter when the digital computer is employed. This was demonstrated at the BARTD test track by Westinghouse Air Brake. The computer was a real-time, general-purpose DDP-24 digital computer built by the Computer Control Company. The machine used a 24-bit word length, had a 3-microsecond access time, and included indexing and indirect addressing features. A memory capacity of 12 000 words was provided to implement both the primary and secondary functions of the central office. The primary function was to maintain automatic line supervision over the entire railroad system. The secondary functions included the provisions of manual supervision and override capabilities, logging and analyzing system data, and operating status and operational displays.

The system, as demonstrated, involved either a relay or a flip-flop indication for each operating function such as track circuit occupancy, switch position, interlocking routes aligned and locked, etc. This equipment was housed at three locations along the track on which the functions were performed. At each location, a digital data link, with a 2000-bit-per-second transmission rate, was connected to all indicators and control input points. This was scanned continuously from the central office.

Man–machine interface. The man–machine interface for the central office consisted of the supervisor's control console and the system display board. All visual indications of system performance were presented by the system display board on a discrete zone basis. Items such as system status, alarms, zone occupancy, station headway, and schedule deviation were presented by combinations of color-coded "line of lights," illuminated signs, and digital readouts. Sectioning of the display at the discrete zones was done on a functional basis; such areas as stations, interlockings, turnbacks, and dispatch points were typical examples.

Functional zones. Each functional zone on the system

display board was further accessible for detailed display and manual control at the supervisor's console on a call-up basis. Console controls were available for implementing manual override of automatic train control, or for total manual operation of the system. The console was functionally sectionalized into five categories:

1. Run-time and dwell-time control and display.
2. Routing control and display.
3. Identification control and display.
4. System alarm annunciation and acknowledgment.
5. System master control.

An important point of hardware implementation was that all console and system display lights and readouts, and console buttons, were under control of the central computer. This meant that all buttons were of the momentary contact type and were continually scanned for contact closure by the computer during each control cycle. Therefore, no storage logic was necessary to store the supervisor's control action. Similarly, all incoming display information was continually decoded, analyzed, and updated by the control computer. The output data were routed to simple storage flip-flops in the control interface rack. Thus no complicated interface logic was necessary to decode incoming display information.

Plan of operation

To operate the system under computer supervision, a timetable called a "plan of operation" was written on the basis of one line per round trip per car or train. All elements of this plan of operation—run time, dwell time, departure time per station, "skip stop," enter and leave revenue service—were controllable. Each of these elements could be entered on a fixed or computer variable basis. Various plans of operation could be stored either on perforated or magnetic tape.

Upon entry and initiation of a plan of operation, the program utilized the typewriter to "ask" a series of questions concerning the starting position of equipment and the time. When all questions were answered and a starting time was entered, the program read the first three lines of the plan of operation, and these were executed on the basis of one round trip at a time.

Upon the completion of each line, one additional line was read in, and data on the actual performance versus the requested performance were printed as two lines on the medium-speed line printer. This information was also stored on tape for later off-line analysis of performance, miles traveled, etc. Limits of automatic correction were built into the computer program, and whenever performance exceeded these limits, alarms were lighted on the display board (which was organized to operate on the principle of management by acceptance).

To demonstrate and investigate computer performance in optimizing schedule, special demonstrations were run with the trains operating at 90 percent of their performance capability. Arbitrary schedule delays were then imposed on the trains, and the computer was allowed to recover schedule by correcting both the run time and a portion of dwell time at each station. Since the normal dwell time was 20 seconds, a minimum limit of 16 seconds per station was used.

In these demonstrations, the computer would react to a delay by imposing 16 seconds dwell time and 100 percent run time until the arrival at a station showed that a choice was available in order for the train to arrive at the next station with a minimum deviation from scheduled time. At this point, the computer would scan its performance data tables and select either a run-time 4 (83 percent of capability), a run-time 2 (95 percent), or a run-time 1 (100 percent) in order to arrive at the next station as nearly exactly on schedule as possible. Corrections for lateness of approximately three minutes were made in 1½ cycles of the test track (a distance of about 14.5 km, encompassing five station stops). Recovery in several instances resulted in departure time from the fifth station within two seconds of scheduled time.

Train identification, routing, and performance. Train identification was established at the beginning of each plan of operation and carried forward by the computer as it tracked train performance. Train routing was displayed on a matrix that was capable of showing any portion of the system. Manual controls were provided to establish any route desired by the supervisor other than that which was automatically established by the program.

Schedule performance was capable of displaying performance at any one station, and it showed, by digital indication, the number of the last train in the station, the remaining seconds of dwell time, and the number of any train that was next approaching the station, with a digital indication of the anticipated number of seconds until arrival. Manual override buttons were provided to vary dwell time, or percent of maximum performance, upon leaving the station.

Computer reliability. It was demonstrated that the digital computer can reliably handle the functions of automatic operation, display of performance, and manual override of automatic operation of any increment of the plan of operation, while the computer continues to handle the balance of the plan.

By recording all indications from the field on magnetic tape on a real-time basis, it was possible to rerun the tape at any time and recreate an operation for purposes of simulation. When this mode was effected, the computer received the information in the same manner as it would from the field and it was processed as if it were fresh information. This afforded a tool not only for complete simulation but also for debugging of any program problems.

Analog computation and station stopping

As part of the velocity control programmer carried on board each car or train, a small analog computer network generated an ideal time–distance profile for station stops from a speed of 128 km/h. Fixed wayside distance markers at 870 and 100 meters before each station stop provided distance references for this curve.

Regardless of the entering velocity of the train, it was permitted to continue at the speed limit until it approached the intersection of the time–distance stopping curve. At this point, an error measurement provided "anticipation" so that braking would start before the actual intersection of the velocity and the time–velocity curve. Actual performance was then fed back and measured against this curve, resulting in a demand for proper braking action to achieve the precision stop within 0.3 meter of the desired station point.

Results and conclusions from the demonstration

The demonstration of computer-controlled automatic train operation by WABCO–U.S. & S. (Westinghouse

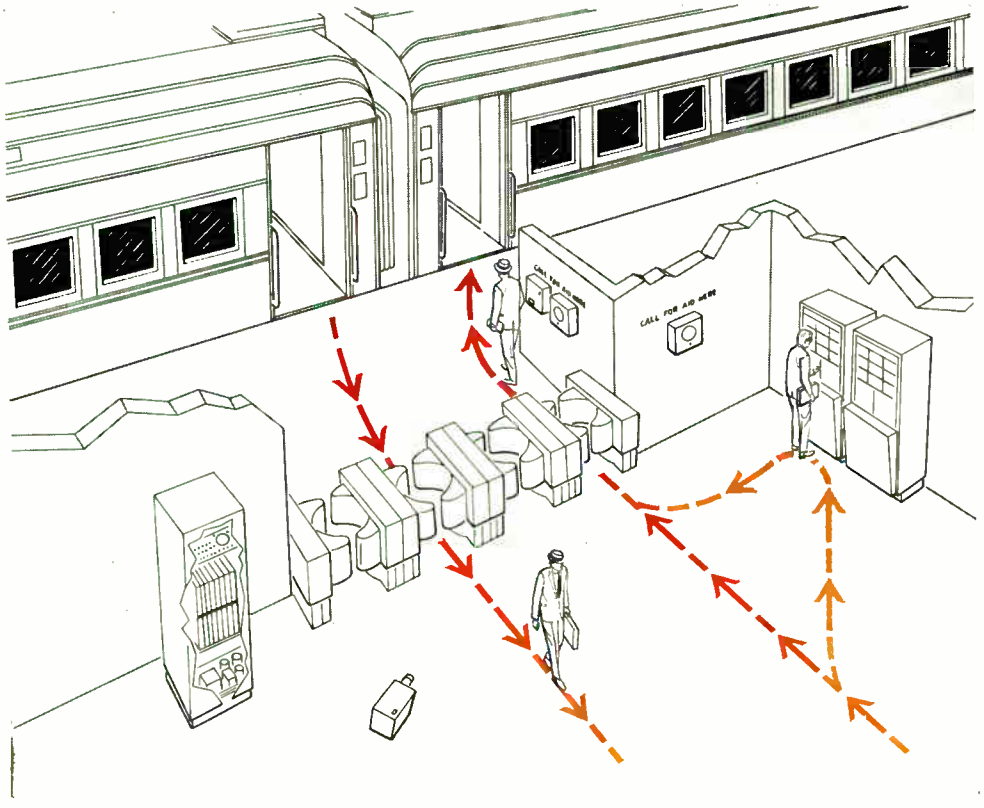


Fig. 8. Diagram of an automatic ticket vendor configuration with four entrance-exit gates.

Air Brake Co.–Union Switch & Signal Div.) proved that the primary function of transit schedule optimization can be readily accomplished by a digital computer. Also, the usefulness of analog techniques in the precise control of station stopping was demonstrated. Thus the importance of electronic computers in the supervisory function of mass transit operations was well established.

Philosophy of automatic fare collection

Among the many technological advances resulting from the resurgence of public transportation is the concept of automatic fare collection. This is the application of computer and data-processing techniques, combined with the total systems approach, to automate the sale and collection of passenger tickets and the handling of money in these processes. For many years the public transportation industry has been plagued by the expense, difficulty, and inconvenience of selling and collecting tickets manually.

Basic types of fare structure. Modern public transportation systems operate with two basic types of fare structure: a flat fare, and a zone—or graduated—fare. In the flat-fare system, the passenger pays a fixed amount when he enters and he is free to ride anywhere on the system since the distance traveled bears no relationship to the rate. The New York City subway system is an excellent example of this type of fare structure.

The graduated fare is related to the distance traveled, and with this type of fare structure the passenger must be checked at the entrance and exit to determine that he rides no farther than the distance allowed for the fare paid.

The graduated fare is equitable to the passenger since it allows flexibility in changing fare for small portions of the system instead of a system-wide fare change. Traditionally, the graduated fare has always brought in a substantially greater revenue than a similar flat-fare system. The graduated fare, however, is difficult and expensive to collect by manual means.

The automatic ticket vendor. To automate the graduated-fare system, the tickets must be "machine readable" and vended by passenger-operated automatic ticket vendors (Fig. 8), which accept money and encode tickets as they are issued to accomplish the selling function. These devices must be controlled by a computer that checks such things as type of ticket, ticket value, time of day, date, station of entry, and station of exit, and any number of variables that relate to a particular ticket. In some instances, the computer must calculate the fare due and subtract this fare from the original value encoded on the ticket and then write the remaining value.

Components of automatic fare-collection system

An automatic fare-collection system is made up of components such as machine-readable tickets, ticket readers, passenger gates, fare vendors, computers, data-collection equipment, ticket generators, and computer fill units.

There are a variety of machine-readable media—punched cards, magnetic ink, and magnetic tape—that can be used as tickets. The magnetic tape, or oxide surface, ticket seems to be the most reliable and flexible ticket system. With this medium, variable information can be read, written, or erased; a high density of information can be encoded; and counterfeiting is very improbable.

The device that actually provides the interface between the public and the equipment is the ticket reader. It will see constant action for several hours a day during peak traffic hours and it must be reliable. The passenger inserts his ticket in the gate as he enters. This action turns on a photocell and causes rollers to grasp the ticket and

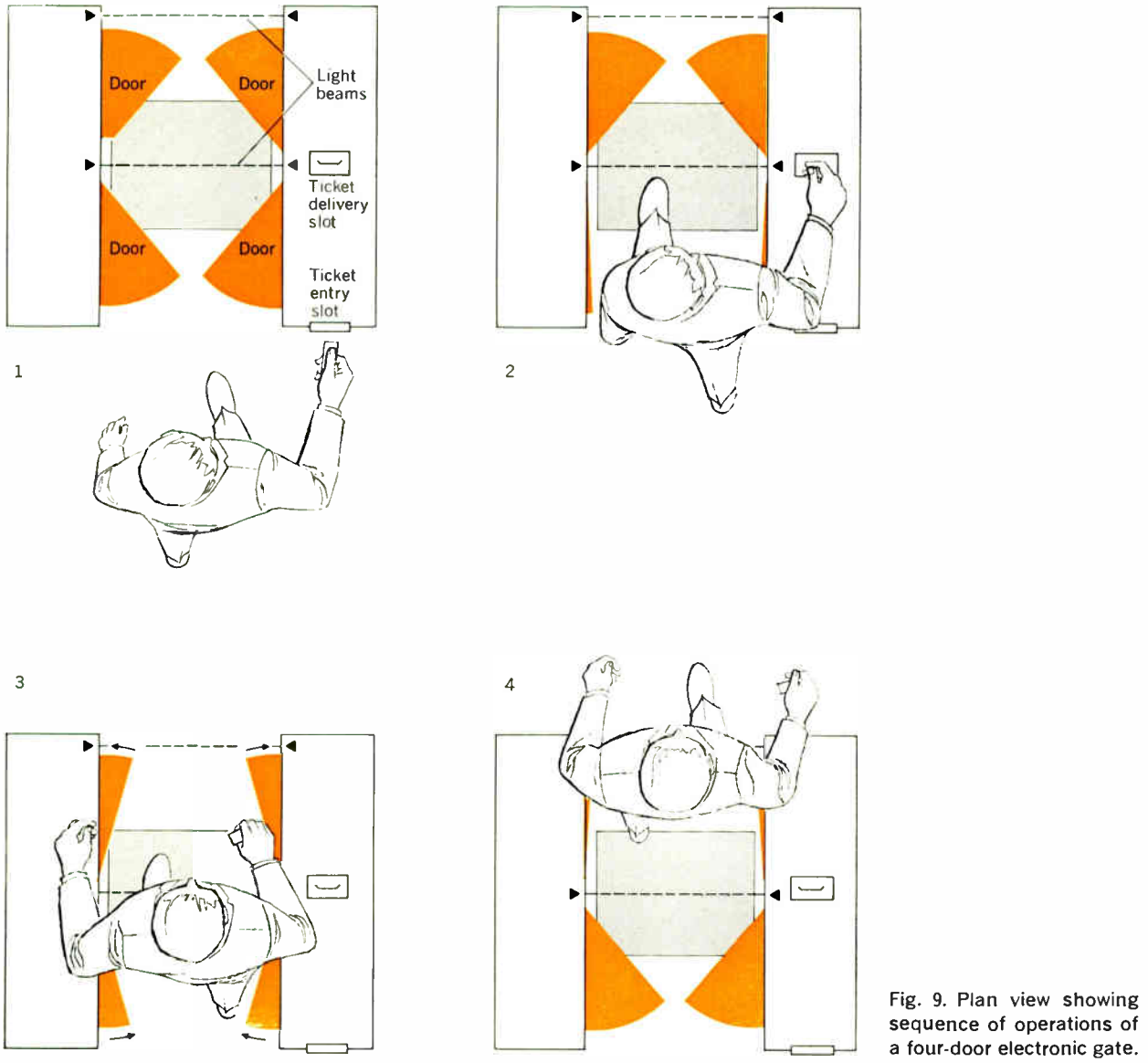


Fig. 9. Plan view showing sequence of operations of a four-door electronic gate.

carry it over read–write stations at 127 centimeters per second. Other photocells in the unit will tell when the ticket is in position to start reading and to start writing. The "bits" of information on the ticket are read, amplified, and transmitted to the computer. The computer will send the results of the transmitted information back to be rewritten. This complete transaction takes place in a fraction of a second from the time of insertion of the ticket and its retrieval by the passenger.

