THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 51 February 1972 Number 2

Multipath Propagation at 4, 6, and 11 GHz

By W. T. BARNETT

(Manuscript received August 24, 1971)

Signals at 4, 6, and 11 GHz, transmitted over a 28.5-mile radio relay path in Ohio, were continuously monitored during the late summer of 1966. Previous publications have reported on the observed 4- and 6-GHz multipath fading statistics, and on the improvements available with space or frequency diversity. This paper presents data for the 11-GHz transmission, and, in combination with the earlier results, establishes an empirical frequency dependence for the amplitude statistics.

A general treatment of the relationships between the factors underlying multipath propagation is intractable. However, based on the results in this and other papers, a general relationship is given for the probability of deep multipath fading which is linear in frequency, cubic in path length, and varies with meteorological-geographical factors.

Temporal aspects of the Ohio data were also investigated at all frequencies, utilizing both a 1-hour and a 1-day clock time interval. It was found that the multipath fade time statistic can be described by a single parameter for either interval. A subset of the multipath fading hours was also analyzed using a 1-minute clock interval, with the result that the difference between the minute median fade and the hourly median fade is frequency independent, and normally distributed with a standard deviation of 5.5 dB.

I. INTRODUCTION

Although it is a relatively rare phenomenon, multipath propagation constitutes a fundamental limitation to the performance of microwave

radio systems. During a period of multipath propagation, the narrowband output from a single receiving antenna can be reduced to equipment noise levels for seconds at a time. Corrective measures such as frequency diversity or space diversity then must be introduced to provide satisfactory commercial operation.[1-3] Propagation data required for economical system design and detailed performance estimates were not available prior to 1966. To fill this need an extensive experimental program was undertaken on a typical radio relay path in Ohio. Previous studies[1,2,4] have reported on the amplitude statistics obtaining during multipath propagation at 4 and 6 GHz, both with and without frequency or space diversity. Data were also obtained for a single frequency in the 11-GHz band. The multipath fading data for this signal have now been analyzed and statistics for the total time faded (P), the number of fades (N), their average duration ($\bar{t}$), and the fade duration distribution are presented in Section IV as functions of fade depth.

Multipath propagation is by its very nature dependent upon the operating microwave frequency; the variation of the fading characteristics with frequency has been considered by many investigators[5,6] with controversial results.* This is not surprising considering the time-variant, nonstationary behavior of the phenomena. However, the data obtained in Ohio were extensive enough to give statistical stability which, with the 11:6:4 frequency sampling, allows a meaningful comparison in Section V of P, N, and $\bar{t}$ as functions of frequency.

It is clear that a great deal is known about one path in Ohio. Generalization of these results to other paths requires an underlying theory. The experimental data show that P, N, and $\bar{t}$ can be quite closely represented by simple, one-term algebraic functions of fade depth. This agrees with predictions by S. H. Lin[7] based on analysis of a simple and plausible analytic model for multipath fading. It is therefore reasonable to assume that the variation of P, N, and $\bar{t}$ with fade depth for all paths subject to multipath fading will have the same functional dependence as did the Ohio path. A general formulation which includes the most important path parameters is proposed in Section VI for the coefficient in the equation relating the total time faded and the fade depth during the so-called worst fading month. This estimate provides necessary information for microwave radio system design in the continental U.S.A.

The intensity of multipath fading varies greatly, even during the normally active summer months; during some days there will be exten-

* Reference 5 gives many references on multipath fading investigations.

sive multipath fading, while on others there will be none. Statistics for time bases shorter than a month—or the entire 68-day period for the test reported here—are also of interest. The time faded characteristic was studied for the 4-, 6-, and 11-GHz signals on both a daily (24-hour) and an hourly basis. Section VII concludes with a study of minute-by-minute variations within an hour for a subset of the multipath fading hours.

All the experimental results mentioned in the preceding paragraphs were obtained from a data base comprising all the time intervals with deep multipath fading.[2] In sum, these intervals were about 15 percent of the total measurement time. The P, N, and $\bar{t}$ statistics for the remaining 85 percent of the time are given in Section VIII for typical 4- and 6-GHz signals. The 11-GHz data for this interval were not included because of the difficulty in identifying rain attenuation data; meteorological measurements were not made in conjunction with this experiment.

II. SUMMARY

Highlights of the results detailed in Sections IV thru VIII are given in this section. A few definitions are needed first:

L: Normalized algebraic value of envelope voltage (fade depth in dB $= -20 \log L$)
P: Fraction of time T that the envelope voltage is $\leqq L$
N: Number of fades (during T) of the envelope voltage below L
t: Duration of a fade below L in seconds ($\bar{t}$ = average duration)
f: Frequency in GHz
D: Path length in miles

The major results are:

(*i*) The 11-GHz amplitude statistics for the data base interval (T) of 5.26×10^6 seconds and for fade depths exceeding 15 dB are $P = 0.69L^2$, $N = 12{,}300L$, $\bar{t} = 330L$. Also $t/\bar{t}$ is log-normal and independent of L with 1 percent of the fades at any level longer than ten times the average.

(*ii*) The P and N statistics for the 4- , 6- , and 11-GHz data are, within experimental error, linear functions of frequency given by $P = 0.078fL^2$ and $N = 1000fL$. The comparable $\bar{t}$ statistic is given by $\bar{t} = 410L$.

(*iii*) An empirical estimate of P for the worst fading month is

$$P = rL^2, \qquad L \leqq 0.1$$

where r is defined as the multipath occurrence factor and is given by

$$r = c\left(\frac{f}{4}\right)(D^3)(10^{-5})$$

with

$$c = \begin{cases} 1 & \text{average terrain} \\ 4 & \text{over-water and Gulf Coast} \\ 0.25 & \text{mountains and dry climate.} \end{cases}$$

(*iv*) Of the days in the 1966 Ohio data base, about 12 had more fading than the average while 54 had less. The worst day contained about 48 percent of the total fade time at or below 40 dB while the worst hour contained some 20 percent.

(*v*) The simple model, $P = aL^2$, can be used to characterize shorter periods with multipath fading.* The cumulative empirical probability distribution (c.e.p.d.) with $a \equiv a_d$ is for daily fading

$$\Pr(a_d \geqq A) \cong \exp[-1.2\sqrt{A(4/f)}]$$

and for the hourly fading with $a \equiv a_h$

$$\Pr(a_h \geqq A) \cong \exp[-0.7\sqrt{A(4/f)}].$$

The hourly median fade depth value exceeded by 1 percent of the hours is 18 dB below free space.

(*vi*) The random variable defined as the difference between the median for a minute in a fading hour and the median for the entire hour was found to be normally distributed with zero mean and a standard deviation of 5.5 ± 1.5 dB.

III. EXPERIMENTAL DESCRIPTION[2]

The data presented were obtained by the MIDAS† measuring equipment at West Unity, Ohio. The basic data consist of measurements of the received envelope voltages of standard TD-2 (4 GHz), TH (6 GHz), and TL (11 GHz) signals; Table I is a list of the center frequencies of each channel. A functional block diagram is shown on Fig. 1. The 4-GHz and 6-GHz channels were standard in-service FM radio channels with nominally constant transmitted power (±0.5 dB). The 11-GHz

* The change in the coefficient from r to a is made to clearly differentiate between the total measurement period and the daily (or hourly) epoch.

† An acronym for Multiple Input Data Acquisition System.

TABLE I—RADIO CHANNELS MEASURED AT WEST UNITY, OHIO

Channel No.*	Frequency (MHz)	Antenna	Polarization
4-7	3750	Horn Reflector	V
4-1	3770		H
4-8	3830		V
4-2	3850		H
4-9	3910		V
4-11	4070		V
4-6	4170		H
6-11	5945.2		H
6-13	6004.5		H
6-14	6034.2		V
6-15	6063.8		H
6-17	6123.1		H
6-18	6152.8		V
11-1	10995		V

* The 4-X channels correspond to standard TD-2 radio system signals; 6-X corresponds to TH; 11-1 corresponds to TL.

channel was added especially for the test program and was unmodulated, with the RF equipment housed in an outdoor cabinet.

West Unity, Ohio, was chosen as the site for this experiment because it is part of a major cross-country route in an area known to suffer multipath fading. The hop monitored was of typical length—28.5

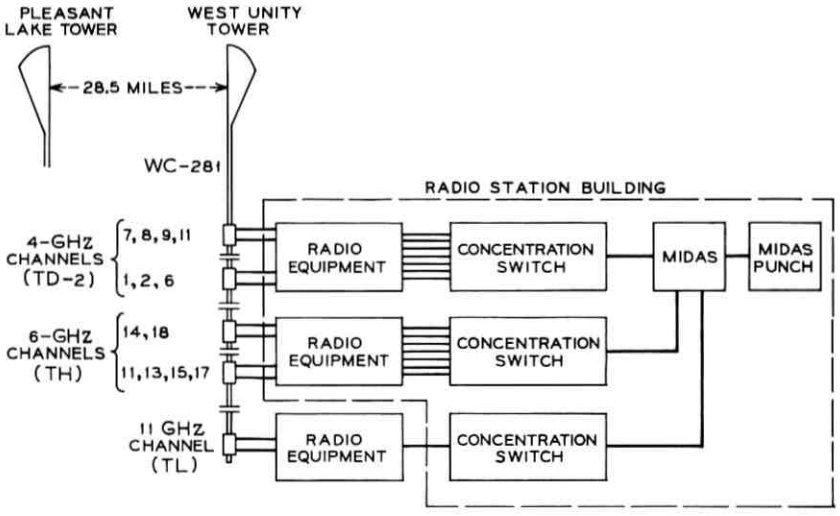


Fig. 1—1966 experimental layout, Pleasant Lake–West Unity.

miles—with negligible ground reflections. The path clearance was adequate even for the extreme of equivalent earth radius (k) equal to two-thirds, as shown on the path profile in Fig. 2. It is believed that this path is typical of those inland paths subject to multipath fading conditions.

