THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIV MARCH 1965 NUMBER 3

New Concepts in Exchange Outside Plant Engineering

By H. S. EDWARDS and H. Z. HARDAWAY

(Manuscript received November 5, 1964)

Concentrated study by the American Telephone and Telegraph Co. and Bell Telephone Laboratories has resulted in several engineering and design innovations that permit more efficient utilization of the exchange cable network. As a first step, a mathematical model of the customer loop plant was developed from survey data. With this model, studies have been made of the transmission properties of the loop plant at both voice-band and carrier frequencies via computer analysis. Results of such studies have been useful in planning plant improvement programs and have also been used to evaluate such new concepts as "dedicated outside plant" and "uniform-gauge customer cable plant."

Other computer programs have been and are being developed to aid in engineering cable routes for future growth and to evaluate alternatives to placing new cable, such as concentrators and exchange carrier systems. Studies to optimize the placement of new switching centers, taking into account existing wire centers and forecasts of growth for the area, have been made by computer analyses. These computer programs aid engineers in making studies in much more depth and in less time than was possible with older cut-and-try methods.

I. INTRODUCTION

Since World War II there have been major changes in exchange outside plant cable networks. Polyethylene has replaced lead for cable sheaths, and in the distribution plant, polyethylene insulated conductor

cable (PIC) and ready-access terminals have replaced paper insulated cable and sealed terminals. Concentrator and carrier systems permit more efficient use of copper in feeder and trunk cables.

In the early stages of these wide-sweeping changes, the major advances were in hardware, as enumerated above; however, exchange plant engineering methods were being analyzed and revised to take full advantage of these innovations. The PIC cable and ready-access terminals had implications on exchange plant installation and maintenance much broader than the purely hardware ones. For instance, the installation of distribution terminals for access to the cable conductors could be postponed economically until required to satisfy a request for service.

Concurrently, unrelated activities were producing results applicable to the plant engineer's problem. Operations research techniques (which aim to optimize an existing system) were being used. More sophisticated electronic computers became available and provided the tools of calculations and machine decision logic on a scale impossible in the past. All of these factors sparked a revolution in the tools and methods used by the engineer to study and evaluate the exchange outside plant as a system, with the ultimate objective of improved service for customers.

It is the purpose of this paper to show how these modern engineering tools and methods are making possible new concepts in the engineering and utilization of the exchange outside plant. Initially, the exchange outside plant was studied as an integrated system. As the work progressed, it was necessary due to the size and complexity of the study to consider each engineering activity as an entity rather than a part of a system. Therefore, for ease of exposition, this paper covers each engineering activity as it was developed during the exchange outside plant engineering study.

II. BACKGROUND

An exchange outside plant cable network (Fig. 1) serves as a medium to connect the central office and station equipment in a manner which is compatible with signaling, supervision and transmission requirements. These requirements usually are stated in terms of circuit resistance and transmission limits. The cable networks are designed to keep within these limits regardless of the distance between the office and the customers. This is accomplished by planning the network around the several options of wire gauges (19, 22, 24 and 26), carrier systems, and the many loading arrangements (H88 and H44, etc.).

Interface problems become quite complex with a cable network that is laid out to connect all customers in an area to a central office. Such a

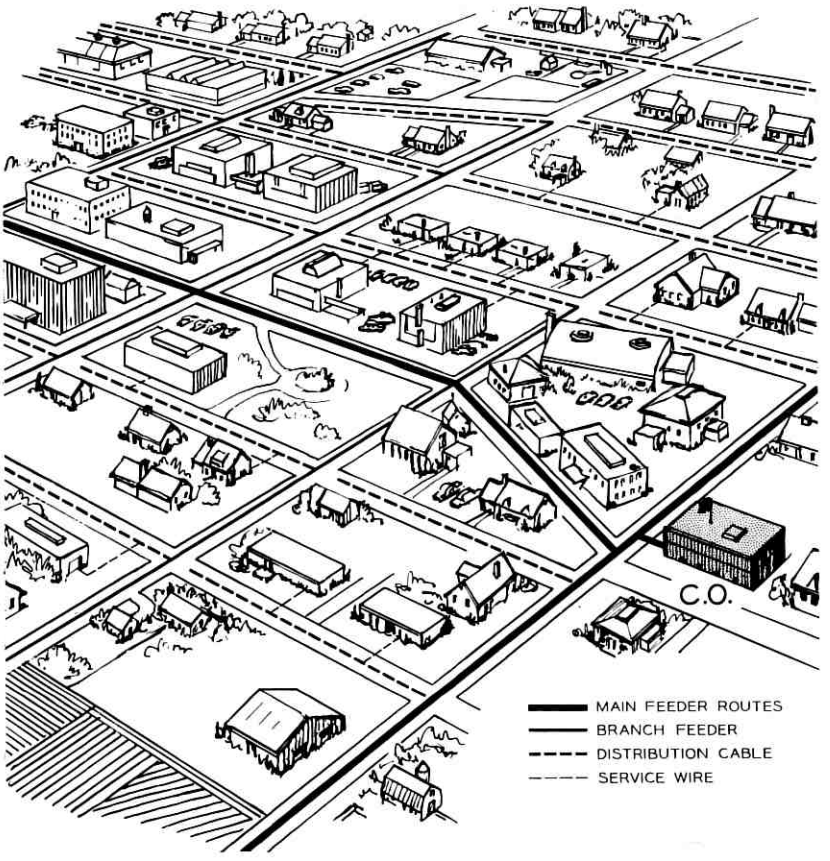

Fig. 1 — Exchange cable network.

network may serve a high-density area of 100,000 customers or more per square mile, as in New York City, or a few hundred customers spread over many square miles, as in some areas of the West. Regardless of area size and type of switching equipment, each network must be carefully designed for the customers it is to serve. Although the size of an area or the number of customers may vary widely, the engineering methods throughout the Bell System are similar. However, these methods reflect the individuality and philosophy of the engineer and the associated company far more than in any other part of the communications system.

The job of engineering facilities in relatively small increments to meet the unique conditions of the area and customer requirements

inherently results in a specific network design. This in turn has made it difficult to both obtain and analyze system-wide data about these networks and to set requirements for new systems with confidence that the proposed system when developed would be of optimal usefulness to all associated companies.

In the past, small segments of the plant judged to be representative of the Bell System were studied in detail. These studies evaluated, from the system viewpoint, transmission improvement characteristics or economic advantages of a new development. This procedure was time-consuming and unsatisfactory, and the time associated with obtaining, processing, and analyzing system-wide data was prohibitive. However, by 1958 the application of computer data reduction and analysis techniques made it possible to obtain a much more comprehensive and accurate picture of the exchange plant than had theretofore been possible. This was one of the most basic steps in the application of the new systems engineering concepts and led to several other surveys of the physical and electrical characteristics of the telephone plant.

III. SUBSCRIBER LOOP SURVEY

In 1960 a sampling survey was designed to yield statistically sound estimates of important characteristics of the customer loop plant. A sample of loops, representative of the facilities provided to the approximately 40 million residential and business customers served by the Bell System, was taken. The loop selections were made from a sampling frame containing a complete list of all central office buildings in the Bell System, together with the central office prefixes assigned in each building, and the total number of customers served from each prefix. From this list, in which each customer was implicitly numbered, 1000 telephone numbers were picked in such a way as to form an optimum stratified random sample,[1] with heavier concentration in office sizes expected to contribute the most variability.

The desired information concerning physical composition of the loop plant was obtained from the outside plant cable and wire records maintained by the associated companies. Fig. 2 represents the kind of information provided for each sampled loop.

To derive transmission properties the computer had to be programmed to reconstruct each loop exactly as it appeared in the physical plant. The computer converted the entire loop between the serving central office and the sample telephone into an equivalent T network at each frequency of interest in the voice band. More details on the

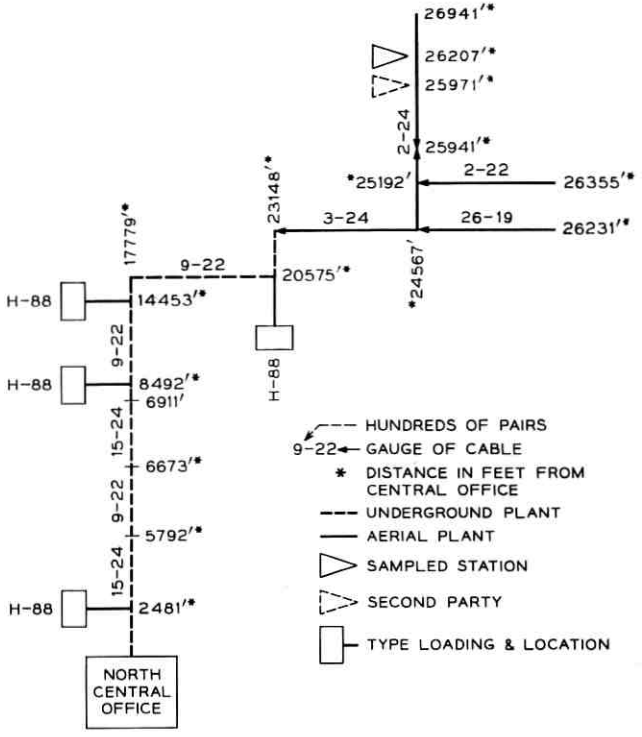

Fig. 2 — Physical composition of loop plant: information from wire and cable records.

method of computing transmission characteristics and specific results are given in Ref. 2.

The more important survey results were summarized as statistical distributions. As examples: (1) the average working distance to a customer in the Bell System was found to be 10,300 feet with 90 per cent confidence limits on this mean value of ±450 feet, as presented in Fig. 3; (2) cumulative insertion losses at seven discrete frequencies in the band from 200 to 3000 cycles are given in Fig. 4. At 1 kc the mean value of insertion loss in loop plant was found to be 3.5 db with 90 per cent confidence limits of ±0.1 db. (3) The degree to which exchange loop input impedance (including station set) matches the toll network is shown by using return loss at each of the six frequency distributions shown in Fig. 5. The best return loss is at midband — around 1 kc, where the mean return loss is 15.0 db with 90 per cent confidence limits of ±0.15 db. The lowest return loss at 3 kc is representative of high

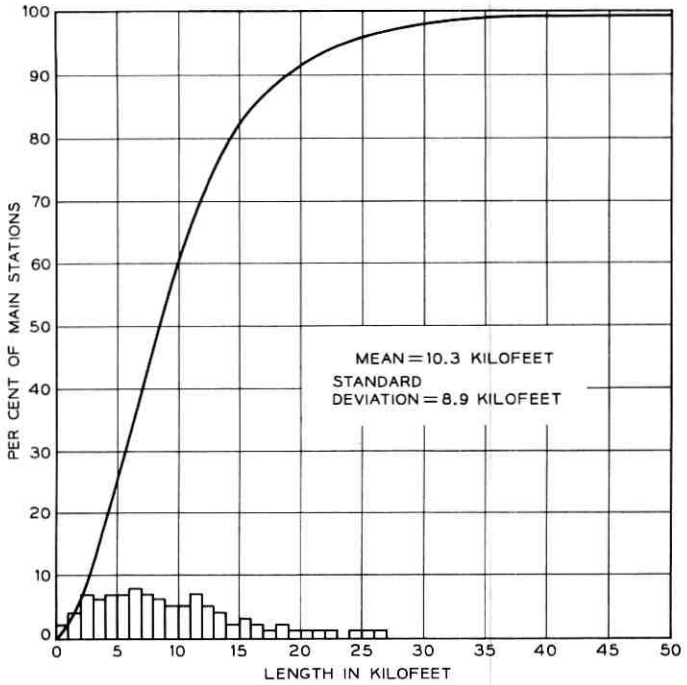

Fig. 3 — Distribution of working lengths to sampled main stations.

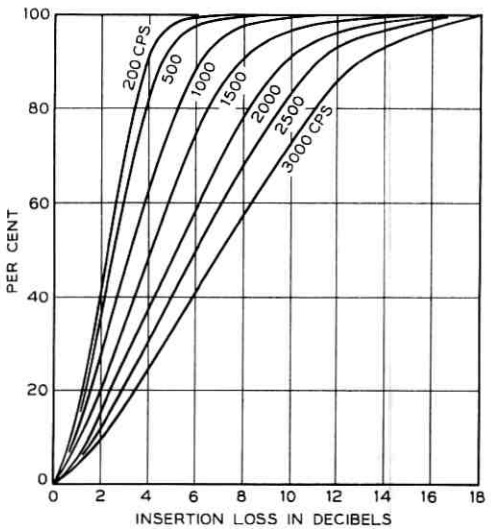

Fig. 4 — Cumulative insertion loss distributions for Bell System loop plant, 200–3000 cps (measured between 900-ohm terminations).

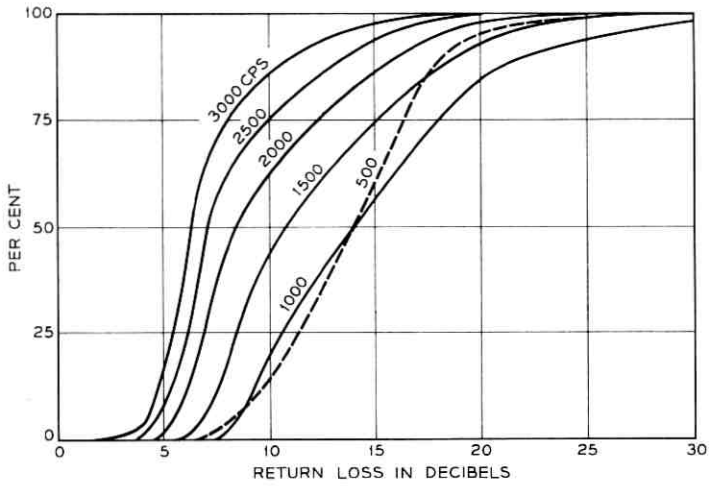

Fig. 5 — Cumulative return loss distributions for Bell System loop plant, 500–3000 cps (measured against 900-ohm + 2 μf compromise balancing network with loop terminated in actual station set impedance).

singing frequencies, where the mean value was found to be 7.2 ± 0.2 db.

One phase of this survey of particular significance was determination of loops that had design irregularities. With these data (composed of the percentage of irregularities by type), the magnitude of the job required to correct these conditions could be estimated and a realistic plant improvement program planned.

In addition to providing guidance for an improvement program, the survey made it possible to develop a statistically sound mathematical model of the existing customer loop plant. This model has been used with considerable manpower savings over the analytical procedures used by Bell Laboratories in the past to determine accurately the transmission effects of new developments designed to be used with the existing plant. With this new tool, studies have been made of (a) effect of cable capacitance variation on transmission properties of loop plant, (b) input impedance of loop plant both at the customer and office end of the loop, (c) need for impedance compensation networks at the central office, (d) characteristics of loop plant at carrier and PICTUREPHONE system frequencies, and (e) the optimum telephone set impedance characteristics. Other uses of these data have to do with the evaluation of new methods of laying out a cable network such as "dedicated outside plant" and "uniform-gauge subscriber cable plant."

IV. DEDICATED PLANT

The dedicated plant concept involves the permanent assignment of a cable pair from the central office to each main station. All party lines are bridged at the central office. This new method of laying out plant was preceded by a gradual but definite change in the composition of telephone plant and customer requirements that began in the early 1950's. Developments such as PIC cable and ready-access terminals provided greater possibilities of circuit availability than did the pulp insulated cables and hermetically sealed terminals previously used. The percentage of households without service was steadily dropping, and at the same time there was an increasing demand for individual line service (see Fig. 6). The labor costs were increasing rapidly for the plant rearrangements necessary to satisfy the changing service requests of customers. All of these favored a more permanent plan of outside plant pair connection than current multiple schemes.

Cable and wire plant is sized on the basis of growth forecasts not only to meet known requirements but to be adequate for some predetermined time in the future. It is difficult to predict the growth pattern, the number of lines, and the type of service for a central office area, and it is even more hazardous to estimate the growth along any given cable route. These uncertainties, along with the inherent difficulty of obtaining access to pairs of pulp insulated distribution cable used in

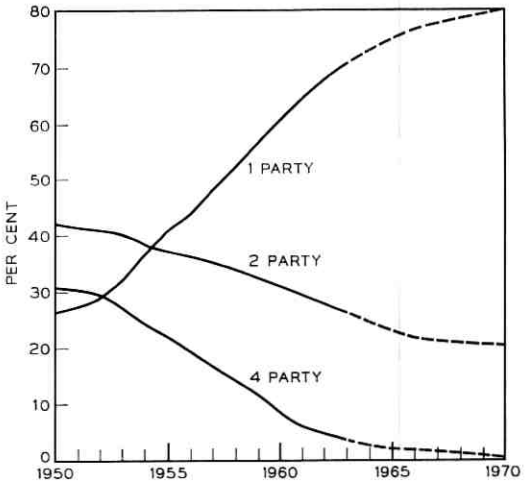

Fig. 6 — Distribution of party line service for all associated companies at end of year shown, with forecast for 1965–1970.

the past, led to the multiple appearance of subscriber cable pairs, not only in several cables, but also at a number of customer terminal locations following a predetermined pattern as shown in Fig. 7. (The multipling of cable pairs is illustrated by the termination of the 1800-pair feeder cable with a 600- and 900-pair branch feeder and a 900-pair main feeder.) Multipling was also necessary to achieve high cable pair utilization and to provide party line association. However, as actual demand does not always match the anticipated growth, it is necessary under this system as growth develops either to rearrange the cable pair layout and unmultiple the pairs, or to leave unused copper in the plant.

The use of multipled cables, cross-connect terminals, and rearrangement of cable complements imposes technical problems and ever-increasing

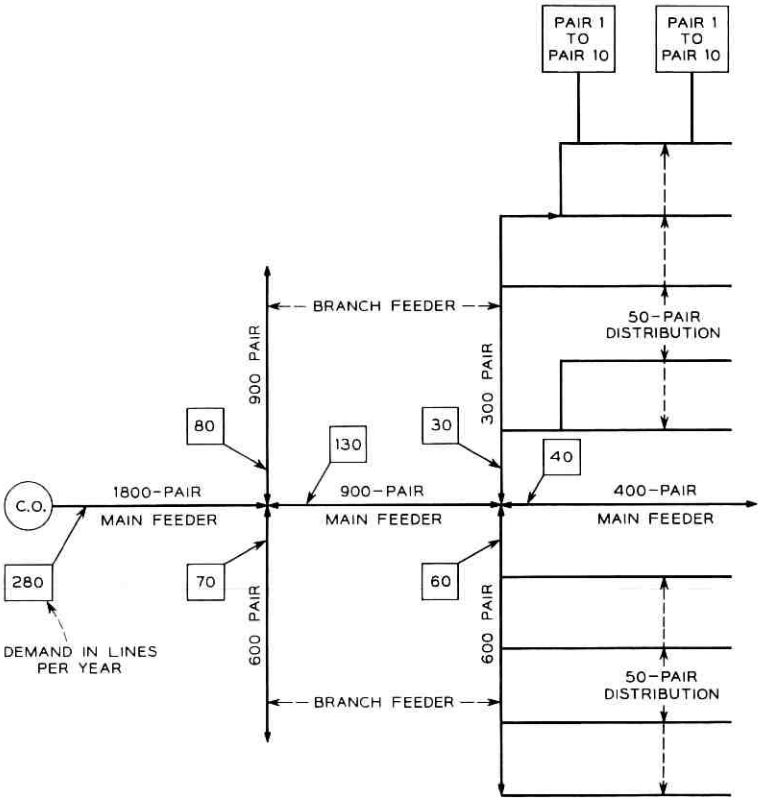

Fig. 7 — Typical feeder route, showing main and branch feeder and distribution cables.

operating costs. For example, when plant rearrangements are required to provide service to a new subscriber cable, pairs must be meticulously searched out, identified, and (often in more than one place) cross connected or spliced. All of these activities require extensive labor. The costs are high and the likelihood of error is always present along with possible interruptions to service.

In view of the recent innovations in hardware and the increasing demand for individual telephone service, this method of laying out plant left something to be desired. Therefore it was necessary to come up with a new scheme that involved completely new cable network design principles capable of virtually eliminating cable rearrangements while retaining the ability to handle growth on increasingly shorter time intervals. Such a scheme was made possible by initially dedicating a percentage of cable feeder pairs to serve a specific area along the route and keeping the remainder in reserve as spares to be dedicated to an area later, as required to satisfy requests for service. The optimum percentage of feeder pairs designated as spares was determined by using computer simulation techniques to study a number of actual plant growth situations. Flexibility points ("control" and "access" points) were conceived to permit ready access to the spare pairs as dictated by future needs along the feeder route.

The use of this dedicated pair concept eliminates the need for multiple appearances of the pairs and permits direct wiring of the customer's residence to the central office while still providing sufficient flexibility (Fig. 8 shows a multipled designed loop by dotted lines and a dedicated loop by solid lines). Once a pair has been assigned to an address, it remains dedicated to that location whether the pair is working or idle, and regardless of class of service. Any required bridging of party lines will of necessity be done at the central office, utilizing switch-like devices to remove the effect of other party stations during conversation. Theoretically, this connection arrangement would result in some advance in capital expenditures for additional feeder cable pairs and other apparatus; however, the savings in the cost of day-to-day operation derived with such a plant design will far outweigh the carrying charges on the advanced capital. Also, ultimately less total capital will be invested due to increased flexibility and the elimination of the multipled portion of all circuits, which also results in an improvement in transmission.

The feasibility of converting existing plant and of installing new plant under the dedicated concept has been studied. The study results indicate that this concept is economically attractive for all residential loops up to approximately 30 kilofeet in length.

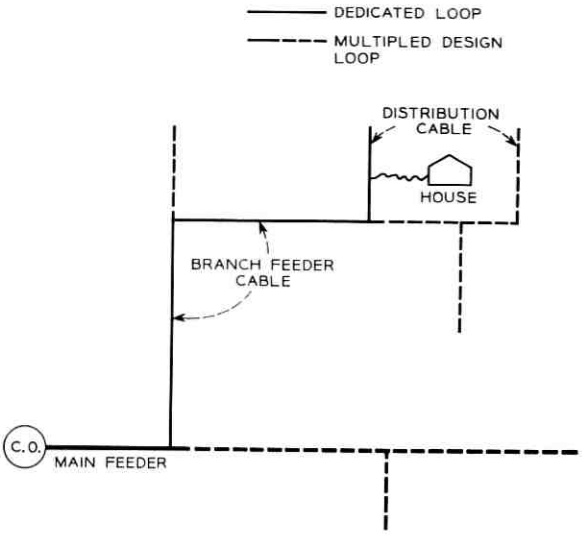

Fig. 8 — Dedicated outside plant, typical feeder route.

The implementation of the dedicated plant technique throughout the Bell System requires familiarizing practically every department with the advantages of such a system and also with how it will affect current methods of operation. Some of the advantages of the new concept borne out by field trial and further confirmed by actual field experience are:

(a) improved efficiency of over-all copper usage,*

(b) reduction in cost of installed cable,†

(c) better transmission through reduction in length of bridge tap,††

(d) virtual elimination of cable, service wire, and central office main frame transfers, thus simplifying assignment and installation procedures (see Table I), and

(e) simplified records, resulting in faster handling of customers' orders.

Actual system application of this concept will be a gradual process, but it is expected that the plant will be converted fully by about 1970, and that large savings will be produced by the elimination of rearrangements and changes in the cable plant, including changes that are at present made for higher cable fills (see Table II).

* Since there are no end sections or multipled connections under the spare pair concept, the entire length of all used cable pairs ultimately will be working.

† Fewer multiple wire connections to make when splicing cables together at junctions of feeder and branch cables.

†† The unused copper in a subscriber's circuit is referred to as bridge tap.

TABLE I — RESULTS OF STUDY AND FIELD TRIAL

Operation	% Reduction	
	Predicted	Actual
Cable pair and service wire transfers	90	97
Cable pair transfers	90	100
Central office main frame transfers	90	97
Service order assignment — residential	40	unknown*
Installation time	substantial	65

* Conversion to simplified records had not been completed during first 7 months.

Dedicated plant will not, however, eliminate rearrangements and changes necessary to reroute customer service to a different central office due to shortage of switching equipment in a specific exchange area or recovery of coarse (19 and 22) gauge cable plant.

V. MULTIGAUGE DESIGN

Before discussing the desirability of further reducing the number of rearrangements and changes in the cable plant, it is appropriate to explore another reason for such activity. The per pair cost of cable conductors varies widely; first as to gauge of the conductors used and second as an inverse function of the total number of pairs included under a given cable sheath. With the coarser gauge the cost per pair rises rapidly due to the increased cost of the copper used. On the other hand, as the size of a cable of given gauge is increased the cost per pair in plant goes down due to both the lesser relative cost of cable sheath and the more or less common placing cost (see Fig. 9)

Because of transmission and resistance limits of both station and central office equipment and the economic factors outlined previously, it is

TABLE II — RESULTS OF FIELD TRIAL CONVERSION TO DEDICATED PLANT

Date	% Pairs in Use at Central Office	
	Working	Assigned
Before dedication		
January 1, 1962	75.8	75.8
After dedication		
March 1, 1962	76.5	92.1
January 1, 1963	78.2	94.0

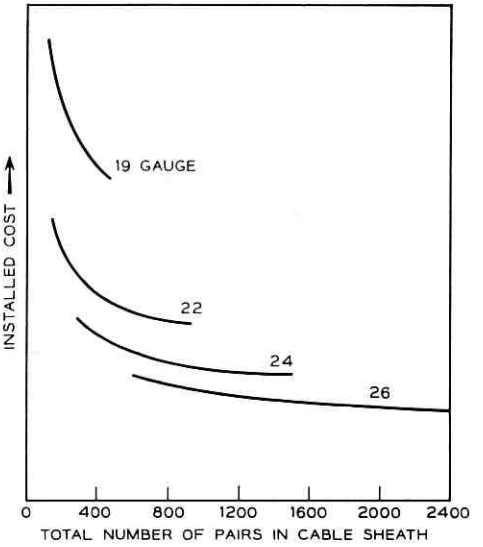

Fig. 9 — Installed costs per circuit mile of underground cable.

a common occurrence to find three or four gauges of cable (from 19-gauge to 26-gauge) in a single exchange cable route (Fig. 2). Frequently, the initial cable placed in the route is coarse-gauge in order to satisfy the requirements of the longer circuits (see Fig. 10). To provide facilities for a reasonable period of time, it contains more total pairs than are currently required in the coarse-gauge area. Then, to postpone the cost of also placing fine-gauge cables, these coarse-gauge pairs are used temporarily for service in areas where fine (26) gauge is sufficient. (This practice is usually followed rather than that of installing composite cables with two gauges of conductors contained in a single sheath.) Later, when customers' requests for service at the extreme end of the route require the remaining coarse-gauge pairs, a finer-gauge cable is placed from the central office, and the circuits which were temporarily served by the initial coarse-gauge cable are transferred to the new cable. Not only is such transfer work costly, but, in addition, the handling of working cable pairs is always at the risk of interference with customer service.

A similar problem arises with special design considerations necessary to meet transmission requirements on the longer loops. Specific pairs of a cable are selected and loaded at discrete distances along the pair. This added complexity results in administration problems, particularly if

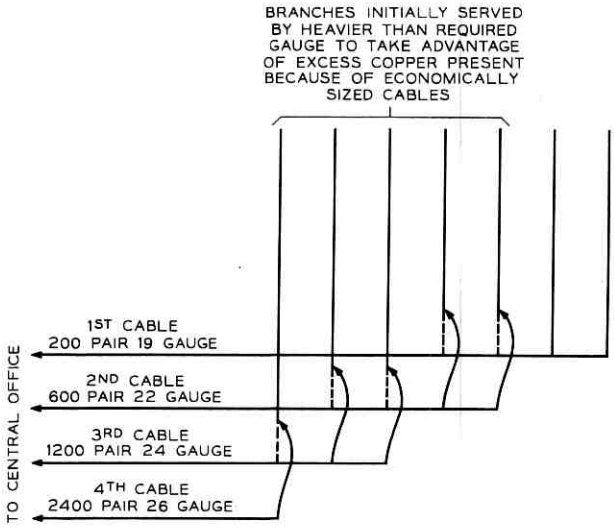

Fig. 10 — Feeder cable route, illustrating gauge requirements.

the pairs are needed in the future for a different service. The loading must then be removed or changed, depending upon proposed circuit requirements.

VI. UNIFORM-GAUGE SUBSCRIBER CABLE PLANT

If it were practical to design and operate a cable feeder route containing only pairs of a single fine gauge and minimum loading, savings would be realized both in capital investment and in operating expense. For example, all the line growth in the route would be absorbed in the single gauge rather than spread over several as is presently the case. An immediate effect of this would be that cables placed in the route would tend to be larger in total number of pairs. Thus not only would the advantage of per-pair cost reduction with the larger cables be realized, but the total number of cables in the route would diminish. Also, fewer ducts would be specified in underground cable structures (conduit) and existing conduit would be used more efficiently. Of course the need to transfer branches to recover coarse gauge would be entirely eliminated, along with its high expense and adverse effect on service.

The engineering of a single-gauge feeder cable relief project would be tremendously simplified, with resultant reductions in engineering costs. Also, the over-all efficiency of the route would be improved, as spare cable pairs would be needed for only a single gauge, as compared

to spare pairs for each of the four gauges which are considered as separate entities under the multigauge plant. Naturally, there will be some offsetting penalties resulting from the need to compensate for the added loss and resistance of the finer-gauge cable.

Studies undertaken to explore the technical and economic factors involved in realizing this objective indicate the feasibility of serving customers within 30 kilofeet of No. 5 crossbar and No. 1 electronic central offices with all 26-gauge cable and less than half the loading now required in loop plant. The customers located beyond 30 kilofeet from their serving central offices would still require coarser-gauge facilities and loading. Requirements for new gain devices and signaling range extension equipment are being developed to implement this concept, including the electronic devices necessary to meet special service transmission objectives. When proven operational, this concept, combined with dedicated plant, will result in a completely new method of laying out customer loop plant, with a great reduction in the multigauge problems mentioned previously.

VII. COMPUTER METHODS

Along with these new engineering concepts, electronic computer programs have been developed and others are being developed to aid the engineer in making studies to determine the optimum plant layout and how best to introduce new engineering and system designs into the exchange network.

To engineer a cable addition to an existing network, data pertaining to the status of each cable pair are gathered from the cable location records. With this information and a forecast of growth requirements, engineering plans are formulated for a number of possible solutions to satisfy the demand for service. The conception of alternate plans and the final decision require engineering judgment which is a function of the engineer's training and knowledge of the area. Having selected a number of plans, the engineer makes a detailed analysis of each possible solution to determine its feasibility and cost. This repetitious analysis is a major time-consuming task, particularly if several solutions appear worth studying.

Careful review of the analysis and evaluation techniques revealed that the modern digital computer was ideally suited to aid the engineer in making these studies. It was possible to formalize parts of the engineering know-how so that data (see Fig. 11) could be entered by simple language into a computer, where its equivalent representation could be manipulated more rapidly and precisely than by the engineer. Thus

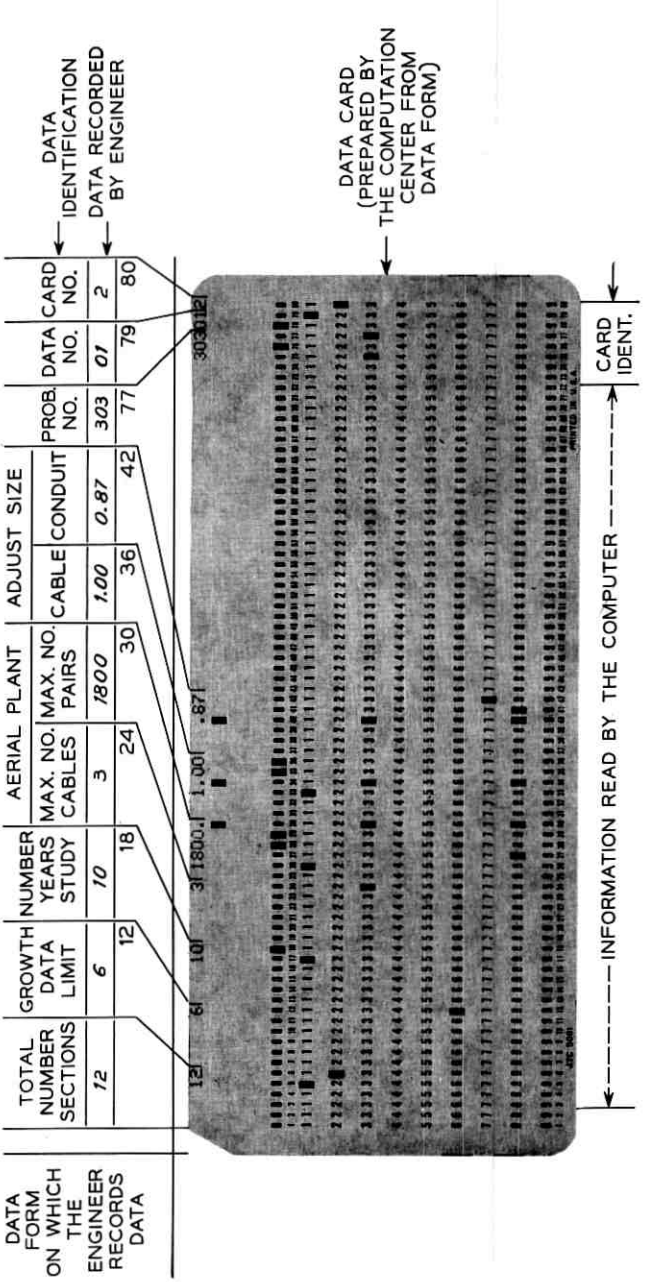

Fig. 11 — Cable relief study data.

freed of the detailed calculations, the engineer could concentrate on the design activities which could not be treated mathematically.

The resulting computer program (see Refs. 3 and 4) produces a complete cable route design for each year of the study period. It is very flexible, permitting the use of local cost data and design rules. Even the most complex cable problems can be handled and an optimum solution obtained for the engineer's final evaluation. Fig. 12 illustrates in schematic form the computer solution to a simple cable relief problem. Although only fourteen sections are shown, the program is capable of handling 99 sections. This program is available on a Bell System basis and is now operational in most of the operating companies.

Of particular interest is the fact that the computer has the capability of exploring the effect of modified rates of growth in a particular cable section more rapidly and economically than possible by the engineer repeating the numerous hand calculations. With this capability, the engineer can consult frequently with the forecaster to obtain his views with respect to areas having unusual growth potential, as well as any substantial deviations from trend which may not have been reflected in the forecast. Forecasting growth in an exchange area will be discussed in more detail later.

In the past when additional pairs were required in the cable route, small increments of cable were added as needed. Now, in addition to cable, new systems such as concentrators and carriers are beginning to play an important role in providing relief facilities. Superimposing electronic equipment on the cable network will have far-reaching effects upon construction and maintenance of the plant. Also, as each new switching or transmission system creates another alternative solution for each cable network growth problem, engineering becomes more involved, time-consuming, and costly.

Therefore a computer program has also been written to explore the cost of using multiplexing systems such as concentrators to postpone cable relief. In addition, this program evaluates all important and unique features of the route and determines an installation and removal date for each of the concentrators.

VIII. WIRE CENTERING

Programs developed for engineering cable routes and evaluating the use of concentrators are also valuable for calculating costs associated with major switching additions to exchange networks. These costs constitute a major factor in deciding where to add additional switching centers in a growing community.

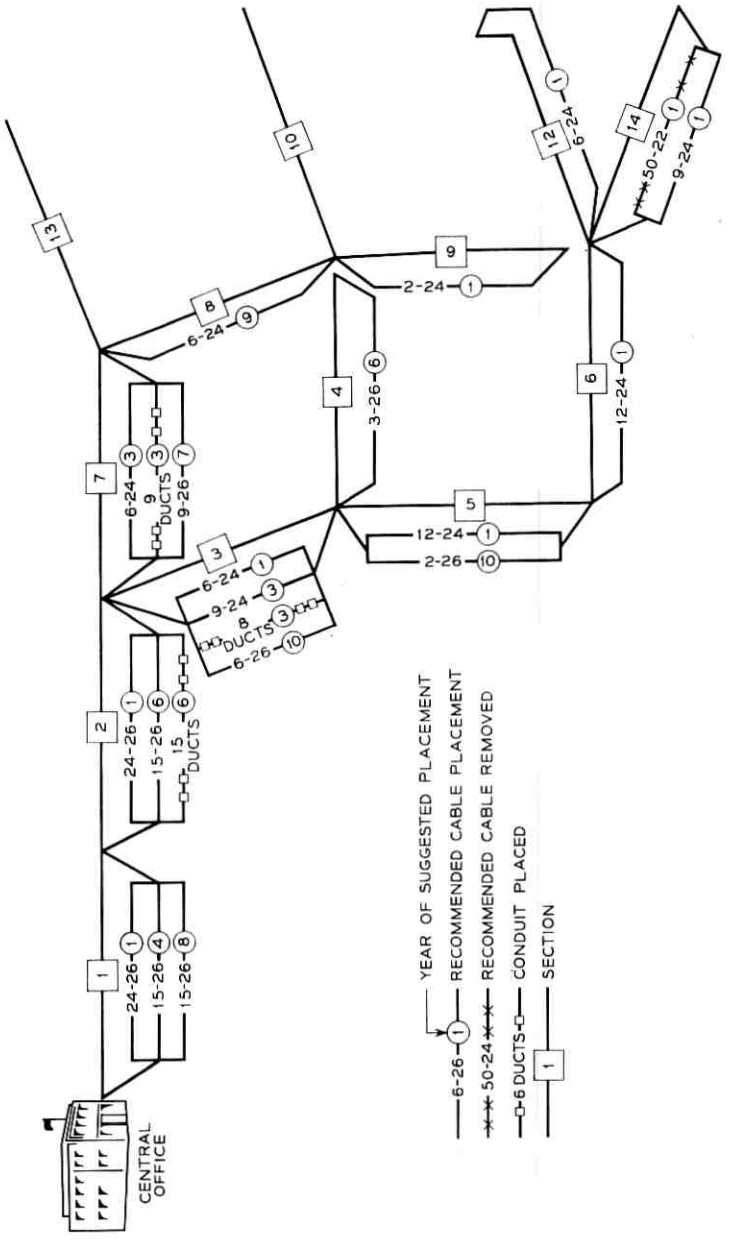

Fig. 12 — Schematic of cable relief study.

Studies made to determine when and where new switching centers should be established are commonly called wire centering studies. The over-all purpose is to obtain the optimum economic combination of outside plant and switching equipment in geographical areas as large as several hundred square miles. More specifically stated, the problem is to determine when and where to add new central offices in the area, considering the configuration of the existing plant, the anticipated growth and the costs associated with reinforcing and extending existing outside plant facilities.

Wire centering studies usually begin with a growth forecast by outside plant cable route sections and an estimate of traffic loads anticipated for the period under study. The first phase is known as a cross-sectional study, an analysis of the situation at a specific future point in time. This type of study gives some idea of the need for additional wire centers and determines approximately their location to satisfy forecasted requirements for customer service.

Usually, the engineer estimates the cost of providing facilities to meet anticipated customer demand from existing wire centers. This estimate is needed as a basis for comparisons of alternate means of providing service. Alternative solutions are then evaluated to determine if the total cost of providing service at this point in time would be less if the area were to be served by additional wire centers at a number of different locations within the area. These cross-sectional studies are repeated using various numbers of switching centers, several growth estimates for the area, and different time intervals. The number of combinations explored can number in the thousands, especially when four, five, or six new wire centers are being considered.

The second phase of the study consists of determining more accurately where and when additional wire centers should be added. This involves making detailed present worth of annual charges (PWAC) comparisons over a 20–30 year study period, first serving a study area by an existing feeder route or routes from one or more existing wire centers and then serving the same area from the combination of existing routes and wire centers with one or more wire centers added. Until now, these studies have been made on a cut-and-try basis and are time-consuming, costly and laborious. Frequently, in fast-growing areas, it is necessary to reach a decision and start the construction of either a new office or additional outside plant before the study is completed.

The cross-sectional method for determining the number and approximate location of new wire centers has been studied and a computer program developed which mechanizes many computations heretofore

tediously performed by the engineer. The engineer is still required to gather the same initial information as needed for a manual wire centering study, except that with the wire centering program the data are used as input to the computer. The required data are as follows:

(1) anticipated number of subscriber lines and their location for each year to be studied (Typically, these data are required for 5, 10, 15 or 20 years into the future.)

(2) number of existing wire centers, their location, and other pertinent characteristics

(3) trunk pattern between existing wire centers

(4) average cost of loops and trunks as a function of length

(5) number of proposed wire centers to be considered.

All of this information must be recorded so that the computer can store and manipulate the data. This is accomplished by superimposing a grid system over the area to be studied and associating all growth and wire center locations with this grid system. The growth of 100 customers at an intersection of grids 5,5 is shown in Fig. 13 as an example.

With the quantity and the location of subscriber lines determined, the engineer is ready to proceed with the study by having the computer

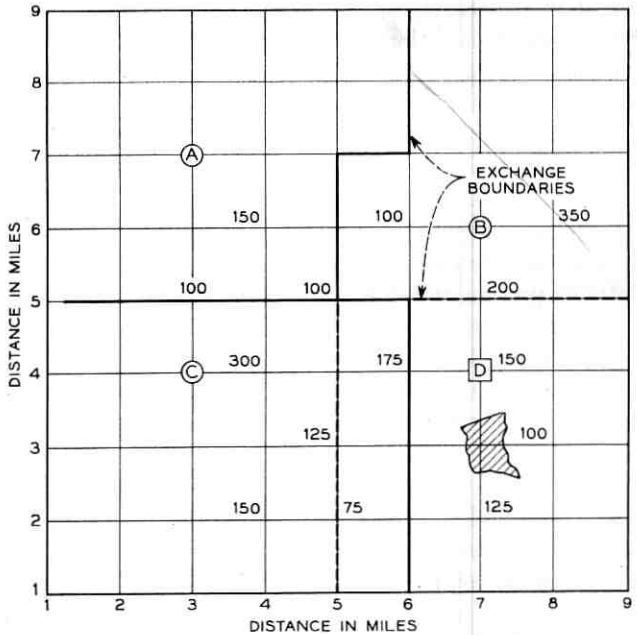

Fig. 13 — Wire centering study area.

calculate the outside plant costs for the study areas considering only the existing offices. After the outside plant costs are calculated for a study area, the building additions, equipment, and additional land costs are estimated by the engineer. Essentially, the entire plant required for serving the customers with existing offices is priced out for each study year. Costs for building additions are needed, since this is usually an important reason for considering a new wire center in the first place. These costs become the reference against which all other possible solutions will be compared. The present computer program will handle eighteen existing wire centers and six proposed centers in a study area. Forty trial locations for each proposed center are possible.

The redistribution of customers to wire centers affects the traffic load within and between centers. This is recognized, and the program distributes the traffic loads in proportion to the customers transferred into and/or out of each wire center.

The output of the computer includes (1) general information regarding the problem, such as the study date, (2) existing wire center data such as present trunk pattern and cost of providing service with existing centers, (3) a list (see Fig. 14) of the ten most economical proposed wire center locations and their associated cost, and (4) a detailed description of the outside plant assignment and the trunk pattern for the best solution. After several field trials of this program, it was accepted as a useful planning tool and has been made available for Bell System adoption.

The primary advantage of using the computer program is its flexibility for examining quickly many alternatives and variations of a given problem which previously could not be examined without complete expensive manual recalculations. If changes occur which were not originally anticipated, a restudy of the area can be made by simply changing the affected information (stored on punched cards) and resubmitting the problem to a computer center.

The program will aid in keeping future plans up to date with a minimum of effort on the part of the planning engineer. Thus with current engineering plans, decisions can be made more deliberately. In addition, the engineer will be in an even better position with a computer program now being developed to aid in determining when a new center can most economically be constructed.

Along with the engineering of cable routes and wire centering studies, other related factors are being considered. As an example, there exists a close economic relationship between the switching techniques, the degree of decentralization of switching equipment, and the configuration of the interoffice trunk cable networks which may be combined with

```
NO WIRE CENTER(S) ADDED
     LOOPS ----------$ 3215722
     TRUNKS ---------$ 1195623
     TOTAL - LOOPS AND TRUNKS ----------$ 4411345
     LAND ------------------------------$       0
     BUILDING --------------------------$       0
     CENTRAL OFFICE EQUIPMENT ----------$       0
     TOTAL -----------------------------------------$ 4411345
4 WIRE CENTER(S) ADDED
     LOOPS ----------$ 2313992
     TRUNKS ---------$ 1329133
     TOTAL - LOOPS AND TRUNKS ----------$ 3643125
     LAND ------------------------------$       0
     BUILDING --------------------------$       0
     CENTRAL OFFICE EQUIPMENT ----------$       0
     TOTAL -----------------------------------------$ 3643125
ECONOMIC ADVANTAGE OF ADDING WIRE CENTERS -------------$  768219

     LIST OF TEN BEST SOLUTIONS AND ANNUAL CHARGES
   LOOPS        TRUNKS         TOTAL        PENALTY      LOCATION(S) OF ADDED WIRE CENTER(S)
$ 2313992.22 $ 1329133.48 $ 3643125.72 $      0.  (16.5,20.0) (18.5,23.0) (12.0,22.0) (13.0,26.0)
$ 2323909.50 $ 1326039.36 $ 3649948.88 $   6823.16 (16.5,20.0) (17.5,24.0) (12.0,22.0) (13.0,26.0)
$ 2326183.09 $ 1335968.23 $ 3662151.34 $  19025.63 (16.5,20.0) (19.5,24.0) (12.0,22.0) (13.0,26.0)
$ 2361164.34 $ 1318489.83 $ 3679654.19 $  36528.47 (15.5,19.0) (18.5,23.0) (12.0,22.0) (13.0,26.0)
$ 2347148.28 $ 1333569.05 $ 3680717.34 $  37591.63 (16.5,20.0) (18.5,23.0) (12.0,22.0) (14.0,27.0)
$ 2357065.59 $ 1330218.39 $ 3687284.00 $  44158.28 (16.5,20.0) (17.5,24.0) (12.0,22.0) (14.0,27.0)
$ 2359339.16 $ 1340834.95 $ 3700174.13 $  57048.41 (16.5,20.0) (19.5,24.0) (12.0,22.0) (14.0,27.0)
$ 2373477.06 $ 1332824.52 $ 3706301.59 $  63175.88 (16.5,20.0) (17.5,24.0) (12.0,22.0) (12.0,27.0)
$ 2394320.44 $ 1322845.81 $ 3717166.25 $  74040.53 (15.5,19.0) (18.5,23.0) (12.0,22.0) (14.0,27.0)
$ 2439273.88 $ 1283318.17 $ 3722592.06 $  79466.34 (14.5,20.0) (18.5,23.0) (12.0,22.0) (13.0,26.0)
```

Fig. 14 — Economic summary — annual changes, year 1990.

subscriber cable facilities. Potentially large savings in copper conductors are possible, particularly through the location of switching equipment near maximum subscriber density. These techniques must, of course, be supplemented with reasonably accurate forecast of future customer requirements.