Passenger gates. The types of passenger gates may range from a tripod turnstile to a four-door gate and may be readily convertible for use as an entrance and exit. The gate is fast operating and its photocell logic determines the passenger's position and then coordinates this information with the logic that governs the ticket insertion (Fig. 9).

An item of prime importance (which will eventually be supplied to relieve the load on agent-operated encoders) is the fare vendor. The agent now encodes the "from" and "to" station information, the ticket type, date, and any other pertinent information. The passenger-operated fare vendor accomplishes the same function but, in addition, will accept bills and coins, total the amounts, and then convert the information to electronic logic for the dispensation of the ticket.

By taking all of the components and putting them together, automatic fare-collection systems can be created to cut the operating costs of the public transportation industry.

This article is based on the four papers of Session 16, "Electronic Systems for Urban Rapid Transportation," presented at the 1966 WESCON, Los Angeles, Calif., Aug. 23–26.

REFERENCES

1. Flory, L. E., "A system for electronic control of highway vehicles," *Proc. SRI Urban Transportation Alternatives Symp.*, Stanford Research Inst., Menlo Park, Calif., pp. 67–69, May 27, 1964.

2. Quintin, W. P., Jr., "Automatic train control for BART," presented at the 1966 Joint Railroad Conf. and Transportation Symp., San Francisco, Calif., May 1–6.

3. Field, R. K., "Electric car makers preparing to rally," *Electron. Des.*, vol. 14, no. 13, pp. 21–22, May 24, 1966.

4. Reid, W. T., "Energy sources for electrically powered automobiles," *Battelle Tech. Rev.*, vol. 14, pp. 9–15, Apr. 1965.

5. Herbert, E., "Transporting people," *Internat'l Sci. Tech.*, vol. 4, p. 32, Oct. 1965.

Deep-space optical communications

Recent investigations have shown that laser systems, particularly the incoherent direct detection and transmitted reference systems, have important potential advantages over local heterodyning techniques for achieving effective deep-space communications

E. Brookner, M. Kolker *Raytheon Company*

R. M. Wilmotte *Consultant*

A major problem in deep-space communication systems is that of obtaining high data rates (of the order of 10^7 bits per second). This article proposes some design concepts that indicate the probable feasibility of achieving wide-band communications by means of the laser. The example selected here is a hypothetical mission to Venus, chosen because of its great brightness and, hence, high background-noise level. Since no earth satellite relay is assumed, the communication channel includes the atmosphere. The down-link is the one considered because of its high-information-rate requirement.

The primary goals of any space subsystem are generally taken to be low power and low weight. For deep-space, wide-band communication, however, another factor may be equally important—namely, the size of the transmitting aperture. A very large aperture, as would undoubtedly be required by a microwave channel, is likely to prove an obstruction to the sensors of the spacecraft and will, therefore, reduce the time available for collecting information or transmitting it. In this respect, the laser has an important advantage over microwave, as will be seen later.

The magnitude of the deep-space communication problem is shown in Table I. At the August 23–27, 1965, conference at Virginia Polytechnic Institute on "The Exploration of Mars and Venus," the total information required in the early Voyager program was suggested to be of the order of 10^8 to 10^{10} bits. A typical imagery requirement of the geologist interested in rough terrain characteristics of Mars or Venus is 20-cm × 20-cm photography, by means of which a single picture could contain as many as 10^8 bits. Based on current capability of the order of 8 bits per second (b/s) of Mariner IV, it could take as long as 11 500 days to transmit 10^{10} bits. At the same conference, it was suggested that 1000 b/s may be achieved in time for the early Voyager program. As many as 115 days would be required, even at this data rate. At the 10-Mb/s data rate suggested as a minimum requirement, the transmission time assumes values in seconds.

To increase the information data rate capability significantly at radio frequencies implies consideration of larger antennas in the spacecraft and on the ground, increased power in the spacecraft, and use of higher frequencies (for example, EHF) with the commensurate development requirements and cost to change from the current NASA deep-space instrument facility S-band system. The laser—with its extremely narrow beam due to its short wavelength, notwithstanding its high quantum and background noise—offers the possibility of surpassing RF techniques in its ability to satisfy deep-space requirements. Should it prove superior to RF at data rates of the order of 10^7 b/s, its growth capability to higher data rates will be much greater than that of RF systems. We can expect the laser communication art to develop in all its component areas, as has been historically achieved in all new technologies.

I. Space communication problem

Early Voyager Requirements, total bits	Transmission Time at 10 b/s, days	at 100 b/s, days	at 1000 b/s, days	at 10^7 b/s, seconds
*10^8	115	11.5	1.15	10
10^{10}	11 500	1150	115	1000

* Approximate requirements for one 20-cm × 20-cm photo with 15 shades of gray at 10 lines/mm.

Candidate optical systems

Three types of systems have been considered:

1. Local heterodyne system (LHS)
2. Direct detection system (DDS)
3. Transmitted reference system (TRS)[1]

Block diagrams of these systems are shown in Fig. 1.

The local heterodyne system often provides the highest signal-to-noise ratio (SNR) of the three systems because the local heterodyne laser can be made sufficiently strong that the shot noise from it dominates all other sources of noise. However, as shall be indicated later, the DDS and TRS can be designed so that the power efficiency is nearly as high as that of the LHS. Moreover, the LHS system suffers from the serious disadvantage of receiver SNR degradation due to spatial dispersive effects of the atmosphere.

For LHS to operate properly, the local laser radiation should maintain spatial coherence with the received signal light over the whole receiving optical aperture. The atmosphere disperses the signal beam so that coherence is lost for much of the time except for very small apertures. For example, based on an atmospheric transmission experiment by Goldstein *et al.*[2] from noon to midnight over a 4-km path at a wavelength of 0.63 μm, a 3-cm dish would experience 70 percent of the time a loss of at least 15 dB greater than that experienced under atmospheric conditions that permit perfect coherence. During the experiment reception was found to be generally poor except shortly after sunset. Smaller losses may be expected for installations on high mountains and with longer wavelengths, but the prospects are not promising for maintaining reliably the spatial coherence across the aperture needed for an LHS. Another alternative for the LHS is to have a receiver system that consists of a large number of small diffraction-limited dishes. The randomness of the phases of the signals from each dish would then be compensated for by some adaptive scheme that adds the signals in phase.

LHS suffers from one other problem. The local laser must be maintained accurately at a specified frequency difference from that of the received signal light, and thus it is necessary that the local laser be continuously corrected for the Doppler shift. For some missions, the Doppler shift can be very large, typically well over 10 GHz.

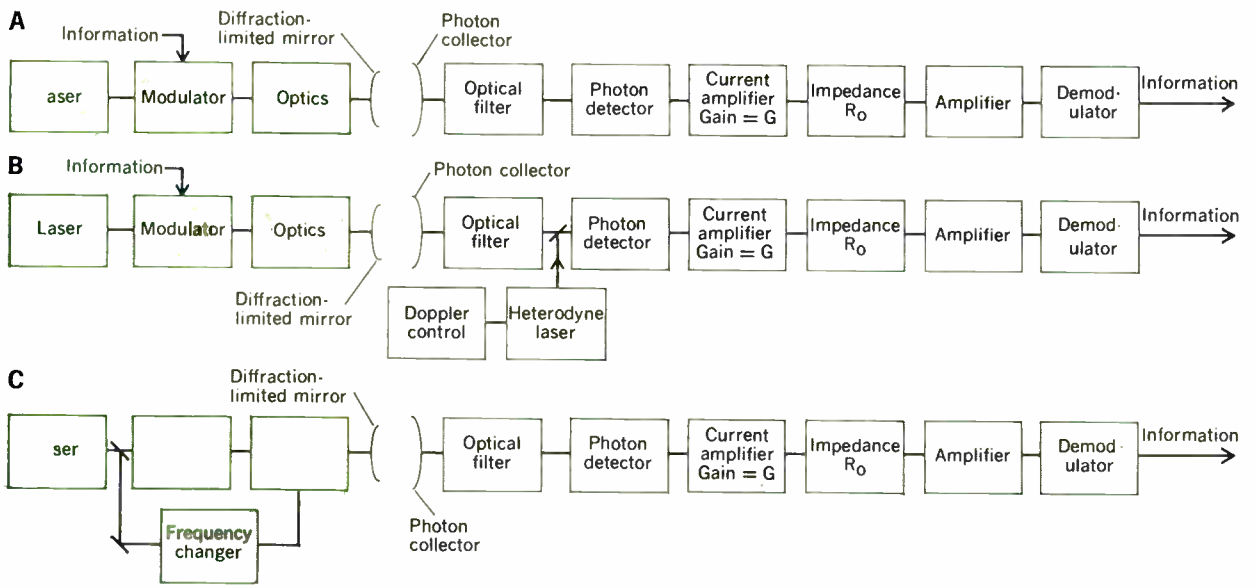
The direct detection system is simply a straightforward transmission and detection system, with a single modulated carrier providing video detection. It has a limitation in the loss of phase information of the carrier. Nevertheless, DDS appears to be the most attractive choice at this time.

The transmitted reference system is a heterodyne system in which the reference is transmitted with the signal from the spacecraft. This technique avoids the Doppler shift problem of LHS, but its SNR is lower than that of either the DDS or LHS. It is lower than that of the DDS principally because only half the power transmitted from the spacecraft is signal power. The successful performance of this system depends on the assumption that the atmosphere will not disperse the very close frequencies of the signal and the reference sufficiently to damage their spatial coherence over the receiver aperture. (The frequency separation is of the order of 0.2–10 GHz.) This system can use almost any form of modulation including phase-shift keying (PSK).

Diffraction-limited optics do not provide increased SNR for the DDS and TRS. Since nondiffraction-limited optics will simply have the effect of producing a focal area larger than the Airy disk, the detector is required to have a larger area. It is, therefore, possible to receive light signals through the atmosphere on extremely large nondiffraction-limited apertures with high efficiency by providing adequate detector area at the focus of the optics. For the greatest accuracy, it is desirable

1. That the optics do not enlarge the focal area beyond a diameter for which the collection of photons by the detector becomes difficult.
2. That the highest frequency of the modulation carried by the light beam remain coherent over the area

Fig. 1. Optical communication system types. A—Direct detection system. B—Local heterodyne system. C—Transmitted reference system.



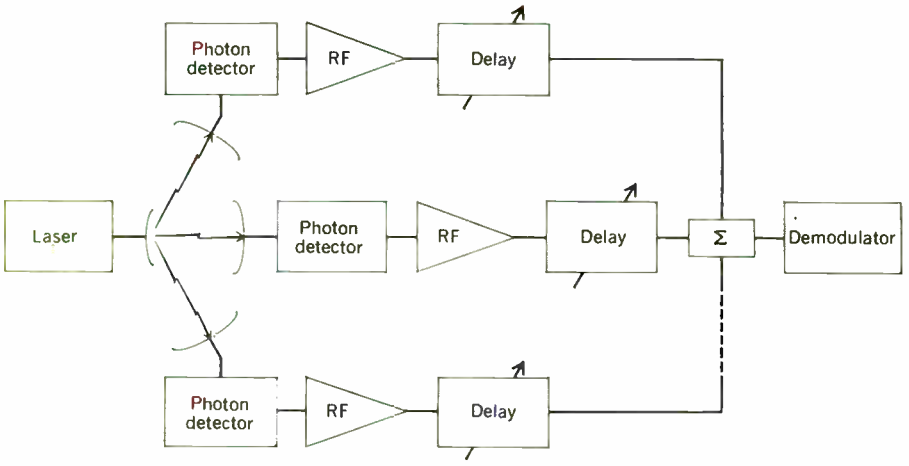


Fig. 2. Multinondiffraction-limited dish system.

of the detector; that is, the mechanical irregularities should not exceed the equivalent of about 1/10 of the wavelength of the highest modulation frequency.

When it is desired to have photon collectors of very large dimensions, the structural requirements may lead to the use of a number of nondiffraction-limited dishes, each with its own detector system; see Fig. 2. The continuously variable delays are introduced to compensate for the changes in path length with change in direction of the received beam.

The configuration of the several receiving apertures must be such that they will not interfere with one another in the directions of interest and such that the compensating delays can be accurately established. In the simplest configuration, all the dishes are installed on a single, very large structure. If it is sufficiently rigid, the whole structure can be moved normal to the direction of the received beam and compensating delays are not required except to correct for the effects of temperature and mechanical stresses.

Comparison of the LHS, DDS, and TRS

In order to compare the performances of the LHS, DDS, and TRS, an analysis was made of the transmitter power required by the three systems for shot-noise-limited conditions. The systems were put on an equal footing for the analysis, by assuming that all three systems have the same parameters—in particular, the same transmitter frequency and area, the same total receiver background noise, and the same detector efficiency. Multiple-dish receiver systems are assumed, with the total receiving area (not necessarily the number of dishes) the same for the systems. For the LHS, it was also assumed that the atmosphere does not degrade the coherence of the incoming field. Equivalently, it was assumed that the spatial incoherence of the received signal is compensated for by an adaptive technique for adding in phase the signals from the various dishes. For the TRS the signals from the outputs of the detectors are added in phase, whereas for the DDS they are added linearly, as indicated in Fig. 2. The pessimistic assumption is made that the background noise is spatially coherent over the total collecting area of the systems. And it is assumed that the background noise radiates from a point source, as would be the case if the background arose from Venus' albedo.

One finds that for these assumptions on the background noise the performances of all three systems are dependent on the receiver collecting area and not on the number of dishes involved. This important property applies theoretically to the TRS, LHS, and DDS systems, regardless of the output SNR per dish. Let

P_{LHS} = transmitted power required for the LHS
P_{DDS} = transmitted power required for the DDS
P_{TRS} = transmitted power required for the TRS
N_b = background noise received by each dish (after optical filtering), photons per second
M = number of dishes
N_{bT} = total background noise received by the M dishes (after optical filtering), photons per second
B_T = signal bandwidth
α = quantum efficiency of the receiver
X_{SN} = power SNR at output of receiver sum point
X_{bT} = $B_T/\alpha N_{bT} = B_T/\alpha M N_b$
= power SNR at output of receiver sum point, if there were collected by the receiver antenna complex one photoelectron per hertz of transmitted signal bandwidth

Figures 3 and 4 give plots of $P_{\mathrm{DDS}}/P_{\mathrm{LHS}}$ and $P_{\mathrm{TRS}}/2P_{\mathrm{LHS}}$ versus X_{bT} for $X_{SN} = 10$. It is noted that the curves are independent of M in accordance with the results already given above. The curves indicate that for high X_{bT}, the incoherent DDS has a power efficiency as high as that of the LHS. The TRS requires four times as much power as the LHS for high values of X_{bT} because half the transmitter power is in the information-carrying part of the signal, which results in a fourfold decrease in the power SNR after detection. By proper design of the system (that is, by the use of a wide bandwidth for the signal when necessary and a narrow-band optical filter and small field of view, the quantity X_{bT} can be made large.