The MIDAS equipment sampled each signal five times per second, converted each measurement to a decibel scale, and recorded the data in digital form for subsequent computer processing (in the absence of fading the recording rate was less than the sample rate). Further equipment details are given in Ref. 2.

The data were obtained during the period from 00:28 on July 22 to 08:38 on September 28, 1966. The total elapsed time was 5.9×10^6 seconds of which 5.26×10^6 seconds was selected for the data base; the balance was unusable mainly because of maintenance of the radio equipment or MIDAS. Within the data base, 7.8×10^5 seconds contained all the multipath fading in excess of approximately 10 dB. The balance of the time, 4.48×10^6 seconds, was categorized as nonfading time.

A natural epoch for multipath fading is the 24-hour period from noon to noon. It was convenient to number these periods from 1 to 69 starting at noon on July 21 and ending at noon on September 28. Here the

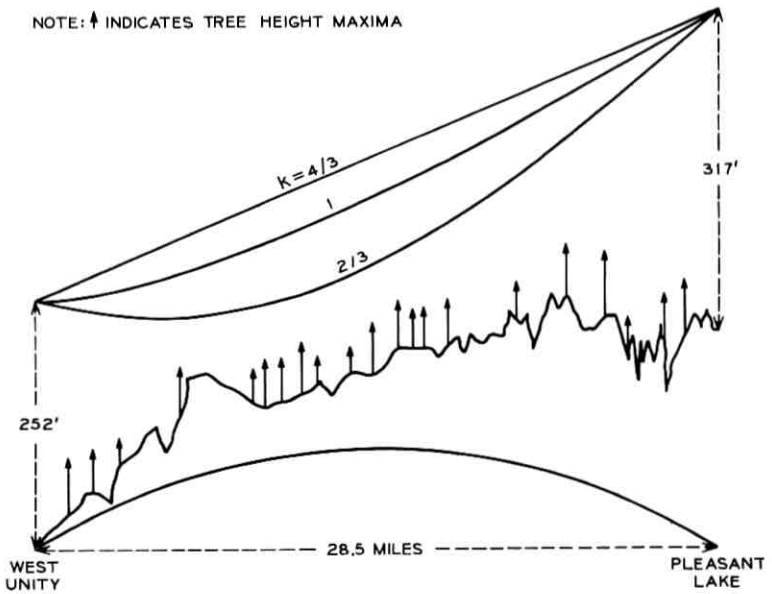


Fig. 2—West Unity–Pleasant Lake path profile.

missing end periods from 12:00, July 21 to 00:28, July 22 and 08:38 to 12:00 on September 28 have been assumed negligible. Most of the multipath fading was found to occur in the period between midnight and 9 A.M. as will be discussed later. These latter time periods were, for all practical purposes, subject to continuous measurement for 66 of the 69 periods. Thus, we reduce the multipath fading data base to 66 nine-hour periods. These were used for channel characterization and for investigating the daily and hourly statistical properties of multipath fading.

All fading distributions will be given in terms of the received voltage relative to the midday normal in dB. The rms variation in the dB reference level was estimated as ± 0.8 dB.[2]

IV. 11-GHz MULTIPATH RESULTS

The 11-GHz data were analyzed in terms of the statistical properties previously reported for the 4- and 6-GHz data.[2,4] These were (*i*) the fraction (P) of 5.26×10^6 seconds that the signal was faded below a given level L, (*ii*) the number of fades (N) below L, (*iii*) the average duration in seconds ($\bar{t}$) of fades below L, and (*iv*) the fade duration distribution. The data were carefully inspected to insure that only multipath fading was included and that rain fading was excluded. This was done by inspection of signal level vs time plots with the determination made by the frequency of the fading and by comparison with the 4- and 6-GHz data. As in the case of the 4- and 6-GHz data, we were most interested in fades greater than 15 dB. However, reliable data for the 11-GHz signal were limited to fade depths of 35 dB because the reference level of received signal strength was 5–10 dB lower than that for the 4- and 6-GHz signals.

The data for the fractional fade time are given in Fig. 3. They are adequately represented by a straight line whose equation is $P = 0.69L^2$. The data for the number of fades are given in Fig. 4 along with the fitted line $N = 12{,}300L$. The data for the average fade duration are obtained from the ratio of the total time faded to the number of fades and are given in Fig. 5 along with the fitted line $\bar{t} = 330L$. These variations of P, N, and $\bar{t}$ with L are in agreement with those previously found for the more extensive 4- and 6-GHz data and are as predicted from a mathematical model of the multipath fading process.[7]

The probability that a fade of depth $-20 \log L$ dB lasts longer than t seconds, i.e., the fade duration distribution, can be estimated by dividing the number of fades of depth L and duration t seconds or longer by the total number of fades of depth L. A normalization is made

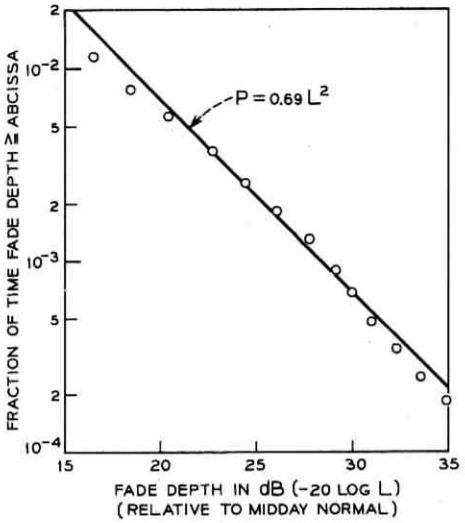


Fig. 3—11-GHz fade depth distribution, 1966 West Unity.

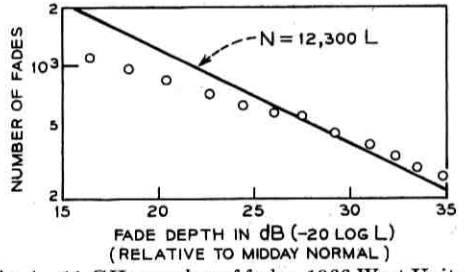


Fig. 4—11-GHz number of fades, 1966 West Unity.

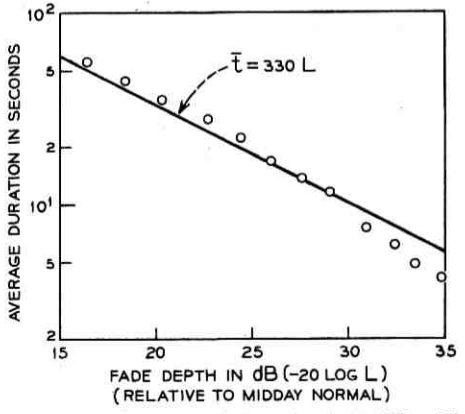


Fig. 5—11-GHz average fade duration, 1966 West Unity.

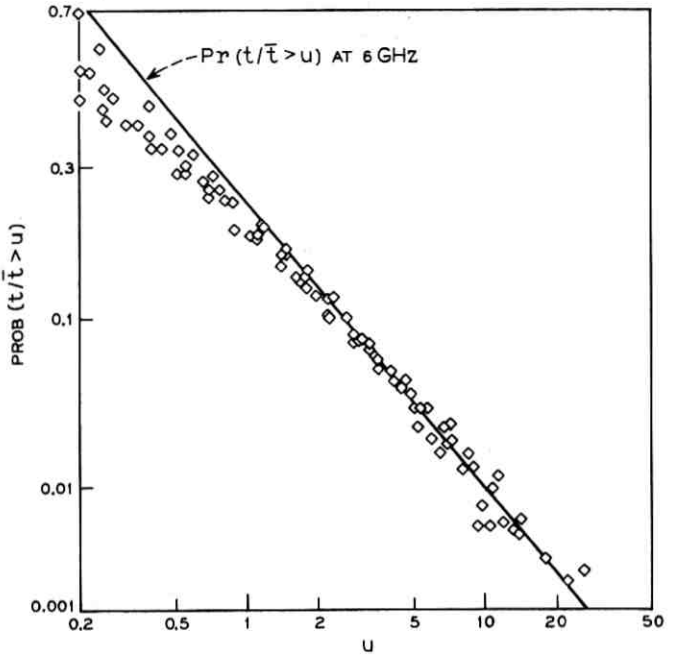


Fig. 6—11-GHz fade duration distribution: probability that the fade duration, normalized to its mean for a given fade depth, is longer than a given number. Data pooled for all fade depths greater than 10 dB.

with respect to the average fade duration. The 11-GHz data are plotted on Fig. 6, using a normal probability scale, for all fades $\geqq 10$ dB. The data indicate that $t/\bar{t}$ is independent of L and that the probability is approximately log normal with 1 percent of the fades being longer than ten times the average fade duration. The line on Fig. 6, taken from Fig. 40 of Ref. 4, represents the fade duration distribution for the corresponding 6-GHz data. Thus, the fade duration distributions, when properly normalized, appear to be invariant with frequency.

V. MULTIPATH EFFECTS AS A FUNCTION OF FREQUENCY

The 11-GHz results of Section IV can be combined with those previously obtained for 4 and 6 GHz[2,4] to obtain an estimate of the variation of the characteristics with microwave frequency. This treatment is valid because all the data were obtained under identical conditions: same path, same antennas,* and same time period.

* The different beamwidths of the horn reflector for the three frequencies play a minor role because the variations in angle-of-arrival of the multipath components are generally less than the smallest beamwidth, which is ± 0.6 degree at 11 GHz.

TABLE II—MULTIPATH FADING CHARACTERISTICS
($L \leqq 0.1$)

Freq (GHz)	P	N	$\bar{l}$
4	$0.25L^2$	$3670L$	$408L$
6	$0.53L^2$	$6410L$	$490L$
11	$0.69L^2$	$12300L$	$330L$

Table II summarizes the 4-, 6-, and 11-GHz results. The tabulated coefficients incorporate the effects of the environment and frequency. Plotting them versus frequency (as in Fig. 7) allows us to observe that the N and P coefficients increase, within experimental error, linearly with f while $\bar{l}$ is longer at 6 GHz and shorter at 11 GHz with respect to 4 GHz. Based upon these data, an approximation that $\bar{l}$ is independent of f is reasonable. The functional dependence is described by:

$$P = 0.078fL^2, \tag{1}$$

$$N = 1000fL, \tag{2}$$

$$\bar{l} = 410L, \tag{3}$$

with f in GHz.

