THE BELL SYSTEM TECHNICAL JOURNAL

DECEMBER 1983
VOL. 62, NO. 10, PART 1

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL (ISSN0005-8580) is published by the American Telephone and Telegraph Company, 195 Broadway, N. Y., N. Y. 10007; C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; T. O. Davis, Secretary.

The Journal is published in three parts. Part 1, general subjects, is published ten times each year. Part 2, Computing Science and Systems, and Part 3, single-subject issues, are published with Part 1 as the papers become available.

The subscription price includes all three parts. Subscriptions: United States—1 year $35; 2 years $63; 3 years $84; foreign—1 year $45; 2 years $73; 3 years $94. Subscriptions to Part 2 only are $10 ($12 foreign). Single copies of the Journal are available at $5 ($6 foreign). Payment for foreign subscriptions or single copies must be made in United States funds, or by check drawn on a United States bank and made payable to The Bell System Technical Journal and sent to Bell Laboratories, Circulation Dept., Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room 1J-319, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

Printed in U.S.A. Second-class postage paid at Short Hills, N. J. 07078 and additional mailing offices. Postmaster: Send address changes to The Bell System Technical Journal, Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 62 December 1983 Number 10, Part 1

Time-Compression Multiplexing (TCM) of Three Broadcast-Quality TV Signals on a Satellite Transponder

By K. Y. ENG,* B. G. HASKELL,* and R. L. SCHMIDT*

(Manuscript received June 15, 1983)

We describe how Time-Compression Multiplexing (TCM) might enable the transmission of three National Television System Committee (NTSC) color TV signals through a satellite transponder of 36-MHz bandwidth. The input TV signals are processed such that three fields from each TV source are compressed into an ordinary field period. This is accomplished by sending one field as is but time compressed; the other two fields are sent as differential signals, also time compressed such that all three fit into a single field period. The resultant compressed waveforms are then time multiplexed between the three sources and have a combined baseband bandwidth of 7.52 MHz for an optimal case, or 8.4 MHz for a practical version. In either case, both the transmitter-multiplexer and the receiver-demultiplexer require only three field memories for (digital) signal processing. Performance is expected to be of network broadcast quality (i.e., weighted signal-to-noise ratio, s/n $\geq$ 56 dB) for the optimal case of 7.52-MHz baseband if 12-meter receive earth stations

*Bell Laboratories.

are employed in a system such as COMSTAR. The practical version, on the other hand, would yield an s/n $\approx$ 54 dB.

I. INTRODUCTION

The problem of transmitting two or more high-quality TV signals through a satellite transponder of 36 MHz continues to be a challenge in optimizing the use of available transponders in current as well as near-future satellites. It was recently proposed[1] that by combining the concepts of Time-Compression Multiplexing (TCM)[2,3] and differential signals,[4] two or more National Television System Committee (NTSC) TV signals can be time multiplexed with bandwidth reduction for transmission with a single FM carrier in a satellite channel. This avoids crosstalk between the pictures. In fact, straightforward TCM alone would permit the transmission of two TVs in a transponder with performance close to broadcast quality [i.e., peak-to-peak video signal to weighted root-mean-square (rms) noise ratio, s/n $\geq$ 56 dB] if 12-meter receive earth stations were used in a satellite system such as COMSTAR. The additional application of time-companded (time-compressed or expanded) differential signals reduces the TCM signal bandwidth and thus can enhance the transmitted picture quality or enable the inclusion of a third TV signal. However, the implementation of such a system as described in Ref. 1 involves converting the input TV scan pattern from interlacing to sequential. This would mean considerable memory needed, particularly in the case of three TVs per transponder. Here, we describe an implementation that offers significant saving in memory, considerable relaxation in timing requirements, and easy adaptation to existing hardware. The technique essentially uses three field memories time shared between the three simultaneous, but *synchronized*, input TV signals to produce differential signals in a proper format for TCM. The receiver, on the other hand, also requires three field memories to reconstruct all three TV signals. Practically all the signal processing could be implemented digitally.

We will describe the details of the present system in the next section. The performance of this could be of broadcast quality if 12-meter receive earth stations were used. Finally, we will discuss the inclusion of audio, up-link synchronization for transmissions from separate earth stations and possible extensions to non-NTSC TV signals.

II. SYSTEM DESCRIPTION

Figure 1 shows the block diagram of a transmitting earth station with three *synchronized* NTSC TV signals that are combined for transmission by a single FM carrier. The use of frame synchronizers

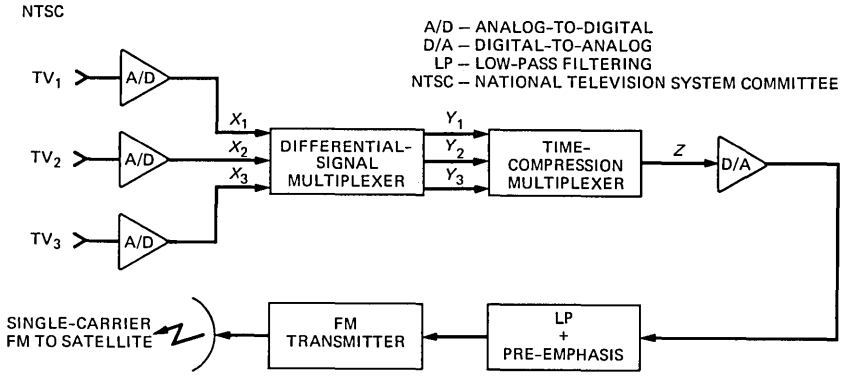


Fig. 1—Transmit earth station for the three TVs per transponder.

is therefore implicit if the three TV sources are not synchronized (the case of noncolocated TV sources is discussed later in Section IV). The TV inputs are first digitized individually. The digitized TVs, denoted by X_i ($i = 1$ to 3), are processed by the differential-signal multiplexer, where various differential signals are formed and multiplexed in its three digital outputs Y_i ($i = 1$ to 3). These signals (Y_i) are then passed through the time-compression multiplexer, which combines them into a single digital stream, Z. The conventional operations of digital-to-analog conversion, low-pass filtering, and preemphasis are performed before transmission to the satellite with a single FM carrier. We will describe the differential-signal multiplexer and the time-compression multiplexer in detail in the following sections.

2.1 The differential-signal multiplexer

We could use three types of differential signals: line differentials, field differentials, and frame differentials.[4,5] Each type in turn can be defined in many ways. They have all been described in the cited references, and only a brief summary is provided here for the purpose of subsequent discussions.

Line differentials can be defined as a difference signal between two successive scan lines in the same field. In their digital implementation, this would mean a difference between more or less vertically adjacent picture elements (pels) from two successive lines in the *same* field, and they are chosen such that their amplitude is much smaller, on the average, than the original signal. But most importantly, the difference signal can be band limited to ≈3 MHz without degrading picture quality. Field differentials are defined essentially in the same way as line differentials except that the difference signal is derived from pels in adjacent scan lines in two successive fields. The bandwidth of field differentials can be further limited to ≈2 MHz without affecting

picture quality. These results were verified and used in a previous experiment.[5]

Frame differentials are merely an extension of the above by using pels from two temporally adjacent (or spatially coincident) lines from two successive frames. They have not been studied so far, either by computer simulation or hardware implementation. Thus, we can only speculate as to their performance. Their amplitudes may be larger than field difference amplitudes for pictures containing movement, whereas for pictures containing no movement they should be smaller. The bandwidth required for frame differentials should be comparable to or smaller than that needed for field differentials. In this regard, much depends on the relative visibility in the picture of distortions occurring at the field rate and the frame rate in detailed or moving areas of the picture. In the following discussion, we will use the field and frame differentials; the use of line differentials will only be a possible, though unlikely, extension of the system.

Our attention returns now to the differential-signal multiplexer, an illustrative implementation of which is shown in Fig. 2. The following explanation will show that the field- and frame-differential generators in this figure could just as well be replaced by two field-differential generators with some connections slightly modified. The key characteristic to note in Fig. 2 is that only three field memories are needed to produce all the differential signals required for the three input TVs.

The three input switches, S_1, S_2 and S_3, move in synchronism from the top position to the middle, to the bottom, and back to the top, etc.

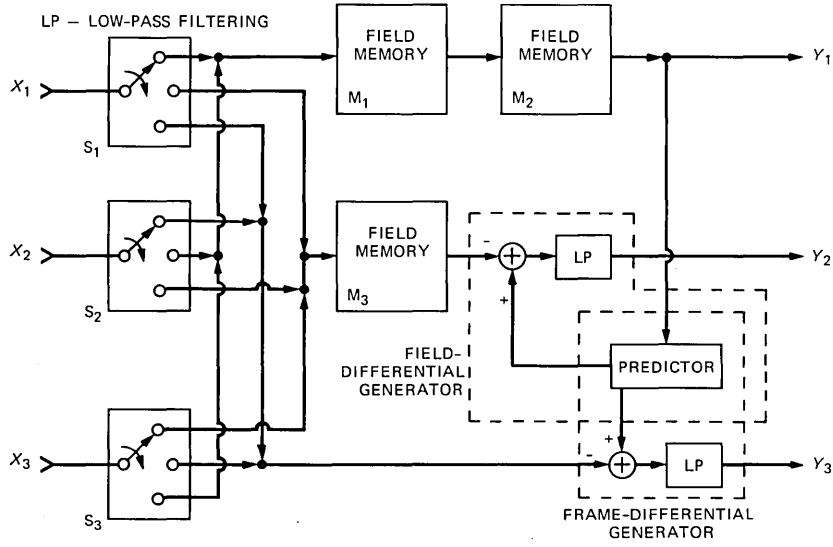


Fig. 2—The differential-signal multiplexer.

They all change position simultaneously sometime during the vertical blanking interval in such a way that complete fields of the input video are routed to either the top, middle, or bottom path.

To demonstrate how this works, we consider Fig. 3, where all the waveforms are digital. In the top of the figure, the three synchronized input TV waveforms are shown with T being a field period ($\approx$1/60 second) and F_{ij} being the jth field from the ith TV source ($i = 1$ to 3). When F_{11} arrives, we assume that S_1, S_2, and S_3 are in the top position, as shown in Fig. 2. F_{11} is written onto field memory M_1 from time zero to T. The switches then change to the middle positions, and F_{12} is written onto M_3 while F_{11}, in M_1, is being transferred to M_2. At the same time, F_{21} is also written onto M_1. Consequently, at the end of $2T$, F_{21} is stored in M_1, F_{11} in M_2, and F_{12} in M_3 before the switches change position again. Now with the switches in the bottom position, F_{13} is routed to the bottom path. It is then used to form a frame differential with F_{11}, from M_2, denoted by $F_{11} - F_{13}$, which is the

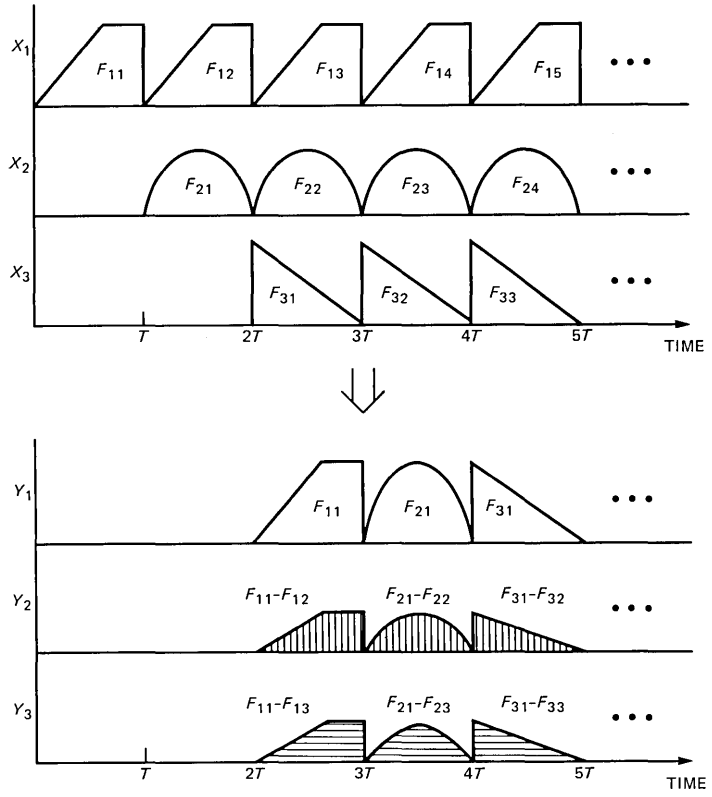


Fig. 3—Input/output waveforms for the differential-signal multiplexer.

output Y_3. The original unchanged signal F_{11} is also read out from M_2 via Y_1. The remaining output Y_2, is a field differential derived from F_{11} (from M_2) and F_{12} (from M_3) and is denoted by $F_{11} - F_{12}$. While all these are taking place, F_{21}, from M_1, is transferred to M_2 with F_{31} being written onto M_1, and F_{22} is written onto M_2. These operations are repeated for all subsequent fields. The output waveforms are illustrated in the bottom of Fig. 3, where a processing delay of $2T$ is incurred. Such a delay enables the conversion from line-multiplexed serial inputs into time-multiplexed parallel outputs. Furthermore, there is flexibility in choosing which of the fields is to be read out as is and which type of differential signal is to be used. For instance, in the above example we could just as well send F_{12} as is, send $F_{11} - F_{12}$ as a field-differential signal, and send $F_{13} - F_{12}$ as another field-differential signal. In any event, in every T-second output interval, there are always one original field plus two differential fields in the three outputs. The bandwidth of the original field is 4.2 MHz, and that of the differential signals is assumed to be 2 MHz.

2.2 The time-compression multiplexer

The purpose of the time-compression multiplexer is to combine the three signals (Y_1, Y_2, and Y_3) from the differential-signal multiplexer into a single signal, Z. In other words, we would like to time compress the three inputs over every T-second interval into a single output with the same duration. This can be achieved by writing the digital words into a memory (say, a RAM) at one speed and reading them out at a faster speed (see Fig. 4). The ratio of the read clock to the write clock is the time-compression factor (>1 for time compression). Since the time compression is to be done over a T-second interval, we could

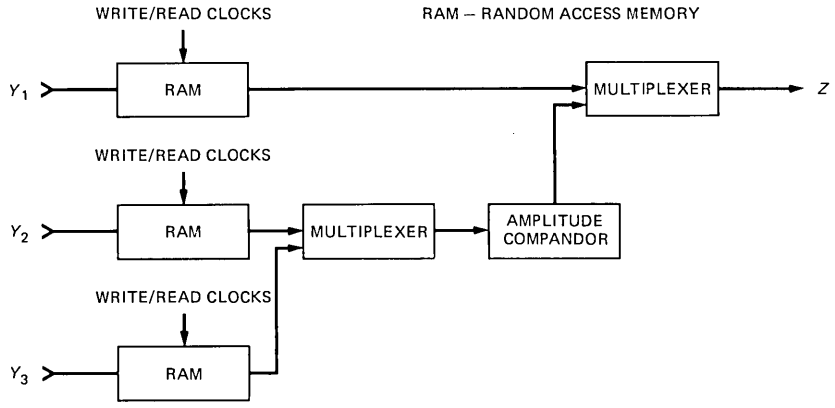


Fig. 4—The time-compression multiplexer.

write all Y_i's into the RAMs for the field period before reading them out appropriately for multiplexing. But this would require the RAMs to accommodate entire fields of signals. Instead of this, we propose that the time compression be done over a line interval ($\approx$63.6 μs) so that only line memories are needed. More specifically, let us consider a line duration T' within a T-second interval shown in Fig. 5. As before, Y_1 is the original 4.2-MHz TV; Y_2 and Y_3 are each a 2-MHz differential signal; and τ in the output Z is the processing delay. We time compress the T'-second line of Y_1 by a factor of $\alpha(\alpha > 1)$ so that the resultant signal occupies T_1 seconds ($T_1 < T'$). Likewise, Y_2 and Y_3 are both compressed by $\beta(\beta > 1)$ so that each of their resultants occupies T_2 seconds ($T_2 < T_1 < T'$). We require that these three time-compressed signals be contained in T', i.e.,

$$\frac{T'}{\alpha} + 2\,\frac{0.83T'}{\beta} = T'. \tag{1}$$

The factor 0.83 is due to the deletion of the differential-signal horizontal blanking intervals, which are identically zero and need not be sent. The above simplifies to

$$\frac{1}{\alpha} + \frac{1.66}{\beta} = 1. \tag{2}$$

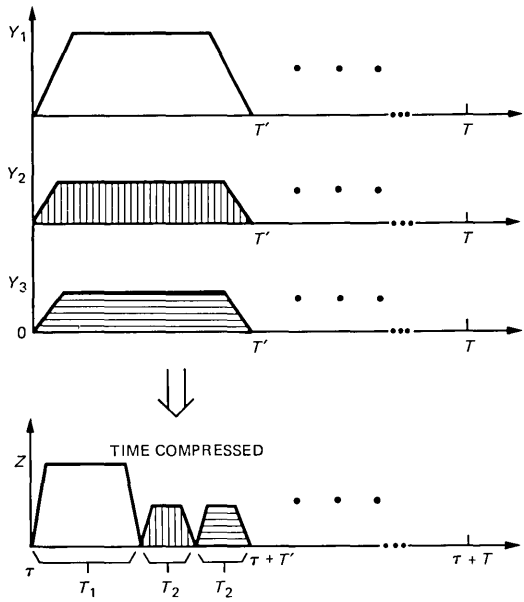


Fig. 5—Input/output waveforms for the time-compression multiplexer.

We further require that the three time-compressed signals have the same bandwidth. This can be written mathematically as[6]

$$\alpha f_1 = \beta f_2 = \beta f_3, \tag{3}$$

where f_1, f_2, and f_3 are the maximum frequencies of Y_1, Y_2, and Y_3, respectively. In this case, $f_1 = 4.2$ MHz, $f_2 = f_3 = 2$ MHz, and the solutions to (2) and (3) are

$$\alpha \approx 1.79;\ \beta = 3.76. \tag{4}$$

This yields $T_1 \approx 0.56T'$ and $T_2 \approx 0.22T'$. The maximum frequency of the combined output is given by (3) and is 7.52 MHz, as compared to 12.6 MHz obtained in a straightforward TCM of the three TVs. We call the above case *optimal* because its bandwidth has been minimized by the deletion of the horizontal blanking intervals in the differential signals. One obvious drawback, however, is that the compression factors required are noninteger, as given in (4). To circumvent this difficulty, we can simply choose $\alpha = 2$ and $\beta = 4$ exactly, i.e., compressing the original signal by two and the differential parts by four, with all their horizontal blankings retained. This *practical* case is much easier to implement with a slightly larger bandwidth of 8.4 MHz.

The last, but not the least, block in the time-compression multiplexer is the amplitude compandor (Fig. 4). As pointed out previously, the differential signals are chosen such that their amplitudes are small compared to the original signal on the average. The amplitude compandor equalizes the voltage levels for the differential signals in the combined output so that the FM link performance can be maximized. This was found to be very useful in a previous experiment[5] to suppress the effect of transmission noise on picture quality.

In summary, the present system takes in three NTSC TV signals and combines them into a 7.52-MHz (or 8.4-MHz) signal for transmission. The multiplexing technique is TCM, and the bandwidth reduction is the result of the use of differential signals. The transmission format is three fields from one TV source compressed into one ordinary field period. Thus, the transmission to the satellite is switched sequentially between the three sources at a rate equal to the field or vertical scanning frequency of ordinary NTSC TV (≈60 Hz). If the three TV sources are synchronized with one another, then the transmitter/multiplexer requires only three field memories. Otherwise, additional memory is needed for synchronization. In either case, the receiver requires only three field memories (see the appendix).

III. PERFORMANCE

Overall performance of time-compression multiplexing of multiple TV signals in a satellite link has never been measured experimentally.

But calculations for estimating TCM performance were shown in Ref. 1. According to these calculations, the optimal case of the present system, which has a baseband-combined bandwidth of 7.52 MHz for the three TV signals, would require a receive earth station with a Gain/Temperature (G/T) of ≈33.7 dB/K to yield a receive baseband TV s/n of 56 dB. Such a G/T value is obtainable from 12-meter earth stations. The practical version (8.4-MHz baseband bandwidth), on the other hand, would require a G/T of 35.9 dB/K to yield s/n = 56 dB. Such a G/T is probably not obtainable with 12-meter stations. However, the degradation in s/n by using 12-meter receive earth stations is only about 2 dB, i.e., the received s/n would be ≈54 dB.

IV. DISCUSSIONS

4.1 Audio

With three TV sources, each producing stereo audio, we must transmit a total of six audio waveforms. We propose sending the stereo audio from each source along with its video by inserting digital audio in either the vertical or horizontal blanking periods. As for the optimal case where the horizontal blanking periods of the differential signals are deleted in transmission, the audio signals may be included in some convenient segment of the vertical blanking period. This of course will lead to a slightly more stringent timing requirement as well as some additional buffer memory.

As for the practical case where the horizontal blankings of the differential signals are retained for transmission, then the insertion of digital audio in these blanking periods can be done quite easily. Within a group of three video lines (one unchanged original plus two differentials), there are two horizontal blankings from the differential lines available. We can use one of these two blankings for one audio and the other blanking for the other audio. Within one of these time-compressed horizontal blanking intervals (≈2.7 μs), we must include the audio samples from three TV scan-line durations. Now each audio signal requires sampling at ≈32 kHz, and with nearly instantaneous companding, 10 bits per sample are sufficient.[7] Thus, we propose sampling the audio at exactly twice the TV line-scan rate, yielding a total of six samples or 60 bits from the three scan lines for transmission in the prescribed time-compressed horizontal blanking period. For this we would use twenty multilevel pulses to represent the 60-bit information. At a baud rate equal to 9/4 × color subcarrier frequency (≈8.06 MHz), the six audio samples from the three lines plus another pulse for bit timing would just fill the 2.7 μs time slot.

There are several ways of mapping the 60+ bits from the three lines into twenty multilevel pulses. More discussions in this regard are

provided in Ref. 5. Because the three TV lines are from three different fields, additional memory is needed to store their audio samples, but this requirement seems trivial compared to the video counterpart.

4.2 Synchronization for multiple up-links

Use of TCM in satellite systems where up-links are not colocated requires that the three TV signals be synchronized, at least to the extent that their vertical blanking periods overlap.[8] This condition is not very stringent compared with that of some digital Time Division Multiple Access (TDMA) systems being proposed or in operation. Other than the additional synchronization hardware required for the transmitters, the only minor imposition in the system is that the receiver be able to demodulate the FM signal subject to short discontinuities in the received carrier at the vertical scanning frequency. Conventional limiter-discriminator receivers should have no problem in dealing with this. Phaselock receivers, on the other hand, might have lockup problems. But then the system is intended for high-quality transmissions with high carrier-to-noise ratios, and threshold extension is not needed.

As an aside, let us note that if the three TV sources are transmitted through noncolocated up-links, then the processing in each transmit earth station needs only two field memories (instead of three) to generate the differential signals required. The input switches in Fig. 2 are also unnecessary. A similar saving in receiver memory is possible too if only one TV is to be received in a down-link earth station.

4.3 Extension to non-NTSC TV signals

Application of this technique to non-NTSC color TV signals may also be feasible. For example, with Phase Alternation Line (PAL) color television the color subcarrier phase is not the same as NTSC. However, with only a slight shift in the sampling pattern from line to line, the same differential signals can be defined and the same transmission system can be used. The same may be true of Sequential With Memory (SECAM) color television, but success is not as likely.

V. CONCLUSION

We have described a method to transmit three NTSC TV signals in a 36-MHz satellite transponder. The technique uses differential signals to reduce the bandwidth and Time-Compression Multiplexing (TCM) to combine the three TVs into a single signal. By the use of novel circuit configurations, the memory requirements are reduced significantly compared with the more naive approach of Ref. 1. By companding the differential signals, the effect of transmission noise on

picture quality is markedly reduced. The estimated performance of the system is at or close to broadcast quality if 12-meter earth stations were to be used in a satellite system such as COMSTAR. Finally, digital audio signals can be sent without interference either to or from the video TCM signal by placing it in the horizontal blanking period. Extensions to up-links from separate earth stations and non-NTSC TVs are also possible.

REFERENCES

1. K. Y. Eng and B. G. Haskell, "TV Bandwidth Compression Techniques Using Time Companded Differentials and Their Applications to Satellite Transmissions," B.S.T.J., *61*, No. 10, Part 1 (December 1982), pp. 2917–27.
2. J. E. Flood and D. I. Urquhart-Pullen, "Time-Compression-Multiplex Transmission," Proc. IEE, *111*, No. 4 (April 1964), pp. 647–68.
3. D. H. Morgen and E. N. Protonotarios, "Time Compression Multiplexing for Loop Transmission of Speech Signals," IEEE Trans. Commun., *COM-22*, No. 12 (December 1972), pp. 1932–9.
4. B. G. Haskell, "Time-Frequency Multiplexing of Two NTSC Color Signals—Simulation Results," B.S.T.J., *60*, No. 5 (May–June 1981), pp. 643–60.
5. R. L. Schmidt and B. G. Haskell, "Transmission of Two NTSC Color Television Signals Over a Single Satellite Transponder Via Time-Frequency Multiplexing," Globecom '82 Conf. Rec., *1*, pp. B6.3.1–.5.
6. K. Y. Eng and O. Yue, "Spectral Properties and Band-Limiting Effects of Time-Compressed TV Signals in a Time-Compression Multiplexing System," B.S.T.J., *60*, No. 9 (November 1981), pp. 2167–85.
7. M. G. Croll et al., "Nearly Instantaneous Digital Compandor for Transmitting Six Programme Signals in a 2.048 Mbit/s Multiplex," Electron. Lett. (July 12, 1973), pp. 298–300.
8. Kai Y. Eng and B. G. Haskell, "Synchronization of Noncolocated TV Signals in a Satellite Time-Compression Multiplexing System," B.S.T.J., this issue.

APPENDIX

Decomposition of the TCM Signal From Three Video Sources

As Fig. 6 shows, the received FM signal from the satellite is demodulated to recover the TCM baseband waveform. It is then digitized to produce Z', which would be identical to Z previously except for the transmission noise and channel distortion added. An amplitude compandor undoes the companding done to the composite waveform. Now the three segments in this waveform, namely the original field and the two differential signals, are then separated by the demultiplexer and written onto three separate memories. They are read out at slower speeds to get time expanded to the full scan-line length. The expansion factors (ratio of write clock to read clock) are precisely the compression factors used in the transmitter. Approximations to Y_1, Y_2, and Y_3, denoted here by Y'_1, Y'_2, and Y'_3, are then obtained. The same predictor as in the transmitter is used to convert the differential signals into the originals. The three output switches, S_4, S_5, and S_6, move in synchronism from the top position to the middle, to the bottom, and back to the top, etc. Their operations are identical to S_1,

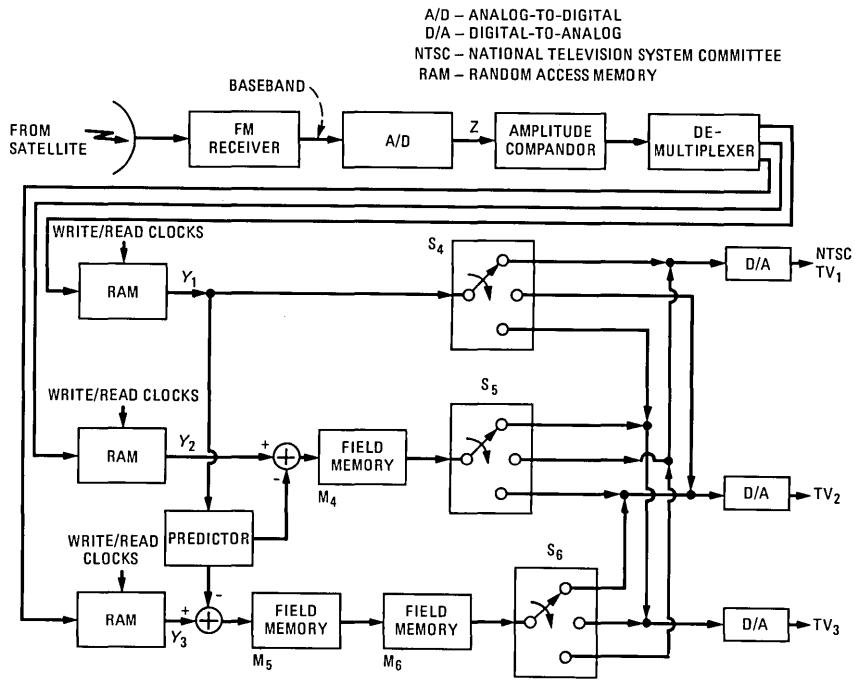


Fig. 6—Receive earth station for three TVs per transponder.

S_2, and S_3 in the transmitter, and they route the output digital signals to their appropriate outputs. The output digital signals may (or may not) then be converted to analog signals for display or local distribution.

AUTHORS

Kai Y. Eng, B.S.E.E. (summa cum laude), 1974, Newark College of Engineering; M.S. (Electrical Engineering), 1976, Dr. Engr. Sc. (Electrical Engineering), 1979, Columbia University; RCA Astro-Electronics, 1974–1979; Bell Laboratories, 1979—. Mr. Eng has worked on various areas of microwave transmission, spacecraft antenna analysis, and communications satellites. He is presently a member of the Radio Research Laboratory, studying TV transmission through satellites. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, Phi Eta Sigma.

Barry G. Haskell, B.S. (Electrical Engineering), 1964, M.S., 1965, and Ph.D., 1968, University of California, Berkeley; University of California, 1965–1968; Bell Laboratories, 1968—; Rutgers University, 1977–1979. Mr. Haskell was a Research Assistant at the University of California Electronics Research Laboratory and a part-time faculty member of the Department of Electrical Engineering at Rutgers University. At Bell Laboratories, he is presently Head of the Radio Communications Research Department, where his research

interests include television picture coding and transmission of digital and analog information via microwave radio. Member, IEEE, Phi Beta Kappa, Sigma Xi.

Robert L. Schmidt, B.S.E.E., 1982, Monmouth College; Bell Laboratories, 1972—. Mr. Schmidt is currently a member of the Radio Communications Research Department, where he has been exploring various techniques in combining multiple television signals onto a single analog radio channel. He has researched television signal encoding, bit rate reduction techniques, and software-controlled coding systems. Member, Eta Kappa Nu.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

Synchronization of Noncolocated TV Signals in a Satellite Time-Compression Multiplexing System

By K. Y. ENG* and B. G. HASKELL*

(Manuscript received June 15, 1983)

We describe here a simple method to synchronize three TV signals originated from noncolocated up-link stations in a satellite Time-Compression Multiplexing (TCM) system. In this system, information in three fields of each TV picture is compressed into a single field time so that the compressed signals from the three sources can be time multiplexed for transmission. The up-link synchronization ensures that the Radio Frequency (RF) bursts from different sources will arrive at the satellite without collision. Our method employs a dynamic master/slave arrangement whereby the first station signing on assumes the role of a master. The other stations subsequently can synchronize their transmissions to the master's by simply monitoring the received RF bursts from the satellite, measuring their respective delays to the spacecraft, and then phase locking their local color subcarrier clocks to the master's transmitted bursts. When the master station stops transmitting, an automatic procedure is provided for the second station to take over as the new master. The worst-case jitter performance is well below 100 ns, and the initial acquisition time can be kept less than one-half second. These are more than adequate for the present TV application, although further improvements are possible if necessary.

I. INTRODUCTION

Time-Compression Multiplexing (TCM) is a method of multiplexing various signals by time compressing their (analog) waveforms into segments in such a way that the compressed segments from different

* Bell Laboratories.

sources can be sent on the same channel in separate time intervals (time-division multiplexing).[1,2] Previous published works[3–5] have discussed various properties and ways to implement TCM in the transmission of multiple high-quality TV signals through a single satellite transponder of 36-MHz bandwidth. More recently, this idea has further been refined to the transmission with practical hardware of three broadcast-quality TVs in a transponder[6] (i.e., the received peak-to-peak video to weighted rms noise ratio $\geq$56 dB.) As with other TCM systems, one requirement in the latter proposal is that the input three TV signals be synchronized, at least to the extent that their vertical-blanking intervals overlap. If the signals are colocated in the same up-link earth station, it merely implies that frame synchronizers be used. However, if they are to be transmitted from separate earth stations, then the up-links have to be synchronized. Of course, the up-link synchronization is needed to ensure that signal bursts from different sources would arrive at the satellite without collision. We show and discuss in this paper how this can be accomplished with simple and easy-to-implement hardware arrangements.

Synchronization techniques in communications satellite systems have been studied extensively in past years,[7–9] mostly in connection with digital Time Division Multiple Access (TDMA) applications. They could all be used in the present problem of synchronizing three TV up-links. However, these previous techniques were designed for performance far exceeding the present requirement and hence tend to be more complicated than what is needed. More importantly, they were meant for digital signals and are not suitable for analog TV where the color subcarrier and various sync pulses must bear strict phase and frequency relationships and thus cannot be advanced or retarded with respect to one another arbitrarily. We will show in the next section how a TV up-link station can synchronize its transmission by simply monitoring the Radio Frequency (RF) bursts sent by other station(s) already on the air. Such an approach enables synchronization between the three stations without a centrally controlled master station or clock, without the knowledge of one another's exact location, without the demodulation of one another's baseband video, and without the use of a separate control channel. The only assumption imposed is that the three up-link stations be within the down-link coverage of the satellite. This is true for satellites similar to Telstar III. The hardware implementation is quite simple (Section III) and can be realized by conventional equipments and digital circuits. Our timing analysis (Section IV) shows that its performance can cover all requirements under a variety of worst-case conditions, and simple procedures for failure recovery are discussed in Section V. Finally, we will make brief comparisons with other methods by showing a number

of practical advantages in using the present technique and also discuss possible extensions to further improve its performance.

II. SYSTEM DESCRIPTION

We outline in this section the basic concept and operation of the present method. Detailed parameters and performance evaluation are left for subsequent discussions. The system configuration is illustrated in Fig. 1, where three up-link earth stations are to transmit their color TV signals to a satellite. The TV pictures are assumed to be National Television System Committee (NTSC) and are to be time compressed with processing prior to transmission so that TCM can be employed. More specifically, three fields of each TV are to be time compressed into one field period, F, ($\approx$1/60s) in a manner previously described in Ref. 6. The resulting waveform of a time-compressed TV contains successive triplets of a field with picture information followed by two blank field periods. The RF transmission of each earth station will

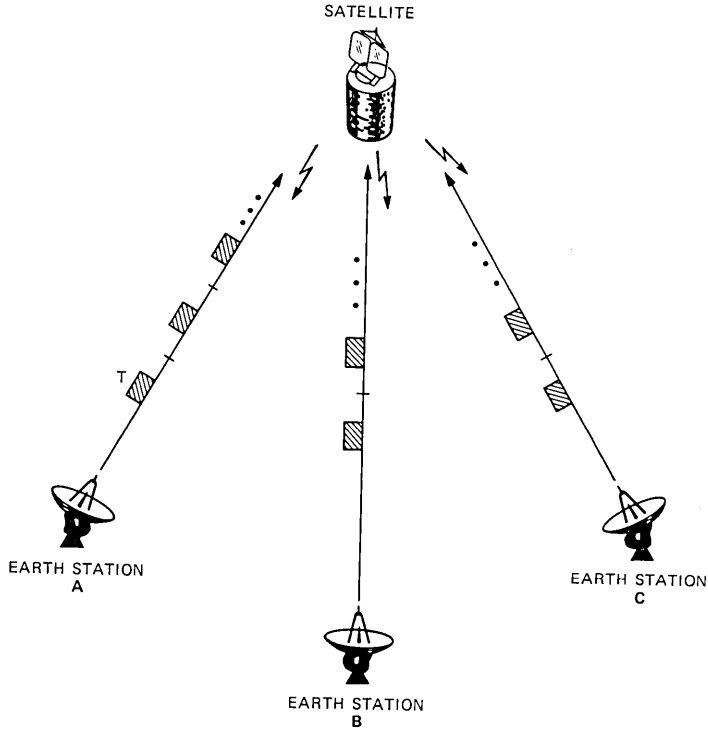


Fig. 1—A three TV/transponder TCM system.

then consist of bursts, each having approximately one field duration, with two blank field periods as separation between successive bursts (Fig. 2). The synchronization problem at hand is to align these bursts from the three stations so that they arrive at the satellite without overlap. All three stations are assumed to be within the down-link coverage of the satellite.

One could design the system, at least in principle, such that the entire portion of the vertical-blanking interval (≈1.4 ms) within each TV burst is used for guard time. This would be sufficient to account for the diurnal drift of the satellite itself (maximum round-trip delay variation of about 500 μs according to Ref. 7). With the exact locations of the stations known, simple open-loop synchronization is then possible. The drawback of such an approach is twofold. First, the deletion of the entire vertical blanking is undesirable in TV transmission because a variety of test signals and nonvideo information are frequently inserted in this time period. Second, the exact known location requirement renders the scheme inflexible for the inclusion of transportable transmit earth stations.

We feel that the deletion of only a portion of a scan line (during vertical blanking, say 15 μs) for interburst guard time is reasonable and would not limit or interfere with picture performance. In addition, we do not assume that locations of the stations are known to one another. Instead, each station is assumed to know only its own approximate location, say to within ±100 km. Note that the latter assumption is not imposing at all since every station needs some location information of its own for antenna pointing purposes anyway.

To illustrate the operation of the present system, the three up-link

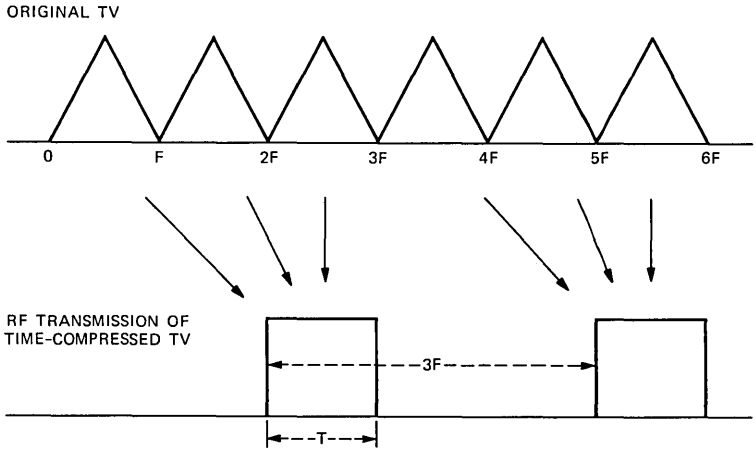


Fig. 2—Time-compression processing of a TV signal (F = Field period, T = F minus a small guard time).

earth stations are labelled A, B, and C. Station A is assumed to be the first to transmit. It can do so at will because no other transmission is taking place, and its transmission is simply synchronized to its own NTSC TV clock.

We now consider the start-up of B after A has been on the air. Station B first monitors the arrivals of the RF bursts from A and records their arrival times. Note that B does not have to demodulate A's signal; it only needs to detect the RF pulses received. (Indeed, A's baseband signal need not be video, as long as its RF timing is otherwise compatible.) The RF pulses from A occur in one out of three fields, the period is perturbed mainly by the time-varying propagation delay between A and B due to the spacecraft motion. Using these arrival times, B can extrapolate for the immediate future arrivals of A's pulses, and with the knowledge of its own approximate location (±100 km), B can compute its propagation delay to the satellite with an accuracy better than ±1.2 ms (including satellite drift). This estimated delay enables the translation of the arrival times of A's bursts from the time reference at B to that at the satellite. Using all this information, B can then position the transmission of a narrow pulse so that it arrives at the satellite in a time window adjacent to a burst from A, but not interfering with it. This narrow pulse is then received back by B, and we have an actual delay measurement, done inband, between B and the satellite. Once the actual delay is obtained, B can derive a windowing signal (frequency = one-third of the TV field rate) that denotes the proper transmission times in order to maintain collision-free synchronization with A.

The derivation of this window signal at B would mean the end of the problem if the system were for digital transmission. However, for TV applications, the picture information cannot be arbitrarily advanced or delayed without regard to the phase and frequency relationships between its color subcarrier and its sync pulses. Therefore, we propose frame (or field) synchronizing the TV picture at B to a local color subcarrier clock that is in turn phase locked to the aforementioned window signal in order to achieve proper transmission timing. This will be explained further when we discuss the hardware implementation.

Note that throughout the above procedure of synchronizing B to A, the up-link delay from A to the satellite remains unknown to B. This is possible because the timing error of B's narrow pulse (as will be shown later) is small compared with the start-up guard time allotted, i.e., a field period. Subsequent synchronization is maintained by B monitoring and updating the delay information and making adjustments accordingly. In this way, A is the master by virtue of being the first comer in the system, and B is locked onto A as a slave.

When station C wants to join in for its transmission, it has to go through the same procedure as B did, except it would lock onto B instead of A. If A drops out of transmission, B would detect that and take over as the master, using its own free-running clock, and C would stay locked to B. When A wants to resume its transmission later, it would have to join in as a slave to C. Therefore, the system assumes a dynamic master/slave arrangement where the first comer assumes the role of the master. Although this arrangement, as described, can only function properly if the three stations join the system sequentially, the time required by a station to establish itself as a slave can be designed to be well within a second, and thus for all practical purposes the initialization can be achieved almost instantaneously. We will show in the next section how all of these operations can be implemented with simple hardware.

III. HARDWARE IMPLEMENTATION

We describe in this section the hardware implementation of the present method. The following discussion will be divided into two major parts. The first part outlines the generation of a window signal that marks the proper transmission time for the time-compressed TV bursts at the local earth station. This window signal is denoted by $r(t)$. The second part explains how $r(t)$ can be used to synchronize the incoming TV picture such that its time-compressed bursts automatically align with the transmission windows.

The window signal, $r(t)$, is a pulse train with pulse width, T, equal to a TV field period minus the guard time and with a repetition rate = 1/3F. It is generated by the window processor depicted in Fig. 3. We assume that an external clock of eight-times B's color subcarrier frequency is made available to the window processor. This (≈28-MHz)

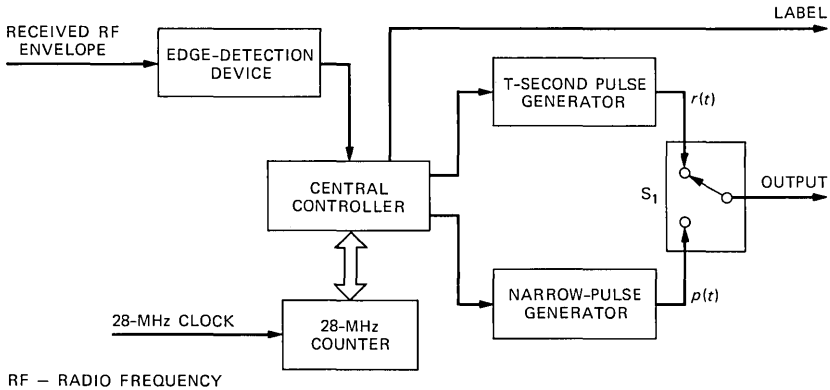


Fig. 3—Window processor for TDM/TCM synchronization.

clock is probably necessary for the time-compression operation itself, and its use here does not impose any additional burden on the system. The other input to the window processor is the received RF envelope from the satellite broadcast. In the trivial case of the first station (A) to go into the system, the window processor does very little because the transmission is free running. Let us consider the operation when the second station (B) wishes to start transmission. The received RF envelope (at B) is simply edge detected, and the arrival times of the bursts from A are recorded using the 28-MHz counter shown in Fig. 3. (Some accommodation for noise may be required, e.g., first detect envelope pulse of duration $\approx$T, then detect edges.) This information is supplied to the central controller, which could be a microprocessor and/or hardwired logic designed to carry out the windowing procedure outlined in the previous section. After acquiring the initial arrival times of the bursts from A, the central controller makes a crude estimate of the future arrival times. Furthermore, based on its location, it can compute an approximate delay to the satellite. Putting all these together, the controller produces a narrow pulse (pulse width $\ll$T) via the narrow pulse generator and sends it via the switch S_1 (in the lower position) to the transmitter. This narrow pulse will arrive at the satellite well within a predetermined time slot without collision with A's transmission. The return of this narrow pulse from the satellite completes a round-trip delay measurement that is then used to refine the arrival-time estimates. After a few cycles of this operation, the proper transmission time windows, $r(t)$, can be established by generating a sequence of pulses from the T-second pulse generator with S_1 switched to the upper position. Note that the pulse width and repetition rate of these T-second pulses are both computed using B's 28-MHz clock. A representative $r(t)$ is shown in Fig. 4a. The label output distinguishes the master from the slaves, and will be discussed later.

Before describing the rest of the hardware implementation, we show in Fig. 4 the conceptual sequence of operations needed to complete the synchronization. The transmission window is established by $r(t)$ in Fig. 4a. We use this to align (or phase lock) a composite TV sync signal, $s(t)$, such that every third, vertical sync pulse in $s(t)$ straddles the beginning of a transmission window (Fig. 4b). The sync signal, $s(t)$, is then used to synchronize an incoming video, resulting in $x_s(t)$, as shown in Fig. 4c. Finally, the video, $x_s(t)$, can be time compressed to obtain $x_c(t)$ (Fig. 4d), which is in synchronism with the transmission windows. The complete hardware to do all these is shown in Fig. 5.

Referring now to Fig. 5, the TV signal, $x(t)$, is passed through a frame synchronizer (and/or time base corrector) whose reference sync signal, $s(t)$, is derived from the TCM synchronizer. The frame synchronizer aligns $x(t)$ to $x_s(t)$ (Fig. 4c). The subsequent time compres-

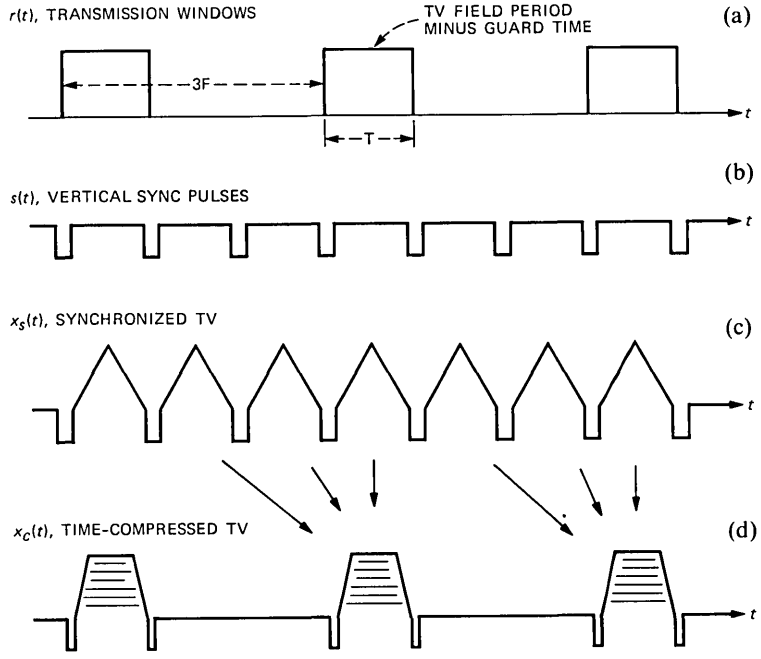


Fig. 4—Illustration of synchronization procedure. Horizontal sync, color burst, etc., are now shown.

sion on $x_s(t)$ is done in the time-compression processor previously described in Ref. 6. In this example, we assume that the time-compression processor requires three clock inputs in addition to the incoming video: a four-times color subcarrier clock (≈14 MHz), an eight-times color subcarrier clock (≈28 MHz), and the transmission window signal, $r(t)$. The time-compressed video, $x_c(t)$, is ready for immediate transmission through the FM modulator and the rest of the system. The *pin* modulator shown after the FM modulator is included to ensure the proper transmission timing as well as to enable the transmission of the narrow pulses at start-up.

As for the TCM synchronizer, its output is $s(t)$, as mentioned previously, and its input is the received RF envelope from the satellite broadcast. From the detected RF pulses, the window processor (Fig. 3) generates either $r(t)$ or the narrow pulses, depending on its state. When it is in the delay measurement mode, i.e., narrow pulses are being generated, the rest of the TCM synchronizer is free running. After $r(t)$ is generated, an internal 3.58-MHz color subcarrier is phase locked onto $r(t)$ via a TV sync generator and appropriate dividers as shown in Fig 3. This simple scheme ensures that the composite sync, $s(t)$, is synchronized with the transmission windows, $r(t)$. The label

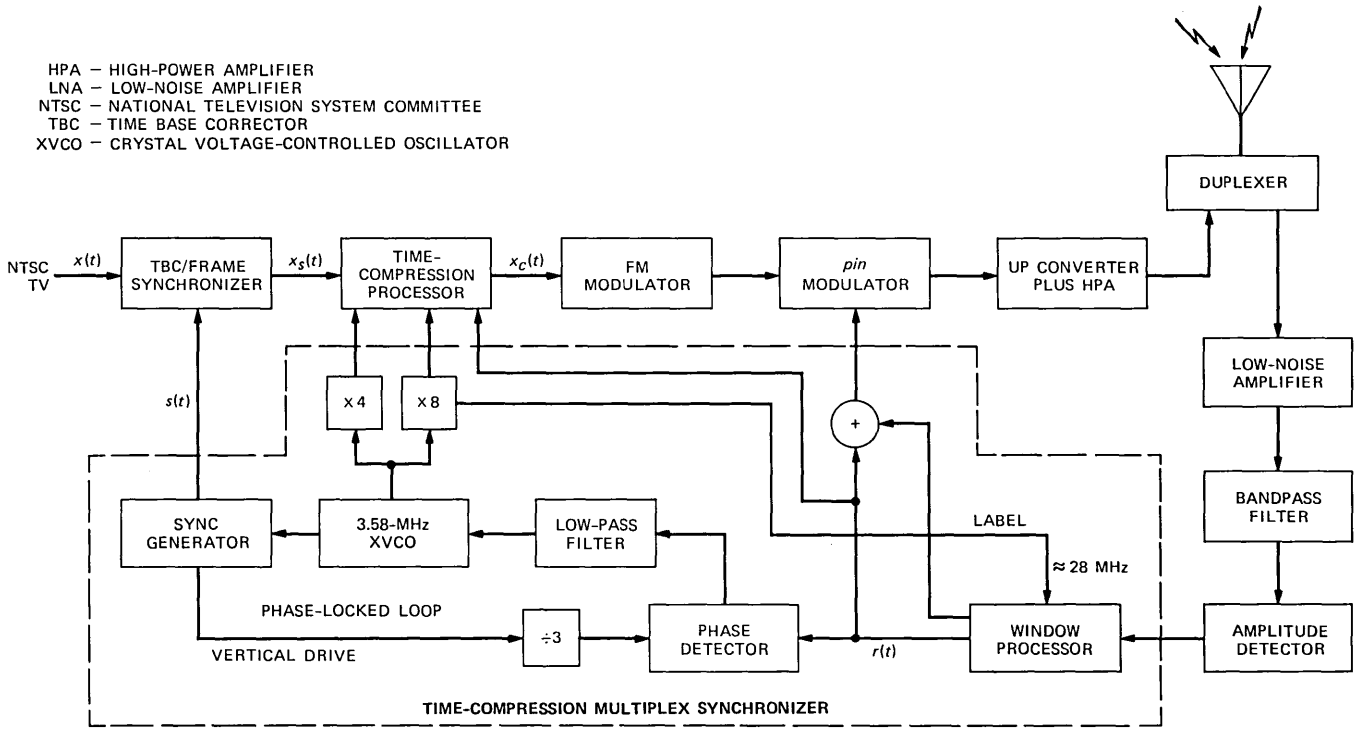


Fig. 5—Block diagram for earth station using present method for synchronization.

output of the window processor causes short RF pulses to be generated in the guard time in order to distinguish master from slaves. More discussion of labels will follow.

IV. TIMING ANALYSIS

Two important timing parameters reflecting the performance of the synchronization method are considered here: the initial acquisition time and the subsequent timing jitter in steady state. In our case of TV broadcasting, transmission is usually planned ahead of time and thus an initial acquisition time of, say, a few seconds should be adequate. However, faster acquisition is probably desirable in the case of failure recovery, as will be discussed later.

The guard time needed between bursts from different users is obviously determined by the timing jitter of the synchronization method and is a rather critical parameter. In the present system, each RF burst has the duration of a TV field time minus guard time, and the TCM synchronizer at the transmit earth station has to detect these bursts individually in order to start, as well as to maintain, lock-up. Therefore, we must ensure that some detectable gap always exists between successive bursts. Since we have the freedom to choose how to segment the original TV into three-field groups before compression, we may as well do it in a way that creates a small gap, and therefore we propose that the segmentation be done during a line of the vertical-blanking interval, which contains no information. Furthermore, we deliberately delete from transmission a portion of that line, thus generating a gap between bursts that could amount to, say, 15 μs. This deletion during vertical blanking does not affect the video quality because it is done where there is no information. The resulting benefits of this are twofold: we have created the necessary time gap between bursts from different stations; and we have a sizable guard time of 15 μs to accommodate the timing jitter (and to include labels to be described in Section V).

The major causes and their effects on the steady-state timing jitter in our system are summarized in Table I. We now discuss briefly the

Table I—Summary of timing jitter performance

Parameter	Jitter (ns)
Delay measurement uncertainty	±22
Up-link delay drift	±11
Down-link delay drift	±1
Clock resolution	±17
Field-rate jitter	±0
Total	±51

meaning of each entry in the table, while the detailed derivation is left to the appendix:

1. Delay measurement uncertainty—Delay measurement is made either via the narrow pulses or by the monitoring of the up-link's own returned TV bursts. In either case, a local 28-MHz clock is used to record the time elapsed, and the clock resolution is limited to half a cycle. It is implicit that fixed delays through the satellite and earth station hardware can be calibrated out from the raw measurement. Since this measurement is done in the communication band, propagation effects are automatically minimized.

2. Up-link delay drift—This refers to the up-link delay variation from station A to the satellite, which is not known to station B. It is time varying because of the spacecraft motion. This cannot be eliminated because we assume station B does not know A's location or have any ranging information on the propagation from A to B.

3. Down-link delay drift—This refers to the down-link delay variation between the satellite and station B. It is also time varying because of the spacecraft motion, but it is trackable via the delay measurement at B. A simple linear prediction should almost eliminate this.

4. Clock resolution—This is the limitation in the TCM synchronizer to time itself for the exact instant to start transmission due to the finite clock resolution (half a cycle in the 28-MHz clock).

5. Field-rate jitter—B and C are trying to lock to the inherent jitter in the RF bursts from A. However, if A's TV source conforms to the NTSC standard, this is so small that it can be dropped for all practical purposes. Otherwise, this item must be included in the table.

As we saw in Table I, the steady-state jitter is so small compared to the 15 μs guard time that under normal circumstances the system can be regarded as jitter free.

We now make a worst-case estimate of the initial acquisition time. It is convenient to make the simplifying assumptions that the system is jitter free and the satellite is truly stationary. The resulting error due to these assumptions is only in the order of less than 100 ns, while the acquisition time, as will be shown below, is in the order of a second. Again we will treat the case of station B trying a cold start after station A has already been on the air.

After turn-on at station B, the window processor needs to monitor a few received bursts from A before it can position its narrow pulses for delay measurements. Since the bursts from A are arriving every 50 ms, the monitoring takes $\approx$150 ms. After this 150-ms listening period, the narrow pulses are sent for delay measurements, and in order to allow for two delay measurements, we need a maximum of $\approx$500 ms. Therefore, after $\approx$650 ms have elapsed since turn-on, the synchronizer

has completed the delay measurements and can compute the near-past, current, and near-future arrival times of A's bursts at the satellite. At this point, transmission can commence in the next available time slot, which in the worst case involves a delay of three field periods (≈50 ms). Putting everything together, we have a worst-case total of ≈700 ms between initial turn-on and the first TV transmission. Such an acquisition time certainly meets our objective of keeping it below a second. In fact, a potential saving of ≈250 ms exists if we do a single narrow pulse delay measurement instead of two. Therefore, we conclude that our acquisition time is less than one-half second with a single delay measurement and less than one second with delay verification.

V. FAILURE RECOVERY

In any prudent system design the possibility of failure of certain components must be taken into account. Here, we desire that the failure of one channel does not disrupt the transmissions of the remaining channels. In order to facilitate this, we provide for a labeling mechanism, in which the window processor causes short RF pulses to be transmitted immediately following the video RF burst, i.e., at the beginning of the guard time. These pulses are then used to distinguish the master from the slaves, as well as to detect anomalies.

For example, station A (being the master) could transmit three pulses. Station B, the next in command, would send two pulses, and station C, one pulse. Additional pulses could identify the up-link station or, alternatively, this information could be embedded in the baseband video.

The window processor keeps track of time and labeling of all received RF bursts, and is ready to accommodate to any change in operating conditions. For example, if A finishes its transmission and goes off the air, B becomes the new master transmitting three pulses, and C becomes second in command transmitting two pulses.

It is never possible to predict all failure modes. The best we can do is accommodate the most likely ones. For example, a brief up-link failure will not be detected at any earth station (including, possibly, the faulty one) for about 240 ms, and during that time it is possible for transmission to resume. Moreover, corrective action by the faulty earth station will not be known to the remaining ones for another 240 ms. Thus, in the case of an up-link failure at master station A, station B should not try to take over as master immediately. Otherwise, there would be the possibility of two masters existing at the same time. In any event, as soon as station A determines that its up-link is unreliable, it should resign as master. This could be done by not transmitting any pulses following its video RF burst. The other stations would recognize

this condition and assume their proper responsibilities, after which station A would begin transmitting a single pulse designating itself as last station aboard.

In the case of a down-link failure, continued operation is not possible unless the faulty station is master. If it were not already the master, it could take over this role by sending, say, four pulses following the video RF burst. The other stations would then recognize this condition and assume their proper responsibilities.

In the case of an earth station power glitch, transmission would have to cease immediately and the start-up procedure would be reinvoked, since the window processor would, in all probability, lose its timing information. Such a restart could be speeded up considerably if nonvolatile memory were provided, however.

VI. COMPARISONS AND DISCUSSIONS

As mentioned previously, a number of synchronization methods are applicable to solve the present problem. The most obvious one is probably that of a centrally controlled station broadcasting a master sync to all three up-link stations. Within this broad class of techniques, a large variety of alternatives are possible. As an example, one fixed station may be assigned as the master and the other stations must lock their transmissions to the master; a master sync marker may be broadcast to all stations by a centrally controlled station, and the marker could contain sufficient information to TV field and color subcarrier synchronizations, as well as ranging data for extremely fast open loop acquisition. In fact, only one such master is needed for the whole satellite system. Its main advantages are that fast acquisition is possible, and the various up-link stations do not have to monitor one another's transmissions, although the hardware implementation at each up-link station is certainly not simpler than our method. The key concern, though, is the reliability of the master station—its maintenance and hardware complexity. A single up-link failure at the master station would immobilize the whole system. In contrast, our method would tolerate quite a combination of different failures because an automatic takeover procedure exists for the master assignment. Any single up-link or down-link failure at a station can interrupt service only at that station and has no bearing on the rest of the system.

It is possible to use a separate channel to perform interstation ranging as proposed in Ref. 9. The bandwidth requirement for this ranging channel is critically determined by the rise time of the ranging pulses, which, in turn, affects the resulting synchronization accuracy. Therefore, the addition of this ranging channel could be an imposing requirement in the system.

Improvements in the jitter performance and the acquisition time in our system are both possible. The up-link drift could be removed if the up-link delay information from each station were inserted into one of the vertical-blanking pulses, and the stations could then demodulate for these data. Higher clock frequencies could be used in the delay measurement, thereby decreasing its uncertainty. This would also increase the time resolution of system and thus enable the synchronizer to time the transmissions more accurately. As for the acquisition time, if an accurate site location plus its up-link delay were provided by the first (or the master) station in one of its vertical-blanking pulses, then the other stations could compute their respective delays to the spacecraft without performing the narrow pulse measurements, resulting in a significant reduction in the acquisition time.

VII. CONCLUSIONS

We have described a method of synchronizing up-link earth stations in a TCM system where the stations take turns transmitting TV information in bursts, each lasting for a field duration. The technique is simple and requires only that the stations receive their own as well as others' transmissions. It has a dynamic master/slave arrangement whereby the first station signing on assumes the role of a master. The other stations subsequently can synchronize their transmissions to the master's by simply monitoring the received RF bursts from the satellite, measuring their respective delays to the spacecraft, and then phase locking their local color subcarrier clocks to the master's transmitted bursts. When the master station stops transmitting, an automatic procedure exists for the second station to take over as the new master. As a result, any single up-link or down-link failure can only affect the station involved, and there is no need to have centralized control. Most of the hardware in the synchronizer can be implemented digitally. The worst-case jitter performance in the system is well below 100 ns, while the initial acquisition time can be kept to less than one-half second. These are more than adequate for the TV application, and we conclude that the proposed method offers a practical means to synchronize the three up-links in our TCM system.

REFERENCES

1. J. E. Flood and D. I. Urquhart-Pullen, "Time-Compression-Multiplex Transmission," Proc. IEE, *111*, No. 4 (April 1964), pp. 647–68.
2. D. H. Morgen and E. N. Protonotarios, "Time Compression Multiplexing for Loop Transmission of Speech Signals," IEEE Trans. Comm., *COM-22*, No. 12 (December 1972), pp. 1932–9.
3. K. Y. Eng and O.-C. Yue, "Spectral Properties and Band-Limiting Effects of Time-Compressed TV Signals in a Time-Compression Multiplexing System," B.S.T.J., *60*, No. 9 (November 1981), pp. 2167–85.
4. K. Y. Eng and O.-C. Yue, "Time Compression Multiplexing of Multiple Television

Signals in Satellite Channels Using Chirp Transform Processors," IEEE Trans. Commun., *COM-29*, No. 12 (December 1981), pp. 1932–40.
5. K. Y. Eng and B. G. Haskell, "TV Bandwidth Compression Techniques Using Time Companded Differentials and Their Applications to Satellite Transmission," B.S.T.J., *61*, No. 10, Part 1 (December 1982), pp. 2917–27.
6. K. Y. Eng, B. G. Haskell, and R. L. Schmidt, "Time-Compression Multiplexing (TCM) of Three Broadcast-Quality TV Signals on a Satellite Transponder," B.S.T.J., this issue.
7. P. P. Nuspl et al., "Synchronization Methods for TDMA," Proc. IEEE, *65*, No. 3 (March 1977), pp. 434–44.
8. J. J. Spilker, Jr., *Digital Communications by Satellite*, Englewood Cliffs, NJ: Prentice-Hall, 1977.
9. A. Acampora, "Synchronizing Transmissions From Two Earth Stations to Satellite," U.S. Patent 4,320,503, applied for August, 1979, issued March 1982.
10. D. D. Slavinskas, private communication.

APPENDIX

Parameters for Timing Jitter

We show briefly here the derivations for the various contributions to the steady-state timing jitter (Table I). The following estimates are, by and large, worst-case and very conservative.

A.1 Delay measurement uncertainty

The slave stations have to measure their respective delays to the satellite in order to start, as well as to keep, synchronized with the master. This is done in the beginning via the narrow pulses, and then it is updated continuously via the monitoring of its own returned bursts. The delay is, of course, measured from edge to edge in the transmitted and received RF bursts. Given a 36-MHz RF channel bandwidth, the fastest RF pulse rise time is in the order of 30 ns. If we have to measure delay from one edge to another, an accuracy of ±5 ns seems reasonable. In addition, the clock used for the measurement is resolution limited due to its finite frequency (≈28 MHz, or eight-times color subcarrier frequency). The uncertainty due to this clock is about ±17 ns, yielding a total uncertainty of ±22 ns.

A.2 Up-link delay drift

In the absence of any knowledge of the master's (or A's) location, a slave station (or B) cannot predict the up-link delay from the master to the satellite. Furthermore, this up-link delay is time varying due to the motion of the spacecraft. The net result is that B's prediction of the near-future burst arrivals from A can never be exact, even though the down-link delay between B and the satellite can be predicted exactly. To illustrate this point, let us consider a burst transmitted from A to B at $t = 0$. The up-link delay (from A to the satellite) is u_0; the down-link delay (from the satellite to B) is d_0; and the delay through the satellite is conveniently chosen to be zero. The arrival time of this burst at B is simply

$$T_0 = u_0 + d_0.$$

Now, at a later instant $t = t_1$, A transmits another burst to B. The corresponding up-link and down-link delays are u_1 and d_1, respectively. Again $u_0 \neq u_1$ and $d_0 \neq d_1$ because of the spacecraft motion. The arrival time at B is then

$$T_1 = u_1 + d_1 + t_1.$$

In order to predict T_1 at B at the time T_0, B has to compute

$$T_1 - T_0 = (u_1 - u_0) + (d_1 - d_0) + t_1,$$

where t_1 is known to B because A is transmitting at a fixed rate; $(d_1 - d_0)$ can be extrapolated based on B's delay measurements; but the quantity $(u_1 - u_0)$ cannot be estimated without knowing A's location. In this example, $(u_1 - u_0)$ is simply the up-link delay variation for A due to the spacecraft displacement in the time interval t_1. As such, an easy upper can be written as

$$|u_1 - u_0| \leq cvt_1,$$

where c is the velocity of light; v is the radial velocity of the spacecraft toward or away from an earth station; and t_1, the time interval, is understood to be small compared to a day. If we replace v by the highest radial velocity of the spacecraft, and t_1 by the round-trip satellite propagation delay ($\approx$300 ms), we have a worst-case estimate on the up-link delay drift. According to an example given in page 149 of Ref. 8 and comparisons to data from more recent communications satellites,[10] a convenient upper bound on the spacecraft radial velocity in geostationary orbit is 10 m/s. Using this, the worst-case up-link delay drift is $\pm$10 ns.

A.3 Down-link delay drift

With the spacecraft radial velocity limited to 10 m/s and continuous updates on the delay measurement, we feel that the down-link delay drift can easily be computed with an accuracy an order of magnitude lower than the up-link delay drift, or about $\pm$1 ns.

A.4 Clock resolution

Using a 28-MHz clock, the resolution is about half a cycle or $\pm$17 ns.

A.5 Field-rate jitter

This refers to the jitter in the rate at which the master station is transmitting its RF bursts. The burst transmission is of course governed by the TV field rate, and only one burst is sent in every three fields. The NTSC standard specifies that the color subcarrier fre-

quency (3,579,545 Hz) must be stable within ±10 Hz and cannot vary more than 0.1 Hz/s. For a worst-case situation, we assume that the color subcarrier is at the lowest value, i.e., 3,579,535 Hz. It then drifts at the maximum rate of 0.1 Hz/s. Thus, at the end of a second, the new frequency is 3579535.1 Hz. The difference in TV field period derived from these two frequencies is only 7.8×10^{-15} s. The net result is that the TV field rate is jitter free over a short period of time, say a few seconds. Moreover, this implies that a much less stable color subcarrier frequency is still quite compatible with our synchronization system.

AUTHORS

Kai Y. Eng, B.S.E.E. (summa cum laude), 1974, Newark College of Engineering; M.S. (Electrical Engineering), 1976, Dr. Engr. Sc. (Electrical Engineering), 1979, Columbia University; RCA Astro-Electronics, 1974–1979; Bell Laboratories, 1979—. Mr. Eng has worked on various areas of microwave transmission, spacecraft antenna analysis, and communications satellites. He is presently a member of the Radio Research Laboratory, studying TV transmission through satellites. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, Phi Eta Sigma.