IX. FORECASTING

A large segment of the Bell System's investment for new construction is spent each year on additions to exchange outside plant. The

quality of techniques for making forecasts and decisions as to where, when, and what additional telephone facilities are required greatly affects the efficiency of the large annual plant construction program and may result in failure to meet customers' telephone needs on time. Present forecast results are not altogether satisfactory in spite of much effort on the part of both commercial forecast and development personnel and plant engineering forces. Errors in prediction are costly, sometimes sufficiently so to offset the advantages of the most carefully engineered project. Outside plant forecasting therefore is an important function worthy of the forecaster's best efforts — certainly it is an area where improvement could result in substantial dollar savings.

Generally, the growth rate of an area is not constant, although there are some patterns of cumulative growth which will be discussed later. Wide fluctuations in the rate of growth can occur for a variety of reasons. These include the location and accessibility of the land, ownership and value of the property, availability of utilities (particularly sewage disposal), development of adjacent areas, penetration of the housing market, political or municipal climate or action, changes in the level of business activity, employment opportunities, zoning restrictions, tax structure, and a host of others. It is important that the forecaster and the engineer recognize these factors. They also present a good argument for considering each forecast section on its own individual merits and against adopting a purely mechanical forecasting procedure which might preclude sound business judgment.

Procedures for maintaining outside plant planning studies covering fundamental feeder routes on a current basis have been implemented. Cable facility charts which graphically display the relationship between existing cable pairs, usable pairs, past trends of working pairs and forecasts of line growth have proved to be invaluable to both the forecaster and the engineer of outside plant in interpretation and analysis (see Fig. 15). Such charts permit more complete and sharper analysis of growth, both past and forecasted, and its relationship to engineering planning and programming.

An ideal forecasting method should be sensitive to the whole spectrum of economic and demographic factors which influence the direction and magnitude of population growth and also the extent of usage of telephone service. Unfortunately, no such comprehensive solution is yet in sight, although its achievement remains a desirable goal towards which to work. In the meantime, work has been done and the search continues for a worthwhile improvement over present methods.

Studies show that cumulative growth of a central office area over a

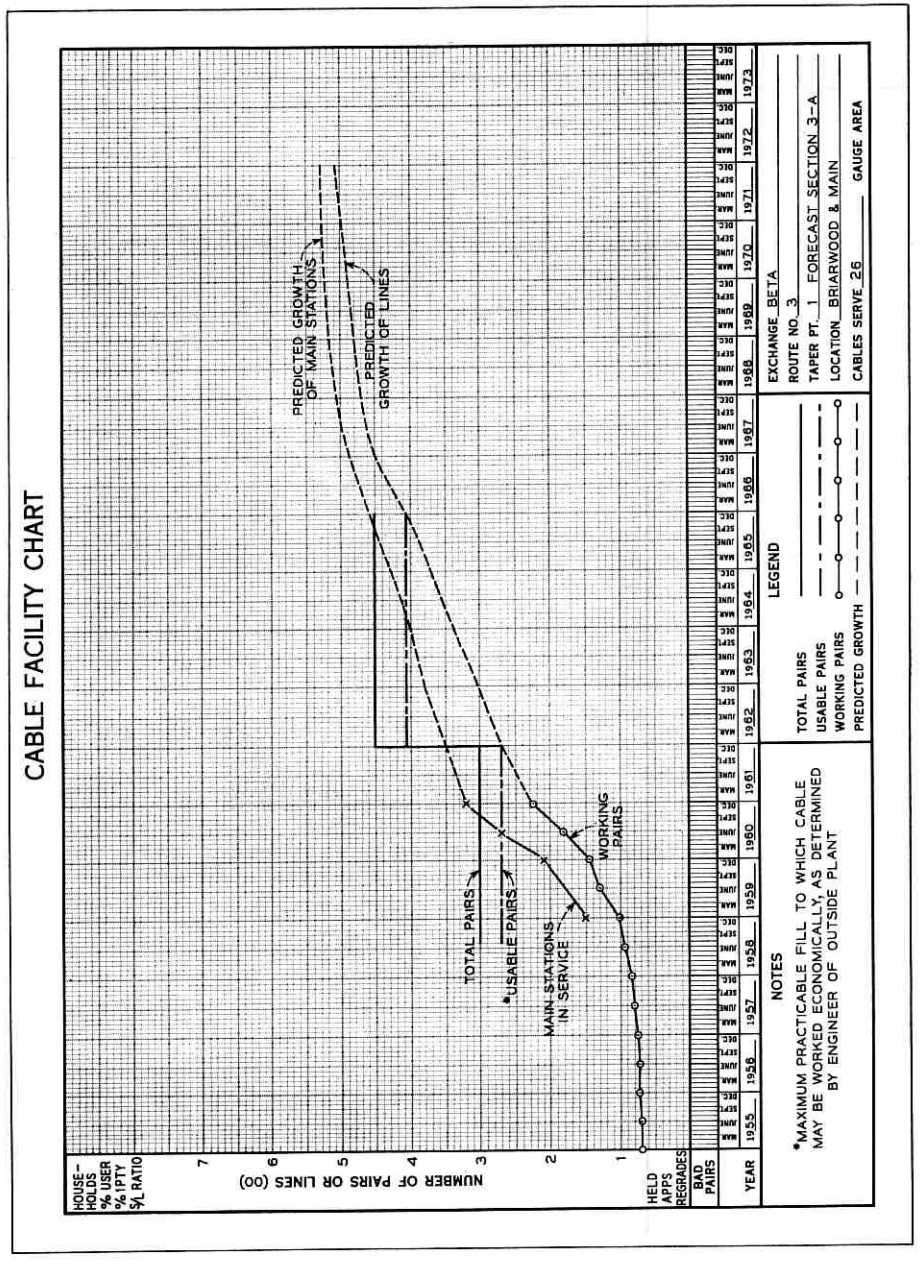

Fig. 15 — Cable facility chart.

period of years exhibits an "S" shape characteristic (Fig. 16). For example, an initial period of slow growth in undeveloped areas is followed by a transition into a period of sharply accelerated growth characteristic in the development of large tracts. Next, the growth tapers off (fill-in development takes place) as smaller developers working on scattered parcels of land tend to predominate. Finally, a terminal condition is reached during which little or no growth occurs, and even some decline may be experienced. At some point during this latter period, land usage may change and a new growth cycle begin in the form of land clearance, rehabilitation, or conversion to higher-density residential or commercial use.

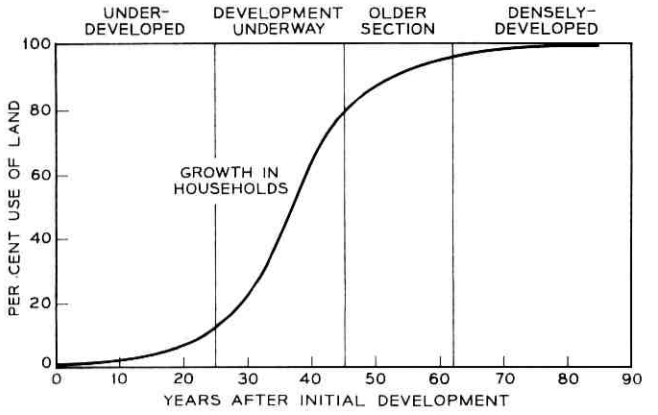

Fig. 16 — Urban land goes through a growth cycle.

The initial effort to improve forecasting was mainly directed toward capitalizing on the existence of these growth patterns. The technique proposed for growth prediction involved selection of a suitable mathematical expression which exhibits the same general "S" shape (so far the simple logistic function* has been used) and estimating the parameters of the function in a particular area from records of growth and the estimated level of the area's maximum development for the present growth cycle.

* Cumulative growth, $G = K/[1 + \exp(\alpha + \beta t)]$, of an area requires making estimates of the three parameters K, α, and β. K corresponds to the maximum development level of the area and can be estimated from knowledge of current and anticipated land usage. Values of the parameters α and β are estimated from the growth records by using the linear transformation, $\log_e[(K - G)/G] = \alpha + \beta t$. In this form a plot of cumulative growth as a function of time appears linear. The values of α and β can be derived by least square methods. This fitted function would then be used to predict future growth.

It should be pointed out that these procedures have proved most useful in growth areas having the following characteristics: boundaries which have remained relatively constant (or growth records which could be readily adjusted to reflect changes), availability of good historical information on households (or main telephones) and at least 20 per cent but not over 80 per cent of ultimate saturation realized. Realistic estimates of ultimate capacity based on sound business judgment and a careful analysis of basic assumptions and growth factors are of course the key to successful forecasts made by this method.

This technique, while of primary use in long-term projections, is also useful in medium- and short-term forecasting. For the latter a forecast may be derived by weighting current experience to reflect short-term trends and to place more emphasis on the more recent growth patterns. A computer program has been written which will allow the exponentially weighted forecast to be programmed along with the logistic function. This allows the weighting to be a function of the actual gain currently being experienced.

Work is planned on an important related factor: timing plant additions requiring a short-term forecast. Of necessity, short-term forecasts should project growth by months or quarters for at least the current year and preferably the following year. To accomplish this, a good short-term forecasting system must be sensitive to fine-grain fluctuations in demand around the long-term trend. However, the forecast cannot be made for an area and forgotten; adjustments are necessary from time to time to reflect new growth data and any changes that affect the saturation level.

As part of the short-term forecasting system, criteria need to be developed for determining whether actual growth falls reasonably close to expected demand, or whether deviations are large enough to warrant review of the forecast.

X. EXCHANGE AREA PLANNING — SUMMARY

A very important function of the engineer in the associated company is medium- and long-range planning of the exchange plant. He must allocate the company's resources in such a manner as to maintain a desirable relationship between cable network and central office equipment investments and also make future additions in each area as needed to meet service requests. To accomplish this task, the exchange feeder route analysis program, the exchange line multiplexing analysis program, the wire centering programs, and forecasting methods combined will aid the engineers and planners in establishing plans for exchange

areas as well as in administering the exchange plant. With these tools the associated company will be better able to estimate current and medium-term construction programs including manpower, materials, and money.

XI. FUTURE WORK

The complete implementation of these new concepts in the Bell System will require extensive training, coordination, and the working out of difficulties that will arise in any program of this magnitude. The people involved will have accomplished an Herculean task if the adoption is complete by the early seventies. The future extending past 1970 is extremely difficult to predict, except that many worthwhile innovations employing more sophisticated engineering skills and programming techniques will probably be superimposed on the concepts discussed in this paper. This is particularly true of some of the analytical techniques used, as in this first application methods were selected to insure that theoretical difficulties would be held to a minimum.

Bell Laboratories can use these same computer techniques in assessing the longer-term requirements for new laboratory developments by extrapolation of the data used by the associated companies for their day-to-day planning. With this capability, systems engineering studies can be completed more quickly and yield results more accurately reflecting future development needs of the Bell System.

XII. ACKNOWLEDGMENTS

The authors wish to give full credit for the contributions of members of the Bell Laboratories outside plant engineering department and the outside plant facilities, metropolitan planning and forecast and development groups at the American Telephone and Telegraph Company whose combined efforts made the writing of this article possible.

REFERENCES

1. Cochran, Wm. G., *Sampling Techniques*, 2nd ed., John Wiley and Sons, New York, 1963, p. 97.
2. Hinderliter, R. G., Transmission Characteristics of Bell System Subscriber Loop Plant, IEEE Trans. Comm. and Elect., Sept., 1963.
3. Amory, R. W., and Trachy, R. A., Computer Techniques Applied to Exchange Outside Plant Engineering, IEEE Trans. Paper No. 63-985.
4. Amory, R. W., Engineering Outside Plant With Computers, Bell Laboratories Record, July–Aug., 1963, pp. 258–266.

Epoch Detection — A Method For Resolving Overlapping Signals

By TZAY Y. YOUNG

(Manuscript received October 20, 1964)

The purpose of this paper is to discuss an epoch detection procedure which is very useful for the resolution and detection of signals overlapping in time. An epoch is the beginning instant of a signal. The epoch detection procedure is based on the following hypotheses: On the null hypothesis H_0 that a certain instant t is not an epoch, analytical continuation exists at t, and one may predict the signal in the future based on past experience or vice versa. On the hypothesis H_1 that t is an epoch, the analytic continuation is disrupted at t.

Based on this idea and the assumption of a Gaussian noise, a test statistic is derived from the maximum likelihood principle. The test statistic may be obtained at the output terminal of a linear filter. The performance of such a system is considered. Also discussed briefly are the cases of overlapping stochastic signals and overlapping radar signals. Some experimental results obtained from a digital computer are shown.

I. INTRODUCTION

Consider a signal composed of a train of overlapping wavelets.* The wavelets may, for one reason or another, arrive at the receiver (or measuring apparatus) delayed by different amounts of time. The time delays of the individual wavelets are unknown, but their differences may be relatively small so that the wavelets overlap. The beginning instant of each wavelet is called an epoch. These signals are corrupted with Gaussian noise. Our problem is to detect the overlapping in time. In other words, we wish to design a practical system which enables us to resolve the received signal train into overlapping wavelets and to describe them individually.

The theory of statistical detection of signals buried in noise has been well established.[1-4] In the field of resolving overlapping wavelets, Hel-

* We use the word "wavelets" for the individual overlapping wavelets, and reserve the word "signal" for the over-all signal train.

strom[5] discussed the optimum detection of two overlapping wavelets. With his assumption that one wavelet is separated from the other by known amount of time, the problem is considerably simplified and relatively easy to handle. Nilsson[6] discussed the problem of resolving N overlapping wavelets by deriving an equation to be maximized in an N-dimensional parameter space. Even in the case $N = 2$, the maximization of this equation is very complex and practically unsolved. Root[7] considered the general resolvability of radar signals, but gave no decision rule. Other studies related to signal resolution place most emphasis on the study of ambiguity functions[8,9] and on the design of a radar waveform which is inherently suitable for signal resolution.[10]

Generally speaking, for N overlapping wavelets, an optimum detection procedure would always involve searching for the maximum value of a likelihood function in an N-dimensional parameter space.[6] For N large this is hardly practical, and furthermore, if the number N is unknown, the problem becomes even more complicated. In a recent memorandum,[11] the author suggested an epoch detection procedure based on the properties of the signal at the epochs. The basic idea was to use a portion of the received signal in the past to predict the signal in the future, and to announce the arrival of a new wavelet if the prediction failed sufficiently badly. The present paper originated from that work. We intend to formalize and to develop the principle of epoch detection.

Consider a signal $f(t)$ consisting of two overlapping wavelets as shown in Fig. 1. The function $f(t)$ is analytic everywhere except at the two epochs t_1 and t_2. For any instant t which is not an epoch, it is possible to use the signal immediately prior to t to predict the signal immediately after. This is indeed the property of analytic continuation. However, at the two epochs, the statement is no longer true. Indeed we may define an epoch as an *instant* at which *analytic continuation is disrupted.*

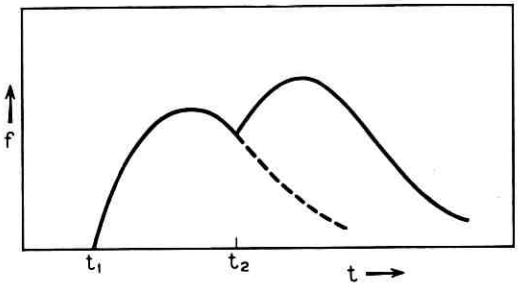

Fig. 1 — Overlapping signals.

It is precisely this disruption of analytic continuation that enables us to detect the epochs. In practice, we shall use the signal representation technique to describe such disruptions.

We make the assumption that any wavelets, though close enough to cause overlapping, are separated by at least T seconds, i.e.,

$$| t_j - t_k | \geqq T, \tag{1}$$

where t_j, t_k are any arbitrary epochs and T is a predetermined quantity. This assumption is necessary for our formulation, since for any instant t we shall utilize the information in the time interval $(t - T, t + T)$ to determine whether t is likely to be an epoch. We further assume that *any* $2T$-second segment of the individual *wavelets* is representable by a set of known component functions. Then the disruption of analytical continuation simply means that if an epoch exists in a certain $2T$ interval the *signal* in that interval is no longer representable by the set of component functions. Likelihood functions may be formulated in accordance with these criteria. The instant $\hat{t}$ which corresponds to a maximum value of a likelihood ratio is then the estimate of the epoch. This is of course the well-known maximum likelihood method of signal extraction, which has some theoretical advantages.[12,13] Other parameters of the wavelet may be estimated simultaneously.

Using the epoch detection scheme, we have in fact reduced an N-dimensional problem to N one-dimensional problems. Undoubtedly, in a process such as this, some information is lost, and one cannot expect optimum signal resolution except for some extreme cases. However, the simplicity and the practicality of the process justify our investigation. The process should be especially useful in the case of strong signals for which the advantage of a simple system outweighs that of optimality. In addition, the concept of epoch detection deserves to be studied and developed on its own right.

II. STATISTICAL EPOCH DETECTION

Let us denote the deterministic signal by $f_s(t)$, the random Gaussian noise by $f_n(t)$, and the noisy signal by $f_{s+n}(t)$. In this section, we shall consider the case that the deterministic signal consists of N overlapping wavelets with each wavelet being of the same waveform. Then we may write

$$\begin{aligned} f_{s+n}(t) &= f_s(t) + f_n(t) \\ &= \sum_{k=1}^{N} A_k f_w(t - t_k) + f_n(t), \end{aligned} \tag{2}$$

where A_k and t_k are the amplitude and the true epoch of the kth wavelet respectively. The function $f_w(\tau)$ represents the waveform of the individual wavelets.

To begin with, let us assume that each wavelet is representable by a set of known component functions. This assumption presents no theoretical difficulty since, by using a set of component functions that constitute a complete set, one may represent any continuous signals to any degree of accuracy.[14] However, practical considerations limit us to use a set of a finite number of component functions. We are particularly interested in the classes of component functions known as *generalized exponentials*, which include real and complex exponentials, sinusoids, polynomials and possible sums of products of such functions. The generalized exponentials have the following important property. A finite and properly-chosen set of generalized exponentials, as a set, goes into itself under the translation of time.[15] As a result, if a wavelet $f_w(\tau)$ is exactly representable by a properly chosen set of m generalized exponentials $\varphi^{(i)}(\tau)$, $i = 1, 2 \cdots m$, i.e.,

$$f_w(\tau) = \begin{cases} 0, & -\infty \leqq \tau < 0, \\ \sum_{i=1}^{m} c_i(0)\varphi^{(i)}(\tau), & 0 \leqq \tau \leqq \infty, \end{cases} \tag{3}$$

then the tail of $f_w(\tau)$ is also exactly representable by the same set of generalized exponentials,

$$f_w(t + \tau) = \sum_{i=1}^{m} c_i(t)\varphi^{(i)}(\tau), \qquad 0 \leqq t \leqq \infty, \quad 0 \leqq \tau \leqq \infty, \tag{4}$$

where $c_i(t)$ is the ith coefficient for a time translation of t seconds. Obviously, under this condition our earlier assumption that every $2T$ segment of the individual wavelets is representable by the set of component functions is fulfilled. The full significance of this property will be appreciated later, when we derive the test statistic for epoch detection.

The assumption of generalized exponentials is not as restrictive as it first appears. For one thing, most physical wavelets may be represented by a few terms of these functions. Furthermore, almost all commonly used functions for signal representation or curve fitting belong to the classes of generalized exponentials, and if we are willing to tolerate some inaccuracies by an approximate representation, practically all waveshapes may be represented by them. It is interesting to note that for the generalized exponentials $\varphi^{(i)}(\tau)$ analytic continuation exists

everywhere except at $\tau = 0$. Consequently, the epoch of a wavelet $f_w(\tau)$ described by (3) satisfies our earlier definition of disruption of analytic continuation.

Next, consider the Gaussian noise $f_n(t)$ having a covariance function $R(\tau)$.[16] The covariance function may be taken as the kernel of an integral equation,

$$\int_0^{2T} R(t - \tau)\psi^{(j)}(\tau)d\tau = \lambda_j\psi^{(j)}(\tau). \tag{5}$$

For our problem, $R(\tau)$ is real and symmetric, the eigenvalues, λ_j, are positive, and the eigenfunctions, $\psi^{(j)}(\tau)$, are orthonormal real functions. Both deterministic and random signals may be expressed in terms of these eigenfunctions.[16,17] Thus, we may write

$$\begin{aligned} f_{s+n}(t + \tau) &= \sum_j v_j(t)\psi^{(j)}(\tau), \\ f_s(t + \tau) &= \sum_j s_j(t)\psi^{(j)}(\tau), \\ f_n(t + \tau) &= \sum_j n_j(t)\psi^{(j)}(\tau), \\ & \qquad 0 \leqq \tau \leqq 2T, \end{aligned} \tag{6}$$

and

$$\varphi^{(i)}(\tau) = \sum_j u_{ij}\psi^{(j)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \tag{7}$$

with

$$\begin{aligned} v_j(t) &= \int_0^{2T} f_{s+n}(t + \tau)\psi^{(j)}(\tau)d\tau, \\ s_j(t) &= \int_0^{2T} f_s(t + \tau)\psi^{(j)}(\tau)d\tau, \\ n_j(t) &= \int_0^{2T} f_n(t + \tau)\psi^{(j)}(\tau)d\tau, \end{aligned} \tag{8}$$

and

$$u_{ij} = \int_0^{2T} \varphi^{(i)}(\tau)\psi^{(j)}(\tau)d\tau. \tag{9}$$

It is essential to note that by this expansion, the random variables n_j (and also v_j) are *independent variables* with variances λ_j. Since we are not interested in the singular case,[16] we assume that

$$\sum_{j=1}^{\infty} \frac{|\, s_j(t)\,|^2}{\lambda_j} < \infty,$$
$$\sum_{j=1}^{\infty} \frac{|\, u_{ij}\,|^2}{\lambda_j} < \infty. \tag{10}$$

We are now in a position to derive the test statistic for epoch detection. Let us start with the simplest case.

2.1 *Known Wavelet at Known Epoch*

In this case, we assume that there are reasons to believe that a wavelet in the form of $A_k f_w(t - t_k)$ may arrive. Both A_k and t_k are assumed known. In assuming known t_k, it is also implied that no epoch other than t_k may appear in the time interval $(t_k - T, t_k + T)$. Let us define a function

$$f_h(\tau) = \begin{cases} f_w(\tau - T), & T \leqq \tau \leqq 2T, \\ 0, & \text{elsewhere.} \end{cases} \tag{11}$$

We wish to test the hypothesis H_1 that the wavelet arrives against the null hypothesis H_0 that it does not. Thus, we write

$$H_0 : f_s(t_k - T + \tau) = \sum_i b_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \tag{12}$$

and

$$H_1 : f_s(t_k - T + \tau) = f_g(\tau) + \sum_i a_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \tag{13}$$

with $f_g(\tau)$ defined as

$$f_g(\tau) \equiv A_k f_h(\tau) - A_k \sum_i r_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T. \tag{14}$$

The constants r_i will be defined later. Let us explain these two hypotheses. In the first place, we notice that in using generalized exponentials as component functions, it is implied that $\varphi^{(i)}(\tau)$ and consequently $f_w(\tau)$ extends from $\tau = 0$ to $\tau = \infty$, as clearly indicated in (3). (The case of overlapping pulses will be treated later.) Therefore, on the null hypothesis H_0, although the new wavelet does not arrive, there will be tails of previously arrived wavelets appearing in the time interval $(t_k - T, t_k + T)$. Since every $2T$-second segment of these previously arrived wavelets is representable by the component functions $\varphi^{(i)}(\tau)$ with $0 \leqq \tau \leqq 2T$, we obtain (12) with the coefficients b_i to be estimated.

On the hypothesis H_1, the wavelet arrives at $t = t_k$. The term $A_k f_h(\tau)$

in (14) simply reflects this fact, since $f_h(\tau) = 0$ for $\tau < T$, as shown in (11). It is also this term that causes the disruption of analytic continuation at t_k. In addition to the kth wavelet, there are also tails of previously arrived wavelets, and we might have written for the hypothesis H_1, $f_s = A_k f_h + \sum q_i \varphi^{(i)}$. For reasons that will be pointed out later, we simply split q_i into two terms, $q_i = a_i - A_k r_i$, and obtain (13).

The random variables are independent when expressed in terms of eigenfunctions, and consequently we expand, similar to (6), $f_h(\tau)$ and $f_g(\tau)$ into

$$\begin{aligned} f_h(\tau) &= \sum_j h_j \psi^{(j)}(\tau), \quad 0 \leqq \tau \leqq 2T, \\ f_g(\tau) &= \sum_j g_j \psi^{(j)}(\tau), \quad 0 \leqq \tau \leqq 2T, \end{aligned} \tag{15}$$

with

$$\begin{aligned} h_j &= \int_0^{2T} f_h(\tau) \psi^{(j)}(\tau) d\tau, \\ g_j &= A_k h_j - A_k \sum_i r_i u_{ij}. \end{aligned} \tag{16}$$

The joint probability density for the null hypothesis may then be written as

$$P_0(v;b_i) = \frac{1}{\prod_j (2\pi\lambda_j)^{\frac{1}{2}}} \exp\left[- \sum_j \frac{(v_j - \sum_i b_i u_{ij})^2}{2\lambda_j} \right] \tag{17}$$

according to (6), (7), and (12). Similarly, we write for hypothesis H_1 the joint probability density

$$P_1(v;a_i) = \frac{1}{\prod_j (2\pi\lambda_j)^{\frac{1}{2}}} \exp\left[- \sum_j \frac{(v_j - \sum_i a_i u_{ij} - g_j)^2}{2\lambda_j} \right]. \tag{18}$$

In the absence of a priori information on the tails of previously arrived wavelets, a reasonable test is the maximum likelihood test which is given by

$$L = \frac{\max_a P_1(v;a)}{\max_b P_0(v;b)} \geqq \exp(\eta) \tag{19}$$

with the threshold η to be determined either by the Bayes criterion or by the Neyman-Pearson criterion. Equation (19) is equivalent to

$$\log L = \max_a \left[- \sum_j \frac{(v_j - \sum_i a_i u_{ij})^2 - 2(v_j - \sum_i a_i u_{ij}) g_j + g_j^2}{2\lambda_j} \right] - \max_b \left[- \sum_j \frac{(v_j - \sum_i b_i u_{ij})^2}{2\lambda_j} \right] \geqq \eta. \tag{20}$$

In order to simplify (20) somewhat, let us write

$$u_{ij}^* = \frac{u_{ij}}{\lambda_j}, \qquad \varphi^{(i)*}(\tau) = \sum_j u_{ij}^* \psi^{(j)}(\tau). \tag{21}$$

We assume that it is possible to write

$$\sum_j \frac{u_{lj} u_{ij}}{\lambda_j} = \delta_{li} \tag{22}$$

where δ_{li} is the Kronecker delta. Remembering the orthonormality of eigenfunctions, (22) may be written as

$$\begin{aligned} \sum_j \frac{u_{lj} u_{ij}}{\lambda_j} &= \sum_j u_{lj} u_{ij}^* \int_0^{2T} \psi^{(j)}(\tau) \psi^{(j)}(\tau)\, d\tau \\ &= \sum_{j,k} \int_0^{2T} u_{lk} \psi^{(k)}(\tau) u_{ij}^* \psi^{(j)}(\tau)\, d\tau \\ &= \int_0^{2T} \varphi^{(l)}(\tau) \varphi^{(i)*}(\tau)\, d\tau \\ &= \delta_{li}. \end{aligned} \tag{23}$$

Thus (22) is simply a consequence of the fact that $\varphi^{(l)}(\tau)$ and $\varphi^{(i)*}(\tau)$ form a biorthonormal system.[18] Furthermore,

$$\begin{aligned} \varphi^{(i)}(\mu) &= \sum_j u_{ij} \psi^{(j)}(\mu) \\ &= \sum_j \lambda_j u_{ij}^* \psi^{(j)}(\mu) \\ &= \sum_j u_{ij}^* \int_0^{2T} R(\mu - \tau) \psi^{(j)}(\tau)\, d\tau \\ &= \int_0^{2T} R(\mu - \tau) \varphi^{(i)*}(\tau)\, d\tau. \end{aligned} \tag{24}$$

Consequently, $\varphi^{(i)*}(\tau)$ is indeed the solution of an integral equation

and may be obtained for given $\varphi^{(i)}(\mu)$ and $R(\mu - \tau)$.[19] Thus it is always possible to achieve the biorthonormalization of a set of known component functions by means of a process similar to the Gram-Schmidt process of orthonormalization. It is appropriate to point out here that the assumption of a biorthonormal system is solely for the purpose of mathematical simplicity.

Now let us define the constant r_i as

$$r_i \equiv \sum_j \frac{u_{ij}}{\lambda_j} h_j . \tag{25}$$

Then, according to (16), we obtain

$$\begin{aligned} \sum_j \frac{u_{ij} g_j}{\lambda_j} &= A_k r_i - A_k \sum_{j,l} \frac{u_{ij} r_l u_{lj}}{\lambda_j} \\ &= A_k r_i - A_k \sum_l r_l \delta_{il} \\ &= 0. \end{aligned} \tag{26}$$

In other words, $f_a(\tau)$ is orthogonal to $\varphi^{(i)*}(\tau)$. Returning to (20), we notice that because of (26), $\log L$ may be simplified into the form of

$$\begin{aligned} \log L = \sum_j \frac{v_j g_j}{\lambda_j} - \sum_j \frac{g_j^2}{2\lambda_j} + \max_a \left[- \sum_j \frac{\left(v_j - \sum_i a_i u_{ij}\right)^2}{2\lambda_j} \right] \\ - \max_b \left[- \sum_j \frac{\left(v_j - \sum_i b_i u_{ij}\right)^2}{2\lambda_j} \right]. \end{aligned} \tag{27}$$

However, the last two terms are indeed identical. Thus,

$$\log L = \sum_j \frac{v_j g_j}{\lambda_j} - \sum_j \frac{g_j^2}{2\lambda_j} . \tag{28}$$

The last term in (28) is only a constant, and we may use the statistic

$$G = \sum_j \frac{v_j g_j}{\lambda_j} \gtrless \xi \tag{29}$$

for testing the arrival of the wavelet at the instant $t = t_k$. Here ξ is the threshold for testing G.

The results may also be expressed in the form of integral equations. Using a procedure similar to that used in (23) and (24), the statistic shown in (29) may be expressed as

$$G = \int_0^{2T} f_{s+n}(t_k - T + \tau) f_a^*(\tau)\, d\tau \gtrless \xi \tag{30}$$

with $f_0^*(\tau)$ being the solution of the integral equation

$$f_0(\mu) = \int_0^{2T} R(\mu - \tau) f_0^*(\tau)\, d\tau. \tag{31}$$

The function $f_0(\mu)$ has been defined in (14) and is rewritten here.

$$f_0(\mu) = A_k f_h(\mu) - A_k \sum_i r_i \varphi^{(i)}(\mu). \tag{14}$$

The constants r_i, written in integral form, become

$$r_i = \int_0^{2T} f_h(\tau) \varphi^{(i)*}(\tau)\, d\tau \tag{32}$$

according to (25). For a white noise with a covariance function $\delta(\tau)$, the results are considerably simpler since in this case,

$$\begin{aligned} \varphi^{(i)*}(\tau) &= \varphi^{(i)}(\tau), \\ f_0^*(\tau) &= f_0(\tau). \end{aligned} \tag{33}$$

The test statistic G shown in (30) may be obtained by a linear filter. If we use a linear filter whose weighting function is characterized by $f_0^*(\tau)$ — or, in other words, if the impulse response of the filter is $f_0^*(-\tau)$ — then with $f_{s+n}(t)$ as input, the output of the filter gives us the desired statistic G with a time delay of T seconds.[14] As examples, we show in Fig. 2 some wavelets and the weighting functions of their corresponding "matched" filters for epoch detection in white noise. (See Appendix.)

The weighting functions shown in the figure are calculated according to (14). It is essential to note the difference between our "matched" filter and the standard matched filter for the detection of non-overlapping signals. Without interfering signals, the matched filter would be $f_w(\tau)$, while in our case, a term in the form of $\sum r_i \varphi^{(i)}(\tau)$ is to be subtracted from the original waveform, as clearly shown in (14). It is indeed the subtraction of this term that enables us to suppress the effect of previously arrived wavelets. It is also this subtraction that represents the price we pay.

We wish to compute the false alarm and detection probabilities for the epoch detection system which is based on the statistic G. Since G is obtained from a linear operation on a Gaussian-distributed variable, G is also Gaussian-distributed.[17] Under the hypothesis H_0, its mean value is

$$E[G \mid H_0] = \int_0^{2T} f_0^*(\tau) \sum_i b_i \varphi^{(i)}(\tau)\, d\tau = 0 \tag{34}$$

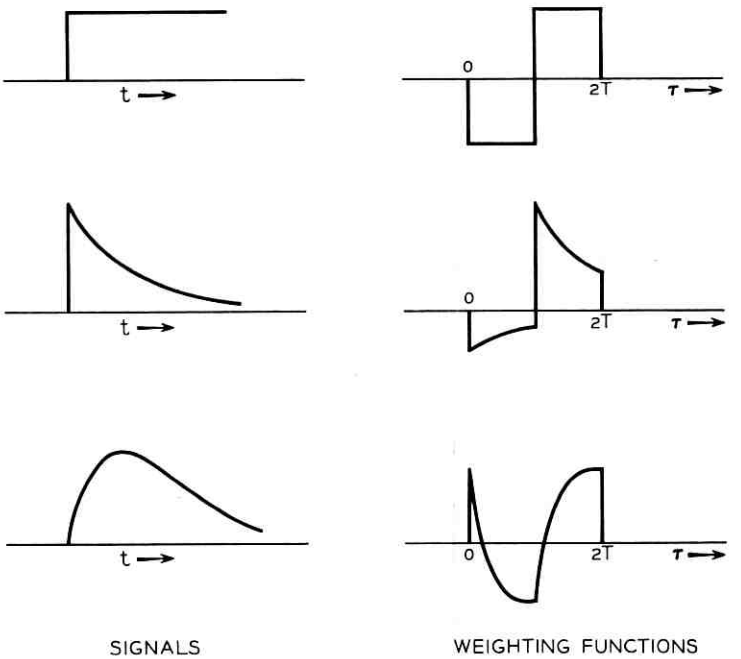

Fig. 2 — "Matched" filters for epoch detection in white noise.

since $f_g^*(\tau)$ and $\varphi^{(i)}(\tau)$ are orthogonal. Under the hypothesis H_1, the mean value becomes

$$\begin{aligned} E[G \mid H_1] &= \int_0^{2T} f_g^*(\tau)[f_g(\tau) + \sum_i a_i \varphi^{(i)}(\tau)]\, d\tau \\ &= \int_0^{2T} f_g^*(\tau) f_g(\tau)\, d\tau. \end{aligned} \tag{35}$$

The variance of G under either hypothesis is

$$\begin{aligned} \text{Var } G &= \int_0^{2T}\int_0^{2T} f_g^*(\tau) f_g^*(\mu) \overline{f_n(\tau) f_n(\mu)}\, d\mu d\tau \\ &= \int_0^{2T}\int_0^{2T} f_g^*(\tau) f_g^*(\mu) R(\tau - \mu)\, d\mu d\tau \\ &= \int_0^{2T} f_g^*(\tau) f_g(\tau) d\tau, \end{aligned} \tag{36}$$

where we have used (31). Thus,

$$d^2 = \int_0^{2T} f_g^*(\tau) f_g(\tau) d\tau, \tag{37}$$

a *dimensionless constant*, plays the role of signal-to-noise ratio (SNR).[3,4] The probability density functions of G are

$$\begin{aligned} p_0(G) &= (2\pi d^2)^{-\frac{1}{2}} \exp\ (-G^2/2d^2), \\ p_1(G) &= (2\pi d^2)^{-\frac{1}{2}} \exp\ [-(G - d^2)^2/2d^2]. \end{aligned} \tag{38}$$

and the false alarm and detection probabilities are, respectively[3]

$$\begin{aligned} Q_0 &= \text{erfc}\ (\xi/d), \\ Q_d &= \text{erfc}\left(\frac{\xi}{d} - d\right), \end{aligned} \tag{39}$$

where erfc (x) is the error-function integral.

2.2 *Unknown Amplitude and Unknown Epoch*

In this case, the two hypotheses become

$$H_0 : f_s(t - T + \tau) = \sum_i b_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \tag{40}$$

$$H_1 : f_s(t - T + \tau) = A(t) f_g(\tau) + \sum_i a_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \tag{41}$$

where

$$f_g(\tau) \equiv f_h(\tau) - \sum_i r_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T. \tag{42}$$

Notice the slight difference between the definition of $f_g(\tau)$ shown in (42) and that of (14). The joint probability densities are, similar to the previous case,

$$\begin{aligned} P_0(v;b_i) &= \frac{1}{\prod_j (2\pi\lambda_j)^{\frac{1}{2}}} \exp\left[- \sum_j \frac{(v_j - \sum_i b_i u_{ij})^2}{2\lambda_j} \right] \\ P_1(v;a_i, A) &= \frac{1}{\prod_j (2\pi\lambda_j)^{\frac{1}{2}}} \exp\left[- \sum_j \frac{(v_j - \sum_i a_i u_{ij} - A g_j)^2}{2\lambda_j} \right]. \end{aligned} \tag{43}$$

Using the principle of maximum likelihood estimation, we first make for each instant t an estimate of the amplitude, $\hat{A}(t)$, and then make

an estimate of the epoch, $\hat{t}$, which corresponds to the maximum value of the likelihood ratio of the hypothesis H_1 against the hypothesis H_0. Thus

$$L(\hat{A},\hat{t}) = \max_t L(\hat{A},t) = \max_t \left[\frac{\max_{a,A} P_1(v;a_i, A)}{\max_b P_0(v;b_i)} \right]. \tag{44}$$

By the same argument leading to (28), we obtain

$$\begin{aligned} \log L(\hat{A},\hat{t}) &= \max_t \log L(\hat{A},t) = \max_t [\max_A \log L(A,t)] \\ &= \max_{t,A} \left[\sum_j \frac{2v_jA(t)g_j - A^2(t)g_j^2}{2\lambda_j} \right]. \end{aligned} \tag{45}$$

Taking the partial derivative with respect to A,

$$\frac{\partial \log L(A,t)}{\partial A} = 0, \tag{46}$$

we get $\hat{A}(t)$, the maximum likelihood estimate of $A(t)$,

$$\hat{A}(t) = \left[\sum_j \frac{v_jg_j}{\lambda_j} \right] \Big/ \left[\sum_j \frac{g_j^2}{\lambda_j} \right]. \tag{47}$$

It should be noted that the random variable v_j is also a function of t, as shown in (8). Substituting (47) into (45) gives us

$$\log L(\hat{A},\hat{t}) = \max_t \left[\frac{1}{2} \hat{A}^2(t) \sum_j \frac{g_j^2}{\lambda_j} \right]. \tag{48}$$

If we normalize function $f_0(\tau)$ such that

$$\begin{aligned} \sum_j \frac{g_j^2}{\lambda_j} &= \int_0^{2T} f_0^*(\tau) f_0(\tau) d\tau \\ &= \int_0^{2T} \int_0^{2T} R(\tau - \mu) f_0(\mu) f_0(\tau) d\mu d\tau \\ &= 1, \end{aligned} \tag{49}$$

then

$$\hat{A}(t) = \sum_j \frac{v_jg_j}{\lambda_j} = \int_0^{2T} f_{s+n}(t - T + \tau) f_0^*(\tau) d\tau, \tag{50}$$

and

$$\log L(\hat{A},\hat{t}) = \max_t \log L(\hat{A}, t) = \max_t \frac{1}{2} |\hat{A}(t)|^2. \tag{51}$$

Thus the instant $\hat{t}$ that corresponds to the maximum value of $\log L(\hat{A},t)$ will be our estimate of the epoch. Indeed we may base our estimate on the maximum value of $|\hat{A}(t)|$. Equation (50) may of course be generated by means of a linear filter. If $f_k(\tau)$ and consequently $f_w(\tau)$ are properly "normalized" in the sense of (49) and (42), the signal-to-noise ratio (SNR) for the kth wavelet is A_k^2. Following the procedure used by Woodward and Davis,[20] it can be shown that, for the strong-signal case, the variance of the epoch estimate $\hat{t}$ is inversely proportional to SNR and to the square of the filter bandwidth.

We notice that the performance of an epoch detection system is related to the component functions only indirectly. It is the signal waveform itself that is important. As a rule of thumb, the smaller the absolute values of the constants r_i are, the more effective the system will be. In fact, we may define a useful figure of merit,

$$\rho \equiv \frac{\text{SNR for epoch detection}}{\text{SNR for the detection of } f_k(\tau)} = \frac{\int_0^{2T} f_0^*(\tau) f_0(\tau) d\tau}{\int_0^{2T} f_k^*(\tau) f_k(\tau) d\tau} \tag{52}$$

as the efficiency of the epoch detection system, where $f_k^*(\tau)$ is defined in the same way as we did for $f_0^*(\tau)$. Using (42) and (49) and the fact that $f_0^*(\tau)$ is orthogonal to $\varphi^{(i)}(\tau)$, we have

$$\rho = \frac{1}{1 + \sum_i r_i^2}. \tag{53}$$

As a result, $\rho < 1$. In the limit as every r_i approaches zero, $\rho \to 1$ and the epoch detection system approaches the optimum detection system for non-overlapping pulses of duration T seconds.

2.3 *Overlapping Pulses*

A pulse of duration T_0 seconds may be regarded as two overlapping wavelets with epochs separated by T_0 seconds. For instance, an exponential pulse, $\exp(-\tau)$ for $0 \leqq \tau \leqq T_0$, may be regarded as the sum of two exponential functions, $\exp(-\tau)$ with $0 \leqq \tau \leqq \infty$ and $-\exp(-\tau)$ with $T_0 \leqq \tau \leqq \infty$. We assume that several pulses may overlap. Thus, arrival of a pulse is characterized by the simultaneous existence of a wavelet $f_w^b(\tau)$ at the beginning epoch t_k^b and a wavelet

$f_w^{\,e}(\tau)$ at the ending epoch $t_k^{\,e}$ where $t_k^{\,e} = t_k^{\,b} + T_0$. For mathematical simplicity, we assume that the Gaussian noise in the time interval $(t_k^{\,b} - T, t_k^{\,b} + T)$ and the noise in the interval $(t_k^{\,e} - T, t_k^{\,e} + T)$ are uncorrelated. In other words, we assume

$$R(\tau) = 0 \quad \text{for} \quad \tau \geqq T_0 - 2T, \tag{54}$$

thus enabling us to treat independently the random variables in the two time intervals.

Again we formulate two hypotheses.

$$\begin{aligned} H_0 : f_s(t - T + \tau) &= \sum_i b_i^{\,b} \varphi^{(i)}(\tau), \\ f_s(t + T_0 - T + \tau) &= \sum_i b_i^{\,e} \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \end{aligned} \tag{55}$$

and

$$\begin{aligned} H_1 : f_s(t - T + \tau) &= A(t) f_g^{\,b}(\tau) + \sum_i a_i^{\,b} \varphi^{(i)}(\tau), \\ f_s(t + T_0 - T + \tau) &= A(t) f_g^{\,e}(\tau) + \sum_i a_i^{\,e} \varphi^{(i)}(\tau), \\ & \qquad 0 \leqq \tau \leqq 2T, \end{aligned} \tag{56}$$

where

$$\begin{aligned} f_g^{\,b}(\tau) &\equiv f_h^{\,b}(\tau) - \sum_i r_i^{\,b} \varphi^{(i)}(\tau), \\ f_g^{\,e}(\tau) &\equiv f_h^{\,e}(\tau) - \sum_i r_i^{\,e} \varphi^{(i)}(\tau), \\ & \qquad 0 \leqq \tau \leqq 2T, \end{aligned} \tag{57}$$

with $f_h^{\,b}(\tau)$ and $f_h^{\,e}(\tau)$ defined in the same way as $f_h(\tau)$, and the constants $r_i^{\,b}$ and $r_i^{\,e}$ defined in the same way as r_i. Let us define

$$f_g(\tau) \equiv f_g^{\,b}(\tau) + f_g^{\,e}(\tau - T_0). \tag{58}$$

Notice that $f_g^{\,e}(\tau - T_0) = 0$ for $\tau < T_0$. The function $f_g(\tau)$ is normalized in the sense that

$$\int_0^{T_0+2T} f_g^{\,*}(\tau) f_g(\tau) d\tau = 1, \tag{59}$$

with

$$f_g^{\,*}(\tau) = f_g^{\,b*}(\tau) + f_g^{\,e*}(\tau - T_0) \tag{60}$$

and

$$f_g^{\,b}(\mu) = \int_0^{2T} R(\mu - \tau) f_g^{\,b*}(\tau) d\tau,$$
$$f_g^{\,e}(\mu) = \int_0^{2T} R(\mu - \tau) f_g^{\,e*}(\tau) d\tau. \tag{61}$$

Equations (58) through (61) can be justified only on the assumption of (54).

Using the maximum likelihood principle, we write

$$L(\hat{A},\hat{t}) = \max_t \left[\frac{\max\limits_{a,A} P_1(v; a_i^{\,b}, a_i^{\,e}, A)}{\max\limits_{b} P_0(v; b_i^{\,b}, b_i^{\,e})} \right]. \tag{62}$$

With a derivation parallel to that of 2.2, we obtain the final results

$$\hat{A}(t) = \sum_j \frac{v_j g_j}{\lambda_j} = \int_0^{T_0+2T} f_{s+n}(t - T + \tau) f_g^{\,*}(\tau) d\tau, \tag{63}$$

and

$$\log L(\hat{A},\hat{t}) = \max_t \tfrac{1}{2} \mid \hat{A}(t) \mid^2. \tag{64}$$

Thus, based on the value of $\mid \hat{A}(t) \mid$, we may obtain the estimate of the epoch $\hat{t}$. Again a simple linear filter with a weighting function $f_g^{\,*}(\tau)$ defined in (60) will suffice to generate $\hat{A}(t)$.