The deviation of the P and N coefficients of Table II from these empirical equations is less than ± 1 dB which is within the bounds of experimental error.[2] The $\bar{l}$ coefficients agree with equation (3) within ± 2 dB. This is satisfactory since the $\bar{l}$ data were originally obtained as the ratio of the P and N data at each fade level; ± 1 dB variation each in P and N corresponds to ± 2 dB variation in $\bar{l}$.

The multiple transmission paths which give rise to the fading effects are generated by irregularities in the refractivity gradient in the volume defined by the beamwidths of the two antennas. As the relative path lengths vary with time the composite received signal may fade due to destructive interference (or be enhanced by constructive interference). It is easy to see that a given change in relative path length will cause more signal variations at higher frequencies because of the proportionally larger phase variations; we have found that the effect in Ohio in 1966 was linear. There is no apparent reason why this variation with frequency does not generally apply for multipath fading for a normal overland path engineered in standard fashion. Also, a linear variation of P with frequency has been theoretically predicted by C. L. Ruthroff[6] from a careful analysis of a simple physical model of multipath fading.*

* The results discussed here predate Ruthroff's analysis.

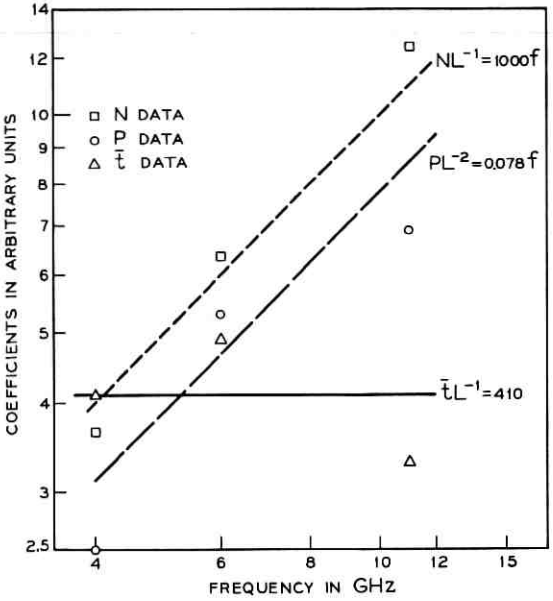


Fig. 7—Coefficients of P, N, and $\bar{t}$ characteristics versus frequency.

VI. OCCURRENCE OF MULTIPATH FADING

6.1 *General*

It is well known that the time (probability) distribution of the envelope of a microwave signal subject to multipath fading depends upon path length, path geometry, terrain clearance, type of terrain, and meteorological conditions in a complex manner. A general treatment of these relationships is intractable. However, based on the results discussed in previous sections and in other papers, an engineering estimate (incorporating the most important factors) of the fade depth distribution can be made for typical microwave paths for the heavy fading time of the year, i.e., the so-called worst month fading. In the results that follow adequate path clearance and negligible ground reflections are assumed.

6.2 *Relation to the Rayleigh Distribution*

Quite often in propagation studies it is assumed that the probability distribution of the envelope (v) of the received signal is given by the Rayleigh formula

$$\Pr(v < L) = 1 - e^{-L^2} \cong L^2 \quad \text{for} \quad L < 0.1. \tag{4}$$

One physical basis of this distribution is the limiting case of the envelope of an infinitely large number of equal amplitude signals of the same frequency, but random phase. Since this is a good approximation in many situations, e.g., tropospheric and mobile radio propagation, this distribution has seen much use. In the case of line-of-sight microwave radio, this is not a good assumption and the distribution is not directly applicable. From Table II the results for the fade depth distribution P vary as L^2 but with different coefficients.* The coefficient is generally not fixed, but depends upon the time base of the data, and upon the particular path parameters. The path parameters can be incorporated in the coefficient by expressing the multipath fade depth distribution as

$$\Pr(v < L) = rL^2 \qquad L < 0.1 \tag{5}$$

where r is defined as the multipath occurrence factor; $r = 1$ is appropriate to the Rayleigh distribution.

6.3 *Path Parameters*

As discussed in Section V, r is directly proportional to frequency; terrain and distance effects have to be incorporated. An engineering estimate for r can be given as a product of three terms†

$$r = c\left(\frac{f}{4}\right)D^3 10^{-5} \tag{6}$$

where: f is frequency in GHz,
D is the path length in miles,

$$c = \begin{cases} 1 & \text{average terrain} \\ 4 & \text{over-water and Gulf Coast} \\ 0.25 & \text{mountains and dry climate.} \end{cases}$$

The terrain effects and the distance dependence are based on applicable (albeit meager) Bell System data, most of which was acquired at 4 GHz on paths of 20–40 miles length. The plot given on Fig. 8 extends beyond this range. Indeed it can be argued that the curves should become parallel to the abscissa as D decreases (no multipath fading for paths sufficiently short[6]) and parallel to the ordinate (saturation) as D increases.

* An analysis of a mathematical model for multipath fading shows that the deep fade region of the distribution will be proportional to L^2 under very general conditions (Ref. 7).

† This empirical result for r is partially supported by British data as reported by K. W. Pearson[8] and is similar to a concise result reported by S. Yonezawa and N. Tanaka.[9]

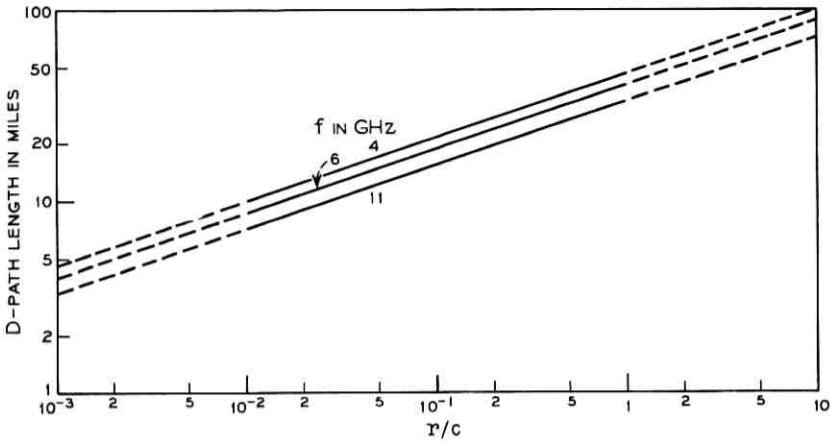


Fig. 8—Worst month multipath fading: $P = rL^2 = c(f/4)\,D^3L^2$.

The plotted values are certainly upper bounds for either extreme. The fD^3 dependence has been theoretically obtained by Ruthroff.[6]

The engineering estimate, equation (6), indicates that on a path of above average length, maintenance of the per-hop fading outage usually obtaining requires compensation for the additional free-space loss ($\propto D^2$) and for increased multipath ($\propto D^3$), which combine to impose a D^5 (15 dB/octave) length dependence.

VII. TIME CONCENTRATION OF DEEP MULTIPATH FADING

7.1 *Introduction*

The results and estimates already given utilize the entire data base, thus averaging temporal effects. It is well established that multipath fading occurs most often at night, with a few nights experiencing considerably more fading than most of the others. Describing this variability statistically is the objective here. We consider the fade time statistic for hourly and for daily periods and the median fade depth during an hour or a minute.

The analysis includes data from four fade depth values,* 9.8 dB, 20.4 dB, 31 dB, and 40.1 dB (henceforth labeled as levels 1 through 4). At each fade depth and for each analysis period the fade time for the seven 4-GHz channels was arithmetically averaged, as was that for the six 6-GHz channels. The fade time for the 11-GHz channel was used

* The unusual numbers are the result of quantization and calibration.[2]

TABLE III—FADE TIME DATA
(Seconds at or Below Given Fade Depth)

Freq Band (GHz)	Fade Depth 1 (9.8 dB)	2 (20.4 dB)	3 (31 dB)	4 (40.1 dB)
4	148,427	13,771	1329	135
6	259,933	27,503	2562	312
11	243,977	32,232	2982	*

* No data was obtained at 11 GHz for fade depth 4; see Section IV for further details.

directly. The resulting data will be referred to as the 4-, 6-, and 11-GHz fade times respectively. The fade time totals for the entire test period (5.26×10^6 seconds) are given in Table III.

7.2 *Distribution by Days—Rank Order Data*

The fade times for fade depths 1–4 were separately compiled for each of the 66 noon-to-noon periods. As expected there is considerable variation. As an example, Fig. 9 shows a plot of the 6-GHz fade time versus day number. Here the value plotted is the ratio of the fade time for the day to the total fade time, given in Table III, for a fixed fade depth. Much of the deep fading (levels 2, 3, 4) occurred on days 10,

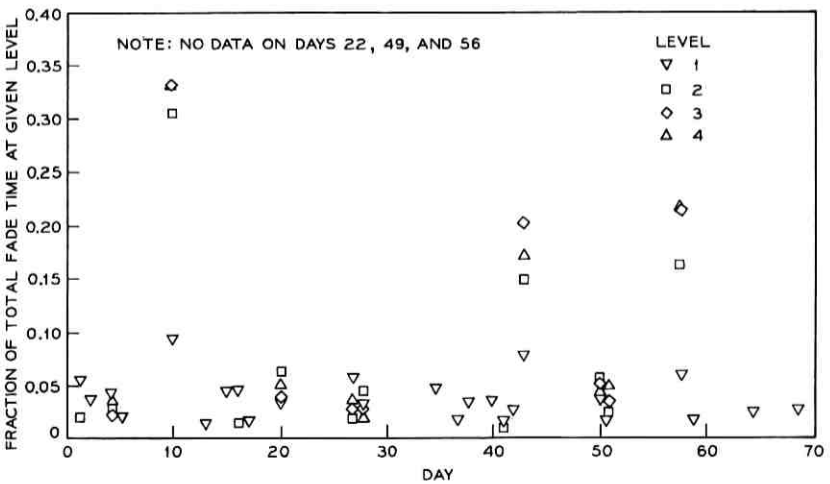


Fig. 9—Daily variation of multipath fading at 6 GHz, 1966 West Unity.