It is found that one generally can achieve, or come close to achieving, shot-noise-limited conditions for the DDS and TRS. A sufficient condition for the DDS and TRS to be shot-noise limited is that

$$\frac{B_T}{B_o}\frac{1}{X_{bT}} = \frac{\alpha N_{bT}}{B_o} \ll 1$$

This condition is met for the systems in Table II that utilize photomultiplier detectors (Systems 1, 4, and 5).

For all these systems $\alpha N_{bT}/B_o \leq 0.1$. It is important to point out, however, that if the systems are partially or completely limited by classical background noise (instead of shot noise), the power performance of the systems is even closer than is indicated in Figs. 3 and 4. The systems become classical background-noise limited when the direction of inequality is reversed in the above equation. Using the results of Fig. 3 one finds then that when

$$\frac{1}{X_{bT}} = \frac{N_{bT}}{B_T} \leqslant 2$$

and $X_{SN} = 10$, the DDS will require less than 1 dB more transmitter power than an equivalent diffraction-limited LHS if the DDS is limited by shot or background noise instead of detector- or receiver-generated noise. These conditions are met for the GaAs DDS using a photomultiplier detector. Moreover, it is found that this DDS requires the same power as the comparable LHS.

So far in the foregoing discussion the comparison has been on the basis that the three systems are operating at the same frequency with no concern being given to the number of receiver dishes required by each system. Now a comparison is made for different frequencies of operation. Table III gives a comparison of the DDS and the LHS, both operating at 0.84 μm and 10 μm. In the derivation of the table, it was assumed that the systems are signal shot-noise limited with X_{bT} large so the DDS has as high a power efficiency as the LHS. To put the systems on the same basis for comparison, they were specified this time to have the same transmitter power, the same receiving area, and the same detector efficiency.

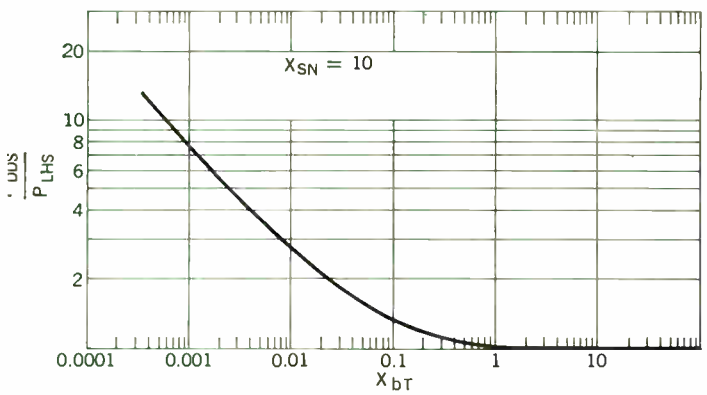
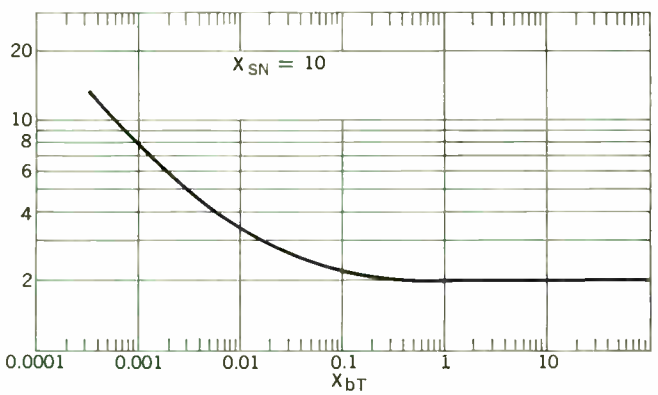


Fig. 3. P_{DDS}/P_{LHS} as a function of X_{bT}.

Fig. 4. $P_{TRS}/2P_{LHS}$ as a function of X_{bT}.



The DDS is assumed to have the same receiver dish configuration as specified for the Venus mission given in Table II; that is, it consists of 25 ten-meter dishes. The assumption is again made that the degrading effects of the atmosphere can be ignored for the LHS.

What was allowed to vary for the systems and serve as a parameter for comparison is the transmitter dish diameter. The transmitter dish diameters were set so as to give the same receiver SNR in all the systems. Also used as a basis of comparison is the number of receiver dishes required for the systems. The sizes of the transmitter dishes required are given in Table III in terms of the dish diameter D_{TO} required for the LHS operated at 0.84 μm. It can be seen that at 0.84 μm the transmitter dish needed for the DDS is the same as for the LHS and hence, on this basis, the systems are equivalent. However, when the systems are compared on the basis of the number of dishes required, this is not the case. In particular, the LHS requires 2.8 million dishes if one uses a dish size of 3 cm in order to attempt to eliminate atmospheric degradation. This is in contrast to requiring only 25 dishes for the DDS system. As was noted previously,[2] the dish size of 3 cm actually is too large, providing serious degradation a large percentage of the time. Hence, the performance of the LHS will actually be worse than indicated.

II. Mission to Venus

System Number	Laser	Detector	Transmitter Aperture Diameter, cm DDS	TRS
1	GaAs	S-1 photomultiplier	29	56
2	GaAs	Diode	209	116
3	GaAs	Avalanche diode	29	31
4	Semiconductor in visible (unavailable)	S-20 photomultiplier ($\lambda = 0.42$ μm)	3.5	6
5	Argon II	S-20 photomultiplier	78	154
6	N_2-CO_2	Cu-doped germanium	30 000	42 300
7	N_2-CO_2	Ideal (not available)	125	221
8	Ho-doped YAG	Ideal (not available)	7	15

Microwave S-band system: diameter = 2000 cm

Notes:
Distance = 180 million km
Power input to transmitter = 30 watts
Information rate = 10^7 b/s Error rate = 10^{-4}
Laser receiver: 25 apertures, each 10 meters in diameter
Microwave receiver: one paraboloid, 50 meters in diameter
Modulator: PCM/PPM, alphabet size of 32, $B_T = 10^8$ Hz

III. Comparison of DDS and LHS

Characteristic	DDS		LHS	
Wavelength (λ), micrometers	0.84	10	0.84	10
Transmitter dish diameter required	D_{TO}	3.5 D_{TO}	D_{TO}	3.5 D_{TO}
Receiver dish diameter, meters	10	10	0.03	0.3
Number of receiver dishes (M)	25	25	2.8×10^6	27 800

Table III also indicates that the transmitter dish diameter for the 10-μm system has to be about 3.5 times that for the 0.84-μm system. Use of a CO_2 laser operating at 10 μm for the LHS will offer considerable improvement as far as the number of dishes necessary; however, as the table shows, an excessively large number still can be expected to be required. The spatial correlation distance is approximately proportional to the signal wavelength; hence, for the 10-μm system, a receiver dish diameter of 0.3 meter should be used. The table indicates that for this dish diameter 27 800 dishes are required. Even if one optimistically assumes that a receiver dish diameter of 2 meters can be used, 625 dishes would be needed for the ground complex.

System analysis

Sources of noise. The four sources of noise in an optical detector are as follows:

1. *Thermal.* Similar to that at microwave frequencies.
2. *Quantum or shot.* Very high compared with that at microwave frequencies. This category includes the shot noise due to the signal photons, the background photons, and the photons that are equivalent to the dark current.
3. *Dark current.* Adds to shot noise. There is no equivalent at microwave frequencies.
4. *Background.* Very high compared with normal operations in the microwave region. The background calculated for the systems listed in Table II was based on irradiance of 10^{-10} W/cm$^2 \cdot \mu$m for Venus, and a radiance from the sun-illuminated atmosphere of 1.3×10^{-4} W/cm$^2 \cdot$sr$\cdot \mu$m.[3]

Other atmospheric effects. Because of the earth's rotation, at least three ground sites are required. In addition, to minimize the attenuation that may occur due to bad weather, these sites must be selected for their high probability of clear weather. There are such areas on the earth.[4] The probability of clear weather can be further increased by providing redundancy with additional sites. The values used for atmospheric attenuation in the systems listed in Table II are for clear weather.[5]

Transmitter optics. Spacecraft transmitter optics must be small and light. A laser should be excellent for this purpose. If the beam from it is perfectly coherent, it can in theory be focused to a point of dimension of the order of a few wavelengths of light, so that it is possible to make full use of the collimating capability of diffraction-limited optics. The limitations for achieving this are (1) the degree to which the laser beam is truly coherent and (2) the stability of the lasing area.

Gas lasers are currently the best lasers for meeting these two conditions. Semiconductor lasers are at present poor in this respect. Although the gallium arsenide laser may not be steady at present (no measurements are known to have been made to determine the lasing area stability), it is expected that it will be when single-mode operation is achieved. It may be necessary, however, to maintain the temperature very constant.

As to the optical mirror, diffraction-limited dishes up to one meter are currently being discussed for deep-space communication transmitters. Such a size appears feasible in view of plans for the Orbiting Astronomical Observatory (OAO) program to orbit a telescope of this diameter. Such large dishes lead to difficult design problems, including an extreme tracking requirement (for example, about ½ μrad), which is also planned in the OAO program, and maintaining alignment of optical system elements and diffraction-limited characteristics after withstanding the launch and space environment during a long mission to the planets. These problems increase rapidly with the size of the aperture. It is, therefore, important to minimize the aperture size.

Optical filter. An important component of the receiving optics is the optical filter, which is incorporated to reduce the background noise. The bandwidth of the filter will usually be large compared with the modulation bandwidth.

Sharp filters operate on the light interference and are sensitive to the angle of incidence. It is, therefore, important to insure that all the signal light is incident on the filter within its angular field of view, and no background light which is incident at larger angles reaches the detector.

Lyott filters are attractive because they provide a wide field of view with narrow bandwidth. A filter of 0.5 Å (0.05 nm) at a wavelength of 0.84 μm, about 5 cm in diameter and 40 cm long, can be made, using calcite and quartz, with a field of view of 5° (0.1 rad). The filter, which could be tuned through ±0.5 Å, is expected to have a transmissivity of 0.15. It is sensitive to temperature changes, which should be maintained within 0.1°K.

Transmitter laser choice. Of the three types of lasers—gas, solid-state, and semiconductor—the most desirable for ultimate development for deep-space communications is the semiconductor type (currently GaAs) because of its small size and weight, its promise of ready capability for wide-band pulsed internal modulation with simple techniques, and its potential for high efficiency (between 0.3 and 0.6) at reasonable temperatures.

In the visible region, gas lasers having a single mode, very narrow bandwidth, and high power can be made; however, the efficiency is low—about 0.1 percent or less for the narrow-band, single-mode operation. In the far-infrared region, molecular gas lasers have recently appeared. The N_2-CO_2 laser (λ = 10.55 μm), which is receiving much attention at present, has been made with an efficiency of about 10 percent. This wavelength falls within a wide atmospheric window. However, compared with GaAs, it has the following disadvantages:

1. It is not possible to obtain high data rate and very narrow pulsed operation. At present, fundamental limitations rule out wide-band internal modulation.
2. For a given diameter of the diffraction-limited dish in the spacecraft, the gain is −22 dB relative to GaAs (because of longer wavelength).

It remains also to develop wide-band detectors that are not detector-noise limited for the direct detection and transmitted reference system. The transmissivity of the atmosphere is about the same in clear weather as at a 0.84-μm/wavelength. Although the N_2-CO_2 laser will operate satisfactory in worse weather conditions than the GaAs laser will, there are weather conditions in which neither laser can operate.

Solid-state lasers, such as ruby, that radiate in the visible region have low power efficiencies (less than 1 percent) and are useful mainly for high-peak power pulses at low repetition rates. A holmium-doped yttrium aluminum garnet (YAG) laser has been made to lase CW in the infrared region at 2.3 μm with a power efficiency of 5 percent at liquid nitrogen temperatures (77°K); a

realizable efficiency of 10 percent seems reasonable. Outputs of the order of a few watts were achieved, with much higher outputs being anticipated. As in the case of N_2-CO_2, there is the problem of developing an efficient modulator for obtaining narrow (nanosecond) pulses at a high data rate of the type desired. High-efficiency pulsed operation with higher repetition rates than tens of kilohertz appears unfeasible because of the lack of a flash lamp that can operate at these higher rates. Also, there remains the equally important problem of developing a good detector for receiver systems that do not use local heterodyning.

Semiconductor lasers (for example, GaAs at 0.84 μm) are at present the most promising, but require considerable development before they can be effectively used for deep-space communication. The present problems with GaAs lasers are concerned with

1. *Multimode operation.* The fluorescence bandwidth is very large, about 200 Å wide; most of the power is within a band of about 20 Å (860 GHz). It comprises a large number of equispaced lines, each about 10 MHz wide. After single-mode operation is achieved, the type of modulation selected must, therefore, be able to operate with a carrier having this bandwidth.
2. *Stability.* The lasing area may shift in position when it is internally modulated. The lasing area must be extremely stable if we are to make full use of the gain capability of the diffraction-limited optics at the transmitter.
3. *Power.* Continuous-wave power of the order of 10 watts requires operation at 4°K, a temperature difficult to reach in a spacecraft. However, it is expected that temperatures of 77°K might be achieved in a spacecraft and that the GaAs laser can be developed to powers in excess of 10 watts at such temperatures or higher. One-watt CW power has been achieved to date at 77°K.

It is anticipated that solving the first two problems will permit optical collimation down to a narrow beam.

Although gas lasers currently have the desirable characteristics that the semiconductor laser has yet to achieve (single-mode operation with spatially coherent, stable output), future developments in the problem areas just mentioned may lead to the choice of a semiconductor laser type for deep-space communications.

Coding and modulation. Table IV[6] is based on Gaussian noise statistics that apply to microwave communications. With the quantum nature of light and the unknown statistics of atmospheric effects, comparable figures for optical communication must still be derived. However, the relative order of magnitude of the required SNR per bit for each type shown is expected to hold for optical communications.

An optical beam can be modulated in phase, amplitude, and polarization. The last has some valuable characteristics. The relative merits of these are not discussed herein. The modulation type selected to compare the systems listed in Table II is PCM/PPM with an alphabet size of 32 (PCM-32 orthogonal). A PCM system of larger alphabet is not used because of the complexities involved. Sequential decoding is not chosen because it is not economically feasible for high data rates. Moreover, the small bit-error rates (of the order of 10^{-6} or less) that can be achieved by using sophisticated codes is not necessary for the high-data-rate video picture communication. Bit-error probability of the order of 10^{-3} or 10^{-2} can produce pictures of satisfactory quality. Of course, engineering data, which would be transmitted at a lower data rate than video data (a rate of 10^5 b/s as compared with 10^7 b/s), need a bit-error probability of about 10^{-5} and hence require error-correcting codes. In Table II, the conservative bit-error probability of 10^{-4} was used for comparison purposes for the high data rate transmission.