III. OVERLAPPING STOCHASTIC SIGNALS

Again we consider a train of overlapping wavelets corrupted with a Gaussian noise. Each wavelet is assumed to be representable by a set of m known generalized exponential functions. However, the wavelets are stochastic in the sense that their exact waveforms are unknown and that each wavelet may differ from the other. As a result, the two hypotheses become

$$H_0 : f_s(t - T + \tau) = \sum_i b_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \tag{65}$$

$$H_1 : f_s(t - T + \tau) = \sum_i c_i(t) \chi^{(i)}(\tau) + \sum_i a_i \varphi^{(i)}(\tau), \qquad 0 \leqq \tau \leqq 2T, \tag{66}$$

where

$$\chi^{(i)}(\tau) = \sum_l \beta_{il} \varphi^{(l)}(\tau - T) - \sum_l \gamma_{il} \varphi^{(l)}(\tau) \tag{67}$$

with

$$\varphi^{(l)}(\tau - T) = 0 \quad \text{for} \quad \tau < T. \tag{68}$$

The constants β_{il} and γ_{il} are constrained by the biorthonormality relationships. Let $\psi^{(j)}(\tau)$ be the eigenfunctions of the covariance function $R(\tau)$. For the sake of the independence of random variables, we expand the component functions, the noise, etc., in terms of $\psi^{(j)}(\tau)$. For $\chi^{(i)}(\tau)$, we then write

$$\chi^{(i)}(\tau) = \sum_j x_{ij}\psi^{(j)}(\tau), \qquad 0 \leqq \tau \leqq 2T. \tag{69}$$

The constraints of biorthonormality are

$$\sum_j \frac{u_{ij}u_{lj}}{\lambda_j} = \delta_{il}\,, \tag{70}$$

$$\sum_j \frac{x_{ij}u_{lj}}{\lambda_j} = 0, \tag{71}$$

and

$$\sum_j \frac{x_{ij}x_{lj}}{\lambda_j} = \delta_{il}\,. \tag{72}$$

A direct result of (70) and (71) is, similar to (25) and (32),

$$\begin{aligned}\gamma_{il} &= \sum_{j,k} \frac{u_{lj}}{\lambda_j}\beta_{ik}\int_0^{2T} \varphi^{(k)}(\tau - T)\psi^{(j)}(\tau)d\tau \\ &= \sum_k \beta_{ik}\int_0^{2T} \varphi^{(k)}(\tau - T)\varphi^{(l)*}(\tau)d\tau.\end{aligned} \tag{73}$$

To illustrate the procedure for formulating $\chi^{(i)}(\tau)$, let us consider the simple case of white noise for which the biorthonormality reduces to orthonormality. The first step is of course to orthonormalize with respect to the time interval $(0,2T)$ the m component functions by means of the Gram-Schmidt procedure. Next we may choose any value of β_{il} for (67) as long as the m functions $\sum \beta_{il}\varphi^{(l)}(\tau - T)$, $i = 1, 2 \cdots m$ are linearly independent. Using (73) for the calculation of γ_{il} guarantees that $\chi^{(i)}(\tau)$ is orthogonal to $\varphi^{(l)}(\tau)$. Finally, by means of the Gram-Schmidt process, we may combine the functions $\chi^{(i)}(\tau)$ linearly to make them orthonormal. In this way, all three conditions, (70), (71) and (72), are satisfied. For colored noise, the procedure is similar.

Under these assumptions of biorthonormality, an application of maximum likelihood principle then gives us

$$\hat{c}_i(t) = \sum_j \frac{v_j x_{ij}}{\lambda_j} = \int_0^{2T} f_{s+n}(t - T + \tau)\chi^{(i)*}(\tau)d\tau, \tag{74}$$

where $\chi^{(i)*}(\tau)$ is the solution of the equation

$$\chi^{(i)}(\mu) = \int_0^{2T} R(\mu - \tau)\chi^{(i)*}(\tau)d\tau, \tag{75}$$

and

$$\log L(\hat{c}_i, \hat{t}) = \max_t \log (\hat{c}_i, t) = \max_t \left[\frac{1}{2}\sum_i |\hat{c}_i(t)|^2\right]. \tag{76}$$

The estimated epoch $\hat{t}$ is the instant that corresponds to the maximum value of $\log L(\hat{c}_i, t)$. The epoch detection system which generates $\log L(\hat{c}_i, t)$ will then consist of a summing amplifier, m squarers and m linear filters characterized by the m weighting functions $\chi^{(i)*}(\tau)$.

For stochastic signals, a proper definition of SNR for the kth wavelet is

$$\begin{aligned} d_k^2 &\equiv \frac{1}{m}\sum_{i=1}^m c_i^2(t_k)\int_0^{2T}\chi^{(i)*}(\tau)\chi^{(i)}(\tau)d\tau \\ &= \frac{1}{m}\sum_{i=1}^m c_i^2(t_k), \end{aligned} \tag{77}$$

where we have used (72). The coefficient, $c_i(t_k)$, is defined by (66) with t_k, the kth epoch, substituted for t.

Let us now show some experimental results obtained from a digital computer. Fig. 3 illustrates the detection of overlapping wavelets, each consisting of two exponentials, $e^{-\tau}$ and $e^{-2\tau}$. Although in our experiment the three overlapping wavelets have the same waveshape, they are regarded as *stochastic* since we do not assume the a priori knowledge of the proportion of the two exponentials that constitute the wavelets. The signals are additively corrupted with white noise as shown in the second row. With the definition of (77), the signal-to-noise ratios for our examples are, in decibels, ∞, 15 and 8, respectively. Since we know the component functions, $e^{-\tau}$ and $e^{-2\tau}$, what we need to do would be simply to estimate the coefficients $\hat{c}_i(t)$ by means of linear filters prescribed by (67) and then calculate $\log L(\hat{c}_i, \hat{t})$ according to (76). What we actually did is based on a more primitive model;[11] nevertheless, the basic philosophy is the same. Using this primitive model, the logarithms of the likelihood ratios, $\log L_f(t)$, are calculated over the noisy signals, and shown in the third row. It is clear from the figure that the estimated epochs $\hat{t}$ which correspond to the maximum value of $\log L_f(t)$, almost coincide with the true epochs. However, for the 8-db

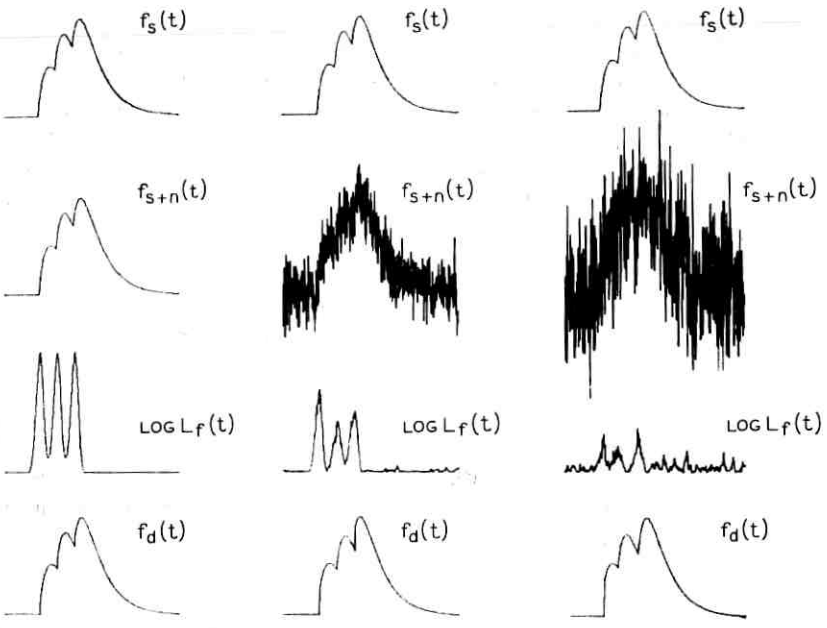

Fig. 3 — Detection of overlapping wavelets.

case, the detection probability is low, and the peak corresponding to the second epoch is almost not distinguishable from the peaks which are due to the random noise alone. With the epochs estimated, one may estimate in a piecewise manner the signal between the epochs, and the signals $f_d(t)$ thus detected are shown in the last row.

IV. OVERLAPPING RADAR SIGNALS

The most important application of statistical detection theory is in radar signal detection. We shall consider typical radar pulses which are sinusoidal signals modulated by square waves. Each pulse, as discussed in Section II, can be characterized by a beginning epoch and an ending epoch. For simplicity, we treat them separately as two epochs.

For overlapping radar signals, we may write

$$f_s(t) \cos (\omega_c t + \theta) = \sum_{k=1}^{N} A_k f_w(t - t_k) \cos (\omega_c t + \theta_k), \quad (78)$$

where we have regarded $f_s(t)$ and $f_w(\tau)$ as envelopes of the sinusoidal signals and θ_k are phase angles. The pulse envelope $f_w(\tau)$ is a square wave. Similarly we consider $f_{s+n}(t)$ as the envelope of the noisy signal.

The epoch detection system developed previously is difficult to implement in this case because it is sensitive to radio-frequency phase. For this reason, we assume the use of a perfect envelope detector to take the signal envelope first.[17] The noise under this condition becomes narrow-band noise. Let us designate, for a certain instant t, the envelope of the sinusoidal signal by α and the envelope of the noisy signal by v. With a variance λ, the probability density for the envelope at time t on the hypothesis that the sampled waveform is sine-wave plus noise is, according to the classic work of Rice,[21]

$$p(v,\alpha) = \begin{cases} \frac{v}{\lambda} \exp\left(-\frac{v^2+\alpha^2}{2\lambda}\right) I_0\left(\frac{\alpha v}{\lambda}\right), & v \geqq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{79}$$

where $I_0(x)$ is the modified Bessel function of the first kind and order zero. It is known as the modified Rayleigh distribution or the Rice distribution. On the hypothesis that the sampled waveform is noise alone, $\alpha = 0$ and $I_0(0) = 1$, and (79) is reduced to the Rayleigh distribution.

$$p_0(v) = \begin{cases} \frac{v}{\lambda} \exp\left(-\frac{v^2}{2\lambda}\right) & v \geqq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{80}$$

We again formulate a null hypothesis H_0 and a hypothesis H_1 that an epoch has arrived. Thus,

$$H_0 : f_s(t - T + \tau) = \alpha_0, \qquad 0 \leqq \tau \leqq 2T. \tag{81}$$

$$H_1 : f_s(t - T + \tau) = \begin{cases} \alpha_1, & 0 \leqq \tau \leqq T, \\ \alpha_2, & T \leqq \tau \leqq 2T. \end{cases} \tag{82}$$

We may look for a coordinate system such that the random variables on these coordinates are statistically independent. However, unlike the Gaussian distribution, it is very difficult to find such a coordinate system. The usual procedure, which we shall follow here, is to use for coordinates samples of the envelope waveform taken at regular intervals and far enough apart so that it is a reasonable approximation to suppose them statistically independent. We take in the time interval $(t - T, t + T)$ $2M$ measurements at $2M$ uniformly spaced instants separated by $\Delta\tau$ seconds apart. Let us write for the instant t

$$v_j(t) = f_{s+n}(t - T + j\Delta\tau). \tag{83}$$

Then on the null hypothesis H_0, the joint probability density is

$$P_0(v;\alpha_0) = \prod_{j=1}^{2M} \frac{v_j}{\lambda} \exp\left(-\frac{v_j^2 + \alpha_0^2}{2\lambda}\right) I_0\left(\frac{\alpha_0 v_j}{\lambda}\right), \tag{84}$$

while on the hypothesis H_1, we have

$$\begin{aligned} P_1(v;\alpha_1, \alpha_2) = &\prod_{j=1}^{M} \frac{v_j}{\lambda} \exp\left(-\frac{v_j^2 + \alpha_1^2}{2\lambda}\right) I_0\left(\frac{\alpha_1 v_j}{\lambda}\right) \\ &+ \prod_{j=M+1}^{2M} \frac{v_j}{\lambda} \exp\left(-\frac{v_j^2 + \alpha_2^2}{2\lambda}\right) I_0\left(\frac{\alpha_2 v_j}{\lambda}\right). \end{aligned} \tag{85}$$

The maximum likelihood principle then requires

$$L(\hat{\alpha}, \hat{t}) = \max_t \left[\frac{\max_\alpha P_1(v;\alpha_1, \alpha_2)}{\max_\alpha P_0(v;\alpha_0)}\right]. \tag{86}$$

By considering the logarithm of the likelihood functions and taking partial derivatives, we can easily show that

$$\begin{aligned} \sum_{j=1}^{M} \left[-\frac{\hat{\alpha}_1}{\lambda} + \frac{v_j}{\lambda} \frac{I_0'\left(\frac{\hat{\alpha}_1 v_j}{\lambda}\right)}{I_0\left(\frac{\hat{\alpha}_1 v_j}{\lambda}\right)}\right] &= 0, \\ \sum_{j=M+1}^{2M} \left[-\frac{\hat{\alpha}_2}{\lambda} + \frac{v_j}{\lambda} \frac{I_0'\left(\frac{\hat{\alpha}_2 v_j}{\lambda}\right)}{I_0\left(\frac{\hat{\alpha}_2 v_j}{\lambda}\right)}\right] &= 0, \end{aligned} \tag{87}$$

and

$$\sum_{j=1}^{2M} \left[-\frac{\hat{\alpha}_0}{\lambda} + \frac{v_j}{\lambda} \frac{I_0'\left(\frac{\hat{\alpha}_0 v_j}{\lambda}\right)}{I_0\left(\frac{\hat{\alpha}_0 v_j}{\lambda}\right)}\right] = 0, \tag{88}$$

where $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_0$ are the maximum likelihood estimates of α_1, α_2, α_0 respectively, and I_0' is the derivative of I_0. The estimated epoch then corresponds to

$$\begin{aligned} \log L(\hat{\alpha}, \hat{t}) = \max_t \Bigg\{ &\sum_{j=1}^{M} \left[-\frac{\hat{\alpha}_1^2}{2\lambda} + \log I_0\left(\frac{\hat{\alpha}_1 v_j}{\lambda}\right)\right] \\ &+ \sum_{j=M+1}^{2M} \left[-\frac{\hat{\alpha}_2^2}{2\lambda} + \log I_0\left(\frac{\hat{\alpha}_2 v_j}{\lambda}\right)\right] \\ &- \sum_{j=1}^{2M} \left[-\frac{\hat{\alpha}_0^2}{2\lambda} + \log I_0\left(\frac{\hat{\alpha}_0 v_j}{\lambda}\right)\right]\Bigg\}. \end{aligned} \tag{89}$$

It should be noted that $\hat{\alpha}$ and v_j are functions of t. Our problem would have been solved if we had been able to solve (87) and (88) for $\hat{\alpha}$, substitute them into (89) and then search for $\hat{t}$ that corresponds to the maximum value of $\log L(\hat{\alpha},t)$. An explicit solution in this case is, to say the least, very difficult. Certain approximations are needed. We shall discuss the case of strong signals and the case of weak signals separately.

In the first place, we notice that if the signal-to-noise ratio is sufficiently high — i.e., $\alpha_1^2/2\lambda \gg 1$ and $\alpha_2^2/2\lambda \gg 1$ [$(\alpha_1^2 - \alpha_2^2)/2\lambda$ may be small] — the Rice distribution approaches the Gaussian distribution and the discussion in Section II is directly applicable. A linear filter may thus be used. On the other hand, if $\alpha_1^2/2\lambda \gg 1$ and $\alpha_2^2/2\lambda$ small or vice versa, we encounter the epoch of a large pulse. Therefore a suboptimal epoch detection scheme may be used, and again a linear filter may be chosen for its simplicity.

Finally, consider the case that $\alpha_1^2/2\lambda \ll 1$ and $\alpha_2^2/2\lambda \ll 1$. It is well-known that for small x (see Ref. 22),

$$\begin{aligned} I_0(x) &= 1 + \left(\frac{1}{2}x\right)^2 + \frac{1}{1^2 \cdot 2^2}\left(\frac{1}{2}x\right)^4 + \cdots \\ \log I_0(x) &= \left(\frac{1}{2}x\right)^2 - \frac{1}{4}\left(\frac{1}{2}x\right)^4 + \cdots. \end{aligned} \tag{90}$$

If we substitute (90) into (87) and (88) and retain only those terms that involve $\hat{\alpha}/\sqrt{\lambda}$ and $(\hat{\alpha}/\sqrt{\lambda})^3$, we obtain as approximations

$$\begin{aligned} \frac{\hat{\alpha}_1^2}{2\lambda} &\approx \sum_{j=1}^{M}\left[\frac{v_j^2}{2\lambda} - 1\right] \Big/ \sum_{j=1}^{M}\frac{v_j^4}{8\lambda^2}, \\ \frac{\hat{\alpha}_2^2}{2\lambda} &\approx \sum_{j=M+1}^{2M}\left[\frac{v_j^2}{2\lambda} - 1\right] \Big/ \sum_{j=M+1}^{2M}\frac{v_j^4}{8\lambda^2}, \end{aligned} \tag{91}$$

and

$$\frac{\hat{\alpha}_0^2}{2\lambda} \approx \sum_{j=1}^{2M}\left[\frac{v_j^2}{2\lambda} - 1\right] \Big/ \sum_{j=1}^{2M}\frac{v_j^4}{8\lambda^2}. \tag{92}$$

Similarly, substituting (90) into (89) gives us

$$\begin{aligned} \log L(\hat{\alpha},\hat{t}) = \max_t \Bigg\{ &\frac{\hat{\alpha}_1^2}{2\lambda}\sum_{j=1}^{M}\left[\frac{v_j^2}{2\lambda} - 1 - \frac{1}{2}\frac{\hat{\alpha}_1^2}{2\lambda}\frac{v_j^4}{8\lambda^2}\right] \\ &+ \frac{\hat{\alpha}_2^2}{2\lambda}\sum_{j=M+1}^{2M}\left[\frac{v_j^2}{2\lambda} - 1 - \frac{1}{2}\frac{\hat{\alpha}_2^2}{2\lambda}\frac{v_j^4}{8\lambda^2}\right] \\ &- \frac{\hat{\alpha}_0^2}{2\lambda}\sum_{j=1}^{2M}\left[\frac{v_j^2}{2\lambda} - 1 - \frac{1}{2}\frac{\hat{\alpha}_0^2}{2\lambda}\frac{v_j^4}{8\lambda^2}\right]\Bigg\}. \end{aligned} \tag{93}$$

An epoch detection system based on (91), (92), and (93) is very difficult to implement. It needs further simplification. We notice that, for M sufficiently large, the random variables take on nearly all possible values of the probability functions. The summations then represent an ensemble average. For example, with $\alpha_1^2/2\lambda \ll 1$,

$$\begin{aligned}\sum_{j=1}^{M} \frac{v_j^2}{2\lambda} &\approx M\,\overline{\left(\frac{v_j^2}{2\lambda}\right)} = M\int_0^\infty \frac{x^2}{2\lambda}\, p(x,\alpha_1)dx \approx M\left(1+\frac{\alpha_1^2}{2\lambda}\right),\\ \sum_{j=1}^{M} \frac{v_j^4}{8\lambda^2} &\approx M\,\overline{\left(\frac{v_j^4}{8\lambda^2}\right)} = M\int_0^\infty \frac{x^4}{8\lambda^2}\, p(x,\alpha_1)dx \approx M\left(1+\frac{\alpha_1^2}{\lambda}\right),\end{aligned} \tag{94}$$

where $p(x,\alpha_1)$ is the Rice distribution shown in (79). We may of course do the same for $\alpha_2^2/2\lambda$ and $\alpha_0^2/2\lambda$. It is indeed from this consideration that we include the term

$$\frac{1}{2}\frac{\hat{\alpha}^2}{2\lambda}\frac{v_j^4}{8\lambda^2}$$

in (93), since it is of the same order as the difference of the remaining two terms.

From the consideration of order of magnitude, it should be obvious that for (91) and (92) the denominators may be replaced by M and $2M$ respectively. As a result, we may write

$$\frac{\hat{\alpha}_0^2}{2\lambda} = \frac{1}{2}\left(\frac{\hat{\alpha}_1^2}{2\lambda}+\frac{\hat{\alpha}_2^2}{2\lambda}\right). \tag{95}$$

Using (91), (92), (94), and (95) for (93) and simplifying, we finally obtain

$$\begin{aligned}\log L(\hat{\alpha},\hat{t}) &= \max_t\left[\frac{M}{2}\left(\frac{\hat{\alpha}_1^2}{2\lambda}\right)^2+\frac{M}{2}\left(\frac{\hat{\alpha}_2^2}{2\lambda}\right)^2-M\left(\frac{\hat{\alpha}_0^2}{2\lambda}\right)^2\right]\\ &= \max_t\frac{M}{4}\left|\frac{\hat{\alpha}_2^2}{2\lambda}-\frac{\hat{\alpha}_1^2}{2\lambda}\right|^2\\ &= \max_t\frac{1}{4M}\left|\sum_{j=M+1}^{2M}\frac{v_j^2(t)}{2\lambda}-\sum_{j=1}^{M}\frac{v_j^2(t)}{2\lambda}\right|^2.\end{aligned} \tag{96}$$

A test may naturally be based on the quantity inside the absolute sign. A large positive value for the quantity indicates the arrival of a beginning epoch, and a large negative value corresponds to an ending epoch. A pulse is of course marked by the arrival of both epochs. The result is consistent with the conventional square-law detector for small, nonoverlapping signals.

V. CONCLUSION

We have investigated the problem of epoch detection. A test statistic, which may be obtained from a simple, linear filter, has been derived for Gaussian noise. In the derivation, we have assumed that each wavelet is representable by a set of known generalized exponentials. This is not as restrictive as it appears, considering the fact that any continuous signal may be represented with a least-square error as small as we wish by using a sufficient number of component functions.

The epoch detection scheme is particularly useful for the resolution and detection of overlapping signals. For N overlapping wavelets, the procedure reduces the resolution problem from an N-dimensional problem to N one-dimensional problems. Some information is lost in this reduction, and consequently it is not a scheme for optimal resolution. However, it has the essential advantage of simplicity and practicality.

The performance of the epoch detection system has been considered briefly. The discussions of overlapping stochastic signals and overlapping radar signals show that the method is applicable to these cases, and the experimental results enhance our confidence in the detection procedure.

VI. ACKNOWLEDGMENTS

The author wishes to express his appreciation to W. H. Huggins of the Johns Hopkins University for many helpful suggestions. He is also grateful to J. F. Kaiser of Bell Telephone Laboratories for his comments on the manuscript. Part of the preliminary work was done while the author was with the Carlyle Barton Laboratory, the Johns Hopkins University, Baltimore, Maryland, under a contract supported by the U. S. Air Force.

APPENDIX

The weighting functions of the "matched" filters are calculated according to the equation

$$f_g(\tau) = f_h(\tau) - \sum_i r_i \varphi^{(i)}(\tau),$$

where r_i's are chosen in such a way that for white noise $f_g(\tau)$ is orthogonal to every $\varphi^{(i)}(\tau)$. This orthogonality (or biorthogonality for Gaussian noise) is the central idea of epoch detection, and has been discussed in the paper.

As an example, for the last signal in Fig. 2, the signal (or wavelet) is of the form

$$f_w(\tau) = e^{-\tau} - e^{-2\tau},$$

and therefore from (11)

$$f_h(\tau) = \begin{cases} e^{-(\tau-T)} - e^{-2(\tau-T)}, & T \leqq \tau \leqq 2T, \\ 0, & \text{elsewhere.} \end{cases}$$

Thus we have

$$f_g(\tau) = \begin{cases} -r_1 e^{-\tau} - r_2 e^{-2\tau}, & 0 \leqq \tau < T, \\ (e^T - r_1)e^{-\tau} - (e^{2T} + r_2)e^{-2\tau}, & T \leqq \tau \leqq 2T, \end{cases}$$

with r_1 and r_2 to be determined by the orthogonality relationships

$$\int_0^{2T} f_g(\tau)e^{-\tau}d\tau = 0,$$

$$\int_0^{2T} f_g(\tau)e^{-2\tau}d\tau = 0,$$

In our example, $T = 0.7$, and then the solution of the above two equations is

$$r_1 = 0.62 \quad \text{and} \quad r_2 = -0.76.$$

A substitution of these values into the equation of $f_g(\tau)$ results in

$$f_g(\tau) = \begin{cases} -0.62e^{-\tau} + 0.76e^{-2\tau}, & 0 \leqq \tau < 0.7, \\ 1.39e^{-\tau} - 3.30e^{-2\tau}, & 0.7 \leqq \tau \leqq 1.4, \end{cases}$$

which, except for a scale factor, is the weighting function shown in Fig. 2.

REFERENCES

1. Woodward, P. M., *Probability and Information Theory with Applications to Radar*, Pergamon Press, New York, 1953.
2. Middleton, D., *An Introduction to Statistical Communication Theory*, McGraw-Hill, New York, 1960.
3. Helstrom, C. W., *Statistical Theory of Signal Detection*, Pergamon Press, New York, 1960.
4. Wainstein, L. A., and Zubakov, V. D., *Extraction of Signals from Noise*, Prentice-Hall, Englewood Cliffs, N. J., 1962.
5. Helstrom, C. W., The Resolution of Signals in White Gaussian Noise, Proc. IRE, *43*, Sept., 1955, p. 1111.
6. Nilsson, N. J., On the Optimum Range Resolution of Radar Signals in Noise, IRE Trans. Information Theory, *IT-7*, Oct., 1961, p. 245.
7. Root, W. L., Radar resolution of closely spaced targets, IRE Trans. Military Electronics, *MIL-6*, April, 1962, p. 197.

8. Sussman, S. M., Least-Square Synthesis of Radar Ambiquity Functions, IRE Trans. Information Theory, *IT-8*, April, 1962, p. 246.
9. Rihaczak, A. W., Radar Resolution Properties of Pulse Trains, Proc. IEEE, *52*, Feb., 1964, p. 153.
10. Klander, J. R., Price, A. C., Darlington, S., and Albersheim, W. L., The Theory and Design of Chirp Radars, B.S.T.J., *39*, July, 1960, p. 745.
11. Young, T. Y., Statistical Epoch Detection of Overlapping Signals, unpublished memorandum, Carlyle Barton Laboratory, Johns Hopkins University, 1963.
12. Slepian, D., Estimation of Signal Parameters in the Presence of Noise, IRE Trans. Information Theory, *PGIT-3*, Mar., 1954, p. 68.
13. Middleton, D., and Van Meter, D., Detection and Extraction of Signals in Noise from the Point of View of Statistical Decision Theory, J. Soc. Indust. Appl. Math., *3*, Dec., 1955, p. 192; *4*, June, 1956, p. 86.
14. Huggins, W. H., Signal Theory, IRE Trans. Circuit Theory, *CT-3*, Dec., 1956, p. 210.
15. Ule, L. A., Weighted Least-Square Smoothing Filters, IRE Trans. Circuit Theory, *CT-2*, June, 1955, p. 197.
16. Kelly, E. J., Reed, I. S., and Root, W. L., The Detection of Radar Echoes in Noise, J. Soc. Indust. Appl. Math., *8*, June, 1960, p. 309; Sept., 1960, p. 481.
17. Davenport, W. B., Jr., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, New York, 1958.
18. Akhiezer, N. I., *Theory of Approximation*, Frederick-Ungar, New York, 1956.
19. Zadeh, L. A., and Ragazzini, J. R., Optimum Filters for the Detection of Signals in Noise, Proc. IRE, *40*, Oct., 1952, p. 1123.
20. Woodward, P. M., and Davis, I. L., A Theory of Radar Information, Phil. Mag., *41*, 1950, p. 1001.
21. Rice, S. O., Mathematical Analysis of Random Noise, B.S.T.J., *23*, 1944, p. 282; *24*, 1945, p. 46.
22. Dwight, H. B., *Tables of Integrals and Other Mathematical Data*, Macmillan, New York, 1957.

Computation of Lattice Sums: Generalization of the Ewald Method

By W. J. C. GRANT

(Manuscript received October 26, 1964)

The Ewald method was originally invented to compute the Madelung constant. In this paper we consider a lattice whose sites are associated with an arbitrary potential function. The "charge," or the scale factor for these potential functions, need not be the same at each site. We consider the evaluation of the resulting lattice sum at an arbitrary point, not necessarily at a lattice site. The method involves two generalizations over previous work: (1) the displacement of the origin off a lattice site and (2) the handling of arbitrary periodic charge distributions by decomposing such distributions into simpler ones involving only $+q$ and $-q$. The method should prove particularly useful for evaluating the expansion coefficients of the crystalline potential when this potential is expanded in the usual spherical harmonic series.

The problem of summing slowly converging series is an old one. One physical context in which the problem has been widely studied is the calculation of the potential due to an ionic crystal lattice. The methods of Madelung[1] and Evjen[2] depend on collecting ions into neutral groups. The convergence obtained in this way, however, is conditional: that is, the result depends on the way in which the neutral groups are chosen. Ewald's[3] method, which hinges on doing part of the summation in reciprocal space, gives rapid convergence and the limit is unique. Subsequent discussions[4-10] of this topic have been extensions and generalizations of these methods. This work too is an extension of the Ewald technique. In particular it is a generalization of the approach taken by Nijboer and DeWette.[9]

For purposes of orientation, we summarize the basic philosophy of the Ewald method. Suppose we have a function $\varphi(r)$ such that the series

$$S = \sum_{n=1}^{\infty} \varphi(\mathbf{r}_n) \tag{1}$$

is slowly converging. The symbol $\sum_n$ is to be understood as a shorthand for $\sum_{n_1} \sum_{n_2} \sum_{n_3}$. It represents independent summation on all three components of the vector $\mathbf{r}$. We now construct some function $g(\mathbf{r})$, which falls off rapidly with r, and its partner

$$f(\mathbf{r}) = 1 - g(\mathbf{r}), \tag{2}$$

which rapidly approaches unity as $\mathbf{r}$ increases. We now write

$$S = \sum_n \varphi(\mathbf{r}_n) g(\mathbf{r}_n) + \sum_n \varphi(\mathbf{r}_n) f(\mathbf{r}_n). \tag{3}$$

The first sum converges rapidly, because of g. The second sum converges like φ, i.e., slowly. Its Fourier transform, however, will converge rapidly. In fact the more slowly this sum converges the more rapidly will its transform converge. To complete the argument we need Parseval's theorem:

If

$$\Phi(\mathbf{h}) = \int \exp(i2\pi\mathbf{h}\cdot\mathbf{r})\varphi(\mathbf{r})\, d\mathbf{r} \tag{4a}$$

and

$$F(\mathbf{h}) = \int \exp(i2\pi\mathbf{h}\cdot\mathbf{r}) f(\mathbf{r})\, d\mathbf{r} \tag{4b}$$

then

$$\int \Phi(\mathbf{h}) F^*(\mathbf{h})\, d\mathbf{h} = \int \varphi(\mathbf{r}) f^*(\mathbf{r})\, d\mathbf{r} \tag{4c}$$

where the symbol * denotes "complex conjugate." The formal passage from sums to integrals can be accomplished by means of Dirac delta functions $\delta(\mathbf{r} - \mathbf{r}_n)$, as we shall see below. Thus Parseval's theorem guarantees that the summation in transform space yields the same result as the summation in the original coordinate space.

We now apply this scheme to the calculation of the potential due to an ionic lattice. To begin with, we consider what we shall call a "primitive" lattice. Such a lattice is generated from primitive translations $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$ in such fashion that

$$\mathbf{r}_n = \sum_{i=1}^{3} n_i \mathbf{c}_i, \tag{5}$$

with n_1, n_2, n_3 taking on independently all integer values from $-\infty$ to ∞; and in addition there is associated with each lattice point $\mathbf{r}_n$ a charge

$$q_n = q_0(-1)^{n_1+n_2+n_3} \tag{6}$$

where q_0 is some constant. A typical primitive lattice is NaCl, in contrast, for instance, to CaF_2, which does not obey (6).

We define reciprocal vectors $\mathbf{h}_i$ by the usual relation

$$\mathbf{h}_i \cdot \mathbf{c}_j = \delta_{ij} \tag{7}$$

and the reciprocal lattice as the aggregate of points

$$\mathbf{h}_n = \sum n_i \mathbf{h}_i \tag{8}$$

where the n's again run from $-\infty$ to ∞. It is trivial to show that if these two lattices are represented respectively as $\sum_n \delta(\mathbf{r} - \mathbf{r}_n)$ and $\sum_n \delta(\mathbf{h} - \mathbf{h}_n)$, then the reciprocal lattice is simply the Fourier transform of the coordinate lattice, the Fourier transform being understood as in (4a) and (4b). If we define the special reciprocal lattice vector

$$\mathbf{k} = \tfrac{1}{2}(\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3), \tag{9}$$

then (6) can be rewritten

$$q_n = q_0 \exp\ (i2\pi \mathbf{k} \cdot \mathbf{r}_n). \tag{10}$$

In addition to q_n, we associate with each lattice point a function $\varphi(\mathbf{r})$. For the present we place no restriction on $\varphi(\mathbf{r})$, except that it possess a Fourier transform. Of course there would be no practical motivation for the calculation unless $\varphi(\mathbf{r})$ fell off slowly with $\mathbf{r}$. We wish to sum the contribution of all the φ's at some arbitrary point $\mathbf{R}$:

$$S = \sum_n \varphi(\mathbf{r}_n - \mathbf{R}) \exp\ (i2\pi \mathbf{k} \cdot \mathbf{r}_n). \tag{11}$$

To change the sum into an integral, as required for the eventual application of (4c), we define

$$w(\mathbf{r}) = \exp\ (i2\pi \mathbf{k} \cdot \mathbf{r}) \sum_n \delta(\mathbf{r} - \mathbf{r}_n), \tag{12}$$

so that

$$S = \int w(\mathbf{r}) \varphi(\mathbf{r} - \mathbf{R}) d\mathbf{r}. \tag{13}$$

In exact analogy to (2) and (3) we can break S into two parts:

$$S = \int w(\mathbf{r}) \varphi(\mathbf{r} - \mathbf{R}) g(\mathbf{r} - \mathbf{R}) d\mathbf{r} + \int w(\mathbf{r}) \varphi(\mathbf{r} - \mathbf{R}) f(\mathbf{r} - \mathbf{R}) d\mathbf{r}. \tag{14}$$

The first integral in (14) corresponds to the first sum in (3):

$$\sum_n \varphi(\mathbf{r}_n - \mathbf{R}) g(\mathbf{r}_n - \mathbf{R}) (-1)^{n_1+n_2+n_3}. \tag{15}$$

The second integral we wish to evaluate in the conjugate domain. For brevity we define

$$\psi(\mathbf{r}) = \varphi(\mathbf{r}) f(\mathbf{r}) \tag{16}$$

$$\Psi(\mathbf{h}) = \int_{-\infty}^{\infty} \exp(i2\pi\mathbf{h}\cdot\mathbf{r})\, \psi(\mathbf{r})\, d\mathbf{r}. \tag{17}$$

Then the Fourier transform of $\varphi(\mathbf{r} - \mathbf{R}) f(\mathbf{r} - \mathbf{R})$ is given by

$$\int_{-\infty}^{\infty} \exp(i2\pi\mathbf{h}\cdot\mathbf{r})\, \psi(\mathbf{r} - \mathbf{R})\, d\mathbf{r} = \exp i2\pi\mathbf{h}\cdot\mathbf{R}\, \Psi(\mathbf{h}). \tag{18}$$

The transform of $w(r)$ is easily evaluated (see Ref. 9) and is given by

$$\int_{-\infty}^{\infty} \exp(i2\pi\mathbf{h}\cdot\mathbf{r})\, w(\mathbf{r})\, d\mathbf{r} = \frac{1}{v_c} \sum_n \delta(\mathbf{h} + \mathbf{k} - \mathbf{h}_n), \tag{19}$$

where v_c equals $\mathbf{c}_1\cdot\mathbf{c}_2 \times \mathbf{c}_3$, or the volume of the coordinate unit cell. By Parseval's theorem [(4a), (4b), (4c)], the second integral of (14) now becomes

$$\frac{1}{v_c} \sum_n \exp[i2\pi(\mathbf{h}_n - \mathbf{k})\cdot\mathbf{R}]\, \Psi(\mathbf{h}_n - \mathbf{k}). \tag{20}$$

This completes the essential derivation, since S is now expressed in terms of the two sums (15) and (20), both of which converge rapidly.

We consider some special cases. If $\mathbf{R} = 0$ — that is, if we are sitting at a lattice point — we presumably will want to exclude the contribution of that point itself. If $\varphi(0)$ is finite, the contribution can be subtracted outright. If $\varphi(0)$ diverges, as is usually the case, one must be clever about picking the functions g and f so that $\psi(0) = f(0)\varphi(0)$ does not diverge. One then simply omits from the sum (15) the term for $n = 0$, and subtracts from the sum (20) the quantity $\psi(0)$. Note that one subtracts $\psi(0)$, not $\Psi(0)$. The Ewald calculation is obtained in this way, if one takes

$$\varphi(\mathbf{r}) = |\mathbf{r}|^{-1} \tag{21a}$$

$$g(\mathbf{r}) = \mathrm{Erfc}\ (|\mathbf{r}|) \tag{21b}$$

$$f(\mathbf{r}) = \mathrm{Erf}\ (|\mathbf{r}|). \tag{21c}$$

We note that if φ is real (for example, any central potential) and if we pick g real, then ψ will be real also. The function w is both real and

symmetric because of our particular definition of the reciprocal vector $\mathbf{k}$. It follows that the sum S as defined in (14) is real. But if we look at the partial sum (20), this reality is not at first sight apparent, since the $\mathbf{R}$ can be chosen arbitrarily. But ψ real implies that Ψ is Hermitian: $\Psi(-\mathbf{h}_n + \mathbf{k}) = \Psi^*(\mathbf{h}_n - \mathbf{k})$, and clearly $\exp[i2\pi(\mathbf{h}_n - \mathbf{k})\cdot\mathbf{R}]$ is Hermitian. Again, because of the peculiar choice of $\mathbf{k}$, the arguments $\pm(\mathbf{h}_n - \mathbf{k})$ are bound to occur in pairs. Since the sum of a function and its complex conjugate is real, the sum (20) is always real, which we may emphasize by rewriting it:

$$\frac{2}{v_c}\,\mathrm{Re}\sum_n^{+}\exp[i2\pi(\mathbf{h}_n - \mathbf{k})\cdot\mathbf{R}]\,\Psi(\mathbf{h}_n - \mathbf{k}). \tag{22}$$

Here $\sum^{+}$ means that we sum only over half the space. This can be accomplished by summing n_3, for instance, from 1 to ∞ instead of from $-\infty$ to ∞.

If $\psi(\mathbf{r})$ is symmetric, $\Psi(\mathbf{h})$ will be real, and in the sum (22) we will obtain only cosine terms. Conversely, if $\psi(\mathbf{r})$ is antisymmetric, the reciprocal sum will contain only sine terms. Again this is independent of the choice of $\mathbf{R}$.

We now also see clearly why the Ewald method "works." The convergence difficulties with the series S of (1) concern its asymptotic behavior. But this behavior is related to the behavior at the origin of the series in reciprocal space.[12] By means of the vector $\mathbf{k}$, we guarantee avoidance of the origin in reciprocal space, regardless of any other conditions in the problem.

The sums that most frequently occur in practice are related to the expansion of the crystalline potential in spherical harmonics:

$$V(x) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} C_{lm}\,|\,x\,|^{l}\,Y_{l,-m}\,(\theta_x\,,\,\varphi_x) \tag{23}$$

$$C_{lm} = \frac{4\pi}{2l+1}\sum_n q_n\,|\,\mathbf{r}_n\,|^{-l-1}\,Y_{l,m}\,(\theta_{\mathbf{r}_n}\,,\,\varphi_{\mathbf{r}_n}) \tag{24}$$

$$Y_{l,m}\,(\theta,\varphi) = \left[\frac{2l+1}{4\pi}\cdot\frac{(l-|\,m\,|)!}{(l+|\,m\,|)!}\right]^{\frac{1}{2}} e^{im\varphi}\,P_{lm}(\theta). \tag{25}$$

The notation is well known and conventional. Our definition of the spherical harmonics Y_{lm} implies ${Y_{lm}}^* = Y_{l-m}$. Also $Y_{lm}(\pi-\theta,\pi+\varphi) = (-1)^l Y_{lm}(\theta,\varphi)$. The evaluation of the crystal sums C_{lm} has been discussed by Nijboer and DeWette.[9] In our notation, $\varphi(\mathbf{r})$ here corresponds to $r^{-l-1}Y_{lm}(\theta,\varphi)$. Nijboer and DeWette's choice of g is the incomplete gamma function[11]

$$g(\mathbf{r}) = \Gamma(n + l,\pi r^2)/\Gamma(n + l) \tag{26a}$$

$$f(\mathbf{r}) = 1 - g(\mathbf{r}) = \gamma(n + l,\pi r^2)/\Gamma(n + l). \tag{26b}$$

They solve the problem subject to the restrictions that (a) the potential is expanded about a lattice point, i.e., $\mathbf{R} = 0$, and (b) the lattice is primitive, in the sense of (5), (6) and (10). Our discussion has made it obvious how to remove the first restriction. Generalizing their result, via (20),

$$\begin{aligned} C_{lm}(\mathbf{r} - \mathbf{R}) = {} & \frac{4\pi}{2l + 1} \cdot \frac{1}{\Gamma(l + \frac{1}{2})} \\ & \cdot [\sum_n |\mathbf{r}_n - \mathbf{R}|^{-l-1}\Gamma(l + \tfrac{1}{2},\pi |\mathbf{r}_n - \mathbf{R}|^2) Y_{lm}(\theta_{\mathbf{r}_n-\mathbf{R}}, \varphi_{\mathbf{r}_n-\mathbf{R}}) \\ & \cdot \exp(i2\pi\mathbf{k}\cdot\mathbf{r}_n) \\ & + i^l \pi^{l-\frac{1}{2}} v_c^{-1} \sum_n \exp(i2\pi(\mathbf{h}_n - \mathbf{k})\cdot\mathbf{R}) |\mathbf{h}_n - \mathbf{k}|^{l-2} \\ & \cdot \Gamma(1,\pi |\mathbf{h}_n - \mathbf{k}|^2) \times Y_{lm}(\theta_{\mathbf{h}_n-\mathbf{k}}, \varphi_{\mathbf{h}_n-\mathbf{k}})]. \end{aligned} \tag{27}$$

The next generalization of the procedure lies in its application to nonprimitive lattices. Such application will clearly be possible if an arbitrary lattice can be decomposed into a sum of component primitive lattices. We illustrate what we mean by a one-dimensional example. Consider a one-dimensional lattice with charges distributed as follows:

$$L = 2\ 1\ 0\ -3\ 2\ 1\ 0\ -3\ 2\ 1\ 0\ -3\ \cdots.$$

Thus we have $q_1 = 2$, $q_2 = 1$, $q_3 = 0$, $q_4 = -3$. If the distance between successive q's is 1 distance unit, then the basic periodicity is 4. We note that $\sum q = 0$, since we must have a neutral lattice, and that zero is itself an allowable q value. Now consider the following sequences of numbers of periodicity 4:

$$L_1 = \tfrac{1}{2}\ -\tfrac{1}{2}\ \tfrac{1}{2}\ -\tfrac{1}{2}\ \cdots$$

$$L_2 = \frac{1}{\sqrt{2}}\ 0\ -\frac{1}{\sqrt{2}}\ 0\ \cdots$$

$$L_3 = 0\ \frac{1}{\sqrt{2}}\ 0\ -\frac{1}{\sqrt{2}}\ \cdots.$$

We can represent L as the sum $L = 2L_1 + \sqrt{2}\,L_2 + 2\sqrt{2}\,L_3$. We

note that L_1, L_2, L_3 are all primitive since they fulfill both (5) and the two equivalent equations (6) and (10). In terms of (5), for L_1, $|\mathbf{c}| = 1$; for L_2 and L_3, $|\mathbf{c}| = 2$, but their origin of coordinates is shifted by one unit.

We can define a dot product for the L's in the following sense. Suppose we have $L_A = \sum_i q_i^{(A)} \delta(\mathbf{r} - \mathbf{r}_i)$ and $L_B = \sum_i q_i^{(B)} \delta(\mathbf{r} - \mathbf{r}_i)$. Then $L_A \cdot L_B = \sum_i q_i^{(A)} q_i^{(B)}$, where the sum runs over a unit cell. We note that, in our example, the components L_i are orthonormal, and that the projection of L on each L_i can be found by taking the dot product.

We also observe that we have three components, which is just enough to account for 4 numbers which are subject to the one constraint $\sum q_i = 0$. Our three components L_i "span the space" of L, and it is clear in general that if L consists of a periodic sequence of P numbers whose sum is zero, we shall need $(P - 1)$ primitive components.

The question is: Is it possible to decompose an arbitrary periodic sequence L into orthonormal primitive components? Can one devise an algorithm for finding these components? Can one do this in three dimensions? In considering these questions, we have collaborated with Dr. R. L. Graham, and we are particularly indebted to him for pointing out the great simplification that results if one confines oneself to sequences of periodicity 2^n in each dimension.

Consider a linear sequence L of periodicity 2^n. If $\mathbf{c}$ is the primitive translation of L, we define $\mathbf{a} = \mathbf{c}/2^n$. The basic vectors for generating primitive component lattices are $\mathbf{b}^{(1)} = \mathbf{a}$, $\mathbf{b}^{(2)} = 2\mathbf{b}^{(1)}$, $\cdots$ $\mathbf{b}^{(n)} = 2\mathbf{b}^{(n-1)}$. Each $\mathbf{b}^{(i)}$ for $i > 1$, will generate a *set* of primitive lattices differing only in their choice of origin. There will evidently be n such sets of components. Primitive component lattices containing 2^m nonzero entries per unit cell will have a basis vector of length $|\mathbf{c}|/2^{n-m}$ and will have 2^{n-m} possible shifts of origin. We note that $\sum_{m=1}^{n} 2^{n-m} = 2^n - 1$, which is the correct total number of components. The number of primitive components of the same periodicity but with shifted origins doubles every time the length of the generating basic vector $\mathbf{b}_i$ doubles, and of course the number of nonzero entries per unit cell halves at the same time. The set of origin positions $\{\mathbf{R}_i\}$ associated with $\mathbf{b}_i$ is clearly the set of all translations $\mathbf{R}$ such that $\mathbf{R}_l - \mathbf{R}_j \neq \mathbf{b}_i$. This includes the set $\{\mathbf{R}_{i-1}\}$ plus a new set formed by adding $\mathbf{b}_{i-1}$ to all $\mathbf{R}$ in $\{\mathbf{R}_{i-1}\}$.

All the primitive component lattices are orthogonal to each other, in the sense that $L_i \cdot L_j = \delta_{ij}$. Within each "phase-shifted" set, each

component has numbers in locations where all the other components have zero, so that the members of such a set are obviously orthogonal. Now consider two sequences L_i and L_{i-1} belonging to sets generated by $\mathbf{b}_i$ and $\mathbf{b}_{i-1}$. Either L_{i-1} will have numbers only where L_i has zero. Otherwise, each alternate number in L_{i-1} will have zero as a partner in L_i; the remaining numbers in L_{i-1} all have the same sign, but their nonzero partners in L_i have alternating signs. Hence once again the dot product is zero. The same argument applies clearly not only to members of contiguous sets L_i and L_{i-1}, but to members of any two sets, L_i and L_j, $i \neq j$. To produce not merely orthogonality, but orthonormality over the unit cell, the normalization factor, or q_0 in (6) and (10), must clearly be $2^{-m/2}$ for a component with 2^m nonzero entries.

In the example we have given, $n = 2$, and all the above arguments can be seen to be rather trivially verified.