43, and 58 while the fading time for level 1 was more widely distributed.

The days were separately rank ordered for each frequency and each level with the variable again being the fraction of the total fade time; the results are given on Figs. 10a–c.* A few observations from these plots:

(*i*) The worst day fraction increases with level.
(*ii*) The data for level 1 do not fall off as rapidly with rank order as for levels 2–4.
(*iii*) Long tails in the rank order are prevalent.

Some of the more pertinent statistics are summarized in Table IV.

As already noted from Fig. 9, the bulk of the deep fading occurred on three days (10, 43, 58). The fraction of the total fading at the sample levels summed for these three days ranges from 0.55 to 0.74. Day 10 was the worst day in all cases. It appears that if a day suffered extensive 20-dB fading it also suffered 30- and 40-dB fading, but this indicator is not valid for 10-dB fading. In fact, about two-thirds of the days had 10-dB fading while only one-third had some 40-dB fading.

The statistical worst night is of particular interest. Figure 11 is based upon the observation that the worst day fraction increases with fade depth. The data points are fairly consistent except for levels 3 and 4 at 6 GHz which, for some unknown reason, do not show the expected increase relative to level 2. The line on Fig. 11 can be used as an estimate of the worst day fraction as a function of level. This estimate predicts that for systems with 40-dB fade margins the worst day will have 48 percent of the total fading within the worst month.†

A different perspective on the daily fading time can be obtained from Figs. 12a–c, which replot the rank order data on a logarithmic scale which has the effect of emphasizing the tail behavior. Generally, the tail is longer for lesser fade depths. It is interesting to compare these data with the result that would obtain for a uniform fade time distribution: a horizontal line at 0.015 (1/66). This line intercepts the level 2, 3, 4 data in the range of 10–15 days which means that this number of days had more fading than the average for the entire period while some 51–56 days have less fading. We shall return to the daily data in a later section where we shall see that they can be reduced to a more

* The data were plotted for all the days such that the cumulative sum of the plotted fade times just exceeded 99 percent of the total; note change of scale at rank order day 5.

† Here we take our statistics as representative of the worst month, the argument being that our results for a late summer—early fall period are generally comparable to the so-called worst fading month in a year.

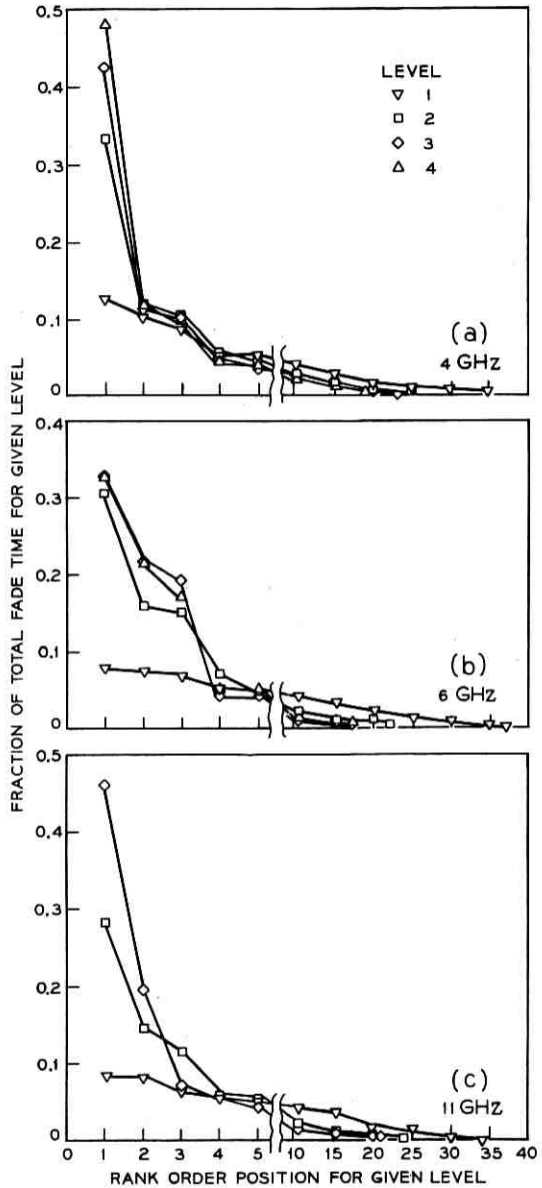


Fig. 10a—Rank order of 4-GHz daily fade times, 1966 West Unity.

Fig. 10b—Rank order of 6-GHz daily fade times, 1966 West Unity.

Fig. 10c—Rank order of 11-GHz daily fade times, 1966 West Unity.

TABLE IV—DAILY FADE TIME STATISTICS

Freq (GHz)	Level	Number of Days With Fade Time >0	Fraction of Total Fade Time		Number of Days to Give 0.99 of Total
			Worst Day	Sum of 3 Worst Days	
4	1	46	0.125	0.31	35
	2	36	0.33	0.56	25
	3	30	0.43	0.64	23
	4	26	0.48	0.70	19
6	1	46	0.077	0.22	37
	2	35	0.30	0.61	22
	3	27	0.33	0.73	17
	4	24	0.33	0.71	16
11	1	43	0.083	0.22	34
	2	35	0.29	0.55	24
	3	30	0.47	0.74	21

meaningful form given the appropriate statistical treatment and mathematical modeling.

7.3 *Distribution by Hours—Rank Order Data*

The preceding treatment on daily fade time is repeated here for hourly fade time. This fade time is expressed as a fraction of all time during the entire measurement period as given in Table III. Of course, greater scatter can be expected in the hourly data than in the daily data.

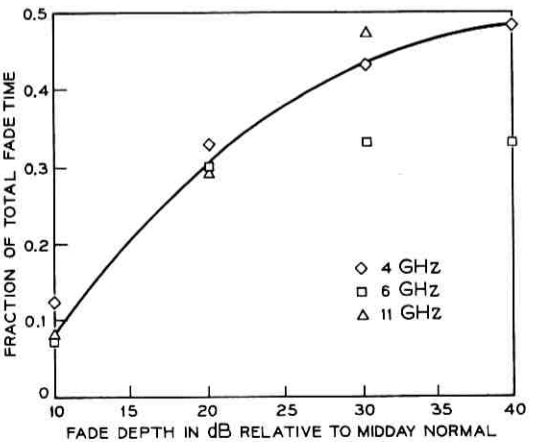


Fig. 11—Fraction of total fade time in worst night, 1966 West Unity.

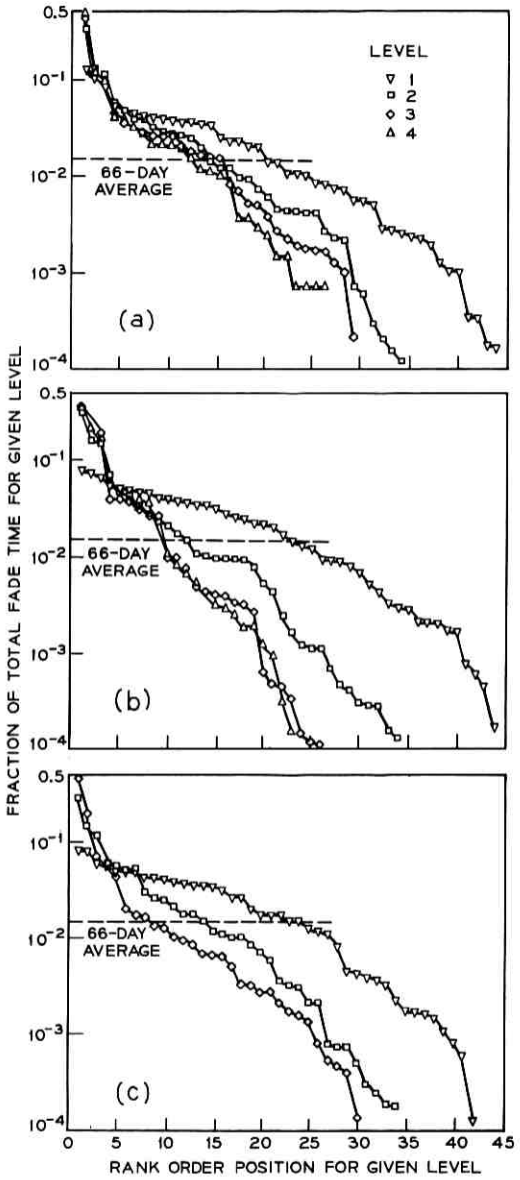


Fig. 12a—4-GHz rank order of daily fade times, 1966 West Unity.

Fig. 12b—6-GHz rank order of daily fade times, 1966 West Unity.

Fig. 12c—11-GHz rank order of daily fade times, 1966 West Unity.

