THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 56 November 1977 Number 9

The Influence of Rain on Design of 11-GHz Terrestrial Radio Relay

By D. C. HOGG, A. J. GIGER, A. C. LONGTON, and E. E. MULLER

(Manuscript received April 25, 1977)

Three salient factors governing attenuation of 11-GHz waves propagating through rain on a terrestrial radio path are discussed: magnitude of attenuation as a function of rain rate, relationship between attenuation and path length, and dependence of attenuation on polarization. Background material is given pertinent to the companion papers in this issue which develop procedures useful for radio system design.

I. INTRODUCTION

Terrestrial radio systems employing the 10.7–11.7 GHz common carrier band have been in use for many years. For example, analog TJ and TL radio are used for short-haul applications, and TL has served in cross-band diversity as protection for a 6-GHz system. However, new emphasis is being placed on autonomous 11-GHz systems, wideband digital implementations such as 3-ARDS[1] being attractive in many applications. It is therefore meaningful to re-examine the effects of rain on the propagation of 11-GHz signals; hop-length limitations imposed by rain have impact on the cost of service. Multipath fading, readily accommodated by antennas operated in space diversity, is not discussed. In the interest of designing reliable systems, our prime intent here and in the companion papers of this issue is to determine the limitation on hop lengths imposed by rain attenuation for given fading margin and annual outage time objectives for systems in the United States.

Three major factors are involved in this determination:

(*i*) The 11-GHz attenuation statistics must be properly associated with the rain rate statistics as determined by measurements at a point. There are two reasons for this: (a) It is only by virtue of long-term measurements of point rain rate, from sources such as the National Climatic Center, that a sufficient quantity of data can be obtained to provide reliable temporal statistics for calculation of the path attenuation. (b) It is only from measurements of point rain rates at numerous locations throughout the country, such as those compiled by the National Climatic Center, that the rain environment at arbitrary locations where radio systems may be installed can be suitably determined.

(*ii*) The spatial extent of rain associated with a given point rain rate must be accounted for; this is necessary, especially in the case of the intense showers which produce large attenuation, because the hop length may or may not exceed the dimension of the shower. This interdependence between storm dimensions and loss has the effect of producing a nonlinear relationship between attenuation and hop length, for a given point rain rate.

(*iii*) Dependence of rain attenuation on the polarization of 11-GHz transmission must be understood. Vertically polarized waves are attenuated less than those which are horizontally polarized because of the oblate shape of large raindrops; this results in different outage durations for vertical and horizontal polarization on a hop. In the design of a system involving many hops, a sequencing of polarization may therefore be desirable to equalize annual channel outage times.

The above three factors are dealt with in detail in companion papers;[2,3] here, the background material that forms the basis for these investigations is discussed.

II. MEASUREMENTS OF ATTENUATION BY RAIN

In 1956, a transmission experiment[4] was mounted at 11 GHz on colinear contiguous paths of 20 and 44 km at Mobile, Alabama. The measured attenuation caused by rain was compared with attenuation calculated[5] from the rain rate measured by 14 gauges along the 44-km path, and, although scatter in the data was large, agreement was fairly satisfactory. Therefore, by scaling annual distributions of 1-hour point rain rates of more than 25 mm/hr, obtained in other regions of the U.S., to the Mobile data, a set of contours defining constant hop length for a fixed outage time was developed. In 1965, an article[6] comparing theoretical calculations with rain attenuation measured at various microwave frequencies in several countries pointed out that serious deficiencies existed in the ability to predict path attenuation from point rain rates.

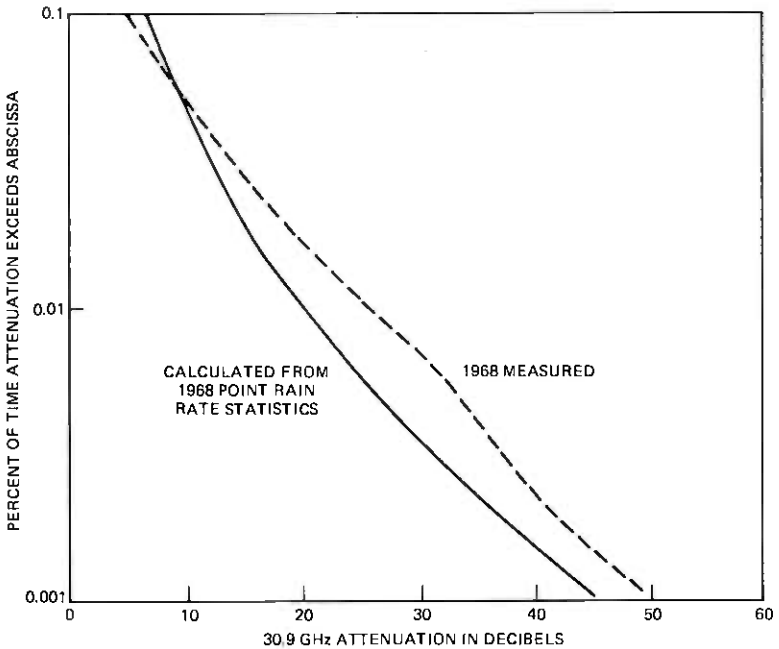


Fig. 1—Rain attenuation measured on a 1.9-km path in New Jersey compared with calculation based upon rain rate measured at a point on the path.

In 1967, an experiment involving 100 rain gauges arranged in a fine-grained network over a 410 km^2 area was performed at Holmdel, N.J.; these data[7] showed that the high microwave attenuation events[8] were produced by intense showers of limited size which resulted in nonlinear dependence of attenuation on path length, as noted in factor (*ii*) above. Simultaneous measurements[9] at 18 and 31 GHz on a 2.6 km path showed that the size distribution of the raindrops in heavy showers is adequately represented by the Laws-Parsons size distribution[10] which has been used for theoretical calculation[12] of attenuation by uniform rain on a path. Direct comparison of attenuation measurements on *short* paths with attenuation calculated from measured point rain rates is fairly satisfactory. In the example shown in Fig. 1, a cumulative distribution of 31 GHz attenuation measured[11] during 1968 on a 1.9-km path in New Jersey is compared with calculated results based upon point rain rates measured near the center of the path. In this early experiment, the predicted values of attenuation are somewhat less than the measured values. However, more favorable comparisons were obtained later in measurements on short paths,[13] and are discussed in a companion paper;[2] these are important in accounting for factor (*i*) of the introduction.

During that same series of experiments, measurements[14] at 31 GHz

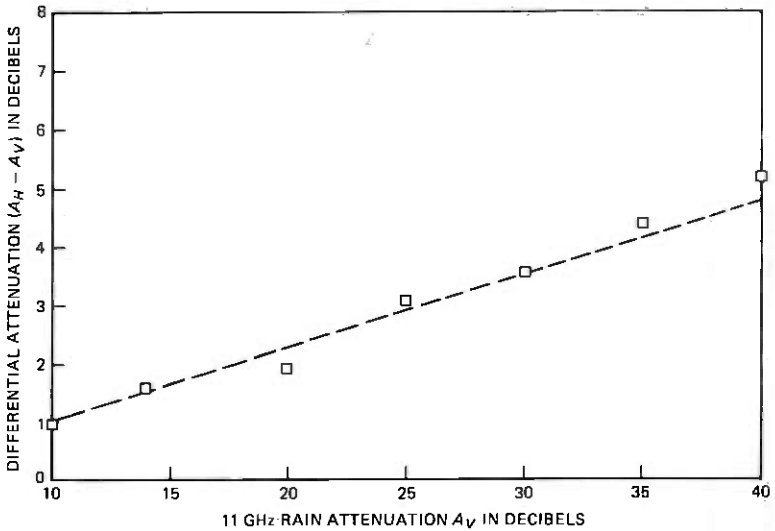


Fig. 2—Difference between attenuation of horizontally and vertically polarized waves measured at 11 GHz on a 23.4-km path in Pennsylvania during 1974. The attenuation on abscissa includes wet radome loss.

showed that horizontal polarization was attenuated about 15 percent (in decibels) more than vertical polarization. More recently, measurements of differential attenuation have been obtained at 11 GHz on a 23 km path at Harrisburg, Pennsylvania. As shown in Fig. 2, the difference amounts to about 12 percent; the effect has been investigated theoretically,[15,16] and good agreement with experimental data is found.[17] These results are used in companion papers, Refs. 2 and 3, to take factor (*iii*), above, into account.

In 1970, a substantial experiment[18] including 11 GHz propagation was started near Atlanta, Georgia. In particular, the 11-GHz rain attenuation was measured simultaneously on three hops of quite different length; cumulative distributions of path attenuation measured over long periods of time could then be compared directly. The comparison was made by choosing a (small) time interval of interest to system outage, 30 minutes per year for example, and noting the rain attenuations on the three distributions at this level of incidence; a certain measured point rain rate is also associated with the chosen time interval. These attenuations, when plotted versus hop length, are found to increase nonlinearly with distance, the slope of the curve decreasing with increasing hop length.[2] If such a plot were to reach an asymptote, attenuation then being constant with hop length, the interpretation would be straightforward, namely, that well-defined uniform storms with dimensions much smaller than the hop length were producing the attenuation. However, this asymptote is never reached because the probability of a shower intersecting a path

continues to increase as the path length is increased. In reality, rain storms are not well defined geometrically, nor is the rain rate uniform throughout, therefore calculations[19] involving spatial correlation of the rain have been made; comparison with the 11-GHz attenuations measured at Atlanta is described in Ref. 3.

III. RAIN ON ANTENNAS

In addition to attenuation on the path per se, rain on various parts of an antenna produces enough attenuation to be of concern. This problem is especially serious in unprotected paraboloids fed from the prime focus, in which case water can form on the aperture of the feed as a layer or as drops. Since water introduces both loss and phase shift[20] at 11 GHz, its presence at the feed aperture where the power density of the wave is high causes considerable degradation in antenna performance by way of loss, reflection, and distortion of the phase of the wavefronts. These antennas therefore are usually protected by weather covers which enclose the paraboloidal reflector as well as the feed. Thus the problem is relegated to transmission through a weather cover wetted by rain; similar considerations apply to horn-reflector antennas.

Measurements have been made at 20 GHz on transmission through the water layer produced by various rain storms on a radome;[21] attenuations from 4 to 8 dB were observed, depending on factors such as the rain rate, wind velocity, and the nature of the surface of the radome. At 11 GHz, these values scale to between 3 and 6 dB. Measurements at 11 GHz have been made on a short hop employing paraboloids with hemispherical weather covers. After the rain on the path was accounted for through calculation, each weather cover was found[22] to contribute at least 3 dB of attenuation during heavy rain. Horn-reflector antennas fortunately are equipped with flat weather covers pitched somewhat beyond vertical; recent measurements during rain storms using a pair of closely spaced horn reflectors indicate that an attenuation of 4 dB per hop should be allowed[23] in the fading margin of an 11-GHz system to account for water on the weather covers. Note that the attenuation caused by water on weather covers is considered part of the fading-margin allowance rather than part of the attenuation caused by rain on the path per se. This allowance is discussed further in Ref. 3.

IV. SUMMARY

If the factors discussed above are taken into account, the 11-GHz rain attenuation statistics on a hop of arbitrary length can be estimated from measured point rain rate statistics as described in the companion papers. In all, estimates of annual attenuation have been carried out for 226 locations in the United States by properly processing point rain rates obtained from the National Climatic Center. The occurrence of intense

rain is, of course, not the same from year to year at a given location; the resulting attenuation behavior will therefore also vary. It has been found necessary to use point rain rates measured over periods of 20 years or more to generate stable statistics. The deviation of data taken over 1- and 5-year periods (from the 20-year value) is discussed in Refs. 2 and 3. Reference 3 deals with application of the results to radio-path engineering, and an illustrative example of a tandem-hop situation is given. In calculating the performance of a system involving two or more hops, it is assumed that no two hops will simultaneously experience very deep fading as a result of rain attenuation. This assumption of lack of correlated outages to rain attenuation leads to slightly pessimistic estimates for the outage of a system.

REFERENCES

1. A. J. Giger and T. L. Osborne, "3A-RDS 11 GHz Digital Radio System," Digest of International Conference on Communications, paper No. 18.1, 1976.
2. S. H. Lin, "Nationwide Long-Term Rain Rate Statistics and Empirical Calculation of 11-GHz Microwave Rain Attenuation," B.S.T.J., this issue, pp. 1581–1604.
3. T. L. Osborne, "Application of Rain Attenuation Data for 11-GHz Radio Path Engineering," B.S.T.J., this issue, pp. 1605–1628.
4. S. D. Hathaway and H. W. Evans, "Rain Attenuation at 11 kmc," B.S.T.J., *38*, No. 1 (January 1959), pp. 73–97.
5. K. L. S. Gunn and T. W. R. East, "The Microwave Properties of Precipitation Particles," Quart. J. Roy. Meterol. Soc.," *80*, 1954, pp. 522–545.
6. R. G. Medhurst, "Rainfall Attenuation of Centimeter Waves," IEEE Trans. Ant. Propag., *AP-13*, July 1965, pp. 550–563.
7. A. E. Freeny and J. D. Gabbe, "A Statistical Description of Intense Rainfall," B.S.T.J., *48*, No. 6 (July 1969), pp. 1789–1851.
8. D. C. Hogg, "Statistics on Attenuation of Microwaves by Intense Rain," B.S.T.J., *48*, No. 9 (November 1969), pp. 2949–2962.
9. R. A. Semplak, "Dual Frequency Measurements of Rain-Induced Microwave Attenuation on a 2.6 Kilometer Propagation Path," B.S.T.J., *50*, No. 8 (October 1971), pp. 2599–2606.
10. J. O. Laws and D. A. Parsons, "The Reduction of Rain Drop Size to Intensity," Trans. Am. Geophys. Union, *25*, 1943, pp. 452–460.
11. D. E. Setzer, "Computed Transmission through Rain at Microwave and Variable Frequencies," B.S.T.J., *49*, No. 8 (October 1970), pp. 1873–1892.
12. R. A. Semplak, "The Influence of Heavy Rain on Attenuation at 18.5 and 30.9 GHz," IEEE Trans. Ant. Propag., *AP-18*, July 1970, pp. 507–511.
13. W. F. Bodtmann and C. L. Ruthroff, "Rain Attenuation on Short Radio Paths: Theory, Experiment and Design," B.S.T.J., *53*, No. 7 (September 1974), pp. 1329–1350.
14. R. A. Semplak, "Effect of Oblate Raindrops on Attenuation at 30.9 GHz," Radio Sci., *5*, March 1970, pp. 559–564.
15. T. Oguchi, "Attenuation of Radio Waves due to Rain with Distorted Drops," J. Radio Res. Lab. (Tokyo), *7*, September 1960, pp. 467–485.
16. J. A. Morrison, M. J. Cross, and T. S. Chu, "Rain-Induced Differential Attenuation and Differential Phase Shift at Microwave Frequencies," B.S.T.J., *52*, No. 4 (April 1973), pp. 599–604.
17. T. S. Chu, "Rain-Induced Cross Polarization at Centimeter and Millimeter Wavelengths," B.S.T.J., *53*, No. 8 (October 1974), pp. 1557–1579.
18. W. T. Barnett and S. H. Lin, to be published.
19. S. H. Lin, "A Method for Calculating Rain Attenuation Distributions on Microwave Paths," B.S.T.J., *54*, No. 6 (July–August 1975), pp. 1051–1086.
20. D. C. Hogg and T. S. Chu, "The Role of Rain in Satellite Communications," Proc. IEEE, *63*, No. 9 (September 1975), pp. 1308–1331.
21. I. Anderson, "Measurement of 20 GHz Transmission through a Wet Radome," IEEE Trans. Ant. Propag., *AP-23*, September 1975, pp. 619–622.
22. S. H. Lin, personal communication.
23. L. J. Morris, personal communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

11-GHz Radio:

Nationwide Long-Term Rain Rate Statistics and Empirical Calculation of 11-GHz Microwave Rain Attenuation

By S. H. LIN

(Manuscript received January 27, 1977)

Two methods are described to obtain long-term (≥20 years) distributions of 5-minute point rain rates from data published by the National Climatic Center for U.S. locations. A set of simple empirical formulas for converting the distribution of 5-minute rain rates into rain attenuation distributions on 11-GHz radio paths has been deduced from data measured in Georgia. Additional data measured in several other locations also support this empirical formulation. These simple formulas and 5-minute point rain rate distributions are useful for path engineering of 11-GHz radio. The work on rain rate distributions discussed in the paper derives from approaches suggested by the late W. Y. S. Chen.[5,9,16]

I. INTRODUCTION

An important problem in designing terrestrial and earth-satellite radio systems at frequencies above 10 GHz is the added path attenuation caused by rain. Long term (≥20 years) rain rate statistics are needed to engineer radio paths for various geographic locations to meet reliability objectives. Section II describes a method to obtain 20-year distributions of 5-minute point rain rates from the excessive short duration rainfall data[1,2] for locations in the relatively wet eastern and midwestern U.S.A. Section III describes another method, employing the theory of extreme value statistics, to obtain 50-year distributions of 5-minute point rain rates from rainfall-intensity-duration-frequency curves[3] for locations in relatively dry locations such as western U.S.A. Section IV discusses the variability of rain rate distributions with observation time base.

Table I — Thresholds* of excessive short-duration rainfalls

Duration τ, minutes	Minimum depth of recorded rainfall, inches	Threshold (τ minute average rain rate), mm/hr
5	0.25	76.2
10	0.30	45.7
15	0.35	35.6
20	0.40	30.5
30	0.50	25.4
45	0.65	22.0
60	0.80	20.3
80	1.00	19.1
100	1.20	18.3
120	1.40	17.8
150	1.70	17.3
180	2.00	16.9

* Exceeding any one of these 12 thresholds is sufficient to qualify a rainstorm as an excessive rainfall. Therefore, this definition does not require that an excessive rainfall exceeds all the 12 thresholds.

In this paper, a "5-minute rain rate" corresponds to the average value of the randomly varying rain rate in a 5-minute interval and is calculated as $\Delta H/\tau$ where ΔH is the 5-minute accumulated depth of rainfall and $\tau = 5$ minutes $= 1/12$ hour is the rain gauge integration time. Similarly, a "τ-minute rain rate" is the average rain rate in a τ-minute interval.

Sections V to VII describe a set of simple empirical formulas deduced from experimental data in Georgia for converting the distribution of 5-minute point rain rates into distributions of rain attenuation on 11-GHz radio paths. Section VIII compares the calculated results with additional data from other locations.

II. 20-YEAR DISTRIBUTIONS OF 5-MINUTE RAIN RATES

The excessive short duration rainfall data[1] record details of those heavy rainfalls which exceed one or more thresholds; these thresholds are dependent upon the rain integration times τ as shown in Table I. For example, the thresholds are 76 and 20 mm/hr for $\tau = 5$ and 60 minutes, respectively. The data, published in tabular form, consist of a storm-by-storm compilation of accumulated depth of rainfall in the most intense 5, 10, 15, 20, 30, 45, 60, 80, 100, 150, and 180 minute periods. For example, Table II shows such data for Newark, New Jersey in 1972. Only three rainstorms exceeded the excessive rainfall thresholds at Newark during 1972. More detailed discussions on the excessive short duration rainfall data can be found in Refs. 1, 5 and 6.

The method for obtaining 5-minute rain rates from this data source is illustrated in Table III for the storm of August 26, 1972 at Newark. In essence, for each storm, the most intense 5-minute accumulation gives

Table II — Excessive short-duration rainfall, year 1972

Station and Date	5	10	15	20	30	45	60	80	100	120	150	180
New Jersey												
Newark												
Jul 13	0.26	0.37	0.48	0.57	0.74	0.86	1.06	1.48	1.65	1.80	2.10	2.37
Jul 17	0.31	0.48	0.49	0.52	0.61	0.70	0.74	0.74	0.74	0.74	0.74	0.74
Aug 26	0.64	1.08	1.54	1.68	1.80	1.87	1.93	1.98	2.02	2.03	2.05	2.06
Trenton												

Duration in minutes

Accumulated depth of rainfall in inches

Table III — Obtaining 5-minute rain rates from excessive short-duration rainfall data, year 1972

	5	10	15	20	30	45	60	80	100	120	150	180	Minutes
Newark													
Aug 26	0.64	1.08	1.54	1.68	1.80	1.87	1.93	1.98	2.02	2.03	2.05	2.06	

$$\frac{0.64 \text{ inch}}{5 \text{ minutes}} = 195 \text{ mm/hr}$$

$$\frac{(1.08 - 0.64) \text{ inch}}{(10 - 5) \text{ minutes}} = \frac{0.44 \text{ inch}}{5 \text{ minutes}} = 134 \text{ mm/hr}$$

$$\frac{(1.68 - 1.54) \text{ inch}}{(20 - 15) \text{ minutes}} = 43 \text{ mm/hr}$$

one sample of 5-minute rain rate; the difference between the 10-minute and 5-minute greatest accumulations gives the second sample of 5-minute rain rate; and so on. By applying this single operation to 20 years (1953 to 1972) of such data for Newark, New Jersey, we obtain a 20-year distribution of 5-minute rain rates in the range above the 76 mm/hr threshold as shown in Fig. 1. The results of this method were tested[5] using the reports of the 20 storms which were available in more detailed form[6] and excellent agreement was found.[5]

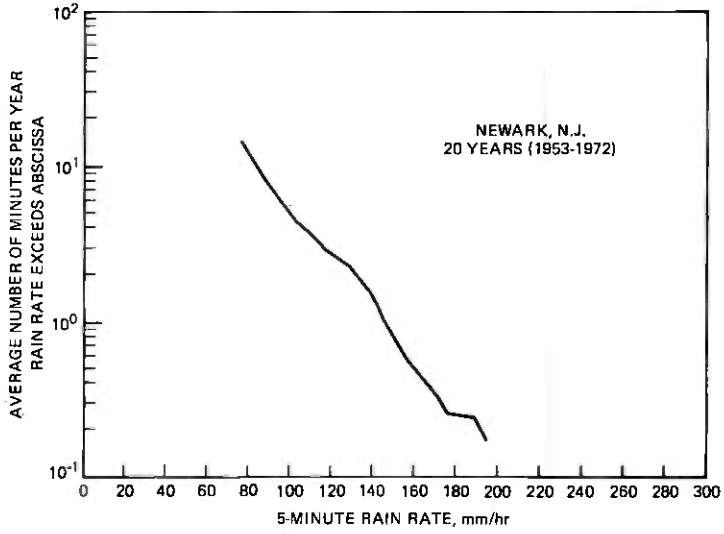


Fig. 1—Twenty-year distribution of 5-minute rain rates above the 76-mm/hr threshold at Newark, New Jersey.

As observed in Table I, the published 5-minute rain rates are given only for those storms which exceed the thresholds in Table I. These data are, therefore, incomplete with respect to 5-minute rain rates of less than 76 mm/hr. However, in the same publication, progressively lower thresholds (see Table I) are used for longer durations; for example, the threshold at 60-minute duration is 20 mm/hr. The method developed to extend the distributions of 5-minute rain rate to very low rain rate employs measured distributions of 1-hour rain rates as the basic data, since these are readily available.[1,2,7] By using long-term ($\geq$10 years) data from New York City, Miami, and McGill Observatory in Canada, a simple empirical formula was deduced for converting the 1-hour rain rate distribution into the 5-minute rain rate distribution in the low rain rate region.[8] A normalizing procedure is used in this conversion formula

to account for the difference in rain characteristics between geographic areas. Twenty-year distributions of 5-minute rain rates for locations in the eastern and midwestern U.S.A. have been obtained by this process. Some results are given in Refs. 8 and 9.

In low rain rate areas such as Oregon and Washington, however, few rainfalls exceed the criterial thresholds, and hence few are included in the excessive short duration rainfall data. For example, at Spokane, Washington, the 5-minute rain rate exceeded the 76 mm/hr threshold only once in the 20-year interval from 1953 to 1972. In such areas, a different data source and a method using statistics of extremes, discussed in the next section, is more suitable.

III. 50-YEAR RAIN RATE DISTRIBUTIONS AND EXTREME VALUE STATISTICS

The statistical behavior of the extremes of a random variable has been extensively investigated.[10–15] In an unpublished work, W. Y. S. Chen and R. L. Lahlum[16] have applied the theoretical distribution of yearly maximum 5-minute rain rates and an empirical extrapolation to obtain the rain rate distribution in the range of interest to radio path engineering. We extend Chen and Lahlum's method by incorporating the theoretical distributions of the yearly K largest 5-minute rain rates for K ranging from 1 to 12. The application of the higher-order statistics of extremes eliminates the need for empirical extrapolation.

Briefly, the average time per year that a rain rate r, measured by a gauge with integration time τ, is exceeded, is given by[9]

$$T(R \geq r) \simeq \tau \sum_{K=1}^{12} \left\{ 1 - e^{-e^{-y}} \sum_{N=0}^{K-1} \frac{e^{-Ny}}{N!} \right\} \tag{1}$$

for the range $T(R \geq r) \leq 50$ minutes/year, where

$$y = \alpha[(\ln r) - U] \tag{2}$$

is called the reduced variate, and α and U are scale and location parameters, respectively. Notice that $T(R \geq r)$ in eq. (1) is uniquely determined by the two parameters α and U. These two parameters can be calculated from the rainfall-intensity-duration-frequency curves[3] which are available for U.S. locations. These rainfall-intensity-duration-frequency curves are derived by the Gumbel method[11,12] using the theory of extreme value statistics and are based on approximately 50 years (1900 to 1950) of rainfall data.

From this data source, we need only the following three numbers for a given location to calculate the long-term distribution of 5-minute rain rates:

M = the number of years of rainfall data from which rainfall-intensity-duration-frequency curves are derived,

r_a = the extreme rain rate with 2-year return period, i.e., the rain rate which is exceeded once in 2 years, on average, by the yearly maximum 5-minute rain rates,

r_b = the extreme rain rate with 10-year return period, i.e., the rain rate which is exceeded once in 10 years, on average, by the yearly maximum 5-minute rain rates.

The formulas for calculating α and U are:[9]

$$\alpha = \alpha_\infty \sigma_z \frac{\sqrt{6}}{\pi} \tag{3}$$

$$U = U_\infty + \frac{1}{\alpha_\infty}\left[\gamma - \frac{\overline{Z}}{\sigma_Z} \cdot \frac{\pi}{\sqrt{6}}\right] \tag{4}$$

where

$$\alpha_\infty = \frac{A_a - A_b}{\ln r_a - \ln r_b} \tag{5}$$

$$U_\infty = \frac{A_a \ln r_b - A_b \ln r_a}{A_a - A_b} \tag{6}$$

$$A_a = -\ln\left(\ln \frac{10 \text{ years}}{2 \text{ years} - 1 \text{ year}}\right) \approx 0.3665 \tag{7}$$

$$A_b = -\ln\left(\ln \frac{10 \text{ years}}{10 \text{ years} - 1 \text{ year}}\right) \approx 2.25 \tag{8}$$

$$\gamma = \text{Euler's Constant} \approx 0.5772 \tag{9}$$

$$Z(j) = -\ln\left(-\ln \frac{j}{M+1}\right) \tag{10}$$

$$\overline{Z} = \frac{1}{M}\sum_{j=1}^{M} Z(j) \tag{11}$$

$$\overline{Z^2} = \frac{1}{M}\sum_{j=1}^{M} [Z(j)]^2 \tag{12}$$

and

$$\sigma_z = \sqrt{\overline{Z^2} - \overline{Z}^2} \tag{13}$$

For example, Fig. 2 shows a portion* of the rainfall-intensity-duration-frequency curves for New York City. The required three numbers read from Fig. 2 are

* The source curves in Ref. 3 cover a wider range for duration τ from 5 to 1440 minutes and return period from 2 to 100 years.

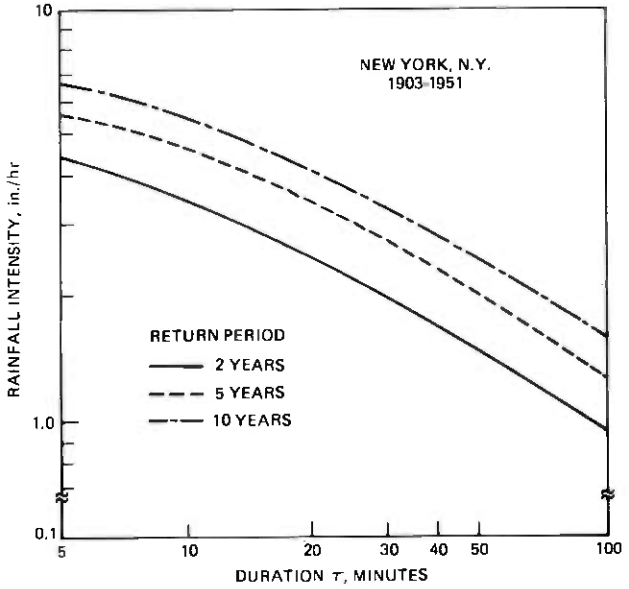


Fig. 2—Rainfall intensity-duration-frequency curve by method of extreme values (after Gumbel) for New York City based on 49 years (1903–1951) of rainfall data.

M = 49 years (1903 to 1951)
r_a = 4.4 inches/hr = 111.8 mm/hr
r_b = 6.5 inches/hr = 165 mm/hr

By substituting these three values into eqs. (3) to (13), we obtain

$$\alpha = 4.363$$

$$U = 4.63$$

Substituting this α, U pair into eqs. (1) and (2) yields the 49-year distribution (dashed line) of 5-minute rain rates as shown in Fig. 3. The 49-year (1903 to 1951) distribution obtained agrees well with the 20-year (1953 to 1972) distribution obtained by the method based on excessive short-duration rainfall data.

Long-term distributions of 5-minute rain rates for U.S. locations can therefore easily be obtained by the extreme value method.

IV. VARIABILITY OF SHORT-TERM DISTRIBUTIONS OF 5-MINUTE RAIN RATES

Figure 4 shows that the 20-year distributions of 5-minute rain rates at Central Park,* La Guardia Airport,* and Newark Airport* agree

* All within the New York Metropolitan area.

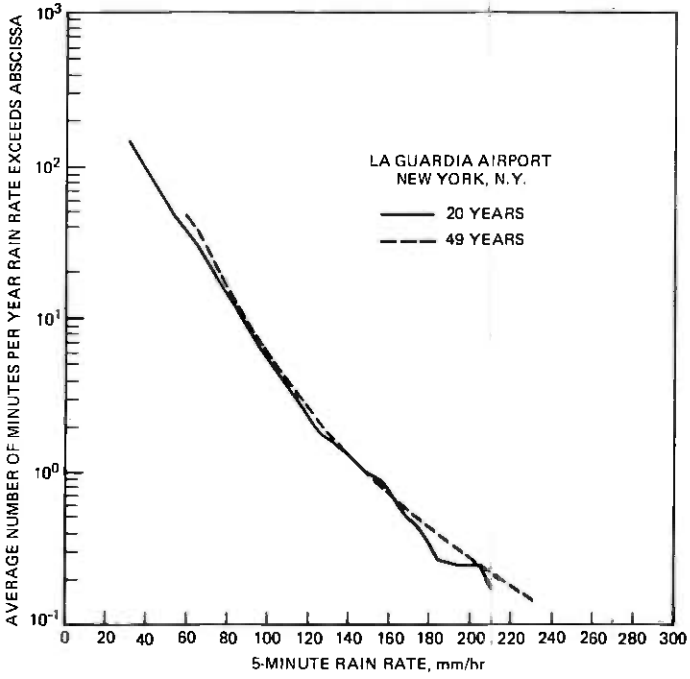


Fig. 3—Comparison of 49-year (1903–1951) distribution of 5-minute rain rates calculated by extreme statistics method with measured 20-year (1953–1972) data at La Guardia Airport, New York City.

closely. On the other hand, Figs. 5 and 6 show that 4-year (1969 to 1972) and 1-year (1972) distributions at these three locations differed significantly. Figure 7 displays the convergence of the rain rate distributions at Newark as the time base is increased from 1 to 20 years. Figure 8 indicates that with a single rain gauge measurement, even a 20-year time base may still be insufficient to provide stable statistics of extremely high rain rates (beyond 160 mm/hr).

Table IV lists the intervals, in minutes per year, that the rain rate exceeded 140 mm/hr at Newark, La Guardia Airport, and Central Park during the 20-year observation period. The last column contains the three-site summation for each year. The last three rows in Table IV indicate the 20-year average $\bar{t}$, the difference Δt, between the worst year and the best year, and the ratio $\Delta t/\bar{t}$ respectively. Figure 9 shows that the normalized range of variations, $\Delta t/\bar{t}$, increases rapidly as the rain rate increases from 80 to 160 mm/hr. This behavior is consistent with the divergence between the upper and lower envelopes in Fig. 7. In Fig. 9, notice that the normalized range of variation, $\Delta t/\bar{t}$, of individual sites are significantly greater than that of three-site summation for high rain rates. Since point rain rate statistics may be representative of a short

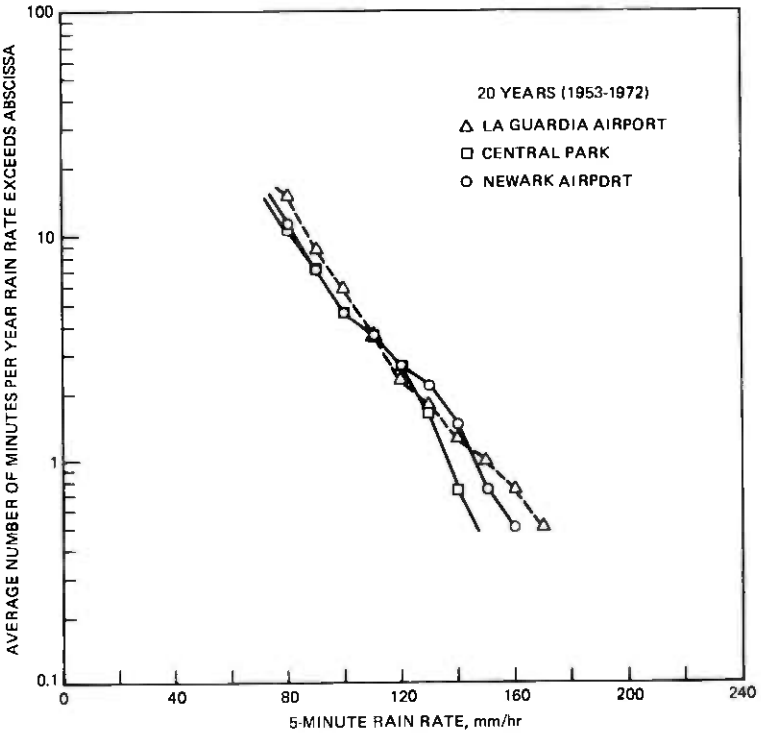


Fig. 4—Twenty-year (1953–1972) distributions of 5-minute rain rates at three locations: La Guardia Airport, Central Park, and Newark Airport, in the New York Metropolitan area.

radio hop, but three-site summation statistics are more representative for a multihop radio route, then, as shown in Fig. 9, the normalized range of variations, $\Delta t/\bar{t}$, of annual outage time of a radio route can be expected to be much smaller than that of a short radio hop. In other words, the annual outage time of a radio route is statistically more stable than that of a short radio hop. Intuitively, the statistical stability of the accumulated outage time of a radio route stems from the partial compensation effect of the incoherent, random variations of individual-hop outage times as displayed in Figs. 5 and 6.

The three-site, 20-year measurements yield 60 samples of annual accumulated time that rain rates exceed 140 mm/hr as listed in Table IV. The average value of these 60 samples is 0.97 minutes per year. Notice that 51 out of these 60 samples are less than the average value. Such nonsymmetric deviations of small sample data from long-term, large-sample average have already been observed and discussed in Ref. 17.

These data indicate that rain rate statistics gathered from a single rain gauge measurement require a very long time base to yield stable statistics

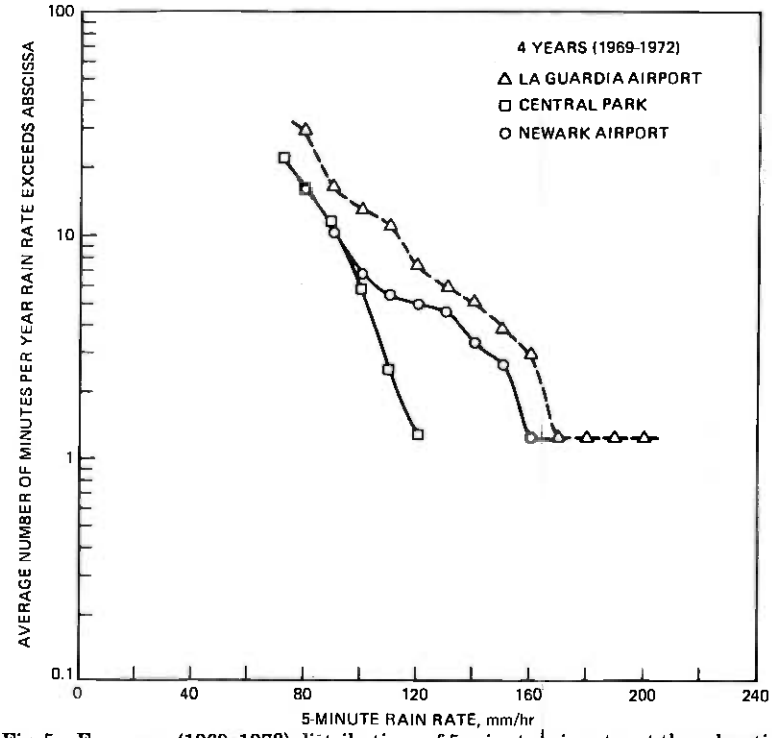


Fig. 5—Four-year (1969–1972) distributions of 5-minute rain rates at three locations: La Guardia Airport, Central Park, and Newark Airport, in the New York Metropolitan area.

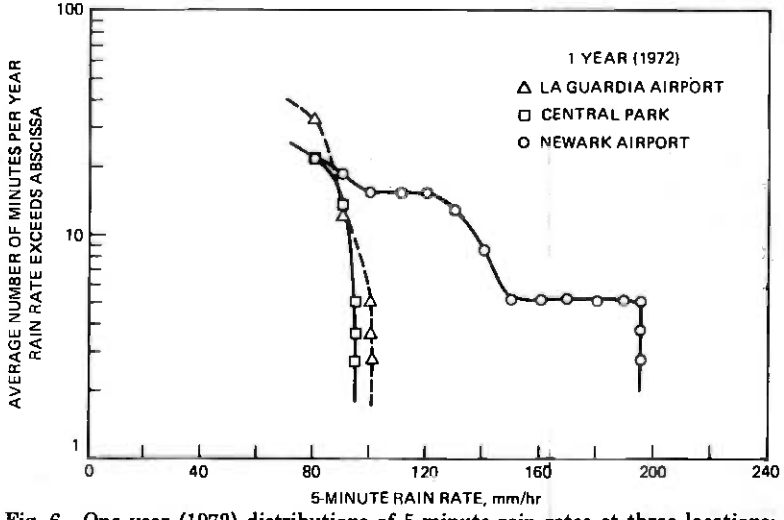


Fig. 6—One-year (1972) distributions of 5-minute rain rates at three locations: La Guardia Airport, Central Park, and Newark Airport, in the New York Metropolitan area.

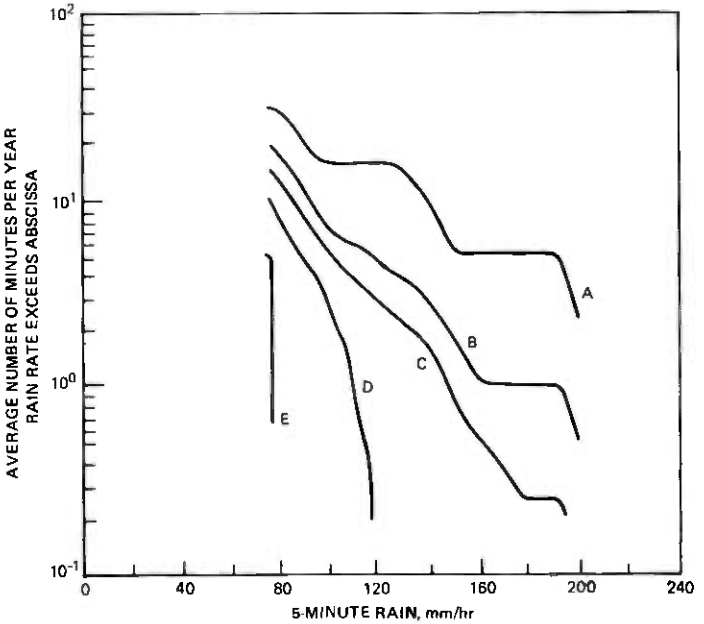


Fig. 7—Variations of yearly distributions of 5-minute rain rates at Newark, New Jersey; A: 1-year upper envelope, B: 5-year upper envelope, C: 20-year average, D: 5-year lower envelope, E: 1-year lower envelope.

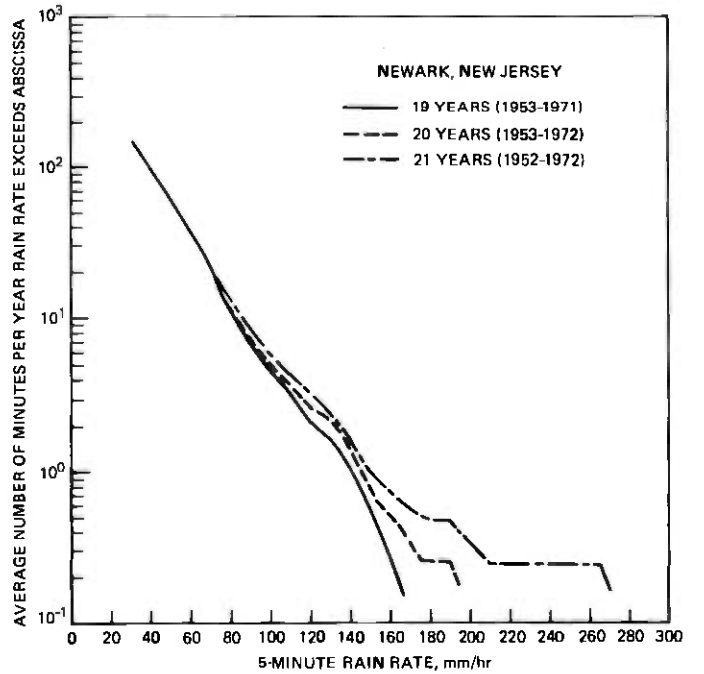


Fig. 8—Comparison of 19-, 20-, and 21-year distributions of 5-minute rain rates at Newark, New Jersey, showing the instability in the extremely high rain rate region.

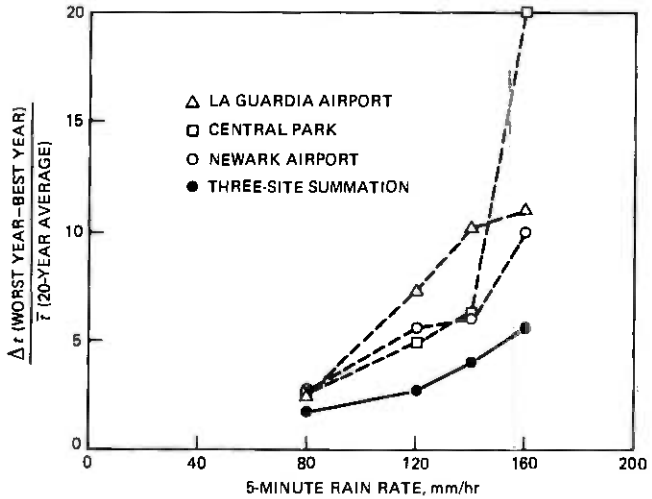


Fig. 9—The effect of three-site summation on the normalized range of variations of annual distributions of 5-minute rain rate. Δt is the difference between the worst year and the best year in number of minutes per year rain rate exceeds abscissa during the 20-year period (1953–1972), and $\bar{t}$ is the number of minutes per year rain rate exceeds abscissa averaged over the 20-year period (see Table IV).

Table IV — Number of minutes per year rain rate exceeds 140 mm/hr

Time base	Newark	La Guardia Airport	Central Park	Three-site summation
1953	0.0	0.0	0.0	0.0
1954	0.0	0.0	0.0	0.0
1955	0.0	0.0	0.0	0.0
1956	0.0	0.0	0.0	0.0
1957	0.0	0.0	0.0	0.0
1958	0.0	0.0	0.0	0.0
1959	0.0	0.0	0.0	0.0
1960	0.0	0.0	5.0	5.0
1961	5.0	0.0	0.0	5.0
1962	0.0	0.0	0.0	0.0
1963	0.0	0.0	0.0	0.0
1964	0.0	0.0	0.0	0.0
1965	5.0	0.0	0.0	5.0
1966	0.0	5.0	0.0	5.0
1967	0.0	0.0	0.0	0.0
1968	0.0	0.0	5.0	5.0
1969	0.0	13.3	0.0	13.3
1970	5.0	0.0	0.0	5.0
1971	0.0	6.7	0.0	6.7
1972	8.3	0.0	0.0	8.3
$\bar{t}$ (20-year average)	1.2	1.3	0.5	2.9
Δt (worst year-best year)	8.3	13.3	5.0	13.3
$\frac{\Delta t}{\bar{t}}$	6.9	10.2	10.0	4.6

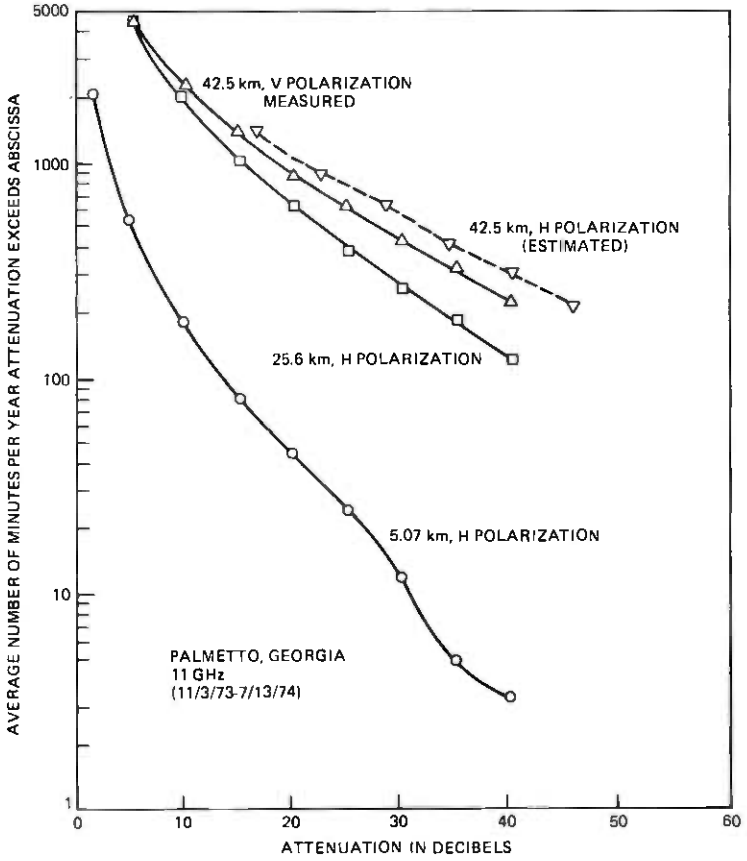


Fig. 10—Distributions of 11-GHz rain attenuation measured on three paths near Atlanta, Georgia.

for engineering a radio route. On the other hand, if the time base is not sufficiently long, the short term results tend to underestimate the long-term, large-sample average.

V. ANALYSES OF MICROWAVE RAIN ATTENUATION

For radio path engineering applications, a procedure is needed to calculate the rain attenuation distributions on microwave radio paths from the available rain rate distributions. Several independent theoretical analyses relating rain attenuation distributions to radio path length have been developed.[18,23,30] One analysis is based upon the approximate log-normality of long-term distributions of rain attenuation and of point rain rates.[17–22] These two distributions are related by suitably derived parameters. Since existing theory for converting rain rate into rain attenuation applies to spatially inform rain rates, whereas

actual rainfalls are almost never uniform over a radio path, the hop is divided into incremental volumes in each of which the rain rate is uniform; the total attenuation is then obtained by integrating over the path. Since the rain rates at various positions along the radio path only partially correlated, the increase of attenuation with radio path length is nonlinear.

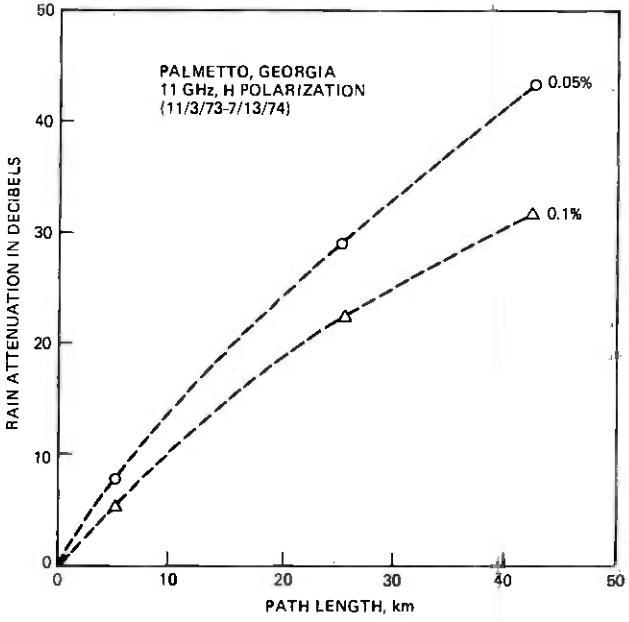


Fig. 11—Nonlinear dependence of 11-GHz rain attenuation on path length measured at Palmetto, Georgia.

In another analysis,[23] rain cells of circular cross section are assumed, allowing calculation of the probable length of path on which rain will fall from the probability of rain occurring at a point. It is found that the increase of attenuation with hop length is nonlinear because of the finite diameter of rain cells. In the limit, for hop lengths smaller than the cell diameter, the attenuation is almost proportional to path length, but for hops much larger than the cell diameter, it is almost independent of path length.

Both analyses indicate that the nonlinearity of the path length dependence is a function of rain rate and rigorous derivations are fairly complex. For path engineering applications, it is desirable to have a simple empirical formula to describe this nonlinear behavior. The following two sections describe the empirical formulas deduced from the available 11-GHz rain attenuation data.

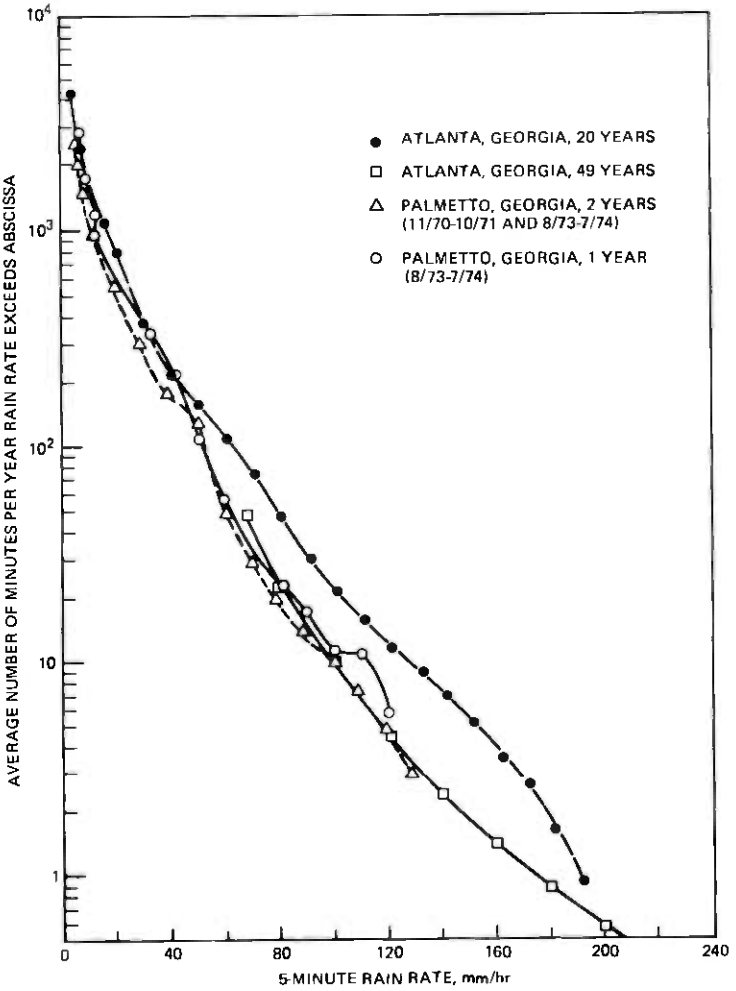


Fig. 12—Rain rate distributions measured at Atlanta and Palmetto, Georgia.

VI. RAIN ATTENUATION AND RAIN RATE DATA AT ATLANTA

The 11-GHz rain attenuation distributions obtained by simultaneous measurement on three paths (5.07, 25.6, and 42.5 km) near Atlanta[24] from November, 1973 to July, 1974 are shown for illustration in Fig. 10. The path-length dependence derived from these data is indicated in Fig. 11 which is a cross-plot relating attenuation observed to path length traversed, for fixed levels of probability. These data demonstrate that, for probability levels of 0.05 and 0.1 percent, rain attenuation increases nonlinearly with increased path length.

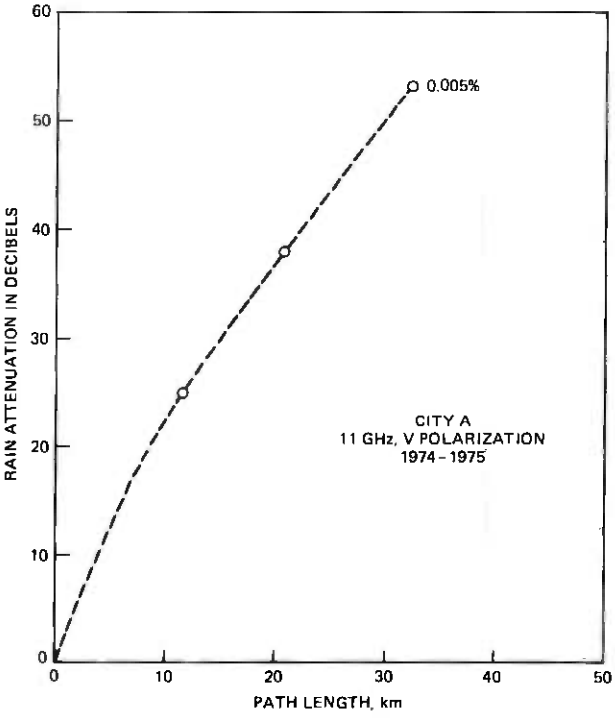


Fig. 13—Nonlinear dependence of 11-GHz rain attenuation on path length measured at another location.

Rain rate data for the above measuring interval are available from a tipping-bucket rain gauge at a common path terminal (Palmetto, Georgia) and from a Weather Bureau rain gauge at Atlanta Airport.[1,2,3] Figure 12 shows the distributions of 5-minute point rain rates obtained from these two rain gauges, as well as long-term results from the same Atlanta station.

Figure 13 shows another example of nonlinear increase of rain attenuation with path length measured at another location. The data in Figs. 13 and 15 are from the same city.

VII. EMPIRICAL PATH-LENGTH DEPENDENCE

Many authors[26–35] have pointed out that the relationship between the rain attenuation gradient, β in dB/km, and the point rain rate, R in mm/hr, can be approximately described by

$$\beta(R) = \rho R^{\eta} \tag{14}$$

where the coefficient ρ and the exponent η depend on the radio fre-

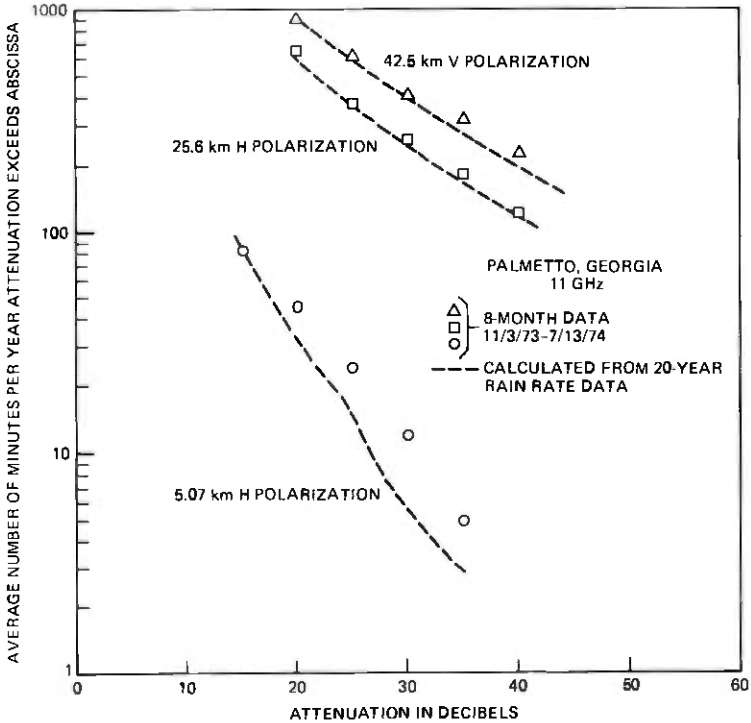


Fig. 14—Comparison of calculated (dashed lines) versus measured 11-GHz rain attenuation distributions on 5.07, 25.6, and 42.5 km paths at Palmetto, Georgia. The calculations are based on 20-year distribution of 5-minute point rain rates at Atlanta, Georgia.

quency, the polarization of the radio signal, and the canting angles of the oblate raindrops. Based on T. S. Chu's theoretical calculation[36] and the experimental data on the polarization dependence of rain attenuation in Refs. 37 and 38, the empirical formula for 11-GHz rain attenuation gradient is

$$\beta_V(R) = 0.0153R^{1.1909} \quad \text{dB/km} \tag{15}$$

for vertically polarized signals and

$$\beta_H(R) = 0.0170R^{1.2012} \quad \text{dB/km} \tag{16}$$

for horizontally polarized signals.

If the rain rates were uniform over a radio path of length L (km), the path rain attenuation $\mu(R,L)$ would be simply $\beta(R)L$, but since actual rainfalls are not uniform, the increase of $\mu(R,L)$ with L is nonlinear. In Ref. 39, a two-parameter empirical formula was shown to describe this nonlinear behavior. A single parameter variation also provides satis-

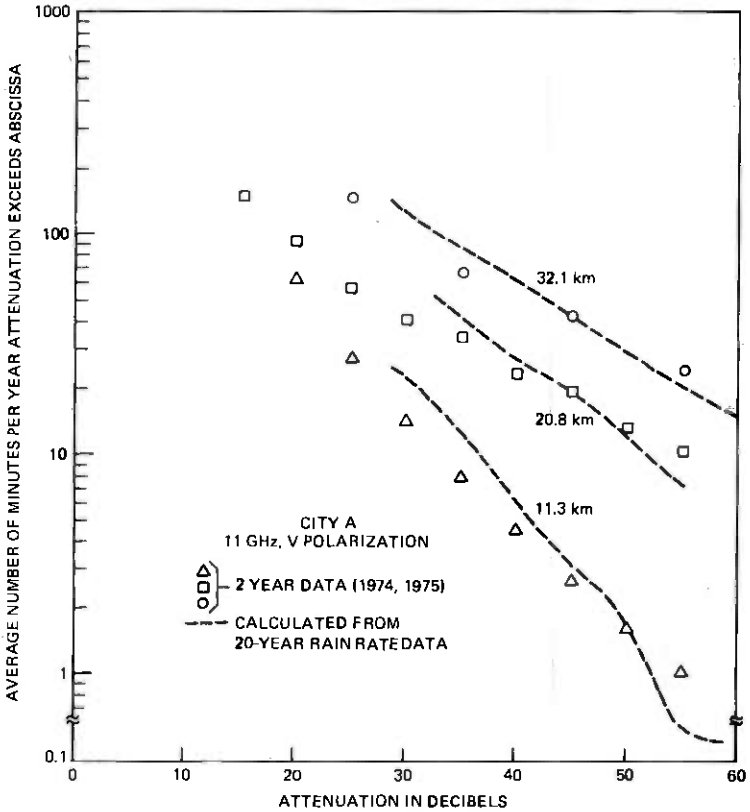


Fig. 15—Comparison of calculated (dashed lines) versus measured 11-GHz rain attenuation distributions on 11.3, 20.8, and 32.1 km paths at City A. The calculations are based on 20-year distribution of 5-minute point rain rates at City A.

factory agreement:

$$\mu(R,L) = \beta(R)L \frac{1}{1 + \frac{L}{\bar{L}(R)}} \tag{17}$$

where the factor

$$\frac{1}{1 + \frac{L}{\bar{L}(R)}} \tag{18}$$

accounts for the partially correlated rain rate variations along the path, and $\bar{L}(R)$ is a characteristic path length such that the nonlinear factor (18) equals one-half when $L = \bar{L}$. $\bar{L}$ is related to the diameter of the rain cell.

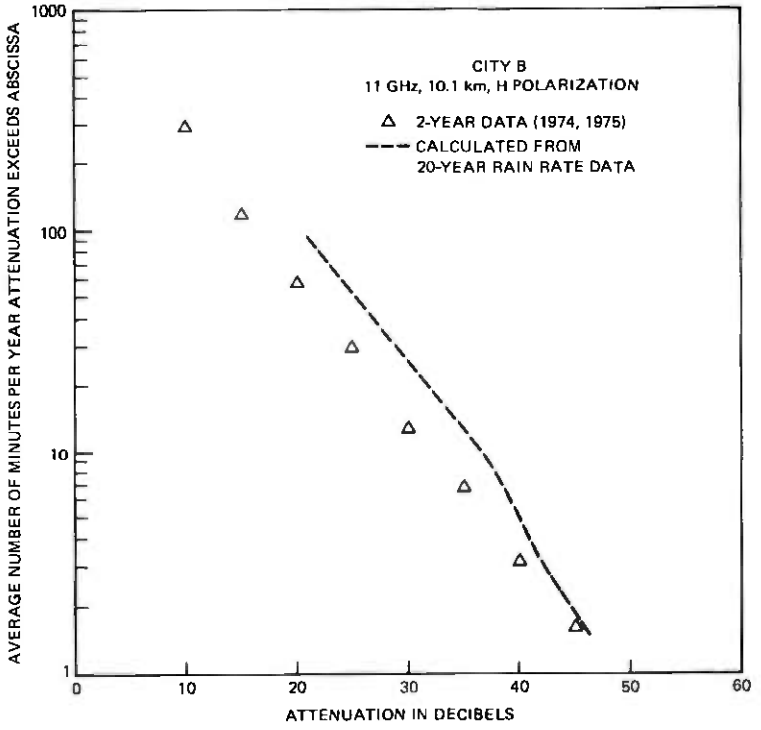


Fig. 16—Comparison of calculated (dashed line) versus measured 11-GHz rain attenuation distribution on a 10.1-km path at City B. The calculation is based on 20-year distribution of 5-minute point rain rate at City B.

By fitting eq. (17) to concurrently measured (August 1973 to July 1974) distributions of 5-minute point rain rates and 11-GHz rain attenuation on the 42.5 km path at Palmetto, Georgia, it is found that $\overline{L}(R)$ can be approximately described by

$$\overline{L}(R) \approx \frac{2636}{R - 6.2} \text{ km} \qquad \text{for} \qquad R > 10 \text{ mm/hr} \tag{19}$$

VIII. COMPARISON OF PREDICTED AND MEASURED 11-GHZ ATTENUATIONS

Measured rain attenuation data on nine 11-GHz paths, listed in Table V, are available for comparison with the calculated results. Figures 14 to 18 show such comparisons. The calculated results (dashed lines) are based on 20-year distributions of 5-minute point rain rates and include path rain attenuation and assumed wet radome loss listed in Table V. It is assumed that, on the average, a wet, flat radome introduces 2 dB loss and a wet, hemispheric radomes introduces 4-dB loss.[18,40,41] Figures

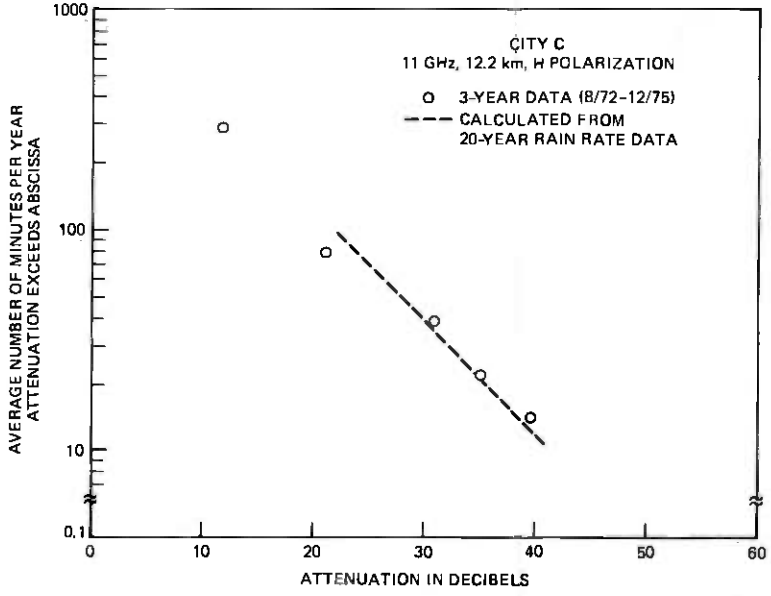


Fig. 17—Comparison of calculated (dashed line) versus measured 11-GHz rain attenuation distribution on a 12.2-km path at City C. The calculation is based on 20-year distribution of 5-minute point rain rates at City C.

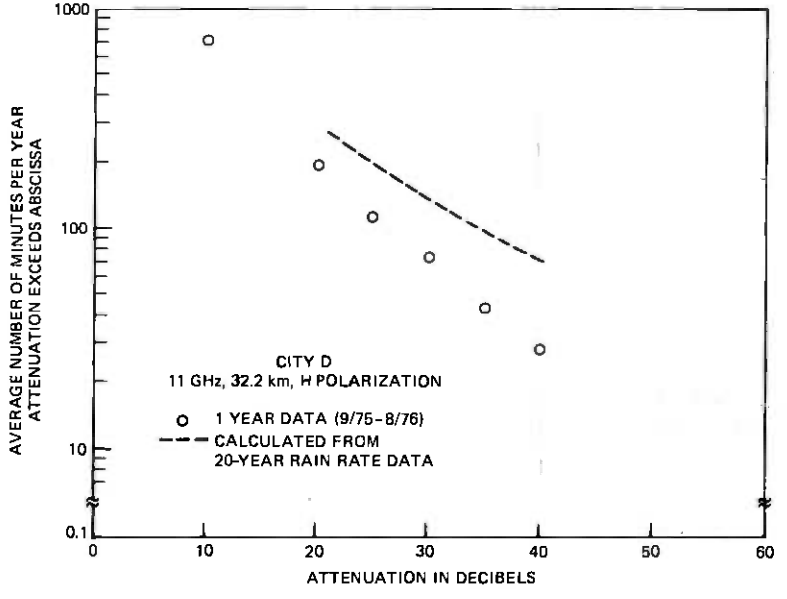


Fig. 18—Comparison of calculated (dashed line) versus measured 11-GHz rain attenuation distribution on a 32.3-km path at City D. The calculation is based on 20-year distribution of 5-minute point rain rates at City D.

Table V — Measured rain attenuation data on 11-GHz paths

Location	Path length, km	Time base	Radomes	Polarization	Assumed attenuation (dB) due to two wet radomes for calculations
Palmetto, Ga.	42.5	8/73–7/74	2 flat radomes	V	4
Palmetto, Ga.	25.6	8/73–7/74	2 flat radomes	H	4
Palmetto, Ga.	5.1	8/73–7/74	2 dishes without radome	H	4
City A	32.1	74–75	2 flat radomes	V	4
City A	20.8	74–75	2 flat radomes	V	4
City A	11.3	74–75	2 flat radomes	V	4
City B	10.1	74–75	1 flat; 1 hemispheric	H	6
City C	12.2	8/72–12/75	2 hemispheric radomes	H	8
City D	32.2	9/75–8/76	2 flat radomes	H	4

14 to 18 indicate that the measured results are comparable with the predicted curves.

IX. CONCLUSION

Two methods have been described to obtain long term ($\geq$20 years) distributions of 5-minute rain rates from data published by the National Climatic Center for U.S. locations. Some typical results are given in Refs. 8 and 9. The variability of distributions based on shorter terms is discussed in Section IV.

A set of simple empirical formulas, for converting the distributions of 5-minute rain rates into the distributions of 11-GHz rain attenuation on any path length, has been deduced from experimental data gathered near Atlanta, Georgia. These formulas are supported by further experimental data from other locations.

ACKNOWLEDGMENT

The author is grateful to the late W. Y. S. Chen for valuable discussions on rain rate statistics and the theory of extreme value statistics. W. C. Y. Lee,[42] R. A. Semplak,[44] P. L. Rice and N. R. Holmberg[43] have separately described three different approximate methods for obtaining rain rate distributions from rainfall data published by the National Climatic Center. The method described in Ref. 5 by W. Y. S. Chen and in Section II of this paper is inspired from this work. The author would also like to thank C. A. Maxon, P. L. Dirner, and E. E. Sorrentino for their assistance in the computer processing of the large volume of rainfall data.

REFERENCES

1. "Climatological Data, National Summary," annual issues since 1950, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Climatic Center, Federal Building, Asheville, North Carolina 28801. The excessive short-duration rainfall data prior to 1950 are published in the Monthly Weather Review, the U.S. Meteorological Yearbook (last published for the period 1943 to 1949), and the Report of the Chief of the Weather Bureau (last published for 1931).
2. "Climatography of the United States No. 82, Decennial Census of United States Climate—Summary of Hourly Observations (1951 to 1960)." National Climatic Center, Federal Building, Asheville, North Carolina 28801.
3. "Rainfall Intensity Duration Frequency Curves," U.S. Department of Commerce, Weather Bureau, Technical Paper No. 25, Washington, D.C., December, 1955, Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.
4. T. L. Osborne, "Application of Rain Attenuation Data to 11-GHz Radio Path Engineering," B.S.T.J., this issue, pp. 1605–1628.
5. W. Y. S. Chen, "A Simple Method for Estimating 5-Minute Rain Rate Distributions Based on Available Climatological Data," B.S.T.J., *55*, No. 1 (January 1976), pp. 129–134.
6. D. L. Yarnell, "Rainfall Intensity Frequency Data," U.S. Department of Agriculture, Miscellaneous Publication No. 204, August, 1935, Washington, D.C.
7. "Hourly Precipitation Data," published for each state in the U.S.A., Superintendent of Documents, Government Printing Office, Washington, D.C. 20402. The hourly precipitation data for approximately 3000 locations are also stored on magnetic tapes

in the Computer Center of National Climatic Center, Federal Building, Asheville, North Carolina 28801.
8. S. H. Lin, "Dependence of Rain Rate Distribution on Rain Gauge Integration Time," B.S.T.J., *55*, No. 1 (January 1976), pp. 135–141.
9. S. H. Lin, "Rain Rate Distributions and Extreme Value Statistics," B.S.T.J., *55*, No. 8 (October 1976), pp. 1111–1124.
10. R. A. Fisher and L. H. C. Tippett, "Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample," Proc. Cambridge Phil. Soc., *24*, 1928, pp. 180–190.
11. E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications*, National Bureau of Standards, Applied Mathematics series No. 33, February 12, 1954, pp. 15, 16, and 28.
12. E. J. Gumbel, *Statistics of Extremes*, New York: Columbia University Press, 1958.
13. H. Cramer, *Mathematical Methods of Statistics*, New Jersey: Princeton University Press, 1967, Section 12.3, pp. 271–272.
14. H. Cramer and M. R. Leadbetter, *Stationary and Related Stochastic Processes*, New York: John Wiley and Sons, Inc., 1967, Section 12.3, pp. 271–272.
15. G. S. Watson, "Extreme Values in Samples from M-Dependent Stationary Stochastic Processes," Ann. Math. Stat., *25*, 1954, pp. 798–800.
16. W. Y. S. Chen and R. L. Lahlum, unpublished work.
17. S. H. Lin, "Statistical Behavior of Rain Attenuation," B.S.T.J., *52*, No. 4 (April 1973), pp. 557–581.
18. S. H. Lin, "A Method of Calculating Rain Attenuation Distributions on Microwave Paths," B.S.T.J., *54*, No. 6 (July–August 1975), pp. 1051–1086, Figs. 10–13.
19. L. Hansson, "General Characteristics of Rain Intensity Statistics in the Stockholm Area," Tele, Swed. ed., *27*, No. 1, 1975, pp. 43–48.
20. L. Hansson, "General Characteristics of Rain Intensity Statistics in the Gothenburg Area," Report USR 75 012, Central Administration of Swedish Telecommunications, January, 1975, S-12386, Farsta, Sweden.
21. B. N. Harden, D. T. LLewellyn-Jones, and A. M. Zavody, "Investigations of Attenuation by Rainfall at 110 GHz in Southeast England," Proc. Inst. Elec. Eng. London, *122*, No. 6 (June 1975), pp. 600–604.
22. P. T. Schickedanz, "Theoretical Frequency Distributions for Rainfall Data," International Symposium on Probability and Statistics in the Atmospheric Sciences, June 1–4, 1971, Honolulu, Hawaii, sponsored by American Meteorological Society and cosponsored by the World Meteorological Organization, 45 Beacon Street, Boston, Massachusetts 02108, U.S.A. Preprints of Symposium Papers, pp. 131–136.
23. A. A. M. Saleh, unpublished.
24. M. V. Pursley, private communication.
25. E. J. Szabocsik, private communication.
26. D. C. Hogg, "Path Diversity in Propagation of Millimeter Waves Through Rain," IEEE Trans. Ant. Propag., *AP-15*, No. 3 (May 1969), pp. 410–415.
27. K. L. S. Gunn and T. W. R. East, "The Microwave Properties of Precipitation Particles," J. Roy. Meteorol. Soc., *80*, 1954, pp. 522–545.
28. B. C. Blevis, R. M. Dohoo, and K. S. McCormick, "Measurements of Rainfall Attenuation at 8 and 15 GHz," IEEE Trans. Ant. Propag., *AP-15*, No. 3 (May 1967), pp. 394–403.
29. G. B. Stracca, "Propagation Tests at 11 GHz and 18 GHz on Two Paths of Different Length," Alta Freq., *38*, No. 5 (May 1969), pp. 345–361.
30. K. Morita and Z. Higuti, "Statistical Studies on Electromagnetic Wave Attenuation Due to Rain," Rev. Electr. Comm. Lab. (Japan), *19*, No. 7–8 (July–August 1971), pp. 798–842.
31. B. J. Easterbrook and D. Turner, "Prediction of Attenuation by Rainfall in the 10.7–11.7 GHz Communication Band," Proc. Inst. Elec. Eng. London, *114*, No. 5, (May 1967), pp. 557–656.
32. R. R. Rogers, "Statistical Rainstorm Models: Their Theoretical and Physical Foundations," IEEE Trans. Ant. Propag., *AP-24*, No. 4 (July 1976), pp. 547–566.
33. R. K. Crane, "Propagation Phenomena Affecting Satellite Communication Systems Operating in the Centimeter and Millimeter Wavelength Bands," Proc. IEEE, *59*, No. 2 (February 1971), pp. 173–188.
34. S. D. Hathaway and H. W. Evans, "Radio Attenuation at 11 kmc and Some Implications Affecting Relay System Engineering," B.S.T.J., *38*, No. 1 (January 1959), pp. 73–97.

35. L. J. Ippolito, "Effects of Precipitation on 15.3- and 31.65-GHz Earth-Space Transmissions with the ATS-V Satellite," Proc. IEEE, *59*, No. 2, (February 1971), pp. 189–205.
36. T. S. Chu, "Rain Induced Cross-Polarization at Centimeter and Millimeter Wavelengths," B.S.T.J., *53*, No. 8 (October 1974), pp. 1557–1579.
37. W. T. Barnett, "Some Experimental Results on 18 GHz Propagation," Conference Record, 1972 National Telecommunications Conference (IEEE Publication 72 CHO 601-5-NTC), pp. 10E-1–10E-4.
38. C. H. Menzel, private communication.
39. T. L. Osborne, "Rain Outage Performance of Tandem and Path Diversity 18 GHz Short Hop Radio Systems," B.S.T.J., *50*, No. 1 (January 1971), pp. 59–79.
40. I. Anderson, "Measurements of 20-GHz Transmission Through a Radome in Rain," IEEE Trans. Ant. Propag., *AP-23*, No. 5 (September 1975), pp. 619–622.
41. D. C. Hogg and T. S. Chu, "The Role of Rain in Satellite Communications," Proc. IEEE (September 1975), *63*, No. 9, pp. 1308–1331.
42. W. C. Y. Lee, "An Approximate Method for obtaining Statistics on Signal Attenuation Due to Rainfall," to be published. This manuscript with a different title was issued as a Bell Laboratory internal technical memorandum on May 10, 1974.
43. P. L. Rice and N. R. Holmberg, "Cumulative Time Statistics of Surface Point Rainfall Rates," IEEE Trans. Comm., *COM-21*, No. 10 (October 1973), pp. 1131-1136.
44. R. A. Semplak, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

11–GHz Radio:

Application of Rain Attenuation Data to 11-GHz Radio Path Engineering

By T. L. OSBORNE

(Manuscript received February 1, 1977)

This paper describes the procedures adopted at Bell Laboratories for using rain attenuation data to engineer 11-GHz microwave radio hops and routes. Rain outage time charts, which show the rain outage time as a function of rain attenuation and hop length, are the basic tools in engineering the hops. The charts, their formulation, and the procedures for using the charts are described and illustrated with several examples. The procedures are used to demonstrate and quantify the sensitivity of allowable hop lengths to the available rain attenuation margin, the effects of a limited rain attenuation margin, and the effects of the variation in the outage in a single year from the 20-year average outage. Guidelines for judging if a hop or route is performing as engineered are developed.