Barry G. Haskell, B.S. (Electrical Engineering), 1964, M.S., 1965, and Ph.D., 1968, University of California, Berkeley; University of California, 1965–1968; Bell Laboratories, 1968—; Rutgers University, 1977–1979. Mr. Haskell was a Research Assistant at the University of California Electronics Research Laboratory and a part-time faculty member of the Department of Electrical Engineering at Rutgers University. At Bell Laboratories, he is presently Head of the Radio Communications Research Department, where his research interests include television picture coding and transmission of digital and analog information via microwave radio. Member, IEEE, Phi Beta Kappa, Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

Theory of Reflection From Antireflection Coatings

By R. H. CLARKE*

(Manuscript received November 24, 1982)

The reflection that occurs when a beam, rather than a plane wave, is incident normally on a quarter-wavelength matching layer can be of vital importance in semiconductor laser design. An analysis in three dimensions is given for the general case of a field of arbitrary form and polarization incident on the matching layer. The field is represented as an angular spectrum of plane waves, each component plane wave being modified by the appropriate Fresnel reflection coefficient to give the field reflected back onto the diode structure. Brown's antenna reciprocity theorem is used to determine the amplitude of the corresponding mode traveling back down the diode.

I. INTRODUCTION

Antireflection coatings are used on one face of superluminescent diodes[1] and on both faces of diode-laser amplifiers.[2] The theoretical performance of such coatings has been analyzed by Clarke[3] using the technique of representing the emerging laser beam as an angular spectrum of plane waves, as originally applied by Reinhart et al.[4] and Gordon[5] to determine the reflectivity of an uncoated facet. Each plane wave was modified by the appropriate reflection coefficient of the uniform coating,[6] and Brown's antenna reciprocity theorem[7] used to calculate the amplitude of the wave coupled back into the device. The previous analysis[3] was restricted to two dimensions, on the grounds

* Work done while at Bell Laboratories. Now at Imperial College of Science and Technology, London, England.

that the active region in the device would be a wide flat stripe, so that the emerging radiation would be a thin fan-shaped beam. Many important laser diode structures, particularly of the refractive index guided type, have relatively narrow active regions, hence the previous restriction is limiting. This restriction is removed in the present work, and the full three-dimensional analysis is presented.

II. FIELDS IN THE DIODE

The transverse electric field of a single mode traveling in the positive-z direction (see Fig. 1) along the length of the active-region stripe in a diode laser can be written in general as

$$\mathbf{E}_t^+(x, y, z) = [\mathbf{u}_x E_{tx}(x, y) + \mathbf{u}_y E_{ty}(x, y)]e^{-j\beta_m z}, \tag{1}$$

where β_m is the phase constant of the mode and the time variation $\exp(j\omega t)$ has been suppressed. The field in this mode reflected back into the diode by the coating is

$$\mathbf{E}_t^-(x, y, z) = \rho[\mathbf{u}_x E_{tx}(x, y) + \mathbf{u}_y E_{ty}(x, y)]e^{+j\beta_m z}. \tag{2}$$

The objective of this paper is to calculate the reflection coefficient ρ for arbitrary thickness h and refractive index n_2 of the coating. (Coupling to other modes is ignored here for the sake of simplicity.) It will be assumed that the beam eventually emerges into air, so that $n_3 = 1$, and that the refractive index of the diode has the effective value n_1, which is that of the active region in which the field is largely confined. (The surrounding bulk material has a refractive index that is some 10 percent below n_1.[1] A better choice of effective refractive index might therefore be a weighted average, as suggested by Kaplan's analysis.[8])

The field incident at the plane $z = 0$ can be represented as an

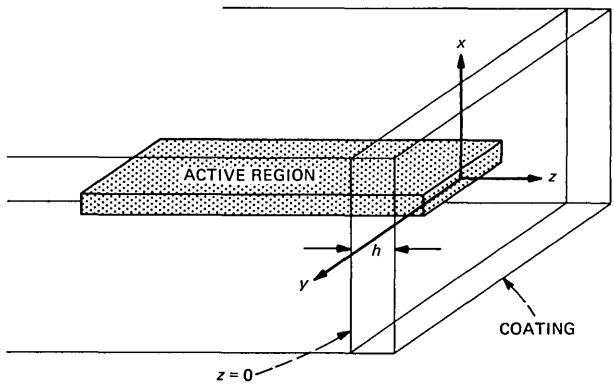


Fig. 1—Diode laser with coating.

angular spectrum of plane waves by the two spectrum functions $F_x(\alpha, \beta)$ and $F_y(\alpha, \beta)$, where (α, β, γ) are the direction cosines in the x-, y-, and z-directions.[9,10] Thus, the elemental plane wave incident in the direction (α, β) is $\mathbf{e}_{\text{inc}}(\alpha, \beta)d\alpha d\beta$, where

$$\mathbf{e}_{\text{inc}}(\alpha, \beta) = F_x(\alpha, \beta)\left(\mathbf{u}_x - \mathbf{u}_z\frac{\alpha}{\gamma}\right) + F_y(\alpha, \beta)\left(\mathbf{u}_y - \mathbf{u}_z\frac{\beta}{\gamma}\right) \tag{3}$$

with

$$F_x(\alpha, \beta) \leftrightarrow E_{tx}(x, y)$$

and

$$F_y(\alpha, \beta) \leftrightarrow E_{ty}(x, y), \tag{4}$$

in which $\leftrightarrow$ symbolizes a Fourier transform, such as

$$F_x(\alpha, \beta) = \frac{1}{\lambda_1^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_{tx}(x, y)\exp\{jk_1(\alpha x + \beta y\}dxdy. \tag{5}$$

The phase constant in the diode is $k_1 = 2\pi/\lambda_1 = n_1 k_0$, where k_0 is the phase constant of free space.

It should be noted for later reference that the above angular spectrum corresponds to a radiation far field (assuming that the subscript 1 region continues indefinitely but the active region stops in the plane $z = 0$), as $k_1 r \to \infty$, of[10]

$$\mathbf{E}(r, \theta, \phi) \simeq \frac{\exp(-jk_1 r)}{k_1 r}\mathbf{e}_{\text{rad}}(\alpha, \beta), \tag{6}$$

where the far-field vector pattern function is given in this case by

$$\mathbf{e}_{\text{rad}}(\alpha, \beta) = j2\pi\gamma\mathbf{e}_{\text{inc}}(\alpha, \beta) \tag{7}$$

and

$$\begin{aligned}\alpha &= \sin\theta\cos\phi\\ \beta &= \sin\theta\sin\phi\\ \gamma &= \cos\theta,\end{aligned} \tag{8}$$

where θ is the polar angle to the z-axis, ϕ is the azimuth angle in the x-y plane, and r is the distance to the point of observation.

III. REFLECTION AT THE COATING

The incident plane wave given by eq. (3) will be reflected by the coating. The amplitude reflection coefficient for a plane wave incident on such a uniform layer with its electric vector perpendicular to its

plane of incidence (see Fig. 2) is[6]

$$R_\perp = \frac{P_1 \cos B + jP_3 \sin B}{P_2 \cos B + jP_4 \sin B} \tag{9}$$

and, with its electric vector parallel to its plane of incidence, the reflection coefficient is

$$R_\parallel = \frac{Q_1 \cos B + jQ_3 \sin B}{Q_2 \cos B + jQ_4 \sin B}, \tag{10}$$

where

$$P_{1,2} = n_2(1 - n_1^2 s^2/n_2^2)^{1/2}[n_1\gamma \mp (1 - n_1^2 s^2)^{1/2}]$$

$$P_{3,4} = n_1\gamma(1 - n_1^2 s^2)^{1/2} \mp n_2^2(1 - n_1^2 s^2/n_2^2) \tag{11}$$

$$Q_{1,2} = n_2(1 - n_1^2 s^2/n_2^2)^{1/2}[n_1(1 - n_1^2 s^2)^{1/2} \mp \gamma]$$

$$Q_{3,4} = n_1(1 - n_1^2 s^2/n_2^2) \mp n_2^2\gamma(1 - n_1^2 s^2)^{1/2} \tag{12}$$

with

$$s^2 = \alpha^2 + \beta^2 = 1 - \gamma^2 \tag{13}$$

and

$$B = (2\pi h/\lambda_2)(1 - n_1^2 s^2/n_2^2)^{1/2}, \tag{14}$$

where $\lambda_2 = \lambda_0/n_2$.

Note that when the magnitude of the sine of the angle of incidence $|s| > (n_1)^{-1}$, the wave will be totally internally reflected. In that case, the magnitude of the reflection coefficient will always be unity, but its phase will vary with the angle of incidence. But note also that the $\exp(j\omega t)$ sign convention adopted here means taking the negative

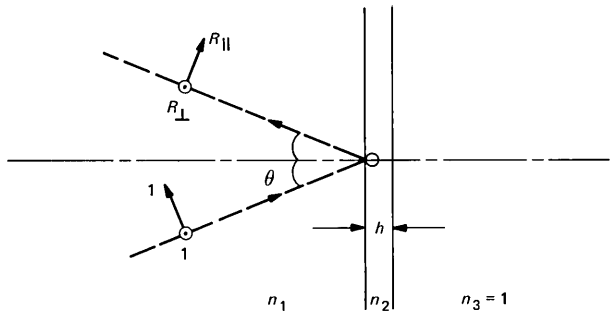


Fig. 2—Definition of the amplitude reflection coefficients $R_\perp$ and $R_\parallel$. Their phases are defined at O.

square root when the round-bracketed quantities in eqs. (11), (12), and (14) become negative.

The elemental plane wave given by the angular spectrum of eq. (3) consists, in general, of the sum of perpendicular and parallel polarized components, such that

$$\mathbf{e}_{\text{inc}} = \mathbf{e}_{\perp} + e_{\parallel}. \tag{15}$$

This resolution can be achieved by noting that the unit vector $\mathbf{u}_n$, which is both normal to the plane of incidence of the plane wave travelling in the direction (α, β) and also parallel to the bounding plane surface xOy, is

$$\mathbf{u}_n = \frac{1}{\sqrt{\alpha^2 + \beta^2}} [\mathbf{u}_x\beta - \mathbf{u}_y\alpha]. \tag{16}$$

Hence we may calculate

$$\mathbf{e}_{\perp} = \mathbf{u}_n[\mathbf{u}_n \cdot \mathbf{e}_{\text{inc}}) \tag{17}$$

and

$$\mathbf{e}_{\parallel} = \mathbf{e}_{\text{inc}} - \mathbf{u}_n(\mathbf{u}_n \cdot \mathbf{e}_{\text{inc}}). \tag{18}$$

The elemental plane wave $\mathbf{e}_{\text{refl}}(\alpha, \beta)d\alpha d\beta$ reflected by the coating is thus given by

$$\mathbf{e}_{\text{refl}}(\alpha, \beta) = R_{\perp}\mathbf{e}_{\perp}(\alpha, \beta) + R_{\parallel}\mathbf{u}_{\parallel}(\alpha, \beta) \tag{19}$$

and comes *from* the direction $(-\alpha, -\beta)$. (To avoid possible confusion it should be noted that the argument of $\mathbf{e}_{\text{refl}}(\alpha, \beta)$ denotes the direction of the incident wave.)

IV. COUPLING BACK INTO THE DIODE

Brown's antenna reciprocity theorem states that if a plane wave of vector amplitude $\mathbf{e}_p$ is incident from the direction $\mathbf{u}_p$ on a linear, reciprocal device, which when radiating has the far-field vector pattern function (see eq. 6) of $\mathbf{e}_{\text{rad}}(\mathbf{u})$, then the coupling ratio

$$c = \frac{\lambda^2}{j4\pi ZP_0} \mathbf{e}_p \cdot \mathbf{e}_{\text{rad}}(\mathbf{u}_p) \tag{20}$$

gives the complex ratio of the single-mode amplitude when receiving to that when transmitting a total power P_0.[7,11] Z and λ are the characteristic impedance and wavelength in the radiating medium. Equation (20) is a precise result, being a consequence ultimately of the Lorentz reciprocity theorem.

In the present instance the incident plane wave $\mathbf{e}_p$ is the elemental reflected plane wave $\mathbf{e}_{\text{ref}}d\alpha d\beta$ given by eq. (19), and so, integrating

over all directions in the forward hemisphere, the reflection coefficient describing the returning mode amplitude is

$$\rho = \frac{\lambda^2}{j4\pi ZP_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{e}_{\text{refl}}(\alpha, \beta) \cdot \mathbf{e}_{\text{rad}}(-\alpha, -\beta) d\alpha d\beta \tag{21}$$

or

$$\rho = \frac{\lambda^2}{2ZP_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{R_{\parallel} \mathbf{e}_{\text{inc}}(\alpha, \beta) + \mathbf{u}_n (R_{\perp} - R_{\parallel}) [\mathbf{u}_n \cdot \mathbf{e}_{\text{inc}}(\alpha, \beta)]\} \cdot \mathbf{e}_{\text{inc}}(-\alpha, -\beta) \gamma d\alpha d\beta \tag{22}$$

with $\mathbf{e}_{\text{inc}}(\alpha, \beta)$ given by eq. (3) and $\mathbf{u}_n$ by eq. (16). The total radiated power P_0, when the radiation is specified by the two spectrum functions $F_x(\alpha, \beta)$ and $F_y(\alpha, \beta)$, is given by[12]

$$P_0 = \frac{\lambda^2}{2Z} \int_D \int \left[\frac{1 - \beta^2}{\gamma} |F_x(\alpha, \beta)|^2 + \frac{1 - \alpha^2}{\gamma} |F_y(\alpha, \beta)|^2 \right] d\alpha d\beta, \tag{23}$$

where D is the domain of (α, β) such that $\alpha^2 + \beta^2 = \leq 1$.

V. APPLICATION TO A *Y*-POLARIZED LASER MODE

In order to see what this result means, consider a guided mode in the laser whose tangential electric field is wholly y-directed, for which therefore $F_x \equiv 0$. Then

$$\mathbf{e}_{\text{inc}}(\alpha, \beta) = F_y(\alpha, \beta) \left(\mathbf{u}_y - \mathbf{u}_z \frac{\beta}{\gamma} \right) \tag{24}$$

and

$$\mathbf{u}_n = \frac{1}{\sqrt{\alpha^2 + \beta^2}} (\mathbf{u}_x \beta - \mathbf{u}_y \alpha). \tag{25}$$

Consequently, the reflection coefficient in this case is

$$\rho = \frac{\lambda^2}{2ZP_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[R_{\parallel} \left(1 + \frac{\beta^2}{\gamma^2} - \frac{\alpha^2}{\alpha + \beta^2} \right) + R_{\perp} \frac{\alpha^2}{\alpha^2 + \beta^2} \right] \gamma F_y(\alpha, \beta) F_y(-\alpha, -\beta) d\alpha d\beta. \tag{26}$$

Then finally, assuming that the beam spread is vanishingly narrow in the y–z plane compared to the x–y plane,

$$\rho = \frac{\lambda^2 K}{2ZP_0} \int_{-\infty}^{\infty} \gamma R_{\perp} F_y(\alpha, 0) F_y(-\alpha, 0) d\alpha, \tag{27}$$

where K depends on the β-dependence of F_y. Equation (27) is the two-dimensional result used previously.[3]

VI. CONCLUSIONS

A complete three-dimensional analysis has been presented for the calculation of reflection from antireflection coatings. It reduces to the two-dimensional result given previously where the incident beam was assumed to be narrowly confined in one of the principal planes. The form and polarization of the field incident on the coating can be arbitrarily specified.

REFERENCES

1. I. P. Kaminow et al., "Lateral Confinement InGaAsP Superluminescent Diode at 1.3 μm," IEEE J. Quantum Electron., *QE-19*, No. 1 (January 1983), pp. 78–81.
2. D. Marcuse, "Computer Model of an Injection Laser Amplifier," IEEE J. Quantum Electron., *QE-19*, No. 1 (January 1983), pp. 63–73.
3. R. H. Clarke, "Theoretical Performance of an Anti-Reflection Coating for a Diode Laser Amplifier," Int. J. Electron., *53*, No. 5 (November 1982), pp. 495–9.
4. F. K. Reinhart, I. Hayashi, and M. B. Panish, "Mode Reflectivity and Waveguide Properties of Double-Heterostructure Injection Lasers," J. Appl. Phys., *42*, No. 11 (October 1971), pp. 4466–79.
5. E. I. Gordon, "Mode Selection in GaAs Injection Lasers Resulting from Fresnel Reflection," IEEE J. Quantum Electron., *QE-9* (July 1973), pp. 772–6.
6. M. Born and E. Wolf, *Principles of Optics*, Fifth Edition, Elmsford, N.Y.: Pergamon, 1975, p. 61.
7. J. Brown, "A Generalized Form of the Aerial Reciprocity Theorem," Proc. IEE, *105*, Part C (1958), pp. 472–5.
8. D. R. Kaplan, private communication.
9. P. C. Clemmow, *The Plane Wave Spectrum Representation of Electromagnetic Fields*, Elmsford, N.Y.: Pergamon, 1966.
10. R. H. Clarke and J. Brown, *Diffraction Theory and Antennas*, N.Y.: Halsted, 1980, p. 85.
11. Reference 10, p. 109.
12. Reference 10, pp. 83–4.

AUTHOR

Richard H. Clarke, B.Sc., 1956, Ph.D., 1960 (Electrical Engineering), University College, London; Assistant Professor, University of California, Berkeley, 1962–1964; Bell Laboratories, 1964–1968, and visiting member of technical staff, summers, 1981 and 1982. Mr. Clarke worked for the NATO ASW Research Centre, La Spezia, Italy, from 1969–1974 in their theoretical studies group. Since 1974 he has been a lecturer in electrical engineering at Imperial College, London. While at Bell Laboratories he worked on the theory of mobile radio and of propagation of random optical fields, and on the design of antireflection coatings for semiconductor laser diodes.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

Equivalent Queueing Networks and Their Use in Approximate Equilibrium Analysis

By A. KUMAR*

(Manuscript received March 14, 1983)

Most Markovian queueing networks that arise as models of stochastic congestion systems (e.g., communication networks and multiprogrammed computer systems) do not have a product form in their stationary probability distributions, and hence are not amenable to the simplicity of product-form analysis. In this paper we suggest an approach for systematically examining the validity of a class of approximation schemes that is based on the idea of equivalent networks and is used for the approximate equilibrium analysis of nonproduct-form networks. We study equivalent networks, and prove a generalization of the so-called "Norton's" Theorem for closed product-form networks in order to study and generalize the equivalent flow method for the approximate analysis of nonproduct-form queueing networks. We then present the results of a study of the approximation scheme as applied to a type of network model (called a central-server model) that arises frequently in modeling multiprogrammed computer systems. In this model the central server uses a priority discipline, so the resulting network is nonproduct form. This study demonstrates the situations under which the approximation can be expected to do well or poorly and the kinds of errors it introduces.

I. INTRODUCTION

Mathematical modeling of stochastic systems frequently gives rise to models in a class referred to as Markovian queueing networks—specifically, queueing networks whose time evolution can be described by a discrete-state, regular Markov stochastic process. Markovian

*Bell Laboratories.

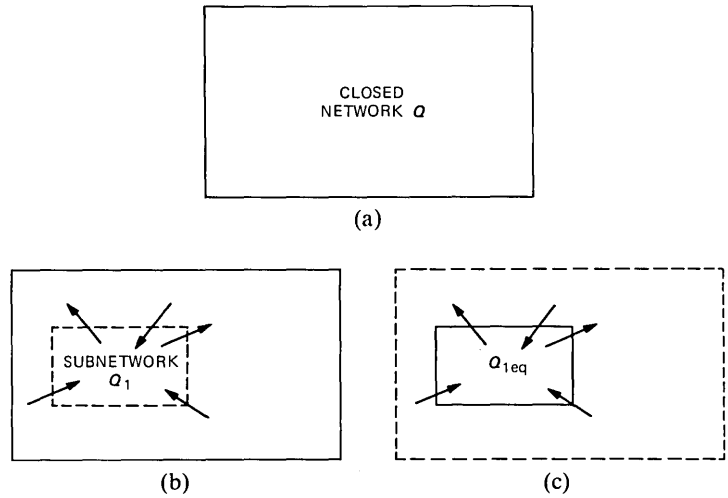


Fig. 1—Notion of an equivalent network; (a) original network, (b) with arrows indicating flows between Q_1 and its complements, and (c) with arrows indicating models of flows between Q_1 and its complement.

queueing network models, known as product-form networks, have been widely studied, owing primarily to their well-understood stochastic behavior, and the simplicity of their analysis in equilibrium. However, the class of product-form queueing network models is far from adequate for modeling many simple real-world congestion systems. The exact equilibrium analysis of nonproduct-form queueing networks is, in most cases, computationally, and often fundamentally, intractable. Much effort has, therefore, been directed towards devising approximation schemes that attempt to reconcile the conflicting requirements of modeling fidelity and the simplicity of product-form analysis. One such class of approximation schemes is based on the idea of equivalent networks. In this paper we systematically study this approximation.

By an equivalent network we mean the following (cf. Fig. 1). Consider a closed queueing network Q constructed from the set of nodes M and the subnetwork Q_1 consisting of nodes $M_1 \subset M$. Let Q_{1eq} be a network constructed from M_1 such that the joint equilibrium (probability) distribution of Q_{1eq} is the same as the marginal joint equilibrium distribution of Q_1 in Q. The network Q_{1eq} is then said to be equivalent to Q_1.* Clearly, to study Q_1 in isolation, one needs to account for the

* This notion of equivalence may appear unduly restrictive. Why not establish a more detailed stochastic equivalence? For the calculation of many performance analysis criteria, the present notion is adequate. However, it is easy to see that the equivalent networks we later identify yield equality in distribution for the entire process in equilibrium.

influence of the nodes in $M - M_1$ on the nodes in Q_1. When Q is product form, the influence of the complementary network on Q_1 takes an especially simple form, and can be determined by analyzing a modified version of the complementary network in isolation! For the case where M_1 consists of a single node, this fact was first recognized by Chandy et al.,[1] who called the equivalent network so obtained a "Norton" equivalent, because of the similarity of this equivalence to Norton equivalence in electrical circuits.

In Section II we study equivalent networks and demonstrate the simplifications that arise for product-form networks. The development yields a generalization of Norton's Theorem to multinode subnetworks of closed product-form networks. Essentially, the same extension to the entire class of closed product-form networks has been obtained independently and concurrently by Kritzinger et al.[2] and Balsamo et al.,[3] through an approach based on verification via detailed computations from the product-form solution. Our approach is substantially different, in that it derives Norton's Theorem directly as a special case of a general result for stochastically equivalent networks. This approach is concise, conceptually and intuitively appealing, gives the result a probabilistic interpretation, and shows up clearly the role played by the product-form solution. It also seems to be the natural approach for the purposes of this study.

This generalization of Norton's Theorem motivates the following approximation scheme. Suppose now that Q is a nonproduct-form network, but for the purposes of studying the subnetwork Q_1, we follow the equivalence procedure for product-form networks. Suppose also that in doing so we find that the version of the complementary network we have to analyze, in order to determine the latter's influence on Q_1, is product form. Let the equivalent network thus obtained be $\hat{Q}_{1eq}$. The approximation scheme, referred to above, approximates the equilibrium distribution of Q_1 with that of $\hat{Q}_{1eq}$ (i.e., approximates Q_{1eq} by $\hat{Q}_{1eq}$). The effort to determine and analyze $\hat{Q}_{1eq}$ will, in general, be considerably less than the effort to exactly analyze Q_1 in Q.

This approximation scheme is an extension of one (often referred to in the literature as an equivalent flow approximation) that has been utilized by several workers, in the field of network performance analysis, with remarkably accurate results. Sauer and Chandy,[4] and Chow and Yu[5] use this idea as the basic step in iterative schemes for approximating central-server models in which the central server is not of product-form type. Schwartz[6] uses the basic scheme directly to approximately analyze a model for a multiple-access communication system. In Section III we draw upon the theoretical development in Section II to study the validity of the approximation scheme when it is applied to a simple test-bed model.

II. EQUIVALENT NETWORKS

Consider a closed Markovian queueing network Q consisting of M congestion nodes. In this section we study the problem of the equilibrium analysis of a subnetwork Q_1 (embedded in Q). To simplify the discussion we shall limit our considerations to networks of First In, First Out (FIFO) nodes. It is easily recognized that the ideas in this section can be extended to apply to more general networks. In Section 2.3 we establish Theorem 1, which explicates the structure of equivalent subnetworks of the networks described in Section 2.1. By combining this result with Theorem 2, we get a generalization of Norton's Theorem.

2.1 Network specifications

Q is a closed queueing network consisting of M FIFO nodes (indexed by $i \in \{1, \cdots, M\}$). There are R classes/types of customers (indexed by $r \in \mathbf{R} = \{1, \cdots, \mathrm{R}\}$) with N_r customers in the rth class. The M x M matrix $P^{(r)} = [p_{ij}^{(r)}]$ is the routing probability matrix of type r customers; customers do not change class as they move from node to node. For each r in $\{1, \cdots, R\}$, $P^{(r)}$ is a stochastic matrix which, when considered as a transition probability matrix, leads to a Markov chain, on the state space $\{1, \cdots, M\}$, with a single positive, communicating class.

Throughout the following discussion, the network state process is assumed to be in equilibrium. The state of the ith node (denoted by S^i) is a finite string drawn from the set $\mathbf{R}$. Given a state vector S^i, $r \in \mathbf{R}$ appearing in the kth position in the string S^i denotes that a customer of type r is in the kth position, in FIFO order, at the node i. Thus, by definition, the customer in service is in the first position. $S = (S^1, \cdots, S^M)$ denotes the state of the entire network Q. The ith node is equipped with an exponential server which, when the state of the network is S, serves a customer of class r at the rate $\nu_{ir}(S)$.

Q_1 is a subnetwork of Q, consisting of M_1 ($<N$) nodes (indexed by $i \in \{1, \cdots, M_1\}$). Q_2 is the complementary network consisting of $M_2 = M - M_1$ nodes (indexed by $i \in \{M_1 + 1, \cdots, M\}$).

Some additional notation is inevitable; this we proceed to describe in the next subsection.

2.2 Notation

$\mathbf{N} = (N_1, \cdots, N_R)$ is the population vector of the network Q, where N_r is the number of customers of class r, $r \in \mathbf{R}$.

Let $\mathbf{R}^* = (\cup_{n \geq 1}\{1, \cdots, R\}^n) \cup \Theta$ where Θ denotes the empty string. For any $s \in \mathbf{R}^*$, denote by $N_r(s)$ the population of class r in the string s, and let

$$N(s) = (N_1(s), \cdots, N_r(s), \cdots, N_R(s)).$$

For K, a positive integer, let

$$S_{\mathbf{N}}^{K} = \{(s^1, \cdots, s^K): (\text{for every } i,\ 1 \leqslant i \leqslant K,\ s^i \in \mathbf{R}^*) \text{ and } \sum_{i=1}^{K} N(s^i) = \mathbf{N}\}.$$

As stated in Section 2.1, $S = (S^1, \cdots, S^M)$ denotes the Q-network state. Let $S_1 = (S^1, \cdots, S^{M_1})$ denote the Q_1-network state and $S_2 = (S^{M_1+1}, \cdots, S^M)$ denote the Q_2-network state. Let $\mathbf{F}_{\mathbf{N}}^{Q}$, and $\mathbf{F}_{\mathbf{N}}^{Q_1}$ and $\mathbf{F}_{\mathbf{N}}^{Q_2}$ denote respectively the sets of feasible states, in equilibrium, of the state process of the networks Q, Q_1, and Q_2, respectively.

A network $Q_{1\text{eq}}$ constructed from the nodes $\{1, \cdots, M_1\}$ is said to be *equivalent* to Q_1 if the joint equilibrium (probability) distribution of the state processes of $Q_{1\text{eq}}$ is the same as the marginal joint equilibrium distribution of the state process of Q_1 in Q.

2.3 Construction of $Q_{1\text{eq}}$

Let π: $\mathbf{F}_{\mathbf{N}}^{Q} \to (0, 1)$ be the equilibrium distribution of the state process of the network Q. Let (for every $(1 \leqslant i \leqslant M)(1 \leqslant r \leqslant R)$) (for every $S \in \mathbf{F}_{\mathbf{N}}^{Q}$) $\nu_{ir}(S) = \nu_{ir}(S_1)$ and (for every $S_1 \in \mathbf{F}_{\mathbf{N}}^{Q_1}$)

$$\rho_{ir}^{S_1} \triangleq \pi \{\text{A customer of type } r \text{ is in service at node } i/Q_1 \ i \text{ is in state } S_1\}.$$

Construct a network $Q_{1\text{eq}}$ from the nodes $\{1, \cdots, M_1\}$ as follows:

1. The routing between the nodes in $Q_{1\text{eq}}$ is the same as in Q_1 (self loops around nodes in Q_1 are included in $Q_{1\text{eq}}$).

2. When the state of $Q_{1\text{eq}}$ is S_1, node $j (1 \leqslant j \leqslant M_1)$ receives an exogenous arrival stream of class r customers with (state dependent) rate $\sum_{i=M_1+1}^{M} \rho_{ir}^{S_1} \nu_{ir}(S_1) p_{ij}^{(r)}$.

3. A customer of class r, after completing service at node i $(1 \leqslant i \leqslant M_1)$, leaves the network $Q_{1\text{eq}}$ with probability $\sum_{j=M_1+1}^{M} p_{ij}^{(r)}$.

Theorem 1: Q_{1eq} as constructed above is equivalent to the subnetwork Q_1 of Q.

Proof: The intuitive appeal of the construction is manifest. In Step 2 of the construction, for every $i(M_1 + 1 \leqslant i \leqslant M)$, $\rho_{ir}^{S_1} \nu_{ir}(S_1)$ is the conditional throughput of type r customers through node i, when Q_1 is in the state S_1. A fraction $p_{ij}^{(r)}$ of this flow through i finds its way into node j of Q_1.

A simple detailed proof can be obtained by summing the Kolmogorov equilibrium equations for Q over the set $\{S \in \mathbf{F}_{\mathbf{N}}^{Q} : S_1 \text{ fixed}\}$, and observing that the resulting equations are exactly the equilibrium equations for $Q_{1\text{eq}}$ described above (cf. Ref. 7). □

Remarks: But for the explosion in notation that occurs in setting up

a detailed proof, it is clear that the construction of Q_{1eq} described above extends easily to networks other than those described in Section 2.1. In this work, however, we continue to restrict our attention to networks of the latter type.

We now turn to the subclass of product-form networks of the class of networks described in Section 2.1. Since we are concerned here with FIFO nodes, the service rates cannot be class dependent. We further assume that the service rates are not state dependent in any way, i.e., we now have

$$(\text{for every } (1 \leqslant i \leqslant M)(1 \leqslant r \leqslant R) \quad \text{and} \quad S \in \mathbf{F}_{\mathbf{N}}^{Q}) \nu_{ir}(S) = \nu_i.$$

Let (for every $r \in \mathbf{R}$) $\mathrm{C}^{(r)} \subset \{1, \cdots, \mathrm{M}\}$ be the subset of nodes of Q that communicate under $P^{(r)}$ (i.e., in queueing-network terminology, the chain corresponding to class r). Let $\mathbf{R}_2 = \{r \in \mathbf{R} : \mathrm{C}^{(r)} \cap \{M_1 + 1, \cdots, M\} \neq \varnothing\}$ be the set of customer types that visit Q_2. Let $\|\mathbf{R}_2\| = R_2$, $\|\mathbf{R} - \mathbf{R}_2\| = R_1$ (where $\| \ \|$ denotes set cardinality), and reindex $\mathbf{R}$ so that the elements of $\mathbf{R}_2$ receive the highest indices. Let $\mathbf{N}^2 = (N_{R_1+1}, \cdots, N_R)$ and if s is a string in $\mathbf{R}^*$, let $N^2(s) = (N_{R_1+1}(s), \cdots, N_R(s))$, i.e., $N^2(s)$ is the population vector, of the string s, restricted to the classes in $\mathbf{R}_2$.

For every $\mathbf{N}' = (N'_{R_1+1}, \cdots, \mathrm{N}'_R) \leqslant \mathbf{N}^2$, consider the network $Q'_2(\mathbf{N}')$ obtained from Q by replacing all servers in Q_1 with infinite speed servers (i.e., by short-circuiting the nodes in Q_1), and placing $\mathbf{N}'$ customers in the resulting network. Let $\pi_{\mathbf{N}'}$ be the equilibrium distribution of the state process of the network $Q'_2(\mathbf{N}')$. Define for every $(M_1 + 1 \leqslant i \leqslant M)$, $r \in \mathbf{R}_2$,

$$\xi_{ir}^{\mathbf{N}'} \triangleq \pi_{\mathbf{N}'} \{\text{A customer of type } r \text{ is in service at node } i \text{ in } Q'_2(\mathbf{N}')\}.$$

2.4 The product-form case

Theorem 2: If Q is a product-form network then

$$\text{for every } (M_1 + 1 \leqslant i \leqslant M)(r \in \mathbf{R}_2) \text{ and every } S_1 \in \mathbf{F}_{\mathbf{N}}^{Q_1}$$

$$\rho_{ir}^{S_1} = \xi_{ir}^{\mathbf{N}^2 - N^2(S_1)} \qquad (\rho_{ir}^{S_1} \text{ as defined earlier}).$$

(Note: it is obvious that (for every $(M_1 + 1 \leqslant i \leqslant M)$, $r \notin \mathbf{R}_2$ and $S_1 \in \mathbf{F}_{\mathbf{N}}^{Q_1}$) $\rho_{ir}^{S_1} = 0$.)

Proof: The proof utilizes a simple lemma and is outlined in the appendix. □

Remarks: Theorem 2, when combined with Theorem 1, yields a generalization of Norton's Theorem[1] to multinode subnetworks. Even though the previous development is specific to the class of networks described in Section 2.1, it is clear that the same approach can be used to extend Norton's Theorem to the entire class of closed product-form networks. The product-form solution continues to play the same role

as it does in Theorem 2, i.e., it allows the rates of the external arrival streams in Q_{1eq} to be computed from an analysis of Q_2' for all possible customer populations in Q_2'.

III. AN APPROXIMATION SCHEME

In an IBM Research Report, Chow and Yu[5] suggest a somewhat ad hoc, iterative approximation scheme for a class of central-server models, with a priority discipline at the central server. As mentioned earlier, the scheme relies on an inexact application of Norton's Theorem to such networks. In Section I we described a natural generalization of this so-called equivalent flow approximation scheme to more general nonproduct-form networks. In this section, we present the results of a detailed study of the application of this approximation to a simple, test-bed, central-server network.

3.1 The test-bed model

Consider the two-node network Q shown in Fig. 2. There are two customer classes, namely 1 and 2, with N_1 and N_2 customers, respectively (i.e., $\mathbf{N} = (N_1,N_2)$). At node 1, the customers of class 1 (high-priority) have preemptive priority over class 2 (low-priority) customers; after being preempted by a class 1 customer, when a class 2 customer reaches the service station again, it resumes service where it left off; class 1 and 2 customers have exponential service times with rates ν_{11} and ν_{12}, respectively. Such a service discipline is commonly referred to as a *preemptive resume* discipline. At node 2, there is no priority; customers are served in the order in which they arrive (irrespective of class), at the class independent exponential service rate ν_2. Customers alternately seek service at nodes 1 and 2 and stay in the network forever. This model belongs to a class of central-server networks that arise as models of computer systems.

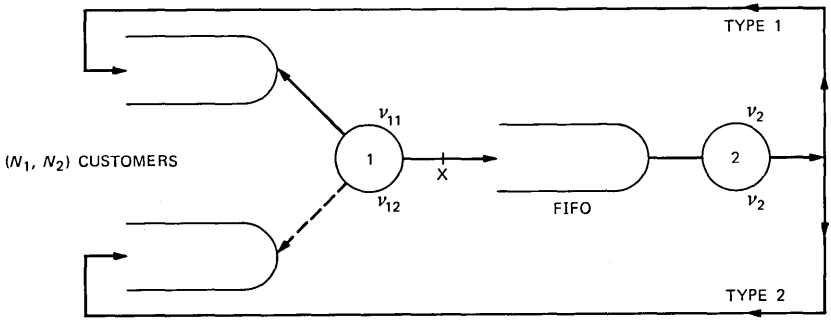


Fig. 2—Q.

3.2 *Approximating the test-bed network*

The network described in Section 3.1 is nonproduct form because of the preemptive resume discipline at node 1. In order to approximate the equilibrium behavior of node 1, we first increase the service rates at node 1 to infinity, thus effectively short circuiting the node. Denote the resulting network by Q_2' (Fig. 3). Then for each $(k_1,k_2) \leq (N_1,N_2)$ analyze Q_2' with k_1 and k_2 customers of types 1 and 2, respectively, in the network. Let (cf. Thm. 2) (for every $(k_1,k_2) \leq (N_1,N_2)$) (for every $r \in \{1, 2\}$) $\xi_{2r}^{(k_1,k_2)}$ = Prob {A customer of type r is in service at node 2 when (k_1,k_2) customers are in Q_2'}.

This probability will not depend on the sequence in which the (k_1,k_2) customers are placed in Q_2'. Since the service rate at node 2 is class independent, it is clear that, in equilibrium, all possible states, for any arrangement of the customers, are equally likely. From this we can directly conclude that

$$\text{(for every } (0, 0) < (k_1,k_2) \leq (N_1,N_2)\text{) (for every } r \in \{1, 2\}\text{)},$$

$$\xi_{2r}^{(k_1,k_2)} = \frac{k_r}{k_1 + k_2}.$$

Now consider the open network $\hat{Q}_{1\text{eq}}$ consisting of the node 1 in isolation. The service rates and discipline remain the same as in Q. When there are n_1 customers of type 1 and n_2 customers of type 2 in $\hat{Q}_{1\text{eq}}$ then customers of type r ($\in \{1, 2\}$) enter the network at the rate $\hat{\lambda}_r^{(n_1,n_2)}$ where

$$\text{(for every } (n_1,n_2) \leq (N_1,N_2))\hat{\lambda}_r^{(n_1,n_2)} = \xi_{2r}^{\mathbf{N}-(n_1,n_2)}\nu_2.$$

When a customer finishes service in $\hat{Q}_{1\text{eq}}$, it leaves the network (Fig. 4).

The evolution of the network $\hat{Q}_{1eq}$ can be described by a regular Markov process on the state space $\{(n_1,n_2) : (n_1,n_2) \leq (N_1,N_2)\}$. The idea is to approximate the equilibrium distribution of customers at node 1 in Q with the equilibrium distribution of customers in $\hat{\mathbf{Q}}_{1\text{eq}}$.

At first glance, the approximation technique described above may seem rather ad hoc. However, we can draw upon the development in

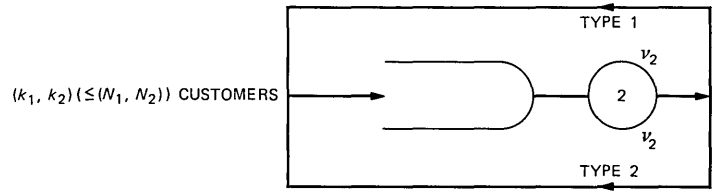


Fig. 3—Q_2.

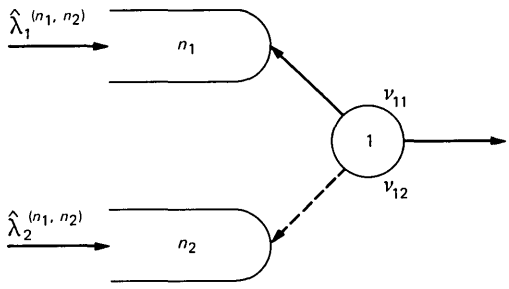


Fig. 4—$\hat{Q}_{1\,\mathrm{eq}}$.

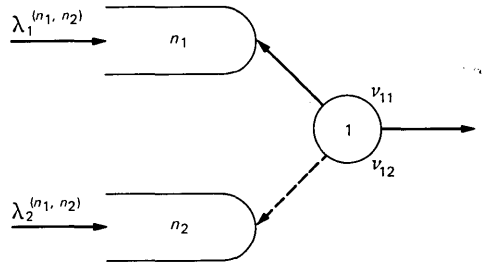


Fig. 5—$Q_{1\,\mathrm{eq}}$.

Section II to understand the inner workings of the test-bed model, and to show that, at least in principle, the approximation scheme is not altogether unreasonable.

It is clear that we can think of node 1, in the test-bed network Q, as comprising two FIFO nodes with service rates that depend just on the joint state of these two nodes. Consider the subnetwork Q_1 of Q consisting only of node 1. Theorem 1 can now be invoked to determine the exact equivalent network $Q_{1\mathrm{eq}}$. Let

(for every $(n_1,n_2) \leqslant (N_1,N_2)$) (for every $r \in \{1, 2\}$),

$$\rho_{2r}^{(n_1,n_2)} = \pi\{\text{A customer of type } r \text{ is in service at node } 2/(n_1,n_2) \text{ customers in } Q_1\}.$$

$Q_{1\mathrm{eq}}$ is then an open network consisting of node 1. When there are n_1 customers of type 1 and n_2 customers of type 2 in $Q_{1\mathrm{eq}}$, then customers of type r ($\in \{1, 2\}$) enter the network at the rate $\lambda_r^{(n_1,n_2)}$ where

$$(\text{for every } (n_1,n_2) \leqslant (N_1,N_2))\lambda_r^{(n_1,n_2)} = \rho_{2r}^{(n_1,n_2)}\nu_2.$$

When a customer finishes service in $Q_{1\mathrm{eq}}$ it leaves the network (Fig. 5).

The equilibrium distribution of customers in $Q_{1\mathrm{eq}}$ is exactly the same

as the equilibrium distribution of customers in Q_1. Observe, though, that the form of Q_{1eq} is the same as that of $\hat{Q}_{1eq}$, the difference lying in the state-dependent input rates. It is in this sense that the approximation scheme is reasonable. The idea now is to compare the exact state-dependent input rates, $\rho_{2r}^{(n_1,n_2)}\nu_2$, with the approximate state-dependent rates,

$$\xi_{2r}^{N-(n_1,n_2)}\nu_2 = \left(\frac{N_r - n_r}{N_1 - n_1 + N_2 - n_2}\,\nu_2\right),$$

i.e., to compare $\rho_{2r}^{(n_1,n_2)}$ with

$$\frac{N_r - n_r}{N_1 - n_1 + N_2 - n_2}$$

for all $(n_1,n_2) \lneqq (N_1,N_2)$ and for $r \in \{1, 2\}$.

3.3 Qualitative evaluation of the approximation

In this section, we present a qualitative evaluation of the approximation scheme as applied to the test-bed model.

Observe that if the service rates for the two FIFO queues comprising node 1, in Q_1, were not state dependent (in the priority scheme they are state dependent), then Q_1 would, in fact, be a product-form network. Theorem 2 would then lead us to conclude that $\hat{Q}_{1eq}$ and Q_{1eq} were the same. Consider what happens if, in Q_1, ν_{11} is allowed to go to infinity. Then, effectively, the high-priority customers do not interfere with the low-priority customers at node 1. With $\nu_{11} = \infty$, the network becomes the one shown in Fig. 6, which is a product-form network. Thus according to our observation above, for values of ν_{11} that are large, compared to ν_{12} and ν_2, the approximation can be expected to yield very good results.

In order to discover the situations in which the approximation can be expected to behave poorly, one needs to understand what aspects

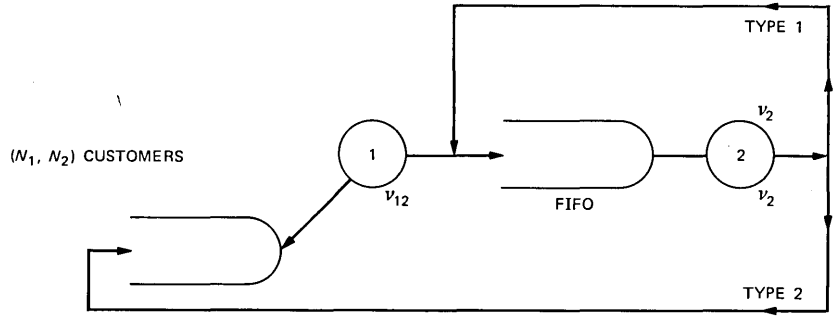


Fig. 6—"Lim" Q "$11^{\to\infty}$.

of the exact network the approximation fails to capture. If Q_1 were a product-form network, then, given that $(k_1,k_2)(\leqslant(N_1,N_2))$ customers were in node 2, all arrangements of customers within the node would be equally likely. As it stands, however, at node 1, priority 1 customers can preempt customers of priority 2. This suggests that given $(k_1,k_2)(\leqslant N_1,N_2)$ customers in node 2, some arrangements of customers would be more likely than others. In fact, we conjecture that priority 1 customers are more likely to be ahead of priority 2 customers, leading to the (conjectured) conclusion that

$$(\text{for every } (0,\,0) < (k_1,k_2) \leqslant (N_1,N_2))\rho_{21}^{\mathbf{N}-(k_1,k_2)} \geq \frac{k_1}{k_1 + k_2},$$

and

$$\rho_{22}^{\mathbf{N}-(k_1,k_2)} \leqslant \frac{k_2}{k_1 + k_2}.$$

Thus $\hat{Q}_{1\text{eq}}$ uses smaller (resp. larger) state-dependent input rates for type 1 (resp. type 2) customers than the exact equivalent $Q_{1\text{eq}}$. This idea is suggestive, but it is difficult to draw any immediate conclusions from this conjecture as to the relationship between exact and approximate performance measures of the network.

Another approach to discovering the direction in which the approximation can be expected to err is to observe that if node 2 in Q is replaced by a processor-sharing node, with class-independent service rate ν_2, then $\hat{Q}_{1\text{eq}}$ becomes the exact equivalent of Q_1 (cf. Fig. 7). (This follows because when node 2 is processor sharing, if $(k_1,k_2)(\leqslant (N_1,N_2))$ customers are present at node 2, then the rate of flow of class $r(\in \{1, 2\})$ customers into node 1 is $(k_r/(k_1 + k_2))\nu_2$.) To fix ideas consider the case $N_1 = n(\geqslant 1)$ and $N_2 = 1$. The throughput of the class 2 customer is simply the reciprocal of the mean successive passage times of the (single) class 2 customer through the point X (cf. Figs. 2 and

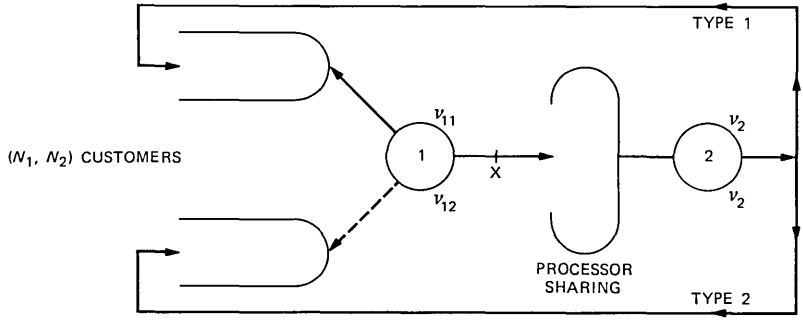


Fig. 7—Q with node 2 processor sharing.

7). In either network, when the class 2 customer crosses the point X to enter node 2, it finds all the class 1 customers receiving service at this node. In the original network, since node 2 is FIFO, the class 2 customer will have to wait for full service completion of the n class 1 customers before it can leave node 2 (and subsequently, at some future time instant, cycle back through X). Thus, the mean sojourn time of the type 2 customer, in node 2 of the original network, is $(n + 1)/\nu_2$. However, if node 2 is processor sharing, then on entering node 2, the customer of type 2 starts receiving service immediately at the rate $\nu_2/(n + 1)$, and continues to receive service at a rate $\nu_2/(k + 1)$ $(0 \leq k \leq n)$ until it finally leaves. Thus, in this case, the sojourn time of the class 2 customer at node 2 is stochastically dominated by an exponentially distributed random variable with mean $n + 1/\nu_2$, and hence has a mean smaller than $(n + 1)/\nu_2$. We further expect, intuitively, that, after completing service at node 2, when the type 2 customer returns to node 1, it expects to find more type 1 customers at node 1 when node 2 is FIFO than when it is processor sharing. Given that the type 2 customer finds $k(0 \leq k \leq n)$ type 1 customers on its arrival at node 1, its sojourn time at node 1 does not depend on whether node 2 is FIFO or processor sharing, and increases with increasing k. Thus, we expect that the mean sojourn time of the type 2 customer at node 1 will be larger if node 2 is FIFO than when it is processor sharing. The conclusion is that the mean passage time of the type 2 customer, through the point X, is larger in the original network than in the approximating network.

To see the magnitude of the error this effect could cause, let $N_1 = 1$ and allow $\nu_{12} \to \infty$, $\nu_2 \to \infty$, and $\nu_2/\nu_{12} \to 0$. Under these assumptions, in the original network, the class 2 customer will be blocked once (and only once) at node 1 each time it cycles through the point X. The average time it spends in the blocked condition is $1/\nu_{11}$. The rest of the time in each cycle tends to 0. Hence, the mean response time for the class 2 customer in the original network is $1/\nu_{11}$. If node 2 is replaced by a processor sharing node, then, each time the class 2 customer cycles through X, it is blocked once (and only once) with probability 1/2. Hence, the mean response time for the class 2 customer in the approximating network is $1/2\nu_{11}$, thus yielding an error of 100 percent.

We do not yet have a simple but rigorous argument that would allow us to say conclusively that the approximation yields higher throughputs for low-priority customers. However, the arguments presented above do make the conclusion plausible.

3.4 Numerical examples

To examine how the approximation works with specific examples,

we wrote a FORTRAN program to solve the equilibrium equations for $\hat{Q}_{1eq}$ using a simple recursive technique.[8] The program was somewhat more general, in that it could accept arbitrary state-dependent input rates and output rates. Thus, the same program could be used to solve the network exactly, if it were given the exact values of $\rho_{12}^{(n_1,n_2)}$ and $\rho_{22}^{(n_1,n_2)}$ for the various feasible (n_1,n_2).

It is not hard to calculate exactly the probabilities $\rho_{21}^{(n_1,n_2)}$ and $\rho_{22}^{(n_1,n_2)}$ for some simple cases. Of course for $(n_1,n_2) \neq (N_1,N_2)$, $\rho_{22}^{(n_1,n_2)} = 1 - \rho_{21}^{(n_1,n_2)}$. Consider, for the purpose of illustration, the case $N_1 = 1$, $N_2 = 1$. Note that the state of the network Q is completely described by the state of node 2. The epochs of entry into the state $S^2 = (12)$ are renewal epochs. The next state is, inevitably, $S^2 = (2)$. The next state is $S^2 = (21)$ with probability $\nu_{11}/(\nu_{11} + \nu_2)$ and $S^2 = (\Theta)$ with probability $\nu_2/(\nu_{11} + \nu_2)$. Because of the preemptive discipline, the next state to be entered in the set $\{(12), (21)\}$ will be $S^2 = (12)$, thus completing a renewal cycle. Since the expected holding time in each state in $\{(12), (21)\}$ is $1/\nu_2$, therefore

$$\rho_{21}^{(0,0)} = \frac{\dfrac{1}{\nu_2}}{\dfrac{1}{\nu_2} + \dfrac{\nu_{11}}{\nu_{11} + \nu_2} \cdot \dfrac{1}{\nu_2}} = \frac{1}{1 + \dfrac{\nu_{11}}{\nu_{11} + \nu_2}},$$

and, of course,

$$\rho_{21}^{(1,0)} = 0 \quad \text{and} \quad \rho_{21}^{(0,1)} = 1.$$

In Table I we list the exact expressions for $\rho_{21}^{(n_1,n_2)}$, for all $(n_1,n_2) \leqslant (N_1,N_2)$, for some values of (N_1,N_2). These were computed in the same fashion as in the above example.

In Tables II(a), (b), and (c), we give several numerical examples of exact and approximate solutions of the test-bed network. The exact solutions and the approximate solutions were obtained using the FORTRAN program described above. The program yields the equilibrium joint-probability distribution of queue lengths at node 1. In Tables II(a), (b), and (c), we display these joint probabilities and the node 1 utilizations.

The following observations are immediate and summarize our conclusions regarding the performance of the approximation scheme when applied to the test-bed network.

1. The numerical computations support our earlier observations that if ν_{11} is large, then the approximation can be expected to yield excellent results [cf. case (1) in each of Tables II(a), (b), and (c)].

2. The low-priority utilizations are consistently higher, again supporting our earlier observations regarding the direction in which the approximation can be expected to err.

Table I—Exact expressions for $\rho_{21}^{(n_1,n_2)}$ in the test-bed network Q. (cf. Thm. 1 and Fig. 2)

N_1	N_2	(n_1,n_2)	$\rho_{21}^{(n_1,n_2)*}$	Comparison with $\hat{\rho}_{21}^{(\cdot,\cdot)}$
1	1	(1,0)	0	
		(0,1)	1	
		(0,0)	$\dfrac{1}{1+\dfrac{\nu_{11}}{\nu_{11}+\nu_2}}$	$\geq \frac{1}{2}$
2	1	(2,0)	0	
		(0,1), (1,1)	1	
		(1,0)	$\dfrac{1}{1+\dfrac{\nu_{11}}{\nu_{11}+\nu_2}+\dfrac{\nu_2}{\nu_{11}+\nu_2}\cdot\dfrac{\nu_{11}}{\nu_{11}+\nu_2}}$	$\geq \frac{1}{2}$
		(0,0)	$\dfrac{1+\dfrac{\nu_{11}}{\nu_{11}+\nu_2}}{1+\dfrac{\nu_{11}}{\nu_{11}+\nu_2}+\left(\dfrac{\nu_{11}}{\nu_{11}+\nu_2}\right)^2+\dfrac{\nu_2}{\nu_{11}+\nu_2}\cdot\left(\dfrac{\nu_{11}}{\nu_{11}+\nu_2}\right)^2}$	$\geq \frac{2}{3}$
1	2	(1,0), (1,1)	0	
		(0,2)	1	
		(0,1)	$\dfrac{1}{1+\dfrac{\nu_2}{\nu_{12}+\nu_2}\cdot\dfrac{\nu_{11}}{\nu_{11}+\nu_2}+\dfrac{\nu_{12}}{\nu_{12}+\nu_2}\cdot\left[\dfrac{\nu_{11}}{\nu_{11}+\nu_2}+\dfrac{\nu_2}{\nu_{11}+\nu_2}\cdot\dfrac{\nu_{11}}{\nu_{11}+\nu_2}\right]}$	$\geq \frac{1}{2}$
		(0,0)	$\dfrac{1}{1+\dfrac{\nu_{11}}{\nu_{11}+\nu_2}+\dfrac{\nu_{11}}{\nu_{11}+\nu_2}\cdot\dfrac{\nu_{12}}{\nu_{12}+\nu_2}\left[1+\dfrac{\nu_2}{\nu_{12}}+\dfrac{\nu_2}{\nu_2+\nu_{11}}\right]}$	$\geq \frac{1}{3}$

*$(\rho_{22}^{(n_1,n_2)} = 1 - \rho_{21}^{(n_1,n_2)}$ if $(n_1,n_2) \neq (N_1,N_2)$; $\rho_{21}^{(N_1,N_2)} = \rho_{22}^{(N_1,N_2)} = 0)$.

3. When ν_{11}, ν_{12} and ν_2 are comparable, then the approximation yields good results with errors in the utilizations in the neighborhood of 10 percent.

4. Considerable errors in the low-priority utilizations can arise, however. Witness case 3 in each of the Tables II(a), (b), and (c). With a very large low-priority service rate at node 1, the approximate low-priority utilization suffers from an error of 20 to 50 percent.

5. For the range of examples studied, the equilibrium probabilities

Table II—Numerical comparisons of exact and approximate solutions of the test-bed network

Case No.	ν_{11}	ν_{12}	ν_2	Equilibrium State Probabilities at Node 1: State (n_1,n_2)	Exact Probability	Approximate* Probability	Node 1 Utilizations: Class	Exact Utilization	Approximate* Utilization
				(a)	($N_1 = 1, N_2 = 1$)				
1	10	1	1	(0,0)	0.614	0.606			
				(1,0)	0.029	0.028	1	0.064	0.063
				(0,1)	0.322	0.331	2	0.322	0.331
				(1,1)	0.035	0.036			
2	1	.5	1	(0,0)	0.231	0.222			
				(1,0)	0.077	0.056	1	0.462	0.444
				(0,1)	0.308	0.333	2	0.308	0.333
				(1,1)	0.385	0.389			
3	1	100	2	(0,0)	0.393	0.488			
				(1,0)	0.196	0.163	1	0.601	0.504
				(0,1)	0.0059	0.0081	2	0.0059	0.0081
				(1,1)	0.405	0.341			
				(b)	($N_1 = 2, N_2 = 1$)				
1	10	1	1	(0,0)	0.681	0.676			
				(1,0)	0.044	0.043			
				(2,0)	0.002	0.002	1	0.0754	0.0748
				(0,1)	0.244	0.249	2	0.244	0.249
				(1,1)	0.027	0.027			
				(2,1)	0.003	0.003			
2	1	.5	1	(0,0)	0.179	0.171			
				(1,0)	0.083	0.065			
				(2,0)	0.024	0.016	1	0.631	0.618
				(0,1)	0.191	0.211	2	0.191	0.211
				(1,1)	0.250	0.260			
				(2,1)	0.274	0.276			
3	1	100	2	(0,0)	0.170	0.217			
				(1,0)	0.116	0.109			
				(2,0)	0.050	0.036	1	0.827	0.780
				(0,1)	0.0022	0.0033	2	0.0022	0.0033
				(1,1)	0.187	0.188			
				(2,1)	0.474	0.447			
				(c)	($N_1 = 1, N_2 = 2$)				
1	10	1	1	(0,0)	0.464	0.457			
				(1,0)	0.015	0.014			
				(0,1)	0.316	0.319	1	0.0495	0.0487
				(1,1)	0.016	0.016	2	0.487	0.494
				(0,2)	0.170	0.175			
				(1,2)	0.019	0.019			
2	1	.5	1	(0,0)	0.116	0.109			
				(1,0)	0.028	0.108			
				(0,1)	0.177	0.182	1	0.437	0.418
				(1,1)	0.070	0.055	2	0.447	0.473
				(0,2)	0.270	0.291			
				(1,2)	0.340	0.346			
3	1	100	2	(0,0)	0.473	0.584			
				(1,0)	0.167	0.130			
				(0,1)	0.008	0.010	1	0.517	0.404
				(1,1)	0.115	0.090	2	0.0101	0.0123
				(0,2)	0.002	0.002			
				(1,2)	0.235	0.184			

*Based on Norton's Equivalent.

are never drastically wrong, and follow trends similar to the exact values.

The test-bed model is hard to analyze exactly for population sizes larger than the ones considered. We have run detailed simulations of the test-bed model for larger population sizes, and the results of these simulations continue to support the qualitative observations we have made above.

IV. CONCLUSIONS

We have demonstrated an approach for systematically analyzing the equivalent flow approximation. Our investigations have (1) revealed the conceptual basis for the approximation scheme, and (2) led to an understanding of the reasons for, and directions of, the errors that such an approximation scheme could introduce when applied to a class of prioritized central-server models. The approximation as described in the paper is of more general applicability, and much work remains to be done to discover its validity (accuracy and computational tractability) for more complicated, nonproduct-form networks. Our work, we think, provides the theoretical understanding and motivation for pursuing more detailed investigations.

V. ACKNOWLEDGMENTS

We are indebted to J. S. Kaufman for introducing us to the problem. We are grateful to him and to B. T. Doshi, B. Melamed, and B. Sengupta for making themselves available for several helpful discussions.

REFERENCES

1. K. M. Chandy, U. Herzog, L. Woo, "Parametric Analysis of Queueing Networks," IBM J. Res. and Develop. *19,* No. 1 (January 1975), pp. 36–42.
2. P. S. Kritzinger, S. Van Wyck, A. E. Krzesinski, "A Generalization of Norton's Theorem for Multiclass Queueing Networks," Performance Evaluation, *2* (July 1982), pp. 98–107.
3. S. Balsamo and G. Iazeolla, "An Extension of Norton's Theorem for Queueing Networks," IEEE Trans. Software Eng., *SE-8,* No. 4 (July 1982), pp. 298–305.
4. C. H. Sauer and K. M. Chandy, "Approximate Analysis of Central Server Models," IBM J. Res. Develop., *19,* No. 3 (May 1975), pp. 301–13.
5. W. Chow and P. S. Yu, "An Approximation Technique for Central Server Queueing Models with a Priority Dispatching Rule," IBM Res. Rpt. (1980), RC8163 (No. 35479).
6. M. Schwartz, "Performance Analysis of the SNA Virtual Route Pacing Control," IEEE Trans. Commun., *COM-30,* No. 1, Part II (January 1982), pp. 172–84.
7. A. Kumar, unpublished work.
8. J. S. Kaufman, unpublished work.
9. F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, Closed and Mixed Networks of Queues with Different Classes of Customers," J. ACM, *22,* No. 2 (April 1975), pp. 248–60.

APPENDIX

Proof of Theorem 2: Index the nodes of Q_2' in the same order in which they were indexed in Q. We need the following lemma.

Lemma: *Let* $T^{(r)}$ *denote the class* r *routing probability matrix for the network* Q_2'. *Partition* $P^{(r)}$ *as follows*:

$$P^{(r)} = \begin{array}{c|c|c} & \overbrace{\phantom{P_{11}^{(r)}}}^{M_1} & \overbrace{\phantom{P_{12}^{(r)}}}^{M_2} \\ M_1 & P_{11}^{(r)} & P_{12}^{(r)} \\ \hline M_2 & P_{21}^{(r)} & P_{22}^{(r)} \end{array}.$$

Then

(1) $r \in \mathbf{R}_2 \Rightarrow T^{(r)} = P_{22}^{(r)} + P_{21}^{(r)}[I - P_{11}^{(r)}]^{-1}P_{12}^{(r)}$

(2) *If* $\Lambda^{(r)}$ *solves* $\Lambda^{(r)}P^{(r)} = \Lambda^{(r)}$ *then, partitioning* $\Lambda^{(r)}$ *as*

$$\Lambda^{(r)} = \left[\begin{array}{c:c} \Lambda_1^{(r)} & \Lambda_2^{(r)} \\ \overbrace{}^{M_1} & \overbrace{}^{M_2} \end{array}\right]$$

(*a*) *if* $r \notin \mathbf{R}_2$ *then* $\Lambda_2^{(r)} = 0$

(*b*) *if* $r \in \mathbf{R}_2$ *then* $\Lambda_2^{(r)} = \Lambda_2^{(r)}T^{(r)}$.

Proof of Lemma: Conclusion 1 follows readily from the fact that, for each $r \in \mathbf{R}$, $P^{(r)}$ is the transition probability matrix of a finite Markov chain with a single, positive communication class that has a nonempty intersection with $\{M_1 + 1, \cdots, M\}$. For details see Ref. 7.

Conclusion 2 follows directly from Conclusion 1. □

Returning to the proof of Theorem 2, we let π: $\mathbf{F}_{\mathbf{N}}^{Q} \to (0, 1)$ be the equilibrium distribution of the state process of network Q.* It is now well known (cf. Ref. 9) that $\pi(\cdot)$ is of the form

$$\pi(S) = \frac{1}{G}\prod_{i=1}^{M} f_i(S^i),$$

where G is a normalization constant and, for each $i \in \{1, \cdots, M\}$, f_i depends only $N(S^i)$, ν_i and (for every r, $(1 \leqslant r \leqslant R)$)$\lambda_i^{(r)}$, where $\Lambda^{(r)} = (\lambda_1^{(r)}, \cdots, \lambda_M^{(r)})$ is any solution of $\Lambda^{(r)}P^{(r)} = \Lambda^{(r)}$.

Hence, (for every $S_1 \in \mathbf{F}_{\mathbf{N}}^{Q_1}$) (for every i, r, $(M_1 + 1 \leqslant i \leqslant M)$, $r \in \mathbf{R}_2$)

$$\rho_{ir}^{S_1} = \frac{\sum_{\{S:S\in\mathbf{F}_{\mathbf{N}}^{Q},(S^1,\cdots,S^{M_1})=S_1,S^i(1)=r\}}\prod_{j=1}^{M} f_j(S^j)}{\sum_{\{S:S\in\mathbf{F}_{\mathbf{N}}^{Q},(S^1,\cdots,S^{M_1})=S_1\}}\prod_{j=1}^{M} f_j(S^j)}$$

$$= \frac{\sum_{\{S_2:S_2\in\mathbf{F}_{\mathbf{N}}^{Q_2}\cap S_{\mathbf{N}^2-N^2(S_1)}^{M_2},S_2^i(1)=r\}}\prod_{j=M_1+1}^{M} f_j(S_2^j)}{\sum_{\{S_2:S_2\in\mathbf{F}_{\mathbf{N}}^{Q_2}\cap S_{\mathbf{N}^2-N^2(S_1)}^{M_2}\}}\prod_{j=M_1+1}^{M} f_j(S_2^j)},$$

*For notation see Section 2.2.

(where $\mathbf{F}_{\mathbf{N}}^{Q_2} \cap S_{\mathbf{N}^2-N^2(S_1)}^{M_2}$ is the set of feasible states of Q_2 when

$$\mathbf{N}^2 - N^2(S_1) \text{ customers are in } Q_1),$$

which, using the above lemma and the fact that the equilibrium distribution of the Q_2' network state process is still product form,

$$= \xi_{ir}^{\mathbf{N}^2-N^2(S_1)}.$$

Remarks: Some care is needed in asserting the last equality in the case where there are classes r_1 and r_2, such that the submatrices of the communicating classes under $T^{(r_1)}$ and $T^{(r_2)}$ are the same permutation matrices (i.e., members of classes r_1 and r_2 cannot overtake each other). In this case the equality follows because for each $\mathbf{N}'$, $\xi_{ir_1}^{\mathbf{N}'}$ and $\xi_{ir_2}^{\mathbf{N}'}$ are independent of the order in which members of these classes circulate in the network Q_2'. □

AUTHOR

Anurag Kumar, B.Tech (Electrical Engineering), 1977, Indian Institute of Technology, Kanpur, India; Ph.D. (Major: Electrical Engineering, Minor: Operations Research/Math), 1981, Cornell University; Bell Laboratories, 1981—. Mr. Kumar joined the Performance Analysis Department in December 1981, and has since been involved in two projects: a study of a class of approximation schemes for queueing-network analysis; and the identification, analysis, and evaluation of possibilities for giving priority to critical user traffic in the Public Switched Network.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

A Model for Special-Service Circuit Activity

By D. R. SMITH*

(Manuscript received April 20, 1983)

We describe a model for special-service circuit activity to assist in forecasting, provisioning, and "churn" studies. We assume that customers order a random number of circuits for an exponentially distributed period of time and that the rate of new connect orders grows exponentially with time. These assumptions yield simple formulae giving the means and variances of the number of active circuits at a future time and the total number of connected and disconnected circuits during a future period. Distributions of these variables can, in principle, also be computed. There are three important parameters characterizing the model: growth rate, disconnect rate, and batchiness; we describe their physical meaning and discuss methods to estimate them. This document describes the analytical portion of an effort to develop a model based on the physics of special-service circuit activity.

I. INTRODUCTION

The purpose of this paper is to describe a model for special-service circuit activity to assist in forecasting, provisioning, and "churn" studies, which can be summarized by a few parameters that have a physical interpretation. The calibration and measurement of the fit of this model to data in a New Jersey Bell database is being pursued simultaneously and will be reported elsewhere.

The model treated here is derived from a priori consideration of the physical behavior of customers. It is based on the assumption that the number of active circuits, although growing, is in some sense in

*Bell Laboratories.

equilibrium as well; that is, certain characteristics of the system are not changing. This is to be contrasted with a model proposed by Nucho in which transient analysis is fundamental.[1] The primary difference between these models is that the demand rate for new circuits is a function of the number of active circuits in the Nucho model, whereas it is considered to be an exogenous variable here. In the Nucho model, the variance to mean ratio of the number of active circuits increases indefinitely with time (since fluctuations tend to feed on themselves); in the model considered here this ratio remains constant. Another difference between the models is that the model described here allows an order to be for more than one circuit.