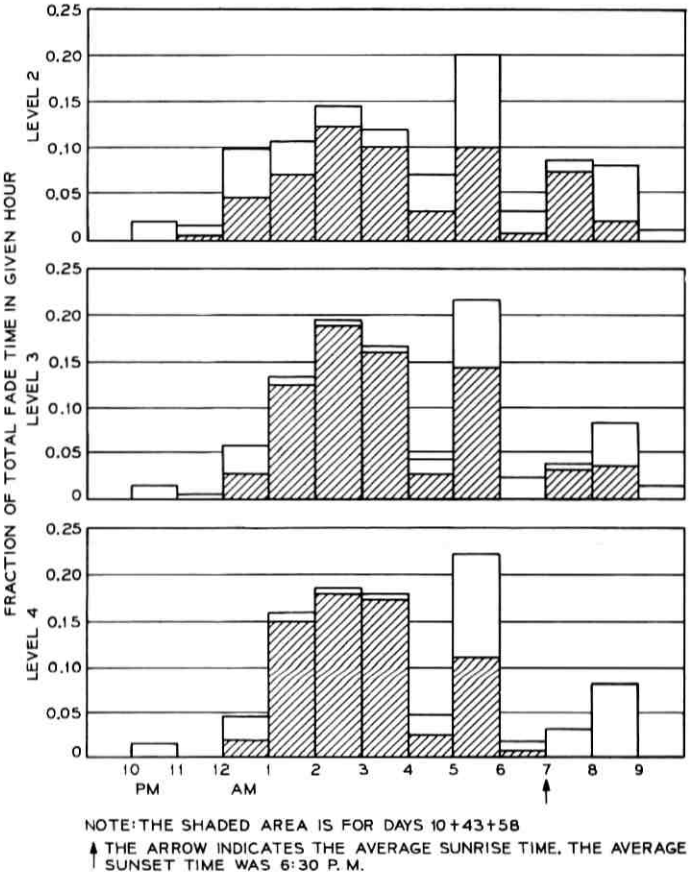


Fig. 13—6-GHz hour-of-day ranking, 1966 West Unity.

Figure 13 shows the distribution of fading for levels 2, 3, and 4 for the 6-GHz channels as a function of the hour of the day. Deep fading was generally within a 9-hour period between 12 P.M. and 9 A.M. The hours were rank ordered by level within a particular frequency band as shown on Figs. 14a–c. The general observations that can be made are similar to the "days" case:

(*i*) The worst night fraction increases with fade depth.
(*ii*) The level 1 fraction does not fall off very rapidly.
(*iii*) Long tails are even more prevalent than for daily fading.

Some of the pertinent statistics are summarized in Table V.

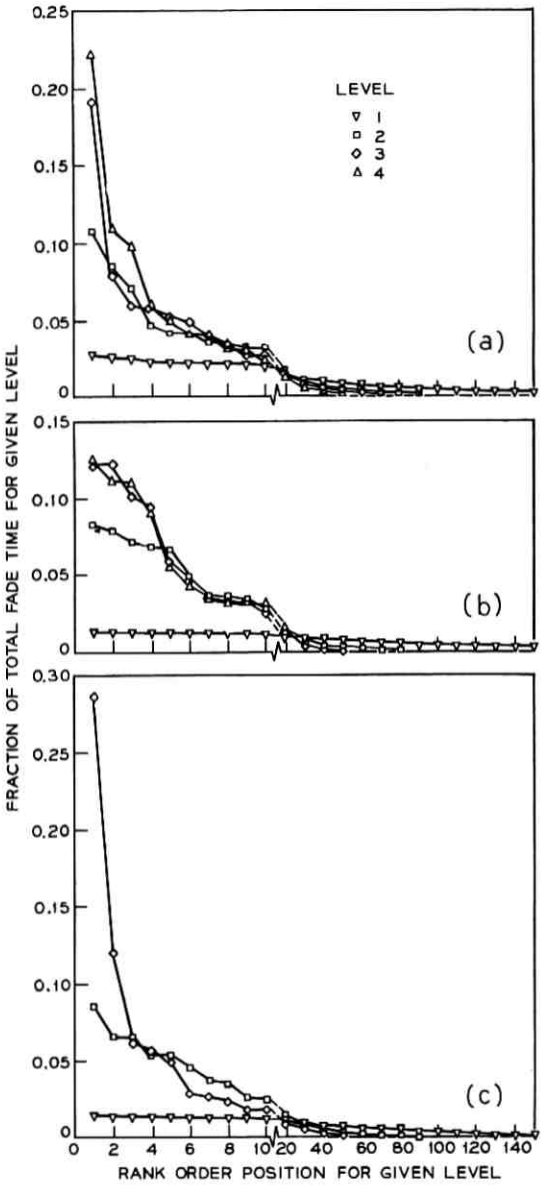


Fig. 14a—4-GHz rank order of hourly fade times, 1966 West Unity.

Fig. 14b—6-GHz rank order of hourly fade times, 1966 West Unity.

Fig. 14c—11-GHz rank order of hourly fade times, 1966 West Unity.

TABLE V—HOURLY FADE TIME STATISTICS

Freq (GHz)	Level	Number of Hours	Fraction of Total Fade Time		Number of Hours to Give 0.50 0.90 0.99 of Total Fade Time		
			Worst Hour	10 Worst Hours			
4	1	220	0.027	0.226	33	103	163
	2	117	0.107	0.525	10	48	80
	3	88	0.192	0.621	7	36	69
	4	61	0.222	0.699	5	24	47
6	1	259	0.014	0.128	50	138	206
	2	123	0.083	0.559	9	39	83
	3	78	0.123	0.681	5	24	49
	4	56	0.126	0.672	6	22	43
11	1	230	0.015	0.136	48	127	189
	2	121	0.085	0.495	11	46	88
	3	70	0.286	0.694	4	28	53

The worst hour for each transmission band is plotted versus fade depth in Fig. 15. The data spread is greater than for the days case (Fig. 11) with 6 GHz again exhibiting the least variation. The line on Fig. 15 can be used as an estimate of the worst hour fraction as a function of level. Thus, the worst day (Fig. 11) and worst hour (Fig. 15) estimates

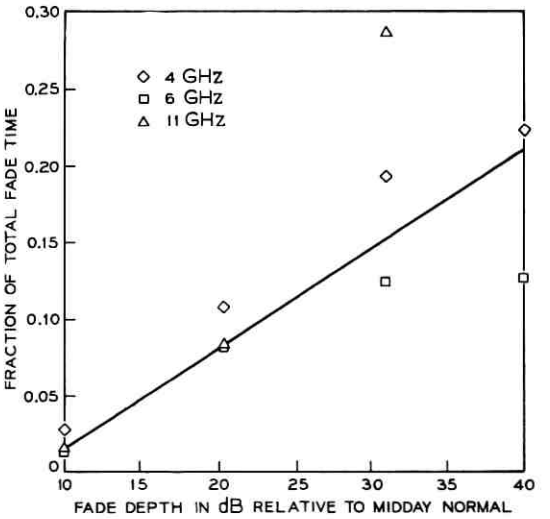


Fig. 15—Fraction of total fade time in worst hour, 1966 West Unity.

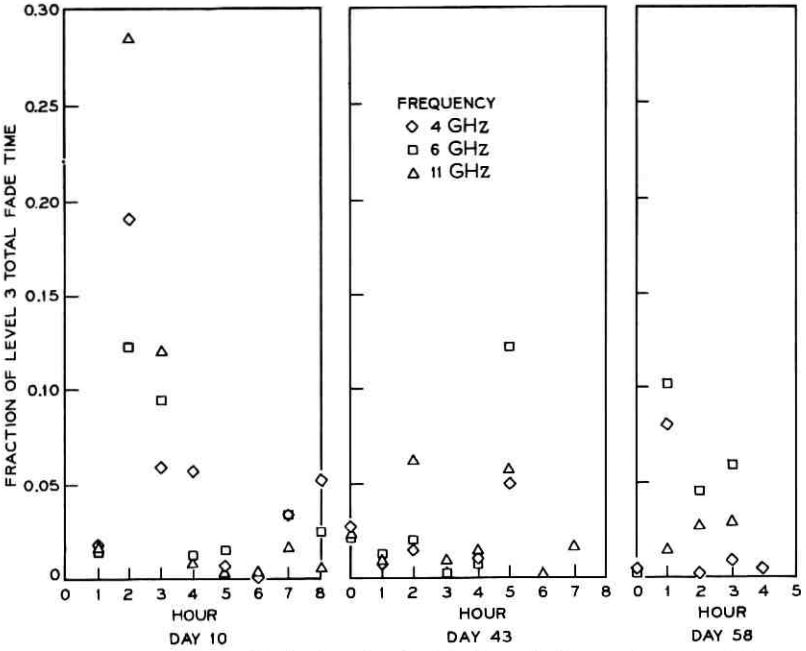


Fig. 16—Fraction of total fade time: level 3 hourly variation on three worst days.

for 40 dB predict 48 percent of the worst month multipath in a single day with 21 percent in the worst hour.

Days 10, 43, and 58 merit special study since they contain a majority of the deep fade time. The hourly variation in fade time for level 3 is given on Fig. 16. It is obvious from these data that there is no fixed relation between the frequency bands on an hourly time scale.* The hour from 2 A.M. to 3 A.M. on day 10 was the worst hour with the fractional fade time ranking with frequency as 11–4–6. However, the hour from 5 A.M. to 6 A.M. on day 43 was also a bad one with the fractional fade time ranking with frequency as 6–11–4. On day 58 the hour from 1 A.M. to 2 A.M., which was also outstanding, had the frequency order 6–4–11. However, the overall statistics show that fading severity increases with frequency.

7.4 *Hourly Median for a 4-GHz Channel*

The data reported in previous sections were in terms of the fraction of time that some fixed fade depth was exceeded; a reversal of these

* This conclusion does not change if absolute fade time is used instead of fractional fade time.

roles is equally valid. The variable examined in this section is the fade depth exceeded for a total of 30 minutes in an hour (hourly median). Figure 17 shows a rank order of the hourly median data for one of the 4-GHz channels as obtained directly from the experimental data for each hour. This particular channel is considered typical. The worst hourly median was 20.5 dB below free space and some 10 hours had hourly medians in excess of 15 dB. The general tendency is quite regular and shows a slowly decreasing median value with 120 hours experiencing hourly median fades in excess of 5 dB.

7.5 *Analytic Model for Hourly Median*

The single-channel fade depth statistics have a common characteristic: the fractional probability that the signal v is at or below L is proportional to L^2 (see Table II). Lin[7] has shown that this is a general property of fading signals under very general conditions, i.e.,

$$P \equiv \Pr(v \leqq L) = aL^2 \equiv \frac{t_L}{T} \qquad L \leqq 0.1 \tag{7}$$

where a is an environmental constant and t_L/T is the fractional fade time for the time period T.