I. INTRODUCTION AND SUMMARY

In engineering a microwave radio system, as in engineering any system, one of the major concerns is the amount of time that the system will not be usable or that its performance will be below an acceptable level; this is known as outage time. For a system to be reliable, the amount of outage time must not exceed some objective and should be controllable and predictable.

In microwave radio systems, outages can be separated into two classes according to source: failure of the system equipment, and anomalous propagation conditions. In modern systems, the equipment outage time can be made negligibly small by using standby equipment and automatic protection switching systems. Protection against some propagation outages can also be achieved by providing an alternate path or channel and automatic switching; multipath fading protected by space or fre-

quency diversity is an example. However, at frequencies above about 10 GHz, attenuation by rain can cause an outage which is not easily protected by providing an alternate path or channel because rain attenuation is relatively constant with frequency across the common carrier bands, attenuates both polarizations, and covers a fairly large area. The use of route diversity to provide an alternate propagation path has been considered, but because of the cost of providing a complete standby system and the uncertainty of the amount of joint fading on the two routes, this has not been practical.

Therefore, the only practical way of achieving a reliable radio system at those frequencies where there is substantial rain attenuation is to engineer the system in such a way that the expected amount of rain outage is below some outage objective. However, reduction of the amount of rain outage is primarily achieved by reducing the radio repeater spacing, which in turn means increasing the number of radio repeaters. Because radio repeaters are very expensive, this can greatly increase the cost of a system. Conversely, in order to reduce the cost of the system as much as possible, the system must be engineered for the longest hops, and consequently the fewest repeaters, for which the rain outage will meet the outage objective. The unavoidable dependence of the economic-versus-reliability trade-off on a statistical occurrence of nature is peculiar to radio systems operating at frequencies above about 10 GHz, and makes a reliable practical method of path engineering crucial to the future use of 11-GHz radio systems.

Prior to the availability of the data described in these papers, most radio paths were engineered using data and methods developed by Hathaway and Evans and published in 1959.[1] While this work did provide a methodology, it was based on only six months of data on two hops in one city which was then related by not-well-established relationships to rain data in other parts of the U.S. By the mid-1960s, as the future use of more and higher-capacity 11-GHz radio systems became apparent, descrepancies and problems with rain attenuation theory in general were pointed out,[2] and complaints of excessive outage in some existing systems were reported, it became obvious that better design information was needed.

Collectively the companion papers in this issue describe the current theories for predicting the amount of rain outage, the underlying substantiating data, and the methodology for engineering 11-GHz radio systems limited by rain attenuation. The paper by Hogg et al.[3] reviews the factors involved in developing the rain attenuation theory and the work which led to the present understanding. The paper by Lin[4] describes the source and processing of the basic rain rate data, and the algorithm for converting the rain rate data to rain attenuation data.

This paper describes the procedures which have been adopted for

using the rain attenuation data to engineer 11-GHz microwave radio hops and routes. Section II describes the rain outage charts which are used for radio path engineering. Section III describes, and illustrates with examples, methods of using the rain outage charts to estimate annual outage time and allowable hop length to meet a given objective. The methods are then used to demonstrate the sensitivity of allowable hop lengths to the available rain attenuation margin and the effect of a limited rain attenuation margin. Section IV uses the methods of Section III to demonstrate the effects of the variation in the outage in a single year relative to the 20-year average. Section V discusses the geographical coverage of the rain attenuation data and considerations in estimating rain outage in areas where no data exist.

In addition to the methodology just described, some quantative data is derived and is summarized as follows.

(*i*) The differential attenuation between the horizontal and vertical polarization is 8.0 to 8.5 dB for 50 dB of attenuation on the vertical polarization and hop lengths from 60 km to 10 km respectively (Section 3.2).

(*ii*) A 5-dB difference in rain attenuation margin results in a 16 to 18 percent difference in allowable hop length, and a 10-dB difference results in a 30 to 35 percent difference in allowable hop length (Section 3.4).

(*iii*) For midcontinental cities in the U.S., the factors by which the maximum 1-year outage exceeds the 20-year average outage range from 2.5 to 7.1; the variability for the western cities is slightly more. Factors are also given for the maximum 5-year averages (Section 4.1).

(*iv*) Guidelines are developed for judging if a system rain outage performance is as engineered. For example, if the route outage time of a route containing three or more hops exceeds the engineered value by more than a factor of 5 in any one year, or by a factor of 2 for any 5-year average, the outage time is excessive and the reason should be determined. (Section 4.1).

(*v*) If hops were engineered on the basis of the maximum 5-year average outage time, 9 to 27 percent more repeaters would be required than if they were engineered on the basis of the 20-year average. From 25 to 77 percent more repeaters would be required to engineer on the basis of the maximum one-year outage time (Section 4.2).

This paper addresses the problem of outage caused by rain attenuation only. In some cases there may be significant amounts of outage due to other causes such as multipath fading or equipment failure. In such cases, the basic procedure is to allocate only part of the total allowable outage to rain attenuation. Detailed considerations of the allocation and com-

putation of outage time from other effects is outside the scope of this paper.

II. RAIN OUTAGE CHARTS

The generation of rain attenuation statistics as described by Lin[4] is a two-step process. First, long-term appropriately averaged, point rain rate distributions are derived from weather bureau data at a given location. These distributions show the number of minutes per year, T, that the 5-minute point rain rate, R, exceeds a given value, and can be described by the functional relation

$$T = g(R) \tag{1}$$

Secondly, the radio path attenuation, α, for a given hop length, L, and polarization is related to the 5-minute point rain rate, R. From eqs. (14) through (19) in Lin's paper,[4] these relations for vertical and horizontal polarizations are

$$\alpha_V = \frac{.0153R^{1.1909}L}{1 + L\left(\frac{R - 6.2}{2636}\right)} \qquad \alpha_H = \frac{.0170R^{1.2012}L}{1 + L\left(\frac{R - 6.2}{2636}\right)} \tag{2}$$

where α is in dB, R is in mm/hr, and L is in km.

In radio path engineering we are interested in the amount of time the path attenuation exceeds the rain attenuation margin, M_R, on the radio path. This can be determined by setting α equal to M_R and solving eq. (2) for the rain rate which gives the marginal attenuation, then using the rain rate distribution to find the amount of time that rain rate is exceeded. Functionally, the solution of (2) for R can be represented by

$$R = f_V\,(M_R, L) \tag{3a}$$

$$R = f_H\,(M_R, L) \tag{3b}$$

for the vertical and horizontal polarizations respectively. Figure 1 and 2 show the rain rate R as a function of rain attenuation margin with hop length as a parameter for 12 dB $\leq M_R \leq$ 70 dB and 10 km $\leq L \leq$ 60 km.

For engineering purposes, a rain outage chart should show the annual expected rain outage time as a function of the rain attenuation margin and hop length for both vertically and horizontally polarized transmission at a given location. Such charts have been devised by solving eqs. (3) and (1) graphically by juxtaposing the rain rate scales of Figs. 1 and 2 with the rain rate scale of the point rain rate distribution. Examples of the resulting charts are shown in Figs. 3 to 9; Fig. 3 shows the rain rate scale for illustration only.

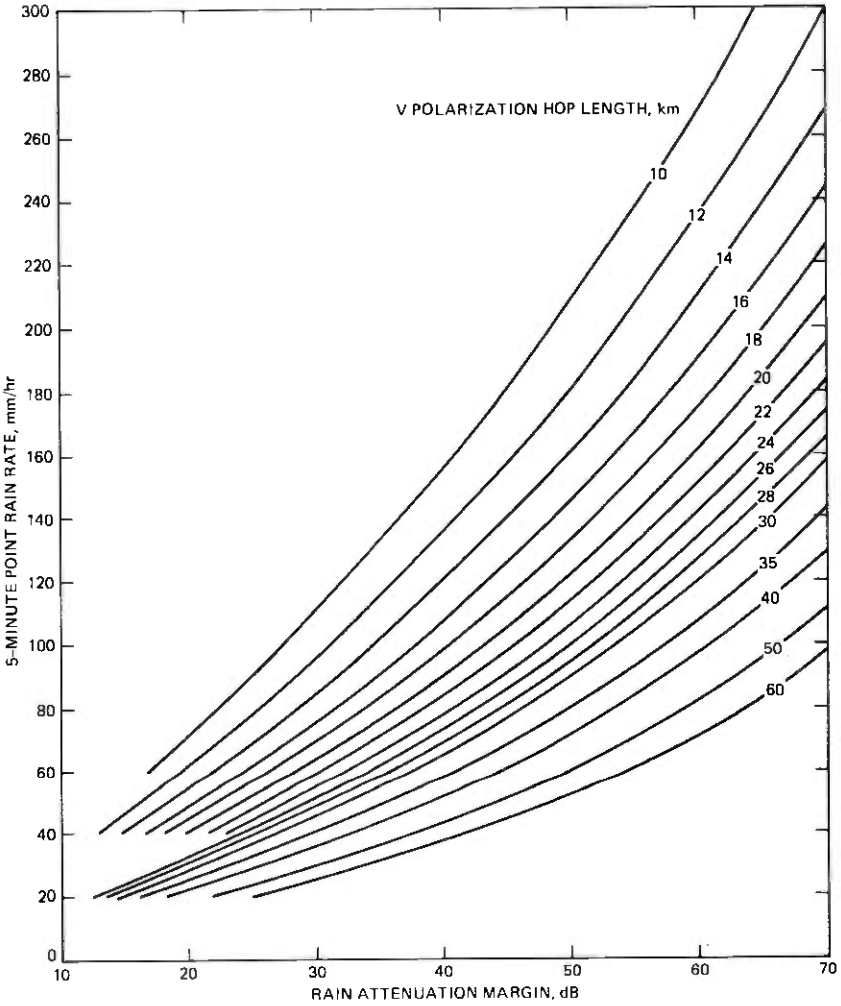


Fig. 1—Graph of the empirical relation between the annual distribution of 5-minute point rain rates and the annual distribution of rain attenuation of the vertical polarization on a radio hop with the hop length as a parameter.

The advantages of this type of rain outage chart are: (*i*) the rain rate distributions are displayed explicitly; (*ii*) the graphical solutions of eqs. (1) and (3) are kept independent, which simplifies the work required in changing the charts should new or revised data become available; (*iii*) the rain rate values are available if needed although not explicitly shown; (*iv*) both horizontal and vertical polarization data is shown on the same chart; (*v*) the rain attenuation margin-hop length scales are the same for every chart.

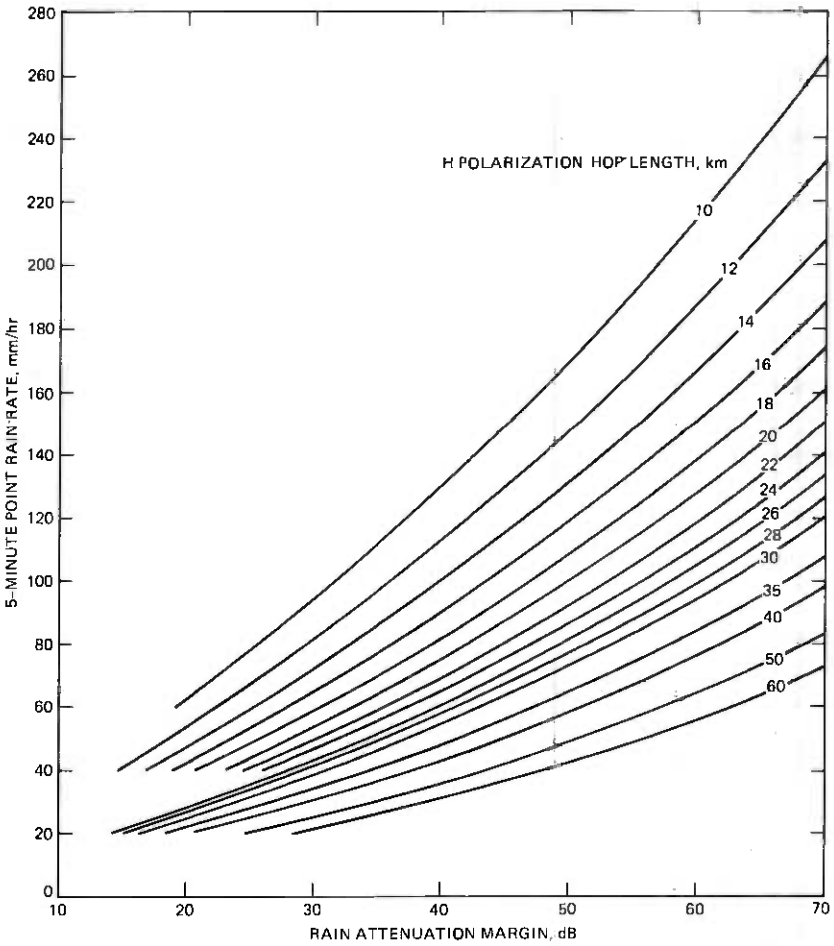


Fig. 2—Graph of the empirical relation between the annual distribution of 5-minute point rain rates and the annual distribution of rain attenuation of the horizontal polarization on a radio hop with the hop length as a parameter.

III. USE OF RAIN ATTENUATION CHARTS

3.1 Determination of rain attenuation margin from system parameters

The first step in using the rain attenuation charts for engineering a radio route is to determine the available rain attenuation margin from the specifications of the equipment used on each hop. This section describes the procedure and gives an example using typical values.

The basic equipment specification is the *system gain*, G_s, for a given performance threshold, which is defined as the dB difference in signal levels between the transmitter bay output and the receiver bay input for the given performance threshold, where the channel combining

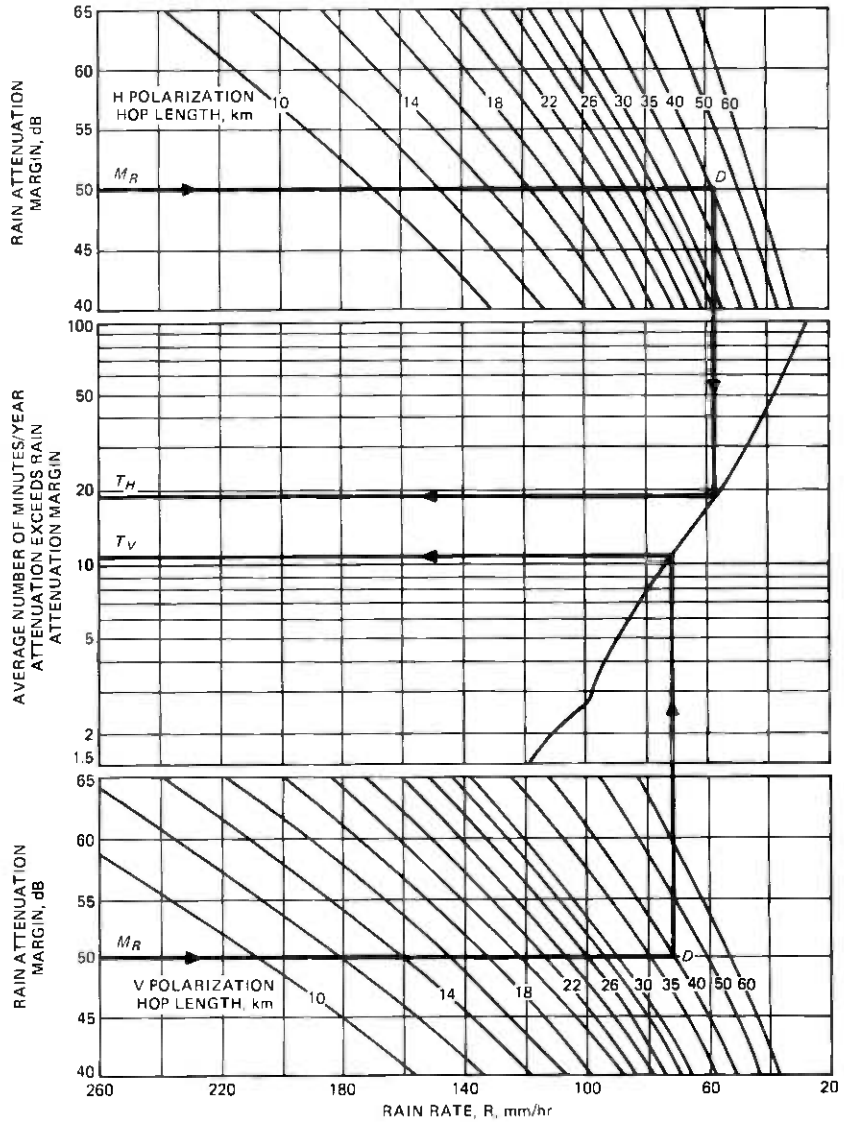


Fig. 3—Illustration of the use of a rain outage chart to find the outage time on the vertical and horizontal polarizations from a known rain attenuation margin and hop length.

networks are assumed to be inside the bays. In digital radio systems the performance threshold is usually specified in terms of bit error rate (BER), whereas in analog radio systems it is specified in terms of voice frequency channel noise. The *total fade margin*, M_T, against rain fading is the amount of flat signal loss that degrades the system performance

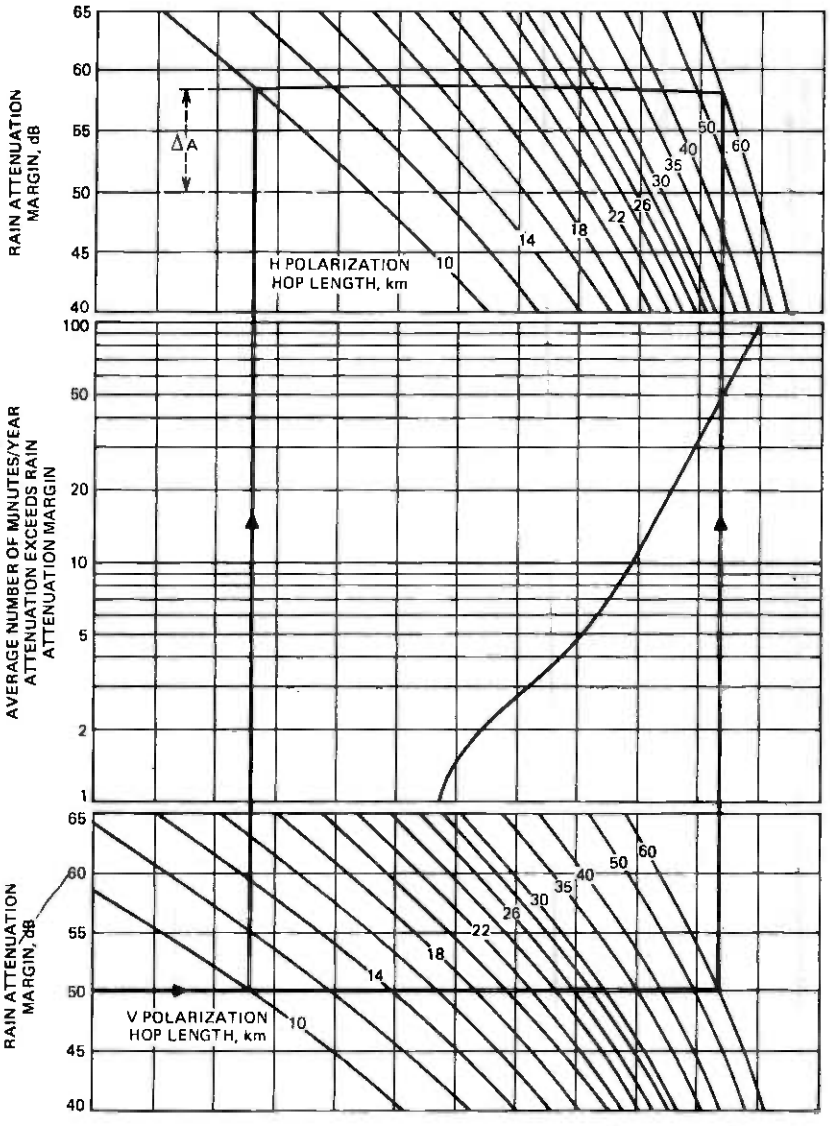


Fig. 4—Illustration of the differential attenuation between the horizontal and vertical polarizations for hop lengths from 10 km to 60 km and an attenuation of 50 dB on the vertical polarization.

to a given threshold in the absence of any other degradations. It can be found from the system gain for the same threshold by subtracting the section loss, L_s, which is the sum of the waveguide, antenna system, and free-space path losses less the antenna gains.

The total fade margin must be allocated to the various losses and

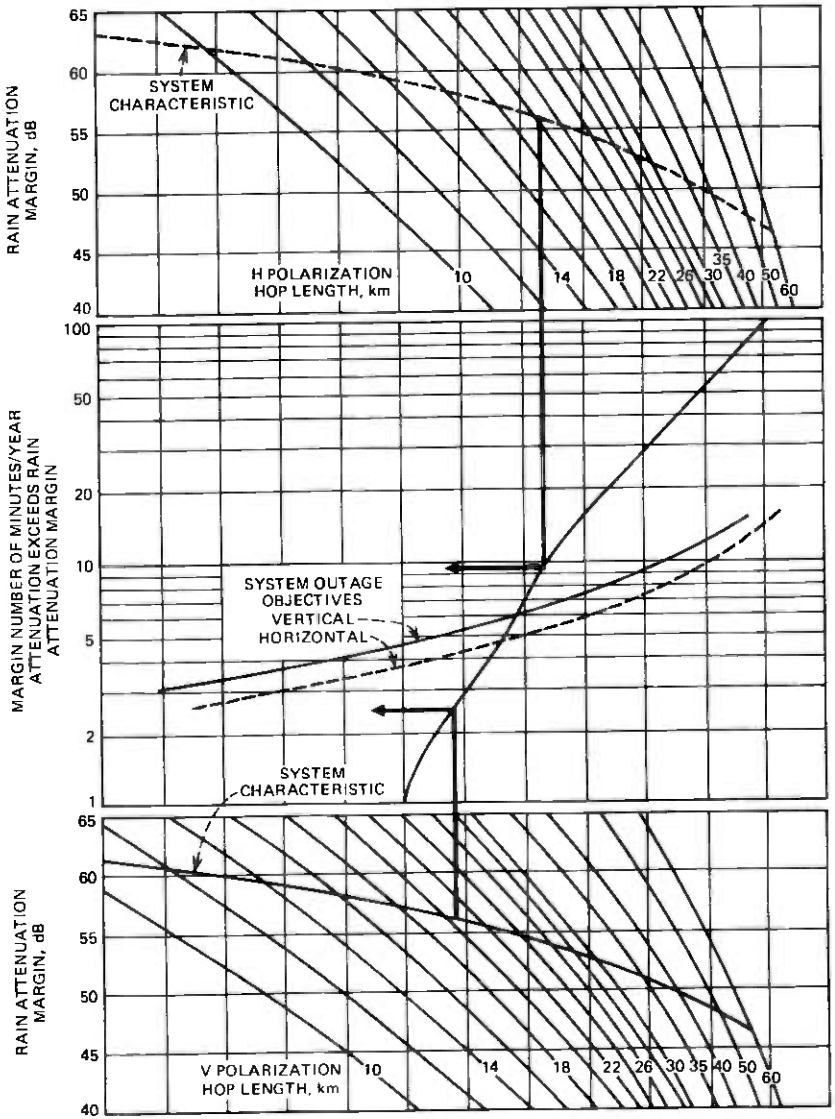


Fig. 5—Illustration of the use of the system characteristic curve and the system objective curves in finding the maximum allowable hop lengths for a particular system.

degradations that occur during rain fading, such as rain attenuation in the aerial path, wet radome loss, depolarization performance degradation, and foreign system interference degradation. The rain attenuation margin, M_R, is that margin which is allowed for *aerial attenuation only*, and is not necessarily equal to the total fade margin.

Wet radome losses have been discussed by Hogg et al.[3] For engineering

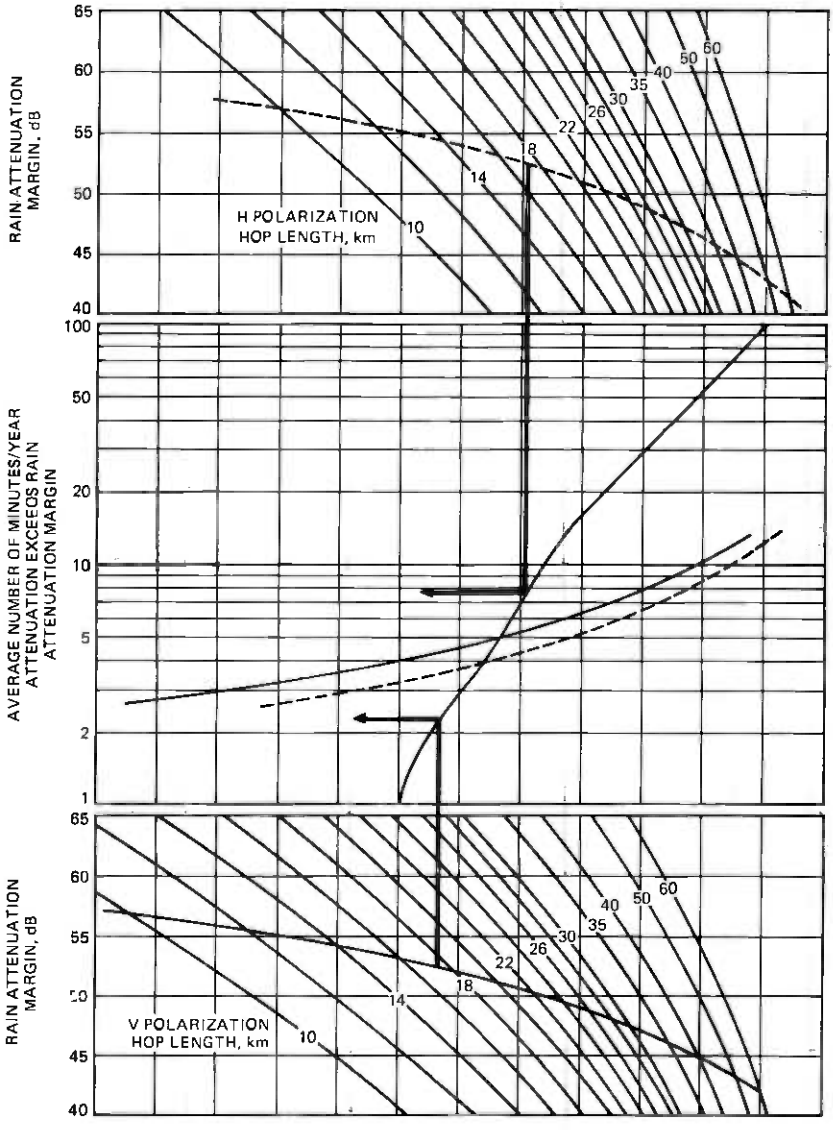


Fig. 6—Illustration of the vertical and horizontal polarization outage times for a system engineered to meet a weighted average outage time.

11-GHz radio systems using antennas with flat vertical radomes, a total loss of 4 dB for both antennas is normally assumed.[5] In digital radio systems using dual polarized frequency channels, depolarization by heavy rain can cause cochannel interference to degrade the system performance by about 2 dB.[5] Foreign system interference can cause a

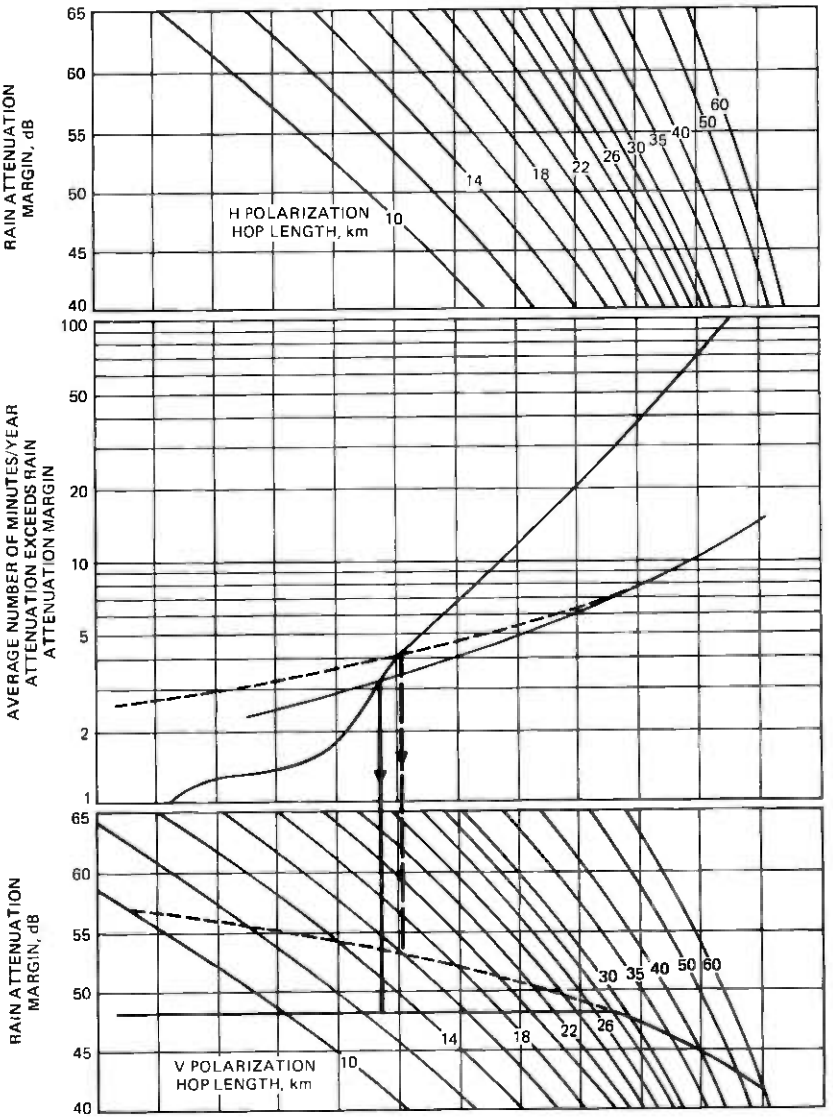


Fig. 7—Illustration of the effect of a rain attenuation margin which is limited by the radio receiver AGC range.

performance degradation during rain fading if the carrier-to-interference (C/I) ratio approaches the fade margin plus the system carrier-to-noise ratio at the performance threshold, and the interference does not fade with the desired signal. Normal frequency coordination practices require C/I ratios so high that this effect is negligible. However, if the desired

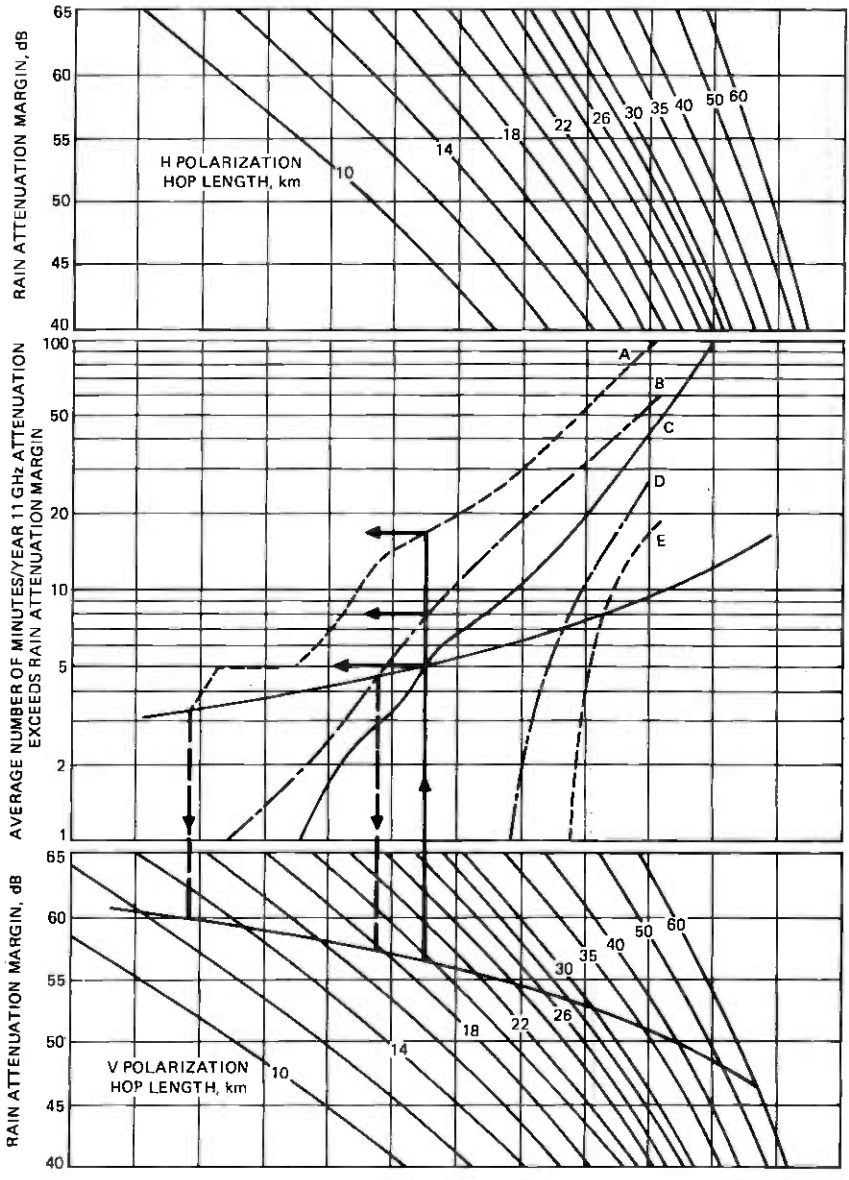


Fig. 8—Illustration of the outage times based on a 20-year average, a maximum 5-year average, and a maximum 1-year average outage time; and of the maximum hop lengths which will meet an objective for the maximum 5-year average and maximum 1-year outage times.

C/I ratio cannot be achieved, then the reduced C/I ratio can be tolerated by engineering with a reduced rain attenuation margin.

Table I shows an example calculation of the rain attenuation margin

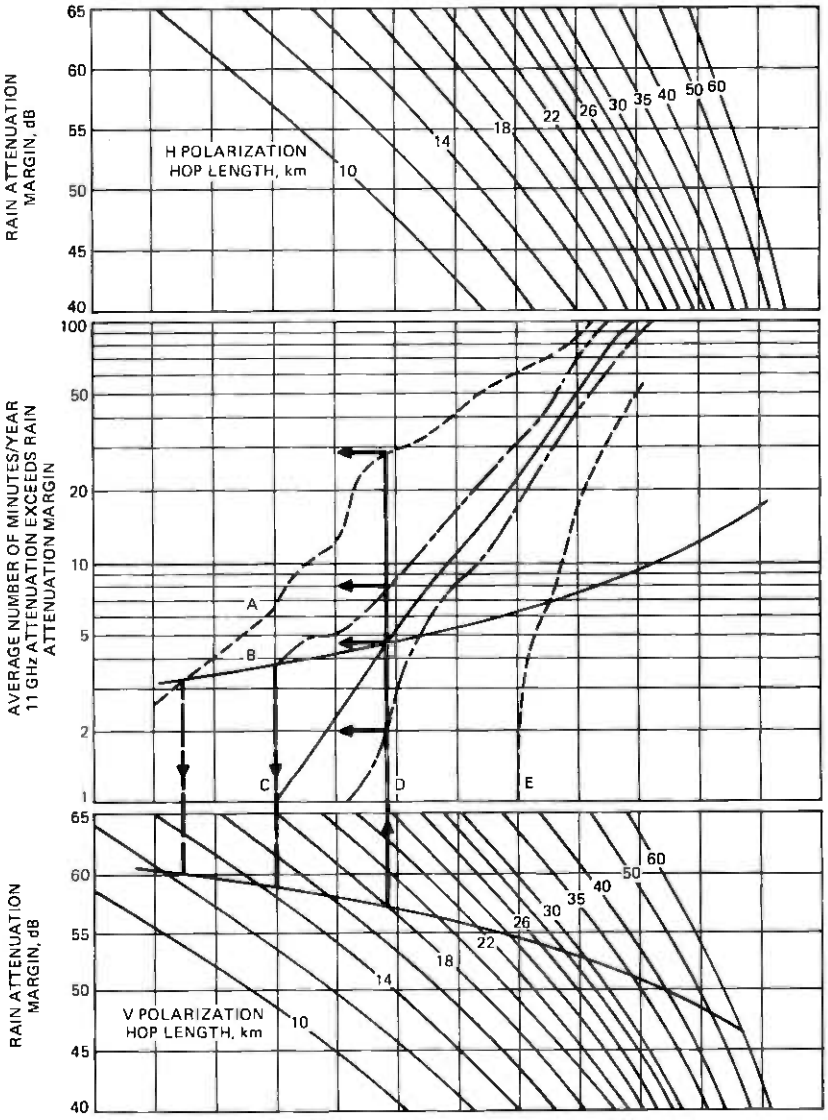


Fig. 9—Illustration of the outage times based on a 20-year average, a maximum 5-year average, and a maximum 1-year average outage time; and of the maximum hop lengths which will meet an objective for the maximum 5-year average and maximum 1-year outage times.

for a typical 11-GHz digital radio system path with a length of 40 km. For this example the rain attenuation margin is 50 dB.

Since the rain attenuation margin for a given type of equipment depends on the components of the section loss, these components can be

Table I — Calculation of rain attenuation margin (example)

Parameters	Decibels
System gain at 10^{-3} BER, G_s	112.0
Waveguide loss, total	6.6
Free-space path loss, 40-km path	145.6
Antenna gain, total for two 10-foot dish antennas	96.6
Section loss, L_s	55.6
Total fade margin available for rain fading ($M_T = G_s - L_s$)	56.4
Wet radome loss, L_r	4.0
Depolarization performance degradation, L_{XPD}	2.0
Foreign system interference degradation, L_{FI}	0.4
Rain attenuation margin, ($M_R = M_T - L_r - L_{XPD} - L_{FI}$)	50.0

chosen to give the optimum reliability versus economic tradeoff for each path. For example, if a path is constrained to be short because of terrain or the need for dropping a channel, then cost can be reduced by using less expensive but more lossy waveguide, or smaller antennas.

However, as is shown in Section 3.4, the allowable path length is quite sensitive to the rain attenuation margin. Because radio repeater site costs are so large in comparison to waveguide and antenna costs, it is usually least expensive to engineer for the longest path possible.

3.2 Determination of per-hop outage from rain attenuation margin

Once the rain attenuation margin and hop length are known, the expected number of minutes per year the hop performance will be below the performance threshold can be read directly from the rain chart. Figure 3 illustrates the use of the chart for the system used in the previous example. The lower scale is used for the outage on the vertical polarization and the upper scale for the outage on the horizontal polarization. For this example, the vertical polarization outage time, T_V, is 11 minutes and the horizontal polarization outage time, T_H, is 19 minutes.

The difference between vertical and horizontal polarization outage times is caused by the differential attenuation between the two polarizations. For 50 dB of attenuation on the vertical polarization, the differential attenuation ranges from 8.0 dB to 8.5 dB for hop lengths from 60 km to 10 km as illustrated in Fig. 4. The differential attenuation is relatively insensitive to hop length but varies substantially with absolute attenuation as shown in Hogg et al.[3]

A short-haul radio annual outage objective of 0.02 percent for a 250-mile system is often used in the Bell System. This amounts to 105 minutes per year for a 400-km system, or 10.5 minutes for a 40-km hop. Thus the expected annual outage for the vertical polarization is just over the objective, while the outage for the horizontal polarization is substantially higher than the objective.

In medium to highly loaded systems, both polarizations must be used

and the hops must be engineered to give adequate reliability with both polarizations utilized. Because some form of polarization frogging can usually be used to average the outage of both polarizations on any one trunk, and because it is almost always possible to use the vertical polarization more than the horizontal, the hop outage is usually taken to be the weighted average outage, weighted 60 percent for the vertical and 40 percent for the horizontal polarization. In this case the weighted average outage is $T_{AV} = 0.6\,T_V + 0.4\,T_H = 14$ minutes, which is still above the objective.

In order to meet the objective the rain attenuation margin must be increased. This can be done by changing the section loss or by shortening the hop length. From this point on, the procedure is by trial and retrial. The next section describes a method for finding the maximum allowable hop length by graphical construction.

The allocation of an objective on a per-hop basis by prorating the route objective on the basis of hop length implicitly assumes that the fading events on each hop are mutually exclusive. Since some simultaneous fading of adjacent or nearly adjacent hops is expected, this procedure leads to pessimistic estimates of the total outage time for tandem hops. However, there is as yet no adequate data for engineering otherwise.

3.3 Determination of allowable hop length to meet an objective

It is often desirable to be able to determine the maximum allowable hop length, for a given set of system parameters and location, for which the expected outage just equals the objective, without doing it by trial and error. This section describes a graphical procedure for determining the allowable hop lengths for the individual polarizations and approximately for the weighted average.

As mentioned in the previous section, it order to reduce the outage time the rain attenuation margin must be increased either by changing the equipment parameters or by changing the hop length or perhaps both. As can be seen from Fig. 3 by using the vertical polarization scale, increasing the fade margin by 3 dB by changing equipment parameters but keeping the hop length at 40 km reduces the outage time from 11 minutes to 8.5 minutes. This is not a significant reduction considering the difficulty involved in gaining 3 dB of margin by changing the waveguide loss or antenna size. The more effective way of increasing the margin is to decrease the hop length because changes in both margin and hop length act together to decrease the outage time. Therefore, in cases where the hops are rain-limited, the best procedure is to use the best practical system parameters and adjust the outage by changing the hop length.

If the system parameters are fixed, the rain attenuation margin varies

with hop length according to the equation

$$M_R = M_{Ro} + 20 \log \frac{D_o}{D} \tag{4}$$

where M_{Ro} is the rain attenuation margin on a hop length D_o. The values of M_{Ro} and D_o therefore become a measure of the equipment performance of a hop, dependent on how that particular equipment has been engineered. In the previous examples, we have shown that $M_{Ro} = 50$ dB on a 40-km hop is typical of the Western Electric 3A-RDS radio system. A plot of eq. (4) on the vertical or horizontal scales of a rain chart is called the *system characteristic curve* for the system with M_{Ro} at D_o. Figure 5 shows such curves for $M_{Ro} = 50$ dB at $D_o = 40$ km. The system characteristic displays the rain outage time as a function of hop length by reading vertically upward from the system characteristic at the desired hop length to the rain outage curve.

The outage objective for a given hop length, assuming the objective is prorated proportionally to the hop length, is

$$T_{\text{OBJ/HOP}} = T_{\text{OBJ/ROUTE}} \left(\frac{D}{\text{route length}} \right) \tag{5}$$

Using an objective of 0.02 percent per 400-km route, eq. (4) becomes

$$T_{\text{OBJ/HOP}} = .26\ D \text{ minutes} \tag{6}$$

when D is in km. If eq. (6) is plotted on the outage time scale of a rain chart as a function of hop length D *along the system characteristic* on either the vertical or horizontal scales, it becomes the *system objective* curve for the system. Figure 5 shows system objective curves for both polarizations and a system with $M_{Ro} = 50$ dB at 40 km.

Since the rain outage curve relative to the system characteristic curve is the hop outage as a function of hop length, and the system objective curve relative to the system characteristic curve is the hop objective, the maximum hop length, outage time, and corresponding fade margin can be found from the intersection of the two curves. Thus from Fig. 5, the values listed in Table IIA are found for the example system.

The maximum allowable hop length for which the weighted average outage time ($0.6\ T_V + 0.4\ T_H$) meets the objective can be found approximately by taking a weighted average of the vertical and horizontal allowable hop lengths:

$$D_{\text{MAX-AVG}} \approx 0.4\ D_{\text{MAX-V}} + 0.6\ D_{\text{MAX-H}} \tag{7}$$

Note that the weighting is reversed because the higher outage of the horizontal polarization contributes more to the weighted average outage even with the 60–40 weighting. Once the maximum hop length has been calculated, the corresponding fade margin can be read from the system

Table II — Allowable hop length calculations*

A. Calculations for V and H polarizations

	Polarization	
	V	H
Maximum hop length	24 km (15.0 mi)	18 km (11.3 mi)
Rain outage time	6.2 min	4.7 min
Rain attenuation margin required	54.5 dB	57.0 dB

B. Calculations for weighted average outage

Approximate maximum hop length [eq. (7)]	20.4 km (12.8 mi)
Rain attenuation margin [eq. (4)]	55.9 dB
Vertical polarization outage	2.5 min
Horizontal polarization outage	9.5 min
Weighted average outage	5.3 min
Objective for 20.4-km hop	5.3 min

* For a system with a rain attenuation margin of 50 dB on a 40-km hop using the rain outage chart in Fig. 5.

characteristic or calculated from (4), and the horizontal and vertical outage times can be read from the rain outage curve. Continuing the previous example gives the results listed in Table IIB, which are indicated in Fig. 5.

3.4 Sensitivity of allowable hop length to rain attenuation margin

Different types of radio systems have different attainable rain attenuation margins not only because the waveguide losses or antenna gains are different, but also because of inherent system performance capabilities. In this section we illustrate the sensitivity of the maximum allowable hop length to the rain attenuation margin by comparing the maximum allowable hop lengths of systems with different rain attenuation margins on a 40-km hop.

Table III and Fig. 6 show the results of calculations paralleling those of Section 3.3 but for a system with M_{Ro} = 45 dB on a 40-km hop. For both systems, the weighted average outage is equal to the outage objective.

The results show that a 5-dB decrease in rain attenuation margin requires a 16 percent decrease in hop length relative to the average of the two hop lengths. Calculations for other values of fade margins and other locations have shown that for fade margins ranging from 40 to 55 dB on a 40-km hop and for all cities where the maximum allowable hop lengths are 50 miles or less, a 5-dB difference in rain attenuation margin results in a difference of 16 to 18 percent in allowable hop length, and a 10-dB difference gives a difference of 30 to 35 percent in allowable hop length. Thus relatively small differences in margins can give substantial savings in system costs by reducing the number of repeaters required.

3.5 Effect of dynamic-range limited rain attenuation margin

Reducing the hop lengths decreases the rain outage time for two reasons: because the rain attenuation margin increases, and because the amount of rain attenuation incurred decreases. In order to actually realize that part of the decrease due to the rain attenuation margin increase, the AGC range of the radio system receiver must be adequate. If the AGC range is inadequate, it will be unable to maintain a constant signal level in the receiver and an outage will be caused by loss of signal level rather than degraded signal-to-noise ratio. In such cases the system characteristic curve does not show a continual increase in margin as the hops are shortened, but remains constant at some limiting value.

Figure 7 shows an example of a system with 45 dB of rain attenuation margin on a 40-km hop, but with a maximum rain attenuation margin of 48 dB, limited by the receiver AGC range. Figure 7 uses the vertical polarization only, but the same principles apply for both horizontal polarization outages and weighted-average outages.

The limited AGC range decreases the maximum allowable hop length substantially. For example, in Fig. 7 if the AGC range were not limited, the maximum allowable hop length would be 15 km (10 miles) and a rain attenuation margin of 53 dB would be required. With the rain attenuation margin limited to 48 dB, the maximum allowable hop length is 12.5 km (7.8 miles), or a decrease of 22 percent relative to 16 km. This would require a 28 percent increase in the number of repeaters.

This effect is much more substantial in the Southeastern U. S. where the hops must be short with correspondingly large margins required.

IV. EFFECTS OF THE VARIATIONS IN THE ANNUAL OUTAGE TIMES

Lin[4] has discussed the variability of the rain rate distributions from year to year and has emphasized the need for stable statistics on which to engineer radio systems. In this section we discuss the implications of this variability on the reliability of systems engineered by the proposed methods and show the penalty for engineering for worst case statistics.

4.1 Estimate of variability of annual outage times

Figures 8 and 9 show rain charts with the worst (A) and best (E) annual distributions, and worst (B) and best (D) 5-year average distributions, in addition to the 20-year average distribution (C). Figures 8 and 9 also show the system characteristic and system objective curves for the example system with 50-dB rain attenuation margin on a 40-km hop. For simplicity, only the vertical polarization will be considered; similar conclusions would be drawn for engineering based on the horizontal polarization outage or the weighted-average outage.

Table III — Allowable hop length calculations*

A. Calculations for V and H polarizations

	Polarization	
	V	H
Maximum hop length	20 km (12.5 mi)	15.5 km (9.7 mi)
Rain outage time	5.2 min	4.0 min
Rain attenuation margin required	51.0 dB	53.2

B. Calculations for weighted average outage

Approximate maximum hop length [Eq. (7)]	17.3 km (10.8 mi)
Rain attenuation margin [Eq. (4)]	52.3 dB
Vertical polarization outage	2.3 min
Horizontal polarization outage	7.8 min
Weighted average outage	4.5 min
Objective for 17.3-km hop	4.5 min

* For a system with a rain attenuation margin of 45 dB on a 40-km hop using the rain outage chart in Fig. 6.

First assume that the system has been engineered for the maximum allowable hop length for which the outage on the vertical polarization will meet a 0.02 percent per 400 km objective. The resulting hop length in Fig. 8 is 19.5 km (12.2 mi) and the average annual outage is 5 minutes per year based on the 20-year average distribution.

The curves shown on the rain outage chart are actually those distributions which were measured over the 20-year base period, 1953 to 1972. Thus the outage times read from the rain chart are those outage times which would have been measured if a system had been operating during the 20-year base period (assuming the rain theory is correct); but they are probably *not* the outage times that will be measured in the next or any other 20-year period. They are, however, the *best estimate* of what similarly averaged outages would be for any 20-year period. Furthermore, the outage time indicated by the 20-year average curve is the *best estimate* of what the annual outage time will be in any one year although we know that it probably will not be that value.

The annual outage times indicated by curves A and E in these figures give some indication of the extreme values that can be expected over a 20-year period. In Fig. 8, the largest outage time is about 17 minutes, a little over 3 times the design value; the smallest outage time is much less than one minute. The largest annual outage time averaged over any 5-year period is expected to be about 8 minutes; and again the smallest is much less than 1 minute. Similar results are obtained from Fig. 9.

A similar ananysis was done for each of 13 representative cities including those in Figs. 3 to 9 and the results are listed in Table IV. Table IV is divided into two parts. For those cities listed in part B the allowable hop lengths are so long, and the corresponding rain rates so low, that meaningful short-term distributions at high rain rates could not be

Table IV Factors by which the 20-year average outage time on the vertical polarization is exceeded*

Rain Outage Chart	Hop length, km	Hop length, miles	20-yr averge outage time, minutes	Factor by which outage time exceeds 20-yr average outage time: 1-yr max	5-yr max	5-yr min	1-yr min
			(A)				
Fig. 4	32.3	20.2	8.4	2.6	1.3	0.6	†
Figs. 5, 6	23.9	14.9	6.2	7.1	1.9	0.3	0.0
Fig. 7	19.2	12.0	5.0	4.8	2.4	0.1	0.0
Fig. 8	19.5	12.2	5.1	3.2	1.6	0.0	0.0
Fig. 9	17.7	11.1	4.6	6.3	1.7	0.5	0.0
(Not shown)	30.8	19.2	8.0	5.0	1.8	0.5	†
(Not shown)	23.1	14.4	6.0	2.5	1.7	0.6	0.0
			(B)				
Fig. 3	36.2	22.6	8.2	3.9	2.3	0.3	†
(Not shown)	36.2	22.6	5.9	5.8	1.8	0.5	†
(Not shown)	36.2	22.6	3.7	4.1	1.8	†	†
(Not shown)	36.2	22.6	2.0	10.0	2.2	†	†
(Not shown)	36.2	22.6	2.8	7.1	2.1	†	†
(Not shown)	36.2	22.6	1.4	7.1	3.4	†	†

* By the maximum and minimum 1-year outage times and 5-year average outage times. Part A uses representative cities for which the hop length listed is the maximum allowable to meet the outage objective. Part B uses representative cities for which the hop length is shorter than the maximum because data was not available at the maximum allowable hop length. The outage time allowable at the 36.2-km hop length is 9.4 minutes. A system with 50-dB margin on a 40-km hop is assumed.

† Data not available.

generated. Consequently, the calculations were made at the longest hop length for which data was available—36.2 km.

The data in Table IV show that for the midcontinent cities in part A the factors by which the maximum 1-year outages exceed the engineered value range from 2.5 to 7.1. Factors by which the maximum 5-year average annual outages exceed the engineered value range from 1.3 to 2.4. At every location there should be at least 1 year out of 20 for which there is no outage. The variability in the outage for the cities in part B is slightly more, the maximum 1-year factors ranging from 3.9 to 10, and the maximum 5-year factors ranging from 1.8 to 3.4.

The question of whether a hop or a route is performing as engineered inevitably arises. Two additional factors which affect the observed outages of a route must be considered. First, as demonstrated by Lin,[4] the outage of a route consisting of several hops should not be as variable as the individual hops themselves. Lin's Fig. 16 shows roughly a factor-of-2 reduction in his $\Delta t/\bar{t}$ factor, which is equivalent to the factor listed for the 1-year maximum in Table IV,* for a route consisting of three

* In Lin's paper Δt is the worst-year minutes less the best-year minutes. However, the best-year minutes are negligible, so the ratio is essentially worst year minutes divided by the 20-year average which is the same as used in Table IV.

hops. Secondly, the route outage should not be as large as the sum of the individual hop outage because of joint fading on tandem hops. This effect should be more important in the midcontinent cities where the hops are shorter.

Based on the foregoing, the following guidelines seem reasonable. First it must be definitely established that the outage in question is caused by *aerial attenuation by rain.* Then, if the route outage time of a route containing three or more hops exceeds the engineered value by more than a factor of 5 in any one year, or by a factor of 2 for any 5-year average, the rain outage is excessive and the reason for the excessive outage should be determined. If the outage time of a single hop or two tandem hops exceeds the engineered value by more than a factor of 10 in any one year, or a factor of 4 for any 5-year average, the reason for the excessive outage should be determined.

4.2 Engineering for worst-case outages

To avoid exceeding the outage objective for any one year, or for any 5-year average, would require that the hops be engineered so that the objective is met for the estimated maximum 1- and 5-year average annual outages respectively. Figures 8 and 9 illustrate the procedure for the example system using only the vertical polarization outage.

In Fig. 8 the hop length is 19.5 km (12.2 miles) based on the 20-year average, 17.7 km (11.1 miles) based on the maximum 5-year average annual outage time, and 12.9 km, (8.1 miles) based on the maximum 1-year annual outage time. These hop lengths are 9.2 and 33.9 percent reductions in hop length, which in turn mean 10.1 and 51.3 percent increases in the number of repeaters, respectively.

Table V lists similar percentages for eight Eastern and Midwestern cities. Such comparisons are not meaningful for the far Western cities because the allowable hop lengths based on 20-year average outage times are much longer than are used in practice. (In other words, the hops are not rain-attenuation limited.) Table V shows that the percentage increase in number of repeaters ranges from 9 to 27 percent if the hop lengths are based on maximum 5-year average outage times, and from 25 to 77 percent if the hop lengths are based on the maximum 1-year outage times. Because radio repeaters are so expensive, such increases in the number of repeaters could make rain-attenuation-limited radio systems very uneconomical.

V. GEOGRAPHICAL COVERAGE

Although charts have been produced for many cities, there are still areas a few hundred miles on a side for which no rain data exists, and so the problem of how to engineer radio systems in these areas still exists.

Table V — Percentage decrease in allowable hop lengths and resulting percentage increase in number of repeaters*

Rain outage chart	Hop length for 20-year average		Percentage decrease in hop length for hops engineered to indicated distributions		Percentage increase in number of repeaters for hops engineered to indicated distributions	
	km	miles	5-yr max	1-yr max	5-yr max	1-yr max
Fig. 3	36.5	22.8	21.1	39.0	26.7	63.8
Fig. 4	32.3	20.2	8.3	40.5	9.1	68.0
Figs. 5, 6	23.9	14.9	11.3	43.6	12.7	77.1
Fig. 7	19.2	12.0	—	20.0	—	25.0
Fig. 8	19.5	12.2	9.2	33.9	10.1	51.3
Fig. 9	17.7	11.1	19.8	30.5	24.7	43.9
(Not shown)	30.8	19.2	16.3	32.5	19.4	48.2
(Not shown)	23.1	14.4	18.1	33.4	21.7	50.2

* Resulting from engineering the maximum 1- and 5-year average outage times to meet a 0.02 percent per 400-km objective.

At the present time there is no definitive proven solution to this problem, but the following approaches seem reasonable.

The problem can be approached in two basic ways: interpolation between locations where data exists, or identification of the unknown location with a known location based on consideration of climates and local judgment. Usually some combination of these two approaches is the most satisfying intuitively.

There is no reason to suppose that other than linear interpolation should be used. Linear interpolation can be used by calculating the outage times for a given system at different locations and interpolating between them, or by calculating the allowable hop lengths and interpolating them. The main advantage interpolation has over judgment is that it is consistent and reproducible.

In using judgment of climatological conditions it is of the utmost importance to remember that it is the *rainfall rate* that determines outage time and *not* the total amount of water that falls. The northwest coast of the United States is the primary example of a very wet region where there is virtually no rain-attenuation-caused outage. Large scale climatological factors which seem to bear some relation to high rain rates are number of thunderstorms, late summer humidity, and total July precipitation. These are probably related because most of the rain rates which are large enough to cause an outage are due to thunderstorms. For example, total July precipitation is related to thunderstorms because in July most of the precipitation is from thunderstorms. Terrain should also be considered, especially in the lee of mountains, because rough terrain and mountains contribute to the formation of thunderstorms.

Finally, local knowledge and judgment should be used in comparing the area in question to a location where data is available.

VI. ACKNOWLEDGMENT

The author wishes to acknowledge A. Hamori as the originator of the idea of using a two-part rain chart to relate fade margin and hop length to outage time.

REFERENCES

1. S. D. Hathaway and H. W. Evans, "Rain Attenuation at 11 kmc," B.S.T.J., *38*, No. 1 (January 1959), pp. 73–97.
2. R. G. Medhurst, "Rainfall Attenuation of Centimeter Waves," IEEE Trans. Ant. Propag., *AP-13*, July 1965, pp. 550–563.
3. D. C. Hogg, A. J. Giger, A. C. Longton, and E. E. Muller, "The Influence of Rain on Design of 11 GHz Terrestial Radio Relay," B.S.T.J., this issue, pp. 1575–1580.
4. S. H. Lin, "Nationwide Long-Term Rain Rate Statistics and Empirical Calculation of 11-GHz Microwave Rain Attenuation," B.S.T.J., this issue, pp. 1581–1604.
5. A. J. Giger and T. L. Osborne, "3A-RDS 11 GHz Digital Radio System," paper No. 18.1, Digest of International Conference on Communications, 1976.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Optimum Digital Filters for Interpolative A/D Converters

By A. N. NETRAVALI

(Manuscript received March 31, 1977)

Interpolative analog-to-digital (A/D) converters allow a fine representation of signals by making many coarse representations and averaging them using a digital filter. In this paper, we give a method of optimizing the characteristics of this digital filter under two different criteria. The first criterion is the well-known signal-to-noise (S/N) ratio, whereas the second criterion is the weighted sum of the signal power, the quantization noise power, and the noise power within a given band of frequencies. We design optimum digital filters and simulate their performance on the computer. We show that the theoretically predicted S/N ratio is in good agreement with the performance obtained by computer simulation. It is seen that about 23 dB improvement in S/N ratio over the S/N ratio attainable by a constant-weight digital filter is possible when the number of coarse quantizations is 256. We also study the effects of changing various parameters of the A/D converter on the S/N ratio.

I. INTRODUCTION

Interpolative A/D converters[1–3] achieve a fine quantization of signals by making several coarse quantizations and averaging them. This requires high-speed operation of that part of the A/D converter which obtains the coarse quantizations. Higher and higher speeds are required for finer and finer ultimate quantization. This trade-off between the speed of operation and amplitude resolution is particularly relevant and important with present-day integrated circuit technology, which provides high-speed operation but no high-amplitude precision.

Several well known methods of obtaining the many coarse quantizations exist. Goodman,[1] and Goodman and Greenstein,[2] have considered the ordinary delta modulator which gives a two-level representation of the signal at a rate many times higher than the Nyquist rate. The output of the delta modulator is filtered by a digital filter and resampled at

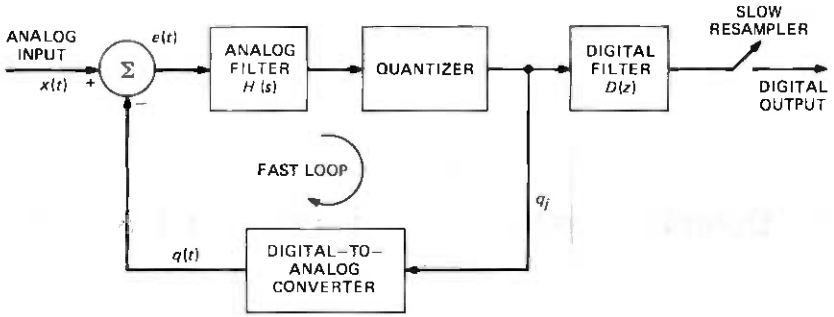


Fig. 1—An interpolative A/D converter.

Nyquist rate to obtain the PCM output. The performance of such an A/D converter depends upon the speed of the delta modulator and the characteristics of the digital filter.

Another method of obtaining the coarse quantization has been proposed recently by Candy.[3] In this method, the coarse quantizations are obtained by a direct feedback encoder shown in Fig. 1. In this encoder a difference between the analog input and a coarsely quantized representation is filtered by an analog filter with characteristics $H(s)$, and then quantized. This is done at speeds higher than the Nyquist rate. The output of the quantizer is represented by binary words and filtered by a digital filter having characteristics $D(z)$. The output of the digital filter is resampled at a slower rate to obtain the final digital output at the Nyquist rate. Use of direct feedback encoding allows shaping of the quantization noise in such a way that the digital filter can be made very simple. Candy[3] has shown that when the analog filter is taken to be a pure integrator the simple digital filter corresponding to "accumulate-and-dump" performs adequately.

Candy et al.[4] have described a method of optimizing the weights of the digital filter when the analog filter in the "fast loop" is a pure integrator. They have shown that the optimum weights can be approximated by a set of triangularly distributed weights and evaluated the improvements in S/N ratio by using these weights. Their results are applicable only when the integrator in the "fast loop" is not reset to zero at the beginning of each slow cycle. In this paper, we first show that when the analog filter in the "fast loop" is reset to zero at the beginning of every slow cycle an advantage in the S/N ratio is obtained when the uniformly distributed weights are used. We then give a different method of optimizing the digital filter characteristics under the assumption that the analog filter is reset. Our method of optimization is applicable to the case of any arbitrary analog filter in the place of the integrator in the fast loop. The resulting optimum weights when the integrator is reset every slow cycle have a different shape than the optimum weights given by Candy[4]

which are applicable when the integrator is not reset. We compare our optimum weights with the triangular weights proposed by Candy as an approximation to his optimum weights. Optimization of the digital filters using a different criterion, which includes a deviation of the digital filter characteristics from desired characteristics is also discussed. In this case it is possible, for example, to shape the discrete Fourier transform of the digital filter weights so that it resembles, as far as possible, an ideal low-pass filter. We evaluate the performance of the A/D converter in terms of S/N ratio by computer simulation for several typical cases.

II. SUMMARY OF RESULTS

Our computer simulations indicate that there is about 3 dB improvement in S/N ratio by resetting the integrator at the beginning of each slow cycle when uniform weights are used for the digital filters. This improvement is independent of the coarseness of the quantizer in the fast loop, the number of fast cycles and the correlations present in the input signal. The use of optimum weights for the digital filter leads to significant improvements in S/N ratio over that obtained by a digital filter with constant weights. This improvement although independent of the coarseness of the quantizers depends on the number of fast cycles; for 32 fast cycles, there is about a 14 dB improvement, whereas for 256 fast cycles, there is a 23 dB improvement. Also, the optimum weights outperform the "triangular" weights used by Candy et al.[4] by about 7.30 dB when the number of fast cycles is 32 and by about 8.80 dB when the number of fast cycles is 256. We also show that there is a good agreement between the theoretically predicted S/N ratio and that obtained from computer simulations of the A/D converter. Changing the analog filter from an integrator to a general analog filter with a given characteristic indicates that there is a gain of a few dB in S/N ratio by choosing the dc gain and the cutoff frequencies judiciously. Our second method of optimizing the digital filter characteristic allows us to minimize the deviation of its frequency characteristics from a given characteristic. Using the desired characteristic to be ideal low-pass, we are able to decrease the noise power in a given band of frequencies. This decrease is about 0.5 to 1.0 dB, but it comes at the expense of an increase in the overall noise power of about 1.0 to 1.5 dB. Thus the digital filter suppresses the noise power in one band of frequencies, but enhances the noise in the rest of the frequency band, resulting consequently in an overall increase in the noise power.

III. DERIVATION OF OPTIMUM DIGITAL FILTER WEIGHTS

In this section, we derive the weights of the optimum digital filter. First we concern ourselves with those digital filters which minimize the S/N ratio, and then derive those weights which can be spectrally shaped.

Let $x(t)$ be the analog input to the A/D converter shown in Fig. 1. Also let $h(\cdot)$ be the impulse response of the time-invariant analog filter in the fast loop; N, the number of fast cycles; T, the fast sampling period; and q_j, the output of the quantizer at the jth fast cycle. We assume that the output of the digital-to-analog converter is given by

$$q(t) = q_j \quad jT \le t < (j+1)T \tag{1}$$

The equation for the fast loop can be written as:

$$\int_0^t h(\tau)[x(t-\tau) - q(t-\tau)]d\tau = q(t+T) + n(t+T) \tag{2}$$

Here we have assumed that the analog filter is reset at the beginning of each slow cycle and that the quantization distortion can be represented by additive random noise $n(\cdot)$. Assuming that $x(t)$ is constant ($=x$) over a slow cycle, then at $t = (i+1)T$,

$$x\int_0^{iT} h(t)dt - \int_0^{iT} h(\tau)q(iT-\tau)d\tau = q[(i+1)T] + n[(i+1)T] \tag{3}$$

Now letting

$$\int_{kT}^{(k+1)T} h(t)dt = h_k \tag{4}$$

eq. (3) can be written as:

$$x\sum_{k=0}^{i-1} h_k - \sum_{k=0}^{i-1} q_{i-k-1}h_k = q_i + n_i \qquad i = 1, \ldots, N \tag{5}$$

where

$$n_i = n(iT)$$

Equation (5) can be written for $i = 1, \ldots, N$, in a matrix form

$$x \cdot A = HQ + N_0 + q_0\begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{N-1} \end{bmatrix} \tag{6}$$

where

$$A = \text{col}\left(h_0, h_0 + h_1, \ldots, \sum_{i=0}^{N-1} h_i\right)$$

$$H = \begin{bmatrix} 1 & 0 & & 0 \\ h_0 & 1 & 0 & 0 \\ h_1 & h_0 & 1 & 0 \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ h_{N-2} & h_{N-3}\cdot\cdot & h_1 h_0 & 1 \end{bmatrix}$$

$$N_0 = \text{col}\,(n_1, n_2, \ldots, n_N)$$

$$Q = \text{col}\,(q_1, q_2, \ldots, q_N)$$

Observe that matrix H has an inverse and therefore eq. (6) can be rewritten as:

$$Q = xH^{-1}A - H^{-1}N_0 - q_0 H^{-1}\begin{bmatrix} h_0 \\ \vdots \\ h_{N-1} \end{bmatrix} \tag{7}$$

The digital filter will process vector Q every slow cycle by multiplying it by a weight vector D, and thus the PCM output will be

$$D^T Q = xD^T H^{-1}A - D^T H^{-1}N_0 - q_0 D^T H^{-1}\begin{bmatrix} h_0 \\ \vdots \\ h_{N-1} \end{bmatrix} \tag{8}$$

Here we assume that $D^T U = 1$, where $U = \text{col}\,(1, 1, \ldots, 1)$. The first term on the right-hand side of eq. (8) is the signal component, whereas the second term is the noise component. The third term results from the initial condition on the D/A, q_0. We assume $q_0 = 0$. In order to maximize the ratio of signal energy to the noise energy, we maximize the following expression:

$$(S/N) \triangleq (D^T G^{-1}A)^2 / \overline{(D^T H^{-1}N_0)^2} \tag{9}$$

where $(\bar{\cdot})$ denotes expectation. Assuming† that the noise components n_i are independent, identically distributed with variance σ^2, we can write eq. (9) as:

$$(S/N) = \frac{1}{\sigma^2}\,[D^T H^{-1}AA^T(H^{-1})^T D]/D^T H^{-1}(H^{-1})^T D \tag{10}$$

We note that since H has an inverse H^{-1}, $(H^{-1})^T$ is positive definite and therefore the denominator of the right hand side of eq. (10) will not be zero unless $D \equiv 0$, a case which we rule out. This implies that (S/N) will

† This assumption is not required. It is easy to extend the following analysis to colored noise.

be bounded from above. Since (S/N) is a ratio of two quadratic forms generated by two symmetric matrices, we can write the optimum D, D^*, as a solution of the eigenvalue problem

$$(H^{-1}A)(H^{-1}A)^T D^* = \lambda_{\max}(H^{-1})(H^{-1})^T D^* \tag{11}$$

or

$$H^T A A^T (H^{-1})^T D^* = \lambda_{\max} D^* \tag{12}$$

where $\lambda_{\max}$ is the maximum eigenvalue. It is easy to see that the only eigenvector for eq. (12) corresponding to a nonzero eigenvalue is given by

$$D^* = H^T A \tag{13}$$

for which

$$(S/N) = \frac{A^T A}{\sigma^2}$$

$$= \frac{1}{\sigma^2}\left[\sum_{i=1}^{N}\left(\sum_{k=0}^{i-1} h_k\right)^2\right] \tag{14}$$

Writing out H and A, we get

$$D^* = \begin{bmatrix} \sum_{j=-1}^{N-2} h_j \left(\sum_{k=0}^{j+1} h_k\right) \\ \vdots \\ \sum_{j=-1}^{N-3} h_j \left(\sum_{k=0}^{j+2} h_k\right) \\ \vdots \\ \sum_{j=1}^{N-(N+1)} h_j \left(\sum_{k=0}^{j+N} h_k\right) \end{bmatrix} \tag{15}$$

where we have assumed for notational convenience that $h_{-1} = 1$. If the filter in the fast loop is a pure integrator, then $h_k = T$, and the optimum digital filter can be written as:

$$D = \text{col}\,(D_1, \ldots, D_j, \ldots, D_N)$$

where

$$D_j = \frac{T(N-j+1)(N+j)}{2} \qquad j = 1, \ldots, N \tag{16a}$$

and for large N the S/N ratio is given (except for a proportionality constant) by

$$S/N = \frac{N(N+1)(2N+1)}{6} \tag{16b}$$

3.1 Optimum digital filter with spectral shaping

Let $D(\omega)$ be the discrete Fourier transform of the samples $\{D_k\}_{k=0,\ldots,N-1}$ and $C(\omega)$ be the transform of the desired response that is obtained from the filter weights $\{C_k\}_{k=0,\ldots,N-1}$. The shaping of the digital filter in the Fourier domain can thus be accomplished by proper choice of $C(\omega)$. We use the following expression for the error between the two:

$$\begin{aligned} E_{RR} &= \int |[D(\omega) - C(\omega)]|^2 \, d\omega \\ &= \int [D(\omega) - C(\omega)][D(\omega) - C(\omega)]^* \, d\omega \\ &= \int D(\omega)D^*(\omega) \, d\omega - \int D(\omega)C^*(\omega) \, d\omega \\ &\quad - \int D^*(\omega)C(\omega) \, d\omega + \int C(\omega)C^*(\omega) \, d\omega \end{aligned} \tag{17}$$

where $(\cdot)^*$ is the complex conjugate. In minimizing E_{RR} with respect to D, we can drop the third term of eq. (17) and rewrite (17) as

$$\begin{aligned} E_{RR} &= \sum_{k=0}^{N-1} D_k^2 - \int \left(\sum_{k=0}^{N-1} D_k e^{-j2\pi\omega k/N} \right) C^*(\omega) \, d\omega \\ &\quad - \int \left[\sum_{k=0}^{N-1} C_k e^{-j2\pi\omega k/N} \right] D^*(\omega) \, d\omega \\ &= \sum_{k=0}^{N-1} D_k^2 - 2 \sum_{k=0}^{N-1} D_k C_k \\ &= D^T D - 2D^T C \end{aligned} \tag{18}$$

The performance function (PF) that we want to maximize can be written as:

$$(\mathrm{PF}) = D^T H^{-1} A (H^{-1}A)^T D - \lambda D^T H^{-1} (H^{-1})^T D - \gamma (D^T D - 2D^T C) \tag{19}$$

where the first term on the right-hand side corresponds to signal energy, second term corresponds to noise energy and the last term is the E_{RR} from eq. (18), and λ and γ are positive constants. Equation (19) can be rewritten as:

$$(\mathrm{PF}) = D^T [H^{-1} A (H^{-1}A)^T - \lambda H^{-1} (H^{-1})^T - \gamma I] D + 2\gamma D^T C \tag{20}$$

The best D which maximizes (PF) is given by

$$D^* = 2\gamma \, [H^{-1} A (H^{-1}A)^T - \lambda H^{-1} (H^{-1})^T - \gamma I]^{-1} C \tag{21}$$

IV. RESULTS OF COMPUTER SIMULATION

In our computer simulations we used uniformly distributed pseudo-random noise as the input signal $x(t)$ to the A/D converter. This was held

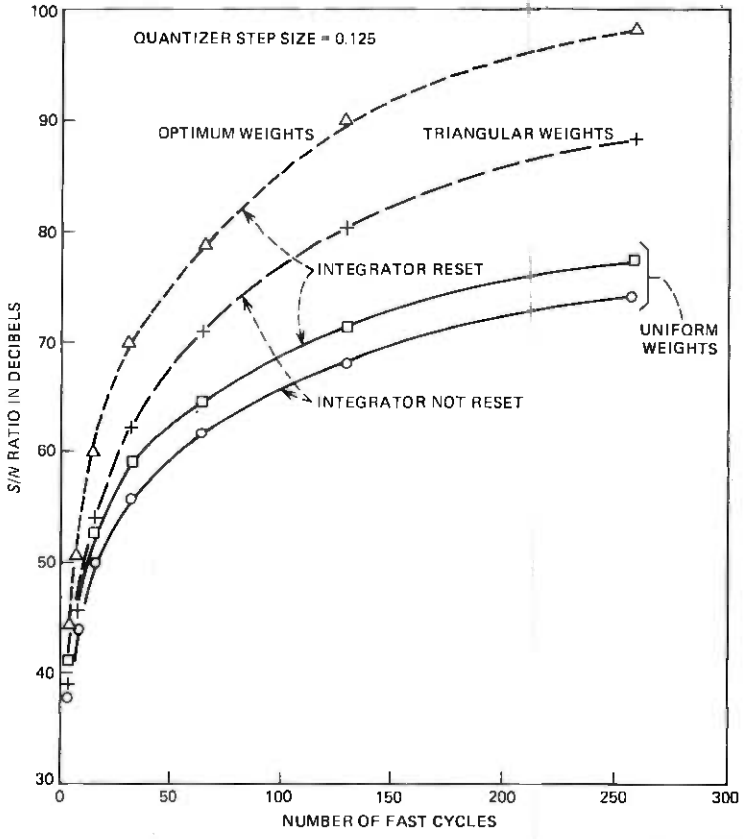


Fig. 2—Performance of an interpolative A/D converter with various digital filter weights.

constant throughout each slow cycle. We also considered cases when the input signal was filtered by an appropriate filter before going into the A/D converter. Simulations were carried out with the quantizer having two different step sizes, namely 0.125 and 0.0625 (signal range 0–1). The quantizer was assumed to have an unlimited number of levels and thus the effects of saturation were neglected. This assumption becomes more restrictive when the gain of the analog filter in the fast loop is increased. To evaluate the dependence of S/N on the number of fast cycles, several (4, 8, 16, . . . ,256) values of fast cycles were used. For the purpose of comparison, we also considered the following cases:

(*i*) Uniform weights, i.e., $D_j = 1, j = 1, \ldots, N$, with integrator not reset.