Detectors. Photodetectors are in effect photon-to-electron converters. In the course of conversion, they provide gain and noise in varying degrees. With the weak light intensity of deep-space communication, the possibility of gain in the photodetector is of prime importance to minimize the effect of thermal noise generated in the output resistor. Two detectors are of special interest in this respect. The first, the photomultiplier, is excellent because of its high gain (of the order of 10^6) with little generation of noise; however, it has only a fair quantum efficiency, which rapidly becomes poor beyond a wavelength of about 0.7 μm. The second detector of interest is the microplasma free avalanche diode, recently developed by Bell Telephone Laboratories.[7,8]

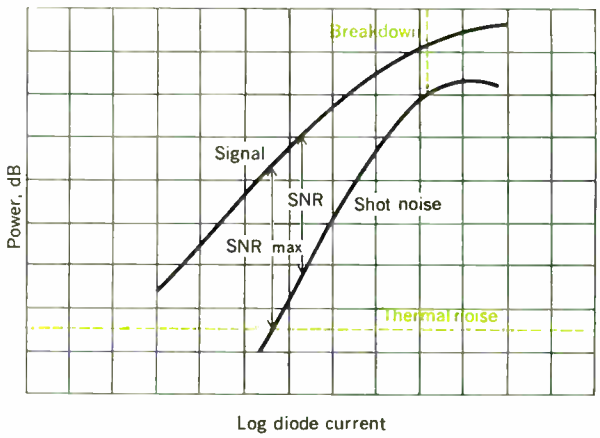
In an avalanche diode recently tested, the shot noise is proportional to the cube of the current gain of the de-

IV. Modulation and coding for bit-error probability $P_e \leq 10^{-4}$

Modulation and coding	SNR per Bit Required, dB	Signal Bandwidth Required, Hz
PAM-DSBSC* Incoherent	19	2×10^7
Coherent	13	2×10^7
PAM-PPM, PAM-FM	6	6×10^7
PCM-Binary (Incoherent)	13	10^7
Orthogonal (DPSK†)	10	10^7
PCM-32 Orthogonal	7	6×10^7
PCM-1024 Orthogonal	5	10^9
Sequential decoding		
Binary PSK, 3-bit quantized detector	2.4	5×10^7
32 orthogonal, list of 8 decoding	5.4	3×10^7
Shannon channel limit	−1.6	—

* Pulse amplitude modulation, double-sideband suppressed carrier.
† Differential phase-shift keying.

Fig. 5. Signal-to-noise ratio for avalanche diode.



tector, whereas the signal is proportional to the square of the gain; see Fig. 5. The ordinate distance between the signal and noise curves is the SNR. It is clear that the smaller the gain, the greater the SNR down to the point where the constant noise—that is, the thermal noise—takes over. For maximum SNR, therefore, the gain should be adjusted so that the shot noise is approximately equal to the thermal noise, and the thermal noise should be kept to a minimum by cooling.

The characteristics obtained to date and predicted on the silicon diode for 0.75 μm (gallium arsenide phosphide laser) are:

1. The shot noise increases as G^m, where m generally lies between 2.5 and 3.0. To date the figure is 3.0. This figure will probably be reduced.
2. Gain–bandwidth product is 10^{11}. It is not likely to be increased materially, possibly up to 2×10^{11}.
3. Avalanche operation is stable.
4. The sensitive area is very small, about 0.005 cm in diameter.
5. Quantum efficiency is high, over 0.5.
6. Dark current between 10^{-10} and 10^{-11} ampere at room temperature. It varies with the gain.

The germanium diode has similar avalanche characteristics at 0.84 μm. At room temperature, however, its dark current noise is too high, although at low temperatures it is likely to be acceptably low.

Laser–detector combination

At this stage of laser and detector development, there is no combination that could be said to be unquestionably superior to microwave. In broad summary, one can say:

1. He-Ne and argon gas lasers are too inefficient.
2. Although the molecular gas lasers have the advantage of reasonably high efficiency, they have a number of disadvantages that must be overcome if they are to be used most effectively for deep-space high-data-rate links. For one thing, it is not possible at present to obtain very wide-band pulse modulation (for example, pulse widths of the order of nanoseconds for a PCM incoherent direct-detection communication link). Also, it is necessary to develop exceptionally good detectors. Finally, the long wavelength has the disadvantage of requiring a larger transmitter aperture, all other things being equal.
3. Solid-state lasers in the visible region are not efficient and operate at low repetition rates. Holmium-doped YAG in the infrared region has a high efficiency but requires the development of efficient wide-band pulsed modulators and detectors for systems that do not use local heterodyning. In addition, it requires cooling to 77°K.
4. Semiconductors are most desirable because of their efficiency, size, weight, and potential ease of modulation; however, the most efficient to date, GaAs, operating in the near infrared, suffers from multimode and possibly spatial instability and, as yet, only one watt has been achieved at 77°K. Developed detectors at this wavelength (photomultipliers having poor quantum efficiency and diodes with too small gains) are of only fair quality.

Size of optics

It is believed that an important criterion of the acceptability of a system will be the size of the aperture in the spacecraft, to minimize obstruction to view. In this respect, the laser appears far superior to other communication systems. For laser operation, small optics in the spacecraft are desirable also for four other reasons:

1. Simpler pointing and tracking equipment.
2. Reduced sensitivity to vibration and distortion due to temperature gradients.
3. Simpler optics in the spacecraft for illuminating the diffraction-limited mirror with the laser.
4. Higher antenna efficiency.

Quantitative comparison of the DDS and TRS

In order to compare possible future systems, a specific communications mission to Venus was considered. Its principal components are listed in Table II. The criterion for the comparison of the several systems shown is the diameter of a diffraction-limited dish in the spacecraft that will provide the communication performance specified for an input power of 30 watts to the laser or microwave transmitter. PCM/PPM with an alphabet size of 32 was chosen, as indicated previously. The results listed below were calculated from the quantities shown in the Appendix and in Table V. Analysis of Table V shows that maximizing the ratio of the transmitted bandwidth to the information bandwidth significantly affects the system SNR. For the system examples presented,

V. Expressions for output signal and noise

	DDS	LHS	TRS
Signal	$\alpha^2 q^2 G^2 R_o N_s^2$	$\alpha^2 q^2 G^2 R_o N_h N_s$	$\alpha^2 q^2 G^2 R_o N_s^2$
Noise			
Thermal	kTB_T	kTB_T	kTB_T
Shot	$2\alpha q^2 G^m R_o (N_s + N_b + N_d) B_T$	$2\alpha q^2 G^m R_o N_h B_T$	$2\alpha q^2 G^m R_o (2N_s + N_b + N_d) B_T$
Background	$\alpha^2 q^2 G^2 R_o N_h (2N_s + N_b) \frac{B_T}{B_o}$	$2\alpha^2 q^2 G^2 R_o N_h N_h \frac{B_T}{B_o}$	$\alpha^2 q^2 G^2 R_o N_h (4N_s + N_b) \frac{B_T}{B_o}$
Noise-signal			
Thermal	$\frac{5.4 \times 10^{11}\, TB_T}{\alpha^2 G^2 R_o N_s^2}$	$\frac{5.4 \times 10^{14}\, TB_T}{\alpha^2 G^2 R_o N_h N_s}$	$\frac{5.4 \times 10^{14}\, TB_T}{\alpha^2 G^2 R_o N_s^2}$
Shot	$\frac{2G^{m-2}(N_s + N_b + N_d)B_T}{\alpha N_s^2}$	$\frac{2G^{m-2}B_T}{\alpha N_s}$	$\frac{2G^{m-2}(2N_s + N_b + N_d)B_T}{\alpha N_s^2}$
Background	$\frac{N_b(2N_s + N_b)B_T}{N_s^2 B_o}$	$\frac{2N_b B_T}{N_s B_o}$	$\frac{N_b(4N_s + N_b)B_T}{N_s^2 B_o}$

Notes:
1. $N_s \propto B_T/B_i$ 2. $N_b \propto B_o$ 3. $m = 2$, except for avalanche diode, where $2.5 < m < 5$. Chosen as 3 in calculations.

a 0.1-GHz bandwidth for pulses of the order of 10 ns was chosen for the PPM. It should be recognized that such a high modulation bandwidth may be difficult to achieve because of modulator limitations, and even if achieved it may reduce the efficiency of the laser.

The following comments refer to the DDS listed in Table II.

1. The GaAs laser with S-1 photomultiplier is attractive despite the low quantum efficiency of the S-1 phosphor. In view of the high gain possible with the photomultiplier, the thermal noise is small relative to shot noise.

2. The GaAs laser with diode detector is worse than System 1, because the gain is unity and thus the thermal noise becomes dominant. The quantum efficiency of 0.5 is excellent.

3. In the case of the GaAs laser with an avalanche diode detector, the results in the table indicate that at the 0.84-μm wavelength, this detector gives efficiencies as great as the photomultiplier detector.

4. The use of a semiconductor laser in the visible region, with an S-20 photomultiplier detector, shows excellent performance potential; however, there are no signs at this time of the possibility of developing a high-power semiconductor laser in the visible part of the spectrum.

5. The argon laser with an S-20 photomultiplier is a bit poorer than System 1 because of the very low efficiency of its laser.

6. The N_2-CO_2 system with Cu-doped germanium is poor, because it is detector-noise limited.

7. When N_2-CO_2 is used with an ideal detector, the results indicate a high loss imposed by the longer wavelength, which would occur even if a good detector were developed.

8. The Ho-doped YAG system may be fairly attractive if the postulated detector (for which a quantum efficiency of 0.3 was assumed in deriving the table) can be developed.

The comments on the first five of these systems apply to TRS as well. In every case, except for the DDS using the thermal-noise-limited diode detector, TRS is worse than DDS, largely because only half the laser power is transmitted as the signal.

The microwave system was calculated for the S-band region (λ = 10 cm) for a single receiver dish 50 meters in diameter (equal in area to the 25 dishes, each 10 meters in diameter, assumed for laser communication).

Appendix. Symbols and quantities of basic configurations for mission to Venus

1. B_i, information bandwidth and information rate = 10^7 Hz.
2. B_T, transmission bandwidth = 10^9 Hz.
3. B_o, optical filter bandwidth, Hz; $0.015/\lambda^2$ for 0.5 Å filter.
4. D_r, diameter of one ground dish, 10 meters. There are 25 such dishes. The angle of view of each is 0.2 mrad, so that effect of radiance of atmosphere is small compared to that of Venus.
5. D_T, diameter of spacecraft dish.
6. Error rate = 10^{-4}.
7. G, gain in detector: 10^6 for photomultiplier; 1 for diode; for avalanche, calculate for condition of shot noise = thermal noise.
8. H_s, irradiance of Venus = 10^{-10} W/cm$^2\cdot\mu$m; H_a, radiance of sunlit atmosphere = 1.3×10^{-4} W/cm$^2\cdot$sr$\cdot\mu$m.
9. l, optical transmissivity; l_o for whole system, 0.05; l_t for transmitter, 0.5; l_p for atmosphere, 0.7; l_r for receiver, 0.15 (includes optical filter).
10. N, number of effective photons incident per second on the detector; N_s for signal, N_h for heterodyne reference, N_b for background, N_d for equivalent dark current (negligible in systems of Table II).
11. m, exponent of G for shot effect; equals 2 except for avalanche diode when it is between 2.5 and 3.
12. P_L, power of laser radiation, watts; P_{si} = power incident in detector, P_{bi} = background power on detector.
13. R, range = 1.8×10^8 km (10^8 nmi).
14. R_o, output impedance = 50 ohms.
15. T, temperature of output impedance = 20°K.
16. α, quantum efficiency = 3.6×10^{-3} for System 1, 0.5 for diodes, 0.18 for Systems 4 and 5.
17. λ, light wavelength = 0.48×10^{-4} cm for argon.
18. h, Planck's constant = 6.62×10^{-34} J·s.
19. k, Boltzmann's constant = 1.38×10^{-23} J/°K.
20. q, electron charge = 1.6×10^{-19} coulomb.
21. The SNR required for PCM/PPM with an alphabet size of 32 for laser communication is 10.
22. Range equation:

$$P_{si} = \left(\frac{\pi}{4}\right)^2 \frac{D_T^2 D_R^2 l_o P_L}{\lambda^2 R^2}$$

23. $$P_{si} = \frac{3 \times 10^{10}}{\lambda} hN$$

24. Effective signal power at receiver = $P_{si} \dfrac{B_T}{B_i} n$

25. Efficiency of microwave antennas = 0.7, of microwave transmitter = 0.33, of semiconductor lasers = 0.33, of argon laser = 0.0005, of N_2-CO_2 laser = 0.1, of Ho-doped YAG = 0.1.

Revised text of a paper presented at the 1966 National Telemetering Conference, Boston, Mass., May 10–12, and published in the Conference Proceedings, pp. 36 41.

REFERENCES

1. "Optical noise discrimination techniques study," Tech. Rept. AFAL-TR-65-149, Wright-Patterson Air Force Base, Ohio, pp. 137–164, Aug. 1965, prepared by Raytheon Company.

2. Goldstein, I., Miles, P. A., and Chabot, A., "Heterodyne measurements of light propagation through atmospheric turbulence," *Proc. IEEE*, vol. 53, pp. 1172–1180, Sept. 1965.

3. Bell, E. E., *et al.*, "Infrared techniques and measurements," Final Eng. Rept., Research Foundation, Ohio State University, Oct. 1957.

4. "Study on optical communications from deep space – interim progress report 27 March through 31 May 1963," NASA Rept. SSD 3166R, DCOCS-4, pp. 1–13, prepared by Hughes Aircraft Company.

5. "Determination of optical technology experiments for a satellite," Perkin-Elmer Eng. Rept. 7846, pp. 3–34, July–Nov. 1964.

6. Jacobs, I., Internal memo., Raytheon Company.

7. Anderson, L. K., McMullin, P. G., D'Asaro, L. A., and Goetzberger, A., "Microwave photodiodes exhibiting microplasma-free carrier multiplication," *Appl. Phys. Letters*, vol. 6, pp. 62 64, Feb. 1965.

8. Melchior, H., and Lynch, W. T., "Signal and noise response of high-speed germanium avalanche photodiode," *IEEE Trans. on Electron Devices*, vol. ED-13, Dec. 1966.

Computing reliable power spectra

With signals at very low or very high frequencies, digital methods of computation appear to offer the most direct route to power spectrum analysis. However, such computations involve considerations that may be all too easily overlooked by the neophyte

Paul I. Richards *Technical Operations Research*

Techniques of harmonic analysis have recently crystallized into modern power-spectrum analysis, the presently accepted "best" tool for uncovering periodicities and near-periodicities hidden in noisy data. Despite the simplicity of this concept—to electrical engineers, at least—computing a meaningful power spectrum involves several subtleties that are too easily overlooked by the nonspecialist. This article presents an elementary review of such pitfalls and of their origins.

For an electrical engineer, it is almost a way of life to think of complicated signals as superpositions of sine waves and there should be no need to justify the application of spectral analysis to an arbitrary signal. Indeed, when a "messy" unknown signal is presented, an engineer's natural reaction is to feed it into a tunable filter and then measure how much of it can get through. An ordinary radio receiver or a narrow-band amplifier, with a wattmeter on the output, will serve adequately for qualitative analysis, although accurate work requires a wave analyzer—essentially the same device specially designed for stability and quantitative accuracy.

Such analog measurements of power spectra will not concern us here because, although these measurements may have their difficulties, the behavior of the measuring instrument will usually warn the user of any problems. However, signals at very low or very high frequencies are often quite inconvenient to treat with analog circuits, and many signals naturally arrive in digital form. Among high-frequency signals, those returned to interplanetary radars are usually manipulated digitally,[1] presumably because this method offers the greatest experimental return for the least investment and effort. At very low signal frequencies, many geophysical observations, for example, are necessarily recorded graphically or as a series of numerical readings, perhaps taken hourly or even as seldom as daily or weekly. Economics data are almost unavoidably digital, and the majority of today's satellites report their findings as digital readings.