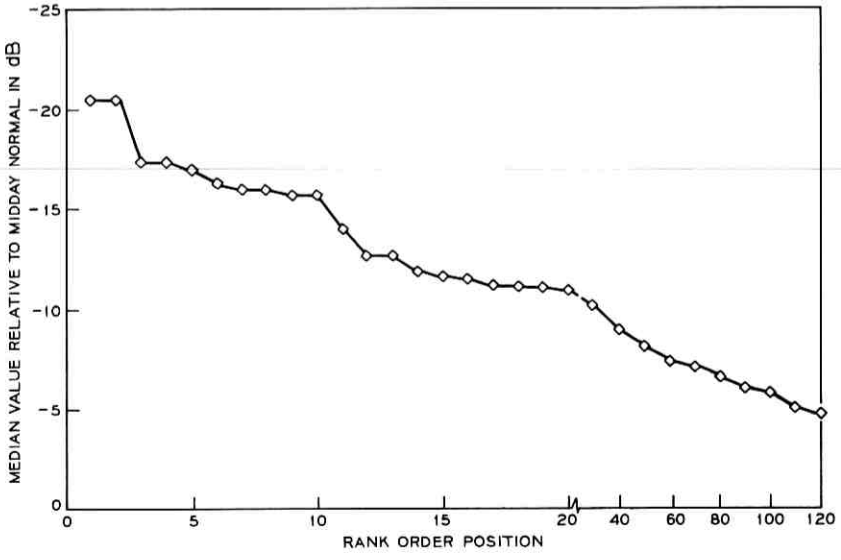


Fig. 17—4-GHz rank order of hourly median (channel 4–7).

This formula will be used here with the following modification for analytical simplicity,

$$P = \Pr(v \leqq L) = \begin{cases} aL^2 = \dfrac{t_L}{T} & 0 \leqq L^2 \leqq \dfrac{1}{a} \\ 1 & L^2 \geqq \dfrac{1}{a} \end{cases}. \tag{8}$$

For this simple model the median value, L_m, is given by

$$L_m^2 = \frac{1}{2a} = \frac{L^2 T}{2t_L}. \tag{9}$$

This relation can be used to calculate values of L_m from the 4-GHz hourly rank order data of Fig. 14a. The results for levels 2 and 3 are shown on Fig. 18 along with the 4-GHz hourly median data from Fig. 17. There is good agreement for the first 20 rank order days. Level 3 predicts a worst hour median 2.5 dB higher and level 2 predicts a worst hour median 1 dB lower than the Fig. 17 data.

The calculated results roll off faster below 10 dB than the Fig. 17 data, which means that the aL^2 model does not hold when the hourly median is less than 10 dB. This is to be expected because the aL^2 model applies for multipath fading while the Fig. 17 data contains a con-

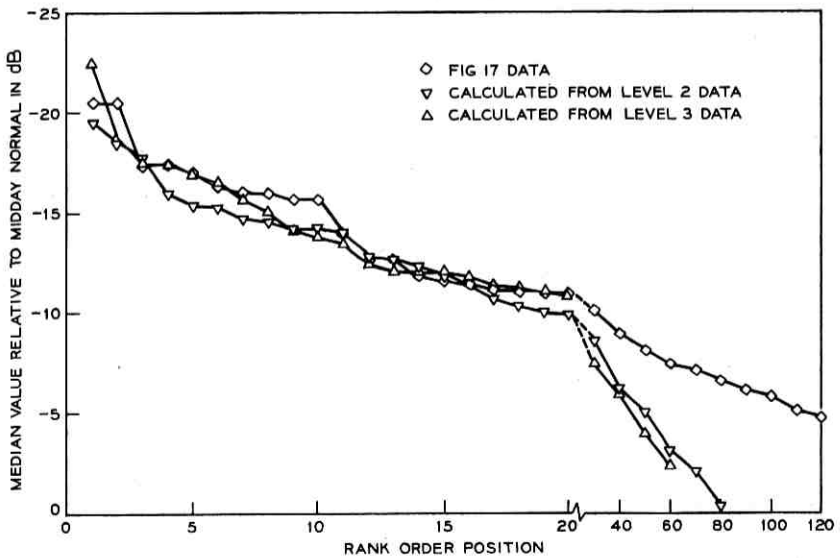


Fig. 18—Comparison of 4-GHz hourly median data of Fig. 17 with calculated values.

siderable number of hours during which the signal is depressed with little multipath fading. In any case, the analytic model (8) is adequate for the higher values of the hourly median which is the region of greatest interest. This model will now be applied to the nightly and hourly rank order data presented in Figs. 10 and 14.

7.6 *Empirical Probability Distribution of Daily Fade Time*

The rank order data (Section 7.2) can be used to estimate the probability distribution for the daily fade time by plotting the value of the ith ordered sample versus the probability estimate $(N)^{-1}\ (i - 0.5)$, defined as the cumulative empirical probability distribution (c.e.p.d.).[10] The random variable t_L is defined as the total amount of time during the 9-hour period for which the signal level is less than or equal to L.* The rank order daily fade time data (Section 7.2) are samples of t_L, with t_{Li} the ith rank ordered sample value. Thus the c.e.p.d. is:

$$P_{Li} \equiv \Pr\,(t_L > t_{Li}) = \frac{i - 0.5}{N_L} \tag{10}$$

where N = number of sample values.

Repeating equation (7) in a form consistent with the above definitions gives

$$\Pr\,(v_i \leqq L) = a_i L^2 = \frac{t_{Li}}{T_d} \tag{11}$$

where v_i is the envelope voltage during the ith interval,
a_i is the environmental constant during the ith interval,
T_d = 9 hours.

Combining (10) with (11) gives

$$P_{Li} = \Pr\left(\frac{t_L}{L^2 T_d} \geqq \frac{t_{Li}}{L^2 T_d}\right) = \Pr\,(a_d \geqq a_i). \tag{12}$$

Thus the c.e.p.d. for t_L is identical to that for the random variable a_d, the daily environmental constant.

In the calculation of P_{Li} for levels 2–4 the values used for N_L will be those given in Table IV. At level 2 there were 36 days with non-zero fade time at 4 GHz and 35 at both 6 and 11 GHz. If the aL^2 model is interpreted in a deterministic sense then all days with level 2 fade time should have level 3 fade time; yet there were only 30 such days at 4 GHz.

* The 9-hour period was chosen because most of the daily fading occurred between 12 P.M. and 9 A.M.

There is no inconsistency because the aL^2 model is statistical so that not all level 2 fades also generate level 3 fades; thus the 30 samples are used to construct a c.e.p.d. which can be compared to that obtained for the 36 samples at level 2. The corresponding procedure was followed for level 4 at 4 GHz and for levels 2–4 for 6 and 11 GHz. Two basic assumptions are made: (*i*) 0.2-second sampling has a negligible effect; (*ii*) the samples at any level are independent. The first assumption will be justified if the level 4 results are consistent with the level 2 results because the sampling interval would have a greater effect on the level 4 results. The second assumption only requires independence from day to day which is plausible.

The daily rank order fade time data have been plotted on Figs. 19a–c according to (12). The probability scale is exponential and the abscissa is logarithmic. The data for all three frequencies appear to be independent of level and approximately linear with increasing scatter above 70 percent. The conclusion is that the aL^2 representation is adequate over the 20–40 dB fade depth range for daily fading.

In Section V we examined the frequency dependence of the environmental constant. Utilizing that relation, and normalizing to 4 GHz, equation (12) becomes:

$$P_{Li} = \Pr\left[\frac{a_d}{\left(\frac{f}{4}\right)} \geqq \frac{a_i}{\left(\frac{f}{4}\right)}\right]. \tag{13}$$

The level 2 data for 4, 6, and 11 GHz has been plotted in Fig. 20 according to (13). The reduced data are consistent for the three frequencies; a straight line whose equation is

$$\Pr\left(\frac{t}{L^2 T_d} = a_d \geqq A\right) = \exp\left(-1.2\sqrt{A\left(\frac{4}{f}\right)}\right) \tag{14}$$

provides a good fit (±2 dB) to the data below 0.9. Similar results are obtained for levels 3 and 4 but with increased scatter.

Figure 20 indicates that the environmental parameter a_d is linearly dependent on frequency on a day-to-day statistical basis for multipath fading. This is a stronger result than that of Section V, where the linear frequency dependence was found valid for the measurement period taken as a whole. The net result of this analysis is that the daily fade time for a day picked at random can be calculated statistically.

The result, (14), can be checked against the results given in Table II for the entire measurement period in the following manner. Equation (11) gives, for the ith fading day out of N,

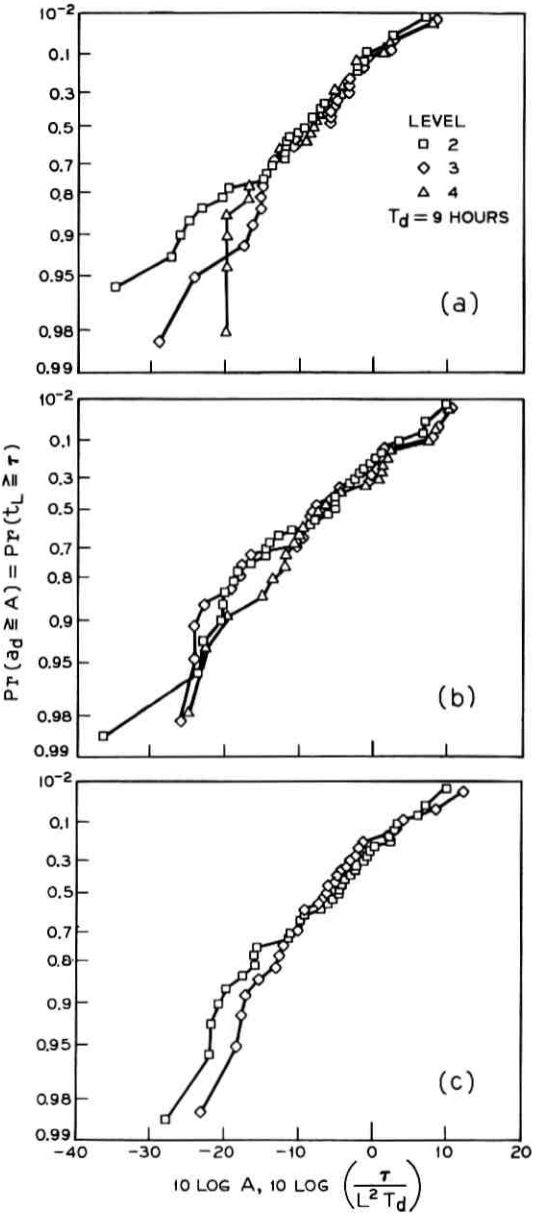


Fig. 19a—4-GHz daily fade time, 1966 West Unity, cumulative empirical probability distribution.
Fig. 19b—6-GHz daily fade time, 1966 West Unity, cumulative empirical probability distribution.
Fig. 19c—11-GHz daily fade time, 1966 West Unity, cumulative empirical probability distribution.