(*ii*) Triangular weights, i.e., $D_j = \min(j, N + 1 - j), j = 1, \ldots, N$, with integrator not reset. Both these weights have been investigated previously.[3,4]

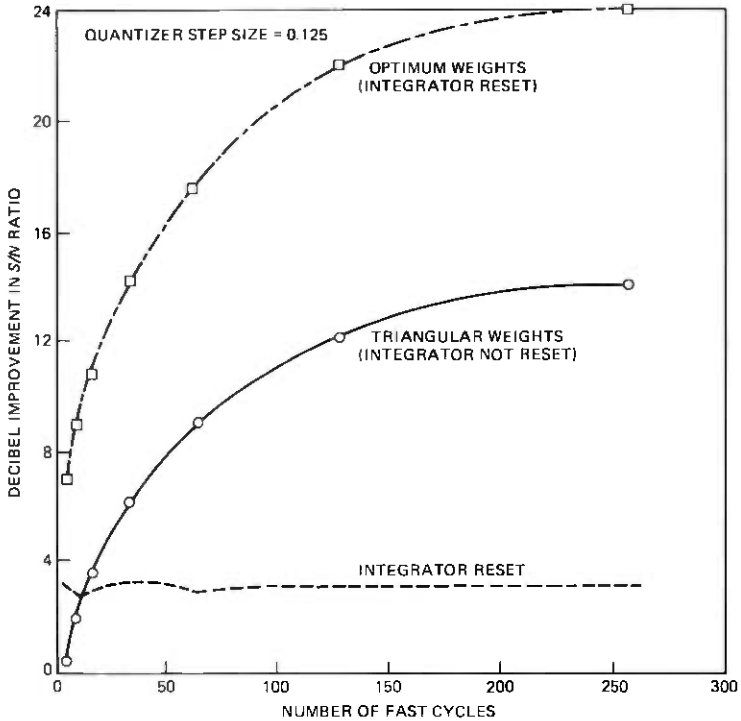


Fig. 3—Improvements in S/N ratio over constant weight digital filter.

4.1 Effect of integrator reset with uniform weights

The effect of resetting the integrator in the fast loop was evaluated by using uniform weights for two cases: (*i*) integrator reset, (*ii*) integrator not reset. The resulting S/N ratios are plotted in Fig. 2. The improvement in S/N ratio by resetting the integrator is plotted in Figs. 3 and 4, for two quantizer step sizes. It is seen that there is about 3 dB improvement by resetting the integrator, and this improvement is somewhat independent of the quantizer step size and the number of fast cycles. This can be easily explained by rewriting eq. (8), for $h_i = 1$ and $D_i = 1$, as:

$$\sum_{i=1}^{N} q_i = Nx - n_N - q_0 \tag{22}$$

Thus there is an extra term on the right-hand side, q_0, if there is no reset. Assuming that it is comparable to n_N, and that it is not correlated with n_N, the S/N ratio would decrease by about 3 dB due to its presence. We also simulated the effects of correlations in the input data, by filtering the pseudorandom noise, and then putting it through the A/D converter. Several low-pass filters were tried, and it was observed that the im-

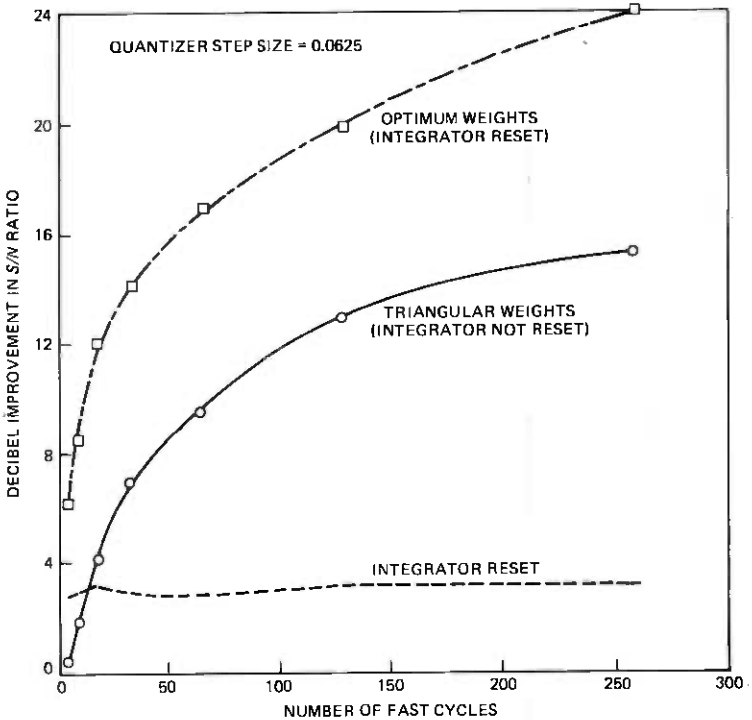


Fig. 4—Improvements in S/N ratio over constant weight digital filter.

provement in S/N ratio was still 3 dB regardless of the amount of low-pass filtering.

4.2 Effect of optimum digital filter weights

Figure 2 shows the effects of optimum digital filter weights on the S/N ratio. The analog filter in the loop is assumed to be a pure integrator with unity gain and it is reset at the beginning of each slow cycle. Figure 2 also shows the advantages of using the triangular weights, proposed by Candy et al., when the integrator is not reset. As observed by Candy et al., triangular weights are significantly better than the uniform weights, and the optimum weights allow a further increase in S/N ratio over the triangular weights. Figures 3 and 4 show the improvements in S/N ratio over those obtainable by the uniform weights when the integrator is not reset. It is seen that the rate of change of S/N ratio depends upon the number of fast cycles and is in close agreement with that predicted by eq. (16b). The S/N ratio using uniform weights when the integrator is not reset is given by (except for a proportionality constant)

$$S/N = \frac{N^2}{2} \tag{23a}$$

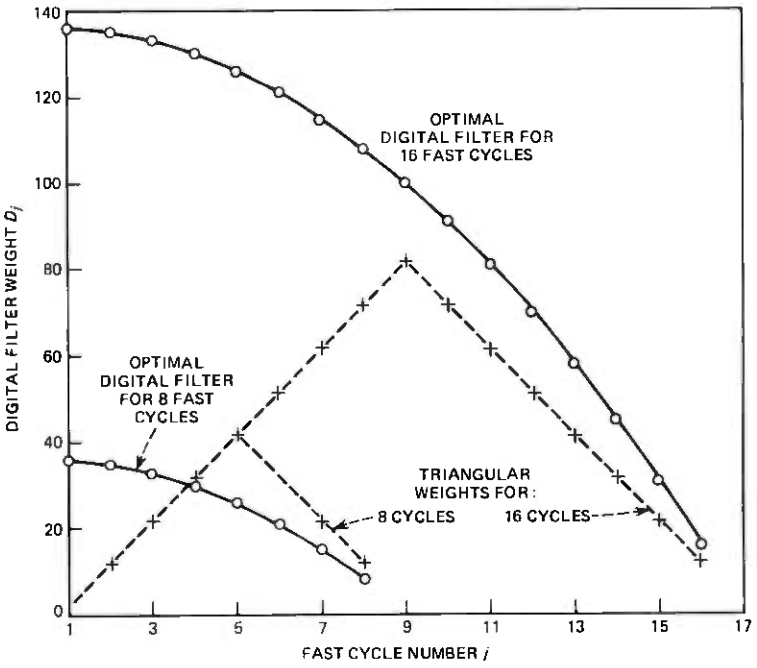


Fig. 5—Weights of optimum- and triangular-weight digital filters.

and with triangular weights

$$S/N = \frac{(N+1)^3}{16} \tag{23b}$$

These are derived by Candy et al.[4] Our simulations are in close agreement with the above equations. Thus when $N = 32$, the improvement in S/N ratio by using triangular weights over uniform weights is 6.55 dB, which is close to 6.40 dB predicted by the above equations. Similar agreement is found at other values of N. Also for large values of N the improvement obtained by our optimum weights, in the presence of integrator being reset, over the triangular weights without resetting the integrator is about 7.25 dB as predicted by eqs. (23b) and (16b). Our simulations indicate that this improvement varies between 5.80 and 10.0 dB with a mean of 7.57 dB. This is a little higher than that predicted by the equations, however the agreement is satisfactory.

The weights of the optimum digital filters are shown in a graphical form in Fig. 5 along with the triangular weights used by Candy et al.[4] It is seen that they have a parabolic shape. Although we have not considered the effect of approximations, for implementational simplicity, the filter shape could be approximated by piecewise straight lines.

4.3 Effect of variation of the analog filter in the fast loop

We attempted to evaluate the effect of varying the analog filter in the fast loop on the S/N ratio. It is known that certain types of analog filters tend to make the fast loop unstable; however, we did not consider questions of stability. Two types of transfer functions for the analog filter were considered:

$$H_1(s) = \frac{\overline{\alpha}}{s + \overline{\beta}}$$

and

$$H_2(s) = \frac{\overline{\alpha}}{s(s + \overline{\beta})}$$

The first case resulted in $h_i = \alpha e^{-i\beta}$ (using eq. 4) and the second case gave $h_i = \alpha - \sigma e^{-i\beta}$, where constants α, β, and σ are related to $\overline{\alpha}$, $\overline{\beta}$. Several simulations were run by varying α, β, and σ. For each of these simulations, the optimum weights were computed by eq. (15), and the resulting S/N ratio was compared with that obtained by using a pure integrator in the fast loop and the optimum digital filter. We considered only 32 fast cycles and a quantizer step size of .0625. In the first case, it was found that larger α and smaller β generally gave better S/N ratio. At $\alpha = 1.2$ and $\beta = 0.01$, the improvement in S/N ratio was about 3.0 dB. For many other cases studied, the improvement was somewhat marginal. For the second case, again, larger values of α, smaller values of β and σ around 1.0 gave the best results. At $\alpha = 1.8$, $\beta = 0.01$, and $\sigma = 1.0$, the improvement in S/N ratio was about 4.2 dB over that obtained by pure integrator in the fast loop. Thus it appears that S/N ratio can be further improved by a proper choice of the analog filter in the loop.

4.4 Effect of spectrally shaped digital filters

Our final simulations used digital filters which resemble a given digital filter as far as possible. For our simulation we obtained the desired digital filter characteristics from an analog function $C(t)$ whose Fourier transform $C(f)$ was 0 outside $|f| > \Omega$ and was constant (=Mag) in the interval $|f| \leq \Omega$. Sampling such a function at N times the Nyquist rate (corresponding to the number of fast cycles) gave

$$C_i = C(i/2\Omega N)$$

$$= \frac{2 \cdot \text{Mag} \cdot \Omega \cdot \text{N}}{\pi i} \sin(i\pi/N) \qquad i = 0, \ldots, N-1$$

Using these weights for the desired filter characteristics and some values of λ, γ (of Section 3.1), optimum digital filters with spectral shaping were obtained for the case when the analog filter in the fast loop was a pure

integrator. Computer simulations were carried out for various values of λ and γ, $N = 32$, and quantizer step size = .0625. Two quantities were measured: (*i*) S/N ratio as before, (*ii*) the noise power in frequency band $-\Omega$ to $+\Omega$. It was found that by giving a high value to γ (i.e., heavily penalizing any deviation of the filter characteristics from the desired characteristics), a decrease of about 1 dB in the noise power in frequency band $-\Omega$ to $+\Omega$ was possible. However, this resulted in a decrease of S/N ratio by about 1.5 dB. Thus it appears that the inband noise could be suppressed to some extent at the expense of decrease of overall S/N ratio.

SUMMARY AND CONCLUSIONS

In this paper, we have given two techniques for optimizing the digital filter characteristics of an interpolative A/D converter. Computer simulations showed that the optimum digital filters with the integrator reset increases the signal-to-noise ratio by as much as 23 dB over that obtainable by a digital filter with uniform weights and no resetting of the integrator. We also showed that by resetting the integrator a 3 dB advantage in signal-to-noise ratio is obtained when uniform weights are used. We varied the transfer function of the analog filter in the fast loop and found that a gain of a few decibels is possible by proper choice of the analog filter. Finally we considered digital filters whose characteristics could be made close to certain desirable characteristics, and found that it is possible to decrease the quantization noise power within a band, but only at the expense of decrease of the overall signal-to-noise ratio. We note that two important factors, which we have not paid attention to, are: (*i*) stability of the fast loop, and (*ii*) simplicity of implementation of the digital filters. These would be crucial in any practical implementation of the interpolative A/D converters.

ACKNOWLEDGMENT

The author wishes to thank J. C. Candy who, through many discussions, contributed to the development of this work.

REFERENCES

1. D. J. Goodman, "The Application of Delta Modulation to Analog-to-Digital PCM Encoding," B.S.T.J., *48*, No. 2, (February 1969), pp. 321–343.
2. D. J. Goodman and L. J. Greenstein, "Quantization Noise of DM/PCM Encoders," B.S.T.J., *52*, No. 2, (February 1973), pp. 183–204.
3. J. C. Candy, "A Use of Limit Cycle Oscillations to Obtain Robust Analog-to-Digital Converter," IEEE Trans. Commun., *COM-22*, March 1974, pages 298–305.
4. J. C. Candy, Y. C. Ching, and D. S. Alexander, "Using Double Interpolation to Get 13-Bit PCM from a Sigma-Delta Modulator," IEEE Trans. Commun., *COM-24*, November 1976, pp. 1268–1275.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Construction for Group-Balanced Connecting Networks

By F. K. HWANG and T. C. LIANG

(Manuscript received April 1, 1977)

We generalize the concept of balanced network to group-balanced network. An s-stage network is called a group-balanced network if its input switches can be partitioned into groups and its output switches into groups such that the connection pattern (called channel graph) between an input group and an output group is independent of which groups we choose. We show by construction that under a simple divisibility condition, a group-balanced network can be constructed satisfying the following requirements: (i) the number of stages is specified, (ii) the size of the switches in each stage is specified, (iii) the channel graph between an input group and an output group is specified.

I. INTRODUCTION

An s-stage (connecting) network satisfies the following conditions:

(i) The network is composed of switches and links. Switches are arranged in a sequence of s stages.

(ii) The switches in a given stage are identical. In particular, they have the same *size*, i.e., the same number of input terminals and output terminals.

(iii) Links can exist only between two switches in adjacent stages.

In this paper, we assume that each switch is a rectangular (matrix) switch; i.e., there is a crosspoint connecting every input terminal with every output terminal of that switch. Figure 1 illustrates a three-stage network.

Consider an s-stage network and let S_i denote a switch in the ith stage. Consider the paths in the network which connect an S_1 (input switch), say, the kth, with an S_s (output switch), say the jth. Taking the union of all such paths and replacing each switch on a path by a node, we have

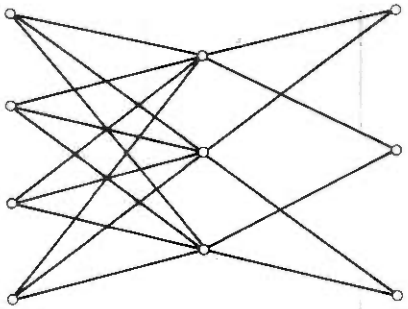


Fig. 1—Three-stage network.

the *channel graph* $G(k, j)$ for that pair of switches. Suppose the collection

$$\{G(k, j): j = 1, 2, \ldots\}$$

is identical for all k, and the collection

$$\{(G(k, j): k = 1, 2, \ldots\}$$

is identical for all j, i.e., the network is symmetric with respect to the switches in the first (last) stage. Then the network is called a *partially balanced network*. If, furthermore, $G(k, j)$ is identical for arbitrary k and j, the network is called a *balanced network*.[4,6]

In this note we generalize the concept of balanced network. An s-stage network is called a *group-balanced network* if its input switches can be partitioned into groups and its output stages into groups such that the connection pattern between an input group and an output group is independent of which groups we choose. This connection pattern will again be referred to as the channel graph between the two groups. When both the input group and the output group contain a single switch, a group-balanced network reduces to a balanced network. Moreover, an s-stage group-balanced network can always be augmented into an $(s + 2)$-stage balanced network by adding a stage before the input stage in such a way that the switches in one input group always connect to the same set of switches in the new stage, and by adding a stage after the output stage with similar connections. We also note that every s-stage network can be viewed as a group-balanced network if all input switches are considered to form one input group and all output switches to form one output group.

The problem of constructing a balanced network with a specified channel graph and given switch sizes has been studied in Refs. 1–8. In this note we give a construction for group-balanced networks under similar conditions.

II. A CONSTRUCTION

For a given node in a channel graph G, we will call the number of links connecting it to a preceding stage its *indegree*, and the number of links connecting it to a succeeding stage its *outdegree*. We assume the specified channel graph is *regular* in the sense that every node in the same stage has the same indegree and outdegree. Let d_i and c_i denote the indegree and outdegree for a node in the ith stage, $i = 1, \ldots, s$. We want to construct a group-balanced network B whose channel graph is the specified one and whose ith stage switches are of given size $n_i \times m_i$ (n_i input terminals and m_i output terminals), $i = 1, \ldots, s$. Note that the number of switches in the ith stage, say l_i, is completely determined from

$$l_i = \prod_{j=0}^{i-1} m_j \prod_{j=i+1}^{s+1} n_j/\lambda, \; i = 1, \ldots, s$$

where $m_0(n_{s+1})$ is defined to he the number of input (output) switches in an input (output) group and λ is the number of paths from the first stage to the last stage in the specified channel graph.

Theorem 1: *Suppose d_i divides n_i and c_i divides m_i for every $i = 1, \ldots, s$. Then the desired B exists.*

Proof: The proof is by construction. Without loss of generality, we may assume that the number of stages s is even. For if s is odd, we can always add an $(s + 1)$th stage, which has a single node, to the channel graph by connecting that node with every node in the sth stage. Since $s + 1$ is now even, we can construct an $(s + 1)$-stage group-balanced network (by defining n_{s+1} and m_{s+1} properly) and then delete the $(s + 1)$th stage. Our construction is by induction on s, $(s = 2, 4, 6, \ldots)$.

Let S_i denote a switch of size $n_i \times m_i$. For $s = 2$, take n_2/d_2 groups of switches S_1 and m_1/c_1 groups of switches S_2. Connect every group of S_1 to every group of S_2 according to the specified channel graph G. The resulting network is the desired one.

Next, consider an s-stage channel graph G for even s. Let f_i be the number of nodes in the ith stage of G. Furthermore, let G^i be the subgraph obtained from G by deleting its first and last $(i - 1)$ stages. Suppose by induction, we have constructed an $(s - 2)$-stage network B' with the specified channel graph G^2. We show show to construct the s-stage network with the specified channel graph G.

Take $(n_s/d_s) \cdot (m_1/c_1)$ copies of B' and label them by $B'(i, j)$ where $i = 1, \ldots, n_s/d_s$ and $j = 1, \ldots, m_1/c_1$. Note that the input (output) switches of B' can be decomposed into $g_2(g_{s-1})$ groups each of which consists of $f_2(f_{s-1})$ switches and the channel graph between every input group and every output group is G^2. Take $g_1 = (n_s/d_s) \cdot (n_2/d_2) \cdot g_2$

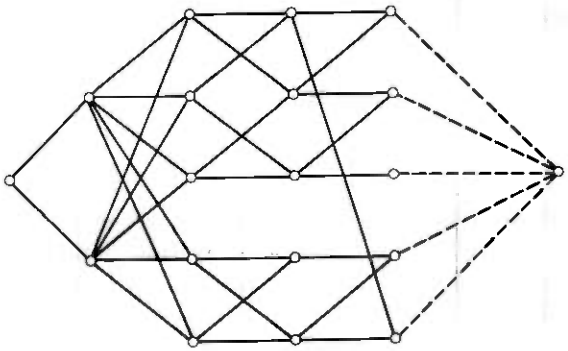


Fig. 2—Specified five-stage channel graph G.

groups (each containing f_1 switches) of S_1 and label the groups by $F(u, v, w)$ where $u = 1, \ldots, n_s/d_s$, $v = 1, \ldots, n_2 d_2$ and $w = 1, \ldots, g_2$. Connect $F(u, v, w)$ to the wth input group of every $B'(u, j)$ according to the connection of nodes in the first two stages of G'. Similarly take $g_s = (m_1/c_1) \cdot (m_{s-1}/c_{s-1}) \cdot g_{s-1}$ groups (each containing f_s switches) of S_s and label the groups by $H(x, y, z)$ where $x = 1, \ldots, m_1/c_1$, $y = 1, \ldots, (m_{s-1}/c_{s-1})$ and $z = 1, \ldots, g_{s-1}$. Connect $H(x, y, z)$ to the zth input group of every $B'(i, x)$ according to the connection of nodes in the last two stages of G'. It is easy to verify that the channel graph between every $F(u, v, w)$ and every $H(x, y, z)$ is the graph G.

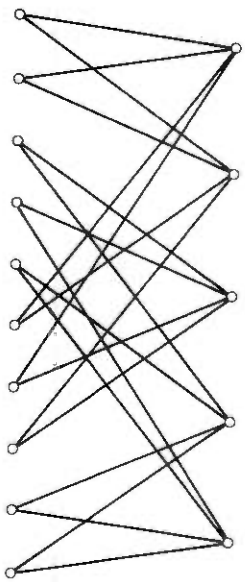


Fig. 3—One of many possible ways of connecting groups, denoted by B^3.

III. EXAMPLES

In this section we illustrate with several examples the scope of applicability of the construction method given in the last section.

Example 1: Let the specified five-stage channel graph G be the one in Fig. 2 (solid lines only), where the specified switch sizes are

$$n_1 = 1, \quad n_2 = 3, \quad n_3 = 2, \quad n_4 = 4, \quad n_5 = 4$$
$$m_1 = 2, \quad m_2 = 5, \quad m_3 = 2, \quad m_4 = 4, \quad m_5 = 1$$

Construction: Since the number of stages is odd, we add an artificial stage (broken lines). We first construct G^3 which has $f_3 = 5$ input nodes and $f_4 = 5$ output nodes. Since

$$\frac{m_3}{c_3} = 1 \quad \text{and} \quad \frac{n_4}{d_4} = 2$$

we take $g_3 = 2$ groups of S_3, $g_4 = 1$ group of S_4 and connect each group of S_3 with the group of S_4 according to G_3. There are many possible ways of connecting, one of which is shown in Fig. 3 and denoted by B^3.

Next we construct G^2 which has two input nodes and five output nodes. Since

$$\frac{m_2}{c_2} = 1, \quad \frac{n_5}{d_5} = 2, \quad \frac{m_4}{c_4} = 2, \quad \frac{n_3}{d_3} = 1$$

we take

$$\frac{n_5}{d_5} \cdot \frac{m_2}{c_2} = 2 \text{ copies of } B^3$$

$$g_2 = \frac{n_5}{d_5} \cdot \frac{n_3}{d_3} \cdot g_3 = 4 \text{ groups of } S_2$$

$$\text{and } g_5 = \frac{m_2}{c_2} \cdot \frac{m_4}{c_4} \cdot g_4 = 2 \text{ groups of } S_5$$

and make connection between groups according to G^2. One possible connection, denoted by B^2, is given in Fig. 4 (solid lines).

Finally, we construct G^1 whose output stage can be ignored (since it is artificial) as long as we define $m_5 = n_6 = 1$. Take

$$\frac{n_6}{d_6} \cdot \frac{m_1}{c_1} = 1 \text{ copy of } B^2$$

$$\text{and } g_1 = \frac{n_6}{d_6} \cdot \frac{n_2}{d_2} \cdot g_2 = 12 \text{ groups of } S_1$$

The final product is given in Fig. 4 with broken lines added.

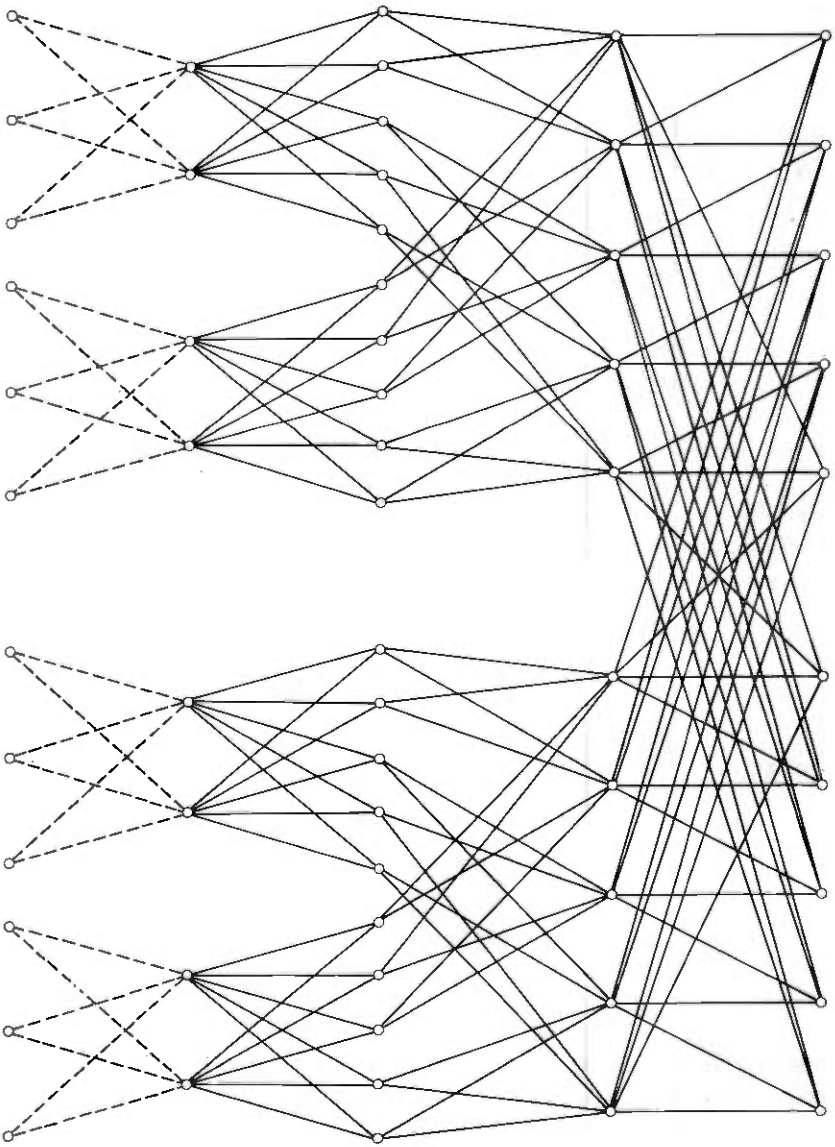


Fig. 4—A possible connection between groups according to G^2.

Example 2. Consider the channel graph in Fig. 5. Suppose we specify $n_1 = 1, n_2 = 1, n_3 = 2, m_1 = 3, m_2 = 2, m_3 = 1$. Then our construction fails since m_1 is not divisible by c_1. However, a balanced network having these parameters and the specified channel graph does exist as shown in Fig. 6.

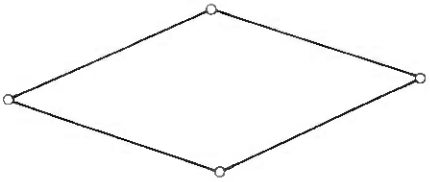


Fig. 5—Channel graph.

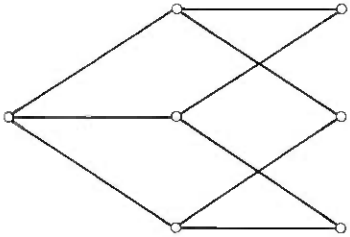


Fig. 6—Specified channel graph.

It turns out that the balanced network in Fig. 6 can be constructed by a method in Ref. 6. However, that method is not able to construct the network in Example 1.

REFERENCES

1. S. Asano, T. Saito, and H. Inose, "An Expression for Structures of Connecting Networks", Electron. Commun. Japan, *54A*, 1974, pp. 68–76.
2. F. R. K. Chung, "On Switching Networks and Block Designs", Conference Records of 10th Annual Asilomar Conference on Circuits, Systems and Computers (1976).
3. K. W. Cattermole, "Graph Theory and the Telecommunication Networks", Bull. Inst. of Math. and its Application, *11*, 1975, pp. 94–106.
4. F. K. Hwang, "Balanced Networks," Conference Record of 1976 International Conference on Communications, pp. 7–13 to 7–16.
5. F. K. Hwang, "Link Designs and Probability Analyses for a Class of Connecting Networks," Technical Memo TM 75-1216-34 (1975).
6. F. K. Hwang and S. Lin, "Construction of Balanced Switching Networks," (to appear).
7. K. Takagi, "Design of Multi-stage Link Systems by Means of Optimal Channel Graphs," Electron. Commun. Japan, *51A*, 1968, pp. 37–46.
8. K. Takagi, "Optimum Channel Graphs of Link System", Electron. Commun. Japan, *54A*, 1971, pp. 1–10.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Automatic Numerical Quadrature

By JAMES L. BLUE

(Manuscript received April 15, 1977)

An automatic numerical quadrature routine (ANQR) attempts to evaluate

$$\int_a^b f(x)\,dx$$

to absolute accuracy ϵ, given only ϵ, a, b, and a user-supplied subroutine which calculates f(x) for any x in [a,b]. An ANQR which guarantees success is impossible to construct, even disregarding the effects of finite computer precision, but the problem is nonetheless of interest. A reliable and efficient ANQR is a necessary part of any mathematical subroutine library. New single- and double-precision ANQRs, QUAD and DQUAD, have been constructed and tested. They are based on adaptive Romberg extrapolation, with cautious error estimation. An important practical feature is the automatic recognition of endpoint singularities, and a change of variable to handle them. QUAD and DQUAD also recognize the presence of noise in the function being integrated, and limit the attempted accuracy accordingly. Since guaranteed ANQRs are impossible, extensive testing of DQUAD is presented to demonstrate its efficiency and robustness. Comparable testing is not available for competitive ANQRs, but performance on a standard set of test integrals is presented for DQUAD and nine other ANQRs. DQUAD is generally better. QUAD and DQUAD are written in PFORT, a subset of American National Standard (ANS) Fortran. Machine-dependent constants are obtained from the PORT library machine-constants programs. A portable package of storage allocation routines is used.

I. INTRODUCTION

The development of automatic numerical quadrature routines (ANQRs) has been a popular research topic for many years (see refs. 1–6, 8, 9, 11, 13, 14). An ANQR is a routine which attempts to calculate

$$\int_a^b f(x)\,dx$$

with absolute error, or perhaps relative error, no larger than ϵ, given ϵ, a, b, and a procedure which calculates $f(x)$ for any desired x in the interval $[a,b]$. It is assumed that no other information about the function f is available. The problem is perhaps the more interesting for being an impossible one. Any numerical quadrature routine must estimate the integral by sampling the function f at a finite number of x's. A guaranteed automatic integration algorithm is clearly impossible for general f, even for analytic f. For example, given any deterministic rule for numerical quadrature, one can readily find constants α and β so that the quadrature rule calculates

$$\sqrt{\alpha} \int_0^1 e^{-\alpha(x-\beta)^2}\, dx$$

to be close to zero. (Choose α to be large and positive, and β to be between sampling points.)

Although the general problem is impossible, one feels that an ANQR which works for "reasonable" functions should be feasible, and much work has been directed at this goal. There has been great confusion and difficulty in comparing the various candidates for ANQRs, partly because the domain of the problem is undefined; a reasonable definition of a "reasonable" function is itself difficult.

In constructing an ANQR, an author is forced to make decisions about the class of "reasonable" functions, in effect to define what is a "reasonable" function. These decisions strongly affect the efficiency and robustness of the ANQR. For example, to avoid completely missing an isolated peak in $f(x)$, the interval $[a,b]$ must be sampled finely. However, a fine sampling is inefficient for easy functions. Another example is a function which is flat over 99 percent of the interval, and which has 200 oscillations in the remaining 1 percent. If an ANQR is able to distinguish this function from one which is merely noisy over 1 percent of the interval, the ANQR is likely to be inefficient on easy functions and very inefficient on noisy functions. An ANQR which gives up relatively quickly on this function, calls it noisy, and returns an error message, may be preferable, especially since many such functions are the result of a user's programing errors.

A compromise strategy, used by QUAD, is to isolate all assumptions about the "reasonable" class of functions in a few parameters. Default values of these parameters can be chosen which will be suitable for most users. More knowledgeable users can use other values. With the default values, QUAD strikes what the author considers to be the proper balance between efficiency and robustness.

Since no *a priori* information about $f(x)$ is available,

$$\int_a^b f(x)\, dx$$

must be evaluated by sampling f in $[a,b]$; the error in the calculated integral is usually estimated by comparing two or more calculated values for the integral. ANQRs typically have a sequence of quadrature rules Q_n, depending on a, b, and the function f, such that

$$\lim_{n\to\infty} Q_n = \int_a^b f(x)\,dx$$

if the calculations are done in infinite precision, and if f is at least piecewise continuous. Most ANQRs have no better error estimation procedure than to accept Q_n whenever $|Q_{n-1} - Q_n| < \epsilon$, a procedure fraught with danger. QUAD has a much more stringent error estimation procedure, described in Section II.

Many functions to be integrated are easy to integrate over some parts of the interval and difficult over other parts. It is frequently more efficient to sample more densely in the difficult regions, if possible. ANQRs which attempt to do this are called *adaptive*—the points at which f is sampled depend on the function being sampled. An adaptive ANQR must include some strategy for how to concentrate the sampling points. Essentially all competitive ANQRs are adaptive.

The usual adaptive procedure is to integrate an interval with quadrature rules Q_n for $n = 1, 2, \ldots, N$, where N is fixed. Q_n may be, for example, Simpson's rule with 2^N intervals, or Gauss-Legendre quadrature with n sampling points. If convergence has not been obtained, the interval is divided in half, and each half considered separately. For efficiency, one wants quadrature rules for which all sampling points for the whole interval are also used for the half-intervals. If the value of N used depends on the results Q_n for $n < N$, the method is sometimes called *doubly-adaptive*.

Most ANQRs do not do well on integrals with endpoint singularities, but users' integrals are frequently of this type. QUAD has a provision for recognizing endpoint singularities and for making a change of variable to facilitate the integration. This feature also works well on another important class of functions, those decaying steeply away from one or both ends of the integration interval. This automatic change of variable technique is a significant improvement over previous ANQRs.

Most ANQRs cannot cope with noisy functions; if there is too much noise in f, most ANQRs fail in an unpleasant, uneconomical way. Convergence will be at best very slow, so that the ANQRs will stop only when their predefined limit on calls to the function evaluation procedure has been exceeded, with no indication that the problem is noise rather than a noise-free but unruly function f. QUAD recognizes noisy functions, sets a warning flag, and integrates only to an accuracy commensurate with the estimated noise.

Finally, there is a large difference between an algorithm for numerical

quadrature and a properly-written ANQR suitable for a program library. Provision must be made, for example, to stop trying to integrate a function if it has been sampled more than some user-defined number of times. The finite machine precision of the computer involved must be taken into account. Temporary storage must not be allowed to overflow. Provision for error returns must be made.

The basic idea behind QUAD is adaptive Romberg extrapolation[5] combined with cautious error estimation[9]. The first such combination was the program CADRE, written by deBoor[6]. CADRE and QUAD are superficially similar, but differ in almost every detail. The major improvements incorporated in QUAD include the following, which will be covered fully in Section II.

(*i*) Noise. QUAD detects noisy functions and quits gracefully.

(*ii*) Endpoint singularities. QUAD detects singularities in $f(x)$ at the endpoints, a and b, and automatically makes a change of variable to reduce the strength of the singularity.

(*iii*) Mesh sequence. QUAD uses the mesh sequence 1, 2, 3, 4, 6, 8, 12, 16, . . . , instead of 1, 2, 4, 8, 16, . . . , giving a higher effective order of convergence.

(*iv*) Portability. QUAD is written in PFORT,[15] a portable subset of ANS Fortran. Machine-dependent quantities are defined with the PORT[7] machine constants. A portable Fortran stack[7] is used for temporary storage.

Section II discusses the algorithm of QUAD and DQUAD more fully. Section III compares the performance of DQUAD and nine competitive ANQRs on a standard set of test integrals, and also presents the results of some more serious testing of QUAD. Section IV discusses the implementation of QUAD, including portability considerations.

II. QUAD

2.1 Romberg extrapolation

QUAD is based on Romberg extrapolation of the composite trapezoidal rule.[5] The formulas are standard, but will be repeated here for completeness. Let $n_1, n_2, \ldots$ be an increasing sequence of positive integers, and let $h_i = (b - a)/n_i$. Then the composite trapezoidal approximation to

$$I = \int_a^b f(x)\,dx$$

is

$$T(h_i) = \frac{1}{2} h_i [f(a) + f(b)] + h_i \sum_{m=1}^{n_i - 1} f(a + m h_i)$$

If f has $2k + 1$ continuous derivatives in $[a,b]$, then the Euler-Maclaurin sum formula shows that

$$T(h_i) = I + \sum_{m=1}^{k} c_m h_i^{2m} + O(h_i^{2k+1})$$

where the c_m depend only on a, b, and f, not on h_i. A higher-order, although not necessarily more accurate, estimate may be obtained by combining two trapezoidal estimates via Richardson extrapolation,[3] eliminating the $c_1 h_i^2$ term. Let $T_0^{(i)} = T(h_i)$.

$$T_1^{(1)} = T_0^{(2)} + \frac{T_0^{(2)} - T_0^{(1)}}{h_1^2/h_2^2 - 1}$$

$$= I + O(h_1^2 h_2^2)$$

Still higher-order estimates may be generated recursively. The general formula for generating $T_k^{(i)}$ is

$$T_k^{(i)} = T_{k-1}^{(i+1)} + \frac{T_{k-1}^{(i+1)} - T_{k-1}^{(i)}}{h_i^2/h_{i+k}^2 - 1}$$

It is customary to think of the T-values as a table, *viz.*

$$\begin{array}{ccccc}
T_0^{(1)} & & & & \\
T_0^{(2)} & T_1^{(1)} & & & \\
T_0^{(3)} & T_1^{(2)} & T_2^{(1)} & & \\
T_0^{(4)} & T_1^{(3)} & T_2^{(2)} & T_3^{(1)} & \\
T_0^{(5)} & T_1^{(4)} & T_2^{(3)} & T_3^{(2)} & T_4^{(1)} \\
\cdot & & & & \\
\cdot & & & & \\
\cdot & & & &
\end{array}$$

The classical Romberg method uses the sequence 1, 2, 4, 8, ... for the n_i's.

Since at any time the right-most element, the "tip," of the T table is of highest order, it is expected to be most accurate, and frequently is so. Early Romberg programs[1] tested for convergence only by checking successive tip elements of the table. There are several arguments against this practice. Firstly, for the tip there is no way to obtain an error estimate with any theoretical foundation. Secondly, highest-order is not the same as most accurate. Even for analytic functions, if the step size used at the beginning of the T table is too large, the tip of the table may not be the most accurate value. Thirdly, for functions $f(x)$ which do not have enough derivatives, the tip of the T table is not of higher order than the elements to the left. For example, if $f(x) = x^\alpha$, for $0 < \alpha < 1$, the lowest-order term in each column of the kth row is $O(h^{\alpha+1})$. In practice it is frequently found that lower-order columns are more accurate than the tip element.

2.2 Cautious error estimation

The idea of cautious error estimation comes from Lynch.[9] It is a simple and seemingly unobjectionable idea, but is not adopted by most authors of ANQRs. A cautious error estimation procedure believes an error estimate only if there is some evidence that the convergence rate of successive quadrature rules is close to the theoretical rate. Cautious error estimation is particularly easy for Romberg extrapolation. The two were first combined by deBoor.[6] QUAD's version of catious error estimation is similar in spirit to deBoor's, but is more cautious and is different in all details.

It has been proven[1] that, if f is merely Riemann-integrable, each column of the T table converges, as does each diagonal. If f has enough continuous derivatives,

$$T_k^{(i)} = I + O(h_i^2 \ldots h_{i+k}^2)$$

These theoretical results provide a basis for cautious error estimation.[6,9] Lynch's suggestion was to consider three successive trapezoidal rule estimates, and form the ratios

$$R_0^{(i)} = \frac{T_0^{(i)} - T_0^{(i+1)}}{T_0^{(i+1)} - T_0^{(i+2)}}$$

$$= \frac{h_i^2 - h_{i+1}^2 + O[(h_i^2 + h_{i+1}^2)^2]}{h_{i+1}^2 - h_{i+2}^2 + O[(h_{i+1}^2 + h_{i+2}^2)^2]}$$

If the step sizes are small enough, the higher-order terms are small compared to the second-order terms, and

$$R_0^{(i)} \approx \frac{h_i^2 - h_{i+1}^2}{h_{i+1}^2 - h_{i+2}^2}$$

The calculated $R_0^{(i)}$ being close to this theoretical value is good evidence that the convergence rate of the column is proper, and that the error in $T_0^{(i)}$ is dominated by $c_1 h_i^2$. Then the Runge estimate of the error,[3]

$$|T_0^{(i+2)} - I| \approx \left| \frac{T_0^{(i+2)} - T_0^{(i+1)}}{h_{i+1}^2/h_{i+2}^2 - 1} \right| = |T_0^{(i+2)} - T_1^{(i+1)}|$$

is likely to be a good estimate. If the calculated $R_0^{(i)}$ is not close to the theoretical value, the Runge estimate of the error is likely to be an underestimate.

Similar calculations are done in higher columns. The general formula for $R_k^{(i)}$ is

$$R_k^{(i)} = \frac{T_k^{(i)} - T_k^{(i+1)}}{T_k^{(i+1)} - T_k^{(i+2)}}$$

As all the h's approach zero,

$$R_k^{(i)} \approx \frac{h_{i+1}^2}{h_{i+k+1}^2} \frac{h_i^2 - h_{i+k+1}^2}{h_{i+1}^2 - h_{i+k+2}^2}$$

The Runge estimate of the error in $T_k^{(i+2)}$ is

$$|T_k^{(i+2)} - I| \approx \left| \frac{T_k^{(i+2)} - T_k^{(i+1)}}{{h_i}^2/{h_{i+k+1}}^2 - 1} \right|$$

A more conservative estimate of the error in $T_k^{(i+2)}$,

$$|T_k^{(i+2)} - I| \approx |T_k^{(i+2)} - T_k^{(i+1)}|$$

is often used.

QUAD calls column k *asymptotic* if $R_k^{(i)}$ is close enough to the theoretical value; the tolerance is 5 percent of the theoretical value for $k = 0$, 10 percent for $k = 1$, 15 percent for $k = 2$, and so on. Column k is *almost asymptotic* if $R_k^{(i)}$ is between 0.25 and 4.0 times the theoretical value, except for column 0, where the criteria are 0.75 and 1.25.

If columns 0 through k are asymptotic, QUAD believes the Runge estimate for the error in the kth column. If columns 0 through $k - 1$ are asymptotic and column k is almost asymptotic, QUAD believes the conservative estimate for the error in the kth column, but no higher columns are believed. (The only exception is that, if the column 0 is only almost asymptotic, the next column is believed if it itself is asymptotic.)

This describes the basic cautious error estimation procedure for QUAD. There are a few more details, however. QUAD does not believe any answer based on less than two extrapolations, or five sampling points, per interval. If two successive entries in a column give the same value to within a few rounding errors, as occurs when integrating a constant function or in doing very accurate integration, then the column does not appear to be asymptotic. The conservative error estimate for the column is believed anyway. If an interval has a singularity, either real or due to rounding errors or truncation errors in the function subprogram for $f(x)$, no column will appear asymptotic. Then a nonasymptotic answer in the first column will be accepted after several extrapolations, with the very conservative error estimate

$$|T_0^{(i+2)} - I| \approx 2|T_0^{(i+2)} - T_0^{(i+1)}| + 2|T_0^{(i+1)} - T_0^{(i)}|$$

Finally, at any stage one column of the T table has only two entries in it, and cannot be judged to be asymptotic or nonasymptotic. The conservative error estimate is accepted for such a column if the previous column is asymptotic.

2.3 Step size sequence

The above discussion applies to any sequence of step sizes. The classical Romberg sequence uses step sizes $(b-a)/n$, with $n = 1, 2, 4, 8, 16, \ldots$, halving the step size and doubling the number of sampling points at each new extrapolation. Several alternative sequences have been suggested which do not cause the number of sampling points to rise so rapidly. QUAD uses n's of 1, 2, 3, 4, 6, 8, 12, . . .; another reasonable possibility is 1, 2, 3, 4, 5, 6, 8, 10, 12, ... These sequences double the number of sampling points every second and third extrapolation, respectively. The classical sequence has the advantages that the bookkeeping is very easy and that all old sampling points are reused if the interval is divided in half. The latter is essential for efficiency.

QUAD's sequence uses fewer sampling points to get the same accuracy, as suggested by Bulirsch and Stoer.[3] The bookkeeping is more complicated than for the classical sequence. For example, it is only convenient to divide the interval in half after the fourth, sixth, eight, . . . extrapolations if all the old sampling points are to be reused.

2.4 Adaptive procedure

For an adaptive Romberg extrapolation routine, it is necessary to decide when to do another trapezoidal rule and another extrapolation, and when to divide the interval. The minimum number of extrapolations for QUAD is 4, using step sizes $(b-a)$ down through $(b-a)/6$, and a total of 9 sampling points. This default lower limit may be raised by the user (see Section IV). Because of roundoff, unlimited extrapolations are impractical—the highest-order columns will not be asymptotic and will not be believed. The maximum number of extrapolations allowed by QUAD is 6; DQUAD allows 8. The default limit may be changed by the user (see Section IV).

If the requested error tolerance for an interval has not been achieved after 4 extrapolations, QUAD goes on to 5 and 6 extrapolations if the first column is asymptotic; after 6, it goes on if the second column is also asymptotic. This procedure is biased in favor of doing more extrapolations, and trying to get higher-order convergence, for smooth functions. Functions which are not smooth, or which do not appear asymptotic because of too large a step size, have the interval divided instead of having more extrapolations done. If QUAD decides to divide an interval, the lower half is stacked, and the upper half is attempted next.

2.5 Change of variable

Functions with singularities are expensive to integrate without special methods. Since the interval containing the singularity will have a nonasymptotic T table, convergence will be limited to the first column.

For example, it can be shown[10] that for

$$I = \int_0^1 f(x)\,dx = \int_0^1 x^{\alpha} g(x)\,dx \tag{1}$$

where g is smooth, and $f(0)$ is set equal to zero,

$$T_1(h) = I + \sum_{m=1}^{\infty} c_m h^{2m} + \sum_{m=1}^{\infty} d_m h^{m+\alpha} \tag{2}$$

The dominant error term for the first column of the T table is likely to be the $d_1 h^{1+\alpha}$ term, so convergence is slow. If $f(0)$ is not zero, another infinite sum is added to (2), like the second sum, but with $\alpha = 0$.

For such an endpoint singularity, the error is of a simple form. It is feasible to recognize this type of singularity in the same way that the cautious error estimation procedure recognizes asymptotic, or h^{2m} behavior. De Boor[6] does exactly this, estimates an $\bar{\alpha}$, and then extrapolates using eq. (2). The success of this procedure depends critically on how accurately α can be estimated. If $f(0)$ is not set equal to zero, de Boor's method will not work well. For logarithmic singularities, the error expansion corresponding to eq. (2) is more complicated; de Boor makes no attempt to recognize logarithmic singularities.

After recognizing an endpoint singularity, QUAD uses a different procedure. Suppose that the integral is as above, where g is well-behaved. Then the leading error term is $O(h^{1+\alpha})$ if $-1 < \alpha < 1$. In the second and higher columns, the $O(h^2)$ term is gone, so the leading term is $O(h^{1+\alpha})$ for $-1 < \alpha < 3$. QUAD looks at ratios of T table entries in the second and third columns to recognize $x^{\bar{\alpha}}$ behavior, and estimates the value of $\bar{\alpha}$. QUAD then makes a change of variable $x = u^n$, where n is the closest integer to $6/(1 + \bar{\alpha})$, giving for eq. (1)

$$\int_0^1 n u^{n-1} f(u^n)\,du = \int_0^1 n u^{n(1+\alpha)-1} g(u^n)\,du$$

(The change of variable is somewhat more complicated if the limits of integration are not 0 and 1.) The new integral has a singularity of the form u^{β}, where β is between 4.5 and 5.5, if $\bar{\alpha}$ is close to the true α. The singularity in the transformed integral is lessened, allowing convergence in the second or third columns of the T table. Convergence is likely to be much quicker. For rapid convergence, the method does not rely on the estimated $\bar{\alpha}$ being close to the true α or upon eq. (2) holding, or indeed on there being any singularity at all at the endpoint.

Steeply decaying integrands such as

$$100 \int_0^1 e^{-100x}\,dx$$

look like step functions when coarsely sampled. A step function is an x^0

singularity, since $f(0) = 0$, so a change of variable is made with $n = 6$. Singularities of the form $x^\beta \log(x)$ will look sufficiently like x^α singularities for some α, so that the transformation will be made. (It is not obvious that this is true, but tests have strongly indicated that it is.)

If $\bar{\alpha}$ is close to -1, n can become large. QUAD requires n to be less than a maximum value determined by the precision of the computer; this value was 22 for tests reported in Section III. The change of variable is not made if $\bar{\alpha}$ is less than -0.99.

To facilitate the change of variable, QUAD starts by dividing the interval $[a,b]$ into three equal intervals, and reverses the upper third. (Three is the default number, and may be changed by the user—see Section IV.) On the lower third and the reversed upper third, a left-hand endpoint singularity is recognized by a pattern of "fail on whole interval, succeed on right half-interval" twice in a row, and is followed by the estimation of $\bar{\alpha}$. If the two estimated values for $\bar{\alpha}$ from the second and third columns of the T table do not agree to within 0.1, no change of variable is made. No change of variable is attempted except at the two endpoints of the originial interval.

2.6 Noisy functions

All procedures for evaluating $f(x)$ are inherently noisy, since they are implemented on finite-precision machines. The value returned is not the exact $f(x)$, but $f(x)[1 + r_1(x)] + r_2(x)$, where $r_1(x)$ and $r_2(x)$ are noise functions. Ideally, r_1 could always be no larger than a few rounding errors, and r_2 could be no larger than a few times the smallest positive machine-representable number. Noise of this size should not affect the performance on an ANQR unless ϵ is very small, of the order of $r_2(x)$ or $f(x)r_1(x)$.

Protecting against this magnitude of noise is quite easy, although few ANQRs bother to do so. QUAD estimates *a priori* the sizes of r_1 and r_2, based on the machine precision, and requires all error tolerances to be at least as large as the estimated rounding error.

Protecting against significantly larger noise is more difficult. A successful ANQR should recognize the presence of noise, estimate its magnitude, and evaluate the integral in a "reasonable" number of function evaluations with an accuracy which is "nearly" as good as possible. (There is of course a trade-off between "reasonable" and "nearly.") If typical values of $r_2(x)$ or of $f(x)r_1(x)$ are much larger than $\epsilon/|b - a|$, most ANQRs will fail in an unpleasant, uneconomical way. Convergence will at best be very slow, so that the ANQRs will stop only when their predefined limit on calls to the function evaluation procedure has been exceeded, with no indication that the problem is noise rather than a noise-free but unruly function f.

SQUANK[11] makes a reasonable effort at recognizing noise. However, it attributes any nonstandard behavior to noise, so that some unruly but noise-free functions are called noisy.

Qualitatively, one may say that a function is noisy if, on a "sufficiently small" interval, the values of the samples of f are "not smooth enough." An algorithm consists of the defining of "sufficiently small" and "not smooth enough," followed by estimation of the magnitude of the noise and by further action to avoid using an excessive number of function values.

In QUAD, no answer is believed unless the function has been sampled with adjacent samples no farther apart than $h_s|b-a|$; h_s is supposedly small enough so that all structure may be seen by sampling with this spacing (h_s is parameter HSAMPL of Section IV). The default value of h_s is $1/8$. Noise is not estimated unless adjacent samples are no farther apart than $h_n|b-a|$, with $h_n = h_s/32$. Choosing h_n smaller would require more function values; choosing h_n larger would increase the risk of calling a function noisy when it is noise-free but rapidly varying.

When a function is noisy, QUAD will usually fail on large intervals, and then attempt smaller and smaller subintervals. When integration on a subinterval has failed, and adjacent samples are no farther apart than $h_n|b-a|$, the noise in f is estimated. First, the second differences of the samples of f on the subinterval are formed, e.g. $f(x) - 2f(x+h) + f(x+2h)$. If the sequence of second differences has no more than two sign changes, noise is not assumed to be present. If there are three or more sign changes, noise is assumed to be present, since the function has too much structure over too small an interval, and the estimated answer and error for that subinterval are accepted as being as good as possible.

When noise has thus been found to be present, the second differences are assumed to be essentially all noise, and the magnitude of the noise is estimated as the average of the absolute values of the second differences. All succeeding subintervals are attempted with accuracy not exceeding the estimated noise magnitude times the length of the subinterval. If other subintervals are found to be noisy, the largest noise magnitude is used.

2.7 Error allocation

QUAD attempts to integrate the upper one-third of $[a,b]$ with error tolerance $\epsilon/3$. Then the following procedure is used to assign an attempted accuracy for each interval. When integration on an interval is attempted with error tolerance ϵ_0 and fails, the upper half is attempted with error tolerance $\epsilon_0/2$. When integration on an interval succeeds, the absolute value of the estimated error is added to a running sum, and the top interval is the stack is attempted. If the top interval is of length δx, the total length of intervals remaining to be integrated is x_1, and the sum

of the absolute values of the error estimates so far is ϵ_1, then the error tolerance assigned to the top interval in the stack is $(\epsilon - \epsilon_1)\ \delta x/x_1$. However, the error assigned to any interval is required to be at least as large as $\epsilon/1000$.

If an answer is returned for an interval with an error estimate which is larger than the requested error, but less than the estimated roundoff or noise, that answer and error are accepted, and a warning flag is set.

III. TESTING AND COMPARISON OF ROUTINES

Testing is necessary to evaluate the efficiency and robustness of an ANQR. Typically the proposer of an ANQR generates an algorithm, codes a simple program, and tests the routine on a few integrals of the proposer's own choosing. It is not unusual for all the test integrals to be done well by the ANQR. As a requlting, the prospective user has no way of evaluating the quality of the ANQR without performing extensive testing.

Some improvement was evidenced in the work of Kahaner,[8] who tested many ANQRs on the same set of 21 test integrals. The same set was used by de Boor[6] for testing his ANQR. At least three of the 21 are not appropriate test integrals, however, because the results are algorithm-dependent in an unrepresentative way.

Two examples will make this clear. First consider

$$\int_0^1 f(x) \cos(\alpha\pi x)\, dx$$

where $f(x)$ is any smooth function. QUAD, which divides according to the 1, ½, ⅓, ¼, ⅙ sequence, will fail for α near 36, but not for α near 32. [For α near 36, the regular sampling procedure of QUAD samples only near the peaks of the cosine, so the integrand looks like $f(x)$.] CADRE,[6] which divides according to the 1, ½, ¼, ⅛ sequence, will fail for α near 32, but not for α near 36. A single test integral with a large α may not compare ANQRs fairly. Test integrals 13 and 17 of Kahaner (see Table I) are of this type, each having about 50 full cycles.

Second, consider

$$\sqrt{\alpha} \int_0^1 e^{-\alpha(x-\beta)^2}\, dx$$

for α large and positive. Depending on the choice of β, this can be either easy or hard for a particular ANQR. If the peak comes sufficiently near a sampling point, adaptive ANQRs can zero in on the peak and integrate it accurately, although many sampling points will be necessary. If the peak does not come sufficiently near a sampling point, the integrand looks like zero, and the ANQR fails. For proper comparison of ANQRs, any single α is insufficient. Test integral 21 of Kahaner is of this type.

Table I — Kahaner's 21 Test Integrals

No.	a	b	answer	$f(x)$
Easy				
12	0	1	+0.7775046341	$x/(e^x - 1)$
11	0	1	+0.3798854930	$1/(1 + e^x)$
1	0	1	+1.7182818284	e^x
10	0	1	+0.6931471806	$1/(1 + x)$
4	−1	1	+0.4794282267	$0.92 \cosh(x) - \cos(x)$
8	0	1	+0.8669729873	$1/(1 + x^4)$
5	−1	1	+1.5822329637	$1/(x^4 + x^2 + 0.9)$
20	−1	1	+1.5643964441	$1/(x^2 + 1.005)$
Steeply decaying				
15	0	10	+1.0000000000	$25e^{-25x}$
14	0	10	+0.5000002112	$\sqrt{50} \exp(-50\pi x^2)$
16	0	10	+0.4993638029	$50/[p(1 + 2500x^2)]$
Singular				
6	0	1	+0.4000000000	$x^{3/2}$
3	0	1	+0.6666666667	$x^{1/2}$
2	0	1	+0.7000000000	$0, x < 0.3; 1, x > 0.3$
7	0	1	+2.0000000000	$0, x = 0; x^{-1/2}, x > 0$
19	0	1	−1.0000000000	$0, x = 0; \ln(x), x > 0$
Oscillatory				
18	0	P	+0.8386763234	$\cos[\cos x + 3 \sin x + 2 \cos 2x + 3 \sin 2x + 3 \cos 3x]$
9	0	1	+0.4794282267	$2/[2 + \sin(10px)]$
17	0.01	1	+0.1121395696	$50[\sin(50px)/(50px)]^2$
13	0.1	1	+0.0090986453	$\sin(100px)/px$
Isolated peak				
21	0	1	+0.2108027354	$\operatorname{sech}^2[10(x - 0.2)] + \operatorname{sech}^4[100(x - 0.4)] + \operatorname{sech}^6[1000(x - 0.6)]$

Note: $p = 3.14159$, $P = 3.1415927$.

Kahaner did not test noisy functions, and did not ask for impossibly small error tolerances.

3.1 Testing on the Kahaner 21

The Kahaner 21 are listed in Table I. They have been grouped according to type; within groups they are ordered approximately by difficulty. Tables II, III, and IV summarize the results of ten ANQRs on the 21 test integrals, for requested error tolerances 10^{-3}, 10^{-6}, and 10^{-9}. In each table, the first column is the integral number; succeeding columns are the number of sampling points used by each ANQR. An F indicates that the ANQR was unsuccessful, and an asterisk that the ANQR used the fewest sampling points of any successful routine. (For these tables only, "successful" means that the true error of an integration is no more than 20% higher than the requested error, since some ANQRs failed by a small amount.) Columns labeled ROMB through RBUN are based on the number of sampling points reported by Kahaner[8] for seven of his highest-quality ANORs, and were obtained on a CDC 6600. Column CADRE is from de Boor,[6] and the integrals were performed on a CDC 6500. Columns QSUBA and DQUAD were computed especially for this com-

Table II — Number of sampling points used by each ANQR; attempted absolute accuracy 10^{-3}

Integral	ROMB	SIMPSN	SQUANK	QNC7	QNC10	QABS	RBUN	CADRE	QSUBA	DQUAD
12	17	19	9	25	37	13	5*	9	7	13
11	17	19	9	25	37	13	5*	5*	7	13
1	17	19	9	25	37	13	5*	9	7	13
10	17	19	9	25	37	13	5*	9	7	13
4	17	19	9	25	37	13	5*	17	7	13
8	17	19	9	25	37	13	5*	9	7	15
5	17	19	9*	25	37	13	11	33	15	17
20	17	31	9*	25	37	13	11	17	15	19
15	513	103	53	85	109	85	527	88	63	52*
14	1025	103	49*	97	127	85	51	62	3F	52
16	2049	115	53	121	163	109	87	81	127	48*
6	17	19	9	25	37	13	5*	9	7	15
3	65	55	9F	49	55	77	211	17	15*	17
2	257F	115	29F	121	163	141	271	53*	771	85
7	8193F	235	105F	241	361	133F	211	33	517	26*
19	4097	175	45	217	307	181	211	137	31	28*
18	129	139	53*	85	73	77	39F	107	63	61
9	33*	163	81	97	145	149	79	183	127	101
17	1025	151	57*	165	307	149	109	512	255	185
13	1025	19F	429	49F	865	573	533	1028	255*	381
21	4097	127	17F	97	127	77	65	108	333F	49*

* = best of all successful results
F = failure

parison, on a Honeywell 6070. Double precision was used so that the relative machine precision would be comparable to that of the CDC machines. QSUBA[14] has provision only for relative error; for these tests a relative error was requested which gave the appropriate absolute error request.

It is important to notice that some of the failures are due to an ANQR deciding to stop because of excessive sampling; these failures are far less reprehensible than the others because an error return could have been made, and an incorrect answer rejected. For RBUN through ROMB, this information can only be inferred since Kahaner did not list any error returns. No such failure occurred for CADRE or DQUAD; QSUBA has no built-in maximum.

ROMB[1] is a standard Romberg extrapolation routine, using the standard 1, ½, ¼, ⅛ sequence. It is not adaptive. It stops, apparently, after 8193 sampling points of $f(x)$. Thus only one of its failures is serious. ROMB requires at least 17 points before believing any answer.

SIMPSN[5] and SQUANK[11] are adaptive Simpson's rule routines, apparently with cutoffs at 5003 and 5001 points, respectively. These routines are decent at low accuracies, but are not of high-enough order to be competitive at high accuracy. SQUANK also assumes that an improper convergence rate is due to noise in the function, rather than to a singu-

Table III — Number of sampling points used by each ANQR; attempted absolute accuracy 10^{-6}

Integral	ROMB	SIMPSN	SQUANK	QNC7	QNC10	QABS	RBUN	CADRE	QSUBA	DQUAD
12	17	19	9	25	37	13	5*	9	7	13
11	17	19	9	25	37	13	11	9	7*	13
1	17	55	17	25	37	13	21	17	7*	19
10	17	55	21	25	37	13*	31	17	15	19
4	17	55	25	25	37	25	5F	33	15*	19
8	33	67	29	25	37	25	41	17	15*	21
5	65	163	65	49	37	49	59	49	31*	41
20	65	163	49	49	37	49	49	33	31*	33
15	2049	343	213	133	145	133	4117	140	63*	98
14	2049	331	169	133	163	109	91	89	3F	86*
16	8193F	511	273	181	181	145	141	145	255	96*
6	129	91	29*	61	73	65	383	65	31	34
3	4097	199	105	157	217	145	423	33	63	32*
2	8193F	235	29F	241	361	261	271	119*	3351	125
7	8193F	1027F	1153F	241F	361F	89F	587F	129	5925	40*
19	8193F	499	257	241	361	105F	403	233	795	38*
18	257	547	301	181	199	205	195	177	63*	117
9	129*	871	377	289	397	313	267F	409	255	285
17	2049	2275	697F	385F	1009	829	697	1237	255*	547
13	2049	19F	2549	1525	1639	1449	2383	1449	255*	757
21	8193F	691	185F	205F	253F	197F	327*	189F	525F	127F

* = best of all successful results
F = failure

larity, and so quits early on some of the test integrals. SIMPSN requires at least 19 points, and SQUANK 9.

QNC7 and QNC10[8] are adaptive Newton-Cotes routines, with 7- and 10-point rules, respectively. (QNC10 was called QUAD in Ref. 8.) They performed quite well, failing only on some of the most difficult integrals. QNC7 requires at least 25 points, and QNC10 at least 37, somewhat excessive for the easiest integrals.

QABS[13] combines Romberg and Curtis-Clenshaw quadrature, and performed quite well, failing only on some of the most difficult integrals. It requires at least 13 points.

RBUN[4] is an adaptive Romberg extrapolation routine, using the standard sequence. It apparently has a cutoff at 5001 points. RBUN requires at least 5 points, and seems to be somewhat unreliable.

Kahaner recommended any of QNC7, QNC10, or QABS as a library routine. He did not have CADRE, QSUBA, or DQUAD available to test.

CADRE[6] is more recent than the ANQRs just discussed. It uses a version of cautious Romberg extrapolation based on the standard sequence. It also includes provision for recognizing a singularity of the form x^{α}, estimates α numerically, and extrapolates using the estimated α. CADRE requires at least 5 points. On the test integrals, it seems somewhat more efficient than QNC7, QNC10, and QABS on nonsingular integrals and

Table IV — Number of sampling points used by each ANQR; attempted absolute accuracy 10^{-9}

Integral	ROMB	SIMPSN	SQUANK	QNC7	QNC10	QABS	RBUN	CADRE	QSUBA	DQUAD
12	17	55	33	25	37	13*	39	17	15	19
11	33	151	33	25	37	25	39	17	15*	19
1	17	163	65	25	37	25	73	17	15*	23
10	65	271	97	37	37	49	123	33	15*	25
4	33	331	105	25	37	85	139	33	15*	25
8	65	463	149	73	73	97	129	65	31*	45
5	129	487	289	97	73	181	239	129	31*	73
20	129	487	249	97	73	145	185	129	31*	49
15	4097	1483	1145	241	217	281	5001F	215	127*	138
14	4097	1123	797	241	253	245	259	202	3F	154*
16	8193F	2467	1649	397	343	397	1435	337	255	192*
6	2049	427	161	133	163	137	1423	529	63	46*
3	8193F	883	513	289	361	289	1595	129	255	42*
2	8193F	235	29F	241F	361F	381	271	173	5931	165*
7	8193F	4279F	5001F	589F	685F	89F	2467F	625	11325	70*
19	8193F	2203	1969F	421F	415F	89F	1571	369F	3495	80*
18	513	2923	1589	409	343	589	753	417	127*	217
9	257*	3967	2525	697	757	893	883	785	765	473
17	4097	5003F	5001F	1345F	1999	2025	2741	2329	255*	1109
13	4097	5003F	5001F	3073	2773	3197	5001F	3505	255*	1161
21	8193F	3751	1657	709	685	633*	1079	661	827F	261F

* = best of all successful results
F = failure

much more efficient on singular ones. In addition, the cautious extrapolation means that CADRE's error estimation procedure has some rationale behind it, and CADRE is more robust than the aforementioned routines. However, CADRE is difficult to understand and to maintain, since its style is the antithesis of structured programming. It is one large program, with no subprograms, but with a liberal and unstructured use of GOTOs.

QSUBA[14] uses a series of 8 whole-interval quadrature rules of increasing order, starting with the 1- and 3-point Gauss-Legendre rules. Succeeding rules are constructed to be of as high order as possible, consistent with using all the previous sampling points. The highest-order rule uses 255 points and is of order 383. If convergence is not obtained after 8 rules, the interval is divided in half, and each half considered anew. Unlike all the other ANQRs under consideration, all function values must be discarded, since none of the sampling points on the half intervals coincides with any on the full interval. QSUBA works well on any integral which can be integrated without dividing the interval, and poorly on integrals which require dividing. It is especially good on easy and oscillatory integrals, since a high-order rule is generally used. QSUBA uses at least 3 points, which is somewhat unsafe, but has *no* maximum built in—it goes on forever, if necessary.

DQUAD takes at least 13 points. It fails only on number 21, for high accuracy, missing the narrow peak at $x = 0.6$. The integral which is the same as 21 except for moving the peak to 0.61 is done properly, using 117, 241, and 447 points for error tolerances 10^{-3}, 10^{-6}, and 10^{-9}, respectively. DQUAD is clearly more efficient and robust, based on these test integrals, than any other ANQR tested except QSUBA. DQUAD is more robust than QSUBA, but is less efficient for integrals where QSUBS does not need to divide the interval.

3.2 Parameter studies (1)

Testing an ANQR on a "random" set of test integrals, while instructive and a good start, is insufficient for a library routine. The testing of an ANQR is incomplete without numerous parameter studies:

$$\int_a^b f(x;\alpha)\,dx$$

with fixed ϵ and varying α (Ref. 12), and with fixed α and varying ϵ. The function $f(x,\alpha)$ should be increasingly difficult to integrate as α approaches some limit, and α should be pushed close enough to that limit so that failure occurs. For error tolerance studies, the requested error should range from the approximate value of the integral to less than typical roundoff on the computer being used.

Several of the first type of parameter study will now be discussed. For all of them, the error requested is 10^{-6}. The first was suggested by de Boor[6]:

$$\int_0^1 \frac{2^\alpha}{1+(2^\alpha x)^2}\,dx$$

For large α, the integrand is highly peaked. The integrand is 2^α at $x = 0$, falls to half that at $x = 2^{-\alpha}$, and is $2^{-\alpha}$ at the endpoints of the interval. There is no danger in missing the narrow spike, since it is exactly at the center of the interval, a normal sampling point for almost all ANQRs. This example demonstrates the power of adaptive ANQR, in that the number of sampling points increases only as $\sqrt{\alpha}$, approximately. The behavior of DQUAD is shown in Fig. 1. The number of sampling points is plotted against α. The meaning of the symbols used for plotting in all the figures is given below.

- · Successful integration; no error flag
- × Unsuccessful integration; error flag
- O Unsuccessful integration; no error flag
- ⊗ Successful integration; error flag

In testing DQUAD, "successful" means that the true error is less than ϵ. DQUAD failed for $\alpha > 38$, but recognized its failure (failure was due to

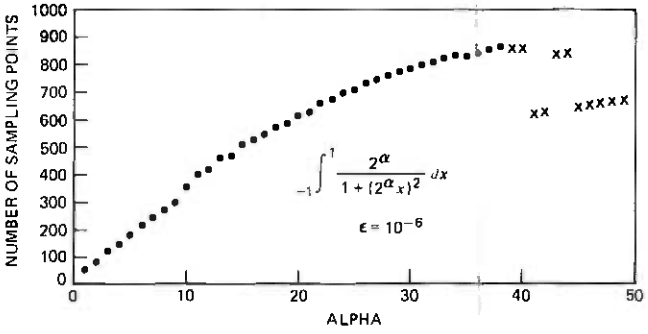


Fig. 1—Performance of DQUAD on function highly peaked at center of interval.

filling of the function value stack). For comparison, CADRE fails starting at $\alpha = 31$, and generally uses about 20 percent fewer sampling points than DQUAD.

This integral should not pose any serious difficulty to a decent ANQR, since the unpleasant behavior occurred exactly at the center of the interval. Usually, though, the user should strive to break up integrals so that any unpleasant behavior happens at one of the endpoints of the interval. It is feasible for ANQRs to recognize such behavior at the endpoints, but difficult if it occurs in the center of the interval. As an example, the previous integral, except with limits 0 and 1, may be considered. DQUAD's performance is shown in Fig. 2. For $\alpha \geq 7$, DQUAD recognizes that the integrand approximates a step function, and makes a change of variable. This change of variable keeps the number of sampling points from growing significantly as α increases.

Figure 3 shows a similar integral, except more steeply decaying away from the endpoint.

$$\int_0^1 2^{\alpha} e^{-2^{\alpha}x}\,dx$$

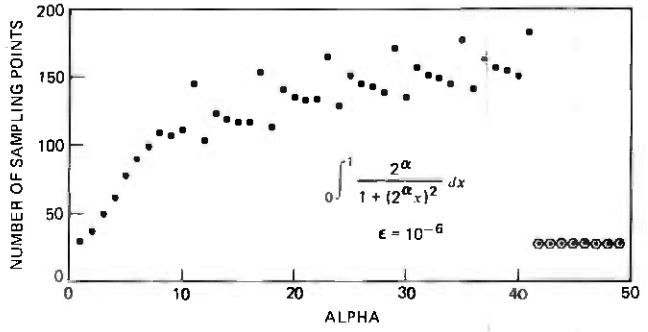


Fig. 2—DQUAD's performance on function highly peaked at end of interval.

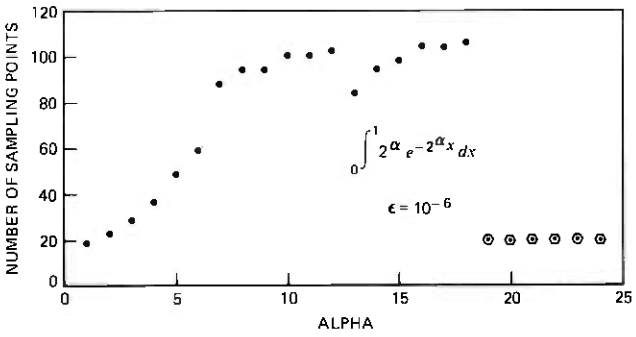


Fig. 3—Performance of DQUAD on steeply decaying function.

Failure eventually occurs, as for the previous integral, when, after the change of variable, the new integrand is a sharply-peaked function with its peak away from an endpoint, and the peak is missed entirely.

Testing of ANQRs on functions with isolated peaks within the range of integration takes more work; α must parameterize the narrowness of the peak, but the position of the peak is also important. A suitable test integral has two parameters. DQUAD was tested on

$$\int_0^1 2^\alpha e^{-4^\alpha(x-\beta)^2}\,dx$$

For each α, 25 integrals were done, with β's of 0.02(0.02.)0.50, for a statistical evaluation. No failures occurred for $\alpha = 1, 2, 3, 4, 5$, or 6; one occurred for $\alpha = 7$, at $\beta = 0.04$. For $\alpha = 8$, 12 out of 25 failed. Figure 4 illustrates the results for $\beta = 0.40$, a typical value.

Another standard test integral, also used by de Boor, is

$$\int_0^1 x^\alpha\,dx$$

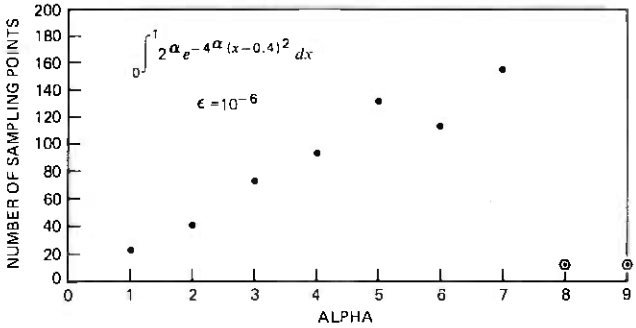


Fig. 4—Performance of DQUAD on highly peaked function.

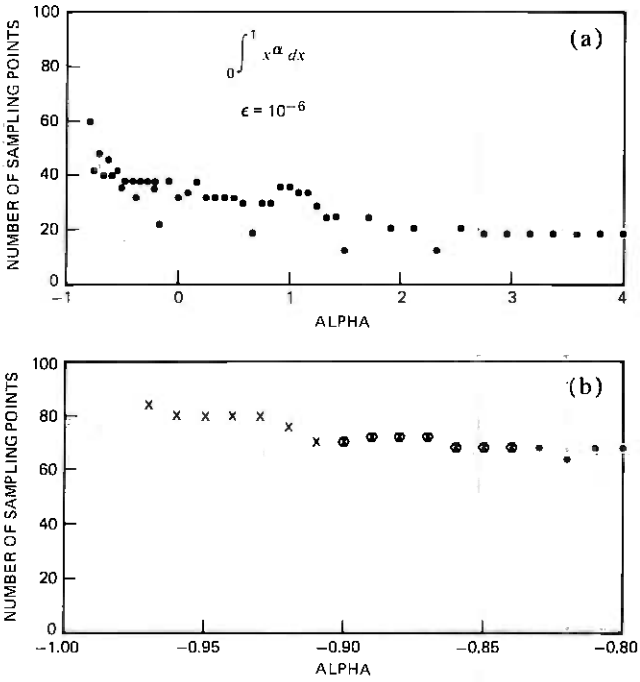


Fig. 5—Performance of DQUAD on function with endpoint singularity.

with the integrand set equal to zero at $x = 0$. Thus for $\alpha = 0$ the integrand is a step function, and for $\alpha < 0$ the integrand has an infinite discontinuity. For $\alpha < 0$, the integrand is "unreasonable" by almost anyone's definition, but users sometimes give such integrals to ANQRs. Figure 5 shows the performance of DQUAD. A change of variable was made automatically by DQUAD for $-0.97 \leqslant \alpha < 1.75$, approximately. DQUAD is designed to reject the change of variable if the estimated α is less than -0.99, and the change is not necessary to achieve the desired accuracy for $\alpha \geqslant 1.75$. For $\alpha < -0.97$, DQUAD fails with an error flag, using about 900 function samples. Machine precision limits the efficiency of the change of variable for α less than about $-3/4$. For comparison, CADRE fails, with an error flag, for α less than $-7/8$, and gives an erroneous error flag for α near, but not at, $\alpha = 1$. CADRE is substantially less efficient than DQUAD for this test integral. Other ANQRs, without special procedures for endpoint singularities, are much worse.

A final type of test integral is one with an oscillatory integrand. Since almost all ANQRs sample the integrand at regularly spaced points, there is a danger of undersampling. As an example, consider

$$\int_0^1 [1 + \cos(\alpha \pi x)] \, dx$$

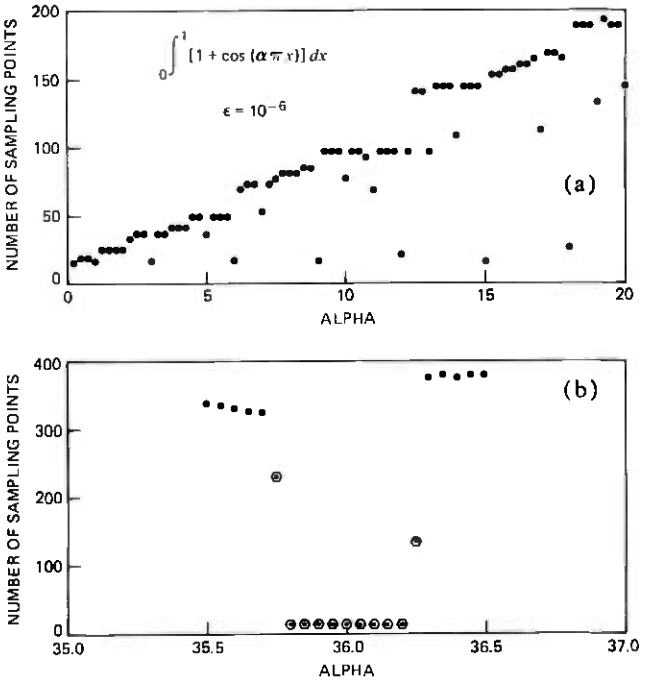


Fig. 6—Performance of DQUAD with oscillatory integrand.

as suggested by de Boor. Figure 6 illustrates the performance of DQUAD on this test integral. The first failure is for α near 36, where the regular sampling causes the cosine to look like unity. The lower part of the figure shows the region near 36. The top part shows the number of sampling points used for α from $\frac{1}{4}$ to 20, with spacing of $\frac{1}{4}$. The trend line is approximately $N = 10\alpha$; CADRE's trend line is approximately $N = 20\alpha$. Besides the general rising trend, there many dots significantly below the trend line. These occur because of resonance between the regular sampling of DQUAD and the regular oscillation of the integrand. If α is integral, or very nearly so, coarse sampling may indicate that the integrand is simpler than it really is. For example, if α is an odd integer, the cosine is odd about $x = 0.5$, and the center third of the integral is integrated (correctly) based on insufficient sampling, since the trapezodial rule correctly integrates odd functions.