In all such cases, if the distribution of signal frequencies is desired or if correlations with other phenomena are to be explored, a natural first step is to determine the power spectrum of the signal; doing so by digital computation is usually the simplest and often the cheapest way. Even when a signal has been recorded in analog form, such as a tracing on a strip chart, digital methods frequently offer the most direct route to spectrum analysis. It is this type of numerical computation that will be considered in the following.

Preliminary concepts

It is usual to think of the signal $x(t)$ as continuous, even though its sampled values $x(t_n)$, at a set of times $t_1, t_2, \ldots$, may be the only data available. For simplicity, assume that the sampling times are uniformly spaced with separation Δ_t. The basic idea of spectral analysis is that a signal $x(t)$, arbitrary to within certain limits, can be represented as a continuous superposition of sine waves. This is expressed mathematically by the familiar Fourier integral formulas. We shall use the notation

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(2\pi i f t)\, dt \qquad (1)$$

and, conversely,

$$x(t) = \int_{-\infty}^{\infty} X(f) \exp(-2\pi i f t)\, df \qquad (2)$$

where f is the frequency of the Fourier component $X(f)$. The integral in (1) is actually finite in any practical case since only a finite segment of the signal will be available. This situation is commonly represented by setting $x(t) = 0$ for $t < 0$ and for $t > T$ when the signal is available over the range $(0, T)$.

The simple spectral components $X(f)$, however, are not particularly helpful in analyzing a previously unknown

signal. For example, any pseudorandom components, such as phase-modulated waves or admixtures of noise, will produce varying $X(f)$, which may simply average to zero over a T of long enough duration.

To avoid such problems, it is expedient to examine, not the "voltage" spectrum of Eq. (1), but the "power" spectrum

$$P(f) = (1/T)\,|X(f)|^2 \qquad (3)$$

where the factor $1/T$ is inserted to make $P(f)$ relatively independent of the duration T of the data. It is customary to speak of $P(f)$ as the power spectrum of $x(t)$, whether or not $x(t)$ is actually a physical voltage or current. For some years, $P(f)$ was called the "periodogram" in statistical literature, but the word fell into some disrepute as the problems we will discuss began to be recognized.

When only sampled values $x(t_n)$ are available—or whenever any numerical computation is undertaken—the integral in (1) must be replaced by a finite sum. If the trapezoidal rule for numerical integration is used in (1), and if the resulting sum is substituted into (3), we obtain, neglecting end corrections,

$$P(f) = (1/N)\,\left|\sum_{0}^{N} x(t_n)\exp(2\pi i f t_n)\right|^2 \qquad (4)$$

where $t_n = n\Delta_t$ and $N\Delta_t = T$ (and the number of samples is $N + 1$). Although more accurate numerical integration schemes might seem desirable, actually they would merely complicate the discussion without offering any greater generality in the final results.

First lesson—aliasing

Perhaps the first indication that all may not be well in these formulas is that Eq. (4) does not approach zero at high frequencies, even when $x(t)$ is known to contain no high-frequency components. In fact, (4) is a finite trigonometric series and must be periodic in f: If $f' = f + (1/\Delta_t)$ then

$$(2\pi i f' t_n) = (2\pi i f t_n) + (2\pi i)n$$

so that each term in (4) has period $1/\Delta_t$ in f.

What has happened? Clearly Eq. (4) would be an exact transcription of (1) and (3) if the signal $x(t)$ actually consisted of a set of successive narrow pulses, the pulse at t_n having an (integrated) amplitude $x(t_n)\Delta_t$. Thus, the high-frequency components of $P(f)$ might be said to arise from modulating the signal onto a train of pulses, as a result of the sampling.

This point of view even offers a quantitative analysis of the situation. The pulse train can be represented by a Fourier series, and if the pulses are regarded as Dirac delta functions, this representation takes the simple (but improper) form

$$\sum_{-\infty}^{\infty} \Delta_t \delta(t - n\Delta_t) = 1 + 2\sum_{1}^{\infty} \cos(2\pi k t/\Delta_t) \qquad (5)$$

(To see this, merely apply the usual formulas for the Fourier series coefficients to the unit impulse $\delta(t)$ with period Δ_t.) When this equation is multiplied by $x(t)$, we obtain a representation of the sampled signal as a series of amplitude-modulated waves (with suppressed carriers):

$$\sum_{-\infty}^{\infty} x(t_n)\Delta_t \delta(t - t_n) = x(t) + 2\sum_{1}^{\infty} x(t)\cos(2\pi k t/\Delta_t) \qquad (6)$$

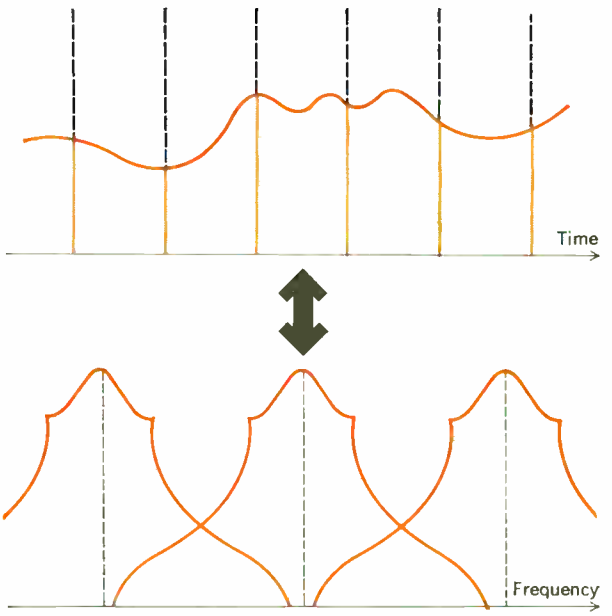


Fig. 1. Aliasing due to inadequate sampling. The signal modulates a train of narrow sampling pulses, producing sidebands at all harmonics of the pulse repetition frequency. If this is too low, sidebands representing the original signal will interfere with each other.

The first term on the right-hand side of this equation, the "dc" term, will clearly yield the correct spectrum of $x(t)$, but the other terms, being sinusoidal carriers modulated in amplitude by $x(t)$, will produce "sideband copies" of the desired spectrum, centered at frequencies $1/\Delta_t$, $2/\Delta_t$, $3/\Delta_t$, and so on, all with the same strength as the desired dc term.

Thus, no harm will be done if the sampling frequency $1/\Delta_t$ is sufficiently high compared with the maximum frequency that is present in $x(t)$, and Eq. (4) could be expected to yield correct results over the frequency range of interest. The fact that $P(f)$ does not vanish at high frequencies merely indicates the presence of the sideband copies of the desired $P(f)$ centered on the sampling frequency and on each of its harmonics.

By the same token, however, if trouble is to be avoided the sampling frequency must be sufficiently high or the sidebands will overlap; in particular, the lower sideband from the lowest carrier wave will add into the true spectrum, which might be called the upper sideband on the dc carrier. Such a superposition will completely destroy the meaning of $P(f)$ in the region of overlap. This argument establishes the rule: *The sampling frequency $1/\Delta_t$ must exceed twice the highest frequency present in $x(t)$.*

One half the sampling frequency is often called the folding or Nyquist frequency, $f_N = 1/2\Delta_t$. It must exceed the highest frequency in $x(t)$. If only the lowest frequencies in $x(t)$ are of interest, some overlap might be allowed, but very careful analysis of the particular case would be necessary to justify this procedure. Thus, although the rule is perhaps not completely general, its violation invites trouble.

If any intermingling of the sidebands takes place at all, it does so immediately upon sampling. Once a signal

is sampled, this intermingling or "aliasing" is already present, and nothing short of returning to resample $x(t)$ more frequently can possibly remove it. The only safe way to reduce the requirement is to filter high frequencies out of $x(t)$ before it is sampled. If this is not possible, although the necessary number of samples $x(t_n)$ cannot be reduced, the subsequent computational burden can sometimes be ameliorated, as will be pointed out.

Statistical difficulties

Nevertheless, Eq. (4) is almost useless for practical computation, except in cases so simple as to be of little interest. Specifically, the power spectrum is of most value to an investigator when the signal is strongly contaminated by noise. However, experience has repeatedly shown that, with a noisy signal, (4) gives a very erratic spectrum $P(f)$, which fails to converge no matter how large T is made or how small Δ_t is chosen.

An acceptable analysis of this behavior would take us too far afield, but the essentials can be traced through in a simple case. Consider white thermal Johnson noise, which has a uniform, flat power spectrum extending down to zero frequency. (It has no dc delta-function spike, but it does have finite spectral density of power at zero frequency.)

In this case, the samples $x(t_n)$ will be independent of one another, and each will have a Gaussian distribution with zero mean. From (4) with $f = 0$ and $P(0)$ abbreviated to P_0,

$$P_0 = S^2/N$$

where

$$S = \sum x(t_n)$$

By the central limit theorem, S also has a Gaussian distribution (for large N) and zero mean. However, P_0 is proportional to S^2, not S, and a simple calculation verifies that if P_{0t} is the average value of P_0, we have

$$\frac{<(P_0 - P_{0t})^2>_{\text{avg}}}{P_{0t}^2} = 2$$

independent of N and $(S^2)_{\text{avg}}$. That is, the fractional error in $P(0)$, for white noise as computed by (4), is always somewhat greater than unity!

A complete calculation of the statistical behavior of (4) can be carried through for any Gaussian signal; i.e., any signal with the following property: If x_1 and x_2 are values of the signal at any two fixed times t_1 and t_2, then the joint probability distribution for x_1 and x_2 is proportional to $\exp[-(x_1^2 + x_2^2 - 2\rho x_1 x_2)/2(1 - \rho^2)]$, with $\rho^2 < 1$ and ρ depending only on $t_2 - t_1$.

The results of such an analysis[2-4] show that the dc value of P is the worst one but that the foregoing situation is perfectly general. If we define

$$\epsilon \equiv \frac{\text{rms } \Delta P}{\text{avg } P} \tag{7}$$

then when (4) is used, $P(f)$ at any f (not too near zero) has

$$\epsilon_{(4)} \simeq 1 \tag{8}$$

for any Gaussian signal. Presumably, the results are equally troublesome for almost any noisy signal.

Nevertheless, it is clear physically that the situation cannot be quite this bad if we are willing to use some reasonable modification of (4). One possible modification might be a formula that duplicates the physical action of a wave analyzer. A conceptually simpler scheme[2,3] would be merely to repeat the entire calculation separately for M independent pieces of the signal and then take the average of their individual $P(f)$. In this way, one might expect to obtain $\epsilon^2 \simeq 1/M$— and a corresponding calculation of errors in $P(0)$ for white noise supports this expectation.

Even if we assume that this would work in general, other questions remain. For instance, would this procedure have any other effect, such as tending to broaden the peaks in $P(f)$, as a wave analyzer must always do to some extent? Also, if only one sample of a signal is available, it can be broken into M independent segments, but what effect might this have on $P(f)$?

To analyze these questions, we must review another theorem of Fourier analysis, one of extraordinary utility in understanding subtleties of power spectra.

A tool—convolution

The product of two Fourier transforms $X(f)$ and $H(f)$ can be written directly from Eq. (1) as a double integral:

$$X(f)H(f) = \int_{-\infty}^{\infty}\int x(t')h(t'') \exp[2\pi i f(t' + t'')]\, dt'\, dt''$$

Replace the integration variable t'' by $t' + t'' = t$, say, and the result is

$$X(f)H(f) = \int_{-\infty}^{\infty} dt \exp(2\pi i f t) \int_{-\infty}^{\infty} x(t')h(t - t')\, dt' \tag{9}$$

This is the convolution theorem, which states that the product of two transforms $X(f)H(f)$ is itself the Fourier transform of the third function of time, called the *convolution* of $x(t)$ and $h(t)$ and defined by

$$x(t) \text{¢} h(t) = \int_{-\infty}^{\infty} x(t')h(t - t')\, dt' \tag{10}$$

This function is symmetric in x and h, as may be seen by changing the variable of integration to $t'' = t - t'$.

The convolution integral was once known as Duhamel's integral in the parlance of electrical engineering; as a matter of fact, Eq. (9) essentially states that when a filter $H(f)$ is driven by a signal with spectrum $X(f)$, the output of the filter will be given by the convolution integral in (10), where $h(t)$ is the impulse response of the filter, the response when x is a delta function. For physical filters, of course, $h(t - t')$ must be zero for all $t' > t$, but the basic theorem (9) applies even to nonphysical filters, which might not satisfy this causality requirement. In any case, the action of the convolution (10) can be envisaged as a superposition of "copies" of $h(t)$, the copy centered at t' having weight or amplitude equal to $x(t')$. By symmetry, the convolution can also be envisaged as a sum of copies of x having weights $h(t')$. In the same way, from (2) we obtain the theorem

$$x(t)h(t) = \int_{-\infty}^{\infty} df \exp(-2\pi i f t) \int_{-\infty}^{\infty} X(f')H(f - f')\, df' \tag{11}$$

This relation states that the simple product $x(t)h(t)$ in the time domain is the Fourier inverse of the (frequency) convolution

$$X(f) \text{¢} H(f) = \int_{-\infty}^{\infty} X(f')H(f - f')\, df \tag{12}$$

which is also symmetric in X and H.

A simple example of how this formula can be applied can be seen in the process of aliasing. Equation (5) for the time domain essentially states that a train of "carrier" pulses spaced at intervals Δ_t has a Fourier transform consisting of delta functions on the frequency axis, spaced at intervals $1/\Delta_t$, with each frequency spike representing one of the sinusoidal carrier waves. Call the train of carrier pulses $h(t)$ and their Fourier transform $H(f)$, the train of frequency spikes. Equation (11) then states that the sampled signal $x(t)h(t)$ has a Fourier transform given by the frequency convolution (12), a superposition of copies of $X(f)$ with weights $H(f)$. Since $H(f)$ is a train of delta functions, this superposition is merely a set of simple copies of $X(f)$, spaced at intervals $1/\Delta_t$. This is the same result that we obtained earlier, and it leads to the sampling rule $f_{max} < 1/2\Delta_t = f_N$ when we note that $X(f)$ is essentially symmetric about $f = 0$; that is, from (1), $X(-f)$ is just the complex conjugate of $X(f)$, since the signal $x(t)$ is a real function.

This last observation is worth carrying a little further: We have $X^*(f) = X(-f)$ when $x(t)$ is real, but it is easy to see from (1) that $X(-f)$ is the Fourier transform of $x(-t)$, and thus

$$X^*(f) = \mathfrak{F}[x(-t)] \tag{13}$$

where $\mathfrak{F}[\ldots]$ denotes "Fourier transform of ...."