Here, we assume that (1) the arrivals of special-service circuit orders are given by a nonhomogeneous Poisson process with exponentially growing intensity, (2) each order is for a random number of circuits (a batch) with arbitrary distribution, and (3) the lifetime of an order is an exponentially distributed random variable, during which time the number of held circuits per order remains constant. Note that the last assumption implies that an order lifetime and a circuit lifetime have the same distribution.

We use three important parameters in special-services modeling, each with its own physical interpretation. These parameters may be described as growth rate, disconnect rate (per circuit), and batchiness.

The growth rate summarizes the rate at which the mean number of active circuits increases with time. It may be expressed in terms of proportion increase per unit of time; we denote it β. Thus the mean number of circuits at time t is proportional to $e^{\beta t}$. We actually assume that connect activity grows at rate β, but it turns out that the number of active circuits, the total connect rate, and the total disconnect rate are all proportional to $e^{\beta t}$ in this model. Of course, for small growth rates or short periods of time, exponential growth is very close to linear growth.

The disconnect rate, denoted μ, is the ratio of the number of disconnects per unit time (i.e., the total disconnect rate) to the number of active circuits. The mean circuit lifetime is then $1/\mu$. The distributions of circuit lifetimes have been shown to be well approximated by negative exponential distributions;[2] thus the disconnect rate does not vary with the age of a circuit.

The batchiness of the arrival process is related to the tendency of special service circuits to be ordered in multiples greater than one. We call the batchiness parameter ν and define it to be the ratio of the second moment to the first moment of the number of circuits in an order.

The ultimate goal of this modeling process is to provide a tool that can be used to predict special-services needs in the future. The model

contained herein should be very useful in this regard. One should remember that the underlying process is stochastic so that there is a fundamental uncertainty even if one has exact specification of the parameters of the model. The standard deviation of future requirements can be quite large compared to the mean for small circuit groupings, and this presents a major problem for provisioning at the most detailed level. This problem cannot be surmounted with a better model and/or additional data collection. The present analysis allows quantification of the fundamental uncertainty of forecasting, an insight which is difficult to obtain purely by statistical methods. The only possible method to further decrease relative uncertainty is to aggregate demand, or to obtain advance knowledge of connect or disconnect activity (sometimes called "deterministic events").

The rest of the paper is organized as follows: Section II summarizes the important results of the paper, giving formulae for the means and variances of the number of active circuits in the future, the total number of connects in a future interval, and the total number of disconnects in a future interval; and giving statistical methods to estimate the fundamental parameters of the model such as growth rate, disconnect rate, and batchiness. The reader not interested in the derivation of these results may stop at this point.

The predictions (summarized in Section II) of the model are derived in Section IV. These derivations are primarily substitutions into formulae given in Section III. Section III describes and analyzes a much more general model than the one described in this introduction (we refer to the latter simply as "the model"). We have chosen to introduce this generalized model for two reasons. First, the analysis required for the treatment of the generalized model is little different in complexity from that required for treatment of the specific model. Second, the general results of Section III allow rapid exploration of the consequences of changes in assumptions of the model. For example, one can explore the effects of linear growth of demand, or the super-exponential growth in demand which follows introduction of a new service. However, we do feel that the original assumptions are appropriate in most circumstances. Thus, the consequences of this model are the only ones summarized in Section II, and it is this specific model which is being verified with respect to the New Jersey Bell Telephone Co. database. Thus, Section III is provided for reference in case of non-typical special service applications.

Section V derives the statistical methods (summarized in Section II) for estimation of the fundamental parameters of the model. Section VI is a summary.

Appendix A gives background information on the compound Poisson random variable, and Appendix B gives background information on

the non-homogeneous Poisson process. These results are needed in Sections III and IV.

Table I presents values of a function useful in estimating growth (see Section II) and Table II lists the notation used in the paper.

II. SUMMARY OF KEY RESULTS

This section provides a summary of the important results of the paper derived in Sections IV and V.

2.1 Churn

Our model depends on three physical parameters: growth rate (β), disconnect rate (μ), and batchiness (ν). The meaning of these parameters is described in Section I. Another physical parameter is "churn," which has been defined in many different ways. For any reasonable definition, the churn is determined by the growth and disconnect rates of the model. We define the churn to be the minimum of the disconnect rate per circuit and the connect rate per circuit, and denote it by γ. With this definition, it can be shown [see (75)] that

$$\gamma = \min(\mu, \mu + \beta). \tag{1}$$

The values of churn under other definitions are also readily available. For example, if one defines churn to be the ratio of the average total connect rate to the average rate of change of net active circuits, then this value of churn is $(1 - \gamma)^{-1}$. Under still another definition, the churn equals $\mu/(\mu + \beta)$.

2.2 Mean and variance of total active circuits at a future time

Here we give the mean $M(t)$ and the approximate variance $V(t)$ of the number of circuits in service at a given time t in the future. The mean and the variance depend on the present (at time $t = 0$) number k of circuits in service, the present instantaneous rate D_o of circuit demand due to new orders, and the three key parameters described previously: β, μ, and ν. We give these two relationships below:

$$M(t) = ke^{-\mu t} + \frac{D_o}{\mu + \beta}(e^{\beta t} - e^{-\mu t}), \tag{2}$$

and

$$V(t) = \nu\left[ke^{-\mu t}(1 - e^{-\mu t}) + \frac{D_o}{\mu + \beta}(e^{\beta t} - e^{-\mu t})\right]. \tag{3}$$

It is interesting that (2) and (3) together imply the relationship

$$V(t) = \nu(M(t) - ke^{-2\mu t}), \tag{4}$$

which relates the variance of a forecast to the mean of the forecast, the number of circuits currently active, k, and the parameters ν and μ.

2.3 Mean and variance of the total number of connected or disconnected circuits in the future

Similar results are available for the mean and variance of the total number of connected circuits (variables subscripted with a C) and the mean and variance of the total number of disconnected circuits (subscripted with a D) in an interval of length t beginning immediately:

$$M_C(t) = \frac{D_o}{\beta}(e^{\beta t} - 1), \tag{5}$$

$$V_C(t) = \nu \frac{D_o}{\beta}(e^{\beta t} - 1) = \nu M_C(t) \tag{6}$$

$$M_D(t) = k(1 - e^{-\mu t}) + \frac{D_o \mu}{\beta(\mu + \beta)} e^{\beta t} + \frac{D_o}{\mu + \beta} e^{-\mu t} - \frac{D_o}{\beta}, \tag{7}$$

and

$$V_D(t) = \nu \left\{ k e^{-\mu t}(1 - e^{-\mu t}) + \frac{D_o \mu}{\beta(\mu + \beta)} e^{\beta t} + \frac{D_o}{\mu + \beta} e^{-\mu t} - \frac{D_o}{\beta} \right\}$$
$$= \nu(M_D(t) - k e^{-2\mu t}). \tag{8}$$

In this case, the total numbers of connected and disconnected circuits are dependent random variables.

We may also obtain the coefficient of correlation ρ between the number of active circuits at different times

$$\rho[Y(t),\ Y(t + \tau)] = e^{-(\mu + \beta/2)\tau}, \tag{9}$$

where $Y(t)$ is the number of active circuits at time t.

2.4 Estimation of the model parameters

To use results such as (2) through (9), we must be able to estimate the parameters β, D_o, ν, and μ. These questions are addressed in Section V; we provide a brief summary here. Suppose that the system has been observed over the interval $[-\Theta, 0]$ and n connect orders are observed at times $t_1, \ldots, t_n$. Form the statistic

$$S = \sum_{i=1}^{n} t_i / n\Theta + 1, \tag{10}$$

and then the maximum likelihood estimator $\hat{\beta}$ for the growth rate β is

$$\hat{\beta} = \frac{1}{\Theta} f^{-1}(S), \tag{11}$$

where f is the function given in (79). Values of f^{-1} are available in Table I. Once $\hat{\beta}$ has been obtained from (11), the estimator $\hat{D}_o$ for the *instantaneous* present demand D_o (assumed to be at the end of the interval of observation $[-\Theta, 0]$) is

$$\hat{D}_o = \frac{n\hat{\beta}}{1 - e^{-\hat{\beta}\Theta}} \hat{N}, \tag{12}$$

where $\hat{N}$ is an estimator for the average number of circuits per order and is equal to the average number of circuits actually observed per order. The estimator $\hat{\nu}$ for the batchiness ν of the order size is

$$\hat{\nu} = \frac{\sum_{k=1}^{\infty} k^2 i_k}{\sum_{k=1}^{\infty} k i_k}, \tag{13}$$

where i_k is the observed number of existing orders of size k. The estimator $\hat{\mu}$ for the parameter μ can be obtained as the average disconnect rate for observed circuits

$$\hat{\mu} = \frac{m}{\tau}, \tag{14}$$

where m is the total number of disconnects observed, and τ is the sum of the observed connection times for all circuits; μ can also be obtained from estimators of the churn and growth rate through the use of (1).

Estimation of these parameters from data supplied by New Jersey Bell Telephone Co. is being investigated. Estimates of the disconnect rate $\hat{\mu}$ by service family are available in the Reed and Smith paper,[2] in which it is shown that the lifetimes of special-service circuits are well approximated by exponential functions with means dependent on the service families.

III. A GENERALIZED MODEL

This section treats a model that is more general than that which we propose for special-service activity in most cases. The analysis presented here will be applied to the specific model in Section IV.

3.1 Description of the generalized model

We examine an arbitrarily defined category of special-service circuits (for example, circuits of a particular service family in a given

wire center) and divide the active circuits into independent groups. Possibly, each group is the demand from a single user, since it is reasonable that the activity of one user does not affect another. To facilitate this method of thinking we shall refer to the groups as "orders." Each order becomes nonzero for the first time at some point in time (referred to as the arrival or connect time of the order) and then has some history of changing size in some arbitrary manner before possibly becoming zero again indefinitely at some time (the departure or disconnect time of the order). The length of the interval between the arrival and departure of an order will be called the lifetime of the order. Obviously, the number of active circuits at any time equals the sum of the sizes of the existing orders at that time.

We assume that there is a large pool of customers (or potential orders) so that the arrival of an order has little effect on the potential arrival of others. Thus, the arrival of orders can be modeled by a nonhomogeneous Poisson process, whose intensity at time t is given by some function $\lambda(t)$. For background on this process see Ross[3] or Karlin and Taylor.[4] Denote the probability that an arriving order at time t is initially of size m as $q_m(t)$, and let $P^*_{mn}(t, x)$ be the probability that an order arriving at time t of initial size m as becomes size n at time $x \geq t$.

3.2 Distribution of the number of active circuits at a given time

Since the orders are noninterfering it can easily be seen (see Appendix B) that the number of orders of size n at time x is Poisson distributed with mean $\alpha_n(x)$, where

$$\alpha_n(x) = \sum_m \int_{-\infty}^{x} \lambda(t) q_m(t) P^*_{mn}(t, x) dt, \tag{15}$$

and that the numbers of orders of different sizes at time x are independent of each other. If $Y(x)$ is the total number of active special-services circuits in the category of interest at time x, then $Y(x)$ has a compound Poisson distribution (see Appendix A), and

$$E[Y(x)] = \sum_{n=1}^{\infty} n\alpha_n(x), \tag{16}$$

and

$$\text{var}[Y(x)] = \sum_{n=1}^{\infty} n^2\alpha_n(x). \tag{17}$$

3.3 Distribution of future active circuits due to present orders

The transient behavior of this model is easily derived if one has knowledge of the distribution of *order sizes* at a given time. We treat

this case first and then consider the more difficult case where only the total number of *active circuits* at a given time is known. In either case, we will find the distribution of the number of active circuits at time y resulting from the orders observed to be active at time x. The total circuits active at time y is the sum of this with the number of circuits at time y resulting from orders arriving between x and y.

Case 1: Order sizes known

Given an order is of size n at time x, the conditional density that it arrived as an order of size m at time t is $\rho_{mn}(t, x)$, where

$$\rho_{mn}(t, x) = \frac{\lambda(t)q_m(t)P^*_{mn}(t, x)}{\alpha_n(x)} \tag{18}$$

and if several orders of size n are present at x their arriving times and sizes may be considered to be conditionally independent (see Appendix B). Thus an order of size n at time x becomes an order of size l at time $y \geq x$ with probability $r_{nl}(x, y)$ where

$$r_{nl}(x, y) = \alpha_n^{-1}(x) \sum_{m=1}^{\infty} \int_{-\infty}^{x} \lambda(t)q_m(t)P^*_{mn}(t, x)q_{mnl}(t, x, y)dt, \tag{19}$$

and $q_{mnl}(t, x, y)$ is the conditional probability that an order arriving as size m at time t which is of size n at time x becomes size l at time y. Note that q_{mnl} is not available solely from P^*.

Equation (19) allows us to compute the distribution of the total number of circuits at time y that were due to orders *observed* at time x, since all orders behave independently. Evaluating these distributions explicitly can be quite difficult. We can, however, easily evaluate the moments. Let $M_n(x, y)$ be the mean order size at time y for an order observed to be size n at time x, and let $V_n(x, y)$ be the mean order size at time y for an order observed to be size n at time x, and let $V_n(x, y)$ be the analogously defined variance. Then

$$M_n(x, y) = \sum_{l=1}^{\infty} l r_{nl}(x, y), \tag{20}$$

$$V_n(x, y) = \sum_{l=1}^{\infty} l^2 r_{nl}(x, y) - M_n^2(x, y). \tag{21}$$

If $i_n(x)$ is the number of orders of size n observed at time x, and $M(x, y)$ and $V(x, y)$ denote the mean and variance of the number of circuits at time y due to orders observed at time x, then

$$M(x, y) = \sum_n i_n(x)M_n(x, y), \tag{22}$$

and

$$V(x, y) = \sum_n i_n(x) V_n(x, y). \tag{23}$$

Note that there is a potential problem if orders can become size zero and then become nonzero later, since determination of i_0, the number of active orders of size 0, may be an impossible task.

Case 2: Order sizes unknown

We now examine the more difficult case where we observe the total number of active circuits at time x (call this k), without observing the distribution of the order sizes. The conditional probability that there are j_1 orders of size 1, j_2 orders of size 2, etc., given that k total circuits are observed at time x, written $\delta_{k,x}(j_1, j_2, \ldots)$, is easily found to be

$$\delta_{k,x}(j_1, j_2, \ldots) = \frac{\prod_i [\alpha_i(x)^{j_i}/j_i!]}{\sum_{j_1+2j_2+\cdots=k} \prod_i [\alpha_i(x)^{j_i}/j_i!]}, \tag{24}$$

provided that $j_1 + 2j_2 + \cdots = k$. Let the conditional first and second moment of the number of circuits at time y due to orders observed at time x, given that a total of k circuits were observed at time x, be $M_{k,x}(y)$ and $M^{(2)}{}_{k,x}(y)$ respectively. Then

$$M_{k,x}(y) = \sum E(J_i) M_i(x, y), \tag{25}$$

where J_i is a random variable with the same distribution as the conditional number of orders of size i at time x, so that the expectation is the expectation with respect to the probability distribution given in eq. (24). Also,

$$M^{(2)}_{k,x}(y) = \sum E(J_i) V_i(x, y) + E((\sum J_i M_i(x, y))^2), \tag{26}$$

where the expectation is in the same sense as before. Needless to say, these expectations with respect to the distribution in (24) are very difficult to evaluate for substantial k.

Things simplify somewhat if

$$M_i(x, y) = i\theta(x, y), \tag{27}$$

that is, if the conditional means are proportional to the size of the order. In this case, (25) and (26) give

$$M_{k,x}(y) = k\theta(x, y), \tag{28}$$

and

$$M^{(2)}_{k,x}(y) = \sum E(J_i) V_i(x, y) + k^2\theta^2(x, y), \tag{29}$$

so that

$$V_{k,x}(y) = \sum E(J_i) V_i(x, y), \tag{30}$$

where $V_{k,x}(y)$ denotes the conditional variance. Equation (30) can be approximated by using the following approximation which is intuitively reasonable for k near $\sum i\alpha_i(x)$,

$$E(J_i) \approx k \frac{\alpha_i(x)}{\sum i\alpha_i(x)}. \tag{31}$$

In this case, (30) and (31) give the following useful approximation:

$$V_{k,x}(y) \approx k \frac{\sum \alpha_i(x) V_i(x, y)}{\sum i\alpha_i(x)}. \tag{32}$$

3.4 Distribution of future active circuits due to future orders

In section 3.3, we found the mean and the variance of the number of circuits at time y from orders observed at time x. To obtain the total number of circuits at time y, we need to add to this the (independent) number of circuits due to orders arriving between time x and time y. The number of orders of size n at time y that arrived between times x and y is easily seen to be Poisson with mean $\alpha_n(x, y)$, where

$$\alpha_n(x, y) = \sum_{m=1}^{\infty} \int_x^y \lambda(t) q_m(t) P^*_{mn}(t, y) dt, \tag{33}$$

and the number of orders of different sizes are independent of each other (see Appendix B). Thus, the number of circuits at time y due to arrivals occurring between x and y has a compound Poisson distribution (see Appendix A) with mean and variance denoted $M^*(x, y)$ and $V^*(x, y)$, where

$$M^*(x, y) = \sum_n n\alpha_n(x, y), \tag{34}$$

and

$$V^*(x, y) = \sum_n n^2\alpha_n(x, y). \tag{35}$$

3.5 Mean and variance of future active circuits

To find expressions for the mean or variance of the total number of active circuits at time y, we merely add together the appropriate means or variances from the circuits active at time y due to orders observed at time x and from the circuits active at time y due to arrivals between x and y, since these are independent. For example, eqs. (28) and (34) give

$$M^T_{k,x}(y) = k\theta(x, y) + \textstyle\sum_n n\alpha_n(x, y), \tag{36}$$

where $M^T_{k,x}(y)$ is the total mean number of circuits observed at time y given k circuits are observed at time x [and assuming relationship (27)]. Also, eqs. (32) and (35) give the following approximation:

$$V^T_{k,x}(y) \approx k \frac{\sum \alpha_i(x) V_i(x, y)}{\sum i\alpha_i(x)} + \sum n^2 \alpha_n(x, y), \tag{37}$$

where $V^T_{k,x}(y)$ is the similarly defined variance.

3.6 Churn

We have previously defined churn as the minimum of the disconnect rate per circuit and the connect rate per circuit. Values of churn from other definitions are also easily obtained. We will here derive the churn, which happens to be a function of time in this case. To compute churn we need to know the probability measure for the individual order histories. Let $U_m(t, x)$ be the expected number of connects for an order of size m arriving at time t in the interval $[t, x]$ (thus $U_m(t, t) = m$). The expected total connect rate at time x, denoted $U(x)$, is then found to be:

$$U(x) = \frac{d}{dx}\left(\sum_m \int_{-\infty}^{x} \lambda(t) q_m(t) U_m(t, x) dt\right), \tag{38}$$

and similarly for the disconnects using the variable D,

$$D(x) = \frac{d}{dx}\left(\sum_m \int_{-\infty}^{x} \lambda(t) q_m(t) D_m(t, x) dt\right), \tag{39}$$

and thus we obtain the churn at time x, $\gamma(x)$:

$$\gamma(x) = \min\{D(x)/E[Y(x)],\ U(x)/E[Y(x)]\}, \tag{40}$$

where $E[Y(x)]$ is given by (16).

IV. THE MODEL FOR SPECIAL-SERVICE CIRCUIT ACTIVITY

Here, we assume that the demand rate grows exponentially and that the behavior of orders is not dependent on the time of arrival. Specifically, we assume,

$$\lambda(t) = \lambda_o e^{\beta t}, \tag{41}$$

$$q_m(t) = q_m, \tag{42}$$

and

$$P^*_{mn}(t, x) = P_{mn}(x - t). \tag{43}$$

Later we will assume a specific form for P_{mn}.

Assumptions (41) through (43) are equivalent to:

1. exponential growth in the rate of new orders at rate β (new orders occur as a nonhomogeneous Poisson process),
2. the probability that a new order is for m circuits is q_m, and
3. an order initially for m circuits requires a total of n circuits after z units of time with probability $P_{mn}(z)$. We will shortly further specify P_{mn} to represent unchanging orders of exponential lifetime.

We now explore the consequences of (41) to (43) in the analysis presented in Section III. Substituting into (15) we find that the number of orders of size n at time x is Poisson distributed with mean $\alpha_n(x)$, where

$$\alpha_n(x) = \alpha_n e^{\beta x}, \tag{44}$$

and

$$\alpha_n = \lambda_o \sum_m q_m \int_o^\infty e^{-\beta y} P_{mn}(y) dy. \tag{45}$$

(The total number of circuits required at any time has a compound Poisson distribution, see Appendix A.) Thus, the mean and variance of the number of circuits at time x, $Y(x)$, are growing exponentially at the same rate, and the ratio remains fixed:

$$E[Y(x)] = e^{\beta x} \sum_{n=1}^{\infty} n\alpha_n, \tag{46}$$

$$\text{var}[Y(x)] = e^{\beta x} \sum_{n=1}^{\infty} n^2\alpha_n, \tag{47}$$

or

$$\text{var}[Y(x)] = \nu E[Y(x)], \tag{48}$$

where

$$\nu = \frac{\sum_{n=1}^{\infty} n^2\alpha_n}{\sum_{n=1}^{\infty} n\alpha_n}. \tag{49}$$

Further results are possible if the behavior for orders over time is specified. We assume that the order size does not change over its lifetime, which has a common distribution with c.d.f. F independent of size. Later we will assume that F is an exponential distribution. Although in practice the number of circuits per order does change with time, it is conceivable that this movement is relatively unimportant; or even if important, that the general form of eqs. (2) and (3)

will hold, although the parameter μ may then have a different physical meaning than we will associate here. In the model suggested here, $P_{mo}(y) = F(y)$; $P_{mm}(y) = 1 - F(y)$; $P_{mn}(y) = 0$, $n \neq 0$, $n \neq m$. In this case we may compute α_n more explicitly. Substituting into (45), we obtain

$$\alpha_n = q_n \left(\frac{\lambda_o}{\beta} [1 - \tilde{F}(\beta)] \right), \tag{50}$$

where

$$\tilde{F}(\beta) = \int_o^\infty e^{-\beta y} dF(y). \tag{51}$$

When the lifetimes are exponentially distributed with mean $1/\mu$, i.e. $F(x) = 1 - e^{-\mu x}$,

$$\alpha_n = q_n \lambda_o (\mu + \beta)^{-1}. \tag{52}$$

Also, the batchiness ν is related to the order-size distribution; substitution into (49) yields

$$\nu = \frac{\sum n^2 q_n}{\sum n q_n}. \tag{53}$$

The assumption that the order size does not change throughout its lifetime also allows more explicit representation of the mean and variance of the future requirements for circuits. Our development here parallels that of Section III. We first compute the probability that an observed order will change size during the period of observation. Recall that $q_{mnl}(t, x, y)$ is the conditional probability that an order is of size l at time y given that it was of size n at time x and arrived as size m at time t. We easily obtain:

$$q_{mmm}(t, x, y) = \frac{\bar{F}(y - t)}{\bar{F}(x - t)}, \tag{54}$$

and

$$q_{mmo}(t, x, y) = 1 - q_{mmm}(t, x, y),$$

where

$$\bar{F}(x) = 1 - F(x),$$

and $q_{mml}(t, x, y) = 0$, $l \neq 0$, $l \neq m$. The value of q is irrelevant for $n \neq m$.

We next find the probability that an order of size n at time x becomes of size l at time y, which we denote $r_{nl}(x, y)$. Substitution into

(19) gives

$$r_{nl}(x, y) = 0, \qquad n \neq l, \qquad n \neq 0; \tag{55}$$

$$r_{nn}(x, y) = \bar{G}(y - x), \tag{56}$$

where

$$\bar{G}(t) = \frac{\int_o^\infty e^{-\beta z}\bar{F}(z + t)dz}{\int_o^\infty e^{-\beta z}\bar{F}(z)dz}; \tag{57}$$

and

$$r_{n0}(x, y) = 1 - \bar{G}(y - x).$$

Note that in the exponential-lifetime case, where $\bar{F}(z) = e^{-\mu z}$,

$$\bar{G}(y) = e^{-\mu y}. \tag{58}$$

We next find the mean and variance of the number of circuits in an order at time y, which was observed to be of size n at time x, denoted $M_n(x, y)$ and $V_n(x, y)$, respectively. Substitutions of (55) and (56) into (20) and (21) give:

$$M_n(x, y) = n\bar{G}(y - x), \qquad \text{and} \tag{59}$$

$$V_n(x, y) = n^2\bar{G}(y - x)[1 - \bar{G}(y - x)]. \tag{60}$$

Notice that the conditional means are proportional to the size of the order, i.e., (59) implies (27).

We now focus on the mean and variance of the number of circuits at time y due to orders which were observed at time x, given that k circuits were observed at time x. These quantities are denoted $M_{k,x}(y)$ and $V_{k,x}(y)$ respectively. Equation (28) gives

$$M_{k,x}(y) = k\bar{G}(y - x). \tag{61}$$

We also conclude that, given the approximation in (32),

$$V_{k,x}(y) \approx k\bar{G}(y - x)[1 - \bar{G}(y - x)]\frac{\sum n^2 q_n}{\sum n q_n}, \tag{62}$$

thus

$$V_{k,x} \approx \nu[1 - \bar{G}(y - x)]M_{k,x}(y), \tag{63}$$

where use has been made of (53) and (61).

We next find the expected number of orders of size n at time y that arrived during the interval (x, y) denoted $\alpha_n(x, y)$. Use of (3.3) yields

$$\alpha_n(x, y) = \lambda_o e^{\beta y} q_n \int_o^{y-x} e^{-\beta z} \bar{F}(z) dz, \tag{64}$$

where use is made of the fact $P^*_{mn}(t, x) = q_{mmn}(t, t, x)$, which follows from (54). When the lifetime distribution for orders is exponential, (64) becomes

$$\alpha_n(x, y) = \lambda_o e^{\beta y} q_n \left(\frac{1 - e^{-(\beta+\mu)(y-x)}}{\beta + \mu} \right). \tag{65}$$

The mean and variance of all circuits at time y due to orders arriving in the interval (x, y), $M^*(x, y)$ and $V^*(x, y)$, respectively, can be obtained by substitution of (64) into (34) and (35) yielding

$$M^*(x, y) = \lambda_o \sum n q_n e^{\beta y} \int_o^{y-x} e^{-\beta z} \bar{F}(z) dz, \tag{66}$$

and

$$V^*(x, y) = \nu M^*(x, y), \tag{67}$$

while for exponential lifetimes,

$$M^*(x, y) = \frac{D_o}{\beta + \mu} (e^{\beta t} - e^{-\mu t}), \tag{68}$$

where

$$D_o = \lambda_o e^{\beta x} \sum n q_n, \tag{69}$$

and

$$t = y - x.$$

Note that (61) and (68) [or (36)] give eq. (2), and (63) and (67) [or (37)] give (3), since the total number of active circuits at time y is the sum of the number of active circuits due to orders present at time x and the number of active circuits due to order arrivals between times x and y, and these random variables are independent. Equations (5) through (9) can easily be derived by the methods described in the paper, although we omit the details here.

Next, turning our attention to churn for the specific model of this section, we find that the expected number of connects in the interval $[t, x]$ for an order arriving at time t of size m, denoted $U_m(t, x)$, is given by

$$U_m(t, x) = m, \tag{70}$$

and similarly,

$$D_m(t, x) = mF(x - t), \tag{71}$$

where the variable D represents disconnects. The total connect rate, total disconnect rate, and churn rate at time x, $U(x)$, $D(x)$, and $\gamma(x)$, respectively, can be obtained from (38) through (40), yielding

$$U(x) = \lambda_o e^{\beta x} \sum m q_m, \tag{72}$$

$$D(x) = \lambda_o e^{\beta x} \sum m q_m \tilde{F}(\beta), \tag{73}$$

and

$$\gamma(x) = \gamma = \beta \tilde{F}(\beta)/(1 - \tilde{F}(\beta)), \qquad \beta \geq 0;$$

$$\gamma(x) = \gamma = \beta/(1 - \tilde{F}(\beta)), \qquad \beta < 0. \tag{74}$$

In the special case where lifetimes are exponentially distributed,

$$\gamma = \mu, \qquad \beta \geq 0;$$

$$\gamma = \mu + \beta, \qquad \beta < 0. \tag{75}$$

V. ESTIMATION OF THE PARAMETERS OF INTEREST

In this section, we describe the methodology that can be used to estimate the three key parameters of the model; β, the growth rate; μ, the disconnect rate; and ν, the batchiness.

5.1 Estimation of β

Suppose that we wish to estimate β on the basis of observed arrivals of orders, which by assumption occur according to a nonhomogeneous Poisson process with intensity $\lambda_o e^{\beta t}$. Suppose that the system is observed over the interval $[-\Theta, 0]$ and arrivals have been noted at times $t_1, \ldots, t_n$. We show how to obtain the maximum-likelihood estimator for β. (For a discussion of maximum-likelihood estimation, see any elementary book on statistics such as Mood & Graybill.)[5] The log-likelihood function, $\ln L(n, t_1, \ldots, t_n)$, is easily seen to be

$$\ln L(n, t_1, \ldots, t_n) = n \ln \lambda_o + \beta \sum_{i=1}^{n} t_i - \lambda_o \left(\frac{1 - e^{\beta\Theta}}{\beta} \right). \tag{76}$$

Differentiating with respect to λ_0 and β we find the necessary conditions for a maximum:

$$n/\lambda_o = \frac{1 - e^{-\beta\Theta}}{\beta}, \tag{77}$$

$$\sum_{i=1}^{n} t_i - \frac{\lambda_o \Theta}{\beta} e^{-\beta\Theta} + \lambda_o \left(\frac{1 - e^{-\beta\Theta}}{\beta^2} \right) = 0. \tag{78}$$

Using (77) to eliminate λ_o in (78), we obtain

$$S = \frac{xe^x - e^x + 1}{x(e^x - 1)} \equiv f(x), \tag{79}$$

where

$$S = \frac{\sum t_i}{n\Theta} + 1 \tag{80}$$

and

$$x = \beta\Theta. \tag{81}$$

The function f defined in (79) can be seen to be strictly monotonic with range between 0 and 1. Therefore, eq. (79) allows us to solve for $\beta\Theta$ as $f^{-1}(S)$, where S is the statistic defined in (80) equal to the proportion of the interval (*after* $-\ \Theta$) at which the average time of arrival occurs. Thus, the maximum-likelihood estimator for β, written $\hat{\beta}$, is given by

$$\hat{\beta} = \frac{f^{-1}(S)}{\Theta}. \tag{82}$$

The function f has the properties:

$$f(-\infty) = 0, f(0) = 1/2, f(\infty) = 1, \quad \text{and} \quad f(x) + f(-x) = 1.$$

Thus,

$$f^{-1}(0) = -\infty,$$
$$f^{-1}(1/2) = 0,$$
$$f^{-1}(1) = \infty,$$

and

$$f^{-1}(1/2 - x) = -f^{-1}(1/2 + x).$$

The function f^{-1} is tabulated in Table I.

For small x, $f(x)$ may readily be expanded in the power series:

$$f(x) = 1/2\left[1 + (1/6)x - \frac{1}{360}x^3 \cdots\right]$$

so that

$$f^{-1}(1/2 + y) = 12y + 28.8y^3 \cdots. \tag{83}$$

Similarly, a large f expansion yields

$$f^{-1}(1 - 1/y) \approx y - y^2 e^{-y}. \tag{84}$$

We may also determine the mean and variance of the statistic S given the correct parameter β and the number of observed arrivals. It is well known that the distribution of the arrival times for a nonhomogeneous Poisson process, conditioned on a given number of arrivals

Table I—Values of the function f^{-1} useful in estimating the growth rate β, and several approximations for the function [see eq. (79) and following]

X	$f^{-1}(X)$	12X−6	Approximation in (83)	Approximation in (84)
0.50	0.0000	0.0000	0.0000	
0.52	0.2402	0.2400	0.2402	
0.54	0.4819	0.4800	0.4818	
0.56	0.7263	0.7200	0.7262	
0.58	0.9751	0.9600	0.9747	
0.60	1.2299	1.2000	1.2288	
0.62	1.4926	1.4400	1.4898	
0.64	1.7654	1.6800	1.7590	
0.66	2.0507	1.9200	2.0380	
0.68	2.3517		2.3280	
0.70	2.6721		2.6308	
0.72	3.0168		2.9467	
0.74	3.3920			
0.76	3.8060			
0.78	4.2703			
0.80	4.8010			4.8316
0.82	5.4219			5.4362
0.84	6.1691			6.1746
0.86	7.1010			7.1025
0.88	8.3164			8.3166
0.90	9.9954			9.9955
0.92	12.4994			12.4994
0.94	16.6667			16.6667
0.96	25.0000			25.0000
0.98	50.0000			50.0000
1.00	∞			∞

in the interval, is the same as the order statistics from n i.i.d. random variables with probability density proportional to the arrival rate. Thus S has the distribution of the average of n i.i.d. random variables, Y_i on [0, 1] with density $g(\rho)$, where

$$g(\rho) = \frac{xe^{\rho x}}{e^x - 1}$$

and

$$x = \beta\Theta.$$

It is easily seen that

$$E(Y) = f(x) \tag{85}$$

and

$$\text{var}(Y) = \frac{1}{x^2} - \frac{1}{e^x + e^{-x} - 2}. \tag{86}$$

Equation (86) is valid if $x \neq 0$; when $x = 0$, $\text{var}(Y) = 1/12$, the limit

of (86) as x goes to 0. The expression for the variance is symmetric in x and takes its maximum value at $x = 0$.

Thus, for a given value of the growth rate β, and a (large) given number of observations n, the observed statistic S is approximately normally distributed with mean $f(x)$ and variance less than or equal to $1/12n$. This observation can be readily translated into confidence intervals through the use of elementary statistical theory. For example, a 95-percent confidence interval for $f(x)$ (assuming normality of the statistic) is

$$S - 1.96\sqrt{\frac{1}{12n}} \leq f(x) \leq S + 1.96\sqrt{\frac{1}{12n}}, \tag{87}$$

which translates to

$$f^{-1}\left(S - 1.96\sqrt{\frac{1}{12n}}\right) \leq \beta\Theta \leq f^{-1}\left(S + 1.96\sqrt{\frac{1}{12n}}\right). \tag{88}$$

If S is close to 0.5, then we can use $f^{-1}(x) \approx 12x - 6$ [see (83)] to obtain for the 95-percent confidence interval for $\beta\Theta$

$$\beta\Theta = 12S - 6 \pm \frac{6.79}{\sqrt{n}}. \tag{89}$$

5.2 Estimation of ν

There are several possible statistics for the measurement of the batchiness ν. We shall take as our starting point eq. (49) which defines the batchiness in terms of the distribution of the order size at (any) point in time. This is preferable and is more robust than using the distribution of the order size on arrival, although the two happen to equal when order sizes do not change with time. Thus, if ν is to be estimated based on observation of the system at a given point in time at which i_n orders of size n are observed, then a reasonable estimator for ν, which we write $\hat{\nu}$, is:

$$\hat{\nu} = \frac{\sum n^2 i_n}{\sum n i_n}. \tag{90}$$

When the number of circuits does change during the lifetime of an order, it is possible that the form of (3) still holds. In this case, it is likely that the parameter ν which multiplies each of the two terms in (3) is different. Equation (90) is a reasonable estimator for the multiplier of the second term. The multiplier of the first term should be estimated by other methods.

Table II—Notation

D_o	Present ($t = 0$) circuit demand rate due to new orders.
$D(x)$	Expected total disconnect rate at time x.
$D_m(t, x)$	Expected number of disconnects in the interval $[t, x]$ for an order of size m arriving at time t.
$i_n(x)$	Observed number of orders of size n at time x.
J_i	Conditional number of orders of size i due to orders observed at time x given that k circuits were observed at time x.
$M(x, y)$	Mean total number of circuits at time y due to orders.
$M^*(x, y)$	Mean of the total number of circuits at time y due to orders arriving between x and y.
$M^T_{k,x}(y)$	Mean of the number of circuits at time y given k are observed at time x.
$M_{k,x}(y)$	Conditional expectation of the number of circuits at time y due to orders observed at time x given that k circuits are observed at time x.
$M^{(2)}_{k,x}(y)$	Conditional second moment of the number of circuits at time y due to orders observed.
$M_n(x, y)$	Mean order size at time y for an order observed to be of size n at time x.
$P_{mn}(t)$	Probability that an order of initial size m becomes of size n, t time units after arrival.
$P^*_{mn}(t, x)$	Probability that an order arriving at time t of initial size m becomes of size n at time x.
q_m	$q_m(t)$ when there is no dependence on t.
$q_m(t)$	Probability that an order at time t is initially for m circuits.
$q_{mnl}(t, x, y)$	Conditional probability that an order arriving as size m at time t and of size n at time x becomes of size l at time y.
$r_{nl}(x, y)$	Probability that an order of size n at time x becomes an order of size l at time y.
$U(x)$	Expected total connect rate at time x.
$U_m(t, x)$	Expected number of connects for an order of size m arriving at time t in the interval $[t, x]$.
$V(x, y)$	Variance of the total number of circuits at time y due to orders observed active at time x.
$V^*(x, y)$	Variance of the total number of circuits at time y due to orders arriving between x and y.
$V^T_{k,x}(y)$	Variance of the number of circuits at time y given that k are observed at time x.
$V_{k,x}(y)$	Conditional variance of the total number of circuits at time y due to orders observed at x given that k circuits are observed at time x.
$V_n(x, y)$	Variance of order size at time y for an order observed to be of size n at time x.
$Y(x)$	Total number of active special service circuits at time x.
α_n	Constant of proportionality for the exponential growth of $\alpha_n(x)$.
$\alpha_n(x)$	Mean number of orders of size n at time x.
$\alpha_n(x, y)$	The number of orders of size n at time y which arrived between times x and y.
β	Growth rate.
γ	Churn.
$\delta_{k,x}(j_1, j_2, \ldots)$	Conditional probability that there are j_1 orders of size 1, j_2 orders of size 2, etc., at time x given that k total circuits are observed at time x.
$\theta(x, y)$	Defined in (27).
Θ	Length of observation period.
λ_o	Present ($t = 0$) arrival rate of orders.
$\lambda(t)$	Instantaneous arrival rate of orders at time t.
μ	Disconnect rate (per circuit).
ν	Batchiness.
$\rho_{mn}(t, x)$	The conditional probability density that an order arrived at time t of initial size m given that it is of size n at time x.

5.3 Estimation of μ

The estimation of μ is relatively straightforward. The maximum-likelihood estimator is given in (14) and further details including estimated values by service family are given in Reed and Smith.[2]

5.4 Estimation of D_o

Equation (77) allows the MLE estimator of λ_o,

$$\hat{\lambda}_o = \frac{n\hat{\beta}}{1 - e^{-\hat{\beta}\Theta}}, \tag{91}$$

where the estimator $\hat{\beta}$ has been previously described in (82). The estimator of D_o (the instantaneous demand at the end of the observation interval of length Θ), $\hat{D}_o$ then is

$$\hat{D}_o = \lambda_o \hat{N}, \tag{92}$$

where $\hat{N}$ is an estimate of the average batch size. The previous expectation can be estimated from the order sizes at arrival epochs, or more crudely from the general distribution of order sizes at an arbitrary point of time.

Interestingly enough, $D_o/(\mu + \beta)$ can be estimated solely from the number of active circuits at a point of time. For simplicity, we assume that the orders are solely of size one, although the analysis could be repeated for other distributions. In this case, the following analysis is applicable.

Suppose that the number of active circuits at time t is a Poisson random variable with mean $\lambda e^{\beta t}$. The time that the mean is between x and $x + dx$ is $dx/x\beta$. The expected time that the mean is between x and $x + dx$ and a total of k active circuits are observed is $\frac{dx}{x\beta}\frac{x^k}{k!}e^{-x}$. Thus, if k active circuits are observed, the conditional distribution of the mean (in our case this is $D_o/(\mu + \beta)$) has density proportional to $x^{k-1}e^{-x}$, or is a standard gamma random variable with k degrees of freedom. This random variable has mean and variance equal to k. Thus, if k circuits are observed, the conditional distribution of $D_o/(\mu + \beta)$ has mean k and variance k, and is approximately normally distributed if k is large. This information can be used to modify eqs. (2) and (3), to take into account the variability of $D_o/(\mu + \beta)$ to obtain:

$$M_k(t) = ke^{\beta t}, \tag{93}$$

and

$$V_k(t) = \nu k[e^{\beta t} - e^{-2\mu t}] + k[e^{\beta t} - e^{-\mu t}]^2, \tag{94}$$

where the subscript k on the variables on the left-hand side indicates

conditional means and variances knowing β, μ, and ν but not knowing D_o. Note that the variance-to-mean ratio is unbounded for increasing t, since errors in estimation of D_o accumulate indefinitely.

VI. SUMMARY

We have described a model for special-services activity useful in forecasting special-services requirements. It requires three physical characterizations of the process (growth rate, disconnect rate, and batchiness) and two instantaneous measurements (the current number of active circuits and the instantaneous rate of new connects). We give means and variances for the numbers of active circuits at a given point in the future and for the total number of connects or disconnects during a future period. The distribution of these variables can be computed by the methodology described in the paper. We also describe general techniques for estimation of the required parameters.

Work is being undertaken to verify and calibrate the model with the New Jersey Bell Telephone Co. database and will be reported elsewhere.

VI. ACKNOWLEDGMENTS

I would like to thank Moshe Segal and Werner Stahel for many helpful suggestions.

REFERENCES

1. R. Nucho, "Transient Behavior of the Kendall Birth-Death Process—Applications to Capacity Expansion for Special Services," B.S.T.J. *60,* No. 1 (January 1981), pp. 57–87.
2. F. F. Reed and D. R. Smith, unpublished work.
3. S. Ross, *Applied Probability Models with Optimization Applications*, San Francisco: Holden-Day, 1970.
4. S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, second edition, New York: Academic Press, 1975.
5. A. Mood and F. Graybill, *Introduction to the Theory of Statistics*, second edition, New York: McGraw-Hill, 1963.

APPENDIX A

The Compound Poisson Random Variable

Briefly, a random variable is said to be a compound Poisson random variable if it can be thought of as the sum of a Poisson number of independent identically distributed positive integer-valued random variables. Thus Y is a compound Poisson random variable if

$$Y = \sum_{i=1}^{N} X_i,$$

where N is a Poisson random variable with mean α, N and the X_i are

independent and

$$P\{X_i = n\} = p_n.$$

We have rather easily

$$E(Y) = \alpha E(X) = \alpha \sum np_n$$

$$\text{Var}(Y) = \alpha E(X^2) = \alpha \sum n^2 p_n.$$

An alternative (and equivalent) way of specifying a compound Poisson random variable is

$$Y = \sum_{i=1}^{\infty} iZ_i,$$

where

Z_i are independent Poisson random variables with $E(Z_i) = \alpha p_i \equiv \alpha_i$. In this case it is convenient to think of Z_i as the number of batches or orders of size n that are aggregated to give the total number denoted Y.

APPENDIX B

The Nonhomogeneous Poisson Process

A process which counts events is a nonhomogeneous Poisson process (see, for example, Ross, Ref. 5) with intensity $\lambda(t) \geq 0$ if the number of events in the interval $[x, y]$ is a Poisson random variable with mean $\int_x^y \lambda(t)dt$, and the number of events in disjoint intervals are independent.

Fact: If events from a nonhomogeneous Poisson process are of two types, and an event at time t is of Type 1 with probability $p(t)$, then the process which counts Type 1 events is a nonhomogeneous Poisson process with intensity $\lambda(t)p(t)$, and it is independent of the counting process which counts Type 2 events [which is a nonhomogeneous Poisson process intensity $\lambda(t)(1 - p(t))$].

Fact: Suppose that a nonhomogeneous Poisson process is observed over the interval $[x, y]$ and n events are observed. If the times of these events are arranged in random order, their distribution is identical to that of n independent identically distributed random variables whose density at t is

$$\frac{\lambda(t)}{\int_x^y \lambda(z)dZ}$$

if $t \in [x, y]$ and is zero otherwise.

AUTHOR

Donald R. Smith, A.B. (Physics), 1969, Cornell University; M.S. (Operations Research), 1974, Columbia University; Ph.D. (Operations Research), 1975, University of California, Berkeley; Bell Laboratories, 1980—. Before joining Bell Laboratories, Mr. Smith was employed at Adaptive Technology, Inc., 1970–1974, and was Assistant Professor in the Department of Industrial Engineering and Operations Research, Columbia University, 1975–1979. At Adaptive Technology, Mr. Smith developed mathematical models for new techniques in statistical multiplexing. At Bell Laboratories he is in the Operations Research Department working on the application of stochastic processes to special service and DACS modeling.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

TELBECC—A Computational Method and Computer Program for Analyzing Telephone Building Energy Consumption and Control

By P. B. GRIMADO*

(Manuscript received February 2, 1983)

Telephone Building Energy Consumption and Control (TELBECC) program has been developed to accurately and efficiently analyze environmental control and energy use in telephone company buildings. The program simulates various operational plans to determine the relative energy and cost savings. By analyzing the operation of the heating, ventilation, and air conditioning system as it regulates a changing environment, TELBECC calculates the heating and cooling load, dry-bulb temperature, and relative humidity in the building. The user specifies the building's dry-bulb temperature limits, which are the control variables for the program analysis. The simplified computational procedure of the program incorporates a recursive scheme using time series to perform the necessary calculations. The results of the computations can be obtained for different periods: the quarter hour, hour, day, or month. Energy consumption and control in several equipment buildings located in three different geographical areas have been analyzed by TELBECC. Analysis and comparison of the resulting data demonstrate the advantages of the program.

I. INTRODUCTION

An ambitious energy-cost savings program has been instituted to reduce energy use in telephone company buildings. In recent years, telephone companies have saved energy mainly by redesigning and

* Bell Laboratories.

retrofitting buildings to operate and maintain environmental control equipment at peak performance, to turn out unneeded lights, to reduce heating and cooling losses, and the like. Further energy and cost savings, although requiring additional capital investment, could be achieved through modification of environmental control systems, purchase of sophisticated microprocessors for more efficient control of building Heating, Ventilating, and Air conditioning (HVAC) systems, and installation of alternative energy sources, such as solar power and wind power. Before adopting any conservation plans that require appreciable capital investment, however, we should make a thorough economic evaluation. Such an evaluation can be carried out by correlating the changing operating characteristics of a building with selected energy conservation plans. This procedure would enable us to pinpoint the most economical operating strategies.

There are several commercially available computer programs to perform this type of analysis, such as DOE II, ESP, and BLAST;[1,2] most, however, are proprietary. These complex programs can analyze any of a broad spectrum of commercial, industrial, and residential buildings. But, because of their versatility, they require large computer systems, extensive data preparation, and high costs. The use of energy in the majority of telephone company equipment buildings, which are small, single-story structures varying in area from 1500 to 10,000 square feet, can be best evaluated by a more focussed computer program.

This paper describes a new computational method and computer program called Telephone Building Energy Consumption and Control (TELBECC). This program simulates building operations and quickly evaluates numerous energy conservation plans and cost-saving strategies under variable weather conditions [according to standard hourly Test Reference Year (TRY) weather data[3]]. The program can evaluate energy consumption for intermittent or proportional HVAC plant operation, economy cycle operation, enthalpy cycle operation, and wideband temperature operation with no heating or cooling between preset room temperature limits. Also, the program can calculate the optimum building orientation and U factor (heat transmission characteristics) of the outside walls and roof, chiller and heater plant size, dry-bulb temperature and Relative Humidity (RH) variations, and quantity of water required to maintain 20-percent minimum RH during economy cycle operations.

II. PROGRAM DESCRIPTION

We can derive the heating or cooling load in an enclosed building from the following considerations:

1. Conduction of heat through the building walls and roof.

2. Permeation of outside air through the building envelope.
3. Internal heat generation from equipment, lights, and people.
4. Direct solar radiation through windows and skylights (fenestration).

Since most operating company equipment buildings have few windows, the program does not consider item 4 in the present version.

A Constant-Air-Volume (CAV) supply fan system typically controls the air temperature of a building. The building engineer normally sizes the fan system using the elementary steady-state heat balance, which takes into account the internal heat loads, outside air temperatures, and solar radiation in conjunction with the U factor of the building envelope. In general, this conservative approach produces fan systems that are oversized and therefore inefficient. To find a smaller, and perhaps more efficient, fan capacity, a TELBECC user selects different fan capacities for analysis by the program. The program generates data on the space temperatures, relative humidity, peak heating and chiller loads, and the hours of system operation for the different fan capacities that can be used to find an optimum air supply fan system.

For comfort, a limit is imposed on the difference between HVAC supply and return air temperatures. For cooling, this temperature difference is −20 degrees F; for heating, +40 degrees F. These default values may be overridden by the user. With an environmental dry-bulb temperature standard specified, the program computes the required operation of the HVAC.

The user can specify one of two basic ways to operate the HVAC: intermittent operation or proportional control. With intermittent operation, the HVAC does not supply any heating or cooling when the dry-bulb air temperature is within the wideband temperature range. Reaching or exceeding either wideband temperature limit activates the HVAC. The HVAC stays on and does not deactivate until the dry-bulb air temperature reaches 3 degrees F above the lower limit of the wideband temperature range for heating and 3 degrees F below the upper limit of the wideband temperature range for cooling. The TELBECC user can reset the numerical values of the throttling range if a different range is appropriate. The proportional control plan operates by continuously adjusting the supply and return air temperature difference in increments of 1 degree F to satisfy the instantaneous building heating or cooling load. This plan follows the building load to closely track the lower and upper limits of the wideband temperature range with essentially no throttling. When selecting a dual or extended wideband temperature standard (that is, one with different wideband limits for occupied and unoccupied times), the HVAC activates before occupancy in order to reach the preset temperature standard.

TELBECC calculates the heat added or removed by the HVAC

system in controlling the dry-bulb air temperature every quarter hour. In particular, when cooling is required, the sensible and latent loads on the chiller plant are simultaneously computed by incorporating any of three standard methods of fan system operation:

1. Conventional operation, which is chiller operation with no economizer.
2. Chiller operation with a dry-bulb economy cycle.
3. Enthalpy cycle.

In conventional operation, the minimum quantity of outside air needed for ventilation is circulated. This mode is also used as a benchmark for the program. The dry-bulb economy cycle uses outside air for cooling whenever the outside dry-bulb air temperature falls below the maximum value. The default value is 55 degrees F, but the user can reset the value. The enthalpy cycle checks the enthalpy of the inside air and the outside air. If the outside air enthalpy is lower, 100 percent outside air is circulated to reduce the load on the chiller regardless of the relative humidity. Otherwise, only the minimum quantity of outside air required for ventilation is circulated.

System control is based on dry-bulb air temperature and is not predicated on maintaining a particular value of relative humidity. Nevertheless, the program computes changes in relative humidity for the three methods of fan-system operation discussed above. The program summarizes the variation in relative humidity for the time period chosen by giving the number of hours the relative humidity is less than 10 percent, between 10 and 15 percent, between 15 and 20 percent, between 20 and 55 percent, between 55 and 60 percent, and greater than 60 percent.

In addition, since dry-bulb economy cycle operation generally calls for bringing in winter air with low humidity, the program calculates the quantity of water required for humidification. The operating company minimum standard of 20 percent RH in the inside air for dry-bulb economy cycle operation in winter is the basis for calculating the amount of water added to the air.

III. TRANSIENT HEAT CONDUCTION THROUGH THE BUILDING ENVELOPE

Weather conditions influence the heating and cooling load of a building by heat conduction through the structural and decorative materials of the exterior walls and roof, as well as by permeation of outside air and direct absorption of solar radiation through window areas. Since, as previously mentioned, most operating company equipment buildings have few windows, only heat conduction and permeation are treated in the computer program. The program must account for the heat storage effects of the structure, as well as the daily and

seasonal variation of the outside air temperature and solar radiation. We can account for these influences on the building by considering the building envelope elements as one-dimensional flat slabs or plates. We then obtain a solution to a partial differential equation with time-dependent boundary conditions. A classical analytical solution of this equation[4] produces a set of equations that require an inordinate quantity of computational effort and time, rendering the whole idea of performing the analysis impractical and uneconomical. However, the analytical solution can be recast into a simpler, though effective, computational scheme with a method first introduced by Mitalas and Stephenson,[5] which is ideally suited to calculation by computer.

The inside-wall and roof-surface temperature $T_{BE}(t)$ and the air temperature of the building $T_a(t)$, which are dependent on time, determine the environmental load due to convection.* $T_{BE}(t)$ is represented in the form of the following time series:

$$T_{BE_t} = -\sum_{i=1}^{m-1} b_i T_{BE_{t-i}} + \sum_{i=1}^{m} a_i T_{O_{t-i-1}} + \sum_{i=1}^{m} a_i' T_{a_{t-i-1}}, \qquad (1)$$

where

t = current time,
T_{BE_t} = inside-wall temperature of the building at the current time,
$T_{BE_{t-i}}$ = inside-wall temperature of the building i time units prior to t,
$T_{O_{t-i}}$ = outside sol-air temperature[6] i time units prior to t, and
$T_{a_{t-i}}$ = air temperture of the building i time units prior to t.

For hourly temperature calculations, the number of terms, m, will rarely exceed 5, and for quarter-hour calculations, m will generally be less than 15. The recursive properties of the calculation make it extremely efficient and economical, especially when the operating characteristics of the building may need to be tracked for an entire calendar year. The coefficients b_i, a_i, and a_i' in eq. (1) are determined from the thermophysical properties of the structure. Only six values are needed to uniquely specify these coefficients: wall thickness, wall U factor, wall-weight density, effective heat-transfer coefficient of the inside- and outside-wall surface, and the time interval between successive calculations. The appendix presents the mathematical procedure to evaluate these coefficients. Figures 1 and 2 show the mathematical and physical models for deriving the coefficients.

We can validate this simplified computational approach by compar-

* Radiative interchange between inside building-wall surfaces is not included in the present version of the program.

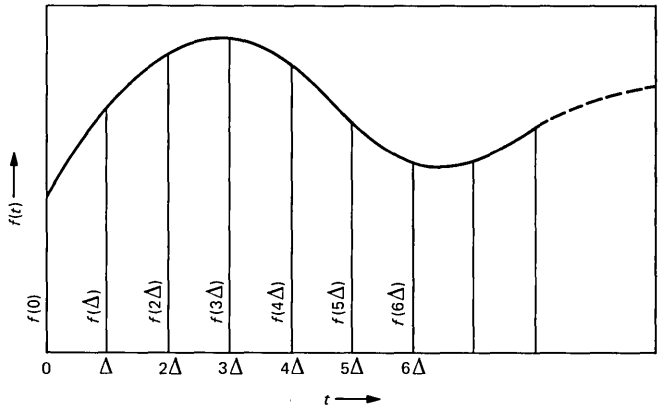


Fig. 1—Discrete and continuous functions.

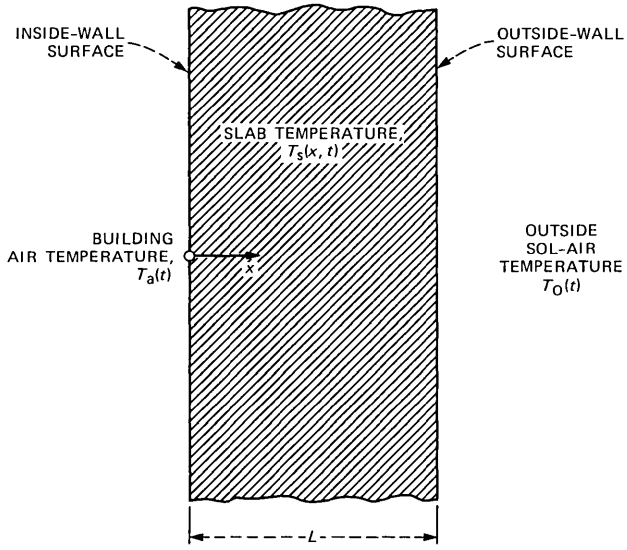


Fig. 2—Homogeneous flat slab.

ing it with an exact solution given in the literature.[7] We can see this in Figs. 3 and 4 for several values of the inside-wall dimensionless convective heat-transfer parameter $B_i \equiv h_i L/k$ and two limiting values of the outside-wall convective heat-transfer parameter $B_O \equiv h_O L/k$. In Fig. 3 the outside-wall convective heat transfer parameter $B_O = 0$; i.e., the surface $x = L$ is insulated. In Fig. 4 the solution corresponds to B_O approaching ∞; i.e., the surface $x = L$ is maintained at a constant temperature. The initial and boundary conditions are indicated in the figures. The solid curves represent the exact condition given in Ref. 7,

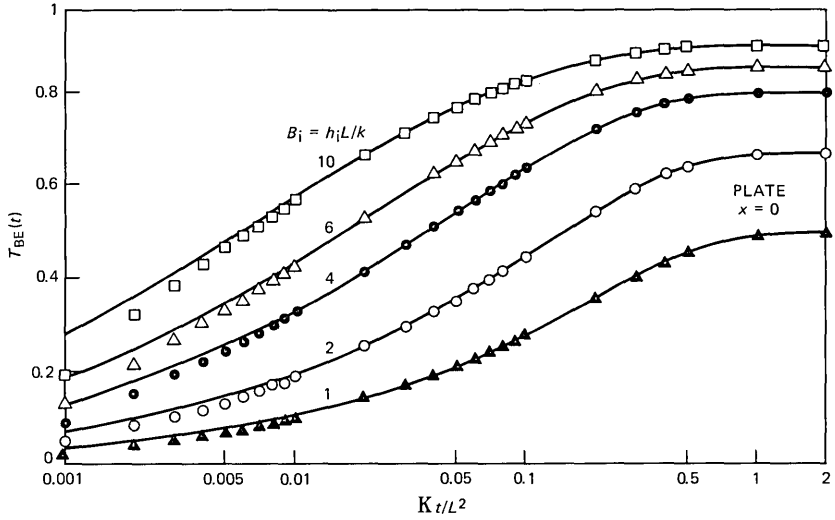


Fig. 3—Temperature response of front face of plate, $0 \leq x \leq L$, with back face $0 \leq x \leq L$, with back face $x = L$ maintained at $T_0 = 0$ degree F after sudden exposure to uniform-temperature convective environment $T_a = 1$ degree F at $x = 0$. Sampling interval Δ is one min.

and the dots represent the numerical values computed by the time series in eq. (1). The sampling interval for this example is $\Delta = 1$ min. or, in terms of dimensionless time, $k\Delta/L^2 = 0.001$. We can see that after some time has elapsed the exact solution and the time-series solution match identically. For the problem considered here at $t = 0$, the ambient convected temperature T_a is suddenly raised from $T_a = 0$ to $T_a = 1$; i.e., the boundary condition is a step function. However, since the development of the time series assumes a linear variation between time intervals, as stated in the appendix, the solution resembles an initial ramp followed by a constant value, as shown in Fig. 5. Once the effect of the initial ramp input diminishes after about 10 sampling intervals, the solution coincides with the exact solution. This characteristic of the time series is not a problem here, since instantaneous changes of air temperatures inside and outside the building do not occur.

IV. CALCULATION OF BUILDING AIR TEMPERATURE AND ENERGY USE

The air temperature of the building is obtained through the following equation for the heat balance within the building:

$$q_{air}(t) = q_{equipt}(t) + q_{lights}(t) + q_{people}(t) + q_{infiltration}(t) + q_{walls}(t) + q_{HVAC}(t), \quad (2)$$

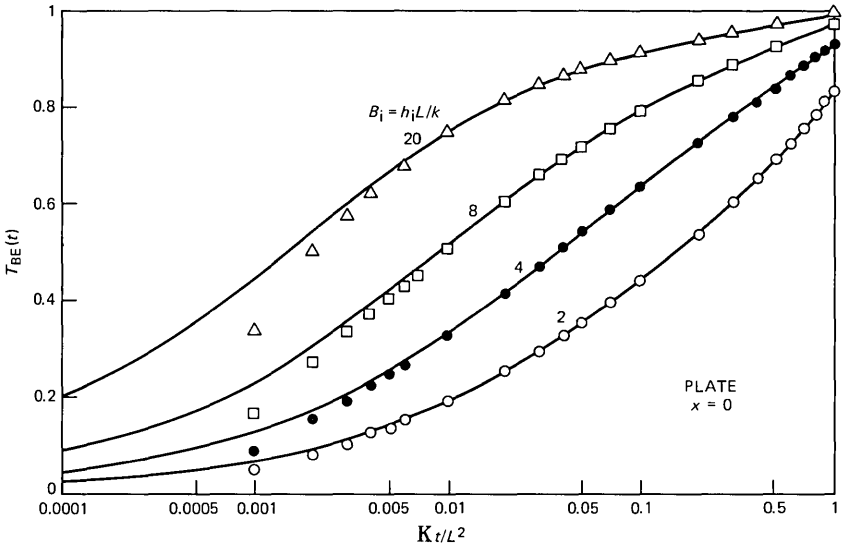


Fig. 4—Temperature response of front face of plate $0 \leq x \leq L$, with insulated back face $x = L$ after sudden exposure to uniform-temperature convective environment T_a = 1 degree F at $x = 0$. Sampling interval Δ in one min.

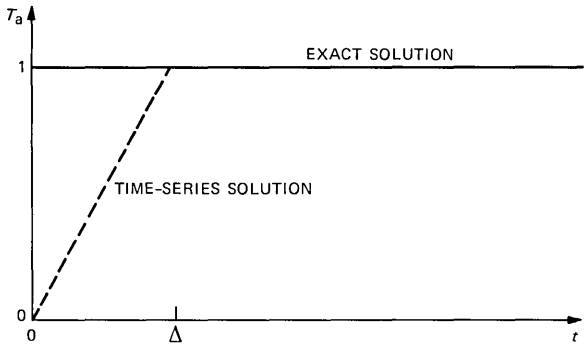


Fig. 5—Convection temperature environment at inside boundary surface $x = 0$.

where $q_{air}(t)$ (Btu/hr) represents the sensible thermal-energy convection rate of the inside air, resulting in a change in the overall dry-bulb air temperature. The other terms on the right represent the rate at which heat is convected to the air from the following: equipment heat dissipation; lighting; people; inadvertent infiltration of outside air; the exterior walls, floor, and roof of the building; and the building's HVAC control system.

To evaluate the air temperature $T_a(t)$ from eq. (2), each term is expressed as the difference in temperatures between the air and the

heat-convecting medium. We can determine $T_a(t)$ from the following differential equation:

$$\rho c v \frac{dT_a(t)}{dt} = H_{EL}A_{EL}[T_{EL}(t) - T_a(t)] + \rho cm[T_O(t) - T_a(t)] + H_{BE}A_{BE}[T_{BE}(t) - T_a(t)] + \rho cQ[DT_{SP}(t)], \quad (3)$$

where

$T_a(t)$ = average building dry-bulb air temperature;
$T_{EL}(t)$ = combined average temperature of the equipment, lights, and people;
$T_O(t)$ = outside dry-bulb air temperature;
$T_{BE}(t)$ = inside surface temperature of the building envelope;
$DT_{SP}(t)$ = difference between the air-supply temperature and the air temperature of the building;
ρ = air density;
c = air specific heat;
v = volume of the building space;
H_{EL} = heat-transfer coefficient between the air and equipment;
A_{EL} = average surface area of equipment;
m = rate at which outside air infiltrates the building;
H_{BE} = heat-transfer coefficient between the building envelope and air;
A_{BE} = surface area of the building envelope; and
Q = air-supply rate of the HVAC system fan units.

Only one of the temperatures in eq. (3) is presumed known: the outside-air dry-bulb temperature, $T_O(t)$. The other four temperatures—$T_a(t)$, $T_{EL}(t)$, $T_{BE}(t)$, and $DT_{SP}(t)$—are coupled; therefore, additional equations are needed for their resolution.

We can consolidate our terms into two other equations. When we combined the heat gain generated by the equipment, lighting, and people into a single term for heat dissipation per unit of building floor area $W(t)(W/ft^2)$, one additional equation can be written as

$$C_{EL}\frac{dT_{EL}(t)}{dt} = 3.41\ A_f W(t) + H_{EL}A_{EL}[T_a(t) - T_{EL}(t)], \quad (4)$$

where

C_{EL} = heat capacity of the equipment, lights, and people;
A_f = building floor area; and
$W(t)$ = combined heat dissipation of the equipment, lights, and people per unit of building floor area.

A third equation coupling the building envelope temperature $T_{BE}(t)$ and the building air temperature $T_a(t)$ is needed. A likely equation

would be the time series given by eq. (1). But before this can be applied, both eqs. (3) and (4) must also be recast in the form of time series. This is easily done by using the properties of the z transform and the procedure already delineated in the appendix. The time-series solution of eqs. (3) and (4) assumes the form

$$T_{a_t} = sT_{a_{t-1}} + \left(\frac{\Delta\tau + s - 1}{\Delta t^2}\right)[\tau_{EL}T_{EL} + \tau_O T_O + \tau_{BE}T_{BE} + \tau_{SP}DT_{SP}]_t$$

$$+ \left(\frac{1 - s(\Delta\tau + 1)}{\Delta\tau^2}\right)[\tau_{EL}T_{EL} + \tau_O T_O + \tau_{BE}T_{BE} + \tau_{SP}DT_{SP}]_{t-1}, \quad (5)$$

and

$$T_{EL_t} = \hat{s}T_{EL_{t-1}} + \frac{(\Delta\hat{\tau} + \hat{s} - 1)}{\Delta\hat{\tau}^2}[\hat{\tau}_w W + \hat{\tau}_a T_a]_t$$

$$+ \frac{[1 - \hat{s}(\Delta\hat{\tau} + 1)]}{\Delta\hat{\tau}^2}[\hat{\tau}_w W + \hat{\tau}_a T_a]_{t-1}, \quad (6)$$

where

Δ = sampling time interval,
$\tau_{EL} = H_{EL}A_{EL}/\rho cv$,
$\tau_O = m/v$,
$\tau_{BE} = H_{BE}A_{BE}/\rho cv$,
$\tau_{SP} = Q/v$,
$\tau = \tau_{EL} + \tau_O + \tau_{BE} + \tau_{SP}$,
$s = \text{EXP}(-\Delta\tau)$,
$\hat{\tau}_w = A_f/C_{EL}$,
$\hat{\tau}_a = H_{EL}A_{EL}/C_{EL}$,
$\hat{\tau} = \hat{\tau}_w + \hat{t}_a$, and
$\hat{s} = \text{EXP}(-\Delta\hat{\tau})$,

and the subscripts t and $t-1$ indicate that the temperature is evaluated at times $t = n\Delta$ and $t = (n-1)\Delta$.

The calculation of the humidity ratio in the space is also formulated in terms of a first-order linear differential equation in time similar to eqs. (3) and (4). This equation is also recast in the form of a time series [see eqs. (5) and (6)]. Knowing the humidity ratio and the dry-bulb air temperature, we can find the relative humidity by employing standard psychometric formulae.[6]

As previously noted in Section II, the system of eqs. (1), (5), and (6) permits controlling the dry-bulb air temperature in two basically different ways, intermittent operation and proportional control, and

determines the hours of operation. We use these equations for intermittent HVAC operation where we assume that the HVAC system holds the supply and return air-temperature difference constant at $DT_{SP} = -20$ degrees F whenever cooling is required and at $DT_{SP} = 40$ degrees F whenever heating is required. The hours of operation, and hence the total quantity of heat removed or added to satisfy the imposed dry-bulb temperature standard, can then be determined. For the proportional control plan, eqs. (1), (5), and (6) are also used to calculate not only the hours of HVAC operation but also the numerical value of the supply and return air-temperature difference $DT_{SP}(t)$, which in general varies continuously for this mode of control. The variation in time of the numerical value of $DT_{SP}(t)$ is determined by just satisfying the instantaneous building environmental load. When the HVAC system is activated, the proportional control plan closely follows the lower limit (for heating) or the upper limit (for cooling) of the building wideband temperature range.