This kind of resonance phenomenon is a difficult problem for an ANQR to surmount. The best way is probably to back off slightly from the notion of a fully automatic ANQR. Usually a prospective user of an ANQR knows if an integrand is oscillatory, and further, knows the approximate period of the oscillation. To avoid problems with possible undersampling, the user could divide the integral into several integrals, each with only

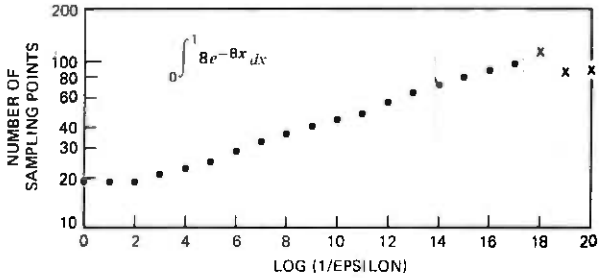


Fig. 7—Parameter test with accuracy varied.

a few periods of the oscillation. An easier way is for the user to require the ANQR to sample the integrand sufficiently finely to see all the oscillations; this assumes the user knows the approximate period. DQUAD has provision for this, using the parameter HSAMPL, discussed in Section IV. If HSAMPL is taken to be $1/\alpha$ in the above example, all the dots fall on the main trend line, and no failures occur.

3.3 Parameter studies (2)

The second type of parameter testing, using a single test integral and varying the requested error, will be considered now. Some ANQRs will fail when the requested error is large compared to the value of the integral, because of insufficiently cautious error estimation. All ANQRs fail if the requested error tolerance is too small, because of roundoff. A properly designed library ANQR will have some mechanism for dealing with too-small error tolerances. Finally, the graph of the number of sampling points versus the requested accuracy is of interest.

Figure 7, for

$$\int_0^1 8e^{-8x}\,dx$$

is typical of DQUAD's performance on easy integrals. The graph of N against $\log 1/\epsilon$ is horizontal at large requested error, with no failures, since DQUAD does not believe any answer until it can be confident of the error estimate. The central portion of the graph is roughly linear. The number of sampling points needed is approximately proportional to $\epsilon^{-1/20}$. DQUAD is effectively functioning as a 20th-order method. The smallest-error part of the graph is also horizontal; accuracy is limited by roundoff.

Figure 8, for

$$\int_0^1 x^{1/2}\,dx$$

is similar except for the smallest requested errors. A change of variable

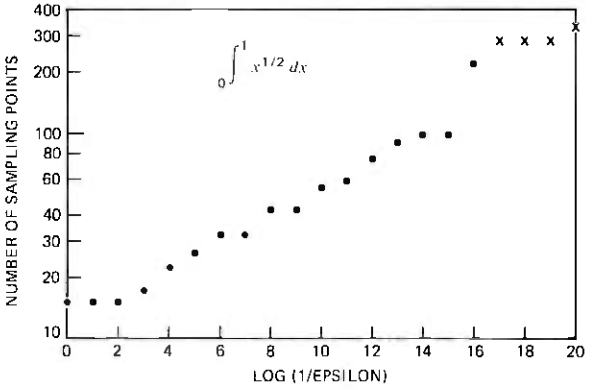


Fig. 8—Parameter test with accuracy varied.

is made only for $\epsilon = 10^{-4}$ through 10^{-16}. For $\epsilon = 10^{-2}$ through 10^{-15}, DQUAD is effectively a 15th-order method. The effective order is less than for the previous example because only three columns of the T table can be asymptotic. Roundoff begins to pollute the answer at $\epsilon = 10^{-16}$.

Figure 9, for

$$\int_0^1 [1 + \cos(\alpha\pi x)]\,dx$$

for two values of α, 1.95 and 17.95, is the last example. For $\alpha = 1.95$, DQUAD is effectively a 20th-order method for $\epsilon = 10^{-3}$ through 10^{-17}. For $\alpha = 17.95$, DQUAD does not begin to be a 20th-order method until $\epsilon = 10^{-6}$.

3.4 Parameter studies (3)

The final type of parameter testing uses a single test integral and error tolerance, but varies the amount of noise in the function. For testing,

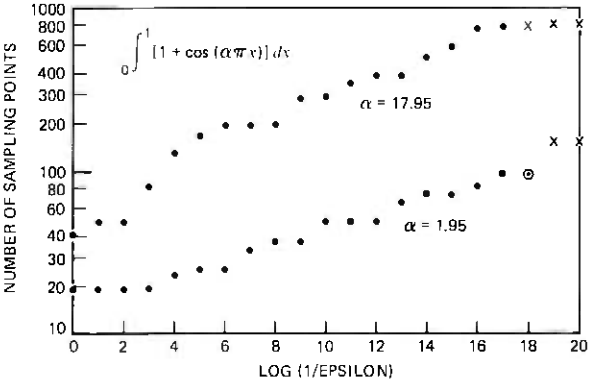


Fig. 9—Parameter test with accuracy varied.

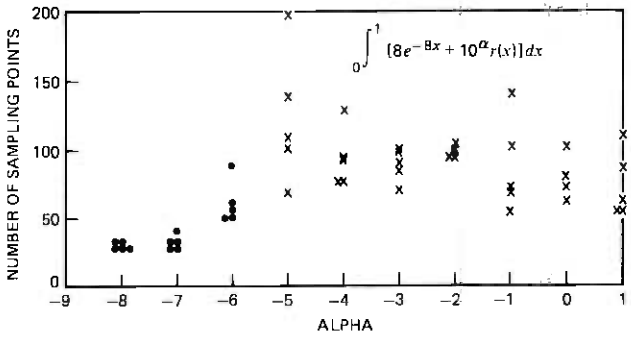


Fig. 10—Parameter test with amount of noise varied.

integrals of the types

$$\int_0^1 [f(x) + 10^{\alpha} r(x)]\,dx$$

$$\int_0^1 f(x)[1 + 10^{\alpha} r(x)]\,dx$$

were done, with requested accuracy $\epsilon = 10^{-6}$. The function $r(x)$ is the output of a pseudorandom number generator with values uniformly distributed on $(-1,1)$. Values used for α were $1, 0, -1, \ldots, -8$. For f, the same four functions were used as in the previous section. For each function and each α, five different starting points of the random number generator were used. Sample results are shown in Fig. 10 and 11. The number of function values used is plotted vs. α, for each of the five tries. (Where two or more of the tries coincide, they are plotted side by side.) Points plotted with · are those for which DQUAD returned with an error estimate less than 10^{-6}; points plotted with × indicate an error estimate greater than 10^{-6}.

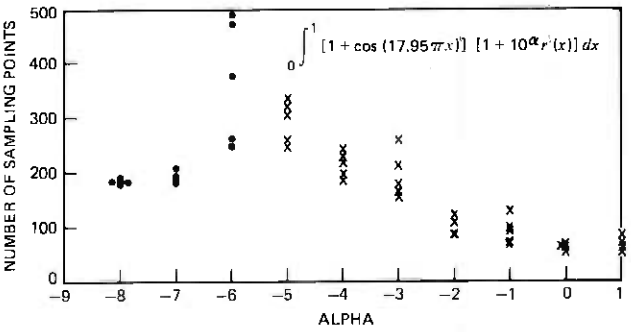


Fig. 11—Parameter test with amount of noise varied.

A summary of the results follows.

For every test with $10^{\alpha} > 10^{-6}$, DQUAD recognized the presence of noise, estimated its magnitude, adjusted accuracy tolerances accordingly, and gave an answer and error estimate. The error estimates given were at most twice 10^{α}, and were generally less than 1.2 times 10^{α}. In only 6 out of 280 integrals was the calculated answer farther from the noise-free value of the integral than the estimated error returned.

For every test with $10^{\alpha} = 10^{-6}$, DQUAD returned an answer with an error estimate less than 10^{-6}. In 3 out of 40 integrals, the presence of noise was recognized. In 2 out of 40, the calculated answer was between 1 and 2 times 10^{-6} away from the noise-free value; in the remaining 38, the answer was less than 10^{-6} away.

For every test with $10^{\alpha} < 10^{-6}$, DQUAD returned an answer with an error estimate of less than 10^{-6}, and the calculated answer was less than 10^{-6} away from the noise-free value. No noise was recognized.

IV. IMPLEMENTATION OF QUAD

QUAD is a Fortran subroutine which attempts to evaluate

$$\int_a^b f(x)\,dx$$

to within absolute accuracy ϵ; the user supplies a, b, ϵ, and a Fortran function subprogram to evaluate $f(x)$. The discussion also applies to the double precision version, DQUAD, except as noted.

QUAD is written in PFORT,[15] a portable subset of ANS Fortran. Temporary storage space is managed by a portable Fortran stack mechanism.[7]

The calling sequence is

```
CALL QUAD (F,A,B,EPS,ANS,ERREST)
```

F is the name of the user's subprogram, A and B are the limits, and EPS the requested accuracy. The estimated value of the integral and an estimate of the accuracy of the estimate are given in ANS and ERREST. If ERREST is larger than EPS, QUAD invokes the PORT library's centralized error handling facility,[7] turning on the error state before returning control to the calling program. If the user has not taken prior action to recover from errors, an error message is printed and the run is terminated. The user who has taken prior action may test for the error state, and continue if desired.

QUAD sets seven parameters to their default values and then calls R1QUAD. The default values should be adequate for almost all users, but the user wishing other values may call R1QUAD directly. R1QUAD also gives more information to the user about problems encountered during the integration. R1QUAD's calling sequence is

CALL R1QUAD(F,A,B,EPS,HSAMPL,NFCALL,LYSTAK,KMAXEX,
KDIVID,JPRINT,NUMINT,ANS,ERREST,KWARN)

The default parameter values used by QUAD are

HSAMPL	0.125
NFCALL	2000
LYSTAK	250
KMAXEX	6
KDIVID	4
JPRINT	0
NUMINT	3

The same parameters are used for DQUAD, except that KMAXEX = 8.

HSAMPL measures how finely $f(x)$ must be sampled. No error estimate is believed unless the trapezoidal rule step size is HSAMPL $|b - a|$ or smaller. Reducing HSAMPL improves the robustness of R1QUAD, but decreases its efficiency. Changing HSAMPL is useful for integrating oscillatory functions, where there is some danger of aliasing. For $\int_0^1 \cos(\alpha\pi x)\, dx$, HSAMPL $= 1/\alpha$ is safe. The minimum number of sampling points of $f(x)$ is roughly 2/HSAMPL.

NFCALL is the upper limit on the number of sampling points of $f(x)$.

LYSTAK is the length of the stack for storing values of the function at the sampling points. Space for the stack is allocated within R1QUAD using a portable storage-allocation package.[7]

KMAXEX is the maximum number of extrapolations allowed. Legal values are 4, 6, 8, 10, and 12.

KDIVID is the minimum number of extrapolations required before dividing an interval. Legal values are 4, 6, 8, . . . , KMAXEX.

JPRINT determines how much printing will be done. With JPRINT < 1, there is none. With JPRINT = 1, the endpoints, attempted error tolerance, answer, and error estimate are printed for each interval attempted. With JPRINT > 1, the T tables are also printed.

NUMINT is the number of equal intervals into which $[a,b]$ is divided to start. If NUMINT > 1, signularities can be recognized at $x = a$ and at $x = b$; if NUMINT = 1, a singularity can be recognized only at $x = a$. Increasing NUMINT generally increases robustness while decreasing efficiency.

KWARN is an integer warning flag, output from R1QUAD. It is zero if all went well. If KWARN is positive, it may have up to 6 digits, each with an independent meaning. Although ERREST may be greater than the requested EPS, it is still reliable. Each digit is zero unless a problem occurred; starting with the right-most digit, the problems are:

(*i*) The error estimate is limited by noise or roundoff, but is above the requested error.

(*ii*) The interval size is as small as is allowed.

(*iii*) As many intervals are stacked as is allowed.

(*iv*) As many function values are stacked as is allowed.

(*v*) As many function sampling points have been used as is allowed.

(*vi*) The error estimate is larger than the requested error.

4.1 Machine-dependent constants

All machine dependency is isolated into four machine-dependent constants, which are set by R1QUAD. No reprogramming is necessary to run QUAD or R1QUAD on another computer. The constants are set using the portable machine constants program R1MACH of Ref. 7.

DLARGE is used as error estimate before any error estimates are considered believable. Its value is set to slightly less than the largest floating-point number, 0.1 R1MACH(2).

DSMALL is used as the default magnitude of $r_2(x)$. Its value is slightly larger than the smallest positive floating-point number, 10 R1MACH(1).

DROUND is used as the default magnitude of $r_1(x)$. Its value is set to 50 times the largest relative rounding error, or 50 R1MACH(4).

HSMALL is the smallest fraction of $|b - a|$ used for a trapezoidal rule step size. Its value is the same as DROUND.

REFERENCES

1. F. Bauer, H. Rutishauser, and E. Stiefel, "New Aspects in Numerical Quadrature," Proc. of Symposia in Applied Mathematics, *15*, pp. 199–218, American Mathematical Society, Providence, 1963.
2. R. Bulirsch and J. Stoer, "Fehlerabschaetzungen and Extrapolation mit Rationalen Funktionen bei Verfahren vom Richardson-Typus," Num. Math., *6*, 1964, pp. 413–427.
3. R. Bulirsch and J. Stoer, "Asymptotic Upper and Lower Bounds for Results of Extrapolation Methods," Num. Math., *8*, 1966, pp. 93–104.
4. W. Bunton, M. Diethelm, and K. Haigler, "Romberg Quadrature Schemes for Single and Multiple Integrals," Jet Propulsion Laboratory Report TM-324-221, 1969.
5. P. J. Davis and P. Rabinowitz, *Numerical Integration*, Waltham, Mass.: Blaisdell, 1967.
6. C. de Boor, "CADRE: An Algorithm for Numerical Quadrature," in *Mathematical Software*, J. R. Rice (ed.), New York: Academic Press, 1971, pp. 417–449.
7. P. A. Fox, A. D. Hall, and N. L. Schryer, "The PORT Mathematical Subroutine Library," Bell Laboratories Computer Science Technical Report No. 47, September, 1976. Accepted for publication in ACM Transactions on Mathematical Software. Inquiries about the PORT library may be addressed to Bell Laboratories Computing Information Service, 600 Mountain Avenue, Murray Hill, New Jersey 07974.
8. D. K. Kahaner, "Comparison of Numerical Quadrature Formulas," in *Mathematical Software*, J. R. Rice (ed.), New York: Academic Press, 1971, pp. 229–259.
9. R. E. Lynch, "Generalized Trapezoidal Formulas and Errors in Romberg Quadrature," Blanch anniversary volume, Aerospace Research Laboratories, Office of Aerospace Research, United States Air Force, pp. 215–229, 1967.

10. J. N. Lyness and B. W. Ninham, "Numerical Quadrature and Asymptotic Expansions," Math. Comp., *21* (1967), pp. 162–178.
11. J. N. Lyness, "Algorithm 379, SQUANK," Comm. ACM, *13*, 1970, pp. 260–263.
12. J. N. Lyness, attribution by C. de Boor (Ref. 6).
13. H. O'Hara and F. Smith, "The Evaluation of Definite Integrals by Interval Subdivision," Nat. Bur. of Standards, Report N69-11541, 1969.
14. T. N. L. Patterson, "Algorithm 468, Algorithm for Automatic Numerical Integration over a Finite Interval," Comm. ACM, *16*, 1973, pp. 694–699.
15. B. G. Ryder, "The FORTRAN Verifier: User's Guide," Bell Telephone Laboratories Computing Science Technical Report No. 12, March 1973.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Criteria for Determining if a High-Order Digital Filter Using Saturation Arithmetic Is Free of Overflow Oscillations

By DEBASIS MITRA

(Manuscript received May 2, 1977)

Recently we found that, among recursive digital filters using saturation arithmetic to contend with overflow, a fundamental difference exists between second and higher order filters: the latter may sustain large-amplitude overflow oscillations. In this paper we have derived a new criterion expressly designed for determining when a given high-order recursive system using saturation arithmetic is free of overflow oscillations. The new criterion, which is easy to use, follows from this result: we associate with the given system two trigonometric polynomials in θ of degree equal to the order of the given system; if any linear combination of the polynomials with nonnegative weights is positive for all θ in $[0,\pi]$ then the system is free of all nontrivial periodic oscillations. We prove that the new criterion subsumes certain well-known criteria, such as Tsypkin's criterion, from the literature on nonlinear systems. To illustrate, three classes of special systems are investigated, and in each case the new criterion gives substantial improvements. Finally, the new test is applied in the synthesis of high order sections for a realistic eighth-order system.

I. INTRODUCTION

Recently[1] we made the unexpected observation that, among recursive filters employing saturation arithmetic, a fundamental difference exists between second and higher order filters, namely, the latter may sustain large-amplitude overflow oscillations. This observation proved to be timely since it coincided with the awareness that economies of scale coupled with various recent developments make highly attractive the use of high-order sections in filter realizations. It has also come to light that the problem of possible large-scale oscillations is of interest not only in data filtering but in other areas where the natural structure is a

high-order recursive system, e.g., code converters (DPCM → PCM) and speech synthesizers.

The economies of scale derive from the fact that the overflow detection and correction circuits, an expensive part of present-day filters, are as many as the number of sections employed; thus if a realization is composed of fourth-order sections rather than the conventional second-order sections, then the number of such circuits may be expected to be halved. The recent developments alluded to earlier refer to the almost simultaneous developments of inexpensive, lower-power-consuming semiconductor read-only memories, and the concept of distributed arithmetic blocks[2,3] in which ROMs are used to implement digital filters. In a pioneering study R. B. Kieburtz[4] recently estimated that in a particular application a saving of about 30 percent in parts may be achieved over the conventional design through the use of fourth-order sections using saturation arithmetic and implemented by ROMs.

Thus there is much to be gained if high-order sections can be used, and for this to happen it is first necessary to ensure that the highly destructive overflow oscillations are not present in a particular design. It is apparent that there is a useful role for an effective criterion for delineating stable systems employing saturation arithmetic. It is possible to conceive of the situation where such a criterion is incorporated in the early design, i.e., the criterion is introduced as a constraint in the approximation problem. The other possibility of an arithmetic different from saturation arithmetic to contend with overflow is not pursued in this paper.

There do exist many such criteria in the literature on the stability of a class of nonlinear feedback systems (i.e., the systems in the Lurie problem[24]) of which the one under consideration here is a member[5–11]; the reader may consult Ref. 8 for a comparative evaluation of some of these criteria. These criteria are in some sense generalizations of Nyquist's criterion for linear feedback systems. The reader will find in Sec. 5.2 a statement, in the context of the present problem, of Tsypkin's criterion and the discrete circle criterion, two well-known examples of such criteria. Unfortunately it is known that when the nonlinearity in the system is the one associated with saturation arithmetic then these criteria, including Barkin's criterion,[7,8] are of limited utility since they are excessively pessimistic. Examples to this effect may be found here. Also telling is the fact that these criteria do not predict that all second-order systems using saturation arithmetic are free of oscillations,[8] a fact proven in Refs. 12–14 by arguments special to second-order systems. This is not totally unexpected in view of the fact that the systems-theoretical criteria apply to large classes of systems and nonlinearities and, concomitantly, use relatively little information (restricted to the sector information, symmetry, and monotonicity) about the nonlinearity.

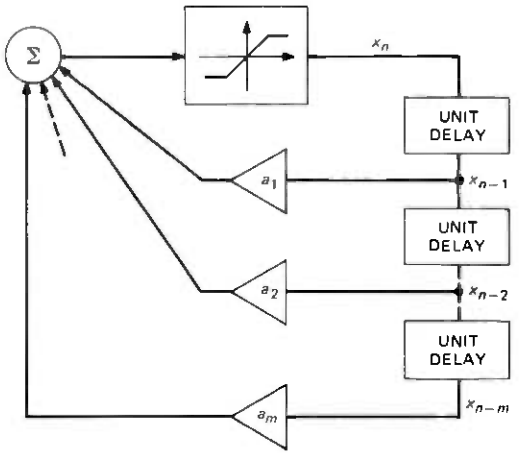


Fig. 1—Schematic of unforced high-order filter employing saturation arithmetic.

In this paper we have derived new criteria expressly designed for the system employing saturation arithmetic. Underlying the new criteria is the observation that certain unique passivity conditions are operative in the case of saturation arithmetic. Both the observations regarding the passivity conditions as well as the technique we use for deriving the criterion are believed to be new. The main ingredient in the derivation is the observation that the expressions associated with the passivity conditions in periodic solutions possess remarkable structure; namely, they are quadratic forms involving circulant matrices.

The systems considered in this paper are of the form (see Figs. 1 and 2)

$$x_n = F\left(\sum_{j=1}^{m} a_j x_{n-j}\right), \quad n = 0,1,\ldots \tag{1}$$

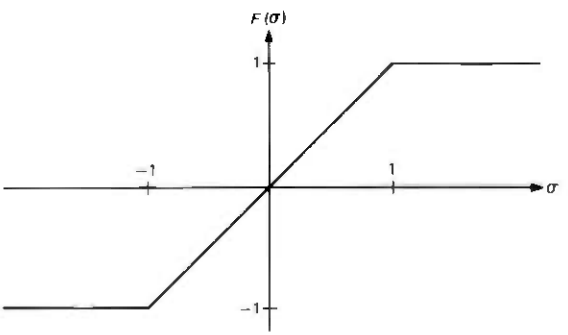


Fig. 2—The saturation arithmetic nonlinearity.

where m is the order of the system, $\{a_j\}$ are the coefficients, and $F(\cdot)$ is a nonlinear function associated with saturation arithmetic, namely

$$F(\sigma) = \sigma \quad \text{if } |\sigma| \leq 1$$
$$= \text{sgn } \sigma \text{ if } |\sigma| \geq 1 \tag{2}$$

It is tacitly assumed that the underlying linear system in eq. (1) is absolutely stable, i.e.,

$$\lambda^m - \sum_{j=1}^{m} a_j \lambda^{m-j} \neq 0 \text{ for } all \ |\lambda| \geq 1. \tag{3}$$

Thus any nontrivial solution of eq. (1) will necessarily have either 1 or -1 as an element and consequently such solutions are referred to as overflow oscillations. Note that we are following convention in ignoring quantization effects in the description of the filter in eq. (1); in investigations of large-scale oscillations it is natural to focus on the gross nonlinearity.

The main result of this paper (Sec. 2.4) is simply stated: for a given system of order m with coefficients $\{a_j\}$, we associate two polynomials of degree m in $\cos \theta$, namely,

$$p_1(\theta) = 1 - \cos \theta - \sum_{j=1}^{m} a_j \{\cos j\theta - \cos (j-1)\theta\} \tag{4}$$

and

$$p_2(\theta) = 1 + \cos \theta - \sum_{j=1}^{m} a_j \{\cos j\theta + \cos (j-1)\theta\}. \tag{5}$$

If any linear combination of the polynomials with nonnegative weights is positive for all θ in $[0,\pi]$ then the system in eq. (1) with arbitrary initial conditions does not admit any nontrivial periodic solutions. Certain generalizations of this criterion are derived in Sec. VI.

In Sec. 2.5 ("How To Use The New Test") we show that the criterion may be used in a straightforward manner by one of two methods. The first method calls for plots of $p_1(\theta)$, $p_2(\theta)$ and $p_1(\theta)/p_2(\theta)$ for θ in $[0,\pi]$. The second requires the consistency of a set of linear inequalities to be checked. The second method may also be used for the generalized criterion in Sec. VI.

In Sec. III we examine three classes of special systems in detail. The results for the following canonical example[1] are typical: in a fourth-order system the poles are taken to be all real and repeated at ρ where $|\rho| < 1$. For $|\rho| \geq 0.669$ overflow oscillations are proven to exist. Tsypkin's criterion and the circle criterion guarantee the absence of oscillations for $|\rho| \leq 0.384$. The new test guarantees the absence of oscillations for $|\rho| \leq 0.610$.

In Sec. IV we apply the test to an eighth-order filter in the TDM-FDM translator, an extensively studied application of digital filtering. We find that this system can be lumped into two fourth-order sections employing saturation arithmetic, neither of which can sustain overflow oscillations. Both sections fail Tsypkin's criterion and the circle criterion tests.

Finally, in Sec. V we prove that the new criterion (i) easily gives the well-known result that overflow oscillations do not exist in second-order sections, and (ii) subsumes Tsypkin's criterion.

II. THE CRITERION

2.1 Passivity properties

On account of the special form of the nonlinearity F the system in eq. (1) possesses certain simply stated but important properties which we interpret as passivity properties. The criterion we derive is a direct consequence of these properties.

We may write eq. (1) as a linear system with a forcing sequence present by defining

$$e_n \triangleq F\left(\sum_1^m a_j x_{n-j}\right) - \left(\sum_1^m a_j x_{n-j}\right), \quad n = 0,1,\ldots \tag{6}$$

so that, from eq. (1),

$$x_n = \sum_{j=1}^m a_j x_{n-j} + e_n, \quad n = 0,1,\ldots \tag{7}$$

The above nonhomogeneous linear recursion is used throughout the paper. Our procedure is to translate the important features of $F(\cdot)$ into tractable constraints on $\{e_n\}$.

Every solution of eq. (1) possesses the following properties.

Proposition 1:

$$e_n(x_n - x_{n-1}) \leqslant 0, \quad n = 1,2,\ldots \tag{8}$$

and

$$e_n(x_n + x_{n-1}) \leqslant 0, \quad n = 1,2,\ldots \tag{9}$$

For proof we observe that if $|\Sigma\, a_j x_{n-j}| \leqslant 1$ then $e_n = 0$ and both conditions are obviously valid. If

$$\sum_{j=1}^m a_j x_{n-j} > 1 \quad \text{then} \quad x_n = 1 \text{ from (1)}$$

$$\text{and} \quad e_n < 0 \text{ from (6)}$$

and, of course, $|x_{n-1}| \leqslant 1$. Thus, in this case eqs. (8) and (9) are valid.

Similarly, if $\Sigma a_j x_{n-j} < -1$ then $x_n = -1$, $e_n > 0$ and as $|x_{n-1}| \leqslant 1$, the relations remain valid.

Equation (8) may be viewed as stating that the forcing term e_n has opposite sign from $(x_n - x_{n-1})$ which we may interpret as the discrete analog of "velocity." Thus locally the forcing term acts to reduce the velocity. Similarly interpreting $(x_n + x_{n-1})/2$ as the local "distance," eq. (9) states that locally the forcing term also acts to reduce the distance. For these reasons we view eqs. (8) and (9) as passivity properties.

Note that the two above conditions imply the following weaker condition:

$$e_n x_n \leqslant 0, \quad n = 1,2,\ldots \tag{10}$$

We find upon reflection that the latter condition is completely equivalent to the nonlinearity F in eq. (1) lying in the sector bounded by lines of slopes 0 and 1:

$$0 \leqslant F(\sigma)/\sigma \leqslant 1 \quad \text{for } all\ \sigma \tag{11}$$

This sector information is exploited in various criteria[5-11] but the additional information in eqs. (8) and (9) is not.

2.2 Equations for an oscillatory solution

We state the equations associated with every oscillatory solution of period N of the nonhomogeneous recursion in eq. (7). We find that the equations in matrix form involve a circulant matrix. We put into perspective some well-known results on circulants which are assembled in Appendix A.

A periodic solution of eq (7) with period N has associated with it the following set of N equations involving the coefficients $\{a_j\}$, the elements of the solution $X_1, X_2, \ldots, X_N$ and the corresponding forcing terms $E_1, E_2, \ldots, E_N$:

$$\begin{aligned} X_1 &= a_1 X_N + a_2 X_{N-1} + \cdots + a_m X_{N-m+1} + E_1 \\ X_2 &= a_1 X_1 + a_2 X_N + \cdots + a_m X_{N-m+2} + E_2 \\ &\vdots \\ X_N &= a_1 X_{N-1} + a_2 X_{N-2} + \cdots + a_m X_{N-m} + E_N \end{aligned} \tag{12}$$

In matrix form,

$$\mathbf{MX} = \mathbf{E}. \tag{13}$$

The interesting feature of the $N \times N$ matrix $\mathbf{M}$ is that it is a circulant[15-17] since it is the following polynomial in the primitive $N \times N$ circulant matrix $\mathbf{P}$ (see Appendix A for definition of $\mathbf{P}$):

$$\mathbf{M} = \mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j \tag{14}$$

The circulant matrices have been extensively studied in the past and we are in the fortunate position of knowing a great deal of their eigenstructure.† In particular, the eigenvalues of **M** are

$$1 - \sum_{j=1}^{m} a_j e^{-ijk2\pi/N}, \quad k = 1,2,\ldots,N \tag{15}$$

The eigenvectors of circulants are also known. The following remarkable property of the eigenvectors of circulants is of utmost importance in the paper (see next section): all $N \times N$ circulants have an identical set of eigenvectors, i.e., the eigenvectors do not depend on the constituents of the matrix. Thus, the eigenvectors of any $N \times N$ circulant are‡:

$$\mathbf{u}_k = \frac{1}{\sqrt{N}} \{e^{-ik(2\pi/N)}, e^{-i2k(2\pi/N)}, \ldots, e^{-i(N-1)k(2\pi/N)}, 1\}^* \tag{16}$$

$k = 1,2,\ldots,N$. Thus the real and imaginary components of the elements of each eigenvector are sequences of equiseparated samples of a sine function. Although these are complex vectors and circulants are not generally symmetric, the eigenvectors of circulants share an important property with eigenvectors of symmetric matrices in that they form an orthonormal set, i.e.,

$$\mathbf{u}_k^* \mathbf{u}_l = \delta_{kl}. \tag{17}$$

In matrix notation,

$$\mathbf{U}^*\mathbf{U} = \mathbf{I} \tag{18}$$

where the eigenvectors $\{\mathbf{u}_k\}$ have been arranged as columns of the matrix **U**.

2.3 Another representation of the passivity properties

We combine the above information with the passivity properties stated in Proposition 1 to obtain a compact and useful representation of the passivity properties that are valid if an oscillatory solution to (1) exists. As in the preceding section an oscillatory solution is assumed to be of period N.

Note that we may write

$$(X_N, X_1, \ldots, X_{N-1}) = (X_1, X_2, \ldots, X_N)\mathbf{P}' = \mathbf{X}'\mathbf{P}' \tag{19}$$

† Recently we have had another occasion[18] to use the eigenstructure of the matrix M. Willson[19] investigates the matrix M from a different angle.

‡ We denote the conjugate transpose by the superscript *. In the case of real matrices it is also denoted by the superscript '.

where $\mathbf{P}$ is the primitive $N \times N$ circulant. Thus

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) = \mathbf{X}'(\mathbf{I} - \mathbf{P}')\mathbf{E}$$

$$= \mathbf{X}'(\mathbf{I} - \mathbf{P}')\mathbf{M}\mathbf{X}, \quad \text{from (13)}$$

$$= \mathbf{X}'\left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j - \mathbf{P}^{N-1} + \sum_{j=1}^{m} a_j \mathbf{P}^{j-1}\right]\mathbf{X} \tag{20}$$

where in the final step we have used, in addition to the expression for $\mathbf{M}$, the relations $\mathbf{P}' = \mathbf{P}^{N-1}$ and $\mathbf{P}^N = \mathbf{I}$. The key observation about eq. (20) is that the matrix there, being a polynomial in $\mathbf{P}$, is a circulant.

We undertake a convenient change of coordinates to diagonalize the matrix in eq. (20). Let

$$\mathbf{Z} \triangleq \mathbf{U}^*\mathbf{X} \tag{21}$$

where $\mathbf{U}$ is, as in Sec. 2.2, the unitary matrix of eigenvectors of $N \times N$ circulants. Denoting the known eigenvalues (see Appendix A) of the matrix in eq. (20) by μ_k, k = 1,. . . ,N, we obtain

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 \operatorname{Re} \mu_k \tag{22}$$

Now,

$$\operatorname{Re} \mu_k = 1 - \cos\{k(2\pi/N)\} - \sum_{j=1}^{m} a_j[\cos\{jk2\pi/N\} - \cos\{(j-1)k2\pi/N\}] \tag{23}$$

To put eqs. (22) and (23) into the most convenient form, define the polynomial $p_1(\cdot)$ where

$$p_1(\theta) \triangleq 1 - \cos\theta - \sum_{j=1}^{m} a_j\{\cos j\theta - \cos(j-1)\theta\} \tag{24}$$

We then have

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N). \tag{25}$$

We proceed in identical fashion to obtain a similar expression corresponding to the other passivity condition in Proposition 1. Note that

$$\sum_{n=1}^{N} E_n(X_n + X_{n-1}) = \mathbf{X}'\left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j + \mathbf{P}^{N-1} - \sum_{j=1}^{m} a_j \mathbf{P}^{j-1}\right]\mathbf{X} \tag{26}$$

Because of the previously mentioned (and crucial) property that all N

$\times N$ circulants have identical sets of eigenvectors, the diagonalizing transformation is same as the one undertaken previously in eq. (21). Hence

$$\sum_{n=1}^{N} E_n(X_n + X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 \operatorname{Re} \lambda_k \tag{27}$$

where we have denoted the eigenvalues of the matrix in eq. (26) by $\{\lambda_k\}$. Here

$$\operatorname{Re} \lambda_k = 1 + \cos \{k2\pi/N\} - \sum_{j=1}^{m} a_j [\cos \{jk2\pi/N\} + \cos \{(j-1)k2\pi/N\}] \tag{28}$$

$k = 1,2, \ldots, N$. Thus for the final form we obtain

$$\sum_{n=1}^{N} E_n(X_n + X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \tag{29}$$

where the polynomial $p_2(\cdot)$ is defined to be

$$p_2(\theta) = 1 + \cos \theta - \sum_{j=1}^{m} a_j \{\cos j\theta + \cos (j-1)\theta\} \tag{30}$$

Now certainly Proposition 1 implies that

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) \leqslant 0 \quad \text{and} \quad \sum_{n=1}^{N} E_n(X_n + X_{n-1}) \leqslant 0 \tag{31}$$

The above, together with eqs. (25) and (27), yields:

Proposition 2: If a periodic solution of period N with elements $(X_1, X_2, \ldots, X_N)$ exists for the recursion in eq. (1) then

$$\sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N) \leqslant 0 \tag{32}$$

and

$$\sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \leqslant 0 \tag{33}$$

where $\mathbf{Z}$, as given in eq. (21), is a transform of $\mathbf{X}$ and $p_1(\theta)$ and $p_2(\theta)$, given in eqs. (24) and (30), are two polynomials in $\cos \theta$ of degree equal to the order of the system of eq. (1).

For a fourth-order system ($m = 4$) the two polynomials are

$$p_1(\theta) = (1 + a_1) - (1 + a_1 - a_2) \cos \theta - (a_2 - a_3) \cos 2\theta - (a_3 - a_4) \cos 3\theta - a_4 \cos 4\theta \tag{34}$$

and

$$p_2(\theta) = (1 - a_1) + (1 - a_1 - a_2) \cos \theta - (a_2 + a_3) \cos 2\theta - (a_3 + a_4) \cos 3\theta - a_4 \cos 4\theta \tag{35}$$

The polynomials for second- and third-order systems are obtained from the above by setting $a_3 = a_4 = 0$ and $a_4 = 0$, respectively.

In Fig. 3a and b we have plotted $p_1(\theta)$ and $p_2(\theta)$ for a particular fourth-order system.

2.4 The main result

It is only a short step from Proposition 2 to the main result which is

Theorem 1: If for any $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$,

$$\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) > 0 \text{ for } all\ \theta\ in\ [0,\pi] \tag{36}$$

then nontrivial periodic oscillations do not exist as solutions to eq. (1).

Proof: The proof is by contradiction. Suppose a nontrivial (i.e., $\mathbf{X} \neq 0$) periodic solution of period N exists and also that the hypothesis of the theorem is valid. Then for such a solution

$$\begin{aligned}\alpha_1 \sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N) + \alpha_2 \sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \\ = \sum_{k=1}^{N} |Z_k|^2 \{\alpha_1 p_1(k2\pi/N) + \alpha_2 p_2(k2\pi/N)\} \\ > 0 \end{aligned} \tag{37}$$

from the hypothesis. However, from the passivity conditions summarized in Proposition 2,

$$\alpha_1 \sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N) + \alpha_2 \sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \leq 0 \tag{38}$$

which is a contradiction. QED.

Note that if it is desirable to know only that oscillations of a particular period N do not exist for eq. (1) then the following is a sufficient condition:

There exist

$$\alpha_1 \geq 0, \quad \alpha_2 \geq 0 \text{ such that } \alpha_1 p_1(k2\pi/N) + \alpha_2 p_2(k2\pi/N) > 0 \tag{39}$$

for $k = 1,2,\ldots,N$.

2.5 How to use the new test

Given an mth-order system, there are two simple and straightforward ways in which the above result may be used to determine whether the system does not admit overflow oscillations.

The first method requires $p_1(\theta)$, $p_2(\theta)$, and $p_1(\theta)/p_2(\theta)$ to be plotted

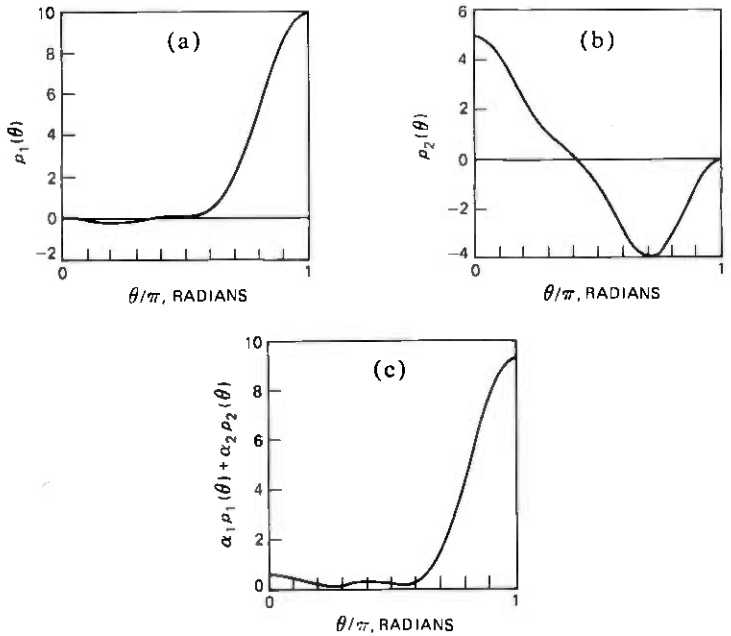


Fig. 3—Plots of polynomials $p_1(\theta)$, $p_2(\theta)$, and $\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta)$ for the following fourth-order system: $a_1 = 1.1015710, a_2 = -1.6571120, a_3 = 0.7733805, a_4 = -0.45135546$. In (c), $\alpha_1 = .92348761, \alpha_2 = .16965636$

for θ in the interval $[0,\pi]$. When such plots are available the first step is to see if there is any θ for which both $p_1(\theta)$ and $p_2(\theta)$ are negative; if this is the case the hypothesis of Theorem 1 obviously cannot be satisfied and the test is automatically failed. Assuming that this is not the case we find upon reflection that the hypothesis of Theorem 1 is satisfied if and only if

$$\max_{\{\theta \mid p_1(\theta)>0, p_2(\theta)\leq 0\}} [p_1(\theta)/p_2(\theta)] < \min_{\{\theta \mid p_1(\theta)\leq 0, p_2(\theta)>0\}} [p_1(\theta)/p_2(\theta)] \tag{40}$$

In fact, if the above is true the interval defined by the left- and right-hand sides of (40) is not empty and the hypothesis of Theorem 1 is satisfied by taking $\alpha_1 = 1$ and $-\alpha_2$ to be any value in the interval.

In summary the procedure is as follows: first check to see if p_1 and p_2 are both negative at the same point. If so, then the test is failed; if not, proceed to determine the intervals where $\{p_1(\theta) > 0, p_2(\theta) \leq 0\}$ and where $\{p_1(\theta) \leq 0, p_2(\theta) > 0\}$. The test is passed (i.e., no overflow oscillations exist) if and only if the maximum of $p_1(\theta)/p_2(\theta)$ in the first interval is less than the minimum of $p_1(\theta)/p_2(\theta)$ in the latter interval.

In the second method we finely discretize the interval for θ, $[0,\pi]$, and evaluate $p_1(\theta)$ and $p_2(\theta)$ at all the discrete points $\{\theta_j\}$. Testing for the

hypothesis of Theorem 1 then amounts to testing for the consistency of the following set of linear inequalities

$$(\alpha_1\ \alpha_2)\begin{bmatrix}1 & 0 & p_1(\theta_1) & p_1(\theta_2)\ \cdots\ p_1(\pi)\\ 0 & 1 & p_2(\theta_1) & p_2(\theta_2)\ \cdots\ p_2(\pi)\end{bmatrix} \geq (1\ 1\ \ldots\ 1) \qquad (41)$$

(There is no loss of generality in assuming that the right-hand side is as specified above.) There are standard procedures[20,21] for testing for consistency of linear inequalities. In any case, Phase 1 of any commercially available linear programming package does precisely this. If a linear programming package is used then the following (dummy) functional may be used in the program: minimize $(\alpha_1 + \alpha_2)$.

The above method is easily adapted to the generalization of the criterion which is developed in Sec. VI.

III. EXAMPLES

We consider three classes of examples in some detail. In each case we tested the criterion by following the second method outlined above. We used a linear programming package (written in machine language) made available to us by A. M. Odlyzko; the interval $[0,\pi]$ was subdivided into 100 intervals. In every case the computation time was of the order of a second.

3.1 Example 1: third-order system with repeated real roots

In this class of examples we take the coefficients to depend on a real number ρ, $|\rho| < 1$, in the following manner:

$$a_1 = -3\rho, \quad a_2 = -3\rho^2, \quad a_3 = -\rho^3 \qquad (42)$$

A third-order system with the above coefficients corresponds to an underlying linear system with characteristic polynomial $(\lambda + \rho)^3$, i.e., the linear system possesses three real roots all repeated at $-\rho$. In the investigation reported in Ref. 1 we found this class of systems to be interesting for various reasons. Also, for $|\rho|$ close to 1 the behavior of the system is to some extent representative, at least with respect to oscillatory behavior, of low-pass systems and high-pass systems, depending upon whether ρ is negative or positive respectively.

In Ref. 1 we showed for system (1) that

$$|\rho| \geq 0.858 \Rightarrow \text{period-3 oscillations exist} \qquad (43)$$

Tsypkin's criterion and the circle criterion (see Sec. 5.2) give

$$|\rho| \leq 0.500 \Rightarrow \text{no overflow oscillations exist} \qquad (44)$$

An application of the new test yields

$$|\rho| \leq 0.785 \Rightarrow \text{no overflow oscillations exist}$$

Thus, in this class of examples the new test makes a substantial contribution in reducing the indeterminate region to $0.785 < |\rho| < 0.858$.

3.2 Example 2: fourth-order system with repeated real roots

The class of examples considered here is a natural extension to a higher order, $m = 4$, of the class considered in the previous example. Again all the coefficients are determined by one real parameter ρ where $|\rho| < 1$:

$$a_1 = -4\rho, \quad a_2 = -6\rho^2, \quad a_3 = -4\rho^3, \quad a_4 = -\rho^4 \tag{45}$$

Thus in this example the underlying linear system possesses four real roots, all repeated at $-\rho$.

By examining the natural set of four equations associated with a periodic solution of period 4, see eq. (12), it is easy to see that a periodic solution with elements $(1,1,-1,-1)$ exists if and only if

$$(a_4 - a_2) \geq 1 + |a_1 - a_3| \tag{46}$$

Thus, we find on substituting for the a's that

$$|\rho| \geq 0.669 \Rightarrow \text{period-4 oscillations exist} \tag{47}$$

Tsypkin's criterion and the circle criterion give

$$|\rho| \leq 0.384 \Rightarrow \text{no overflow oscillations exist} \tag{48}$$

Application of the new criterion gives

$$|\rho| \leq 0.610 \Rightarrow \text{no overflow oscillations exist} \tag{49}$$

Thus we find that in this example too the new criterion makes an effective contribution in determining the region of stability.

3.3 Example 3: fourth-order filter for sample rate conversion

The example we consider now, a fourth-order system, was designed originally for interpolation and filtering for a terminator in a local digital switch.[4] We have reported previously[1] that in its original form the filter using saturation arithmetic sustained overflow oscillations. Here we vary one of the parameters in the design in order to estimate the modification required to guarantee the absence of oscillations. We find that the requisite variation is large. However, in the process we obtain a measure of the effectiveness of the new criterion.

The example we consider has two pairs of complex poles

$$\lambda_{1,2} = \rho_1 e^{\pm i\theta_1}, \quad \lambda_{3,4} = \rho_2 e^{\pm i\theta_2} \tag{50}$$

(The coefficients of the system are not of much interest; however, they

may be obtained from the information given below.) Also

$$\rho_1 = 0.786427817, \quad \theta_1 = 37.309784226 \text{ degrees} \tag{51}$$

and

$$\theta_2 = 39.675296075 \text{ degrees}$$

We vary ρ_2 keeping ρ_1, θ_1, and θ_2 fixed at the above values; in the original design $\rho_2 = 0.952851183$.

In (46) we have given a condition for the existence of limit cycles of period 4 with elements (1,1,−1,−1). Translating (46) to the present examples gives

$$\rho_2 \geq 0.671 \Rightarrow \text{period-4 oscillations exist} \tag{52}$$

Tsypkin's criterion and the circle criterion give

$$\rho_2 \leq 0.070 \Rightarrow \text{no oscillations exist} \tag{53}$$

An application of the new criterion gives

$$\rho_2 \leq 0.665 \Rightarrow \text{no oscillations exist} \tag{54}$$

This is a rather striking example of the effectiveness of the new criterion.

IV. AN APPLICATION

Here we examine a particular eighth-order system* which has been used in an applied research project[22] on a TDM/FDM translator.[23] The latter, a system for translating between analog frequency-division and digital time-division signals, is an extensively studied application of digital filtering. The eighth-order system has been designed to function as a low-pass filter with a sampling frequency of 8 kHz and a cutoff frequency of 2 kHz. Our object here is to demonstrate through an application of the new criterion that it is possible to design the filter as a cascade of two fourth-order sections both employing saturation arithmetic such that no overflow oscillations are sustained in either section. At least as far as overflow oscillations are concerned the margin of safety is adequate so that small changes in the coefficients due to quantization of coefficients, for example, are not going to cause overflow oscillations to appear. It should be emphasized that the result here is not a substitute for a design study and the structure suggested may well turn out to be unacceptable on grounds not related to overflow oscillations.

The system has four pairs of complex poles; the modulus (ρ_i) and

* I am grateful to V. B. Lawrence for bringing this system to my attention.

argument ($\pm\theta_i$) of each pair is as follows:

$$\rho_1 = 0.5115846\,, \quad \theta_1 = 32.870 \text{ degrees}$$

$$\rho_2 = 0.980274196, \quad \theta_2 = 80.828 \text{ degrees}$$

$$\rho_3 = 0.75259969\text{-}, \quad \theta_3 = 64.482 \text{ degrees}$$

$$\rho_4 = 0.892679\,, \quad \theta_4 = 75.297 \text{ degrees}$$

We group the first and second pairs of poles together to form one fourth-order section and the remaining pairs to form the second fourth-order section. The resulting coefficients of the two sections are, respectively,

$$a_1 = 1.1718731, \quad a_2 = -1.4912153, \quad a_3 = 0.9075846, \quad a_4 = -0.2514954 \tag{56}$$

$$a_1 = 1.1015710, \quad a_2 = -1.6571120, \quad a_3 = 0.7733805, \quad a_4 = -0.45135546 \tag{57}$$

Both sections pass the new test. For the first section it may be ascertained that with

$$\alpha_1 = 6.0819413 \text{ and } \alpha_2 = 0.07538601 \tag{58}$$

the hypothesis of Theorem 1 is satisfied. In fact, the polynomial $p_1(\theta)$ is itself positive everywhere except at $\theta = 0$, where its value is 0. However, $p_2(0) > 0$. Thus, any positive choice of α_1 and α_2 chosen suitably small will satisfy the hypothesis of Theorem 1.

For the second section (57), a choice of α_1 and α_2 for which $\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) > 0$ for all θ is

$$\alpha_1 = .92348761 \text{ and } \alpha_2 = .16965636 \tag{59}$$

Plots of $p_1(\theta)$, $p_2(\theta)$, and $\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta)$ for the second section are displayed in Fig. 3.

It is noteworthy that both sections fail Tsypkin's criterion and the circle criterion.

V. SOME IMPLICATIONS OF THE MAIN RESULT (THEOREM 1)

5.1 Overflow oscillations do not exist in second-order systems

It is well known[12,13,14] that when the order of the system in eq. (1) is two, then overflow oscillations are not sustained. The proofs of this result are rather special to second-order systems and to the saturation arithmetic. On the other hand, there are the frequency-domain criteria[5-11] for stability which are systems-theoretical results applicable to large classes of nonlinearities and systems of arbitrary order. However, we may

infer from the results in Ref. 8 that these criteria do not give the result that all second-order systems are free from overflow oscillations.

We show that the criterion in Theorem 1 does give the well-known result on second-order systems. Our result is given in Proposition 3. [It is assumed that $|a_2| < 1$ and $1 - |a_1| - a_2 > 0$; these relations are equivalent to eq. (3), i.e., the underlying linear system is stable.]

Proposition 3: Let $m = 2$ in eq. (1). Also let

$$\alpha_1 = (1 + a_1 - a_2) > 0 \text{ and } \alpha_2 = (1 - a_1 - a_2) > 0 \tag{60}$$

Then,

$$\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) > 0 \text{ for } all\ \theta \tag{61}$$

The proof of this result is in Appendix B. The above in conjunction with Theorem 1 shows that oscillations are not sustained in second-order systems.

5.2 Tsypkin's criterion and discrete circle criterion are subsumed by new criterion

The object here is to show that the new criterion subsumes both Tsypkin's criterion[5] and the discrete circle criterion[11] when the latter criteria are used to determine the nonexistence of oscillations in eq. (1). The two closely related frequency-domain criteria are identical when applied to the system in eq. (1).

Tsypkin's criterion[5] is as follows in applications to systems like eq. (1) where the nonlinearity F satisfies

$$K_{\min} \leqslant F(\sigma)/\sigma \leqslant K_{\max} \quad \text{for } all\ \sigma \tag{62}$$

If

$$(i)\ \sum a_j z^{-j}/[1 - K_{\min} \sum a_j z^{-j}] \text{ is finite for } all\ |z| \geqslant 1 \tag{63}$$

$$(ii)\ \frac{1}{K_{\max} - K_{\min}} - \text{Re}\,[\sum a_j e^{-ij\theta}/(1 - K_{\min} \sum a_j e^{-ij\theta})] > 0 \text{ for } all\ \theta \text{ in } [0,2\pi] \tag{64}$$

then $\lim_{n\to\infty} x_n = 0$; in particular, oscillations do not exist.

In the case of eq. (1) where F is the saturation nonlinearity,

$$K_{\min} = 0 \quad \text{and} \quad K_{\max} = 1 \tag{65}$$

so that the effective restriction is (64) which reduces to

$$1 - \sum_{j=1}^{m} a_j \cos j\theta > 0 \quad \text{for } all\ \theta \tag{66}$$

When (65) holds the discrete time version of the circle criterion[11] is identical to the above condition.

In Theorem 1 let $\alpha_1 = \alpha_2 = \frac{1}{2}$. Then from the defining relations for $p_1(\theta)$ and $p_2(\theta)$ in (24) and (30) respectively, we find that

$$\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) = 1 - \sum_{j=1}^{m} a_j \cos j\theta \tag{67}$$

Thus, as previously asserted, if either Tsypkin's criterion or the circle criterion is satisfied, i.e., (66) is valid, then the hypothesis of Theorem 1 is also satisfied.

VI. A GENERALIZATION OF THE MAIN RESULT

The reader will recall that the main result, Theorem 1, is a direct consequence of the rather special passivity properties, stated in Sec. 2.1, which are implied by the special features of the saturation nonlinearity F. Another key ingredient is that the passivity conditions imply inequalities on quadratic forms involving circulants. We show here that many conditions akin to the ones in Proposition 1 are valid by virtue of the properties of the saturation nonlinearity. All or some of these may be used to augment the passivity conditions used so far so as to obtain improved criteria for the nonexistence of oscillations.

The following generalized passivity conditions exist* for any $l \geq 1$:

$$e_n(x_n - x_{n-l}) \leqslant 0 \quad n = l, l+1, \ldots \tag{68}$$

$$e_n(x_n + x_{n-l}) \leqslant 0 \quad n = l, l+1, \ldots \tag{69}$$

where $\{x_n\}$ is any solution of eq. (1) and $\{e_n\}$ is obtained from the solution through eq. (6). The proof is similar to that of Proposition 1. Thus in Proposition 1 we have used only a very small subset ($l = 1$) of all the above conditions.

The interesting fact is that each of the expressions in the above conditions summed over N, where N is the period of any periodic solution of (1), is equivalent to a quadratic form involving a circulant. Thus if $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$ are the elements of the periodic solution and $(E_1, E_2, \ldots, E_N)$ are the corresponding forcing terms, see eqs. (12) and (13), then

$$\sum_{n=1}^{N} E_n(X_n - X_{n-l}) = \mathbf{X}' \left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j - \mathbf{P}^{-l} - \sum_{j=1}^{m} a_j \mathbf{P}^{-l+j} \right] \mathbf{X} \tag{70}$$

* The generalized passivity conditions are also valid for negative values of l although we do not make any explicit use of this fact.

$$\sum_{n=1}^{N} E_n(X_n + X_{n-l}) = \mathbf{X}' \left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j + \mathbf{P}^{-l} + \sum_{j=1}^{m} a_j \mathbf{P}^{-l+j} \right] \mathbf{X} \tag{71}$$

for $l = 1,2, \ldots$. Hence by transforming $\mathbf{X}$ to $\mathbf{Z}$ where $\mathbf{Z} = \mathbf{U}^*\mathbf{X}$, $\mathbf{U}$ being the unitary matrix of eigenvectors of $N \times N$ circulants, we obtain for $l = 1,2, \ldots$

$$\sum_{n=1}^{N} E_n(X_n - X_{n-l}) = \sum_{k=1}^{N} |Z_k|^2 p_l(k2\pi/N) \tag{72}$$

and

$$\sum_{n=1}^{N} E_n(X_n + X_{n-l}) = \sum_{k=1}^{N} |Z_k|^2 p'_l(k2\pi/N) \tag{73}$$

where

$$p_l(\theta) \triangleq 1 - \cos l\theta - \sum_{j=1}^{m} a_j \{\cos j\theta + \cos (j - l)\theta\} \tag{74}$$

and

$$p'_l(\theta) \triangleq 1 + \cos l\theta - \sum_{j=1}^{m} a_j \{\cos j\theta - \cos (j - l)\theta\} \tag{75}$$

Thus $p_1(\theta)$ and $p_2(\theta)$ defined in Sec. 2.3 correspond to $p_1(\theta)$ and $p'_1(\theta)$ respectively in the present notation.

Certainly the generalized passivity condition in (68) and (69) imply that the expressions in (72) and (73) are nonpositive. We thus arrive at the following generalization of Theorem 1:

Theorem 2: If any convex linear combination of $p_1(\theta), p'_1(\theta), p_2(\theta), p'_2(\theta), \ldots$ is positive for all θ in $[0,\pi]$, then the system in eq. (1) does not have any nontrivial periodic solutions.

In experiments involving fourth-order systems of practical interest we have not found the use of the above generalized criterion to make any material difference in delineating the stable systems. In these investigations we used a linear programming package (Sec. 2.5) to apply the test in Theorem 2 with up to six polynomials (the leading six polynomials of Theorem 2) being used. However, it is quite possible for substantial improvements to exist in other cases.

ACKNOWLEDGMENTS

We gratefully acknowledge our debt to A. M. Odlyzko for making available and in assisting us in the use of a linear programming package. We are grateful to V. B. Lawrence for bringing the filter in Section IV to our attention.

APPENDIX A

Circulant matrices

For completeness we collect here some of the well-known properties of circulants which are used in the paper. The interested reader may refer to Muir[16] and Grenander and Szego[17] for further details and applications; Ref. 15 concisely lists some of the main properties.

We let $\mathbf{P}$ denote the primitive $N \times N$ circulant:

$$\mathbf{P} = \begin{bmatrix} 0 & — & — & — & — & 0 & 1 \\ 1 & 0 & — & — & — & — & 0 \\ 0 & 1 & 0 & — & — & — & 0 \\ — & — & — & — & — & — & — \\ 0 & — & — & — & — & 1 & 0 \end{bmatrix} \tag{76}$$

Note that

$$\mathbf{P}^N = I \tag{77}$$

and that

$$\mathbf{P}' = \mathbf{P}^{N-1} = \mathbf{P}^{-1} \tag{78}$$

A polynomial of arbitrary degree in $\mathbf{P}$ is a circulant. An $N \times N$ circulant $\mathbf{C}$,

$$\mathbf{C} = \sum_{j=0}^{N-1} c_j P^j \tag{79}$$

has as its eigenvalues

$$\sum_{j=0}^{N-1} c_j e^{-ijk2\pi/N} \quad k = 1,2,\ldots,N \tag{80}$$

All $N \times N$ circulants have as eigenvectors $\mathbf{u}_k$, $k = 1, \ldots, N$, given in eq. (16). The matrix $\mathbf{U}$ with the eigenvectors as columns is unitary, i.e.,

$$\mathbf{U}^*\mathbf{U} = I \tag{81}$$

APPENDIX B

Proof of proposition 3

We prove here the assertion in Proposition 3, namely, for second-order systems

$$q(\theta) \triangleq (1 + a_1 - a_2)p_1(\theta) + (1 - a_1 - a_2)p_2(\theta) > 0 \quad \text{for } all\ \theta \tag{82}$$

For second-order systems

$$p_1(\theta) = (1 + a_1) - (1 + a_1 - a_2)\cos\theta - a_2 \cos 2\theta \tag{83}$$

and,

$$p_2(\theta) = (1 - a_1) + (1 - a_1 - a_2)\cos\theta - a_2\cos 2\theta \tag{84}$$

We find upon substitution that

$$q(\theta) = -4a_2(1 - a_2)\cos^2\theta - 4a_1(1 - a_2)\cos\theta + 2(1 + a_1^2 - a_2^2) \tag{85}$$

First observe that at the corner points q is positive:

$$q(0) = 2(1 - a_1 - a_2)^2 > 0 \text{ and } q(\pi) = 2(1 + a_1 - a_2)^2 > 0 \tag{86}$$

Through differentiation we find that minima of $q(\theta)$ occur in the interior of the region $[0,\pi]$ if and only if

$$|a_1| \leqslant -2a_2 \tag{87}$$

and that at a minimum $\hat{\theta}$,

$$\cos\hat{\theta} = -a_1/2a_2 \tag{88}$$

Evaluating q at such a point we obtain

$$q(\hat{\theta}) = \frac{-(1 + a_2)}{a_2}[-2a_2(1 + a_2) + (4a_2^2 - a_1^2)] \tag{89}$$

$$> 0$$

REFERENCES

1. Debasis Mitra, "Large Amplitude, Self-Sustained Oscillations in Difference Equations Which Describe Digital Filter Sections Using Saturation Arithmetic," IEEE Trans. Acoustics, Speech, Signal Proc., April 1977.
2. A. Croisier, D. J. Esteban, M. E. Levilion and V. Rizo, "Digital Filter for PCM Encoded Signals," U. S. Patent 3777130, December 3, 1973.
3. A. Peled and B. Liu, "A New Hardware Realization of Digital Filters," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-22*, December 1974, pp. 456–462.
4. R. B. Kieburtz, "Interpolation and Filtering for the T-Line Terminator," Bell Laboratories internal memorandum, March 1975. Also, "Digital Interpolation Interface Between Two Systems at Slightly Different Sampling Rates," presented at IEEE Workshop on Signal Processing, Arden House, 1976.
5. Ya. Z. Tsypkin, "Fundamentals of the Theory of Non-Linear Pulse Control Systems," Proc. Second Intl. Cong., Intl. Fed. of Automatic Control, Basel, 1963, pp. 172–180.Also, "A Criterion for Absolute Stability of Automatic Pulse-Systems with Monotonic Characteristics of the Non-Linear Element," Sov. Phys. Dokl., *9*, October 1964, pp. 263–266.
6. E. D. Garber, "Frequency Criteria for the Absence of Periodic Responses," Automat. Remote Contr., *28*, No. 11, November 1967.
7. A. I. Barkin, "Sufficient Conditions for the Absence of Auto-Oscillations In Pulse Systems," Automat. Remote Contr., *31*, June 1970, pp. 942–946.
8. T. Claasen, W. F. G. Mecklenbraüker, and J. B. H. Peek, "Frequency Domain Criteria for the Absence of Zero-Input Limit Cycles in Nonlinear Discrete-Time Systems, With Applications to Digital Filters," IEEE Trans. Circuits Syst., *CAS-22*, No. 3 (March 1975) pp. 232–239.
9. I. W. Sandberg, "On the Boundedness of Solutions of Nonlinear Integral Equations," B.S.T.J., *44*, No. 3 (March 1965), pp. 439–453.
10. G. P. Szego, "On the Absolute Stability of Sampled-Data Control Systems," Proc. Nat. Acad. Sci., *50*, 1963, pp. 558–560.
11. J. L. Willems, *Stability Theory of Dynamical Systems*, New York: John Wiley, 1970; Ch. 6.

12. P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow Oscillations in Digital Filters," B.S.T.J., *48*, No. 9 (November 1969), pp. 2999–3020.
13. I. W. Sandberg, "A Theorem Concerning Limit Cycles in Digital Filters," Proc. 7th Ann. Allerton Conf. Circuits and Systems Theory, pp. 63–68, 1969.
14. A. N. Willson, Jr., "Limit Cycles Due to Adder Overflow in Digital Filters," IEEE Trans. Circuit Theory, *CT-19*, 1972, pp. 342–346.
15. M. Marcus, "Basic Theorems in Matrix Theory," Nat. Bur. Stds., Applied Math. Series 57, January 1960, p. 9.
16. T. Muir, *A Treatise On The Theory of Determinants*, Dover Publications, New York, 1960; Ch. 12.
17. U. Grenander and G. Szego, *Toeplitz Forms And Their Applications*, Berkeley: Univ. of Calif. Press, 1958; Ch. 8.
18. Debasis Mitra, "A Bound on Limits in Digital Filters which Exploits a Particular Structural Property of the Quantization." A summary appears in Proc. 1977 IEEE Intl. Conf. on Acoustics, Speech, and Signal Proc., May 1977. The full paper is due for publication in IEEE Trans. Circuits Syst., Nov. 1977.
19. A. N. Willson, Jr., "Computation of the Periods of Forced Overflow Oscillations in Digital Filters," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-24*, No. 1 (February 1976).
20. D. Gale, "How to Solve Linear Inequalities," Amer. Math. Monthly, *76*, 1969, pp. 589–599.
21. G. Debreu, "Non-negative Solutions of Linear Inequalities," Internat. Economic Review, Vol. 5, 1964.
22. R. B. Kieburtz, V. B. Lawrence, and K. V. Mina, "Control of Limit Cycles in Recursive Digital Filters by Randomized Quantization," Bell Laboratories memorandum; also talk presented at Intl. Symp. Circuits and Systems, Munich, Germany, April 1976.
23. S. L. Freeny, R. B. Kieburtz, K. V. Mina, and S. K. Tewksbury, "Design of Digital Filters for an all Digital Frequency Division Multiplex-Time Division Multiplex Translator," IEEE Trans. Circuit Theory, *CT-18* No. 6 (November 1971).
24. S. Lefschetz, *Stability of Nonlinear Control Systems*, Academic Press, New York, 1965; Ch. 1 and 2.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Tandem Connections of Wideband and Narrowband Speech Communications Systems: Part 1—Narrowband-to-Wideband Link

By R. E. CROCHIERE, D. J. GOODMAN,
L. R. RABINER, and M. R. SAMBUR

(Manuscript received March 25, 1977)

The performance of a tandem connection of narrowband and wideband speech communication systems is evaluated. Specifically, the narrowband system consists of a conventional Linear Predictive Coding (LPC) vocoder operating at a bit-rate of 2.4 kb/s and the wideband system consists of a Continuously Variable Slope Delta modulator CVSD operating at a bit rate of 16 kb/s. In Part 1 of this paper the properties of the narrowband-to-wideband link are investigated and in Part 2 the properties of the wideband-to-narrowband link are investigated. In part 1 the SNR (signal-to-quantizing noise ratio) of the CVSD coder is analyzed over a 50-dB variation of the input signal levels and for a variety of source excitations for the LPC synthesizer. It is shown that SNR improvements in the CVSD coder of 2 to 2.5 dB are possible in the slope overload region of the coder by modifying the source excitation of the LPC synthesizer and by preprocessing the input signal to the coder with an allpass filter. Both methods aid in reducing the peak factor (peak-to-RMS level) of the input speech to the coder. Subjectively, however, only slight improvements in quality, if any, were observed with these modifications.

I. INTRODUCTION

Agencies of the United States government are currently formulating plans for an extensive digital secure voice communication network. In this network, a substantial fraction of the signals will be transmitted over "wideband" circuits at 16 kb/s. Owing to severe bandwidth constraints in some parts of the network, however, there will also be "narrowband" speech links in which the transmission rate is 2.4 kb/s. In preliminary plans, the wideband code format is CVSD (Continuously Variable Slope

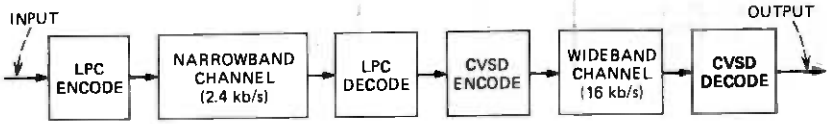


Fig. 1—Narrowband-to-wideband link.

Delta modulation) and the narrowband code format is LPC (Linear Predictive Coding).

Both of these coding methods have been studied extensively, and their performance over single transmission links (involving one encoding operation and one decoding operation) is now well understood.[1,2] In addition to creating single links, however, the proposed communication network will establish tandem connections containing both narrowband and wideband links. It is not clear a priori that two systems, each designed for single-link operation, will interact in tandem to provide acceptable overall quality. Existing knowledge of LPC and CVSD is of limited value in predicting tandem performance and yet the viability of the proposed network depends on adequate performance of tandem as well as single circuits. It is the purpose of this paper to describe the properties of CVSD and LPC that influence the performance of the narrowband-to-wideband connection shown in Fig. 1. A companion paper deals with the complementary wideband-to-narrowband connection.

Our study focuses on issues that arise in tandem links but not in individual circuits. In particular, in this paper we investigate the effects of the narrowband channel on CVSD signal-to-noise ratio (SNR). In doing so we have measured the SNR of the CVSD coder with an original speech input and compared it with the SNR when the CVSD input is LPC synthesized speech with a conventional (impulse) excitation source (during voiced intervals). With a view to improving the quality of tandem circuits, we have also investigated the effect of allpass filtering the LPC output and of using broadened excitation sources for voiced sounds in the LPC synthesizer.

The studies have been carried out by means of computer simulations on a Honeywell DDP 516 computer. In the narrowband-to-wideband tandem we have measured SNR as a function of CVSD input level for a variety of interface and LPC synthesizer source configurations. For each condition (i.e., a given input level, synthesizer source and interface) we have recorded two sentences transmitted through the tandem link. The SNR measurements as well as informal listening experience suggest that CVSD is a critical element in this tandem connection. It has been shown that combinations of interface filter and modified synthesizer source are effective (to some extent) in improving overall quality when the CVSD input level is high. In this case the delta modulator is subject to substantial slope overload. The overload is reduced both by prefiltering and

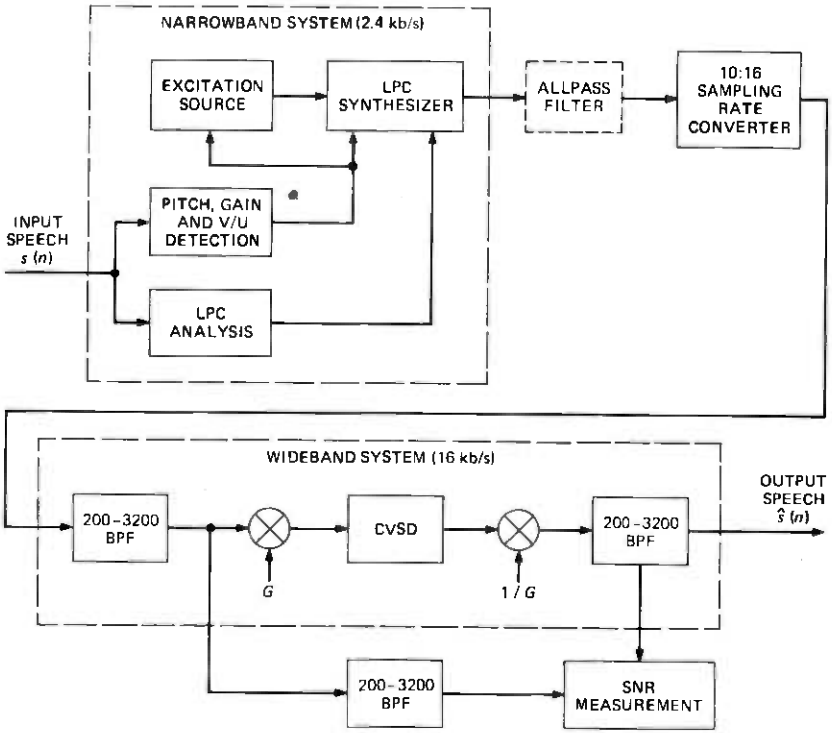


Fig. 2—Block diagram of LPC-to-CVSD link.

by broadening of the LPC synthesizer source because both of these methods reduce the ratio of peak-to-rms level of the signal at the CVSD input. Of the two methods, adjustment of the LPC excitation is more effective but does require some modification of the LPC link. All-pass filtering the LPC output signal has the advantage of being external to both CVSD and LPC.

II. OVERVIEW OF THE NARROWBAND TO WIDEBAND LINK

In this section we discuss the elements of the narrowband-to-wideband tandem connection. We will first review the basic operation of the LPC vocoder and the CVSD coder and will then discuss issues involved in connecting these two systems in a tandem link. After establishing a basic understanding of the various elements in this link, we will discuss factors which affect the performance of this connection and ways in which this performance can be improved.

Figure 2 shows a more detailed block diagram of the overall tandem connection. The narrowband system consists of an LPC analyzer and a pitch and voiced/unvoiced detector. The parameters from these two

analyses are used by the LPC synthesizer to resynthesize the speech waveform. An allpass network may be used for further post processing of this waveform. The details of this network will be discussed in a later section.

As shown in Fig. 2, the wideband system consists of a bandpass filter to prevent aliasing, the CVSD coder and another bandpass filter that suppresses CVSD quantizing noise. Gains G and $1/G$ are used in measuring the dynamic range (i.e., variations in performance as a function of signal level) of the CVSD coder.

The basic sampling rate of the narrowband system is 10 kHz and the sampling rate of the wideband system is 16 kHz. In order to interface these two systems a sampling rate converter is used. The details of this conversion will also be discussed in this section as well as the conversion from 16 kHz to 10 kHz which is required in the wideband-to-narrowband tandem connection.

2.1 The wideband system (CVSD)

Figure 3a is a block diagram of the CVSD coding process. The input speech signal is called $x(t)$. An approximation signal $y(t)$ is generated in the encoder feedback loop and at the kth sampling instant ($t = kT$, $T = 1/16000$ sec), the transmitted signal is $b_k = 1$ if

$$x(kT) = x_k > y_k = y(kT) \tag{1}$$

Otherwise $b_k = -1$. A positive output causes $y(t)$ to increase during the next sampling interval making y_{k+1} attain the value

$$y_{k+1} = \alpha y_k + H(1 - \alpha)\Delta_k \tag{2a}$$

where α is the leakage of the approximation signal integrator and $\Delta_k = \Delta(kT)$ is the kth step size. A negative output, $b_k = -1$, results in

$$y_{k+1} = \alpha y_k - H(1 - \alpha)\Delta_k \tag{2b}$$

The step size is obtained from another integrator which processes the output of an overload detector. The overload detector has output V when the three previous CVSD outputs are identical (all 1 or all -1). Otherwise the output of the overload detector is 0. To ensure that the minimum step size is nonzero a small quantity V_1 is added to the output of the overload detector. Thus, the step size satisfies the relation

$$\Delta_{k+1} = \beta\Delta_k + (1 - \beta)(V + V_1) \tag{3}$$

when three previous outputs are identical where β is the leakage of the step size integrator. Otherwise,

$$\Delta_{k+1} = \beta\Delta_k + (1 - \beta)V_1 \tag{4}$$

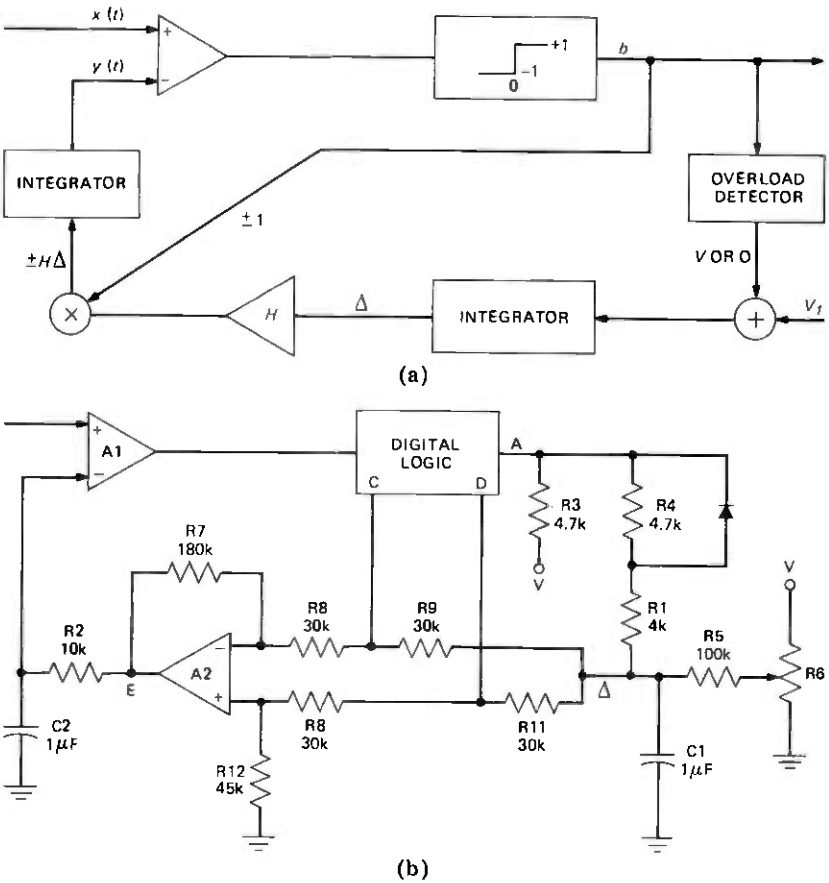


Fig. 3—(a) Block diagram of CVSD coder and (b) circuit implementation of CVSD coder.

Figure 3b shows the circuit implementation of these operations. In the digital logic, point A is the output of the overload detector. It is an open circuit when the three previous bits are all 1 or all −1. This open circuit condition allows the 1 μF capacitor to charge toward $+V$ through R1 and R3. When the last 3 bits are not identical, point A is grounded and the capacitor discharges to ground through R1 and R4. The potentiometer R6 establishes V_1 the minimum voltage on C1.

When $b_k = 1$, point C is grounded and point D is an open circuit. The gain of amplifier A2 is $H = 3$ and the voltage at point E is 3Δ. When $b_k = -1$, C is open circuited and D is grounded causing the voltage at E to be -3Δ. Thus the integrator R2-C2 charges toward ± 3 times the voltage on capacitor C1, depending on whether $b_k = \pm 1$.

The time constant of the step size integrator is 5.69 ms (1 μF × R1 +

R3 or R4 in parallel with R5 and with the input impedance of A2 which is 20 kΩ). The step size coefficient is therefore

$$\beta = \exp\left(-\frac{1}{16{,}000}\Big/ 5.69 \times 10^{-3}\right) = .99 \tag{5}$$

The time constant of the approximation signal integrator is 1 ms which gives

$$\alpha = \exp\left(-\frac{1}{16{,}000}\Big/ 10^{-3}\right) = .94 \tag{6}$$

In the computer simulation, speech is represented as a 16-bit integer between −32,768 and 32,767, so that a value of $V = 32{,}767$ is equivalent in hardware to a peak speech input equal to the supply voltage. In our studies we have provided for a wide dynamic range of step sizes, $V/V_1 = 200$ so that $V_1 = 164$.

Thus eqs. (3) and (4) are, numerically,

$$\Delta_{k+1} = .99\Delta_k + 329 \tag{7}$$

when three outputs are identical and

$$\Delta_{k+1} = .99\Delta_k + 1.64 \tag{8}$$

otherwise. Similarly eqs. (2a) and (2b) are,

$$y_{k+1} = .94y_k + .18\Delta_k \tag{9}$$

if $x_k > y_k$ and

$$= .94y_k - .18\Delta_k \tag{10}$$

otherwise.