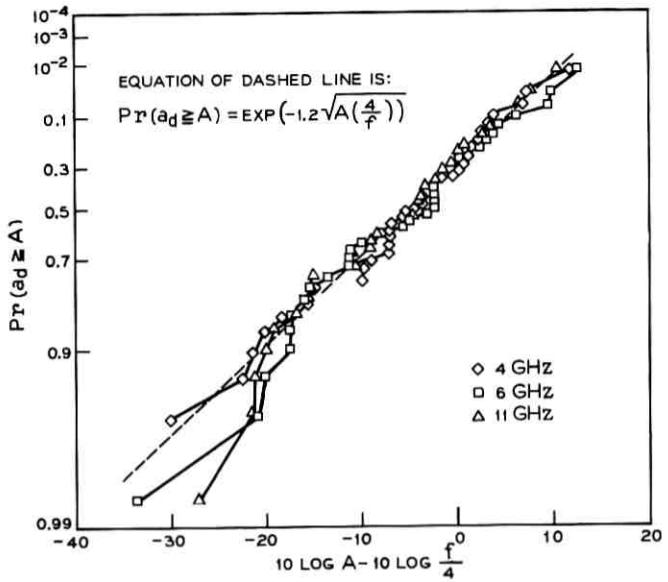


Fig. 20—Daily fade time for level 2, cumulative empirical probability distribution, 4, 6, and 11 GHz, 1966 West Unity.

$$\Pr(v_i \leqq L) = a_i L^2 = \frac{t_{Li}}{T_d}. \tag{11}$$

The fractional fade time accumulated over the N days is [equation (5)]

$$\Pr(v \leqq L) = r_N L^2 = \frac{\sum_{i=1}^{N} t_{Li}}{NT_d} = \frac{\sum_{i=1}^{N} a_i L^2 T_d}{NT_d}. \tag{15}$$

Thus

$$r_N = \frac{\sum_{i=1}^{N} a_i}{N} \tag{16}$$

so that r_N is the average value of the a_i's which in turn can be calculated from

$$P(a_d \leqq A) = 1 - \exp\left(-1.2\sqrt{A\left(\frac{4}{f}\right)}\right). \tag{17}$$

Thus

$$r_N = \int_0^\infty ap(a)\,da = \int_0^\infty a\frac{dP}{dA}\,da \tag{18}$$
$$= 1.4\left(\frac{f}{4}\right).$$

To convert from N periods of 9 hours each to the entire measurement period of 5.26×10^6 seconds the above result must be multiplied by $[(N)(32,\ 400)/5.26 \times 10^6]$. Substitution of the number of days with nonzero level 2 fade time (Table IV) gives the results shown in Table VI. The coefficients obtained from the daily fade times are in fair agreement with the overall coefficients which is a reassuring check on the consistency of the results.

As a digression it is to be noted that the usual Rayleigh assumption for modeling the propagation medium corresponds to $A = 1$. Equation (18) shows that the average value of a_d corresponds to $A = 1.4$. It appears that the Rayleigh assumption is reasonable on the average but it should be recognized that some 30 percent of the days will have greater fading.

The calculation of the daily median is the last topic in this section. As noted in Section 7.5, the median value L_m for the aL^2 distribution model is given as

$$L_m^2 = \frac{1}{2a} \tag{19}$$

or

$$20 \log L_m = -10 \log a - 3 \text{ dB}. \tag{20}$$

Values for 20 log L_m can be read off directly from Fig. 20, e.g., at 4 GHz the 90-percent point is —8 dB relative to midday normal, while the 1-percent point is −14 dB. This calculation is valid only for median values less than some −10 dB because as the value of a gets small the

TABLE VI—FADE TIME COEFFICIENT OF L^2

Freq (GHz)	Calculated from Daily Fade Time	Measured (Table II)
4	0.3	0.25
6	0.45	0.53
11	0.82	0.69

calculated median values will be much too high. This occurs because the range of validity of the aL^2 representation certainly does not extend above -10 dB relative to midday normal. As a matter of fact, the daily median is uninteresting and is included here only for completeness. The next section will take up the matter of the hourly variation for which the median calculation is more meaningful.

7.7 *Empirical Probability Distribution of Hourly Fade Time*

The treatment of the daily fade time in Section 7.6 will be applied to the hourly fade time in this section. As in Section 7.6, we define*

- t_L total time during an hour for which the signal level is less than or equal to L,
- t_{Li} ith rank ordered sample value,
- N_L number of samples,
- v_i envelope voltage during ith hour,
- a_i environmental constant for the ith hour,
- T_h one hour (3600 seconds).

The cumulative empirical probability distribution for the hourly data is constructed according to (see Section 7.6)

$$P_{Li} \equiv \frac{i - 0.5}{N_L} = \Pr\left(\frac{t_L}{L^2 T_h} \geqq \frac{t_{Li}}{L^2 T_h}\right) = \Pr\,(a_h \geqq a_i) \tag{21}$$

with

$$\Pr\,(v_i \leqq L) = a_i L^2 = \frac{t_{Li}}{T_h}. \tag{22}$$

The hourly rank order data on Figs. 14a–c are replotted on Figs. 21a–c according to equation (21). The probability scale is exponential and the abscissa is logarithmic. The 4-GHz results on Fig. 21a are consistent with less than 3 dB scatter from 0.8 to 0.01 and increasing scatter for smaller data values. The cutoff value imposed by the 0.2-second sampling rate is -22.2 dB for level 2, -11.6 dB for level 3, and -2.5 dB for level 4. Since the 4- and 6-GHz data is averaged for 7 and 6 channels respectively, the actual cutoff point is some 8 dB lower. In any case increased scatter is to be expected for smaller sample values.

The 6-GHz results on Fig. 21b are consistent for levels 2 and 3 but the level 4 data is offset. If all the sample hours had the same amount of fade time at a given level then the c.e.p.d. would be a vertical line on Fig. 21b. One possible explanation, therefore, is that the level 4 hours

* The hourly data utilizes similar notation to that for the daily data.

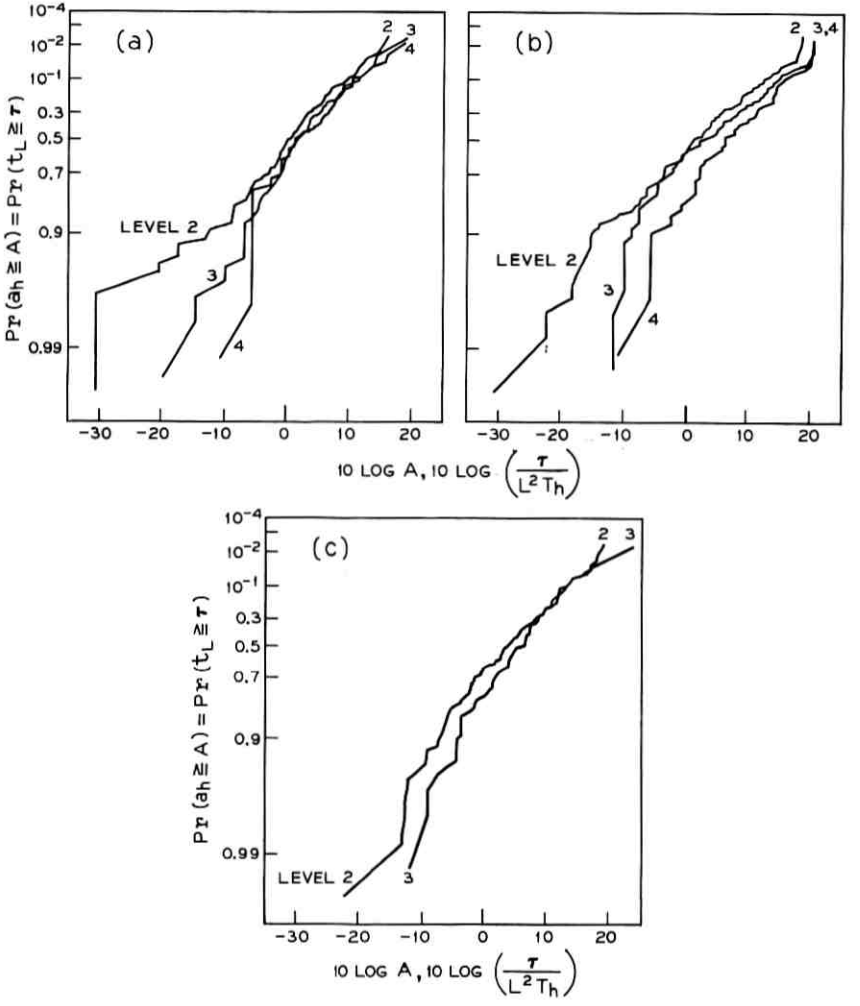


Fig. 21a—4-GHz hourly fade time, 1966 West Unity, cumulative empirical probability distribution.
Fig. 21b—6-GHz hourly fade time, 1966 West Unity, cumulative empirical probability distribution.
Fig. 21c—11-GHz hourly fade time, 1966 West Unity, cumulative empirical probability distribution.

at 6 GHz tended to be more alike than the level 2 and level 3 hours. This behavior was also noted in conjunction with Figs. 14b and 15. We assume that the 6-GHz hourly data for level 4 is atypical.