Once $DT_{SP}(t)$ and the hours of operation are known, the program calculates the heat added or removed from the building by the HVAC system during every quarter hour and for whatever other period is of interest, e.g., monthly. As a corollary, we can estimate the environmental control system energy use, assuming the following HVAC system characteristics:

1. For chiller operation, a constant Coefficient Of Performance (COP) supplied by the user, together with the quantity of heat removed from the building air, characterizes its energy requirements.

Table I—Equipment buildings analyzed

Energy Consumption and Control				
	Wideband Temperature (°F)			
Case	Occupied Times	Unoccupied Times	Control	Geographic Location
1	65–80	65–80	Intermittent	New York City
2	65–80	65–80	Proportional	New York City
3	65–80	60–85	Intermittent	New York City
4	65–80	65–80	Intermittent	New Orleans
5	65–80	65–80	Intermittent	Phoenix

Building Parameters	
Factor	Parameter
Size (L × W × H)	60 ft × 40 ft × 13 ft
Average heat transmission	U = 0.25 Btu/hr − ft^2
Occupancy time	8 a.m. to 6 p.m.
Fan support rate	7400 CFM
Ventilation capacity	150 CFM
Static fan pressure	2 in. of water
Internal heat load	15W/ft^2
Economy cycle temperature limit	≤65 degrees F

Table IIa—Intermittent control of building space air temperature—New York City, case 1 in Table I

Month	Degree-Days		Space Temp		Max Load (tons)		Total Load (MBtu)		Number of Hours			Heating (kWh Elect)	Cooling (kWh)		Water (gal) to Maintain 20% Min RH
	Heat	Cool	Min	Max	Heat	Cool	Heat	Cool	Heat	Cool	Econ		No Econ	Econ	
Jan	858	0	76.1	80.5	0.0	13.3	0.0	49.2	0	319	314	0.0	5291.2	1229.3	1352.7
Feb	797	0	76.3	80.5	0.0	13.3	0.0	44.2	0	256	281	0.0	4754.4	1114.7	1143.5
Mar	703	0	76.5	80.5	0.0	13.3	0.0	56.1	0	361	356	0.0	6021.7	1391.1	1026.2
Apr	358	0	76.6	80.6	0.0	13.6	0.0	65.1	0	414	334	0.0	6967.9	2582.9	321.7
May	118	38	76.6	80.6	0.0	13.9	0.0	77.5	0	486	179	0.0	8271.2	5901.2	101.2
Jun	23	147	76.6	80.9	0.0	14.2	0.0	83.1	0	512	36	0.0	8833.2	8348.4	0.0
Jul	0	325	76.7	80.8	0.0	14.2	0.0	92.6	0	563	0	0.0	9819.7	9819.7	0.0
Aug	0	268	76.7	81.1	0.0	14.2	0.0	90.4	0	550	0	0.0	9590.8	9587.5	0.0
Sep	32	116	76.8	80.9	0.0	14.1	0.0	80.1	0	495	42	0.0	8521.6	7960.9	20.2
Oct	192	21	76.4	80.6	0.0	13.9	0.0	71.7	0	450	207	0.0	7659.2	4928.8	79.3
Nov	625	0	76.1	80.6	0.0	13.6	0.0	53.7	0	344	306	0.0	5755.8	1764.3	1124.1
Dec	793	0	76.4	80.5	0.0	13.4	0.0	50.2	0	323	292	0.0	5385.9	1593.2	1245.5
Totals	4499	915	76.1	81.1	0.0	14.2	0.0	814.0	0	5107	2351	0.0	86873.0	56222.0	6414.0

Notes: Fan supply rate = 7400 CFM, ventilation = 150 CFM, wideband temperature limits for occupied and unoccupied times = 65° to 80°F, economizer temperature limit = 65°F, time period = 1 to 365 days, total hours = 8760.

2. For fan operation, fan power is calculated from the following equation:

$$\text{Fan power (kW)} = 2.487 \times 10^{-4} \times Q \times \Delta P, \tag{7}$$

where

$Q(CFM)$ = air flow delivered by the fan; and
ΔP = static pressure head of the fan in inches of water.

By multiplying the fan power by the total hours of fan operation, we can obtain the total energy use (kWh).

3. Humidification costs are based on supplying energy at the rate of 1000 Btu per pound of water added to the supply air stream. Costs are derived from the unit cost of energy, such as electricity ($/kWh), natural gas ($/1000 ft^3), and fuel oil ($/gal), which is supplied by the user. An 80 percent efficiency rate is assumed for these energy sources.

4. Heating costs are similarly calculated by the unit cost. An 80 percent efficiency rate is also used in these calculations.

V. ILLUSTRATIVE EXAMPLES OF ENERGY CONSUMPTION AND CONTROL

Different geographic loctions of equipment buildings, dual or extended wideband temperature limits, and the method of HVAC control (intermittent or porportional) are considered in several variations. Table I gives this information along with some of the more salient building parameters. The results of the calculation are summarized by month in Tables II through VI for the cases specified in Table I. We assume here that a conventional cooling system, consisting of a chiller

Table IIb—Intermittent control of building space air temperature—New York City, case 1 in Table I

	Number of Hours at Specified Relative Humidity (No Humidity Control)					
	<10%	10–15%	15–20%	20–55%	55–60%	>60%
Conv (no econ)	697.00	1090.00	1086.75	5886.25	0.0	0.0
Economy	815.25	1048.75	1029.75	5866.25	0.0	0.0
Enthalpy	697.50	1101.00	1144.00	5817.50	0.0	0.0

Estimated Operating Cost for Cooling at $0.10/kWh for Electricity (Chiller COP = 3.50)		
Conv (no econ)	$9687 for 86870 kWh	(Fans = 18669 kWh, chiller = 68201 kWh)
Economy	$8622 for 56221 kWh	(Fans = 18669 kWh, chiller = 37552 kWh)
Enthalpy	$7534 for 75342 kWh	(Fans = 18669 kWh, chiller = 56673 kWh)

Estimated Operating Cost for Humidification (20% min) and Heating at $0.10/kWh for Electricity	
Humidification	$1959 for 19580 kWh
Heating	$0 for 0 kWh

Notes: Min space temp occurred on day 300, max space temp occurred on day 213, max cooling load occurred on day 224.

Table IIIa—Proportional control of building space air temperature—New York City, case 2 in Table I

Month	Degree-Days		Space Temp		Max Load (tons)		Total Load (MBtu)		Number of Hours			Heating (kWh Elect)	Cooling (kWh)		Water (gal) to Maintain 20% Min RH
	Heat	Cool	Min	Max	Heat	Cool	Heat	Cool	Heat	Cool	Econ		No Econ	Econ	
Jan	8.58	0	78.9	80.0	0.0	7.8	0.0	44.7	0	744	741	0.0	6465.4	2734.1	4261.7
Feb	797	0	79.9	80.0	0.0	8.3	0.0	40.4	0	672	667	0.0	5839.7	2487.0	3737.9
Mar	703	0	79.9	80.0	0.0	8.5	0.0	52.6	0	744	744	0.0	7126.1	2719.8	2885.7
Apr	358	0	79.9	80.0	0.0	10.0	0.0	62.3	0	720	652	0.0	7855.6	3182.1	837.9
May	118	38	79.9	80.0	0.0	11.4	0.0	75.4	0	744	434	0.0	9034.2	5461.4	262.5
Jun	23	147	79.9	80.0	0.0	12.2	0.0	81.4	0	720	139	0.0	9451.8	8252.0	0.0
Jul	0	325	79.9	80.0	0.0	13.0	0.0	91.0	0	744	0	0.0	10342.3	10342.3	0.0
Aug	0	268	79.9	80.0	0.0	12.3	0.0	88.9	0	744	10	0.0	10172.1	10080.5	0.0
Sep	32	116	79.9	80.0	0.0	11.4	0.0	78.2	0	720	135	0.0	9187.3	8078.6	76.3
Oct	192	21	79.9	80.0	0.0	11.2	0.0	69.2	0	744	514	0.0	8520.7	4672.8	241.7
Nov	625	0	79.9	80.0	0.0	8.6	0.0	50.1	0	720	696	0.0	6828.1	2797.4	3312.4
Dec	793	0	79.9	80.0	0.0	8.5	0.0	46.1	0	744	731	0.0	6584.7	2806.8	4105.5
Totals	4499	915	79.9	80.0	0.0	13.0	0.0	780.3	0	8960	5466	0.0	97408.0	63615.0	19721.0

Notes: Fan supply rate = 7400 CFM, ventilation = 150 CFM, wideband temperature limits for occupied and unoccupied times = 65° to 80°F, economizer temperature limit = 65°F, time period = 1 to 365 days, total hours = 8760.

and air-handling unit, provides cooling. By comparing the different cases, we find some interesting results.

Case 1 differs from case 2 in that the HVAC is intermittently controlled in case 1, but proportionally controlled in case 2. We see from Tables IIb and IIIb, for example, that the maximum cooling loads for both cases occur close in time [August 12 (day 224) and July 31 (day 212)]. However, the maximum cooling load of 13 tons for the proportional control plan (Table IIIa), which compares favorably with the load of 14.2 tons for the intermittent control plan (Table IIa), reduces the required size of the chiller plant by 9 percent. We would expect such a reduction from using a control sequence that follows the load closely and minimally overshoots the dry-bulb air temperature. Also, a control plan that matches the fan capacity to the load would compare favorably in energy use with on-off fan operation.

We can see in Table IIIa (in the column labeled "Space Temp") that the proportional control plan regulates the temperature to within one-tenth of a degree of the wideband temperature limit for the entire calendar year. For the economy cycle operation, the yearly electrical use of case 1 in Table IIb is 56,221 kWh, and that of case 2 in Table IIIb is 63,614 kWh. The chiller energy consumption for case 1, 37,522 kWh, is larger than that for use 2, 31,591 kWh. The proportional control plan, which modulates the air-supply temperature, requires the fan to run continuously at maximum power for the entire year. This maximum use of the fan creates larger overall energy requirements in spite of lower chiller energy use. However, a variable-air-volume system that modulates the fan supply rate to match loads should decrease the required fan power and significantly reduce total energy use.

Table IIIb—Proportional control of building space air temperature—New York City, case 2 in Table I

Number of Hours at Specified Relative Humidity (No Humidity Control)						
	<10%	10–15%	15–20%	20–55%	55–60%	>60%
Conv (No Econ)	869.75	1090.75	1118.75	4393.25	1011.25	276.50
Economy	911.25	1092.50	1110.50	4354.00	1015.00	276.75
Enthalpy	869.75	1095.50	1134.50	4337.75	938.75	383.75

Estimated Operating Cost for Cooling at $0.10/kWh for Electricity (Chiller COP = 3.50)		
Conv (no econ)	$9741 for 97400 kWh	(Fans = 32023 kWh, chiller = 65383 kWh)
Economy	$6361 for 63614 kWh	(Fans = 32023 kWh, chiller = 31591 kWh
Enthalpy	$8930 for 88296 kWh	(Fans = 32023 kWh, chiller = 56273 kWh)

Estimated Operating Cost for Humidification (20% min) and Heating at $0.10/kWh for Electricity	
Humidification	$6022 for 60219 kWh
Heating	$0 for 0 kWh

Notes: Min space temp occurred on day 100, max space temp occurred on day 10, max cooling load occurred on day 212.

Table IVa—Intermittent control of building space air temperature—New York City, case 3 in Table I

Month	Degree-Days		Space Temp		Max Load (tons)		Total Load (MBtu)		Number of Hours			Heating (kWh Elect)	Cooling (kWh)		Water (gal) to Maintian 20% Min RH
	Heat	Cool	Min	Max	Heat	Cool	Heat	Cool	Heat	Cool	Econ		No Econ	Econ	
Jan	858	0	74.7	85.4	0.0	13.2	0.0	46.8	0	304	299	0.0	5033.1	1171.1	1590.2
Feb	797	0	75.1	85.4	0.0	13.3	0.0	42.2	0	274	268	0.0	4534.1	1074.7	1341.3
Mar	703	0	76.0	85.4	0.0	13.3	0.0	53.8	0	347	343	0.0	5776.8	1326.7	1302.5
Apr	358	0	76.3	85.6	0.0	13.6	0.0	62.8	0	400	326	0.0	6725.0	2452.0	446.6
May	118	38	76.3	85.6	0.0	13.9	0.0	75.4	0	473	225	0.0	8047.1	5073.1	158.9
Jun	23	147	76.3	85.6	0.0	14.1	0.0	81.3	0	502	70	0.0	8643.7	7711.2	0.0
Jul	0	325	76.3	85.7	0.0	14.2	0.0	90.7	0	553	0	0.0	9625.4	9625.4	0.0
Aug	0	268	76.2	85.8	0.0	14.2	0.0	88.6	0	541	3	0.0	9406.7	9356.9	0.0
Sep	32	116	75.7	85.6	0.0	14.1	0.0	78.3	0	485	72	0.0	8336.3	7384.6	45.6
Oct	192	21	76.0	85.5	0.0	13.9	0.0	69.5	0	437	245	0.0	7422.9	4189.6	148.0
Nov	625	0	75.1	85.5	0.0	13.6	0.0	51.7	0	332	307	0.0	5545.3	1556.7	1332.1
Dec	793	0	75.5	85.4	0.0	13.4	0.0	48.0	0	310	285	0.0	5152.6	1457.1	1429.6
Totals	4499	915	74.7	85.8	0.0	14.2	0.0	789.0	0	4962	2447	0.0	84249.0	52379.0	7795.0

Notes: Fan supply rate = 7400 CFM, ventilation = 150 CFM, wideband temperature limits for occupied times = 65° to 80°F, wideband temperature limits for unoccupied times = 60° to 85°F, economizer temperature limit = 65°F, time period = 1 to 365 days, total hours = 8760.

Tables IVa and IVb display monthly energy use for case 3 of Table I. Differing from case 1, this plan imposes dual wideband temperature limits. The wideband temperature limits for unoccupied times increase to 60 degrees F and 85 degrees F. Tables IVa and IVb show that this simple change with economy cycle operation reduces annual cooling energy use by 7 percent, from 56,222 kWh to 52,379 kWh. We can attribute this saving mainly to the lower chiller energy requirement, from 37,552 kWh to 34,237 kWh, and to a lesser extent to the smaller fan energy requirements, from 18,669 kWh to 18,141 kWh.

Tables V and VI show the results of the simulation for the buildings located in New Orleans and Phoenix, cases 4 and 5 of Table I. We can see that the maximum chiller load, 14.4 tons, is the same for these diverse locations. The total energy required in these buildings for economy cycle cooling is also nearly equal, 85,027 kWh and 85,123 kWh. Since the cooling load includes both sensible and latent energy, we can surmise from the degree-day totals that the dominant load on the system for Phoenix is sensible heat, and on that for New Orleans latent heat. The distribution of the relative humidity and the costs for humidification shown in Tables Vb and VIb tends to support these observations.

We can see the advantages in running the TELBECC program to compare different control plans. For example, although the intermittent control plan appears to use less energy overall, the proportional control plan actually reduces the size of the chiller plant by 9 percent in the same locale under similar conditions. The higher overall costs can be attributed to continuous operation of the fan, which, if operated to match the load, would consume much less power and make the

Table IVb—Intermittent control of building space air temperature—New York City, case 3 in Table I

Number of Hours at Specified Relative Humidity (No Humidity Control)						
	<10%	10–15%	15–20%	20–55%	55–60%	>60%
Conv (No Econ)	857.50	1205.75	1213.00	5483.75	0.0	0.0
Economy	976.75	1115.75	1138.25	5529.25	0.0	0.0
Enthalpy	857.50	1241.00	1212.00	5449.50	0.0	0.0

Estimated Operating Cost for Cooling at $0.10/kWh for Electricity (Chiller COP = 3.50)		
Conv (no econ)	$8425 for 84247 kWh	(Fans = 18141 kWh, chiller = 66106 kWh)
Economy	$5238 for 52378 kWh	(Fans = 18141 kWh, chiller = 34237 kWh)
Enthalpy	$7278 for 72776 kWh	(Fans = 18141 kWh, chiller = 54635 kWh)

Estimated Operating Cost for Humidification (20% min) and Heating at $0.10/kWh for Electricity	
Humidification	$2380 for 23802 kWh
Heating	$0 for 0 kWh

Notes: Min space temp occurred on day 22, max space temp occurred on day 224, max cooling load occurred on day 224.

Table Va—Intermittent control of building space air temperature—New Orleans, case 4 in Table I

Month	Degree-Days		Space Temp		Max Load (tons)		Total Load (MBtu)		Number of Hours			Heating (kWh Elect)	Cooling (kWh)		Water (gal) to Maintain 20% Min RH
	Heat	Cool	Min	Max	Heat	Cool	Heat	Cool	Heat	Cool	Econ		No Econ	Econ	
Jan	502	0	76.4	80.6	0.0	13.5	0.0	60.8	0	388	341	0.0	6518.0	2044.9	386.5
Feb	465	0	76.4	80.6	0.0	13.8	0.0	55.9	0	356	287	0.0	5984.1	2231.2	865.2
Mar	204	8	76.6	80.7	0.0	14.0	0.0	72.8	0	458	223	0.0	7776.9	4836.3	74.3
Apr	21	133	76.6	80.9	0.0	14.2	0.0	81.1	0	499	36	0.0	8624.8	8136.8	0.2
May	0	296	76.7	81.1	0.0	14.3	0.0	91.7	0	555	2	0.0	9716.7	9690.2	0.0
Jun	0	459	76.8	81.1	0.0	14.4	0.0	95.7	0	570	0	0.0	10103.2	10103.2	0.0
Jul	0	457	76.8	81.1	0.0	14.4	0.0	98.3	0	583	0	0.0	10369.4	10369.4	0.0
Aug	0	477	76.7	80.8	0.0	14.4	0.0	98.6	0	586	0	0.0	10402.5	10402.5	0.0
Sep	0	397	76.8	81.0	0.0	14.4	0.0	92.8	0	553	0	0.0	9797.1	9797.1	0.0
Oct	25	147	76.7	80.9	0.0	14.1	0.0	83.9	0	516	31	0.0	8915.4	8503.2	0.0
Nov	123	62	76.6	80.9	0.0	14.2	0.0	73.5	0	456	141	0.0	7824.2	5953.5	83.7
Dec	383	1	76.1	80.6	0.0	13.9	0.0	64.6		409	300	0.0	6909.0	2960.9	228.1
Totals	1723	2437	76.1	81.1	0.0	14.4	0.0	969.7	0	5934	1364	0.0	102941.0	85029.0	1638.0

Notes: Fan supply rate = 7400 CFM, ventilation = 150 CFM, wideband temperature limits for occupied and unoccupied times = 65° to 80°F, economizer temperature limit = 65°F, time period = 1 to 365 days, total hours = 8760.

Table Vb—Intermittent control of building space air temperature—New Orleans, case 4 in Table I

Number of Hours at Specified Relative Humidity (No Humidity Control)						
	<10%	10–15%	15–20%	20–55%	55–60%	>60%
Conv (No Econ)	14.25	395.50	441.25	7909.00	0.0	0.0
Economy	60.0	383.00	474.75	7842.25	0.0	0.0
Enthalpy	14.25	419.00	455.25	7871.50	0.0	0.0

Estimated Operating Cost for Cooling at $0.10/kWh for Electricity (Chiller COP = 3.50)					
Conv (no econ)	$10294 for 102939 kWh	(Fans = 21693 kWh, chiller = 81246 kWh)			
Economy	$8503 for 85027 kWh	(Fans = 21693 kWh, chiller = 63334 kWh)			
Enthalpy	$9295 for 92949 kWh	(Fans = 21693 kWh, chiller = 71256 kWh)			

Estimated Operating Cost for Humidification (20% min) and Heating at $0.10/kWh for Electricity	
Humidification	$500 for 5002 kWh
Heating	$ 0 for 0 kWh

Notes: Min space temp occurred on day 345, max space temp occurred on day 147, max cooling load occurred on day 165.

proportional control plan much more attractive. We can conclude that TELBECC has great potential for pinpointing significant energy reductions and cost savings before a building's HVAC system is purchased.

VI. CONCLUSIONS

The TELBECC program analyzes more efficiently and quickly than any method used heretofore the possible telephone building environmental energy use and control options. To pinpoint the most economical energy-conservation plan, the program analyzes multiple plans at minimal cost and with minimal expenditure of time. The program calculates the energy consumed every quarter hour by the HVAC in regulating the environment under changing weather conditions. It computes the required energy from the physical characteristics of the building envelope, such as the U factor, internal heat generation, geographic location, orientation of the building, and the dry-bulb temperature standard. In order to make it feasible to calculate by computer, we employ a simplified recursive computation procedure using time series. For each of the illustrative problems in Tables II through VI, the procedure produced monthly projections; yet it took less than 40 seconds to calculate results on an IBM/3033 computer. From the examples, we see the advantages and disadvantages of both the intermittent and proportional control plans, as well as the significant savings obtained from increasing the range of the dual or

Table VIa—Intermittent control of building space air temperature—Phoenix, Ariz., case 5 in Table I

Month	Degree-Days		Space Temp		Max Load (tons)		Total Load (MBtu)		Number of Hours			Heating (kWh Elect)	Cooling (kWh)		Water (gal) to Maintain 20% Min RH
	Heat	Cool	Min	Max	Heat	Cool	Heat	Cool	Heat	Cool	Econ		No Econ	Econ	
Jan	432	0	76.3	80.6	0.0	13.5	0.0	62.8	0	401	300	0.0	6724.2	2812.0	1132.0
Feb	257	1	76.5	80.7	0.0	13.6	0.0	61.5	0	390	249	0.0	6578.6	3313.1	455.0
Mar	139	12	76.6	80.7	0.0	13.5	0.0	74.0	0	468	212	0.0	7908.0	5125.0	1705.1
Apr	38	122	76.6	81.1	0.0	13.7	0.0	79.6	0	497	104	0.0	8486.9	7105.4	336.4
May	16	403	76.7	81.2	0.0	13.9	0.0	92.1	0	570	38	0.0	9807.0	9301.8	64.9
Jun	0	560	76.7	81.1	0.0	14.1	0.0	95.8	0	589	2	0.0	10181.0	10148.0	256.0
Jul	0	847	76.9	81.2	0.0	14.4	0.0	109.5	0	657	0	0.0	11574.5	11574.5	0.0
Aug	0	679	76.7	81.2	0.0	14.4	0.0	104.1	0	626	0	0.0	11016.5	11016.5	0.0
Sep	0	521	76.8	80.8	0.0	14.2	0.0	94.4	0	574	0	0.0	10010.3	10010.3	0.0
Oct	14	201	76.6	81.0	0.0	14.0	0.0	84.5	0	525	86	0.0	9003.7	7860.2	12.0
Nov	196	0	76.5	80.6	0.0	13.6	0.0	67.8	0	428	236	0.0	7243.1	4138.2	161.3
Dec	413	0	76.5	80.6	0.0	13.4	0.0	63.0	0	401	307	0.0	6742.1	2719.4	458.0
Totals	1505	3346	76.3	81.2	0.0	14.4	0.0	989.0	0	6130	1536	0.0	105276.0	85124.0	4582.0

Notes: Fan supply rate = 7400 CFM, ventilation = 150 CFM, wideband temperature limits for occupied and unoccupied times = 65° to 80°F, economizer temperature limit = 65°F, time period = 1 to 365 days, total hours = 8760.

Table VIb—Intermittent control of building space air temperature—Phoenix, Ariz., case 5 in Table I

	Number of Hours at Specified Relative Humidity (No Humidity Control)					
	<10%	10–15%	15–20%	20–55%	55–60%	>60%
Conv (No Econ)	168.50	611.50	1430.00	6550.00	0.0	0.0
Economy	199.00	627.00	1432.25	6501.75	0.0	0.0
Enthalpy	185.50	648.50	1375.00	6551.00	0.0	0.0

Estimated Operating Cost for Cooling at $0.10/kWh for Electricity (Chiller COP = 3.50)		
Conv (no econ)	$10527 for 105274 kWh	(Fans = 22412 kWh, chiller = 82862 kWh)
Economy	$8512 for 85123 kWh	(Fans = 22412 kWh, chiller = 62711 kWh)
Enthalpy	$9205 for 92054 kWh	(Fans = 22412 kWh, chiller = 69642 kWh)

Estimated Operating Cost for Humdification (20% min) and Heating at $0.10/kWh for Electricity	
Humidification	$1399 for 13991 kWh
Heating	$0 for 0 kWh

Notes: Min space temp occurred on day 18, max space temp occurred on day 225, max cooling load occurred on day 204.

extended wideband temperature limits for unoccupied times. In the larger view, we can understand how TELBECC can significantly contribute toward the operating companies' energy-conservation plan for future savings.

ACKNOWLEDGMENT

The author wishes to express his gratitude to Mr. P. P. Gwozdz for the many fruitful and informative discussions during the development of the computer program.

REFERENCES

1. C. Kalasinsky and F. Ferreira, "Energy Simulation with ESP-II," ASHRAE J., *23*, No. 8 (August 1981), pp. 26–9.
2. T. Kusuda, "A Comparison of Energy Calculation Procedures," ASHRAE J., *23*, No. 8 (August 1981), pp. 21–4.
3. *Test Reference Year (TRY) Manual*, National Climatic Center (Dept. of Commerce, National Oceanic and Atmospheric Admin., National Environmental Satellite, Data, and Information Service), Asheville, NC.
4. H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids.* 2d ed. London: Clarendon Press, 1959.
5. G. P. Mitalas and D. G. Stephenson, *Calculation of Heat Conduction Transfer Functions for Multi-Layer Slabs*," ASHRAE Publications, No. 2203, August 1971.
6. *ASHRAE Handbook & Product Directory, 1977 Fundamentals*, New York: American Society of Heating, Refrigerating, and Air-Conditioning Engineers, 1977.
7. P. J. Schneider, *Temperature Response Charts*, New York: John Wiley, 1963.
8. E. I. Jury, *Theory and Application of the Z-Transform Method.* Reprint. Melbourne, FL: R. E. Krieger, 1973.

APPENDIX

Calculating Inside-Wall Temperature Time Series Coefficients b_i, a_i, a_i' of Eq. (1)

The basis for computing the time-series coefficients is the z transform,[8] a discrete function transformation. This transformation is applied to time functions sampled at regular intervals of time. The z transform has the same role in discrete systems that the Laplace transform has in continuous systems analysis.

Let us consider a continuous function of time $f(t)$. When the function is sampled at regular intervals Δ, the output consists of a train of pulses, as illustrated in Fig. 1. We defined the z transform of this output as a polynomial in powers of z^{-1} in the following:

$$f(0) + f(\Delta)z^{-1} + f(2\Delta)z^{-2} + f(3\Delta)z^{-3} + \cdots . \tag{8}$$

The successive outputs of the sampler are the coefficients of the successive powers of z^{-1} in the z transform.

A linear system is characterized when its response to an elementary input (such as a pulse, a unit step, or, as will be adopted here, a unit ramp) is ascertained. This is nothing more than obtaining a transfer function of the system. If both input and output of the system are expressed in terms of their z transforms, the ratio of output/input is the z transform of the system. If we assume that such a transfer function, $G(z)$, can be found and that it can be expressed as the quotient of two polynomials in z^{-1}, then

$$G(z) = \frac{N(z)}{D(z)} = \frac{a_0 + a_1z^{-1} + a_2z^{-2} + \cdots a_jz^{-j}}{b_0 + b_1z^{-1} + b_2z^{-2} + \cdots b_pz^{-p}}. \tag{9}$$

It follows that the z-transform of the output $O(z)$ resulting from an arbitrary input $I(z)$ is represented by

$$O(z) = G(z)I(z) \quad \text{or} \tag{10}$$

$$O(z)D(z) = N(z)I(z). \tag{11}$$

Since both sides of (11) are polynomials, the coefficients of the various powers of z^{-1} must be the same on both sides of the equation. If, say, the coefficients of z^{-n} are equated, eq. (11) yields

$$b_0O_n = a_0I_n + a_1I_{n-1} + a_2I_{n-2} + \cdots + a_jI_{n-j} - [b_1O_{n-1} + b_2O_{n-2} + \cdots + b_pO_{n-p}]. \tag{12}$$

The subscript n on O and I indicates the value of the function at $t = n\Delta$; i.e., $O_n \equiv O(n\Delta)$, the coefficient of z^{-n} in the z transform of $O(z)$. This expression relates the output at any time $t = n\Delta$ to the input at that time and the input and output at earlier times. The coefficients $a_0, \cdots a_j$ and $b_0, \cdots b_p$ contain all the characteristics of

the system. With the properties of the z transform described above, a method for determining the z transform or time-series coefficients for the inside building wall temperature follows.

If we consider the outside building walls and roof structure as homogeneous flat slabs (Fig. 2), the temperature in the slab adheres to the following equations:

$$\kappa \frac{\delta^2 T_s}{\delta x^2} = \frac{\delta T_s}{\delta t},$$

$$T_s(x, 0) = 0,$$

$$k \frac{\delta T_s}{\delta x}(L, t) = -h_O[T_s(L, t) - T_O(t)],$$

$$k \frac{\delta T_s}{\delta x}(0, t) = h_i[T_s(0, t) - T_a(t)], \tag{13}$$

where

$T_s(x, t)$ (°F) = temperature in the slab,
L(ft) = slab thickness,
k(Btu/hr − ft − °F) = thermal conductivity,
$k \equiv \kappa/\rho c$(ft^2/hr) = diffusivity,
ρc(Btu/ft^3) = volumetric heat capacity,
h_O(Btu/hr − ft^2 − °F) = outside-wall heat transfer coefficient,
h_i(Btu/hr − ft^2 − °F) = inside-wall heat transfer coefficient,
$T_O(t)$ = outside sol-air temperature, and
$T_a(t)$ = inside building air temperature.

It is convenient to use the Laplace transform

$$\bar{T}_s(x, p) = \int_0^\infty T_s(x, t) e^{-pt} dt$$

to eliminate the independent time variable t in eq. (13). Then the solution for the inside wall surface ($x = 0$) temperature in terms of the transform parameter p assumes the form:

$$\bar{T}_s(0, p) = \frac{h_i[kq \cosh(qL) + h_O \sinh(qL)]\bar{T}_a(p) + h_O kq \bar{T}_O(p)}{h_i[kq \cosh(qL) + h_O \sinh(qL)] + kq[kq \sinh(qL) + h_O \cosh(qL)]}, \tag{14}$$

where $q = (p/\kappa)^{1/2}$.

Letting $T_a(t)$ and $T_O(t)$ be unit ramp functions and inverting eq. (14) back to the real-time domain by using standard residue theory in the complex plane, the solution for $T_s(0, t)$, the temperature of the inside surface, is expressed as

$$T_s(0, t) = T_s^{(1)}(0, t) + T_s^{(2)}(0, t), \tag{15}$$

where $T_s^{(1)}(0, t)$ is the portion of the solution dependent on the outside sol-air temperature, and $T_s^{(2)}(0, t)$ the part dependent on the building space-air temperature. These temperatures are explicitly:

$$T_s^{(1)}(0, t) = B_O \left[\frac{1}{B_i + B_O B_i + B_O} \left(t - \frac{L^2}{6\kappa} \frac{(3B_i + B_O B_i + 3B_O + 6)}{(B_i + B_O B_i + B_O)} \right) \right.$$

$$\left. - \frac{2L^2}{\kappa} \sum_{n=1}^{\infty} \frac{e^{-\alpha_n^2 \kappa t/L^2}}{\alpha_n^2 \{[(B_i + B_O B_i + B_i) - \alpha_n^2] \cdot \cos \alpha_n - \alpha_n (2 + B_O + B_i) \sin \alpha_n\}} \right], \tag{16}$$

and

$$T_s^{(2)}(0, t) = B_i \left[\frac{1}{B_i + B_O B_i + B_O} \left(\frac{L^2}{6\kappa} (3 + B_O) \right. \right.$$

$$\left. \left. - \frac{L^2}{6\kappa} \frac{(1 + B_O)(3B_i + B_i B_O + 3B_O + 6)}{(B_i + B_O B_i + B_O)} + (1 + B_O) t \right) \right.$$

$$\left. - \frac{2L^2}{\kappa} \sum_{n=1}^{\infty} \frac{[\alpha_n \cos \alpha_n + B_O \sin \alpha_n] e^{-\alpha_n^2 \kappa t/L^2}}{\sigma_n^3 [(B_i + B_i B_O + B_O - \alpha_n^2] \cdot \cos \alpha_n - \alpha_n [2 + B_O + B_i] \sin \alpha_n} \right], \tag{17}$$

where $B_O = \dfrac{h_O L}{k}$, $B_i = \dfrac{h_i L}{k}$, and α_n are roots of the transcendental equation

$$\cot \alpha_n = \frac{\alpha_n^2 - B_O B_i}{\alpha_n (B_O + B_i)}, \qquad n = 1, 2, \cdots.$$

Equations (15) through (17) contain all the ingredients for forming the z-transform transfer functions for the inside-wall surface temperature. These are obtained by forming the ratio of output/input z transforms as per eq. (9):

$$G^{(1)}(z) = \frac{T_s^{(1)}(0, z)}{T_O(z)} \quad \text{and} \quad G^{(2)}(z) = \frac{T_s^{(2)}(0, z)}{T_a(z)}. \tag{18}$$

$T_O(t)$ and $T_a(t)$ were taken as unit ramp functions, and therefore their z transform from Ref. 8 is given as

$$T_O(z) = T_a(z) = \frac{\Delta}{z(1 - z^{-1})^2}. \tag{19}$$

The sampling interval is Δ. The use of the input ramp function amounts to linear interpolation between the discrete values given by the z-transform coefficients.

The z transforms of both $T_s^{(1)}(0, z)$ and $T_s^{(2)}(0, z)$ are similar in form and, with the aid of the table of z transforms given in Ref. 8, can be expressed as

$$T_s^{(1,2)}(0, z) = \frac{A^{(1,2)}}{1 - z^{-1}} + \frac{B^{(1,2)}\Delta}{z(1 - z^{-1})^2} + \sum_{j=1}^{\infty} \frac{C_j^{(1,2)}}{1 - s_j z^{-1}}, \tag{20}$$

where

$$A^{(1)} = \frac{-B_0 L^2}{6\kappa(B_i + B_0 B_i + B_0)^2}(3B_i + B_0 B_i + 3B_0 + 6),$$

$$B^{(1)} = \frac{B_0}{B_i + B_0 B_i + B_0},$$

$$C_j^{(1)} = -\frac{2L^2 B_0}{\kappa}\left(\frac{1}{\alpha_j^2\{[(B_i + B_0 B_i + B_0) - \alpha_j^2] \cdot \cos\alpha_j - \alpha_j(2 + B_0 + B_i)\sin\alpha_j\}}\right),$$

$$A^{(2)} = \frac{B_i}{B_i + B_0 B_i + B_0}\left(\frac{L^2}{6\kappa}(3 + B_0) - \frac{L^2}{6\kappa}\frac{(1 + B_0)(3B_i + B_i B_0 + 3B_0 + 6)}{(B_i + B_0 B_i + B_0)}\right),$$

$$B^{(2)} = \frac{B_i(1 + B_0)}{B_i + B_0 B_i + B_0},$$

$$C_j^{(2)} = -\frac{2B_i L^2}{\kappa}\left(\frac{\alpha_j \cos\alpha_j + B_0 \sin\alpha_j}{\alpha_j^3[(B_i + B_i B_0 + B_0 - \alpha_j^2)] \cdot \cos\alpha_j - \alpha_j(2 + B_0 + B_i)\sin\alpha_j}\right), \quad \text{and}$$

$$s_j = e^{-\alpha_j^2 \Delta\kappa/L^2}$$

Equation (18) can now be expressed in the form of a ratio of polynomials in z^{-1}:

$$G^{(1,2)}(z) = \frac{N^{(1,2)}(z)}{D(z)}, \tag{21}$$

where, from the results of eqs. (19) and (20),

$$N^{(1,2)}(z) = \left(A^{(1,2)}\frac{z(1 - z^{-1})}{\Delta} + B^{(1,2)}\right)\prod_{j=1}^{\infty}(1 - s_j z^{-1}) + \frac{z(1 - z^{-1})^2}{\Delta}\sum_{n=1}^{\infty} C_n^{(1,2)} \prod_{j=n}^{\infty}(1 - s_j s^{-1}),$$

$$D(z) = \prod_{j=1}^{\infty}(1 - s_j z^{-1}).$$

Equation (21) is the form of eq. (9); consequently, the b_i coefficients that are derived, as in eq. (12), from the coefficients of the polynomial $D(z)$ can be generated by a recursive scheme given as

$$b_O = 1,\ b_i^{(1)} = 0,\ b_i^{(n+1)} = b_i^{(n)} - s_{i-n+1} b_{i-1}^{(n)}\ n = 1, \cdots N \qquad (22)$$

The number of s_j terms needed to obtain the desired degree of accuracy for the b_i coefficients is indicated by the index $n = N$, which in most instances should not exceed 20.

The a_i and a_i' coefficients in eq. (1) came from $N^{(1)}(z)$ and $N^{(2)}(z)$, respectively, by expanding these functions into polynomials in powers of z^{-1}; i.e.,

$$N^{(1)}(z) = a_1 z^{-1} + a_2 z^{-2} + \cdots$$

$$N^{(2)}(z) = a_1' z^{-1} + a_2' z^{-2} + \cdots$$

The desired coefficients are sorted out.

AUTHOR

Philip B. Grimado, B.S. (Civil Engineering), 1961, City University of New York; M.S. (Applied Mechanics), 1962, Columbia University; Ph.D. (Applied Mechanics), 1968, Columbia University; Bell Laboratories, 1968—. Mr. Grimado's responsibilities include vulnerability studies of antiballistic missile systems, fire protection studies involving fire risk analyses, heat-transfer calculations, development of standard fire testing methods for operating company equipment, and development of algorithms for optimum control of building environmental equipment. He is currently engaged in developing computer software for automatic generation of optimum layout configurations of administrative office space.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

Recursive Fixed-Order Covariance Least-Squares Algorithms

By M. L. HONIG*

(Manuscript received May 6, 1983)

This paper derives fixed-order recursive Least-Squares (LS) algorithms that can be used in system identification and adaptive filtering applications such as spectral estimation, and speech analysis and synthesis. These algorithms solve the sliding-window and growing-memory covariance LS estimation problems, and require less computation than both unnormalized and normalized versions of the computationally efficient order-recursive (lattice) covariance algorithms previously presented. The geometric or Hilbert space approach, originally introduced by Lee and Morf to solve the prewindowed LS problem, is used to systematically generate least-squares recursions. We show that combining subsets of these recursions results in prewindowed LS lattice and fixed-order (transversal) algorithms, and in sliding-window and growing-memory covariance lattice and transversal algorithms. The paper discusses both least-squares prediction and joint-process estimation.

I. INTRODUCTION

Computationally efficient recursive Least-Squares (LS) algorithms have recently attracted attention in applications such as adaptive equalization,[1-4] echo cancellation,[5] and speech analysis and synthesis[6,7] because of their fast convergence properties when compared to older least-mean-square or gradient adaptation techniques.[8-10] Since the work on computationally efficient LS algorithms by Morf and others first appeared in Refs. 11 and 12, numerous papers have followed that produce computationally efficient algorithms that solve different types

* AT&T Information Systems.

of autoregressive LS estimation problems.[13-17] In general, these algorithms fall into four categories: (1) prewindowed recursive LS, (2) sliding-window recursive LS, (3) growing-memory covariance recursive LS, and (4) nonrecursive LS algorithms. Each of the first three categories has two subcategories: fixed-order, or transversal, algorithms; and order-recursive, or lattice, algorithms.

References 1 and 12 present a prewindowed LS algorithm that satisfies a transversal filter structure (fast Kalman algorithm). Subsequent Refs. 7, 18, and 19 describe prewindowed and growing-memory covariance LS algorithms that satisfy a lattice structure. Normalized prewindowed LS lattice algorithms that involve fewer recursions than the original unnormalized versions, and which have the important advantage that all internal variables are less than or equal to unity in magnitude are presented in the more recent Ref. 13. Reference 14 extends the normalized lattice algorithms to solve the sliding-window and growing-memory covariance LS problems. The recursive algorithms mentioned so far require order N arithmetic operations per iteration to update the filter parameters, where N is the order of the filter. A computationally efficient order-recursive algorithm that solves the set of linear equations for the covariance LS prediction problem has been presented in Ref. 11, and extended to the joint-process-estimation case in Ref. 17. These algorithms require order N^2 operations to compute the LS prediction coefficients and are nonrecursive in the sense that the solution generated at time interval i is not used to generate the solution at time interval $i + 1$.

This paper attempts to unify and extend the previous work by (1) systematically generating all of the recursions needed to derive all of the previously mentioned algorithms, and (2) using these recursions to derive new recursive fixed-order sliding-window and growing-memory covariance LS algorithms. These new algorithms solve directly for the prediction- or autogressive-model coefficients, and involve significantly less computation than both the unnormalized and normalized versions of the order-recursive or covariance lattice algorithms presented in Ref. 14. In addition, in some applications it may be advantageous to work directly with the prediction- or autogressive-model coefficients, rather than the set of reflection coefficients produced by lattice algorithms. The algorithms mentioned in the previous paragraph, along with the new ones derived here, are obtained by appropriately arranging subsets of least-squares recursions. The geometric or Hilbert space approach originally used by Lee and Morf[20] to derive the prewindowed LS lattice algorithm is used to derive all of the basic least-squares recursions in a cohesive manner. In this paper, however, only scalar-valued data are considered.

The next section defines the sliding-window and growing-memory

covariance LS problems to be solved. Then Section III reviews the geometric approach to LS estimation. Fundamental order and time updates for the least-squares projection operator are given in Section IV, with derivations in Appendix A. In Section V these projection updates systematically derive least-squares recursions. Section VI gives fixed-order covariance algorithms and Section VII extends the preceding discussion to the joint-process-estimation case. Appendix B lists subsets of recursions in Sections V and VII that constitute other LS algorithms.

II. PROBLEM STATEMENT

We start by defining a sequence of data values $y_0, y_1, \cdots, y_i$, where i is the current time index. A linear least-squares *forward predictor* of order n chooses the coefficients $f_{j|n}$ to minimize

$$\epsilon_f(i|n) = \sum_{m=i'}^{i} \left(y_m - \sum_{j=1}^{n} f_{j|n} y_{m-j} \right)^2, \tag{1}$$

where i' to i is the time window of interest. The coefficients $f_{j|n}$, $1 \leq j \leq n$, are called the forward-prediction coefficients. A linear least-squares *backward predictor* of order n chooses the backward-prediction coefficients $b_{j|n}$, $1 \leq j \leq n$, to minimize

$$\epsilon_b(i|n) = \sum_{m=i'}^{i} \left(y_{m-n} - \sum_{j=1}^{n} b_{j|n} y_{m-j+1} \right)^2. \tag{2}$$

Minimization of (1) rather than (2) is generally desired for a given application. The forward and backward prediction problems stated above are closely related, however, and the LS algorithms to be presented use the backward prediction coefficients to solve for the forward prediction coefficients in a computationally efficient manner.

If, instead of estimating future values of the *same* process, we wish to estimate another related process $\{x_i\}$, the least-squares cost function becomes

$$\epsilon_x(i|n) = \sum_{m=i'}^{i} \left(x_m - \sum_{j=1}^{n} c_{j|n} y_{m-j+1} \right)^2, \tag{3}$$

where tap coefficients $c_{j|n}$ replace the prediction coefficients $f_{j|n}$ and $b_{j|n}$. The cost function (3) is relevant to joint-process-estimation problems such as channel equalization and echo cancellation. In the case of channel equalization, y_i is the ith sample of the channel output, and x_i is the ith channel symbol.

Setting the derivatives of the cost functions (1), (2), and (3) with respect to the prediction (tap) coefficients equal to zero results in the

following linear equations:

$$\Phi_{i'-1,i-1|n}\mathbf{f}(i|n) = \sum_{j=i'}^{i} y_j \mathbf{y}_{j-1|n}, \tag{4a}$$

$$\Phi_{i',i|n}\mathbf{b}(i|n) = \sum_{j=i'}^{i} y_{j-n} \mathbf{y}_{j|n}, \tag{4b}$$

and

$$\Phi_{i',i|n}\mathbf{c}(i|n) = \sum_{j=i'}^{i} x_j \mathbf{y}_{j|n}, \tag{4c}$$

where

$$\mathbf{f}^T(i|n) \equiv [f_{1|n} f_{2|n} \cdots f_{n|n}], \tag{5a}$$

$$\mathbf{b}^T(i|n) \equiv [b_{1|n} b_{2|n} \cdots b_{n|n}], \tag{5b}$$

$$\mathbf{c}^T(i|n) \equiv [c_{1|n} c_{2|n} \cdots c_{n|n}], \tag{5c}$$

$$\mathbf{y}_{j|n}^T \equiv [y_j y_{j-1} \cdots y_{j-n+1}], \tag{6}$$

and the covariance matrix

$$\Phi_{i',i|n} \equiv \sum_{j=i'}^{i} \mathbf{y}_{j|n} \mathbf{y}_{j|n}^T. \tag{7}$$

Suppose now that $i' = 0$, and that y_0 is the first available data sample. The least-squares solutions for **f**, **b**, and **c**, obtained by solving (4), are undefined since they depend on the unspecified data values $y - 1, y - 2, \cdots, y_{-n}$. The simplest, and perhaps most popular, technique for circumventing this problem is to assume all data values $y_j, j < 0$, are zero. The least-squares solutions resulting from this so-called prewindowed estimation are then well defined as long as the covariance matrix is nonsingular. In applications such as speech modelling, however, where estimates of the prediction coefficient vector $\mathbf{f}(i|n)$ are desired given relatively few data samples, prewindowed estimation may result in undesirable edge effects from assuming data is zero outside a given window. For these types of applications, it is desirable to estimate the prediction coefficients without any assumptions concerning the data outside the time window of interest.

Covariance least-squares estimation replaces the lower time limit i' in (1), (2), and (3) by n, so that only known data values are used to compute the LS prediction (tap) coefficients. The improved estimates so obtained are not without cost, however. The resulting covariance LS algorithms derived in this paper and elsewhere[7] involve more computation than prewindowed LS algorithms. Notice that at each

iteration i, the prediction coefficients are computed from $i + 1$ data values. Because the number of data samples entering the least-squares computation grows with time, this type of estimation has been called growing-memory covariance estimation.[16]

Finally, another windowing technique that has attracted attention recently is the sliding-window technique, in which the lower time limit i' in (1) and (2) is replaced by $i - M + n + 1$, and in (3) by $i - N + n$, where M is a predetermined constant. At each iteration the least-squares prediction coefficients are therefore computed from a fixed number (M) of data samples. Notice that data samples outside the time window $i - M + 1$ to i have no effect on the least-squares solution for $\mathbf{f}$, $\mathbf{b}$, and $\mathbf{c}$ at time i, i.e., they are totally forgotten. This is in contrast to more conventional exponential forgetting techniques that reduce the effects of past data samples in a more continuous fashion.[16] The sliding window is therefore useful in applications where the autoregressive model changes abruptly with time, or where undesirable transients periodically affect the data samples. In the former case, when the model parameters change, the sliding window eventually discards data values corresponding to previous model parameters. In the latter case, the sliding window eventually discards corrupted data values.

Computationally efficient recursive algorithms that solve the growing-memory covariance and sliding-window LS estimation problems will be derived in Sections V through VII. The next section develops the necessary mathematical background by reviewing the geometric interpretation of linear least-squares estimation.

III. MATHEMATICAL BACKGROUND

Given two vectors $\mathbf{X}$ and $\mathbf{Y}$ having the same dimension i, the inner product of $\mathbf{X}$ and $\mathbf{Y}$ is defined to be

$$\langle \mathbf{X}, \mathbf{Y} \rangle \equiv \mathbf{X}^T \mathbf{W} \mathbf{Y}, \tag{8}$$

where $\mathbf{W}$ is some prespecified $i \times i$ weighting matrix. As an example, a typical weighting matrix is the exponential weighting matrix

$$\mathbf{W}_i = [1 \; w \; w^2 \; \cdots \; w^{i-1}]\mathbf{I}, \tag{9}$$

where $\mathbf{I}$ is the $i \times i$ identity matrix. For convenience, we will assume that $\mathbf{W}$ is the identity matrix. Modification of the results in this paper to the case where $\mathbf{W}$ is arbitrary is straightforward. The distance between two vectors $\mathbf{X}$ and $\mathbf{Y}$ with the same dimension is therefore the regular Euclidean distance,

$$\mathrm{d}(\mathbf{X}, \mathbf{Y}) = \| \mathbf{Y} - \mathbf{X} \| \equiv \langle \mathbf{Y} - \mathbf{X}, \mathbf{Y} - \mathbf{X} \rangle^{1/2}. \tag{10}$$

The (nth order) projection of a vector $\mathbf{Y}$ onto a subspace (or

manifold) M, which is spanned by the n vectors $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$, is denoted as $P_M\mathbf{Y}$. The orthogonal projection of $\mathbf{Y}$ onto M is defined as

$$P_M^{\perp}\mathbf{Y} \equiv \mathbf{Y} - P_M\mathbf{Y}, \tag{11}$$

and is orthogonal to the subspace M. This implies that

$$\langle \mathbf{X}_j, \mathbf{Y} - P_M\mathbf{Y} \rangle = 0, \qquad \text{for } j = 1, \cdots, n. \tag{12}$$

Since $P_M\mathbf{Y}$ lies in M, there exist constants, or regression coefficients $f_1, f_2, \cdots, f_n$ such that

$$P_M\mathbf{Y} = \sum_{j=1}^{n} f_j\mathbf{X}_j = \mathbf{S}\mathbf{f}, \tag{13}$$

where $\mathbf{S} = [\mathbf{X}_1 \cdots \mathbf{X}_n]$ and $\mathbf{f}^T = [f_1 \cdots f_n]$. Using (12) and (13), it is easy to show that

$$\mathbf{f} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{X} \tag{14}$$

and

$$P_M\mathbf{Y} = \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{Y}, \tag{15}$$

assuming $\mathbf{S}^T\mathbf{S}$ is nonsingular.

The linear least-squares estimate of $\mathbf{Y}$, based upon the vectors $\mathbf{X}_1, \cdots, \mathbf{X}_n$, is formed by choosing $f_1, \cdots, f_n$ such that

$$\| \underline{\epsilon} \|^2 \equiv \| \mathbf{Y} - \sum_{j=1}^{n} f_j\mathbf{X}_j \|^2 \tag{16}$$

is minimized. Differentiating this quantity with respect to f_j and setting the result equal to zero gives

$$\hat{\mathbf{Y}} \equiv \sum_{j=1}^{n} f_j\mathbf{X}_j = P_M\mathbf{Y}, \tag{17}$$

and the vector of estimation errors,

$$\underline{\epsilon} = \mathbf{Y} - \sum_{j=1}^{n} f_j\mathbf{X}_j = P_M^{\perp}\mathbf{Y}. \tag{18}$$

We have identified the operator P as a least-squares projection.

IV. PROJECTION-OPERATOR UPDATE FORMULAS

In this section some fundamental relationships satisfied by the least-squares projection operator are presented. These projection updates fall into two main categories: order updates and time updates. Under time updates are two subcategories, forward and backward time updates. We point out in advance that a total of three projection-operator updates will be used throughout this paper: one order update, one

forward time update, and one backward time update. In addition, one forward and one backward time update for inner products will be needed.

4.1 Order updates

Given two vectors, $\mathbf{Y}$ and $\mathbf{X}$, and a linear space M spanned by the vectors $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$, all in $\mathbf{R}^i$, suppose we wish to calculate the least-squares estimate of $\mathbf{Y}$ based upon the vectors $\mathbf{X}_1, \cdots, \mathbf{X}_n$ and $\mathbf{X}$. In particular, we wish to find coefficients a_j, $j = 1, \cdots, n$, and b such that $\| \mathbf{Y} - (\sum_{j=1}^{n} a_j \mathbf{X}_j + b\mathbf{X}) \|^2$ is minimized. From the discussion in the last section, we know that the least-squares estimate of $\mathbf{Y}$ is

$$\sum_{j=1}^{n} a_j \mathbf{X}_j + b\mathbf{X} = P_{\{M+\mathbf{X}\}}\mathbf{Y}, \tag{19}$$

where $\{M + \mathbf{X}\}$ denotes the space spanned by M and $\mathbf{X}$. We can write the following orthogonal decomposition of the space $\{M + \mathbf{X}\}$,[21]

$$\{M + \mathbf{X}\} = M \oplus \{P_M^{\perp}\mathbf{X}\}. \tag{20}$$

By the Hilbert space projection theorem,[21] we have that for any vector $\mathbf{Y} \in \mathbf{R}^i$,

$$P_{\{M+\mathbf{X}\}}\mathbf{Y} = P_M\mathbf{Y} + P_{\{P_M^{\perp}\mathbf{X}\}}\mathbf{Y}. \tag{21}$$

Figure 1 illustrates this equation for the special case $n = 1$. The projection of $\mathbf{Y}$ onto the space spanned by two vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ is shown as the sum of the two projections $P_{X_1}\mathbf{Y}$ and $P_{\{P_{X_1}^{\perp}\mathbf{X}_2\}}\mathbf{Y}$.

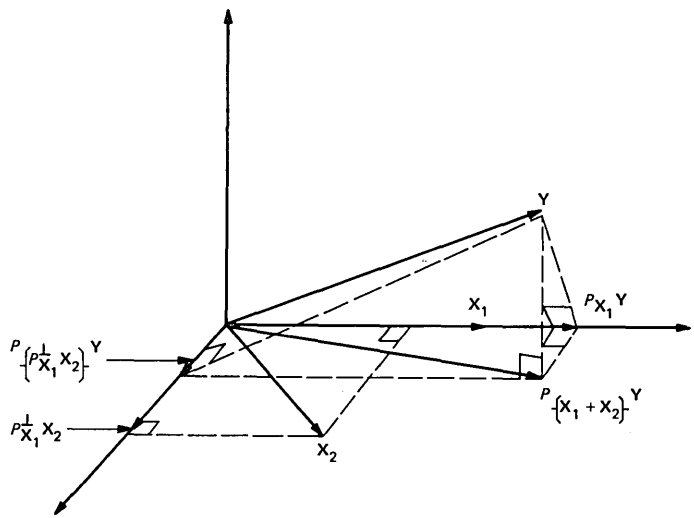


Fig. 1—Decomposition of $P_{\{\mathbf{X}_1+\mathbf{X}_2\}}\mathbf{Y}$.

Equation (21) constitutes a fundamental order update for the least-squares projection operator. The $(n+1)$st-order projection $P_{\{M+\mathbf{X}\}}$ is expressed as the sum of the nth order projection P_M and the first order projection $P_{\{P_M^{\perp}\mathbf{X}\}}$. By subtracting both sides of (21) from $\mathbf{Y}$, we obtain the following order update for the *orthogonal* projection operator $P^{\perp}$,

$$P^{\perp}_{\{M+\mathbf{X}\}}\mathbf{Y} = P^{\perp}_{M}\mathbf{Y} - P_{\{P^{\perp}_{M}\mathbf{X}\}}\mathbf{Y}. \tag{22}$$

4.2 Forward time updates

The forward time updates derived in this section compute a least-squares projection at time i given the same least-squares projection at time $i-1$. These recursions, when combined with the order recursions in the last subsection, can be used to derive prewindowed LS algorithms. We first consider the following vectors $\mathbf{X}_{i_0,i}$ and $\mathbf{Y}_{i_0,i}$, which are composed of data samples from time i_0 to i, i.e.,

$$\mathbf{X}^T_{i_0,i} = [x_i \; x_{i-1} \; \cdots \; x_{i_0}], \tag{23a}$$

and

$$\mathbf{Y}^T_{i_0,i} = [y_i \; y_{i-1} \; \cdots \; y_{i_0}]. \tag{23b}$$

For notational convenience, in this section only we will omit the lower time subscript on the data vectors and assume it to be i_0. Our objective is to compute the linear least-squares estimate of $\mathbf{Y}_i$, given $\mathbf{X}_i$ in terms of a least-squares estimate that does not use the most recent value y_i. With this in mind we define the unit vector

$$\mathbf{u}_i^T = [1 \; 0 \; \cdots \; 0 \; 0], \tag{24}$$

which has the same dimension as $\mathbf{Y}_i$, i.e., $\mathbf{u}_i \in \mathbf{R}^{i-i_0+1}$. Associated with $\mathbf{u}_i$ is the space spanned by $\mathbf{u}_i$, or the space of most recent data values, denoted as U_i. Note that $P_{U_i}\mathbf{Y}_i = y_i\mathbf{u}_i$. For notational convenience we define a tilde operator as follows,

$$\tilde{\mathbf{Y}}_i \equiv P^{\perp}_{U_i}\mathbf{Y}_i = [0 \; y_{i-1} \; y_{i-2} \; \cdots \; y_{i_0+1} \; y_{i_0}], \tag{25}$$

i.e., $\tilde{\mathbf{Y}}_i$ is the projection of $\mathbf{Y}_i$ onto the subspace of past data values.

The basic prediction problem is illustrated in Fig. 2, where $\mathbf{Y}_i$ is a vector having its endpoint in back of the plane of the paper and $\mathbf{X}_i$ has its endpoint in front of the plane of the paper. We are given the vector $\mathbf{X}_i$, from which the least-squares estimate of $\mathbf{Y}_i$, $P_{\mathbf{X}_i}\mathbf{Y}_i$, is to be recursively obtained. At time i we therefore assume a regression coefficient a computed at time $i-1$ (i.e., $P_{\mathbf{X}_{i-1}}\mathbf{Y}_{i-1} = a\mathbf{X}_{i-1}$, or equivalently, $P_{\tilde{\mathbf{X}}_i}\tilde{\mathbf{Y}}_i = a\tilde{\mathbf{X}}_i$), which we wish to modify using the most recent data values y_i and x_i. Figure 2 therefore shows $P_{\mathbf{X}_i}\mathbf{Y}_i$ decomposed into the two vectors $a\mathbf{X}_i$ and $P_{\mathbf{X}_i}(\mathbf{Y}_i - a\mathbf{X}_i)$. Figure 3 illustrates the plane spanned by $\mathbf{X}_i$, $\tilde{\mathbf{X}}_i$, and U_i. Since ABC and ADE are similar

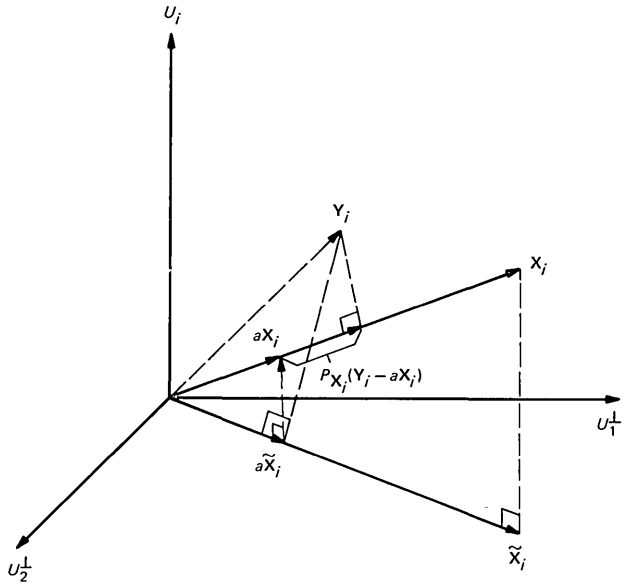


Fig. 2—Decomposition of $P_{\mathbf{X}_i}\mathbf{Y}_i$.

triangles,

$$\frac{\text{AB}}{\text{AD}} = \frac{\text{AC}}{\text{AE}} = a, \tag{26}$$

so that $\overline{\text{AC}} = a\mathbf{X}_i$. Figure 4 attempts to include vectors not shown in Fig. 2 and again illustrates the decomposition of $P_{\mathbf{X}_i}\mathbf{Y}_i$. (Only the endpoint of $\mathbf{Y}_i$ is in Fig. 4.)

Assume now that the vector $\mathbf{X}_i$ is replaced by the *subspace* M_i spanned by the vectors $\mathbf{X}_{1,i}, \mathbf{X}_{2,i}, \cdots, \mathbf{X}_{n,i}$. Let

$$\mathbf{S}_i = [\mathbf{X}_{1,i}\ \mathbf{X}_{2,i}\ \cdots\ \mathbf{X}_{n,i}] \tag{27}$$

and

$$\tilde{\mathbf{S}}_i = [\tilde{\mathbf{X}}_{1,i}\ \tilde{\mathbf{X}}_{2,i}\ \cdots\ \tilde{\mathbf{X}}_{n,i}]. \tag{28}$$

We define the projection

$$P_{M_{i|i-1}}\mathbf{Y}_i \equiv \mathbf{S}_i[\tilde{\mathbf{S}}_i^T\tilde{\mathbf{S}}_i]^{-1}\tilde{\mathbf{S}}_i^T\tilde{\mathbf{Y}}_i = \mathbf{S}_i\mathbf{f}, \tag{29}$$

i.e., $P_{M_{i|i-1}}\mathbf{Y}_i$ lies in M_i, but uses regression coefficients computed at time $i-1$. Referring to Fig. 3, $P_{\mathbf{X}_{i|i-1}}\mathbf{Y}_i = a\mathbf{X}_i$. Appendix A shows that

$$P_{M_i}\mathbf{Y}_i = P_{M_{i|i-1}}\mathbf{Y}_i + P_{M_i}\mathbf{u}_i\langle\mathbf{u}_i, P_{M_i}^\perp\mathbf{Y}_i\rangle\sec^2\theta_i, \tag{30}$$

where

$$\sin^2\theta_i = \langle\mathbf{u}_i, P_{M_i}\mathbf{u}_i\rangle = \|P_{M_i}\mathbf{u}_i\|^2$$

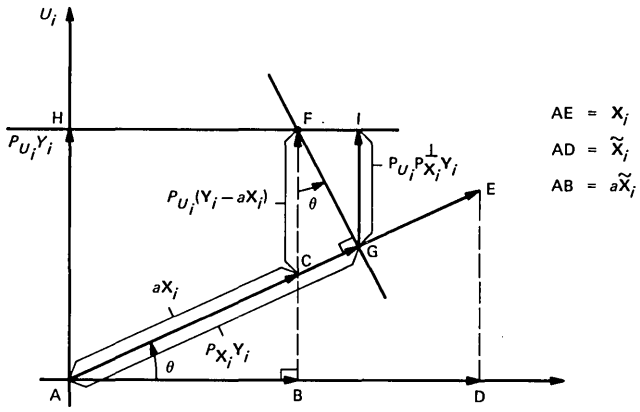


Fig. 3—Plane spanned by $\mathbf{X}_i$ and $\tilde{\mathbf{X}}_i$.

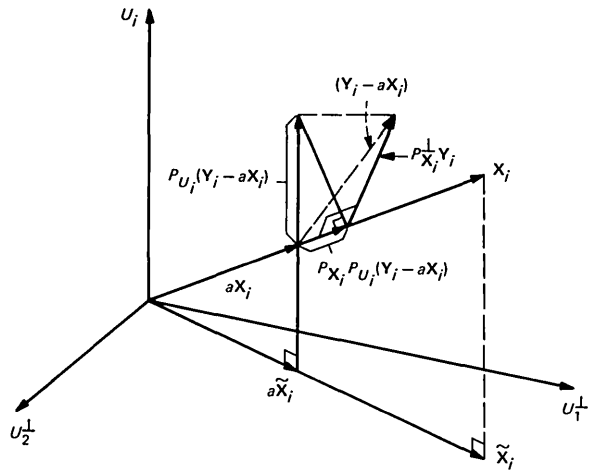


Fig. 4—Rotated view of Fig. 2.

$$= (\mathbf{u}_i^T\mathbf{S}_i)(\mathbf{S}_i^T\mathbf{S}_i)^{-1}(\mathbf{S}_i^T\mathbf{u}_i), \tag{31}$$

and

$$\sec^2\theta_i = \frac{1}{1 - \sin^2\theta_i}. \tag{32}$$

The variable θ_i can be interpreted as the angle between the spaces spanned by the matrices of basis vectors $\mathbf{S}_i$ and $\tilde{\mathbf{S}}_i$. Referring to Fig. 3, the angle θ is given by

$$\sin^2\theta = 1 - \frac{\|\tilde{\mathbf{X}}_i\|^2}{\|\mathbf{X}_i\|^2} = \frac{x_i^2}{\|\mathbf{X}_i\|^2}, \tag{33}$$

and measures the unexpectedness of the data received at time i. Notice that (33) can be rewritten as (31), where M_i and $\mathbf{S}_i$ are replaced by $\mathbf{X}_i$.

We obtain the following time update for the orthogonal projection operator by subtracting both sides of (30) from $\mathbf{Y}_i$,

$$P_{M_i}^{\perp}\mathbf{Y}_i = P_{M_{i|i-1}}^{\perp}\mathbf{Y}_i - P_{M_i}\mathbf{u}_i\langle \mathbf{u}_i,\ P_{M_i}^{\perp}\mathbf{Y}_i\rangle \sec^2\theta_i. \tag{34}$$

One more relation that will be useful in the following section is a recursive equation for the inner product $\langle \mathbf{v}_i,\ P_{M_i}^{\perp}\mathbf{Y}_i\rangle$, where $\mathbf{v}_i$ is an arbitrary vector in $\mathbf{R}^{i-i_0+1}$. This recursion, which is derived in Appendix A, is

$$\langle \mathbf{v}_i,\ P_{M_i}^{\perp}\mathbf{Y}_i\rangle = \langle \tilde{\mathbf{v}}_i,\ P_{\tilde{M}_i}^{\perp}\tilde{\mathbf{Y}}_i\rangle + \langle \mathbf{u}_i,\ P_{M_i}^{\perp}\mathbf{v}_i\rangle\langle \mathbf{u}_i,\ P_{M_i}^{\perp}\mathbf{Y}_i\rangle \sec^2\theta_i, \tag{35}$$

where $\tilde{M}_i$ is the space spanned by $\tilde{\mathbf{S}}_i$.

4.3 Backward time updates

Consider again the data vectors $\mathbf{X}_i$ and $\mathbf{Y}_i$ defined by (23). Suppose we wish to compute the linear least-squares estimate of $\mathbf{Y}_i$ given $\mathbf{X}_i$ in terms of a least-squares estimate that does not use the most *distant* or *past* values y_{i_0} and x_{i_0}. Clearly, this problem can be solved in exactly the same fashion as the time-update problem stated at the beginning of the last section. By turning the vectors $\mathbf{Y}_i$ and $\mathbf{X}_i$ upside down, and assuming that y_{i_0} and x_{i_0} are the most recent samples, one can solve this problem by using time updates already derived. The same argument holds when $\mathbf{X}_i$ is replaced by the subspace M_i spanned by vectors $\mathbf{X}_{1,i},\ \mathbf{X}_{2,i},\ \cdots,\ \mathbf{X}_{n,i}$. In this case we wish to calculate the projection $P_{M_i}\mathbf{Y}_i$ in terms of a projection onto the space spanned by the matrix of basis vectors $\mathbf{S}_i$ in which the *bottom row* has been replaced by zeros. This is in contrast to the previous time updates, which expressed $P_{M_i}\mathbf{Y}_i$ in terms of a projection onto the space spanned by $\mathbf{S}_i$ in which the *top* row has been replaced by zeroes (i.e., $\tilde{M}_i$).