2.2 The narrowband system (LPC)

The narrowband system consists of a Linear Predictive Coding (LPC) system based on an all-pole model of the speech production mechanism. The all-pole model implies that within a frame of speech, the output speech sequence is given by

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + Gu_n \tag{11}$$

where p is the number of poles, u_n is the appropriate input, G is the gain, and the a_k's are the LPC coefficients that represent the spectral characteristics of the speech frame. For a voiced speech segment, u_n is a sequence of pulses separated by the pitch period. If the segment is unvoiced, pseudorandom noise is used as input.

In our study, the LPC coefficients were calculated by the autocorrelation method with $p = 12$ (Ref. 2). The analysis was performed every

20 msec (50 times/sec) across overlapping 300 sample (30 msec) Hamming windowed speech frames. The pitch detection and V/U (voiced/unvoiced) decision is based on the modified autocorrelation method.[3] The effects of pitch and V/U analysis do not in general influence the performance of the narrowband-to-wideband link. In the reverse link (wideband to narrowband), however, the pitch and V/U analysis is strongly affected by the performance of the wideband system. Therefore we will discuss the pitch and V/U analysis in the accompanying paper on the wideband-to-narrowband link.

Since the stability and characterization of the LPC synthesizer is extremely sensitive to small perturbations in the LPC coefficients, it is not possible to achieve low-bit-rate coding by transmitting the LPC coefficients.[2] However, by transmitting either the log area coefficients or the parcor coefficients, a 2.4-kb/s vocoder is readily achieved.[4] The log area coefficients are related to the LPC coefficients by

$$g_i = \log \frac{1 + k_i}{1 - k_i} \tag{12}$$

where the k_i's are termed the parcor coefficients.[2] If we denote $a_i^{(j)}$ as the ith linear prediction coefficient for a jth-order linear-prediction model then

$$k_i = a_i^{(i)} \tag{13}$$

The parcor coefficients have the very important property that if

$$|k_i| < 1 \qquad i = 1, \ldots, p \tag{14}$$

then it is guaranteed that the linear prediction synthesizer is stable.[2] Thus, small perturbations in the parcor coefficients or log area coefficients will not affect the stability of the synthesizer, and morever these small perturbations will not seriously alter the spectral characterization of the speech segment.[5] Since the log area coefficients are slightly less sensitive to quantization error[5] they were transmitted in the narrowband system.

The quantization of the LPC control signals (pitch, gain, and the g_i's) was accomplished by ADPCM (Adaptive Differential PCM) techniques.[6] In this scheme, the value of a particular control parameter in the nth frame is initially estimated as equal to the transmitted values of the parameter in the (n − 1)st frame. The difference between this predicted value and the actual parameter value is then quantized using a gamma or laplace quantizer with an adaptive step size.[4,6] Complete details of the adaptation scheme and the quantization method are given in ref. 4.

The bit allotment in the narrowband link is as follows. The pitch and gain information is encoded with 3 bits/sample each. The first six log area

ratios $g_1, g_2, \ldots, g_6$ are each encoded with 4 bits/sample and $g_7, g_8, \ldots, g_{12}$ are encoded with 2 bits/sample. One bit/frame is used to transmit the V/U decision. This gives a total of 43 bits/frame or a transmission rate of 2.15 kb/s. Another 5 bits/frame are used for transmission of initialization information and frame synchronizing information giving a total of 48 bits/frame or a total transmission rate of 2.4 kb/s for the narrowband system.

2.3 Bandpass filters

The bandpass filters in Fig. 2 are all identical and are used to limit the bandwidth of the signal to the range 200 Hz to 3200 Hz. The third bandpass filter below the block diagram of the wideband system is used for compensating the group delay of the input signal of the CVSD in order to make meaningful SNR measurements on the CVSD.

The bandpass filters are eighth-order recursive elliptic filters with a passband ripple of 0.25 dB and a stopband attenuation greater than 35 dB. The average group delay of the filters is 0.325 msec in the passband and it peaks to 7 msec in the lower transition band. Figure 4 shows the log magnitude response (dB), group delay, and impulse response of these filters.

2.4 Sampling rate conversion

In the tandem connections it is necessary to convert the sampling rate of the signal from 10 kHz to 16 kHz and from 16 kHz to 10 kHz (in the opposite connection). One way of achieving this conversion is to convert the signal to analog form and then resample it at the new sampling rate. This approach is susceptible to electronic noise in the analog circuitry and is limited by the dynamic range of the analog components.

A more attractive approach to the sampling rate conversion process is to do a direct digital-to-digital conversion of the sampling rate. This conversion can be done as accurately as desired and is not prone to extraneous noise from electronic components. The digital-to-digital conversion is accomplished with the aid of a linear phase FIR digital interpolating filter whose output sample values are computed at a different sampling rate than the incoming samples.[7]

Figure 5a shows the frequency response of a 119-tap FIR lowpass filter which was used in the 10 kHz to 16 kHz conversion. Although the length of the filter is 119 samples, only 15 multiplications and additions per output sample are required in the conversion process because only a subset of the filter coefficients are needed in computing each output sample.[7] Similarly, Fig. 5b shows the frequency response of a 127-tap linear phase FIR filter used in the 16 kHz to 10 kHz sampling rate conversion. In this case 26 multiplies and adds are used in computing each output sample.

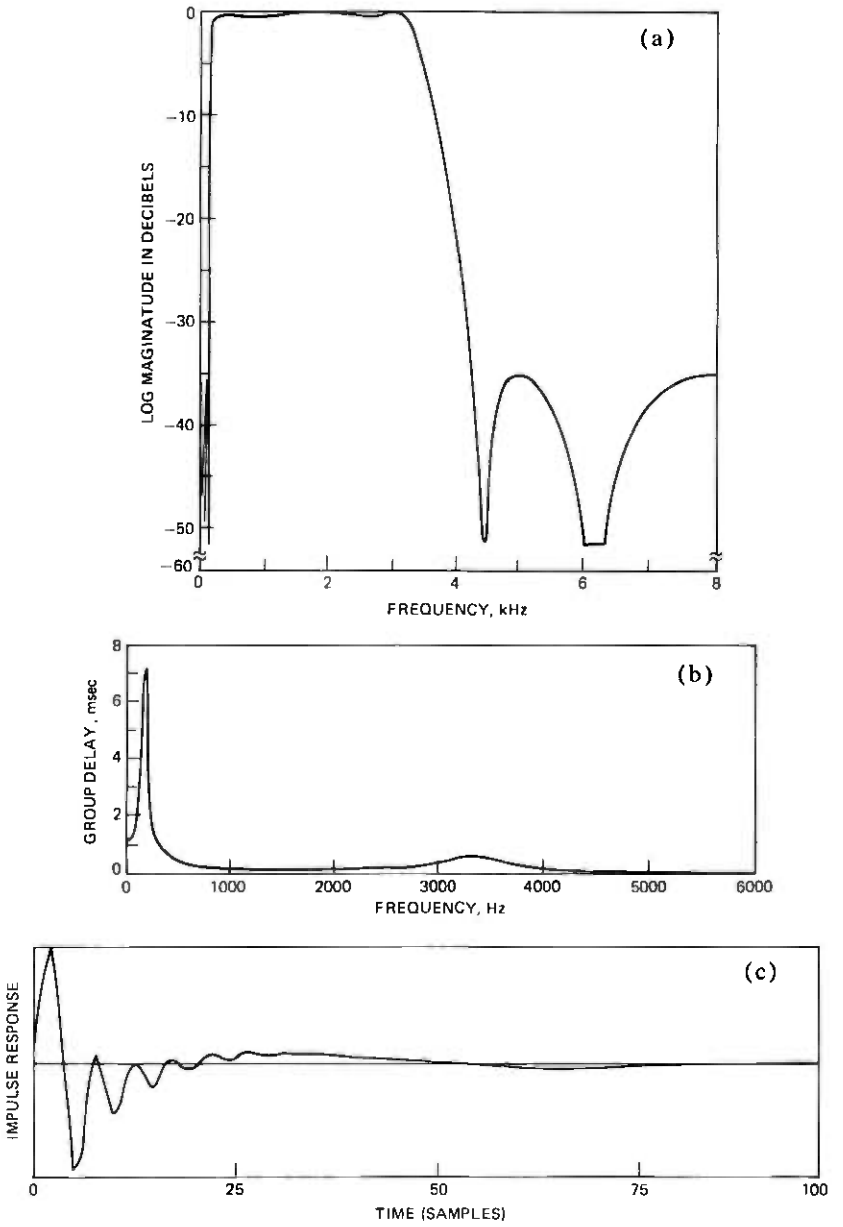


Fig. 4—(a) Log magnitude response. (b) group delay, and (c) impulse response of bandpass filters.

III. FACTORS AFFECTING THE TANDEM LINK

The performance of the LPC to CVSD link is affected by several parameters. Since the LPC vocoder analyzes and then synthesizes the

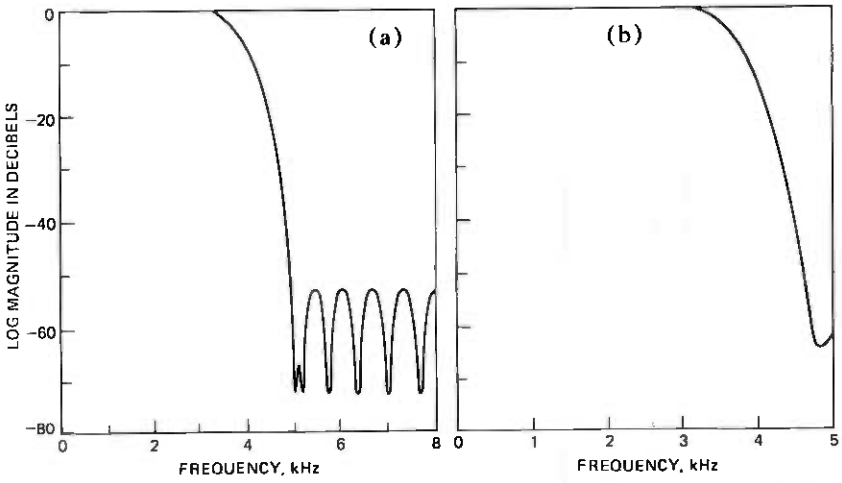


Fig. 5—Frequency response of (a) 119-tap FIR filter for 10:16 sampling rate conversion and (b) 127-tap FIR filter for 16:10 sampling rate conversion.

speech signal, the performance of the CVSD coder will be affected by the manner in which the speech waveform is synthesized. Primarily the performance of the CVSD coder can be affected by factors such as the input level of the speech and the peak factor (ratio of peak-to-RMS value) of the synthesized waveform. Alternatively, parameters in the narrowband system relating to the pitch and the coefficients of the all-pole filter in the synthesis model have little bearing on the performance of the CVSD coder. Therefore, our investigation of the LPC to CVSD link concentrates primarily on the first effects (input level and peak factor).

The input level of the speech waveform determines the operating mode of the CVSD coder. If the input level is too low the coder will be operating in the region in which its performance is determined primarily by granular noise. If the input level is too high the coder will operate in a slope overload condition. Typical waveforms for these coder conditions are shown in Figs. 6–8. Figure 6 shows a complete sequence of waveforms for the wideband system in Fig. 2 under normal (maximum SNR) operating conditions. Figure 6a shows 100 msec of speech appearing at the output of the 10 kHz to 16 kHz sampling rate converter. In Fig. 6b the speech waveform has passed through the first bandpass filter (see Fig. 2) and the effects of bandlimiting and phase distortion can be observed. Figure 6c shows the output waveform of the CVSD coder with the gain $G = 0.158$ which results in maximum SNR. The effects of quantization are clearly noticeable. Finally, Fig. 6d shows the CVSD coder output after bandpass filtering (i.e., the output of the wideband system). Figure 7 shows waveforms for the coder operating in the granular noise region ($G < 0.158$). In Fig. 7a and 7b, waveforms of the unfiltered and band-

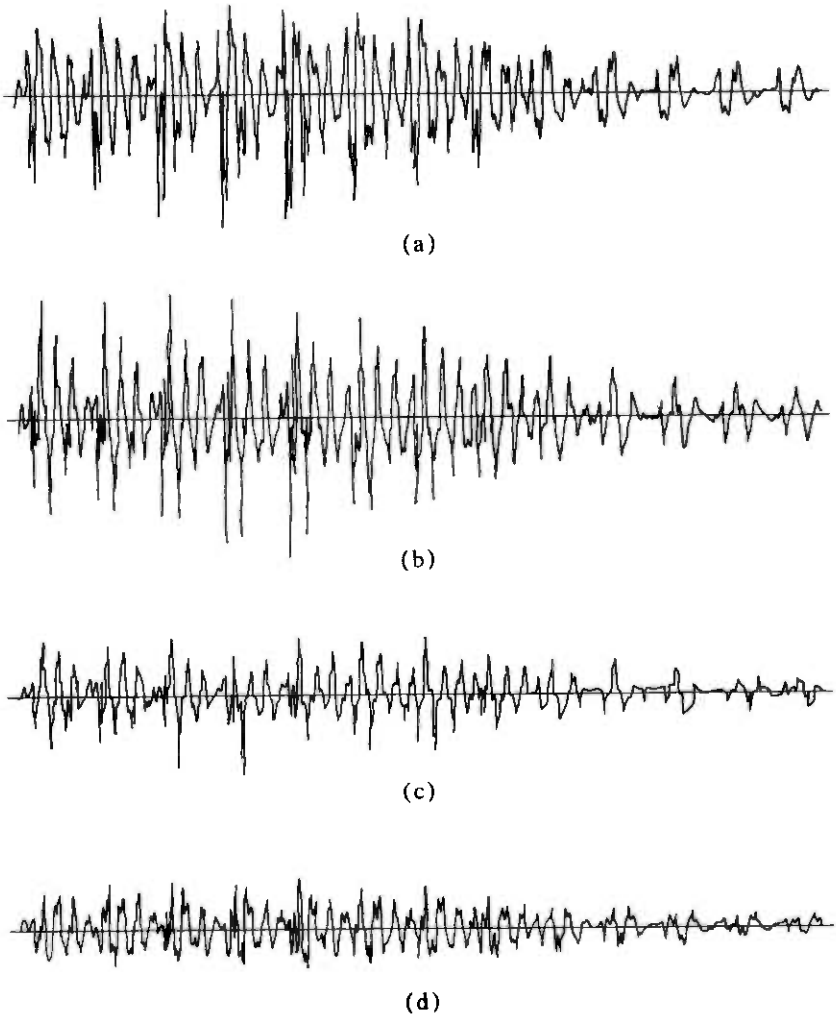


Fig. 6—Speech waveforms for the wideband system. (a) Waveform after 10:16 sampling rate conversion. (b) Waveform after first BPF (input to CVSD). (c) CVSD output waveform (G = 0.158). (d) CVSD output waveform after BP filtering (output of wideband system).

pass-filtered CVSD coder output are shown for a gain setting of G = .009375 or about 25 dB below the maximum SNR operating point. The effects of severe distortion are clearly visible and the speech was completely unintelligible at this point. In Fig. 7c and 7d, waveforms are shown for unfiltered and filtered coder outputs with G = .0395 or about 12 dB below the maximum SNR operating point. Figure 8 shows examples of waveforms for the coder operating in the slope overload region ($G >$ 0.158). In Fig. 8a and b the unfiltered and bandpass-filtered CVSD coder

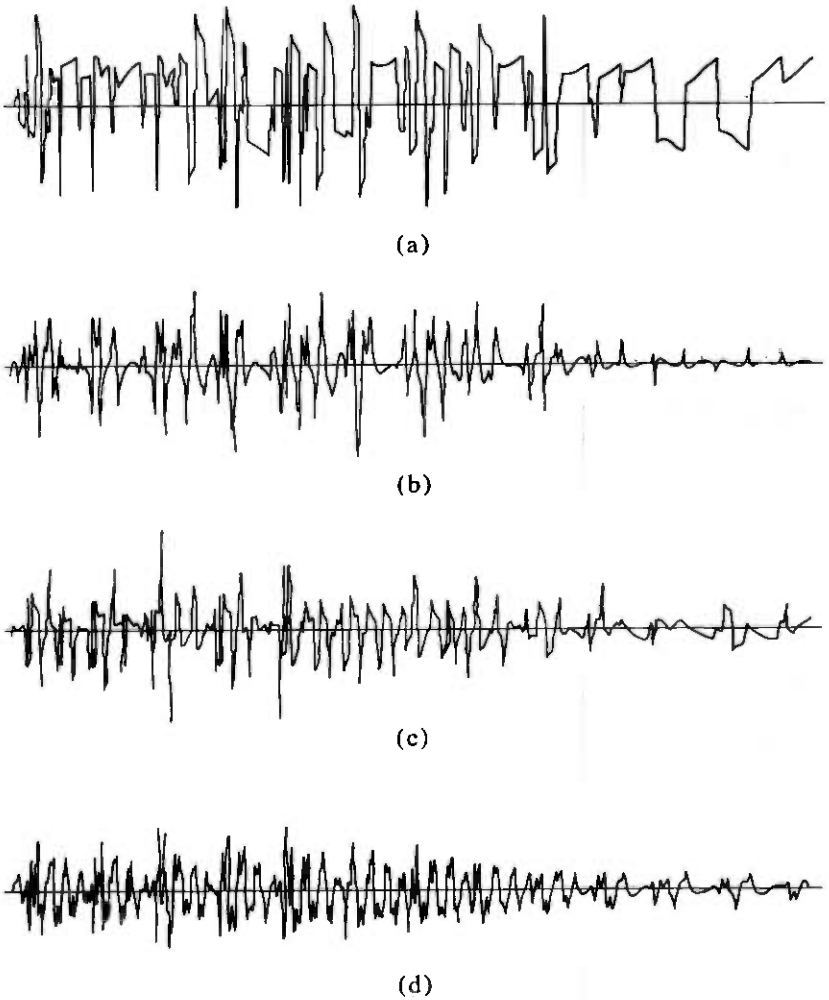


Fig. 7—Output waveforms of the CVSD coder in the granular noise region. (a) Coder output for $G = 0.009375$ and (b) the output after BP filtering. (c) Coder output for $G = 0.0395$ and (d) the same output after BP filtering.

output is shown for $G = 2.528$ or about 24 dB above the maximum SNR operating point. Although the effects of severe slope overload are apparent, the intelligibility of the coder in the slope overload region is not greatly reduced from that at the maximum SNR. Finally, Fig. 8c and Fig. d show unfiltered and filtered output waveforms of the CVSD coder for $G = 0.632$ or about 12 dB above the maximum SNR operating point.

One measure of coder performance is signal to quantizing noise ratio (SNR). The range of input signal level over which the coder maintains an acceptable SNR is often used as a measure of the dynamic range of

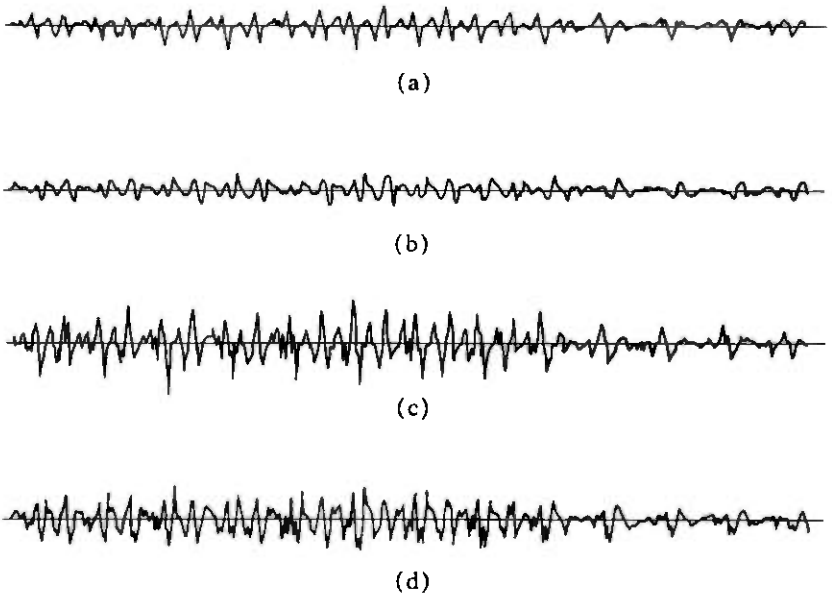


Fig. 8—Output waveforms of the CVSD coder in the slope overload region. (a) Coder output for $G = 2.528$ and (b) the output after BP filtering. (c) Coder output for $G = 0.632$ and (d) BP filtered output.

the coder. The point of optimum SNR is achieved when the coder is on the verge of slope overload.[6] Unfortunately this operating point is not the same as the optimum operating point observed on the basis of subjective performance.[6] Subjectively the noise due to slope overload is less objectionable than the granular noise. Therefore, SNR by itself is not a reliable means for determining the optimum operating region of the coder. More will be said about this in the next section, and in Part 2 (the accompanying paper) another measure of coder performance is proposed which correlates better with subjective performance than the SNR measure.

An important factor affecting the performance of the CVSD, at least in terms of its SNR, is the peak factor of the LPC synthesized speech. The step size of the coder tends to track the RMS level of the input and, if the speech waveform has a large peak-to-RMS ratio, slope overload will cause the peaks to be clipped giving the speech a hoarse sound. If the clipping is severe, intelligibility is degraded.

The peak factor of the synthesized speech can be reduced in several ways to make it more amenable to waveform coding. In one technique the standard pitch source excitation to the LPC synthesizer (an impulse), is modified to spread the energy of the pitch pulse over a larger portion of the pitch period.[8] A pulse which is spread over about 7 percent of the pitch period has been found to be effective for this purpose.[9] Two pulse

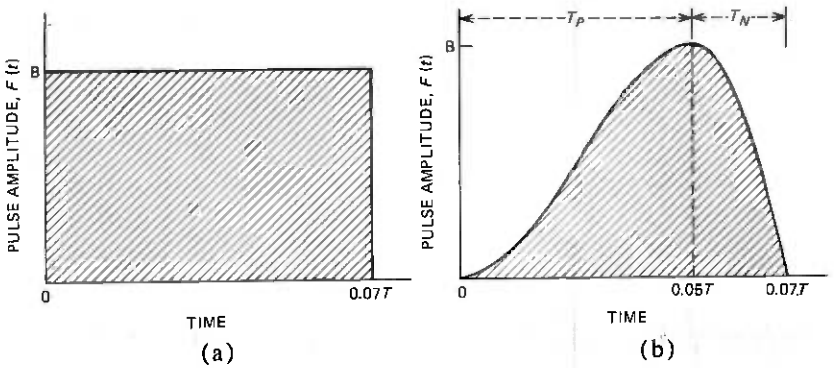


Fig. 9—Pulse excitation sources used for the LPC synthesizer. (a) Rectangular pulse shape. (b) Rounded pulse shape.

shapes were tried in this experiment—a rounded pulse shape and a rectangular pulse shape. The rectangular pulse shape is shown in Fig. 9a. The energy in the pulse is normalized to that of an equivalent impulse. The second pulse shape, shown in Fig. 9b, is a rounded shape proposed by Rosenberg[8] (pulse shape B) to approximate the shape of an actual glottal pulse. T_p is defined as the opening time and T_N is defined as the closing time of the pulse. The pulse shape is then defined by the relation

$$F(t) = B\left[3\left(\frac{t}{T_p}\right)^2 - 2\left(\frac{t}{T_p}\right)^3\right] \text{ for } 0 \leqslant t \leqslant T_p$$

$$F(t) = B\left[1 - \left(\frac{t - T_p}{T_N}\right)^2\right] \text{ for } T_p \leqslant t \leqslant T_p + T_N \qquad (15)$$

where $F(t)$ is the height of the pulse and B is its peak amplitude. Values of T_p and T_N used in the experiment are $T_p = 0.05T$ and $T_N = 0.02T$ where T is the pitch period. The width of the pulse therefore expands or contracts dynamically with the pitch period. The rounded pulse shape was found to give the most natural sound for the LPC synthesized speech.[9]

A second technique that can be used to reduce the peak factor of the LPC sythesized speech is to filter the speech with an allpass filter which disperses the energy of pitch peaks in the waveform. One approach to designing such an allpass filter has been proposed by Rabiner and Crochiere[10] in which the parameters of an allpass filter were optimized to spread the energy of an impulse signal under the limitations of a maximum peak amplitude. This allpass network has been effective in reducing the peak factor of the LPC synthesized speech.

The allpass filter which was used in our experiments was an eighth-order filter which was cascaded three times to give a total allpass filtering

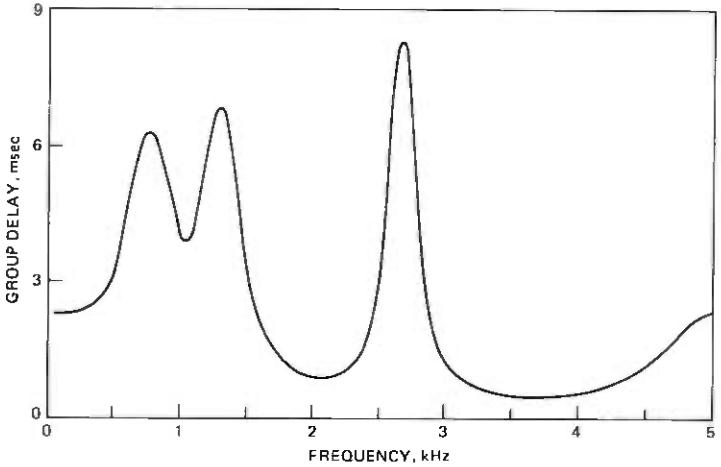


Fig. 10—Group delay of the allpass filter used for preprocessing the CVSD coder input.

equivalent to that of a 24th order filter. The z-transform of each eighth-order filter is of the form

$$H(z) = \prod_{i=1}^{4} H_i(z) \tag{16}$$

where

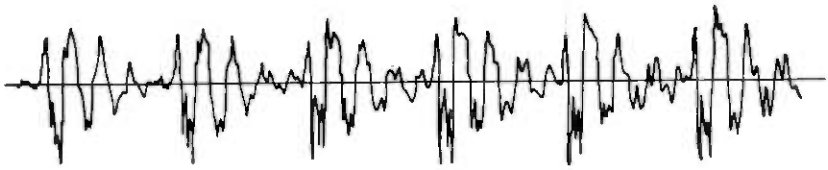
$$H_i(z) = \frac{b_i - c_i z^{-1} + z^{-2}}{1 - c_i z^{-1} + b_i z^{-2}} \tag{17}$$

and the coefficients are

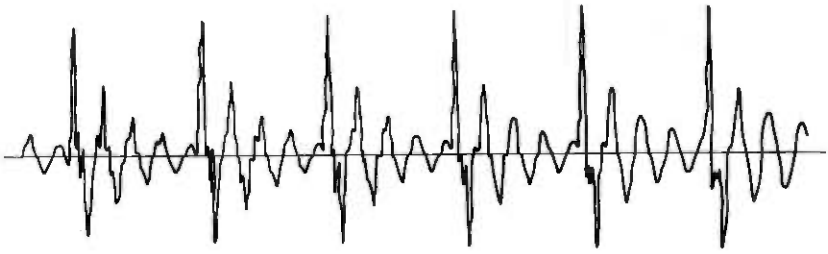
$$b_1 = 0.8149 \qquad c_1 = 1.2308$$
$$b_2 = -0.4970 \qquad c_2 = -0.1060$$
$$b_3 = 0.8621 \qquad c_3 = -0.2135$$
$$b_4 = 0.7870 \qquad c_4 = 1.5727$$

The total group delay of the 24th order all-pass filter is given in Fig. 10. It is seen that the group delay is dispersed between 5 and 90 samples (0.5 to 9 msec) across the frequency band (0 to 5 kHz).

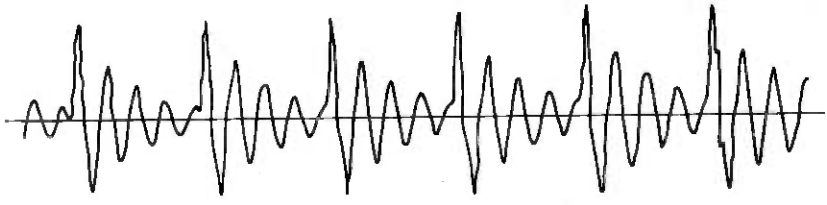
Fig. 11 shows the effects of pitch pulse modifications and allpass filtering on a voiced region of speech. Figure 11a shows the natural speech waveform and Fig. 11b shows an equivalent section of LPC synthesized speech using an impulse excitation. In Fig. 11c and d waveforms are given for LPC synthesized speech with the rectangular and rounded pulse excitations respetively. Figure 11e shows the waveform for the LPC impulse excited speech which was allpass filtered. Finally, Fig. 11f and g show the combination of both allpass filtering and rectangular and rounded source excitations respectively. It is seen that the rectangular or rounded



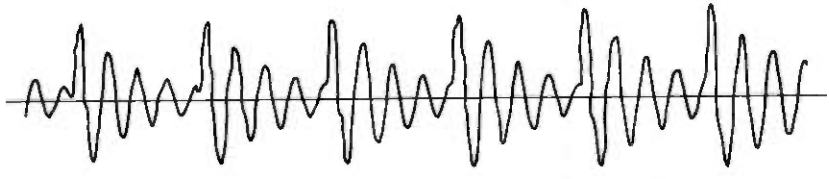
(a)



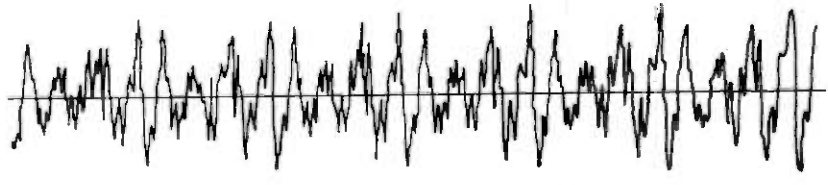
(b)



(c)

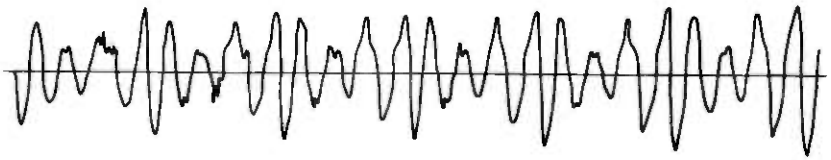


(d)



(e)

(f)



(g)

Fig. 11—Waveforms of the LPC synthesized speech. (a) Natural speech input. (b) LPC synthesized speech with impulse excitation. (c) LPC synthesized speech with rounded source excitation. (d) LPC synthesized speech with rectangular source excitation. (e) Allpass filtered waveform of LPC speech with impulse excitation. (*f*) Allpass filtered waveform of LPC speech with rounded source excitation. (g) Allpass filtered waveform of LPC speech with rectangular source excitation.

excitation source modifications do improve the peak factor of the speech as does the allpass filtering. The combination gives a further improvement. In the next sections we investigate the effects of these modifications on the performance of the CVSD system.

IV. SNR MEASUREMENTS OF THE CVSD SYSTEM

In this section we report on the performance of the CVSD coder in the tandem link as a function of the signal gain and the modifications of the peak factor of the LPC synthesized speech. Computer simulations were made for the system shown in Fig. 2. Two sentences were used for the simulations. The first sentence, "Every salt breeze comes from the sea," was spoken by a low-pitched male and was recorded off a conventional telephone line. The second sentence, "I know when my lawyer is due," was spoken by another male into a high-quality microphone.

The signal-to-quantizing noise ratio (SNR) of the CVSD coder was measured across the entire sentence. The CVSD noise was obtained by subtracting the filtered output from the CVSD input (also filtered) as shown in Fig. 2. The gain G of the signal was varied from 0.009375 to 2.528 or over a range of approximately 50 dB.

Table I shows the resulting SNRs for the first sentence, "Every salt breeze . . . ." Column 1 corresponds to results for natural speech input to the CVSD coder. Columns 2, 3, and 4 are for LPC synthesized speech using an impulse source, a rounded source, and a rectangular source excitation respectively. Table II gives corresponding SNR's measured with the all-pass filter preceding the CVSD. Tables III and IV pertain

Table I — SNR of CVSD coder vs. gain and source excitation

Gain G	Original speech	Coder SNR* (dB) LPC synthesized speech: Impulse source	Rounded source	Rectangular source
0.009375	2.00	1.84	1.77	1.78
0.0395	7.30	7.35	7.89	
0.158	9.29	8.89	10.47	10.62
0.316	7.29	7.23	9.00	
0.632	5.06	5.13	6.39	6.76
1.264	3.31	3.34	4.17	4.43

* For sentence "Every salt breeze comes from the sea."

Table II — SNR of CVSD coder vs. gain and source excitation for allpass filtered inputs

Gain G	Original speech	Coder SNR* (dB) LPC synthesized speech: Impulse source	Rounded source	Rectangular source
0.009375	1.57	1.63	1.74	1.84
0.0395	7.53	7.51	7.91	7.94
0.158	9.26	9.67	10.33	10.79
0.316	7.95	8.37	9.41	9.68
0.632	5.83	6.10	7.18	7.53
1.264	3.60	3.84	4.68	4.92
2.528	2.00	2.15	2.63	2.76

* For sentence "Every salt breeze comes from the sea."

Table III — SNR of CVSD coder vs. gain and source excitation

Gain G	Original speech	Coder SNR* (dB) LPC synthesized speech: Impulse source	Rounded source	Rectangular source
0.009375	2.52	2.37	2.28	2.31
0.0395	8.93	8.80	8.85	9.06
0.158	11.14	10.77	11.61	12.01
0.316	9.48	8.90	10.01	10.46
0.632	7.07	6.61	7.54	7.81
1.264	4.50	4.38	4.96	5.10
2.528	2.52	2.64	2.95	3.03

* For sentence "I know when my lawyer is due."

to the sentence "I know when . . ." and show measurements corresponding to those in Tables I and II, respectively.

The data indicate that, with or without the allpass filter, CVSD SNR with natural speech input is quite similar to SNR with speech derived from an LPC synthesizer with impulse excitation. (In all four tables the greatest difference between an entry in Column 1 and the corresponding

Table IV — SNR of CVSD coder vs. gain and source excitation for allpass filtered inputs

Gain G	Original speech	Coder SNR* (dB) LPC synthesized speech Impulse source	Rounded source	Rectangular source
0.009375	1.98	1.42	1.69	1.67
0.0395	8.33	8.48	8.91	9.06
0.158	10.59	10.44	11.67	11.49
0.316	9.53	9.02	10.42	10.16
0.632	7.22	6.89	7.84	7.86
1.264	4.59	4.76	5.44	5.55
2.528	2.54	2.75	3.20	3.31

* For sentence "I know when my lawyer is due."

entry in Column 2 is 0.6 dB; most differences are less than 0.4 dB.) Comparing Columns 3 and 4 with Column 2 in the tables we see that broadened pitch pulses lead to 1–2 dB improvements in measured CVSD performance in the slope overload region. As a rule the rectangular pulses result in a slightly higher SNR than rounded ones.

The benefits of allpass filtering are less pronounced than the benefits of broadened pitch pulses. Comparing Column 2 entries (impulse excitation) in Table I and Table II, we see that the allpass filter offers improvements of about 1 dB in SNR at high levels for one sentence. Tables III and IV show virtually no improvement with the other sentence. When the synthesizer uses broadened pitch pulses (Columns 3 and 4) the allpass filter adds 0.5 to 1 dB to CVSD performance with the first sentence and little or nothing to the SNR of the second sentence.

Figure 12 displays the range of possible improvements in CVSD SNR relative to the conventional tandem configuration which includes an LPC synthesizer with an impulse source and no allpass filter at the narrowband-wideband interface. The lower curve in Fig. 12a and b shows CVSD for this configuration for the two sentences in our study. The upper curve in Fig. 12a pertains to the most successful modification of the sentence recorded form a telephone line. This modification involves rectangular pitch pulses and an allpass filter. With the sentence recorded from a high-quality microphone, the best SNR performance, plotted in Fig. 12b, was obtained with the rectangular excitation and no allpass filter.

V. SPEECH QUALITY

Informal judgments of the processed speech suggest that the predominant distortions of tandem circuits are those of CVSD. However, the quality of a vocoder such as LPC depends on speaker and utterance while a waveform coder such as CVSD is relatively insensitive to speech material. Although the utterances used in this work were amenable to

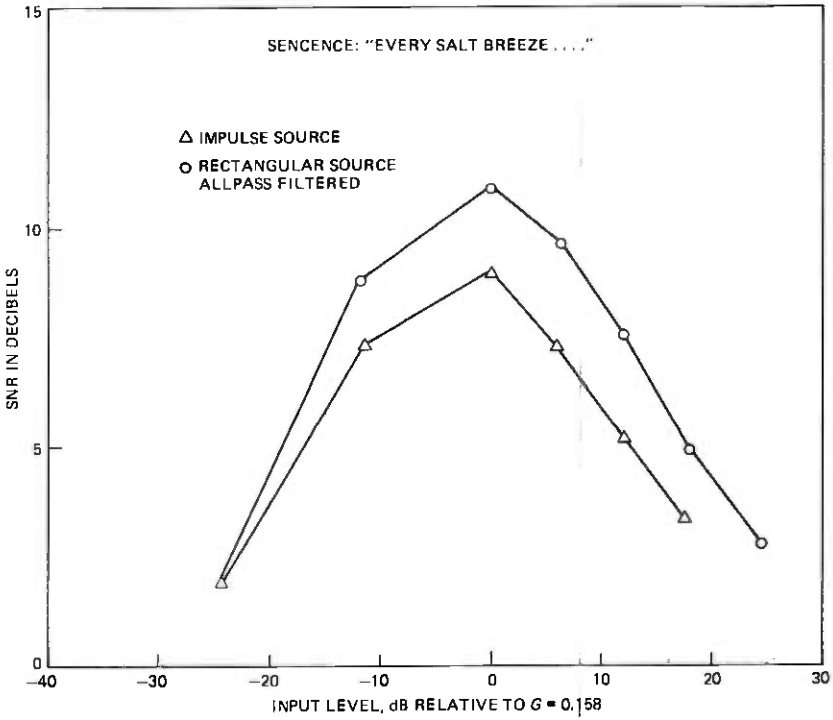


Fig. 12—Summary of the results of the SNR measurements of the CVSD coder, for sentences (a) "Every salt breeze comes from the sea" and (b) "I know when my lawyer is due".

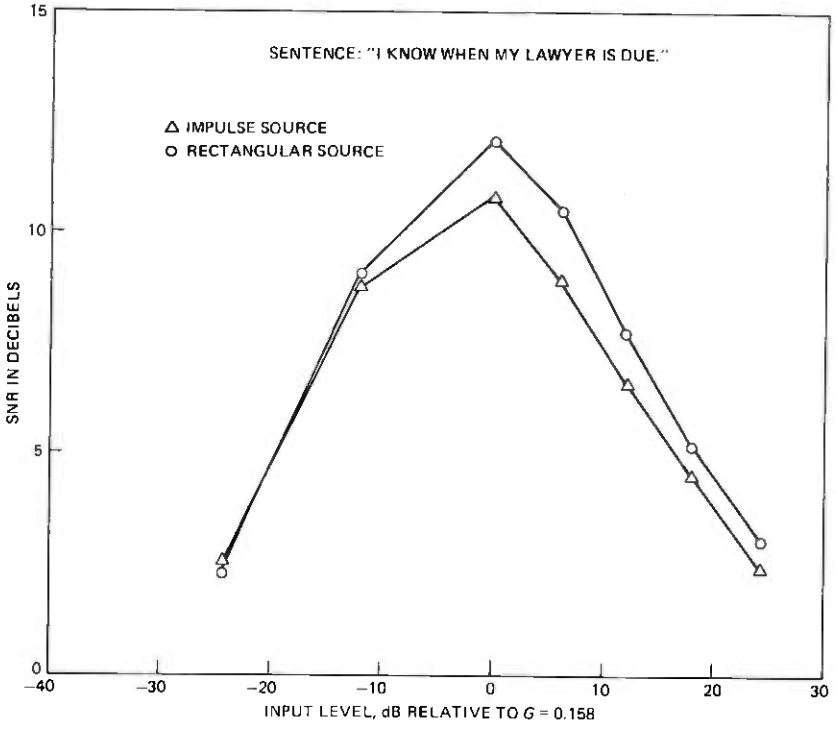
LPC, we anticipate that for certain speakers LPC would be the weaker link in a tandem connection.

As a function of input level, CVSD quality appears to be much lower with weak inputs, which lead to substantial granular quantizing noise, than with strong inputs, for which the main distortion is slope overload. This subjective effect is at variance with SNR indications which show rapidly declining quality as the input level rises into the coder overload range.

The use of broadened LPC excitation pulses lends a more natural quality to the resynthesized speech as well as improving CVSD SNR in the overload region. An allpass filter which also improves SNR for one sentence seems to offer little, if any, enhancement of subjective quality of tandem circuits.

VI. DISCUSSION

Although the conclusions of the previous section must be regarded as tentative, pending formal subjective evaluation of speech processed



(b)

Fig. 12 (*continued*)

in tandem connections, it does appear that efforts to improve the quality of the wideband link would be justified. The CVSD encoder is a 9-year-old design with values of circuit elements chosen to withstand transmission errors occurring at rates as high as 10 percent. If this very demanding requirement is relaxed somewhat and recent advances in delta modulation are incorporated, it may be possible to modify the CVSD to produce higher stand-alone and tandem quality. Alternatively other 16 kb/s wideband coding schemes such as adaptive PCM, adaptive differential PCM or sub-band coding may offer even greater advantages than improved CVSD.[6,11]

REFERENCES

1. R. Steele, "Delta Modulation Systems," London: Peutech Press, 1975 (see Scale system, pp 206).
2. J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, Berlin: Springer-Verlag, 1976.
3. J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-24* February 1976, pp. 2–8.
4. M. R. Sambur, "An efficient linear prediction vocoder," B.S.T.J., *54*, No. 10 (December 1975), pp 1693–1723.

5. A. H. Gray and J. D. Markel, "Quantization and Bit Allocation in Speech Processing," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-24*, December 1976, pp. 459–473.
6. N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," Proc. IEEE, May 1974, pp. 611–632.
7. R. E. Crochiere and L. R. Rabiner, "Optimum FIR digital filter implementations for decimation, interpolation, and narrowband filtering," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-23*, October 1975, pp. 444–456.
8. A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Amer., *50*, February 1971, pp. 637–644.
9. M. R. Sambur, A. E. Rosenberg, L. R. Rabiner, and C. A. McGonegal "On removing the buzz in LPC Synthesis," Proc. 1977 IEEE Int. Conf. on Acoust., Speech, Sig. Proc., pp 401–404.
10. L. R. Rabiner and R. E. Crochiere, "On the design of allpass signals with peak amplitude constraints," B.S.T.J., *55*, No. 4 (April 1976), pp. 395–407.
11. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital coding of speech in subbands," B.S.T.J., *55*, No. 8 (October 1976), pp. 1069–1086.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Tandem Connections of Wideband and Narrowband Speech Communication Systems Part 2—Wideband-to-Narrowband Link

By L. R. RABINER, M. R. SAMBUR, R. E. CROCHIERE, and D. J. GOODMAN

(Manuscript received March 25, 1977)

In this paper the tandem link of a 16 kb/s Continuously Variable Slope Delta modulator (CVSD) waveform coder and a 2.4 kb/s Linear Predictive Coding (LPC) vocoder is studied. Of prime concern are the effects of the CVSD coder on the LPC vocoder analyzer. In particular the problems involved in making a reliable voiced-unvoiced decision, estimating pitch period, and estimating LPC coefficients from the coder output are studied. It is shown that LPC coefficient estimation from the CVSD output is highly inaccurate. An analytical distortion measure (an LPC distance) is used to show the magnitude of the distortion introduced by the coder as a function of the signal gain into the CVSD coder. Although the remainder of the LPC analysis (i.e., pitch detection, voiced-unvoiced decision, and gain calculation) can be performed reasonably accurately, the magnitude of the distortions in estimating the LPC coefficients is sufficiently large to make the vocoded speech barely intelligible and of poor quality.

I. OVERVIEW OF THE TANDEM LINK OF CVSD TO LPC

In the first part of this paper we discussed the effects of the narrowband system (the LPC vocoder operating at 2400 b/s) on the wideband system (the CVSD waveform coder).[1] There it was shown that one of the major issues was tailoring the signal characteristics of the vocoded speech to reduce the peak factor, thereby reducing the amount of slope overload noise generated in the CVSD. When we consider the tandem link of CVSD and LPC, far more serious problems are encountered since we must estimate the basic speech production parameters (i.e., pitch, voiced-unvoiced, LPC coefficients) from a severely degraded signal. Since speech parameter estimation is an imperfect process, even on high-quality speech, the effects of the CVSD coder, which include quantization noise

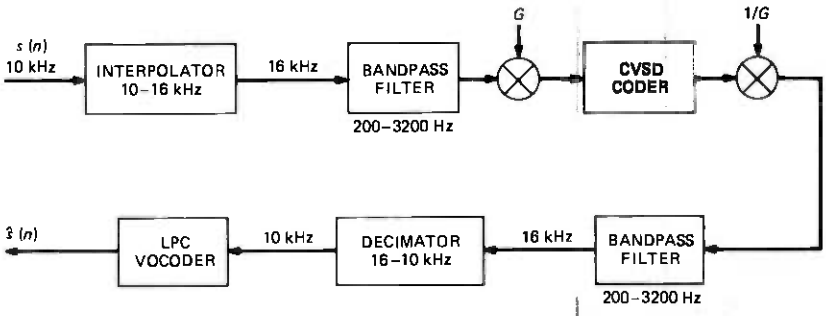


Fig. 1—Block diagram of signal processing operations in tandem link of a CVSD coder and an LPC vocoder.

as well as slope overload noise, could potentially make the tandem link totally unacceptable.

In this paper we discuss several aspects of a tandem link consisting of a CVSD waveform coder, and an LPC vocoder. Our purpose is to demonstrate the range of signal levels over which the LPC can operate reasonably well in tandem with the CVSD coder. Figure 1 shows a block diagram of the signal processing used in implementing and testing a CVSD-LPC tandem link. The speech signal $s(n)$ is assumed to be sampled at a 10-kHz rate. Thus the first block in Fig. 1 is an interpolator to raise the sampling rate of the signal to 16 kHz. The interpolator described in Part 1 of this paper was used here.[1] The 16-kHz signal was then sharply bandpass-filtered from 200 Hz to 3200 Hz using the 8th-order elliptic bandpass filter described in Part 1 of this paper.[1] To simulate variations in overall signal level into the CVSD coder, a variable gain G was applied to the filtered 16-kHz signal. The gain G was varied from 0.009375 to 2.5 in the simulations which gave about a 50-dB variation in signal level over which the system was studied. To compensate for the input scaling, a gain of $1/G$ was used at the output of the CVSD coder. The output of the coder was again sharply bandpass-filtered from 200 to 3200 Hz to remove the wideband quantization noise generated in the CVSD coder. For compatibility with the LPC system the signal was then decimated to a 10-kHz sampling rate using the decimator described in Part 1 of this paper.

Figure 2 shows a block diagram of the processing required for the LPC vocoder. The LPC analyzer estimates the following control parameters:

(*i*) Pitch period
(*ii*) Voiced-unvoiced decision
(*iii*) Signal gain
(*iv*) LPC parameters

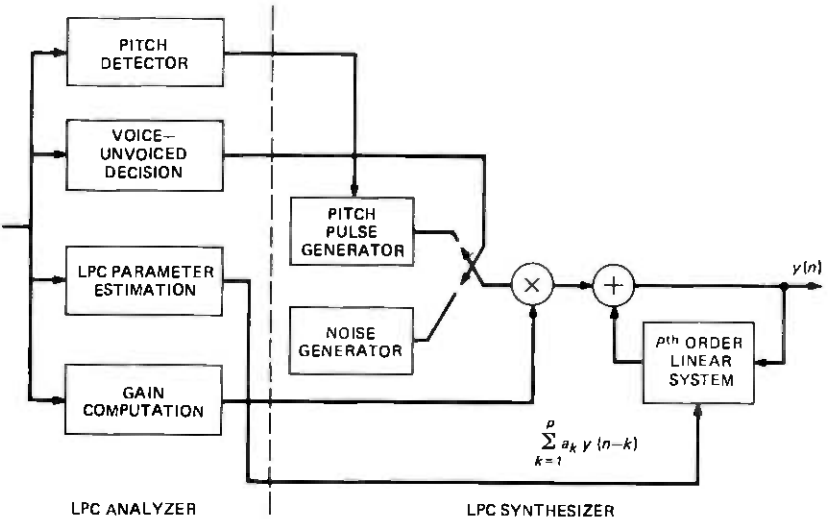


Fig. 2—Block diagram of LPC analyzer and synthesizer.

The LPC synthesizer uses the estimated parameters to recreate the speech in the manner shown in Fig. 2. The details of the analysis and synthesis methods are described in Part 1 of this paper.

Based on our knowledge of both the techniques used in LPC analysis and the degradations introduced by the CVSD coder, it was anticipated that the voiced-unvoiced decision and the LPC parameter estimation algorithms would be most affected by the CVSD coder. Thus, in the next two sections we discuss the specific algorithms used for voiced-unvoiced detection (along with pitch detection) and show results on how the algorithms performed in the tandem link as a function of the signal level into the CVSD coder. In Section IV we present results on the accuracy with which the LPC parameters were estimated from the coder output. For a measure of similarity between coder input and output, the LPC distance measure proposed by Itakura is used. Finally, in Section V we discuss the interactions between the CVSD coder and the LPC vocoder and suggest some possible ways to improve the performance of a tandem link of a wideband and a narrowband system.

II. PITCH DETECTOR AND VOICED-UNVOICED DETECTOR USED IN THE TANDEM LINK

As discussed in the preceding section, the choice of an appropriate pitch detector and voiced-unvoiced detector is critical to the proper operation of the LPC vocoder. Based on a series of intensive investigations into both objective and subjective rankings of a variety of pitch detectors,[2,3] it was shown that simple waveform pitch detectors would be in-

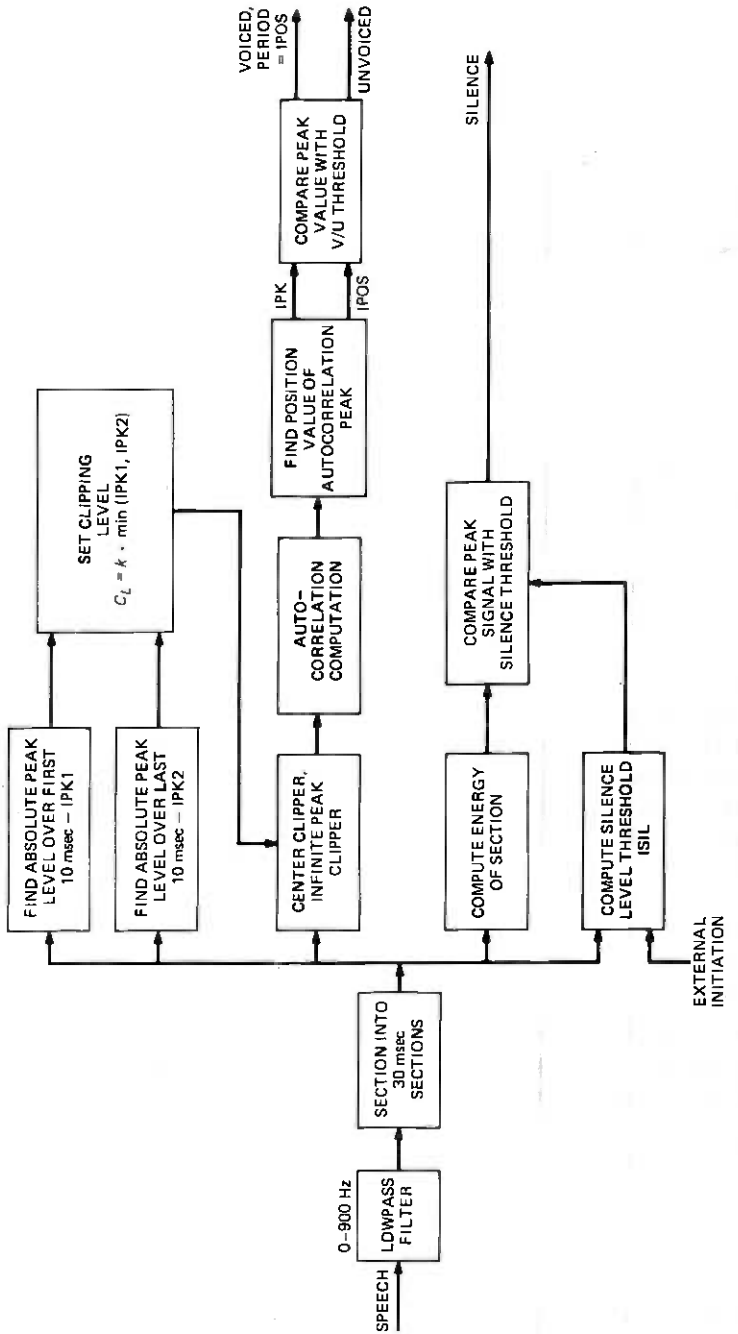


Fig. 3—Block diagram of modified autocorrelation pitch detector.

adequate for a severely degraded waveform such as obtained at the output of a CVSD coder. Thus either a sophisticated correlation-type pitch detector, or a spectral-type pitch detector is required for this application. From this class of pitch detectors both the AMDF[4] and AUTOC[5] pitch detectors were found to be moderately fast, and sufficiently robust over a wide variety of transmission conditions and pitch range of the speaker. Because of the familiarity of the authors with the AUTOC pitch detector, this method was finally selected.

Before the method of operation of this pitch detector is reviewed, some comments must be made about the selection of the voiced-unvoiced detector. Ideally one would prefer to make a voiced-unvoiced decision prior to, and independent of, the pitch detection. In this manner the role of the pitch detector is strictly to make the best estimate of pitch period, given a priori that the segment is accurately classified as voiced. For unvoiced segments, the pitch detector is not used at all. There have been at least three proposed methods for making a voiced-unvoiced decision prior to and independent of any pitch detection.[6–8] However, all three methods suffer from the necessity of having a training set of data that characterizes the signal classes. For CVSD coding, the variability of the signals due to variations in gain is exceedingly large—i.e., a 40-dB variation in input level can change the signal from one with a large amount of granular noise to one with a large amount of slope overload noise. Therefore, making a voiced-unvoiced decision accurately without a periodicity measurement (pitch detector) to aid the decision is extremely difficult. Thus, the voiced-unvoiced decision is combined with the pitch detection in the AUTOC method.

A block diagram of the AUTOC pitch detector is given in Fig. 3. The method requires that the speech be lowpass-filtered to 900 Hz. Thus a 99-point linear phase, FIR digital filter is used here.[9] The lowpass-filtered speech is sectioned into overlapping 30-msec (300 samples at 10 kHz) sections for processing. Since the pitch period computation for all pitch detectors is performed 100 times/second—i.e., every 10 msec—adjacent sections overlap by 20 msec or 200 samples.

The first stage of processing is the computation of a clipping level C_L for the current 30-ms section of speech. The clipping level is set at a value which is 64 percent of the smaller of peak absolute sample values in the first and last 10-ms portions of the section. Following the determination of the clipping level, the 30-ms section of speech is center clipped, and then infinite-peak-clipped, resulting in a signal which assumes one of three possible values, 1 if the sample exceeds the positive clipping level, -1 if the sample falls below the negative clipping level, and 0 otherwise.

Following clipping, the autocorrelation function for the 30-ms section is computed over a range of lags from 20 samples to 200 samples (i.e.,

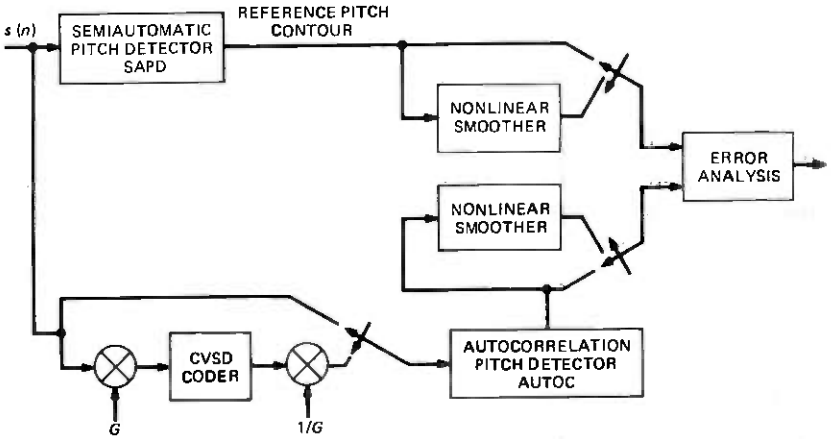


Fig. 4—Block diagram of system used to compare pitch contours from two pitch detectors and to perform an appropriate error analysis.

2-msec to 20-msec period). Additionally, the autocorrelation at 0 delay is computed for appropriate normalization purposes. The autocorrelation function is then searched for its maximum (normalized) value. If the maximum (normalized value) exceeds 0.25, the section is classified as voiced and the location of the maximum is the pitch period. Otherwise, the section is classified as unvoiced.

In addition to the voiced-unvoiced classification based on the autocorrelation function, a preliminary test is carried out on each section of speech to determine if the peak signal amplitude within the section is sufficiently large to warrant the pitch computation. If the peak signal level within the section is below a threshold computed from the background noise level, the section is classified as unvoiced (silence) and no pitch computations are made.

III. EFFECTS OF CVSD CODING ON PITCH DETECTION

To investigate the effects of CVSD coding on pitch detection, two sentences were used whose pitch contours were known extremely accurately.[9] Figure 4 shows a block diagram of the experimental arrangement used to show pitch detection errors in the tandem link. The speech, $s(n)$, is analyzed by the SAPD method[9] to give the reference pitch contour, $p_r(m)$, $m = 1, 2, \ldots, M$, where M is the number of 10-msec frames in the utterance, and $p_r(m) = 0$ if the frame is classified as unvoiced. Otherwise $p_r(m)$ is the estimated pitch period. Extensive tests have shown the SAPD method to be a reliable and robust procedure for obtaining the reference pitch contour.[9]

The test pitch contours are obtained by sending the speech either directly to the pitch detector, or first through the CVSD coder where the

signal level is determined by the gain G. We denote the test pitch contour as $p_t(m)$, $m = 1, 2, \ldots, M$. The error analysis compares $p_r(m)$ and $p_t(m)$ over the utterance and makes the following measurements:

(*i*) Average pitch period error during voiced regions, $\bar{P}$, defined as

$$\bar{P} = \frac{1}{N_v} \sum_{\substack{m=1 \\ p_r(m)\neq 0 \\ p_t(m)\neq 0 \\ |p_t(m)-p_r(m)|\leq 10}}^{M} [p_r(m) - p_t(m)] \tag{1}$$

where N_v is the number of voiced regions satisfying the conditions that the reference pitch contour indicates a voiced region ($p_r(m) \neq 0$), the test pitch contour indicates a voiced region ($p_t(m) \neq 0$), and the difference in estimated pitch period is less than or equal to 10 samples ($|p_t(m) - p_r(m)| \leq 10$).

(*ii*) Standard deviation of the pitch period during voiced regions, σ_p, defined as

$$\sigma_p = \left[\frac{1}{N_v} \sum_{\substack{m=1 \\ p_r(m)\neq 0 \\ p_t(m)\neq 0 \\ |p_t(m)-p_r(m)|\leq 10}}^{M} (p_r(m) - p_t(m))^2 - \bar{P}^2 \right]^{1/2} \tag{2}$$

(*iii*) Number of voiced-to-unvoiced errors, N_{vu}, defined as

$$N_{vu} = \sum_{m=1}^{M} g(p_r(m), p_t(m)) \tag{3}$$

where

$$\begin{aligned} g(x,y) &= 1 \quad \text{if} \quad x > 0 \text{ and } \quad y = 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{4}$$

(*iv*) Number of unvoiced-to-voiced errors, N_{uv}, defined as

$$N_{uv} = \sum_{m=1}^{M} g(p_c(m), p_r(m)) \tag{5}$$

(*v*) Number of gross pitch period errors, N_G, defined as

$$N_G = \sum_{m=1}^{M} f(p_r(m), p_t(m)) \tag{6}$$

Table I — Error analysis for utterance "Every salt breeze comes from the sea"

(a) Analysis on raw pitch data

Signal	$\overline{P}$	σ_p	N_{vu}	N_{uv}	N_G
Original speech	0.142	0.786	8	7	1
CVSD-G = 0.009375	1.154	1.925	69	73	29
CVSD-G = 0.0395	0.221	0.901	22	18	6
CVSD-G = 0.158	0.252	0.874	7	8	4
CVSD-G = 0.316	0.288	0.961	6	8	5
CVSD-G = 0.632	0.294	0.952	3	12	4
CVSD-G = 1.264	0.397	1.037	5	23	4
CVSD-G = 2.528	0.397	1.159	4	37	10

(b) Analysis on nonlinearly smoothed pitch data

Signal	$\overline{P}$	σ_p	N_{vu}	N_{uv}	N_G
Original speech	0.156	0.756	7	1	0
CVSD-G = 0.009375	1.589	1.236	91	24	1
CVSD-G = 0.0395	0.556	1.029	15	2	0
CVSD-G = 0.158	0.426	1.073	7	0	0
CVSD-G = 0.316	0.282	0.922	6	0	0
CVSD-G = 0.632	0.356	0.920	2	1	0
CVSD-G = 1.264	0.367	1.155	1	4	0
CVSD-G = 2.528	0.490	1.253	0	6	1

where

$$f(p_r(m), p_t(m)) = 1 \quad \text{if} \quad p_r(m) \neq 0, p_t(m) \neq 0, |p_r(m) - p_t(m)| > 10$$
$$= 0 \quad \text{otherwise} \tag{7}$$

Since many of the errors made in pitch detection are easily corrected by a nonlinear median-type smoother,[10] the test arrangement in Fig. 4 also shows the capability of passing both the reference and test pitch contours through such a smoother prior to the error analysis. Results will be presented on both the raw data and the smoothed data.

Results obtained on two different sentences are presented in Tables I and II, and some of the key results are summarized in Figs. 5–8. Utterance 1 was the sentence "Every salt breeze comes from the sea" spoken by a low-pitched male and recorded off a conventional telephone line. The utterance had 256 frames (i.e., it was 2.56 seconds long), of which 108 were unvoiced and 148 were voiced. Table I shows values of $\overline{P}$, σ_p, N_{vu}, N_{uv}, and N_G as a function of the gain G, for both the raw data and the nonlinearly smoothed pitch contours. Figure 5 shows plots of N_{vu} versus G (plotted in dB on a normalized scale) for both the raw and smoothed data, and Fig. 6 shows plots of N_{uv} versus G. Results obtained on the original utterance (uncoded) are also presented as a means of comparison.

As seen in Table I, values of $\overline{P}$ for the coded speech were about 2 to 3 times larger than for the original speech (except for G = 0.009375).

Table II — Error analysis for utterance "I know when my lawyer is due"

(a) Analysis on raw pitch data

Signal	$\overline{P}$	σ_p	N_{vu}	N_{uv}	N_G
Original speech	0.304	0.796	1	3	0
CVSD-G = 0.009375	0.192	2.722	21	12	63
CVSD-G = 0.0395	0.304	0.738	17	2	7
CVSD-G = 0.158	0.193	0.660	10	1	2
CVSD-G = 0.316	0.209	0.639	10	1	4
CVSD-G = 0.632	0.228	0.812	9	2	4
CVSD-G = 1.264	0.225	0.922	6	3	5
CVSD-G = 2.528	0.221	0.993	8	4	9

(b) Analysis on nonlinearly smoothed pitch data

Signal	$\overline{P}$	σ_p	N_{vu}	N_{uv}	N_G
Original speech	0.323	0.617	1	2	0
CVSD-G = 0.009375	1.247	2.922	25	10	40
CVSD-G = 0.0395	0.382	0.656	18	1	0
CVSD-G = 0.158	0.172	0.573	11	1	0
CVSD-G = 0.316	0.213	0.549	12	1	0
CVSD-G = 0.632	0.252	0.711	11	1	0
CVSD-G = 1.264	0.257	0.823	10	0	0
CVSD-G = 2.528	0.329	0.985	10	0	0

However, values of $\overline{P}$ were all less than 0.5 samples (except for G = 0.009375) indicating that the average pitch period errors, due to the coder, were still relatively insignificant. For a gain of G = 0.009375 (large amounts of granular noise) the pitch detection process broke down en-

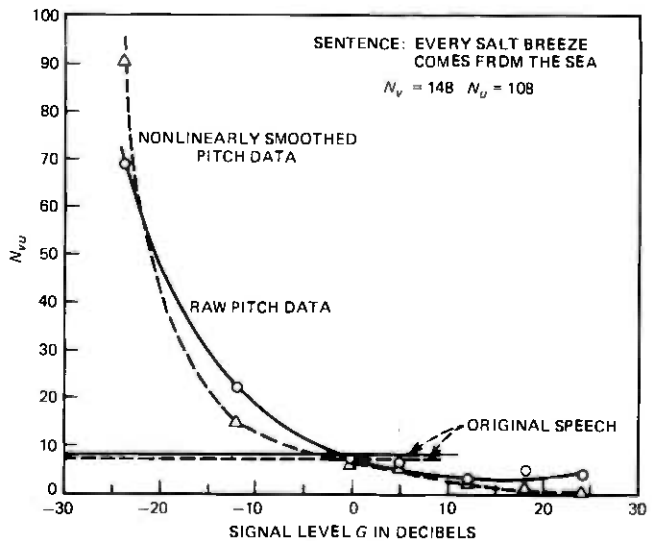


Fig. 5—Plot of number of voiced-to-unvoiced errors versus CVSD signal level for utterance "Every salt breeze comes from the sea."

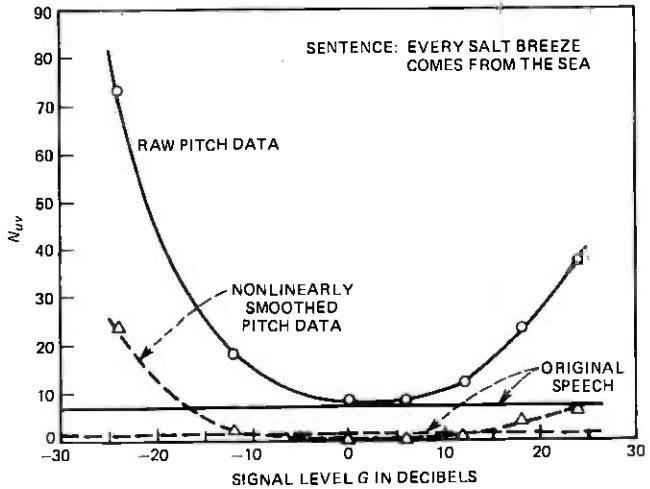


Fig. 6—Plot of number of unvoiced-to-voiced errors versus CVSD signal level for utterance "Every salt breeze comes from the sea."

tirely. Thus, at this extreme the LPC vocoder cannot possibly operate. However, as was shown previously, for this value of gain the CVSD coder produced unintelligible speech; hence we need not be concerned with this result.

Values for σ_p for the coded speech were essentially identical to those obtained for the original utterance. Also the number of gross pitch period errors was small for all values of G except $G = 2.528$ and $G = 0.009375$,

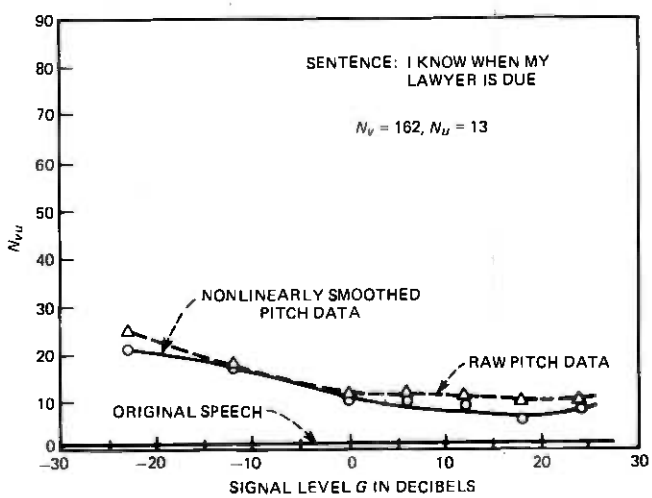


Fig. 7—Plot of number of voiced-to-unvoiced errors versus CVSD signal level for utterance "I know when my lawyer is due."

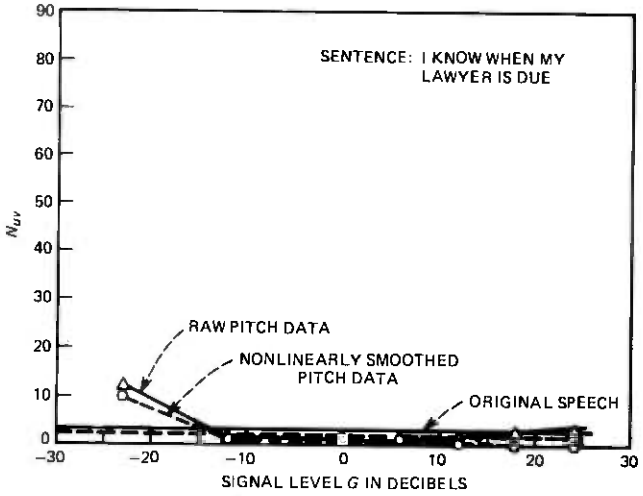


Fig. 8—Plot of number of unvoiced-to-voiced errors versus CVSD signal level for utterance "I know when my lawyer is due."

and all these errors were correctable by the nonlinear smoother, as shown in Table Ib. Thus, one can conclude that for cases in which both the reference and test pitch contours were classified as voiced, the coder did not impede accurate determination of the pitch period—i.e., pitch is well preserved in the CVSD output.

Now the major question is how well the voiced-unvoiced decision could be made on the coder output. An examination of Table I and Figs. 5 and 6 shows that, for several values of G, a substantial number of unvoiced-to-voiced errors occurred. However most of these errors were easily correctable by the nonlinear smoother since the estimated pitch periods (when such errors occur) are essentially random, and are automatically "smoothed" to zero (i.e., unvoiced). Also some of the voiced-to-unvoiced errors are corrected by the smoother.

For this sentence it is concluded that over a fairly large variation in coder input gain, the deterioration of the signal is not so large so as to make pitch detection unreliable.

A second set of results is given for the utterance "I know when my lawyer is due" spoken by another male speaker over a high-quality microphone. This sentence had 175 frames (1.75 seconds) of which only 13 were unvoiced and 162 were voiced. Thus this utterance was essentially all voiced. Results obtained on this utterance are given in Table II and Figs. 7 and 8. Again it is seen that, except for $G = 0.009375$, values of $\overline{P}$, σ_p and N_G (smoothed) are essentially the same for the coder output as for the original. Since there were very few unvoiced frames, the number of unvoiced-to-voiced errors is also the same for the coded speech as for the original. However, the number of voiced-to-unvoiced errors for the

coded speech is much larger than for the original speech. Most of these errors occur in the region of the /z/ in "is due," and as such are not correctable by the nonlinear smoother. However, the errors in this low-intensity region are not very preceptible and therefore such errors are not overly crucial.

In summary we have shown that the CVSD coder preserves the pitch of the speech over a reasonably large signal range and that the voice-unvoiced decision can also be reliably made over a fairly large dynamic range of coder inputs.

IV. EFFECTS OF CVSD CODING OF ESTIMATION OF LPC COEFFICIENTS

The next issue to consider is the effects of the CVSD coder on the estimation of the LPC parameters. The LPC coefficients model the combined transfer function of the vocal tract, glottal source, and radiation load. Incorrect estimates of the coefficients can seriously perturb the frequency spectrum of the modeled speech signal and, hence, affect the intelligibility of the synthesized sound.[11]

4.1 Distance measure

To evaluate objectively the spectral distortion introduced by the CVSD coder, an LPC distance measure proposed by Itakura was employed.[12] The LPC distance measure is defined as

$$d_n = \log\left[\frac{a_n V a_n^t}{b_n V b_n^t}\right] \tag{8}$$

where

a_n = LPC coefficient vector $(1, a_1, \ldots, a_p)$ measured in the nth frame of the original uncoded speech signal.

b_n = LPC coefficient vector measured in the nth frame of the CVSD coded speech signal

and V is the speech correlation matrix with elements V_{ij} defined as

$$V_{ij} = v(|i-j|) = \sum_{n=1}^{N-|i-j|} x(n)x(n+|i-j|) \tag{9}$$

where $x(n)$ is the speech signal and N is the number of samples in the frame.

Figure 9 shows examples which illustrate how the measured d_n is useful in measuring the degree of spectral deviation of the coded sound from that of the original.* Although the measure d_n is not the only possible indicator of spectral distortion,[13] it has been shown to closely

* The quantitative significance of d_n is discussed in detail in Ref. 14.

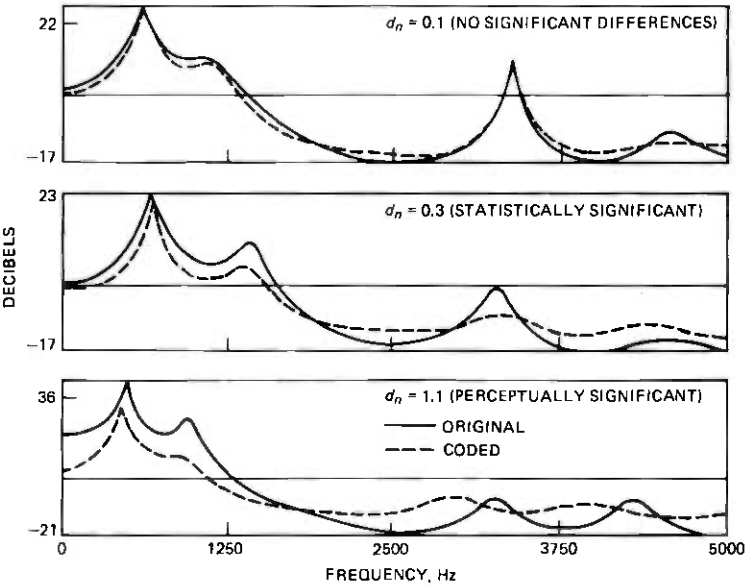


Fig. 9—Plots of typical spectra and the resulting values of d_n for three examples.

correspond to perceptual judgments.[14] In addition, the measure has been effectively applied in problems of speech recognition,[12] speaker recognition,[15] and variable frame rate synthesis.[16] Before discussing the results of the LPC distance evaluation of the CVSD coder, it is important to emphasize that d_n is not a perfect measure of perceptual changes in the character of the sound.[11,17] However, it is a good measure of spectral deviations, which is a useful indicator of intelligibility loss.[14]

4.2 Evaluation

The two sentences utilized in the investigation of pitch detection accuracy were also employed in the evaluation of the effects of CVSD distortion on the estimation of the LPC coefficients. For each sentence, the LPC coefficients for the uncoded, original speech are first calculated. The LPC parameters are calculated 50 times per second at a uniform rate using the autocorrelation method[18] with a 30-msec Hamming window. The speech is preemphasized using a first order digital network with transfer function

$$H(z) = 1 - 0.95z^{-1} \tag{10}$$

prior to LPC analysis in order to minimize the effects of performing the LPC analysis at a uniform rate (i.e., pitch asychronously).[19] The results of this analysis provide the reference LPC coefficients (the a_n's) for each 20-msec frame.

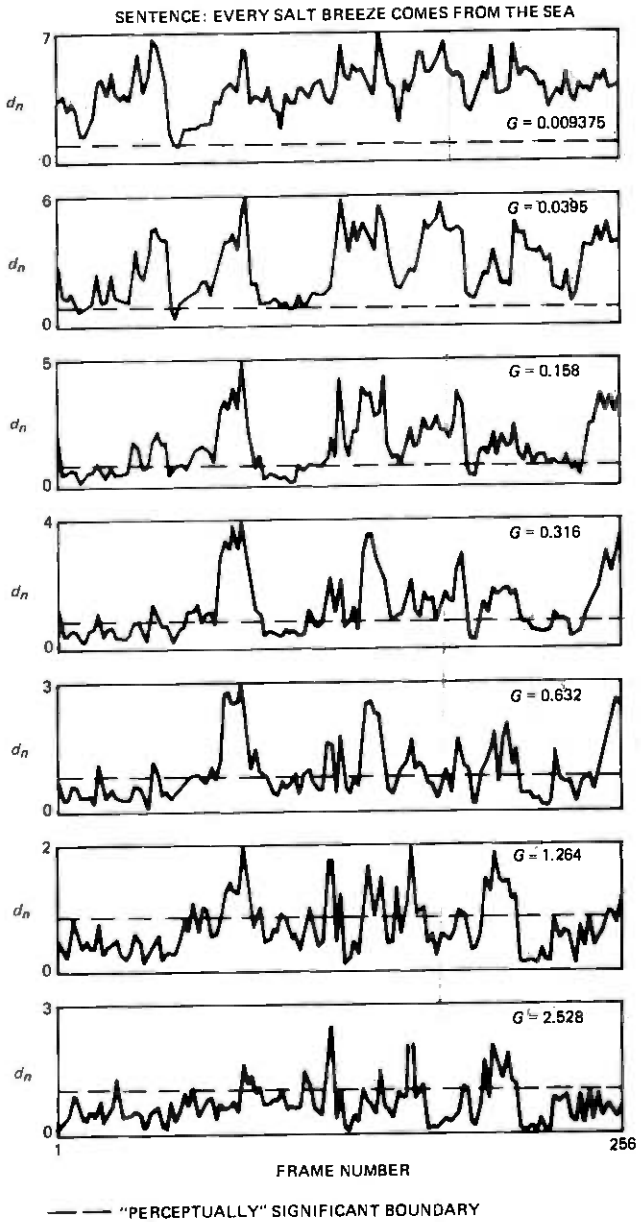


Fig. 10—Values of d_n versus frame number as a function of CVSD signal level for utterance "Every salt breeze comes from the sea."