An alternate formulation

The theorems in (9) and (13) enable us to reformulate $P(f)$ for the ideal, unsampled signal as it was originally defined in (3), which can be rewritten in the form

$$P(f) = (1/T)\, X(f)\, X^*(f)$$

From (9) and (13),

$$P(f) = (1/T)\, \mathfrak{F}\left[\int_{-\infty}^{\infty} x(-t')x(t - t')\, dt'\right]$$

Thus, if we define

$$R(\tau) = (1/T) \int_{-\infty}^{\infty} x(t'')x(\tau + t'')\, dt'' \tag{14}$$

then we see (setting $t'' = -t'$) that $P(f)$ is just the ordinary Fourier transform of $R(\tau)$:

$$P(f) = \int_{-\infty}^{\infty} R(\tau) \exp(2\pi i f\tau)\, d\tau \tag{15}$$

The function $R(\tau)$ is called the "autocorrelation function" of the continuous signal $x(t)$. That it is an even function, $R(-\tau) = R(\tau)$, is easily seen by letting $t'' = t' - \tau$ in (14); it can be regarded as the convolution (times $1/T$) of $x(t)$ with the mirror image $x(-t)$, or as the average (when $T \to \infty$) of all so-called lagged products $x(t')x(\tau + t')$ having a given lag τ. Equation (15) is the famous result that the power spectrum of a signal is the Fourier transform of its autocorrelation function. (The foregoing "proofs" are purely formal; rigorous proofs, originally due to Wiener, are far more complicated.[5])

It might seem that we should next employ numerical integration in (14) and (15), but the result would then be exactly equivalent to the discredited Eq. (4). This is not surprising, for we have been simply recasting the formulas for continuous, unsampled signals without facing the issue of statistical behavior. The new formulas will be useful for this purpose, however.

Analysis of segmental averaging

In the discussion of the discouraging statistical result (8), it was suggested that a signal might be broken into M segments of duration T/M, each segment being treated as a different signal and their power spectra being averaged. We can now analyze this suggestion.

Since the Fourier transform (15) is a linear relation, the average of the individual power spectra of the segments will be the transform of the average of the $R(\tau)$ functions for each segment. For the kth segment of $x(t)$, the autocorrelation function (14) becomes (assuming $\tau \gtrless 0$ for the moment)

$$R_k(\tau) = (M/T) \int_{(k-1)T/M}^{k(T/M)-\tau} x(t')x(t' + \tau)\, dt' \tag{16}$$

where the integration limits are determined from the requirement that the kth segment of the signal must vanish for times less than $(k-1)\,T/M$ or greater than kT/M.

For the averaged power spectrum, we require the average of the R_k, namely,

$$R_A(\tau) = (1/M)\, \Sigma_k R_k(\tau)$$

If the upper limit of the integration in (16) were kT/M, this sum would become the integral of $x(t')x(t' + \tau)$ over the full available range $(0, T - \tau)$. The actual ranges of integration for all the R_k, however, add up to a total length of only $T - M\tau$. On the average, therefore,

$$R_A(\tau) = \frac{T - M|\tau|}{T - |\tau|} \left(\frac{1}{T}\right) \int_0^{T-|\tau|} x(t')x(t' + |\tau|)\, dt' \tag{17}$$

where the actual limits of integration have been explicitly displayed—and τ has been replaced by $|\tau|$ to make the result correct for $\tau < 0$. In view of (16), note that $R_A(\tau)$ must be understood to vanish when $|\tau| > T/M$.

Upon comparing (17) with (14) we find that the proposed segmental averaging is statistically equivalent to multiplying the original autocorrelation function $R(\tau)$ by a "window" function, given by

$$\begin{aligned} g(\tau) &= (T - M|\tau|)/(T - |\tau|) && \text{when } |\tau| < T/M \\ &= 0 && \text{otherwise} \end{aligned} \tag{18}$$

If M is large, the denominator will be nearly constant when $|\tau| < T/M$, and $g(\tau)$ can be approximated by a simple triangular window

$$\begin{aligned} g_M(\tau) &= 1 - (|\tau|M/T) && \text{when } |\tau| < T/M \\ &= 0 && \text{otherwise} \end{aligned} \tag{19}$$

In any case, the convolution theorem shows what the corresponding effect on $P(f)$ must be. Specifically, the averaged spectrum $P_A(f)$ is the Fourier transform of

$$R_A(\tau) = g(\tau)R(\tau) \tag{20}$$

and the convolution theorem then shows that

$$P_A(f) = P(f) \,\text{¢}\, G(f) = \int_{-\infty}^{\infty} P(f')G(f - f')\, df' \tag{21}$$

where $G(f)$ is the Fourier transform of $g(\tau)$. In particular, the approximation $g_M(\tau)$ has the simple transform

$$G_M(f) = (T/M)[(M/\pi fT)\sin(\pi fT/M)]^2 \tag{22}$$

If no segmenting is done ($M = 1$), then Eq. (19) is not

a good approximation, but (18) reduces to

$$g_0(\tau) = 1 \qquad \text{when } |\tau| < T$$
$$= 0 \qquad \text{otherwise} \tag{23}$$

and (22) is replaced by

$$G_0(f) = (1/\pi f)\sin(2\pi fT) \tag{24}$$

Even when $x(t)$ is undefined outside the range $(0, T)$, (20) and (21) are valid for the unsegmented signal. In such a situation (20) obviously is a trivial statement, but (21) states that $P(f)$ satisfies an integral equation with the kernel function (24). In this sense, the final $P(f)$ is still given by (21).

When we examine (21) and (22) we find that segmental averaging introduces some spectral averaging. That is, according to (21), $P(f)$ is "smeared out" by $G_M(f)$ to produce the final average spectrum $P_A(f)$. According to (22), the smearing function—or resolution function G_M—has a central peak at $f' = f$, which vanishes at $|f - f'| = M/T$ (and it has small sidelobes at odd multiples of $M/2T$). In short, the technique for reducing statistical variability has also introduced a smoothing of $P(f)$, a reduction in spectral resolution. In particular, if an attempt were made to decrease ϵ too much, by increasing M without increasing T, the resolution would become so poor that $P(f)$ would be virtually meaningless.

Resolution was not mentioned previously but we can see from (21) and (24) that some smoothing occurs even without segmenting, at least in the sense that values of $P(f)$ for closely neighboring frequencies are not completely independent. Without a precise formal definition of spectral resolution, exact statements cannot be developed; however, for most engineering purposes the simple relation

$$\Delta_f = M/T \tag{25}$$

is an adequate estimate of the frequency range Δ_f over which $P(f)$ has been spread as a result of averaging the power spectra of M segments of a signal with total duration T.

The uncertainty principle

We have been presuming that segmental averaging will reduce the statistical variations in $P(f)$, thus making them more acceptable than those in (8) for the simple formula (4). Specifically, we hypothesized that the mean-square fractional error ϵ^2 would be approximately $1/M$ for an average of M segments. If true, this and (25) lead to an "uncertainty" relation:

$$\epsilon^2 T\Delta_f \simeq 1 \tag{26}$$

This relation proves to be quite general. Not only can it be verified for segmental averaging, but Eqs. (14)–(21) can be statistically analyzed[4] for Gaussian signals and noise, with an arbitrary $g(\tau)$ or $G(f)$. If the spectral resolution Δ_f is "defined" as the approximate width of the resolution function $G(f)$, then the uncertainty relation (26) emerges in full generality.

Unlike uncertainty relations in quantum mechanics, (26) involves three physical variables, and it is an (approximate) equality, not merely a lower bound. In particular, a desired fractional error and spectral resolution completely determine the minimum signal duration. Signals shorter than this cannot yield both the same stability and the same resolution in $P(f)$.

In practice, the uncertainty relation is most useful for estimating the statistical error, which would be very difficult to calculate directly.

$$\epsilon = \frac{\text{rms } \Delta P}{\text{avg } P} \simeq \sqrt{\frac{1}{T\Delta_f}} \tag{27}$$

Of course, these results have been proved only for Gaussian signals, but they are commonly assumed to be approximately valid for any signal generally encountered in science and engineering. Naturally, if a signal is suspected to be statistically pathological, a specific analysis of the errors should be attempted.

Taking stock

Enough results have now been developed to show that segmental averaging is just one of many methods for controlling error by an appropriate sacrifice of spectral resolution, as required by the uncertainty relation (26) or (27). One way of doing this would be to use the original (4) followed by a smoothing operation on $P(f)$, as in (21), with some appropriate $G(f)$; but the modern custom is to employ the alternate route—through the autocorrelation function—for the following reasons.

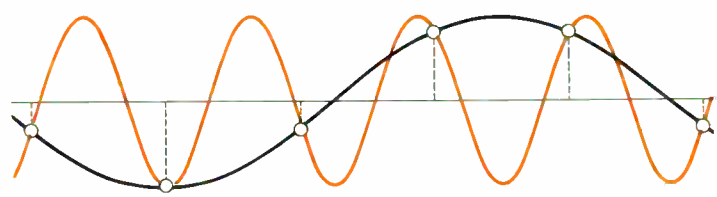
In terms of computational labor there is little to recommend one route over the other; however, the autocorrelation function (which does converge as the samples accumulate) supplies a bonus—an additional "statistical summary" of the signal. This fact alone makes it attractive. Moreover, in order to smooth a "raw" $P(f)$ computed with (4), it would be necessary to sample this $P(f)$ in frequency space, in the process of convolving it numerically with the chosen $G(f)$. This additional sampling (at intervals $D_f = 1/T$ or less) would be equivalent to convolving the entire signal with a train of time pulses widely spaced at intervals of T, or more. In other words, it would introduce an assumption that the signal can be treated as periodic with period T; or, more generally, $1/D_f$. This may be acceptable in some cases but it is entirely unnecessary when the autocorrelation function is used.

For numerical computation, of course, the sampled signal (6) must be employed. The autocorrelation function is then already sampled because (14) reduces to

$$R(\tau_n) = (1/N)\sum_{k=0}^{N-|n|} x(t_k)x(t_k + |\tau_n|) \tag{28}$$

where $\tau_n = n\Delta_t$ for $n = 0, \pm 1, \pm 2, \ldots$, and $R(\tau) = 0$ when $\tau \neq$ some τ_n.

Fig. 2. Another view of aliasing. If the samples are spaced at intervals of Δ_t, then the sine waves whose frequencies differ by multiples of $1/\Delta_t$ cannot be distinguished.



Incidentally, if a thousand or more signal points $x(t_k)$ must be processed, new techniques[6,7] for computing "fast Fourier transforms" will usually be more efficient than evaluating (28) directly. Two transforms are required: first, Eq. (4) is evaluated for $f = mD_f$ with $m = 0, 1, \ldots, N$ and with $D_f = (2T + \Delta t)^{-1}$; then, an inverse transform on the raw $P(\pm f)$ yields $R(\tau_n)$. The specified value of D_f avoids aliasing of $R(\tau)$, and when the outlying sideband copies of $R(\tau)$ are ignored, the method is exactly equivalent to (28) except possibly for round-off errors, which might favor either method.

Once $R(\tau)$ has been obtained, spectral smoothing is introduced by multiplying $R(\tau)$ with some suitable $g(\tau)$ so chosen that the spectral width and sidelobes of the associated $G(f)$ are satisfactory for the problem being considered and will give an acceptable statistical error ϵ in Eq. (27). Since the product $g(\tau)R(\tau)$ is nonzero only at τ_n, the Fourier transform for the final spectrum (15) also reduces to a sum

$$P_g(f) = \sum_{-N}^{N} g(\tau_n)R(\tau_n)\exp(2\pi i f\tau_n)$$

which, for even functions, further reduces to

$$P_g(f) = g(0)R(0) + 2\sum_{1}^{N} g(\tau_n)R(\tau_n)\cos(2\pi f\tau_n) \quad (29)$$

The subscript g merely indicates that this final estimate of the power spectrum depends somewhat on the smoothing function $g(\tau)$. Specifically, P_g is related as follows to the aliased but otherwise true spectrum $P_{al}(f)$ of the portion of the signal in $(0, T)$:

$$P_g(f) = \int_{-\infty}^{\infty} P_{al}(f')G(f - f')\,df' \quad (30)$$

where $G(f)$ is the Fourier transform of $g(\tau)$. In particular, the frequency resolution Δ_f of P_g is approximately the width of the main lobe of $G(f)$.

We must not forget that omitting $g(\tau)$ in (29) is not equivalent to using $G(f) \simeq \delta(f)$; rather, it is equivalent to using g_0 and G_0 as defined in (23) and (24). Only for very large T would G_0 approximate a delta function in any practical sense whereas ϵ would be unity for any T. Some better choice of g is always advisable, and usually is vital.

Equations (28) and (29), together with (27) and (30) for their interpretation, are all that is really needed to compute a power spectrum. However, these formulas can be used and interpreted in various ways.

Choosing a smoothing function

The function $g(\tau)$ in Eq. (29) is arbitrary to some extent, but, in view of (30), it should be chosen to have an associated resolution function $G(f)$ with a strong central peak and preferably with small sidelobes. The sidelobes cannot be entirely eliminated because $g(\tau)$ vanishes (effectively, in these formulas) when $|\tau| > T$, if not before, and a fundamental theorem of Fourier analysis implies that $G(f)$ then must have nonzero values at arbitrarily high frequencies.

The triangular windows g_M defined in (19) yield the G_M of (22); these have strong central peaks with fairly small, positive sidelobes, the largest of which is some 4.5 percent of the main peak. These, perhaps the most commonly used windows, are often called "Bartlett windows." They are named for M.S. Bartlett, who first suggested[2,3] segmental averaging and worked out its connection with $g_M(\tau)$ and the spectral resolution function $G_M(f)$.

Another commonly used window, suggested by R.W. Hamming, is

$$\begin{aligned} g_h(\tau) &= 0.54 + 0.46\cos(\pi\tau/T_M) && \text{for } |\tau| < T_M \\ &= 0 && \text{otherwise} \end{aligned} \quad (31)$$

where the "maximum lag" T_M is analogous to T/M in g_M. The $G_h(f)$ associated with g_h has $\Delta_f \simeq 1/T_M$ with sidelobes of alternating sign, but the largest lobes (the third and fourth) have absolute values a little less than one percent of the main peak.

A window that generates only very small sidelobes, none of them negative, can be constructed by taking the inverse Fourier transform of

$$G_p(f) = \frac{3T_M}{4}\left(\frac{\sin(\tfrac{1}{2}\pi f T_M)}{\tfrac{1}{2}\pi f T_M}\right)^4$$

which has $\Delta_f \simeq 1.5/T_M$ and leads to

$$\begin{aligned} g_p(\tau) &= 1 - 6u^2 + 6u^3 && \text{when } u \leqslant \tfrac{1}{2} \\ &= 2(1-u)^3 && \text{when } \tfrac{1}{2} \leqslant u \leqslant 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

where

$$u = |\tau|/T_M$$

This window was originally suggested by Parzen.[8] The largest sidelobe of G_p (the first) is about 0.2 percent of the main peak, and the later ones fall off very rapidly as $1/(f - f')^4$.

If the Hamming window (31) is convolved with itself, the corresponding G_h will be squared, and all the sidelobes will then be positive and less than 10^{-4} of the main peak. The resulting $g(\tau)$, rescaled to $g(0) = 1$, is

$$\begin{aligned} g(\tau) &= (0.74)(1 - |\tau|/T_M)[1 + 0.35\cos(2\pi|\tau|/T_M] \\ &\quad + 0.157\sin(2\pi|\tau|/T_M) && \text{when } |\tau| \leqslant T_M \\ &= 0 && \text{otherwise} \end{aligned}$$

The Δ_f associated with this g is also about $1.5/T_M$. Of an infinity of possible $g(\tau)$, these seem to be the most useful.