In analogy with the notation defined in the last section, we define the unit vector

$$\mathbf{u}_{i_0}^T = [0\ 0\ \cdots\ 0\ 1] \in \mathbf{R}^{i-i_0+1}, \tag{36}$$

and the space spanned by $\mathbf{u}_{i_0}$ as U_{i_0}. We also define the following asterisk operator in analogy with the previous tilde operator,

$$\mathbf{Y}_i^* = P_{U_{i_0}}^{\perp}\mathbf{Y}_i = [y_i\ y_{i-1}\ \cdots\ y_{i_0+1}\ 0]^T. \tag{37}$$

Similarly,

$$\mathbf{S}_i^* = [\mathbf{X}_{1,i}^*\ \mathbf{X}_{2,i}^*\ \cdots\ \mathbf{X}_{n,i}^*]. \tag{38}$$

The projection of $\mathbf{Y}_i$ onto M_i using regression coefficients computed from $\mathbf{S}_i^*$is defined as

$$P_{M_{i_0|i_0+1}}\mathbf{Y}_i \equiv \mathbf{S}_i[\mathbf{S}_i^{*T}\mathbf{S}_i^*]^{-1}[\mathbf{S}_i^{*T}\mathbf{Y}_i]. \tag{39}$$

The regression coefficients that multiply the basis vectors of M_i are in this case elements of the vector $[\mathbf{S}_i^{*T}\mathbf{S}_i^*]^{-1}[\mathbf{S}_i^{*T}\mathbf{Y}_i]$.

The derivation of (30) can be repeated with $\mathbf{u}_i$ replaced by $\mathbf{u}_{i_0}$, tildes replaced by asterisks, and $P_{M_{i|i-1}}$ replaced by $P_{M_{i_0|i_0+1}}$ to give the following projection decomposition,

$$P_{M_i}\mathbf{Y}_i = P_{M_{i_0|i_0+1}}\mathbf{Y}_i + P_{M_i}\mathbf{u}_{i_0}\langle \mathbf{u}_{i_0}, P_{M_i}^{\perp}\mathbf{Y}_i\rangle \sec^2\theta_i^*, \tag{40}$$

where

$$\begin{aligned}\sin^2\theta_i^* &= \| P_{M_i}\mathbf{u}_{i_0} \|^2 \\ &= (\mathbf{u}_{i_0}^T\mathbf{S}_i)(\mathbf{S}_i^T\mathbf{S}_i)^{-1}(\mathbf{S}_i^T\mathbf{u}_{i_0}) \\ &= \langle \mathbf{u}_{i_0}, P_{M_i}\mathbf{u}_{i_0}\rangle,\end{aligned} \tag{41}$$

and

$$\sec^2\theta_i^* = \frac{1}{1 - \sin^2\theta_i^*}. \tag{42}$$

Subtracting both sides of (40) from $\mathbf{Y}_i$ gives

$$P_{M_i}^{\perp}\mathbf{Y}_i = P_{M_{i_0|i_0+1}}^{\perp}\mathbf{Y}_i - P_{M_i}\mathbf{u}_{i_0}\langle \mathbf{u}_{i_0}, P_{M_i}^{\perp}\mathbf{Y}_i\rangle \sec^2\theta_i^*. \tag{43}$$

Finally, the following update for inner products is analogous to (35),

$$\langle \mathbf{v}_i, P_{M_i}^{\perp}\mathbf{Y}_i\rangle = \langle \mathbf{v}_i^*, P_{M_i^*}^{\perp}\mathbf{Y}_i^*\rangle + \langle \mathbf{u}_{i_0}, P_{M_i}^{\perp}\mathbf{v}_i\rangle\langle \mathbf{u}_{i_0}, P_{M_i}^{\perp}\mathbf{Y}_i\rangle \sec^2\theta_i^*. \tag{44}$$

This completes the presentation of projection-operator recursions needed to derive the least-squares recursions in Sections V and VII. All order updates for variables entering the least-squares algorithms to be presented can be derived from (22). Similarly, all forward and backward time updates for vectors entering these algorithms can be derived from (34) and (43), respectively, and all forward and backward time updates for inner products can be derived from (35) and (44), respectively.

V. LEAST-SQUARES RECURSIONS

5.1 Notation

Referring to the definition (23), a shift operator z^{-j} is defined by

$$z^{-j}\mathbf{Y}_{i_0,i}^T = [y_{i-j}\ y_{i-j-1}\ \cdots\ y_{i_0-j}]. \tag{45}$$

Equations (1) and (2) can now be rewritten as

$$\epsilon_f(i \mid n) = \| \mathbf{Y}_{i_0+n,i} - \sum_{j=1}^{n} f_{j\mid n}(z^{-j}\mathbf{Y}_{i_0+n,i}) \|^2 \tag{46a}$$

and

$$\epsilon_b(i \mid n) = \| z^{-n}\mathbf{Y}_{i_0+n,i} - \sum_{j=1}^{n} b_{j\mid n}(z^{-j+1}\mathbf{Y}_{i_0+n,i}) \|^2, \tag{46b}$$

where i' has been replaced by $i_0 + n$. A matrix of shifted data vectors is denoted as

$$\mathbf{S}_{i_0+n,i}(l, n) = [z^{-l}\mathbf{Y}_{i_0+n,i}\ z^{-l-1}\mathbf{Y}_{i_0+n,i} \cdots z^{-n}\mathbf{Y}_{i_0+n,i}], \tag{47}$$

where $l < n$. The space spanned by the columns of $\mathbf{S}_{i_0+n,i}(l, n)$, which is a subspace generated by past data values, is denoted as $M_{i_0+n,i}(l, n)$. For notational convenience we will omit the lower time index of $\mathbf{S}$ and M and assume that it is always $i_0 + n$. Notice that we can write the covariance matrix defined by (7) as

$$\Phi_{i_0+n,i\mid n} = \mathbf{S}_i^T(0, n-1)\mathbf{S}_i(0, n-1). \tag{48}$$

Two types of updates exist for least-squares parameters: order updates and time updates. The time updates in this section generally fall into two categories. Given some LS parameter ξ (i.e., the forward prediction vector $\mathbf{f}$ or the prediction residual), we wish to find (1) a recursion for ξ computed from the data samples $\{y_{i_0}, y_{i_0+1}, \cdots, y_i\}$ in terms of ξ computed from the data samples $\{y_{i_0}, y_{i_0+1}, \cdots, y_{i-1}\}$ (forward time update), and (2) a recursion for ξ computed from the data samples $\{y_{i_0}, y_{i_0+1}, \cdots, y_i\}$ in terms of ξ computed from the data samples $\{y_{i_0+1}, y_{i_0+2}, \cdots, y_i\}$ (backward time update). Associated with the variable ξ is therefore an order index n and the time indices of the data used in the least-squares computation. If the data values $\{y_{i_0}, y_{i_0+1}, \cdots, y_i\}$ are used to compute ξ, then the indices i_0 and i must be specified. This is in contrast to the prewindowed case where only i need be specified since i_0 is always zero.

Throughout the rest of this paper, the starting-time index of the generic parameter ξ will appear as a subscript, and the current-time index will appear as a function argument. As an example $\xi_{i_0}(i \mid n)$ implies that the data values $\{y_{i_0}, y_{i_0+1}, \cdots, y_i\}$ are used to compute the nth order variable ξ. The following variables are needed to derive the LS algorithms in the next section:

1. Forward and backward prediction vectors [from (4)],

$$\mathbf{f}_{i_0}(i \mid n) = \Phi^{-1}_{i_0+n-1,i-1\mid n}[\mathbf{S}_i^T(1, n)\mathbf{Y}_{i_0+n,i}] \tag{49a}$$

and

$$\mathbf{b}_{i_0}(i \mid n) = \Phi^{-1}_{i_0+n,i\mid n}[\mathbf{S}_i^T(0, n-1)(z^{-n}\mathbf{Y}_{i_0+n,i})]. \tag{49b}$$

2. Forward and backward prediction residual vectors,

$$\mathbf{E}_{f,i_0}(i \mid n) \equiv \mathbf{Y}_{i_0+n,i} - \mathbf{S}_i(1, n)\mathbf{f}_{i_0}(i \mid n) \tag{50a}$$

and

$$\mathbf{E}_{b,i_0}(i \mid n) \equiv z^{-n}\mathbf{Y}_{i_0+n,i} - \mathbf{S}_i(0, n-1)\mathbf{b}_{i_0}(i \mid n). \tag{50b}$$

3. Forward and backward prediction residuals (scalars),

$$e_{f,i_0}(i \mid n) \equiv \langle \mathbf{u}_i, \mathbf{E}_{f,i_0}(i \mid n) \rangle = y_i - \mathbf{f}_{i_0}^T(i \mid n)\mathbf{y}_{i-1 \mid n} \tag{51a}$$

and

$$e_{b,i_0}(i \mid n) \equiv \langle \mathbf{u}_i, \mathbf{E}_{b,i_0}(i \mid n) \rangle = y_{i-n} - \mathbf{b}_{i_0}^T(i \mid n)\mathbf{y}_{i \mid n}. \tag{51b}$$

4. Forward and backward cost functions,

$$\epsilon_{f,i_0}(i \mid n) \equiv \| \mathbf{E}_{f,i_0}(i \mid n) \|^2, \qquad \epsilon_{b,i_0}(i \mid n) \equiv \| \mathbf{E}_{b,i_0}(i \mid n) \|^2. \tag{52}$$

5. PARtial CORrelation (PARCOR) coefficient,

$$k_{n,i_0}(i) \equiv \langle \mathbf{E}_{f,i_0+1}(i \mid n-1), \mathbf{E}_{b,i_0}(i-1 \mid n-1) \rangle. \tag{53}$$

6. Auxiliary variables, or gains,

$$\mathbf{g}_{i_0+1}(i \mid n) \equiv \Phi_{i_0+n,i \mid n}^{-1}\mathbf{y}_{i \mid n}, \tag{54a}$$

$$\mathbf{h}_{i_0+1}(i \mid n) \equiv \Phi_{i_0+n,i \mid n}^{-1}\mathbf{y}_{i_0+n \mid n}, \tag{54b}$$

$$\gamma_{i_0+1}(i \mid n) \equiv \langle \mathbf{u}_i, P_{M_i(0,n-1)}\mathbf{u}_i \rangle = \mathbf{y}_{i \mid n}^T \Phi_{i_0+n,i \mid n}^{-1}\mathbf{y}_{i \mid n}, \tag{55a}$$

$$\gamma_{i_0+1}^*(i \mid n) \equiv \langle \mathbf{u}_{i_0}, P_{M_i(0,n-1)}\mathbf{u}_{i_0} \rangle = \mathbf{y}_{i_0+n \mid n}^T \Phi_{i_0+n,i \mid n}^{-1}\mathbf{y}_{i_0+n \mid n}, \tag{55b}$$

and

$$\alpha_{i_0+1}(i \mid n) \equiv \langle \mathbf{u}_{i_0}, P_{M_i(0,n-1)}\mathbf{u}_i \rangle = \mathbf{y}_{i \mid n}^T \Phi_{i_0+n,i \mid n}^{-1}\mathbf{y}_{i_0+n \mid n}. \tag{55c}$$

Notice that

$$\mathbf{S}_i(0, n-1)\mathbf{g}_{i_0+1}(i \mid n) = P_{M_i(0,n-1)}\mathbf{u}_i; \tag{56a}$$

and

$$\mathbf{S}_i(0, n-1)\mathbf{h}_{i_0+1}(i \mid n) = P_{M_i(0,n-1)}\mathbf{u}_{i_0}; \tag{56b}$$

and that

$$\gamma_{i_0+1}(i \mid n) = \mathbf{g}_{i_0+1}^T(i \mid n)\mathbf{y}_{i \mid n}, \tag{57a}$$

$$\gamma_{i_0+1}^*(i \mid n) = \mathbf{h}_{i_0+1}^T(i \mid n)\mathbf{y}_{i_0+n \mid n}, \tag{57b}$$

and

$$\alpha_{i_0+1}(i \mid n) = \mathbf{g}_{i_0+1}^T(i \mid n)\mathbf{y}_{i_0+n \mid n} = \mathbf{h}_{i_0+1}^T(i \mid n)\mathbf{y}_{i \mid n}. \tag{57c}$$

Using the notation in the last section, the gains γ and γ^* are, respectively, $\sin^2\theta_i$ and $\sin^2\theta_i^*$, where θ_i and θ_i^* are, respectively, the angles between $M_i(0, n-1)$ and $\tilde{M}_i(0, n-1)$, and between $M_i(0, n-1)$ and $M_i^*(0, n-1)$.

At each time instant our objective is to minimize the cost functions

$\epsilon_{f,i_0}(i \mid n)$ and $\epsilon_{b,i_0}(i \mid n)$. From the discussion in Section III it follows that

$$\mathbf{E}_{f,i_0}(i \mid n) = P^{\perp}_{M_i(1,n)} \mathbf{Y}_{i_0+n,i} \tag{58a}$$

and

$$\mathbf{E}_{b,i_0}(i \mid n) = P^{\perp}_{M_i(0,n-1)}(z^{-n}\mathbf{Y}_{i_0+n,i}). \tag{58b}$$

The following variables, which are closely related to the prediction residuals, are also needed:

$$\begin{aligned}\mathbf{E}'_{f,i_0}(i \mid n) &\equiv P^{\perp}_{M_{i\mid i-1}(1,n)} \mathbf{Y}_{i_0+n,i} \\ &= \mathbf{Y}_{i_0+n,i} - \mathbf{S}_i(1, n)\mathbf{f}_{i_0}(i-1 \mid n),\end{aligned} \tag{59a}$$

and

$$\begin{aligned}\mathbf{E}'_{b,i_0}(i \mid n) &\equiv P^{\perp}_{M_{i\mid i-1}(0,n-1)}(z^{-n}\mathbf{Y}_{i_0+n,i}) \\ &= z^{-n}\mathbf{Y}_{i_0+n,i} - \mathbf{S}_i(0, n-1)\mathbf{b}_{i_0}(i-1 \mid n),\end{aligned} \tag{59b}$$

i.e., $\mathbf{E}'_f$ and $\mathbf{E}'_b$ are the forward and backward residual vectors obtained by using prediction vectors computed at the *previous* time interval. The top components of $\mathbf{E}'_{f,i_0}(i \mid n)$ and $\mathbf{E}'_{b,i_0}(i \mid n)$ are, respectively,

$$\begin{aligned}e'_{f,i_0}(i \mid n) &\equiv \langle \mathbf{u}_i, \mathbf{E}'_{f,i_0}(i \mid n)\rangle \\ &= y_i - \mathbf{f}^T_{i_0}(i-1 \mid n)\mathbf{y}_{i-1\mid n}\end{aligned} \tag{60a}$$

and

$$\begin{aligned}e'_{b,i_0}(i \mid n) &\equiv \langle \mathbf{u}_i, \mathbf{E}'_{b,i_0}(i \mid n)\rangle \\ &= y_{i-n} - \mathbf{b}^T_{i_0}(i-1 \mid n)\mathbf{y}_{i\mid n}.\end{aligned} \tag{60b}$$

The nth order forward prediction residual computed at time $i_0 + n$ using the tap vector $\mathbf{f}_{i_0}(i \mid n)$ is

$$\begin{aligned}e^*_{f,i_0}(i \mid n) &\equiv \langle \mathbf{u}_{i_0}, \mathbf{E}_{f,i_0}(i \mid n)\rangle \\ &= y_{i_0+n} - \mathbf{f}^T_{i_0}(i \mid n)\mathbf{y}_{i_0+n-1\mid n}.\end{aligned} \tag{61}$$

The forward residual vector at time i using the forward prediction vector calculated from the data samples $\{y_{i_0+1}, \cdots, y_i\}$ is

$$\begin{aligned}\mathbf{E}_{f,i_0\mid i_0+1}(i \mid n) &\equiv P^{\perp}_{M_{i_0\mid i_0+1}(1,n)} \mathbf{Y}_{i_0+n,i} \\ &= \mathbf{Y}_{i_0+n,i} - \mathbf{S}_i(1, n)\mathbf{f}_{i_0+1}(i \mid n).\end{aligned} \tag{62}$$

The variables $e^*_{b,i_0}(i \mid n)$ and $\mathbf{E}_{b,i_0\mid i_0+1}(i \mid n)$ are similarly defined.

Notice that the time indices associated with a residual vector change in accordance with the projection space, i.e.,

$$P^{\perp}_{M_i(1,n-1)} \mathbf{Y}_{i_0+n,i} = \mathbf{E}_{f,i_0+1}(i \mid n-1), \tag{63a}$$

and

$$P^{\perp}_{M_i(1,n-1)}(z^{-n}\mathbf{Y}_{i_0+n,i}) = \mathbf{E}_{b,i_0}(i-1\,|\,n-1). \tag{63b}$$

The recursions needed to derive the algorithms in the next section are now generated systematically. By appropriately defining the vectors and subspaces entering the projection order update (22), order updates are derived for all of the basic variables defined by (49) through (55). We then use the forward and backward time updates (34) and (43) to obtain forward and backward time updates for the basic vectors defined by (49), (50), and (54). Finally, the forward and backward time updates for inner products (35) and (44) are applied to $k_{n,i_0}(i)$, $\epsilon_{f,i_0}(i\,|\,n)$, and $\epsilon_{b,i_0}(i\,|\,n)$. It would take up too much space to explicitly define the vectors and subspaces that must be substituted in the projection update used to derive each recursion. Consequently, only the results are stated, with a few representative examples worked out in more detail.

5.2 Order updates

The following order updates are obtained by using the projection order update (22) [or equivalently (21)]. The lth through the mth component of $\mathbf{f}_{i_0}(i\,|\,n)$ is denoted by $[\mathbf{f}_{i_0}(i\,|\,n)]_{l,m}$, and $[\mathbf{f}_{i_0}(i\,|\,n)]_j$ is the jth component of $\mathbf{f}_{i_0}^{j}(i\,|\,n)$. The same notation is used for the backward prediction vector $\mathbf{b}_{i_0}(i\,|\,n)$ and the gain vectors $\mathbf{g}_{i_0}(i\,|\,n)$ and $\mathbf{h}_{i_0}(i\,|\,n)$.

$$\mathbf{E}_{f,i_0}(i\,|\,n) = \mathbf{E}_{f,i_0+1}(i\,|\,n-1) - \frac{k_{n,i_0}(i)}{\epsilon_{b,i_0}(i-1\,|\,n-1)} \cdot \mathbf{E}_{b,i_0}(i-1\,|\,n-1), \tag{64a}$$

$$\mathbf{E}_{b,i_0}(i\,|\,n) = \mathbf{E}_{b,i_0}(i-1\,|\,n-1) - \frac{k_{n,i_0}(i)}{\epsilon_{f,i_0+1}(i\,|\,n-1)} \cdot \mathbf{E}_{f,i_0+1}(i\,|\,n-1), \tag{64b}$$

$$\epsilon_{f,i_0}(i\,|\,n) = \epsilon_{f,i_0+1}(i\,|\,n-1) - \frac{k^2_{n,i_0}(i)}{\epsilon_{b,i_0}(i-1\,|\,n-1)}, \tag{65a}$$

$$\epsilon_{b,i_0}(i\,|\,n) = \epsilon_{b,i_0}(i-1\,|\,n-1) - \frac{k^2_{n,i_0}(i)}{R_{f,i_0+1}(i\,|\,n-1)}, \tag{65b}$$

$$[\mathbf{f}_{i_0}(i\,|\,n)]_n = \frac{k_{n,i_0}(i)}{\epsilon_{b,i_0}(i-1\,|\,n-1)}, \tag{66a}$$

$$[\mathbf{f}_{i_0}(i\,|\,n)]_{1,n-1} = \mathbf{f}_{i_0+1}(i\,|\,n-1) - [\mathbf{f}_{i_0}(i\,|\,n)]_n \mathbf{b}_{i_0}(i-1\,|\,n-1), \tag{66b}$$

$$[\mathbf{b}_{i_0}(i \mid n)]_1 = \frac{k_{n,i_0}(i)}{\epsilon_{f,i_0+1}(i \mid n-1)}, \tag{67a}$$

$$[\mathbf{b}_{i_0}(i \mid n)]_{2,n} = \mathbf{b}_{i_0}(i-1 \mid n-1) - [\mathbf{b}_{i_0}(i \mid n)]_1 \mathbf{f}_{i_0+1}(i \mid n-1), \tag{67b}$$

$$[\mathbf{g}_{i_0}(i \mid n+1)]_{n+1} = \frac{e_{b,i_0}(i \mid n)}{\epsilon_{b,i_0}(i \mid n)}, \tag{68a}$$

$$[\mathbf{g}_{i_0}(i \mid n+1)]_{1,n} = \mathbf{g}_{i_0+1}(i \mid n) - [\mathbf{g}_{i_0}(i \mid n+1)]_{n+1} \mathbf{b}_{i_0}(i \mid n), \tag{68b}$$

$$[\mathbf{g}_{i_0}(i \mid n+1)]_1 = \frac{e_{f,i_0}(i \mid n)}{\epsilon_{f,i_0}(i \mid n)}, \tag{69a}$$

$$[\mathbf{g}_{i_0}(i \mid n+1)]_{2,n+1} = \mathbf{g}_{i_0}(i-1 \mid n) - [\mathbf{g}_{i_0}(i \mid n+1)]_1 \mathbf{f}_{i_0}(i \mid n), \tag{69b}$$

$$[\mathbf{h}_{i_0}(i \mid n+1)]_{n+1} = \frac{e^*_{b,i_0}(i \mid n)}{\epsilon_{b,i_0}(i \mid n)}, \tag{70a}$$

$$[\mathbf{h}_{i_0}(i \mid n+1)]_{1,n} = \mathbf{h}_{i_0+1}(i \mid n) - [\mathbf{h}_{i_0}(i \mid n+1)]_{n+1} \mathbf{b}_{i_0}(i \mid n), \tag{70b}$$

$$[\mathbf{h}_{i_0}(i \mid n+1)]_1 = \frac{e^*_{f,i_0}(i \mid n)}{\epsilon_{f,i_0}(i \mid n)}, \tag{71a}$$

$$[\mathbf{h}_{i_0}(i \mid n+1)]_{2,n+1} = \mathbf{h}_{i_0}(i-1 \mid n) - [\mathbf{h}_{i_0}(i \mid n+1)]_1 \mathbf{f}_{i_0}(i \mid n), \tag{71b}$$

$$\gamma_{i_0}(i \mid n+1) = \gamma_{i_0+1}(i \mid n) + \frac{e^2_{b,i_0}(i \mid n)}{\epsilon_{b,i_0}(i \mid n)}, \tag{72a}$$

$$\gamma_{i_0}(i \mid n+1) = \gamma_{i_0}(i-1 \mid n) + \frac{e^2_{f,i_0}(i \mid n)}{\epsilon_{f,i_0}(i \mid n)}, \tag{72b}$$

$$\gamma^*_{i_0}(i \mid n+1) = \gamma^*_{i_0+1}(i \mid n) + \frac{e^{*2}_{b,i_0}(i \mid n)}{\epsilon_{b,i_0}(i \mid n)}, \tag{73a}$$

$$\gamma^*_{i_0}(i \mid n+1) = \gamma^*_{i_0}(i-1 \mid n) + \frac{e^{*2}_{f,i_0}(i \mid n)}{\epsilon_{f,i_0}(i \mid n)}, \tag{73b}$$

$$\alpha_{i_0}(i \mid n+1) = \alpha_{i_0+1}(i \mid n) + \frac{e_{b,i_0}(i \mid n) e^*_{b,i_0}(i \mid n)}{\epsilon_{b,i_0}(i \mid n)}, \tag{74a}$$

and

$$\alpha_{i_0}(i \mid n+1) = \alpha_{i_0}(i-1 \mid n) + \frac{e_{f,i_0}(i \mid n) e^*_{f,i_0}(i \mid n)}{\epsilon_{f,i_0}(i \mid n)}. \tag{74b}$$

As an example, (64a) is derived from (22), where M is replaced by $M_i(1, n-1)$, $\mathbf{X}$ is replaced by $z^{-n}\mathbf{Y}_{i_0+n,i}$, and $\mathbf{Y}$ is replaced by $\mathbf{Y}_{i_0+n,i}$.

By observing that $\mathbf{E}_{b,i_0}(i-1\,|\,n-1)$ is orthogonal to $M_i(1, n-1)$, it is clear that

$$\langle \mathbf{Y}_{i_0+n,i}, \mathbf{E}_{b,i_0}(i-1\,|\,n-1)\rangle = \langle \mathbf{E}_{f,i_0+1}(i\,|\,n-1), \cdot \mathbf{E}_{b,i_0}(i-1\,|\,n-1)\rangle = k_{n,i_0}(i). \tag{75}$$

Recursions (65) are obtained by taking norms of (64) respectively. The recursions (68) through (71) are obtained from (21), where $\mathbf{Y}$ is replaced by $\mathbf{u}_i$ and $\mathbf{u}_{i_0}$, respectively. Making the same substitutions in (21) and then taking inner products with $\mathbf{u}_i$ or $\mathbf{u}_{i_0}$ gives recursions (72) through (74).

5.3 Forward time updates

The following forward time updates are obtained from the (orthogonal) projection operator forward time update (34):

$$\mathbf{E}_{f,i_0}(i\,|\,n) = \mathbf{E}'_{f,i_0}(i\,|\,n) - [P_{M_i(1,n)}\mathbf{u}_i]\frac{e_{f,i_0}(i\,|\,n)}{1-\gamma_{i_0}(i-1\,|\,n)}, \tag{76a}$$

$$\mathbf{E}_{b,i_0}(i\,|\,n) = \mathbf{E}'_{b,i_0}(i\,|\,n) - [P_{M_i(0,n-1)}\mathbf{u}_i]\frac{e_b(i\,|\,n)}{1-\gamma_{i_0+1}(i\,|\,n)}, \tag{76b}$$

$$\mathbf{f}_{i_0}(i\,|\,n) = \mathbf{f}_{i_0}(i-1\,|\,n) + \mathbf{g}_{i_0}(i-1\,|\,n)\frac{e_{f,i_0}(i\,|\,n)}{1-\gamma_{i_0}(i-1\,|\,n)}, \tag{77a}$$

$$\mathbf{b}_{i_0}(i\,|\,n) = \mathbf{b}_{i_0}(i-1\,|\,n) + \mathbf{g}_{i_0+1}(i\,|\,n)\frac{e_{b,i_0}(i\,|\,n)}{1-\gamma_{i_0+1}(i\,|\,n)}, \tag{77b}$$

$$\mathbf{h}_{i_0}(i\,|\,n) = \mathbf{h}_{i_0}(i-1\,|\,n) - \mathbf{g}_{i_0}(i\,|\,n)\frac{\alpha_{i_0}(i\,|\,n)}{1-\gamma_{i_0}(i\,|\,n)}, \tag{78}$$

$$\gamma^*_{i_0}(i\,|\,n) = \gamma^*_{i_0}(i-1\,|\,n) - \frac{\alpha^2_{i_0}(i\,|\,n)}{1-\gamma_{i_0}(i\,|\,n)}, \tag{79}$$

and

$$\alpha_{i_0}(i\,|\,n) = \mathbf{y}^T_{i|n}\mathbf{h}_{i_0}(i-1\,|\,n)[1-\gamma_{i_0}(i\,|\,n)]. \tag{80}$$

Equation (78) is obtained from (34), where M_i is replaced $M_i(0, n)$ and $\mathbf{Y}_i$ is replaced by $\mathbf{u}_{i_0}$. Equations (79) and (80) are obtained by making the same substitutions in (34) and then taking inner products with $\mathbf{u}_{i_0}$ and $\mathbf{u}_i$, or by premultiplying (78) by $\mathbf{y}^T_{i_0+n-1|n}$ and $\mathbf{y}^T_{i|n}$, respectively. Taking the inner product of (76) with $\mathbf{u}_i$ and $\mathbf{u}_{i_0}$, respectively gives the following recursions:

$$e'_{f,i_0}(i\,|\,n) = \frac{e_{f,i_0}(i\,|\,n)}{1 - \gamma_{i_0}(i-1\,|\,n)}, \tag{81a}$$

$$e'_{b,i_0}(i\,|\,n) = \frac{e_{b,i_0}(i\,|\,n)}{1 - \gamma_{i_0+1}(i\,|\,n)}, \tag{81b}$$

$$e^*_{f,i_0}(i\,|\,n) = e^*_{f,i_0}(i-1\,|\,n) - e_{f,i_0}(i\,|\,n)\,\frac{\alpha_{i_0}(i-1\,|\,n)}{1 - \gamma_{i_0}(i-1\,|\,n)}, \tag{82a}$$

and

$$e^*_{b,i_0}(i\,|\,n) = e^*_{b,i_0}(i-1\,|\,n) - e_{b,i_0}(i\,|\,n)\,\frac{\alpha_{i_0+1}(i\,|\,n)}{1 - \gamma_{i_0+}(i\,|\,n)}. \tag{82b}$$

5.4 Backward time updates

The following backward time updates are obtained from the projection operator backward time update (43):

$$\mathbf{E}_{f,i_0}(i\,|\,n) = \mathbf{E}_{f,i_0|\,i_0+1}(i\,|\,n) - [P_{M_i(1,n)}\mathbf{u}_{i_0}]\,\frac{e^*_{f,i_0}(i\,|\,n)}{1 - \gamma^*_{i_0}(i-1\,|\,n)}, \tag{83a}$$

$$\mathbf{E}_{b,i_0}(i\,|\,n) = \mathbf{E}_{b,i_0|\,i_0+1}(i\,|\,n) - [P_{M_i(0,n-1)}\mathbf{u}_{i_0}]\,\frac{e^*_{b,i_0}(i\,|\,n)}{1 - \gamma^*_{i_0+1}(i\,|\,n)}, \tag{83b}$$

$$\mathbf{f}_{i_0+1}(i\,|\,n) = \mathbf{f}_{i_0}(i\,|\,n) - \mathbf{h}_{i_0}(i-1\,|\,n)\,\frac{e^*_{f,i_0}(i\,|\,n)}{1 - \gamma^*_{i_0}(i-1\,|\,n)}, \tag{84a}$$

$$\mathbf{b}_{i_0+1}(i\,|\,n) = \mathbf{b}_{i_0}(i\,|\,n) - \mathbf{h}_{i_0+1}(i\,|\,n)\,\frac{e^*_{b,i_0}(i\,|\,n)}{1 - \gamma^*_{i_0+1}(i\,|\,n)}, \tag{84b}$$

$$\mathbf{g}_{i_0}(i\,|\,n) = \mathbf{g}_{i_0+1}(i\,|\,n) - \mathbf{h}_{i_0}(i\,|\,n)\,\frac{\alpha_{i_0}(i\,|\,n)}{1 - \gamma^*_{i_0}(i\,|\,n)}, \tag{85}$$

$$\gamma_{i_0}(i\,|\,n) = \gamma_{i_0+1}(i\,|\,n) - \frac{\alpha^2_{i_0}(i\,|\,n)}{1 - \gamma^*_{i_0}(i\,|\,n)}, \tag{86}$$

and

$$\alpha_{i_0}(i\,|\,n) = \mathbf{y}^T_{i_0+n-1|\,n}\mathbf{g}_{i_0+1}(i\,|\,n)[1 - \gamma^*_{i_0}(i\,|\,n)]. \tag{87}$$

Equation (85) is obtained by replacing $\mathbf{Y}_i$ by $\mathbf{u}_i$ in (43). Equations (86) and (87) are obtained by premultiplying (85) by $\mathbf{y}^T_{i|\,n}$ and $\mathbf{y}^T_{i_0+n-1|\,n}$, respectively. The following recursions are obtained by taking the inner product of (83) with $\mathbf{u}_i$, respectively:

$$e_{f,i_0+1}(i\,|\,n) = e_{f,i_0}(i\,|\,n) + e^*_{f,i_0}(i\,|\,n)\,\frac{\alpha_{i_0}(i-1\,|\,n)}{1 - \gamma^*_{i_0}(i-1\,|\,n)}, \tag{88a}$$

and

$$e_{b,i_0+1}(i \mid n) = e_{b,i_0}(i \mid n) + e^*_{b,i_0}(i \mid n) \frac{\alpha_{i_0+1}(i \mid n)}{1 - \gamma^*_{i_0+1}(i \mid n)}. \tag{88b}$$

The recursions that result from taking inner products with $\mathbf{u}_{i_0}$ will not be used and are therefore omitted.

5.5 *Inner product updates*

The following recursions are obtained from the forward time update for inner products (35):

$$k_{n,i_0}(i) = k_{n,i_0}(i-1) + e_{f,i_0+1}(i \mid n-1) e_{b,i_0}(i-1 \mid n-1) \cdot \frac{1}{1 - \gamma_{i_0+1}(i-1 \mid n-1)}, \tag{89}$$

$$\epsilon_{f,i_0}(i \mid n) = \epsilon_{f,i_0}(i-1 \mid n) + e^2_{f,i_0}(i \mid n) \frac{1}{1 - \gamma_{i_0}(i-1 \mid n)}, \tag{90a}$$

and

$$\epsilon_{b,i_0}(i \mid n) = \epsilon_{b,i_0}(i-1 \mid n) + e^2_{b,i_0}(i \mid n) \frac{1}{1 - \gamma_{i_0+1}(i \mid n)}. \tag{90b}$$

The following recursions are obtained from the backward time update for inner products (44):

$$k_{n,i_0}(i) = k_{n,i_0+1}(i) + e^*_{f,i_0+1}(i \mid n-1) e^*_{b,i_0}(i-1 \mid n-1) \cdot \frac{1}{1 - \gamma^*_{i_0+1}(i-1 \mid n-1)}, \tag{91}$$

$$\epsilon_{f,i_0}(i \mid n) = \epsilon_{f,i_0+1}(i \mid n) + e^{*2}_{f,i_0}(i \mid n) \frac{1}{1 - \gamma^*_{i_0}(i-1 \mid n)}, \tag{92a}$$

and

$$\epsilon_{b,i_0}(i \mid n) = \epsilon_{b,i_0+1}(i \mid n) + e^{*2}_{b,i_0}(i \mid n) \frac{1}{1 - \gamma^*_{i_0+1}(i \mid n)}. \tag{92b}$$

Equations (89) and (91) are obtained by using (35) and (44), where $\mathbf{v}_i$ is replaced by $\mathbf{Y}_{i_0+n,i}$, $\mathbf{Y}_i$ is replaced by $z^{-n}\mathbf{Y}_{i_0+n,i}$, and M_i is replaced by $M_i(1, n-1)$, respectively. The previous set of recursions (64) through (92) are complete in the sense that any existing least-squares alogrithm can be derived by manipulating suitable subsets of these recursions.

VI. RECURSIVE FIXED-ORDER COVARIANCE ALGORITHMS

6.1 Sliding-window algorithm

Combining (60), (61), (68) through (73), (77), (81a), (84), (90a), and (92a), gives the following sliding-window LS algorithm for the prediction coefficients. Where unspecified, the order of the variable is assumed to be N, the order of the least-squares filter. Also, the starting time index is denoted as i_0. If the sliding window contains M data values, then $i_0 = i - M + 1$,

$$e'_{f,i_0}(i) = y_i - \mathbf{f}_{i_0}^T(i-1)\mathbf{y}_{i-1}, \tag{93a}$$

$$\mathbf{f}_{i_0}(i) = \mathbf{f}_{i_0}(i-1) + \mathbf{g}_{i_0}(i-1)e'_{f,i_0}(i), \tag{93b}$$

$$e_{f,i_0}(i) = e'_{f,i_0}(i)[1 - \gamma_{i_0}(i-1)], \tag{93c}$$

$$e^*_{f,i_0}(i) = y_{i_0+N} - \mathbf{f}_{i_0}^T(i)\mathbf{y}_{i_0+N-1}, \tag{93d}$$

$$\epsilon_{f,i_0}(i) = \epsilon_{f,i_0}(i-1) + e'_{f,i_0}(i)e_{f,i_0}(i), \tag{93e}$$

$$\mathbf{g}_{i_0}(i \mid N+1)]_1 = \frac{e_{f,i_0}(i)}{\epsilon_{f,i_0}(i)}, \tag{93f}$$

$$[\mathbf{g}_{i_0}(i \mid N-1)]_{2,N+1} = \mathbf{g}_{i_0}(i-1) - [\mathbf{g}_{i_0}(i \mid N+1)]_1\mathbf{f}_{i_0}(i), \tag{93g}$$

$$[\mathbf{h}_{i_0}(i \mid N+1)]_1 = \frac{e^*_{f,i_0}(i)}{\epsilon_{f,i_0}(i)}, \tag{93h}$$

$$[\mathbf{h}_{i_0}(i \mid N+1)]_{2,N+1} = \mathbf{h}_{i_0}(i-1) - [\mathbf{h}_{i_0}(i \mid N+1)]_1\mathbf{f}_{i_0}(i), \tag{93i}$$

$$\mathbf{f}_{i_0+1}(i) = \mathbf{f}_{i_0}(i) - \mathbf{h}_{i_0}(i-1)\frac{e^*_{f,i_0}(i)}{1 - \gamma^*_{i_0}(i-1)}, \tag{93j}$$

$$\epsilon_{f,i_0+1}(i) = \epsilon_{f,i_0}(i) - \frac{e^{*2}_{f,i_0}(i)}{1 - \gamma^*_{i_0}(i-1)}, \tag{93k}$$

$$e'_{b,i_0}(i) = y_{i-N} - \mathbf{b}_{i_0}^T(i-1)\mathbf{y}_i, \tag{93l}$$

$$\mathbf{b}_{i_0}(i) = \frac{\mathbf{b}_{i_0}(i-1) + e'_{b,i_0}(i)[\mathbf{g}_{i_0}(i \mid N+1)]_{1,N}}{1 - e'_{b,i_0}(i)[\mathbf{g}_{i_0}(i \mid N+1)]_{N+1}}, \tag{93m}$$

$$e^*_{b,i_0}(i) = y_{i_0} - \mathbf{b}_{i_0}^T(i)\mathbf{y}_{i_0+N}, \tag{93n}$$

$$\mathbf{g}_{i_0+1}(i) = [\mathbf{g}_{i_0}(i \mid N+1)]_{1,N} + [\mathbf{g}_{i_0}(i \mid N+1)]_{N+1}\mathbf{b}_{i_0}(i), \tag{93o}$$

$$\mathbf{h}_{i_0+1}(i) = [\mathbf{h}_{i_0}(i \mid N+1)]_{1,N} + [\mathbf{b}_{i_0}(i \mid N+1)_{N+1}\mathbf{b}_{i_0}(i), \tag{93p}$$

$$\gamma_{i_0+1}(i) = \frac{\gamma_{i_0}(i-1) + e_{f,i_0}(i)[\mathbf{g}_{i_0}(i \mid N+1)]_1 - e'_{b,i_0}(i)[\mathbf{g}_{i_0}(i \mid N+1)]_{N+1}}{1 - e'_{b,i_0}(i)[\mathbf{g}_{i_0}(i \mid N+1)]_{N+1}}, \tag{93q}$$

$$\gamma^*_{i_0+1}(i) = \gamma^*_{i_0}(i-1) + e^*_{f,i_0}(i)[\mathbf{h}_{i_0}(i \mid N+1)]_1 - e^*_{b,i_0}[\mathbf{h}_{i_0}(i \mid N+1)]_N, \tag{93r}$$

and

$$\mathbf{b}_{i_0+1}(i) = \mathbf{b}_{i_0}(i) - \mathbf{h}_{i_0+1}(i) \frac{e^*_{b,i_0}(i)}{1 - \gamma^*_{i_0+1}(i)}. \tag{93s}$$

The recursions (93m) and (93q) were not listed in the previous section, but are easily obtained by solving (77b) and (68b) simultaneously for $\mathbf{b}_{i_0}(i)$, and by substituting (68a), (69a), and (81b) into (72), and solving for $\gamma_{i_0+1}(i \mid n)$. Notice that all data samples in the sliding window $(y_{i-M+1}, \ldots, y_i)$ must be stored. This is also true of the order-recursive sliding-window algorithm presented in Ref. 14. If division is counted as multiplication, then the algorithm (93) requires $12N + 16$ multiplies and $12N + 12$ additions at each iteration. In contrast, the unnormalized sliding-window lattice predictor (see Appendix B) requires $16N$ multiplies and $10N$ additions per iteration, and the normalized lattice predictor[16] requires $30N$ multiplies, $18N$ additions, and $6N$ square roots per iteration.

Because sliding-window algorithms have finite memory, initialization for these algorithms is basically the same as for the prewindowed case, i.e., the data y_i can be assumed to be zero for $i < 0$. After M iterations, where M is the window length, these data points are discarded. The algorithm (93) is therefore initialized by setting the gains γ and γ^*, and the elements of the vectors **f**, **b**, **g**, and **h** equal to zero, and letting

$$\epsilon_{f,i_0}(0) = \delta, \tag{94}$$

where δ is chosen to ensure that the algorithm remains stable. It is easily verified that for time $i < M - N - 1$, where M is the length of the sliding window, the algorithm (93) becomes a modified version of the prewindowed LS transversal (fast Kalman) algorithm.[1,22]

6.2 *Growing-memory covariance algorithm*

The following fixed-order growing-memory covariance algorithm is obtained by combining (60), (68b), (69), (77), (78), (80), (85), (87), and (90a). The lower index of the window i_0 is assumed to be zero. For notational convenience we define the following variables,

$$\beta_{i_0}(i \mid n) \equiv \frac{\alpha_{i_0}(i \mid n)}{1 - \gamma_{i_0}(i \mid n)} \tag{95a}$$

and

$$\beta^*_{i_0}(i \mid n) \equiv \frac{\alpha_{i_0}(i \mid n)}{1 - \gamma^*_{i_0}(i \mid n)}. \tag{95b}$$

Where unspecified, the lower time index and the order of the variables are equal to zero and N, respectively,*

$$e'_f(i) = y_i = \mathbf{f}^T(i-1)\mathbf{y}_{i-1}, \tag{96a}$$

$$\mathbf{f}(i) = \mathbf{f}(i-1) + \mathbf{g}(i-1)e'_f(i), \tag{96b}$$

$$e_f(i) = y_i - \mathbf{f}^T(i)\mathbf{y}_{i-1}, \tag{96c}$$

$$\epsilon_f(i) = \epsilon_f(i-1) + e_f(i)e'_f(i), \tag{96d}$$

$$[\mathbf{g}(i \mid N+1)]_1 = \frac{e_f(i)}{\epsilon_f(i)}, \tag{96e}$$

$$[\mathbf{g}(i \mid N+1)]_{2,N+1} = \mathbf{g}(i-1) - [\mathbf{g}(i \mid N+1)]_1\mathbf{f}(i), \tag{96f}$$

$$e'_b(i) = y_{i-N} - \mathbf{b}^T(i-1)\mathbf{y}_i, \tag{96g}$$

$$\mathbf{b}(i) = \frac{\mathbf{b}(i-1) + e'_b(i)[\mathbf{g}(i \mid N+1)]_{1,N}}{1 - e'_b(i)[\mathbf{g}(i \mid N+1)]_{N+1}}, \tag{96h}$$

$$\mathbf{g}_1(i) = [\mathbf{g}(i \mid N+1)]_{1,N} + [\mathbf{g}(i \mid N+1)]_{N+1}\mathbf{b}(i), \tag{96i}$$

$$\beta(i) = \mathbf{y}_i^T\mathbf{h}(i-1), \tag{96j}$$

$$\beta^*(i) = \mathbf{y}_{N-1}^T\mathbf{g}_1(i), \tag{96k}$$

$$\mathbf{g}(i) = \frac{\mathbf{g}_1(i) - \beta^*(i)\mathbf{h}(i-1)}{1 - \beta(i)\beta^*(i)}, \tag{96l}$$

and

$$\mathbf{h}(i) = \mathbf{h}(i-1) - \beta(i)\mathbf{g}(i). \tag{96m}$$

Notice that this algorithm can be applied only if $i > N$. Otherwise, the least-squares variables of order N are undefined and cannot be used to compute the same least-squares variables at the successive time interval. Initialization of this algorithm can be performed, however, by using an order-recursive algorithm for $i < N$ to increase the order of the filter by one at each successive time iteration. An order-recursive algorithm for the prediction coefficients is obtained by combining (89), (66a), (67a), top components of (64a) and (64b), (66b), (67b), (65a) and (65b), (82a), (88a), (84a), (92a), (71), (73b), (72), and (74b). This algorithm is basically the same as the covariance lattice

* The author recently discovered that this algorithm has been independently derived in Ref. 23 using an algebraic approach.

algorithm presented in Refs. 7 and 19, except that additional order recursions for the prediction vectors **f** and **b** have been added. It is not explicitly stated in an effort to conserve space. Order-recursive computation of **f** and **b** requires order N^2 arithmetic operations per iteration, rather than order N operations per iteration, as required by the fixed-order algorithm. Not all N components of the vectors **f** and **b** need to be updated at each iteration for $i < N$, however. If data is first received at time $i = 0$, the recursions listed above can be used for $n = 0$ up to $n = i$. At time $i = N$ all of the variables that enter the fixed-order algorithm (96) have been computed by the order-recursive algorithm except for $\mathbf{g}(i)$, $\beta(i)$ and $\beta^*(i)$. The gain $\mathbf{g}(i)$ is the only variable needed at the next iteration of the fixed-order algorithm and can be computed by first using (96j) to calculate $\beta(i)$ and then using (96m) to solve for $\mathbf{g}(i)$.

Derivation of initial conditions for the order-recursive initialization routine is significantly more complicated than for the sliding-window algorithm. This is because for $i = n$, the matrix $\Phi_{n,i|n}$ is guaranteed to be singular, and hence all variables are technically undefined. Reference 14 gives a convenient solution to this startup problem. By using a generalized inverse of a singular or nonsingular matrix, the least-squares projection operator P, given by (15), can be defined even when the matrix $\mathbf{S}^T\mathbf{S}$ is singular. If this generalized inverse is defined appropriately, it can be shown that the projection updates in Section IV hold even when the covariance matrix is singular. This implies that all of the recursions listed in the last paragraph that constitute the order-recursive initialization routine can be used starting from $i = 0$ with the following initial conditions:

$$\mathbf{f}(0\,|\,0) = \mathbf{b}(0\,|\,0) = \mathbf{f}_1(-1\,|\,0) = \mathbf{h}(-1\,|\,0) = \mathbf{0}, \tag{97a}$$

$$k_n(-1) = 0, \qquad 1 \leqslant n \leqslant N, \tag{97b}$$

$$\gamma(-1\,|\,0) = \gamma^*(-1\,|\,0) = \alpha(-1\,|\,0) = 0, \tag{97c}$$

and

$$e_f^*(0\,|\,n) = \begin{cases} y_0 & \text{for} \quad n = 0 \\ 0 & \text{for} \quad n > 0. \end{cases} \tag{97d}$$

At each iteration $i < N$,

$$e_f(i\,|\,0) = e_b(i\,|\,0) = y_i \tag{97e}$$

and

$$\epsilon_f(i\,|\,0) = \epsilon_b(i\,|\,0) = \epsilon_f(i-1\,|\,0) + y_i^2. \tag{97f}$$

Counting division as multiplication, the algorithm (96) requires

$11N + 7$ multiplies and $11N + 1$ additions per iteration. To compare, the unnormalized growing-memory lattice predictor (see Appendix B) requires $22N$ multiplies and $12N$ additions per iteration. The normalized lattice algorithm requires $30N$ multiplies, $18N$ additions, and $6N$ square roots per iteration. We point out that the fixed-order covariance algorithm specified by (96) is not unique. In particular, equation (96c) can be replaced by (81a). The extra recursion (93q) must be added to compute γ, however. This type of modification has been applied to the fast Kalman algorithm, and has resulted in improved numerical properties.[22]

VII. EXTENSIONS TO JOINT-PROCESS ESTIMATION

The algorithms presented so far solve the LS prediction problem wherein the sums (1) and (2) are minimized. In applications such as channel equalization, echo and noise cancellation, and adaptive line enhancement, two processes, $\{x_j\}$ and $\{y_j\}$, are given, and our objective is to estimate the $\{x_j\}$ process in terms of the $\{y_j\}$ process. The vector of estimation errors is denoted as

$$\begin{aligned}\mathbf{E}_{x,i_0+1}(i \mid n) &\equiv \mathbf{X}_{i_0+n,i} - \sum_{j=0}^{n-1} c_{j+1\mid n}(z^{-j}\mathbf{Y}_{i_0+n,i}) \\ &= \mathbf{X}_{i_0+n,i} - \mathbf{S}_i(0,\, n-1)\mathbf{c}_{i_0+1}(i \mid n), \end{aligned} \tag{98}$$

where $\mathbf{X}_{i_0,i}$ is defined by (23a), $\mathbf{c}_{i_0+1}(i \mid n)$ is the n-dimensional vector of regression coefficients at time i used to estimate $\mathbf{X}_{i_0+n,i}$ [given by (4c)] where $i' = i_0 + n$), and the lower time subscript of $\mathbf{E}_x$ and $\mathbf{c}$ denotes the time index of the starting value from the y sequence (i.e., y_{i_0+1}), which is used in the least-squares computation. Our objective is to choose $\mathbf{c}_{i_0+1}(i \mid n)$ such that

$$\epsilon_{x,i_0+1}(i \mid n) \equiv \| \mathbf{E}_{x,i_0+1}(i \mid n) \|^2 \tag{99}$$

is minimized. The discussion in Section III implies that

$$\mathbf{E}_{x,i_0+1}(i \mid n) = P^{\perp}_{M_i(0,n-1)}\mathbf{X}_{i_0+n,i}. \tag{100}$$

We now use the projection recursions in Section IV to derive order and time updates for $\mathbf{E}_{x,i_0}(i \mid n)$ and $\mathbf{c}_{i_0}(i \mid n)$. Details are again omitted since they are basically the same as before. Combining recursions in this section with the prediction algorithms of the last section results in recursive algorithms that solve the LS joint-process-estimation problem.

The following notation, which is analogous to the notation in Section 5.1, is first defined:

1. Cross-correlation coefficient,

$$k^{(x)}_{n+1,i_0}(i) \equiv \langle \mathbf{X}_{i_0+n,i},\ \mathbf{E}_{b,i_0}(i \mid n)\rangle$$
$$= \langle \mathbf{E}_{x,i_0+1}(i \mid n),\ \mathbf{E}_{b,i_0}(i \mid n)\rangle. \tag{101}$$

2. Current residual (scalar),

$$e_{x,i_0}(i \mid n) = \langle \mathbf{u}_i,\ \mathbf{E}_{x,i_0}(i \mid n)\rangle$$
$$= x_i - \mathbf{c}^T_{i_0}(i \mid n)\mathbf{y}_{i \mid n}. \tag{102}$$

3. Past residual (scalar),

$$e^*_{x,i_0+1}(i \mid n) = \langle \mathbf{u}_{i_0},\ \mathbf{E}_{x,i_0+1}(i \mid n)\rangle$$
$$= x_{i_0+n} - \mathbf{c}^T_{i_0+1}(i \mid n)\mathbf{y}_{i_0+n \mid n}. \tag{103}$$

4. Oblique residual

$$e'_{x,i_0}(i \mid n) \equiv x_i - \mathbf{c}^T_{i_0}(i-1 \mid n)\mathbf{y}_{i \mid n}. \tag{104}$$

The following order recursions are obtained from (22):

$$\mathbf{E}_{x,i_0}(i \mid n+1) = \mathbf{E}_{x,i_0+1}(i \mid n) - \frac{k^{(x)}_{n+1,i_0}(i)}{\epsilon_{b,i_0}(i \mid n)}\,\mathbf{E}_{b,i_0}(i \mid n), \tag{105}$$

$$\epsilon_{x,i_0}(i \mid n+1) = \epsilon_{x,i_0+1}(i \mid n) - \frac{k^{(x)2}_{n+1,i_0}(i)}{\epsilon_{b,i_0}(i \mid n)}, \tag{106}$$

$$[\mathbf{c}_{i_0}(i \mid n+1)]_{n+1} = \frac{k^{(x)}_{n+1,i_0}(i)}{\epsilon_{b,i_0}(i \mid n)}, \tag{107a}$$

and

$$[\mathbf{c}_{i_0}(i \mid n+1)]_{1,n} = \mathbf{c}_{i_0+1}(i \mid n) - [\mathbf{c}_{i_0}(i \mid n+1)]_{n+1}\mathbf{b}_{i_0}(i \mid n). \tag{107b}$$

Derivation of the following forward time updates involves a straightforward application of (34) and (35), where $\mathbf{Y}_i$ is replaced by $\mathbf{X}_{i_0+n,i}$ and M_i is replaced by $M_i(0, n-1)$:

$$\mathbf{c}_{i_0}(i \mid n) = \mathbf{c}_{i_0}(i-1 \mid n) + e'_{x,i_0}(i \mid n)\mathbf{g}_{i_0}(i \mid n), \tag{108}$$

$$k^{(x)}_{n+1,i_0}(i) = k^{(x)}_{n+1,i_0}(i-1) + e_{x,i_0+1}(i \mid n)e_{b,i_0}(i \mid n)\,\frac{1}{1-\gamma_{i_0+1}(i \mid n)}, \tag{109}$$

$$\epsilon_{x,i_0}(i \mid n) = \epsilon_{x,i_0}(i-1 \mid n) + e^2_{x,i_0}(i \mid n)\,\frac{1}{1-\gamma_{i_0}(i \mid n)}, \tag{110}$$

$$e^*_{x,i_0}(i \mid n) = e^*_{x,i_0}(i-1 \mid n) - e_{x,i_0}(i \mid n)\,\frac{\alpha_{i_0}(i \mid n)}{1-\gamma_{i_0}(i \mid n)}, \tag{111}$$

and

$$e'_{x,i_0}(i \mid n) = \frac{e_{x,i_0}(i \mid n)}{1 - \gamma_{i_0}(i \mid n)}. \tag{112}$$

Similarly, the following backward time updates are obtained from (43) and (44):

$$\mathbf{c}_{i_0}(i \mid n) = \mathbf{c}_{i_0+1}(i \mid n) + \mathbf{h}_{i_0}(i \mid n) \frac{e^*_{x,i_0}(i \mid n)}{1 - \gamma^*_{i_0}(i \mid n)}, \tag{113}$$

$$e_{x,i_0+1}(i \mid n) = e_{x,i_0}(i \mid n) + e^*_{x,i_0}(i \mid n) \frac{\alpha_{i_0}(i \mid n)}{1 - \gamma^*_{i_0}(i \mid n)}, \tag{114}$$

$$k^{(x)}_{n+1,i_0+1}(i) = k^{(x)}_{n+1,i_0}(i) - e^*_{x,i_0+1}(i \mid n) e^*_{b,i_0}(i \mid n) \frac{1}{1 - \gamma^*_{i_0+1}(i \mid n)}, \tag{115}$$

and

$$\epsilon_{x,i_0+1}(i \mid n) = \epsilon_{x,i_0}(i \mid n) - e^{*2}_{x,i_0}(i \mid n) \frac{1}{1 - \gamma^*_{i_0}(i \mid n)}. \tag{116}$$

Combining (104), (108), (103), and (113) (in that order) with the fixed-order sliding-window algorithm (93) gives the corresponding sliding-window joint-process-estimation algorithm. Adding these additional recursions results in a total computational complexity of $16N + 17$ multiplies and $16N + 13$ additions per iteration. This should be compared with $23N$ multiplies and $14N$ additions per iteration required by the unnormalized sliding-window lattice joint-process estimator. Initialization of these additional recursions is accomplished in a fashion analogous to the prediction recursions. In particular, the data y_i and x_i is assumed to be zero for $i < 0$, and $\mathbf{c}_{i_0}(-1 \mid n) = \mathbf{0}$.

The fixed-order growing-memory algorithm (96) is extended to the joint-process-estimation case by adding the recursions (104) and (108). The order-recursive prediction algorithm listed in Section 6.2 is extended to the joint-process-estimation case by adding the recursions (105) (top component only), (109), (107), (111), and (113). In each case the variable $i_0 = 0$. Adding (104) and (108) to (96) results in a total computational complexity of $13N + 7$ multiplies and $13N + 1$ additions per iteration. This should be compared with $28N$ multiplies and $16N$ additions per iterations required by the growing-memory covariance lattice joint-process estimator. The following accomplishes the initialization of the additional recursions for the order-recursive algorithm:

$$k^{(x)}_n(-1) = 0, \qquad 1 \leq n \leq N, \tag{117a}$$

$$\mathbf{c}(-1 \mid n) = \mathbf{0}, \qquad 0 \leq n \leq N, \tag{117b}$$

and

$$e_x^*(0 \mid n) = \begin{cases} x_0 & \text{for } n = 0 \\ 0 & \text{for } n > 0. \end{cases} \tag{117c}$$

The fixed-order algorithm is initialized by using the order-recursive algorithm for $i < N$.

VIII. CONCLUSIONS

We have presented new fixed-order algorithms that recursively solve the sliding-window and growing-memory covariance least-squares estimation problems. The fixed-order growing-memory algorithm requires approximately one half the number of multiplies and divides required by the analogous unnormalized order-recursive or lattice algorithm. The fixed-order sliding-window algorithm requires approximately 70 percent of the number of multiplies and divides required by the analogous lattice algorithm. These fixed-order algorithms also help complete the list of computationally efficient LS algorithms currently available. In particular, each type of windowing technique that has been proposed for the LS computation (i.e., prewindowed, growing-memory covariance, and the sliding window) has resulted in both computationally efficient fixed-order and order-recursive algorithms. The order-recursive algorithms offer the advantage of being able to dynamically choose the order of the autoregressive model, while the fixed-order algorithms require less computation.

Associated with the algorithms mentioned in this paper are performance issues such as the relative convergence speed of each algorithm given different types of stationary and nonstationary random inputs, and the evaluation of finite word-length effects. As an example, the relative performance improvement offered by LS covariance algorithms over LS prewindowed algorithms has yet to be ascertained in applications where the prediction coefficients must be estimated from relatively few data samples. These issues will play a crucial role in determining the practical value of the LS algorithms presented in this paper.

REFERENCES

1. D. D. Falconer and L. Ljung, "Application of Fast Kalman Estimation to Adaptive Equalization," IEEE Trans. Commun., *COM-26* (October 1978), pp. 1439–46.
2. E. Satorius and J. Pack, "Application of Least Squares Lattice Algorithms to Adaptive Equalization," IEEE Trans. Commun., *COM-29* (February 1981), pp. 136–42.
3. T. L. Lim and M. S. Mueller, "Rapid Equalizer Start-Up Using Least Squares Algorithms" 1980 Proc. IEEE ICC, Seattle, WA.
4. M. Mueller, "On the Rapid Initial Convergence of Least Squares Equalizer Adjustment Algorithms," B.S.T.J., *60*, No. 10 (December 1981), pp. 2345–58.

5. F. K. Soong and A. M. Pederson, "Fast Least-Squares (LS) in the Voice Echo Cancellation Application," Proc. 1982 IEEE ICASSP, Paris, France, May 1982.
6. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.
7. M. Morf and D. T. L. Lee, "Fast Algorithms for Speech Modeling," Technical Report No. M308-1, Information Systems Laboratory, Stanford University, Stanford, CA, December 1978.
8. B. Widrow, "Adaptive Filters," in *Aspects of Network and System Theory*, R. Kalman and N. De Claris, Eds. New York, NY: Holt, Rinehart, and Winston, 1971, pp. 563–87.
9. R. D. Gitlin, J. E. Mazo, and M. G. Taylor, "On the Design of Gradient Algorithms for Digitally Implemented Adjustment Filters," IEEE Trans. Circuit Theory, *CT-20* (March 1973), pp. 125–36.
10. L. J. Griffiths, "A Continuously-Adaptive Filter Implemented as a Lattice Structure," Proc. 1977 IEEE ICASSP, Hartford, CT (May 1977), pp. 683–6.
11. M. Morf, B. Dickinson, T. Kailath, and A. Vieira, "Efficient Solution of Covariance Equations for Linear Prediction," IEEE Trans. ASSP, *ASSP-25* (October 1977), pp. 429–33.
12. L. Ljung, M. Morf, and D. Falconer, "Fast Calculation of Gain Matrices for Recursive Estimation Schemes," Int. J. Contr. *27*, No. 1 (1978), pp. 1–19.
13. D. T. Lee, B. Friedlander, and M. Morf, "Recursive Least Squares Ladder Estimation Algorithms," IEEE Trans. ASSP, *ASSP-29* (June 1981), pp. 627–41.
14. B. Porat, B. Friedlander, and M. Morf, "Square-Root Covariance Ladder Algorithms," IEEE Trans. Aut. Control, *AC-27*, No. 4 (August 1982), pp. 813–29.
15. C. Samson, "A Unified Treatment of Fast Kalman Algorithms for Identification," Int. J. Control, *35*, No. 5 (May 1982), pp. 909–34.
16. B. Friedlander, "Lattice Filters for Adaptive Processing," Proc. IEEE, *70*, No. 8 (August 1982), pp. 829–67.
17. S. L. Marple, Jr., "Efficient Least Squares FIR System Identification," IEEE Trans. ASSP, *ASSP-29* (Feb. 1981), pp. 62–73.
18. M. Morf, D. T. Lee, J. R. Nickolls, and A. Vieira, "A Classification of Algorithms for ARMA Models and Ladder Realizations," Proc. 1977 IEEE Conf. ASSP, Hartford, CT (April 1977), pp. 13–9.
19. M. Morf, D. T. Lee, "Recursive Least Squares Ladder Forms for Fast Parameter Tracking," Proc. 1978 IEEE Conf. D&C, San Diego, CA (Jan. 12, 1979), pp. 1326–67.
20. D. T. L. Lee and M. Morf, "Recursive Square-Root Ladder Estimation Algorithms," Proc. 1980 IEEE ICASSP, Denver, CO, April 1980.
21. A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*, New York, NY: Holt, Rinehart and Winston, Inc., 1971.
22. J. Cioffi and T. Kailath, "Fast, Fixed-Order, Least-Squares Algorithms for Adaptive Filtering," Proc. 1983 IEEE ICASSP, Boston MA, April 1983.
23. C. C. Halkias et al., "A New Generalized Recursion for the Fast Computation of the Kalman Gain to Solve the Covariance Equations," Proc. 1982 IEEE ICASSP, Paris, France (April 1982), pp. 1760–3.
24. M. L. Honig, "Performance of FIR Adaptive Filters Using Recursive Algorithms," Ph.D. Dissertation, Univ. of California, Berkeley, April 1981.

APPENDIX A

Derivation of (30)

We wish to prove (30). By definition,

$$P_{M_i|i-1}\mathbf{Y}_i = P_{\tilde{M}_i}\mathbf{Y}_i + P_{U_i}P_{M_i|i-1}\mathbf{Y}_i, \tag{118}$$

where $\tilde{M}$ is the subspace spanned by the column vectors of $\tilde{\mathbf{S}}_i$. Projecting both sides of (118) onto M_i gives

$$P_{M_i}P_{M_i|i-1}\mathbf{Y}_i = P_{M_i}P_{\tilde{M}_i}\mathbf{Y}_i + P_{M_i}P_{U_i}P_{M_i|i-1}\mathbf{Y}_i. \tag{119}$$

Now $P_{M_i|i-1}\mathbf{Y}_i$ lies in M_i, and hence

$$P_{M_i}P_{M_i|i-1}\mathbf{Y}_i = P_{M_i|i-1}\mathbf{Y}_i. \tag{120}$$

Also,

$$P_{M_i}P_{\tilde{M}_i}\mathbf{Y}_i = \mathbf{S}_i(\mathbf{S}_i^T\mathbf{S}_i)^{-1}(\mathbf{S}_i^T\tilde{\mathbf{S}}_i)(\tilde{\mathbf{S}}_i^T\tilde{\mathbf{S}}_i)^{-1}\tilde{\mathbf{S}}_i^T\mathbf{Y}_i$$

$$= \mathbf{S}_i(\mathbf{S}_i^T\mathbf{S}_i)^{-1}\tilde{\mathbf{S}}_i^T\mathbf{Y}_i$$

$$= P_{M_i}(\mathbf{Y}_i - P_{U_i}\mathbf{Y}_i). \tag{121}$$

Combining (118) through (121) gives

$$P_{M_i}\mathbf{Y}_i = P_{M_{i|i-1}}\mathbf{Y}_i + P_{M_i}P_{U_i}P^{\perp}_{M_{i|i-1}}\mathbf{Y}_i$$

$$= P_{M_{i|i-1}}\mathbf{Y}_i + (P_{M_i}\mathbf{u}_i)\langle \mathbf{u}_i, P^{\perp}_{M_{i|i-1}}\mathbf{Y}_i\rangle. \tag{122}$$

Subtracting both sides of (122) from $\mathbf{Y}_i$, and then taking inner products of both sides with $\mathbf{u}_i$ gives

$$\langle \mathbf{u}_i, P^{\perp}_{M_i}\mathbf{Y}_i\rangle = \langle \mathbf{u}_i, P_{M_{i|i-1}}\mathbf{Y}_i\rangle[1 - \langle \mathbf{u}_i, P_{M_i}\mathbf{u}_i\rangle]. \tag{123}$$

Combining (122) and (123), and using the definition (31) gives (30). [Ref. 24 gives a purely geometric proof of (30) for the case where M_i is spanned by one vector (as illustrated in Fig. 2).]

To derive the inner product update (35), we first rewrite (34) as

$$P^{\perp}_{M_i}\mathbf{Y}_i = P^{\perp}_{U_i}P^{\perp}_{M_{i|i-1}}\mathbf{Y}_i + P_{U_i}P^{\perp}_{M_{i|i-1}}\mathbf{Y}_i - P_{M_i}\mathbf{u}_i\langle \mathbf{u}_i, P^{\perp}_{M_i}\mathbf{Y}_i\rangle\sec^2\theta_i$$

$$= P^{\perp}_{U_i}P^{\perp}_{\tilde{M}_i}\mathbf{Y}_i + \mathbf{u}_i\langle \mathbf{u}_i, P^{\perp}_{M_{i|i-1}}\mathbf{Y}_i\rangle - P_{M_i}\mathbf{u}_i\langle \mathbf{u}_i, P^{\perp}_{M_i}\mathbf{Y}_i\rangle\sec^2\theta_i$$

$$= P^{\perp}_{U_i}P^{\perp}_{\tilde{M}_i}\mathbf{Y}_i + P^{\perp}_{M_i}\mathbf{u}_i\langle \mathbf{u}_i, P^{\perp}_{M_i}\mathbf{Y}_i\rangle\sec^2\theta_i. \tag{124}$$

Taking the inner product of both sides with $\mathbf{v}_i$ and using the fact that

$$\langle \mathbf{v}_i, P^{\perp}_{M_i}\mathbf{u}_i\rangle = \langle \mathbf{u}_i, P^{\perp}_{M_i}\mathbf{v}_i\rangle \tag{125}$$

gives (35).

APPENDIX B

Other Recursive Least-Squares Algorithms

The recursions in Section V and VII are complete in the sense that any of the existing computationally efficient LS prediction or joint-process-estimation algorithms can be derived from suitable subsets of these recursions. The purpose of this appendix is to illustrate this point by listing the recursions that enter the prewindowed LS transversal (fast Kalman) and lattice algorithms, the unnormalized sliding-window and growing-memory covariance lattice algorithms,[14] and the nonrecursive LS algorithm presented in Refs. 11 and 17. The list of recursions presented below does not completely describe each algorithm. For example, initialization is not discussed. Consistent steady-state algorithms can be formulated, however, by choosing the time indices and order of the variables in each recursion appropriately. The

following algorithms apply to the more general joint-process-estimation case (eliminating the recursions from Section VII gives the analogous prediction algorithm):

1. Prewindowed transversal (fast Kalman) algorithm:[1] (60a), (77a), (51a), (90a), (69), (60b), (93m), (68b), (104), and (108).
2. Prewindowed lattice algorithm* (See Refs. 18 and 16): (89), (64a)† and (64b), (65a) and (65b), (72a), (109) and (105).
3. Sliding-window lattice algorithm:[14,16] (89), (64a) and (64b), (65a) and (65b), (72a), (73a), (91), (109), (105), and (115).
4. Growing-memory covariance lattice algorithm:[14,16] (89), (64a) and (64b), (65a) and (65b), (82a), (88a), (92a), (72b) and (72a), (73b), (74b), (109), (105), (111), and (114).
5. Nonrecursive LS algorithm:[11,17]

The following set of recursions, which represents a modified version of the algorithm presented in Ref. 17, can be used to compute $\mathbf{f}(i \mid N)$, $\mathbf{b}(i \mid N)$, and $\mathbf{c}(i \mid N)$, given by (4), in an order-recursive fashion starting with first-order least squares variables at time i. Initialization consists of computing these first-order variables via the definitions given in Section V.