A similar LPC analysis is performed for each of the various CVSD coded versions of the original sentences. These analyses provide the b_n's for use in the calculation of distance (d_n) between the original sentence and

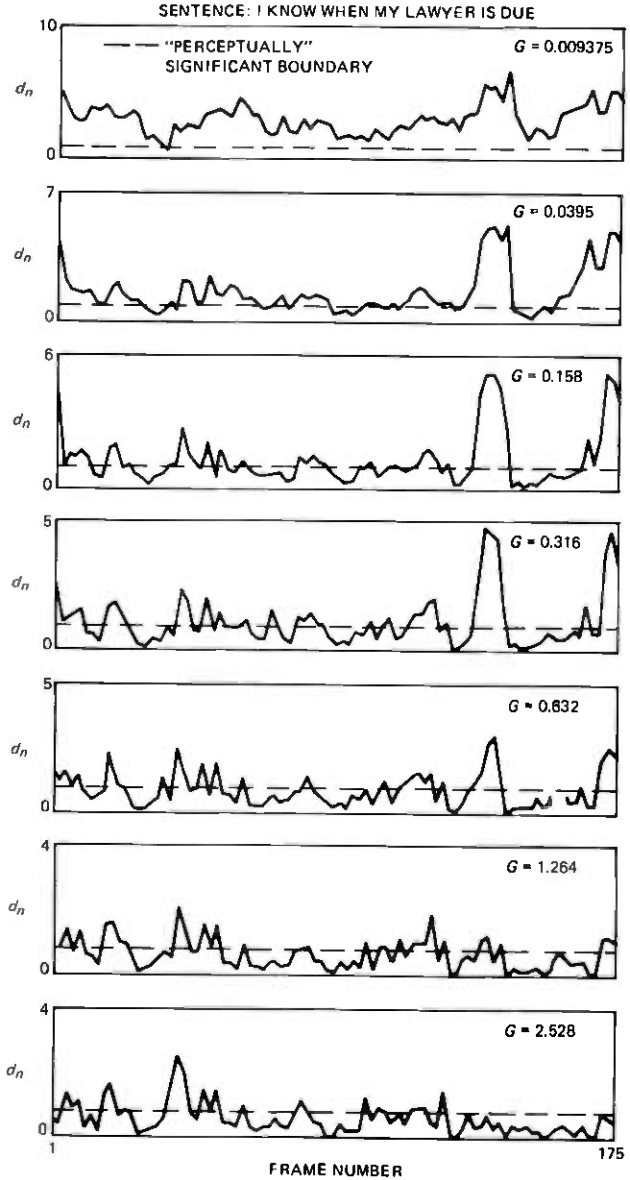


Fig. 11—Values of d_n versus frame number as a function of CVSD signal level for utterance "I know when my lawyer is due."

the particular CVSD-coded sentence. Figures 10 and 11 show the frame-by-frame LPC distance measured for each CVSD-coded version of the two original sentences. The dashed line in the figures refers to a suggested threshold of $d_n = 0.9$ for a just-perceptible difference.[14] Figure

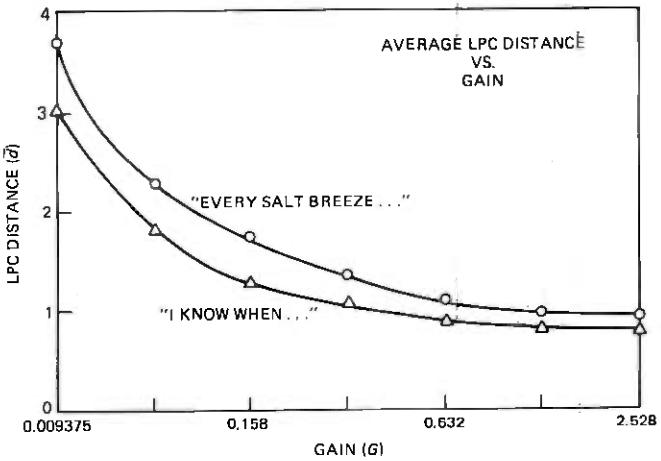


Fig. 12—Plots of average LPC distance ($\bar{d}$) as a function of CVSD signal level (G) for both test sentences.

12 shows the average LPC distance as a function of gain G. The average distance is defined as

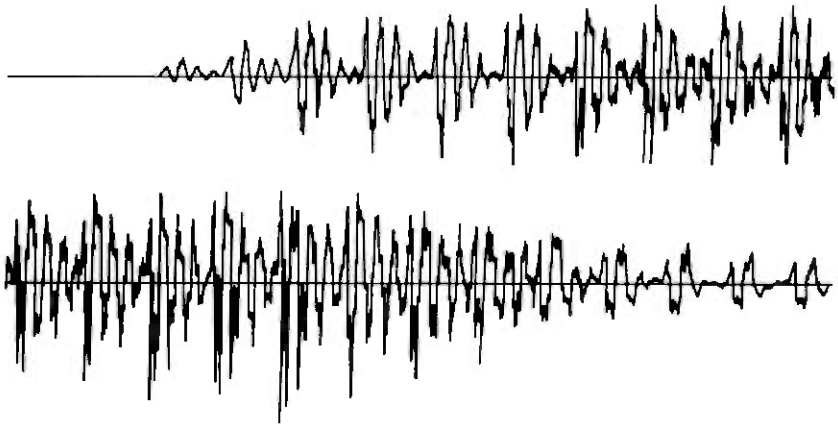
$$\bar{d} = \frac{1}{M} \sum_{n=1}^{M} d_n \tag{11}$$

where M is the number of frames in the sentence.

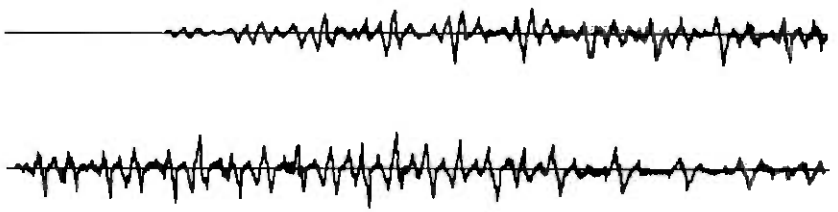
The results of the LPC distance analyses are striking in that the distance uniformly decreases as the gain G increases. This result is in direct opposition to the SNR findings discussed in the first part of this paper.[1] According to the LPC distance measure, the CVSD-coded sentence is improving in quality (i.e., closer in distance to the original) as the gain increases. However, according to the SNR measurements, the similarity between the original and the CVSD-coded sentence is decreasing as the gain increases beyond $G = 0.158$. Although the dissimilarity between the waveforms of the original and the CVSD-coded version with $G = 1.264$ is apparent from Fig. 13, it is interesting to note that informal perceptual experiments indicate that the quality of the CVSD coder is actually improving as the gain G increases. Since the LPC distance measure is sensitive to spectral distortions, it is (in this case) a better measure of quality than SNR. The use of the LPC distance measure as an indication of speech quality has been suggested by other authors.[14]

V. COMPATIBILITY OF CVSD WITH LPC

As a final check on the performance of the entire system, an informal perceptual evaluation of the CVSD-LPC tandem link depicted in Fig. 1 was performed. The LPC vocoder was efficiently designed for a bit rate



(a) ORIGINAL WAVEFORM



(b) CVSD–CODED WAVEFORM (G = 1.264)

Fig. 13—Waveform plots of one section of an utterance and the resulting output of the CVSD coder for G = 1.264.

of 2.4 kb/s[20] and the CVSD was designed for 16 kb/s operation using the various gains G. For the smallest gain, G = 0.009375, the speech was unintelligible. For the higher gains, the output speech was intelligible, but the quality was significantly worse than the quality of the 2.4 kb/s LPC synthesis. The quality of the tandem link appeared to saturate (or even become slightly worse due to the poorer estimates of pitch and gain) for $G \geq 0.158$. Even for the best-quality output, the combination of CVSD noise and the parametric distortions of the LPC vocoder rendered the tandem a marginal communications link.

VI. SUMMARY

In the tandem link of a wideband and narrowband speech communication system in which the wideband system was a 16 kb/s CVSD coder and the narrowband system was a 2.4 kb/s LPC vocoder, the CVSD coder was shown to be the weak link. The major distortion introduced by the

CVSD coder was spectral distortion as measured using an appropriate LPC distance measure. This distortion was sufficiently severe to make the LPC output, although intelligible, of poor quality. It was further shown that the waveform distortion in the CVSD coder was not so severe so as to make pitch detection unreliable, and even a reliable voiced-unvoiced decision could be made on the CVSD-coded speech.

The major conclusion from this study is that alternative 16-kb/s coders be considered as the wideband communication system for such communication links. Possible alternatives include ADPCM systems,[21] subband coders,[22] and transform coders.[23]

REFERENCES

1. R. E. Crochiere, D. J. Goodman, L. R. Rabiner, and M. R. Sambur, "Tandem Connections of Wideband and Narrowband Speech Communications Systems: Part 1—Narrowband-to-Wideband Link," B.S.T.J., this issue, pp. 0000–0000.
2. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch-Detection Algorithms," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-24*, No. 5 (October 1976), pp. 399–418.
3. C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-25*, No. 3 (June 1977), pp. 221–229.
4. M. J. Ross et al., "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-22* (October 1974), pp. 353–362.
5. J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," IEEE Trans. Acoustics, Speech Signal Proc., *ASSP-24* (February 1976), pp. 2–8.
6. B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-24*, No. 3 (June 1976), pp. 201–121.
7. R. J. McAulay, "Optimum Classification of Voiced Speech, Unvoiced Speech and Silence in the Presence of Noise and Interference," Lincoln Laboratory Technical Note 1976–7, June 1976.
8. L. R. Rabiner and M. R. Sambur, "Application of an LPC Distance Metric to the Voiced-Unvoiced-Silence Detection Problem," submitted for publication.
9. C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A Semi-Automatic Pitch Detector," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-23* (December 1975), pp. 570–574.
10. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-23*, No. 6 (December 1975), pp. 552–557.
11. M. R. Sambur and N. S. Jayant, "Speech Encryption by Manipulations of LPC Parameters," B.S.T.J., *55*, No. 9 (November 1976), pp. 1373–1388.
12. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 67–72.
13. A. H. Gray Jr. and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-24* (October 1976), pp. 380–391.
14. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis From Speech Inputs Containing Noise or Additive White Noise," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-24* (December 1976), No. 6.
15. H. Wakita, "On the Use of Linear Prediction Error Energy for Speech and Speaker Recognition," J. Acoust. Soc. Amer., *57*, Supplement No. 1, Spring 1975 (A).
16. D. T. Magill, "Adaptive Speech Compression for Packet Communication Systems," Telecommunications Conference Record, IEEE Publ. 73, CH0805-2, 29D 1-5.
17. J. R. Makhoul, L. Viswanathan, L. Cosel, and W. Russel, "Natural Communication with Computers: Speech Compression Research at BBN," BBN Report No. 2976, *II*, Bolt Beranek and Newman, Inc., Cambridge, Massachusetts, December 1974.

18. J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech,* Springer Verlag, 1976.
19. L. R. Rabiner, B. S. Atal, and M. R. Sambur, "LPC Prediction Error—Analysis of its Variation with the Position of the Analysis Frame," submitted for publication.
20. M. R. Sambur, "An Efficient Linear Prediction Vocoder," B.S.T.J., *54,* No. 10 (December 1975), pp. 1693–1723.
21. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52,* No. 7 (September 1973), pp. 1105–1118.
22. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Subbands," B.S.T.J., *55,* No. 8 (October 1976), pp. 1069–1086.
23. R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals," to appear in IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-25,* No. 4 (August 1977), pp. 299–309.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Discrete-Time Single Server Queues with Correlated Inputs

By B. GOPINATH and J. A. MORRISON

(Manuscript received April 15, 1977)

A wide variety of queueing systems with a single server can be modeled by the equation $b_{n+1} = (b_n - 1)^+ + z_n$, where b_n denotes queue length and z_n the input. The usual assumption about the sequence $\{z_n\}$ is that it be a sequence of independent identically distributed (i. i. d.) random variables. However, in many applications, this is not really the case; $\{z_n\}$ is a sequence of correlated random variables. We show that with the help of a transformation, a $(k+1)$-dimensional Markov process that suffices to describe the queueing system may be found, where k is the memory of the input process. We derive an equation for the steady-state generating function corresponding to the joint distribution of this vector process. We find that a simple set of equations can be obtained for the marginal distributions. In particular, the steady-state distribution of b_n, the queue length, can be obtained without solving for the joint distribution.

I. INTRODUCTION

Several computer systems and networks involve queueing models with single server queues. We consider a discrete-time queueing system, with service time normalized to unity, modeled by the equation

$$\begin{aligned} b_{n+1} &= b_n - 1 + z_n \text{ if } b_n \geq 1 \\ &= z_n \qquad \text{if } b_n = 0 \end{aligned}$$

or equivalently

$$b_{n+1} = (b_n - 1)^+ + z_n \tag{1}$$

Here b_n denotes queue length[1] and the nonnegative integer valued sequence z_n is the input.

A vast majority of literature in queueing theory deals with the case when $\{z_n\}$ is a sequence of independent identically distributed random

variables. In this situation, when the average value $Ez_n < 1, b = \lim_{n\uparrow\infty}{}^{*} b_n$ is a well-defined random variable, and various authors have analyzed the distribution of b; see Ref. 1.

An interesting approach is due to Spitzer[2] who uses a simple consequence of eq. (1): when $b_0 = 0$ then

$$b_{n+1} = \max_r \left\{ \sum_{i=0}^{r} z_{n-i} - r \right\} \tag{2}$$

to derive an integral equation for the distribution of b. However, we will follow the approach that models $\{b_n\}$ as a Markov process as in Ref. 3. Here the theory of Markov chains can be used to derive formulas for the equilibrium distribution of b_n, that is, the distribution of b.

The literature dealing with models where $\{z_n\}$ are not necessarily independent is relatively scant. Recently Ali Khan[4] and Herbert[5] have analyzed the case when z_n is the state of a denumerable Markov chain. In this case (b_n, z_n) forms a Markov process, thus relaxing somewhat the condition that $\{z_n\}$ are independent identically distributed (i. i. d.) random variables.

The queueing process that motivated the work presented in this paper arose in a data communications system. Messages are temporarily stored in a buffer before they are sent across the communications network. It is assumed that the buffer transmits one packet, the basic unit of data, in a unit time interval, provided that it is not empty. In this context, then, z_n is the number of packets that arrive at the buffer in the time interval $(n, n+1]$. It is assumed that the inputs are correlated and z_n is taken to be a sum of moving averages.

In order to illustrate the techniques, the particular example $z_n = x_n^1 + x_{n-2}^1 + x_n^2$ is first analyzed. This corresponds to the arrival of two kinds of messages. The first kind of message consists of two packets which are spread apart in time, the second packet being transmitted two units of time after the first packet. The number of such messages generated in the $(n+1)$st time unit is denoted by x_n^1. The second kind of message consists of just one packet, and the number of such messages generated in the $(n+1)$st time unit is denoted by x_n^2. It is assumed that (x_n^1, x_n^2), $n = 0, 1, 2, \ldots$, are independent identically distributed vector random variables. However, for each n, x_n^1 and x_n^2 may be dependent. In particular, if

$$E(t_1^{x_n^1} t_2^{x_n^2}) = \Phi[(1-\rho)t_1 + \rho t_2]$$

with $0 \le \rho \le 1$ fixed, then the probability that a message is of the first kind is $1 - \rho$, and the probability that it is of the second kind is ρ.

* We mean here limit in distribution: for each j,

$$\lim_{n\uparrow\infty} Pr\{b_n \le j\} = Pr\{b \le j\}$$

There are several other examples where such a model for the input process z_n is more appropriate than the usual one. We give two examples. Consider a queueing system where each request for service may consist of a sequence of tasks to be completed by the same server. However, these tasks may not be available for completion in the same time interval; instead they are spread out in time. Hence the random variables corresponding to the number of tasks arriving at the server may be correlated as in the above example. This model may apply to a scheduler in a computer processing system. Another example, that of a dam fed by rivers that originate at geographically distant points, motivated the model considered by Herbert.[6] When rainfall occurs, affecting the flow in all of the rivers, the increase in flow to the dam is spread out in time since the origins of the rivers are at different distances from the dam. A discrete time model of the dam process, similar to the one in the packet network example above, can be solved by the method presented in this paper.

In general we assume that

$$z_n = \sum_{i=1}^{\ell} \sum_{j=0}^{k} \alpha_j^i x_{n-j}^i \tag{3}$$

where the nonnegative integer valued random variables in the sequence $\{(x_n^1, x_n^2, \ldots, x_n^\ell)\}$ are independent and identically distributed, and α_j^i are nonnegative integers with $\alpha_0^i > 0$ for each i. For each n the random variables $x_n^1, x_n^2, \ldots, x_n^\ell$ may be dependent on each other. Notice that z_n by itself is not necessarily a Markov process. As far as we know there is only one work dealing with a special case of eq. (3) which is related to ours. Herbert[6] considers the case when

$$z_n = \sum_{j=0}^{k} \alpha_j x_{n-j} \tag{4}$$

where $\{x_n\}$ are i. i. d. random variables and α_j are positive integers. In this case whenever $x_n \neq 0, b_{n+i} \neq 0, i = 1, \ldots, k+1$, hence b_{n+r} is linearly related to $b_{n+1}, r = 2, \ldots, k+2$ from eq. (1). From this property, formulas can be derived for the equilibrium distribution for b_n given $x_{n-1}, x_{n-2}, \ldots, x_{n-k}$. However, even in this special case our approach gives formulas for

$$b = \lim_{n \uparrow \infty} b_n$$

itself more simply than the method of Ref. 6.

In the general case b_n is not a Markov process, but it is shown that, with the help of a transformation, a $(k+1)$-dimensional Markov process that suffices to describe the queueing system may be found. The first component of this Markov process is just b_n. An equation is derived for the steady-state generating function corresponding to the joint distri-

butions. This equation involves a multinomial, which corresponds to zero queue length. It is shown that a finite system of linear equations can be obtained to solve for the coefficients in this multinomial. A simple set of equations for the marginal distributions is then derived, leading to the calculation of the steady-state generating function of the queue length.

In Sec. II we review the case when z_n are i. i. d. random variables. An example for a system where z_n is a moving average is worked out in Sec. III to illustrate our method. In Sec. IV we introduce the model considered in this paper and describe the transformation that leads to the simplification in the solution. The generating function of the underlying vector Markov process is derived in Sec. V. The method of solving for certain parameters that occur in Sec. V is described in Sec. VI. The isolation of marginals and the derivation of a simple set of equations for them is the subject of Sec. VII. A pair of limiting cases of the input process is analyzed in Sec. VIII. Finally, for a special class of problems, some formulas relating the limiting cases are also derived in Sec. VIII. The terminology of Markov chains used in this paper is consistent with that of Ref. 3.

II. QUEUE WITH INDEPENDENT INPUTS

When $\{z_n\}$ is a sequence of independent identically distributed random variables, it follows that b_n is a Markov process. The number of packets waiting to be transmitted, b_n, serves as the state for a Markov chain S. The state space of S is the set of nonnegative integers. The transition probabilities for S are generated by eq. (1) as follows:

$$P_i^{n+1} \triangleq Pr\{b_{n+1} = i\} = \sum_{j \geq 0} Pr\{b_{n+1} = i \mid b_n = j\} P_j^n$$

$$= \sum_{j \geq 0} Pr\{z_n = i - (j-1)^+\} P_j^n \quad (5)$$

Let $Pr\{z_n = i\} = p_i$ for $i = 0,1, \ldots$. Then, since z_n is a nonnegative integer,

$$P_i^{n+1} = p_i P_0^n + \sum_{j=1}^{i+1} p_{i-j+1} P_j^n \quad (6)$$

When $1 > p_0 > 0$, S is irreducible and aperiodic. The following theorem gives conditions under which S is positive recurrent.

Theorem 1: The Markov-chain S is positive recurrent when $Ez_n < 1$.

When S is positive recurrent then

$$\lim_{n \uparrow \infty} b_n = b$$

is a well-defined random variable and the equilibrium distribution of b_n, that is, the distribution of b, is such that

$$\lim_{n\uparrow\infty} Pr\{b_n = j\} = \pi_j > 0 \text{ for } j = 0,1,\ldots \tag{7}$$

Furthermore, if $P_j^n = \pi_j$, $j = 0,1,\ldots$, so is P_j^{n+1}, and π_j are the unique nonnegative solution to the infinite system of linear equations:

$$\sum_{j=0}^{\infty} \pi_j = 1$$

and, for $i = 0,1,\ldots$,

$$\pi_i = p_i\pi_0 + \sum_{j=1}^{i+1} p_{i-j+1}\pi_j \tag{8}$$

These are obtained from eq. (6) by substituting $P_i^n = P_i^{n+1} = \pi_i$.

For a proof of the above results see Karlin.[3]

In order to solve for the equilibrium distribution we will employ the method of generating functions. For any random variable x, the generating function of x, $\phi_x(s)$, is defined as

$$\phi_x(s) = Es^x, \ |s| \leq 1 \tag{9}$$

Let

$$\phi_n(s) = Es^{b_n} = \sum_{i=0}^{\infty} P_i^n s^i.$$

Then using eq. (1) and the independence of b_n,z_n we have

$$Es^{b_{n+1}} = Es^{(b_n-1)^+} Es^{z_n} \tag{10}$$

From the definition of ϕ_n it follows that

$$\phi_{n+1}(s) = (s^{-1}\phi_n(s) + (1 - s^{-1})P_0^n)\phi_z(s) \tag{11}$$

where $\phi_z(s) = Es^{z_n}$. Assuming that $Ez_n < 1$ and $1 > p_0 > 0$, let the generating function of

$$b = \lim_{n\uparrow\infty} b_n$$

be

$$\phi(s) = \sum_{i=0}^{\infty} \pi_i s^i$$

[see eq. (7)]. Then from above it is clear that if $\phi_n(s) = \phi(s)$ then $\phi_{n+1}(s) = \phi(s)$. So from eq. (11) we get

$$\phi(s) = \frac{(1 - s^{-1})\pi_0\phi_z(s)}{1 - s^{-1}\phi_z(s)} \tag{12}$$

To find π_0 we take expectations of both sides of eq. (1) and take the limit as $n \uparrow \infty$. Then

$$\pi_0 = 1 - Ez_n \tag{13}$$

So

$$\phi(s) = \frac{(1-s)\phi_z(s)(1-Ez_n)}{\phi_z(s) - s} \tag{14}$$

This gives the generating function of b in terms of $\phi_z(s)$. However, to get π_j, we need not invert the generating function $\phi(s)$. Treating ϕ_z, ϕ as formal power series, using Π_j to denote $\Sigma_{i=0}^{j} \pi_i$, and equating like powers of s in eq. (14), we can show:

$$\begin{aligned} \Pi_0 &= \pi_0 = 1 - Ez_n \\ \Pi_1 &= (\pi_0 p_1 + \Pi_0 - p_1\Pi_0)/p_0 \\ &\vdots \\ \Pi_j &= \left(\pi_0 p_j + \Pi_{j-1} - \sum_{i=1}^{j} p_i \Pi_{j-i}\right) \Big/ p_0 \end{aligned} \tag{15}$$

Equations (15) give explicitly the formulas needed to solve for π_j or Π_j. Notice that any finite number of the π_j's can be determined by solving a finite number of linear equations. Informally we refer to such a situation as being finitely solvable.

III. AN EXAMPLE OF A QUEUING PROCESS WITH CORRELATED INPUTS

In the context of the application discussed in 1, there are instances when the data arriving at the buffer form a sequence of correlated random variables. For an example we consider here a case when there are two classes of sources that generate data. The first kind generates two packets whenever it transmits a message. However, these packets are not generated simultaneously; instead they are spread apart in time, the second packet being transmitted two seconds after the first one. The number of such messages generated in the $(n+1)$st second is denoted by x_n^1. The second class of sources generates messages of one packet each and the number of such messages generated in the $(n+1)$st second is denoted by x_n^2. (x_n^1, x_n^2), $n = 0, 1, 2, \ldots$, are assumed to be independent identically distributed vector random variables. Note that, for each n, x_n^1 and x_n^2 may be dependent. Then the number of packets arriving at the buffer in the $(n+1)$st second is

$$z_n = x_n^1 + x_{n-2}^1 + x_n^2 \tag{16}$$

So the number of packets in the buffer at the end of the $(n+1)$st second

is given as in eq. (1) by

$$b_{n+1} = (b_n - 1)^+ + x_n^1 + x_{n-2}^1 + x_n^2 \tag{17}$$

It is clear b_n is not a Markov process. However, $(b_n, x_{n-1}^1, x_{n-2}^1, x_{n-1}^2, x_{n-2}^2)$ is a five-dimensional Markov process. We will derive another Markov process from eq. (17) that is only three-dimensional and suffices to describe the queueing process. Define

$$\begin{aligned} y_{0n} &= b_n \\ y_{1n} &= y_{0n} + x_{n-2}^1 \\ y_{2n} &= y_{1n} + x_{n-1}^1 \end{aligned} \tag{18}$$

Then from eq. (17) we have

$$\begin{aligned} y_{0,n+1} &= [(y_{0n} - 1)^+ - y_{0n}] + y_{1n} + x_n^1 + x_n^2 \\ y_{1,n+1} &= [(y_{0n} - 1)^+ - y_{0n}] + y_{2n} + x_n^1 + x_n^2 \\ y_{2,n+1} &= [(y_{0n} - 1)^+ - y_{0n}] + y_{2n} + 2x_n^1 + x_n^2 \end{aligned} \tag{19}$$

Let $v_{0n} = v_{1n} = x_n^1 + x_n^2$ and $v_{2n} = 2x_n^1 + x_n^2$. Then (v_{0n}, v_{1n}, v_{2n}) is independent of (y_{0j}, y_{1j}, y_{2j}) for $j \leq n$ by assumptions about x_n^1, x_n^2. Hence (y_{0n}, y_{1n}, y_{2n}) is a three-dimensional Markov process. The state space of the corresponding Markov-chain S can naturally be indexed by a triple of nonnegative integers. Let

$$P_{i_0,i_1,i_2}^n = Pr\{y_{0n} = i_0, y_{1n} = i_1, y_{2n} = i_2\} \tag{20}$$

Then

$$P_{i_0,i_1,i_2}^{n+1} = \sum_{j_0,j_1,j_2} Pr\{y_{0,n+1} = i_0, y_{1,n+1} = i_1, y_{2,n+1} = i_2 | y_{0n} = j_0, y_{1n} = j_1, y_{2n} = j_2\} P_{j_0,j_1,j_2}^n \tag{21}$$

These form the equations for transition probabilities. Notice that not all states (i_0, i_1, i_2) communicate with (0, 0, 0). For example, we can show that when $i_0 = 0$, the only states that communicate with (0, 0, 0) are (0, 0, 0) and (0, 1, 1). Suppose $y_{0,n+1} = 0$, $y_{1,n+1} = i_1$ and $y_{2,n+1} = i_2$. Then $b_{n+1} = 0$. Hence, from eq. (17), $b_n \leq 1, x_n^1 = 0$. But $x_n^1 = 0$ implies $y_{2,n+1} = y_{1,n+1}$. Also, $b_n \leq 1$ implies $x_{n-1}^1 \leq 1$. Further, $y_{0,n+1} = 0$ and $x_{n-1}^1 \leq 1$ imply $y_{1,n+1} \leq 1$. However, it can be shown that states that do not communicate with (0, 0, 0) are transient (see Sec. IV). So we will restrict the state space by allowing it to consist only of those states, denoted by $\mathcal{A}$, that communicate with zero. We will continue to denote by S the Markov chain on the restricted state space $\mathcal{A}$. Then S is irreducible and aperiodic (see Sec. IV). Notice that for every state at time n

$$y_{0n} \leq y_{1n} \leq y_{2n} \tag{22}$$

Later in this paper we will show that S is positive recurrent when $Ez_n < 1$. For now we will assume this is so. Interpreting the sums over j_0, j_1, j_2 to extend only over $\mathcal{A}$ we have from eqs. (19) and (21), and the definitions of v_{0n}, v_{1n}, v_{2n},

$$P^{n+1}_{i_0,i_1,i_2} = \sum'_{j_1-i_0=j_2-i_1} Pr\{v_{0n} = i_0 - j_1, v_{2n} = i_2 - j_2\} P^n_{0,j_1,j_2}$$

$$+ \sum'_{j_0>0, j_1-i_0=j_2-i_1} Pr\{v_{0n} = 1 + i_0 - j_1, v_{2n} = 1 + i_2 - j_2\} P^n_{j_0,j_1,j_2} \quad (23)$$

The equilibrium distribution of S:

$$\lim_{n\uparrow\infty} P^n_{i_0,i_1,i_2} = P_{i_0,i_1,i_2}$$

has the property that if $P^n_{i_0,i_1,i_2} = P_{i_0,i_1,i_2}$ for $(i_0, i_1, i_2) \in \mathcal{A}$, so does $P^{n+1}_{i_0,i_1,i_2}$. So P_{i_0,i_1,i_2} satisfies:

$$P_{i_0,i_1,i_2} = \sum'_{j_1-i_0=j_2-i_1} Pr\{v_{0n} = i_0 - j_1, v_{2n} = i_2 - j_2\} P_{0,j_1,j_2}$$

$$+ \sum'_{j_0>0, j_1-i_0=j_2-i_1} Pr\{v_{0n} = 1 + i_0 - j_1, v_{2n} = 1 + i_2 - j_2\} P_{j_0,j_1,j_2} \quad (24)$$

$$\sum_{(i_0,i_1,i_2)\in\mathcal{A}} P_{i_0,i_1,i_2} = 1$$

P_{i_0,i_1,i_2} is the unique nonnegative solution of eq. (24) (see Ref. 3). In principle, solving the infinite system of linear eq. (24) determines P_{i_0,i_1,i_2}, hence the equilibrium distribution of (y_{0n}, y_{1n}, y_{2n}). However we will see a much simpler way to find equilibrium distributions of the components y_{0n}, y_{1n}, y_{2n}, without computing P_{i_0,i_1,i_2}. Denote $Es^{y_{in}}$ by $\phi_{ni}(s)$ and $Es^{v_{in}}$ by $\phi_{iv}(s)$. Then from eq. (19) we can derive the following equations paralleling eq. (11):

$$\begin{aligned}
\phi_{n+1,0}(s) &= [s^{-1}\phi_{n1}(s) + (1 - s^{-1})c_{1n}(s)]\phi_{0v}(s) \\
\phi_{n+1,1}(s) &= [s^{-1}\phi_{n2}(s) + (1 - s^{-1})c_{2n}(s)]\phi_{1v}(s) \\
\phi_{n+1,2}(s) &= [s^{-1}\phi_{n2}(s) + (1 - s^{-1})c_{2n}(s)]\phi_{2v}(s)
\end{aligned} \quad (25)$$

Here

$$c_{in}(s) = \sum_{j\geq 0} Pr\{y_{0n} = 0, y_{in} = j\}s^j, \; i = 1,2 \quad (26)$$

For any n the only admissable states in $\mathcal{A}$ that have $y_{0n} = 0$ are $(0, 0, 0)$ and $(0, 1, 1)$. So $c_{in}(s)$, $i = 1,2$ are polynomials of degree 1, and $c_{1n}(s) =$

$c_{2n}(s)$. Let $\phi_i(s)$ denote the generating function of

$$y_i = \lim_{n\uparrow\infty} y_{in}$$

and

$$c_i(s) = \lim_{n\uparrow\infty} c_{in}(s)$$

for $i = 1,2$. Then

$$\begin{aligned}
\phi_0(s) &= [s^{-1}\phi_1(s) + (1 - s^{-1})c_1(s)]\phi_{0v}(s) \\
\phi_1(s) &= [s^{-1}\phi_2(s) + (1 - s^{-1})c_1(s)]\phi_{1v}(s) \\
\phi_2(s) &= [s^{-1}\phi_2(s) + (1 - s^{-1})c_1(s)]\phi_{2v}(s)
\end{aligned} \tag{27}$$

From eq. (27)

$$\phi_2(s) = \frac{(1 - s^{-1})c_1(s)\phi_{2v}(s)}{1 - s^{-1}\phi_{2v}(s)} \tag{28}$$

Since ϕ_{0v}, ϕ_{1v}, ϕ_{2v} are known directly from the distribution of x_n^1, x_n^2, eq. (27) gives ϕ_0,ϕ_1 in terms of $c_1(s)$, the only unknown. Let $c_1(s) = k_0 + k_1 s$. Then

$$\begin{aligned}
c_1(1) &= Pr\{y_0 = 0, y_1 = 0\} + Pr\{y_0 = 0, y_1 = 1\} \\
&= Pr\{y_0 = 0\} = k_0 + k_1
\end{aligned}$$

As in eq. (13)

$$\begin{aligned}
k_0 + k_1 &= 1 - E(x_n^1 + x_{n-2}^1 + x_n^2) \\
&= 1 - Ez_n
\end{aligned} \tag{29}$$

In order to derive another equation for k_0,k_1 we go back to the original equations for $P_{i_0i_1i_2}$, eq. (24). From eq. (24) we can derive the following: for $i_0 = i_1 = i_2 = 0$, since v_{in} are nonnegative, $P_{000} = Pr\{v_{0n} = 0, v_{2n} = 0\}P_{000} + Pr\{v_{0n} = 0, v_{2n} = 0\}P_{111}$. However, $v_{2n} = 2x_n^1 + x_n^2 = 0$ implies $x_n^1 = x_n^2 = 0$, so $v_{0n} = 0$. Therefore

$$k_0 = P_{000} = Pr\{v_{0n} = 0\}(P_{000} + P_{111}) \tag{30}$$

Similarly

$$\begin{aligned}
k_1 = P_{011} &= Pr\{v_{0n} = 0\}P_{112} \\
P_{112} &= Pr\{v_{0n} = 1, v_{2n} = 2\}P_{111} \\
&\quad + Pr\{v_{0n} = 1, v_{2n} = 2\}P_{000}
\end{aligned}$$

Hence

$$k_1 = Pr\{v_{0n} = 0\}Pr\{v_{0n} = 1, v_{2n} = 2\}(P_{000} + P_{111}) \tag{31}$$

Notice that the various probabilities occurring on the right-hand sides of eqs. (30), (31) can be calculated from the distribution of (x_n^1, x_n^2). For example:

$$Pr\{v_{0n} = 1, v_{2n} = 2\} = Pr\{x_n^1 = 1, x_n^2 = 0\}$$

Therefore using eq. (29) we can determine k_0, k_1, hence $c_1(s)$. From eq. (27), therefore, it is easy to derive the formula for $\phi_0(s)$, namely

$$\phi_0(s) = (1 - s^{-1})c_1(s)\left[\frac{s^{-2}\phi_{2v}(s)\phi_{1v}(s)\phi_{0v}(s)}{1 - s^{-1}\phi_{2v}(s)} + s^{-1}\phi_{1v}(s)\phi_{0v}(s) + \phi_{0v}(s)\right]$$

$$= (1 - s^{-1})c_1(s)\phi_{0v}(s)\left[1 + \frac{s^{-1}\phi_{1v}(s)}{1 - s^{-1}\phi_{2v}(s)}\right] \tag{32}$$

To solve for the equilibrium distribution of b_n, i.e., distribution of y_0, we do not have to invert $\phi_0(s)$. It turns out that eq. (27) can be translated to linear recursions for marginal distributions for y_0, y_1, y_2. Hence, as in Sec. II, the distributions of y_0, y_1, y_2 are finitely solvable. That this is so, in the general case, is shown in Sec. VII.

IV. QUEUEING PROCESSES WITH MOVING AVERAGE INPUTS

The most general input process that we will consider in this paper is a finite sum of moving averages, i.e.,

$$z_n = \sum_{i=1}^{\ell}\sum_{j=0}^{k} \alpha_j^i x_{n-j}^i \tag{33}$$

Equation (1) in this setting is

$$b_{n+1} = (b_n - 1)^+ + \sum_{i=1}^{\ell}\sum_{j=0}^{k} \alpha_j^i x_{n-j}^i \tag{34}$$

The integer k is referred to as memory of the input process z_n. Under the assumptions below, the $(k\ell + 1)$ dimensional vector process $(b_n, x_{n-1}^1, x_{n-2}^1, \ldots, x_{n-k}^1, x_{n-1}^2, \ldots, x_{n-k}^2, \ldots, x_{n-1}^\ell, \ldots, x_{n-k}^\ell)$ is Markov as in the example of Sec. III.

However, by a transformation we will find a $(k + 1)$ dimensional Markov process that suffices to describe the queueing system. Define:

$$y_{0n} = b_n$$

and, for $r = 0, 1, \ldots, k - 1$,

$$y_{r+1,n} = y_{rn} + \sum_{i=1}^{\ell}\sum_{j=r+1}^{k} \alpha_j^i x_{n-j+r}^i \tag{35}$$

Let

$$\sum_{j=0}^{r} \alpha_j^i = \mu_r^i \text{ and } \sum_{i=1}^{\ell} \mu_r^i x_n^i = v_{rn}$$

for $r = 0, 1, \ldots, k$. Then using eq. (34) we can verify:

$$y_{r,n+1} = [(y_{0n} - 1)^+ - y_{0n}] + y_{r+1,n} + v_{rn}, \; r = 0, 1, \ldots, k-1 \quad (36)$$

$$y_{k,n+1} = [(y_{0n} - 1)^+ - y_{0n}] + y_{kn} + v_{kn} \quad (37)$$

We make the following assumptions for the rest of this paper: The α_j^i are assumed to be nonnegative integers and, for each i, $\alpha_0^i > 0$. We will assume that the vector, nonnegative integer valued random variables $(x_n^1, x_n^2, \ldots, x_n^\ell)$ are independent and identically distributed, though for each n, $x_n^1, x_n^2, \ldots, x_n^\ell$ will be allowed to be dependent on each other. We will also assume that $Pr\{\mathbf{v}_n = 0\} > 0$ and $Pr\{v_{rn} > 1\} > 0$ for some r.

From the assumptions about x_n^i, $(v_{0n}, v_{1n}, \ldots, v_{kn})^t \equiv \mathbf{v}_n$ is independent of $\mathbf{y}_j \equiv (y_{0j}, y_{1j}, \ldots, y_{kj})^t$ for $j \leq n$. Hence $\mathbf{y}_n$ is a $(k+1)$ dimensional Markov process. The state space corresponding to this Markov process is indexed naturally by a $(k+1)$ triple of nonnegative integers. Furthermore by definition of y_{in}, $i = 0, 1, \ldots, k$, $n = 0, 1, 2, \ldots$, $y_{0n} \leq y_{1n} \leq y_{2n} \leq \cdots \leq y_{kn}$. Hence we can assume that if $(i_0, i_1, \ldots, i_k)$ denotes a state then

$$i_0 \leq i_1 \leq i_2 \leq \cdots \leq i_k \quad (38)$$

Let $\mathcal{A}'$ denote the set of vectors satisfying (38) and S' the Markov chain with state space $\mathcal{A}'$. Of the states in $\mathcal{A}'$ let $\mathcal{A}$ denote the set of states that communicate with the state $\mathbf{0} = (0, 0, \ldots, 0)^t$. Using the following theorem, we will be able to restrict our attention to only those states that are in $\mathcal{A}$, and to the irreducible Markov chain S, with state space $\mathcal{A}$, derived from S'.

Theorem 2:

(*i*) Every state in $\mathcal{A}'$ of the form $(m, m, \ldots, m)^t$ belongs to $\mathcal{A}$.

(*ii*) Every state of S' transitions to a state belonging to $\mathcal{A}$ in at most k steps.

(*iii*) Every state in $\mathcal{A}$ is accessible from a state of the form $(m, m, \ldots, m)^t$ in at most k steps.

(*iv*) S is irreducible and aperiodic.

(*v*) For each i_0, the number of states in $\mathcal{A}$ which are of the form $(i_0, i_1, \ldots, i_k)^t$ is finite.

Proof: Let F denote the $(k+1) \times (k+1)$ matrix with elements $F_{i,i+1} = 1$, for $i = 0, \ldots, k-1$, $F_{kk} = 1$, and $F_{ij} = 0$ otherwise. Also, let $\mathbf{1}$ denote the vector $(1, 1, \ldots, 1)^t$. We note that $F\mathbf{1} = \mathbf{1}$ and, for any $\mathbf{y} = (y_0, y_1, \ldots, y_k)^t$, $F^r\mathbf{y} = (y_r, \ldots, y_k, y_k, \ldots, y_k)^t$ by induction. Equations (36) and

(37) can then be written in vector form as follows, using σ_n to denote $y_{0n} - (y_{0n} - 1)^+$. Note that $\sigma_n = 1$ if $y_{0n} > 0$, and $\sigma_n = 0$ if $y_{0n} = 0$.

$$\mathbf{y}_{n+1} = F\mathbf{y}_n - \sigma_n \mathbf{1} + \mathbf{v}_n, n = 0, 1, 2, \ldots$$

We can then show that for $n - 1 \geq i \geq 0$

$$\mathbf{y}_n = F^{n-i}\mathbf{y}_i - \sum_{j=i}^{n-1} \sigma_j \mathbf{1} + \sum_{j=i}^{n-1} F^{n-1-j}\mathbf{v}_j \tag{39}$$

Hence, if $\mathbf{v}_j = 0$ for $j = n - 1, \ldots, n - k$, then it follows from (39), with $i = n - k$, that $\mathbf{y}_n$ is a vector of the form $m\,\mathbf{1} = (m, m, \ldots, m)^t$ for some nonnegative integer m. Therefore, since $Pr\{\mathbf{v}_j = 0\} > 0$ by assumption, $\mathbf{0}$ is accessible from any state by allowing $\mathbf{v}_j$ to be zero for as many consecutive j's as needed. We assumed earlier that $Pr\{v_{rn} > 1\} > 0$ for some r. Hence, since $v_{kn} \geq v_{rn}$, $Pr\{v_{kj} = M\} > 0$ for some integer $M > 1$ and all j. Therefore, if $\mathbf{y}_0 = \mathbf{0}$, $v_{kj} = M$ for $j = 0, 1, \ldots, n - 1$ implies $y_{kn} > nM - n$. Hence $Pr\{y_{kn} > nM - n, \mathbf{y}_0 = 0\} > 0$. For every sequence $\mathbf{v}_j$ such that $\mathbf{v}_j = \mathbf{0}$ for $j = n, n + 1, \ldots, n + k, \ldots$, $\mathbf{y}_{n+k+i}$ remains proportional to 1 for all $i \geq 0$. Therefore, for each m, $Pr\{\mathbf{y}_{n+k+i} = m\,\mathbf{1}, \mathbf{y}_0 = 0\}$ is greater than zero for some n,i dependent on m. From (39) we can therefore show that any state of the form $m\,\mathbf{1}$ communicates with $\mathbf{0}$ and hence belongs to $\mathcal{A}$.

If $\mathbf{y}_0 \in \mathcal{A}'$, then we will prove, irrespective of what $\mathbf{v}_j$'s are for $j = 0, 1, \ldots, k - 1$, that $\mathbf{y}_k \in \mathcal{A}$, by showing that $\mathbf{y}_k$ is accessible from a state of the form $m\,\mathbf{1}$ in at most k steps, where

$$m = k + y_{k0} - \sum_{j=1}^{k} \sigma_{k-j} \tag{40}$$

Let $\mathbf{y}'_j$, $j = 0, 1, \ldots, k$ be the sequence of states traced by S' if $\mathbf{y}_0$ is set to zero but $\mathbf{v}_j$, $j = 0, 1 \ldots, k - 1$ are left unchanged. If $\sigma'_j = y'_{0j} - (y'_{0j} - 1)^+$ for $j = 0, 1, 2 \ldots$ then (39) holds with primes on $\mathbf{y}_j$'s and σ_j's, and $\mathbf{y}'_0 = 0$. We will first prove that for each k, $\mathbf{y}_k \geq \mathbf{y}'_k$ by showing that for each i:

$$\begin{aligned}\mathbf{y}_i - \mathbf{y}'_i \geq 0, F(\mathbf{y}_i - \mathbf{y}'_i) - (\mathbf{y}_i - \mathbf{y}'_i) \geq 0 \\ \Rightarrow \mathbf{y}_{i+1} - \mathbf{y}'_{i+1} \geq 0, F(\mathbf{y}_{i+1} - \mathbf{y}'_{i+1}) - (\mathbf{y}_{i+1} - \mathbf{y}'_{i+1}) \geq 0\end{aligned} \tag{41}$$

Suppose the assumptions in (41) hold as they do for $i = 0$. Then $y_{ki} - y'_{ki} \geq y_{k-1,i} - y'_{k-1,i} \geq \cdots \geq y_{0i} - y'_{0i} \geq 0$. Therefore if $y_{0i} > y'_{0i}$, then $(\mathbf{y}_i - \mathbf{y}'_i) + (\sigma'_i - \sigma_i)\,\mathbf{1} \geq 0$, which is trivially so if $y_{0i} = y'_{0i}$. Hence, using (39) and corresponding equations for $\mathbf{y}'_{i+1}$,

$$\mathbf{y}_{i+1} - \mathbf{y}'_{i+1} = F(\mathbf{y}_i - \mathbf{y}'_i) + (\sigma'_i - \sigma_i)\mathbf{1} \geq (\mathbf{y}_i - \mathbf{y}'_i) + (\sigma'_i - \sigma_i)\mathbf{1} \geq 0$$

Furthermore

$$F(\mathbf{y}_{i+1} - \mathbf{y}'_{i+1}) - (\mathbf{y}_{i+1} - \mathbf{y}'_{i+1}) = F[F(\mathbf{y}_i - \mathbf{y}'_i) - (\mathbf{y}_i - \mathbf{y}'_i)] \geq 0$$

since all elements of F are nonnegative. Hence in particular,

$$y_{kk} - y'_{kk} = y_{k0} + \sum_{j=1}^{k} \sigma'_{k-j} - \sum_{j=1}^{k} \sigma_{k-j} \geq 0$$

Therefore from (40), with

$$s = \sum_{j=1}^{k} \sigma'_{k-j}, \; s + m \geq k.$$

Now let $\mathbf{y}''_i$ be the sequence of states traced by S' if $\mathbf{y}_0$ were set to $m\mathbf{1}$ while $\mathbf{v}_j$, $j = 0, 1, \ldots, k-1$ were left unchanged. Then, with $\sigma''_i = y''_{0i} - (y''_{0i} - 1)^+$,

$$\mathbf{y}''_i - \mathbf{y}'_i = \left(m + \sum_{j=1}^{i} \sigma'_{i-j} - \sum_{j=1}^{i} \sigma''_{i-j}\right)\mathbf{1} \tag{42}$$

for each i. As before, using (41) we can show $\mathbf{y}''_i - \mathbf{y}'_i \geq 0$. From (42) we have $\mathbf{y}''_{i+1} - \mathbf{y}'_{i+1} = \mathbf{y}''_i - \mathbf{y}'_i + (\sigma'_i - \sigma''_i)\mathbf{1}$, hence $\mathbf{y}''_i = \mathbf{y}'_i \Rightarrow \sigma''_i = \sigma'_i \Rightarrow \mathbf{y}''_{i+1} = \mathbf{y}'_{i+1}$. If $\sigma''_i = 0$ for some $i < k$, then $y''_{0i} = 0$, but $y''_{0i} \geq y'_{0i}$, hence $y''_{0i} = y'_{0i}$ and $\mathbf{y}''_i = \mathbf{y}'_i$ from (42). So $\sigma''_i = 0 \Rightarrow \mathbf{y}''_n = \mathbf{y}'_n$ for $k \geq n \geq i$. So in particular

$$\mathbf{y}''_k - \mathbf{y}'_k = \left(m + \sum_{j=1}^{k} \sigma'_{k-j} - \sum_{j=1}^{k} \sigma''_{k-j}\right)\mathbf{1} = 0 \tag{43}$$

However, we noted earlier that

$$s + m = m + \sum_{j=1}^{k} \sigma'_{k-j} \geq k$$

Hence (43) can only hold if

$$s + m = k = \sum_{j=1}^{k} \sigma''_{k-j}$$

Therefore we have shown that for each $i < k$, $\sigma''_i = 1$. Now we use eqs. (40) and (41) to show

$$\mathbf{y}''_k - \mathbf{y}_k = (m - y_{k0})\mathbf{1} - \sum_{j=1}^{k} \sigma''_{k-j}\mathbf{1} + \sum_{j=1}^{k} \sigma_{k-j}\mathbf{1} = \mathbf{0} \tag{44}$$

Since $\mathbf{y}''_k$ belongs to $\mathcal{A}$, being accessible from $m\mathbf{1}$ belonging to $\mathcal{A}$, we have shown that, starting from any state in $\mathcal{A}'$, S' transitions into a state in $\mathcal{A}$ in at most k steps. Furthermore, every state of $\mathcal{A}$ is accessible from some state of the form $m\mathbf{1}$ in at most k steps. It is clear from the definition of $\mathcal{A}$ that S is irreducible. To show that S is aperiodic, we merely note that $\mathbf{y}_j$ can equal zero for arbitrarily many consecutive j's with positive probability.

We will now prove that the set of states in $\mathcal{A}$ which are of the form $(0, i_1, \ldots, i_k)^t$ is finite. This result is used later to derive conditions for the

positive recurrence of S. We just showed that every state in $\mathcal{A}$ is accessible from a state of the form $m\mathbf{1}$ in k steps. In particular, if a state of the form $(0, i_1, \ldots, i_k)^t$ is $\mathbf{y}_k$ with $\mathbf{y}_0 = m\mathbf{1}$ for some m then

$$y_{0k} = m - \sum_{j=1}^{k} \sigma_{k-j} + \sum_{j=1}^{k} v_{k-j,j-1} = 0$$

$$y_{kk} = m - \sum_{j=1}^{k} \sigma_{k-j} + \sum_{j=1}^{k} v_{k,j-1} \tag{45}$$

Hence

$$m - \sum_{j=1}^{k} \sigma_{k-j} \leq 0$$

and

$$\sum_{j=1}^{k} v_{k-j,j-1} = \sum_{i=1}^{\ell} \sum_{j=1}^{k} \mu_{k-j}^{i} x_{j-1}^{i} \leq k$$

Therefore

$$\sum_{i=1}^{\ell} \left(\sum_{j=1}^{k} x_{j-1}^{i} \right) \leq k$$

since, for $i = 1, 2, \ldots, \ell$, $\alpha_0^i = \mu_0^i \geq 1$ and $\mu_j^i \geq \mu_0^i$ for $j = 0, 1, \ldots, k$. Therefore, from eq. (45),

$$\begin{aligned} y_{kk} &\leq \sum_{j=1}^{k} v_{k,j-1} \\ &\leq \sum_{i=1}^{\ell} \mu_k^i \sum_{j=1}^{k} x_{j-1}^{i} \\ &\leq \left(\sum_{i=1}^{\ell} \mu_k^i \right) k \end{aligned} \tag{46}$$

Hence for every state

$$(0, i_1, \ldots, i_k)^t \in \mathcal{A}, i_k \leq k \sum_{i=1}^{\ell} \mu_k^i,$$

hence such states are finite in number from eq. (38). In a similar way we can show for any integer j the states $(i_0, i_1, \ldots, i_k)^t \in \mathcal{A}$ such that $i_0 \leq j$ is a finite set.

The transition probabilities for S can be derived from eqs. (36) and (37). Let $P_{\mathbf{i}}^n = Pr\{y_{0n} = i_0, y_{1n} = i_1, \ldots, y_{kn} = i_k\}$. Then

$$P_{\mathbf{i}}^{n+1} = \sum_{\mathbf{j} \in \mathcal{A}} Pr\{\mathbf{y}_{n+1} = \mathbf{i} | \mathbf{y}_n = \mathbf{j}\} P_{\mathbf{j}}^n \tag{47}$$

$$P_{\mathbf{i}}^{n+1} = \sum_{\substack{j_0=0 \\ \mathbf{j}\epsilon\mathcal{A}}} Pr\{v_{0n} = i_0 - j_1, \ldots, v_{k-1,n} = i_{k-1} - j_k, v_{kn}$$

$$= i_k - j_k\} P_{\mathbf{j}}^n + \sum_{\substack{j_0>0 \\ \mathbf{j}\epsilon\mathcal{A}}} Pr\{v_{0n} = i_0 - j_1 + 1, \ldots, v_{k-1,n}$$

$$= i_{k-1} - j_k + 1, v_{kn} = i_k - j_k + 1\} P_{\mathbf{j}}^n \quad (48)$$

If the equilibrium probabilities

$$P_{\mathbf{i}} = \lim_{n\uparrow\infty} P_{\mathbf{i}}^n$$

exist, then $P_{\mathbf{i}}^n = P_{\mathbf{i}}$ for every $\mathbf{i} \epsilon \mathcal{A}$ implies $P_{\mathbf{i}}^{n+1} = P_{\mathbf{i}}$, $\mathbf{i} \epsilon \mathcal{A}$. Furthermore $P_{\mathbf{i}}$ is the unique nonnegative solution of

$$\sum_{\mathbf{i}\epsilon\mathcal{A}} P_{\mathbf{i}} = 1$$

$$P_{\mathbf{i}} = \sum_{j_0=0,\mathbf{j}\epsilon\mathcal{A}} p_{i_0-j_1,i_1-j_2,\ldots,i_k-j_k} P_{\mathbf{j}}$$

$$+ \sum_{j_0>0,\mathbf{j}\epsilon\mathcal{A}} p_{i_0-j_1+1,i_1-j_2+1,\ldots,i_k-j_k+1} P_{\mathbf{j}} \quad (49)$$

Here $p_{i_0,i_1,\ldots,i_k} = Pr\{v_{0n} = i_0, \ldots, v_{kn} = i_k\}$.

We show next that S is positive recurrent when $Ez_n < 1$.

Theorem 3: S is positive recurrent if

$$E \sum_{i=1}^{\ell} \mu_k^i x_n^i < 1$$

Proof: Define a new process c_n as follows:

$$c_{n+1} = (c_n - 1)^+ + \sum_{i=1}^{\ell} \mu_k^i x_n^i$$

$$\equiv (c_n - 1)^+ + v_{kn} \quad (50)$$

We know that the Markov chain corresponding to c_n is positive recurrent if $Ev_{kn} < 1$ from Theorem 1. In particular if $Ev_{kn} < 1$ then

$$\lim_{n\uparrow\infty} Pr\{c_n = 0\} > 0$$

The processes c_n, b_n as defined by eqs. (50), (34) are related by x_n^i, $i = 1, \ldots, \ell$. Let $b_0 = 0$ and let r be such that $b_{n-i} > 0$ for $i = 0, 1, \ldots, r-1$ and $b_{n-r} = 0$. Then eq. (34) implies that

$$b_{n+1} = \sum_{m=0}^{r} \sum_{i=1}^{\ell} \sum_{j=0}^{k} \alpha_j^i x_{n-j-m}^i - r \quad (51)$$

We also know from eq. (2), assuming $c_0 = 0$, that

$$c_{n+1} \geq \sum_{m=0}^{r+k} \sum_{i=1}^{\ell} \mu_k^i x_{n-m}^i - (r+k)$$

From the definition of μ_k^i it can be easily verified that

$$c_{n+1} \geq b_{n+1} - k$$

Hence

$$Pr\{b_{n+1} \leq k\} \geq Pr\{c_{n+1} = 0\}$$

When $Ev_{kn} < 1$ we know from Theorem 1 that

$$\lim_{n \uparrow \infty} Pr\{c_{n+1} = 0\} > 0$$

hence

$$\liminf_{n \uparrow \infty} Pr\{b_{n+1} \leq k\} > 0$$

Let the set of states $(i_0, i_1, \ldots, i_k)$ in $\mathcal{A}$ with $i_0 \leq k$ be denoted by $\mathcal{A}_k$. Let P_{ij}^n be the probability of S being in state **i** at time n starting from **j** at time 0. Then, we have shown that

$$\liminf_{n \uparrow \infty} \sum_{i \in \mathcal{A}} P_{i0}^n > 0$$

Since cardinality of $\mathcal{A}_k$ is finite

$$\liminf_{n \uparrow \infty} P_{i0}^n > 0$$

for some $\mathbf{i} \in \mathcal{A}_k$. We also know that **0** is accessible from **i**. So $P_{0i}^r > 0$ for some r. Therefore

$$\liminf_{n \uparrow \infty} P_{00}^{n+r} \geq \liminf_{n \uparrow \infty} P_{0i}^r P_{i0}^n > 0$$

Hence **0** is positive recurrent. Therefore, since S is irreducible and aperiodic, S is positive recurrent.

V. GENERATING FUNCTIONS FOR JOINT DISTRIBUTIONS

We will now derive expressions for joint distributions of $(y_0, y_1, \ldots, y_k)$ assuming $Ez_n < 1$, so S is positive recurrent. Let

$$E\left(\prod_{r=0}^{k} s_r^{y_{rn}}\right) = \phi_n(s_0, s_1, \ldots, s_k)$$

and

$$E\left(\prod_{r=0}^{k} s_r^{v_{rn}}\right) = \phi_v(s_0, s_1, \ldots, s_k), \quad |s_i| \leq 1$$

From eqs. (36) and (37) we have, using independence of $\mathbf{v}_n$ and $\mathbf{y}_n$,

$$\phi_{n+1}(s_0, s_1, \ldots, s_k) = E\left(\left(\prod_{r=0}^{k} s_r\right)^{(y_{0n}-1)^+ - y_{0n}} \prod_{r=0}^{k-1} s_r^{y_{r+1,n}} s_k^{y_{kn}}\right) \times \phi_v(s_0, s_1, \ldots, s_k) \quad (52)$$

Proceeding as in eq. (11) we can show

$$\phi_{n+1}(s_0, s_1, \ldots, s_k) = \left[\phi_n(1, s_0, s_1, \ldots, s_{k-2}, s_{k-1}s_k) \prod_{i=0}^{k} s_i^{-1} + \left(1 - \prod_{i=0}^{k} s_i^{-1}\right) \phi_n(0, s_0, s_1, \ldots, s_{k-1}s_k)\right] \times \phi_v(s_0, s_1, \ldots, s_k) \quad (53)$$

When ϕ_n is the generating function of the equilibrium distribution i.e., when

$$\phi_n(s_0, s_1, \ldots, s_k) = \phi(s_0, s_1, \ldots, s_k) = E \prod_{i=0}^{k} s_i^{y_i} \quad (54)$$

then $\phi_{n+1} = \phi$. Therefore ϕ satisfies

$$\phi(s_0, s_1, \ldots, s_k) = \left[\phi(1, s_0, s_1, \ldots, s_{k-2}, s_{k-1}s_k) \prod_{i=0}^{k} s_i^{-1} + \left(1 - \prod_{i=0}^{k} s_i^{-1}\right) \phi(0, s_0, s_1, \ldots, s_{k-2}, s_{k-1}s_k)\right] \times \phi_v(s_0, s_1, \ldots, s_k) \quad (55)$$

We note that $\phi(0, t_1, \ldots, t_k)$ is a polynomial of finite degree since the set of states $(0, i_1, \ldots, i_k)$ is finite. Knowledge of $\phi(0, t_1, \ldots, t_k)$ determines $\phi(s_0, s_1, \ldots, s_k)$ as follows. If we set $s_0 = s_1 = \cdots = s_{k-1} = 1$ then (55) becomes

$$\phi(1, 1, \ldots, 1, s_k) = [s_k^{-1}\phi(1, 1, \ldots, 1, s_k) + (1 - s_k^{-1})\phi(0, 1, 1, \ldots, 1, s_k)]\phi_v(1, 1, \ldots, 1, s_k)$$

This determines $\phi(1, 1, \ldots, 1, s_k)$ in terms of $\phi(0, 1, 1, \ldots, 1, s_k)$:

$$\phi(1, 1, \ldots, 1, s_k) = \frac{(1 - s_k^{-1})\phi(0,1,1, \ldots, 1, s_k)\phi_v(1, 1, \ldots, 1, s_k)}{1 - s_k^{-1}\phi_v(1, 1, \ldots, 1, s_k)} \quad (56)$$

For $r = 0, 1, \ldots, k$ set

$$\phi^r(s_r, \ldots, s_k) = \phi(1,1, \ldots, 1, s_r, s_{r+1}, \ldots, s_k) \quad (57)$$

Then eq. (56) determines ϕ^k in terms of $\phi(0, 1, \ldots, 1, s_k)$. Using eq. (55) yields:

$$\phi^r(s_r, \ldots, s_k) = \left[\prod_{i=r}^{k} s_i^{-1} \, \phi^{r+1}(s_r, \ldots, s_{k-2}, s_{k-1} s_k) + \left(1 - \prod_{i=r}^{k} s_i^{-1}\right) \phi(0, 1, \ldots, s_r, \ldots, s_{k-1} s_k) \right] \times \phi_v(1, 1, \ldots, 1, s_r, \ldots, s_k) \quad (58)$$

So starting with ϕ^k, k applications of (58) yields $\phi^0(s_0, \ldots, s_k) = \phi(s_0, \ldots, s_k)$ in terms of $\phi(0, s_1, \ldots, s_k)$. Equations (15) have a counterpart here. These can be derived in mechanical fashion using formal power series expressions for ϕ and ϕ_v. we will not go into the details here. The derivation is analogous to that given in Sec. VII for the marginals.

In the case $\ell = 1$, an alternate generating function was considered in Ref. 8. The corresponding generating function is obtained by setting

$$u_j = \prod_{i=j}^{k} s_i, \; j = 0, \ldots, k$$

and defining

$$\phi(s_0, s_1, \ldots, s_k) = \Phi(u_0, u_1, \ldots, u_k) = \lim_{n \to \infty} E\left(u_0^{\gamma_{0n}} \prod_{r=1}^{k} u_r^{\gamma_{rn} - \gamma_{r-1,n}}\right)$$

Then corresponding to eq. (55),

$$\Phi(u_0, u_1, \ldots, u_k) = [u_0^{-1} \Phi(u_0, u_0, u_1, \ldots, u_{k-1}) + (1 - u_0^{-1}) \Phi(0, u_0, u_1, \ldots, u_{k-1})] \Phi_v(u_0, u_1, \ldots, u_k) \quad (55')$$

where

$$\Phi_v(u_0, u_1, \ldots, u_k) = E\left(\prod_{r=0}^{k} u_r^{w_{rn}}\right)$$

with

$$w_{rn} = \sum_{i=1}^{\ell} \alpha_r^i \, x_n^i$$

It follows from eq. (55) that, for j = 0, 1, . . . , $k - 2$, $(k \geq 2)$,

$$\Phi(s, \ldots, s, u_1, \ldots, u_{k-j}) = [s^{-1} \Phi(s, \ldots, s, u_1, \ldots, u_{k-j-1}) + (1 - s^{-1}) \Phi(0, s, \ldots, s, u_1, \ldots, u_{k-j-1})] \times \Phi_v(s, \ldots, s, u_1, \ldots, u_{k-j}) \quad (58')$$

and

$$\Phi(s, \ldots, s, u_1) = [s^{-1}\Phi(s, \ldots, s) + (1 - s^{-1})\Phi(0, s, \ldots, s)]\Phi_v(s, \ldots, s, u_1) \quad (58'')$$

These equations are equivalent to eq. (58). If we set $u_1 = s$ in (58″) then we may solve for $\Phi(s, \ldots, s) = \phi(1, \ldots, 1, s)$, as in (56).

VI. FINDING $\phi(0, s_1, \ldots, s_k)$

We will show here that a finite system of linear equations can be obtained to solve for the coefficients of the polynomial of finite degree that represents $\phi(0, s_1, \ldots, s_k)$. Let

$$\psi_j(s_1, \ldots, s_k) = \mu \prod_{i=1}^{k} s_i^{j_i}$$

where $\mu = Pr\{y_0 = 0\}$ and let $\theta_j(s_0, s_1, \ldots, s_k)$ be related to ψ_j as $\phi(s_0, s_1, \ldots, s_k)$ is to $\phi(0, s_1, \ldots, s_k)$ in eq. (58). That is, if θ_j^r is defined as in (57), $\theta_j^r(s_r, \ldots, s_k) = \theta_j(1, 1, \ldots, 1, s_r, s_{r+1}, \ldots, s_k)$, then θ_j^r satisfies the set of equations equivalent to (58): for $r = 0, 1, \ldots, k-1$

$$\theta_j^r(s_r, \ldots, s_k) = \left[\prod_{i=r}^{k} s_i^{-1}\,\theta_j^{r+1}(s_r, \ldots, s_{k-2}, s_{k-1}s_k) + \left(1 - \prod_{i=r}^{k} s_i^{-1}\right)\psi_j(1, 1, \ldots, 1, s_r, \ldots, s_{k-2}, s_{k-1}s_k)\right] \times \phi_v(1, 1, \ldots, 1, s_r, \ldots, s_k) \quad (59)$$

and (56) corresponds to

$$\theta_j^k(s_k) = \frac{(1 - s_k^{-1})\psi_j(1, 1, \ldots, 1, s_k)\phi_v(1, 1, \ldots, 1, s_k)}{1 - s_k^{-1}\phi_v(1, 1, \ldots, 1, s_k)} \quad (60)$$

From the definition of ϕ_v, $\phi_v(1, 1, \ldots, 1, s_k) = Es_k^{y_{kn}}$ [see above (52)]. Hence from (50) and applying (13), (14) we have

$$\theta_j^k(s_k) = \frac{1}{\mu}\psi_j(1, 1, \ldots, s_k)\phi_c(s_k) \quad (61)$$

where $c = \lim_{n\uparrow\infty} c_n$ and $Es_k^c = \phi_c(s_k)$. Hence whenever $\mu > 0$,

$$\frac{1}{\mu}\phi_c(\cdot)$$

is a generating function. Now, it can be easily verified that corresponding to each j the unique solution $\theta_j^0(s_0, \ldots, s_k) = \theta_j(s_0, s_1, \ldots, s_k)$ satisfies

an equation similar to (55):

$$\theta_{\mathbf{j}}(s_0, s_1, \ldots, s_k) = \left[\theta_{\mathbf{j}}(1, s_0, s_1, \ldots, s_{k-2}, s_{k-1}s_k) \prod_{i=0}^{k} s_i^{-1} + \left(1 - \prod_{i=0}^{k} s_i^{-1}\right) \psi_{\mathbf{j}}(s_0, s_1, \ldots, s_{k-2}, s_{k-1}s_k) \right] \times \phi_v(s_0, s_1, \ldots, s_k) \quad (62)$$

The family of such solutions $\theta_{\mathbf{j}}$ are linearly independent. If the generating function of $P_{\mathbf{i}}$ (the equilibrium distribution of S), $\phi(s_0, s_1, \ldots, s_k)$, is such that

$$\phi(0, s_1, \ldots, s_k) = \mu \sum_{\mathbf{j}}{}' c_{\mathbf{j}} \prod_{i=1}^{k} s_i^{j_i} \quad (63)$$

where the sum on the right is over all indices $\mathbf{j} = (0, j_1, \ldots, j_k)$ which are in $\mathcal{A}$, then $\phi(s_0, s_1, \ldots, s_k)$ has the unique representation

$$\phi(s_0, s_1, \ldots, s_k) = \sum_{\mathbf{j}}{}' c_{\mathbf{j}} \theta_{\mathbf{j}}(s_0, s_1, \ldots, s_k) \quad (64)$$

Notice that corresponding to each $\mathbf{j}$ there exists a sequence $P_{\mathbf{i}}(\mathbf{j})$, not necessarily nonnegative, such that

$$\sum P_{\mathbf{i}}(\mathbf{j}) s_0^{i_0} s_1^{i_1} \ldots s_k^{i_k} = \theta_{\mathbf{j}}(s_0, s_1, \ldots, s_k) \quad (65)$$

Hence $P_{\mathbf{i}}$ itself has the representation

$$P_{\mathbf{i}} = \sum_{\mathbf{j}}{}' c_{\mathbf{j}} P_{\mathbf{i}}(\mathbf{j}) \quad (66)$$

Furthermore $\theta_{\mathbf{j}}^k(s_k)$ from (61) corresponds to a nonnegative summable sequence. From (59), starting with $r = k - 1$ and going backwards to $r = 0$, we can show that for each $\mathbf{j}$, $P_{\mathbf{i}}(\mathbf{j})$ is the convolution of absolutely summable sequences and hence

$$\sum |P_{\mathbf{i}}(\mathbf{j})| < \infty \quad (67)$$

From eqs. (61), (65), and (66), when $\Sigma_{\mathbf{j}}{}' c_{\mathbf{j}} = 1$,

$$\sum_{\mathbf{i} \in \mathcal{A}} P_{\mathbf{i}} = \sum_{\mathbf{j}}{}' c_{\mathbf{j}} \theta_{\mathbf{j}}^k(1) = \sum_{\mathbf{j}}{}' c_{\mathbf{j}} = 1 \quad (68)$$

We will now show that there is a finite number of linear equations derived by substitution of (66) into (49) which uniquely determine $\{c_{\mathbf{j}}\}$ and hence $P_{\mathbf{i}}$. Let us denote the elements of the transition probability matrix of S, $Pr\{\mathbf{y}_{n+1} = \mathbf{i} | \mathbf{y}_n = \mathbf{j}\}$, by $T_{\mathbf{ij}}$. Then from (40)

$$P_i = \sum_{j \in \mathcal{A}} T_{ij} P_j \tag{69}$$

Substitution of (66) yields

$$\sum_{\mathbf{m}}{}' c_{\mathbf{m}} P_i(\mathbf{m}) = \sum_{j \in \mathcal{A}} T_{ij} \sum_{\mathbf{m}}{}' c_{\mathbf{m}} P_j(\mathbf{m}) \tag{70}$$

which is a set of linear equations for $c_{\mathbf{m}}$. Hence any solution $\{d_{\mathbf{m}}\}$ of (70) has the property that $Q_i = \Sigma_{\mathbf{m}}{}' d_{\mathbf{m}} P_i(\mathbf{m})$ satisfies (69), hence

$$Q_i = \sum_{j \in \mathcal{A}} T_{ij} Q_j \tag{71}$$

Now let T_{ij}^n be the n-step transition matrix of S. Then using (71) we have

$$Q_i = \sum_{j \in \mathcal{A}} T_{ij}^n Q_j \tag{72}$$

Since S is positive recurrent

$$\lim_{n \uparrow \infty} T_{ij}^n = P_i \tag{73}$$

Furthermore since $\Sigma_i |P_i(j)| < \infty$ for each j, $\Sigma_i |Q_i| < \infty$. Hence taking limits of both sides of (72) and interchanging limit and sum on the right hand side of (72) we have

$$\begin{aligned} Q_i &= \lim_{n \uparrow \infty} \sum_j T_{ij}^n Q_j \\ &= P_i \sum_j Q_j \end{aligned} \tag{74}$$

However, if $\Sigma_j{}' d_j = 1$ then, from (68), $\Sigma_i Q_i = 1$. Therefore

$$Q_i = P_i \tag{75}$$

Since $P_i(\mathbf{j})$ are linearly independent, $c_j = d_j$ for each $\mathbf{j}$, and $\{c_j\}$ are the unique solution of (70).

Remark 1: A similar set of equations can be obtained by substituting (64) into (55) and equating the coefficients of like powers on both sides of (55).

Remark 2: Note that $\phi(0, s_1, \ldots, s_k) = \mu$ in the case when, for each i, j, $\alpha_j^i > 0$. Hence (58) may be used repeatedly to obtain an expression for $\phi(s_0, s_1, \ldots, s_k)$. Herbert[6] considered this model when $\ell = 1$.

In the alternate formulation $\Phi(0, u_1, \ldots, u_k)$ is a multinomial. Moreover, from (34) and (35), since $\alpha_0^i > 0$, $i = 1, \ldots, \ell$, $y_{0n} = 0 \Rightarrow x_{n-1}^i = 0$, $i = 1, \ldots, \ell$, which implies $y_{kn} = y_{k-1,n}$. Hence $\Phi(0, u_1, \ldots, u_k)$ is

independent of u_k. From (58′) and (58″), $\Phi(s, u_1, \ldots, u_k)$ may be expressed in terms of $\Phi(0, s, \ldots, s, u_1, \ldots, u_{k-j-1}), j = 0, \ldots, k-2$, and $\Phi(0, s, \ldots, s)$. If we let $s \to 0$ in this expression, and equate $\Phi(0, u_1, \ldots, u_k)$ with the finite part, we obtain a system of homogeneous linear equations for the coefficients in the multinomial. In general, we also obtain a (consistent) set of homogeneous linear equations from finiteness conditions.

VII. GENERATING FUNCTIONS FOR MARGINALS AND FINITE SOLVABILITY

The joint distributions of $(y_0, y_1, \ldots, y_k)$ have $(k+1)$ arguments. We will see that we can reduce the problem to "$k+1$ one-dimensional problems" when we are only interested in the marginal distributions of $y_0, y_1, \ldots, y_k$. Let us denote the generating functions of y_i by $\Phi_i(s)$ and those of v_{rn} by $\phi_{rv}(s)$. Then

$$\phi_i(s) = \phi(1, 1, \ldots, \overset{i}{s}, 1, \ldots, 1)$$

$$= \Phi(\overset{i+1}{s}, \ldots, s, \overset{k-i}{1}, \ldots, 1), i = 0, \ldots, k \tag{76}$$

From (55) we then obtain for $r = 0, \ldots, k-1$

$$\phi_r(s) = \left[s^{-1}\phi_{r+1}(s) + (1 - s^{-1})\phi(0, 1, \ldots, \overset{r+1}{s}, \ldots, 1) \right] \phi_{rv}(s)$$

and

$$\phi_k(s) = [s^{-1}\phi_k(s) + (1 - s^{-1})\phi(0, 1, \ldots, 1, s)]\phi_{kv}(s) \tag{77}$$

Note that

$$\phi(0, 1, \ldots, \overset{r}{s}, \ldots, 1)$$

$$= \Phi(0, \overset{r}{s}, \ldots, s, \overset{k-r}{1}, \ldots, 1), r = 1, \ldots, k$$

Therefore once the c_j have been determined from the method presented above, Eq. (77) gives the marginal distributions. Once again we can translate (77) into linear equations for the distributions themselves as in (15). The marginals are finitely solvable in the sense that a finite number of components of the marginal distributions can be solved for from a finite number of linear equations.

For each $r = 1, 2, \ldots, k$ let γ_{rj} be the coefficient of s^j in the polynomial

$$\phi(0, 1, \ldots, \overset{r}{s}, \ldots, 1)$$

denoted by $c_r(s)$. Equating coefficients of like powers of s_r^j on both sides of (63) after setting $s_i = 1$ for $i \neq r$ yields

$$\gamma_{rj} = \mu \sum_{j_r=j}{}' c_j \tag{78}$$

Therefore since the c_j's can be determined as the solutions to a finite system of linear equations, so can the γ_{rj}'s.

Let

$$\phi_r(s) = \sum_{j=0}^{\infty} \pi_{rj} s^j, \quad \Pi_{rj} = \sum_{i=0}^{j} \pi_{ri}$$

and $F_r(s) = \Sigma_{j=0}^{\infty} \Pi_{rj} s^j$ for $|s| < 1$ and $r = 0, 1, \ldots, k$. Then $F_r(s) = \phi_r(s)/(1-s)$, and eqs. (77) become

$$F_r(s) = s^{-1}[F_{r+1}(s) - c_{r+1}(s)]\phi_{rv}(s), \; r = 0, 1, \ldots, k-1 \tag{79}$$

$$F_k(s) = \frac{\phi_{kv}(s) c_k(s)}{\phi_{kv}(s) - s} \tag{80}$$

From (79) we can show that, for each $r = 0, 1, \ldots, k-1$, $\{\Pi_{rj}\}_{j=0}^{N+r}$ are determined from $\{\Pi_{r+1,j}\}_{j=0}^{N+r+1}$ by a finite set of linear equations, for any N. Let the sequence $\{\delta_{rj}\}$ correspond to $s^{-1}[F_{r+1}(s) - c_{r+1}(s)]$. From the definition of $F_{r+1}(s)$ and $c_{r+1}(s)$ it follows that $\Pi_{r+1,0} = \gamma_{r+1,0}$. Therefore $\delta_{rj} = 0$ for $j < 0$ and

$$\begin{aligned} \delta_{rj} &= \Pi_{r+1,j+1} - \gamma_{r+1,j+1} \text{ for } 1 \leq j+1 \leq \text{degree of } c_{r+1}(s) \\ &= \Pi_{r+1,j+1} \text{ for } j+1 > \text{degree of } c_{r+1}(s) \end{aligned} \tag{81}$$

From (79), the sequence $\{\Pi_{rj}\}_{j=0}^{N+r}$ is the convolution of $\{\delta_{rj}\}$ with $\{p_{rj}\}$—the sequence of probabilities corresponding to the characteristic function $\phi_{rv}(s)$. Therefore, since the sequence $\{p_{rj}\}$ is known a priori, we can find Π_{rj} as:

$$\Pi_{rj} = \sum_{i=0}^{j} \delta_{r,j-i} p_{ri}, \; j = 0, 1, \ldots, N+r \tag{82}$$

Hence, we observe that $\{\Pi_{0j}\}_{j=0}^{N}$ can be determined as solutions to a finite system of linear equations using $\{\Pi_{kj}\}_{j=0}^{N+k}$.