Extension of the preceding to two dimensions is straightforward. We consider a two-dimensional array, doubly periodic with periodicity $2^n \times 2^n$. The primitive translations $\mathbf{c}_1$ and $\mathbf{c}_2$ carry any q into the corresponding q in another cell. Within each cell, the different q's are separated by multiples of $\mathbf{c}_1/2^n$ and $\mathbf{c}_2/2^n$, and we shall call these vectors $\mathbf{a}_1$ and $\mathbf{a}_2$. As before, we define a set of components of equal periodicity by defining the basic vectors which will generate the primitive lattice. The translations by which components within a set differ we denote by $\mathbf{R}$. The algorithm for producing a complete orthonormal set of components is simple:

$$\mathbf{b}_1^{(1)} = \mathbf{a}_1 \tag{28a}$$

$$\mathbf{b}_2^{(1)} = \mathbf{a}_2 \tag{28b}$$

$$\mathbf{b}_1^{(2)} = \mathbf{a}_1 + \mathbf{a}_2 \tag{28c}$$

$$\mathbf{b}_2^{(2)} = \mathbf{a}_1 - \mathbf{a}_2 \tag{28d}$$

$$\mathbf{b}_1^{(n+2)} = 2\mathbf{b}_1^{(n)} \tag{28e}$$

$$\mathbf{b}_2^{(n+2)} = 2\mathbf{b}^{(n)} \tag{28f}$$

$$q^{(1)} = 2^{-n} \tag{29a}$$

$$q^{(n)} = q^{(n-1)}\sqrt{2} \tag{29b}$$

$$\{\mathbf{R}^{(1)}\} = 0 \tag{30a}$$

$$\{\mathbf{R}^{(n+1)}\} = \{\mathbf{R}^{(n)}\} + \{\mathbf{R}^{(n)} + \mathbf{b}_1^{(n)}\}. \tag{30b}$$

The set labels appear as superscripts in (28–30). The vectors $\mathbf{b}$ are the basis of a primitive lattice, and $\{\mathbf{R}\}$ indicates the set of all translations $\mathbf{R}$ which relate primitive lattices having the same basis. Equation (30b) says that the translations for set n include all the translations for set $n-1$, plus all translations formed by adding any *one* of the vectors $\mathbf{b}^{n-1}$ to all the $\mathbf{R}$'s of set $n-1$. Again all components are orthonormal over the unit cell, so that the decomposition of any $2^n \times 2^n$ dimensional cell can simply be found by taking dot products. Normalization is insured by our definition of q_n, and orthogonality becomes clear from the same type of reasoning as in the linear case, which we shall not repeat here.

In three dimensions, we consider an array triply periodic with periodicity $2^n \times 2^n \times 2^n$. We define $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$, and $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ in analogy with the two-dimensional case. We present the basic vectors for the primitive lattices, plus the associated translations:

$$\mathbf{b}_1^{(1)} = \mathbf{a}_3 \tag{31a}$$

$$\mathbf{b}_2^{(1)} = \mathbf{a}_2 \tag{31b}$$

$$\mathbf{b}_3^{(1)} = \mathbf{a}_2 + \mathbf{a}_1 \tag{31c}$$

$$\mathbf{b}_1^{(2)} = \mathbf{a}_3 + \mathbf{a}_2 \tag{31d}$$

$$\mathbf{b}_2^{(2)} = -\mathbf{a}_3 + \mathbf{a}_2 \tag{31e}$$

$$\mathbf{b}_3^{(2)} = \mathbf{a}_1 \tag{31f}$$

$$\mathbf{b}_1^{(3)} = \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3 \tag{31g}$$

$$\mathbf{b}_2^{(3)} = \mathbf{a}_1 - \mathbf{a}_2 + \mathbf{a}_3 \tag{31h}$$

$$\mathbf{b}_3^{(3)} = \mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_3 \tag{31i}$$

$$\mathbf{b}_1^{(n+3)} = 2\mathbf{b}_1^{(n)} \tag{31j}$$

$$\mathbf{b}_2^{(n+3)} = 2\mathbf{b}_2^{(n)} \tag{31k}$$

$$\mathbf{b}_3^{(n+3)} = 2\mathbf{b}_3^{(n)} \tag{31l}$$

$$q^{(1)} = 2^{-3n/2} \tag{32a}$$

$$q^{(n+1)} = q^n \sqrt{2} \tag{32b}$$

$$\{\mathbf{R}^{(1)}\} = 0 \tag{33a}$$

$$\{\mathbf{R}^{(n+1)}\} = \{\mathbf{R}^n\} + \{\mathbf{R}^{(n)} + \mathbf{b}_1^{(n)}\}. \tag{33b}$$

The choice of $\mathbf{b}$'s is not unique, and we have given a set that seems to bear a maximum analogy to the two-dimensional case.

Many lattices can be represented as $2^n \times 2^n \times 2^n$ dimensional arrays. We note that the primitive translations $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$ need not be orthogonal. Even if a lattice does not lend itself to such a representation, by choosing a fine enough grid (with a correspondingly large number of zero charges), one can approximate any lattice to any desired degree of accuracy.*

We now return to the problem of the lattice sum. We will obtain a sum of the type (20), or more specifically of the type (27a), for each primitive component. For a particular set of components which are translated from each other by $\{\mathbf{R}_n\}$, it clearly makes no difference whether we sum the contribution of each component at the origin or the contribution of any one component at the points given by $\{\mathbf{R}_n\}$. (The difference between shifting the function and shifting the coordinate system is merely a conceptual one.) Hence the shift vectors $\mathbf{R}_n$ correspond to the position vectors $\mathbf{R}$ in expressions (20) and (27a).

In the present approach, we have completely split the geometrical character of the lattice from the charges assigned to each lattice point. This suggests the possibility of computing the lattice sums arising from all the primitive components of commonly occurring grids, once and for all. The problem of computing a particular lattice sum would then involve only the decomposition of the given lattice into primitive component lattices, whose contribution to the sum would already be known. Work along this line is in progress.

ACKNOWLEDGMENTS

I am indebted to Dr. R. L. Graham and Dr. H. O. Pollak for several helpful discussions.

REFERENCES

1. Madelung, E., Z. Physik, *19*, 1918, p. 528.
2. Evjen, H. M., Phys. Rev., *39*, 1932, p. 675.
3. Ewald, P. P., Ann. Physik, *64*, 1921, p. 253.
4. Misra, R. D., Proc. Camb. Phil. Soc., *36*, 1940, p. 173.

* The restriction of $2^n \times 2^n \times 2^n$ dimensional arrays is in fact too rigid. It is possible to formulate necessary and sufficient conditions for the decomposability of an arbitrary grid; to prove the existence of a complete set of orthogonal primitive lattices when the decomposition is possible; and to provide simple algorithms for constructing these lattices. These more general results have been derived by Dr. R. L. Graham, and we are grateful for access to them prior to publication.

5. Frank, F. C., Phil. Mag., *41*, 1950, p. 1287.
6. Placzek, G., Nijboer, B. R. A., and Van Hove, L., Phys. Rev., *82*, 1951, p. 392.
7. Roy, S. K., Can. J. Phys., *32*, 1954, p. 509.
8. Kanamori, J., Moriya, T., Motizuki, K., and Nagamiya, T., J. Phys. Soc. Japan, *10*, 1955, p. 93.
9. Nijboer, B. R. A., and DeWette, F. W., Physica, *23*, 1957, p. 309.
10. Nijboer, B. R. A., and DeWette, F. W., Physica, *24*, 1958, p. 422.
11. *Higher Transcendental Functions*, ed. Erdélyi, A., Vol. II, McGraw-Hill, New York, 1953.
12. Grant, W. J. C., Physica, *30*, 1964, p. 1433.

On the Boundedness of Solutions of Nonlinear Integral Equations

By I. W. SANDBERG

(Manuscript received November 5, 1964)

Sufficient conditions are presented for the boundedness of the solutions of a vector nonlinear Volterra integral equation of the second kind that frequently arises in the study of automatic control systems containing an arbitrary finite number of time-varying nonlinear elements. Similar conditions are given for the boundedness of the solutions of the discrete analog of the integral equation.

A direct application of the results yields a Nyquist-like frequency-domain condition for the "bounded-input implies bounded-output stability" of a large class of feedback systems containing a single time-varying nonlinear element.

I. NOTATION AND DEFINITIONS

Let M denote an arbitrary matrix. We shall denote by M', M^*, and M^{-1}, respectively, the transpose, the complex-conjugate transpose, and the inverse of M. The positive square root of the largest eigenvalue of M^*M is denoted by $\Lambda\{M\}$, and 1_N denotes the identity matrix of order N.

The set of real measurable N-vector-valued functions of the real variable t defined on $[0, \infty)$ is denoted by $\mathfrak{C}_N(0, \infty)$ and the jth component of $f \in \mathfrak{C}_N(0, \infty)$ is denoted by f_j.

The sets $\mathfrak{L}_{\infty N}(0, \infty)$ and $\mathfrak{L}_{2N}(0, \infty)$ are defined by

$$\mathfrak{L}_{\infty N}(0, \infty) = \{f \mid f \in \mathfrak{C}_N(0, \infty), \sup_{t \geq 0} [f'(t)f(t)] < \infty\}$$

$$\mathfrak{L}_{2N}(0, \infty) = \left\{f \mid f \in \mathfrak{C}_N(0, \infty), \int_0^\infty f'(t)f(t)dt < \infty\right\}$$

The norm of $f \in \mathfrak{L}_{2N}(0, \infty)$ is denoted by $\| f \|$ and is defined by

$$||f|| = \left(\int_0^\infty f'(t)f(t)dt\right)^{\frac{1}{2}}.$$

With this norm $\mathfrak{L}_{2N}(0,\infty)$ is a Banach space.

Let $y \varepsilon (0,\infty)$ and define f_y by

$$f_y(t) = f(t) \quad \text{for} \quad t \varepsilon [0,y]$$
$$= 0 \quad \text{for} \quad t > y$$

for any $f \varepsilon \mathfrak{IC}_N(0,\infty)$, and let

$$\mathcal{E}_N = \{f \mid f \varepsilon \mathfrak{IC}_N(0,\infty), f_y \varepsilon \mathfrak{L}_{2N}(0,\infty) \text{ for } 0 < y < \infty\}.$$

With A an arbitrary real measurable $N \times N$ matrix-valued function of t with elements $\{a_{nm}\}$ defined on $[0,\infty)$, let $\mathfrak{K}_{pN}$ $(p = 1,2)$ denote

$$\left\{A \mid \int_0^\infty |a_{nm}(t)|^p dt < \infty \ (n,m = 1, 2, \cdots, N)\right\}.$$

Let $\psi[f(t),t]$ denote

$$(\psi_1[f_1(t),t],\psi_2[f_2(t),t], \cdots, \psi_N[f_N(t),t])', \qquad f \varepsilon \mathfrak{IC}_N(0,\infty)$$

where $\psi_1(w,t)$, $\psi_2(w,t)$, $\cdots$, $\psi_N(w,t)$ are real-valued functions of the real variables w and t for $-\infty < w < \infty$ and $0 \leqq t < \infty$ such that

(*i*) $\psi_n(0,t) = 0$ for $t \varepsilon [0,\infty)$ and $n = 1, 2, \cdots, N$

(*ii*) there exist real numbers α and β with the property that

$$\alpha \leqq \frac{\psi_n(w,t)}{w} \leqq \beta \qquad (n = 1, 2, \cdots, N)$$

for $t \varepsilon [0,\infty)$ and all real $w \neq 0$.

(*iii*) $\psi_n[w(t),t]$ $(n = 1, 2, \cdots, N)$ is a measurable function of t whenever $w(t)$ is measurable.

The symbol s denotes a scalar complex variable with $\sigma = \text{Re}[s]$ and $\omega = \text{Im}[s]$.

II. INTRODUCTION AND SUMMARY

In the study of physical systems such as nonlinear automatic control systems containing an arbitrary finite number of time-varying nonlinear elements, attention is frequently focused on the properties of the equation

$$g(t) = f(t) + \int_0^t k(t-\tau)\psi[f(\tau),\tau]d\tau, \qquad t \geqq 0$$

in which $g \varepsilon \mathcal{E}_N$, $f \varepsilon \mathcal{E}_N$, $k(\cdot) \varepsilon \mathcal{K}_{1N}$, and $\psi[\cdot,\cdot]$ is as defined in the previous section.

In Ref. 1, the following theorem is proved.

Theorem 1: Let $k \varepsilon \mathcal{K}_{1N}$, and let

$$v(t) = u(t) + \int_0^t k(t-\tau)\psi[u(\tau),\tau]d\tau, \qquad t \geqq 0$$

where $v \varepsilon \mathcal{L}_{2N}(0,\infty)$ and $u \varepsilon \mathcal{E}_N$. Let

$$K(s) = \int_0^\infty k(t)e^{-st}dt, \qquad \sigma \geqq 0.$$

Suppose that

(i) $\det [1_N + \frac{1}{2}(\alpha+\beta)K(s)] \neq 0$ for $\sigma \geqq 0$

(ii) $\frac{1}{2}(\beta-\alpha) \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \frac{1}{2}(\alpha+\beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

Then $u \varepsilon \mathcal{L}_{2N}(0,\infty)$, and there exists a positive constant ρ which depends only on k, α, and β such that

$$\| u \| \leqq \rho \| v \|.$$

The primary purpose of this paper is to prove the following related result.

Theorem 2: Let $t^p k \varepsilon \mathcal{K}_{1N} \cap \mathcal{K}_{2N}$ for $p = 0, 1, 2$. Let

$$g(t) = f(t) + \int_0^t k(t-\tau)\psi[f(\tau),\tau]d\tau, \qquad t \geqq 0$$

where $g \varepsilon \mathcal{L}_{\infty N}(0,\infty)$ and $f \varepsilon \mathcal{E}_N$. Let

$$K(s) = \int_0^\infty k(t)e^{-st}dt, \qquad \sigma \geqq 0.$$

Suppose that

(i) $\det [1_N + \frac{1}{2}(\alpha+\beta)K(s)] \neq 0$ for $\sigma \geqq 0$

(ii) $\frac{1}{2}(\beta-\alpha) \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \frac{1}{2}(\alpha+\beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

Then $f \varepsilon \mathcal{L}_{\infty N}(0,\infty)$, there exists a positive constant c which depends only on k, α, and β such that

$$\max_j \sup_{t\geqq 0} |f_j(t)| \leqq c \max_j \sup_{t\geqq 0} |g_j(t)|,$$

and $f_j(t) \to 0$ as $t \to \infty$ for $j = 1, 2, \cdots, N$ whenever $g_j(t) \to 0$ as $t \to \infty$ for $j = 1, 2, \cdots, N$.

A direct application of Theorem 2 yields a *frequency-domain* condition for the $\mathfrak{L}_\infty$-stability[2] of a well-known type of feedback system. This is discussed in Section IV. In Section V, sufficient conditions are stated for the boundedness of the solutions of the discrete analog of the nonlinear integral equation considered in Theorem 2. In Section VI, we describe some additional results that can be proved by combining the methods of this paper with the $\mathfrak{L}_{2N}(0,\infty)$ arguments of Ref. 1 and another earlier paper.

III. PROOF OF THEOREM 2

Assume throughout this section that the hypotheses of Theorem 2 are satisfied.

Let $q_j(t)$ be defined on $[0,\infty)$ by

$$\begin{aligned} q_j(t) &= \frac{\psi_j[f_j(t),t]}{f_j(t)}, \qquad t \in \{t \mid t \geqq 0, f_j(t) \neq 0\} \\ &= \tfrac{1}{2}(\alpha + \beta), \qquad t \in \{t \mid t \geqq 0, f_j(t) = 0\} \end{aligned}$$

for $j = 1, 2, \cdots, N$; and let $q(t)$ denote the diagonal matrix diag $[q_1(t), q_2(t), \cdots, q_N(t)]$. Then

$$g(t) = f(t) + \int_0^t k(t-\tau)q(\tau)f(\tau)d\tau, \qquad t \geqq 0.$$

Let a be an arbitrary positive number, and for each nonnegative integer n let $g^{(n)}(t)$ be defined on $[0,\infty)$ by

$$\begin{aligned} g^{(n)}(t) &= g(t), \quad na \leqq t < (n+1)a \\ &= 0, \qquad 0 \leqq t < na \text{ and } t \geqq (n+1)a. \end{aligned}$$

Lemma 1: For each integer $n \geqq 0$, $\mathfrak{L}_{2N}(0,\infty)$ contains a unique element $f^{(n)}$ such that

(i) $f^{(n)}(t) = 0, \qquad 0 \leqq t < na$

(ii) $g^{(n)}(t) = f^{(n)}(t) + \int_0^t k(t-\tau)q(\tau)f^{(n)}(\tau)d\tau, \qquad t \geqq 0.$

Proof of Lemma 1:

Clearly $g^{(n)} \in \mathfrak{L}_{2N}(0,\infty)$ for $n \geqq 0$. Let $\mathbf{I}$ denote the identity operator on $\mathfrak{L}_{2N}(0,\infty)$, and let $\mathbf{K}$ and $\mathbf{Q}$ denote the mappings of $\mathfrak{L}_{2N}(0,\infty)$ into itself defined by[3]

$$(\mathbf{K}h)(t) = \int_0^t k(t-\tau)h(\tau)d\tau, \qquad t \geqq 0$$

$$(\mathbf{Q}h)(t) = (q_1(t)h_1(t), q_2(t)h_2(t), \cdots, q_N(t)h_N(t))', \qquad t \geqq 0$$

where h is an arbitrary element of $\mathfrak{L}_{2N}(0,\infty)$.

According to Lemma 5 of Ref. 1, the operator $[\mathbf{I} + \frac{1}{2}(\alpha+\beta)\mathbf{K}]$ possesses an inverse on $\mathfrak{L}_{2N}(0,\infty)$. Thus the functional equation

$$g^{(n)} = h^{(n)} + \mathbf{K}\mathbf{Q}h^{(n)}, \qquad h^{(n)} \varepsilon\, \mathfrak{L}_{2N}(0,\infty)$$

can be written as $h^{(n)} = \mathbf{T}h^{(n)}$, in which $\mathbf{T}$ is defined by

$$\mathbf{T}h^{(n)} = [\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}g^{(n)} - [\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}\mathbf{K}[\mathbf{Q} - \tfrac{1}{2}(\alpha+\beta)\mathbf{I}]h^{(n)}.$$

Using the bounds of Lemma 5 of Ref. 1, and the fact that $\alpha \leqq q_j(t) \leqq \beta$ for $j = 1, 2, \cdots, N$ and $t \geqq 0$, it can easily be shown that $\mathbf{T}$ is a *contraction* mapping of $\mathfrak{L}_{2N}(0,\infty)$ into itself. Thus, it follows from the contraction-mapping fixed-point theorem that $\mathfrak{L}_{2N}(0,\infty)$ contains a unique element $f^{(n)}$ which satisfies condition (*ii*) of the lemma.

Since $[\mathbf{I} + \frac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}$ is necessarily causal, and

$$f^{(n)} = \lim_{m\to\infty} \mathbf{T}^m\theta,$$

in which θ is the zero-element of $\mathfrak{L}_{2N}(0,\infty)$, we see that $f^{(n)} = 0$ for $0 \leqq t < na$ and $n > 0$.

Lemma 2: Let $f^{(n)}$ be the associate of $g^{(n)}$ in accordance with Lemma 1. Then

$$f(t) = \sum_{n=0}^{\infty} f^{(n)}(t), \qquad t \geqq 0.$$

Proof of Lemma 2:

Let

$$\hat{f}(t) = \sum_{n=0}^{\infty} f^{(n)}(t), \qquad t \geqq 0.$$

Then

$$g(t) = \hat{f}(t) + \int_0^t k(t-\tau)q(\tau)\hat{f}(\tau)\,d\tau, \qquad t \geqq 0$$

and hence

$$0 = [f(t) - \hat{f}(t)] + \int_0^t k(t-\tau)q(\tau)[f(\tau) - \hat{f}(\tau)]\,d\tau, \qquad t \geqq 0. \quad (1)$$

Theorem 1 implies that $(f - \hat{f}) \in \mathcal{L}_{2N}(0,\infty)$ and that $\| f - \hat{f} \| = 0$. Since the integral in (1) must therefore vanish for $t \geqq 0$, we have

$$f(t) = \hat{f}(t) \quad \text{for} \quad t \geqq 0.$$

Lemma 3: Let $f^{(n)}$ be the associate of $g^{(n)}$ in accordance with Lemma 1. Then there exists a positive constant Ω which depends only on k, α, and β such that

$$|f_j^{(n)}(t)| \leqq |g_j^{(n)}(t)| + (1 + t - na)^{-2}\Omega(1 + a)^2(Na)^{\frac{1}{2}} \max_j \sup_{t \geqq 0} |g_j^{(n)}(t)|, \qquad t \geqq na$$

for $j = 1, 2, \cdots, N$ and every $n \geqq 0$.

Before proceeding to the proof of Lemma 3, it is convenient to state the following result, which is easily provable with the aid of Parseval's identity, the well-known extremal property of the largest eigenvalue of a Hermitian matrix, and the Schwarz inequality.

Lemma 4: Let $w \in \mathcal{K}_{1N} \cap \mathcal{K}_{2N}$, $z \in \mathcal{L}_{2N}(0,\infty)$, and

$$y(t) = \int_0^t w(t-\tau)z(\tau)\,d\tau \quad \text{for} \quad t \geqq 0.$$

Then:

(*i*) $y \in \mathcal{L}_{2N}(0,\infty)$

(*ii*) with $W(j\omega) = \int_0^\infty w(t)\, e^{-i\omega t}\, dt \qquad (-\infty < \omega < \infty)$,

$$\| y \| \leqq \sup_{-\infty<\omega<\infty} \Lambda\{W(j\omega)\} \| z \|$$

(*iii*) $$| y_n(t) | \leqq \left(\sum_{m=1}^{N} \int_0^\infty | w_{nm}(t) |^2\, dt \right)^{\frac{1}{2}} \| z \|$$

for $t \geqq 0$ and $n = 1, 2, \cdots, N$.

Proof of Lemma 3

Let n denote an arbitrary nonnegative integer.

Since

$$g^{(n)}(t) = f^{(n)}(t) + \int_0^t k(t-\tau)q(\tau)f^{(n)}(\tau)\,d\tau, \qquad t \geqq 0$$

it is certainly true that for each positive integer p

$$(1 + t - na)^p g^{(n)}(t) = (1 + t - na)^p f^{(n)}(t) + \int_0^t k(t - \tau)[(1 + \tau - na) + (t - \tau)]^p q(\tau) f^{(n)}(\tau)\, d\tau, \qquad t \geqq 0$$

or, what is the same thing,

$$h(p,t) = (1 + t - na)^p f^{(n)}(t) + \int_0^t k(t - \tau) q(\tau)(1 + \tau - na)^p f^{(n)}(\tau)\, d\tau, \qquad t \geqq 0 \tag{2}$$

in which

$$h(p,t) = (1 + t - na)^p g^{(n)}(t) - \sum_{m=0}^{p-1} \frac{p!}{(p - m)!m!} \int_0^t (t - \tau)^{p-m} k(t - \tau) \cdot q(\tau)(1 + \tau - na)^m f^{(n)}(\tau)\, d\tau, \qquad t \geqq 0.$$

From Lemma 4, our assumption that $tk \in \mathcal{K}_{1N}$, and the fact that $f^{(n)} \in \mathcal{L}_{2N}(0, \infty)$, it is clear that $h(1,\cdot) \in \mathcal{L}_{2N}(0, \infty)$. A direct application of Theorem 1 to (2) with $p = 1$ shows [recall that $\alpha \leqq q_j(t) \leqq \beta$ for $t \geqq 0$ and $j = 1, 2, \cdots, N$] that $(1 + t - na)f^{(n)} \in \mathcal{L}_{2N}(0, \infty)$ and that there exists a positive constant c_1 that depends only on k, α, and β such that

$$\| (1 + t - na)f^{(n)} \| \leqq c_1 \| h(1,\cdot) \|.$$

Since by assumption $t^2 k \in \mathcal{K}_{1N}$, this argument can be repeated for $p = 2$. Thus, $(1 + t - na)^2 f^{(n)} \in \mathcal{L}_{2N}(0, \infty)$ and

$$\| (1 + t - na)^2 f^{(n)} \| \leqq c_1 \| h(2,\cdot) \|.$$

Using Lemma 4, our assumption that $t^r k \in \mathcal{K}_{2N}(r = 0,1,2)$, (2) with $p = 2$, our bounds on $\| (1 + t - na)f^{(n)} \|$ and $\| (1 + t - na)^2 f^{(n)} \|$, and the fact that $\| f^{(n)} \| \leqq c_1 \| g^{(n)} \|$, it is a simple matter to show that there exist positive constants c_2, c_3, and c_4, each depending only on k, α, and β, such that

$$| f_j^{(n)}(t) | \leqq | g_j^{(n)}(t) | + (1 + t - na)^{-2}[c_2 \| (1 + t - na)^2 g^{(n)} \| + c_3 \| (1 + t - na) g^{(n)} \| + c_4 \| g^{(n)} \|], \qquad t \geqq na$$

for $j = 1, 2, \cdots, N$.

Since

$$\| g^{(n)} \| \leqq \| (1 + t - na) g^{(n)} \| \leqq \| (1 + t - na)^2 g^{(n)} \|,$$

and

$$\begin{aligned} \| (1 + t - na)^2 g^{(n)} \| &\leqq (1 + a)^2 \| g^{(n)} \| \\ &= (1 + a)^2 \left(\int_{na}^{(n+1)a} g'(t) g(t)\, dt \right)^{\frac{1}{2}} \\ &\leqq (1 + a)^2 (Na)^{\frac{1}{2}} \max_j \sup_{t \geqq 0} | g_j^{(n)}(t) | , \end{aligned}$$

we have, with $\Omega = c_2 + c_3 + c_4$,

$$\begin{aligned} | f_j^{(n)}(t) | \leqq & | g_j^{(n)}(t) | \\ & + (1 + t - na)^{-2} \Omega (1 + a)^2 (Na)^{\frac{1}{2}} \max_j \sup_{t \geqq 0} | g_j^{(n)}(t) |, \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t \geqq na \end{aligned}$$

for $j = 1, 2, \cdots, N$. This proves Lemma 3.

Let t satisfy $ma \leqq t < (m + 1)a$ where m is an arbitrary nonnegative integer. Then, by Lemmas 1, 2, and 3

$$f(t) = \sum_{n=0}^{\infty} f^{(n)}(t) = \sum_{n=0}^{m} f^{(n)}(t),$$

and

$$\begin{aligned} | f_j(t) | &\leqq \sum_{n=0}^{m} | f_j^{(n)}(t) | \\ &\leqq | g_j^{(m)}(t) | \\ &\quad + c_5(a) \max_j \sup_{t \geqq 0} | g_j(t) | \sum_{n=0}^{m} (1 + ma - na)^{-2} \end{aligned}$$

for $j = 1, 2, \cdots, N$, in which

$$c_5(a) = \Omega (1 + a)^2 (Na)^{\frac{1}{2}}.$$

Let

$$c_6(a) = \sum_{n=0}^{\infty} (1 + na)^{-2}.$$

Since

$$\sum_{n=0}^{m} (1 + ma - na)^{-2} < \sum_{n=0}^{\infty} (1 + na)^{-2},$$

we have

$$|f_j(t)| \leqq \sup_{t\geqq 0} |g_j(t)| + c_5(a)c_6(a) \max_j \sup_{t\geqq 0} |g_j(t)|$$

for *every* integer $m \geqq 0$ (and hence *every* $t \geqq 0$) and $j = 1, 2, \cdots, N$. Therefore

$$\max_j \sup_{t\geqq 0} |f_j(t)| \leqq [1 + c_5(a)c_6(a)] \max_j \sup_{t\geqq 0} |g_j(t)|.$$

Now suppose that $g_j(t) \to 0$ as $t \to \infty$ for $j = 1, 2, \cdots, N$. We will show that for each $\epsilon > 0$ there exists a $t_\epsilon > 0$ such that $|f_j(t)| < \epsilon$ for $t > t_\epsilon$ and $j = 1, 2, \cdots, N$.

Let $\epsilon > 0$ be given, and again consider the relation

$$f(t) = \sum_{n=0}^{\infty} f^{(n)}(t).$$

Since

$$\sum_{n=n_1}^{\infty} |f_j^{(n)}(t)| \leqq \max_j \sup_{t\geqq n_1 a} |g_j(t)| \left[1 + c_5(a) \sum_{n=n_1}^{n_2} (1 + n_2 a - na)^{-2}\right]$$
$$\leqq \max_j \sup_{t\geqq n_1 a} |g_j(t)|[1 + c_5(a)c_6(a)]$$

for $n_1 a \leqq n_2 a \leqq t < (n_2 + 1)a$, with n_1 and n_2 positive integers, it is clear that there exists a positive integer n_3 such that

$$\sum_{n=n_3}^{\infty} |f_j^{(n)}(t)| < \tfrac{1}{2}\epsilon \quad \text{for} \quad t \geqq n_3 a \quad \text{and} \quad j = 1, 2, \cdots, N.$$

From the inequality

$$\sum_{n=0}^{(n_3-1)} |f_j^{(n)}(t)| \leqq c_5(a) \max_j \sup_{t\geqq 0} |g_j(t)| \sum_{n=0}^{(n_3-1)} (1 + t - na)^{-2}, \qquad t \geqq n_3 a$$

it is evident that there exists a positive integer $n_4 > n_3$ such that

$$\sum_{n=0}^{(n_3-1)} |f_j^{(n)}(t)| < \tfrac{1}{2}\epsilon \quad \text{for} \quad t > n_4 a \quad \text{and} \quad j = 1, 2, \cdots, N.$$

Thus

$$|f_j(t)| \leqq \sum_{n=0}^{\infty} |f_j^{(n)}(t)| < \epsilon \quad \text{for} \quad t > n_4 a \quad \text{and} \quad j = 1, 2, \cdots, N$$

This completes the proof of Theorem 2.

Remarks:

With regard to the hypotheses of Theorem 2, it can easily be verified that if the elements of k are uniformly bounded on $[0, \infty)$, then the assumption that $f \in \mathcal{E}_N$ can be replaced by $f \in \mathcal{K}_N(0, \infty)$ with locally integrable elements.

In most cases of interest the elements of $t^p k$ are uniformly bounded on $[0, \infty)$ for $p = 0,1,2$. In such cases $t^p k \in \mathcal{K}_{1N} \cap \mathcal{K}_{2N}$ for $p = 0,1,2$ provided that $t^2 k \in \mathcal{K}_{1N}$.

IV. AN APPLICATION: A FREQUENCY-DOMAIN CONDITION FOR THE $\mathcal{L}_\infty$-STABILITY OF FEEDBACK SYSTEMS CONTAINING A SINGLE TIME-VARYING NONLINEAR ELEMENT

In a recent brief,[2] a two-part sufficient condition is given for the $\mathcal{L}_\infty$-stability of a well-known type of feedback system containing a single time-varying nonlinear element. In another publication,[4] conditions are presented for the $\mathcal{L}_2$-stability of the same type of feedback system. Unlike the conditions for $\mathcal{L}_2$-stability of Ref. 4, which are expressed entirely in the frequency domain, the key condition of Ref. 2 for $\mathcal{L}_\infty$-stability is that the integral of the modulus of a certain function be less than unity.

A direct application of Theorem 2 shows that under somewhat stronger assumptions than those of Ref. 2 or Ref. 4 concerning $k(\cdot)$, there the impulse-response function of the linear time-invariant portion of the forward path, the conditions given for $\mathcal{L}_2$-stability are also sufficient conditions for $\mathcal{L}_\infty$-stability. Specifically, the following result is a direct consequence of Theorem 2.

Theorem 3: The feedback system described in Ref. 2 is $\mathcal{L}_\infty$-stable if

$$(i) \int_0^\infty |t^p k(t)|\, dt < \infty \text{ and } \int_0^\infty |t^p k(t)|^2\, dt < \infty \text{ for } p = 0, 1, 2$$

$$(ii) \text{ with } K(s) = \int_0^\infty k(t) e^{-st}\, dt \quad \text{for } \sigma \geqq 0,$$

$$(a)\ 1 + \tfrac{1}{2}(\alpha + \beta) K(s) \neq 0 \quad \text{for } \sigma \geqq 0$$

$$(b)\ \tfrac{1}{2}(\beta - \alpha) \max_{-\infty < \omega < \infty} |K(i\omega)[1 + \tfrac{1}{2}(\alpha + \beta) K(i\omega)]^{-1}| < 1.$$

Part (b) of (ii) above is a weaker condition than the condition of the theorem of Ref. 2 that it replaces [i.e., (ii) of Ref. 2]. From an engineer-

ing viewpoint condition (*ii*) above possesses an interesting frequency-domain interpretation.[4]†

V SUFFICIENT CONDITIONS FOR THE BOUNDEDNESS OF SOLUTIONS OF THE DISCRETE ANALOG OF THE INTEGRAL EQUATION CONSIDERED IN THEOREM 2

Sufficient conditions for the boundedness of the solutions of the discrete analog of the nonlinear integral equation considered in Theorem 2 can be obtained by modifying in a straightforward manner both the arguments presented in Section III and the arguments of Ref. 1 that lead to Theorem 1. In order to state the result (Theorem 2′, below) we need some notation.

Let Ξ denote the set of *nonnegative* integers. Let $\tilde{\mathcal{H}}_N$ be the set of real N-vector-valued functions defined on Ξ, and let the jth component of $f \in \tilde{\mathcal{H}}_N$ be denoted by f_j. Let

$$\tilde{\mathcal{L}}_{\infty N} = \{f \mid f \in \tilde{\mathcal{H}}_N, \sup_{n \geqq 0} [f'(n)f(n)] < \infty\},$$

$$\tilde{\mathcal{L}}_{2N} = \{f \mid f \in \tilde{\mathcal{H}}_N, \sum_{n=0}^{\infty} f'(n)f(n) < \infty\},$$

and

$$\| f \|_{\sim} = (\sum_{n=0}^{\infty} f'(n)f(n))^{\frac{1}{2}} \quad \text{for} \quad f \in \tilde{\mathcal{L}}_{2N}.$$

With B an arbitrary real $N \times N$ matrix-valued function of n with elements $\{b_{lm}(n)\}$ defined on Ξ, let $\tilde{\mathcal{K}}_{pN}(p = 1,2)$ denote

$$\{B \mid \sum_{n=0}^{\infty} | b_{lm}(n)|^p < \infty \ (l,m = 1, 2, \cdots, N)\}.$$

Let $\varphi[f(n),n]$ denote

$$(\varphi_1[f_1(n),n],\varphi_2[f_2(n),n], \cdots, \varphi_N[(f_N(n),n])', \qquad f \in \tilde{\mathcal{H}}_N$$

where $\varphi_1(w,n),\varphi_2(w,n), \cdots, \varphi_N(w,n)$ are real-valued functions of w and n for $-\infty < w < \infty$ and $n \in \Xi$ such that

(*i*) $\varphi_m(0,n) = 0$ for $n \in \Xi$ and $m = 1, 2, \cdots, N$

(*ii*) there exist real numbers α and β with the property that

† We take this opportunity to correct the result of a typographical error: In the first inequality on page 1606 of Ref. 4 the "$<$" sign should be replaced by "$\leqq$".

$$\alpha \leqq \frac{\varphi_m(w,n)}{w} \leqq \beta \qquad (m = 1, 2, \cdots, N)$$

for all real $w \neq 0$ and $n \varepsilon \Xi$.

Theorem 2′: Let $n^2 k \varepsilon \tilde{\mathcal{K}}_{1N}$. *Let*

$$g(n) = f(n) + \sum_{m=0}^{n} k(n - m)\varphi[f(m),m], \qquad n \varepsilon \Xi$$

where $g \varepsilon \tilde{\mathcal{L}}_{\infty N}$ *and* $f \varepsilon \tilde{\mathcal{K}}_N$. *Let*

$$K(s) = \sum_{n=0}^{\infty} k(n)e^{-sn}, \qquad \sigma \geqq 0.$$

Suppose that

(*i*) *det* $[1_N + \frac{1}{2}(\alpha + \beta)k(0)] \neq 0$, *and*

det $[1_N + \frac{1}{2}(\alpha + \beta)K(s)] \neq 0$ *for* $\sigma \geqq 0$

(*ii*) $\frac{1}{2}(\beta - \alpha) \sup_{-\pi \leqq \omega \leqq \pi} \Lambda\{[1_N + \frac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

Then $f \varepsilon \tilde{\mathcal{L}}_{\infty N}$, *there exists a positive constant* c *which depends only on* k, α, *and* β *such that*

$$\max_j \sup_{n \geqq 0} |f_j(n)| \leqq c \max_j \sup_{n \geqq 0} |g_j(n)|,$$

and $f_j(n) \to 0$ *as* $n \to \infty$ *for* $j = 1, 2, \cdots, N$ *whenever* $g_j(n) \to 0$ *as* $n \to \infty$ *for* $j = 1, 2, \cdots, N$.

In the statement of Theorem 2′ we have used the fact that $n^p k \varepsilon \tilde{\mathcal{K}}_{1N} \cap \tilde{\mathcal{K}}_{2N}$ for $p = 0, 1, 2$ provided that $n^2 k \varepsilon \tilde{\mathcal{K}}_{1N}$.

The result analogous to Theorem 1 is the following theorem.

Theorem 1′: Let $k \varepsilon \tilde{\mathcal{K}}_{1N}$, *and let*

$$g(n) = f(n) + \sum_{m=0}^{n} k(n - m)\varphi[f(m),m], \qquad n \varepsilon \Xi$$

where $g \varepsilon \tilde{\mathcal{L}}_{2N}$ *and* $f \varepsilon \tilde{\mathcal{K}}_N$. *Let*

$$K(s) = \sum_{n=0}^{\infty} k(n)e^{-sn}, \qquad \sigma \geqq 0.$$

Suppose that

(*i*) *det* $[1_N + \frac{1}{2}(\alpha + \beta)k(0)] \neq 0$, *and*

det $[1_N + \frac{1}{2}(\alpha + \beta)K(s)] \neq 0$ *for* $\sigma \geqq 0$.

(*ii*) $\frac{1}{2}(\beta - \alpha) \sup_{-\pi \leqq \omega \leqq \pi} \Lambda\{[1_N + \frac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

Then $f \in \tilde{\mathfrak{L}}_{2N}$, *and there exists a positive constant* ρ *which depends only on* k, α, *and* β *such that*

$$\| f \|_{\sim} \leqq \rho \| g \|_{\sim} .$$

VI. SOME ADDITIONAL RESULTS

Arguments very similar to those of Section III and the proof of the lemma of Ref. 5 can be used to establish the following result, which is of direct interest in the study of the properties of solutions of systems of differential equations.

Theorem 3: Let $t^p k \in \mathfrak{K}_{1N} \cap \mathfrak{K}_{2N}$ *for* $p = 0, 1, 2$. *Let* $Q(\cdot)$ *denote a real measurable* $N \times N$ *matrix-valued function of* t *defined on* $[0, \infty)$, *and let the elements of* $Q(t)$ *be uniformly bounded on* $[0, \infty)$. *Let*

$$g(t) = f(t) + \int_0^t k(t - \tau) Q(\tau) f(\tau)\, d\tau, \qquad t \geqq 0$$

where $g \in \mathfrak{L}_{\infty N}(0, \infty)$ *and* $f \in \mathcal{E}_N$. *With*

$$K(i\omega) = \int_0^{\infty} k(t) e^{-i\omega t}\, dt \quad \textit{for} \quad -\infty < \omega < \infty,$$

let

$$\sup_{t \geqq 0} \Lambda\{Q(t)\} \sup_{-\infty < \omega < \infty} \Lambda\{K(i\omega)\} < 1.$$

Then $f \in \mathfrak{L}_{\infty N}(0, \infty)$, *there exists a positive constant* c *which depends only on* $k(\cdot)$ *and* $Q(\cdot)$ *such that*

$$\max_j \sup_{t \geqq 0} |f_j(t)| \leqq c \max_j \sup_{t \geqq 0} |g_j(t)|,$$

and $f_j(t) \to 0$ *as* $t \to \infty$ *for* $j = 1, 2, \cdots, N$ *whenever* $g_j(t) \to 0$ *as* $t \to \infty$ *for* $j = 1, 2, \cdots, N$.

Theorem 3 remains valid if the sets $\mathfrak{K}_{1N}$, $\mathfrak{K}_{2N}$, $\mathfrak{L}_{\infty N}(0, \infty)$, and $\mathcal{E}_N$ are replaced with their natural complex extensions, and $Q(\cdot)$ is permitted to be complex valued.

A result that can easily be proved with the aid of Theorem 3 (see the proofs of the theorem and corollary of Ref. 5) is as follows.

Theorem 4: Let $\psi(\cdot, \cdot)$ *be as defined in Section I with* $N = 1$ *and* $\alpha > 0$, *and let* f *be any real-valued function of* t *defined and twice differentiable on* $[0, \infty)$ *such that*

$$\frac{d^2 f}{dt^2} + a \frac{df}{dt} + \psi[f, t] = g, \qquad t \geqq 0$$

where $g(t)$ is uniformly bounded on $[0,\infty)$. Suppose that a is a real constant such that $a > \sqrt{\beta} - \sqrt{\alpha}$. Then $f(t)$ is uniformly bounded on $[0,\infty)$, and $f(t) \to 0$ as $t \to \infty$ if $g(t) \to 0$ as $t \to \infty$.

The following theorem, which can be proved with arguments very similar to those of Section III and the proof of Theorem 5 of Ref. 1, is of immediate interest in the theory of stability of electrical networks containing time-varying capacitors.[6]

Theorem 5: Let $t^p k \in \mathcal{K}_{1N} \cap \mathcal{K}_{2N}$ for $p = 0, 1, 2$. Let B denote a constant real $N \times N$ matrix, and let $a_1(t), a_2(t), \cdots, a_N(t)$ denote real-valued measurable functions of the real variable t for $t \geqq 0$ with the property that there exist real constants α and β such that

$$\alpha \leqq a_n(t) \leqq \beta \qquad (n = 1, 2, \cdots, N)$$

for $t \geqq 0$. Let $A(t) = diag\,[a_1(t), a_2(t), \cdots, a_N(t)]$ for $t \geqq 0$, and let

$$g(t) = A(t)f(t) + Bf(t) + \int_0^t k(t-\tau)f(\tau)\,d\tau, \qquad t \geqq 0$$

where $g \in \mathcal{L}_{\infty N}(0,\infty)$ and $f \in \mathcal{E}_N$. Suppose that

(i) $det\,[\frac{1}{2}(\alpha+\beta)1_N + B] \neq 0$, $det\,[A(t) + B] \neq 0$ for $t \geqq 0$, and $\sup_{t \geqq 0} \Lambda\{[A(t) + B]^{-1}\} < \infty$;

and that, with

$$K(s) = \int_0^\infty k(t)e^{-st}\,dt \quad for \quad \sigma \geqq 0,$$

(ii) $det\,[\frac{1}{2}(\alpha+\beta)1_N + B + K(s)] \neq 0$ for $\sigma \geqq 0$

(iii) $\frac{1}{2}(\beta - \alpha) \sup_{-\infty<\omega<\infty} \Lambda\{[\frac{1}{2}(\alpha+\beta)1_N + B + K(i\omega)]^{-1}\} < 1$.

Then $f \in \mathcal{L}_{\infty N}(0,\infty)$, there exists a positive constant c which depends only on $A(\cdot)$, B, and k such that

$$\max_j \sup_{t \geqq 0} |f_j(t)| \leqq c \max_j \sup_{t \geqq 0} |g_j(t)|,$$

and $f_j(t) \to 0$ as $t \to \infty$ for $j = 1, 2, \cdots, N$ whenever $g_j(t) \to 0$ as $t \to \infty$ for $j = 1, 2, \cdots, N$.

Theorem 5 remains valid if the sets $\mathcal{K}_{1N}$, $\mathcal{K}_{2N}$, $\mathcal{L}_{\infty N}(0,\infty)$ and $\mathcal{E}_N$ are replaced with their natural complex extensions and B is permitted to be complex valued.

REFERENCES

1. Sandberg, I. W., On the $\mathfrak{L}_2$-Boundedness of Solutions of Nonlinear Functional Equations, B.S.T.J., *43*, July, 1964, p. 1581.
2. Sandberg, I. W., A Condition for the $\mathfrak{L}_\infty$-Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element, B.S.T.J., *43*, July, 1964, p. 1815.
3. Bochner, S., and Chandrasekharan, K., *Fourier Transforms*, Princeton University Press, Princeton, New Jersey, 1949, p. 99.
4. Sandberg, I. W., A Frequency-Domain Condition for the Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element, B.S.T.J., *43*, July, 1964, p. 1601.
5. Sandberg, I. W., On the Solutions of Systems of Second-Order Differential Equations with Variable Coefficients, to appear in the SIAM Journal on Control, Vol. 2, No. 2.
6. Sandberg, I. W., A Stability Criterion for Linear Networks Containing Time-Varying Capacitors, to appear in IEEE-PTGCT, March, 1965.

Imaging of Optical Modes — Resonators with Internal Lenses

By HERWIG KOGELNIK

(Manuscript received November 10, 1964)

This paper discusses the modes of optical resonators, and optical modes of propagation or Gaussian beams of light. The passage of Gaussian beams through lenses, telescopes, sequences of lenses, and lenslike media is studied. Mode matching formulae are derived. A complex beam parameter is introduced for which the law of transformation by any given optical structure can be written in the simple form of a bilinear transformation (ABCD law). Resonators with internal optical elements and their transmission line duals are also investigated. Effective Fresnel numbers and curvature parameters are determined which allow one to infer the diffraction losses, the resonant conditions, and the mode patterns of the various systems. Results are obtained for resonators with internal lenses, sequences of lenses with irises inserted between the lenses, resonators with internal lenslike media, transmission lines consisting of a lenslike medium with periodically spaced irises, and resonators with one very large mirror.

I. INTRODUCTION

The theory of Fresnel diffraction is the basis for an understanding of optical resonators[1-5] and of optical modes of propagation.[2,3,4] Fresnel diffraction explains the mode patterns and diffraction losses of optical resonators, and the beam waist and spreading of the modes of propagation or "Gaussian beams." In this paper we will discuss how these Gaussian beams of light are transformed on their passage through lenses, telescopes, various lens combinations, and lenslike (guiding) media, and how these optical systems affect the properties of optical resonators when inserted between the resonator mirrors.

We will assume that no additional aperture diffraction effects are introduced by these optical systems, i.e., that the apertures of the internal lenses can be regarded as infinitely large. The imaging laws of geometrical optics are therefore expected to apply, and we will use them

wherever possible, as they generally simplify the algebraic derivations and at the same time provide some physical insight.

Some of the problems to be investigated here in greater detail have already been treated in the literature. Goubau[6] has given some mathematical relations between the parameters of Gaussian beams transformed by a thin lens. The recently published mode matching formulae[7] are the result of a computation which will now be presented. Resonators with internal lenses have also been discussed in the literature,[8-11] and we have used the concept of an effective distance[9,10] in a previous publication.[9] In several cases an alternative to our algebraic approach is the graphical method of Collins,[11] who introduced the circle diagram[11,12] for Gaussian beams.

In the following we will first establish the rules of imaging for Fresnel diffraction with attention to the imaging of the phase fronts which are of particular importance for optical modes. Then we will list expressions for the focal length and the principal planes of various optical systems of interest, because these parameters are needed later for application of the imaging rules. This listing includes the parameters of the telescope, of sequences of lenses, and of sections of lenslike medium. Armed with these tools we will study the passage of Gaussian beams through lenses and various optical systems. The paper is concluded by an investigation of optical resonators with internal optical elements and their transmission line duals. Effective Fresnel numbers and curvature parameters are determined which allow one to infer the diffraction losses, the resonant conditions, and the mode patterns of the various systems. Results are obtained for resonators with internal lenses, sequences of lenses with irises inserted between the lenses, resonators with internal lenslike media, transmission lines consisting of a guiding medium with periodically spaced irises, and resonators with one very large mirror.