(78) (for computing $\mathbf{h}(i-1 \mid n)$), (85) (for computing $\mathbf{g}_1(i \mid n)$), (79), (86), (84a), (77b), (92a), (90b), (53), (66), (67), (65a) and (65b), (51b), (61), (68), (71), (57c), (72a), (73b), (103), (113), (101), (107).

Assuming that the covariance matrix $\Phi_{N,i\mid N+1}$ has been computed, a more convenient form for (53) is

$$\begin{aligned} k_n(i) &= \langle \mathbf{Y}_{n,i},\, \mathbf{E}_b(i-1 \mid n-1)\rangle \\ &= \mathbf{Y}_{n,i}^T[z^{-n}\mathbf{Y}_{n,i} - \mathbf{S}_i(1,\, n-1)\mathbf{b}(i-1 \mid n-1)] \\ &= R_n - \sum_{j=1}^{n-1} R_{n-j}[\mathbf{b}(i-1 \mid n-1)]_j, \end{aligned} \tag{126}$$

where $R_j = \mathbf{Y}_{n,i}^T(z^{-j}\mathbf{Y}_{n,i})$, and is the $(1,\, j+1)$st element of $\Phi_{n,i\mid n+1}$. Equation (101) can be similarly modified.

AUTHOR

Michael L. Honig, B.S. (Electrical Engineering), 1977, Stanford University; M.S. and Ph.D. (Electrical Engineering), University of California, Berkeley, 1978 and 1981, respectively; AT&T Bell Laboratories, 1981–1982, AT&T

---

* The normalized lattice algorithms presented in Refs. 13 and 14 can be obtained by making substitutions for the variables entering the unnormalized recursions presented here.[16,20,24]

† In the prewindowed and growing-memory covariance cases, the inner products of (64) and (105) with $\mathbf{u}_i$ are required. In the sliding-window case, the inner products of (64) and (105) with both $\mathbf{u}_i$ and $\mathbf{u}_{i_0}$ are required.

Information Systems, 1983—. At AT&T Bell Laboratories and AT&T Information Systems Mr. Honig has worked on modulation, coding, and echo cancellation of voiceband data signals; and performance analysis of local area networks. He is currently working on office information networks. Member, IEEE, Tau Beta Pi, Phi Beta Kappa.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

On the Average Product of Gauss-Markov Variables

By B. F. LOGAN, Jr.,* J. E. MAZO,* A. M. ODLYZKO* and L. A. SHEPP*

(Manuscript received April 8, 1983)

Let x_i be members of a stationary sequence of zero mean gaussian random variables having correlations $Ex_ix_j = \sigma^2\rho^{|i-j|}$, $0 < \rho < 1$, $\sigma > 0$. We address the behavior of the averaged product $q_m(\rho, \sigma) \equiv Ex_1x_2 \cdots x_{2m-1}x_{2m}$ as m becomes large. Our principal result when $\sigma^2 = 1$ is that this average approaches zero (infinity) as ρ is less (greater) than the critical value $\rho_c = 0.563007169\ldots$. To obtain this we introduce a linear recurrence for the $q_m\cdot(\rho, \sigma)$, and then continue generating an entire sequence of recurrences, where the $(n + 1)$-st relation is a recurrence for the coefficients that appear in the nth relation. This leads to a new, simple continued fraction representation for the generating function of the $q_m(\rho, \sigma)$. The related problem with $\tilde{q}_m(\rho, \sigma) = E|x_1 \cdots x_m|$ is studied via integral equations and is shown to possess a smaller critical correlation value.

I. INTRODUCTION

The problem that we consider in this paper is as follows: Let $\{x_i\}_1^\infty$ be a stationary sequence of zero mean, gaussian random variables with covariances

$$\rho_{ij} \equiv Ex_ix_j = \sigma^2\rho^{|i-j|},\ 0 < \rho < 1,\ \sigma > 0; \quad i, j = 1, 2, \cdots, \quad (1)$$

where $E(\cdot)$ denotes mathematical expectation. What is the behavior of

*Bell Laboratories.

$$q_m(\rho, \sigma) \equiv Ex_1x_2 \cdots x_{2m-1}x_{2m} \tag{2}$$

as m becomes large?

In other words, the product in (2) is formed from samples of a gauss-markov process that are taken at regular intervals. Only an even number of samples is considered in (2) since an odd number would result in a zero average.

Originally, the problem was conceived as a simple model for averages of multiplicative structures having infinite memory between the factors of the product. Such products arise in the analysis of learning curves for many adaptive systems, and for these problems one encounters products whose factors are noncommuting matrices. We felt that the analysis of a simple problem, such as that described above, would serve as a valuable guide to what results might be achievable for more realistic situations. However as one may readily imagine, as soon as the problem described in (1) and (2) was written down it became of interest in its own right, consisting as it does of a simple question about long familiar quantities.

Our principal result is that for large m the behavior of the average product $q_m(\rho, \sigma)$ in (2) depends on the relationship of ρ to a critical value, $\rho_c = \rho_c(\sigma)$. If $\rho < \rho_c$, then $q_m(\rho, \sigma)$ will approach zero exponentially fast; if $\rho > \rho_c$, $q_m(\rho, \sigma)$ approaches infinity exponentially fast; finally, if $\rho = \rho_c$, $q_m(\rho, \sigma) \to q_\infty(\sigma)$. We find for $\sigma = 1$, $\rho_c(1) = 0.563007169391816 \cdots$, and $q_\infty(1) = 0.50900853 \ldots$. A plot of $\rho_c(\sigma)$ is given in Fig. 1. All of these results were obtained from a continued fraction representation for the generating function

$$Q(z, \rho, \sigma) = \sum_{m=0}^{\infty} q_m(\rho, \sigma) z^m. \tag{3}$$

Since $q_m(\rho, \sigma) = \sigma^{2m} q_m(\rho, 1)$, we have

$$Q(z, \rho, \sigma) = Q(z\sigma^2, \rho, 1), \tag{4}$$

so it is without loss of generality that we will set $\sigma = 1$, $Q(z, \rho) = Q(z, \rho, 1)$, and $q_m(\rho) = q_m(\rho, 1)$. By introducing a sequence of generating functions, we show in Section II that

$$Q(z, \rho) = \cfrac{1}{1 - \cfrac{\rho z}{1 - \cfrac{2\rho^3 z}{1 - \cfrac{3\rho^5 z}{1 - \ddots}}}} \tag{5}$$

The value $\rho_c^{(\sigma)}$ is then the smallest ρ for which $Q(\sigma^2, \rho) = \infty$, while the value $q_\infty(\sigma)$ is the limit as $z \to 1$ of $(1 - z)Q(z\sigma^2, \rho_c)$.

Since methods are as interesting as results, Section III presents another approach involving integral equations for discussing the $q_m(\rho)$ behavior. Although this method is not rigorously justified for the present problem due to a non-hermitian kernel, it is applicable to a

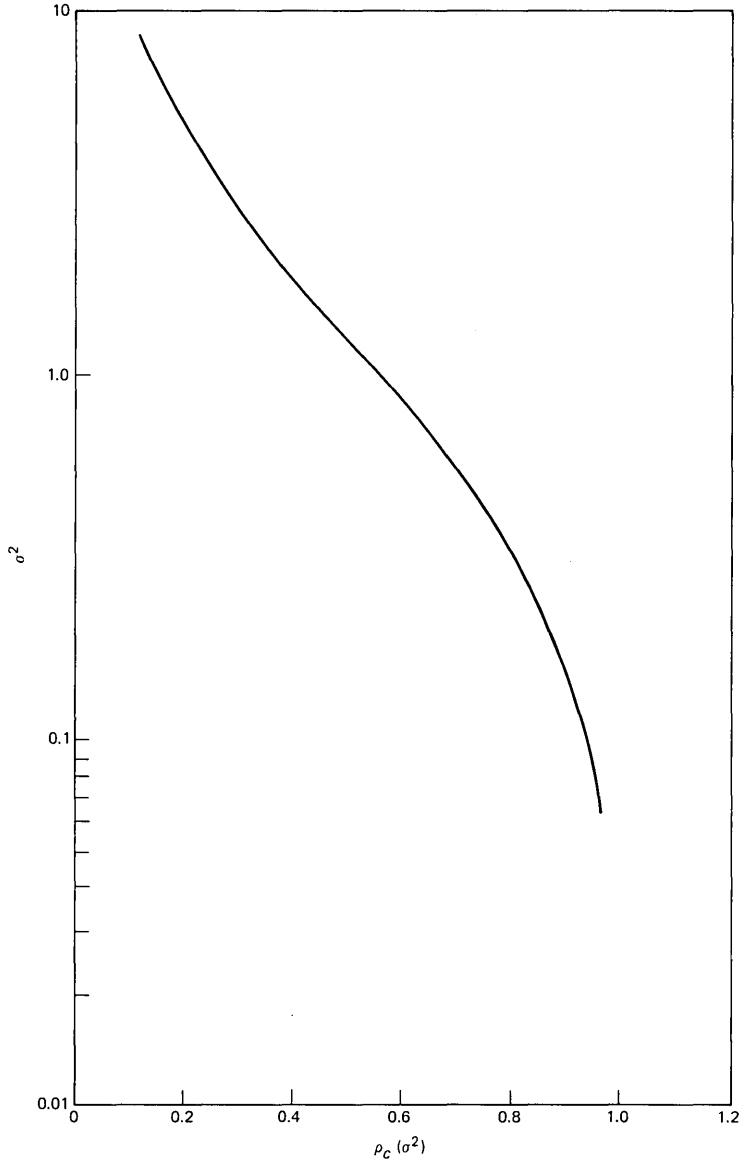


Fig. 1—Critical correlation value $\rho_c(\sigma^2)$ vs. variance σ^2.

related problem, the behavior of $E|x_1 \cdots x_m|$ as $m \to \infty$ [still assuming (1)]. Using the integral equation we show that $\bar{\rho}_c$, the critical value of ρ for this new problem, is strictly smaller than the ρ_c defined above. This is of interest since it shows that the behavior of $q_m(\rho)$ is determined both by how large $|x_1 \cdots x_{2m}|$ is on the average, and by the extent of cancellation between positive and negative values of q_m.

Although we do not give the details here, it is not difficult to show that for all $\rho < 1$, $q_m(\rho)$ approaches zero with probability one as m becomes large.

II. LINEAR RECURRENCES AND GENERATING FUNCTIONS

Given $2m$ zero-mean jointly gaussian random variables x_i of unit variance and correlations $Ex_ix_j = \rho_{ij}$, then a known formula[1] states that

$$Ex_1 \cdots x_{2m} = \sum_{\substack{\text{all} \\ \text{pairs}}} \rho_{i_1 i_2}\rho_{i_3 i_4} \cdots \rho_{i_{2m-1} i_{2m}}, \tag{6}$$

where the unordered set $\{i_1, \cdots, i_{2m}\}$ is equal to the unordered set $\{1, 2, \cdots, 2m\}$. The sum in (6) is over all distinct, unordered pairs of subscripts. That is, we do not count twice terms which differ only by interchanging the values within one or more subscript pairs, nor do we count twice terms which differ only by permuting subscript pairs. Thus there are $(2m)!/(2^m m!)$ terms in the sum (6).

If we denote permutations of $2m$ objects by $\sigma(i)$: $i \to \sigma(i)$, $i = 1, 2, \cdots, 2m$, then a succinct way of writing (6) when (1) holds is

$$q_m(\rho) = \frac{1}{2^m m!} \sum_{\sigma \in S_{2m}} \rho^{\sum_{j=1}^{m} |\sigma(2j) - \sigma(2j-1)|}, \tag{7}$$

the sum in (7) being over all $(2m)!$ permutations of S_{2m}, the group of permutations of $2m$ symbols. Formula (7) shows immediately that $q_m(\rho) > 0$ if $\rho > 0$.

Now define $q_0(\rho) = 1$ and write

$$q_m(\rho) = \sum_{s=1}^{m} b_s(\rho) q_{m-s}(\rho), \qquad m = 1, 2, \cdots. \tag{8}$$

We evaluate a few of the $b_s(\rho)$, writing for convenience $b_i(\rho) = b_i$, $q_i(\rho) = q_i$. The evaluation is done from (8) by explicitly evaluating the $q_m(\rho)$ as needed. A partial list of $b_i(\rho)$ follows:

$$b_1 = \rho$$

$$b_2 = 2\rho^4$$

$$b_3 = 4\rho^7 + 6\rho^9$$

$$b_4 = 8\rho^{10} + 24\rho^{12} + 18\rho^{14} + 24\rho^{16}$$

$$b_5 = 16\rho^{13} + 72\rho^{15} + 108\rho^{17} + 150\rho^{19} + 144\rho^{21} + 96\rho^{23} + 120\rho^{25}. \tag{9}$$

Equation (9) suggests the possibility that, for small ρ, only a few terms in (8) would need to be kept for an accurate description of $q_m(\rho)$. For example, keeping only one term yields

$$q_m = \rho q_{m-1}, \tag{10}$$

or $q_m = \rho^m$. Since $Ex_1x_2 = \rho$, this approximation corresponds to treating the successive pairs of gaussian variables which determine $q_m(\rho)$, via (2), as independent. The next step after (10) would be to write

$$q_m = b_1 q_{m-1} + b_2 q_{m-2}. \tag{11}$$

This equation, involving b_2 as well, would be a correction to the "independence assumption," but one involving only up to fourth-order correlations, since, from (8) the highest average appearing in b_2 is $E(x_1x_2x_3x_4)$. Further corrections are obtained by including more terms of (8), with higher order correlations entering.†

Assuming the $b_i(\rho)$ to be known, the natural procedure would be to "solve" (8) using generating functions. We define these as follows: if $y_0, y_1, y_2, \cdots$ is a bounded sequence of numbers, then the generating function, $Y(z)$, of the sequence is defined for complex z, $|z| < 1$, by

$$Y(z) = \sum_{i=0}^{\infty} y_i z^i. \tag{12}$$

Given $Y(z)$, the y_i are, in principle, uniquely determined. We assume that the reader is familiar with the use of generating functions. If not, consult Chapters XI and XIII of Feller.[2]

We define $b_0(\rho) = 0$, $q_0(\rho) = 1$, and call the generating functions of the $b_i(\rho)$, and $q_i(\rho)$ sequences $B(z; \rho)$ and $Q(z; \rho)$, respectively. The ρ dependence is explicitly indicated.

If we multiply (8) by z^m and sum from $m = 1$ to ∞ (treating $q_m = 0$, $m < 0$ and $b_m = 0$, $m < 0$), we obtain the basic relation

$$Q(z; \rho) = \frac{1}{1 - B(z; \rho)}. \tag{13}$$

Equation (13) thus allows us to determine, in principle, the q_m from the b_m. In particular, we have

---

†The above interpretation prompts us to advocate consideration of the ideas represented by (8) for analyzing more complex multiplicative structures, particularly when connections to some sort of independence approximation are a natural thing to seek.

$$\sum_{m=1}^{\infty} q_m(\rho) = Q(1; \rho) = \frac{1}{1 - B(1; \rho)}, \tag{14}$$

and the critical value ρ_c will be given by the equation

$$B(1; \rho_c) = \sum_{1}^{\infty} b_k(\rho_c) = 1. \tag{15}$$

Although we could work with the $b_m(\rho)$ themselves, a more convenient approach for finding ρ_c numerically is to set up a continued fraction representation for the generating functions $Q(z; \rho)$, or equivalently, $B(z; \rho)$. It is this approach that we follow now.

Recall (8) defining $b_s(\rho)$. Since these b_s coefficients are a numerical sequence themselves, we can use the same reasoning that took us from the q_k to the b_s and use it to suggest going from the b_s to a new set of coefficients, $b_k^{(2)}$, via the following recurrence

$$b_k(\rho) = \sum_{s=1}^{k} b_s^{(2)}(\rho) b_{k-s}(\rho), \qquad k = 2, 3, \cdots, \tag{16}$$

where we define $b_0(\rho) = 0$. The recurrence (16) yields

$$B(z; \rho) = \frac{b_1(\rho) z}{1 - B^{(2)}(z; \rho)}, \tag{17}$$

$B^{(2)}(z; \rho)$ being the generating function for the $b_k^{(2)}(\rho)$. To continue this procedure with a uniform notation, we define

$$\begin{aligned} b_s^{(1)}(\rho) &= b_s(\rho) \\ b_0^{(m)}(\rho) &= 0, \qquad m = 1, 2, \cdots \end{aligned} \tag{18}$$

and write

$$b_k^{(m)}(\rho) = \sum_{s=1}^{k} b_s^{(m+1)}(\rho) b_{k-s}^{(m)}(\rho), \qquad \begin{matrix} m = 1, 2, \cdots \\ k = 2, 3, \cdots \end{matrix}. \tag{19}$$

The corresponding sequence of generating functions are related by

$$B^{(m)}(z; \rho) = \frac{b_1^{(m)}(\rho) z}{1 - B^{(m+1)}(z; \rho)}. \tag{20}$$

We use this repeatedly in (13) and obtain the continued fraction representation†

† The fact that this continued fraction does not terminate implies that $Q(z; \rho)$ is not a rational function of z, and thus one cannot find a (finite-order) difference equation for the $q_m(\rho)$. See Ref. 3, Theorem 99.1, p. 400.

$$Q(z,\rho)=\cfrac{1}{1-\cfrac{b_1^{(1)}z}{1-\cfrac{b_1^{(2)}z}{1-\cfrac{b_1^{(3)}z}{1-\ddots}}}} \tag{21}$$

In (21) we have, for simplicity, written $b_1^{(m)}(\rho)=b_1^{(m)}$.

A relation which will be used later to aid in finding the $b_1^{(m)}$ follows by setting $k=2$ in (19), to obtain

$$b_1^{(m+1)}(\rho)=\frac{b_2^{(m)}(\rho)}{b_1^{(m)}(\rho)}. \tag{22}$$

We can calculate some of the $b_1^{(m)}(\rho)$ using the partial list of the $b_k(\rho)$ given in (9) to derive several $b_s^{(m)}(\rho)$ from (19). Using (22) we then obtain

$$\begin{aligned} b_1^{(1)} &= \rho \\ b_1^{(2)} &= 2\rho^3 \\ b_1^{(3)} &= 3\rho^5 \\ b_1^{(4)} &= 4\rho^7 \\ b_1^{(5)} &= 5\rho^9. \end{aligned} \tag{23}$$

The obvious guess that

$$b_1^{(m)}=m\rho^{2m-1}, \qquad m=1,2,\cdots \tag{24}$$

follows from a direct proof of the continued fraction given in the appendix. Assuming (24) to hold yields the simple representation

$$Q(z;\rho)=\cfrac{1}{1-\cfrac{\rho z}{1-\cfrac{2\rho^3 z}{1-\cfrac{3\rho^5 z}{1-\ddots}}}} \tag{25}$$

The accurate numerical value

$$\rho_c=0.563007169391816\cdots \tag{26}$$

was obtained by using this representation along with (14) and (15).

When $\rho = \rho_c$,

$$q_\infty = \lim_{k\to\infty} q_k(\rho_c) = \frac{1}{\sum_1^\infty k b_k(\rho_c)}. \tag{27}$$

Using the computed value of ρ_c and the definition of the $b_k(\rho)$, we found numerically that

$$q_\infty = 0.50900853 \cdots. \tag{28}$$

It was quite surprising to us that $Q(z, \rho)$ turned out to be a new, but simple, continued fraction.

III. INTEGRAL EQUATION METHOD

The purpose of this section is to introduce the integral equation method and to show that $\bar{\rho}_c < \rho_c$, where $\bar{\rho}_c$ is the critical correlation value for the related problem involving $E|x_1 \cdots x_n|$.

We begin by developing an expression for $Ex_1 \cdots x_n$. We have, from the Markov property of the x_i sequence,

$$Ex_1 \cdots x_n = \int \cdots \int x_n p(x_n|x_{n-1}) \cdots x_1 p(x_1|x_0)\phi(x_0)dx_0 \cdots dx_n, \tag{29}$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \tag{30}$$

is the standard normal density and

$$p(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp[-(y-\rho x)^2/2(1-\rho^2)] \tag{31}$$

is the generic form of the conditional densities occurring in (29). Define a kernel $K(x, y)$ by

$$K(x, y) = yp(y|x),$$

$$Kf(x) = \int_{-\infty}^{\infty} K(x, y)f(y)dy.$$

Then (29) may be written in the inner product notation of Hilbert Space

$$Ex_1 \cdots x_n = (K^n \underline{1}, \phi), \tag{32}$$

where ϕ is the normal density (30), $\underline{1}$ is the unit constant function,

and K^n is the nth iterated kernel. Now assume, heuristically, that K^n has the usual expansion

$$K^n(x, y) = \sum_{j=1}^{\infty} \lambda_j^n \psi_j(x)\psi_j(y) \tag{33}$$

in terms of eigenfunctions $\psi_j(x)$ and eigenvalues λ_j of K. Then $Ex_1 \cdots x_n$ would remain bounded, if, and only if, the largest eigenvalue $\lambda_1 = \lambda_1(\rho)$ is less or equal to one; thus $\lambda_1(\rho_c) = 1$ would determine ρ_c. Unfortunately there is no general eigenexpansion theory available for K since it is not symmetric and is not symmetrizable.

Fortunately symmetry holds for the integral equation method when one expresses $E|x_1 \cdots x_n|$ via kernels. Define, in analogy to (32),

$$\bar{K}(x, y) = |y| p(y|x). \tag{34}$$

If we further define

$$J(x, y) = \frac{h(x)}{h(y)} \bar{K}(x, y), \tag{35}$$

where

$$h(x) = \sqrt{|x|}\exp(-x^2/4), \tag{36}$$

then

$$J(x, y) = \frac{1}{\sqrt{2\pi(1 - \rho^2)}} \sqrt{|xy|}\exp\left[-\frac{1}{4}\frac{1 + \rho^2}{1 - \rho^2}(x^2 + y^2)\right] \cdot\exp\left[\frac{\rho xy}{1 - \rho^2}\right] \tag{37}$$

is a symmetric kernel.

As in (29),

$$\begin{aligned} E|x_1 \cdots x_n| &= \int \cdots \int \bar{K}(x_{n-1}, x_n)\bar{K}(x_{n-2}, x_{n-1}) \\ &\qquad \cdots \bar{K}(x_0, x_1)\, \phi(x_0)dx_0 \cdots dx_n \\ &= \int \cdots \int dx_0 \cdots dx_n\, J(x_0,x_1) \\ &\qquad \cdots J(x_{n-1}, x_n)\frac{h(x_n)}{h(x_0)}\, \phi(x_0) \\ &= \left(J^n h, \frac{\phi}{h}\right). \end{aligned} \tag{38}$$

Since J is symmetric and square-integrable, it is a Hilbert-Schmidt kernel and so has a discrete spectrum. Further its maximum eigenvalue, λ, is given by

$$\lambda = \sup_f \frac{(Jf, f)}{(f, f)}. \tag{39}$$

Since $J(x, y) \geq 0$, we see that the maximum eigenfunction $g = g(x)$ is nonnegative and $\lambda > 0$ as well. Further since h and ϕ/h are nonnegative, $(h, g) > 0$ and $(\phi/h, g) > 0$ so that $E|x_1 \cdots x_n| = (J^n h, \phi/h) \to \infty$ if and only if $\lambda > 1$.

Define f_α by

$$f_\alpha(x) = \sqrt{|x|}\exp(-\alpha x^2/4), \tag{40}$$

and note that from (39),

$$\lambda > (Jf_\alpha, f_\alpha)/(f_\alpha, f_\alpha). \tag{41}$$

Now

$$(f_\alpha, f_\alpha) = \int |x|\exp(-\alpha x^2/2)dx = 2/\alpha \tag{42}$$

and

$$(Jf_\alpha, f_\alpha) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |xy| \cdot \exp\left[-\frac{c}{2}(x^2+y^2) + \frac{\rho xy}{1-\rho^2}\right] dxdy, \tag{43}$$

where

$$c = \frac{1}{2}\left(\frac{1+\rho^2}{1-\rho^2} + \alpha\right). \tag{44}$$

Set $y = xu$ and integrate over x to obtain

$$(Jf_\alpha, f_\alpha) = \frac{4}{\sqrt{2\pi(1-\rho^2)}} \int_{-\infty}^{\infty} \frac{|u|du}{(c(1+u^2) - \beta u)^2},$$

$$\beta = \frac{2\rho}{1-\rho^2}. \tag{45}$$

Using Ref. 4 (p. 68, 2.175) we evaluate the last integral as

$$(Jf_\alpha, f_\alpha) = \frac{4}{\sqrt{2\pi(1-\rho^2)}}\left[\frac{4}{\Delta} + \frac{4\beta}{\Delta^{3/2}}\tan^{-1}\frac{\beta}{\sqrt{\Delta}}\right],$$

$$(\Delta = 4c^2 - \beta^2. \quad (46)$$

Setting $\alpha = 0.5$, $\rho = 0.55$, we find that $c = 1.18369$, $\beta = 1.57706$, $\Delta = 3.11738$, $(Jf_\alpha, f_\alpha) = 4.0824$, $(f_\alpha, f_\alpha) = 4$, so that $\lambda > 1.012$. Thus for $\rho = 0.55$, $E|X_1 \cdots X_n| \to \infty$, and $\bar{\rho}_c < 0.55$. We have seen that $\rho_c > 0.563$ so the claim is proven.

REFERENCES

1. D. Middleton, *An Introduction to Statistical Communication Theory*, New York: McGraw-Hill, 1960, p. 343.
2. W. Feller, *An Introduction to Probability Theory and Its Applications*, Volume 1, 2nd ed, New York: John Wiley and Sons, 1957.
3. H. S. Wall, *Analytic Theory of Continued Fractions*, New York: Van Nostrand Co., 1948.
4. I. S. Gradshteyn and I. R. Ryzhik, *Table of Integrals, Series, and Products*, New York: Academic Press, 1965.

APPENDIX

Combinatorial Derivation of Continued Fraction

In this appendix we give a direct combinatorial proof of the continued fraction representation (25) of $Q(z, \rho)$. This derivation is complete in itself, but we preferred the method of the text for showing where the continued fraction comes from. Our starting point is the formula (7). Let us define, for $\sigma \in S_{2m}$,

$$V(\sigma) = \sum_{i=1}^{m} |\sigma(2i) - \sigma(2i-1)|. \quad (47)$$

For $1 \leqslant k \leqslant m$, let

$$\begin{aligned} S(m, k) = \{\sigma \in S_{2m} : &\sigma(2m) = 2m, \\ &\sigma(2m-2) = 2m-1, \\ \sigma(2m-4) &= 2m-2, \cdots, \\ &\sigma(2m-2k+2) = 2m-k+1\}. \end{aligned} \quad (48)$$

For $k = 0$, we adopt the convention that $S(m, 0) = S_{2m}$. We also define

$$u(m, k) = \frac{1}{2^{m-k}(m-k)!}\sum_{\sigma \in S(m,k)} \rho^{V(s)}, \quad (49)$$

so that $u(m, 0) = q_m$. (We take $u(0, 0) = q_0 = 1$, and $u(m, k) = 0$ for $k < 0$ and $k > m$.) Our key result is:

Lemma. If $m \geqslant 1$, $k \geqslant 0$, *then*

$$u(m, k) = k\rho^{2k-1}u(m-1, k-1) + u(m, k+1). \tag{50}$$

Proof. We will prove this for $1 \leqslant k \leqslant m-1$, as the other cases are easy. Let

$$\begin{aligned} S' = \{\sigma \in S(m, k) : & 2m-k \\ & \in \{\sigma(2m-1), \sigma(2m-3), \cdots, \sigma(2m-2k+1)\}\}, \\ S'' = S(m, k) - S'. & \end{aligned} \tag{51}$$

If $\sigma \in S''$, we construct a permutation $\sigma^* \in S(m, k+1)$ by changing the action of σ on four letters in such a way that $V(\sigma) = V(\sigma^*)$ and $\sigma^*(2m-2k) = 2m-k$. To define σ^* precisely, let p and r be such that $\{r, 2m-k\} = \{\sigma(2p-1), \sigma(2p)\}$. Then, if we associate to σ the vector $A(\sigma) = (\sigma(1), \sigma(2), \cdots, \sigma(2m))$, the vector $A(\sigma^*)$ is obtained from $A(\sigma)$ by interchanging the pairs $\{\sigma(2m-2k-1), \sigma(2m-k)\}$ and $\{r, 2m-k\}$ so as to keep the same ordering in the first pair, but possibly reversing it in the second, so as to have $\sigma^*(2m-2k) = 2m-k$. As an example, if $m = 5$, $k = 3$, and $A(\sigma) = (7, 2, 4, 1, 6, 8, 3, 9, 5, 10)$, then $A(\sigma^*) = (4, 1, 2, 7, 6, 8, 3, 9, 5, 10)$. It is clear that $\sigma^* \in S(m, k+1)$ and $V(\sigma^*) = V(\sigma)$. Moreover, every $\tau \in S(m, k+1)$ can be represented in exactly $2(m-k)$ ways as $\tau = \sigma^*$, $\sigma^* \in S''$. Therefore,

$$\frac{1}{2^{m-k}(m-k)!} \sum_{\sigma \in S''} \rho^{V(\sigma)} = u(m, k+1). \tag{52}$$

Suppose now that $\sigma \in S'$. Then $2m-k = \sigma(2m-2r+1)$ for some r, $1 \leqslant r \leqslant k$. We now define a permutation $\sigma' \in S(m-1, k-1)$ as follows: In $A(\sigma)$, delete $a = \sigma(2m-2r+1)(= 2m-k)$ and $b = \sigma \cdot (2m-2r+2)$ and reduce the remaining entries that are between a and b by 1, and those that are larger than $\max(a, b) = a$ by 2. As an example, if $m = 5$, $k = 3$, and $A(\sigma) = (2, 1, 6, 3, 5, 8, 7, 9, 4, 10)$, then $A(\sigma') = (2, 1, 6, 3, 5, 7, 4, 8)$. The resulting vector clearly equals $A(\sigma')$ for some $\sigma' \in S(m-1, k-1)$, and each $\tau \in S(m-1, k-1)$ has exactly k such representations. Further, $V(\sigma)$ equals the sum of (i) $V(\sigma')$, (ii) $a-b$ for the pair that was dropped, (iii) 2 for each of the $r-1$ pairs $(\sigma(2m-2j+1), \sigma(2m-2j+2))$ for $1 \leqslant j \leqslant r-1$, since in each such pair $\sigma(2m-2j+2) > a$, $\sigma(2m-2j+1) < b$, and finally (iv) 1 for each of the $k-r$ pairs $(\sigma(2m-2j+1), \sigma(2m-2j+2))$, $r+ \leqslant j \leqslant k$, since in each of them $\sigma(2m-j+1) < b$, $b < \sigma(2m-2j+2) < a$. Hence,

$$V(\sigma) = V(\sigma') + a - b + 2(r-1) + k - r. \tag{53}$$

But $a = 2m-k$ and $b = \sigma(2m-2r+2) = 2m-r+1$ from the definitions of $S(m, k)$, so

$$V(\sigma) = V(\sigma') + 2k - 1. \tag{54}$$

Hence, we have

$$\frac{1}{2^{m-k}(m-k)!}\sum_{\sigma\in S'}\rho^{V(\sigma)} = k\rho^{2k-1}u(m-1,\,k-1), \tag{55}$$

which proves the lemma.

We now can use the recurrence of the Lemma to derive the continued fraction expansion of the generating function. Let

$$f_k = f_k(z) = \sum_{m=0}^{\infty} u(m,\,k)z^m, \qquad k = 0,\,1,\,\cdots,$$

which for the moment we regard as formal power series in z. Then the Lemma gives us

$$f_1 = f_0 - 1, \tag{56}$$

and for $k \geqslant 2$,

$$\begin{aligned} f_k &= \sum_m u(m,\,k-1)z^m - (k-1)\rho^{2k-3}\sum_m u(m-1,\,k-2)z^m \\ &= f_{k-1} - (k-1)\rho^{2k-3}zf_{k-2}. \end{aligned} \tag{57}$$

Relations (56) and (57) show that for $k \geqslant 0$,

$$f_k = s_k f_0 - r_k, \tag{58}$$

where $s_0 = s_1 = 1$, $r_0 = 0$, $r_1 = 1$, and for $k \geqslant 2$ both s_k and r_k satisfy the recurrence

$$x_k = x_{k-1} - (k-1)\rho^{2k-3}zx_{k-2}.$$

Hence the quotients r_k/s_k are the partial quotients of the continued fraction $R(z,\,\rho)$ on the right side of (25), and s_k and r_k converge as $k \to \infty$ to power series (in z) $s(z,\,\rho)$ and $r(z,\,\rho)$, respectively, for which

$$R(z,\,\rho) = \frac{r(z,\,\rho)}{s(z,\,\rho)}. \tag{59}$$

On the other hand, since f_k starts with a term involving z_k, we conclude that f_k converges to 0 in the ring of formal power series as $k \to \infty$. Therefore, from (58),

$$f_0 = \frac{r(z,\,\rho)}{s(z,\,\rho)} = R(z,\,\rho). \tag{60}$$

Since $f_0 = Q(z,\,\rho)$, we obtain the relation (25), at least in the ring of formal power series in z. However, the continued fraction (25) is clearly a meromorphic function of z for ρ fixed, $0 < \rho < 1$, and it is analytic at 0. Hence (25) holds as an equality among meromorphic functions, and we can obtain from this the exponential decrease of the $q_m(\rho)$ for $\rho < \rho_c$ and the exponential increase for $\rho > \rho_c$.

AUTHORS

Benjamin F. Logan, Jr., B.S. (Electrical Engineering), 1946, Texas Technological College; M.S., 1951, Massachusetts Institute of Technology; Eng.D.Sc. (Electrical Engineering), 1965, Columbia University; Bell Laboratories, 1956—. While at MIT, Mr. Logan was a research assistant in the Research Laboratory of Electronics, investigating characteristics of high-power electrical discharge lamps. Also at MIT he engaged in analog computer development at the Dynamic Analysis and Control Laboratory. From 1955 to 1956 he worked for Hycon-Eastern, Inc., where he was concerned with the design of airborne power supplies. He joined Bell Laboratories as a member of the Visual and Acoustics Research Department, where he was concerned with the processing of speech signals. Currently, he is a member of the Mathematical Research Department. Member, Sigma Xi, Tau Beta Pi.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. (Physics), 1963, Syracuse University; Research Associate, Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964—. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now working in the Mathematics and Statistics Research Center. Member, IEEE.

Andrew M. Odlyzko, Ph.D., 1975, Massachusetts Institute of Technology; Bell Laboratories, 1975—. Mr. Odlyzko works in various areas of mathematics, including error-correcting codes, combinatorics, analysis, probability theory, and analysis of algorithms.

Lawrence A. Shepp, B.S. (Applied Mathematics), 1958, Brooklyn Polytechnic Institute; M.A., Ph.D. (Mathematics), 1961, Princeton; Bell Laboratories, 1962—. Mr. Shepp has worked mainly on problems involving probability theory. Since 1972 he has also worked in reconstruction of images from X-ray (and other) projections. He won the Paul Levy prize in probability, an IEEE (Nuclear Science) Distinguished Scientist award, and the Bell Laboratories Distinguished Scientist award. He is Professor of Radiology at Columbia University, biomathematician at The Neurological Institute, NYC, and a member of the Mathematics and Statistics Research Center at Bell Laboratories.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

Series Solutions of Companding Problems

By B. F. LOGAN, Jr.*

(Manuscript received March 10, 1982)

A formal power series solution (i) $x(t) = \sum_1^\infty m^k x_k(t)$ is given for the companding problem (ii) $Bf\{x(t)\} = my(t)$, $B\{x(t)\} = x(t)$, where B is the bandlimiting operator defined by $Bg = (Bg)(t) = \int_{-\infty}^{\infty} g(s)[\sin \lambda(t - s)]/[\pi(t - s)]ds$ and $f(t)$ has a Taylor series with $f(0) = 0, f'(0) \neq 0$. Expressions for the x_k are given in terms of the coefficients of f, and operations on y, and in a different form in terms of the coefficients of the inverse function ϕ, $\phi\{f(x)\} = x$. A series development is given for a bandlimited $z(t)$, $Bz = z$, such that the solution of (ii) is given by $x = B\phi(z)$. Also a series development is given for the "approximate identity", $x \doteq B\phi\{Bf(x)\}$, where $x = x(t)$, $Bx = x$, which is shown to be a good approximation to x for fairly linear $f(x)$, not necessarily having a Taylor series expansion. As an example of one application of the results, a few terms are given for correction of the "inband" distortion arising in envelope detection of "full-carrier" single-sideband signals. The results should prove useful in correcting small distortions in other transmission systems. Finally, it is shown that the formal series solution (*i*) actually converges for sufficiently small $|m|$. This involves proving that the companding problem (*ii*) has a unique solution for arbitrary complex-valued $y(t)$ and complex m of sufficiently small magnitude, the solution $x(t; m)$ being, for each t, an analytic function of the complex variable m in a neighborhood of the origin. It is a curious fact, as shown by an interesting example, that the series (*i*) may converge for values of m for which it is not a solution of (*ii*).

I. INTRODUCTION

Suppose $x(t)$ is a bandlimited signal whose Fourier transform vanishes outside the interval $[-\lambda, \lambda]$. If such a signal is instantaneously

* Bell Laboratories.

distorted by a nonlinear (companding) function $f(x)$, the distorted signal $f\{x(t)\}$ will, in general, have frequency components outside the interval $[-\lambda, \lambda]$. If the out-of-band components of the distorted signal are removed by ideal low-pass filtering, the result is a bandlimited signal $y(t)$ whose Fourier transform agrees with that of $f\{x(t)\}$ over $(-\lambda, \lambda)$. How, and under what conditions, may $x(t)$ be recovered from $y(t)$? When the signals are real-valued, this is known as the *companding problem* of Landau and Miranker (Refs. 1 and 2), hereafter referred to as the *r.v. companding problem.* Before stating their result, and our purpose, we introduce some notation.

The symbol $\mathscr{B}_2(\lambda)$ will denote the subspace of $L_2 = L_2(-\infty, \infty)$ whose elements are those (square-integrable) functions whose Fourier transforms vanish outside $[-\lambda, \lambda]$. Associated with this subspace is the *bandlimiting operator* B_λ, defined for g in L_2 by

$$B_\lambda g = (B_\lambda g)(t) = \int_{-\infty}^{\infty} g(s) \frac{\sin \lambda(t-s)}{\pi(t-s)}\, ds.$$

The Fourier transforms of g and $B_\lambda g$ agree over $(-\lambda, \lambda)$, the transform of the latter vanishing outside $[-\lambda, \lambda]$. In the language of Hilbert space, $B_\lambda g$ is the projection of g on $\mathscr{B}_2(\lambda)$, being the best approximation to g in the subspace $\mathscr{B}_2(\lambda)$. In case g belongs to $\mathscr{B}_2(\lambda)$, we have

$$B_\lambda g = g.$$

It follows that

$$B_\lambda^n g = B_\lambda g, \qquad g \text{ in } L_2, \qquad n = 1, 2, \cdots.$$

The operator B_λ may be applied also to functions belonging to L_p, $1 \leqslant p < \infty$; i.e., to functions g satisfying

$$\|g\|_p = \left\{\int_{-\infty}^{\infty} |g(t)|^p dt\right\}^{1/p} < \infty \ (1 \leqslant p < \infty).$$

Here the notation $\|g\|_p$ designates the norm of g in L_p, or simply the L_p-norm of g. The space L_∞ consists of those functions g whose magnitude is bounded on the real line, their norm $\|g\|_\infty$ being the "essential supremum" of $|g(t)|$, which for functions we will be dealing with here, is simply the maximum value of $|g(t)|$. The operator B_λ may not be applied to an arbitrary bounded function, since the associated integral may not converge. However, the integral may converge conditionally for a large class of functions; in particular, $B_\lambda g = g$, for any constant function g.

The operator B_λ is a "contraction" operator on L_2; i.e.,

$$\|B_\lambda g\|_2 \leqslant \|g\|_2,$$

with equality attaining only for g in $\mathscr{B}_2(\lambda)$. This follows from Parseval's theorem and the definition of B_λ.

Applying Schwarz's inequality to the integral equation denoted by $B_\lambda g = g$, we obtain the useful inequality

$$\|g\|_\infty \leqslant \sqrt{\lambda/\pi}\ \|g\|_2, \qquad g \quad \text{in} \quad \mathscr{B}_2(\lambda).$$

Also, it is easy to show from the integral equation that

$$\lim_{t\to\pm\infty} g(t) = 0, \qquad g \quad \text{in} \quad \mathscr{B}_2(\lambda).$$

We shall also make use of the *high-pass operator* H_λ, defined by

$$H_\lambda = I - B_\lambda,$$

where I is the identity operator. H_λ is an identity operator for functions h of L_2 whose Fourier transforms vanish over $(-\lambda, \lambda)$, and it is also a contraction operator on L_2,

$$\|H_\lambda g\|_2 \leqslant \|g\|_2,$$

with equality attaining only for $H_\lambda g = g$, i.e., for $B_\lambda g = 0$. (In these operational equations, 0 is interpreted as the null function.)

It is clear from the operator definitions and the associated Fourier transform relations that any function f in L_2 has the decomposition

$$f = g + h,$$

where

$$g = B_\lambda f, \qquad h = H_\lambda f.$$

Since λ will be fixed throughout the paper, we will, except where emphasis is desired, simply write B, H, and $\mathscr{B}_2$ for B_λ, H_λ, $\mathscr{B}_2(\lambda)$, respectively.

Now, using our notation, we may state the important result of Landau and Miranker as follows:

Theorem (Landau and Miranker): *Let $f(x)$ be a real-valued function of the real variable x, satisfying*

(*i*) $f(0) = 0$

(*ii*) $0 < m_1 \leqslant f'(x) \leqslant m_2 < \infty, \qquad (-\infty < x < \infty).$

Then to each real-valued y in $\mathscr{B}_2$ there corresponds a unique x in $\mathscr{B}_2$, also real-valued, satisfying

(*iii*) $Bf(x) = y.$

The solution x of (iii) may be obtained as the limit of the sequence of approximants $\{x_n\}$ defined iteratively by

(*iv*) $x_{n+1} = x_n - cB\{f(x_n) - y\},$

provided only that x_1 is a real-valued function in $\mathscr{B}_2$ and the real constant c is so chosen that

(v) $\max_x |1 - cf'(x)| \leq r < 1.$

The beauty of this result is that, under the hypotheses on f, every (r.v.) y in $\mathscr{B}_2$ has the representation (*iii*) where x is a unique (r.v.) function in $\mathscr{B}_2$. In some r.v. companding problems of interest, $f(x)$ may not be defined outside some interval and/or the condition on $f'(x)$ may not be satisfied over the whole real axis, but rather over some interval including the origin. Then the conclusion will apply only to y of sufficiently small norm. In such cases, the companding problem has two essentially different interpretations. The first is the *recovery problem*: y is known to be of the form (*iii*); recover x. The second is the *design problem*: y is a prescribed (desired) signal; find x, if possible, so that y is given by (*iii*). In this case, one is faced with the problem of determining for what y the problem has a solution.

The speed of covergence of the iterative solution of Landau and Miranker is a matter of practical concern. They show that

$$\| x_{n+1} - x_n \|_2 \leq r \| x_n - x_{n-1} \|_2.$$

Then the constant c in (v) should be chosen to make r as small as possible. Assuming that equality may attain on both sides in (*ii*), one should choose

$$c = \frac{2}{m_1 + m_2}, \qquad \text{giving} \quad r = \frac{m_2 - m_1}{m_2 + m_1}.$$

Thus rapid convergence is assured if (m_2/m_1) is not much larger than 1. If this is not the case, a large number of iterations are, in general, required to obtain a close approximation to the solution of the problem. In a practical implementation of the iterative scheme of solution (Ref. 1), the ideal bandlimiting operator is replaced by an approximate operator, incurring a certain delay, in addition to (eventually) significant spectral distortions, with the result that the sequence $\{x_n\}$ will not converge to the solution x. Thus, in practice, the number of iterations to be performed is limited both by practical and theoretical considerations. The conclusion is that good approximate solutions to companding problems may be conveniently obtained in practice only in those cases where the companding function $f(x)$ is fairly linear over the range of $x(t)$.

We should remark at this point that there is only one known (nonlinear) r.v. companding problem (see Ref. 3) admitting of an explicit noniterative solution; viz.,

$$B\{\log(1 + x)\} = y, \qquad x > -1, \qquad x \text{ in } \mathscr{B}_2,$$

which has a solution if, and only if, the function†

† Here we are applying the B operator to a function not in L_2, the proper interpretation being $w = 1 + B\{-1 + \exp(\cdot)\}$.

$$w(t) = B\{\exp\ 1/2[y(t) + \hat{y}(t)]\},$$

where $\hat{y}$ is the Hilbert transform of y, extends as a function zero-free in the upper half-plane, which will be the case if $\| y \|_2$ is sufficiently small. Then the solution is given by

$$x(t) = |w(t)|^2 - 1.$$

Motivated by the above considerations, pure curiosity, and the fact that in many cases of practical interest the companding function and/or its inverse can be well approximated by a polynomial of low degree over the range of interest, we are led to consider the case where the companding function has a Taylor series expansion, allowing the possibility of developing a corresponding series solution to the problem. To obtain the terms (1st order, 2nd order, etc.) in the series solution it is convenient to multiply y by a scalar parameter m, and consider the problem

$$Bf(x) = my \tag{1}$$

to be solved for x in $\mathscr{B}_2$, given y in $\mathscr{B}_2$, for companding functions

$$f(x) = \sum_1^\infty b_k x^k, \qquad |x| < R_0 \tag{2}$$

$$b_1 \neq 0.$$

For sufficiently small $|x|$, f will have an inverse ϕ,

$$x = \phi\{f(x)\}$$

$$\phi(y) = \sum_1^\infty a_k y^k, \qquad |y| \leqslant R_0^*. \tag{3}$$

We assume that the solution $x = x(t; m)$ of (1) has a series expansion in the parameter m,

$$x(t; m) = \sum_1^\infty m^k x_k(t), \tag{4}$$

where the $x_k(t)$, aptly described as kth order corrections, (not to be confused with the Landau-Miranker approximants) depend only on $y(t)$ and f. Presumably, in cases of small distortion, a few terms of the series would give a satisfactory approximation to the solution.

Explicit expressions for the first five of the $x_k(t)$ are given in the sequel, first in formulas involving the coefficients of ϕ, and next, the coefficients of f, together with certain operations on y. These formulas reveal how the Fourier transforms of the $x_k(t)$ may be calculated from the Fourier transform of $y(t)$, if this be given.

Next, we find a series development of z in $\mathscr{B}_2$,

$$z(t) = z(t; m) = \sum_1^\infty m^k z_k(t) \tag{5}$$

such that the solution to (1) is given (presumably for sufficiently small $|m|$) by

$$x = B\phi(z). \tag{6}$$

We find that $z_1 = y$, $z_2 = 0$, and in case a_2 in (3), or b_2 in (2), vanishes, we have, in addition, $z_3 = z_4 = 0$. That is, under certain conditions, $z \doteq my$, implying that $B\phi\{Bf(x)\}$ is an "approximate identity" ($\doteq x$) for x in $\mathscr{B}_2$, especially if f is odd and fairly linear, or if $x(t)$ is a predominantly low-frequency function.

To further investigate the approximate identity, we introduce the parameter m again, and obtain expressions for u_k in

$$B\phi\{Bf(mx)\} = \sum_1^\infty m^k u_k, \qquad x \text{ in } \mathscr{B}_2. \tag{7}$$

To see how interchanging f and ϕ affects the approximate identity, we compare u_k with v_k in

$$Bf\{B\phi(mx)\} = \sum_1^\infty m^k v_k, \qquad x \text{ in } \mathscr{B}_2. \tag{8}$$

As expected from the series development of z, we find $u_1 = v_1 = x$, and $u_2 = v_2 = 0$. Further comparisons [with the same m in (7) and (8)] should be made for the case $f'(0) = \phi'(0) = 1$. For this case, we find $u_3 = v_3 = 2b_2^2 B(x \cdot Hx^2)$, which may be small if b_2 is small or if Hx^2 is small. In case $b_2 = 0$, we find $u_k = v_k = 0$ for $k = 2, 3, 4$, and $u_5 = v_5 = 3b_3^2 B(x^2 \cdot Hx^3)$.

These series developments of the approximate identity suggest that it would be useful in obtaining an approximate solution to the r.v. companding problem for fairly linear companding functions, not necessarily having a Taylor series expansion, but merely satisfying $f(0) = 0$ and

$$0 < m_1 \leqslant f'(x)\ m_2 < \infty, \qquad (-\infty < x < \infty). \tag{9}$$

Compelled by this suggestion, we digress in the Appendix to show for such f that

$$\| x - B\phi\{Bf(x)\} \|_2 \leqslant \gamma \| x \|_2, \qquad x \text{ in } \mathscr{B}_2, \tag{10}$$

where

$$\gamma = \frac{\epsilon^2}{4(1+\epsilon)}, \qquad \epsilon = \frac{m_2}{m_1} - 1.$$

(Note that $\gamma = 1/8$ for $m_2/m_1 = 2$.) Thus in many companding

problems, $B\phi(y)$, involving only one filtering operation, would be an adequate approximation to the solution x. We go on to define an iterative procedure, involving both f and its inverse ϕ, obtaining approximants converging to x for $\gamma < 1$, offering an alternative to the solution of Landau and Miranker in cases where $(m_2/m_1) < 3 + 2\sqrt{2}$. In any case, $B\phi(y)$ is suggested as a good choice for x_1 in their iterative solution. We note, in leaving this topic, that the inequality (10) is invariant to the interchange of f and its inverse ϕ.

Returning to the series solution, we apply the results to the problem of compatible single-sideband transmission (Ref. 4), obtaining a few terms for correction of the "in-band" distortion arising in envelope detection of "full-carrier" single-sideband signals.

Although the original intent of the work here was to obtain expressions for the first few terms of the series solution (4), supposedly adequate for correcting small distortions, the mathematical question naturally arises in the end as to whether the series actually converges for sufficiently small $|m|$ (or equivalently, for $|m| = 1$ and $\|y\|_2$ sufficiently small), or whether it is merely an asymptotic series. It is indeed a pertinent mathematical question, since the expressions for the $x_k(t)$ were obtained by purely formal manipulations of power series and application of the operators B and H. The resulting expressions become progressively cumbersome and complicated, with no obvious general form, offering no possibility of establishing (from them) bounds on $|x_k(t)|$ which would ensure convergence of the series. The remainder of the paper is addressed to the problem of establishing the convergence of the series.

If we suppose in the r.v. companding problem that the series converges for real-valued m of sufficiently small magnitude, then it would also converge for similar complex m, suggesting that the companding problem [for fixed $y(t)$] would have a solution for all complex m of sufficiently small magnitude. This, in turn, suggests that the problem would have a solution for arbitrary *complex-valued* $y(t)$ in $\mathscr{B}_2$ and all complex m of sufficiently small magnitude, depending on f and the norm of y. That this is a fact has been established previously (Ref. 3) only for complex-valued $y(t)$ whose Fourier transforms vanish outside $[0, \lambda]$, (or $[-\lambda, 0]$), the Fourier transform of the solution $x(t)$ having the same property. In this case (with $m = 1$), the solution is given by $x = B\phi(y)$ for y of sufficiently small norm; i.e., in case the Fourier transform of $x(t)$ vanishes outside $[0, \lambda]$, (or $[-\lambda, 0]$) the "approximate identity" is an exact identity,

$$x = B\phi\{Bf(x)\} = \phi\{f(x)\}$$

for x of sufficiently small norm. This result can be explained, roughly, by the fact that nonlinear (analytic) distortion of such $x(t)$ does not

produce both "sum and difference" frequency components, but only "sum" components.

In order to prove that the series solution actually converges for sufficiently small $|m|$, we show that the companding problem (1), where $y(t)$ is an arbitrary *complex-valued* function in $\mathscr{B}_2$, has a solution $x(t; m)$ for all complex m of sufficiently small magnitude, this solution being, for each fixed t, an anaytic function of the complex variable m, from which it follows that the solution has a Taylor series expansion in m; i.e.,

$$x(t;m) = \sum_{1}^{\infty} m^k x_k(t), \qquad |m| < m_0. \tag{11}$$

To obtain this result, we first have to establish for the complex-valued (c.v.) companding problem the analogue of A. Beurling's uniqueness theorem (see Ref. 1) for the r.v. companding problem.

We then examine in detail a specific problem illustrative of the theory and some of its nuances; viz., the problem (taking $\lambda = 2$ for convenience)

$$B\left\{\frac{x}{1-x}\right\} = m\,\frac{\sin 2t}{2t} \tag{12}$$

for which the solution is (at least for sufficiently small $|m|$)

$$x = x(t;m) = 2\beta\,\frac{\sin 2t}{2t} - \beta^2\left(\frac{\sin t}{t}\right)^2, \tag{13}$$

where

$$\beta = m/(2 + m).$$

The rather surprising revelation of this example is that, although the series expansion in m of $x(t;m)$ converges, uniformly in t, for $|m| < 2$, it is not a solution of (12) for all such m. Furthermore, one might reasonably assume that (13) is a solution of (12) for all m other than -2, but this is not true either. As an illuminating exercise, we determine precisely the set of m for which (13) is a solution of (12).

II. THE INVERSE SERIES METHOD

To obtain a series solution to (1), we first think of recovering from $y(t)$ the out-of-band components of $f\{x(t)\}$, so that we might apply the inverse function ϕ to the whole in order to recover $x(t) = x(t;m)$. We have

$$f\{x(t)\} = my(t) + h(t), \tag{14}$$

where $h(t) = h(t;m)$ is some unknown "high-pass" function satisfying

$$Bh(t) = 0 \tag{15}$$

and hence

$$x(t) = \phi\{my(t) + h(t)\}. \tag{16}$$

It is convenient at this point to introduce the high-pass operator defined by

$$H = I - B, \tag{17}$$

where I is the identity operator. Thus applying H to (16) we have

$$H\phi\{my(t) + h(t)\} = 0. \tag{18}$$

We would like to solve (18) for $h(t)$, which we think of as small in cases of interest.

Now we assume that

$$\phi(y) = \sum_1^\infty a_k y^k \quad \text{for sufficiently small} \quad |y|, \tag{19}$$

$$a_1 \neq 0$$

and that in (16)

$$h(t) = h(t;m) = \sum_2^\infty m^k h_k(t), \qquad Hh_k = h_k, \qquad k \geqslant 2, \tag{20}$$

$$x(t) = x(t;m) = \sum_1^\infty m^k x_k(t), \qquad Bx_k = x_k, \qquad k \geqslant 1, \tag{21}$$

where $h_k(t)$ and $x_k(t)$ do not depend on m.

We want to expand $\phi\{my(t) + h(t)\}$ as a power series in m. To do this it is convenient to write

$$my(t) + h(t) = \sum_1^\infty m^k h_k(t), \tag{22}$$

where we identify

$$y(t) = h_1(t) \quad \text{in} \quad \mathscr{B}_2. \tag{23}$$

Then we write

$$\phi\left\{\sum_1^\infty m^k h_k(t)\right\} = F(m;t) = \sum_1^\infty m^k F_k(t). \tag{24}$$

For convenience we suppress the variable t and write simply x_k, h_k, F_k. In terms of the coefficients a_k in

$$\phi(y) = \sum_1^\infty a_k y^k,$$

we find, equating coefficients of m^k in (24),

$$F_1 = a_1 h_1 \tag{25.1}$$

$$F_2 = a_1 h_2 + a_2 h_1^2 \tag{25.2}$$

$$F_3 = a_1 h_3 + a_2(2h_1 h_2) + a_3 h_1^3 \tag{25.3}$$

$$F_4 = a_1 h_4 + a_2(2h_1 h_3 + h_2^2) + a_3(3h_1^2 h_2) + a_4 h_1^4 \tag{25.4}$$

$$F_5 = a_1 h_5 + a_2(2h_1 h_4 + 2h_2 h_3) + a_3(3h_1^2 h_3 + 3h_1 h_2^2) + a_4(4h_1^3 h_2) + a_5 h_1^5 \tag{25.5}$$

$$F_6 = a_1 h_6 + a_2(2h_1 h_5 + 2h_2 h_4 + h_3^2) + a_3(3h_1^2 h_4 + 6h_1 h_2 h_3 + h_2^3) + a_4(4h_1^3 h_3 + 6h_1^2 h_2^2) + a_5(5h_1^4 h_2) + a_6 h_1^6 \tag{25.6}$$

$$\begin{aligned} F_7 = {} & a_1 h_7 + a_2(2h_1 h_6 + 2h_2 h_5 + 2h_3 h_4) \\ & + a_3(3h_1^2 h_5 + 6h_1 h_2 h_4 + 3h_1 h_3^2 + 3h_2^2 h_3) \\ & + a_4(4h_1^3 h_4 + 12h_1^2 h_2 h_3 + 4h_1 h_2^3) \\ & + a_5(5h_1^4 h_3 + 10h_1^3 h_2^2) + a_6(6h_1^5 h_2) + a_7 h_1^7 \end{aligned} \tag{25.7}$$

$$\begin{aligned} F_8 = {} & a_1 h_8 + a_2(2h_1 h_7 + 2h_2 h_6 + 2h_3 h_5 + h_4^2) \\ & + a_3(3h_1^2 h_6 + 6h_1 h_2 h_5 + 6h_1 h_3 h_4 + 3h_2^2 h_4 + 3h_2 h_3^2) \\ & + a_4(4h_1^3 h_5 + 12h_1^2 h_2 h_4 + 6h_1^2 h_3^2 + 12h_1 h_2^2 h_3 + h_2^4) \\ & + a_5(5h_1^4 h_4 + 20h_1^3 h_2 h_3 + 10h_1^2 h_2^3) \\ & + a_6(6h_1^5 h_3 + 15h_1^4 h_2^2) + a_7(7h_1^6 h_2) + a_8 h_1^8 \end{aligned} \tag{25.8}$$

$$\begin{aligned} F_9 = {} & a_1 h_9 + a_2(2h_1 h_8 + 2h_2 h_7 + 2h_3 h_6 + 2h_4 h_5) \\ & + a_3(3h_1^2 h_7 + 6h_1 h_2 h_6 + 6h_1 h_3 h_5 + 3h_1 h_4^2 + 3h_2^2 h_5 \\ & + 6h_2 h_3 h_4 + h_3^3) + a_4(4h_1^3 h_6 + 12h_1^2 h_2 h_5 + 12h_1^2 h_3 h_4 \\ & + 12h_1 h_2^2 h_4 + 12h_1 h_2 h_3^2 + 4h_2^3 h_3) + a_5(5h_1^4 h_5 \\ & + 20h_1^3 h_2 h_4 + 10h_1^3 h_3^2 + 30h_1^2 h_2^2 h_3 + 5h_1 h_2^4) + a_6(6h_1^5 h_4 \\ & + 30h_1^4 h_2 h_3 + 20h_1^3 h_2^3) + a_7(7h_1^6 h_3 + 21h_1^5 h_2^2) \\ & + a_8(8h_1^7 h_2) + a_9 h_1^9 \end{aligned} \tag{25.9}$$

$$\begin{aligned} F_{10} = {} & a_1 h_{10} + a_2(2h_1 h_9 + 2h_2 h_8 + 2h_3 h_7 + 2h_4 h_6 + h_5^2) \\ & + a_3(3h_1^2 h_8 + 6h_1 h_2 h_7 + 6h_1 h_3 h_6 + 6h_1 h_4 h_5 + 3h_2^2 h_6 \\ & + 6h_2 h_3 h_5 + 3h_2 h_4^2 + 3h_3^2 h_4) + a_4(4h_1^3 h_7 + 12h_1^2 h_2 h_6 \end{aligned}$$

$$+ 12h_1^2h_3h_5 + 6h_1^2h_4^2 + 12h_1h_2^2h_5 + 24h_1h_2h_3h_4 + 4h_1h_3^3$$
$$+ 4h_2^3h_4 + 6h_2^2h_3^2) + a_5(5h_1^4h_6 + 20h_1^3h_2h_5 + 20h_1^3h_3h_4$$
$$+ 30h_1^2h_2^2h_4 + 30h_1^2h_2h_3^2 + 20h_1h_2^3h_3 + h_2^5) + a_6(6h_1^5h_5$$
$$+ 30h_1^4h_2h_4 + 15h_1^4h_3^2 + 60h_1^3h_2^2h_3 + 15h_1^2h_2^4)$$
$$+ a_7(7h_1^6h_4 + 42h_1^5h_2h_3 + 35h_1^4h_2^3) + a_8(8h_1^7h_3$$
$$+ 28h_1^6h_2^2) + a_9(9h_1^8h_2) + a_{10}h_1^{10}. \tag{25.10}$$

In general the coefficient of a_m in the expansion of F_n consists of sums of products of the h_k corresponding to partitions of n into m parts. The coefficient of the product is $m!$ divided by the product of the factorials of the exponents of the h_k (the multinomial theorem). For example, in F_{10} the coefficient of a_5 is found by writing down the partitions of 10 into 5 parts and proceeding thus (see Table 24.2, Ref. 5):

$1^4, 6$	$\rightarrow$ $h_1^4h_6$	coef. $= 5!/4!$	$= 5$
$1^3, 2, 5$	$h_1^3h_2h_5$	$5!/3!$	$= 20$
$1^3, 3, 4$	$h_1^3h_3h_4$	$5!/3!$	$= 20$
$1^2, 2^2, 4$	$h_1^2h_2^2h_4$	$5!/2!2!$	$= 30$
$1^2, 2, 3^2$	$h_1^2h_2h_3^2$	$5!/2!2!$	$= 30$
$1, 2^3, 3$	$h_1h_2^3h_3$	$5!/3!$	$= 20$
2^5	h_2^5	$5!/5!$	$= 1.$

Now we may obtain a formal series solution (21) by successively solving for the h_k by requiring

$$HF_k(t) = 0, \qquad k = 1, 2, \cdots \tag{26}$$

and setting

$$x_k(t) = BF_k(t) = F_k(t), \qquad k = 1, 2, \cdots. \tag{27}$$

Recall that $h_1 = y$, the given bandlimited (low-pass) function, and all the other h_k are high-pass functions.† We have $Hh_k = h_k$, $k \geq 2$ and

$$HF_1 = a_1Hh_1 = 0 \tag{28.1}$$
$$HF_2 = a_1h_2 + a_2Hh_1^2 = 0$$
$$h_2 = -\frac{a_2}{a_1}Hh_1^2 \tag{28.2}$$

† Actually, for $k \geq 2$, h_k is a bandpass function whose Fourier transform vanishes over $(-\lambda, \lambda)$ and outside $[-k\lambda, k\lambda]$. This can be seen from (28.1)–(28.k).

$$HF_3 = a_1h_3 + 2a_2H(h_1h_2) + a_3Hh_1^3 = 0$$

$$h_3 = \frac{2a_2^2}{a_1^2} H(h_1 \cdot Hh_1^2) - \frac{a_3}{a_1} Hh_1^3 \tag{28.3}$$

$$HF_4 = a_1h_4 + a_2(2H(h_1h_3) + Hh_2^2) + 3a_3H(h_1^2h_2) + a_4Hh_1^4 = 0$$

$$\begin{aligned} h_4 = &- \frac{4a_2^3}{a_1^3} H[h_1 \cdot H(h_1 \cdot Hh_1^2)] + \frac{2a_2a_3}{a_1^2} H(h_1 \cdot Hh_1^3) \\ &- \left(\frac{a_2}{a_1}\right)^3 H[(Hh_1^2)^2] + \frac{3a_2a_3}{a_1^2} H(h_1^2 \cdot Hh_1^2) - \frac{a_4}{a_1} Hh_1^4 \end{aligned} \tag{28.4}$$

$$\begin{aligned} HF_5 = {} & a_1h_5 + a_2[2H(h_1h_4) + 2H(h_2h_3)] + a_3[3H(h_1^2h_3) \\ & + 3H(h_1h_2^2)] + 4a_4H(h_1^3h_2) + a_5Hh_1^5 = 0 \end{aligned}$$

$$\begin{aligned} h_5 = {} & \frac{8a_2^4}{a_1^4} H\{h_1 \cdot H[h_1 \cdot H(h_1 \cdot Hh_1^2)]\} - \frac{4a_2^2a_3}{a_1^3} H[h_1 \cdot H(h_1 \cdot Hh_1^3)] \\ & + \frac{2a_2^4}{a_1^4} H[h_1 \cdot (Hh_1^2)^2] - \frac{6a_2^2a_3}{a_1^3} H[h_1 \cdot H(h_1^2 \cdot Hh_1^2)] \\ & + \frac{2a_2a_4}{a_1^2} H(h_1 \cdot Hh_1^4) + \frac{4a_2^4}{a_1^4} H[(Hh_1^2) \cdot H(h_1 \cdot Hh_1^2)] \\ & - \frac{2a_2^2a_3}{a_1^3} H[(Hh_1^2) \cdot Hh_1^3] - \frac{6a_2^2a_3}{a_1^3} H[h_1^2 \cdot H(h_1 \cdot Hh_1^2)] \\ & + \frac{3a_3^2}{a_1^2} H(h_1^2 \cdot Hh_1^3) - \frac{3a_2^2a_3}{a_1^3} H[h_1 \cdot (Hh_1^2)^2] \\ & + \frac{4a_2a_4}{a_1^2} H(h_1^3 \cdot Hh_1^2) - \frac{a_5}{a_1} Hh_1^5. \end{aligned} \tag{28.5}$$

Now replacing h_1 by y we have from (25), (27), and (28)

$$x_1 = BF_1 = a_1y \tag{29.1}$$

$$x_2 = BF_2 = a_2By^2 \tag{29.2}$$

$$\begin{aligned} x_3 = BF_3 &= 2a_2B(yh_2) + a_3By^3 \\ &= -\frac{2a_2^2}{a_1} B(y \cdot Hy^2) + a_3By^3 \end{aligned} \tag{29.3}$$

$$\begin{aligned} x_4 = BF_4 &= a_2(2B(yh_3) + Bh_2^2) + 3a_3B(y^2h_2) + a_4By^4 \\ &= \frac{4a_2^3}{a_1^2} B[y \cdot H(y \cdot Hy^2)] - \frac{2a_2a_3}{a_1} B(y \cdot Hy^3) \end{aligned}$$

$$+\frac{a_2^3}{a_1^2}B(Hy^2)^2-\frac{3a_2a_3}{a_1}B(y^2\cdot Hy^2)+a_4By^4 \tag{29.4}$$

$$\begin{aligned}x_5=BF_5&=a_2[2B(yh_4)+2B(h_2h_3)]+3a[B(y^2h_3)+B(yh_2^2)]\\&\quad+4a_4B(y^3h_2)+a_5By^5\\&=-\frac{8a_2^4}{a_1^3}B\{y\cdot H[y\cdot H(y\cdot Hy^2)]\}+\frac{4a_2^2a_3}{a_1^2}B[y\cdot H(y\cdot Hy^3)]\\&\quad-\frac{2a_2^4}{a_1^3}B[y\cdot H(Hy^2)^2]+\frac{6a_2^2a_3}{a_1^2}B[y\cdot H(y^2\cdot Hy^2)]\\&\quad-\frac{2a_2a_4}{a_1}B(y\cdot Hy^4)-\frac{4a_2^4}{a_1^3}B[(Hy^2)\cdot H(y\cdot Hy^2)]\\&\quad+\frac{2a_2^2a_3}{a_1^2}B[(Hy^2)\cdot(Hy^3)]+\frac{6a_2^2a_3}{a_1^2}B[y^2\cdot H(y\cdot Hy^2)]\\&\quad-\frac{3a_3^2}{a_1}B(y^2\cdot Hy^3)+\frac{3a_2^2a_3}{a_1^2}B[y\cdot(Hy^2)^2]\\&\quad-\frac{4a_2a_4}{a_1}B(y^3\cdot Hy^2)+a_5By^5.\end{aligned} \tag{29.5}$$

If in (29) we replace the H operator by $I-B$ and collect terms we obtain

$$x_1=a_1y \tag{30.1}$$

$$x_2=a_2By^2 \tag{30.2}$$

$$x_3=\frac{2a_2^2}{a_1}B(y\cdot By^2)+\left(a_3-\frac{2a_2^2}{a_1}\right)By^3 \tag{30.3}$$

$$\begin{aligned}x_4&=\frac{4a_2^3}{a_1^2}B[y\cdot B(y\cdot By^2)]-\left(\frac{4a_2^3}{a_1^2}-\frac{2a_2a_3}{a_1}\right)B(y\cdot By^3)\\&\quad+\frac{a_2^3}{a_1^2}B(By^2)^2-\left(\frac{6a_2^3}{a_1^2}-\frac{3a_2a_3}{a_1}\right)B(y^2\cdot By^2)\\&\quad+\left(a_4-\frac{5a_2a_3}{a_1}+\frac{5a_2^3}{a_1^2}\right)By^4\end{aligned} \tag{30.4}$$

$$\begin{aligned}x_5&=\frac{8a_2^4}{a_1^3}B\{y\cdot B[y\cdot B(y\cdot By^2)]\}+\left(\frac{4a_2^2a_3}{a_1^2}-\frac{8a_2^4}{a_1^3}\right)B[y\cdot B(y\cdot By^3)]\\&\quad+\left(\frac{6a_2^2a_3}{a_1^2}-\frac{12a_2^4}{a_1^3}\right)B[y\cdot B(y^2\cdot By^2)]+\frac{2a_2^4}{a_1^3}\end{aligned}$$

$$\cdot B[y\cdot B(By^2)^2] + \left(\frac{3a_2^2a_3}{a_1^2} - \frac{6a_2^4}{a_1^3}\right) B[y\cdot(By^2)^2]$$

$$+ \left(\frac{2a_2a_4}{a_1} - \frac{10a_2^2a_3}{a_1^2} + \frac{10a_2^4}{a_1^3}\right) B(y\cdot By^4) + \frac{4a_2^4}{a_1^3} B[(By^2)$$

$$\cdot B(y\cdot By^2)] + \left(\frac{6a_2^2a_3}{a_1^2} - \frac{12a_2^4}{a_1^3}\right) B[y^2\cdot B(y\cdot By^2)]$$

$$+ \left(\frac{2a_2^2a_3}{a_1^2} - \frac{4a_2^4}{a_1^3}\right) B[(By^2)\cdot(By^3)]$$

$$+ \left(\frac{3a_3^2}{a_1} - \frac{12a_2^2a_3}{a_1^2} + \frac{12a_2^4}{a_1^3}\right) B(y^2\cdot By^3)$$

$$+ \left(\frac{4a_2a_4}{a_1} - \frac{20a_2^2a_3}{a_1^2} + \frac{20a_2^4}{a_1^3}\right) B(y^3\cdot By^2)$$

$$+ \left(a_5 - \frac{6a_2a_4}{a_1} + \frac{21a_2^2a_3}{a_1^2} - \frac{14a_2^4}{a_1^3} - \frac{3a_3^2}{a_1}\right) By^5. \tag{30.5}$$

Note that if y belongs to $\mathscr{B}_2(\lambda/n)$, then $By^k = y^k$ for $k = 1, 2, \cdots, n$. In this case we will have $x_n = a_n y^n$. So the sum of all coefficients in the expressions for x_n must be a_n. If ϕ is an odd function these formulas simplify considerably. It is rather curious that if $a_2 = 0$, $a_3 \neq 0$, the coefficient of $B(y^3\cdot By^2)$ vanishes, whereas the coefficient of $B(y^2\cdot By^3)$ does not. The coefficients in (30) are more simply expressed in terms of the coefficients in the power series for f as we see below.