In order to find $\{\Pi_{kj}\}_{j=0}^{N+k}$ we proceed as in (15). Equating the coefficients of like powers of s^j in

$$\sum_{j=0}^{\infty} \Pi'_{kj} s^j = \frac{\phi_{kv}(s)}{\phi_{kv}(s) - s} \tag{83}$$

yields:

$$\begin{aligned} \Pi'_{k0} &= 1 \\ \Pi'_{kj} &= \left(p_{kj} + \Pi'_{k,j-1} - \sum_{i=1}^{j} p_{ki} \Pi'_{k,j-i}\right) \Big/ p_{k0} \\ & \qquad j = 1, 2, \ldots\ldots, N+k \end{aligned} \tag{84}$$

Therefore $\{\Pi'_{kj}\}_{j=0}^{N+k}$ can be determined uniquely as solutions of (84). From (80) and (83), $\{\Pi_{kj}\}_{j=0}^{N+k}$ is the convolution of $\{\Pi'_{kj}\}$ with $\{\gamma_{kj}\}$,

$$\Pi_{kj} = \sum_{i=0}^{j} \Pi'_{k,j-i}\gamma_{ki}, \; j = 0, 1 \ldots . N + k \tag{85}$$

Therefore we have shown that each of the Π_{rj}, and hence the marginal distributions $\pi_{rj}, r = 0, 1, \ldots, k, j = 0, 1, \ldots, N + r$, can be found, for any finite N, as solutions to a finite system of linear equations.

VIII. A LIMITING CASE

For each m let d_{jm} be a nondecreasing sequence of nonnegative integers such that

$$(i) \quad d_{j0} = 0, d_{j1} = j; j = 0, 1, \ldots, k$$

$$(ii) \quad \lim_{m \uparrow \infty} d_{jm} - d_{j-1,m} = \infty, j = 1, \ldots, k \tag{86}$$

We define a sequence of processes $\{z_n^m\}$ which will be time-scaled versions of z_n. Let

$$z_n^m = \sum_{i=1}^{\ell} \sum_{j=0}^{k} \alpha_j^i x_{n-d_{jm}}^i$$

We observe that z_n^1 is the same as z_n, and z_n^0 is the "fastest" version of z_n, in the sense that all the packets triggered by x_n^i are bunched together and arrive at the same time. As m increases the different delayed contributions of x_n^i are spread farther and farther apart in time. The limiting case can then be interpreted as the "slowest"; see Ref. 7.

Let $\{\eta_n\}$, $n = 0, 1, 2, \ldots$, be a sequence of independent identically distributed random variables such that for each n the distribution of η_n is the same as that of z_n. We will show that η_n then corresponds to the slowest case: the finite dimensional distributions of the processes $\{z_n^m\}$, $m = 0, 1, \ldots$ converge to the corresponding distributions of $\{\eta_n\}$ as $m \uparrow \infty$. Indeed, let $n_1 < n_2 < \cdots < n_s$ be nonnegative integers. Then

$$Pr\{z_{n_1}^m = i_1, z_{n_2}^m = i_2, \ldots, z_{n_s}^m = i_s\} = \prod_{j=1}^{s} Pr\{z_{n_j}^m = i_j\} \tag{87}$$

for large enough m, in particular for every m such that $d_{jm} - d_{j-1,m} > n_s - n_1, j = 1, \ldots, k$. However from the definition of η_n, $Pr\{z_{n_j}^m = i_j\} = Pr\{\eta_{n_j} = i_j\}$. Therefore from the independence of η_n and (87)

$$\begin{aligned} &Pr\{z_{n_1}^m = i_1, z_{n_2}^m = i_2, \ldots, z_{n_s}^m = i_s\} \\ &= P\{\eta_{n_1} = i_1, \eta_{n_2} = i_2, \ldots, \eta_{n_s} = i_s\} \end{aligned} \tag{88}$$

for large enough m.

We now define a sequence of processes b_n^m, b_n^∞ corresponding to z_n^m, η_n respectively. Formally let

$$b_{n+1}^m = (b_n^m - 1)^+ + z_n^m$$
$$b_{n+1}^\infty = (b_n^\infty - 1)^+ + \eta_n \tag{89}$$

Since $Ez_n^m = E\eta_n$, if $Ez_n < 1$ then for each m, $\lim_{n\uparrow\infty} b_n^m = b^m$ is a well-defined random variable, and so is $b^\infty = \lim_{n\uparrow\infty} b_n^\infty$. We can then show that

$$\lim_{m\uparrow\infty} b^m = b^\infty \tag{90}$$

from Theorem 22 in Ref. 9, since z_n^m, η_n are nonnegative. Hence the distribution of b^∞ approximates the distribution of b^m for sufficiently large m. Therefore for each j

$$\lim_{m\uparrow\infty} Pr\{b^m \le j\} = Pr\{b^\infty \le j\} \tag{91}$$

Therefore b^∞ is the steady-state queue size corresponding to the "slowest" version of z_n. Let

$$\phi_x(s_1, \ldots, s_\ell) = E\left(\prod_{i=1}^{\ell} s_i^{x_n^i}\right)$$

Then it is easy to verify that $Es^{z_n^0}$ *and* Es^{η_n} are given by

$$\phi_x^0 = \phi_x(s^{\mu_k^1}, s^{\mu_k^2}, \ldots, s^{\mu_k^\ell}) \text{ and } \phi_x^\infty = \prod_{i=0}^{k} \phi_x\,(s^{\alpha_i^1}, s^{\alpha_i^2}, \ldots, s^{\alpha_i^\ell})$$

respectively. If $\phi^0 = Es^{b^0}$ and $\phi^\infty = Es^{b^\infty}$ then

$$\phi^0 = \frac{(1 - s^{-1})\phi_x^0(s)\mu}{1 - s^{-1}\phi_x^0(s)}$$
$$\phi^\infty = \frac{(1 - s^{-1})\phi_x^\infty(s)\mu}{1 - s^{-1}\phi_x^\infty(s)} \tag{92}$$

In the special case when $\ell = 1$, and (omitting the superscript) $\alpha_j = 0$ or 1 for each j, we have an interesting special relationship between ϕ^0 and ϕ^∞. Let $f_n^0 = Pr\{b^0 \le n\}$, $f_n^\infty = Pr\{b^\infty \le n\}$ and $F^0 = \Sigma f_n^0 s^n$, $F^\infty = \Sigma f_n^\infty s^n$. Then F^0 and F^∞ are $\phi^0/1 - s$ and $\phi^\infty/1 - s$ respectively for $|s| < 1$. We will show that

$$f_n^\infty = f_{n\mu_k}^0 \tag{93}$$

equivalently

$$Pr\{b^\infty \le n\} = Pr\{b^0 \le n\mu_k\} \tag{94}$$

Let ω be a primitive μ_kth root of unity. Then for $|s| < 1$

$$\frac{1}{\mu_k}\sum_{i=0}^{\mu_k-1} F^0(\omega^i s) = \frac{\phi_x(s^{\mu_k})}{\mu_k}\sum_{i=0}^{\mu_k-1}\frac{\mu}{\phi_x(s^{\mu_k})-\omega^i s}$$

$$= \frac{\mu[\phi_x(s^{\mu_k})]^{\mu_k}}{[\phi_x(s^{\mu_k})]^{\mu_k} - s^{\mu_k}} = F^{\infty}(s^{\mu_k}) \tag{95}$$

Therefore

$$\sum_{n=0}^{\infty} f_n^{\infty} s^{n\mu_k} = \frac{1}{\mu_k}\sum_{i=0}^{\mu_k-1}\sum_{m=0}^{\infty} f_m^0 (\omega^i s)^m$$

$$= \sum_{n=0}^{\infty} f_{n\mu_k}^0 s^{n\mu_k} \tag{96}$$

Since f_n^0 and f_n^{∞} are both increasing and bounded by 1, (96) shows that (93) holds.

ACKNOWLEDGMENTS

The authors are indebted to J. McKenna for a careful reading of the manuscript, and for many helpful suggestions for improving the presentation. They are also grateful to A. G. Fraser for bringing the data communications problem to their attention.

REFERENCES

1. A. Ghosal, "Some Aspects of Queueing and Storage Systems," Lecture Notes in Operations Research and Mathematical Systems, *23*, Springer-Verlag, 1970.
2. F. Spitzer, "A Combinatorial Lemma and its Application to Probability Theory," Trans. Amer. Math. Soc., *82* (1950), 449–461.
3. S. Karlin, "A First Course in Stochastic Processes," Academic Press, 1966.
4. M. S. Ali Khan, "Finite Dams with Inputs Forming a Markov Chain," J. Appl. Prob., *7* (1970), 291–303.
5. H. G. Herbert, "A Note on First Emptiness in a Discrete Storage System with Markovian Inputs," SIAM J. Appl. Math., *28* (1975), 657–661.
6. H. G. Herbert, "An Infinite Discrete Dam with Dependent Inputs," J. Appl. Prob., *9* (1972), 404–413.
7. A. G. Fraser, B. Gopinath, and J. A. Morrison, "Buffering of Slow Terminals," to be published.
8. B. Gopinath and J. A. Morrison, "A Discrete Queueing Problem Arising in Packet Switching," Analyse et Contrôle de Systèmes, 201–210 (1976), Séminaires IRIA, Rocquencourt.
9. A. A. Borovkov, "Stochastic Processes in Queueing Theory," Springer-Verlag, 1976.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Spectrum Estimation Techniques for Characterization and Development of WT4 Waveguide—I

By DAVID J. THOMSON

(Manuscript received April 7, 1977)

Techniques for reliably estimating the power spectral density function for both small and large samples of a stationary stochastic process are described. These techniques have been particularly successful in cases where the range of the spectrum is large. The methods are resistant to a moderate amount of contaminated or erroneous data and are well suited for use with auxiliary tests for stationarity and normality. Part I is concerned with background and theoretical considerations while examples from the development and analysis of the WT4 waveguide medium will be discussed in Part II, next issue.

I. INTRODUCTION

The problem of estimating the spectrum of a stationary time series has appeared frequently in the scientific literature and myriad approaches have been suggested. Nonetheless it became apparent during the course of the development of the WT4 waveguide system that these methods were inadequate for many of the data sets of interest. The techniques presented here were therefore developed.

It is commonly stated that the method selected to estimate a spectrum depends on the ultimate use of the estimate, and unfortunately to some extent this is true. The method described below is felt to represent an advance in that the basic technique works well in a variety of cases which previously would have required individual treatment. The loss calculations reported in Anderson *et al.*[1] are indicative of its accuracy.

The procedure which has evolved for estimating spectra can best be described as robust adaptive prewhitening. Such methods have three distinct stages: formation of a *pilot* spectrum estimate, using this estimate to design a *prewhitening* filter, and finally giving the result as the *ratio* of the spectrum of the filtered data to the power transfer function of the filter. This method is potentially both efficient and robust. The

efficiency of a statistical estimation procedure is the fraction of the information, in the sense of Fisher,[2] conveyed by the estimate about the parameter being estimated to the total information on this parameter inherent in the data. An estimation procedure is *robust* if it remains efficient over a wide range of conditions and is relatively immune to a small fraction of outlying or erroneous data.

For the sequential method described here to be efficient, the pilot estimate must be designed to have a large dynamic range at the expense of frequency resolution. The second spectrum estimate, which works on the filtered data, uses the opposite choice and so is chosen on the basis of frequency resolution. This can be done without incurring a large penalty in loss of effective dynamic range as this information, acquired by the pilot estimate, has been transferred to the filter specification. In one meaning of the term this procedure is robust in that it can normally handle situations where either estimate alone would fail. By using a nonlinear filter for the prewhitening operation the procedure may also be made robust in the sense that it is resistant to moderate amounts of erroneous or contaminated data.

In this method the pilot estimate of spectra is a combination of several direct estimates of spectra computed on subsets of the data using a window defined by a prolate spheroidal wave function. Using this estimate as a basis an autoregressive model of the process is formed. This model is then used to generate a *nonlinear* prediction error filter. The output of this filter consists of prediction residuals from a modified data sequence and is quite immune to occasional isolated errors in the data.

Section II gives an overview of the complete estimation procedure so that the descriptions of the individual stages of the process are taken in the proper perspective. Section III is a review of properties of direct estimates of spectra which are used for both the pilot and final estimation procedures. Sections IV to VIII describe the several stages of the estimation procedure in detail. While these sections contain some examples they are primarily concerned with theory and background. Part II will consist primarily of examples and comparisons with standard techniques.

It should be emphasized that the same approach is used for *both* short and long data sets and that the only difference between these cases is one of detail and not philosophy. We define a *short* time series as one which cannot be subdivided into subsets having almost uncorrelated spectrum estimates.

Since this technique is basically nonparametric, it is frequently asked whether a parameterized estimate of spectrum might not give better results. It has been shown by Arato[3] that *only* for the autoregressive case can a process be described by a fixed number of sufficient statistics and

that in general the number of sufficient statistics increases with the sample size. As a result, efficient parametric estimates are not likely to be even conceptually simpler than the nonparametric estimates used here.

It is also asked why maximum-likelihood techniques are not used directly, and, while asymptotic results on parametric maximum likelihood estimates of spectra are available in Whittle,[4] constructive procedures for obtaining nonparametric maximum-likelihood estimates of the spectrum of a stationary Gaussian time series are unknown. It is, however, possible to check if a *given* estimate is maximum likelihood or not. This test, described in Thomson,[5] depends on the Karhunen-Loève expansion of a random process (see Loève[6]). In this test the data is expanded in terms of the sample eigenfunctions of the spectrum estimate, and, if this estimate is maximum likelihood, the expansion coefficients, $\hat{a}_n$, will satisfy the conditions $\hat{a}_n^2 = \hat{\lambda}_n$ in which $\hat{\lambda}_n$ are the corresponding sample eigenvalues. By the Szegö theorem (see Grenander and Szegö[7]) this comparison is asymptotically equivalent to comparisons on the spectrum at a frequency spacing of $1/T$. This agrees with the conventional Rayleigh resolution and heuristically a spectrum estimate with this resolution and low bias is likely to be efficient. This argument provides the motivation for the present technique. Simple data windows with frequency resolution close to $1/T$ do not provide enough bias protection. Moreover this is not just a result of not having chosen the right "simple" data window but the result of fundamental characteristics of the Fourier transform (see Landau and Pollak[8]). Data windows like the 4π prolate spheroidal wave function which provide the protection from bias have frequency resolution on the order of $4/T$ and so are inefficient from this viewpoint. It must be emphasized that the sequential approach used here potentially has both limitations since it cannot resolve details spaced by $1/T$ in frequency when their levels are more than 4 or 5 decades apart. On the other hand if the spectrum is not quite so pathological and varies "slowly" over 10 to 15 decades then the method can provide frequency resolutions approaching $1/T$ with relatively low bias.

II. SUMMARY OF THE ESTIMATION PROCEDURE

2.1 Data preparation

At the beginning the data is plotted, and serious outliers, missing values, etc., edited by use of either interpolation or successive prediction and interpolation. These predictors and interpolators are the optimum linear forms based on previous spectra estimates of a similar process or on assumed valid data from the current sample. It is also frequently necessary to remove the mean value function of the "cleaned" data. This

is always done in the analysis of individual tubes to eliminate the curvature resulting from gravitational sag.

2.2 Pilot spectrum estimate

For the remainder of this paper we assume that the available data is a sequence of samples $\{x_t\}$, $t = 0, 1 \cdots L$, and that the sampling interval has been normalized to 1. Consequently the normalized Nyquist frequency is $\frac{1}{2}$. Both because the notation is more compact and also as a reminder that the basic processes are continuous† most operations will be denoted by integrals. In the actual computations most of these integrals are replaced by simple sums but on occasion spline approximations to the integrals (see Aronson[10]) are used. The frequency variable will be denoted by f with $\omega = 2\pi f$.

The initial spectrum estimate is normally computed using a variation of Welch's[11] method: the basic data set is divided into k overlapping *subsets* each of length T and offset from the previous one by a distance b. The data from each subset is tapered using a zero order *prolate spheroidal wave function*, with parameter $c = 4\pi$ and the Fourier transform of the result computed. The raw estimate of spectra on the jth subset, $\hat{S}_j(\omega)$, is then the squared magnitude of the transform so that its univariate distribution is proportional to a χ^2_2. The use of the prolate data window guarantees, under simple conditions, that the bias of the estimate within each subset is of purely *local* origin and that estimates separated by more than $2c/T$ in frequency are essentially uncorrelated. However, to account for the correlation induced by the tapering the total number of degrees of freedom must be reduced. These effects and the bivariate distribution of the estimates is discussed in Section 3.2.

Because the raw estimates, $\hat{S}_j(\omega)$, are very volatile it is often desirable to smooth the different subset estimates. These estimates, smoothed to have ν degrees of freedom, will be denoted by $\bar{S}_j(\omega)$. In the original Welch technique the pilot estimate of spectrum, $\bar{S}(\omega)$, is the arithmetic average of the subset estimates. When the data contains outliers it is advantageous to replace the simple average with a robust combination of the subset estimates as discussed in Section V. Both because it is based on subsets of the data and because of the heavy tapering implied by the use of the parameter $c = 4\pi$ (see Section III) the pilot spectrum estimate has poor frequency resolution compared to the final estimate of spectra. For reasons discussed below excessive resolution in the pilot estimate is frequently counterproductive and this technique produces a stable estimate with adequate bias protection in situations where the range of the spectrum is very large.

† The paper by Dzhaparidze and Yaglom[9] contains information on the complexities induced by sampling basically continuous records.

2.3 Tests for stationarity

Large data sets can be tested for stationarity using the method described in Thomson.[12] Briefly the approach compares the different subset estimates, $\overline{S}_j(\omega)$, using Bartlett's M statistic for heteroscedasticity of variance between subsets at constant frequency. Equally spaced samples of the test statistic, $M(\omega_j)$, are then pooled and tested for conformance to the distribution expected for homogeneous samples.

2.4 Construction of autoregressive models

Stationary time series have four generally accepted canonical representations; Cramér's orthogonal increment spectral representation, the Karhunen-Loève expansion, the moving average, and autoregressive models. Of these the autoregressive model is perhaps the most useful for making inferences on the structure of the process. For further information see the review paper by Kailath.[13]

Most autoregressive methods either begin with a sample autocorrelation function and solve the Yule-Walker equations directly (Makhoul[14]) or else resort to a variation of Wiener spectral factorization (Whittle[15]) applied to an estimate of spectra; neither approach is entirely satisfactory. For the estimation of waveguide spectra both methods have been used and in Section VI a method combining the better features of both is discussed. In cases where the range[‡] of the spectrum is relatively small, solving of the Yule-Walker equations using Durbin's modification of the Levinson algorithm (see Section VI) is satisfactory. In this case the autocorrelations used are obtained by Fourier-transforming the pilot estimate of spectra. When the range of the spectrum is larger the Wiener technique is more stable but results in a very long predictor. Backward application of the Levinson algorithm may then be used to generate a more compact representation. In both cases the order p of the autoregressive representation has usually been chosen on the basis of Parzen's[16] stopping rule and the innovations variance corresponding to the pilot spectrum $\overline{S}(\omega)$. Details of the procedure are given in Section VI.

The autoregressive representation has an intuitive explanation in waveguide applications in which the prediction can be thought of as analogous to a local "warped normal mode" representation and the innovations process the changes required in the field configuration to maintain the "local" character. The casual nature of the autoregressive representation corresponds to propagation in the forward direction so that the field configuration at a given point reflects distortions which have been passed but not those in the future.

‡ The *range* of a spectrum refers to the *logarithmic* range or the ratio $\max\{S\}/\min\{S\}$.

2.5 Prewhitening, robust filtering

The autoregressive model formulation gives the casual filter which, for fixed impulse duration, p, has minimum output power. The residual sequence or output of such a filter (known as a *prediction error filter*) is the difference between the observed and predicted values of a data sequence *using the previous p data points as a base for the prediction.* When the autoregressive model is correct the residual sequence will be serially uncorrelated and have a white spectrum. When the data contains outliers the effect of such filtering is to contaminate the p residuals following each erroneous point.

The *robust filter algorithm* is a nonlinear procedure based on an autoregressive model which is designed to reduce the effects of occasional outliers. The output of this filter or the *modified data sequence* is an estimate of the uncontaminated process. This sequence is formed by comparing successive input data points with the value predicted from the modified sequence. In regions where the prediction errors are "small" relative to the innovations variance, the modified sequence is essentially a copy of the input data. When the prediction errors are "large," the corresponding points of the modified data sequence are the predictions rather than the data and for intermediate prediction errors the behavior depends on a weight function. When the modified data sequence is used as a basis for the final estimate of spectra, the prediction error sequence is the difference between the predictions and the value of the modified data sequence. For uncontaminated data this corresponds to the output of the linear prediction error filter but when a large error is present the algorithm has two effects: first, the large output residual is replaced by a zero; second, because of the feedback nature of the method, propagation of the error into subsequent predictions is greatly reduced. As with all methods which alter or ignore extreme observations a compromise must be drawn between rejecting some valid data and accepting occasional errors and, in the robust filter algorithm, this compromise is reflected in the choice of weight function. In Section VII a weight function motivated by the normal extreme value distribution which has both intuitive appeal and desirable mathematical properties is described.

2.6 Final estimate of spectrum

The prediction residuals, or output of the prediction error filter, are the sequence which has minimum power for a filter whose impulse response has duration p. Consequently in the frequency domain the effect of such an operation is necessarily to reduce the highest parts of the spectrum first. As the complexity of the filter is increased the residual spectrum approaches a constant at which point further improvement is impossible. In practive finite order autoregressive filters seldom attain

this limit but rather have the effect of reducing the range of the spectrum, usually without following any fine structure which is present and, as a result, information describing the fine structure is left in the residual process. On occasions when the autoregressive fit is forced to follow too fine structure in the spectrum the spectrum of the residuals may be locally more complex than the spectrum of the original process.†

Since the range of the spectrum has been reduced the procedure used to estimate the spectrum of the residuals is designed to have high-frequency resolution at the expense of sidelobe suppression.

When the nonlinear version of the prediction error filter is used it is commonly observed that the pilot spectrum, estimated from the contaminated data, is considerably higher than the final estimated spectrum at frequencies where the spectrum is small. So that these differences are not obscured with bias the pilot final taper must be such that the corresponding spectral window decays significantly with frequency and consequently tapers such as the Taylor equiripple design (see Rife and Vincent[17]) are inadvisable. The window which has been used most for this purpose is Tukey's spliced cosine taper. For long data sets this window is satisfactory but with very short sets, for example individual waveguide tubes, the first sidelobe of this window is too high and a more complex window described by a series expansion in prolate spheroidal wave functions is used.

The final estimate of spectrum is based on an approximation introduced in Grenander and Rosenblatt,[18] which is that the predictor and prediction residuals are statistically independent. Under this assumption the final estimate of spectrum will be the spectrum of the residuals divided by the power transfer function of the prediction error filter.

2.7 Smoothing

One of the most commonly recommended operations in spectrum estimation is that of smoothing the raw estimates by means of local averaging over frequency. Contrary to these recommendations the final estimates of spectra are almost never smoothed. Moreover, in cases where "smoothed" estimates of spectra are used, the smoothing is frequently the result of nonlinear and adaptive procedures. Such smoothing is useful in plotting applications, and for improving the stability of pilot spectrum estimates from short data sets. Certain nonlinear smoothers are also very useful for finding low level lines in complex spectra.

The general philosophy of these methods has been to test the raw spectrum for local homogeneity: when the local spectrum appears to be

† For spectrum estimation problems a good measure of complexity is $\left|\frac{\partial^2 S(\omega)}{\partial \omega^2}\right| \Big/ S(\omega)$.

homogeneous it is smoothed, but in cases where the raw spectrum exhibits variations greater than normal, a typical response is to reduce the width of the smoother. A second approach which is used is to initially "smooth" the raw spectrum using a robust nonsymmetric location estimate and then to put the peaks back on the basis of "inverse influence."

III. DIRECT ESTIMATES OF SPECTRA

In both the pilot and final phases, the spectrum is estimated by the so-called *direct method* and while the parameters and application of the estimator are different in the two cases, the basic form is the same. Information on direct estimates is available from several sources, for example Blackman and Tukey,[19] Jones,[20–22] Tukey,[23] Koopmans,[24] Brillinger.[25] In this section properties of the direct estimate are reviewed and compared to the indirect estimate; the role of prolate spheroidal wave functions as a means of reducing the bias of the estimate is described and compared to standard data windows. The next subsection describes the variance of the estimates with emphasis on characteristics of prolate windows and smoothing when the estimates included in the smoother are correlated. The final subsection is concerned with Welch estimates and a technique for choosing the optimum subset spacing.

The *direct* estimate of spectrum is defined by

$$\hat{S}_D(\omega) = \left| \int_0^T e^{i\omega t} D(t) x(t) dt \right|^2 \tag{1}$$

In this definition the data, x, is defined on the domain $[0, T]$, ω is radian frequency, and D is a *data window* or *taper*. The data window is normalized according to the convention

$$\int_0^T D^2(t) dt = 1 \tag{2}$$

so that the resulting spectrum is interpretable in physical units.

Almost all of the published estimates of spectra are either direct estimates, smoothed direct estimates, or rational fits to direct estimates. When D is constant $\hat{S}_D$ is the periodogram. Smoothing the *extended* periodogram† with appropriate weights corresponds to the various *indirect* estimates. Similarly an autoregressive or "maximum entropy" estimate may be regarded as an all-pole rational fit to the extended periodogram and Pisarenko[26] estimates constitute a generalization of

† In the simple periodogram estimates are computed at a frequency spacing of $1/T$ and the corresponding autocorrelations are circularly defined. A frequency spacing $<1/2T$ is used in the extended periodogram and its Fourier transform yields the common autocorrelations.

this idea.[‡] The notable exceptions are the Whittaker periodogram and the Burg estimate (see Section 6.4).

The application of smoothers or curve-fitting procedures to the basic estimate conceals its true nature and the fact that the properties of these estimates are controlled primarily by the data window D. For example it is commonly stated that the fundamental uncertainty in spectrum estimation is between resolution and variance and more papers than it is convenient to list have worked on better "lag windows" to minimize this conflict. Unfortunately the emphasis on this secondary problem has masked the primary uncertainty between resolution and bias. The basic problem with indirect estimates and the lag window approach is that it represents an attempt to patch the periodogram. A more logical approach is to start with a better basic spectrum estimate.

Despite its simplicity the direct estimate is not well understood. In particular the differences between direct estimates using *data windows* and indirect estimates using *lag windows* are frequently confused.

The expected value of the direct estimate (1) may be written

$$\mathrm{E}\{\hat{S}_D(\omega)\} = \int_0^T \int_0^T e^{i\omega(t-u)} D(t) D(u) \mathrm{E}\{x(t)x(u)\}\, dt\, du \tag{3}$$

For second order or covariance stationary processes the autocovariance function is defined by

$$R(\tau) = \mathrm{E}\{x(t)x(t+\tau)\} \tag{4}$$

and may be represented in terms of the spectral density function by using the Wiener-Khintchine relation

$$R(\tau) = \frac{1}{2\pi} \int e^{i\omega\tau} S(\omega)\, d\omega \tag{5}$$

and denoting the Fourier transform of the data window D by $\tilde{D}$ one obtains

$$\mathrm{E}\{\hat{S}_D(\omega)\} = S(\omega) * |\tilde{D}(\omega)|^2 \tag{6}$$

where S is the true spectrum of the process and * indicates convolution. Since D is a *time-limited* function $\tilde{D}$ is an entire function of ω so that the direct estimate is biased for all spectra which are not white. The function $|\tilde{D}(\omega)|^2$ is known as the *spectral window* of the estimate.

An alternative description results from expressing eq. (3) in terms of the autocovariance function, R, of the process as

$$\mathrm{E}\{\hat{S}_D(\omega)\} = \int_{-T}^{T} e^{-i\omega\tau} L_D(\tau) R(\tau)\, d\tau \tag{7}$$

[‡] The Capon[27] estimate, while superficially similar, is intended for estimating the magnitude of periodic components in a background of a known covariance structure.

where the convolution $D*D$ has been identified as an "equivalent" *lag window*, $L_D(\tau)$. Because of this identification characteristics of the *indirect* estimate of spectra,

$$\hat{S}_L(\omega) = \int_{-T}^{T} e^{-i\omega\tau} L_D(\tau) \hat{R}_u(\tau)\, d\tau \tag{8}$$

using an unbiased estimate of autocovariance

$$\hat{R}_u(\tau) = \frac{1}{T - |\tau|} \int_0^{T-|\tau|} x(t)x(t + |\tau|)\, dt \tag{9}$$

are often used incorrectly to describe the *direct* estimate, $\hat{S}_D(\omega)$. Except for their first moments these two estimates have few properties in common: one very important difference is that *the direct estimate is positive* while the "equivalent" indirect form need not be. Also, because their common spectral window enters the estimate in fundamentally different ways, the variances of the two estimates are different.

3.1 Minimum bias estimates and prolate spheroidal wave functions

The most convenient description of bias induced by the data window is through the spectral window $|\tilde{D}|^2$ as expressed in eq. (6). The effect of this convolution is to change the apparent distribution of power in a complex manner and, since all windows cause some redistribution, a minimal requirement is that the indicated power be left "close" to its original location. Defining "close" to be within a tolerance Ω of ω we require that the *broadband bias*, i.e., bias from outside $(\omega - \Omega, \omega + \Omega)$, be small. Denoting this bias by $B_B(\omega)$ and the integral over frequency with the section $(\omega - \Omega, \omega + \Omega)$ excluded by $\oint$ we have

$$B_B(\omega) = \frac{1}{2\pi} \oint S(\omega - \zeta)|\tilde{D}(\zeta)|^2\, d\zeta \tag{10}$$

From the definition of a direct estimate and the convolution theorem the broadband bias in a particular sample is

$$\hat{B}_B(\omega) = \left| \oint \tilde{x}(\omega - \zeta)\tilde{D}(\zeta)\, d\zeta \right|^2 \tag{11}$$

where $\tilde{x}(\omega)$ is the spectral representation of x (see Doob,[28] chapter 10). By the Cauchy inequality this bias may be bounded so that

$$\hat{B}_B(\omega) \leq \frac{1}{2\pi} \oint |\tilde{x}(\omega - \zeta)|^2\, d\zeta \frac{1}{2\pi} \oint |\tilde{D}(\omega - \zeta)|^2\, d\zeta \tag{12}$$

The first factor of this inequality depends only on the process and, as the integrand is positive, is simply bounded by adding the integral from $\omega - \Omega$ to $\omega + \Omega$ and identifying the result using Parseval's theorem. The second factor in the inequality depends only on the data window, D, and

expresses the energy in $\tilde{D}$ outside $-\Omega, \Omega$. D is a time-limited function of unit energy and this inequality is minimized when D is a *prolate spheroidal wave function.* The fundamental role of these functions in relation to Fourier transforms and related problems have been described in a remarkable series of papers by Slepian and Pollak,[29] Landau and Pollak,[8,30] and Slepian.[31] When the bounds for both integrals are combined the result is that

$$\hat{B}_B(\omega) \leq \hat{\sigma}^2 T(1 - \lambda_{00}(c)) \tag{13}$$

where $\hat{\sigma}^2$ is the sample variance, $c = \Omega T/2$, and $\lambda_{00}(c)$ is the largest eigenvalue of the integral equation

$$\lambda_n \psi_n(t) = \int_{-1}^{1} \frac{\sin c(t-s)}{\pi(t-s)} \psi_n(s)\, ds \tag{14}$$

Tables of the eigenvalues of this equation have been published in Slepian and Sonnenblick[32] and asymptotic descriptions given by Slepian.[33] From the latter reference

$$1 - \lambda_{00}(c) \approx 4\sqrt{\pi c} e^{-2c} \tag{15}$$

As the width of the guard band, Ω, increases this bound decreases rapidly. For exploratory time series analysis and the formation of pilot spectrum estimates a very convenient value of c is 4π for which $1 - \lambda_{00} \approx 3 \times 10^{-10}$. In Thomson *et al.*[34] empirical studies show that direct estimates using this window are generally superior to several other spectral estimates in common use. Other examples are contained in Thomson.[5] Windows using approximations to prolate spheroidal wave functions have been described by Kaiser,[35] Eberhard,[36] and in fact the Parzen[37] window can be considered as a fourth-order successive approximation to the 4π prolate window.

Figure 1 shows the 4π prolate data window (and several other windows described below) and the low weighting given near the ends of the data are evident. The corresponding spectral windows are shown in Fig. 2, and here the reason for using the 4π prolate taper in situations where the spectrum varies over large ranges is most evident. The frequency scale of this plot has been normalized to units of $1/T$ so that by a frequency of $4/T$ the spectral window corresponding to the 4π taper has decayed by more than 10 decades. It should be noted that the curves for the other windows represent *envelopes* of the spectral windows. The actual spectral windows are similar to that shown for the compound prolate window and decay in an oscillatory manner.

When the range of the spectrum is known to be small it is clear that the use of this window is inefficient in that the frequency resolution is much less than it is for windows with higher sidelobes, and several al-

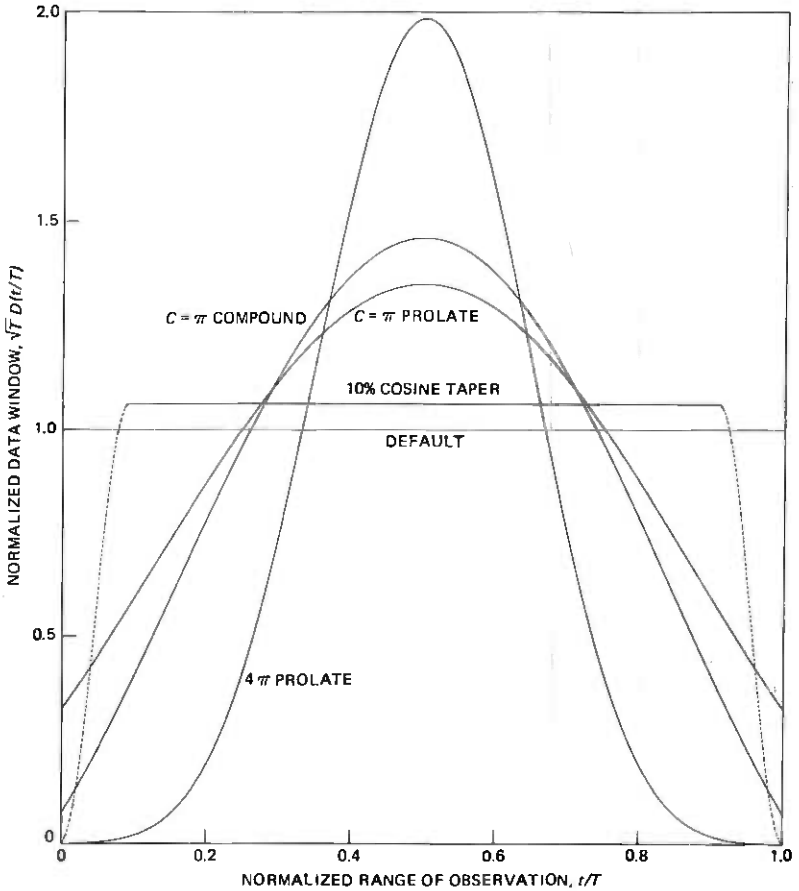


Fig. 1—Comparison of data windows.

ternatives are available: no tapering, ad-hoc tapers, and prolate tapers with lower values of c.

Very few spectra resulting from physical processes are so uninteresting that the "elimination" of tapering is ever advisable; in this case the taper actually used is $1/\sqrt{T}$ over $(0, T)$ and 0 elsewhere. This "default taper" has

$$T\left(\frac{\sin \omega T/2}{\omega T/2}\right)^2$$

as a spectral window so that, as shown in Fig. 2, the first sidelobe is only ~13 dB down from the central maxima.

Of the various ad-hoc techniques, Tukey's[23] spliced cosine taper is perhaps the most useful and it has been used for many of the final esti-

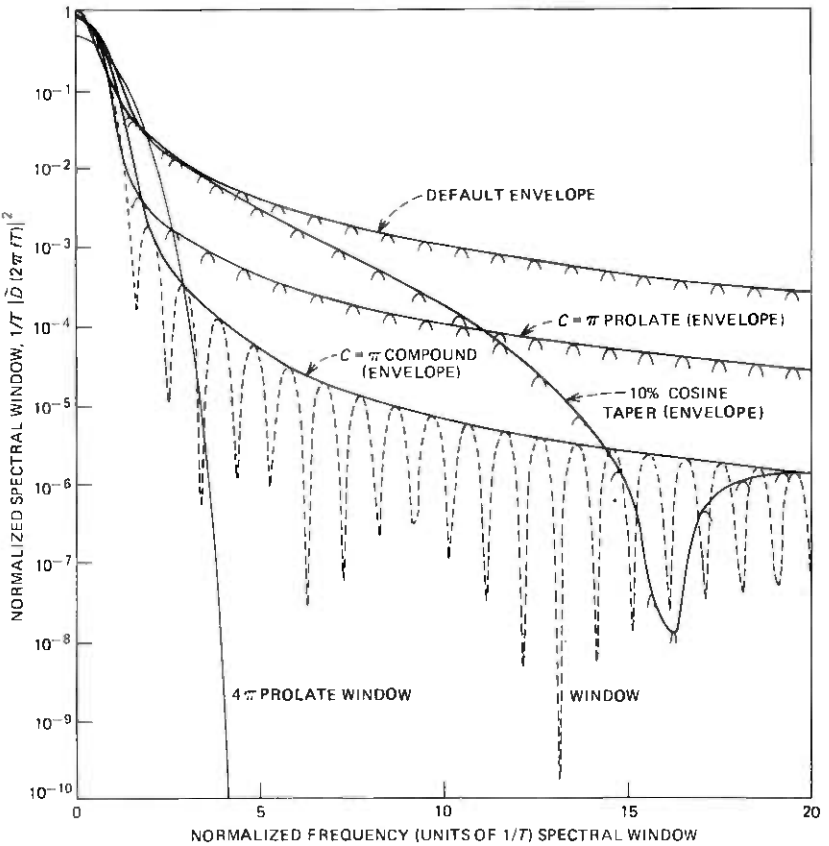


Fig. 2—Envelope of spectral windows.

mates of spectra used in the WT4 project. Since this taper, shown in Fig. 1, weights the data in a much more uniform manner the corresponding spectral window, Fig. 2, has a narrower center lobe than the 4π window and the sidelobes decay much faster than those of the default window.

Unfortunately the first few sidelobes of this window are too high for it to be usable in many applications where accurate estimates of the fine structure of a spectrum are required. This leads to considering the spheroidal wave functions again and for maximum concentration in a bandwidth $1/T$ the appropriate value of the parameter c is π. As before the maximum concentration is achieved by using the function of order 0 but for the present application a better compromise can be obtained by using a linear combination of the functions of order 0 and 2 with the coefficients determined by the additional constraint imposed by requiring that the first two sidelobes be minimized. This "compound

prolate" taper is also plotted in Fig. 1 and it is clear that the weight is less extreme than the 4π taper but distinctly different from the spliced cosine form. From the plot of the spectral windows it can be seen that the main lobe of the compound window is almost as narrow as that of the spliced cosine and also that the first sidelobes are down 27 dB instead of 13 dB. Since the widths of the main lobes are all very close to the same width, this gain in performance is essentially free and results from the superior characteristics of the prolate functions. It might also be mentioned that the usual objection to the use of the prolate spheroidal wave functions, namely that they are "impossible" to compute, is false and that by using Horner's rule together with the expansion given in Flammer,[38] Section 3.2, they may be computed very rapidly. Appendix A gives expansion formulae for the π and 4π prolate data windows.

In anticipation of Section VI it is also interesting to compare the bias of the estimates of autocorrelation obtained by transforming the various spectrum estimates. From eq. (7) it is apparent that such estimates of the autocorrelation function at lag τ will be biased by the factor $L_D(\tau)$. These lag windows are plotted in Fig. 3. From this figure it can be seen that the bias imposed on the low-order autocorrelations by the windowing techniques is much less than that resulting from the common positive definite estimate [obtained by replacing the factor $T - |\tau|$ in eq. (9) with T] corresponding to the simple extended periodogram. It should be noted that if this factor is divided out the resulting unbiased estimate is not positive definite and frequently results in negative "prediction variances." For fitting autoregressive models, the low-order autocorrelations are crucial and, as can be seen from the insert in Fig. 3, for $\tau/T < 0.01$ the bias obtained using the 4π prolate window is lower than that obtained from the extended periodogram on data sets *10 times as long*. The scale of such comparisons can be best judged by noting that the one-step autocorrelation in the field evaluation test curvature data is about 0.99983.

3.2 The distribution of direct spectrum estimates: liftering[†]

The preceding sections were adressed primarily to the problem of bias in direct spectrum estimates without particular attention being paid to their variances or distributions. Since reliable interpretation of spectrum estimates requires understanding of both their distributions and the correlations between estimates, the following sections treat these and the closely related problem of smoothing. Because of the correlations induced by the data window, the variance of smoothed direct estimates depends *both* on the smoothing weights *and* on the data window.

As mentioned in the introduction, the final estimate is rarely

[†] See Bogert, Healy, and Tukey[39] for definitions of these terms.

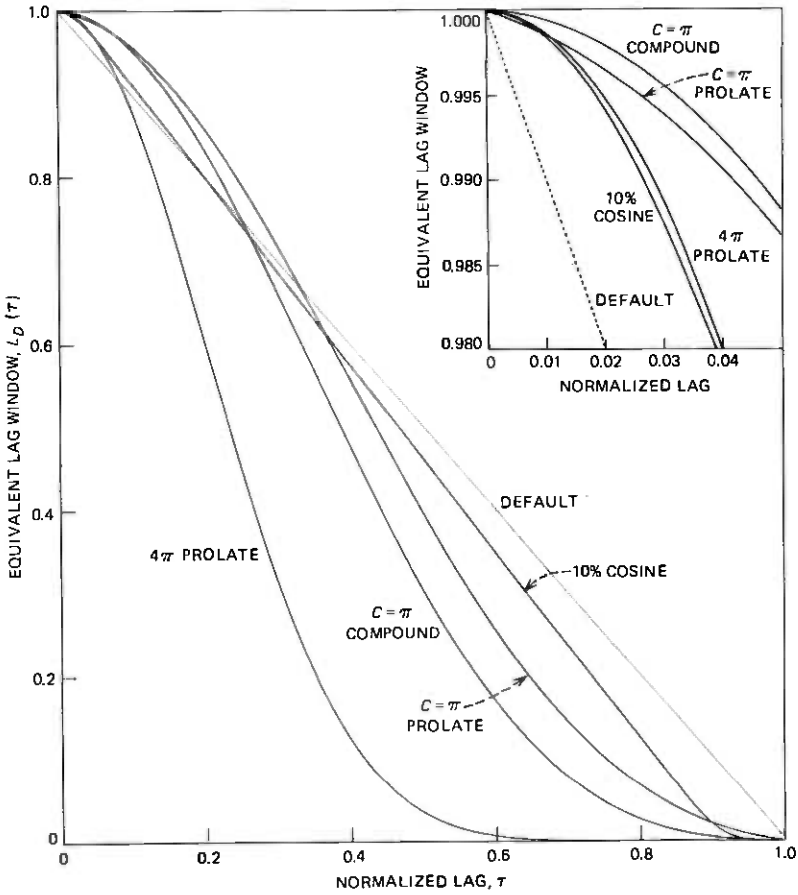


Fig. 3—Comparison of lag windows.

smoothed, but in the formation of the pilot estimate smoothing is a critical step. The primary reason for smoothing the pilot spectrum estimate is to obtain a more accurate autoregressive model. Because the direct estimate is inconsistent, that is, its first-order distribution and variance are independent of sample size, T, smoothing is imperative when spectral factorization is employed and experience has shown that serious errors are obtained when unsmoothed estimates are used with the other autoregressive modeling techniques. Also, when the robust filter algorithm is used, the prediction residuals are measured on the scale of the estimated innovations variance. The accuracy of this estimate, and hence the reliability of the procedure, depends *both* on the actual stability of the pilot estimate *and* on what we estimate that stability to be. The latter effect enters in the form of a bias correction factor, a

function of the "equivalent degrees of freedom" of the pilot estimate, on the innovations variance estimate. With "short" data sets stability can only be obtained through liftering but even with long data sets where the Welch technique is applicable, liftering is used to improve the sensitivity of the stationarity test.

3.2.1 The distribution of direct spectrum estimates

Excluding the neighborhoods of the origin and the Nyquist frequency, direct estimates of spectra are approximately distributed as a χ^2_2. For Gaussian data this result is exact and even in cases where the original data is reasonably nonnormal it is known (Fisher,[40] Bartlett[41]) to be a remarkably good approximation. Because the variance of such estimates is given by the square of their expected value, this fact emphasizes the need to start with a better estimate of spectra than the periodogram; estimates with low bias will have lower variance than estimates with high bias.

The bivariate probability density function of direct spectrum estimates can be obtained from those given by Miller *et al.*[42] for Rayleigh processes

$$p(s_1, s_2) = \frac{1}{1-\Lambda} e^{-(s_1+s_2)/(1-\Lambda)} I_0\left(\frac{2\sqrt{\Lambda s_1 s_2}}{1-\Lambda}\right) \tag{16}$$

where both s_1 and s_2 have been standardized to unit level, I_0 is the usual modified Bessel function, and Λ is the correlation between s_1 and s_2 given below by eq. (18).

The characteristics of this distribution are most easily seen by considering the conditional distribution $p(s_1|s_2)$. For this distribution a critical point is given by $s_2 = (1-\Lambda)/\Lambda$; at this point $\partial p(s_1|s_2)/\partial s_1|_{s_1=0} = 0$ which for lower values of s_2 resembles the univariate distribution and has its maximum at 0, while for larger values of s_2 the mode approaches s_2.

Figure 4 shows plots of the conditional distribution for $s_2 = 0.5$ and 1 and for values of Λ appropriate for the 4π prolate window at frequency spacings of $0.25/T$, $0.5/T$, $0.75/T$, $1/T$, and $2/T$.

3.2.2 Smoothing and frequency correlations of spectrum estimates

There is a considerable literature on smoothing spectrum estimates (see for example Blackman and Tukey,[19] Parzen,[37,43] Papoulis,[44]) and the variance and distribution of smoothed estimates (Jones,[20] Grenander *et al.*[45]) but much of this work is specialized to estimates based on the periodogram and cases where the different raw estimates included in the smoothing operation are uncorrelated. For the prolate data windows the latter assumption is unwarranted (as indeed it is even for the ex-

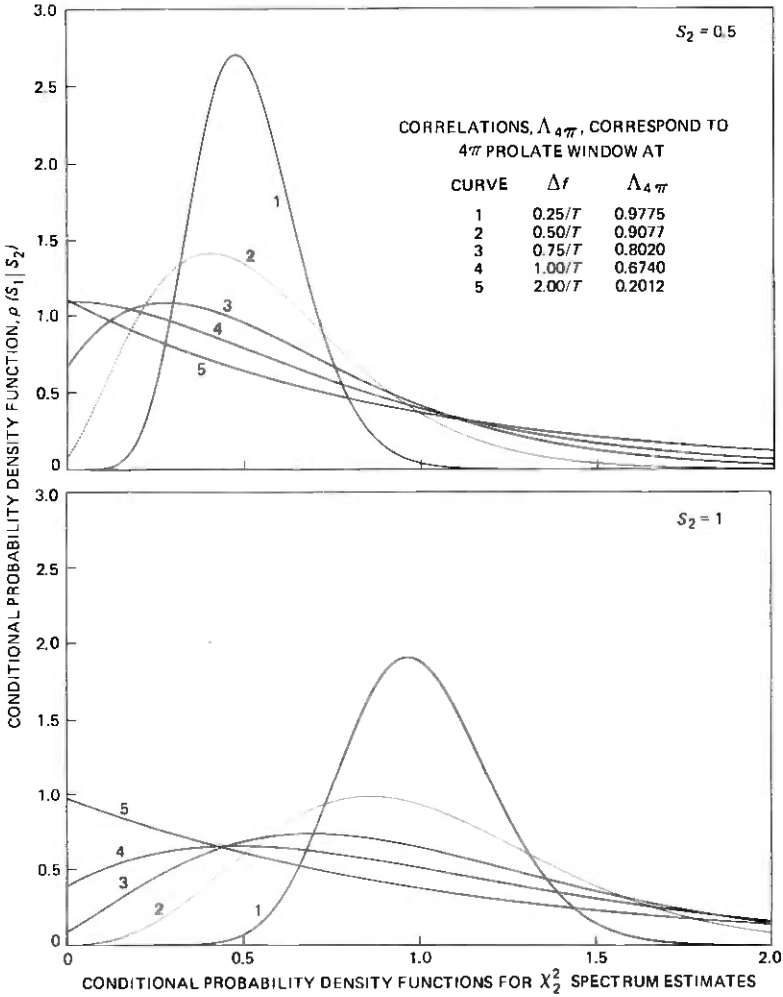


Fig. 4—Conditional probability density functions $p(S_1/S_2)$ for bivariate χ^2_2 distribution. Top set of curves, conditioning variable $S_2 = 0.5$; lower set, $S_2 = 1$.

tended periodogram) since, for the 4π window, the bias is only localized within a band of $\pm 4/T$. In the more general case where the raw spectrum estimates are correlated, smoothing over a fixed bandwidth is less effective and conventional smoothing techniques will be characterized by fewer "equivalent degrees of freedom" than given by the usual estimate. Most of the work on smoothing assumes that the true spectrum does not vary appreciably over the width of the smoother and under this approximation the influence of smoothers on direct estimates is fairly simple to evaluate.

To assess the effects of smoothing correlated spectrum estimates it

is necessary to examine their correlation properties. By expanding the fourth moment formula it can be shown that for *Gaussian* processes the covariance of the direct estimate at different frequencies is given by

$$\mathrm{Cov}\{\hat{S}_D(\omega+\zeta), \hat{S}_D(\omega-\zeta)\} = \left|\frac{1}{2\pi}\int S(\zeta-\xi)\tilde{D}^*(\xi+\omega)\tilde{D}(\xi-\omega)d\xi\right|^2 + \left|\frac{1}{2\pi}\int S(\omega-\xi)\tilde{D}^*(\xi+\zeta)\tilde{D}(\xi-\zeta)\,d\xi\right|^2 \quad (17)$$

In this equation the first term is large only in the neighborhood of the origin ($\omega = 0$) while the second term is a convolution which, for $\zeta = 0$, equals $\mathrm{E}\{\hat{S}_D(\omega)\}^2$. If, on the other hand, we set $\zeta = \Delta/2$, the second term gives the covariance of estimates with a frequency separation of Δ in the vicinity of ω. It is helpful to view the direct estimate, $\hat{S}_D(\omega)$, as a nonstationary time series with a *known* covariance structure and in regions where the spectrum is locally white as a *stationary* series. As with other stationary series the second-order properties of the direct estimate in such regions are described by an autocorrelation function, which for unit spectrum is given by

$$\Lambda(\Delta) = |\tilde{D}^*\tilde{D}^*|^2 \quad (18)$$

Figure 5 shows the autocorrelation functions of the different direct spectrum estimates as a function of frequency separation, and again the local properties of the prolate tapers are striking by comparison to the very poor properties of the other estimates. It should be noted, however, that for the 4π window at the usual frequency mesh spacing of $1/2T$ the correlation between estimates is 0.9077 so that, as shown in Fig. 4, the distribution of estimates at this spacing is quite different than it is for independent estimates.

It is frequently more convenient to work with the Fourier transform, Ξ_D, of this autocorrelation which we call the *antespectrum* of the estimator. Thus Ξ_D is defined by

$$\Xi_D(Q) = D^2(Q)*D^2(Q) \quad (19)$$

and is the *spectrum of the spectrum estimate*, $S_D(\omega)$. The antespectrum is a function of *quefrency*, Q, which is a lag or time-like variable and its Fourier transform is the autocorrelation function, Λ, *of the spectrum estimate* expressed as a function of ordinary frequency separation.

Defining a smoothed direct estimate $S_{D,W}(\omega)$ as

$$\hat{S}_{D,W}(\omega) = \frac{1}{2\pi}\int \hat{S}_D(\omega-\zeta)W(\zeta)\,d\zeta \quad (20)$$

in which the weight, W, is usually considered to be symmetric, positive, and integrating to 1. Since the spectral window of the direct estimate, $S_D(\omega)$, is $|\tilde{D}(\omega)|^2$ the spectral window of the smoothed estimate is clearly

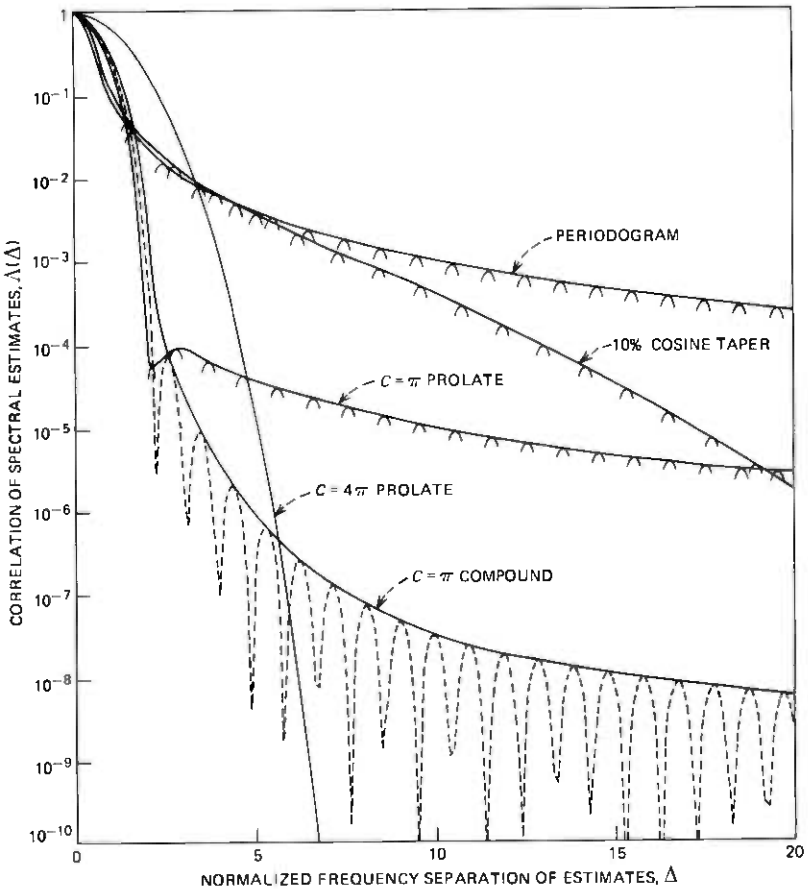


Fig. 5—Envelopes of autocorrelation functions of direct spectrum estimates.

the convolution $|\bar{D}|^{2} * W$ but the variance does not correspond to the usual interpretation of a spectral window in the literature on indirect estimates. Using the above definitions the influence of a smoothing operation, or *lifter*, may be described in the quefrency domain as a linear filter so that the antespectrum, $\Xi_{D,W}$, or the *spectrum of the smoothed spectrum estimate* is the product of the antespectrum, Ξ_D, and the power transfer function of the lifter. The variance of the smoothed spectrum estimate is the integral, over quefrency, of its antespectrum so that the estimate $\hat{S}_{D,W}$ will have an approximately χ^2 distribution with

$$\nu_{D,W} = 2\left[\int_{-T}^{T} |\bar{W}(Q)|^{2}\Xi_D(Q)dQ\right]^{-1} \tag{21}$$

equivalent degrees of freedom. For direct estimates the antespectrum

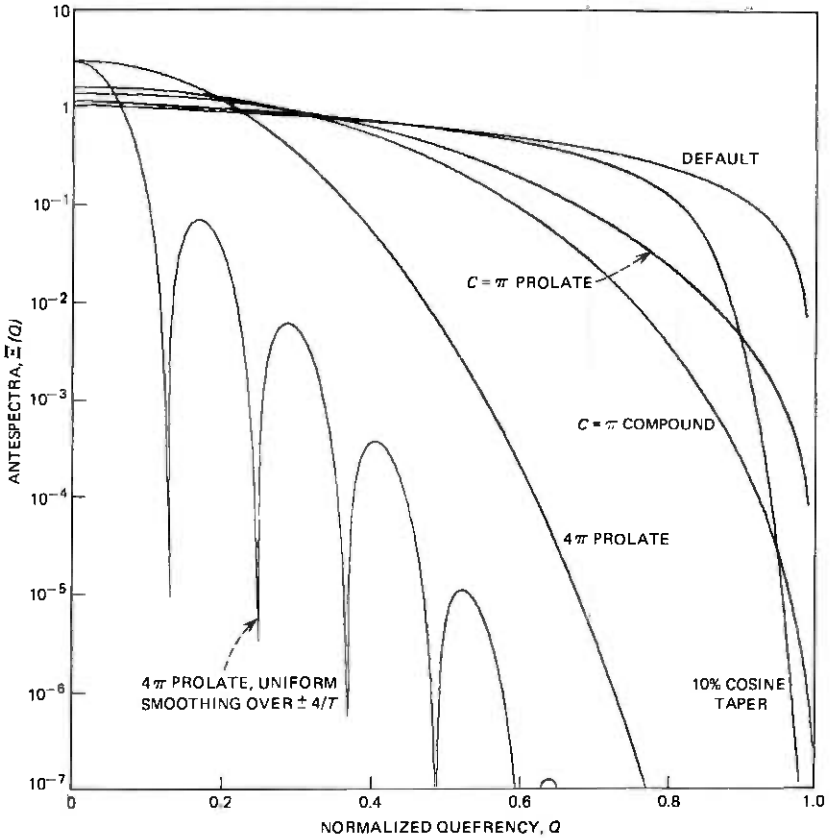


Fig. 6—Antespectra corresponding to various data windows (the antespectrum is the spectrum of the spectrum estimate.

is symmetric with a global maximum at zero quefrency. From eq. (21) it is clear that for liftering to be effective $|\tilde{W}|^2$ should be small when Ξ_D is large.

As indicated above the choice of weights is a complex subject which depends to a large extent on the intended application with perhaps the best linear smoothers obtained by modifying the technique described by Papoulis[44] to account for the data window. When this is done the Sturm-Liouville equation

$$\frac{\partial}{\partial t}(D^2(t)y') + \lambda y = 0 \tag{22}$$

is obtained corresponding to his eq. (22) and can be solved by standard techniques.

Figure 6 shows the antespectra corresponding to the various data windows. From these curves it is apparent that estimates using the 4π

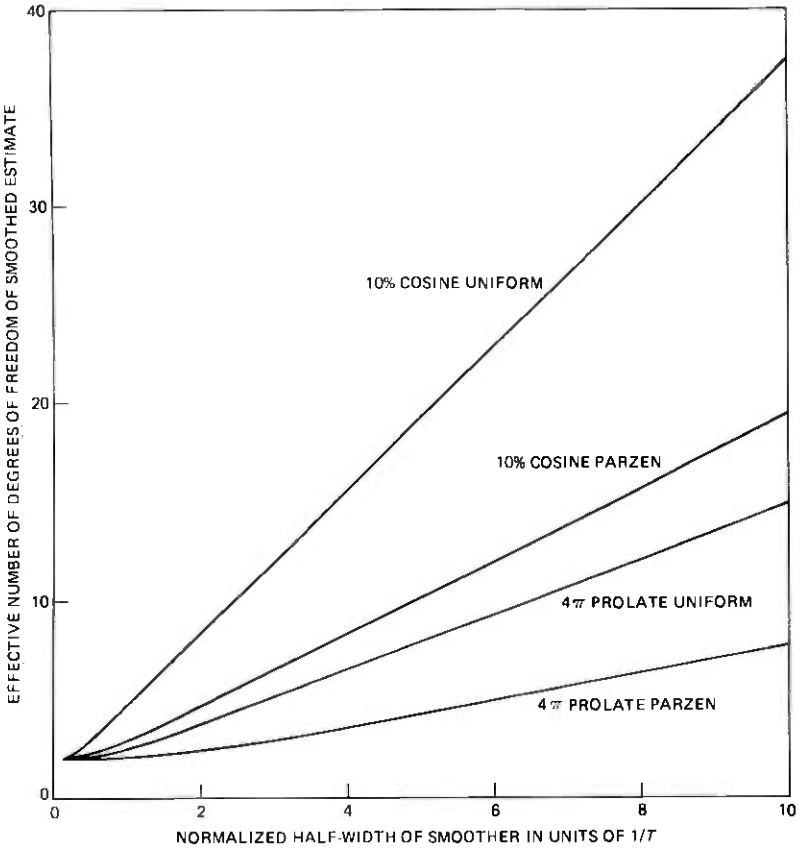


Fig. 7—Effectiveness of smoothing direct estimates.

taper have more of their variance at low frequencies than the other estimates. The bottom curve in this figure shows the antespectrum of a 4π estimate smoothed with uniform weights over a bandwidth of $\pm 4/T$ which, when integrated, gives only 6.7 equivalent degrees of freedom. Figure 7 is a plot of the equivalent degrees of freedom resulting when direct estimates are smoothed. Two smoothers are used: simple moving averages which are useful for calibration purposes and the modified Parzen weights (Cleveland and Parzen[46]). *From these curves it is obvious that the correlation induced in the raw spectrum estimate by the data window can result in significantly fewer degrees of freedom than expected on the basis of uncorrelated spectrum estimates.* This effect is particularly noticeable with the 4π taper where, when bias considerations are excluded, the asymptotic efficiency is only 36 percent when only a single direct estimate is computed. As will be seen in the next section, the use of overlapped subsets results in variance efficiency.

3.3 The Welch technique and optimum subset spacings

An alternative to smoothing across frequency is Welch's method[11] (see also Cooley *et al.*[47]) in which the data is divided into several overlapping subsets, direct estimates computed on each subset, and the results combined. The individual subset estimates have the usual statistical properties of direct estimates but when used jointly one must also account for the correlation between subsets. For the same reasons given in the previous section the effectiveness of this averaging must be accurately determined. Clearly spacing the subsets too close results in computational inefficiency while if they are spaced too far apart the procedure is statistically inefficient.

Consider two direct estimates of spectra $\hat{S}_1$ and $\hat{S}_2$ made on the domains $(0, T)$ and $(b, b + T)$ respectively. For Gaussian processes the covariance between these estimates is given by

$$\mathrm{Cov}\{\hat{S}_1(\omega), \hat{S}_2(\omega)|b\} = \left|\frac{1}{2\pi}\int S(\zeta)e^{i\zeta b}\tilde{D}(\omega-\zeta)\tilde{D}^*(\omega+\zeta)\,d\zeta\right|^2 + \left|\frac{1}{2\pi}\int S(\zeta)e^{i\zeta b}|\tilde{D}(\omega-\zeta)|^2\,d\zeta\right|^2 \quad (23)$$

The first term of this expression is large near the frequency origin but elsewhere the second term dominates. Since $\tilde{D}$ is an entire function it is clear that the covariance between the two estimates is governed primarily by the characteristics of the actual spectrum S, in the vicinity of ω. In particular spectra having very narrow resonances or discontinuous characteristics will result in the subsets being correlated for large values of the offset b. The effect of this correlation is that averaging the different subset estimates does not give the usual reduction of variance so that the autoregressive model is unstable when only a few subsets are available. When the correlation between subsets is low the distribution of the average of k subsets is nearly χ^2_{2k}.

When the spectrum is locally smooth estimates of this type depend, in addition to the data window, on the two parameters T and b. The length of the individual subsets depends primarily on the fine structure of the process and will be discussed in Section V. The relative spacing of subsets, however, depends largely on the choice of the data window and in general there is an optimum spacing. Under the usual approximation that the true spectrum is locally constant or linear and that we are interested in frequencies away from the origin, eq. (23) simplifies, and the correlation between subsets becomes the square of the equivalent lag window, $L_D(b)$.

As a measure of effectiveness of this procedure, assume that sufficient data is available to compute k subsets. Standardizing the local spectral level to 1, the variance of the averaged estimate is

$$V_k(b) = \frac{1}{k} + \frac{2}{k}\sum_{s=1}^{k-1}\left(1 - \frac{s}{k}\right) L_D^2(sb) \tag{24}$$

We now consider the effect of adding sufficient new data to compute $k + 1$ subsets and, by analogy with Fisher information, we measure the relative gain in information by

$$\Delta I_k(b) = \frac{1}{b}\left[\frac{1}{V_{k+1}(b)} - \frac{1}{V_k(b)}\right] \tag{25}$$

As k becomes large ΔI_k rapidly approaches the limit

$$\Delta I_\infty(b) = \frac{1}{b}\frac{1}{1 + 2\sum_{s=1}^{[T/b]} L_D^2(sb)} \tag{26}$$

This function is plotted in Fig. 8 for the different windows discussed earlier. When the subsets are spaced very closely relative to their length, no information is "missed" by falling between adjacent subsets, but on the other hand the subsets are highly correlated with each other so that the addition of a subset does not decrease the variance very much. For the 4π prolate window this situation remains true until the spacing between subsets becomes about 0.25 to 0.30 of their length, after which the information recovery becomes rapidly less efficient. Because the computational burden rapidly increases as the offset is decreased, a subset spacing of about 0.29 of the subset length is used. For the less concentrated windows this effect is less important. It should also be noted that the higher information recovery of the 4π prolate window evident here is consistent with the fact that it has a broader frequency response, so that, apart from bias considerations, the overall efficiencies of the techniques are similar. When bias considerations are included the efficiency of the prolate window is much higher.

IV. DATA PREPARATION

Assuming that aliasing and noise effects have been properly kept at a minimum there are usually two steps of data preparation necessary in time series work. The first is the elimination of *gross* errors and the second is the removal of *deterministic* mean value functions.

Gross errors are inevitable in very large data sets, see Hampel,[48] and experience has shown that the WT4 project is no exception to this rule.

Errors which are large and easily visible are best removed at an early stage in the processing. A simple strategy which works for both large errors and missing values is as follows:

(*i*) Data points in serious error are tagged, either on the basis of vi-

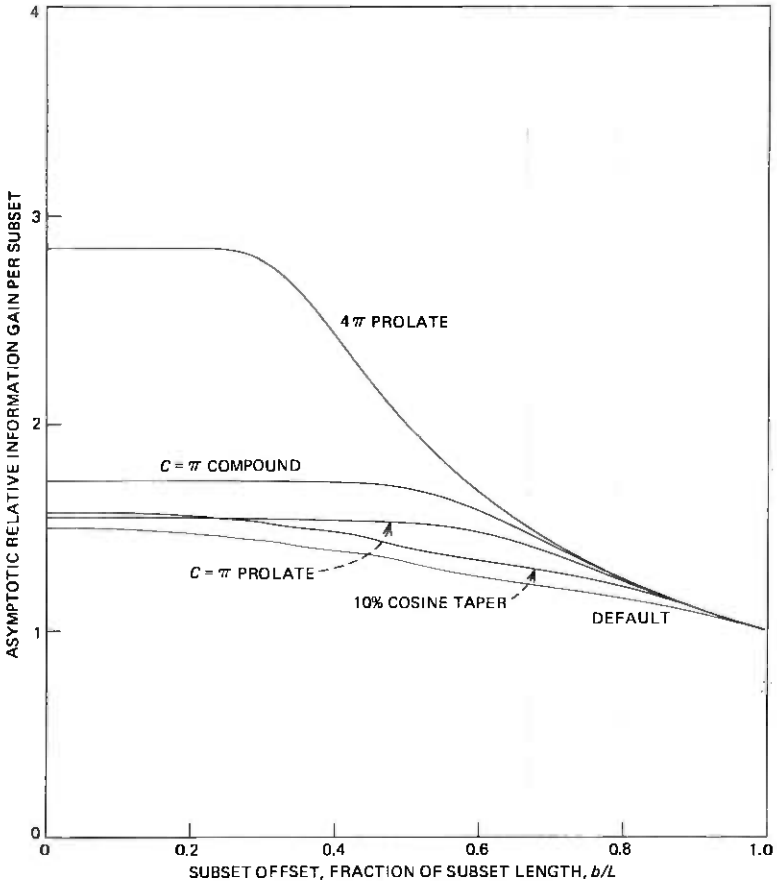


Fig. 8—Asymptotic relative information gain as a function of subset base offset.

sual examination or automatically on the basis of built in validity checks.†

(*ii*) Predictors and interpolators were generated either from untagged data or, if similar data sets were available, from the autocorrelations of these sets.

(*iii*) When the error is isolated it is replaced by interpolation. If the errors could not be considered isolated those adjacent to the longest stretch of good data were corrected first by prediction from the good section. After all the tagged points had been initially corrected by prediction, the corrections were recomputed by interpolation using the initial correction as a basis for the interpolations.

† Records of long range mouse data were coded to provide indications of hardware malfunction.

However, not all errors in time series data are obvious from a plot. In situations where the spectrum covers many decades an error may be insignificant on the scale of the *process* variance but catastrophic compared to the *innovations* variance. At the data preparation stage these errors are neither easily detectable nor troublesome and so their correction is deferred to the prewhitening part of the process where they are effectively eliminated by the robust filtering process.

The other stage of data preparation is the removal of deterministic mean value functions from the data. A simple example of a series with a nonconstant mean value function is given by the axis curvature of a waveguide following a planned route bend.

The usual approach in time series analysis with problems of this type is to remove the "trend" using orthogonal polynomial regression techniques. This approach has proved unsatisfactory primarily because such a high-degree polynomial is required to approximate the mean value function that the residuals bear little resemblance to the stochastic part of the process.

A method of removing trends in data which has proved generally effective is based on the use of polynomial B-splines. A B-spline of *order* k is a piecewise continuous polynomial of degree $k - 1$ defined by an array of *knots*, some of which may be multiple. The continuity properties of these functions are controlled by the knots; the spline is discontinuous at a knot of multiplicity k, has a discontinuous derivative at knots of multiplicity $k - 1$, and so on. At simple knots, or knots of multiplicity 1, the spline has $k - 2$ continuous derivatives. Details of the theory of B-splines are contained in a paper by Curry and Schoenberg,[49] a recent paper by de Boor[50] describes computational aspects, and Horowitz[51] discusses the characteristics of splines with equispaced simple knots in terms of their frequency domain characteristics.

Figure 9 shows a plot of the measured elevation of a waveguide line and an approximate mean value function generated through the use of B-splines. By choosing a spline with few knots, indicated on the figure, a simple fit to the gross topology of the run is obtained so that the "roughness" of the installation is readily apparent.

A second example of the use of polynomial spline mean value functions is shown in Fig. 10, which is a plot of the vertical output from a measurement of axis curvature on a waveguide tube supported on Airy point supports (see Fox *et al.*[52]). In this case most of the indicated curvature is a result of the tube sagging under its own weight and this effect is readily calculable and is shown by the dashed line. As a check that this removal is not distorting the spectrum of the actual distortions in the tube the ratio, an F statistic, of the average of 10 estimates of the spectrum of the detrended vertical curvature to the spectrum of the hori-

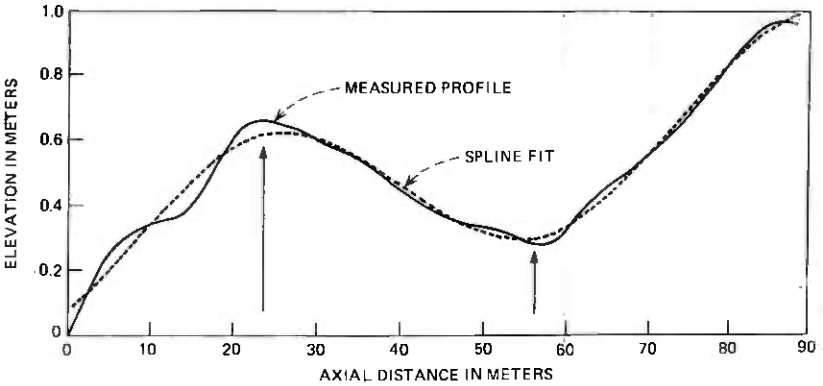


Fig. 9—Example of a B-spline fit to waveguide elevation knots. The spline is order 4 with knots of multiplicity 4 at −0.1 and 90 meters, simple knots at 23.5 and 56.5 meters.

zontal curvature was computed. No significant differences could be detected between the two sets of spectra.

V. PILOT SPECTRUM ESTIMATE

The actual process used to generate the pilot spectrum estimate is a combination of the two smoothing approaches described in Section III. The use of subsets allows the test for stationarity and, because this test is more sensitive when applied to smoothed data, a logical step is to smooth the subset estimates individually. However for the stationarity test to be effective it is necessary that the different subset estimates be essentially uncorrelated at any given frequency. This requirement results in the base offset between adjacent subsets being more than about 57 percent of the subset length, which is larger than is desirable for the most effective use of the data from an information recovery viewpoint. The obvious solution is to compute the subsets with the 29 percent offset mentioned above and use every other subset in the stationarity test.

A further advantage of the use of subsets is that a significant improvement in the accuracy of the pilot estimate can often be obtained by combining the different subset estimates in a robust manner instead of by the usual arithmetic average. Denoting the ordered subset estimates by $\vec{S}_j(\omega)$ with $\vec{S}_1(\omega) \leqslant \vec{S}_2(\omega) \leqslant \cdots \leqslant \vec{S}_{k'}(\omega)$, a robust estimate $\vec{S}(\omega)$ may be formed as

$$\vec{S}(\omega) = \sum_{j=1}^{k'} \Theta_j \vec{S}_j(\omega) \tag{27}$$

where the weights, $\{\Theta_j\}$, which depend on k', are chosen so that $\vec{S}$ is a minimum variance unbiased estimate of S. General techniques for forming such estimates are given in Lloyd[53] and the specific means and

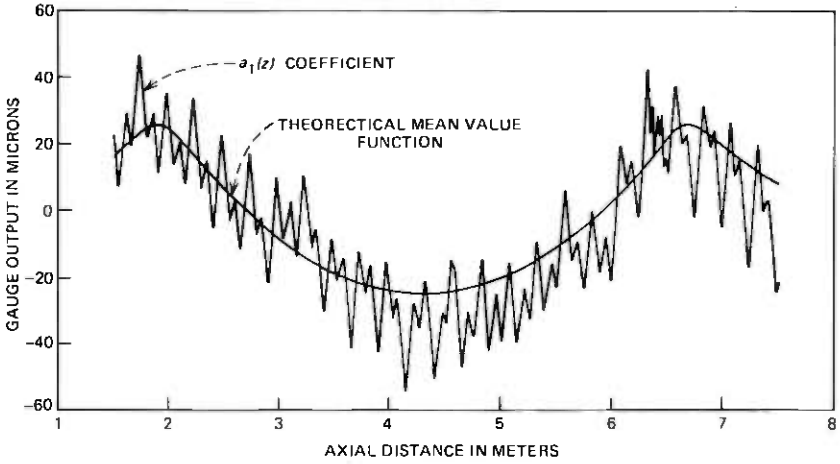


Fig. 10—Example of a spline mean value function. Data represent vertical curvature output from a rotating-head mouse.

covariances of the order statistics for gamma distributions required are given in Sarhan and Greenberg[54] and Prescott.[55]

For the unsmoothed subset estimates the means and covariances are particularly simple and the weights are given explicitly by

$$\Theta_j = \frac{1}{k'} \qquad j < k' \tag{28}$$

$$\Theta_{k'} = \frac{k + 1 - k'}{k'}$$

and the variance of $\vec{S}(\omega) = \mathrm{E}|\hat{S}(\omega)|^2/k'$ so that the efficiency, relative to an uncensored estimate, is just k'/k. It is shown by Mehrata and Nanda[56] that this estimate is maximum likelihood. This procedure is most effective for eliminating the effects of the occasional gross outlier missed in the data preparation stage but, unlike the robust filter algorithm, is ineffective against numerous small outliers.

As mentioned earlier, the length of the individual subsets is dependent on the fine structure of the spectrum to be estimated. A simple method of estimating this length (which within fairly broad bounds is not critical since the final estimate is primarily responsible for fine structure) is to compute a moving average representation of the process. For this purpose the Wiener canonical spectral factorization approach is ideally suited and, if in eq. (42) below, the sign of the summation is reversed and the expression Fourier-transformed, a moving average representation‡

‡ This moving average representation is the minimum delay causal nonrecursive (transverse) filter generating the observed process from white noise. The convolution of the moving average with itself gives the autocovariance function of the process.