II. IMAGING RULES

While geometrical optics deals with rays, the theory of Fresnel diffraction deals with (scalar) fields. To describe the field distribution, we use complex amplitudes $E(x,y,z)$ and a Cartesian (x,y,z) coordinate system. We consider a wave that propagates in the direction of the optic axis (z axis). Within the assumptions of Fresnel diffraction an ideal thin lens of focal length f transforms the incoming wave with a field $E_{\text{left}}(x,y,z = \text{const})$ immediately to the left of the lens into a wave with the field

$$E_{\text{right}}(x,y,z = \text{const}) = E_{\text{left}}(x,y,z = \text{const}) \exp\left(-jk\,\frac{x^2 + y^2}{2f}\right) \quad (1)$$

immediately to the right of the lens. Here k is the propagation constant. The thin lens produces a phase shift which is proportional to the square of the distance to the optic axis, while the intensity distribution is the same on both sides of the lens.

Consideration of spherical waves provides a link between (1) and the laws of geometrical optics according to which a spherical wave with a radius of curvature R_1 at the left of the lens is transformed into a wave with curvature radius R_2 as shown in Fig. 1. The radii R_1 and R_2 are related by

$$(1/R_1) + (1/R_2) = 1/f. \tag{2}$$

For Fresnel diffraction the transverse field distribution of a wave with a spherical phase front of radius R is given[2,3] by

$$E(x,y,z = \text{const}) = \exp(-jkr^2/2R) \tag{3}$$

where

$$r^2 = x^2 + y^2, \tag{4}$$

and R is counted positive for a phase front that is concave if observed from the left. For spherical phase fronts of radius R_1 on the left and $-R_2$ on the right of the lens (where the phase front curvature is negative, as shown in Fig. 1) we can express E_{left} and E_{right} with the help of (3), compare the exponents in (1), and find the same relation (2) between R_1, R_2, and f as for the spherical waves of geometrical optics.

To discuss imaging consider an object, i.e., the field $E_1(x_1, y_1)$ in an object plane, and its image $E_2(x_2, y_2)$ in the corresponding image plane (see Fig. 2). The distances d_1 and d_2 between the lens and the two planes are related by

$$(1/d_1) + (1/d_2) = 1/f. \tag{5}$$

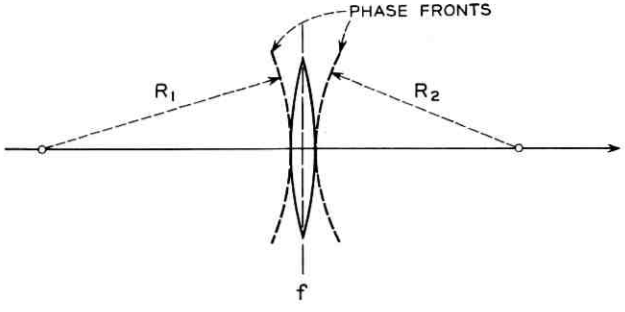

Fig. 1 — Lens transforming phase front of spherical wave.

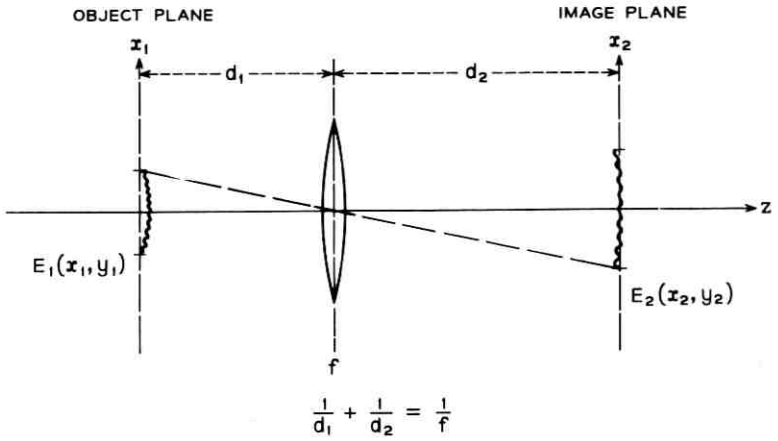

Fig. 2 — Imaging of field distribution by a thin lens.

We know from geometrical optics that the intensity distributions of the object and the image are similar. This is, of course, still true for Fresnel diffraction by any field aperture in the object plane. Assuming that no aperture diffraction effects are introduced by the thin lens, one can use the Fresnel diffraction formula to relate E_1 and E_2 (see Appendix A) and arrive at

$$E_2(x_2, y_2) = -\frac{d_1}{d_2} E_1\left(-\frac{d_1}{d_2} x_2, -\frac{d_1}{d_2} y_2\right) \cdot \exp - jk\left(d_1 + d_2 + \frac{r_2^2}{2f}\frac{d_1}{d_2}\right) \tag{6}$$

with $r_2^2 = x_2^2 + y_2^2$. The factor d_1/d_2 in this equation follows from conservation of energy; the arguments $-(d_1/d_2)x_2$ and $-(d_1/d_2)y_2$ indicate that the image is inverted and magnified by d_2/d_1. The first two terms in the exponent are simply due to the phase shift $k(d_1 + d_2)$ which the light wave suffers in propagating from the object to the image plane, while the third and last term is of particular importance for our considerations. It describes an additional phase shift proportional to r_2^2 which appears in the field distribution of the image. Apart from this additional phase shift the amplitude and phase distribution of the image and the object are scales of each other.

The expression for the additional phase shift follows also from geometrical optics (see e.g. Appendix B), and it is related to the thick-mirror formulae,[13] as we shall see later. It is also obtained by studying

the passage of Gaussian beams through a lens.[6] The additional phase shift does not appear in Abbe's theory of imaging; he finds that the image is strictly similar to the object, both as regards the amplitude and phase distribution.[14] But Abbe used the Fraunhofer diffraction theory, where phase terms proportional to r^2 are neglected.

For Fresnel diffraction the r^2 dependence of the additional phase shift suggests that one should use spherical reference surfaces instead of plane ones, as shown in Fig. 3. By proper choice of the curvature of these surfaces tangential to the image and object planes, one can achieve an image field on one surface that strictly reproduces the object on the other surface in amplitude and phase. For an object reference surface of radius R_1 and an image reference surface of radius R_2 one gets for the fields additional phase factors of exp $(-jkr_1^2/2R_1)$ and exp $(-jkr_2^2/2R_2)$, respectively. These phase factors cancel the additional phase shift in (6) if

$$\frac{1}{R_2} = \frac{1}{R_1}\frac{d_1^2}{d_2^2} + \frac{1}{f}\frac{d_1}{d_2}. \tag{7}$$

After some algebraic manipulations involving (5) this relation can be rewritten as

$$\frac{1}{d_1 + R_1} + \frac{1}{d_2 - R_2} = \frac{1}{f}. \tag{8}$$

This simply means that the center of curvature C_1 of the object surface

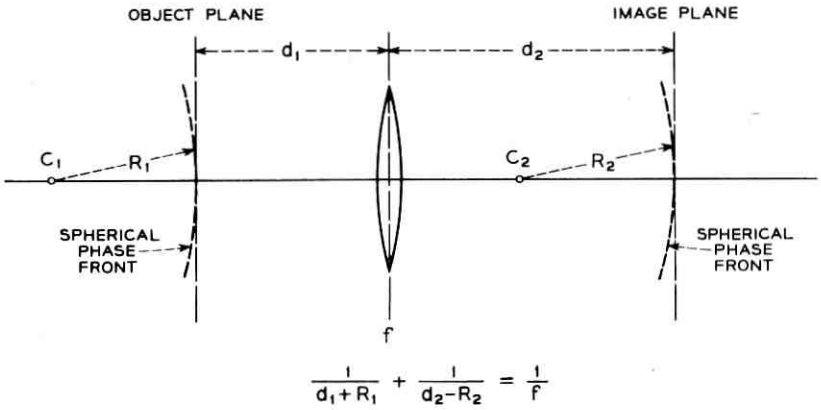

Fig. 3 — Imaging of fields with spherical wave fronts; centers of curvature are images of each other. The corresponding spherical reference surfaces are used when fields with nonspherical phase fronts are imaged.

is imaged onto the center of curvature C_2 of the image surface. Thus, whenever the centers of curvature of the image and object surfaces are images of each other we have an image which is a strict (scaled) reproduction of the object as regards both the amplitude and phase distribution, with no additional phase shift.

The imaging rules discussed above can also be used to study imaging by a combination of lenses (or by any optical system that can be regarded as such). It is not necessary to apply the rules step by step to each individual thin lens of the combination. It is generally simpler to determine the parameters of the equivalent thick lens as usual in geometrical optics. The place of f is then taken by the combined focal length of the system, and object and image distances (d_1 and d_2) are measured from the principal planes of the thick lens.

III. FOCAL LENGTHS OF VARIOUS OPTICAL SYSTEMS

3.1 *The Ray Matrix*

When one traces a paraxial ray through combinations of lenses and lenslike media, the quantities of interest are the position x_1 and the slope x_1' of the ray in the input plane, and the corresponding quantities x_2 and x_2' in the output plane (see Fig. 4). There is in general a linear relation[15,16,17] between the output and input quantities which can be written in matrix form as

$$\begin{vmatrix} x_2 \\ x_2' \end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \begin{vmatrix} x_1 \\ x_1' \end{vmatrix}. \tag{9}$$

We will call this $ABCD$ matrix the ray matrix of the system. Because of reciprocity the determinant of the ray matrix is generally unity:

$$AD - BC = 1. \tag{10}$$

It is easy to determine the focal length and the principal planes from the elements of the ray matrix of an optical system. By tracing a beam that leaves the output plane parallel to the optic axis ($x_2' = 0$) we find the location of the focal point on the input side. Its distance s_1 from the input plane is obtained as

$$s_1 = \frac{x_1}{x_1'}\bigg|_{x_2'=0} = -\frac{D}{C} \tag{11}$$

where we refer to Fig. 4. Similarly, we find for the distance s_2 between

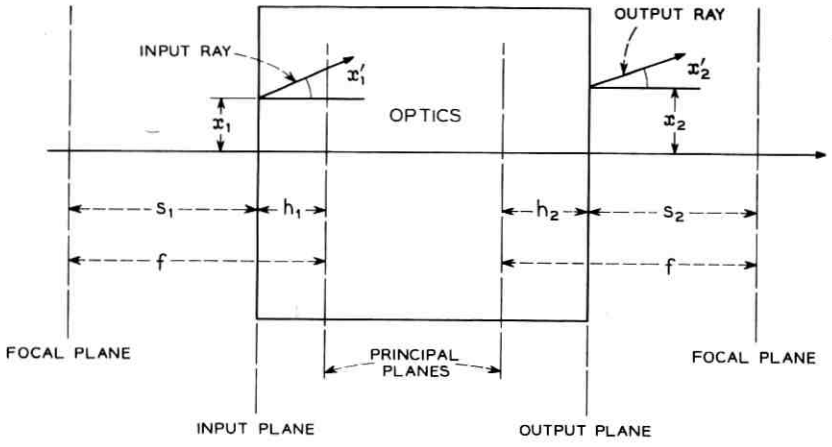

Fig. 4 — Reference planes for optical system.

the output plane and the corresponding focal point

$$s_2 = -\frac{x_2}{x_2'}\bigg|_{x_1'=0} = -\frac{A}{C}. \tag{12}$$

To find the principal plane on the input side we follow an input ray from the focal point until its distance from the axis is equal to the position x_2 of the corresponding output ray and have

$$h_1 = \frac{x_2 - x_1}{x_1'}\bigg|_{x_2'=0} \tag{13}$$

where the distance h_1 between the principal plane and the input plane is measured positive as shown in Fig. 4. On the output side we find similarly

$$h_2 = \frac{x_2 - x_1}{x_2'}\bigg|_{x_1'=0}. \tag{14}$$

The focal length f of the system is obtained by calculating the distance between a principal plane and the corresponding focal point

$$f = s_1 + h_1 = s_2 + h_2 = \frac{x_2}{x_1'}\bigg|_{x_2'=0} = -\frac{x_1}{x_2'}\bigg|_{x_1'=0}. \tag{15}$$

Using the linear relations of (9) together with the last three expressions, one finally gets

$$f = -(1/C) \tag{16}$$

$$h_1 = (D - 1)/C \tag{17}$$

$$h_2 = (A - 1)/C \tag{18}$$

where the thick-lens parameters are expressed in terms of the elements of the ray matrix. For later use we also write down the matrix elements as functions of the lens parameters which follow from the last expressions

$$A = 1 - (h_2/f) \tag{19}$$

$$B = h_1 + h_2 - (h_1h_2/f) \tag{20}$$

$$C = -(1/f) \tag{21}$$

$$D = 1 - (h_1/f). \tag{22}$$

3.2 *The Two-Lens Combination—Telescope*

The lens parameters of a combination of two lenses are well known and are listed here for completeness and for later use. The combination is shown in Fig. 5. For lenses of focal lengths f_1 and f_2 spaced at a distance d we have

$$1/f = (1/f_1) + (1/f_2) - (d/f_1f_2) \tag{23}$$

$$h_1 = \frac{df}{f_2} \tag{24}$$

$$h_2 = \frac{df}{f_1} \tag{25}$$

where the lens planes are used as input and output planes.

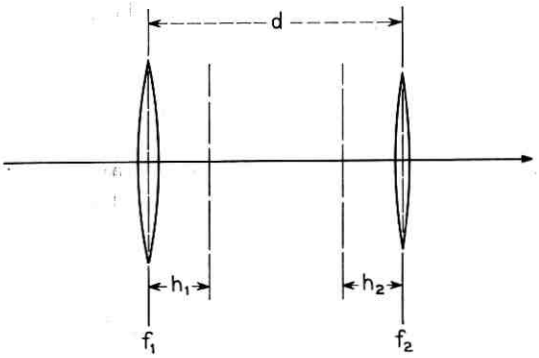

Fig. 5 — The two-lens combination.

For a slightly misadjusted telescope the lens spacing is

$$d = f_1 + f_2 - \Delta d \tag{26}$$

where Δd measures the misadjustment. The lens parameters of the telescope can be written as

$$f = \frac{f_1 f_2}{\Delta d} \tag{27}$$

$$h_1 = \frac{f_1 d}{\Delta d} \tag{28}$$

$$h_2 = \frac{f_2 d}{\Delta d}. \tag{29}$$

3.3 *Sequence of Lenses*

A periodic sequence of lenses of equal focal length f_0 and lens spacing d is shown in Fig. 6. The reference planes are chosen just to the right of each lens. The elements of the ray matrix $\hat{S}$ of one section of the sequence (i.e., one lens spaced at a distance d from the input plane) are well known[15,17] and are given by

$$\hat{S} = \begin{vmatrix} 1 & d \\ -\dfrac{1}{f_0} & 1 - \dfrac{d}{f_0} \end{vmatrix}. \tag{30}$$

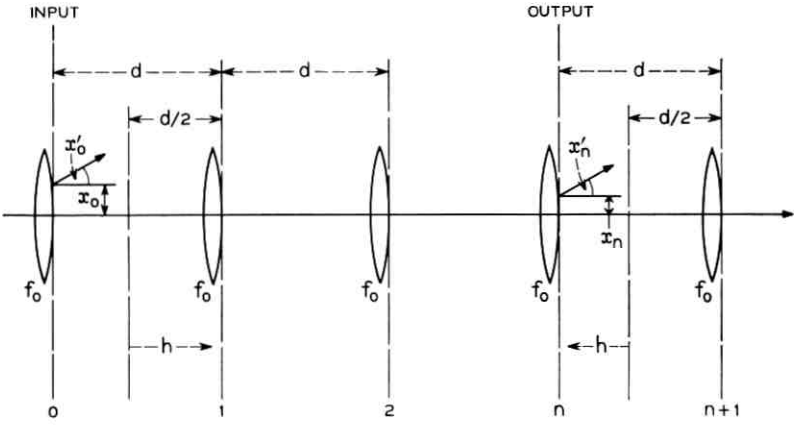

Fig. 6 — Sequence of lenses of equal focal length.

They relate the position and slope (x_1 and x_1') of the ray just after the first lens to the ray position and slope just after the zeroth lens

$$\begin{vmatrix} x_1 \\ x_1' \end{vmatrix} = \hat{S} \begin{vmatrix} x_0 \\ x_0' \end{vmatrix}. \tag{31}$$

The ray to the right of the nth lens is related to the input ray by the nth power of the ray matrix of one section

$$\begin{vmatrix} x_n \\ x_n' \end{vmatrix} = \hat{S}^n \begin{vmatrix} x_0 \\ x_0' \end{vmatrix}. \tag{32}$$

The matrix elements of $\hat{S}^n$ can be computed with the help of Sylvester's theorem[18] and are well known.[15,16] One has*

$$\hat{S}^n = \frac{1}{\sin\theta} \cdot \begin{vmatrix} \sin n\theta - \sin(n-1)\theta & d \sin n\theta \\ -\frac{1}{f_0}\sin n\theta & \left(1 - \frac{d}{f_0}\right)\sin n\theta - \sin(n-1)\theta \end{vmatrix} \tag{33}$$

with

$$\cos\theta = 1 - (d/2f_0). \tag{34}$$

We can now employ (16) and obtain for the focal length f of n sections of a periodic sequence of lenses

$$f = f_0(\sin\theta/\sin n\theta). \tag{35}$$

The distance of the two principal planes from the input and output planes (zeroth and nth lens) follows also from (33) with the help of (17) and (18). One finds

$$h_1 = (d/2) + f(1 - \cos n\theta), \tag{36}$$

and

$$h_2 = -(d/2) + f(1 - \cos n\theta). \tag{37}$$

If we measure the distance h of the principal planes from the midplanes between the lenses as shown in Fig. 6 we have

$$h = f(1 - \cos n\theta). \tag{38}$$

* These matrix elements can be written in terms of Chebyshev polynomials of the second kind of the variable $[1 - (d/2f_0)]$.

A more complicated sequence of lenses is shown in Fig. 7. Here a lens of focal length f_1 is followed by a lens of focal length f_2 and vice versa. The lens spacings are d_1 and d_2 in sequence as shown in the figure. This sequence of lenses can be reduced to the simpler type discussed above. We can regard it as a sequence of thick lenses formed by lens pairs of focal lengths f_1 and f_2. The focal length f_0 of the thick lens is, according to (23), given by

$$1/f_0 = (1/f_1) + (1/f_2) - (d_1/f_1f_2), \tag{39}$$

and expressions for the principal planes are given in (24) and (25). The distance d between the output principal plane of a thick lens and the input principal plane of the consecutive thick lens is obtained as

$$d = d_2 + h_1 + h_2 = d_2 + f_0d_1\left(\frac{1}{f_1} + \frac{1}{f_2}\right). \tag{40}$$

With the principal planes as reference planes, rays passing through this sequence of thick lenses behave the same way as rays passing through a sequence of lenses of equal focal length that are equally spaced. We can therefore use the expressions (33) and (34) obtained above. With (39) and (40) the latter becomes

$$\cos\theta = 1 - \frac{d_1 + d_2}{2}\left(\frac{1}{f_1} + \frac{1}{f_2}\right) + \frac{d_1d_2}{2f_1f_2}. \tag{41}$$

3.4 *Lenslike Medium*

A lenslike medium or "guiding medium" is one whose refractive index n varies near the optic axis as in

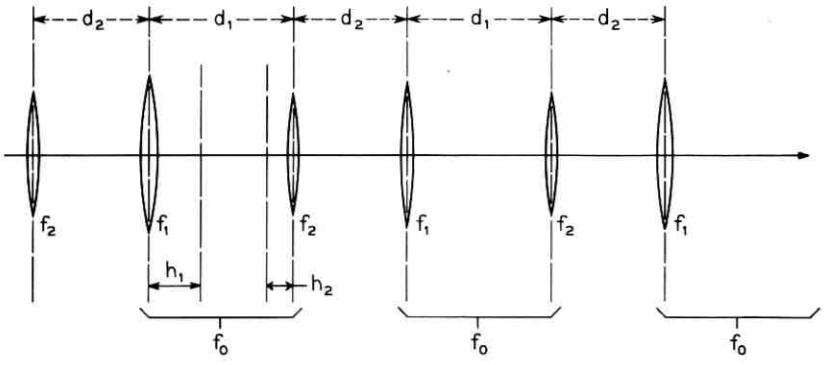

Fig. 7 — Sequence of lenses of alternating focal length with alternating lens spacings.

$$n = n_0\left(1 - 2\frac{r^2}{b^2}\right) \tag{42}$$

where n_0 is a constant, r is the distance from the optic axis, and b measures the degree of the variation of n. A medium of this kind can be produced by inhomogeneities in laser crystals[19,20] or by a radial variation of the gain in high-gain gaseous lasers.[21] Another important example is the medium of the recently reported gas lens.[22,23,24]

To trace rays in a lenslike medium one uses the differential equation for light rays.[25] For paraxial rays this ray equation has the form

$$n_0\frac{d^2x}{dz^2} = \frac{\partial}{\partial x}n = -4n_0\frac{x}{b^2} \tag{43}$$

for the distance $x(z)$ of the ray from the z axis. A corresponding relation holds for $y(z)$. The solution is, again, a linear relation between the ray position and slope in the output plane (x and x') and the corresponding input quantities x_0 and x_0'

$$\begin{vmatrix} x \\ x' \end{vmatrix} = \begin{vmatrix} \cos 2\frac{z}{b} & \frac{b}{2}\sin 2\frac{z}{b} \\ -\frac{2}{b}\sin 2\frac{z}{b} & \cos 2\frac{z}{b} \end{vmatrix} \begin{vmatrix} x_0 \\ x_0' \end{vmatrix}. \tag{44}$$

A typical ray path is shown in Fig. 8. To calculate the optical parameters for a section of lenslike medium immersed in a medium with a refractive index of unity (vacuum), we invoke Snell's law to relate the ray slopes

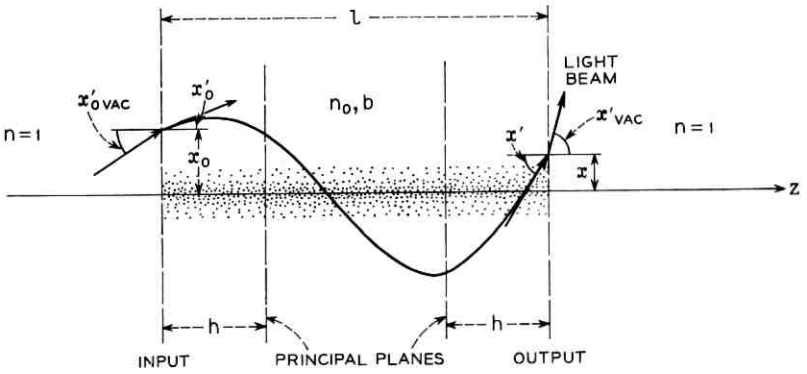

Fig. 8 — Ray path in lenslike medium.

at the section boundaries. For paraxial rays we have approximately

$$x_{\text{vac}} = n_0 x'; \qquad x_{0\text{vac}} = n_0 x_0'. \tag{45}$$

Now we use (16) and find for the focal length of a section of length l

$$f = \frac{b}{2n_0 \sin 2\dfrac{l}{b}} \tag{46}$$

(for $n_0 = 1$ this formula has been given in Refs. 23 and 26, for example). The distance h of the principal planes from the input and output planes respectively (see Fig. 8) is computed with the help of (17) and (18). One obtains

$$h = \frac{b}{2n_0} \tan \frac{l}{b}. \tag{47}$$

The above expressions have been derived for a focusing medium where $b^2 \geqq 0$. For a defocusing medium we have $b^2 \leqq 0$, and the expression for the focal length becomes

$$f = -\frac{|b|}{2n_0 \sinh 2\dfrac{l}{|b|}}. \tag{48}$$

The location of the corresponding principal planes is described by

$$h = \frac{|b|}{2n_0} \tanh \frac{l}{|b|}. \tag{49}$$

IV. OPTICAL TRANSFORMATION OF GAUSSIAN BEAMS

4.1 *Light Propagation in Free Space*

Near the optic axis an optical mode of propagation or Gaussian beam is regarded as a TEM wave with a spherical phase-front and a transverse field distribution that is described by Laguerre-Gaussian[2] or Hermite-Gaussian[3] functions. The two beam parameters of interest are the "spot size" or beam radius $w(z)$ and the radius of the phase front $R(z)$. In any beam cross section of a fundamental mode the field varies as

$$\exp\left(-\frac{r^2}{w^2} - jk\frac{r^2}{2R}\right)$$

and is specified by w and R. The light beam expands as it propagates through space as shown in Fig. 9. The law of expansion is[2,3,6,26,27]

$$w^2 = w_0^2\left[1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2\right]. \tag{50}$$

Here the z is measured from the beam waist where the phase front is plane and the beam reaches its minimum radius w_0. For $R(z)$ we have[2,3,6,26,27]

$$R = z\left[1 + \left(\frac{\pi w_0^2}{\lambda z}\right)^2\right]. \tag{51}$$

Dividing (50) by (51) we find

$$\frac{\pi w^2}{\lambda R} = \frac{\lambda z}{\pi w_0^2} \tag{52}$$

which we can use to rewrite the terms in the round brackets, and express w_0 and z in terms of w and R

$$w_0^2 = \frac{w^2}{1 + \left(\frac{\pi w^2}{\lambda R}\right)^2} \tag{53}$$

$$z = \frac{R}{1 + \left(\frac{\lambda R}{\pi w^2}\right)^2}. \tag{54}$$

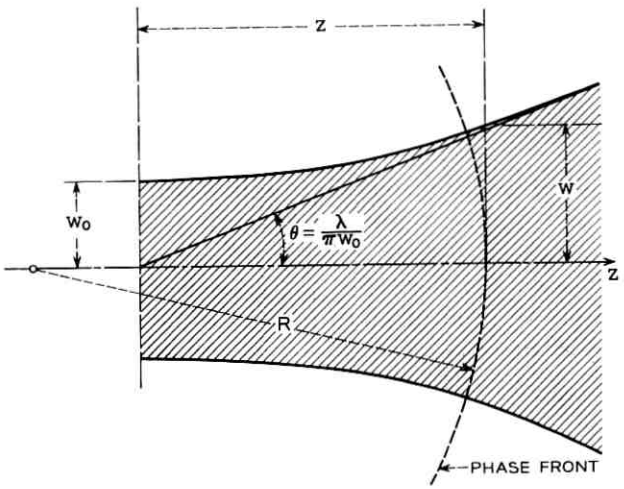

Fig. 9 — Contour of Gaussian beam of light.

4.2 *Beam Transformation by a Lens*

When a Gaussian beam passes through a lens a new beam waist is formed, and the parameters in the expansion laws are changed. Assume the light beam is propagating to the right. Before passing through the lens it has a beam waist a distance d_1 away from the lens with a beam radius w_1 as shown in Fig. 10. The lens produces another beam waist a distance d_2 away with a beam radius w_2. The distances d_1 and d_2 are measured positive as shown in the figure (for a negative d_1 one has a virtual waist). In the following we will establish some relationships between beam parameters of the incoming beam (identified by the subscript 1) and the parameters of the transformed beam (subscript 2).

The far field angles[27] θ_1 and θ_2 of the two beams are computed from (50) as

$$\theta_1 = \lambda/\pi w_1 ; \qquad \theta_2 = \lambda/\pi w_2 . \tag{55}$$

From these two angles follow immediately the beam radii w_{1f} and w_{2f} in the two focal planes of the lens where the image of the far field appears

$$w_{1f} = f\theta_2 = \lambda f/\pi w_2 \tag{56}$$

$$w_{2f} = f\theta_1 = \lambda f/\pi w_1 . \tag{57}$$

The beam radius in one of the focal planes is, of course, independent of the spacing between the lens and the beam waist of the other beam. It follows from (51) that the center of curvature of the far field phase front is in the beam waist. According to the imaging rules of Section II, corresponding centers of curvature are images of each other (where we take the phase fronts as reference surfaces). We therefore have to determine the image of a beam waist to find the curvature center of the phase front in the focal plane on the other side of the lens. The distance d_2' between the lens and the image of the waist w_2 follows from

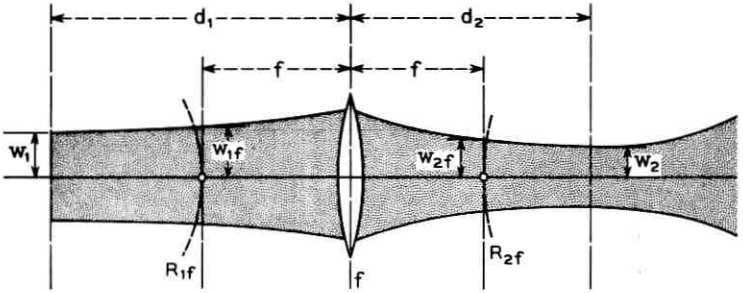

Fig. 10 — Gaussian beam transformed by a lens.

$$(1/d_2') + (1/d_2) = 1/f, \tag{58}$$

and the radius of curvature R_{1f} in the left focal plane is equal to the spacing between that image and the focal plane

$$R_{1f} = d_2' - f. \tag{59}$$

Combining (58) and (59) we have

$$R_{1f} = \frac{f^2}{d_2 - f} \tag{60}$$

and correspondingly

$$R_{2f} = \frac{f^2}{d_1 - f} \tag{61}$$

for the radius of curvature in the right focal plane. R_{1f} and R_{2f} are independent of the beam radii w_2 and w_1, respectively, a fact that can be used for mode matching into confocal resonators.

To relate the beam waists we use (56) and (60) to write

$$\frac{\pi {w_{1f}}^2}{\lambda R_{1f}} = \frac{\lambda(d_2 - f)}{\pi {w_2}^2} \tag{62}$$

and similarly

$$\frac{\pi {w_{2f}}^2}{\lambda R_{2f}} = \frac{\lambda(d_1 - f)}{\pi {w_1}^2}. \tag{63}$$

To express w_2 in terms of w_1 and d_1 we insert (57) and (63) into (53) and find

$$\frac{1}{{w_2}^2} = \frac{1}{{w_1}^2}\left(1 - \frac{d_1}{f}\right)^2 + \frac{1}{f^2}\left(\frac{\pi w_1}{\lambda}\right)^2. \tag{64}$$

This relation, first given by Goubau,[6] relates the beam radius of the waist of the transformed beam to the parameters of the incoming beam. A corresponding relation for the spacing d_2 between the lens and the beam waist w_2 is found by inserting (61) and (63) in (54). The result is

$$d_2 - f = (d_1 - f)\frac{f^2}{(d_1 - f)^2 + \left(\frac{\pi {w_1}^2}{\lambda}\right)^2}. \tag{65}$$

The above expressions were derived with the help of the imaging rules of Section II. As mentioned before, these rules apply not only to thin lenses but also to thick lenses and lens combinations. Therefore,

if d_1 and d_2 are measured from the principal planes the results given above are valid for the transformation of Gaussian beams by thick lenses.

4.3 *Mode Matching*

In experiments with optical modes one often wants to transform a beam with a given beam radius w_1 at the waist into another beam of waist radius w_2. One wants to "match" the modes of one optical system (like a laser resonator) to the modes of another one (an optical transmission line for example). This can be done by selecting a suitable lens and by properly adjusting the waist spacings d_1 and d_2, where we refer to Fig. 10. The proper spacings are given by the mode matching formulae[7] derived below.

We combine (62) or (63) with (52) and obtain

$$\frac{d_1 - f}{d_2 - f} = \frac{w_1^2}{w_2^2}. \tag{66}$$

This is used to rewrite (64) in the form

$$\frac{1}{w_2^2} = \frac{1}{w_2^2 f^2}(d_1 - f)(d_2 - f) + \frac{1}{f^2}\left(\frac{\pi w_1}{\lambda}\right)^2. \tag{67}$$

Multiplying (67) by $w_2^2 f^2$ we arrive at

$$(d_1 - f)(d_2 - f) = f^2 - f_0^2 \tag{68}$$

where we have defined

$$f_0 = \pi(w_1 w_2/\lambda). \tag{69}$$

To arrive at the mode-matching formulae we multiply or divide (68) by (66), extract the square root, and find

$$d_1 - f = \pm \frac{w_1}{w_2}\sqrt{f^2 - f_0^2} \tag{70}$$

or

$$d_2 - f = \pm \frac{w_2}{w_1}\sqrt{f^2 - f_0^2}. \tag{71}$$

As discussed in Ref. 7, one achieves mode matching by choosing a lens (or lens combination) with a focal length f that is larger than f_0 or equal to it. For a given lens there are generally two ways open to match the modes. One can choose either the plus sign in both (70) and

(71), or the minus sign. For $f = f_0$ there is only one set of proper spacings $d_1 = d_2 = f = f_0$.

4.4 *Complex Beam Parameter — ABCD Law*

In the foregoing we have used two parameters to characterize a Gaussian beam in a given beam cross section: the spot size or beam radius w, and the radius of phase front curvature R. We define now a more abstract complex beam parameter q

$$1/q = (1/R) - j(\lambda/\pi w^2). \tag{72}$$

The propagation and transformation laws for this beam parameter are particularly simple and allow one to trace Gaussian beams through more complicated optical structures. The old parameters R and w can, of course, be recovered from the real and imaginary parts of $1/q$. Note that we can regard the circle diagram of Collins[11] as plotted in the complex plane of the variable j/q, and the circle diagram of Li[12] as plotted in the complex plane of jq^*.

In terms of the complex beam parameter the laws of propagation (50) and (51) have the simple and compact form†

$$q = q_0 + z \tag{73}$$

as one easily verifies by inserting (50) and (51) into (72). Here

$$q_0 = j(\pi w_0^2/\lambda) \tag{74}$$

is the complex beam parameter at the beam waist. Because of the linearity of (73) the parameters q_1 and q_2 of two arbitrary beam cross sections are related by

$$q_2 = q_1 + d \tag{75}$$

where d is the distance between the two planes of interest measured positive in the direction of the optic axis.

The beam parameters q_1 and q_2 to the left and to the right of a lens are related by

$$1/q_2 = (1/q_1) - (1/f) \tag{76}$$

which simultaneously states the transformation of the phase fronts as in (1) or (2), and the fact that the beam radii (widths) are the same on both sides of the lens [compare (1)].

† Similar propagation laws for optical modes have been used independently by D. A. Kleinman, A. Ashkin, and G. D. Boyd in an analysis of second-harmonic generation in crystals and by G. A. Deschamps and P. E. Mast in their recent paper in Proc. Symp. Quasi-Optics, Polytechnic Inst. Brooklyn, 1964, p. 379.

The imaging law (6) applied to Gaussian beams takes the form

$$\frac{1}{q_2} = \frac{d_1^2}{d_2^2} \cdot \frac{1}{q_1} + \frac{1}{f}\frac{d_1}{d_2} \tag{77}$$

if written in terms of the complex parameters q_1 and q_2 of the beam in the object or image planes, respectively. Comparing with (7) and (8) one can also write this relation between the parameters of the object and the image as

$$\frac{1}{d_1 + q_1} + \frac{1}{d_2 - q_2} = \frac{1}{f}. \tag{78}$$

Using (75) and (76) one can easily determine how an incoming beam with the parameter q_1 at a distance d_1 from a lens is transformed. The parameter q_2 of the transformed beam at a distance d_2 from the lens is obtained as

$$q_2 = \frac{\left(1 - \frac{d_2}{f}\right) q_1 + \left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)}{-\frac{q_1}{f} + \left(1 - \frac{d_1}{f}\right)}. \tag{79}$$

To establish a link to the transformation laws for the real parameters developed before, we multiply both sides of (79) with the denominator of the right side. Then we postulate that we have beam waists at d_1 and d_2 by putting $q_1 = j\pi w_1^2/\lambda$ and $q_2 = j\pi w_2^2/\lambda$. If we compare the real parts of the resulting expression, we obtain relation (68), and comparing the imaginary parts we find (66).

Let us now regard q_1 and q_2 as the beam parameters in the input and output planes of an optical system described by its ray $(ABCD)$ matrix as in Section III. This system is also described by its focal length and its principal planes as calculated from (16), (17), and (18). To relate q_1 and q_2 we use (79) and put $d_1 = h_1$, and $d_2 = h_2$. Comparing with (19), (20), (21), and (22) we see that

$$q_2 = \frac{Aq_1 + B}{Cq_1 + D} \tag{80}$$

which we shall call the $ABCD$ law. The q parameters of the input and the output are related by this bilinear transformation. The $ABCD$ law says that the constants of this transformation are equal to the elements of the ray matrix. The ray matrices of several optical structures are given in Section III, and we shall use the $ABCD$ law to study the passage of Gaussian beams through some of these structures.

There appears to be a very close connection between Gaussian light

beams and the spherical waves of geometrical optics. In fact, all the important laws of this chapter are formally the same for a spherical wave with a radius of curvature q. One is therefore tempted to regard a Gaussian beam as a spherical wave with a complex radius of curvature. For the limit of infinitely small wavelengths the curvature radius becomes real and one has a spherical wave of geometrical optics.

The $ABCD$ law allows also a kind of "black box" approach to the study of optical modes. One can, for example, inquire about the mode parameters of a sequence of equal black boxes, i.e., optical structures characterized by their ray matrix elements A, B, C, and D. For a mode the beam parameter at the output of a black box is equal to the parameter at the input ($q_1 = q_2 = q$). From (80) follows a quadratic equation for the mode parameter q

$$Cq^2 + (D - A)q - B = 0. \tag{81}$$

The solution of this equation can be written as

$$\frac{1}{q} = \frac{D - A}{2B} \mp \frac{j}{2B} \sqrt{4 - (A + D)^2} \tag{82}$$

from which one can obtain the beam radius or spot size of the mode and the radius of curvature of its phase front.

4.5 *Beam Transformation by a Telescope*

In this chapter we shall study the passage of a Gaussian beam through a telescope consisting of two lenses of focal length f_1 and f_2, respectively, spaced at a distance $d = f_1 + f_2 - \Delta d$. The "misadjustment" Δd is assumed small. The focal length and the location of the principal planes of the telescope are given in (27), (28), and (29). We consider an incoming beam with a beam radius w_1 at its waist, and the waist spaced at a distance s_1 from the first lens as shown in Fig. 11. We want to determine the location s_2 of the waist of the outgoing beam and its beam radius w_2.

The distances of the waists to the corresponding principal planes are

$$d_1 = s_1 + h_1; \qquad d_2 = s_2 + h_2. \tag{83}$$

From this we find with (24) and (27)

$$\frac{d_1}{f} - 1 = \frac{f_1}{f_2} + \frac{\Delta d}{f_2}\left(\frac{s_1}{f_1} - 1\right). \tag{84}$$

Inserting this expression together with (27) in (64) we get for the beam

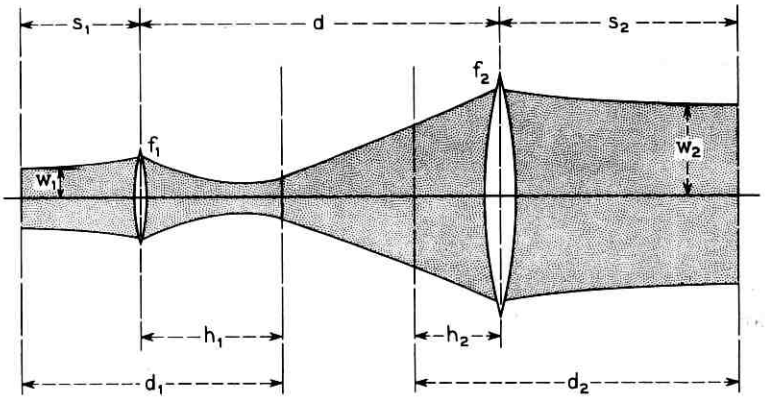

Fig. 11 — Gaussian beam passing through a telescope.

waist w_2

$$w_2 = w_1 \frac{f_2}{f_1}\left[1 + \frac{\Delta d}{f_1}\left(1 - \frac{s_1}{f_1}\right)\right] \tag{85}$$

which is correct to first order in Δd. We see that the ratio of the beam waists is more or less equal to the ratio of the focal lengths of the lenses. There is only a slight dependence on the position of the input beam waist for $\Delta d \neq 0$.

To determine the location of the output beam waist we use (84) and the corresponding expression for $(d_2/f) - 1$ to rearrange (68) as

$$\frac{s_1 - f_1}{f_1^2} + \frac{s_2 - f_2}{f_2^2} = -\frac{\Delta d}{f_1 f_2}\left[\left(\frac{s_1}{f_1} - 1\right)\left(\frac{s_2}{f_2} - 1\right) + \frac{1}{f_1 f_2}\left(\pi \frac{w_1 w_2}{\lambda}\right)^2\right]. \tag{86}$$

Inserting (85) and expanding to first order in Δd this becomes

$$\frac{s_1 - f_1}{f_1^2} + \frac{s_2 - f_2}{f_2^2} = \frac{\Delta d}{f_1^4}\left[(s_1 - f_1)^2 - \left(\frac{\pi w_1^2}{\lambda}\right)^2\right]. \tag{87}$$

For a well-adjusted telescope we have $\Delta d = 0$, and the distances between the beam waists and the focal planes of the corresponding lenses (i.e., $s_1 - f_1$ and $s_2 - f_2$) scale like the squares of the focal lengths of the two lenses. The signs in (87) indicate that for an input waist which lies to the left of the focal plane of the input lens one has an output beam waist to the left of the focal plane of the output lens, and conversely for an input waist to the right of the input focal plane.

4.6 *Beam Transformation by a Sequence of Lenses*

Consider now a sequence of lenses of equal focal length f spaced at a distance d as shown in Fig. 6. Immediately to the right of each lens an optical mode of this structure has a phase front with a radius of curvature of $-2f$ and a beam radius w_m given[3,4,26] by

$$\frac{\lambda}{\pi w_m^2} = \frac{1}{2f}\sqrt{4\frac{f}{d} - 1} = \frac{\sin\theta}{d} \tag{88}$$

where θ is defined in (34). To the right of each lens the complex beam parameter (72) of a mode is therefore

$$\frac{1}{q_m} = -\frac{1}{2f} - j\frac{\sin\theta}{d}. \tag{89}$$

Assume that a Gaussian beam is injected into the lens sequence, and call its complex beam parameter in the input plane q_1. If $q_1 = q_m$, then we have launched a mode of the system, and the parameter of the beam to the right of every lens is q_m. For $q_1 \neq q_m$ we use the $ABCD$ law (80) to compute the beam parameter q_2 to the right of the nth lens. The elements of the ray matrix of n sections of the lens sequence are given in (33), and we use them to apply the $ABCD$ law. We have

$$q_2 = \frac{[\sin n\theta - \sin(n-1)\theta]q_1 + d\sin n\theta}{-\frac{1}{f}\sin n\theta \cdot q_1 + \left(1 - \frac{d}{f}\right)\sin n\theta - \sin(n-1)\theta}. \tag{90}$$

From (34) it follows that

$$\sin n\theta - \sin(n-1)\theta = (d/2f)\sin n\theta + \sin\theta\cos n\theta \tag{91}$$

which can be used together with (89) to rewrite (90) as

$$\frac{1}{q_2} - \frac{1}{q_m} = \frac{\sin\theta \cdot e^{jn\theta}}{\sin\theta \cdot e^{-jn\theta} + d\left(\frac{1}{q_1} - \frac{1}{q_m}\right)\sin n\theta}\left(\frac{1}{q_1} - \frac{1}{q_m}\right). \tag{92}$$

After some further rearranging this can be written in the form

$$\frac{1}{\frac{1}{q_2} - \frac{1}{q_m}} + \frac{1}{\frac{2}{q_m} + \frac{1}{f}} = \left[\frac{1}{\frac{1}{q_1} - \frac{1}{q_m}} + \frac{1}{\frac{2}{q_m} + \frac{1}{f}}\right]\exp(-2jn\theta). \tag{93}$$

For the case where the q parameter of the injected beam does not differ too much from the parameter q_m of a mode we can put

$$\Delta = (1/q_1) - (1/q_m) \tag{94}$$

and assume that Δ is small. Developing (92) in powers of Δ we obtain

$$\frac{1}{q_2} - \frac{1}{q_m} = \Delta \cdot e^{2jn\theta} - j\Delta^2 \frac{\pi w_m^2}{2\lambda} (e^{2jn\theta} - e^{4jn\theta}) + 0(\Delta^3). \quad (95)$$

If we neglect all but the first-order term in (95) and compare the real and imaginary parts, we arrive at approximate formulae* for the output parameters R_2 and w_2 :

$$\begin{aligned} \frac{1}{R_2} + \frac{1}{2f} &= \left(\frac{1}{R_1} + \frac{1}{2f}\right) \cos 2\, n\theta + \frac{\lambda}{\pi}\left(\frac{1}{w_1^2} - \frac{1}{w_m^2}\right) \sin 2\, n\theta, \\ \frac{1}{w_2^2} - \frac{1}{w_m^2} &= -\frac{\pi}{\lambda}\left(\frac{1}{R_1} + \frac{1}{2f}\right) \sin 2n\theta + \left(\frac{1}{w_1^2} - \frac{1}{w_m^2}\right) \cos 2\, n\theta. \end{aligned} \quad (96)$$

Comparing these expressions with (33) we see that the beam radius w_2 varies in z direction with a period that is half the period with which a ray displacement varies. This fact has already been seen experimentally,[7] and has also been noted for other optical structures.[28]

The formulae (96) are valid for cases where the parameters w_1 and R_1 of the input beam do not differ much from the parameters of the mode of the lens sequence (i.e. for small Δ). For cases where this condition is not fulfilled we have to go back to (90). Using (72) we re-express the q parameters in terms of w_1 , R_1 , w_2 , and R_2 and compare the imaginary parts of $1/q_2$ as given by (90). After some algebra, where (91) is used to make simplifications, we obtain

$$\begin{aligned} \frac{w_2^2}{w_1^2} = {} & \frac{1}{2}\left[1 + \frac{w_m^4}{w_1^4} + \left(\frac{\pi w_m^2}{\lambda}\right)^2 \left(\frac{1}{R_1} + \frac{1}{2f}\right)^2\right] \\ & + \frac{1}{2}\left[1 - \frac{w_m^4}{w_1^4} - \left(\frac{\pi w_m^2}{\lambda}\right)^2 \left(\frac{1}{R_1} + \frac{1}{2f}\right)^2\right] \cos 2n\,\theta \\ & + \left(\frac{\pi w_m^2}{\lambda}\right)\left(\frac{1}{R_1} + \frac{1}{2f}\right) \sin 2n\,\theta. \end{aligned} \quad (97)$$

In this exact expression for w_2 we find the same periodicity in z direction as in (96). As n is varied w_2 goes through maximum values $w_{\max}$ and minimum values $w_{\min}$. It is easy to show from (97) that

$$w_{\max}\, w_{\min} = w_m^2.$$

Note that $w_{\max}$ and $w_{\min}$ are the extrema of the envelope curve obtained for continuously variable n. The extremal values of w_2 actually occur at a lens only if the corresponding n is an integer.

* In a recent publication by J. Hirano and Y. Fukatsu in Proc. IEEE, *52*, Nov., 1964, p. 1284, similar expressions were derived by means of a perturbation technique in which the real beam parameters were used directly.