III. FORWARD SERIES METHOD

We can also solve (1) in the "forward" direction by writing

$$Bf(mx_1 + m^2x_2 + m^3x_3 + \cdots) = my, \tag{31}$$

where

$$f(x) = \sum_1^\infty b_k x^k. \tag{32}$$

Then applying the expansion (24) and (25) to $f(\sum m^k x_k)$ we have, equating coefficients of m^k,

$$Bb_1x_1 = y \tag{33.1}$$

$$Bb_1x_2 + Bb_2x_1^2 = 0 \tag{33.2}$$

$$Bb_1x_3 + Bb_2(2x_1x_2) + Bb_3x_1^3 = 0 \tag{33.3}$$

$$Bb_1x_4 + Bb_2(2x_1x_3 + x_2^2) + Bb_3(3x_1^2x_2) + Bb_4x_1^4 = 0 \tag{33.4}$$

$$Bb_1x_5 = Bb_2(2x_1x_4 + 2x_2x_3) + Bb_3(3x_1^2x_3 + 3x_1x_2^2)$$
$$+ Bb_4(4x_1^3x_2) + Bb_5x_1^5 = 0. \tag{33.5}$$

Then solving (33) successively for x_k, $(Bx_k = x_k)$, we have

$$x_1 = y/b_1 \tag{34.1}$$

$$x_2 = -\frac{b_2}{b_1^3}By^2 \tag{34.2}$$

$$x_3 = \frac{2b_2^2}{b_1^5}B(y\cdot By^2) - \frac{b_3}{b_1^4}By^3 \tag{34.3}$$

$$x_4 = -\frac{4b_2^3}{b_1^7}B[y\cdot B(y\cdot By^2)] + \frac{2b_2b_3}{b_1^6}B(y\cdot By^3)$$
$$- \frac{b_2^3}{b_1^7}B(By^2)^2 + \frac{3b_2b_3}{b_1^6}B(y^2\cdot By^2) - \frac{b_4}{b_1^5}By^4 \tag{34.4}$$

$$x_5 = \frac{8b_2^4}{b_1^9}B\{y\cdot B[y\cdot B(y\cdot By^2)]\} + \frac{4b_2^4}{b_1^9}B[(By^2)\cdot B(y\cdot By^2)]$$
$$+ \frac{2b_2^4}{b_1^9}B[y\cdot B(By^2)^2] - \frac{4b_2^2b_3}{b_1^8}B[y\cdot B(y\cdot By^3)]$$
$$- \frac{6b_2^2b_3}{b_1^8}B[y\cdot B(y^2\cdot By^2)] - \frac{2b_2^2b_3}{b_1^8}B[(By^2)\cdot(By^3)]$$
$$- \frac{6b_2^2b_3}{b_1^8}B[y^2\cdot B(y\cdot By^2)] - \frac{3b_2^2b_3}{b_1^8}B[y\cdot(By^2)^2]$$
$$+ \frac{2b_2b_4}{b_1^7}B(y\cdot By^4) + \frac{4b_2b_4}{b_1^7}B(y^3\cdot By^2)$$
$$+ \frac{3b_3^2}{b_1^7}B(y^2\cdot By^3) - \frac{b_5}{b_1^6}By^5. \tag{34.5}$$

These correspond to the solutions for the h_k in (28) with H and B, and a's and b's interchanged, except here $x_1 = y/b_1$ as compared by $h_1 = y$ in (28). They agree with the formulas in (30) according to the identities in reversion of series (see 3.6.25, Ref. 5)

$$a_1b_1 = 1 \tag{35.1}$$

$$a_1^3b_2 = -a_2 \tag{35.2}$$

$$a_1^5b_3 = 2a_2^2 - a_1a_3 \tag{35.3}$$

$$a_1^7b_4 = 5a_1a_2a_3 - a_1^2a_4 - 5a_2^3 \tag{35.4}$$

$$a_1^9 b_5 = 6a_1^2 a_2 a_4 + 3a_1^2 a_3^2 + 14a_4^2 - a_1^3 a_5 - 21a_1 a_2^2 a_3 \tag{35.5}$$

$$a_1^{11} b_6 = 7a_1^3 a_2 a_5 + 7a_1^3 a_3 a_4 + 84a_1 a_1^3 a_3 - a_1^4 a_6 - 28a_1^2 a_2 a_3^2 - 42a_2^5 - 28a_1^2 a_2^2 a_4 \tag{35.6}$$

$$a_1^{13} b_7 = 8a_1^4 a_2 a_6 + 8a_1^4 a_3 a_5 + 120a_1^2 a_2^3 a_4 + 180a_1^2 a_2^2 a_3^2 + 132a_2^6 - a_1^5 a_7 - 36a_1^3 a_2^2 a_5 - 72a_1^3 a_2 a_3 a_4 - 12a_1^3 a_2 a_3 a_4 - 12a_1^3 a_3^3 - 330a_1 a_1^4 a_3. \tag{35.7}$$

Of course the a's and b's may be interchanged in (35). Actually the expressions for x_k in (34) are analogous in a way to the coefficients a_k expressed in terms of the b_k as given by (35) (with the a's and b's interchanged). That is, if it were not for the B operator in eq. (33.k), we would have simply $x_k = a_k y^k$ according to the determining equations for the inverse coefficients. Because of the B operator, successive solutions for the x_k generate powers of y interposed with B operators and powers of $\{B(\cdot)\}$ in all combinations so that, for example, in x_5 we have a number of terms with coefficients $b_2^2 b_3 / b_1^8$ that combine only when B is replaced by I to give $-21 b_2^2 b_3 y^5 / b_1^8$, corresponding to the last term in (35.5). Note if all the $b_k = 1$, the sum of the integer coefficients in x_n is $(-1)^{n+1}$ as this is the case $y = f(x) = x/(1 - x)$, $x = \phi(y) = y/(1 + y)$. Note also that in the expression (34.5) for x_5, for example, there are five groups of functions with common "b" coefficients. These combine with certain weights, depending only on the coefficients of $f(x)$, to give x_5. Similarly, in the expression (29.5) obtained from the inverse function, there are again five groups of functions with common "a" coefficients that combine with certain weights, still depending only on the coefficients of $f(x)$, to give x_5. The interesting and rather puzzling fact is that the groups are not identical but overlap.

IV. A SOLUTION OF THE FORM $x(t) = B\phi\{z(t)\}$, z in $\mathscr{B}_2$

For the solution to (1) we have $x = \sum_1^\infty m^k x_k$, where according to (25) and (29), we have

$$x_1 = a_1 y$$

$$x_2 = a_1 h_2 + a_2 y^2 = a_2 B y^2$$

$$x_3 = a_1 h_3 + 2a_2 y h_2 + a_3 y^3 = 2a_2 B(y h_2) + a_3 B y^3.$$

Since

$$\phi(my) = a_1 my + a_2 m^2 y^2 + a_3 m^3 y^3 + \cdots,$$

we see that $x = B\phi(my) + \mathscr{O}(m^3)$, $m \to 0$. Then setting $m = 1$ (with y sufficiently small) we can conclude that

$$x = B\phi(y) + \mathscr{O}(y^3), \qquad y \to 0. \tag{36}$$

At least as $y \to 0$, $B\phi(y)$ is a better approximation than $\phi(y) = x + \mathscr{O}(y^2)$. Also $B\phi(y)$ could be a good approximation to x without y being small, as would be the case if y were a predominately low-frequency function (compared with its top frequency). This suggests that given y in (1) we determine a bandlimited function z, perhaps close to y, such that the solution to (1) is given by

$$x = B\phi(z). \tag{37}$$

To determine a series solution for z we set

$$z = \sum_1^\infty z_k m^k, \qquad Bz_k = z_k \tag{38}$$

and expand $\phi(z)$ in a power series in m:

$$\phi(mz_1 + m^2 z_2 + m^3 z_3 + \cdots) = \sum_1^\infty m^k F_k. \tag{39}$$

The F_k are given by (25) with z_k replacing h_k. The difference now is that the F_k are bandlimited to $[-k\lambda, k\lambda]$, i.e., $\phi(z)$ is not bandlimited (in general) with z in $\mathscr{B}_2$. However, we must have

$$BF_k = x_k, \tag{40}$$

where in terms of operations on y the x_k are given conveniently by (29). We have

$$Ba_1 z_1 = x_1 = a_1 y \tag{41.1}$$

$$B(a_1 z_2 + a_2 z_1^2) = x_2 = a_2 B y^2 \tag{41.2}$$

$$B(a_1 z_3 + 2a_2 z_1 z_2 + a_3 z_1^3) = x_3 = -\frac{2a_2^2}{a_1} B(y \cdot Hy^2) + a_3 B y^3 \tag{41.3}$$

$$\begin{aligned} B[a_1 z_4 + a_2(2z_1 z_3 + z_2^2) + a_3(3z_1^2 z_2) + a_4 z_1^4] &= x_4 \\ &= \frac{4a_2^3}{a_1^2} B[y \cdot H(y \cdot Hy^2)] - \frac{2a_2 a_3}{a_1} B(y \cdot Hy^3) \\ &\quad + \frac{a_2^3}{a_1^2} B(Hy^2)^2 - \frac{3a_2 a_3}{a_1} B(y^2 \cdot Hy^2) + a_4 B y^4 \end{aligned} \tag{41.4}$$

$$\begin{aligned} &B[a_1 z_5 + a_2(2z_1 z_4 + 2z_2 z_3) + a_3(3z_1^2 z_3 + 3z_1 z_2^2) \\ &\quad + a_4(4z_1^3 z_2) + a_5 z_1^5] = x_5 = -\frac{8a_2^4}{a_1^3} B\{y \cdot H[y \cdot H(y \cdot Hy^2)]\} \\ &\quad + \frac{4a_2^2 a_3}{a_1^2} B[y \cdot H(y \cdot Hy^3)] - \frac{2a_2^4}{a_1^3} B[y \cdot H(Hy^2)^2] \end{aligned}$$

$$+\frac{6a_2^2a_3}{a_1^2}B[y\cdot H(y^2\cdot Hy^2)]-\frac{2a_2a_4}{a_1}B(y\cdot Hy^4)$$

$$-\frac{4a_2^4}{a_1^3}B[(Hy^2)\cdot H(y\cdot Hy^2)]+\frac{2a_2^2a_3}{a_1^2}B[(Hy^2)\cdot(Hy^3)]$$

$$+\frac{6a_2^2a_3}{a_1^2}B[y^2\cdot H(y\cdot Hy^2)]-\frac{3a_3^2}{a_1}B(y^2\cdot Hy^3)$$

$$+\frac{3a_2^2a_3}{a_1^2}B[y\cdot(Hy^2)^2]-\frac{4a_2a_4}{a_1}B(y^3\cdot Hy^2)+a_5By^5. \tag{41.5}$$

Solving these equations successively for z_k we find

$$z_1=y \tag{42.1}$$

$$z_2=0 \tag{42.2}$$

$$z_3=-\frac{2a_2^2}{a_1^2}B(y\cdot Hy^2) \tag{42.3}$$

$$z_4=\frac{4a_2^3}{a_1^3}B[y\cdot H(y\cdot Hy^2)]-\frac{2a_2a_3}{a_1^2}B(y\cdot Hy^3)+\frac{a_2^3}{a_1^3}B(Hy^2)^2$$
$$-\frac{3a_2a_3}{a_1^2}B(y^2\cdot Hy^2)+\frac{4a_2^3}{a_1^3}B[y\cdot B(y\cdot Hy^2)]. \tag{42.4}$$

The first and last terms in (42.4) combine ($H+B=I$) to give $\frac{4a_2^3}{a_1^3}B(y^2\cdot Hy^2)$, which then combines with the fourth term to give

$$z_4=\frac{a_2}{a_1^3}(4a_2^2-3a_1a_3)B(y^2\cdot Hy^2)-\frac{2a_2a_3}{a_1^2}B(y\cdot Hy^3)$$
$$+\frac{a_2^3}{a_1^3}B(Hy^2)^2 \tag{42.4a}$$

$$z_5=-\frac{8a_2^4}{a_1^4}B\{y\cdot H[y\cdot H(y\cdot Hy^2)]\}$$
$$-\frac{8a_2^4}{a_1^4}B[y\cdot B(y^2\cdot Hy^2)]+\frac{a_3}{a_1^3}(4a_2^2-3a_1a_3)B(y^2\cdot Hy^3)$$
$$+\frac{a_2^2}{a_1^4}(3a_1a_3-2a_2^2)B[y\cdot(Hy^2)^2]$$
$$+\frac{4a_2}{a_1^3}(3a_2a_3-a_1a_4)B(y^3\cdot Hy^2)-\frac{2a_2a_4}{a_1^2}B(y\cdot Hy^4)$$

$$-\frac{4a_2^4}{a_1^3}B[(Hy^2)\cdot H(y\cdot Hy^2)]+\frac{2a_2^2a_3}{a_1^3}B[(Hy^2)\cdot(Hy^3)]. \tag{42.5}$$

The expressions for the third- and fourth-order terms, z_3 and z_4, are quite simple, owing to the fact that $z_1 = y$, $z_2 = 0$. Note in eq. (41.n) the "diagonal" term $a_nBz_1^n$ on the left is cancelled by the term a_nBy^n appearing in x_n on the right. Also in case $a_2 = 0$ we have

$$z_1 = y \tag{43.1}$$

$$z_2 = 0 \tag{43.2}$$

$$z_3 = 0 \tag{43.3}$$

$$z_4 = 0 \tag{43.4}$$

$$z_5 = -\frac{3a_3^2}{a_1^2}B(y^2\cdot Hy^3). \tag{43.5}$$

V. THE APPROXIMATE IDENTITY

The series development in the previous section suggests that as a practical expedient one might take

$$x \doteq B\phi(y) = B\phi\{Bf(x)\} \doteq \phi\{f(x)\} = x.$$

That is, what appears to be the naive thing to do may in fact be quite good, especially for odd functions ϕ (or f) that are not severely nonlinear. The interposition of the bandlimiting operator between a nonlinear function and its inverse and then subsequent bandlimiting is an interesting "approximate identity" that we examine further in the Appendix. One might ask how the interchange in the order of a particular function and its inverse in the transformation affects the approximate identity. The series expansion of the approximate identity may shed some light on the general problem. To keep track of the various orders, it is convenient to introduce the parameter m as before. We have

$$Bf(mx) = \sum_1^\infty m^k b_k Bx^k, \tag{44}$$

$$\phi\{Bf(mx)\} = \sum_1^\infty m^k F_k, \tag{45}$$

where the F_k are given by (25) with b_kBx^k replacing h_k, and $Bx = x$. Thus,

$$F_1 = a_1b_1x = x \tag{46.1}$$

$$F_2 = a_1b_2Bx^2 + a_2b_1^2x^2 \tag{46.2}$$

$$F_3 = a_1b_3Bx^3 + 2a_2b_1b_2x\cdot Bx^2 + a_3b_1^3x^3 \tag{46.3}$$

$$F_4 = a_1b_4Bx^4 + 2a_2b_1b_3x\cdot Bx^3 + a_2b_2^2(Bx^2)^2 + 3a_3b_1^2b_2x^2\cdot Bx^2 + a_4b_1^4x^4 \tag{46.4}$$

$$F_5 = a_1b_5Bx^5 + 2a_2b_1b_4x\cdot Bx^4 + 2a_2b_2b_3(Bx^2)\cdot(Bx^3) + 3a_3b_1^2b_3x^2\cdot Bx^3 + 3a_3b_1b_2^2x\cdot(Bx^2)^2 + 4a_4b_1^3b_2x^3\cdot Bx^2 + a_5b_1^5x^5. \tag{46.5}$$

Now we set

$$B\phi\{Bf(mx)\} = \sum_1^\infty m^k u_k, \tag{47}$$

where

$$u_k = BF_k. \tag{48}$$

Now note in (46) that if all the terms in BF_k involving x were of the form Bx^k then BF_k would vanish identically for $k \geqslant 2$ because $\phi\{f(x)\} = x$. So we will introduce the high-pass operator $H = I - B$ to collect the terms Bx^k that cancel. For example, to collect terms Bx^3 in BF_3 we write

$$B(x\cdot Bx^2) = B(x^3 - x\cdot Hx^2) = Bx^3 - B(x\cdot Hx^2).$$

Thus,

$$u_1 = BF_1 = x \tag{49.1}$$

$$u_2 = BF_2 = a_1b_2Bx^2 + a_2b_1^2Bx^2 = 0 \tag{49.2}$$

$$\begin{aligned} u_3 = BF_3 &= a_1b_3Bx^3 + 2a_2b_1b_2B(x\cdot Bx^2) + a_3b_1^3Bx^3 \\ &= a_1b_3Bx^3 + 2a_2b_1b_2B(x^3 - x\cdot Hx^2) + a_3b_1^3Bx^3 \\ &= -2a_2b_1b_2B(x\cdot Hx^2) \end{aligned} \tag{49.3}$$

$$\begin{aligned} u_4 = BF_4 &= a_1b_4Bx^4 + 2a_2b_1b_3B(x\cdot Bx^3) + a_2b_2^2B(Bx^2)^2 \\ &\quad + 3a_3b_1^2b_2B(x^2\cdot Bx^2) + a_4b_1^4Bx^4 \\ &= a_1b_4Bx^4 + 2a_2b_1b_3B(x^4 - x\cdot Hx^3) + a_2b_2^2B(x^2 - Hx^2)^2 \\ &\quad + 3a_3b_1^2b_2B(x^4 - x^2\cdot Hx^2) + a_4b_1^4Bx^4 \\ &= -2a_2b_1b_3B(x\cdot Hx^3) - (2a_2b_2^2 + 3a_3b_1^2b_2)B(x^2\cdot Hx^2) \\ &\quad + a_2b_2^2B(Hx^2)^2 \end{aligned} \tag{49.4}$$

$$
\begin{aligned}
u_5 = BF_5 &= a_1b_5Bx^5 + 2a_2b_1b_4B(x\cdot Bx^4) \\
&\quad + 2a_2b_2b_3B[(Bx^2)\cdot(Bx^3)] \\
&\quad + 3a_3b_1^2b_3B(x^2\cdot Bx^3) + 3a_3b_1b_2^2B[x\cdot(Bx^2)^2] \\
&\quad + 4a_4b_1^3b_2B(x^3\cdot Bx^2) + a_5b_1^5Bx^5 \\
&= a_1b_5Bx^5 + 2a_2b_1b_4B(x^5 - x\cdot Hx^4) \\
&\quad + 2a_2b_2b_3B[(x^2 - Hx^2)\cdot(x^3 - Hx^3)] \\
&\quad + 3a_3b_1^2b_3B(x^5 - x^2\cdot Hx^3) \\
&\quad + 3a_3b_1b_2^2B[x\cdot(x^2 - Hx^2)^2] + 4a_4b_1^3b_2B(x^5 - x^3\cdot Hx^2) \\
&\quad + a_5b_1^5Bx^5 \\
&= -2a_2b_1b_4B(x\cdot Hx^4) - (2a_2b_2b_3 + 3a_3b_1^2b_3)B(x^2\cdot Hx^3) \\
&\quad - (2a_2b_2b_3 + 6a_3b_1b_2^2 + 4a_4b_1^3b_2)B(x^3\cdot Hx^2) \\
&\quad + 2a_2b_2b_3B[(Hx^2)\cdot(Hx^3)] + 3a_3b_1b_2^2B[x\cdot(Hx^2)^2]. \qquad (49.5)
\end{aligned}
$$

Now in order to assess the symmetry or lack of symmetry in interchanging ϕ and f we can use the identities (35) to express the mixed coefficients of u_n in terms of the b_k or the a_k alone. We have

$$u_1 = x \qquad (50.1)$$

$$u_2 = 0 \qquad (50.2)$$

$$u_3 = {}_3C_1B(x\cdot Hx^2) \qquad (50.3)$$

$$u_4 = {}_4C_1B(x\cdot Hx^3) + {}_4C_2B(x^2\cdot Hx^2) + {}_4C_3B(Hx^2)^2 \qquad (50.4)$$

$$
\begin{aligned}
u_5 &= {}_5C_1B(x\cdot Hx^4) + {}_5C_2B(x^2\cdot Hx^3) + {}_5C_3B(x^3\cdot Hx^2) \\
&\quad + {}_5C_4B[(Hx^2)\cdot(Hx^3)] + {}_5C_5B[x\cdot(Hx^2)^2], \qquad (50.5)
\end{aligned}
$$

where

$${}_3C_1 = -2a_2b_1b_2 = \frac{2b_2^2}{b_1^2} = \frac{2a_2^2}{a_1^4} \qquad (50.3a)$$

$${}_4C_1 = -2a_2b_1b_3 = \frac{2b_2b_3}{b_1^2} = \frac{2a_2a_3}{a_1^5} - \frac{4a_2^3}{a_1^6} \qquad (50.4a)$$

$${}_4C_2 = -(2a_2b_2^2 + 3a_3b_1^2b_2) = -\frac{4b_2^3}{b_1^3} + \frac{3b_2b_3}{b_1^2} = -\frac{2a_2^3}{a_1^6} + \frac{3a_2a_3}{a_1^6}$$

$${}_4C_3 = a_2b_2^2 = -\frac{b_2^3}{b_1^3} = \frac{a_2^3}{a_1^6}$$

$$
\begin{aligned}
{}_5C_1 &= -2a_2b_1b_4 = \frac{2b_2b_4}{b_1^2} = -\frac{10a_2^2a_3}{a_1^7} + \frac{2a_2a_4}{a_1^6} + \frac{10a_2^4}{a_1^8} \qquad (50.5a)\\
{}_5C_2 &= -(2a_2b_2b_3 + 3a_3b_1^2b_3) = -\frac{4b_2^2b_3}{b_1^3} + \frac{3b_3^2}{b_1^2}\\
&= (2a_2^2 - a_1a_3)(2a_2^2 - 3a_1a_3)/a_1^8\\
{}_5C_3 &= -(2a_2b_3 + 6a_3b_1b_2^2 + 4a_4b_1^3b_2)\\
&= \frac{8b_2^4}{b_1^4} - \frac{12b_2^2b_3}{b_1^3} + \frac{4b_2b_4}{b_1^2} = \frac{4a_2^4}{a_1^7} - \frac{8a_2^2a_3}{a_1^7} + \frac{4a_2a_4}{a_1^3}\\
{}_5C_4 &= 2a_2b_2b_3 = -\frac{2b_2^2b_3}{b_1^3} = \frac{2a_2^2a_3}{a_1^7} - \frac{4a_2^4}{a_1^8}\\
{}_5C_5 &= 3a_3b_1b_2^2 = \frac{6b_2^4}{b_1^4} - \frac{3b_2^2b_3}{b_1^3} = \frac{3a_2^2a_3}{a_1^7}.
\end{aligned}
$$

Now if we interchange the order of ϕ and f in (45) and write

$$Bf\{B\phi(mx)\} = \sum_1^\infty m^k v_k, \qquad (51)$$

we obtain the v_k by replacing u_k by v_k in (49.k) and then interchanging the a's and b's. We have $u_1 = v_1 = x$ and $u_2 = v_2 = 0$. We should compare u_k and v_k, $k \geqslant 3$, for $f'(0) = 1 = a_1 = b_1$. Then we have

$$u_k = v_k \qquad k = 1, 2, 3. \qquad (52)$$

But we have, for example (with $a_1 = b_1 = 1$),

$$u_4 = 2b_2b_3B(x\cdot Hx^3) + (3b_2b_3 - 4b_2^3)B(x^2\cdot Hx^2) - b_2^3B(Hx^2)^2 \qquad (53.1)$$

$$v_4 = (2b_2b_3 - 4b_2^3)B(x\cdot Hx^3) + (3b_2b_3 - 2b_2^3)B(x^2\cdot Hx^2) + b_2^3B(Hx^2)^2. \qquad (53.2)$$

If, however, $b_2 = 0$ ($a_2 = 0$) we have

$$u_k = v_k = 0, \qquad k = 2, 3, 4, \qquad (53.3)$$

and if $a_1 = b_1 = 1$,

$$u_5 = v_5 = 3b_3^2B(x^2\cdot Hx^3). \qquad (53.4)$$

In case $b_2 = b_4 = b_6 = 0$, ($a_2 = a_4 = a_6 = 0$), we have

$$u_7 = \frac{3b_3b_5}{b_1^2}B(x^2\cdot Hx^5) + \left(\frac{5b_3b_5}{b_1^2} - \frac{9b_3^3}{b_1^3}\right)B(x^4\cdot Hx^3) - \frac{3b_3^3}{b_1^3}B[x\cdot(Hx^3)^2], \qquad (54)$$

where the coefficients expressed in terms of the a's are

$$\frac{3b_3b_5}{b_1^2} = \frac{3a_3a_5}{a_1^8} - \frac{9a_3^3}{a_1^9},$$

$$\frac{5b_3b_5}{b_1^2} - \frac{9b_3^3}{b_1^3} = \frac{5a_3a_5}{a_1^8} - \frac{6a_3^3}{a_1^9} - \frac{3b_3^3}{b_1^3} = \frac{3a_3^3}{a_1^9}.$$

So we do not have, in general, $u_7 = v_7$ for odd functions $f(x)$ with $f'(0) = 1$.

VI. APPLICATION TO COMPATIBLE SINGLE-SIDEBAND TRANSMISSION

The mathematical problem of compatible single-sideband transmission was formulated in Ref. 4. Given a signal $y(t)$ in $\mathscr{B}_2$ the problem is to determine m such that the equation

$$B\{\sqrt{(1 + s(t))^2 + \hat{s}(t)^2}\} = my(t) + 1 \tag{55}$$

has a solution $s(t)$ in $\mathscr{B}_2$. In (55) $\hat{s}(t)$, sometimes called the quadrature signal, is the Hilbert transform of $s(t)$, and B is operating on the envelope of the single-sideband signal. The idea is to transmit a single-sideband signal that is compatible with receivers designed for double-sideband (AM) reception. Setting

$$2s(t) + s^2(t) + \hat{s}^2(t) = x(t), \tag{56}$$

we may write (55) as

$$Bf\{x(t)\} = my(t), \tag{57}$$

where

$$f(x) = \sqrt{1 + x} - 1, \qquad x \geqslant -1. \tag{58}$$

Then $s(t)$ may be found from the solution $x(t)$ of (57). (This requires factoring $1 + x(t)$ in the form $g(t)\bar{g}(t)$, where the bandwidth of g is half the bandwidth of x.) Then with $y = f(x)$ we have for the inverse

$$x = \phi(y) = 2y + y^2. \tag{59}$$

Setting

$$x = \sum_1^\infty m^k x_k, \tag{60}$$

we have from (29) with $a_1 = 2$, $a_2 = 1$, $a_k = 0$ for $k \geqslant 3$,

$$x_1 = 2y \tag{61.1}$$

$$x_2 = By^2 \tag{61.2}$$

$$x_3 = -B(y \cdot Hy^2) \tag{61.3}$$

$$x_4 = B[y \cdot H(y \cdot Hy^2)] + 1/4B(Hy^2)^2 \tag{61.4}$$

$$x_5 = -B\{y \cdot H[y \cdot H(y \cdot Hy^2)]\} - 1/4B[y \cdot H(Hy^2)^2] - 1/2B[(Hy^2) \cdot H(y \cdot Hy^2)]. \tag{61.5}$$

Replacing H by $I - B$ we have from (30) the alternate forms

$$x_3 = B(y \cdot By^2) - By^3 \tag{61.3a}$$

$$x_4 = B[y \cdot B(y \cdot By^2)] - B(y \cdot By^3) + 1/4B(By^2)^2 - 3/2B(y^2 \cdot By^2) + 5/4By^4 \tag{61.4a}$$

$$\begin{aligned} x_5 = {} & B\{y \cdot B[y \cdot B(y \cdot By^2)]\} - B[y \cdot B(y \cdot By^3)] \\ & - 3/2B[y \cdot B(y^2 \cdot By^2)] + 1/4B[y \cdot (By^2)^2] \\ & - 3/4B[y \cdot (By^2)^2] + 5/4B(y \cdot By^4) \\ & + 1/2B[By^2) \cdot B(y \cdot By^2)] \\ & - 3/2B[y^2 \cdot B(y \cdot By^2)] - 1/2B[(By)^2 \cdot (By^3)] \\ & + 3/2B(y^2 \cdot By^3) \\ & + 5/2B(y^3 \cdot By^2) - 7/4By^5. \end{aligned} \tag{61.5a}$$

The factoring of $1 + x$ can be avoided by developing a series solution for s, $Bs = s$. We have

$$x = 2s + s^2 + \hat{s}^2$$

$$x = mx_1 + m^2x_2 + m^3x_3 + \cdots.$$

Then setting

$$s = ms_1 + m^2s_2 + m^3s_3 + \cdots \tag{62}$$

$$\hat{s} = m\hat{s}_1 + m^2\hat{s}_2 + m^3\hat{s}_3 + \cdots, \tag{63}$$

we have

$$s^2 = m^2s_1^2 + m^3 2s_1s_2 + m^4(2s_1s_3 + s_2^2) + m^5(2s_1s_4 + 2s_2s_3) + \cdots \tag{64}$$

$$\hat{s}^2 = m^2\hat{s}_1^2 + m^3 2\hat{s}_1\hat{s}_2 + m^4(2\hat{s}_1\hat{s}_3 + \hat{s}_2^2) + m^5(2\hat{s}_1\hat{s}_4 + 2\hat{s}_2\hat{s}_3) + \cdots. \tag{65}$$

Note that if s belongs to $\mathscr{B}_2(\lambda)$, then the Fourier transform* of $(s + i\hat{s})$ vanishes outside $[0, \lambda]$ and that of its complex conjugate $(s - i\hat{s})$

*Here the Fourier transform of $\hat{s}$ is $-i(sgn\omega)S(\omega)$ where $S(\omega)$ is the Fourier transform of s.

vanishes outside $[-\lambda, 0]$. Thus the Fourier transform of $e^{-i\lambda t/2}(s + i\hat{s})$ vanishes outside $[-(\lambda/2), \lambda/2]$. It follows that the Fourier transform of $s^2 + \hat{s}^2$ vanishes outside $[-\lambda, \lambda]$, and hence that the sums of the coefficients of m^n in (64) and (65) are functions whose Fourier transforms vanish outside $[-\lambda, \lambda]$.

Now we can solve successively for s_k. It is convenient to introduce the Hilbert transform (Quadrature) operator Q

$$\hat{g} = Qg \tag{66}$$

to indicate the "hat" of complicated expressions.

Equating coefficients of m_k in

$$x = 2s + s^2 + \hat{s}^2,$$

we have

$$s_1 = x_1/2 = y \tag{67.1}$$

$$\begin{aligned} s_2 &= 1/2x_2 - 1/2(s_1^2 + \hat{s}_1^2) \\ &= 1/2By^2 - 1/2(y^2 + \hat{y}^2), \end{aligned} \tag{67.2}$$

which may be written, using $Bs_2 = s_2$, as

$$s_2 = -1/2B\hat{y}^2 \tag{67.2a}$$

$$\begin{aligned} s_3 &= 1/2x_3 - (s_1 s_2 + \hat{s}_1\hat{s}_2) \\ &= -1/2B(y\cdot Hy^2) + 1/2yB\hat{y}^2 + 1/2\hat{y}\cdot QB\hat{y}^2. \end{aligned} \tag{67.3}$$

Here we may write

$$yB\hat{y}^2 = y\cdot\hat{y}^2 - y\cdot H\hat{y}^2$$

and then use $Bs_3 = s_3$ to obtain

$$\begin{aligned} s_3 &= -1/2B(y\cdot Hy^2) - 1/2B(y\cdot H\hat{y}^2) + 1/2B(y\cdot\hat{y}^2) + 1/2B(\hat{y}\cdot QB\hat{y}^2) \\ &= -1/2B[y\cdot H(y^2 + \hat{y}^2)] + 1/2B(y\cdot\hat{y}^2) + 1/2B(\hat{y}\cdot QB\hat{y}^2). \end{aligned}$$

Then since $H(y^2 + \hat{y}^2) = 0$, we have

$$s_3 = 1/2\ B(y\cdot\hat{y}^2) + 1/2\ B(\hat{y}\cdot QB\hat{y}^2) \tag{67.3a}$$

$$\begin{aligned} s_4 &= 1/2\ x_4 - (s_1 s_3 + 1/2\ s_2^2) - (\hat{s}_1\hat{s}_3 + 1/2\ \hat{s}_2^2) \\ &= 1/2\ B[y\cdot H(y\cdot Hy^2)] + 1/4\ B(Hy^2)^2 \\ &\quad - 1/2\ yB(y\cdot\hat{y}^2) - 1/2\ \hat{y}QB(y\cdot\hat{y}^2) \\ &\quad - 1/8\ (B\hat{y}^2)^2 - 1/8\ (QB\hat{y}^2)^2. \end{aligned} \tag{67.4}$$

There appears to be no simplification here. One may prefer the alternate expression (61.4a) for x_4 to eliminate the H operator. Note

that QB can be replaced by the bandlimiting quadrature operator $\hat{B}$ where

$$\hat{B}g = (\hat{B}g)(t) = \int_{-\infty}^{\infty} g(s) \frac{1 - \cos \lambda(t - s)}{\pi(t - s)} ds. \tag{68}$$

VII. THE COMPLEX-VALUED COMPANDING PROBLEM

The c.v. companding problem is considerably more complicated than the r.v. companding problem, even for the same analytic companding function. For example, if

$$f(x) = \tan^{-1}x + \epsilon x, \qquad (\epsilon > 0),$$

we know from the Landau-Miranker theory that the r.v. companding problem

$$Bf(x) = y$$

has a unique solution x in $\mathscr{B}_2$ corresponding to every real-valued y in $\mathscr{B}_2$. However, in the case of complex-valued y, this may not be true because x must then take complex values which, if the norm of y is not restricted, may be singularities of f. In addition, we are confronted with the problem of establishing the uniqueness of the solution, which may require still more severe restrictions on the norm of y.

Beurling's uniqueness proof (see Ref. 1) for the r.v. companding problem is elegant and simple: Suppose $f(x) = \mathscr{O}(|x|)$, $|x| \to 0$, and is monotone increasing, and further that

$$Bf(x_1) = y \qquad \text{and} \qquad Bf(x_2) = y,$$

with x_1, x_2, (and y) in $\mathscr{B}_2$. Then $f(x_1)$ and $f(x_2)$ belong to L_2 and $B\{f \cdot (x_1) - f(x_2)\} = 0$, i.e., the Fourier transform of $\{f(x_1) - f(x_2)\}$ vanishes over $(-\lambda, \lambda)$, and therefore $\{f(x_1) - f(x_2)\}$ must be orthogonal to $(x_1 - x_2)$. But this is impossible unless $x_1 \equiv x_2$, for otherwise $(x_1 - x_2) \cdot \{f(x_1) - f(x_2)\}$, which is everywhere non-negative, will be positive everywhere on the real axis, except at the isolated zeros (if any) of $(x_1 - x_2)$.

For establishing uniqueness in the c.v. companding problem, it would seem that the weakest analogue of monotocity should be "schlichtness" of f, i.e., that x should be confined to a region, where $f(x_1) = f(x_2)$ implies $x_1 = x_2$. This suffices to establish uniqueness of the solution in the special case where x has a one-sided Fourier transform, but we are not able to see that it suffices in the general case. We can establish the following analogue of Buerling's theorem, where, without loss of generality, we assume $f'(0)$ is positive.

Theorem 1: *Suppose* $f(0) = 0$, $f'(0) > 0$, *and* $f(z)$ *is analytic in a convex region* G *including the origin, wherein*

$$Re\{f'(z)\} > 0.$$

If $x_1(t)$ *and* $x_2(t)$ *belong to* $\mathscr{B}_2$ *and are confined to* G *for all real* t, *then*

$$Bf(x_1) = y \qquad \text{and} \qquad Bf(x_2) = y$$

imply

$$x_1(t) \equiv x_2(t).$$

Proof of Theorem 1: Since G is convex, any two points x_1 and x_2 in G can be connected by a straight line segment in G. Suppose $(x_1 - x_2) = re^{i\theta}$, where $r > 0$. Then integrating f' along the connecting line segment, we have

$$f(x_1) - f(x_2) = e^{i\theta} \int_0^r f'(x_2 + se^{i\theta})ds,$$

and hence

$$\operatorname{Re} \frac{f(x_1) - f(x_2)}{x_1 - x_2} = \operatorname{Re} \frac{1}{r} \int_0^r f'(x_2 + se^{i\theta})ds > 0.$$

In case $x_1 \to x_2$, the limit is $\operatorname{Re} f'(x_2) > 0$.

Now $\{f[x_1(t)] - f[x_2(t)]\}$ belongs to L_2 and must be orthogonal to all members of $\mathscr{B}_2$, in particular to $\{x_1(t) - x_2(t)\}$; i.e., setting

$$P(t) = (\bar{x}_1 - \bar{x}_2)\{f(x_1) - f(x_2)\} = |x_1 - x_2|^2 \frac{\{f(x_1) - f(x_2)\}}{x_1 - x_2},$$

the integral of $P(t)$ must vanish. However, we see that the real part of $P(t)$ is non-negative everywhere on the real axis, and vanishes only where $x_1 = x_2$. Since the integral of P is zero, its real part vanishes a.e. Thus the function $\{x_1(t) - x_2(t)\}$ in $\mathscr{B}_2$ vanishes a.e., and hence everywhere. □

Now we can establish that the c.v. companding problem,

$$Bf(x) = y,$$

will have a solution x, which will take values in a disk centered on the origin, wherein $\operatorname{Re}\{f'(z)\} > 0$, provided $\|y\|_2$ is sufficiently small. Then the uniqueness of the solution follows from Theorem 1.

An objectionable, but inherent, feature of companding problems (as formulated here) is that a restriction on $\|y\|_\infty$ is not sufficient to give a corresponding restriction on $\|x\|_\infty$. We can, however, establish that $\|x\|_2$ will be small if $\|y\|_2$ is small, and hence that $\|x\|_\infty$ will be small, according to the inequality (given in the introduction) for a function g in $\mathscr{B}_2(\lambda)$,

$$\|g\|_\infty \leq \sqrt{\lambda/\pi}\,\|g\|_2. \tag{69}$$

In the sequel, we assume, for convenience and without loss of generality, that $f'(0) = 1$,

$$f(z) = z + \sum_{2}^{\infty} b_k z^k, \qquad |z| < R_0, \tag{70}$$

where R_0 (perhaps ∞) is the radius of convergence of the series. We exclude the trivial case $f(z) = z$, and define

$$M(r) = \max_{|z|=r} |f'(z) - 1| \leq \sum_{2}^{\infty} k|b_k| r^{k-1}, \qquad r < R_0, \tag{71}$$

which increases steadily from 0 to ∞, allowing us to define p uniquely by

$$M(p) = 1. \tag{72}$$

Then it is clear that

$$\mathrm{Re}\{f'(z)\} > 0 \qquad \text{for} \quad |z| < p. \tag{73}$$

We are now able to establish the following result.

Theorem 2: *Let $y(t)$ be any complex-valued function in $\mathscr{B}_2 = \mathscr{B}_2(\lambda)$, satisfying*

$$\sqrt{\lambda/\pi}\, \|y\|_2 \leq \max_{0<r<p} \{r[1 - M(r)]\} = r_0[1 - M(r_0)]. \tag{74}$$

Then the companding problem

$$Bf(x) = y$$

has a unique solution $x = x(t)$ in $\mathscr{B}_2$.

Proof of Theorem 2: We can use the method of Landau and Miranker to obtain a Cauchy sequence $\{x_n\}$ converging to the solution x, provided we restrict $\|y\|_2$, in the end, to be sufficiently small that all the approximants satisfy $|x_n| < p$.

Assuming the norm of y to be sufficiently small, we take

$$x_1 = y = By, \tag{75}$$

which should be a good approximation to x for small y. Then we set, following Landau and Miranker,

$$x_{n+1} = x_n + y - Bf(x_n), \qquad n \geq 1, \tag{76}$$

so that, by induction, $Bx_n = x_n$, i.e., x_n belongs to $\mathscr{B}_2$. We have, writing the same equation for $n - 1$ and subtracting,

$$\begin{aligned} x_{n+1} - x_n &= x_n - Bf(x_n) - \{x_{n-1} - Bf(x_{n-1})\} \\ &= B[x_n - f(x_n) - \{x_{n-1} - f(x_{n-1})\}]. \end{aligned} \tag{77}$$

Now we write

$$f(x_n) - x_n - \{f(x_{n-1}) - x_{n-1}\} = \int_{x_{n-1}}^{x_n} \{f'(z) - 1\}dz. \tag{78}$$

Then assuming that

$$|x_n| \leqslant r < p \qquad \text{for all} \quad n \geqslant 1 \tag{79}$$

[and all t, suppressed in the notation $x_n = x_n(t)$] we have in (78)

$$\left| \int_{x_{n-1}}^{x_n} \{f'(z) - 1\}dz \right| \leqslant M(r)\,|x_n - x_{n-1}|, \tag{80}$$

where

$$M(r) < M(p) = 1.$$

Substituting in (77) the inequality (80) for the magnitude of the function in (78), we obtain

$$\|x_{n+1} - x_n\|_2 \leqslant M(r)\,\|x_n - x_{n-1}\|_2. \tag{81}$$

So, under the assumption (79), $\{x_n\}$ forms a Cauchy sequence converging to x in $\mathscr{B}_2$ [cf. Landau, Ref. 1]. It follows from (76) that

$$Bf(x) = y. \tag{82}$$

Now we would like to see how large $\|y\|_2$ may be in order that (79) hold, giving the conclusion in (82). We write

$$x_n = x_1 + (x_2 - x_1) + (x_3 - x_2) + \cdots + (x_n - x_{n-1}) \tag{83}$$

from which follows

$$\|x_n\|_2 \leqslant \sum_1^n \|x_k - x_{k-1}\|_2, \tag{84}$$

where

$$x_1 = y \qquad \text{and} \qquad x_0 = 0.$$

Applying (81) to (84), we have

$$\|x_n\|_2 \leqslant \frac{1 - \alpha^n}{1 - \alpha}\,\|y\|_2, \tag{85}$$

where $\alpha = M(r) < 1$, provided (79) holds. This will be the case, according to (69), if

$$\|x_n\|_2 \leqslant \sqrt{\pi/\lambda}\,r \qquad \text{for all} \quad n \geqslant 1, \tag{86}$$

which, in turn, will hold if in (85) we have

$$\sqrt{\pi/\lambda}\,\|y\|_2 \leq r[1 - M(r)]. \tag{87}$$

Here we are free to take the maximum over r. Thus the problem will have a solution x satisfying the hypotheses, provided the norm of y satisfies

$$\sqrt{\pi/\lambda}\,\|y\|_2 \leq \max_{0<r<p}\{r(1 - M(r)\} = r_0[1 - M(r_0)], \tag{88}$$

and the solution is unique according to Theorem 1. □

We note that if (88) is satisfied, then the solution x satisfies, according to (85) and (69),

$$\|x\|_\infty \leq r_0 < p. \tag{89}$$

So, in fact, the restriction (88) on the norm of y is too severe. We obtain a slightly better result later, using a different method.

We now wish to show that if y_1 and y_2 are close to each other, then the corresponding solutions, x_1 and x_2, are also close to each other.

Lemma: *Let y_1 and y_2 satisfy the hypotheses on y in Theorem 2. Then the solutions of*

$$Bf(x_1) = y_1 \qquad \textit{and} \qquad Bf(x_2) = y_2$$

satisfy

$$\|x_1 - x_2\|_2 \leq \frac{\|y_1 - y_2\|_2}{1 - M(r_0)} \leq 2\sqrt{\pi/\lambda}\, r_0 \tag{90}$$

$$\|x_1 - x_2\|_\infty \leq \sqrt{\pi/\lambda}\,\frac{\|y_1 - y_2\|_2}{1 - M(r_0)} \leq 2r_0. \tag{91}$$

Proof of the Lemma: We have

$$x_1 - x_2 = y_1 - y_2 - B[f(x_1) - x_1 - f(x_2) + x_2], \tag{92}$$

giving

$$\|x_1 - x_2\|_2 \leq \|y_1 - y_2\|_2 + \|B[\cdot]\|_2 \leq \|y_1 - y_2\|_2 + \|[\cdot]\|_2. \tag{93}$$

Also, since, according to (89),

$$|x_1| \leq r_0 < p \qquad \text{and} \qquad |x_2| \leq r_0 < p,$$

we have from (78) and (80),

$$\|f(x_1) - x_1 - f(x_2) + x_2\|_2 \leq M(r_0)\,\|x_1 - x_2\|_2, \tag{94}$$

which with (93) gives

$$\|x_1 - x_2\|_2 \leq \frac{\|y_1 - y_2\|_2}{1 - M(r_0)}. \tag{95}$$

This, with (69) and the assumptions on y_1 and y_2, establishes the lemma. □

With this Lemma and Theorem 2 we can show that the problem

$$Bf\{x(t;m)\} = my(t),$$

for fixed y in $\mathscr{B}_2$, has a unique solution in $\mathscr{B}_2$ for all complex m of sufficiently small magnitude, the solution $x(t; m)$ being a continuous function of the complex variable m in a certain disk centered on the origin. To establish for each t that $F(m;t) = x(t;m)$ is an analytic function of m in that disk, we show that $F(m)$ has a derivative (nondirectional) there. Working with the derivative we are able to improve on Theorem 2. It is convenient now to set $\sqrt{\lambda/\pi}\,\|y\|_2 = 1$ so that $|y(t)| \leqslant 1$.

Theorem 3: *Let $y(t)$ be any complex-valued function in $\mathscr{B}_2 \equiv \mathscr{B}_2(\lambda)$ satisfying*

$$\sqrt{\lambda/\pi}\,\|y\|_2 = 1.$$

Then the problem

$$Bf\{x(t;m)\} = my(t)$$

has a unique solution $x(t;m)$ in $\mathscr{B}_2$ for all complex m satisfying

$$|m| \leqslant \alpha(p) = \int_0^p [1 - M(r)]dr, \tag{96}$$

where $M(r)$ and p are defined in (71) *and* (72). *Furthermore, for each fixed real t, $x(t;m)$ is an analytic function of m, $|m| \leqslant \alpha(p)$, and hence, since $x(t; 0) \equiv 0$,*

$$x(t;m) = \sum_1^\infty m^k x_k(t), \qquad |m| \leqslant \alpha(p), \qquad (-\infty < t < \infty) \tag{97}$$

where the $x_k(t)$ depend only on $y(t)$ and f.

We note, before proving Theorem 3, that in Theorem 2, $0 < r_0 < p$, and in Theorem 3

$$\alpha(p) = \int_0^{r_0} [1 - M(r)]dr + \int_{r_0}^p [1 - M(r)]dr,$$

where $M(r)$ increases from 0 to 1 over $(0, p)$. Thus

$$\int_0^{r_0} [1 - M(r)]dr > r_0[1 - M(r_0)],$$

and hence

$$r_0[1 - M(r_0)] < \alpha(p) < p. \tag{98}$$

Then, according to Theorem 3, the c.v. companding problem

$$Bf(x) = y$$

has solutions for y of larger norm than specified in Theorem 2.

Proof of Theorem 3: We first consider the solutions $x_1 = x(t;m_1)$, $x_2 = x(t;m_2)$, corresponding to $y_1 = m_1 y(t)$, $y_2 = m_2 y(t)$, where

$$m_2 = m_1 + \epsilon, \qquad \epsilon = |\epsilon| e^{i\theta} \tag{99}$$

and

$$|m_1| + |\epsilon| \leq r_0[1 - M(r_0)], \tag{100}$$

so that Theorem 2 and the Lemma apply.

We have

$$y_2 - y_1 = \epsilon y(t), \tag{101}$$

and hence, from the Lemma,

$$\| x_2 - x_1 \|_2 \leq \frac{|\epsilon|}{1 - M(r_0)} \cdot \| y \|_2 = \frac{|\epsilon| \sqrt{\pi/\lambda}}{1 - M(r_0)}. \tag{102}$$

Now

$$B\{f(x_2) - f(x_1)\} = \epsilon y, \tag{103}$$

which we rewrite as

$$\frac{x_2 - x_1}{\epsilon} = y - B\left\{\frac{f(x_2) - f(x_1) - (x_2 - x_1)}{\epsilon}\right\}. \tag{104}$$

We intend to let $\epsilon \to 0$ (with any argument) and show that the quantity on the left tends to a limit, independent of $\arg(\epsilon)$; viz., $F'(m_1;t)$, where

$$F'(m;t) = \frac{\partial}{\partial m} x(t;m), \qquad |m| \leq r_0[1 - M(r_0)]. \tag{105}$$

From (78), (80), and (102) we have

$$\| f(x_2) - f(x_1) - (x_2 - x_1) \|_2 \leq M(r_0) \| x_2 - x_1 \|_2 \leq \frac{|\epsilon| \sqrt{\pi/\lambda} M(r_0)}{1 - M(r_0)}. \tag{106}$$

So

$$\frac{f(x_2) - f(x_1) - (x_2 - x_1)}{\epsilon} \quad \text{belongs to} \quad L_2. \tag{107}$$

We also have from (102), or the Lemma,

$$\| x_2 - x_1 \|_\infty \leqslant \frac{|\epsilon|}{1 - M(r_0)}. \tag{108}$$

Thus we may write

$$x_2 = x_1 + \epsilon u, \tag{109}$$

where

$$u = u(t;m_1, \epsilon) \quad \text{belongs to } \mathscr{B}_2 \quad \text{and} \quad |u| = \mathscr{O}(1) \quad \text{as} \quad \epsilon \to 0. \tag{110}$$

Then

$$\begin{aligned} f(x_2) - f(x_1) &= f(x_1 + \epsilon u) - f(x_1) \\ &= \epsilon u f'(x_1) + \mathscr{O}(\epsilon^2 u^2). \end{aligned}$$

So (104) may be rewritten as

$$u = y - B\{uf'(x_1) - u + \mathscr{O}(\epsilon u^2)\}. \tag{111}$$

Now letting $\epsilon \to 0$ and replacing m_1 by m and x_1 by $x(t;m)$, we obtain, setting $u(t;m, o) = u(t;m)$, the equation

$$\begin{aligned} u(t;m) &= y(t) - B\{u(t;m)[f'\{x(t;m)\} - 1]\}, \\ |m| &\leqslant A < r_0[1 - M(r_0)]. \end{aligned} \tag{112}$$

Here we make the identification

$$u(t;m) = F'(m;t) = \frac{\partial}{\partial m} x(t;m), \qquad |m| \leqslant A \tag{113}$$

by verifying that (112) has a solution $u(t;m)$ in $\mathscr{B}_2$, in fact, for $|m|$ larger than $r_0[1 - M(r_0)]$. We observe, since $x(t;0) \equiv 0$, and $f'(0) = 1$, that

$$u(t;0) = y(t). \tag{114}$$

Actually, we can obtain better estimates for $|x(t;m)|$ by integrating its partial derivative from 0 to m.

We consider the equation for u,

$$u = y - B\{u \cdot [f'(x) - 1]\}, \tag{115}$$

assuming $x = x(t; m)$ is known and satisfies

$$\| x \|_\infty \leqslant r < p, \tag{116}$$

so that

$$|f'(x) - 1| \leqslant M(r) < M(p) = 1. \tag{117}$$

Using this inequality in (115) we obtain

$$\| u \|_2 \leq \| y \|_2 + M(r) \| u \|_2, \tag{118}$$

or

$$\| u \|_2 \leq \frac{\| y \|_2}{1 - M(r)} = \frac{\sqrt{\pi/\lambda}}{1 - M(r)}. \tag{119}$$

The last inequality implies that (115) has a solution u in $\mathscr{B}_2$ (obtained iteratively), in fact, for $\| x \|_\infty < p$, since $M(r) < 1$ for $r < p$. We note further that the inequality (119) is crude, with equality possible only for $r = 0$, for we cannot have

$$| f'\{x(t;m)\} - 1 | \equiv M(r), \qquad (-\infty < t < \infty)$$

unless $x(t;m) \equiv 0$. Therefore, in (115) we have

$$\| B\{u[f'(x) - 1]\|_2 < M(r) \| u \|_2 \quad \text{for} \quad 0 < \| x \|_\infty \leq r < p.$$

So we have strict inequality in (119) for $0 < r < p$. Hence,

$$\| u \|_\infty < \frac{1}{1 - M(r)} \quad \text{for} \quad 0 < \| x \|_\infty \leq r < p. \tag{120}$$

Now let us set

$$m = \alpha e^{i\theta}, \qquad \alpha > 0 \tag{121}$$

and

$$r(\alpha) = \max_\theta \| x(t;\alpha e^{i\theta}) \|_\infty. \tag{122}$$

We want to see how large we can make α, say $\alpha(p)$, and have $r(\alpha) < p$. Using (120) in

$$x(t;m) = \int_0^m u(t;\xi) d\xi, \tag{123}$$

we obtain the inequality

$$r(\alpha) < \int_0^\alpha \frac{d\xi}{1 - M[r(\xi)]}, \qquad 0 < \alpha \leq \alpha(p). \tag{124}$$

Then, after defining

$$s(\alpha) = \int_0^\alpha \frac{d\xi}{1 - M[s(\xi)]}, \qquad 0 < \alpha \leq \alpha(p), \tag{125}$$

it is clear, since $M(r)$ is an increasing function of r, that we will have

$$r(\alpha) < s(\alpha), \qquad 0 < \alpha \leq \alpha(p). \tag{126}$$

Differentiating (125) with respect to α, we obtain the simple equation

$$s'(\alpha)\{1 - M[s(\alpha)]\} = 1,$$

or, considering α as a function of s,

$$\frac{d\alpha}{ds} = 1 - M(s).$$

Thus

$$\alpha(s) = \int_0^s [1 - M(r)]dr, \qquad 0 < s \leqslant p. \tag{127}$$

We have $s(\alpha(p)) = p$ and $r(\alpha) < s(\alpha)$ for $\alpha > 0$. So we will have

$$\| x(t;m) \|_\infty < p \quad \text{for} \tag{128}$$

$$| m | \leqslant \alpha(p) = \int_0^p [1 - M(r)]dr. \tag{129}$$

According to (120) and (128), the partial derivative $u(t;m)$ will exist for $| m |$ somewhat larger than $\alpha(p)$. This completes the proof of Theorem 3. □

VIII. AN ILLUSTRATIVE EXAMPLE

It will be shown in a future paper that the r.v. companding problem

$$B\left\{\frac{x}{1 - x}\right\} = y, \qquad x < 1, \qquad x, y \text{ in } \mathscr{B}_2(\lambda), \tag{130}$$

is equivalent to finding the reproducing kernel for a certain Hilbert space of bandlimited functions. The specific problem with $\lambda = 2$ (for convenience) and

$$y = m \frac{\sin 2t}{2t} \tag{131}$$

is quite easily solved. For real $m > -2$, the solution is

$$x(t;m) = 2\beta \frac{\sin 2t}{2t} - \beta^2 \left(\frac{\sin t}{t}\right)^2, \tag{132}$$

where

$$\beta = m/(2 + m).$$

We need not be concerned here with the derivation of this solution, as we will later show directly that it satisfies

$$B\left\{\frac{x(t;m)}{1 - x(t;m)}\right\} = m \frac{\sin 2t}{2t} \tag{133}$$

for all m in a certain region of the complex plane, but for no other m.

We know from Theorem 3 that (133) has a unique solution for sufficiently small $|m|$, and that the solution is an analytic function of m. It follows that (132) is the solution of (133) for all complex m, $|m| < |m_0|$, for some $|m_0| > 0$. Since $m = -2$ is the only point where $x(t;m)$ is not analytic, we might suppose $|m_0| = 2$. The series expansion of $x(t;m)$ certainly converges (uniformly in t) for all $|m| < 2$, but it is not a solution of (133) for all such m. For example, we have

$$x\left(\pm\frac{\pi}{2};m\right) = -\left(\frac{2\beta}{\pi}\right)^2$$

and

$$m = \frac{2\beta}{1-\beta}.$$

Then for $\beta = \pm i\pi/2$, we have

$$m = \frac{-2}{1 \pm i\dfrac{2}{\pi}}, \qquad \text{and} \quad x\left(\pm\frac{\pi}{2};m\right) = 1.$$

Therefore, the meromorphic function of t,

$$f\{x(t;m)\} = \frac{x(t;m)}{1 - x(t;m)}$$

will have poles at $t = \pm\pi/2$ for $m = -2/(1 \pm i2/\pi)$. Thus we have here an example, $|m| < 2$, for which (132) is not a solution of (133). However, it is, according to Theorem 3, for all m satisfying $|m| \leqslant 3 - 2\sqrt{2}$. This, as it turns out, is an overly conservative upper bound for $|m|$.

We now turn to the problem of determining precisely those m for which (132) is a solution of (133).

First we can easily show that the r.v. problem has no solution for $m \leqslant -2$ by convolving both sides of (133) with $1/\pi K(t)$, where $K(t) = (\sin t)^2/t^2$. The result is

$$\int_{-\infty}^{\infty} \frac{x(s;m)}{1 - x(s;m)} \cdot \frac{1}{\pi} K(t-s)ds = \frac{m}{2} K(t). \tag{134}$$

Since $x/(1-x) > -1$, and $K(t) \geqslant 0$, we have

$$\int_{-\infty}^{\infty} \frac{x(s;m)}{1 - x(s;m)} \cdot \frac{1}{\pi} K(t-s)ds > -\frac{1}{\pi}\int_{-\infty}^{\infty} K(t)dt = -1$$

This gives, setting $t = 0$ in (134),

$$m > -2. \tag{135}$$

Now to proceed towards our stated goal, we first write

$$1 - x(t;m) = \left(1 - \beta e^{it} \frac{\sin t}{t}\right)\left(1 - \beta e^{-it} \frac{\sin t}{t}\right). \tag{136}$$

Then in order for $x(t;m)$ to be a solution of the problem, the Fourier transform of the function

$$h(t;m) = \frac{x(t;m)}{1 - x(t;m)} - m \frac{\sin 2t}{2t} \tag{137}$$

must vanish over $(-2,2)$. With a bit of manipulation we arrive at the expression

$$h(t;m) = g(t;m) + g(-t;m), \tag{138}$$

where

$$g(t;m) = \frac{-\beta^2}{1 - \beta} \frac{e^{2it}\left(1 - e^{it} \frac{\sin t}{t}\right)}{2it\left(1 - \beta e^{it} \frac{\sin t}{t}\right)} \tag{139}$$

and

$$\beta = \frac{(m/2)}{1 + (m/2)} \neq 1.$$

We now introduce the complex variable $\tau = t + iu$, and observe that

$$|g(t + iu;\, m)| = \mathscr{O}\left\{\frac{e^{-2u}}{|t + iu|}\right\}, \qquad u \to +\infty. \tag{140}$$

Then if the denominator satisfies the condition,

$$\left(1 - \beta e^{i\tau} \frac{\sin \tau}{\tau}\right) \textit{ is zero-free} \quad \text{for} \quad u \geqslant 0, \tag{141}$$

it is easy to see (by contour integration in the upper half-plane) that

$$G(\omega;m) = \int_{-\infty}^{\infty} g(t;m)e^{-i\omega t}dt = 0 \quad \text{for} \quad \omega < 2. \tag{142}$$

On the other hand, if the function in (141) has zeros τ_k in the upper half plane $u > 0$, (it must have no real zeros in order for g to have a Fourier transform) we will have, by the calculus of residues,

$$G(\omega;m) = \sum_1^n c_k e^{-i\omega\tau_k}, \qquad c_k \neq 0, \tag{143}$$

where n, depending on β, is finite, since it is clear that the function in (141) can have only a finite number of zeros in the upper half-plane. Since the Fourier transform of $h(t;m)$ is given by

$$H(\omega, m) = G(\omega;m) + G(-\omega;m) \tag{144}$$

it will vanish over $(-2,2)$ if, and only if, the condition (141) is satisfied.

Now the values taken by $e^{i\tau} \sin \tau/\tau$ in the upper half-plane, $u \geq 0$ are precisely those values on the boundary and interior of the cardioid-like region whose boundary is described parametrically by

$$e^{it} \frac{\sin t}{t}, \qquad -\pi \leq t \leq \pi.$$

(Some values are taken more than once.) Then $x(t;m)$ will be a solution to the problem except for those values of m such that $1/\beta$ is a point on the boundary or in the interior of the cardioid-like region. By the mapping

$$m = \frac{2}{\frac{1}{\beta} - 1},$$

$x(t;m)$ is a solution to the problem for precisely those (finite) m lying in the region to the right of the boundary line described parametrically by

$$m = \frac{2}{e^{it} \frac{\sin t}{t} - 1}, \qquad -\pi \leq t \leq \pi. \tag{145}$$

This region (see Fig. 1) includes the half-plane $\mathrm{Re}\{m\} \geq -4/3$, its boundary indenting more to the left near the real axis, having a cusp at its leftmost point, $m = -2$, where it is tangent to the real axis. It is found that the distance from the origin to the boundary is minimal (see circle in Fig. 1) at the point m_0 and its conjugate, where

$$\begin{aligned} m_0 &= (-2 + \xi) + i\xi, \\ \xi &= \frac{1}{t_0} \doteq .4895273114 \doteq (2.42786943)^{-1}, \end{aligned} \tag{146}$$

t_0 is the smallest positive root of $\sin t/t = \cos t + \sin t$,

$$|m_0| = \frac{\sqrt{2}}{\sin t_0} \doteq 1.58781760, \ \arg\{m_0\} = \frac{\pi}{4} + t_0.$$

So $|m_0|$ is the largest number such that $x(t;m)$ is a solution for *all* m

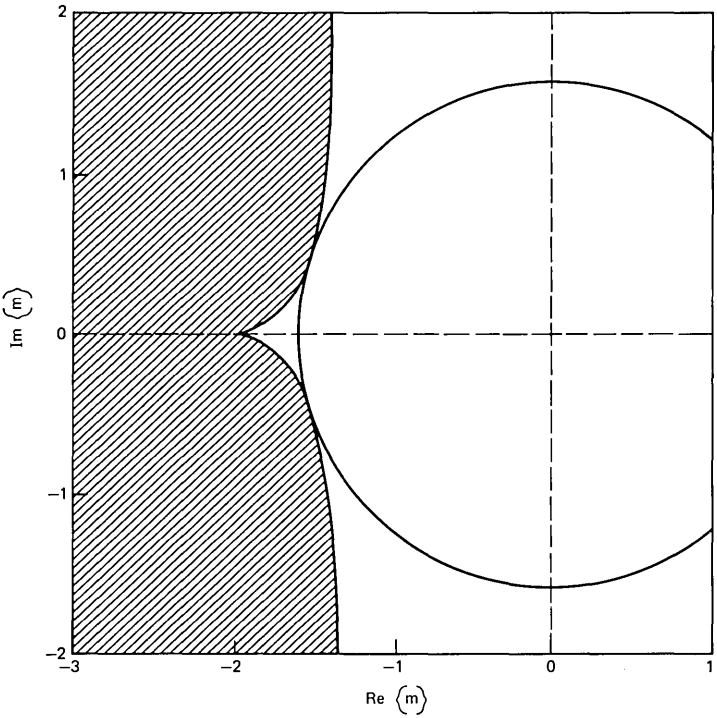


Fig. 1—Open region in m-plane (unshaded) for which eq. (132) is a solution of eq. (133). Shown is the largest disk (centered on the origin) contained in the region.

satisfying $|m| < |m_0|$. Also,

$$|m_0| = \min_t R_0(t),$$

where $R_0(t)$ is the radius of convergence of the series

$$\frac{x(t;m)}{1 - x(t;m)} = m \frac{\sin 2t}{2t} + m^2 h_2(t) + m^3 h_3(t) + \cdots,$$

the minimum occurring for $t = \pm t_0$.