Perhaps more important than the choice of a window is the assurance that the signal duration T is great enough to support a desired resolution and accuracy. No amount of ingenuity in selecting a window can evade the limitations expressed in the uncertainty relation (27).

Some minor points should be mentioned for completeness. First, $G(f)$ is convolved with the entire power spectrum of the sampled signal, including all the sideband copies near the sampling harmonics. Thus, in interpreting the convolution (30) for sampled signals, it should be noted that the distant sidelobes of $G(f)$ can pick up contributions from the sideband copies of the true $P(f)$ and can in principle affect the computed spectrum in this way. With any reasonable choice of window, however, this effect will be entirely negligible.

Second, when a copy of $G(f)$ is placed near zero frequency in (30), the sidelobes that fall on the negative f axis are not "chopped off" in any sense; rather, they are multiplied by $P(-f)$, and thus they contribute to the convolution $P_g(f)$, even down to $f = 0$.

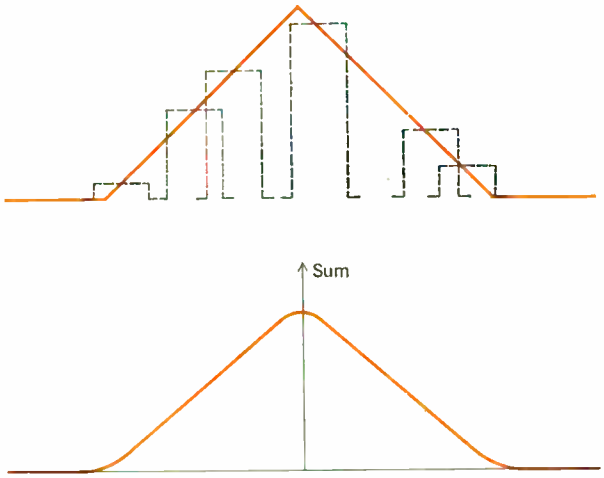


Fig. 3. Example of convolution. Summing the contributions of all possible copies of the rectangle, with heights proportional to the "roof" function, produces the bottom curve. This curve is called the convolution of the rectangle and roof function.

Finally, one feature of Eq. (28) for $R(\tau)$ is sometimes confusing, and is largely resolved by considering the choice of $g(\tau)$. The summation limits in (28) are determined by the convention that $x(t)$ vanishes outside the range $(0, T)$ where the samples are taken. When the signal does not vanish outside this range but instead continues for some time more or less unchanged, it might seem advisable to divide the sum in (28) by a factor $1 - (|\tau|/T)$ to compensate for missing portions of the signal. (In all of our formulas, $P(f)$ has been the power spectrum of the truncated signal.)

However, this unbiased estimate of $R(\tau)$ for the persisting, untruncated signal must be smoothed, like any other $R(\tau)$, with some $g_u(\tau)$. If we choose $g_u(\tau) = g_1(\tau)\, g(\tau)$, where g_1 is the widest possible triangular window (19), then the g_1 will simply remove the supposed compensating factor, and we arrive back at the simple biased form (28), smoothed with $g(\tau)$. In short, (28) is already somewhat smoothed for long-duration signals, and there is really little distinction between the biased and unbiased estimates of $R(\tau)$ in the sense that any reasonably small statistical error ϵ will always require much more drastic smoothing than $g_1(\tau)$ can accomplish. Of the two estimates, the simple biased one (28) has the smaller mean-square error.

More detailed discussions of smoothing windows and their effects on the spectrum can be found in the literature.[4,5,8,9]

The dc spike, prewhitening

In the foregoing discussions we uncovered most of the pitfalls in computing power spectra, but a few snares remain that might delay the unwary.

If $P(f)$ has a very large peak that stands far above a more uniform background, this peak will emerge as essentially a copy of $G(f)$ in the computed $P_g(f)$, and the wings of $G(f)$ can easily obscure large parts of the background. Although this situation might be recognized fairly quickly for what it is, it is largely a waste of computing effort and should be avoided if possible. Little general advice about strong sinusoidal components can be offered, aside from urging that they be subtracted from the signal (or filtered out) as soon as their presence is discovered. After the spectrum has been computed, they can be added back in. (Although subtraction leaves the phase relative to the background undetermined, the corresponding uncertainty as to how to add the spike back in will be small if the spike is large enough.)

One large peak, the dc peak, is almost always present and should be removed before Eqs. (28) and (29) are applied lest the low frequencies be distorted. Whereas the true dc level can easily be set to zero by subtracting the mean of the signal, there is an equally important "near-dc" contribution, the treatment of which is less straightforward: If the signal has a linear trend (in the sense that $x(t_n) \simeq a + bt_n +$ smaller terms) this slow drift could be regarded as a small section of a sine wave with a very low frequency f_0. Thus a linear drift can also produce a large "dc" peak with its obscuring wings and minor lobes. In this case, the peak has the form $G(f-f_0)+G(f+f_0)$, where f_0 is so low that the negative-frequency part $G(f+f_0)$ will spill over onto low postive frequencies.

One way to suppress these peaks would be to remove the true dc level and subtract a low-frequency sine wave from the remainder of the signal. However, this scheme has not been tested adequately in practice and so we prefer to stress one that is in more common use.[4] It can be shown that if

$$a_- = \text{average of } x \text{ in } (0, T/3)$$

$$a_+ = \text{average of } x \text{ in } (2T/3, T)$$

$$a = \text{average of } x \text{ in } (0, T)$$

then the modified autocorrelation function

$$R_{\text{cor}}(\tau) = R(\tau) - a^2 - \frac{3}{16}(a_+ - a_-)^2\left(1 - \frac{2|\tau|}{T} - \frac{2\tau^2 + \Delta_t^2}{T^2}\right) \quad (32)$$

is largely corrected, both for true dc level and for any linear drift. Although this correction is not perfect, the simplicity of the method and the lack of clearly better ones have encouraged its widespread use.

If we carry the idea of removing spectral peaks to its logical conclusion, we arrive at the concept of *prewhitening*—which would be followed, of course, by recoloring at the end of the computation. That is, if the power spectrum were made perfectly flat during the computation, any interference between different frequency components would presumably be minimized and a more accurate picture of the true spectrum should emerge after recoloring. Whether such refinements would be worth all the exploratory calculations usually needed to determine the necessary filtering functions, only the economics and importance of a problem can determine. If the signal is available in analog form, the filtering to prewhiten its spectrum might be done before it is sampled. Otherwise, the filtering can be performed on the samples themselves, as will be described in the following.

Abbreviating the calculation

We have found that the highest frequency in the signal determines the minimum sampling rate for avoiding

aliasing whereas the desired accuracy and resolution determine the duration of the signal that must be sampled. These factors can require many hundreds or thousands of samples in a practical situation, resulting in a computing requirement that can be quite burdensome.

Frequently, however, fine-scaled resolution in the spectrum is actually needed only at the lower frequencies. In such cases, it may be entirely feasible to obtain such information with fewer samples (over the same duration) provided that the higher frequencies are first filtered out of the signal to avoid aliasing at the lower sampling rate. Then a shorter portion of the original signal, sampled at the full rate, can yield an analysis of the higher frequencies but at coarser resolution.

If the high frequencies can be filtered out before part of the signal is recorded, so much the better. However, where this is not practicable, the sampled signal can be filtered by applying the convolution theorem (Duhamel's integral) in Eqs. (9) and (10). This type of filtering can be analyzed in various ways but the simplest approach is to observe that, with a sampled signal, only sampled values of the filter response $h(t_n)$ are pertinent. Thus we can regard the filter as having a sampled response function

$$h(t) = \sum a_n \, \delta(t - t_n) \tag{33}$$

where the $a_n = h(t_n)$ are sufficient to characterize the filter completely for the sampled signal $x(t_n)$. Substitution into (10) then shows that the filtered signal is given by a convolution that reduces to

$$x_f(t_n) = \sum a_{n-k} x(t_k) \tag{34}$$

This smoothing operation must be performed using all the samples of the unfiltered signal on the right-hand side of (34) although t_n on the left side need only assume values appropriate to the coarser sampling.

The appropriate coarseness in sampling will be determined by the needs of the problem and by the high-frequency filtering afforded by the filter. To examine the latter, take the Fourier transform of (33) to obtain the frequency response of the filter. Thus,

$$H(f) = \sum a_n \exp(2\pi i f t_n) \tag{35}$$

This is the function that is multiplied onto $X(f)$ by the operation (34); $P(f)$, of course, is multiplied by $|H(f)|^2$.

The details of adjusting the coefficients a_n and the relative sampling rates for x and x_f are too dependent on individual problems to permit much general discussion. The principles, however, are simple: (1) Equations (34) and (35) determine the effect of a given filter. (2) The aliasing theorem then determines the minimum sampling rate for the filtered signal. (3) The desired resolution Δ_f and error ϵ determine the total signal duration T from the uncertainty relation (26).

Incidentally, we now see why complicated schemes for numerical integration offer no increase in generality. Using such a scheme is completely equivalent to using the trapezoidal rule after applying a filter to the signal, as in Eq. (34)—except that no samples would be deleted afterwards. It seems best to continue to regard this possibility as a separate filtering operation such as prewhitening, which should be tailored to the needs of a specific problem, if it is done at all.

Conclusion

One curse of a survey article is that it has no definite stopping point. Perhaps a good way to conclude this one is to cite a surprising but simple theorem, originally due to Van Vleck. In effect, this theorem states that the power spectrum of a signal can be determined by recording only the algebraic signs of successive samples.

Specifically, if $x(t)$ is a Gaussian signal, and if its fully clipped form is defined as

$$\begin{aligned} x_{cl}(t) &= 1 \qquad \text{if } x(t) \geqslant 0 \\ &= -1 \qquad \text{if } x(t) \leqslant 0 \end{aligned} \tag{36}$$

(the value for $x(t) = 0$ is immaterial), then the autocorrelation function of the clipped signal

$$R_{cl}(\tau) = (1/T) \int_0^{T-|\tau|} x_{cl}(t') x_{cl}(t' + |\tau|) \, dt'$$

completely determines the autocorrelation function of the original signal (on the average), through the relation[10]

$$R(\tau) = \sin[\tfrac{1}{2}\pi R_{cl}(\tau)] \tag{37}$$

The proof, surprisingly enough, is a mere matter of evaluating the average of $x_{cl}(0)x_{cl}(\tau)$ for a Gaussian signal.

This simple result is entirely practical for computation and can save significant amounts of computer time. The only other modification required in the theory is a moderate increase of the constant in the uncertainty relation, indicating a slight loss of stability (or resolution) for a clipped signal of the same duration as the unclipped signal. For clipped signals[1]

$$(\epsilon^2 T \Delta_f)_{cl} \simeq 2.5$$

or

$$\epsilon_{cl} \simeq \sqrt{\frac{2.5}{T\Delta_f}} \tag{38}$$

Despite this compensating change, the total computing burden can often[1] be reduced by a factor of from five to seven if the longer signal duration can be tolerated.

REFERENCES

1. Muhleman, D. O., Goldstein, R., and Carpenter, R., "A review of radar astronomy," *IEEE Spectrum*, vol. 2, pp. 44–55, Oct. 1965; pp. 78–89, Nov. 1965.
2. Bartlett, M. S., "Smoothing periodograms from time-series with continuous spectra," *Nature*, vol. 161, pp. 686–687, 1948.
3. Bartlett, M. S., "Periodogram analysis and continuous spectra," *Biometrika*, vol. 37, pp. 1–16, 1950.
4. Blackman, R. B., and Tukey, J. W., *The Measurement of Power Spectra from the Point of View of Communications Engineering*. New York: Dover, 1958.
5. Lee, Y. W., *Statistical Theory of Communication*. New York: Wiley, 1960, pp. 93–96.
6. Cooley, J. S., and Tukey, J. W., "An algorithm for the machine calculation of complex Fourier series," *Math. Computation*, vol. 19, no. 90, pp. 297–301, 1965.
7. Gentleman, W. M., and Sande, G., "Fast Fourier transforms—for fun and profit," *Proc. 1966 Fall Joint Computer Conf.*
8. Parzen, E., "Mathematical considerations in the estimation of spectra," *Technometrics*, vol. 3, pp. 167–190, May 1961.
9. Jenkins, G. M., "General considerations in the analysis of spectra," *Technometrics*, vol. 3, pp. 133–166, May 1961.
10. Van Vleck, J. H., and Middleton, D., "The spectrum of clipped noise," *Proc. IEEE*, vol. 54, pp. 2–19, Jan. 1966. (This previously unpublished report was originally prepared in 1943.)
11. Goodman, N. R., "Some comments on spectral analysis of time series," *Technometrics*, vol. 3, pp. 221–228, May 1961.

Planning and operation of a large power pool

Reliability of service and economy of operation are the main purpose for utility system interconnections, from the simple emergency tie line to the most sophisticated form, the fully coordinated power pool as exemplified by the PJM Interconnection

R. G. Rincliffe *Philadelphia Electric Company*

Most of the electric power supply systems in the United States are interconnected to provide more reliable service and economy of operation. The majority of these interconnections consist, generally, of tie lines between neighboring utility systems, with contracts between the parties providing for sharing of benefits or for operating procedures. This article discusses how interconnections evolved into fully coordinated power pools, and the benefits resulting from this coordination. Of particular interest is the Pennsylvania–New Jersey–Maryland Interconnection, the largest such pool in the country. Dating back to 1927, it established a successful pattern for the coordination of utility systems.

In an effort to maintain reliability of service together with economy of operation, some 97 percent of all generating capacity in the United States today has been interconnected to some degree. The simplest form of interconnection between two systems is a low-capacity tie line intended only for emergency use. But although such interconnections do exist, and do reduce the investment required to provide reliable service, a more common type of interconnection is that with high-capacity ties. This interconnection additionally provides for economic energy transactions with resultant savings in fuel costs, the sharing of reserves, and support in time of emergencies. Its applications are numerous throughout the United States.

An even more sophisticated type of interconnection provides for the joint planning of capacity additions by both systems. This planning permits the installation of larger generating units without increasing overall reserve requirements and results in lower unit costs of investment and lower operating expenses. Contractual arrangements may provide for the sale of excess capacity from one system to the other, or for jointly owned units.

A combination of interconnected systems is often referred to as a pool. Agreements among the systems may vary widely from pool to pool; for example, the Interconnected Systems Group (ISG) is composed of some 100 electric systems covering a 32-state area (see Fig. 1). Coordination among the systems is achieved through bilateral agreeements, some of which are written and some of which are merely understandings. Because each system does not have contracts with all other systems, participation is voluntary; however, the magnitude of the benefits from the pool is so great that the required cooperation is generally forthcoming. Although an informal pool such as this yields many of the benefits of interconnecting several systems, the maximum benefits result from a fully coordinated power pool that operates under a single formal agreement.