An exact expression for R_2 is obtained by comparing the real parts of $1/q_2$ in (90) in a similar way.

4.7 *Beam Transformation by a Lenslike Medium*

The passage of Gaussian beams through a lenslike medium as described by (42) has been discussed by several investigators.[19,28,29,30,31] We assume here for simplicity that $n_0 = 1$, or a refractive index given by

$$n = 1 - 2(r^2/b^2). \tag{98}$$

It is easy to show[19,28,29,30,31] that for a Gaussian beam that is injected with a plane wave front and a beam radius w_0 given by

$$w_0^2 = \lambda b/2\pi \tag{99}$$

the wave front remains plane, and the beam radius remains constant as the wave propagates. These light beams are called the modes of the lenslike medium. If the beam is injected with a beam radius $w_1 \neq w_0$, the wave front and the beam radius will change as a function of z. This problem has been treated by Tien et al.[28] with the help of a differential equation, and by Marcatili[29] who expanded the field distribution of the input beam in terms of the modes of the lenslike medium. We will show here that one can get the desired results in a rather simple fashion by employing the *ABCD* law (80).

The elements of the ray matrix of a medium section of length z are given in (44). Using these together with (80) one computes for a beam with the complex parameter q_1 in the input plane a beam parameter

$$q_2 = \frac{q_1 \cos 2\frac{z}{b} + \frac{b}{2} \sin 2\frac{z}{b}}{-q_1 \frac{2}{b} \sin 2\frac{z}{b} + \cos 2\frac{z}{b}} \tag{100}$$

in the output plane a distance z away from the input. Assuming an input beam with a plane wave front and a beam radius w_1 we have

$$\frac{1}{q_1} = -j \frac{\lambda}{\pi w_1^2}. \tag{101}$$

Inserting this and (99) in (100) we obtain

$$\frac{1}{q_2} = -\frac{\lambda}{\pi w_0^2} \frac{\sin 2\frac{z}{b} + j \frac{w_0^2}{w_1^2} \cos 2\frac{z}{b}}{\cos 2\frac{z}{b} - j \frac{w_0^2}{w_1^2} \sin 2\frac{z}{b}}. \tag{102}$$

If we compare the imaginary parts in this expression we get

$$w_2^2 = w_1^2\left(\cos^2 2\frac{z}{b} + \frac{w_0^4}{w_1^4}\sin^2 2\frac{z}{b}\right) \tag{103}$$

which agrees with the results of Refs. 28 and 29. A comparison of the real parts yields an expression for the curvature of the wave front.

V. RESONATORS WITH INTERNAL OPTICAL ELEMENTS

5.1 *The Basic Resonator Parameters*

A resonator consisting of two spherical mirrors spaced at a distance d is shown in Fig. 12. R_1 and R_2 are the radii of curvature of the two mirrors, measured positive as shown in the figure. The mirror diameters or widths are $2a_1$ and $2a_2$, respectively. The three basic parameters of such a resonator are[10,32,33]

$$N = \frac{a_1 a_2}{\lambda d}, \tag{104}$$

$$G_1 = \frac{a_1}{a_2}\left(1 - \frac{d}{R_1}\right), \tag{105}$$

$$G_2 = \frac{a_2}{a_1}\left(1 - \frac{d}{R_2}\right). \tag{106}$$

Within the Fresnel diffraction theory of optical resonators these three

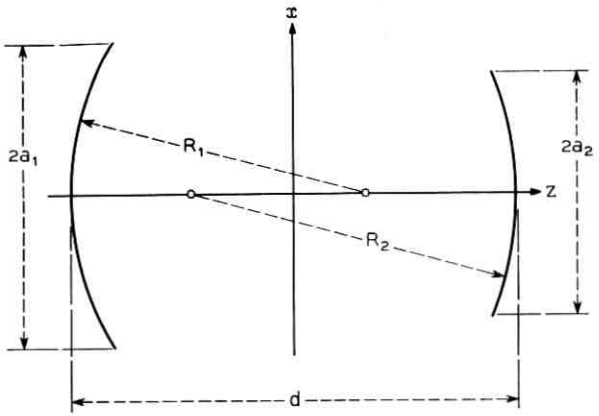

Fig. 12 — Empty spherical mirror resonator.

parameters determine completely the diffraction losses, the resonant frequencies, and the mode patterns of the resonator.[32]

In the following we will show that resonators in which lenses or similar optical structures are inserted between the resonator mirrors are equivalent to an empty resonator of the type shown in Fig. 12. By equivalent resonators we mean here resonators with the same diffraction losses, the same mode patterns except for a scaling factor, and the same resonant conditions. To specify an empty resonator equivalent to a resonator with internal optical elements we will compute its parameters N, G_1, and G_2.

5.2 *Resonators with an Internal Lens*

A resonator with an internal lens is shown in Fig. 13. A lens of focal length f is spaced a distance d_1 away from the left mirror and d_2 away from the right mirror. As before we call the radii of curvature of the two mirrors R_1 and R_2, and their diameters $2a_1$ and $2a_2$ as shown. The internal lens is assumed to be so large that no additional aperture diffraction effects are introduced.

Suppose now that we know the modes of the resonator. We can apply the imaging rules of Section II and choose the mirror surface of the right mirror, say, as reference surface. The image of the mode pattern on this mirror appears a distance

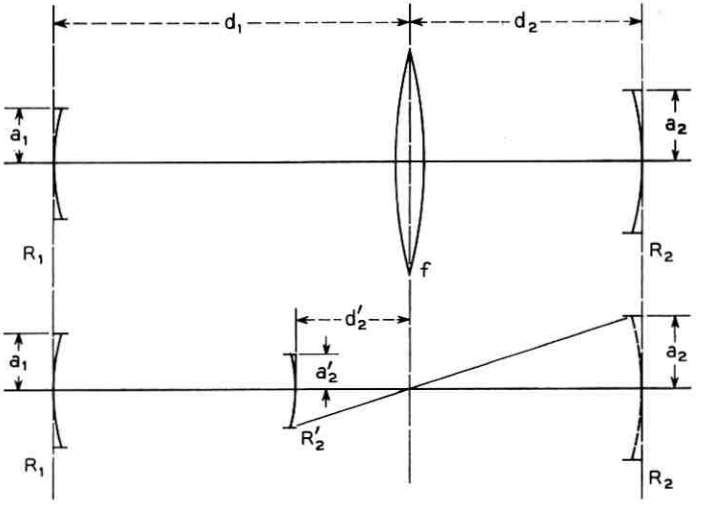

Fig. 13 — Resonator with internal lens and equivalent empty resonator.

$$d_2' = \frac{fd_2}{d_2 - f} \tag{107}$$

away from the lens as shown in the figure. The field of the wave reflected from the mirror is zero outside the mirror aperture a_2. The field of the corresponding image is therefore zero outside an aperture a_2' given by the magnification

$$\frac{a_2}{a_2'} = -\frac{d_2}{d_2'} = 1 - \frac{d_2}{f}. \tag{108}$$

The image is a scaled reproduction of the mode pattern on the mirror which is exact in amplitude and phase if a spherical reference surface is chosen. The correct curvature of this surface is found in accordance with (6) and (8) by imaging the center of curvature of the mirror on the right.

Consider now a mirror of diameter $2a_2'$ placed at the location of the image a distance

$$d = d_1 - d_2' \tag{109}$$

away from the original left mirror as shown in the lower part of Fig. 13. The mirror curvature is chosen to be the same as the curvature of the reference surface for the image. This mirror may be called the image mirror of the original mirror on the right. Apart from a phase difference of $2k(d_2 + d_2')$ it reflects a wave coming in from the left in exactly the same way as the original mirror combined with the lens. The incoming wave produces the same (magnified) complex amplitude distribution or field pattern on the image mirror as on the original mirror on the right. The outgoing wave reflected by the image mirror has a field pattern at d_2' that is identical to the field pattern at d_2' of the outgoing wave reflected by the combination of the original mirror and the lens. The field patterns of the two outgoing waves are thus also identical in any other beam cross section, and in particular across the left mirror. Therefore the modes of the empty resonator formed by the image of the right mirror and the original left mirror as shown in the figure are equivalent to the original resonator with the internal lens. The mode patterns on the left mirror are identical in both cases, and the mode patterns of the corresponding mirrors on the right are scales of each other. The diffraction losses of the two systems are also the same, and there is only a small difference in the corresponding resonant conditions due to the difference in phase shift of $k(d_2 + d_2')$ per transit.

The basic parameters of the equivalent empty resonator are easily

obtained. According to (104) we have a Fresnel number of

$$N = a_1 a_2'/\lambda d \tag{110}$$

and with (105)

$$G_1 = \frac{a_1}{a_2'}\left(1 - \frac{d}{R_1}\right). \tag{111}$$

With (107), (108), and (109) these expressions can be written in terms of the dimensions of the resonator with the internal lens. One obtains

$$N = \frac{a_1 a_2}{\lambda\left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)}, \tag{112}$$

and

$$G_1 = \frac{a_1}{a_2}\left\{1 - \frac{d_2}{f} - \frac{1}{R_1}\left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)\right\}. \tag{113}$$

By interchanging subscripts one gets

$$G_2 = \frac{a_2}{a_1}\left\{1 - \frac{d_1}{f} - \frac{1}{R_2}\left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)\right\}. \tag{114}$$

These three parameters determine the properties of the modes of the internal lens resonator. In the above expression one notes the appearance of the term

$$d_0 = d_1 + d_2 - (d_1 d_2/f) \tag{115}$$

which one might call the effective distance between the mirrors. It is modified by the presence of the lens.

In Refs. 4 and 32 approximate expressions are given for the resonant condition and the beam radii (spot size) of the fundamental mode at the mirrors of an empty resonator that is stable. Recall that for a stable resonator there holds

$$0 \leqq G_1 G_2 \leqq 1. \tag{116}$$

We can apply these formulae to our equivalent resonator and obtain by imaging the corresponding expressions for the resonant wavelength λ and the beam radii w_1 and w_2 on the mirrors of our resonator with an internal lens. Using the parameters discussed above we get

$$\frac{2(d_1 + d_2)}{\lambda} = q + \frac{1}{\pi}(m + n + 1)\cos^{-1}\sqrt{G_1G_2} \qquad (117)$$

where q is the longitudinal mode number, and m and n are the transverse mode numbers. The sign of the square root should be chosen equal to the sign of G_1 (or G_2). For the beam radii we get

$$w_1w_2 = \frac{\lambda d_0}{\pi}(1 - G_1G_2)^{-\frac{1}{2}}, \qquad (118)$$

and

$$\frac{w_1}{w_2} = \frac{a_1}{a_2}\left(\frac{G_2}{G_1}\right)^{\frac{1}{2}}. \qquad (119)$$

The image mirror discussed above can also be obtained from the concept of a "thick mirror." A thick mirror[13] is a combination of a spherical mirror and a lens. The optical characteristics of this combination are represented by a combined focal length and a principal plane.[13] A mirror of this focal length located at the principal plane is equivalent to the thick mirror combination. This equivalent mirror is the same as our image mirror.

The equivalence of the empty resonator and the internal lens resonator can also be shown by using the Fresnel diffraction formula in the manner of Appendix A. One obtains integral equations for the modes of an internal lens resonator. After performing the integration over the lens plane which involves infinite Fresnel integrals, the equivalence to the empty resonator is easily seen.

For cases where the effective distance d_0 as given by (115) is very small, ray angles of interest become rather large and the theory of Fresnel diffraction is no longer expected to apply. We have to exclude these cases from our considerations.

Our discussion includes internal lens resonators with flat mirrors as shown in Fig. 14. The basic parameters of this resonator type can be obtained from (112), (113), and (114) by putting $R_1 = R_2 = \infty$. Burch and Toraldo di Francia[8] have discussed the confocal system of this resonator family where $G_1 = G_2 = 0$. The transmission line dual of an internal lens resonator with flat mirrors is also shown in Fig. 14. It is a sequence of lenses and irises spaced as shown. In this sequence the lenses are large and the irises inserted between them control the modes of the system. For a symmetric system of this kind where $d_1 = d_2 = d$ and $a_1 = a_2 = a$ the above expressions simplify, and we have

$$G_1 = G_2 = 1 - (d/f), \qquad (120)$$

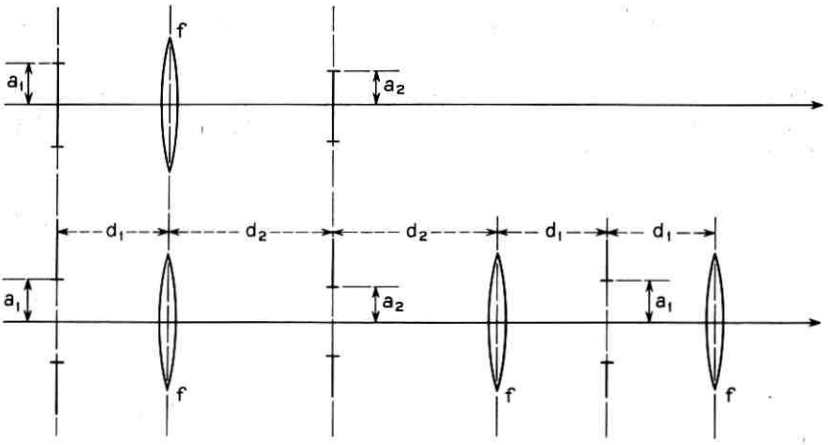

Fig. 14 — Internal lens resonator with flat mirrors and its transmission line dual, a sequence of lenses with irises placed between the lenses.

and a Fresnel number of

$$N = \frac{a^2}{\lambda d\left(2 - \frac{d}{f}\right)}. \tag{121}$$

5.3 *Resonators with an Internal Optical System*

As discussed before, the imaging rules of Section II apply not only to thin lenses but to any optical system that can be characterized by a focal length and by principal planes. The expressions derived above for internal lens resonators can therefore be applied also to spherical mirror resonators with an internal optical system. All one has to do is to interpret f as the focal length of the system and put

$$d_1 = h_1\,; \qquad d_2 = h_2 \tag{122}$$

where h_1 and h_2 measure the distances between the two principal planes and the corresponding mirrors.

We can also characterize the internal optical system by its $ABCD$ or ray matrix as in (9). Inserting (122) in (112), (113), and (114) we compare the resulting expressions with (19) through (22). We note immediately that the three basic resonator parameters can be written in terms of the elements of the ray matrix in the form

$$N = \frac{a_1 a_2}{\lambda B}, \tag{123}$$

$$G_1 = \frac{a_1}{a_2}\left(A - \frac{B}{R_1}\right), \tag{124}$$

$$G_2 = \frac{a_2}{a_1}\left(D - \frac{B}{R_2}\right). \tag{125}$$

5.4 *Internal Lenslike Medium — Guiding Medium with Apertures*

In this section we consider a spherical mirror resonator with a lenslike medium inserted between the resonator mirrors. The optical properties of a lenslike medium have been discussed in Sections 3.4 and 4.7. The refractive index of this medium changes with the square of the distance from the optic axis and is described by (98) if we assume $n_0 = 1$. The degree of this index variation is measured by the parameter b. As shown in Fig. 15, we assume that the medium fills the space between the resonator mirrors which are spaced at a distance l. The mirror diameters are $2a_1$ and $2a_2$, respectively, and the corresponding radii of curvature are R_1 and R_2.

The three basic resonator parameters which describe the modal properties of this resonator with an internal lenslike medium are easily computed by using the results obtained before. The elements of the ray matrix for a medium section of length l are given in (44). Inserting these in (123), (124), and (125) we obtain for the Fresnel number of the system

$$N = \frac{2a_1a_2}{\lambda b \sin 2\dfrac{l}{b}}, \tag{126}$$

and for the G parameters

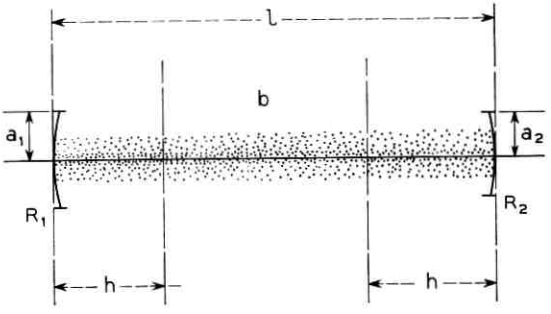

Fig. 15 — Resonator with an internal lenslike medium.

$$G_1 = \frac{a_1}{a_2}\left(\cos 2\frac{l}{b} - \frac{b}{2R_1}\sin 2\frac{l}{b}\right), \tag{127}$$

$$G_2 = \frac{a_2}{a_1}\left(\cos 2\frac{l}{b} - \frac{b}{2R_2}\sin 2\frac{l}{b}\right). \tag{128}$$

A special case of the above system is shown in Fig. 16, where the mirrors are flat, i.e., $R_1 = R_2 = \infty$. The transmission line dual of this resonator is also shown in the figure. It is the interesting case of a lenslike medium or a gas lens with periodically spaced irises as shown. For irises of equal diameter ($a_1 = a_2 = a$) the above expressions simplify, and we obtain for the Fresnel number of the system

$$N = \frac{2a^2}{\lambda b \sin 2\dfrac{l}{b}}, \tag{129}$$

and

$$G_1 = G_2 = \cos 2\frac{l}{b}. \tag{130}$$

This system is confocal for $l = (\pi/4)b$. When the value of $2l/b$ approaches a multiple of π, N gets very large and we have a case where the effective distance between the mirrors is very small [compare (115)]. As discussed before, the theory of Fresnel diffraction is no longer expected to apply under these circumstances.

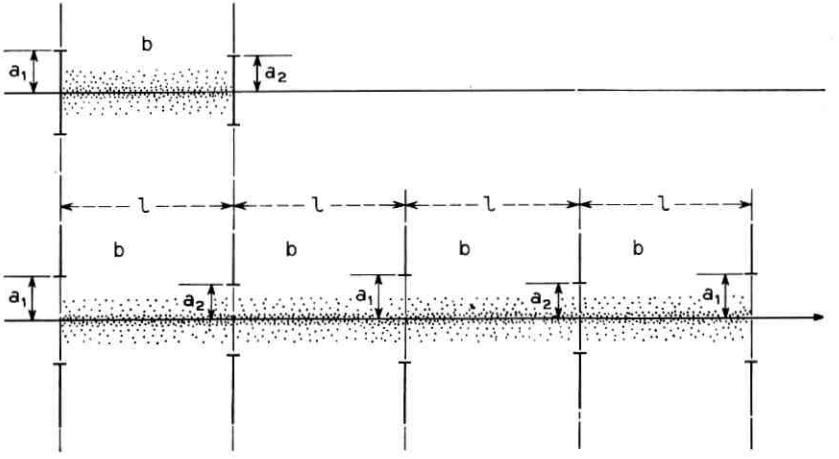

Fig. 16 — Resonator with flat mirrors and an internal lenslike medium, and its transmission line dual, a gas lens with periodically spaced irises.

In high-gain lasers the parameter l/b can become rather large for certain frequencies.[21] For frequencies where the laser medium is focusing we have $b^2 > 0$, while for frequencies where the medium is defocusing $b^2 < 0$ and b is imaginary. It is interesting to study the stability[4] of a resonator with an internal lenslike medium allowing for positive and negative values of b^2. For simplicity we assume that the radii of curvature of the two resonator mirrors are equal and put $R_1 = R_2 = 2f$. With (127) and (128) we obtain

$$G_1 G_2 = G^2 = \left(\cos 2\frac{l}{b} - \frac{b}{4f}\sin 2\frac{l}{b}\right)^2 \tag{131}$$

and write the stability condition (116) in the form

$$-1 \leqq G \leqq 1. \tag{132}$$

One can plot a stability diagram in which each resonator with given parameters l, f, and b is represented by a point. Such a diagram is shown in Fig. 17, where l/f is plotted as ordinate and l/b and jl/b are plotted as abscissae. Resonators with $b^2 > 0$ are represented by points to the

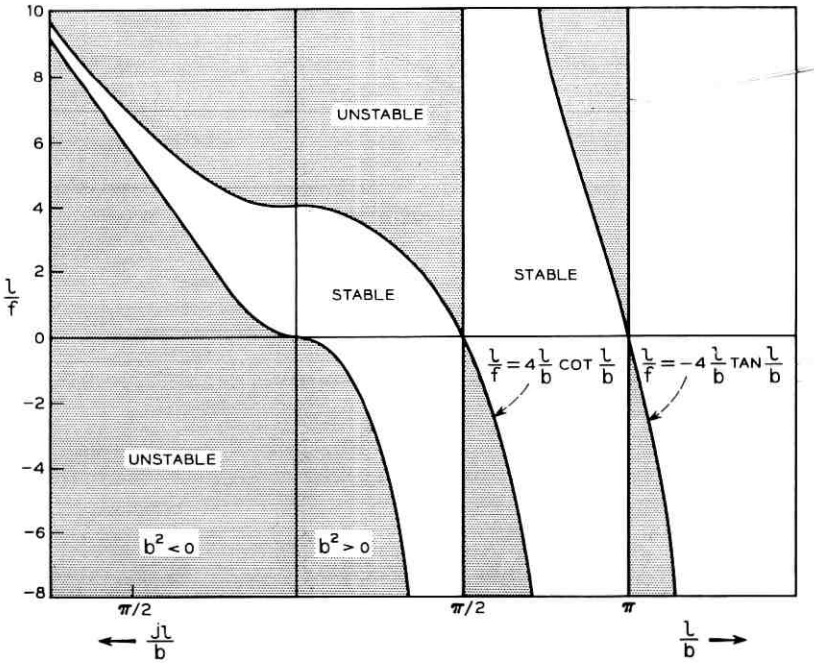

Fig. 17 — Stability diagram for a resonator with an internal lenslike medium.

right of the l/f axis, and resonators with $b^2 < 0$ are represented by points to the left. Points in the shaded regions correspond to unstable resonators, and resonators represented by points in the unshaded regions are stable. The boundaries between the stable and unstable regions follow from (131) and (132). They are described by the equations

$$\frac{l}{b} = \nu \frac{\pi}{2}, \qquad \nu \text{ integer}, \tag{133}$$

$$\frac{l}{f} = 4 \frac{l}{b} \cot \frac{l}{b}, \tag{134}$$

$$\frac{l}{f} = -4 \frac{l}{b} \tan \frac{l}{b}. \tag{135}$$

For $b^2 < 0$, where b is imaginary, the trigonometric functions of (134) and (135) become hyperbolic functions as in (48) and (49). For $b^2 > 0$ one gets periodically stable and unstable regions as l/b is increased.

We have not discussed in detail cases where the lenslike medium occupies only a part of the space between the resonator mirrors. However, one can compute easily the basic parameters for resonators of this kind with the help of the matrix elements of (44), and the formulae (123), (124), and (125).

5.5 *Resonators with One Very Large Mirror*

Let us return to the case of an empty resonator. In some practical arrangements the diameter of one of the two mirrors, say $2a_2$, is so large that diffraction by its aperture can be neglected. The resonator modes are then more or less controlled by the aperture a_1 of the other mirror. This statement is not true for resonators of the degenerate confocal type where the diffraction losses at each mirror are equal[4] for any aperture ratio a_2/a_1. We exclude resonators of this type from our present discussion.

The properties of the resonator modes are generally determined by the three basic parameters given in (104), (105), and (106). But for an infinitely large a_2 the Fresnel number N and the parameter G_2 become infinitely large, and $G_1 = 0$. The resonator parameters are now quite meaningless. It is, however, possible to construct an equivalent resonator with parameters of finite value, as we will show below.

Consider Fig. 18. An empty resonator with one mirror of large diameter is shown schematically at the top. Below it we have drawn its transmission line dual. It consists of a sequence of lenses where an apertured lens

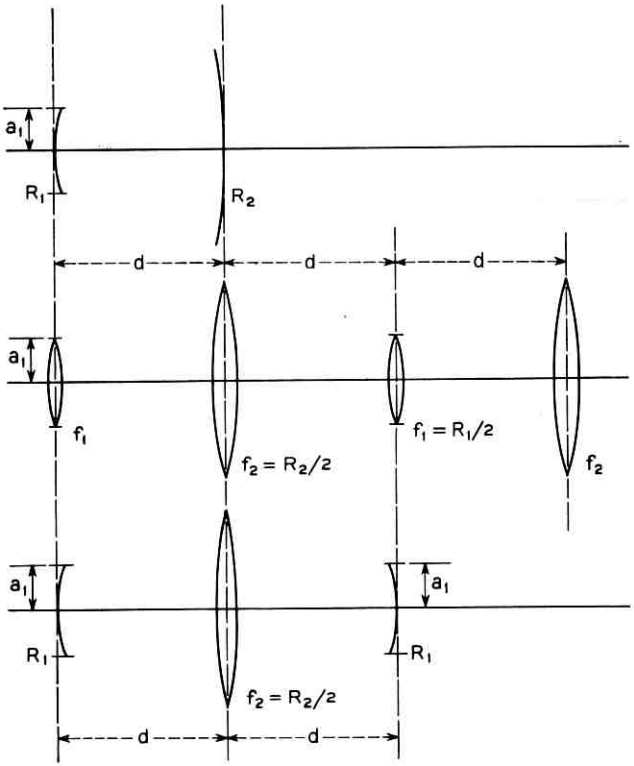

Fig. 18 — Empty resonator with one very large mirror, its transmission line dual, and its equivalent internal lens resonator.

follows an unapertured lens of large diameter. But this transmission line is also the dual of the resonator shown at the bottom of the figure. This is a resonator formed by apertured mirrors of finite diameter $2a_1$ with an internal lens of focal length $f = R_2/2$. The lens is unapertured. Internal lens resonators of this type have been considered before. We can compute the Fresnel number of this system from (112) and obtain

$$N = \frac{a_1^2}{2\lambda d\left(1 - \frac{d}{R_2}\right)}. \tag{136}$$

Equations (113) and (114) are used to calculate the G parameters with the result

$$G_1 = G_2 = 1 - 2d\left(\frac{1}{R_1} + \frac{1}{R_2} - \frac{d}{R_1R_2}\right). \tag{137}$$

These parameters determine the properties of the modes of the internal lens resonator shown in Fig. 18. The mode patterns at the apertured mirrors of this resonator are, of course, equal to the mode pattern at the apertured mirror of the empty resonator. The one-trip diffraction loss of a mode of the internal lens resonator is equal to the return-trip diffraction loss of an empty resonator mode, as there are no diffraction losses at the infinitely large mirror.

For the special case where the large mirror is flat ($R_2 = \infty$) the above discussed equivalences are well known. They follow from symmetry considerations.

VI. ACKNOWLEDGMENTS

Stimulating discussions with E. I. Gordon, J. P. Gordon, R. Kompfner, T. Li, and P. K. Tien in various phases of this work are gratefully acknowledged.

APPENDIX A

Imaging for Fresnel Diffraction

The purpose of this appendix is to show how the imaging relation (6) of the main text is derived within the formalism of scalar Fresnel diffraction theory. Assume a light wave traveling in z direction and refer to Fig. 19. Call the object field $E_1(x_1, y_1)$ and the image field $E_2(x_2, y_2)$. The distances d_1 and d_2 between the lens and the object and image planes, respectively, are related by

$$(1/d_1) + (1/d_2) = 1/f. \tag{138}$$

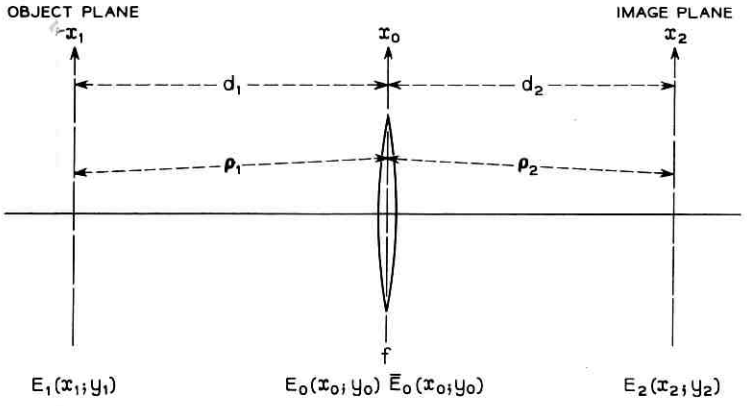

Fig. 19 — Dimensions of interest for Fresnel diffraction theory of imaging.

The field immediately to the left of the lens is denoted $E_0(x_0, y_0)$ and the field to the right of the lens is $\bar{E}_0(x_0, y_0)$. According to (1) of the main text we have for a large, ideal lens

$$\bar{E}_0 = E_0 \exp\left(-jk \frac{x_0^2 + y_0^2}{2f}\right) \tag{139}$$

where $k = 2\pi/\lambda$ is the propagation constant in the medium. With the help of the Fresnel diffraction formula the fields E_0 and E_2 can be expressed as

$$E_0 = \frac{jk}{2\pi d_1} \int_{A_1} dx_1 dy_1 E_1 \exp(-jk\rho_1) \tag{140}$$

and

$$E_2 = \frac{jk}{2\pi d_2} \int_{-\infty}^{+\infty} dx_0 dy_0 \bar{E}_0 \exp(-jk\rho_2) \tag{141}$$

where

$$\rho_1 = d_1 + \frac{1}{2d_1}(x_1 - x_0)^2 + \frac{1}{2d_1}(y_1 - y_0)^2 \tag{142}$$

and

$$\rho_2 = d_2 + \frac{1}{2d_2}(x_2 - x_0)^2 + \frac{1}{2d_2}(y_2 - y_0)^2. \tag{143}$$

The integration in (140) is performed over the aperture area A_1 of the object field, and the integration limits in (141) are extended to infinity with the assumption that the lens is so large that no additional aperture diffraction effects are introduced.

Combining (139), (140), and (141) we obtain by interchanging the order of integration

$$E_2 = -\frac{k^2}{4\pi^2 d_1 d_2} \int_{A_1} dx_1 dy_1 E_1 \int_{-\infty}^{+\infty} dx_0 dy_0 \cdot \exp\left[-jk\left(\rho_1 + \rho_2 + \frac{r_0^2}{2f}\right)\right] \tag{144}$$

where

$$r_0^2 = x_0^2 + y_0^2. \tag{145}$$

Now the expressions (142) and (143) for ρ_1 and ρ_2 are inserted. One finds that in the exponential the terms proportional to r_0^2 cancel because of (138). The integration with respect to x_0 and y_0 can be performed by noting that

$$\int_{-\infty}^{+\infty} dx_0 \exp\left[jkx_0\left(\frac{x_1}{d_1}+\frac{x_2}{d_2}\right)\right] = 2\pi\delta\left(k\left[\frac{x_1}{d_1}+\frac{x_2}{d_2}\right]\right) \tag{146}$$

where δ is the Dirac delta function.[34] With this (144) becomes

$$\begin{aligned} E_2 = &-\frac{k^2}{d_1 d_2}\exp\left[-jk\left(d_1+d_2+\frac{r_2^2}{2d_2}\right)\right] \\ &\cdot\int_{A_1} dx_1 dy_1 E_1 \exp\left(-jk\frac{r_1^2}{2d_1}\right) \\ &\cdot\delta\left(k\left[\frac{x_1}{d_1}+\frac{x_2}{d_2}\right]\right)\cdot\delta\left(k\left[\frac{y_1}{d_1}+\frac{y_2}{d_2}\right]\right). \end{aligned} \tag{147}$$

This simplifies immediately with the help of the formalism of the delta function[34] to

$$\begin{aligned} E_2(x_2, y_2) = &-\frac{d_1}{d_2}E_1\left(-\frac{d_1}{d_2}x_2, -\frac{d_1}{d_2}y_2\right) \\ &\cdot\exp\left[-jk\left(d_1+d_2+\frac{r_2^2}{2d_2}\left(1+\frac{d_1}{d_2}\right)\right)\right]. \end{aligned} \tag{148}$$

Multiplying (138) by d_1/d_2 one finds that

$$\frac{1}{d_2}\left(1+\frac{d_1}{d_2}\right) = \frac{1}{f}\frac{d_1}{d_2} \tag{149}$$

which is used to write (148) in the form of (6) of the main text.

APPENDIX B

Principle of Equal Optical Path Leading to Additional Phase Shift in Image Plane

The process of imaging the field distribution in the object plane into the image plane can be understood in terms of the rays leaving each point (say P_1) in the object plane at various angles as shown in Fig. 20. All rays originating from P_1 are collected at a corresponding point P_2 in the image plane. A form of the principle of equal optical path[35] says that the optical path lengths from P_1 to P_2 are the same for all rays regardless of initial slope.

To obtain an image which is an exact reproduction of the original amplitude *and phase* distribution it would be necessary for the various optical paths which connect corresponding points, say P_1 and P_2 or Q_1 and Q_2, to be equally long for all points regardless of their distance from the optic axis. That these path lengths are not the same for all

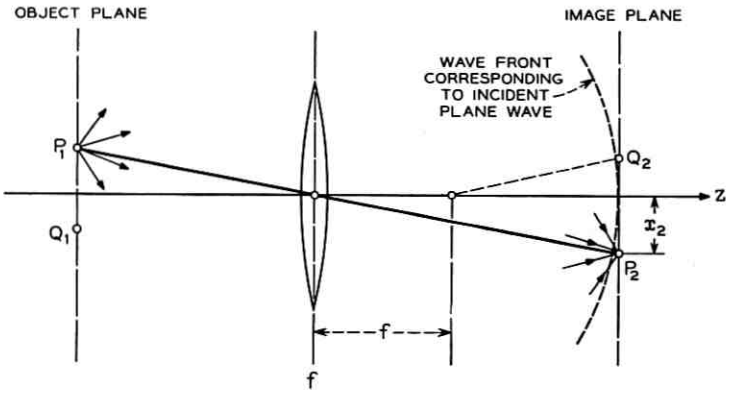

Fig. 20 — Rays emerging from a point of the object collected at the image.

points but increase with increasing distance between the imaged point and the optic axis can be seen from the simple example of an ideal plane wave coming in from the left. This case furnishes an expression for the path length difference as a function of the distance between the imaged point and the axis. As the path length is independent of the field distribution imaged, this expression is valid for the general case. To derive it we recall that an ideal plane wave is transformed by an ideal lens into an ideal spherical wave with the focal point of the lens as its center. The rays connecting points which lie on corresponding wave fronts are equally long for all points on the wave front.[35] Therefore all path lengths measured from the object plane to the spherical wave front which touches the image plane are equal. Paraxial rays (which are practically parallel to the optic axis) need an additional length equal to $r_2^2/2f$ to reach a point (P_2) in the image plane which is a distance r_2 away from the axis. This additional ray length accounts for the additional phase shift given in (6) in the main text.

REFERENCES

1. Fox, A. G., and Li, Tingye, Resonant Modes in a Maser Interferometer, B.S.T.J., *40*, March, 1961, p. 453; Proc. IRE, *48*, 1960, p. 1904.
2. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, Trans. IRE, *AP-9*, May, 1961, p. 248.
3. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, B.S.T.J., *40*, March, 1961, p. 489.
4. Boyd, G. D., and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., *41*, July, 1962, p. 1347.
5. Fox, A. G., and Li, Tingye, Modes in a Maser Interferometer with Curved and Tilted Mirrors, Proc. IEEE, *51*, Jan., 1963, p. 80.
6. Goubau, G., Optical Relations for Coherent Wave Beams, in *Electromagnetic Theory and Antennas*, ed. Jordan, E. C., Pergamon Press, 1963, p. 907.
7. Kogelnik, H., Matching of Optical Modes, B.S.T.J., *42*, Jan., 1964, p. 334.

8. Burch, J. M., Design of Resonators, p. 1187 in *Quantum Electronics III*, ed. Grivet, P., and Bloembergen, N., Columbia University Press, New York, 1964; Toraldo di Francia, G., On the Theory of Optical Resonators, Proc. Symp. Optical Masers, New York, 1963, p. 157, Polytechnic Press, Brooklyn, New York.
9. Kogelnik, H., and Yariv, A., Considerations of Noise and Schemes for Its Reduction in Laser Amplifiers, Proc. IEEE, *52*, Feb., 1964, p. 165.
10. Gloge, D., Analysis of Fabry-Perot Laser-Resonators by Means of Scattering Matrices, Arch. El. Ü., *18*, March, 1964, p. 197.
11. Collins, S. A., Analysis of Optical Resonators Involving Focusing Elements, Appl. Opt., *3*, Nov., 1964, p. 1263.
12. Li, T., Dual Forms of the Gaussian Beam Chart, Appl. Opt., *3*, Nov., 1964, p. 1315.
13. Jenkins, F. A., and White, H. E., *Fundamentals of Optics*, 3rd ed., McGraw-Hill Book Company, New York, 1957, p. 89.
14. Born, M., and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, p. 420.
15. Pierce, J. R., *Theory and Design of Electron Beams*, D. Van Nostrand Company, Inc., New York, 1954, p. 194.
16. Herriott, D. R., Kogelnik, H., and Kompfner, R., Off-Axis Paths in Spherical Mirror Interferometers, Appl. Opt., *3*, April, 1964, p. 523.
17. Bertolotti, M., Matrix Representation of Geometrical Properties of Laser Cavities, Nuovo Cimento, June, 1964, p. 1242; O'Neill, E. L., *Introduction to Statistical Optics*, Addison-Wesley Publ. Co., Reading, Mass., 1963; Brouwer, W., *Matrix Methods in Optical Instrument Design*, Benjamin, New York, 1964.
18. Richards, P. I., *Manual of Mathematical Physics*, Pergamon Press, London, 1959, p. 312.
19. Tonks, L., Filamentary Standing-Wave Pattern in a Solid-State Maser, J. Appl. Phys., June, 1962, p. 1980.
20. Gudzenko, L. I., Concentration of a Light Wave in a Weakly Inhomogeneous Dielectric, J. Exp. Theor. Phys. (USSR), *44*, April, 1963, p. 1298; Sov. Phys. JETP, *17*, Oct., 1963, p. 875.
21. Kompfner, R., Talk at the Electron Device Research Conference, Salt Lake City, Utah, 1963.
22. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., *43*, July, 1964, p. 1469; A Gas Lens Using Unlike, Counter-Flowing Gases, B.S.T.J., *43*, July, 1964, p. 1476.
23. Marcuse, D., and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., *43*, July, 1964, p. 1759.
24. Beck, A. C., Thermal Gas Lens Measurements, B.S.T.J., *43*, July, 1964, p. 1818; Gas Mixture Lens Measurements, B.S.T.J., *43*, July, 1964, p. 1821.
25. Born and Wolf, op. cit., p. 121.
26. Kogelnik, H., Modes in Optical Resonators; in *Advances in Lasers*, ed. Levine, A. K., Dekker Publishers, New York.
27. Yariv, A., and Gordon, J. P., The Laser, Proc. IEEE, *51*, Jan., 1963, p. 4.
28. Tien, P. K., Gordon, J. P., and Whinnery, J. R., Focusing of a Light Beam of Hermite-Gaussian Distribution in Continuous and Periodic Lenslike Media, to be published in Proc. IEEE.
29. Marcatili, E. A. J., Modes in a Sequence of Thick Astigmatic Lens-Like Focusers, B.S.T.J., *43*, November, 1964, p. 2887.
30. Gloge, D., Focusing of Coherent Light Rays in a Space Dependent Dielectric; Arch. El. Ü., *18*, July, 1964, p. 451.
31. Pierce, J. R., unpublished work.
32. Gordon, J. P., and Kogelnik, H., Equivalence Relations among Spherical Mirror Optical Resonators; B.S.T.J., *43*, November, 1964, p. 2873.
33. Streifer, W., and Gamo, H., On the Schmidt Expansion for Optical Resonator Modes, Proc. Symp. on Quasi-Optics, Polytechnic Inst. of Brooklyn, June, 1964.
34. Born and Wolf, op. cit., p. 752.
35. Born and Wolf, op. cit., p. 126.

Geometrical Optics of Magnetoelastic Wave Propagation in a Nonuniform Magnetic Field

By B. A. AULD

(Manuscript received July 21, 1964)

The propagation of magnetoelastic waves in a magnetic insulator having a nonuniform internal magnetic field is examined in the geometrical optics approximation. Hamilton's ray path equations are obtained from the slowness relation for the medium, and it is shown that for YIG there is a substantial focusing action in the rod configuration commonly used for magnetic delay line experiments. When external field shaping is used to produce a minimum internal field at the midpoint of the rod it is found that divergence of the magnetoelastic waves is to be expected.

I. INTRODUCTION

In a number of experiments,[1,2,3] propagation of magnetoelastic waves has been observed in discs and rods of yttrium iron garnet. Coupling is provided through an internal field variation along the direction of propagation, radially in a disc and axially in a rod. This permits excitation of the wave in a region of small wave vector,[4,5] where the magnetic field can couple to the magnetization, with subsequent tapering into the magnetoelastic crossover region. The demagnetizing field also varies in magnitude and in orientation across the direction of propagation. In regions where the wave vector is large it is appropriate to consider the effects of this field inhomogeneity in terms of geometrical optics, and it is to be expected that refraction of the magnetoelastic waves will occur.

II. THE SLOWNESS RELATION AND GROUP VELOCITY

In a cubic crystal with a dc magnetic field applied along a [100] axis x_3 and, for simplicity, assumed elastically isotropic propagation of magnetoelastic waves is governed by the set of equations[6]

$$
\begin{aligned}
i\omega M_{x_1} + \omega_{H_k} M_{x_2} - i\gamma bk \cos\theta\, R_{t'} &= 0 \\
(\omega_{H_k} + \omega_M \sin^2\theta) M_{x_1} - i\omega M_{x_2} - i\gamma bk(\cos 2\theta\, R_t + \sin 2\theta\, R_l) &= 0 \\
(\omega^2 - c_t^2 k^2) R_t - i(bk/\rho M) \cos 2\theta\, M_{x_1} &= 0 \qquad (1) \\
(\omega^2 - c_t^2 k^2) R_{t'} - i(bk/\rho M) \cos\theta\, M_{x_2} &= 0 \\
(\omega^2 - c_l^2 k^2) R_l - i(bk/\rho M) \sin 2\theta\, M_{x_1} &= 0,
\end{aligned}
$$

where

$$
\begin{aligned}
\omega_{H_k} &= \gamma(H + H_{ex} a^2 k^2) \\
\omega_M &= \gamma 4\pi M.
\end{aligned}
$$

The wave vector $\bar{\mathbf{k}}$ is assumed to lie in a (100) plane at an angle θ with the dc field. M_{x_1}, M_{x_2} are transverse components of the magnetic moment referred to axes along [100] directions, and $R_{t'}$, R_t, R_l are transverse and longitudinal components of elastic displacement. The saturation magnetization is denoted by M and the mass density by ρ. Transverse and longitudinal elastic wave velocities are represented by c_t and c_l respectively, and b is the second magnetoelastic constant, generally designated by b_2. H is the internal dc magnetic field, H_{ex} is the exchange field, and a the lattice constant. In what follows it will be assumed that the crystal has sufficient magnetoelastic isotropy ($b_1 \approx b_2$) that (1) is valid for a magnetic field applied at a small angle to the [100] axis and for propagation in any azimuthal direction.

Upon elimination of variables in (1), the secular equation is found to be

$$
\begin{aligned}
\Omega(\omega,k,\theta) = (\omega_s^2 - \omega^2) - \frac{(bk)^2}{\rho M H_k} \\
\cdot\left\{\frac{\omega^2 \cos^2\theta}{\omega_{td}^2 - \omega^2} + \frac{\omega_{H_k} \cos^2 2\theta}{\omega_t^2 - \omega^2} + \frac{\omega_{H_k}^2 \sin^2 2\theta}{\omega_l^2 - \omega^2}\right\} = 0
\end{aligned} \qquad (2)
$$

where

$$
\begin{aligned}
\omega_s^2 &= \omega_{H_k}(\omega_{H_k} + \omega_M \sin^2\theta) \\
\omega_{td}^2 &= \omega_t^2 - \frac{(bk)^2}{\rho M H_k} \cos^2\theta \\
\omega_t^2 &= c_t^2 k^2 \\
\omega_l^2 &= c_l^2 k^2.
\end{aligned}
$$

This equation relates the wave vector $\mathbf{k}$ or the wave slowness vector $\mathbf{k}/\omega = \mathbf{k}/kv_{ph}$ to ω and θ, a relation which may be displayed graphically by a dispersion diagram (see Fig. 1). In coordinates k_1, k_2, k_3 the slowness relation (2) defines a "wave vector" [7] or "slowness" [8] surface for each value of ω. Since the magnetoelastic dispersion curves (see Fig. 1) have four branches and the dispersion relation is independent of azimuthal angle, the wave vector surface is a surface of revolution about x_3 and comprises four sheets. For example, a vertical section through the sheet of the wave vector surface corresponding to branch III appears, at $\omega \approx \omega_H$, as shown in Fig. 2. The group velocity vector[7]

$$\mathbf{V}_g = \nabla_k \omega = -\frac{\nabla_k \Omega}{\partial \Omega / \partial \omega} \tag{3}$$

is proportional to the gradient of Ω and is therefore normal to the wave vector surface, as shown in Fig. 2, the sense of the normal being determined by the requirement that the angle between $\mathbf{V}_g$ and $\mathbf{k}$ be less than $\pi/2$. This means that except in the special cases $\theta = 0$ or $\pi/2$, the group velocity vector is not exactly parallel to the wave vector; and a wave packet does not move in a direction normal to its phase fronts, a phenomenon which is characteristic of anisotropic media.

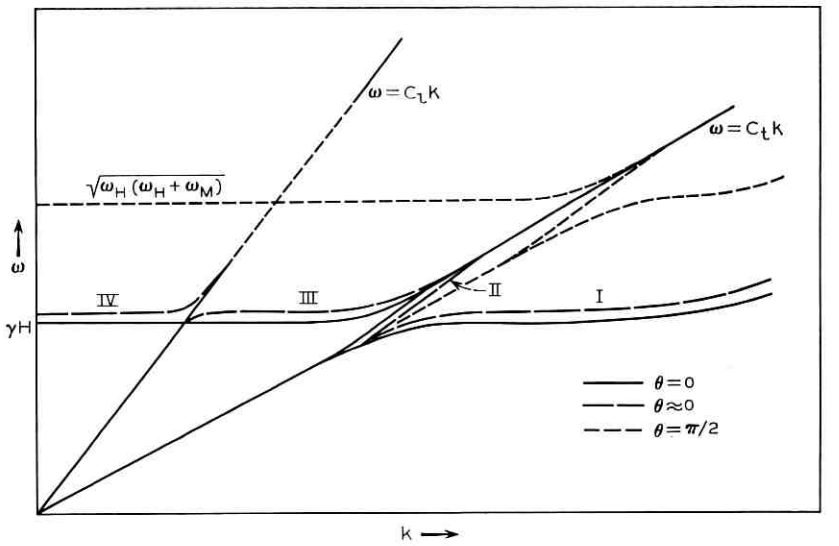

Fig. 1 — Magnetoelastic dispersion diagram.