IX. CONCLUSION

The expressions for the nth order components $x_n(t)$ of the series solution to the companding problem become so complicated that, for practical purposes, only the first few are of interest. These should be useful in correcting small distortions in nonlinear transmission systems which fit the companding model. It would appear that the corrections applied internally to the inverse function (the z_k in Section V) would be more effective for correcting larger distortions, especially if the lower frequency components are predominant in the signals. In

this connection, the simpler "approximate identity" should be quite effective for correcting small to moderate distortions of a more general nature, as evidenced by the inequality given in the appendix. Experimental evidence of the effectiveness of these correction schemes would be desirable.

The question of the convergence of the series solution is a matter of little practical concern, but the fact that it does attaches more mathematical significance to the results. To settle this question we had to show that the complex-valued companding problem has solutions for functions of sufficiently small norm. This generalizes the result for functions of one-sided spectra; and whether or not the general result will ever find practical application, it is an interesting addition to a theory, though still incomplete in many respects. For example, it is doubtful that the condition that $x(t)$ be confined to a convex region G where $\mathrm{Re}\{f'\} > 0$ is a necessary condition for uniqueness of the solution. In this connection, one could probably use analytic continuation arguments to show that the specific problem examined in Section VIII has solutions only for those values of m for which the (particular) solution given is the only solution, this being unique for sufficiently small $|m|$, and being an analytic function of m having no branch points. Also there is the difficult question of determining for what $y(t)$ the companding problem has a solution, where particular interest is attached to the real-valued problem with analytic companding functions. It can be shown for the case $f(x) = x/(1 - x)$, $x < 1$, that the problem has a solution for every (r.v.) y in $\mathscr{B}_2$ satisfying $y > -1$. This suggests (conjecture) that the r.v. companding problem with $f(x) = x/(1 - x^2)$, $-1 < x < 1$, has a solution for every (r.v.) y in $\mathscr{B}_2$, or more generally for monotone $f(x)$ defined on $(-1, 1)$ having singularities as strong as poles at ± 1. In general, it is not enough for $f(x)$ to increase from $-\infty$ to $+\infty$ over its range of definition in order to draw the same conclusion; e.g., $f(x) = \log(1 + x)$, $x > -1$. The questions raised here are certainly worthy of future consideration.

In connection with the series solution, one naturally inquires whether an explicit formula (albeit complicated and involving partitions of various kinds) can be given for the general term x_n. Perhaps the combinatorics experts will consider this question.

We note that the solution $x = B\phi(y)$, valid for y (of sufficiently small norm) whose Fourier transforms vanish outside $[0, \lambda]$, is verified by the fact that in (29.n) the expression reduces to $x_n = a_n B y^n$, the other terms vanishing because B is operating on functions whose Fourier transforms vanish over $(-\infty, \lambda)$. The same reduction occurs in the expression (34.n), because, in this case, B is operating on functions whose Fourier transforms vanish over $(-\infty, 0)$ and agree over $[0, \lambda]$; i.e.,

$$B[\cdots] = By^n$$

holds for each term in (34.n), the sum of all the coefficients being a_n.

Some interesting identities are obtained by equating the expressions for x_n in the series solution of the general problem to those obtained from the explicit solution (given in the introduction) to the special problem

$$B \log(1 + x) = my,$$

which involve $\hat{y}$, the Hilbert transform of y, which does not appear in the more general expressions. For example, we find from (34.3) that x_3 in the series solution of this problem is given by

$$x_3 = 1/2B(y \cdot By^2) - 1/3By^3,$$

and, from the series expansion of the explicit solution, by

$$x_3 = 1/8yB(y^2 - \hat{y}^2) - 1/8B(y\hat{y}^2) + 1/24\,By^3 + 1/4\hat{y}B(y\hat{y}).$$

It is an interesting exercise to show directly that these two expressions are identical.

Finally, since truly bandlimited signals exist only as mathematical abstractions, some attention should be given to developing a mathematical theory of practical companding problems,

$$\int_{-\infty}^{\infty} f\{x(s)\}k(t - s)ds = y(t),$$

where $k(t)$ is the (absolutely-integrable) impulse response of a practical low-pass filter, so that the theory may be extended to signals that are merely bounded. Here one may not be interested, for various reasons, in the exact solution of this problem but, instead, a compromise problem, where the equation is nearly satisfied with both $x(t)$ and $y(t)$ being close to bandlimited functions. For example, in many cases $f\{x(t)\}$ is given (say) by an nth order differential operator acting on $y(t)$. Then the (exact) solution $x(t) = \phi\{f(x)\}$ may be far from a bandlimited function. However, if $y(t)$ is close to a bandlimited function there should be an approximate solution which is also close to a bandlimited function. A case in point is found in Landau's simulation of the iterative solution of the companding problem (Ref. 2), where, in fact, the equation he was obtaining approximate solutions to was the case $y(t) = k(t)$, (approximately bandlimited) for which the unique solution is (a multiple of) the Dirac delta function.

X. ACKNOWLEDGMENT

I am grateful to H. J. Landau for helpful discussions, to J. C.

Lagarias for aid in preparation of the manuscript, and to A. M. Odlyzko for the production of Fig. 1.

REFERENCES

1. H. J. Landau, "On the Recovery of a Band-Limited Signal, After Instantaneous Companding and Subsequent Band-Limiting," B.S.T.J., *39,* No. 2 (March 1960), pp. 351–64.
2. H. J. Landau and W. L. Miranker, "The Recovery of Band-Limited Signals," J. Math. Anal. and App., *2,* No. 1 (February 1961), pp. 97–104.
3. B. F. Logan, "Theory of Analytic Modulation Systems," B.S.T.J., *57,* No. 3 (March 1978), pp. 491–576.
4. B. F. Logan and M. R. Schroeder, "A Solution to Problem of Compatible Single-Sideband Transmission," IRE Trans. Info. Th., *IT-8,* No. 5 (September 1962), pp. 252–9.
5. M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions,* New York: Dover, 1965.

APPENDIX

Suppose $f(x)$ is a monotone increasing function of the real variable x, satisfying

$$(i) \qquad f(0) = 0$$

$$(ii) \qquad 0 < m_1 \leqslant f'(x) \leqslant m_2 < \infty, \qquad (-\infty < x < \infty).$$

Then f has an inverse ϕ

$$(iii) \qquad x = \phi\{f(x)\}, \qquad (-\infty < x < \infty),$$

satisfying, since $1 = f'(x)\phi'\{f(x)\}$,

$$(iv) \qquad 0 < \frac{1}{m_2} \leqslant \phi'(y) \leqslant \frac{1}{m_1} < \infty, \qquad (-\infty < y < \infty).$$

Now let $x = x(t)$ be any function in $\mathscr{B}_2$. We wish to establish

$$\| x - B\phi\{Bf(x)\}\|_2 \leqslant \gamma \| x \|_2, \tag{147}$$

where

$$\gamma = \frac{\epsilon^2}{4(1+\epsilon)}, \qquad \epsilon = \frac{m_2}{m_1} - 1.$$

Set

$$y(t) = y = Bf(x). \tag{148}$$

Then

$$f(x) = y + h, \qquad Bh = 0. \tag{149}$$

Now set

$$x_1 = B\phi(y) \tag{150}$$

so that, since $x = \phi(y + h) = Bx = B\{\phi(y + h)\}$,

$$x - x_1 = B\{\phi(y + h) - \phi(y)\}. \tag{151}$$

Since $|\phi(y+h)-\phi(y)| \leq \frac{1}{m_1} h$, we see that

$$\|\phi(y+h)-\phi(y)\|_2 \leq \frac{1}{m_1}\|h\|_2$$

and hence that

$$\|x-x_1\|_2 \leq \frac{1}{m_1}\|h\|_2,$$

but we can improve this inequality by writing

$$\phi(y+h)-\phi(y) = \int_y^{y+h} \{\phi'(\xi)-\alpha\}d\xi + \alpha h, \tag{152}$$

where

$$\alpha = \frac{1}{2}\left(\frac{1}{m_1}+\frac{1}{m_2}\right).$$

Then, since

$$|\phi'(\xi)-\alpha| \leq \frac{1}{2}\left(\frac{1}{m_1}-\frac{1}{m_2}\right), \tag{153}$$

we have

$$\phi(y+h)-\phi(y) = u + \alpha h, \tag{154}$$

where $u = u(t)$ and

$$|u| \leq \frac{1}{2}\left(\frac{1}{m_1}-\frac{1}{m_2}\right)|h|. \tag{155}$$

Thus

$$x - x_1 = B(u+\alpha h) = Bu, \tag{156}$$

and hence

$$\|x-x_1\|_2 \leq \|u\|_2 \leq \frac{1}{2}\left(\frac{1}{m_1}-\frac{1}{m_2}\right)\|h\|_2. \tag{157}$$

Now we need an inequality of the form $\|h\|_2 \leq c\|x\|_2$. We have

$$h = Hf(x), \tag{158}$$

where $H = I - B$ is the high-pass operator. So, clearly

$$\|h\|_2 \leq \|f(x)\|_2 \leq m_2\|x\|_2.$$

We can improve this inequality by setting

$$v(t) = v = f(x) - \beta x, \tag{159}$$

where

$$\beta = 1/2(m_1+m_2).$$

Then

$$|v| \leqslant 1/2(m_2 - m_1)|x|, \tag{160}$$

and, since $Hx = 0$,

$$h = Hf(x) = H\{f(x) - \beta x\} = Hv. \tag{161}$$

Hence

$$\|h\|_2 \leqslant \|v\|_2 \leqslant 1/2(m_2 - m_1)\|x\|_2. \tag{162}$$

This, with (157) gives

$$\|x - x_1\|_2 \leqslant \frac{1}{4}\left(\frac{1}{m_1} - \frac{1}{m_2}\right)(m_2 - m_1)\|x\|_2 = \gamma\|x\|_2, \tag{163}$$

which is the result (147). The number γ in the inequality is invariant under the interchange of ϕ and f in the approximate identity (147), as we would expect from using only (*ii*) and (*iv*).

So $x_1 = B\phi\{Bf(x)\}$ is a good approximation to x if γ is small. The manipulations leading to the inequality (163) suggest an iterative scheme for solving, given y in $\mathscr{B}_2$,

$$BF(x) = y, \qquad x \quad \text{in} \quad \mathscr{B}_2, \tag{164}$$

provided $\gamma < 1$, which will be the case if $(m_2/m_1) < 3 + 2\sqrt{2}$.

We set

$$x_n = B\phi(y + h_{n-1}), \qquad n \geqslant 1, \tag{165}$$

where

$$\mathrm{h}_n = \mathrm{Hf}(\mathrm{x}_n), \qquad \mathrm{n} \geqslant 0, \tag{166}$$

and $\mathrm{x}_0 = \mathrm{h}_0 = 0,$

giving $\mathrm{x}_1 = \mathrm{B}\phi(\mathrm{y})$ as in (150).

Now we wish to show that

$$\|x - x_n\|_2 \leqslant \gamma^n\|x\|_2, \qquad n \geqslant 1. \tag{167}$$

We have

$$x - x_n = B\{\phi(y + h) - \phi(y + h_{n-1})\}. \tag{168}$$

Following the previous pattern we write

$$\phi(y + h) - \phi(y + h_{n-1}) = u_n + \alpha(h - h_{n-1}), \tag{169}$$

where

$$u_n = \int_{y+h_{n-1}}^{y+h} \{\phi'(\xi) - \alpha\}d\xi,$$

and hence

$$|u_n| \leq \frac{1}{2}\left(\frac{1}{m_1} - \frac{1}{m_2}\right)|h - h_{n-1}|. \tag{170}$$

Then

$$x - x_n = B\{u_n + \alpha(h - h_{n-1})\} = Bu_n, \tag{171}$$

giving

$$\|x - x_n\|_2 \leq \frac{1}{2}\left(\frac{1}{m_1} - \frac{1}{m_2}\right)\|h - h_{n-1}\|_2. \tag{172}$$

Now

$$h - h_{n-1} = H\{f(x) - f(x_{n-1})\}, \quad n \geq 1, \quad \text{with } x_0 = h_0 = 0. \tag{173}$$

Here we write

$$\begin{aligned} f(x) - f(x_{n-1}) &= \int_{x_{n-1}}^{x} \{f'(\xi) - \beta\}d\xi + \beta(x - x_{n-1}) \\ &= v_n + \beta(x - x_{n-1}), \end{aligned} \tag{174}$$

where

$$|v_n| \leq \frac{m_2 - m_1}{2}|x - x_{n-1}|. \tag{175}$$

Then

$$h - h_{n-1} = H\{v_n + \beta(x - x_{n-1})\} = Hv_n, \tag{176}$$

giving, with (175),

$$\|h - h_{n-1}\|_2 \leq \frac{m_2 - m_1}{2}\|x - x_{n-1}\|_2. \tag{177}$$

This, with (172), gives

$$\|x - x_n\|_2 \leq \gamma\|x - x_{n-1}\|_2, \tag{178}$$

whence follows, with $x_0 = 0$,

$$\|x - x_n\|_2 \leq \gamma^n\|x\|_2, \qquad n \geq 1. \tag{179}$$

Note that there is a bonus attached to $x_1 = B\phi(y)$, in that only one filtering operation is required to obtain it. Thereafter, two filtering operations are required to obtain x_n from y and x_{n-1}.

AUTHOR

Benjamin F. Logan, Jr., B.S. (Electrical Engineering), 1946, Texas Technological College; M.S., 1951, Massachusetts Institute of Technology;

Eng.D.Sc. (Electrical Engineering), 1965, Columbia University; Bell Laboratories, 1956—. While at MIT, Mr. Logan was a research assistant in the Research Laboratory of Electronics, investigating characteristics of high-power electrical discharge lamps. Also at MIT he engaged in analog computer development at the Dynamic Analysis and Control Laboratory. From 1955 to 1956 he worked for Hycon-Eastern, Inc., where he was concerned with the design of airborne power supplies. He joined Bell Laboratories as a member of the Visual and Acoustics Research Department, where he was concerned with the processing of speech signals. Currently, he is a member of the Mathematical Research Department. Member, Sigma Xi, Tau Beta Pi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 10, December 1983
Printed in U.S.A.

Bandwidth-Conserving Independent Amplitude and Phase Modulation

By B. F. LOGAN, Jr.*

(Manuscript received March 26, 1982)

Given two baseband signals $f(t)$ and $g(t)$, suitably restricted in amplitude and bandlimited to $[\lambda, \mu]$ and $[-\mu, -\lambda]$, $0 < \lambda < \mu < \infty$, it is shown how to generate a carrier signal, $s(t) = A(t)\cos\{ct + \phi(t)\}$, bandlimited to $[c - \beta, c + \beta]$ and $[-(c + \beta), -(c - \beta)]$, where β need be only sightly larger than μ, and such that $f(t)$ and $g(t)$ may be recovered by bandlimiting $\log A(t)$ and $\phi(t)$, respectively. The restriction $\lambda > 0$, i.e., that the baseband signals be bandpass, is not essential but it is a practical constraint in approximating the required operations. Also a modification is given for conserving bandwidth in case the signals $f(t)$ and $g(t)$ are of disparate bandwidths.

I. INTRODUCTION

Double-sideband amplitude modulation is wasteful of bandwidth, but it offers the advantage of envelope detection (with full carrier). A simple way to utilize the same bandwidth in transmitting two independent signals, $f(t)$ and $g(t)$, is the so-called in-phase and quadrature modulation

$$S_1(t) = f(t)\cos ct - g(t)\sin ct,$$

where synchronous demodulation is required to recover f and g. A modification that allows f to be recovered (approximately) by envelope detection is

$$S_2(t) = \{1 + f(t)\}\cos ct - g(t)\sin ct.$$

* Bell Laboratories.

The envelope of $S_2(t)$ is

$$A_2(t) = \sqrt{\{1 + f(t)\}^2 + g^2(t)}.$$

Then if g is made small (compared to $\min\{1 + f(t)\}$), we have

$$A_2(t) \cong 1 + f(t).$$

The phase of S_2 (i.e., the part due to signals) is

$$\phi_2(t) = \tan^{-1} \frac{g(t)}{1 + f(t)} \cong \frac{g(t)}{1 + f(t)}.$$

So making g small allows envelope and phase detection to be used so as to approximately recover f and g (multiplying the phase output by the envelope output).

A still further modification is

$$S_3(t) = \{1 + x_1(t)\}\cos ct - \{1 + x_2(t)\}\sin ct,$$

where x_1 and x_2 are both small. The envelope of S_3 is

$$A_3(t) = \sqrt{(1 + x_1)^2 + (1 + x_2)^2} = \sqrt{2 + 2x_1 + 2x_2 + x_1^2 + x_2^2}$$
$$\cong \sqrt{2}\left(1 + \frac{x_1 + x_2}{2}\right).$$

The phase of S_3 is

$$\phi_3(t) = \tan^{-1} \frac{1 + x_2}{1 + x_1} \cong \frac{\pi}{4} + \frac{x_2 - x_1}{2}.$$

So if

$$\frac{x_1 + x_2}{2} = f \quad \text{and} \quad \frac{x_1 - x_2}{2} = g,$$

i.e.,

$$x_1 = f + g$$
$$x_2 = f - g,$$

then envelope and phase detection of S_3 will give (approx.) the desired independent signals f and g.

An exact result of this type may be obtained using log of the envelope, rather than the envelope, and then *bandlimiting* the phase and log of the envelope to obtain the desired independent signals f and g. A slight increase in bandwidth is required to allow a guard band in the filtering operations. Also $|f|$ and $|g|$ cannot be too large if the increase in bandwidth is to be small. The basic theory is that of

Exponential Single-Sideband Modulation (ESSB) developed in Ref. 1.

II. THE EXACT METHOD

We assume that the desired signals, f and g, are bounded band-pass signals whose Fourier transforms vanish (in the sense detailed in Ref. 1) outside $[\lambda, \mu]$ and $[-\mu, -\lambda]$, $0 < \lambda < \mu < \infty$, which then (automatically) have bounded Hilbert transforms, $\hat{f}$ and $\hat{g}$. The band-pass assumption is not essential to the theory, but affords important practical simplifications in approximating the Hilbert transform operations as well as in effecting the subsequent analytic exponential modulation.

Now suppose $z_1(t)$ and $z_2(t)$ are bandlimited "analytic signals" whose Fourier transforms vanish outside $[0, \beta]$ and which signals are zero-free in the upper half-plane with

$$|z_{1,2}(t + iu)| \geq \epsilon \quad \text{for} \quad u \geq 0, \qquad -\infty < t < \infty. \tag{1}$$

Then $\log z_1$ and $\log z_2$ are analytic and bounded in the upper half-plane, and hence their Fourier transforms vanish over $(-\infty, 0)$.

Writing

$$z_1(t) = A_1(t)e^{i\phi_1(t)}, \qquad A_1 = |z_1| \tag{2}$$

$$z_2(t) = A_2(t)e^{i\phi_2(t)}, \qquad A_2 = |z_2|, \tag{3}$$

we have

$$\log z_1(t) = \log A_1(t) + i\phi_1(t) \tag{4}$$

$$\log z_2(t) = \log A_2(t) + i\phi_2(t). \tag{5}$$

Under further assumptions on z, e.g.,

$$z_{1,2}(t + iu) = 1 + 0(e^{-\lambda u}), \qquad u \to \infty, \tag{6}$$

$\log A$ and ϕ will be Hilbert transform pairs:

$$\phi_1(t) = \log\hat{} \; A_1(t), \qquad \log A_1(t) = -\hat{\phi}_1(t) \tag{7}$$

$$\phi_2(t) = \log\hat{} \; A_2(t) \qquad \log A_2(t) = -\hat{\phi}_2(t). \tag{8}$$

Now we consider the product

$$z_1(t)\overline{z_2(t)} = A_1(t)A_2(t)e^{i[\phi_1(t)-\phi_2(t)]},$$

where the bar denotes the complex conjugate. The F.T. (Fourier transform) of $z_1(t)$ vanishes outside $[0, \beta]$ and the F.T. of $\overline{z_2(t)}$ vanishes outside $[-\beta, 0]$. Therefore, the F.T. of $z_1(t)\overline{z_2(t)}$ vanishes outside $[-\beta, \beta]$. Then we form the signal

$$s(t) = \text{Re } e^{ict} z_1(t)\overline{z_2(t)}$$

$$= A(t)\cos\{ct + \phi(t)\}, \tag{9}$$

where $c > \beta$,

$$A(t) = A_1(t)A_2(t) \tag{9a}$$

$$\phi(t) = \phi_1(t) - \phi_2(t), \tag{9b}$$

and the spectrum of $s(t)$ is confined to $[c - \beta, c + \beta]$ and $[-c - \beta, -c + \beta]$.

We require

$$\mathbf{B}_{\mu,\alpha}\{\log A(t)\} = f(t) \tag{10}$$

$$\mathbf{B}_{\mu,\alpha}\{\phi(t)\} = g(t), \tag{11}$$

where $\mu < \alpha < \beta$, and, in general, $\mathbf{B}_{p,q}$ is any bandlimiting operator [with passband $(-p, p)$ and cut-off frequency $\pm q$] defined by

$$\mathbf{B}_{p,q}\{x(t)\} = \int_{-\infty}^{\infty} x(s)K_{p,q}(t - s)ds \tag{12}$$

and

$$\overset{>}{K}_{p,q}(\omega) = \int_{-\infty}^{\infty} K_{p,q}(t)e^{-i\omega t}dt = 1, \qquad -p < \omega < p$$

$$= 0, \qquad |\omega| > q. \tag{12a}$$

$$0 < p < q < \infty. \tag{12b}$$

The definition of $\overset{>}{K}_{p,q}(\omega)$ in the cut-off region (p, q) and $(-q, -p)$ is not important, but $\overset{>}{K}_{p,q}(\omega)$ must be sufficiently smooth to give

$$\int_{-\infty}^{\infty} |K_{p,q}(t)|dt < \infty \tag{12c}$$

so that $\mathbf{B}_{p,q}\{x(t)\}$ is defined for any bounded $x(t)$.

Writing (10) as

$$\mathbf{B}_{\mu,\alpha}\{\log A_1(t) + \log A_2(t)\} = f(t)$$

and taking Hilbert transforms of both sides of (10) and (11), using (7) and (8), we have

$$\mathbf{B}_{\mu,\alpha}\{\phi_1(t) + \phi_2(t)\} = \hat{f}(t) \tag{13}$$

$$\mathbf{B}_{\mu,\alpha}\{\phi_1(t) - \phi_2(t)\} = g(t) \tag{14}$$

or

$$\mathbf{B}_{\mu,\alpha}\{\phi_1(t)\} = \tfrac{1}{2}\{\hat{f}(t) + g(t)\} \tag{15}$$

$$\mathbf{B}_{\mu,\alpha}\{\phi_2(t)\} = \tfrac{1}{2}\{\hat{f}(t) - g(t)\}, \tag{16}$$

which according to (7) and (8) is equivalent to

$$\mathbf{B}_{\mu,\alpha}\{\log A_1(t)\} = \tfrac{1}{2}\{f(t) - \hat{g}(t)\} \tag{17}$$

$$\mathbf{B}_{\mu,\alpha}\{\log A_2(t)\} = \tfrac{1}{2}\{f(t) + \hat{g}(t)\}. \tag{18}$$

Setting

$$h_1(t) = \tfrac{1}{2}\{f(t) - \hat{g}(t)\}; \qquad \hat{h}_1(t) = \tfrac{1}{2}\{\hat{f}(t) + g(t)\}, \tag{19}$$

$$H_1(t) = h_1(t) + i\hat{h}_1(t), \tag{19a}$$

$$h_2(t) = \tfrac{1}{2}\{f(t) + \hat{g}(t)\}; \qquad \hat{h}_2(t) = \tfrac{1}{2}\{\hat{f}(t) - g(t)\}, \tag{20}$$

$$H_2(t) = h_2(t) + i\hat{h}_2(t), \tag{20a}$$

the four equations (15), (16), (17), and (18) are equivalent to the two equations, implying (6),

$$\mathbf{B}_{\mu,\alpha}\{\log z_1(t)\} = H_1(t) \tag{21}$$

$$\mathbf{B}_{\mu,\alpha}\{\log z_2(t)\} = H_2(t), \tag{22}$$

where H_1 and H_2 are given "analytic" band-pass signals whose Fourier transforms vanish outside the single interval $[\lambda, \mu]$ and z_1 and z_2 are bandlimited "analytic" signals whose Fourier transforms vanish outside the single interval $[0, \beta]$. The problem of finding z_1 and z_2 has been solved (see Ref. 1):

$$z_1(t) = \mathbf{B}_{\alpha,\beta}\{\exp H_1(t)\} \tag{23}$$

$$z_2(t) = \mathbf{B}_{\alpha,\beta}\{\exp H_2(t)\}, \tag{24}$$

where $\mathbf{B}_{\alpha,\beta}$ is any bandlimiting operator with passband $(-\alpha, \alpha)$ and cut-off frequency $\pm\beta$.

Now z_1 and z_2 given by (23) and (24) satisfy (21) and (22), *provided* $z_1(\tau)$ and $z_2(\tau)$, $\tau = t + iu$, are zero free in the upper half-plane $u > 0$. The filter characteristic $\overset{>}{K}_{\alpha,\beta}(\omega)$ in the cut-off region (α, β) becomes important, but not critical, in this respect. From theoretical considerations the linear cut-off characteristic is desirable (see Ref. 1):

$$\overset{>}{K}_{\alpha,\beta}(\omega) = \frac{\beta - \omega}{\beta - \alpha}, \qquad \alpha < \omega < \beta. \tag{25}$$

For a given α and β, and a smooth cut-off characteristic, z_1 and z_2 will be zero free in the upper half-plane provided $|H_1|$ and $|H_2|$ are not too large. In practice this means that the levels of f and g must not be

too large if α and β are not much larger than μ, the top signal frequency, i.e., in the bandwidth conserving case. The results in Ref. 1 may be used as a rough guide. For example, if α is only slightly larger than μ and $\beta/\alpha = 1.1$ (relatively sharp cut-off), then z_1 and z_2 will be zero free in the upper half-plane if $|H_1|$ and $|H_2|$ are less than 0.6. (See the appendix for a modification of signals f and g of disparate bandwidths.)

III. IMPLEMENTATION

The block diagram of an implementation is shown in Fig. 1. The transmitter is shown in Fig. 1a. The inputs are labeled $f(t + T)$ and $g(t + T)$ to account for a delay T incurred in the Hilbert transform filters. The delay T need not be more than one or two periods of the lower signal frequency λ to obtain a good approximation to the Hilbert transforms, $\hat{f}(t)$ and $\hat{g}(t)$. (The inputs $f(t + T)$ and $g(t + T)$ must be delayed accordingly to obtain $f(t)$ and $g(t)$.) The signals $f(t)$, $\hat{f}(t)$, $g(t)$, and $\hat{g}(t)$ are summed to obtain

$$h_1 = \tfrac{1}{2}(f - \hat{g})$$

$$\hat{h}_1 = \tfrac{1}{2}(\hat{f} + g)$$

$$h_2 = \tfrac{1}{2}(f + \hat{g})$$

$$\hat{h}_2 = \tfrac{1}{2}(\hat{f} - g)$$

in accord with (19) and (20). (The gain factor of the summing networks, shown as 1/2, may be any constant, which may be simply reflected as a gain factor on the inputs.) Then these outputs are fed to two analytic exponential modulators that furnish outputs

$$X_1 = e^{h_1}\cos \hat{h}_1 = \mathrm{Re}\{\exp H_1\}$$

$$Y_1 = e^{h_1}\sin \hat{h}_1 = \mathrm{Im}\{\exp H_1\}$$

$$X_2 = e^{h_2}\cos \hat{h}_2 = \mathrm{Re}\{\exp H_2\}$$

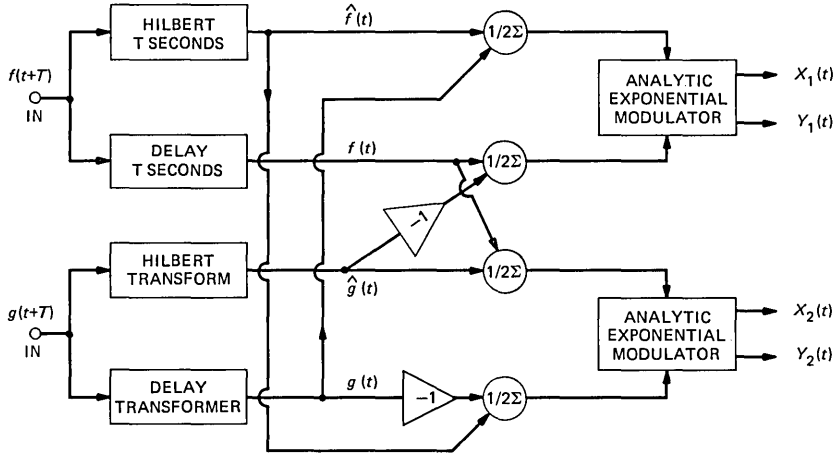
$$Y_2 = e^{h_2}\sin \hat{h}_2 = \mathrm{Im}\{\exp H_2\}.$$

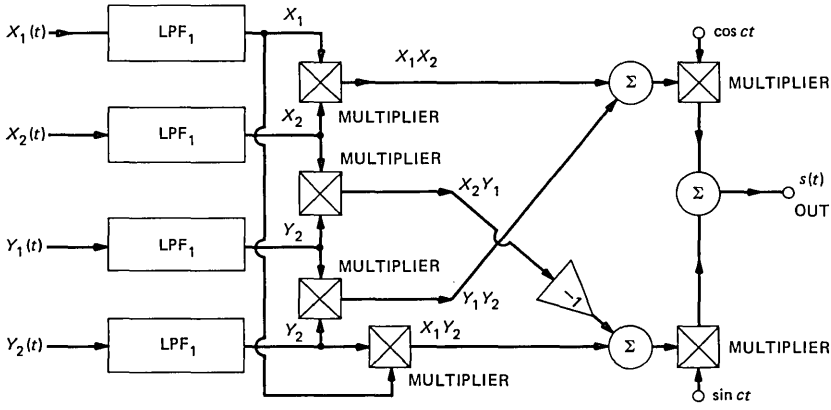
A feedback circuit for accomplishing the analytic exponential modulation is described in Ref. 2. The outputs of the modulators are then bandlimited with identical low-pass filters LPF_1 having the characteristic shown in Fig. 2a to obtain

$$x_1 = \mathrm{Re}\{z_1\}, \qquad y_1 = \mathrm{Im}\{z_1\}$$

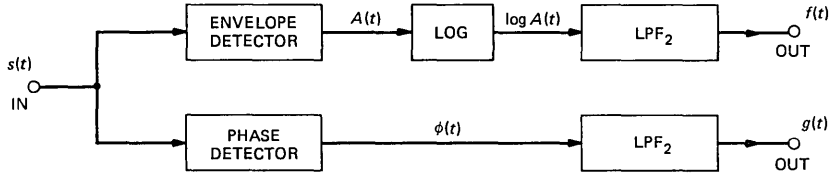
$$x_2 = \mathrm{Re}\{z_2\}, \qquad y_2 = \mathrm{Im}\{z_2\}.$$

These outputs are then combined to form





(a)



(b)

Fig. 1a—Transmitter.
Fig. 1b—Receiver.

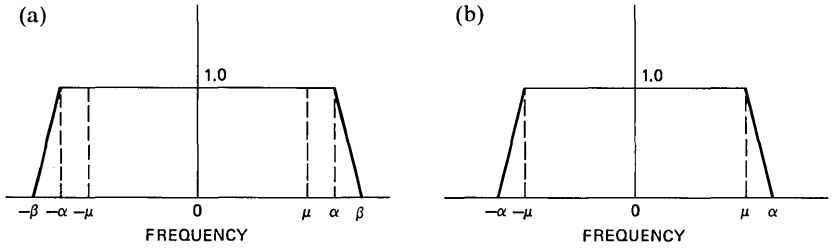


Fig. 2a—Characteristic of LPF_1.
Fig. 2b—Characteristic of LPF_2.

$$\begin{aligned} s(t) &= \mathrm{Re}\{[x_1(t) + iy_1(t)][x_2(t) - iy_2(t)]e^{ict}\} \\ &= \mathrm{Re}\{[x_1x_2 + y_1y_2 + i(y_1x_2 - y_2x_1)]e^{ict}\} \\ &= (x_1x_2 + y_1y_2)\cos ct - (y_1x_2 - y_2x_1)\sin ct \\ &= A(t)\cos\{ct + \phi(t)\}. \end{aligned}$$

The signal $s(t)$ is then transmitted to the receiver, Fig. 1b, where an envelope detector is used to obtain the envelope $A(t)$, which is then fed to a device having a logarithmic characteristic to furnish the output $\log A(t)$. This output is then filtered with LPF_2 to obtain $f(t)$. A phase detector, e.g., a phase-locked loop, is used to detect the phase $\phi(t)$, which is subsequently filtered with another LPF_2 to obtain $g(t)$. The characteristic of the filters LPF_2 is shown in Fig. 2b.

Note that $\phi(t)$ is high pass with lower frequency λ; so $\phi(t)$ may be recovered from $\{\phi'(t) + c\}$, if desired.

REFERENCES

1. B. F. Logan, "Theory of Analytic Modulation Systems," B.S.T.J., *57*, No. 3 (March 1978), pp. 491–576.
2. B. F. Logan, "Click Modulation," unpublished work.

APPENDIX

Modification for Signals of Disparate Bandwidths

Note that the bandwidth of the transmitted signal is the sum (or twice the sum, counting positive and negative frequencies) of the bandwidths of the analytic signals $z_1(t)$ and $z_2(t)$, which need be only slightly larger than the sum of the bandwidths of the analytic signals $H_1(t)$ and $H_2(t)$. Owing to the linear combinations in (19) and (20), the bandwidths of $H_1(t)$ and $H_2(t)$ will be the same, equal to the larger of the bandwidths of $f(t)$ and $g(t)$. In case the bandwidth of, say, $g(t)$ is (considerably) larger than that of $f(t)$, the bandwidth of the transmitted signal may be reduced by setting

$$H_1(t) = f(t) + i\hat{f}(t) \tag{26}$$

$$H_2(t) = g(t) + i\hat{g}(t). \tag{27}$$

Here we assume that the Fourier transforms of $H_1(t)$ and $H_2(t)$ vanish outside the intervals $[\lambda, \mu_1]$ and $[\lambda_2, \mu_2]$, respectively. Now we set

$$z_1(t) = \mathbf{B}_{\alpha_1,\beta_1}\{\exp H_1(t)\}, \qquad \mu_1 < \alpha_1 < \beta_1 \tag{28}$$

$$z_2(t) = \mathbf{B}_{\alpha_2,\beta_2}\{\exp H_2(t)\}, \qquad \mu_2 < \alpha_2 < \beta_2, \tag{29}$$

where β_1 and β_2 need be only slightly larger than μ_1 and μ_2 (respectively), the top frequencies of $f(t)$ and $g(t)$ (respectively). The Fourier transform of $z_1(t)\overline{z_2(t)}$ now vanishes outside the interval $[-\beta_2, \beta_1]$, which is smaller than would obtain in the previous scheme. Thus the Fourier transform of the transmitted signal,

$$s(t) = \text{Re } e^{ict} z_1(t)\overline{z_2(t)}, \qquad c > \beta_2 \tag{30}$$

vanishes outside the interval $[c - \beta_2, c + \beta_1]$ (and its reflection about the origin). The price paid for the saving in bandwidth is another Hilbert transform operation required in separating the signals at the receiver.

We have

$$s(t) = A(t)\cos[ct + \phi(t)], \tag{31}$$

where

$$A(t) = |z_1(t)z_2(t)|$$

$$\phi(t) = \phi_1(t) - \phi_2(t).$$

Then, assuming as before that z_1 and z_2 are zero free in the upper half-plane, we have

$$L(t) = \log A(t) = L_1(t) + L_2(t), \tag{32}$$

where $L_1(t) = \log |z_1(t)|$, $L_2(t) = \log |z_2(t)|$ and $L(t)$ is related to $\phi(t)$ by

$$\phi(t) = \phi_1(t) - \phi_2(t) = \hat{L}_1(t) - \hat{L}_2(t) \tag{33}$$

$$\hat{\phi}(t) = \hat{\phi}_1(t) - \hat{\phi}_2(t) = -L_1(t) + L_2(t). \tag{34}$$

In accord with (28) and (29) and the zero-free hypothesis, we have (as shown in Ref. 1)

$$\mathbf{B}_{\mu_1,\alpha_1}\{L_1(t)\} = f(t) \tag{35}$$

$$\mathbf{B}_{\mu_2,\alpha_2}\{L_2(t)\} = g(t) \tag{36}$$

To obtain $L_1(t)$ and $L_2(t)$ from $\log A(t)$ and $\phi(t)$, we need the Hilbert transform $\hat{\phi}(t)$, where according to (32) and (34),

$$L_1(t) = \tfrac{1}{2} \log A(t) - \tfrac{1}{2}\hat{\phi}(t) \tag{37}$$

$$L_2(t) = \tfrac{1}{2} \log A(t) + \tfrac{1}{2}\hat{\phi}(t). \tag{38}$$

However, to recover $f(t)$ and $g(t)$, we may use a modified version of $\hat{\phi}(t)$. We define

$$\hat{\phi}_{\lambda,\mu}(t) = \mathbf{H}_{\lambda,\mu}\{\phi(t)\}, \tag{39}$$

where $\mathbf{H}_{\lambda,\mu}$ is a modified (e.g., band-pass) Hilbert transform operator defined by

$$\mathbf{H}_{\lambda,\mu}\{\phi(t)\} = \int_{-\infty}^{\infty} h_{\lambda,\mu}(t - x)\phi(x)dx \tag{40a}$$

$$h_{\lambda,\mu}(\omega) = \int_{-\infty}^{\infty} h_{\lambda,\mu}(t)e^{i\omega t}dt = -i \operatorname{sgn} \omega, \tag{40b}$$

$$\text{for} \quad 0 < \lambda \leqslant |\omega| \leqslant \mu.$$

Now $\phi_1(t)$ and $\phi_2(t)$ are high-pass functions with lower frequencies λ_1 and λ_2 (respectively). Thus, if we require

$$0 < \lambda \leqslant \min(\lambda_1, \lambda_2), \qquad \mu \geqslant \max(\mu_1, \mu_2), \tag{41}$$

then we have

$$f(t) = \tfrac{1}{2}\mathbf{B}_{\mu_1,\alpha_1}\{\log A(t) - \hat{\phi}_{\lambda,\mu}(t)\} \tag{42}$$

$$g(t) = \tfrac{1}{2}\mathbf{B}_{\mu_2,\alpha_2}\{\log A(t) + \hat{\phi}_{\lambda,\mu}(t)\}. \tag{43}$$

AUTHOR

Benjamin F. Logan, Jr., B.S. (Electrical Engineering), 1946, Texas Technological College; M.S., 1951, Massachusetts Institute of Technology; Eng.D.Sc. (Electrical Engineering), 1965, Columbia University; Bell Laboratories, 1956—. While at MIT, Mr. Logan was a research assistant in the Research Laboratory of Electronics, investigating characteristics of high-power electrical discharge lamps. Also at MIT he engaged in analog computer development at the Dynamic Analysis and Control Laboratory. From 1955 to 1956 he worked for Hycon-Eastern, Inc., where he was concerned with the design of airborne power supplies. He joined Bell Laboratories as a member of the Visual and Acoustics Research Department, where he was concerned with the processing of speech signals. Currently, he is a member of the Mathematical Research Department. Member, Sigma Xi, Tau Beta Pi.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Beeri C., Fagin R., Maier D., Yannakakis M., **On the Desirability of Acyclic Database Schemes.** J ACM 30(3):479–513, 1983.

Cargill T. A., **A Robust Distributed Solution to the Dining Philosophers Problem—Response (Letter)** Software 13(6):549, 1983.

Coffman, E. G., **An Introduction to Combinatorial Models of Dynamic Storage-Allocation.** SIAM Rev 25(3):311–325, 1983.

Coffman, E. G., Sethi R., **Instruction Sets for Evaluating Arthimetic Expressions.** J ACM 30(3):457–478, 1983.

Du D. Z., Hwang F. K., **A New Bound for the Steiner Ratio.** T Am Math S 278(1):137–148, 1983.

Du D. Z., Hwang F. K., Weng J. F., **Steiner Minimal-Trees on Zig-Zag Lines.** T Am Math S 278(1):149–156, 1983.

Gehani N. H., **An Electronic Form System—An Experience in Prototyping.** Software 13(6):479–486, 1983.

Heck W. S., **Large Families of Incomparable *A*-Isols.** J Symb Log 48(2):250–252, 1983.

Hwang F. K., Weng J. F., Du D. Z., **A Class of Full Steiner Minimal-Trees.** Discr Math 45(1):107–112, 1983.

Lagarias J. C., **Sets of Primes Determined by Systems of Polynomial Congruences.** Ill J Math 27(2):224–239, 1983.

Liebesma B. S., Saperstein B., **A Proposed Attribute Skip-Lot Sampling Program (Review).** J Qual Tech 15(3):130–140, 1983.

Matula R. A., **The Need for an Online Compilation of Compilations.** Online 7(4):61–63, 1983.

Murphy B. T., **Microcomputers—Trends, Technologies, and Design Strategies.** IEEE J Soli 18(3):236–244, 1983.

Ramsey H. R., Atwood M. E., Vandoren J. R., **Flowcharts Versus Program Design Languages—An Experimental Comparison.** Comm ACM 26(6):445–449, 1983.

Raoult J. C., Sethi R., **Properties of a Notation for Combining Functions.** J ACM 30(3):595–611, 1983.

Sleator D. D., Tarjan R. E., **A Data Structure For Dynamic Trees.** J Comput Syst 26(3):362–391, 1983.

Slepian D., **Some Comments on Fourier-Analysis, Uncertainty and Modeling.** Siam Rev 25(3):379–393, 1983.

Tarjan R. E., **Citation Classic—Depth-1st Search and Linear Graph Algorithms.** CC/Eng Tech 1983(30):20, 1983.

Venkataraman K. N., Yasuhara A., Hawrusik F. M., **A View of Computability on Term Algebras.** J Comput Sy 26(3):410–471, 1983.

ENGINEERING

Agrawal G. P., Joyce W. B., Dixon R. W., Lax M., **Beam-Propagation Analysis of Stripe-Geometry Semiconductor-Lasers—Threshold Behavior.** Appl Phys L 43(1):11–13, 1983.

Anthony P. J., Pawlik J. R., Swaminathan V., Tsang W. T., **Reduced Threshold Current Temperature-Dependence in Double Heterostructure Lasers Due to Separate P-N and Heterojunctions.** IEEE J Q El 19(6):1030–1035, 1983.

Antler M., **Effect of Lubricants on Frictional Polymerization of Palladium Electrical Contacts.** ASLE Trans 26(3):376–380, 1983.

Auston D. H., Freeman R. R., Smith P. R., Mills D. M., Siemann R. H., **High-Speed X-Ray Sensitive Photoconducting Detector.** Appl Phys L 42(12):1050–1052, 1983.

Beni G., Hackwood S., **Robotic Electroplating of Gold on Quaternary Semiconductors.** J Appl Elec 13(4):531–533, 1983.

Bokor J., Eichner L., Storz R. H., Bucksbaum P. H., Freeman R. R., **Wavelength Conversion With Excimer Lasers.** AIP Conf PR1983(100):143–150, 1983.

Bowers J. E., Coldren L. A., Hemenway B. R., Miller B. I., Martin R. J., **1.55 μm Multisection Ridge Lasers.** Electr Lett 19(14):523–525, 1983.

Campbell J. C., Dentai A. G., Qua G. J., Ferguson J. F., **Avalanche InP/InGaAs Heterojunction Photo-Transistor.** IEEE J Q El 19(6):1134–1138, 1983.

Capasso F., **Band-Gap Engineering Via Graded Gap, Superlattice, and Periodic Doping Structures: Applications to Novel Photodetectors and Other Devices.** J Vac Sci B 1(2):457–461, 1983.

Chen C. Y., Cho A. Y., Bethea C. G., Garbinski P. A., Pang Y. M., Levine B. F., Ogawa K., **Ultrahigh Speed Modulation-Doped Heterostructure Field-Effect Photodetectors.** Appl Phys L 42(12):1040–1042, 1983.

Cheng J., Thurston R. N., Boyd G. D., Meyer R. B., **A New Low-Voltage Electrically-Addressed Bistable Nematic Liquid-Crystal Boundary-Layer Display.** IEEE Device 30(5):520–525, 1983.

Coldren L. A., Ebeling K. J., Miller B. I., Rentschler J. A., **Single Longitudinal Mode-Operation of Two-Section GaLnAsP/InP Lasers Under Pulsed Excitation.** IEEE J Q El 19(6):1057–1062, 1983.

Cox D. C., **Antenna Diversity Performance in Mitigating the Effects of Portable Radiotelephone Orientation and Multipath Propagation.** IEEE Commun 31(5):620–628, 1983.

Dvorak C. A., **Text Sharpness, Its Components and Text Quality.** J Appl Phot 9(3):109–111, 1983.

Dvorak C. A., Hamerly J. R. **Just-Noticeable Differences for Text Quality Components.** J Appl Phot 9(3):97–100, 1983.

Finegan S. N., Swartz R. G., McFee J. H., **A UHV-Compatible Round Wafer Heater for Silicon Molecular-Beam Epitaxy.** J Vac Sci B 1(2):497–500, 1983.

Geyling F. T., Walker K. L., Csencsits R., **The Viscous Collapse of Thick-Walled Tubes.** J Appl Mech 50(2):303–310, 1983.

Gilbert J. A. et al., **The Monomode Fiber—A New Tool for Holographic-Interferometry.** Exp Mech 23(2):190–195, 1983.

Henry C. H., Levine B. F., Logan R. A., Bethea C. G., **Minority-Carrier Lifetime and Luminescence Efficiency of 1.3-μm InGaAsP-InP Double Heterostructure Layers.** IEEE J Q El 19(6):905–912, 1983.

Henry C. H., Logan R. A., Merritt F. R., Luongo J. P., **The Effect of Intervalence Band Absorption on the Thermal-Behavior of InGaAsP Lasers.** IEEE J Q El 19(6):947–952, 1983.

Henry C. H., Logan R. A., Temkin H., Merritt F. R., **Absorption, Emission, and Gain Spectra of 1.3-μm InGaAsP Quaternary Lasers.** IEEE J Q El 19(6):941–946, 1983.

Howard R. E., Liao P. F., Skocpol W. J., Jackel L. D., Craighead H. G., **Microfabrication as a Scientific Tool.** Science 221(4606):117–121, 1983.

Johnson L. F., Ingersoll K. A., Dalton J. V., **Planarizing of Phosphosilicate Glass-Films on Patterned Silicon Wafers.** J Vac Sci B 1(2):487–489, 1983.

Kmetz A. R., Penz P. A., **Joint Special Issue on Displays—Foreword (Editorial).** IEEE Device 30(5):429–430, 1983.

Meloni A., Lanzerotti L. J., Gregori G. P., **Induction of Currents in Long Submarine Cables by Natural Phenomena (Review).** Rev Geophys 21(4):795–803, 1983.

Miller B., Rosamilia J. M., **Hydrodynamically Modulated Rotating-Disk Electrode Analysis in Derivative Mode (Review).** Analyt Chem 55(8):1281–1285, 1983.

Miller D. A. B., **Dynamic Non-Linear Optics in Semiconductors—Physics and Applications.** Laser Focus 19(7):61+, 1983.

Nygren S. F., **A Non-Destructive Method for Predicting Laser-Emission Wavelength From Photocurrent Spectra of GaAlAs Double Heterostructure Wafers.** IEEE J Q El 19(6):898–905, 1983.

Phillips J. C., **Realization of a Zachariasen Glass.** Sol St Comm 47(3):203–206, 1983.

Phillips J. C., **Why Localized and Extended Impurity Band States Can Coexist and Be Separated.** Sol St Comm 47(3):191–193, 1983.

Pini R. et al., **Ultraviolet Stimulated Raman-Scattering in Multimode Silica Fibers Pumped with Excimer Lasers.** Appl Phys L 43(1):6–8, 1983.

Smith J. S., Chiu L. C., Margalit S., Yariv A., Cho A. Y., **A New Infrared Detector Using Electron-Emission.** J Vac Sci B 1(2):376–378, 1983.

Sorace R., **Trellis Coding for a Multiple-Access Channel (Letter).** IEEE Info T 29(4):606–611, 1983.

Spencer E. G., **Citation Classic—Dielectric Materials for Electrooptic, Elastooptic, and Ultrasonic Device Applications.** CC/Eng Tech1983(31):22, 1983.

Tomita A., **Cross Talk Caused by Stimulated Raman-Scattering In Single-Mode Wavelength-Division Multiplexed Systems.** Optics Lett 8(7):412–414, 1983.

Tsang W. T., Olsson N. A., Logan R. A., **Stable Single-Longitudinal-Mode Operation Under High-Speed Direct Modulation in Cleaved-Coupled-Cavity GaInAsP Semiconductor-Lasers.** Electr Lett 19(13):488–490, 1983.

Tsang W. T., Olsson N. A., Logan R. A., **Threshold-Wavelength and Threshold-Temperature Dependences of GaInAsP-InP Lasers with Frequency Selective Feedback Operating in the 1.3-μm and 1.5-μm Regions.** Appl Phys L 43(2):154–156, 1983.

Tucker R. S., Eisenstein G., Kaminow I. P., **10 GHz Active Mode-Locking of a 1.3-μm Ridge-Waveguide Laser in an Optical-Fibre Cavity.** Electr Lett 19(14):552–553, 1983.

Weschler C. J., Kelty S. P., Lingousky J. F., **The Effect of Building Fan Operation on Indoor-Outdoor Dust Relationships.** J Air Pollu 33(6):624–628, 1983.

White J. C., Henderson D., **Tuning and Saturation Behavior of the Anti-Stokes Raman Laser.** AIP Conf PR1983(100):121–127, 1983.

White J. C., Miller T. A., Heaven M., **Anti-Stokes Raman-Scattering of Excimer and Tunable UV Lasers in Chemically Pumped CO.** AIP Conf PR1983(100):195–199, 1983.

Wilson T., Boyd G. D., Thurston R. N., Cheng J., Storz F. G., Westerwick E. H., **A Matrix Addressable Bistable Nematic Liquid-Crystal Display with Electric-Field Writing and Thermal Erasure.** IEEE Device 30(5):513–520, 1983.

Wilson T. G., Whelan E. W., Rodriguez R., Dishman J. M., **DC-to-DC-Converter Power-Train Optimization for Maximum Efficiency.** IEEE Aer El 19(3):413–427, 1983.

MANAGEMENT/ECONOMICS

Gordon R. H., Slemrod J., **A General Equilibrium Simulation Study of Subsidies to Municipal Expenditures.** J Finance 38(2):585–594, 1983.

Heiser W. J., Meulman J., **Analyzing Rectangular Tables by Joint and Constrained Multidimensional-Scaling.** J Economet 22(1–2):139–167, 1983.

Hill S. et al., **Putting a Lid on Pension Costs (Editorial).** Inst Invest 17(6):87+, 1983.

PHYSICAL SCIENCES

Agrawal G. P., **Nonperturbative Analysis of Zeeman-Coherence Effects on Resonant Phase Conjugation.** Optics Lett 8(7):359–361, 1983.

Alavi K., Petroff P. M., Wagner W. R., Cho A. Y., **Substrate Rotation-Induced Compositional Oscillation in Molecular-Beam Epitaxy (MBE).** J Vac Sci B 1(2):146–148, 1983.

Anderson P. W. et al., **Theory of the Universal Degradation of Tc in High-Temperature Superconductors.** Phys Rev B 28(1):117–120, 1983.

Ballman A. A., Glass A. M., Nahory R. E., Brown H., **Double Doped Low Etch Pit Density InP With Reduced Optical-Absorption.** J Cryst Gr 62(1):198–202, 1983.

Bates F. S., Baker G. L., **Polyacetylene Single-Crystals (Letter).** Macromolec 16(6):1013–1015, 1983.

Becker G. E., Cardillo M. J., Serri J. A., Hamann D. R., **He-Ag {001} – *C* (2×2)Cl Attractive Potential From Resonance Scattering.** Phys Rev B 28(2):504–514, 1983.

Blasie J. K. et al., **The Location of Redox Centers in the Profile Structure of a Reconstituted Membrane Containing a Photosynthetic Reaction Center-Cytochrome-C Complex by Resonance X-Ray-Diffraction.** Bioc Biop A 723(3):350–357, 1983.

Brasen D., Karlicek R. F., Donnelly V. M., **TEM Observations of Laser-Induced Pt and Au Deposition on InP.** J Elchem So 130(7):1473–1475, 1983.

Brillson L. J., Shapira Y., Heller A., **InP Surface-States and Reduced Surface Recombination Velocity.** Appl Phys L 43(2): 174–176, 1983.

Brummell M. A., Nicholas R. J., Portal J. C., Cheng K. Y., Cho A. Y., **Two-Dimensional Magnetophonon Resonance 2. GaInAs-AlInAs Heterojunctions (Letter).** J Phys C 16(17):L579–L584, 1983.

Bucksbaum P. H., Bokor J., **Time Resolved Amorphous-Silicon Formation From Laser Melted Liquid Silicon Films.** AIP Conf PR1983(100):279–287, 1983.

Cardillo M. J., Becker G. E., Hamann D. R., Serri J. A., Whitman L., Mattheiss L. F., **Geometry of the Ag{001} – *C* (2×2)Cl Structure as Determined by He Diffraction.** Phys Rev B 28(2):494–503, 1983.

Cassanho A., Guggenheim H., Walstedt R. E., **Ionic-Conductivity and NMR Study of Fluorite-Structured $K_{0.4}Bi_{0.6}F_{2.2}$.** Phys Rev B 27(11):6587–6592, 1983.

Cheng A. F., Maclennan C. G., Lanzerotti L. J., Paonessa M. T., Armstrong T. P., **Energetic Ion Losses Near Io's Orbit.** J Geo R-S P 88(NA5):3936–3944, 1983.

Chin A. K., Zipfel C. L., Chin B. H., Digiuseppe M. A., **Degradation of 1.3-μm InP InGaAsP Light-Emitting Diodes With Misfit Dislocations.** Appl Phys L 42(12):1031–1033, 1983.

Chiu L. C., Smith J. S., Margalit S., Yariv A., Cho A. Y., **Application of Internal Photoemission From Quantum-Well and Heterojunction Super-Lattices to Infrared Photodetectors.** Infrar Phys 23(2): 93–97, 1983.

Cohen J. D., Lang D. V., Harbison J. P., Sergent A. M., **Photoinduced Changes in the Bulk-Density of Gap States in Hydrogenated Amorphous-Silicon Associated With the Staebler-Wronski Effect.** Solar Cells 9(1–2):119–131, 1983.

Dubois L. H., Somorjai G. A., **Why CO_2 Does Not Dissociate on RH At Low-Temperature—Comment (Letter).** Surf Sci 128(2–3):L231–L235, 1983.

Dutt B. V., Ludwig R. A., Ermanis F., **The Origin and Elimination of Si Pyramids in (Ga, Al) As-Si LEDS.** J Cryst Gr 62(1):21–26, 1983.

Englert T., Maan J. C., Uihlein C., Tsui D. C., Gossard A. C., **Cyclotron-Resonance of Two-D-Electrons in $GaAs/Al_xGa_{1-x}As$ Heterostructures at Low-Densities.** J Vac Sci B 1(2):427–430, 1983.

Fiory A. T., **Brownian-Dynamics Simulations of Melting of Finite Two-Dimensional Systems With Logarithmic Interaction.** Phys Rev B 28(1):236–243, 1983.

Gibbs H. M. et al., **Optical Bistability, Regenerative Pulsations, and Transverse Effects in Room-Temperature GaAs-AlGaAs Super-Lattice Etalons.** J Physique 44(NC-2):195–204, 1983.

Gill P. S., Graedel T. E., Weschler C. J., **Organic Films on Atmospheric Aerosol-Particles, Fog Droplets, Cloud Droplets, Raindrops, and Snowflakes (Review).** Rev Geophys 21(4):903–920, 1983.

Gilmer G. H. et al., **Simulation-Models of the Crystal Vapor Interface.** J Vac Sci B 1(2):298–304, 1983.

Greenside H. S., Schulter M., **Pseudopotentials for the Three-D Transition-Metal Elements.** Phys Rev B 28(2):535–543, 1983.

Greis N. P., **Flood Frequency-Analysis—A Review of 1979–1982.** Rev Geophys 21(3):699–706, 1983.

Griscom D. L., Friebele E. J., Long K. J., Fleming J. W., **Fundamental Defect Centers in Glass-Electron-Spin Resonance and Optical-Absorption Studies of Irradiated Phosphorus-Doped Silica Glass and Optical Fibers.** J Appl Phys 54(7):3743–3762, 1983.

Gurvitch M., **Experimental-Evidence Versus Exchange Theory of Resistivity Saturation.** Phys Rev B 28(2):544–549, 1983.

Haddon R. C., Hirani A. M., Kroloff N. J., Marshall J. H., **1,3,4,9-Tetramethoxyphenalenyl system.** J Org Chem 48(12):2115–2117, 1983.

Hilinski E. F., Rentzepis P. M., **Chemical Applications of Picosecond Spectroscopy (Review).** Acc Chem Re 16(6):224–232, 1983.

Jelinski L. W., Dumais J. J., Stark R. E., Ellis T. S., Karasz F. E., **Interaction of Epoxy-Resins With Water—A Quadrupole Echo Deuterium NMR-Study (Letter).** Macromolec 16(6):1019–1021, 1983.

Kinsbron E., Murarka S. P., Sheng T. T., Lynch W. T., **Oxidation of Arsenic Implanted Polycrystalline Silicon.** J Elchem So 130(7):1555–1560, 1983.

Klauder J. R., **A Langevin Approach to Fermion and Quantum Spin Correlation-Functions (Letter).** J Phys A 16(10):L317–L319, 1983.

Kleinman D. A., **Binding-Energy of Biexcitons and Bound Excitons in Quantum Wells.** Phys Rev B 28(2):871–879, 1983.

Klem J., Fischer R., Drummon T. J., Morkoc H., Cho A. Y., **Incorporation of Sb in $GaAs_{1-x}Sb_x$ (X-Less-Than-0.15) By Molecular-Beam Epitaxy.** Electr Lett 19(12):453–455, 1983.

Lake G., Norman C., **Stellar and Gaseous Dynamics of Triaxial Galaxies.** Astrophys J 270(1):51–70, 1983.

Lifshitz N., **Study of Breakdown Fields of Oxides Grown on Reactive Ion Etched Silicon Surface—Improvement of Breakdown Limits by Oxidation of the Surface.** J Elchem So 130(7):1549–1550, 1983.

Luongo J. P., **Infrared Characterization of Alpha-Crystalline and Beta-Crystalline Silicon-Nitride.** J Electrochem Soc 130(7):1560–1562, 1983.

Lyons A. M., Wilkins C. W., Robbins M., **Thin Pinhole-Free Carbon-Films.** Thin Sol Fi 103(4):333–341, 1983.

Malyj M., Griffiths J. E., **Stokes Anti-Stokes Raman Vibrational Temperatures—Reference Materials, Standard Lamps, and Spectrophotometric Calibrations.** Appl Spectr 37(4):315–333, 1983.

McWhan D. B., Gurvitch M., Rowell J. M., Walker L. R., **Structure and Coherence of NbAl Multilayer Films.** J Appl Phys 54(7):3886–3891, 1983.

Olego D., Pinczuk A., Gossard A. C., Wiegmann W., **Plasma-Oscillations of Layered Electron Gases in Semiconductor Heterostructures.** J Vac Sci B 1(2):412–414, 1983.

People R., Wecht K. W., Alavi K., Cho A. Y., **Measurement of The Conduction-Band Discontinuity of Molecular-Beam Epitaxial Grown $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$, N-N Heterojunction by C-V Profiling.** Appl Phys L 43(1):118–120, 1983.

Petroff P. M., Wilson R. J., **Surface Dislocation Process for Surface Reconstruction and Its Application to the Silicon (111) 7×7 Reconstruction.** Phys Rev L 51(3):199–202, 1983.

Phillips J. M., Feldman L. C., Gibson J. M., Mcdonald M. L., **Rutherford Backscattering Channeling and Transmission Electron-Microscopy Analysis of Epitaxial BaF_2 Films on Ge and InP.** J Vac Sci B 1(2):246–249, 1983.

Pian T. R. et al., **Electron and Photon Stimulated Desorption of Positive-Ions From Alkali-Halide Surfaces.** Surf Sci 128(1):13–21, 1983.

Schiavone J. A., **Microwave Radio Meteorology—Seasonal Fading Distributions.** Radio Sci 18(3):369–380, 1983.

Schneemeyer L. F., Cohen U., **Electrochemical Synthesis of Photoactive MoS_2.** J Elchem So 130(7):1536–1539, 1983.

Schubert R., Augis J. A., **Sub-PPMA Gas-Analysis in 40-μl Volumes by rf Mass Spectrometry.** J Vac Sci A 1(2):248–251, 1983.

Schwartz G. P., **Analysis of Native Oxide-Films and Oxide Substrate Reactions on III-V-Semiconductors Using Thermochemical Phase-Diagrams.** Thin Sol Fi 103(1–2):3–16, 1983.

Schwartz G. P., Bondybey V. E., English J. H., Gualtieri G. J., **Thermal and Pulsed Laser Evaporation of Single Phase As_xP_{1-x} Alloys.** Appl Phys L 42(11):952–954, 1983.

Schwartz G. P., Dutt B. V., Malyj M., Griffiths J. E., Gualtieri G. J., **Thermal-Oxidation and Anodic Film Substrate Reactions on $In_xGa_{1-x} - As_yP_{1-y}$.** J Vac Sci B 1(2):254–259, 1983.

Shaw E. D. et al., **Helical Undulatory Study.** J Physique 44(NC-1):153–161, 1983.

Silberg E., Chang T. Y., Caridi E. A., Evans C. A., Hitzman C. J., **Manganese and**

Germanium Redistribution in $In_{0.53}Ga_{0.47}As$ Grown by Molecular-Beam Epitaxy. J Vac Sci B 1(2):178–181, 1983.

Simmons J. G., Taylor G. W., **Generalized Theory of Conduction in Schottky Barriers.** Sol St Elec 26(7):705–709, 1983.

Stall R. A., **Growth of Refractory Oxide-Films Using Solid Oxygen Sources in a Molecular-Beam Epitaxy Apparatus.** J Vac Sci B 1(2):135–137, 1983.

Stavola M. et al., **Symmetry Determination For Deep States in Semiconductors From Stress-Induced Dichroism of Photocapacitance.** J Appl Phys 54(7):3897–3901, 1983.

Stormer H. L., Haavasoja T., Narayanamurti V., Gossard A. C., Wiegmann W., **Observation of the Dehaas-Vanalphen Effect in a Two-Dimensional Electron-System.** J Vac Sci B 1(2):423–426, 1983.

Stormer H. L., Schlesinger Z., Chang A., Tsui D. C., Gossard A. C., Wiegmann W., **Energy Structure and Quantized Hall-Effect of Two-Dimensional Holes.** Phys Rev L 51(2):126–129, 1983.

Subramaniam P., Zemanian A. H., **Analysis of $Delta_2Phi$-$c_2Phi = g$ in a Semi-Infinite Medium With an Irregular Boundary by Means of Network Manipulations.** IEEE Circ S 30(5):300–307, 1983.

Swaminatthan V., Wagner W. R., Anthony P. J., Henein G., Koszi L. A., **Bonding Pad Induced Stresses in (Al, Ga)As Double Heterostructure Lasers.** J Appl Phys 54(7):3763–3768, 1983.

Tennant D. M., **Metal on Polymer Ion-Implantation Mask.** J Vac Sci B 1(2):494–496, 1983.

Teo B. K., Antonio M. R., Averill B. A., **Molybdenum K-Edge Extended X-Ray Absorption Fine Structure (EXAFS) Studies of Synthetic Mo-Fe-S Clusters Containing the MoS_4 Unit: Development of a Fine Adjustment Technique Based on Models.** J Am Chem S 105(12):3751–3762, 1983.

Tompkins H. G., Allara D. L., Pasteur G. A., **Molecular-Orientation and Structure of Copper Benzimidazole Films.** Surf Int An 5(3):101–104, 1983.

Tsang W. T., Olsson N. A., **New Current Injection 1.5-μm Wavelength $Ga_xAl_yIn_{1-x-y}As/InP$ Double-Heterostructure Laser Grown by Molecular Beam Epitaxy.** Appl Phys L 42(11):922–924, 1983.

Tsang W. T., Olsson N. A., **Preparation of 1.78-μm Wavelength $Al_{0.2}Ga_{0.8}Sb$/GaSb Double-Heterostructure Lasers by Molecular-Beam Epitaxy.** Appl Phys L 43(1):8–10, 1983.

Tsang W. T., Olsson N. A., Logan R. A., **Transient Single-Longitudinal Mode Stabilization in Double Active Layer GaInAsP InP Laser Under High-Bit Rate Modulation.** Appl Phys L 42(12):1003–1005, 1983.

Varma C. M., Simons A. L., **Strong-Coupling Theory of Charge-Density-Wave Transitions.** Phys Rev L 51(2):138–141, 1983.

Verter F., Knapp G. R., Stark A. A., Wilson R. W., **Regions of Low-Molecular Column Density Near the Galactic Plane.** Astroph J S 52(3):289–292, 1983.

Weber T. A. et al., **Simulation of Polyethylene (Review).** Adv Chem SE1983(204):487–500, 1983.

Wilson L. O., Nelson T. J., **Magnetization Distributions in Ion-Implanted Bubble Garnet-Films.** J Appl Phys 54(7):4163–4167, 1983.

Wilson R. J., Mills A. P., **Electron and Positron Work-Functions of Cr(100).** Surf Sci 128(1):70–80, 1983.

Wilson T., Boyd G. D., Westerwick E. H., Storz F. G., **Alignment of Liquid-Crystals on Surfaces With Films Deposited Obliquely at Low and High-Rates.** Molec Cryst 94(3–4):359–366, 1983.

Wudl F. et al., **Tetramethyldithiadiselenafulvalene (TMDTDSF)—Properties of its Hexafluorophosphate Salt and Alloys With Tetramethyltetraselenafulvalene.** J Chem Phys 79(2):1004–1012, 1983.

SOCIAL AND LIFE SCIENCES

Hopfield J. J., Feinstein D. I., Palmer R. G., **Unlearning Has a Stabilizing Effect in Collective Memories.** Nature 304(5922):158–159, 1983.

Pierrehumbert J., Liberman M., **On Finding the Iguana (Letter).** Cont Psycho 28(7):569–570, 1983.

SPEECH AND ACOUSTICS

Atal B. S., **Speech Coding—Recognizing What We Do Not Hear in Speech.** Ann NY Acad 405(Jun):18–32, 1983

Pols L. C. W., Olive J. P., **Intelligibility of Consonants in CVC Utterances Produced by Dyadic Rule Synthesis.** Speech Comm 2(1):3–13, 1983.

CONTENTS, JANUARY 1984

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/ Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System