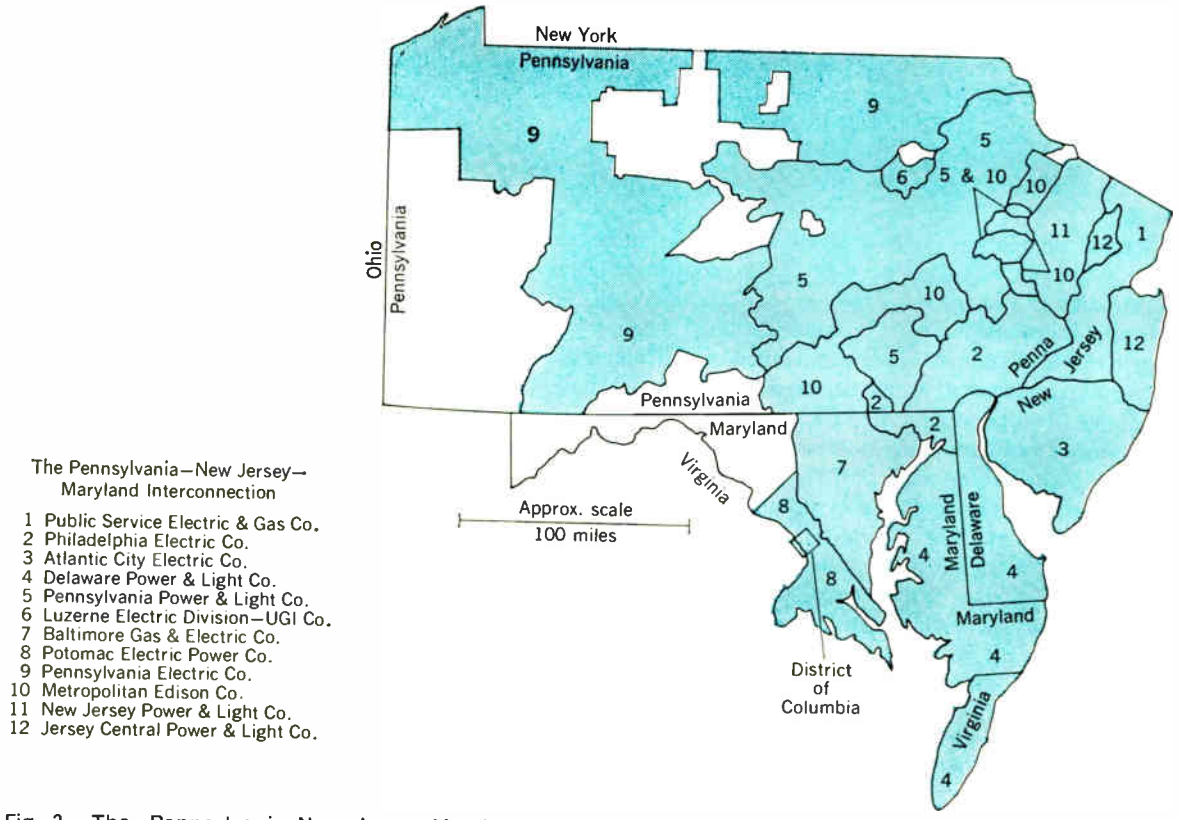


Fig. 2. The Pennsylvania–New Jersey–Maryland Interconnection.

member systems—the Atlantic City Electric Company (3), the Delmarva Power and Light Company (4), and the Luzerne Electric Division of the United Gas Improvement Company (6).

The combined area served by these 12 investor-owned electric power systems of PJM is almost 80 000 square kilometers. To serve the 20 million people living in this area, the combined installed generating capacity at the time of the peak load in 1966 was in excess of 19 000 MW operating in over 60 major generating stations. The peak load was 17 852 MW.

The agreement. The PJM agreement provides for the establishment of a Management Committee to administer all matters pertaining to the interconnection. The Management Committee presently consists of one corporate officer from each of the member systems, and all its decisions or directions must be unanimous. The agreement also provides for the establishment of an Office of the Interconnection.

Whereas policies are formulated and engineering studies are conducted by representatives of the member systems on the committees subsequently described, day-to-day operation is the responsibility of the Office of the Interconnection. In addition to a manager, personnel of this office include seven other graduate engineers, five dispatchers, and a total of seven technical assistants and clerks. Although these people work for all the member systems, for convenience they are on the payroll of Philadelphia Electric Company, in whose territory the office is located. All expenses of the office and its personnel are shared equally among the member systems. The manager reports to the Philadelphia Electric Company representative on the Management Committee on personnel matters and to the entire Management Committee on all other matters.

This office has broad responsibilities. It must schedule sufficient operating capacity over the peak load periods of the day to supply the estimated load and to provide sufficient reserve, as computed on a probability basis, to protect against forced outages and deviations from the load estimate. It is the interconnection dispatcher's duty to make instantaneous checks of the actual operating reserve and to order corrective measures as necessary to assure reliable operation of PJM. These may include (1) transferring units that are on ready standby to operating reserve, (2) ordering systems to maximum generation on all units in operation, (3) ordering voltage reductions to reduce load, and (4) in extreme cases, requesting actual load shedding in accordance with a prearranged schedule. The interconnection dispatcher is also responsible for scheduling capacity and energy transactions with neighboring systems or pools on an hourly basis.

The manager, in addition to his other duties, shares in the responsibility for negotiating agreements with other pools. The engineering personnel monitor the automatic load-frequency control equipment, supervise the accounting for transactions among the systems, develop and maintain computer programs relating to the operation and accounting, serve as members of various committees, and currently are actively engaged in the developmental stages of the new PJM computer for scheduling and monitoring discussed later in this article. The load schedulers are responsible for (1) obtaining the daily peak load estimates and minimum capacity requirements of the systems, (2) combining the loads with the estimated diversity to obtain an estimated peak for the period and expected operating

reserve, and (3) scheduling additional units as necessary to assure economical operation and adequate operating reserve.

An Operating Committee composed of members' representatives is provided to establish, by majority vote, operating and accounting practices. A Maintenance Committee coordinates the scheduling of maintenance for all generating units on PJM. A Planning and Engineering Committee coordinates the planning and engineering of major facilities on the interconnection, performs system studies on a pool basis, and coordinates planning with other pools.

The agreement requires that each member make available to PJM all generating and transmission facilities in excess of the member's own requirements. This is the provision on which operation as one system depends and which permits economic dispatch of the lowest cost unit to supply the next increment of load regardless of the system on which it is located. The agreement states also that each member must provide, within its system or by purchase, a specified level of installed capacity. If a member system does not have the required capacity installed, it must share its savings in carrying charges at a specified rate with those members having excess installed capacity.

The formula used to determine the installed capacity requirements of PJM and its members is unique. It recognizes that the installed capacity must be no less than the sum of the kilowatts of annual peak load plus the kilowatts of reserve that a probability study has shown to be adequate for emergencies. Even more important, the formula recognizes that the installed capacity on any day must be no less than the sum of the kilowatts of the peak of that day, the kilowatts of a specified reserve, and the kilowatts of capacity unavailable for service because of maintenance or other reasons. This second measure is made each week for the day on which the peak of the week occurs. Its importance is shown by the fact that, on the average, the PJM weekly capacity requirement *exceeds* the capacity required at the time of the annual peak load. Thus there is an installed capacity requirement determined by the annual peak, and an average requirement determined by the seasonal load curve of PJM and the ability of its members to carry out their maintenance programs with a minimum of outage time.

This unique formula has enabled PJM to pinpoint the benefits obtained from coordination of maintenance schedules. Maintenance is scheduled to produce minimum operating costs throughout the year. Each week the performance of the past 52 weeks is evaluated. A system whose capacity is inadequate on either an annual or weekly basis shares its savings in carrying charges as mentioned above. Thus full advantage is taken of the diversity in the shape of the seasonal load curves of the systems and their diversity in equipment outages for maintenance or other reasons.

The rate for installed capacity deficiencies is based on (1) a saving in carrying charges for generating capacity and its associated transmission, plus a saving in fixed operating expenses, less (2) a loss in operating economy as a result of not installing the needed capacity. The carrying charges reflect representative costs of capacity on PJM at the time the agreement was prepared and are not based on the experience of any one system. Fixed operating expenses are principally labor costs but include other fixed expenses. The split-savings principle is applied, with the deficient system paying a rate equal to one half the calculated price times the number of kilowatts of deficiency. The rate for capacity deficiency does not include any associated energy.

Since the interconnection is loaded on the basis of economic dispatch, energy transactions are accounted for, after the fact, for each hour and are based on the principle that supplying and receiving systems shall share the savings equally. The savings are then allocated within the two groups in proportion to the manner in which such savings are created. Operating capacity transactions are similarly accounted for during the high load periods of each day.

The essential features of the PJM agreement may be summarized as follows:

1. A Management Committee requiring unanimous consent.
2. A central Interconnection Dispatching Office.
3. The mutual availability of facilities.
4. An installed reserve requirement and payment for failure to meet this requirement.
5. Energy scheduling at incremental rates on a continuous basis.

Operating policies. The operating policies in use today are essentially the same as those established more than 30 years ago. The formal PJM agreement contains the basic operating principles. Maximum benefits accrue to the members because the systems are operated as though they were one system under one management. With this one-system concept, accounting procedures have been set up to allocate savings as equitably as possible to the systems that create the savings, always keeping in mind that it takes two parties to complete a transaction. However, it should be noted that the operators of the individual systems still are responsible for maintaining service continuity on their own systems.

The success of the one-system concept may be credited to the central dispatching office, which coordinates the operation and accounting for the systems. Monitoring from a central office assures coordination of the operation of thermal units, hydro units, and pumped storage units that are actually dispatched from the control centers of the various systems. Operating reserve capacity requirements are developed on an interconnection basis rather than on a system basis, resulting in appreciable savings to the systems.

The principle of area regulation of generation to match the interconnection load requirements is followed, rather than individual system regulation to match individual system load requirements, resulting in a benefit to the systems by a reduction in their regulating burden. Savings are realized by maintaining free-flowing ties within the interconnection on the basis of continuous economic dispatch to meet total interconnection load requirements and by interchanging power with neighboring pools on the basis of scheduled external tie-line flows. These latter schedules are based on verbal contacts and agreements between the interconnection dispatcher and his counterparts in the power pools surrounding PJM.

Cooperation with other pools

For many years, PJM operated as an isolated pool, with no permanent connected tie lines to other systems other than a few low-capacity ties to the New York State utilities, which were closed only for emergency assistance or to

effect economies under special circumstances. The principal reason for this relative isolation was the different operating philosophy of PJM as compared with that governing surrounding systems. As has been indicated, the members of many interconnections schedule transactions with their neighbors on an hourly basis and thus must control both their system frequency and tie-line interchange very closely. Systems in such interconnections, therefore, have depended on the development of automatic frequency and tie-line control equipment. On the other hand, PJM, by providing high-capacity intersystem ties and by permitting free-flowing interchange among its members, found it unneccessary to control internal tie-line flow and manual regulation of generation to control frequency was satisfactory.

Automatic control equipment was not needed until it became economical to interconnect PJM with the other pools. That time arrived in 1962, and since then various member systems have built tie lines in New York, Ohio, Maryland, and Virginia. PJM is now interconnected with its neighbors and is in the process of strengthening its interconnections by means of 500-kV lines. It has been necessary, of course, for the PJM Central Dispatch Office and the member systems to install automatic load control equipment. Flows on the tie lines from PJM to neighboring pools are regulated, but internal flows among the member systems are not.

Ties to other pools do not, of themselves, create benefits. No benefits are automatically available on any interconnection; a contract is required between the parties to recognize the benefits and to allocate them equitably. PJM is now pioneering contractual relationships with other pools, and since 1951 has had a contract with two large New York State power companies acting as a second group. This agreement provides for installed capacity obligations and the exchange of operating capacity and energy on both an economy and emergency basis and requires payments between the groups for installed capacity deficiencies.

Three additional pool-to-pool agreements, providing for coordinated planning and operation between the groups, were executed in 1965 between PJM and its neighbors to the west and south.

Joint projects of member systems

The close coordination of the planning on the Pennsylvania–New Jersey–Maryland Interconnection has fostered a number of important joint projects that will make significant contributions in maintaining low-cost electrical supply. Chief among these are: (1) Keystone 1800-MW Mine-Mouth Generating Station, (2) Conemaugh 1800-MW Mine-Mouth Generating Station, (3) an extensive 500-kV transmission system, and (4) a large-scale digital computer and data transmission system to schedule and monitor PJM operation. All of these projects have been jointly planned and will be cooperatively owned and operated with the freedom of action of the individual systems being maintained. Still other significant projects are in the negotiating stage.

The Keystone Station, to be located in the bituminous coal fields of western Pennsylvania, will consist of two 900-MW coal-fired turbogenerators. The plant will be owned as tenants in common by seven systems with shares ranging from 2.47 to 22.84 percent. A member system of PJM will operate the plant as a contractor for the owners.

Conemaugh Station, now in the preliminary planning stage, will also be located in western Pennsylvania. To be essentially a duplicate of Keystone Station, it will involve the largest individual coal-mining operation in the United States. Conemaugh will be owned by nine systems as tenants in common.

A 500-kV transmission system, owned by six systems, is now being constructed to bring power from the mine-mouth stations to the load centers and to provide high-capacity interpool tie lines. All of the PJM systems will contribute toward the annual costs of the transmission system. The total cost of all the 500-kV lines and substations and the 500-kV to 230-kV substations has been divided into an interarea tie function and a generation delivery function. The interarea function is allocated to all PJM members and associated systems in proportion to their sizes as measured by peak loads. The generation delivery function is allocated to the owners of Keystone and Conemaugh in proportion to their ownership of the combined capacity of these stations.

Much of this extra-high-voltage system will be completed in 1967 when the first Keystone unit is placed in service. Additional lines will be extended as high-capacity ties to surrounding power pools. The 500-kV system will be greatly expanded with the erection of Conemaugh Station in 1970, when the total circuit length will be about 1500 kilometers.

The large-scale digital computer and an extensive data transmission system will be placed in service in 1967. A communications network will link the central computer to dispatch offices of the member and associated systems where automatic equipment controls the output of the generators throughout the individual company service area. The computer will be used further to improve operating economy through closer coordination of the scheduling and operation of thermal, run-of-river, storage, and pumped storage hydro plants, and through inclusion of transmission loss effects in the incremental loading of equipment. It is expected to provide more reliable operation by giving consideration to transmission limitations in the scheduling of equipment to be operated, and by monitoring actual transmission line loadings. The computer will also perform data logging functions and will produce essential operating reports as well as perform accounting functions associated with the transfer of power among companies.

Conclusion

We have seen that although the benefits of interconnection are more readily attainable on a fully coordinated power pool, there is a point beyond which these benefits are outweighed by the complications involved in such coordination. The Pennsylvania–New Jersey–Maryland Interconnection, the largest fully coordinated power pool in the United States, has been cited as establishing a successful pattern for such coordination.

The PJM Interconnection has also pioneered pool-to-pool agreements as a further step in minimizing the cost of electric energy to the public it serves. Close coordination of the planning among the systems has resulted in a number of projects that will be of significant value in maintaining a low-cost electrical supply.

Essentially full text of a paper presented at the World Power Conference, Tokyo Sectional Meeting, Tokyo, Japan, October 16–20, 1966.