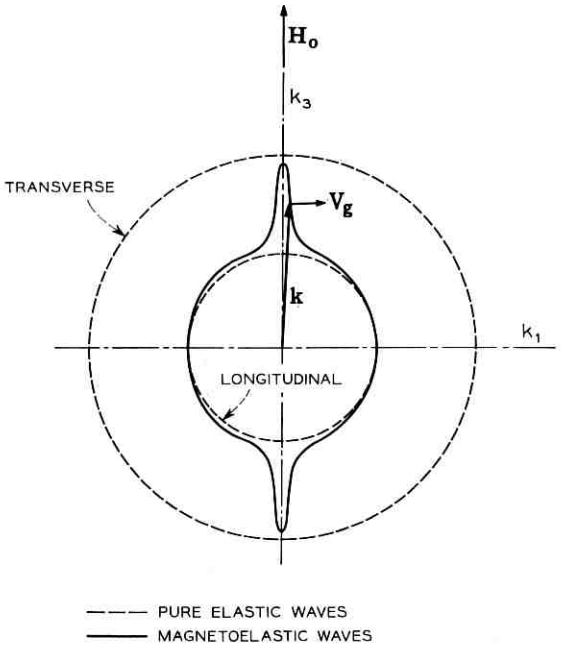

Fig. 2 — Wave vector surface for branch III of the dispersion diagram, at $\omega \approx \omega_H$.

III. THE EIKONAL EQUATION AND THE RAY EQUATIONS

It is assumed that the magnetic field varies in both magnitude and direction from point to point in the medium, but is sufficiently strong at all points to saturate the magnetization. The geometrical optics approximation is appropriate when the magnetic field is almost constant over regions comparable with a wavelength in dimension, so that a solution to the magnetoelastic equations having the form

$$\mathbf{M}(\mathbf{r})e^{i\psi(\mathbf{r})}$$

$$\mathbf{R}(\mathbf{r})e^{i\psi(\mathbf{r})}$$

appears over a small region as a plane wave with relatively slowly varying amplitude. If the assumed solutions are introduced into the equations of motion and spatial derivatives of **R** and **M** are neglected, equations (1) are obtained with $|\nabla\psi|$ substituted for $|k|$, and the angle between the "local" wave vector $\nabla\psi$ and the local magnetic field is substituted for θ. With the same substitutions, the slowness relation (2) reduces to a first-order partial differential equation for the phase function ψ,

$$\Omega(\omega, p_i, x_i) = 0$$

$$p_i = \frac{\partial \psi}{\partial x_i}, \qquad i = 1, 2, 3. \tag{4}$$

This is the eikonal equation or equation of geometrical optics. In the neighborhood of a singularity of the medium the approximations made in deriving the eikonal equation sometimes break down.[9] For the case of electromagnetic propagation in a ferrite, Seidel[10] has shown that singularities of this kind occur because of the appearance of logarithmic derivate coefficients in the field equation. It is not clear whether similar singularities exist for the magnetoelastic equations, and no attempt will be made to justify rigorously the use here of the geometrical optics approximation.

The standard method of solving (4) is by means of the characteristic or ray equations[11]

$$dx_i/dw = \partial\Omega/\partial p_i \tag{5a}$$

$$dp_i/dw = -\partial\Omega/\partial x_i, \tag{5b}$$

where w is a parameter. For any set of initial values of p_i, x_i satisfying (4) these equations, which form the basis of Hamiltonian optics,[8,12] define a unique curve in the space x_1, x_2, x_3. The significance of this curve becomes clear when the equivalence of p_i to the component k_i of the "local" wave vector is recalled. This shows [from (3)] that the tangent,

$$(dx_1 : dx_2 : dx_3) = \left(\frac{\partial\Omega}{\partial p_i} : \frac{\partial\Omega}{\partial p_2} : \frac{\partial\Omega}{\partial p_3}\right),$$

to any curve defined by (4) and a set of initial conditions is always collinear with the group velocity vector. Therefore the curve, or *ray path*, obtained by integrating (5) describes the trajectory of a wave packet launched at a specified point x_i with a specified "local" wave vector $k_i = p_i$. The value of the phase function ψ at any point on the ray path is obtained implicitly from

$$\frac{d\psi}{dw} = p_1 \frac{\partial\Omega}{\partial p_1} + p_2 \frac{\partial\Omega}{\partial p_2} + p_3 \frac{\partial\Omega}{\partial p_3}, \tag{6}$$

where [from (5a)] dw is related to the increment in ray path length ds through the relation

$$ds = \left\{\left(\frac{\partial\Omega}{\partial p_1}\right)^2 + \left(\frac{\partial\Omega}{\partial p_2}\right)^2 + \left(\frac{\partial\Omega}{\partial p_3}\right)^2\right\}^{\frac{1}{2}} dw.$$

When the wave vector surface has more than one sheet there are several initial group velocities corresponding to the same initial wave vector *direction*. These are distinguished by the magnitude of the initial wave vector, and a wave packet will therefore trace out different ray paths according to the magnitude of the initial wave vector, each path corresponding in a local sense to propagation in a mode associated with a particular branch of the dispersion diagram. When there is only a slow spatial variation of the magnetic field there will be little coupling between modes and the different ray paths will maintain their distinct identities.

The present discussion will be concerned with ray paths corresponding to branches I and III of the dispersion diagram. In this case Schlömann[6] has shown that, in the region where the uncoupled magnetic and transverse elastic waves cross, the slowness relation can be written approximately as

$$(\omega - \omega_s)(\omega - c_t k) - (\sigma/2)\omega_{cr} f(\theta) = 0,$$

where $\sigma = \gamma b^2/\alpha M$, ω_{cr} is the crossover frequency for the uncoupled waves, α is the elastic stiffness c_{44}, and

$$f(\theta) = \tfrac{1}{2}[(2 - 5\sin^2\theta + 4\sin^4\theta)(1 + \tfrac{1}{4}\omega_M{}^2\omega_s{}^{-2}\sin^4\theta)^{\frac{1}{2}} + \tfrac{1}{2}\omega_M\omega_s{}^{-1}\sin^4\theta(3 - 4\sin^2\theta)].$$

At the lower microwave frequencies it can be shown that the k dependence of ω_s due to exchange has a very much smaller effect on the slope of the dispersion curves than does the magnetoelastic coupling. If exchange is neglected

$$\omega_{cr} = \omega_s = \omega_H\left(1 + \frac{\omega_M}{\omega_H}\sin^2\theta\right)^{\frac{1}{2}},$$

in which $\omega_H = \gamma H$ and the eikonal equation takes the form

$$\Omega(\omega, p_i, x_i) = (p_1{}^2 + p_2{}^2 + p_3{}^2)^{\frac{1}{2}} + \frac{\sigma}{2c_t}\frac{\omega_s f(\theta)}{\omega - \omega_s} - \frac{\omega}{c_t} = 0, \quad (7)$$

where θ is the angle between the vector $(p_1{:}p_2{:}p_3)$ and the local magnetic field.

In the following section attention will be directed toward rotationally symmetric systems, with rays travelling in meridian planes. It is appropriate, then, to use a cylindrical coordinate system, and (5) and (7), which are written in Cartesian coordinates, must be transformed. Since the φ component of $\nabla\psi$ is zero, (7) becomes

$$\Omega(\omega,p_r\,,\,p_z\,,\,r,z) = (p_r^{\,2} + p_z^{\,2})^{\frac{1}{2}} + \frac{\sigma}{2c_t}\frac{\omega_s f(\theta)}{\omega - \omega_s} - \frac{\omega}{c_t} = 0 \tag{8}$$

$$\theta = \eta - \xi$$

where $\eta = \tan^{-1} p_r/p_z$ and ξ is the polar angle of the local dc magnetic field (see Fig. 3). In a rotationally symmetric system the ray equations (5) transform into

$$dr/dw = \partial\Omega/\partial p_r, \qquad dz/dw = \partial\Omega/\partial p_z \tag{9a}$$

$$dp_r/dw = -\partial\Omega/\partial r, \qquad \partial p_z/dw = -\partial\Omega/\partial z \tag{9b}$$

IV. PARAXIAL RAY EQUATIONS AND REFRACTION IN CONVERGING AND DIVERGING MAGNETIC FIELDS

The discussion will now be restricted to the paraxial case; that is, only rays lying close to the symmetry axis and traveling almost parallel to it will be considered. Then

$$\eta \approx p_r/p_z\,.$$

Furthermore, the rotationally symmetric magnetic field will be assumed to be almost parallel to the axis ($\xi \ll 1$). Since $\theta \ll 1$,

$$f(\theta) \approx 1 - 2.5\,\theta^2;$$

and, when $\omega_M/\omega_H < 10$,

$$\omega_s \approx \omega_H + \omega_M(\theta^2/2).$$

If $\omega \approx \omega_s$ the denominator of the second term in (8) is small and

$$\frac{\partial}{\partial p_i}\frac{\sigma}{2c_t}\frac{\omega_s f(\theta)}{\omega - \omega_s} \approx \frac{\sigma}{2c_t}\frac{\omega_s f(\theta)}{(\omega - \omega_s)^2}\frac{\partial\omega_s}{\partial\theta}\frac{\partial\theta}{\partial p_i}$$

$$\frac{\partial}{\partial x_i}\frac{\sigma}{2c_t}\frac{\omega_s f(\theta)}{\omega - \omega_s} \approx \frac{\sigma}{2c_t}\frac{\omega_s f(\theta)}{(\omega - \omega_s)^2}\left(\frac{\partial\omega_s}{\partial H}\frac{\partial H}{\partial x_i} + \frac{\partial\omega_s}{\partial\theta}\frac{\partial\theta}{\partial x_i}\right),$$

where $p_i = p_r\,,\,p_z$ and $x_i = r,z$. Then

$$\frac{\partial\Omega}{\partial p_r} = \frac{p_r}{p_z} + \frac{\sigma\omega\omega_M}{2c_t p_z}\frac{\left(\dfrac{p_r}{p_z} - \xi\right)}{(\omega - \omega_H)^2} \tag{10a}$$

$$\frac{\partial\Omega}{\partial p_z} = 1, \tag{10b}$$

where only terms linear in p_r/p_z and ξ have been retained, in accordance

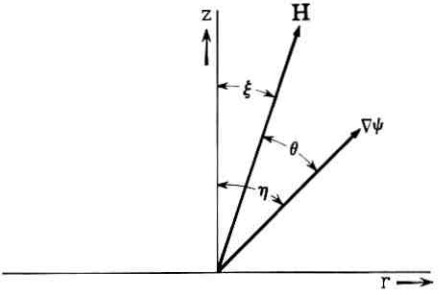

Fig. 3 — Orientation of the "local" wave vector $\nabla\psi$ relative to the local magnetic field.

with the paraxial approximation, and $\omega_s \approx \omega_H$ has been replaced by ω in the numerator of the second term in (10a). Similarly

$$\frac{\partial\Omega}{\partial r} = \frac{\sigma\omega}{2c_t}\frac{\gamma\dfrac{\partial H}{\partial r}}{(\omega - \omega_H)^2}. \tag{11}$$

From (9), (10) and (11) the paraxial ray equations are

$$\frac{dr}{dz} = \frac{1}{p_z}\left(p_r + \frac{\sigma\omega\omega_M}{2c_t}\frac{(p_r/p_z - \xi)}{(\omega - \omega_H)^2}\right) \tag{12}$$

and

$$\frac{dp_r}{dz} = -\frac{\sigma\omega}{2c_t}\frac{\gamma\dfrac{\partial H}{\partial r}}{(\omega - \omega_H)^2}. \tag{13}$$

In the paraxial approximation p_z is obtained directly from (8),

$$p_z \approx \frac{\omega}{c_t}\left(1 - \frac{\sigma}{2(\omega - \omega_H)}\right). \tag{14}$$

Consider now the case of a composite magnetic rod, the middle and outer sections having saturation magnetizations M and M' respectively, which is magnetized along its axis (see Fig. 4). The potential function for the dipolar field on the center line of the middle section, assuming

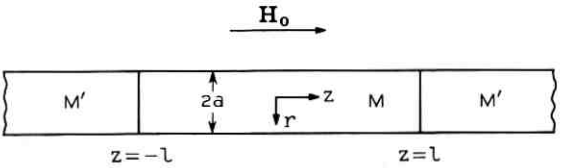

Fig. 4 — Composite rod configuration.

uniform magnetization, is[13]

$$V_D(z) = 2\pi(M - M')[\{(l - z)^2 + a^2\}^{\frac{1}{2}} - \{(l + z)^2 + a^2\}^{\frac{1}{2}} + 2z]$$

if the end effects of the outer sections are assumed to be negligible. Close to the axis the potential function is[14]

$$M_D(r,z) = V_D(z) - \frac{r^2}{4}\frac{\partial^2}{\partial z^2} V_D(z).$$

Assuming $(a/l)^2 \ll 1$, this leads to an internal field in the central region of the middle section

$$\begin{aligned} H &\approx H_0 - 6\pi(M - M')\frac{a^2}{l^2}\left(\frac{z^2}{l^2} - \frac{r^2}{2l^2}\right) \\ \xi = \frac{H_r}{H_z} &\approx \frac{6\pi(M - M')}{H_0}\frac{a^2}{l^2}\frac{zr}{l^2}, \end{aligned} \tag{15}$$

where only terms up to second order in z/l, r/l have been retained. Equation (15) shows that the internal field diverges with increasing $|z|$ when $M' < M$ and converges when $M' > M$. This result has been derived under the assumption of uniform magnetization. Actually the magnetization in the rod will itself be nonuniform, and nonuniformity of the field will be greater than is shown in (15).

A plane magnetoelastic wave is assumed to be propagating in the $+z$ direction at the midpoint of the rod, $z = 0$. Since p_r is then zero at this point and $\xi = 0$ from (15), it follows from (12) that $dr/dz = 0$. This means that the ray paths are parallel to the axis at $z = 0$. Elimination of p_r and p_z from (12), (13) and (14) leads to a second-order differential equation with variable coefficients for $r(z)$, and the ray path trajectories are obtained by solving this equation, subject to the assumed initial conditions. In this case a numerical integration is required for a complete description of the ray paths. If only the direction of refraction is required, the following simpler procedure may be used.

Substitution of (15) into (13) leads to

$$\frac{dp_r}{dz} = -\frac{3\sigma\omega}{4c_t}(\omega_M - \omega_{M'})\frac{a^2}{l^2}\frac{\dfrac{r}{l^2}}{\left(\omega - \omega_{H_m} + \dfrac{3}{2}(\omega_M - \omega_{M'})\dfrac{a^2}{l^2}\dfrac{z^2}{l^2}\right)^2}$$

where

$$\begin{aligned} \omega_M - \omega_{M'} &= \gamma 4\pi(M - M') \\ \omega_{H_m} &= \gamma\left(H_0 + \frac{3(M - M')}{4}\frac{a^2 r^2}{l^4}\right). \end{aligned}$$

Over a small range of z close to $z = 0$ the value of r will not change appreciably along a ray path and dp_r/dz may be integrated directly, giving

$$p_r = -\frac{3\sigma\omega}{4c_t}\frac{(\omega_M - \omega_{M'})}{(\omega - \omega_{Hm})^2}\frac{a^2}{l^2}\frac{zr}{l^2} \tag{16}$$

to second order in z/l. This shows that the "local" wave vector deflects toward the axis with increasing z when $M' < M$ and away from the axis when $M' > M$. The corresponding slope of the ray path is found by substituting p_z and p_r from (14) and (16) into (12). That is

$$\frac{dr}{dz} = -A\left\{1 + \frac{\sigma\omega_M}{2(\omega - \omega_{H_m})^2\left(1 - \frac{\sigma}{2(\omega - \omega_{H_m})}\right)} + \frac{\omega_H}{\omega_{H_m}}\right\}\frac{a^2}{l^2}\frac{zr}{l^2} \tag{17}$$

up to terms of second order in z/l and r/l, where

$$A = \frac{3\sigma(\omega_M - \omega_{M'})}{4\left(1 - \frac{\sigma}{2(\omega - \omega_{H_m})}\right)(\omega - \omega_{H_m})^2}.$$

Equation (17) shows that when the field diverges ($M' < M$) magnetoelastic ray paths which are axial at $z = 0$ will converge, and vice versa. This is easily understood in terms of simple physical concepts. When the internal field decreases with increasing $|z|$ it increases with increasing r, as shown by (15). For a fixed frequency and propagation angle it is clear from Fig. 1 that the "refractive index" kc_t/ω decreases as H increases. If the anisotropy of the dispersion relation is ignored for the moment, this means that off-axis rays curve toward the region of higher "refractive index" closer to the axis. This isotropic effect is enhanced by anisotropy in the dispersion relation. When the wave vector is deflected from the magnetic field direction the ray path (defined by the group velocity) is deflected even further, as shown in Fig. 2, leading to an increased bending of the ray path.

This enhancement of the refractive effect by anisotropy is represented in (17) by the second and third terms under the bracket. In order to estimate the magnitude of these effects consider a YIG rod with $a/l = 0.1$ and $\omega_{H_0} = 2\pi \times 10^9$. For YIG $\sigma = 4 \times 10^6$ sec^{-1} and $\omega_M = 3.08 \times 10^{10}$. According to Schlömann[6] one-half the minimum frequency separation of the transverse magnetoelastic branches is

$$\omega_{\min} = (\sigma\omega_{cr}/2)^{\frac{1}{2}} \approx (\sigma\omega_{H_0}/2)^{\frac{1}{2}} = 1.12 \times 10^8.$$

If this value is assumed for $\omega - \omega_{H_m}$ the approximations used in obtaining (16) and (17) are valid when $z < 0.1l$, $r < 0.01l$. At $z = 0.1l$, $r =$

$0.01l$ the anisotropic terms in (17) are found to be an order of magnitude larger than the isotropic term, and the slope of the ray path is

$$\frac{dr}{dz} = -7.7 \times 10^{-4}.$$

On the basis of an extrapolation at this slope, the ray path should intersect the axis at $z \approx 10l$. The actual intersection would be closer than this because the ray path slope changes continuously with z. When the signal frequency is shifted closer to ω_{H_m}, an increased refraction results. For example, if

$$\omega - \omega_{H_m} = \omega_{\min}/3$$

the phase velocity of the magnetoelastic wave is, from (8), still within a few per cent of the acoustic velocity; but the slope of the ray path at $z = 0.1l$, $r = 0.01l$ is now

$$\frac{dr}{dz} = -3.2 \times 10^{-2}$$

and the extrapolated intersection point occurs at $z \approx 0.3l$. This large change in refractive power with decreasing $\omega - \omega_{H_m}$ is due to the resonance denominators in (17) and is an indication of the steep slopes of the wave vector surface, Fig. 2, in the vicinity of $\theta = 0$. The approximations used in obtaining (17) are, of course, not valid at resonance but are still at least marginally valid in the case considered here.

V. CONCLUSIONS

It has been shown that in a uniformly magnetized medium the phase and group velocities of a magnetoelastic wave are not collinear except

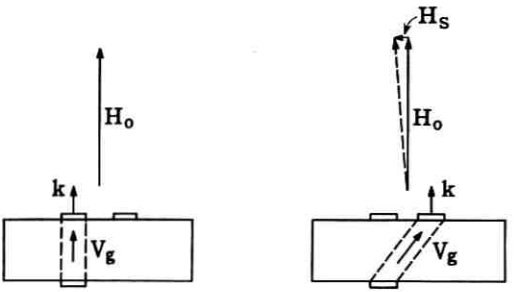

Fig. 5 — Beam steering by means of an auxiliary field.

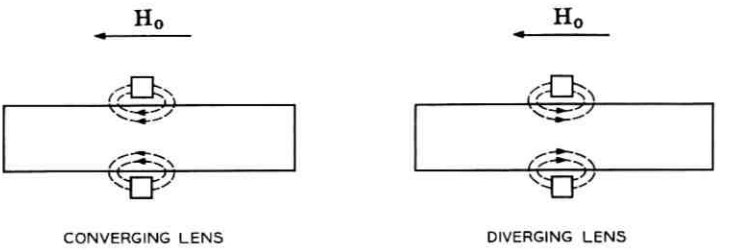

Fig. 6 — Examples of lens configurations.

when the wave vector is either parallel or normal to the magnetic field. This effect might be used for steering or switching an ultrasonic beam by means of an auxiliary field (see Fig. 5). Since the direction of the wave vector remains constant, the phase fronts remain parallel to the transducer faces.

Substantial refraction effects have been shown theoretically to occur in a nonuniformly magnetized medium. For the case of a magnetized rod it is found that paraxial magnetoelastic rays at frequency $\omega \approx \omega_H$

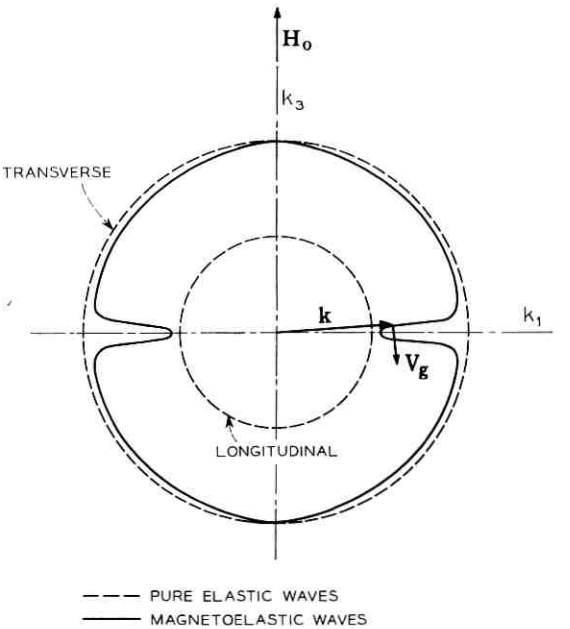

Fig. 7 — Wave vector surface for branch III of the dispersion diagram, at $\omega = \{\omega_H(\omega_H + \omega_M)\}^{\frac{1}{2}}$.

converge when the dipolar field and the applied field are opposing and diverge when the fields are aiding. In arriving at these results, the effects of losses and scattering due to imperfections have been ignored. By a similar analysis it can be shown that an annular permanent magnet or a circular coil encircling the rod will act as a converging lens if its field aids the applied field and as a diverging lens if the fields are opposing (Fig. 6). Paraplanar ray equations can be derived for radial propagation in an axially magnetized thin disk at a frequency $\omega \approx \{\omega_H(\omega_H + \omega_M)\}^{\frac{1}{2}}$. In this case the anisotropic refraction effect is found to oppose the isotropic effect and can even cause the net refraction to change sign. The physical reason for this can be seen by examining the wave vector surface for this case; see Fig. 7. This shows that a deflection of the wave vector away from $\theta = \pi/2$ produces a deflection of the group velocity in the opposite direction.

VI. ACKNOWLEDGMENTS

It is a pleasure to acknowledge many stimulating discussions with H. Seidel about propagation and ray optics in anisotropic media. This investigation was motivated by the work of W. Strauss, H. Matthews and R. T. Denton, with whom a number of helpful discussions have been held.

REFERENCES

1. Eshbach, J. R., J. Appl. Phys., *34*, 1963, p. 1298.
2. Strauss, W., J. Appl. Phys., *35*, 1964, p. 1022.
3. Damon, D., Bull. Amer. Phys. Soc. Ser. II, *9*, 1964, p. 260.
4. Schlömann, E., *Advances in Quantum Electronics*, Columbia University Press, New York, 1961, p. 437.
5. Joseph, R. I., and Schlömann, E., Bull. Amer. Phys. Soc. Ser. II, *9*, 1964, p. 260.
6. Schlömann, E., J. Appl. Phys., *31*, 1960, p. 1647.
7. Landau, L. D., and Lifshitz, E. M., *Electrodynamics of Continuous Media*, Pergamon Press, London, 1960, pp. 317–318.
8. Synge, J. L., *Geometrical Mechanics and de Broglie Waves*, Cambridge University Press, 1964, p. 56.
9. Landau, L. D., and Lifshitz, E. M., op. cit., p. 284.
10. Seidel, H., *Electromagnetic Theory and Antennas*, Pergamon Press, London, 1963, p. 565.
11. Burington, R. S., and Torrance, C. C., *Higher Mathematics*, McGraw-Hill, New York, 1939, p. 756.
12. Synge, J. L., Proc. Roy. Irish Acad., *63A*, 1963, p. 1.
13. Somerfeld, A., *Electrodynamics*, Academic Press, New York, 1952, p. 82.
14. Cosslett, V. E., *Introduction to Electron Optics*, Oxford University Press, London, 1950, p. 35.

Properties of Random Traffic in Nonblocking Telephone Connecting Networks

By V. E. BENEŠ

(Manuscript received November 16, 1964)

Some of the properties of random traffic in nonblocking connecting networks are described and proved. Even though nonblocking networks are rare, they represent an important limiting case, approached as blocking is reduced by adding switches. For many purposes they provide a useful first approximation in the calculation of system parameters. The number of calls in progress is extensively studied in both equilibrium and transient regimes, and its properties are used to distinguish between the wide and strict senses of "nonblocking."

I. INTRODUCTION

In the continuing effort to understand the nature of congestion in telephone connecting networks, it is important to have a thorough knowledge of the special case of *no congestion*, exemplified by traffic in a nonblocking network. Such knowledge is useful not merely as a guide to theoretical investigations, but also in answering questions that are of immediate practical import in the design of networks with small congestion.

It is the purpose of this paper to describe some results concerning random traffic in *nonblocking* connecting networks; these results have important applications to traffic in networks that are not nonblocking. For although nonblocking networks are rare in present telephone practice, and are therefore of limited immediate interest to engineers, they form an important limiting case that is approached as the probability of blocking is reduced by the addition of links and switches to the network. Moreover, many parameters descriptive of the traffic can be calculated with ease for a nonblocking network, and only arduously or not at all for a network that has a nonzero probability of blocking. Hence for low blocking, certain results pertaining to the nonblocking case can be used to approximate those in the blocking case.

In other words, for many purposes the nonblocking case serves as a useful first approximation, as a guide for intuition and computation, in the general case. It is important not to misconstrue our claim. We are not making the banal and useless point that zero is a good first approximation to the probability of blocking when the blocking is small. We are making the point that if the blocking is small then various interesting parameters of the system, other than blocking, are very nearly related as they would be in the nonblocking case. This point has direct practical value.

The present work is, nevertheless, restricted to depicting the properties of nonblocking systems, and no attempt is made here to apply the results to systems with low blocking. Such applications are to appear in later papers, e.g., Ref. 1.

II. THEORETICAL MODEL

Let S be the set of permitted (i.e., physically meaningful) states of the one-sided connecting network ν (of T terminals) under study.† The set S is *partially ordered* by inclusion $\leqq$, where

$$x \leqq y$$

means that state x can be obtained from state y by removing zero or more calls. If x is a state, the notation $|\, x \,|$ will denote the number of calls in progress in state x, while if X is a set, $|\, X \,|$ will denote the number of elements of X. We also use, for a state x, the notations

A_x = set of states accessible from x by *adding* one call
B_x = set of states accessible from x by *removing* one call.

The following two probabilistic assumptions are made:

(*i*) Holding times of calls are mutually independent random variables, each with the negative exponential distribution of unit mean.

(*ii*) If u is an inlet idle in state x and $v \neq u$ is any outlet, there is a probability

$$\lambda h + o(h), \qquad \lambda > 0$$

that u attempt a call to v in the next interval of time of length h, as $h \to 0$.

The choice of unit mean for the holding times merely means that the mean holding time is being used as the unit of time, so that only the one parameter λ need be specified.

We can complete the description of the traffic model to be used by

† A given (network) graph can give rise to several *networks* ν depending on what states are permitted, i.e., belong to S.

indicating how routes for calls are chosen. For this purpose we introduce a routing matrix $R = (r_{xy})$, with these properties: For each $x \in S$ let Π_x be the partition of A_x induced by the equivalence relation of "having the same calls in progress"; then, for each $Y \in \Pi_x$, r_{xy} is a probability distribution over $y \in Y$; in all other cases $r_{xy} = 0$. As in Ref. 2, the interpretation of R is this: any $Y \in \Pi_x$ represents all the ways in which some call c not blocked in x could be completed when ν is in state x; for $y \in Y$, r_{xy} is the chance that if c is attempted in x, it will be routed through the network so as to take the system to state y. Evidently,

$$\sum_{y \in A_x} r_{xy} = \text{number of calls each of which could actually be put up in state } x$$
$$= s(x), \text{ ("successes" in } x)$$

the second equality defining $s(\cdot)$ on S.

A Markov process x_t based on the preceding assumptions has been studied in previous work,[2] and is used here again as a mathematical description of an operating connecting network subject to random traffic.

We restrict attention entirely to the important case of "one-sided" networks in which all inlets are outlets.[2] Analogous results are valid for two-sided, and other, cases.

III. SUMMARY

The wide and strict senses of "nonblocking" are reviewed in Section IV, where it is also pointed out that for most of our purposes it will not be necessary to distinguish them. The equilibrium distribution of the number of calls in progress is calculated in Section V. The terms of the distribution are proportional to the (corresponding) terms of the Poisson distribution with parameter λ, the factors of proportionality indicating the "finite source effect" that is present.

In Section VI various relations among the moments of the distribution of calls in progress are explored. It is noted that the mean determines the variance, and that, as functions of λ, successive moments are related by a difference-differential equation, and can be obtained by logarithmic differentiation of the generating function of the number of assignments of k inlets to k outlets. An extremal property of the distribution of the number of calls in progress, closely related to the author's "thermodynamic" model[3] for telephone traffic, is studied in Section VII. In Section VIII it is shown that the number of calls in progress assumes a Poisson distribution in the limit as $\lambda \to 0$ and the number T of terminals becomes large, with λT^2 constant.

The remainder of the paper is concerned with the transient behavior

of the process x_t representing network operation. The principal result of Section IX is that the past of the process (prior to 0) and the actual state at 0 are both irrelevant to the number of calls in progress at $t > 0$, if it is known how many calls are in progress at $t = 0$. It follows from this that the number $|x_t|$ of calls in progress at t is actually a Markov process, indeed, even a birth-and-death process. These results make it possible to calculate the covariance of $|x_t|$ in terms of $1 + [\frac{1}{2}T]$ characteristic values rather than the astronomical $|S|$ associated with x_t, and to give natural approximations (Sections X and XI). This covariance, it is to be recalled, is the essential ingredient in estimates of sampling error in traffic time-averages. In Section XII, finally, we conclude with characterizations of both the wide and the strict sense of "nonblocking" in terms of the stochastic properties of $|x_t|$.

IV. WIDE AND STRICT SENSES OF "NONBLOCKING"

In a previous paper[4] we have distinguished between a wide sense and a strict sense of the word "nonblocking," as follows: a network ν is nonblocking in the wide sense if there exists a routing matrix R which confines the trajectory of the operating system to nonblocking states, i.e., such that use of the rule R makes the system nonblocking; and ν is nonblocking in the strict sense if no call is ever blocked in *any* of its states. Topological equivalents of these properties were derived in the cited paper.

It is apparent that if ν is nonblocking in the wide sense, then for each rule R that makes ν nonblocking there exists another network ν' whose states are exactly those of ν that are accessible from the zero state under R, and ν' is nonblocking. For this reason most of our results can be (and are) stated for nonblocking networks without specifying whether the sense is wide or strict. The only excepted results are in Section XII, where the stochastic properties of $|x_t|$ are used to distinguish the wide sense of "nonblocking" from the strict.

V. THE NUMBER OF CALLS IN PROGRESS

The equation of statistical equilibrium for the stochastic process x_t is[2]

$$[|x| + \lambda s(x)]p_x = \sum_{y \epsilon A_x} p_y + \lambda \sum_{y \epsilon B_x} p_y r_{yx}, \qquad x \epsilon S. \tag{1}$$

We let

$$p_k = \sum_{|x|=k} p_x, \qquad k = 0, 1, \cdots, \max_{x \epsilon S} |x|,$$

be the probability that k calls are in progress. Our first result is the observation that the $\{p_k\}$ depend only on λ and T, if ν is nonblocking. Let α_x = number of idle inlet-outlet pairs of state x.

Theorem 1: Let ν be nonblocking. For $k = 1, \cdots, \max_{x \in S} |x| = [\frac{1}{2}T]$,

$$\begin{aligned} p_k &= p_0 \frac{\lambda^k}{k!} \prod_{j=0}^{k-1} \binom{T-2j}{2} \\ &= p_0 \frac{(\frac{1}{2}\lambda)^k}{k!} \frac{T!}{(T-2k)!} . \end{aligned} \tag{2}$$

Proof: We sum (1) over $|x| = k$. Since (with the third equality a definition)

$$s(x) = \alpha_x = \binom{T-2|x|}{2} = \alpha_{|x|}$$

if ν is nonblocking, we obtain

$$(k + \lambda\alpha_k)p_k = \sum_{|x|=k} \sum_{y \in A_x} p_y + \lambda \sum_{|x|=k} \sum_{y \in B_x} p_y r_{yx}.$$

In the first sum on the right, each p_y gets counted $(k+1)$ times, because if $|y| = (k+1)$, then $y \in A_x$ for exactly $(k+1)$ values of x. Thus this sum has the value

$$(k+1) \sum_{|y|=(k+1)} p_y = (k+1)p_{k+1} .$$

The second sum is

$$\sum_{|x|=k} \sum_{|y|=k-1} p_y r_{yx} = \sum_{|y|=k-1} p_y \sum_{|x|=k} r_{yx} .$$

However, by the definition of the routing matrix R,

$$\begin{aligned} \sum_{|x|=k} r_{yx} &= \sum_{x \in A_y} r_{yx} \\ &= s(y) \\ &= \alpha_{|y|} , \end{aligned}$$

because ν is nonblocking. Hence the second sum is

$$p_{k-1}\alpha_{k-1} ,$$

and we have shown that

$$(k + \lambda\alpha_k)p_k = (k+1)p_{k+1} + \lambda\alpha_{k-1}p_{k-1} ,$$

with the convention $p_k = 0$ if $k < 0$ or $k > [\frac{1}{2}T]$. Thus

$$kp_k = \lambda\alpha_{k-1}p_{k-1} \qquad k = 1, \cdots, [\tfrac{1}{2}T].$$

By iteration, the theorem follows.

We remark that the probability p_0 that no calls are in progress, determined from the normalization

$$\sum_{k=0}^{[\frac{1}{2}T]} p_k = 1,$$

is just

$$p_0 = \frac{1}{1 + \sum_{k=1}^{[\frac{1}{2}T]} \frac{(\frac{1}{2}\lambda)^k}{k!} \frac{T!}{(T-2k)!}}. \tag{3}$$

VI. MOMENTS OF THE NUMBER OF CALLS IN PROGRESS

From the formulas (2) and (3) giving the distribution of the number of calls in progress, any moment of the distribution of calls in progress can be calculated in principle. More important, though, are the several systematic relationships that obtain among the moments and the parameters λ and T of the system. To these we now turn our attention.

We use the abbreviations

$$a_k = \begin{cases} \dfrac{T!}{2^k k!(T-2k)!} & k = 0, \cdots, [\tfrac{1}{2}T], \\ 0, & k > [\tfrac{1}{2}T] \end{cases}$$

$$m_i = \sum_{x \in S} |\, x \,|^i p_x \qquad i = 1, 2, \cdots,$$

$$= i\text{th moment of } \{p_k\},$$

$$\Phi(\lambda) = \sum_{k \geq 0} \lambda^k a_k,$$

$$\sigma^2 = m_2 - m_1^2 = \text{variance of calls in progress}$$

and $m_1 = m$.

First, it has been shown[2] that whether ν is nonblocking or not, a stochastic process x_t based on our assumptions has the property that the probability Pr {bl} of blocking, the mean m and variance σ^2 of the number of calls in progress, and the parameters λ and T, are all related by the formula, for one-sided networks ν,

$$1 - \Pr\{\text{bl}\} = \frac{1}{\lambda} \frac{2m}{(T-2m)^2 - T + 2m + 4\sigma^2}.$$

(A similar, but different, formula obtains for two-sided ν.) It follows that when ν is nonblocking, the mean and variance of calls in progress are related by

$$(T - 2m)^2 - T + 2m + 4\sigma^2 = 2m/\lambda, \tag{4}$$

and thus determine each other uniquely when λ and T are specified. This means that for a nonblocking ν the important parameters m and σ^2 cannot assume just any values, but must lie on the curve defined by (4).

Second, it is intuitively obvious that, for many networks ν, $m = m(\lambda)$ should be an increasing function of λ. The rationale for this claim is, of course, that if the calling rate per idle pair λ increases, the network will carry a greater (equilibrium) load. For nonblocking networks ν, the claim is a consequence of

Theorem 2: For nonblocking ν, and $i = 1, 2, \cdots$,

$$\frac{d}{d\lambda} m_i = \frac{1}{\lambda} (m_{i+1} - m_i m_1).$$

Proof: We have

$$m_i = \frac{\sum_{k>0} k^i \lambda^k a_k}{\Phi(\lambda)}$$

$$\frac{d}{d\lambda} m_i = \frac{(\sum_{k>0} k^{i+1} \lambda^{k-1} a_k)\Phi(\lambda) - (\sum_{k>0} k_i \lambda^k a_k)(\sum_{k>0} k\lambda^{k-1} a_k)}{\Phi^2(\lambda)}$$

$$= \frac{1}{\lambda} (m_{i+1} - m_i m_1).$$

In particular

$$\frac{dm}{d\lambda} = \frac{\sigma^2(\lambda)}{\lambda} \tag{5}$$

and so m is a strictly increasing function of λ.

Corollary 1: The mean number m of calls in progress as a function of λ satisfies the differential equation

$$\frac{dm}{d\lambda} = \frac{m}{2\lambda^2} - \frac{(T - 2m)^2 - T + 2m}{4\lambda}$$

with the initial conditions $m(0) = 0$, $m'(0) = \binom{T}{2}$.

Proof: We substitute (5) in (4) with $\Pr\{\text{bl}\} = 0$. The initial conditions follow from

$$m(\lambda) = \lambda \binom{T}{2} + o(\lambda), \quad \text{as} \quad \lambda \to 0.$$

It can be verified that Theorem 2 can be rephrased as saying that all the moments of $\{p_k\}$ can be obtained from the logarithmic derivatives of the generating function $\Phi(\cdot)$ of the numbers $\{a_k\}$. Thus for example

$$m = m_1 = \lambda \frac{d}{d\lambda} \log \Phi,$$

$$\sigma^2 = \lambda^2 \frac{d^2}{d\lambda^2} \log \Phi + \lambda^2 \frac{d}{d\lambda} \log \Phi.$$

Indeed, it now becomes apparent that $\{p_k\}$ has the same relationship to the function $\Phi(\cdot)$ as the distribution of calls in progress in the "thermodynamic" model of Ref. 3 had to the generating function of the number of ways of having k calls in progress. It will turn out in the next section that $\Phi(\cdot)$ is actually the generating function of the number of *assignments* of k inlets to k outlets, without reference to how many states of ν, if any, actually realize a given assignment.

VII. AN EXTREMAL PROPERTY OF THE DISTRIBUTION OF CALLS IN PROGRESS

With X the set of T terminals of the network ν, let us consider the set A of all fixed-point free maps of X into itself, together with all submaps thereof. The physical significance of A is that it consists of all the possible "assignments" of k inlets to k outlets with $0 \leqq k \leqq [\frac{1}{2}T]$. The fixed-point free restriction reflects the physically realistic circumstance that no customer will request connection to himself. It is readily seen that the set A of assignments is partially ordered by inclusion, and in fact forms a semilattice. Also there is a natural map of S onto A, the map $\gamma(\cdot)$ of Ref. 4, which takes every state of ν into the assignment it realizes. It can be seen that $\gamma(\cdot)$ preserves order and intersections, so that $\gamma(\cdot)$ is a semilattice homomorphism of S onto A.

Let us now pose the problem of finding a probability distribution $\{p_a, a \in A\}$ which maximizes the entropy functional

$$H(p) = -\sum_{a \in A} p_a \log p_a$$

subject to the condition that

$$\sum_{a \in A} |a| p_a = m,$$

where m is a given positive number with $0 < m < [\frac{1}{2}T]$, and $|\, a \,|$, the norm of a, is the number of inlets mapped into outlets by a, i.e., the number of "intended calls in progress" called for by the assignment a. It follows from Lemma 1 of Ref. 3 that this maximum is achieved by

$$p_a = \frac{\lambda^{|a|}}{\sum_{a \epsilon A} \lambda^{|a|}}$$

i.e., the "canonical" distribution of thermodynamics, with $|\cdot|$ playing the role of energy (see Ref. 3), and $\lambda > 0$ determined uniquely by

$$m = \lambda \frac{d}{d\lambda} \log \sum_{a \epsilon A} \lambda^{|a|}.$$

It follows that the probability assigned by $\{p_a , a \epsilon A\}$ to the set of assignments with k "intended calls in progress" is just

$$\frac{\lambda^k \sum_{|a|=k} 1}{\sum_{x \epsilon A} \lambda^{|a|}} = p_k ,$$

since there are exactly

$$\sum_{|a|=k} 1 = \frac{T!}{k! 2^k (T - 2k)!} \qquad 0 \leqq k \leqq [\tfrac{1}{2}T]$$

fixed-point free maps of k elements out of a set of T into k others from the set, so that $\Phi(\lambda) = \sum_{a \epsilon A} \lambda^{|a|}$.

Thus the distribution $\{p_k\}$ of the number of calls in progress in a nonblocking network arises naturally from maximizing the entropy functional for a probability distribution over the set A of assignments subject to a given average value for $|\, a \,|$, and then calculating the probability of the set of assignments of k calls.

In a similar way, it can be shown that $\{p_k\}$ maximizes the entropy functional $-\sum_k p_k \log p_k$, subject to

$$m = \sum k p_k ,$$

over all distributions having the form $b_k a_k$.

VIII. A POISSON LIMIT THEOREM

It is intuitively reasonable to expect that a nonblocking network with a very large number T of inlets (= outlets, here) and a very small

calling rate λ per idle inlet pair will behave roughly like Palm's "infinite trunk" model for telephone traffic.[5] In particular, if λ becomes small and T becomes large in the right way, the distribution of the number of calls in progress in equilibrium should become Poisson. That this occurs is the content of

Theorem 3: Let a be a positive number, and let $\lambda \to 0$ and $T \to \infty$ in such a way that

$$a = \lambda T^2/2.$$

Then

$$p_k \to e^{-a}(a^k/k!), \qquad k = 0, 1, 2, \cdots.$$

Proof: We have

$$p_k/p_0 = \frac{\left(\frac{\lambda T^2}{2}\right)^k}{k!}\left(1 - \frac{1}{T}\right)\cdots\left(1 - \frac{2k-1}{T}\right) \to \frac{a^k}{k!}.$$

Since

$$p_0^{-1} = 1 + \sum_{k=1}^{[\frac{1}{2}T]} p_k/p_0,$$

the result follows.

The reason why λT^2, and not, e.g., λT, must be of the order of the average carried load, is that λ is the calling rate per pair of idle inlets (= outlets, here), so that if all are idle, this calling rate is just

$$\lambda\binom{T}{2},$$

omitting attempts by a customer to himself. Indeed, the load carried by one customer's line is

$$q = (2m/T) = \lambda T((1-q)^2 - T^{-1}(1-q) + T^{-2}4\sigma^2)).$$

It is easily seen that q and $T^{-2}\sigma^2$ are bounded independently of λ and T, so that

$$q \sim \lambda T \to 0$$

in the limit taken.

IX. TIME-DEPENDENT BEHAVIOR OF THE NUMBER OF CALLS IN PROGRESS

So far all our results have concerned only the equilibrium behavior of the process x_t representing the operation of a nonblocking connecting network. We now turn to the transient or time-dependent behavior.

The matrix $Q = (q_{xy})$ of transition rates of x_t is given by

$$q_{xy} = \begin{cases} 1, & x \in A_y \\ \lambda r_{xy}, & x \in B_y \\ -|x| - \lambda s(x), & x = y \\ 0 & \text{otherwise.} \end{cases}$$

The matrices $P(t) = (p_{xy}(t))$, t real, of transition probabilities, i.e., such that

$$p_{xy}(t) = \Pr\{x_t = y \mid x_0 = x\},$$

satisfy the Kolmogorov equations

$$P'(t) = QP(t) = P(t)Q, \qquad P(0) = I.$$

We let

$$p_{ij}(t) = \Pr\{\,|x_t| = j \,\big|\, |x_0| = i\}$$
$$p_{xj}(t) = \Pr\{\,|x_t| = j \mid x_0 = x\}.$$

Intuitively, if ν is nonblocking and $|x_t| = j$, then the (conditional) probabilities of the possible changes in the number of calls in progress in the next interval of time of length h are

$$\begin{cases} jh + o(h), & \text{for a hangup,} \\ \lambda\binom{T-2j}{2} + o(h), & \text{for a new call,} \end{cases}$$

as $h \to 0$. Indeed, one expects that these evaluations remain true even if information about x_s for $s < t$ is added to what is known at time t, for the reason that only the fact that $|x_t| = j$ is relevant to what happens to $|x_s|$ for $s > t$. In other words it is natural to expect that for nonblocking ν,

$$|x_t|$$

is itself a Markov process, indeed, a birth-and-death process. It will be shown that these conjectures are true, and that they have important practical consequences.

Theorem 4: If ν is nonblocking, then knowledge of the actual state x_0 is irrelevant to $|x_t|$ if $|x_0|$ is known, i.e.,

$$p_{xk}(t) = p_{|x|k}(t), \qquad \text{for all } x.$$

Proof: The backward Kolmogorov equation for the process is

$$\frac{d}{dt}p_{xy} = -[|x| + \lambda s(x)]\,p_{xy} + \sum_{u \in B_x} p_{uy} + \lambda \sum_{u \in A_x} r_{xu}p_{uy}\,.$$

Summing on $|y| = k$ gives

$$\frac{d}{dt}p_{xk} = -[|x| + \lambda s(x)]p_{xk} + \sum_{u \in B_x} p_{uk} + \lambda \sum_{u \in A_x} r_{xu}p_{uk}\,.$$

Since $u \in B_x$ for exactly $(|x| - 1)$ values of u, and since

$$\sum_{x \in A_u} r_{xu} = s(x) = \binom{T - 2|x|}{2},$$

it is enough to show that the result is true in a neighborhood of $t = 0$. Evidently, though,

$$p_{xk}(0) = \begin{cases} 1 & |x| = k \\ 0 & |x| \neq k \end{cases}$$

$$\frac{d}{dt}p_{xk}(0) = \begin{cases} -\left[|x| + \lambda\binom{T - 2|x|}{2}\right] & |x| = k \\ 0 & |x| \neq k \end{cases}$$

and

$$p_{xk}^{(n)}(0) = \begin{cases} -\left[|x| + \lambda\binom{T - 2|x|}{2}\right] p_{xk}^{(n-1)}(0) + \sum_{u \in B_x} p_{uk}^{(n-1)}(0) \\ \qquad + \lambda \sum_{x \in A_u} r_{xu}p_{uk}^{(n-1)}(0), & |x| = k \\ 0 & |x| \neq k. \end{cases}$$

Since $p_{xk}(\cdot)$ is analytic in a neighborhood of $t = 0$, the theorem follows.

Theorem 5: If ν is nonblocking, then

$$|x_t|$$

is a Markov stochastic process.