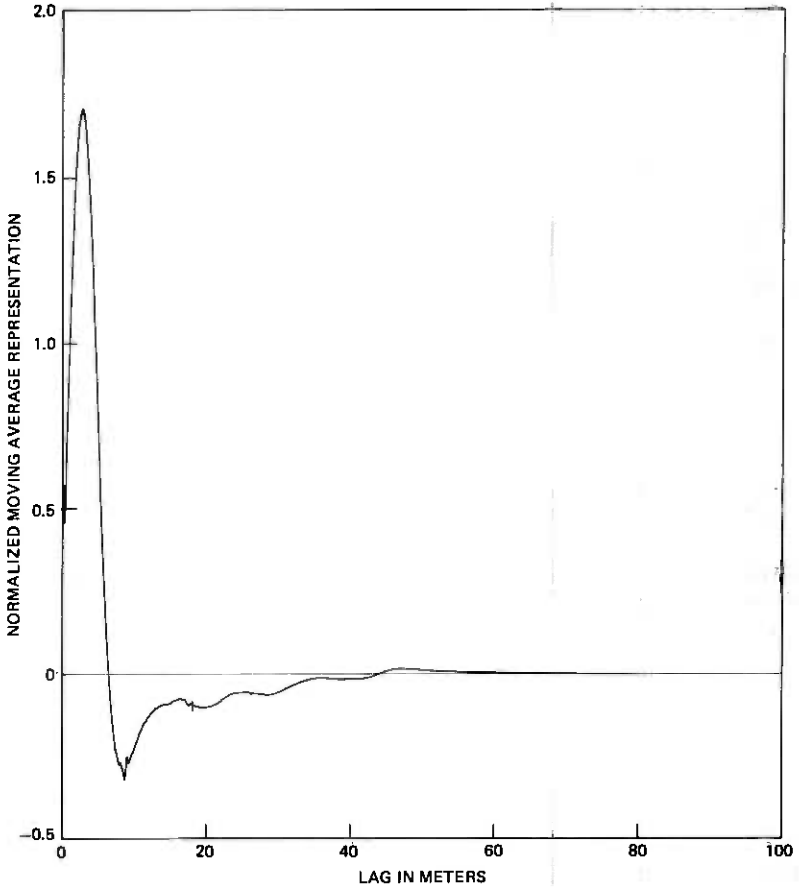


Fig. 11—Canonical moving average representation of vertical curvature gauge output. Representation based on the average of 102 450-meter subsets.

is obtained instead of an autoregression. Figure 11 shows such a model for the vertical gauge output for the Netcong field trial data and it is apparent that most of the weight is concentrated within a 60-meter range.† To allow for the heavy tapering effect of the 4π prolate window a subset length of 160 meters was used.

VI. CONSTRUCTION OF AUTOREGRESSIVE MODELS

The basic reason for computing a pilot spectrum estimate is to permit the design of an accurate prewhitening filter and the subsequent use of

† The discontinuities visible in this plot near 9 and 18 meters are due to the couplings but due to the randomized guide lengths this effect is rapidly suppressed with increasing separation.

a high-resolution spectrum estimation technique on the filtered data. Prewhitening filters are subject to several constraints; for example their transfer function must have no zeroes in the frequency range of interest [see eq. (61)], their design must be readily automated, they must have finite impulse response, they must be numerically well conditioned, and they must be absolutely stable. The prediction error filter satisfies these requirements.

Moreover, since the prediction error filter is causal and depends on the canonical autoregressive representation of a stationary time series it has the further major advantage that it may be readily robustified as described in Section VII. The use of the prediction error filter is therefore conditioned on one's ability to estimate the parameters of an autoregressive model of the process and this problem is the subject of the present section. Current work on autoregressive modeling uses two distinct approaches; direct solution of the Yule-Walker equations such as described by Pagano,[57] Ulrych and Bishop,[58] Makhoul,[59] and spectral factorization as described in Bhansali.[60,61] Following a brief review of these two approaches a composite technique is described which exploits features of both. The section concludes by considering three alternative methods of computing prewhitening filters.

In the *autoregressive* representation of a discrete time process the value, x_t, of the series at time t is given by the sum of a regression on the past values of the series and an independent random component, ξ_t,

$$x_t = \xi_t + \sum_{j=1}^{p} \alpha_j x_{t-j} \tag{29}$$

An equivalent description is to regard the regression on the past of the series as a *prediction* of the value of the process at time t so that the random component, ξ_t, represents the "new" information or *innovations* of the process. The length of the predictor or order of the regression is denoted by p which may be infinite. Such processes and questions related to them are discussed extensively in the literature; see for example Hannan,[62] Koopmans,[24] Box and Jenkins,[63] or Doob.[28] The papers by Kailath,[64] Kailath and Frost[65] are also relevant to these problems.

6.1 Yule-Walker equations and the Levinson algorithm

The basic equations determining the autoregressive coefficients are derived by minimizing the one-step prediction variance

$$\sigma_p^2 = \mathrm{E}\left\{\left(x_t - \sum_{j=1}^{p} \alpha_j^{(p)} x_{t-j}\right)^2\right\} \tag{30}$$

with respect to α_j for $j = 1, 2 \cdots p$ and are known as the *Yule-Walker equations*

$$\rho_k = \sum_{j=1}^{p} \alpha_j^{(p)} \rho_{|j-k|} \qquad k=1, 2, \cdots p \tag{31}$$

In these equations ρ_k is the autocorrelation function of the $\{x\}$ process at lag k. These equations are not only *linear* in the α's but also *Toeplitz*, so that the matrix elements depend only on their distance from the main diagonal and for real series the $p \times p$ matrix has only p distinct elements. Since these equations are linear, they may be solved using standard techniques such as the QR algorithm (see Dahlquist *et al.*[66]). However, because of their special structure, special procedures are available for their solution which require only p^2 operations instead of the p^3 required with general linear equation techniques. Also, because fewer operations are required, roundoff errors are reduced and the faster algorithms can be more accurate.

Generally these fast algorithms are similar in structure to the recursive solution of the Yule-Walker equations discovered by Levinson.[67] One convenient and numerically stable variant is due to Durbin,[69] which in the notation of Ramsey,[70] is initiated using

$$\phi_1 = \alpha_1^{(1)} = \rho_1 \tag{32}$$

$$\sigma_1^2 = 1 - \phi_1^2 \tag{33}$$

and continued for $k = 1, 2, \cdots, p-1$ by

$$\phi_{k+1} = \alpha_{k+1}^{(k+1)} = \left\{ \rho_{k+1} - \sum_{j=1}^{k} \alpha_j^{(k)} \rho_{k+1-j} \right\} \Big/ \sigma_k^2 \tag{34}$$

$$\alpha_j^{(k+1)} = \alpha_j^{(k)} - \phi_{k+1} \alpha_{k+1-j}^{(k)} \qquad j = 1, 2 \cdots, k \tag{35}$$

$$\sigma_{k+1}^2 = \sigma_k^2 (1 - \phi_{k+1}^2) \tag{36}$$

In these equations the $\alpha_j^{(k)}$'s are the autoregressive or prediction coefficients for k step prediction, the ϕ sequence is known as the *partial autocorrelation function*, and σ_k^2 is the k step relative prediction error. In the original Levinson algorithm the expansion

$$\sigma_k^2 = 1 - \sum_{j=1}^{k} \alpha_j^{(k)} \rho_j \tag{37}$$

obtained by substituting eq. (31) into (30) was used in place of eq. (36). Analytically these equations are identical but the latter is both slower and also has much poorer numerical properties than Durbin's form.

6.2 Spectral factorization

One drawback of the Toeplitz matrix formulation is that it does not provide much insight into the actual minimization process and it is helpful to rewrite the equations in terms of a *prediction error filter* where

we define

$$\alpha_0^{(p)} = -1 \tag{38}$$

and the negative error sequence

$$z_i = -(x_i - \hat{x}_i) \tag{39}$$

$$= \sum_{k=0}^{p} \alpha_k^{(p)} x_{i-k}$$

Note that the $\{z\}$ sequence is a result of a linear causal convolution, or filtering operation, applied to the $\{x\}$ sequence. The transfer function of this filter is

$$A^{(p)}(\omega) = \sum_{k=0}^{p} \alpha_k^{(p)} e^{-i\omega k} \tag{40}$$

so that the spectrum of $\{z\}$ is $S_x(\omega)|A^{(p)}(\omega)|^2$ with the corresponding variance

$$\sigma_p^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega)|A^{(p)}(\omega)|^2 \, d\omega \tag{41}$$

where $S_x(\omega)$ is the spectrum of the $\{x\}$ process. As $p \to \infty$ the spectrum of the error sequence approaches a constant so that the problem is to choose the causal filter in such a way that $|A|^2$ is small whenever S_x is large. For A a trigonometric polynomial of degree p the problem has been completely solved by Szegö[71] and the recursion formulae for the orthogonal polynomials obtained are essentially similar to those above.

An alternative solution is provided by Wiener's[68] canonical spectral factorization where the filter transfer function A is represented as

$$A(\omega) = -\exp\left\{-\sum_{k=1}^{\infty} c_k e^{-i\omega k}\right\} \tag{42}$$

so that the variance of $\{z\}$ is given by

$$\sigma_z^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) e^{-2\sum_{k=1}^{\infty} c_k \cos\omega k} \, d\omega \tag{43}$$

Direct minimization of this expression as a function of the c_k's is impractical due to the complexity of the resulting equations. Wiener's approach is to identify the c_k with the Fourier series coefficients of $\ln S_x$, that is

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos \omega k \, \ln \{S_x(\omega)\} \, d\omega \tag{44}$$

The sequence c_k is referred to as the *cepstrum*.

It is important to notice that the series in eq. (42) does not include a c_0 term because the constraint imposed by eq. (38) implies that c_0 defines

the minimum. This is most easily seen by substituting the Fourier series representation (44) into eq. (41) with the result

$$\sigma_f^2 = \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\{S_x(\omega)\}\, d\omega\right] \tag{45}$$

σ_f^2 is the *innovations* variance of the process. For this procedure to be valid the spectral distribution function must be absolutely continuous which implies that autoregressive representations are invalid for process containing periodic components. The procedure is formal and as mentioned in Wiener and Masani,[72] the sense in which it converges is unknown.

D. Preston[73] has observed that in practice these convergence problems may be avoided by evaluating the formula (Rozanov[74])

$$A(\omega) = \lim_{\mu \to 1} \frac{-1}{\sigma_f} \exp\left\{-\frac{1}{4\pi}\int_{-\pi}^{\pi} \ln S(\lambda) \frac{e^{-i\lambda} + \mu e^{-i\omega}}{e^{-i\lambda} - \mu e^{-i\omega}} d\lambda\right\} \tag{46}$$

inside it's radius of convergence, *ie* for $\mu < 1$, rather than on the radius of convergence as does the Wiener approach. With this modification one obtains

$$A_\mu(\omega) = -\exp\left\{-\sum_{k=1}^{\infty} c_k \mu^k e^{-i\omega k}\right\} \tag{47}$$

so that the coefficients α_j may be computed by Fourier transforming $A_\mu(\omega)$ and dividing by μ^j.

These two techniques have been described as if the actual spectrum were known. When applied to an estimate of the spectrum, things are more complex and neither technique has a clear advantage over the other. The disadvantage of the first approach is that it works explicitly with the autocorrelation function. The range of spectra common in waveguide work is so large that use of the autocorrelation function is numerically undesirable in that information corresponding to the lower parts of the spectrum may be lost due to numerical roundoff errors. The second method is numerically stable but produces a filter with a very long impulse response which reproduces all the details of the spectrum on which it is based, including those due to sampling. Since the robust filter algorithm works in the time domain the shortest autoregressive model which retains the statistically significant features of the spectrum is desirable.

Cleveland[75] and Bartholomew[76] have described several sources of error in prediction problems. Of these the most critical appears to be a result of sampling variability in the spectrum estimate. As an example consider the estimate of innovations variance, $\hat{\sigma}_f^2$ obtained by using the pilot spectrum estimate, $\overline{S}$, in place of the spectrum, S_x, in eq. (45). This estimate is described by Davis and Jones[77] except that their bias correction

is not used in steps involving the model formulation. (The bias correction is used to set the scale of the residuals for the robust filter algorithm.) Grenander and Rosenblatt[18] give a formula for prediction error in the case when the predictor is based on an estimate of spectrum, $\hat{S}$, rather than on the true spectrum, S, of the process

$$\hat{\sigma}_f^2 = \sigma_f^2 \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \frac{\hat{S}(\omega)}{S(\omega)}\, d\omega\right\} \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(\omega)}{\hat{S}(\omega)}\, d\omega \tag{48}$$

Periodogram estimates are distributed as χ_2^2 so that $\mathrm{E}\{\hat{S}^{-1}(\omega)\} = \infty$. While this result is based on the Wiener spectral factorization method and so applies to prediction using the entire past, it appears to give a good indication of the behavior of autoregressive fits even for relatively compact predictors. Further information on the effects of smoothing on the estimated innovations variance is available in Jones.[78]

6.3 Reduced factorization

A method which exploits many of the advantages of both of the preceding approaches without having the fatal flaws of either is to reduce the result of spectral factorization. In the reduced factorization approach one begins by creating a long autoregressive model using Wiener's spectral factorization method described above and then converts it to a shorter representation using the Levinson recursion formulae. In this reduction the key equation is (35) which, by combining the updates for $\alpha_j^{(k)}$ and $\alpha_{k+1-j}^{(k)}$, may be written backwards. When written for use in a downwards recursion this formula becomes

$$\alpha_j^{(k)} = \frac{\alpha_j^{(k+1)} + \phi_{k+1}\alpha_{k+1-j}^{(k+1)}}{1 - \phi_{k+1}^2} \tag{49}$$

Similarily the k-step prediction variance, σ_k^2 may be obtained from σ_{k+1}^2 by using eq. (36) backwards starting from the estimate of innovations variance given by eq. (45).

The major disadvantage of the reduced factorization technique is that it is somewhat slower than either of the standard techniques individually. Of these the Levinson recursion is the faster: it requires only a single Fourier transform to convert the pilot spectrum to a sample autocorrelation function and then p^2 operations for the actual solution. In practice it is necessary to "search" for the correct order of the autoregression. Since this search is never carried past $p_{\max} = \sqrt{T}$ the total computation time is $\sim T \ln T$. Since spectral factorization requires three Fourier transform operations, its speed is comparable with that of the Levinson technique. Reduced factorization requires an additional T^2 operations and is therefore considerably slower when very large data subsets are being used.

As mentioned earlier the most serious flaw with the Levinson approach is a result of roundoff errors *in the Fourier transform* used to convert the pilot spectrum estimate to autocovariances and is only serious when the range of the pilot spectrum estimate is large. Roundoff characteristics of fast Fourier transform algorithms are well understood (see Kaneko and Liu[79]) and consequently the characteristics of the pilot spectrum estimate relative to the computer precision may be used to select the "best" procedure: when the range of the pilot estimate is low the Levinson-Durbin algorithm is used, but in cases where the range is large reduced factorization is preferred.

With either approach the order of the autoregressive representation, p, has been chosen as the value of τ for which Parzen's[16] criterion

$$P(\tau) = 1 - \frac{\hat{\sigma}_I^2}{\hat{\sigma}_\tau^2} + \frac{\tau}{T} \tag{50}$$

attains its minimum. Within reasonable bounds the actual order selected is not critical as the autoregressive model is used as a prewhitening filter and not as a spectrum estimate. (The function $\hat{\sigma}_p^2/|A^{(p)}(\omega)|^2$ is known as an *autoregressive spectrum estimate*. See Akaike,[80] Gersch and Sharpe.[81]) Berk[82] gives conditions on the order, p, for obtaining a consistent model of the process.

The actual method used to determine the prediction error filter is a combination of the two methods discussed in Sections 6.2 and 6.3 as shown in the flow diagram, Fig. 12. In its general form this spectrum estimation technique is an iterative process and intermediate estimates are used to update the pilot estimate of spectrum and the prediction error filter. In cases when iteration is used it is stopped when the estimated innovations variance stabilizes.

6.4 Alternatives

Since the sequences of steps which is being used here to generate an autoregressive model is by no means obvious it is worthwhile to briefly examine the alternatives. The obvious technique of eliminating the pilot spectrum estimation and transformation to autocorrelations procedure and estimating the sample autocorrelations directly is not done because of the high bias, discussed in Section IV, of this estimate.

The second possibility is to form the pilot estimate of spectra and then design a conventional digital filter for the prewhitening operation. Details of this approach using Gegenbauer filters are given in Thomson.[83] The drawback is that such filters are incompatible with the robust filter algorithm.

A third alternative is to directly estimate the partial correlations using Burg's[84] algorithm. In this approach an autoregressive model is estimated

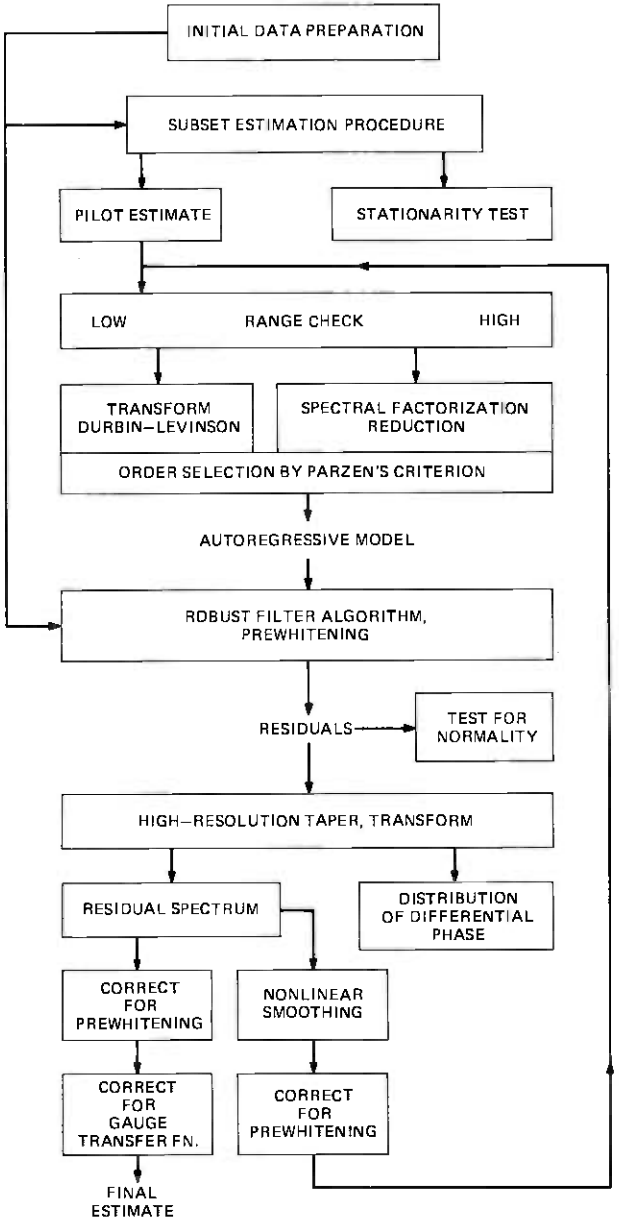


Fig. 12—Iterative estimation procedure.

by minimizing the sum of the forward and reverse prediction errors

$$e^{2(p)} = \sum_{n=1}^{T-p} \left(x_n - \sum_{k=1}^{p} \alpha_k^{(p)} x_{n+k} \right)^2 + \sum_{n=p+1}^{T} \left(x_n - \sum_{k=1}^{p} \alpha_k^{(p)} x_{n-k} \right)^2 \quad (51)$$

successively for $p = 1, 2 \cdots$ under the constraint that the covariance matrix is Toeplitz so that the autoregressive coefficients, $\alpha_k^{(p)}$, are updated using eq. (35). This method, also referred to as the "maximum entropy" approach, gives the partial autocorrelations directly and constrains them to be less than 1 in magnitude.

Limited Monte-Carlo studies indicate that spectra based on autoregressive representations obtained in this way have exceptionally high variance whenever the order, p, of the autoregression is carried far enough to reveal details of the spectrum. Other problems with the Burg algorithm as a *spectrum analysis technique* are described in Chen and Stegun.[85] As a *prewhitening algorithm* it has been used on individual tubes with reasonable success. These "partial Burg" routines have been effective in situations where a *low* order autoregressive representation is adequate for the prewhitening filter, the range of the spectrum is low, and very compact code is required.

VII. ROBUST FILTERING AND PREWHITENING

One of the most useful data analysis tools developed during the course of this work is the *robust filter algorithm.* This is a nonlinear technique designed to eliminate the effects of occasional "outliers" in the data from the final spectrum estimate, where, as mentioned earlier, outliers are measured on the scale of the innovations process. As an example of the magnitude of this problem, the typical output of a tubing curvature gauge is about 15 microns rms whereas the scale of the innovations process is about 0.5 microns. This is considerably smaller than the size of typical dust particles and, since this gauge operates primarily in a tubing mill, the probability of some dust particles being measured is quite high and the need for robust filtering is evident.

The robust filter algorithm differs from linear filtering in that most of the data passes through the "filter" without modification and only those points which are basically unpredictable from past values of the series are changed. The characteristics of the filtering algorithm are controlled by providing (*i*) an autoregressive model of the process, (*ii*) an estimate of the innovations variance, and (*iii*) an influence function. In practice the autoregressive model and innovations variance must be estimated from the data and it has been found that the algorithm works well even with surprisingly inacurate models. Further details and examples on this procedure are available in Kleiner *et al.*[86] The steps of this procedure, as it is currently implemented, are listed below. Section 7.2 summarizes results (see Kleiner *et al.*[87] for details) relevant to the choice of influence function, and in Section 7.3 an example of the action on contaminated data is given. Further information on robust procedures is available in Huber[88] and Hampel.[48]

7.1 The robust filter algorithm

We assume that the observations, $\{y\}$, consist of the process of interest, $\{x\}$, plus occasional outliers, $\{v\}$

$$y_k = x_k + v_k \tag{52}$$

Based on this contaminated data the robust filter algorithm produces an estimate, $\{\hat{x}\}$, of the "core" process by the following steps:

(*i*) A prediction, $\tilde{x}_n$, is made from the filtered sequence using the autoregressive coefficients obtained by the methods discussed in Section VI.

$$\tilde{x}_n = \sum_{k=1}^{p} \alpha_k^{(p)} \hat{x}_{n-k} \tag{53}$$

(*ii*) A weight is defined which depends on the difference, $y_n - \tilde{x}_n$, between the actual observation, y_n, and the prediction. This difference is normalized by the scale of the innovations process, σ_p. (The scale is the square root of the prediction variance estimate, $\hat{\sigma}_p^2$, with the bias correction given in Davis and Jones.[77])

$$w_n = W\left(\frac{y_n - \tilde{x}_n}{\sigma_p}\right) \tag{54}$$

In the applications described here W is an even function with $W(0) = 1$ and $W(\infty) = 0$. When multiple errors are encountered the scale, σ_p, used in this formula is replaced by an approximation of the k-step prediction variance.

(*iii*) The output of the robust filter algorithm is an estimate of the core process, $\hat{x}_n$, formed by the weighted average of observation and prediction

$$\hat{x}_n = w_n y_n + (1 - w_n)\tilde{x}_n \tag{55}$$

The effect of this procedure is to leave the data unmodified where the prediction errors are small and to replace the data with its prediction at points where the prediction errors are gross. The action taken when the prediction errors are near the expected extreme for the given sample size depends on the weight function which will be discussed below. In spectrum estimation applications the desired output is usually not the filtered sequence but rather the prewhitening residuals

$$z_n = \hat{x}_n - \tilde{x}_n \tag{56}$$

The z_n may be described in terms of an influence function (see Hampel[89])

$$\psi(e) = eW(e) \tag{57}$$

applied to the relative prediction error, $(y_n - \tilde{x}_n)/\sigma_p$, but the notation is deceptive in that it deemphasizes the fact that the weighting procedure also influences the prediction for subsequent steps.

To use this algorithm the filtered sequence must be initialized on the first p points. When the data is only slightly contaminated the raw data has been used to start the process but when the contamination is more severe special precautions must be taken.

The detailed behavior of the algorithm depends on the choice of weight function, and this represents a compromise between rejecting valid outliers of the innovations and accepting the occasional erroneous data point. Considerable information is available on the choice and characteristics of influence functions for robust estimates of location (see Andrews *et al.*[90]), but this information is of limited utility in time series applications since in location estimates there is no concern with frequency response characteristics. It must be remembered that this operation is *nonlinear* and that nonlinear operations on time series generally change the spectrum in complex ways. Because of this the weight or influence function must be chosen in such a way that the spectral content due to the induced nonlinearities is much less than that due to the presence of errors in the data.

Several different weights have been used. Of these the best found to date is a result of motivation by the extreme value distribution for distributions of exponential type (see Kendall and Stuart[91]) and is defined by

$$W(u) = \exp\{-e^{u_0(|u|-u_0)}\} \tag{58}$$

in which

$$u_0 = \Phi^{-1}\left(1 - \frac{1}{N}\right) \tag{59}$$

Φ being the normal cumulative distribution function and N the sample size. This influence function, shown in Fig. 13 for $N = 1000$, is very linear in the center and, at about $\pm 3\sigma$, decreases rapidly to zero.

7.2 Spectral distortions resulting from robust filtering

In its most general form, use of the robust filter algorithm is alternated with the model formation process as shown in Fig. 12. In this iterative mode the output from the filter is used to generate a better autoregressive model which is used to filter the data and so on. This kind of iterative procedure has been used for some difficult data sets and was found to converge to a stable estimate of spectrum very rapidly. Typically two or three iterations are required on short series (for example, some distortions in individual tubes) where the range of the spectrum is very large

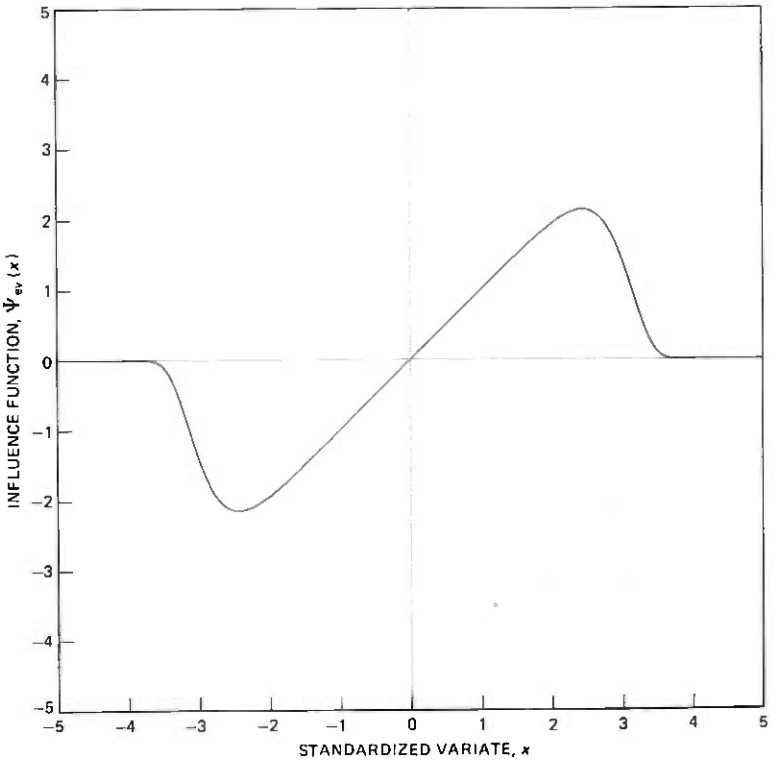


Fig. 13—Extreme value influence function.

and the outliers are small relative to the scale of the process but large compared to the scale of the innovations. With very large data sets, such as those from complete mode filter sections of the field trial (which average 80,000 data points), a single iteration has been used and found satisfactory.

If one assumes that this iterative process has converged, it is possible to describe the distortions introduced into the spectral density estimate. At convergence the autoregressive parameters, $\hat{\alpha}_k$, describe the estimated process, $\{\hat{x}\}$, and are solutions of the Yule-Walker equations based on the estimated process. Then by computing the expectation of $\hat{x}_{n-k}$ with $\hat{x}_n$ using the representation eq. (55) for the latter, it is found that the $\hat{\alpha}_k$'s also are solutions of a set of *robust* Yule-Walker equations

$$\mathrm{E}\left\{\hat{x}_{n-k}\psi\left[\frac{y_n - \sum_{j=1}^{p}\hat{\alpha}_j\hat{x}_{n-j}}{\sigma_p}\right]\right\} = 0 \qquad k = 1, \cdots, p \tag{60}$$

An alternative viewpoint is to regard the algorithm as the solution of minimizing a nonquadratic loss function of $y_n - \tilde{x}_n$ with respect to the $\hat{\alpha}_k$'s and it can be shown that the solution to this problem also yields the robust Yule-Walker equations provided that the influence function is such that the error sequences, $\partial \hat{x}_n / \partial \hat{\alpha}_j$ are small. This is a reasonable requirement: small changes in the process specification should not result in large changes in the filter output. The satisfaction of this condition depends on the choice of influence function, ψ. It can be shown that the scale of the error sequences depends on $1 - \psi'$ so that influence functions having very high curvature in regions where the probability density function of the innovations process is large result in larger errors than influence functions which are more linear in such regions. The most important property of the algorithm, however, is that, for reasonable influence functions, the effect of the nonlinearities on the spectrum estimate, is proportional to the spectrum so that the net effect is a slight downwards bias. The scale of the bias factor is $E\{\xi\psi(\xi)\}$ and, for the dominant error terms, is independent of frequency.

7.3 Action of the robust filter on contaminated data

The intent of the robust filter algorithm is to reduce the effects of outliers and erroneous data from the final estimate of spectra. Since the choice of influence function is to some extent distribution dependent, it is also of interest to observe the effect of this algorithm in a direct manner. It is also interesting to check to what extent a normal assumption on the basic data is warranted. Since the high serial correlations existing in most time series in the physical sciences make the usual tests for goodness-of-fit to a given distribution inapplicable this must be done cautiously. A very conservative approach is to find some lag, τ_0, such that the autocorrelations at multiples of this lag are small and test samples taken at this spacing for normality. Since the spacing required to obtain uncorrelated data may be large, this approach is rather inefficient. An alternative is to consider the residuals from the prewhitening operation. Since these residuals are generally very small, usually only a few times the quantization level, this method is very sensitive to outliers and measurement errors. Figure 14 shows a Q-Q plot of the residuals from a *linear* prewhitening operation and it is clear that the apparent distribution has very heavy tails. If the actual residuals are plotted as a time series, Fig. 15, it is clear that at least part of the long tailed characteristics are a simple consequence of the fact that in linear prewhitening each outlier in the original series is converted into $p + 1$ outliers in the residual series. In Fig. 16 a Q-Q plot of the residuals from the robust prewhitening algorithm is given and the contrast is striking. In this case the residuals are quite close to normal and in agreement with tests made on other sections of the line.

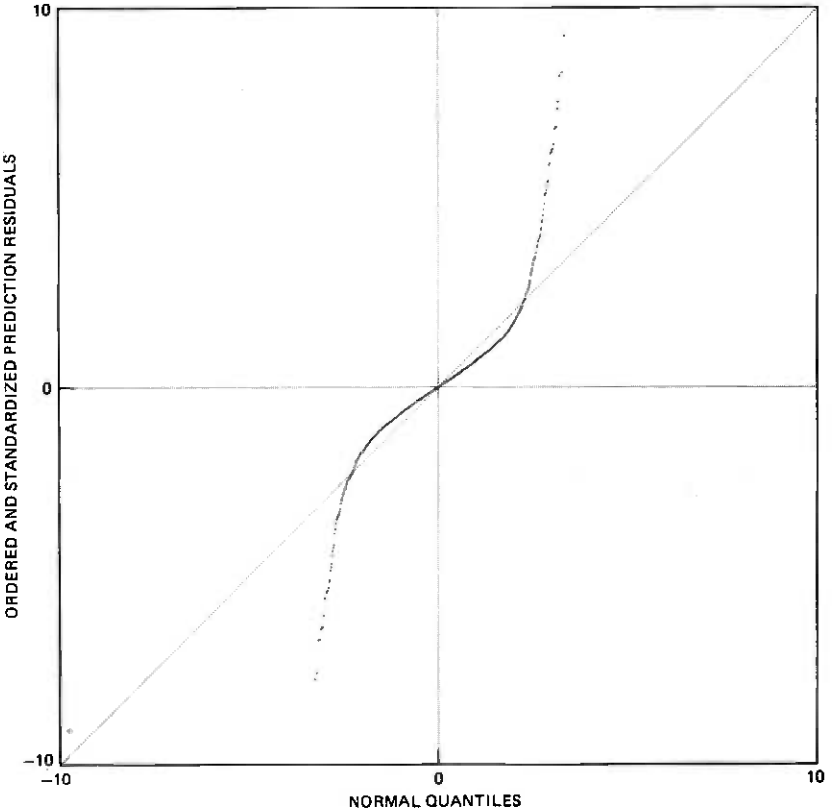


Fig. 14—WT4 field evaluation test horizontal curvature gauge output. Residuals from a linear prediction error filter.

VIII. FINAL ESTIMATE OF SPECTRUM

Prewhitening converts the data from a highly correlated into an almost uncorrelated form whose spectrum has a low dynamic range. Estimates of such spectra are best made with windows which have high frequency resolution and do not need the extreme sidelobe suppression used for the pilot estimate and the Tukey spliced cosine window has been used for most such applications.

The final estimate is intended primarily to extract *details* of the process: consequently the data is *not* split into subsets and the estimate is not smoothed by liftering. In cases where "smoothing" is done it has been by the nonlinear methods discussed in Section 2.7. These techniques might be described as "inverse influence" in that individual points are lumped into a moving average *except* when they are outliers in which case they are used instead of the average. This procedure is a useful aid for spotting peaks and other low level features in the spectrum.

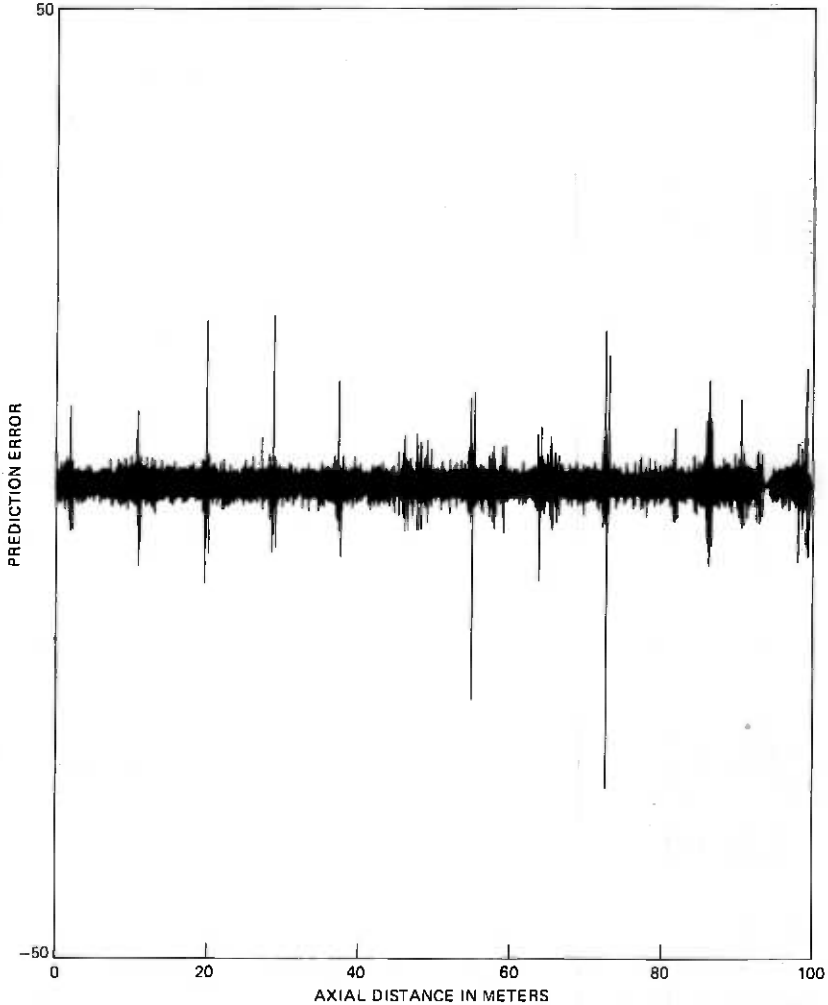


Fig. 15—WT4 field evaluation test horizontal curvature gauge output. Residuals from a linear prediction error filter.

In the final estimate of spectrum it is necessary to correct for the prewhitening operation so that the result is expressed as the ratio

$$\hat{S}(\omega) = \frac{S_z(\omega)}{\left| \sum_{k=0}^{p} \alpha_k e^{-i\omega k} \right|^2} \tag{61}$$

in which $S_z(\omega)$ is a direct estimate of the spectrum of the prewhitened residuals [eq. (56)], and the denominator is the power transfer function of the prediction error filter defined in eq. (39).

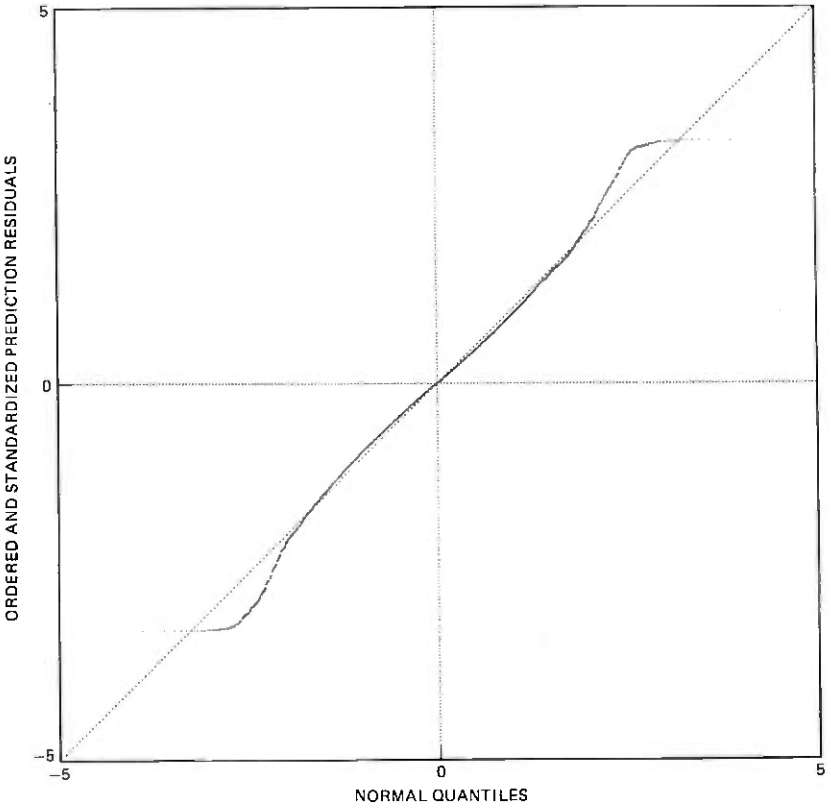


Fig. 16—WT4 field evaluation trial curvature gauge output. Prediction residuals from the robust filter algorithm.

The validity of this form depends on the assumption of the independence of the filter and the residuals and is discussed briefly in Grenander and Rosenblatt.[18] With the prediction error filter this is a reasonable assumption in that the filter depends on the partial autocorrelation functions up to lag p while any residual structure is primarily the contribution of the partial correlations for higher lags. This assumption is also supported by Whittle's[15] observation that the information matrix splits into one part describing the structure of the process and a second part describing the innovations sequence.

IV. CONCLUSIONS OF PART I

A technique for estimating the power spectral density function of a stationary time series has been described which is robust, accurate, and computationally straightforward. Part II of this paper will give examples of its use and comparisons with standard techniques.

APPENDIX A

Formulae for prolate spheroidal data windows

A convenient expansion of the prolate spheroidal wave function data windows is given in Flammer[38] Section 3.2. This expansion is a power series in terms of

$$U = (1 - x)(1 + x)$$

Computationally it is advantageous to rewrite the power series using Horner's rule and for $c = 4\pi$ the expansion is:

$$D_{00}(4\pi, x) = \sqrt{\frac{2\Delta x}{.5081255481474 97T}}\,(((((((((((((((((((((($$

$$+2.6197747176990866d - 11\ U + 2.9812025862125737d - 10)\ U$$

$$+3.0793023552299688d - 09)\ U + 2.8727486379692354d - 08)\ U$$

$$+2.4073904863499725d - 07)\ U + 1.8011359410323110d - 06)\ U$$

$$+1.1948784162527709d - 05)\ U + 6.9746276641509466d - 05)\ U$$

$$+3.5507361197109845d - 04)\ U + 1.5607376779150113d - 03)\ U$$

$$+5.8542015072142441d - 03)\ U + 1.8482388295519675d - 02)\ U$$

$$+4.8315671140720506d - 02)\ U + 1.0252816895203814d - 01)\ U$$

$$+1.7233583271499150d - 01)\ U + 2.2242525852102708d - 01)\ U$$

$$+2.1163435697968192d - 01)\ U + 1.4041394473085307d - 01)\ U$$

$$+5.9923940532892353d - 02)\ U + 1.4476509897632850d - 02)\ U$$

$$+1.5672417352380246d - 03)\ U + 4.2904633140034110d - 05)$$

The expansion for the higher resolution window with $c = \pi$ is:

$$D_{00}(\pi, x) = \sqrt{\frac{2\Delta x}{T}}\,(((((((((($$

$$+5.3476939016920851d - 11\ U + 2.2654256220146656d - 09)\ U$$

$$+7.8075102004229667d - 08)\ U + 2.1373409644281953d - 06)\ U$$

$$+4.5094847544714943d - 05)\ U + 7.0498957221483167d - 04)\ U$$

$$+7.7412693304064753d - 03)\ U + 5.5280627452077586d - 02)\ U$$

$$+2.2753754228751827d - 01)\ U + 4.3433904277546202d - 01)\ U$$

$$+2.2902051859068017d - 01)$$

In the forms given here both functions have been normalized for use as

data windows. In this application x takes on values

$$x_t = \frac{2t-1}{T} - 1; \qquad t = 1, 2, \ldots, T$$

REFERENCES

1. J. C. Anderson *et al.*, B.S.T.J., to be published.
2. R. A. Fisher, *Statistical Methods and Scientific Inference* (3d ed), Hafner Press, 1973.
3. M. Arato, "On the Sufficient Statistics for Stationary Gaussian Processes," Theory Probab. Appl., *6* (1961), pp. 199–201.
4. P. Whittle, "Estimation and Information in Stationary Time Series," Arkiv Fōr Matermatik, *2* (1953), pp. 423–434.
5. D. J. Thomson, "Spectral Analysis of Short Series," thesis, Polytechnic Institute of Brooklyn, 1971.
6. M. Loève, *Probability Theory*, D. Van Nostrand, 1963.
7. U. Grenander and G. Szegō, *Toeplitz Forms and Their Applications*, Univ. of Cal. Press, 1958.
8. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—II," B.S.T.J., *40*, No. 1 (January 1961), pp. 65–84.
9. K. O. Dzhaparidze and A. M. Yaglom, "Asymtotically Efficient Estimation of the Spectrum Parameters of Stationary Stochastic Processes," Proc. Prague Symp. on Asymptotic Statistics, *1*, Prague: Charles Univ. Press, 1974.
10. E. A. Aronson, "Fast Fourier Integration of Piecewise Polynomial Functions," Proc. IEEE, *57* (1969), pp. 691–692.
11. P. D. Welch, "The Use of the Fast Fourier Transform for Estimation of Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," IEEE Trans. Audio Electroacoust., *AU-15* (1967), pp. 70–74.
12. D. J. Thomson, "A Test for Stationarity," 1977.
13. T. Kailath, "A View of Three Decades of Linear Filtering Theory," IEEE Trans., *IT-20* (1974), pp. 146–180.
14. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, *63* (1975), pp. 563–580.
15. P. Whittle, *Prediction and Regulation by Linear Least-Squares Methods*, D. Van Nostrand, 1963.
16. E. Parzen, "Some Recent Advances in Time Series Modelling," IEEE Trans., *AC-19* (1974), pp. 723–730.
17. D. C. Rife and G. A. Vincent, "Use of the Discrete Fourier Transform in the Measurement of Frequencies and Levels of Tones," B.S.T.J., *49*, No. 2 (February 1970), pp. 197–228.
18. U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, New York: Wiley, 1975.
19. R. B. Blackman and J. W. Tukey, "The Measurement of Power Spectra," B.S.T.J., *37*, Nos. 1 and 3 (January and March 1958), pp. 185–282, 485–569. (Reprinted by Dover.)
20. R. H. Jones, "Spectral Estimates and Their Distributions," Skandinavisk Aktuatietidskrift, *45* (1962), pp. 39–69, 135–153.
21. R. H. Jones, "A Reappraisal of the Periodogram in Spectral Analysis," Technometrics, *7* (1965), pp. 531–542.
22. R. H. Jones, "Spectrum Estimation with Missing Observations," Ann. Inst. Stat. Math., *23* (1971), pp. 387–398.
23. J. W. Tukey, "An Introduction to the Calculations of Numerical Spectrum Analysis," *Spectral Analysis of Time Series*, B. Harris, ed., New York: Wiley, 1967.
24. L. H. Koopmans, *The Spectral Analysis of Time Series*, Academic Press, 1974.
25. D. Brillinger, *Times Series, Data Analysis and Theory*, Holt, Rinehart & Winston, 1975.
26. V. F. Pisarenko, "On the Estimation of Spectra by Means of Non-Linear Functions of the Covariance Matrix," Geophysical J., Royal Astronomical Soc., *28* (1972), pp. 511–531.
27. J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," Proc. IEEE, *57* (1969), pp. 1408–1418.
28. J. L. Doob, *Stochastic Processes*, New York: Wiley, 1953.

29. D. Slepian and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—I," B.S.T.J., *40*, No. 1 (January 1961), pp. 43–64.
30. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—III," B.S.T.J., *41*, No. 4 (July 1962), pp. 1295–1336.
31. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—IV," B.S.T.J., *43*, No. 9 (November 1964), pp. 3009–3057.
32. D. Slepian and E. Sonnenblick, "Eigenvalues Associated with Prolate Spheroidal Wave Functions of Zero Order," B.S.T.J., *44*, No. 8 (October 1965), pp. 1745–1759.
33. D. Slepian, "Some Asymptotic Expansions for Prolate Spheroidal Wave Functions," J. Math. Physics, *44* (1965), pp. 99–140.
34. D. J. Thomson, M. F. Robbins, C. G. MacLennan, and L. J. Lanzerotti, "Spectral and Windowing Techniques in Power Spectral Analysis of Geomagnetic Data," Physics of the Earth and Planetary Interiors, *12* (1976), pp. 217–231.
35. J. F. Kaiser, "Nonrecursive Digital Filter Design Using the I_0 – sinh Window Function," IEEE Inter. Symp. Circuits & Systems Proc. (1974), pp. 20–23.
36. A. Eberhard, "An Optimal Discrete Window for the Calculation of Power Spectra," IEEE Trans., *AU-21* (1973), pp. 37–43.
37. E. Parzen, "Mathematical Considerations in the Estimation of Spectra," Technometrics, *3* (1961), pp. 167–190.
38. C. Flammer, *Spheroidal Wave Functions*, Stanford Univ. Press, 1967.
39. R. P. Bogert, M. J. Healy, and J. W. Tukey, "The Frequency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking," *Time Series Analysis*, M. Rosenblatt, ed., pp. 209–243, New York: Wiley, 1962.
40. R. A. Fisher, "Tests of Significance in Harmonic Analysis," Proc. Royal Soc. London, *125 A* (1929), pp. 54–59.
41. M. S. Bartlett, "Some Remarks on the Analysis of Time Series," Biometrika, *54* (1967), pp. 25–38.
42. K. S. Miller, R. I. Bernstein, and L. E. Blumenson "Generalized Rayleigh Processes," Quart. Jour. Math., *16* (1958), pp. 137–145.
43. E. Parzen, "On Consistent Estimates of the Spectrum of a Stationary Time Series," Ann. Math. Stat., *28* (1957), pp. 329–348.
44. A. Papoulis, "Minimum-Bias Windows for High-Resolution Spectral Estimates," IEEE Trans., *IT-19* (1973), pp. 9–12.
45. U. Grenander, H. O. Pollak, and D. Slepian, "The Distribution of Quadratic Forms in Normal Variates: A Small Sample Theory with Applications to Spectral Analysis," J. SIAM, *7* (1959), pp. 374–401.
46. W. S. Cleveland and E. Parzen, "The Estimation of Coherence, Frequency Response, and Envelope Delay," Technometrics, *17* (1975), pp. 167–172.
47. J. W. Cooley, P. A. W. Lewis, and P. D. Welch, "The Application of the Fast Fourier Transform Algorithm to the Estimation of Spectra and Cross-Spectra," *Computer Processing in Communications*, Polytechnic Institute of Brooklyn Microwave Research Institute Symposia Series, *19* (1969), pp. 5–20.
48. F. R. Hampel, "Robust Estimation: A Condensed Partial Survey," Z. Wahrscheinlichkeitstheorie verw., *27* (1973), pp. 87–104.
49. H. B. Curry and I. J. Schoenberg, "On Polya Frequency Functions IV; The Fundamental Spline Functions and Their Limits," Jour. Analyse Math., *17* (1966), pp. 71–107.
50. C. de Boor, "Package for Calculating with B-Splines," SIAM Jour. Numer. Analysis, *14* (1977), pp. 441–472.
51. L. L. Horowitz, "The Effects of Spline Interpolation on Power Spectra Density," IEEE Trans., *ASSP-22* (1974), pp. 22–27.
52. P. E. Fox, S. Harris, and D. J. Thomson, "Mechanical Gauging Techniques," B.S.T.J., to be published.
53. E. H. Lloyd, "Least Squares Estimation of Location and Scale Parameters Using Order Statistics," Biometrika, *39* (1952), pp. 88–95.
54. A. E. Sarhan and B. G. Greenberg, *Contributions to Order Statistics*, New York: Wiley, 1962.
55. P. Prescott, "Variances and Covariances of Order Statistics from the Gamma Distribution," Biometrika, *61* (1974), pp. 607–613.
56. K. G. Mehrata and P. Nanda, "Unbiased Estimation of Parameters by Order Statistics in the Case of Censored Samples," Biometrika, *61* (1974), pp. 601–606.
57. M. Pagano, "An Algorithm for Fitting Autoregressive Schemes," Jour. Royal Stat. Soc., C *21* (1972), pp. 274–281.
58. T. J. Ulrych and T. N. Bishop, "Maximum Entropy Spectral Analysis and Autoregressive Decomposition," Rev. Geophys. Space Phys., *13* (1975), pp. 183–200.

59. J. Makhoul, "Spectral Linear Prediction: Properties and Applications," IEEE Trans., *ASSP-23* (1975), pp. 283–296.
60. R. J. Bhansali, "A Monte-Carlo Comparison of the Regression Method and the Spectral Methods of Prediction," Jour. Amer. Stat. Assoc., *68* (1973), pp. 621–625.
61. R. J. Bhansali, "Asymptotic Properties of the Wiener-Kolmogorov Predictor," Jour. Royal Stat. Soc., 1974, pp. 61–73.
62. E. J. Hannan, *Multiple Time Series*, New York: Wiley, 1970.
63. G. E. P. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, Holden-Day, 1970.
64. T. Kailath, "An Innovations Approach to Least-Squares Estimation—Part I: Linear Filtering in Additive White Noise," IEEE Trans., *AC-13* (1968), pp. 645–655.
65. T. Kailath and P. Frost, "An Innovations Approach to Least-Squares Estimation—Part II: Linear Smoothing in Additive White Noise," IEEE Trans., *AC-13* (1968), pp. 655–660.
66. G. Dahlquist, Å. Björk, and N. Anderson, *Numerical Methods*, New York: Prentice-Hall, 1974.
67. N. Levinson, "The Wiener RMS Error Criterion in Filter Design and Prediction," Jour. Math. Physics, *25* (1947), pp. 261–278. (Reprinted as Appendix B of Wiener.[68])
68. N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, M.I.T. Press, 1949.
69. J. Durbin, *Distribution Theory for Tests Based on the Sample Distribution Function*, SIAM, 1973.
70. F. L. Ramsey, "Characterization of the Partial Autocorrelation Function," Ann. Stat., 2 (1974), pp. 1296–1301.
71. G. Szegő, *Orthogonal Polynomials*, third ed., American Math. Society, 1967.
72. N. Wiener and P. Masani, "The Prediction Theory of Multivariate Stochastic Processes, II—The Linear Predictor," Acta Mathematica, *99* (1958), pp. 93–137.
73. D. B. Preston, "private communication," 1977.
74. Y. A. Rozanov, *Stationary Stochastic Processes*, Holden Day, 1967.
75. W. S. Cleveland, "Fitting Time Series Models for Prediction," Technometrics, *13* (1971), pp. 713–723.
76. D. J. Bartholomew, "Errors of Prediction for Markov Chain Models," Jour. Royal Stat. Soc., *B 37* (1975), pp. 444–456.
77. H. T. Davis and R. H. Jones, "Estimation of the Innovations Variance of a Stationary Time Series," Jour. Amer. Stat. Assoc., *63* (1968), pp. 141–149.
78. R. H. Jones, "Estimation of the Innovation Generalized Variance of a Multivariate Stationary Time Series," Jour. Amer. Stat. Assoc., *71* (1976), pp. 386–388.
79. T. Kaneko and B. Liu, "Accumulation of Round-off Error in Fast Fourier Transforms," J. Assoc. Comp. Mach., *17* (1970), pp. 637–654.
80. H. Akaike, "Power Spectrum Estimation Through Autoregressive Model Fitting," Ann. Inst. Stat. Math., *21* (1969), pp. 407–419.
81. W. Gersch and D. R. Sharpe, "Estimation of Power Spectra with Finite-Order Autoregressive Models," IEEE Trans., *AC-18* (1973), pp. 367–369.
82. K. N. Berk, "Consistent Autoregressive Spectral Estimates," Ann. Stat., 2 (1974), pp. 489–502.
83. D. J. Thomson, "Generation of Gegenbauer Prewhitening Filters by Iterative Fast Fourier Transforming," *Computer Processing in Communications*, Polytechnic Institute of Brooklyn Press, 1969.
84. J. P. Burg, "Maximum Entropy Spectral Analysis," thesis, Stanford Univ., 1975.
85. W. Y. Chen and G. R. Stegen, "Experiments with Maximum Entropy Power Spectra of Sinusoids," J. Geophysical Res., *79* (1974), pp. 3019–3022.
86. B. Kleiner, R. D. Martin, and D. J. Thomson, "Three Approaches Towards Making Power Spectra Less Vulnerable to Outliers," Bus. & Econ. Stat. Sect., Proc. Amer. Stat. Assoc. (1976), pp. 386–391.
87. B. Kleiner, R. D. Martin, and D. J. Thomson, "Robust Estimates of Spectra," in preparation.
88. P. J. Huber, "Robust Statistics: a Review," Ann. Math. Stat., *43* (1972), pp. 1041–1067.
89. F. R. Hampel, "The Influence Curve and its Role in Robust Estimation," Jour. Amer. Stat. Assoc., *69* (1974), pp. 383–393.
90. D. F. Andrews *et al.*, *Robust Estimates of Location*, Princeton Univ. Press, 1972.
91. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics, 1*, New York: Hafner, 1963.

Contributors to This Issue

James L. Blue, A.B., 1961, Occidental College; Ph.D., 1966, California Institute of Technology; Bell Laboratories, 1966—. Mr. Blue has done research in noise theory for avalanche diodes and in modeling of semiconductor devices, and was involved in the development of computer aids for testing of integrated circuits. He is now a member of the Computing Mathematics Research Department, where he is involved in mathematical modeling, research in numerical methods, and the development of numerical software.

Ronald E. Crochiere, B.S., (E.E.) 1967, Milwaukee School of Engineering; M.S. (E.E.) and Ph.D. (E.E.), 1968 and 1974, Massachusetts Institute of Technology; Raytheon Co., 1968–1970; Bell Laboratories, 1974–. Mr. Crochiere is presently engaged in research activities in speech communications, speech coding, and digital signal processing. Member, IEEE, Sigma Xi, ASSP-DSP Subcommittee.

Adolf J. Giger, Diploma E. E., 1950, and Dr. Sc. Techn., 1956, Swiss Federal Institute of Technology; Bell Laboratories, 1956—. He has worked in the field of microwave communications, including work on TH-1 and its associated protection switching system, on low-noise receivers, waveguide circuits, antennas and autotracking for Telstar, and on the digital systems and circuit aspects of the WT-4 millimeter waveguide system. As head of a microwave radio department he has been responsible for the development of new analog and digital radio equipment and the current engineering on existing radio systems. Senior member, IEEE.

David J. Goodman, B.E.E., 1960, Rensselaer Polytechnic Institute; M.E.E., 1962, New York University; Ph.D. (E.E.), 1967, Imperial College London; Bell Laboratories, 1967–. Mr. Goodman has studied various aspects of digital communications, including analog-to-digital conversion, digital signal processing, assessment of the quality of digitally coded speech, and error mechanisms in digital transmission lines. In 1974 and 1975, he was a Senior Research Fellow at Imperial College, London, England. Member, IEEE.

B. Gopinath, M.Sc. (Mathematics), 1964, University of Bombay; Ph.D. (E.E.), 1968, Stanford University; research associate, Stanford University, 1967–1968; Alexander von Humbolt research fellow, University of Göttingen, 1971—1972; Bell Laboratories, 1968—. Mr. Gopinath is engaged in applied mathematics research in the Mathematics and Statistics Research Center.

D. C. Hogg, B.Sc., 1949, University of Western Ontario; M.Sc., 1950, Ph.D., 1953, McGill University; Bell Laboratories, 1953—February 1977 (retired). Mr. Hogg's work included studies of artificial dielectrics for microwaves, diffraction of microwaves, and over-the-horizon, millimeter wave, and optical propagation and antenna research. Mr. Hogg is now Chief, Environmental Radiometry, at the Wave Propagation Laboratory, Environmental Research Laboratories, National Oceanic and Atmospheric Administration, Boulder, Colorado. Fellow, IEEE, and Union de Radio Scientifique Internationale.

Frank K. Hwang, B.A., 1960, National Taiwan University; M.B.A., City University of New York; Ph.D. (statistics), 1968, North Carolina State University; Bell Laboratories, 1967—. Mr. Hwang visited the Department of Mathematics of National Tsing-Hua University in 1970, and the Institute of Mathematics, Academia Sinica and Telecommunication Laboratories in 1976. He has been engaged in research in statistics, computing algorithms, discrete mathematics, and switching networks.

T. C. Liang, B.S. (math.), 1972, and M.S. (applied math.), 1976, National Tsing-Hua University, Telecommunication Laboratories, 1976—. Mr. Liang has been engaged in research in statistics and switching networks.

Sing-Hsiung Lin, B.S.E.E., 1963, National Taiwan University; M.S.E.E., 1966, and Ph.D., 1969, University of California, Berkeley; Bell Laboratories, 1969—. At the Electronics Research Laboratory, University of California at Berkeley, Mr. Lin was engaged in research on antennas in plasma media and numerical solutions of antenna problems. Mr. Lin is presently working on wave propagation problems on terrestrial radio systems and earth-satellite radio systems. Member, IEEE, Sigma Xi, AIAA.

A. C. Longton, B.S.E.E., 1954, Tufts University, M.S.E.E., 1962, Northeastern University; Bell Laboratories, 1954—. Mr. Longton has worked on the exploratory development of PCM systems. He supervised the development of channel units with built-in interoffice signaling for the T1 carrier system. He has supervised the development of IF circuits for the TD-3 microwave radio system and held a number of responsibilities on radio system developments. In 1966 he became head of a department responsible for short-haul radio and the development of radio protection switching systems. In 1969 he became responsible for exploratory studies of digital radio systems and currently heads the Digital Radio Department.

Debasis Mitra, B.Sc. (E.E.), 1964, and Ph.D. (E.E.) 1967, London University; Bell Laboratories, 1967—. Mr. Mitra has worked on the stability analysis of nonlinear systems, semiconductor networks, analysis of queues in communication systems, computer memory management, growth models for new communication services, and speech waveform coding. Most recently he has been involved in the analysis of adaptive systems and digital filters. Member, IEEE and SIAM.

John A. Morrison, B.Sc., 1952, King's College, University of London; Sc.M., 1954, and Ph.D., 1956, Brown University; Bell Telephone Laboratories, 1956—. Mr. Morrison has done research in various areas of applied mathematics and mathematical physics. His recent interests have included stochastic differential equations and propagation in random media, electromagnetic scattering by raindrops, and the high-frequency propagation of surface waves. He was a visiting professor of mechanics at Lehigh University during the fall semester 1968. Member, American Mathematical Society, SIAM, Sigma Xi.

Erwin E. Muller, B.S., 1952, Stevens Institute of Technology; M.S., 1954, University of California; Bell Laboratories, 1954—. Mr. Muller has worked on design of ballistic missile guidance computers, single-sideband long-haul radio systems, and satellite communications systems. He is head of the Transmission Systems Characterization Department, concerned with describing the operational environment of radio and wire-pair transmission systems. Senior member, IEEE.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, and Ph.D. (E.E.), 1970, Rice

University; Optimal Data Corporation, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on problems related to filtering, guidance, and control for the space shuttle. At Bell Laboratories, he has worked on various aspects of signal processing. Member, Tau Beta Pi, Sigma Xi.

Thomas L. Osborne, B.S.E.E., 1961, M.S.E.E., 1963, Auburn University; Bell Laboratories, 1963—. Initially, Mr. Osborne was involved in research on microwave radio systems and related topics including microwave mixers, microwave integrated circuits, injection-locked oscillators, and rain attenuation. Since 1972, he has been supervisor of a group involved in circuit and system development of digital microwave radio systems. Member, Sigma Xi, Phi Kappa Phi, Tau Beta Pi, Eta Kappa Nu, Pi Mu Epsilon.

Lawrence R. Rabiner, S.B. and S.M., degrees, 1964, and Ph.D. (electrical engineering), 1967, Massachusetts Institute of Technology. Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques. Member, of Eta Kappa Nu, Sigma Xi, Tau Beta Pi; fellow, the Acoustical Society of America and IEEE.

Marvin R. Sambur, B.E.E., City College of New York, 1968; S.M., 1969, and Ph.D., 1972, Massachusetts Institute of Technology; Bell Laboratories, 1968–1977. Mr. Sambur was engaged in automatic speaker verification and automatic speech recognition research in the Acoustics Research Department. He is currently with the Defense Communications Division of ITT. Member, MPA-TC subcommittee on Speech Recognition and Understanding, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

David J. Thomson, B.S., 1965, Acadia University; MS, 1967, Ph.D., 1971, Polytechnic Institute of Brooklyn; Bell Laboratories, 1965—. Mr. Thomson has been involved in analysis of multipair and coaxial cable. He also worked on WT4 in measurement, analysis, and specification of geometric imperfections in addition to time series analysis and spectrum estimation techniques. Currently Mr. Thomson is working on the high-capacity mobile telephone system. Member, IEEE, IMS, SIAM.

Papers by Bell Laboratories Authors

BIOLOGY

^{31}P Nuclear Magnetic Resonance Studies O-Ehrlich Ascites Tumor Cells. G. Navon, S. Ogawa, R. G. Shulman, T. Yamane, Proc. Natl. Acad. Sci., *74* (January 1977), pp. 87–91.

CHEMISTRY

Picosecond Dynamics of Azulene. E. P. Ippen, C. V. Shank, R. L. Woerner, Chem. Phys. Lett., *46*, No. 1 (February 15, 1977), pp. 20–23.

Is Samarium Metal in an Intermediate Valence State? G. K. Wertheim, M. Campagna, Chem. Phys. Lett., *47* (1977), pp. 182–184.

The Radiofrequency Discharge Chemistry of Benzene and Mixtures with Helium, Argon, and Xenon. G. Smolinsky, M. J. Vasile, Int. J. Mass Spectrom. Ion Phys., *24* (1977), pp. 311–322.

The Nature and Effects of Platinum in Perovskite Catalysts. D. W. Johnson, Jr., P. K. Gallagher, G. K. Wertheim, E. M. Vogel, J. Catal., *48* (1977), pp. 87–97.

Oxidation of Ni_3Sa_4, Metastable SaNi, Ni_3Sa_2 and Ni_3Sa_1. H. G. Tomkins, J. E. Bennett, J. Electrochem. Soc., *124*, No. 4 (April 1977), pp. 621–623.

X-Ray Absorption Studies of Halide Binding to Carbonic Anhydrase. G. E. Brown, G. Navon, R. G. Shulman, Proc. Nat. Acad. Sci., *74* (May 1977), pp. 1794–1797.

ELECTRICAL AND ELECTRONIC ENGINEERING

Solid State Electronics: Scientific Basis for Future Advances. J. A. Giordmaine, Sci. Compendium, *195* (March 18, 1977), pp. 1235–1240.

Noise Performance of Microwave GaAs FET Amplifiers at Low Temperatures, R. E. Miller, T. C. Phillips, D. E. Iglesias, R. H. Knerr, Electron Lett., *13* (January 6, 1977), pp. 10–11.

Image Contour Extraction with Analog MOS Circuit Techniques: P. I. Siciu, D. A. Hodges, IEEE J. Solid State Circuits, *SC-12* (February 1977), pp. 65–72.

On Quantizers for DPCM Coding of Picture Signals. A. N. Netravali, IEEE Trans. Inform. Theory, *IT-23*, No. 3 (May 1977), pp. 360–370.

Geometrical Factors in Avalanche Punch-Through Erase. J. R. Brews, IEEE Trans. Electron Dev., *ED-24* (August 1977), pp. 1108–1116.

Simultaneous Phase Tracking and Detection in Data Transmission Over Noisy Dispersive Channels. F. R. Magee, Jr., IEEE Trans. Commun., *COM-25* (July 1977), pp. 712–715.

Rain-Scatter Interference on an Earth-Space Path. T. S. Chu, IEEE Trans. Ant. Propag., *AP-25* (March 1977), pp. 287–288.

DC Thermal Model of Semiconductor Device Produces Current Filaments as Stable Current Distributions. Hilding M. Olson, IEEE Trans. Electron Dev., *ED-24* (September 1977), pp. 1177–1184.

Laser Coding of Bipolar Read Only Memories. J. C. North, W. W. Weick, IEEE J. of Solid State Circuits, *SC-11*, No. 4 (August 1976), pp. 500–505.

Rank Reduction of Ill-Conditioned Matrices in Waveguide Junction Problems. Douglas N. Zuckerman, Paul Diament, IEEE Trans. Micro. Theory Tech., *MTT-25* (July 1977), pp. 613–619.

On Low-Sensitivity Realizations of Band-Elimination Active Filters. Renato N. Gadenz, IEEE Trans. Circuits Syst., *CAS-24*, No. 4 (April 1977), pp. 175–183.

MATERIALS SCIENCE

A Simplified High Frequency MOS Capacitance Formula. J. R. Brews, Solid State Electron., *20* (1977), pp. 607–608.

Giant Core-Exciton Effects on Si(111) 7 × 7 Surfaces. G. Margaritondo, J. E. Rowe, Phys. Lett., *59A* (January 10, 1977), pp. 464–466.

Near Infrared Sources in the 1–1.3 μm Region by Efficient Stimulated Raman Emission in Glass Fibers. C. Lin, L. G. Cohen, R. H. Stolen, G. W. Tasker, and W. G. French, Opt. Commun., *20*, No. 3 (March 1977), pp. 426–428.

Dispersion of Electronic Surface Resonances and Crystal Surface Structure. E. G. McRae, J. M. Landwehr, C. W. Caldwell, Phys Rev Lett, *38* (June 13, 1977), pp. 1422–1425.

Theoretical Modeling of the Simultaneous Exposure and Development (SED) Process of a Positive Photoresist. W. T. Tsang, Appl. Opt., *16* (July 1977), pp. 1918–1930.

Dependence of Residual Damage on Temperature During Ar^+ Sputter Cleaning of Silicon. J. C. Bean, G. E. Becker, P. M. Petroff, T. E. Seidel, J. Appl. Phys., *48* (March 1977), pp. 907–913.

The Loading Effect in Plasma Etching. C. J. Mogab, J. Electrochem. Soc., *124* (August 1977), pp. 1262–1268.

Demagnetization of the Ni(100) Surface by Hydrogen Absorption. M. Landolt, M. Campagna, Phys. Rev. Lett., *39* (August 29, 1977), pp. 568–570.

The Anomalous Behavior of $TiSe_2$ and the Excitonic Insulator Mechanism. J. A. Wilson, S. Mahajan, Commun. Phy., *2* (1977), pp. 23–29.

Computer Model of Spin Glasses. L. R. Walker, R. E. Walstedt, *38*, No. 9 (February 28, 1977), pp. 514–518.

The Copolymerization of Styrene with Sulfur Dioxide. Determination of the Monomer Sequence Distribution by Carbon-13 NMR. R. E. Cais, J. H. O'Donnell, F. A. Bovey, Macromolecules, *10*, No. 2 (March–April 1977), pp. 254–260.

A Model for the Formation of Oxidation Induced Stacking Faults In Silicon. S. Mahajan, G. A. Rozgonyi, D. Brasen, Appl. Phys. Lett., *30* (January 1977), pp. 73–75.

A Quantitative Model for the Diffusion of Phosphorus in Silicon and the Emitter Dip Effect. R. B. Fan, J. C. C. Tasi, J. Electrochem. Soc., *124* (July 1977), pp. 1107–1118.

A Method of Tungsten Dopant Deposition for Dual-Dielectric Charge-Storage Cells. J. R. Ligenza, D. Kahng, M. P. Lepselter, E. Labate, IEEE Trans. Electron. Dev., *ED-24*, No. 5 (May 1977), pp. 581–583.

The Tetragonal Deformation of the TiO_6 Octahedron in Ferroelectric $PbTiO_3$. J. G. Bergman, G. R. Crane, E. H. Turner, J. Solid State Chem., *21* (1977) pp. 127–133.

Absolute Configuration of Piezoelectrics: A Bibliography II. D. T. Hawkins, Ferroelectrics, *15* (1977) pp. 77–96.

PHYSICS

Relaxation of Tunneling States in Fused Silica. J. E. Graebner, Brage Golding, Bull. Amer. Phys. Soc., *22* (March 1977) p. 310.

Feedback Stabilization of Optically Levitated Particles. A. Ashkin, J. M. Dziedzic, Appl. Phys. Lett., *30,* No. 4 (February 15, 1977), pp. 202–204.

Polarization Rotation Effects in Atomic Sodium Vapor. P. F. Liao, G. C. Bjorklund, Phys. Rev., *15,* No. 5 (May 1977), pp. 2009–2018.

Condensation of Optically Excited Carriers in CdS: Determination of EHL Phase Diagram. R. F. Leheny, Jagdeep Shah, Phys. Rev. Lett., *38,* No. 9 (February 28, 1977), pp. 511–514.

Observation of the Two-Dimensional Plasmon in Silicon Inversion Layers. S. J. Allen, Jr., D. C. Tsui, R. A. Logan, Phys. Rev. Lett., *38,* No. 7 (April 25, 1977), pp. 980–983.

AC Stark Splitting of Two-Photon Spectra. J. E. Bjorkholm, P. F. Liao, Opt. Commun., *21* No. 1 (April 1977) pp. 132–136.

Critical Nuclear Magnetic Relaxation in a Strong Itinerant. M. Shaham, J. Barak, U. El-Hanany, W. W. Warren, Jr., Phys. Rev. Lett., *39* (August 29, 1977), pp. 570–574.

Torsional Waves in a Sapphire Rod at Low Frequencies. Robert N. Thurston, Lynn O. Wilson, IEEE Trans. Son. Ultrason., *SU-24* (September 1977), pp. 305–312.

Normal Photoemission Demonstration of Two and Three Dimensionality of Electron States in Layer Compounds. P. K. Larsen, M. Schluter, N. V. Smith, Solid State Commun., *21* (February 1977), pp. 775–778.

Measurement of the Fine Structure Splitting of the 4F State in Sodium Using Two-Photon Spectroscopy with a Resonant Intermediate State. P. F. Liao, J. E. Bjorkholm, Phys. Rev. Lett., *36,* No. 26 (June 28, 1977), pp. 1543–1545.

Bell Laboratories Scientist Named Nobel Laureate

Philip W. Anderson, consulting director in the Physical Research Division at Bell Labs, Murray Hill, and a Professor at Princeton University, has been named one of three joint recipients of the 1977 Nobel Prize in Physics.

Anderson shares the award from the Royal Swedish Academy of Sciences with John H. Van Vleck of Harvard University and Sir Neville Mott of Cambridge University, England.

According to the Academy, "The three prize winners are theoreticians within the field of solid state physics, the branch of physics which lies behind current technical developments, particularly in electronics."

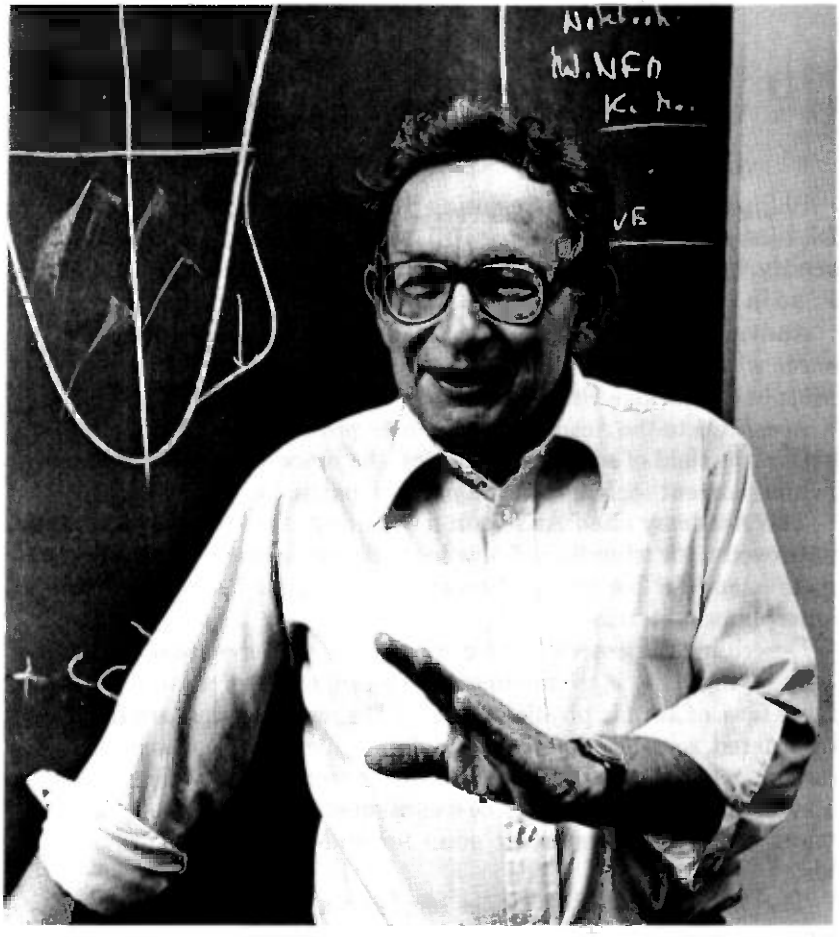
The Academy cited Anderson for his central role in developing an improved understanding of how local magnetic moments can occur in metals, and for his explanation of how electrons become localized in amorphous materials.

Describing the general nature of his work, Anderson said, "What I've been trying to do is to understand the properties of matter, given the basic laws of atomic physics. Many of the real materials around us are disordered, so-called amorphous materials like glass, in which the atoms have no regular arrangement. Many of the properties that we understand about solids are supposed to be a consequence of that regular arrangement. So we had no starting point for understanding disordered solids.

The work the Academy cites has been a starting point for understanding how electrons move in irregular solids. It consists of classifying two kinds of situations: One in which the electrons can move freely in the whole of the material and the other in which the electrons are pinned down, or localized, in one particular place.

Anderson's work resulted in a better understanding of why certain atoms such as iron are magnetic when dissolved in nonmagnetic host metals, why other atoms that might be expected to be magnetic are not, and why certain amorphous materials (such as glass) do not conduct electricity.

In 1958 Anderson published a paper in which he showed under what conditions an electron in a disordered system can either move through the system as a whole or be more or less tied to a specific position as a localized electron. This paper, according to the Royal Swedish Academy, "has become one of the cornerstones in our understanding of, among other things, the electronic conductivity of disordered systems." These



Philip W. Anderson in his office at Bell Laboratories, Murray Hill, N.J.

ideas, the Academy said, "have been experimentally verified and they have in this way laid the foundations for important technical developments."

Explaining further, Anderson said, "One example that particularly interests me is window glass; everyone knows that ordinary window glass is a good electric insulator. It's not a metal. It doesn't conduct electricity well and is used for insulators in power lines and things like that. If you look at the reasons in standard physics textbooks for why a substance like glass is an insulator, you won't find answers. This is because these materials all depend on the irregularity of the structure, and glass is a totally irregular structure. You need this concept of localization to understand something as simple as window glass being an insulator."

Anderson also was cited by the Academy for his contributions to the basic understanding of local magnetic phenomena. One emerging practical application of this theoretical work is increasing use of magnetic materials in telecommunications systems and commercial computers.

Explaining the relationship of his work to bubble technology, Anderson said, "I was part of the group many years ago that worked in magnetism at Bell Laboratories. The Bell Labs group was a codiscoverer of the garnets. Before that, I had formulated a theory which explained the kind of magnetism we have in the garnets, and certainly that set the stage for understanding these materials. There was even a magnetic material that was discovered as a consequence of my theory . . . . My work has almost always been to propose the theoretical background for work others do in developing technology."

Anderson will be the fifth Bell Labs scientist to be awarded the Nobel Prize in Physics. In 1937, Clinton C. Davisson shared the award for discovery of the wave nature of matter, which was vital to the subsequent development of modern physics and its impact on technology from atomic energy to the transistor. The Nobel Prize for the transistor was awarded in 1956 to John Bardeen, William H. Brattain, and William B. Shockley.