Proof: Set $y_t = |x_t|$. Since x_t is a Markov process, for $t_1 < t_2 < \cdots < t_n < t$ we have a.e.

$$\Pr\{y_t = k \mid x_{t_i}, i = 1, \cdots, n\} = \Pr\{y_t = k \mid x_{t_n}\}, \\ = \Pr\{y_t = k \mid y_{t_n}\}$$

by Theorem 4.

X. TRANSITION PROBABILITIES OF $|x_t|$

It follows from Theorem 5 and the *forward* Kolmogorov equation for for x_t that the transition probabilities $p_{ij}(\cdot)$ of $y_t = |x_t|$ satisfy the equations

$$\frac{d}{dt}p_{ij} = -\left[j + \lambda\binom{T-2j}{2}\right]p_{ij} \\ + (j+1)p_{i(j+1)} + \lambda\binom{T-2j+2}{2}p_{i(j-1)}, \tag{6}$$

with obvious conventions at the (reflecting) boundaries $j = 0$ and $j = [\frac{1}{2}T]$. These are the equations of a birth-and-death process on a finite number of states, and so the known results of Karlin and McGregor[6] can be carried over at once, as summarized below.

The matrix $A(T,\lambda)$ governing the system (6) is given by

$$a_{ij} = \begin{cases} 0 & |i-j| > 1 \\ i & j+1 = i \\ -i - \lambda\binom{T-2i}{2} & i = j \\ \lambda\binom{T-2i}{2} & i+1 = j. \end{cases} \tag{7}$$

With

$$\pi_k = \lambda^k a_k \qquad k = 0, 1, \cdots, [\tfrac{1}{2}T],$$

and

$$Q_0(x) \equiv 1,$$

$$-xQ_0(x) = -\lambda\binom{T}{2}Q_0(x) + \lambda\binom{T}{2}Q_1(x),$$

$$-xQ_k(x) = kQ_{k-1}(x) - \left[k + \lambda\binom{T-2k}{2}\right]Q_k(x) \\ + \lambda\binom{T-2k}{2}Q_{k+1}(x), \qquad 1 < k < [\tfrac{1}{2}T],$$

there is a unique[6,7] positive regular measure ψ on $0 \leqq x < \infty$ such that

$$\int_0^\infty Q_i(x)Q_j(x)d\psi(x) = \frac{\delta_{ij}}{\pi_j} \qquad i,j = 0, 1, \cdots, [\tfrac{1}{2}T].$$

The transition probabilities of $|\, x_t \,|$ are represented by the formula

$$p_{ij}(t) = \pi_j \int_0^\infty e^{-xt} Q_i(x)Q_j(x)d\psi(x) \tag{8}$$

XI. THE COVARIANCE OF $|\, x_t \,|$

As has been pointed out,[3,8] the covariance function of the number of calls in progress is of great practical interest in connection with estimates of sampling error in telephone traffic averages. This covariance is defined as

$$R(t) = E\{\, |\, x_{t+s} \,|\, |\, x_s \,|\, \} - E^2\{\, |\, x_s \,|\, \},$$

and does not depend on s, since it is understood that x_s has its equilibrium distribution. The variance of the continuous time-average

$$\frac{1}{T}\int_0^T |\, x_t \,| dt$$

is then

$$2T^{-2}\int_0^T (T - t)R(t)dt,$$

while that of the periodic scanned average

$$\frac{1}{n}\sum_{j=1}^{n} |\, x_{j\tau} \,|, \qquad \tau > 0,$$

with scanning interval τ is

$$\sum_{j=-n}^{n} (n - |\, j \,|\,)R(j\tau).$$

It is easily seen from the integral representation (8) that the covariance of $|\, x_t \,|$ is

$$R(t) = \sum_{i,j=1}^{w} ij\pi_i\pi_j \int_0^\infty e^{-xt} Q_i(x)Q_j(x)d\psi(x) - m^2,$$

$$w = [\tfrac{1}{2}T] = \max_{x \epsilon S} |\, x \,|.$$

The orthogonality of the $Q_i(\cdot)$ with respect to $\psi(\cdot)$ allows the simplification of this formula to

$$R(t) = \int_0^\infty e^{-xt} \left[\sum_{i=1}^{w} i\pi_i Q_i(x)\right]^2 d\psi(x) - m^2.$$

It is easily verified that for $k > 0$

$$Q_k(0) = 1$$

and that

$$\psi(0+) - \psi(0-) = \left(\sum_{j=0}^{w} \pi_j\right)^{-1}.$$

Hence the contribution of $\psi(\cdot)$ at the origin (to the first term on the right of $R(t)$) gives precisely m^2, and we have proved the important result that

$$R(t) \geqq 0.$$

We note next that the matrix $A(T,\lambda)$ of the differential equations for $p_{ij}(\cdot)$ is symmetrizable, and so has real nonpositive characteristic values. In a standard way[3,8] it is deduced that one of these is zero, and that the dominant characteristic value r_1 satisfies

$$\begin{aligned} -(m/\sigma^2) &\leqq r_1 < 0, \\ R(t) &\leqq \sigma^2 e^{r_1 t}. \end{aligned} \tag{9}$$

As in the theory[8] of the finite trunk group, it is expected that this upper bound for $R(\cdot)$ will be a good approximation for low to moderate traffic levels. Together, the two inequalities suggest the alternative estimate

$$R(t) \sim \sigma^2 \exp -\left(\frac{m}{\sigma^2} t\right),$$

also used in Ref. 8.

Since the equilibrium distribution $\{p_k\}$ of the number of calls in progress approaches Poisson's as $\lambda \to 0$ and $T \to \infty$ with λT^2 constant, it is to be expected that the characteristic values of the matrix $A(T,\lambda)$ of the system (6) will concentrate at the nonpositive integers in this same limit. In this connection it is instructive to see how the lower bound $-m/\sigma^2$ to r_1 behaves in the above limit. With $\lambda T^2 \equiv 2a > 0$, we find

$$\begin{aligned} \frac{m}{\sigma^2} &= 2\lambda + \frac{\lambda}{2}\frac{(T - 2m)^2}{\sigma^2} + \frac{\lambda}{2}\frac{T - 2m}{\sigma^2} \\ &= \frac{1}{1 + 2\lambda T + \lambda}\left[2\lambda + \frac{a}{\sigma^2}(1 + T^{-1}) + 2\lambda\left(\frac{m}{\sigma}\right)^2\right]. \end{aligned} \tag{10}$$

Since the variance of a Poisson distribution equals its mean, $\sigma^2 \to a$,

and it is easily verified that σ^2/a depends only on T and not on λ so that

$$\sigma^2/a = 1 + o(1)$$

with $o(1)$ depending only on T. It follows that for any $a > 0$,

$$\liminf_{\substack{\lambda \to 0 \\ T \to \infty \\ \lambda T^2 = 2a}} r_1 \geqq -1,$$

i.e., the lower limit of the dominant characteristic value is at least -1. If we retain only terms of order λT in (10) we obtain

$$-1 - \frac{4a - 1}{T} \tag{11}$$

as an approximate lower bound for r_1, indicating that r_1 actually approaches -1 from above or below according as $a < \frac{1}{4}$ or $a > \frac{1}{4}$, the latter case being overwhelmingly prevalent in practice.

Actually it is not necessary that $T \to \infty$ in order that the lower bound in (9) approach -1. It suffices that λ be small, for with T fixed, as $\lambda \to 0$,

$$-\frac{m}{\sigma^2} = -\frac{m}{\lambda \dot{m}} = -\frac{\lambda\binom{T}{2} + o(\lambda)}{\lambda\binom{T}{2} - \lambda^2\binom{T-2}{2} + o(\lambda)}$$

$$= -1 + \lambda\binom{T-2}{2} + o(\lambda).$$

We note that the correction term is quite different from that in (11).

XII. STOCHASTIC CHARACTERIZATION OF WIDE AND STRICT SENSES OF "NONBLOCKING"

In the following, we regard the process x_t defined in Section II as a function of ν, λ and the routing matrix R, $T = T(\nu)$, etc.

Theorem 6: ν is nonblocking in the wide sense if and only if for some routing matrix R, $|x_t|$ is a birth-and-death process whose semigroup of transition probabilities is generated by $A(T,\lambda)$.

Proof: The necessity follows from Theorem 5. For the sufficiency we argue that if ν is not nonblocking in the wide sense then any choice of R gives rise to a nonzero probability of blocking. Thus by the basic

formula (4)

$$\frac{1}{\lambda}\frac{2m}{(T-2m)^2 - T + 2m + 4\sigma^2} < 1$$

for any R, which contradicts the condition that for some R, $p = \{p_k\}$ satisfies

$$Ap = 0,$$

with the convention $(Ap)_j = \sum_i a_{ij}p_i$. In a similar way we can prove

Theorem 7: ν is nonblocking in the strict sense if and only if for every R, $|x_t|$ is a birth-and-death process whose semigroup of transition probabilities is generated by $A(T,\lambda)$.

The proof is a minor modification of that of Theorem 6, and is omitted.

REFERENCES

1. Beneš, V. E., Some Inequalities in the Theory of Telephone Traffic, to appear.
2. Beneš, V. E., Markov Processes Representing Traffic in Connecting Networks, B.S.T.J., *42*, 1963, p. 2795.
3. Beneš, V. E., A Thermodynamic Theory of Traffic in Connecting Networks, B.S.T.J., *42*, 1963, p. 567.
4. Beneš, V. E., Algebraic and Topological Properties of Connecting Networks, B.S.T.J., *41*, 1962, p. 1249.
5. Palm, C., Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, *44*, 1943.
6. Karlin, S., and McGregor, J., The Classification of Birth-and-Death Processes, Trans. Amer. Math. Soc., *86*, 1957, p. 366.
7. Shohat, J. A., and Tamarkin, J. D., The Problem of Moments, Mathematical Surveys, *1*, 1943.
8. Beneš, V. E., The Covariance Function of a Simple Trunk Group, with Applications to Traffic Measurement, B.S.T.J., *40*, 1961, p. 117.

Gain of Electromagnetic Horns

By T. S. CHU and R. A. SEMPLAK

(Manuscript received December 11, 1964)

The absolute gain of a standard horn is often measured by determining the transmission loss versus separation between two identical standard horns. Correction ratios are needed because the usual criterion for separation ($2a^2/\lambda$) may not justify the use of the far-zone power transmission formula. Using the near-field power transmission formula, the ratio between the Fraunhofer and Fresnel gain of a pyramidal electromagnetic horn has been computed as a function of horn dimensions and separation distance.

The calculated corrections have been applied in the absolute gain measurement of a standard horn which was used as a calibration reference in a recent 4080-mc gain measurement of a large horn-reflector antenna. The measured gain of the standard horn at 4080 mc is 20.11 db with an accuracy of ± 0.035 db. The calculated gain is 20.15 db.

I. INTRODUCTION

Recently, a standard horn was used as a calibration reference in measuring the gain of a 400-square foot aperture horn-reflector antenna at 4080 mcs.[1] Since the horn-reflector antenna is currently being used for precision measurement of the absolute flux of stellar radio sources, it is desirable that the gain of the standard horn be known as accurately as possible. From previous work,[2] the calculated gain[7] of a standard horn was believed to be within ± 0.1 db of its true gain. Our purpose was to measure the absolute gain of the standard horn to an accuracy better than that previously achieved.

The gain of a standard horn can be determined by measuring the transmission loss versus separation between two identical standard horns. In the technique of measurement commonly used, the separation distance is not large, and it is well known that the far-zone power transmission formula

$$P_R/P_T = (G\lambda/4\pi r)^2 \tag{1}$$

is not valid if the separation r between the apertures of the two horns

is not great enough. Therefore the gain formula

$$G = \frac{4\pi r}{\lambda} (P_R/P_T)^{\frac{1}{2}} \tag{2}$$

may introduce considerable error when the far-zone gain of pyramidal electromagnetic horns is measured at relatively short distances. Even an aperture-to-aperture separation r of about $2a^2/\lambda$ between two optimum horns, where a is the large dimension of the aperture, introduces an error of the order of 1 db. Jakes[2] suggested the junction of the horn with the feeding waveguide as the reference point for optimum horns. He demonstrated empirically that the error in gain may be reduced to about 0.1 db if r is measured between the reference points of two optimum horns. Braun[3] calculated the error in the gain of electromagnetic horns measured at short distances. However, his assumptions about the received power are questionable, since the power in the transmitted wave was averaged over the receiving aperture. Although the near-field power transmission formula appeared in the literature,[4] to our knowledge it has not been applied to the gain measurement of electromagnetic horns. With the aid of the digital computer, the near-field power transmission formula easily yields the required correction ratios for the far-zone gain of pyramidal electromagnetic horns measured at relatively short distances.

II. CALCULATION OF THE CORRECTIONS

Using the Lorentz reciprocity theorem, it has been shown[4] that the ratio of the received to transmitted power between two antennas at any separation is

$$\frac{P_R}{P_T} = \frac{1/4 \left| \int_S (\mathbf{H}_2 \times \mathbf{E}_1 + \mathbf{E}_2 \times \mathbf{H}_1) \cdot \hat{n}\, ds \right|^2}{\left\{ \mathrm{Re} \int_{S_1} (\mathbf{E}_1 \times \mathbf{H}_1^*) \cdot \hat{n}_1 ds \right\} \left\{ \mathrm{Re} \int_{S_2} (\mathbf{E}_2 \times \mathbf{H}_2^*) \cdot \hat{n}_2\, ds \right\}} \tag{3}$$

where $\mathbf{E}_1$, $\mathbf{H}_1$ are the fields when antenna 1 is transmitting, $\mathbf{E}_2$, $\mathbf{H}_2$ are the fields when antenna 2 is transmitting; and $\hat{n}$, $\hat{n}_1$, and $\hat{n}_2$ are the unit normals of the surfaces. The surface S can be either one of the two antenna apertures. Equation (3) is an exact formula if all the field quantities are evaluated with both antennas in place and under matched conditions. In the following calculation the reflections between the antennas will be neglected; that is, in evaluating $\mathbf{E}_1$, $\mathbf{H}_1$ antenna 2 will be removed, and in evaluating $\mathbf{E}_2$, $\mathbf{H}_2$ antenna 1 will be removed. We also neglect any mismatch between antennas and their transmission lines.

Furthermore, we assume that the tangential components of **E** and **H** are related by the free-space impedance at each point:

$$\hat{n} \times \mathbf{E}^t = \sqrt{\frac{\mu}{\epsilon}}\, \mathbf{H}^t.$$

With these approximations, we can write down the power transmission formula between two electromagnetic horns at any separation.

$$\frac{P_R}{P_T} = \frac{\left| \int_{s_1} \int_{s_2} E_1{}^t(P)\, \frac{e^{-jkr}}{r}\, E_2{}^t(P')\, ds\, ds' \right|^2}{\lambda^2 \int_{s_1} |\, E_1{}^t(P)\,|^2 ds \int_{s_2} |\, E_2{}^t(P')\,|^2 ds'} \tag{5}$$

where P and P' are points on the aperture surfaces S_1 and S_2 respectively. Assuming the field at the aperture of the transmitting horn is the same as though the horn were continued (i.e., the usual Kirchhoff approximation), the tangential electric fields in the aperture are given by

$$E_1{}^t = E_1{}^0 \cos \frac{\pi y}{a} \exp -\left[jk \left(\frac{x^2}{2l_E} + \frac{y^2}{2l_H} \right) \right] \tag{6}$$

$$E_2{}^t = E_2{}^0 \cos \frac{\pi \eta}{a} \exp -\left[jk \left(\frac{\zeta^2}{2l_E} + \frac{\eta^2}{2l_H} \right) \right] \tag{7}$$

where l_E and l_H are the E- and H-plane slant heights respectively. The distance r may be approximated by

$$\begin{aligned} r &= [R^2 + (x - \zeta)^2 + (y - \eta)^2]^{\frac{1}{2}} \\ &\approx R + \frac{(x - \zeta)^2 + (y - \eta)^2}{2R}. \end{aligned} \tag{8}$$

All pertinent dimensions are illustrated in Fig. 1. Since the gain measurements usually involve two identical horns, $S_1 = S_2$, and substituting (6), (7), and (8) into (5), (2) reduces to the near-field gain in the Fresnel approximation:

$$G_N = \frac{4\pi \left| \int_s \int_{s'} \cos \frac{\pi y}{a} \cos \frac{\pi \eta}{a} \exp -\left[jk \left\{ \frac{x^2 + \zeta^2}{2l_E} + \frac{y^2 + \eta^2}{2l_H} + \frac{(x - \zeta)^2}{2R} + \frac{(y - \eta)^2}{2R} \right\} \right] ds\, ds' \right|}{\lambda^2 \int_s \cos^2 \frac{\pi y}{a}\, ds} \tag{9}$$

while the Fraunhofer gain is

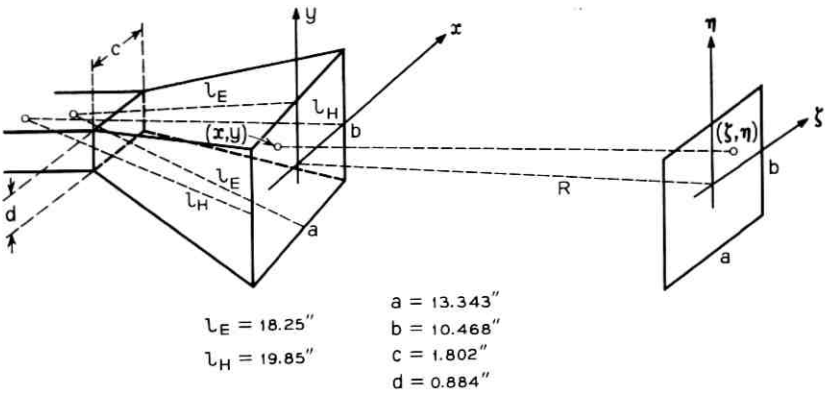

Fig. 1 — Physical dimensions for transmission between two electromagnetic horns.

$$G = \frac{4\pi \left| \int_s \int_{s'} \cos \frac{\pi y}{a} \cos \frac{\pi \eta}{a} \cdot \exp - \left[jk \left(\frac{x^2 + \zeta^2}{2l_E} + \frac{y^2 + \eta^2}{2l_H} \right) \right] ds\, ds' \right|}{\lambda^2 \int_s \cos^2 \frac{\pi y}{a}\, ds} . \quad (10)$$

Dividing (10) by (9) yields the required correction ratio. It is convenient to split this ratio into the E-plane correction and the H-plane correction

$$C = (G/G_N) = C_E C_H \quad (11)$$

where

$$C_E = \frac{\left| \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} \exp - \left[jk \left(\frac{x^2 + \zeta^2}{2l_E} \right) \right] dx\, d\zeta \right|}{\left| \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} \exp - \left[jk \left(\frac{x^2 + \zeta^2}{2l_E} \right) \right] \cdot \exp - \left[jk \frac{(x - \zeta)^2}{2R} \right] dx\, d\zeta \right|} \quad (12)$$

and

$$C_H = \frac{\left| \int_{-a/2}^{a/2} \int_{-a/2}^{a/2} \cos \frac{\pi y}{a} \cos \frac{\pi \eta}{a} \exp - \left[jk \left(\frac{y^2 + \eta^2}{2l_H} \right) \right] dy\, d\eta \right|}{\left| \int_{-a/2}^{a/2} \int_{-a/2}^{a/2} \cos \frac{\pi y}{a} \cos \frac{\pi \eta}{a} \exp - \left[jk \left(\frac{y^2 + \eta^2}{2l_H} \right) \right] \cdot \exp - \left[jk \frac{(y - \eta)^2}{2R} \right] dy\, d\eta \right|} . \quad (13)$$

The numerators in the above expressions may be identified as Fresnel integrals. After normalizing the parameters, we have

$$C_E = \frac{M\left[C^2\left(\frac{2}{\sqrt{M}}\right) + S^2\left(\frac{2}{\sqrt{M}}\right)\right]}{\left\{\left[\int_{-1}^{1}\int_{-1}^{1} \cos 2\pi\left(\frac{\omega^2+\zeta^2}{M} + \frac{(\omega-\zeta)^2}{H}\right) d\omega\, d\zeta\right]^2 + \left[\int_{-1}^{1}\int_{-1}^{1} \sin 2\pi\left(\frac{\omega^2+\zeta^2}{M} + \frac{(\omega-\zeta)^2}{H}\right) d\omega\, d\zeta\right]^2\right\}^{\frac{1}{2}}} \tag{14}$$

and

$$C_H = \frac{\frac{N}{4}\{[C(f) - C(g)]^2 + [S(f) - S(g)]^2\}}{\left\{\left[\int_{-1}^{1}\int_{-1}^{1} \cos\frac{\pi}{2}u \cos\frac{\pi}{2}v \cos 2\pi\left(\frac{u^2+v^2}{N} + \frac{(u-v)^2}{P}\right) du\, dv\right]^2 + \left[\int_{-1}^{1}\int_{-1}^{1} \cos\frac{\pi}{2}u \cos\frac{\pi}{2}v \sin 2\pi\left(\frac{u^2+v^2}{N} + \frac{(u-v)^2}{P}\right) du\, dv\right]^2\right\}^{\frac{1}{2}}} \tag{15}$$

where

$$M = 8\lambda l_E/b^2 \qquad H = 8\lambda R/b^2$$

$$N = 8\lambda l_H/a^2 \qquad P = 8\lambda R/a^2$$

$$f = \frac{1}{\sqrt{2}}\left(\sqrt{\frac{N}{8}} + \frac{1}{\sqrt{\frac{N}{8}}}\right) \qquad g = \frac{1}{\sqrt{2}}\left(\sqrt{\frac{N}{8}} - \frac{1}{\sqrt{\frac{N}{8}}}\right).$$

The Fresnel integrals are defined as

$$C(u) = \int_0^u \cos\frac{\pi}{2}t^2\, dt \qquad \text{and} \qquad S(u) = \int_0^u \sin\frac{\pi}{2}t^2\, dt$$

Equations (14) and (15) have been programmed for a digital computer; the results are summarized in Tables I and II.

It is interesting to notice that there exists substantial discrepancy between our correction ratios and those in Braun's article,[3] especially at short separations. In addition to the approximations made here, Braun employed an averaging process in which the power of the transmitted wave is integrated over the effective receiving aperture area $(\lambda^2/4\pi)G$. Therefore the correction ratios presented here are expected to be much more accurate than Braun's data and they should be useful for precision gain measurement of pyramidal electromagnetic horns.

TABLE I — E-PLANE CORRECTIONS (db)

H / M	8	16	32	64	128	256
2.0	1.740	0.997	0.520	0.263	0.132	0.066
2.5	1.585	0.856	0.426	0.210	0.104	0.051
3.0	1.490	0.757	0.362	0.175	0.085	0.042
3.5	1.418	0.684	0.317	0.150	0.073	0.036
4.0	1.359	0.627	0.284	0.133	0.064	0.031
5.0	1.268	0.547	0.237	0.108	0.051	0.025
6.0	1.201	0.492	0.207	0.092	0.043	0.021
8.0	1.109	0.423	0.168	0.072	0.033	0.016
10.0	1.050	0.381	0.145	0.060	0.027	0.013
32.0	0.870	0.261	0.081	0.028	0.011	0.005
∞	0.779	0.205	0.052	0.010	0.003	0.001

$M = 8\lambda l_E/b^2$ $\qquad$ $H = 8\lambda R/b^2$

III. MEASUREMENT TECHNIQUE

The standard horn was mounted in a wooden structure suitably covered with hairflex absorber; a sketch of the horn and its physical dimensions are shown in Fig. 1. A level monorail track was installed along the center line of the floor of an anechoic chamber. A stable, wooden equipment cart was designed to move smoothly along the monorail. One of two identical standard horns with hairflex baffle was mounted on the equipment cart (Fig. 2), the other being mounted in the end wall of the chamber. The equipment set-up is quite conventional and is shown schematically in Fig. 3.

The following procedure was used in the measurements: a reference level was set by removing the standard horns and connecting the waveguides directly (Fig. 3). With the standard horns in place and separated by $r \geqq 2a^2/\lambda$, a series of measurements of received power versus increas-

TABLE II — H-PLANE CORRECTIONS (db)

P / N	8	16	32	64	128	256
2.0	0.833	0.422	0.209	0.104	0.052	0.026
2.5	0.772	0.376	0.181	0.089	0.044	0.022
3.0	0.717	0.336	0.159	0.077	0.038	0.019
3.5	0.671	0.304	0.141	0.067	0.033	0.016
4.0	0.633	0.279	0.127	0.060	0.029	0.014
5.0	0.575	0.242	0.107	0.049	0.024	0.012
6.0	0.533	0.216	0.093	0.042	0.020	0.010
8.0	0.478	0.183	0.075	0.033	0.015	0.007
10.0	0.443	0.162	0.064	0.027	0.013	0.006
32.0	0.340	0.103	0.033	0.012	0.005	0.002
∞	0.291	0.071	0.019	0.005	0.001	0.0002

$N = 8\lambda l_H/a^2$ $\qquad$ $P = 8\lambda R/a^2$

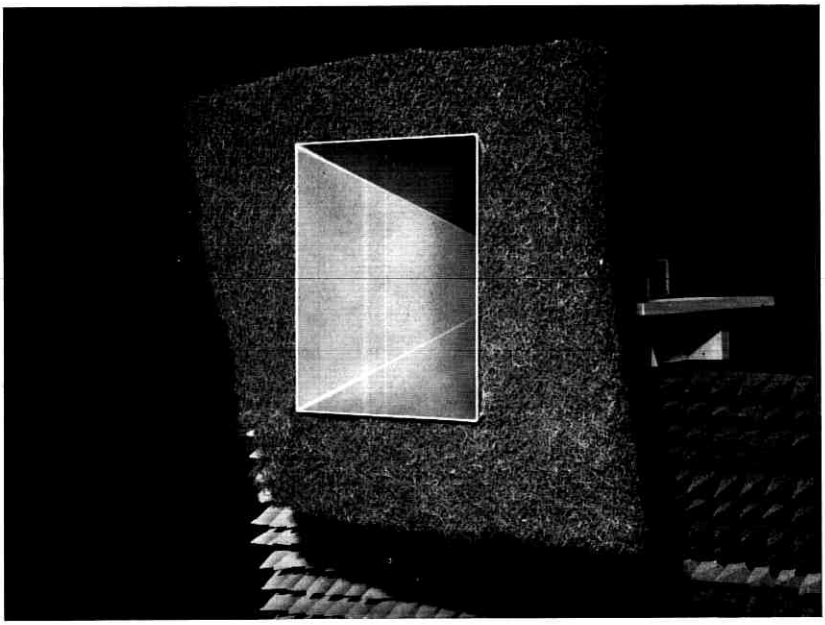

Fig. 2 — Standard horn mounted on equipment cart.

ing (r) were made. After completion of such a series, the reference level was rechecked by removing the standard horns and connecting the waveguides together. The above procedure was repeated several times for vertical and horizontal polarizations.

IV. RESULTS OF MEASUREMENT

The distribution of all of the measured gains at 4080 mc has been plotted as a histogram in Fig. 4. The near-field correction discussed above has been applied to these data. It should be pointed out that occurrences falling on the boundary lines of the columns have been evenly divided between the two neighboring columns; this accounts for the half occurrences which appear in the heights of some of the columns. The mean value of this sample distribution is 20.11 db, and its standard deviation is 0.05 db. The central limit theorem of probability theory indicates a 99.7 per cent confidence interval of $\bar{X} \pm (3\sigma/\sqrt{n})$ for the true mean, where $\bar{X}$ is the sample mean, n is the sample size, and σ is the population standard deviation.[5] Since the present sample size is 90, the population standard deviation should be close to the above sample standard devia-

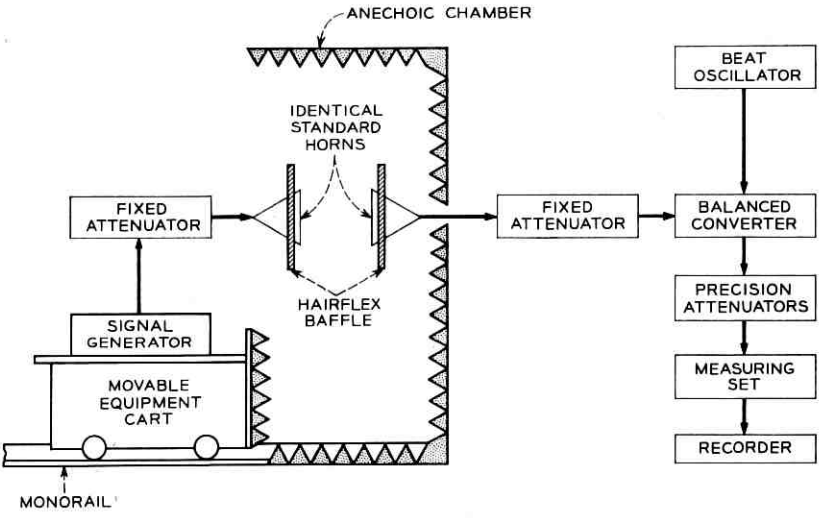

Fig. 3 — Equipment set-up.

tion, 0.05 db; therefore the random error in the mean value 20.11 db is of the order of ± 0.016 db ($3 \times 0.05/\sqrt{90}$).

The spread in the measured gain may be attributed to the following factors:

1. measuring system stability	± 0.01 db
2. precision attenuator readings	± 0.015 db
3. repeatability of electrical connections	± 0.015 db
4. imperfection of the anechoic chamber	± 0.02 db
5. interaction between the transmitting horn and the receiving horn	± 0.04 db.

The figures for the above factors are the estimates for one horn; they are half the probable random errors in the transmission between two horn antennas. Half of the measured gains were obtained when the horn apertures were vertically polarized, and half when horizontally polarized; when compared, the difference between the means of the two samples is only 0.01 db. This comparison implies only small errors due to the anechoic chamber.

The interaction effect is clearly demonstrated by the measured ($\lambda/2$)-period oscillation versus separation shown in Figs. 5(a) and 5(b). The amplitude of the oscillation is of the order of 0.05 db, and agrees fairly well with the qualitative calculation of Silver.[6]

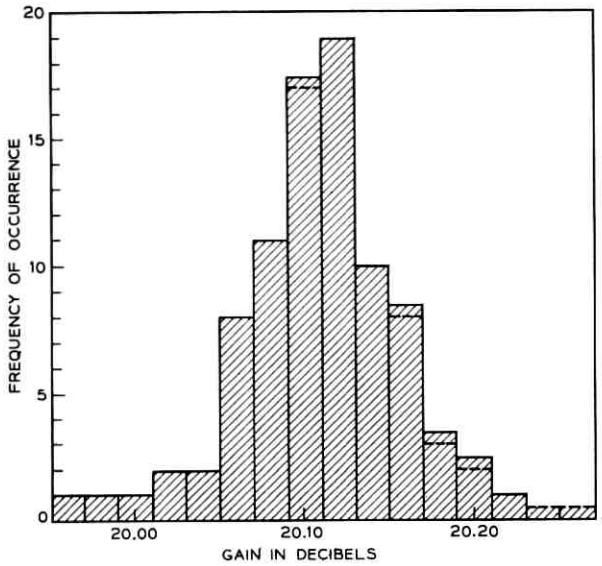

Fig. 4 — Histogram for the measured gains.

In addition to the random errors discussed above, the calibrated precision attenuators hide an absolute error which is constant for all measured gains. The probable value of this error is ± 0.04 db in the power transmission measurement, which contributes ± 0.02 db to the gain error. It follows that the total possible error of the measured gain (which includes the random error and the absolute attenuator error) is about ± 0.035 db. The calculated gain[7] of the standard horn is 20.15 db at 4080 mc. The discrepancy between the calculated value and the measured gain (20.11 db) is 0.04 db.

It should be pointed out that both transmitting and receiving horns in this gain measurement are isolated by 10-db fixed attenuators. However the mismatch at the horn-waveguide junction is not tuned out, because this same mismatch was not tuned out when the standard horn was used as a calibration reference for the gain measurement of the large horn-reflector antenna. A VSWR measurement revealed a reflection coefficient of -25 db, which represents a transmission loss of 0.015 db.

V. SUMMARY AND CONCLUSIONS

Using the near-field power transmission formula, the ratio between the Fraunhofer and Fresnel gain of a pyramidal electromagnetic horn

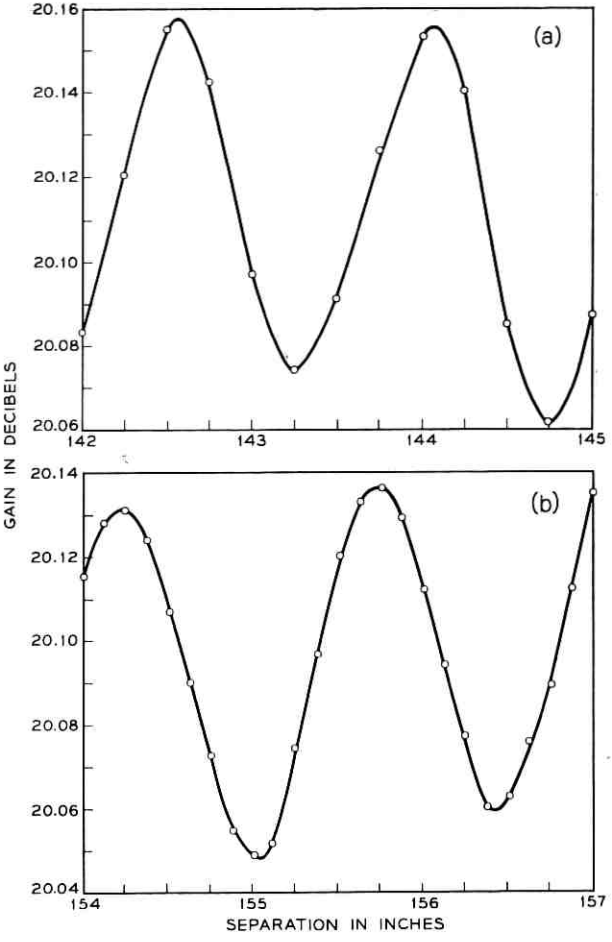

Fig. 5 — Measured gain variation due to interaction.

has been computed as a function of horn dimensions and separation distance. Our computations are expected to be much more accurate than previous data and should be very useful for precision gain measurement of pyramidal electromagnetic horns.

An application of the calculated corrections was made in the absolute gain measurement of a standard horn. The measured gain of the standard horn at 4080 mc is 20.11 db with an accuracy of ± 0.035 db; the calculated gain is 20.15 db. The interaction between two standard horns may introduce an error of the order of 0.05 db in the gain measurement

at a separation distance of $2a^2/\lambda$; however, it is reduced considerably by taking the average of several measurements. The averaging procedure can also reduce other random errors due to environment, measuring system stability, attenuator readings, etc. Using the corrections presented above, together with other careful considerations, it is possible to achieve an accuracy well below 0.1 db in the gain measurement of pyramidal electromagnetic horns.

VI. ACKNOWLEDGMENT

The authors are indebted to Mrs. C. L. Beattie for programming the computations.

REFERENCES

1. Hogg, D. C., and Wilson, R. W., A Precise Measurement of the Gain of a Large Horn-Reflector Antenna, to be published.
2. Jakes, W. C., Gain of Electromagnetic Horns, Proc. IRE, *39*, Feb., 1951, pp. 160–162.
3. Braun, E. H., Gain of Electromagnetic Horns, Proc. IRE, *41*, Jan., 1953, pp. 109–115.
4. Hu, M. K., Near-Zone Power Transmission Formulas, IRE National Convention Record, *6*, Pt. 8, 1958, pp. 128–135.
5. Anderson, R. L., and Bancroft, T. A., *Statistical Theory in Research*, McGraw-Hill, p. 71.
6. Silver, S., *Microwave Antenna Theory and Design*, McGraw-Hill, p. 592.
7. Schelkunoff, S. A., *Electromagnetic Waves*, D. Van Nostrand, p. 364.

Contributors to This Issue

B. A. Auld, B.A.Sc. (EE), 1946, University of British Columbia; MS, 1949, an Phd.D. (EE), 1952, Stanford University; Electrical and Musical Industries Ltd., London, 1953–1955; staff, University of British Columbia, 1955–1958; research staff, Stanford University, 1958—; Visiting Fellow, Bell Telephone Laboratories, 1963–1964. His work has related to the theory of microwave circuits and interactions of microwave fields with spin waves and acoustic waves.

Vaclav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory, and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. He is the author of *General Stochastic Processes in the Theory of Queues* (Addison-Wesley, 1963). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mind Association and Phi Beta Kappa.

Ta-Shing Chu, B.S., 1955, National Taiwan University; M.S., 1957, Ph.D., 1960, Ohio State University; Bell Telephone Laboratories, 1963—. He has worked in the field of electromagnetics with emphasis on surface waves and microwave antennas. At present he is working on optical and infrared wave propagation through the atmosphere. Member, IEEE, American Physical Society, Sigma Xi, and Pi Mu Epsilon.

Harold S. Edwards, B.S.E.E., 1925, Yale University; The Southern New England Telephone Company, 1925–1946; American Telephone and Telegraph Company, 1946—. At the Southern New England Telephone Company he was concerned with various assignments in the Plant and Engineering Departments. At the present time he is Plant Facilities Design Engineer, with responsibilities for the design of the outside plant subscriber network as well as methods and administration of the outside plant engineering job.

WALTER J. C. GRANT, A.B., 1951, M.A., 1952, Boston College; B.S., 1958, Ph.D., 1962, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1962—. He has been engaged in theoretical study of paramagnetic and electro-optic solid-state devices. Member, American Physical Society and Sigma Xi.

HENRY Z. HARDAWAY, B.S.M.E., 1940, University of Iowa; Southern Bell Telephone and Telegraph Company, 1940–42; Bell Telephone Laboratories, 1942—. He was first engaged in military equipment design work on airborne and submarine radar and navigational equipment. He is currently directing systems engineering studies which involve basic design concepts of the outside plant exchange cable network and which have important interfaces with transmission and switching equipment. Also, he is involved in the application of operations analysis techniques to the Associated Companies' operating and engineering problems.

HERWIG KOGELNIK, Dipl.-Ing., 1955, Dr. techn., 1958, Technische Hoschschule Wien, Austria; D.Phil., 1960, Oxford University, England; Bell Telephone Laboratories, 1961—. He is engaged in optical maser research. Member, American Physical Society, IEEE, Elektrotechnischer Verein Osterreichs (Austria).

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. He has been concerned with analysis of military systems, particularly radar systems, and with synthesis and analysis of active and time-varying networks. He is currently involved in a study of the signal-theoretic properties of nonlinear systems. Member, IEEE, SIAM, Eta Kappa Nu, Sigma Xi and Tau Beta Pi.

R. A. SEMPLAK, B.S., 1961, Monmouth College; Bell Telephone Laboratories, 1955—. He has been engaged in beyond-the-horizon radio propagation and three satellite communications projects: Project Echo, Telstar I and Telstar II. He has also participated in studies of the effects of rain on sky noise temperatures at 6-gc frequency and has recently completed an experimental study of the near-field Cassegrainian antenna. He is currently engaged in measuring the scattered radiation from various surfaces at 0.6-micron wavelength.

TZAY Y. YOUNG, B.S., 1955, National Taiwan University; M.S., 1959, University of Vermont; D.E.E., 1962, John Hopkins University;

Bell Telephone Laboratories, 1963—. He has been engaged in the investigation of the statistical extraction and detection of signals overlapping in time. Currently he is on leave from Bell Laboratories, teaching as an Assistant Professor at the Carnegie Institute of Technology. Member, IEEE, AAAS and Sigma Xi.

B.S.T.J. BRIEFS

Modulation of Laser Beams by Atmospheric Turbulence

By M. SUBRAMANIAN and J. A. COLLINSON

(Manuscript received January 11, 1965)

When laser beams are propagated through the air, they are modulated with a noise-like spectrum[1,2] having a baseband width the order of hundreds of cycles and a nearly exponential frequency distribution.[1] Hogg[1] used a 2.6-km path; Hinchman and Buck[2] used paths of 9 and 90 miles. In each case the optics and range were such that the receiver collected a small fraction of the total beam. Since atmospheric refraction causes twinkling and tearing of the beam, one would expect amplitude modulation of the signal received under these conditions even for constant intensity of the total beam.

We report here that the shape of the noise spectrum is unchanged when all of the detectable beam is received. Moreover, the spectrum is unaffected by changes in the diameter or geometrical divergence of the transmitted beam, by whether the receiver is in the near or far field of the transmitter, or by a threefold change in transmission distance. The general spectrum characteristics appear to be determined by atmospheric conditions.

We have transmitted a horizontally polarized 6328-Å laser beam over a 120-meter path 8 meters above black-top pavement. The beam was detected through a 3Å-wide interference filter by an RCA 7265 photomultiplier tube. The frequency spectrum of the signal was analyzed and displayed on a CRT by a Singer Metrics TA-2 spectrum analyzer. The resolution of the analyzer was 70 cycles, and its low frequency limit was 20 cycles. Each spectral analysis took one second. Generally, 120 successive spectral patterns were recorded, and thus averaged, in a single photograph of the CRT screen. The laser[3] oscillated in a single transverse and axial mode and provided about one milliwatt of power in a diffraction-limited Gaussian beam one millimeter in diameter. Measurements were taken with the direct beam, so that the receiver was very much in the far field of the transmitter. Telescopes of 9, 20, and 38 powers were used to enlarge the diameter of the transmitted beam, reduce diffraction spreading, and put the receiver in the near field. The telescopes were focused to vary the geometrical beam divergence, or convergence, thus greatly varying the size of the received beam. With the 38-power tele-

scope focused on the receiver, the beam diameter was 4 cm as transmitted, about 1 cm as received. The fraction of the total beam collected varied from one to much less than one with these variations in the transmitter. In no case did any of these changes produce a detectable change in the spectrum.

To investigate more carefully the effect of collecting part of the beam, the beam was transmitted plane parallel and 2 cm in diameter, diverging to 3 cm as received. (No signal above shot noise could be found beyond a 3-cm diameter.) An iris was placed before the receiver, and its diameter was adjusted from 4 to 1 cm, again with no effect on the spectrum.

To assess the possibility that the noise spectrum is produced by the product of the sensitivity profile of the photocathode by the time-varying intensity profile of the beam, a 4-inch, diffraction-limited lens was used at the receiver to focus the beam to a spot about one millimeter in diameter. (One millimeter is smaller than the scale of the structure of the photocathode sensitivity.) The spectrum was unchanged from that with no lens.

The effect of transmission distance was observed by splitting the beam at the receiving station, returning (with a corner reflector) part to the transmitting station and (with a flat mirror) reflecting that part again to the receiving station. Thus there were available to the receiver two beams, otherwise similar, which had traversed 120 m and 360 m of air. When the beams were switched on the receiver alternately with successive one-second scans by the analyzer, no change in the spectrum was seen. This was true at different times and under different conditions. Table I summarizes only the single most extensive run, lasting 5 hours and including thousands of individual spectra. The data entered are the widths, in cycles per second, from the maximum (at the 20-cycle cutoff of the analyzer) to the reduction from maximum shown. It is readily seen that there is no significant effect of distance on spectral width. Indeed, the agreement seems surprisingly good in view of the errors listed. The reason is that the errors are mean deviations of spectra which changed steadily over 5 hours, the spectra for the two distances changing together.

TABLE I — SHAPE OF SPECTRUM AT TWO DISTANCES

Distance	Width of Spectrum			
	Power level below 20-cycle peak by			
	−2½ db	−5 db	−7½ db	−10 db
120 meters	39 ± 8 cps	84 ± 14 cps	126 ± 19 cps	176 ± 25 cps
360 meters	38 ± 10	82 ± 22	124 ± 29	168 ± 40

This change in spectra correlated with changes in atmospheric conditions. As conditions changed in a pronounced way from one day to the next, so too did the spectra change in a pronounced way. Noise spectra obtained under five widely different weather conditions are plotted in Fig. 1. The curves are of the form $P(f) = \text{const} \exp(-\alpha f)$, and the value of α accompanies each curve. $P(f)$ is relative modulated power in the beam, and f is frequency. The data for curve A were obtained at 6:30 a.m. under an overcast sky. Thus the ground had cooled overnight and had not yet been warmed by the morning sun. Although a steady 10-mph wind was blowing, this was by far the narrowest spectrum observed. For B, the wind was steady, and there was sun on the pavement. For C, the wind was gusty, but there was no sun. For D, the wind was gusty, and there was sun. For E, the wind was violently gusty and there was heavy rain. The departure from exponential dependence probably was caused by the rain.

It appears that the spectrum is broadened to the extent that atmospheric conditions produce refractive gradients along the path. While the other variables may have affected the amplitude of the spectrum, they did not alter the shape. We are further investigating the effect of range, since at zero distance the amplitude is known to reduce to zero.

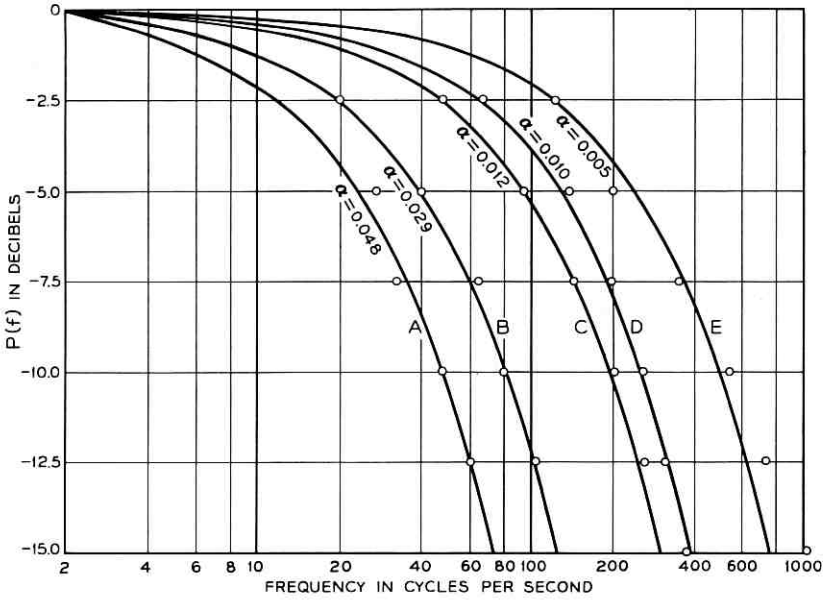

Fig. 1 — Dependence of the modulation spectrum on weather conditions. Refractive gradients increase from curve A to curve E as described in text.

The dependence of both spectral width and per cent modulation on range are of particular interest from a theoretical point of view. Spectral width should be independent of distance in the single-scatter regime, and per cent modulation should be small. The extent of the single-scatter regime depends upon the scale size of the refractive structure and upon the amplitude of variations in the refractive index. More extensive measurements are being made to allow a definitive comparison of theoretical expectations with the observations.

REFERENCES

1. Hogg, D. C., On the Spectrum of Optical Waves Propagated through the Atmosphere, B.S.T.J., *42*, Nov., 1963, pp. 2967–2969.
2. Hinchman, W. R., and Buck, A. L., Fluctuations in a Laser Beam over 9- and 90-Mile Paths, Proc. IEEE, *52*, March, 1964, pp. 305–306.
3. This is a single-frequency RF-excited laser to be described in a forthcoming publication by J. A. Collinson.