The 11-GHz results on Fig. 21c are reasonably consistent. Since

there was only one 11-GHz channel, the effect of the 0.2-second cutoff is clearly discernible.

The level 2 data from Figs. 21a–c is cross-plotted on Fig. 22 where the frequency has been normalized to 4 GHz. Thus, assuming that the level 2 data is typical, it is found that the distribution of the hourly environmental constant a_h for hours containing level 2 fades is approximately given by

$$P(a_h \leqq A_h) = 1 - \exp(-0.7[A(4/f)]^{1/2}). \tag{23}$$

The square-root function in the exponent was arbitrarily chosen to agree with the result for the daily data, e.g., (14). A slightly larger value than 0.5 would give a better fit for the smaller sample values but this was considered unimportant.

From equation (23), the 50-percent point for 4 GHz falls at $A_h = 1$, with the 99-percent point at $A_h = 30$. This means that for a fading hour the level 2 fade time will exceed 1080 seconds with 1 percent probability.

The hourly median can now be obtained based on the aL^2 model (see Section 7.3):

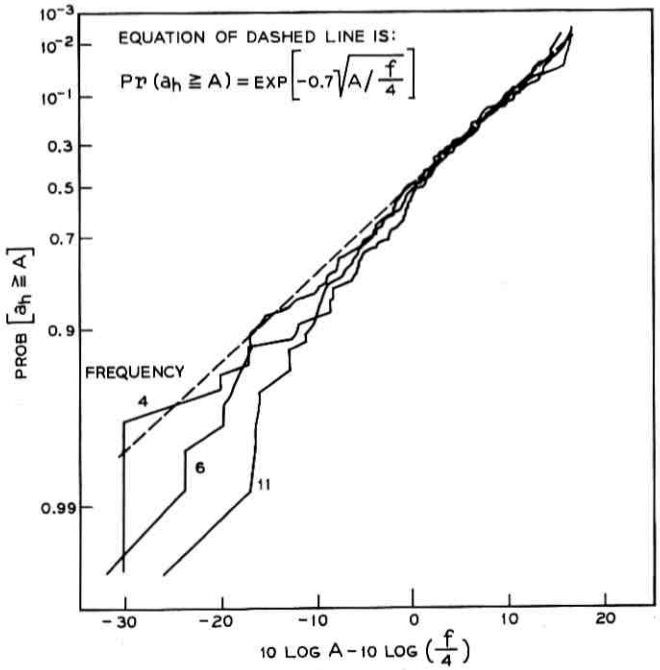


Fig. 22—Hourly fade time for level 2, cumulative empirical probability distribution, 4, 6, and 11 GHz, 1966 West Unity.

$$20 \log L_{mh} = -10 \log a_h - 3 \text{ dB}. \tag{24}$$

Values for L_{mh} can be obtained from Fig. 22 using (24). For example, at 4 GHz the 1-percent median is -18 dB relative to midday normal. The actual maximum data point shown, however, falls at 10 log A = 17 dB which gives a median of -20 dB. This is in good agreement with the minimum median of -20.5 dB for the data given on Fig. 17 for one of the 4-GHz channels. This points up the problems of using a best fit line to estimate tail probabilities. Within such limitations it appears that the simple aL^2 model for the hourly and daily variations of multipath fade time is adequate.

7.8 *Empirical Probability Distribution of the Median of the Fade Depth Distribution for a Minute in a Fading Hour*

In preceding sections, the multipath fading data have been examined on a daily basis (Sections 7.2 and 7.6) and an hourly basis (Sections 7.3, 7.4, 7.5, and 7.7). Finer scale variations also are of interest. The sampling rate for a single radio channel varies from 0.2 second to 30 seconds depending on the amount of activity. This suggests that the smallest consecutive time interval that can be used in the construction of fade depth distributions is one minute. The measurement technique guarantees that if the 30-second rate is being used then the difference between any two 30-second samples is less than 2 dB.

The previous section (7.7) gave an estimate of the probability distribution of the hourly median fade depth of a fading hour. It is logical then to consider the median of the fade depth distribution for each minute within a clock hour. One channel in each of the three bands, 4, 6, and 11 GHz, was selected for study during five hours with multipath activity. The hourly medians in dB for each combination are given in Table VII. Four of the hours selected were drawn from among the ten having the most fading, with one lesser fading hour (day 10, 5–6 A.M.) included for comparison.

The data analysis for the five hours proceeds as follows:

(*i*) Construct the experimental fade depth distribution for each minute within the hour and for the entire hour.

(*ii*) Estimate the 50-percent dB point from the fade depth distribution for: (a) each minute within the hour: m_i dB, $1 \leqq i \leqq 60$; (b) the entire hour: h dB.

(*iii*) Calculate the difference in minute and hour medians:

$$d_i = h - m_i \text{ dB}. \tag{25}$$

(*iv*) Rank order the d_i values from largest to smallest (i is then

TABLE VII—MEDIAN VALUES OF THE HOURLY FADE DEPTH DISTRIBUTION

Day	Hour	4 GHz	6 GHz	11 GHz
10	2–3 A.M	①* −20.5 dB	② −23.5 dB	① −27.5 dB
	3–4 A.M.	③ −17.4 dB	④ −21.0 dB	② −22.7 dB
	5–6 A.M.	㉕ −11.5 dB	⑰ −13.0 dB	⑳ −14.2 dB
28	0–1 A.M.	⑦ −16.4 dB	⑩ −16.2 dB	⑦ −17.4 dB
43	5–6 A.M.	⑥ −17.5 dB	① −22.8 dB	③ −21.0 dB

* The circled numbers give the hourly rank order position of the fade time at or below level 3 (−31.0 dB) in the hour.

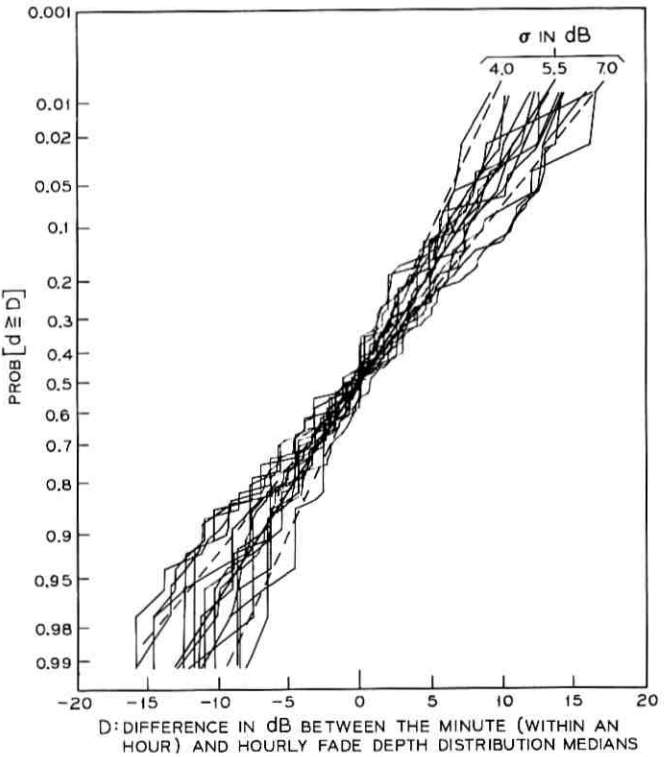


Fig. 23—Cumulative empirical probability distribution for the difference between the minute and hourly fade depth distribution medians. Data samples of five hours for 4, 6, and 11 GHz.

redefined as the rank order index with $i = 1$ for the worst 50-percent minute median fade value as normalized to the hourly median).

(*v*) Construct the cumulative empirical probability distribution for d, that is,

$$\Pr [d \geqq d_i] \cong \frac{i - 0.5}{60}. \tag{26}$$

The c.e.p.d. for d is plotted on Fig. 23 for all five hours and the three radio channels. This single plot suffices because there is no consistent difference between the different hours for a particular channel or between the different channels in a particular hour. As expected, the 50-percent point falls at the 0-dB difference point (within ± 1 dB). The entire set of data appears to be normal with a mean of 0 dB and a standard deviation of 5.5 $\pm$ 1.5 dB. It can be seen that, for a multipath fading hour, the minute medians vary considerably as compared to the hourly median. This is not surprising since the average duration of a multipath fade varies from 4 seconds at a -40-dB fade depth to 40 seconds at a -20-dB fade depth.[4]

To recapitulate, the hourly median can be estimated from Fig. 22 using equation (24) and the difference in the hourly and minute median calculated using a normal distribution with a mean of 0 dB and σ = 5.5 dB.

VIII. AMPLITUDE STATISTICS FOR ENTIRE TEST PERIOD

8.1 *Introduction*

The effects of multipath propagation are most important in the deep fade region, because the received signal can be rendered unusable. The signal statistics for shallow fade depths also are of interest if only because the signal amplitude resides in this range for the vast majority of time. At West Unity an elapsed time of 5.26×10^6 seconds (T_o) was the total data base; of this total 0.78×10^6 seconds (T_A) contained all the deep multipath fading and was subjected to detailed analysis.[1,2,4] In this section, statistics for the remaining 4.48×10^6 seconds (T_B) will be presented for two 4-GHz and two 6-GHz channels. The data for the 11-GHz channel was not included in this analysis because of the difficulty of separating out the effects of rain attenuation.

8.2 *Fade Depth Distribution*

The fade depth distributions for 4 and 6 GHz are given on Figs. 24 and 25, respectively, for the three time bases T_A, T_B, and $T_o = T_A + T_B$.

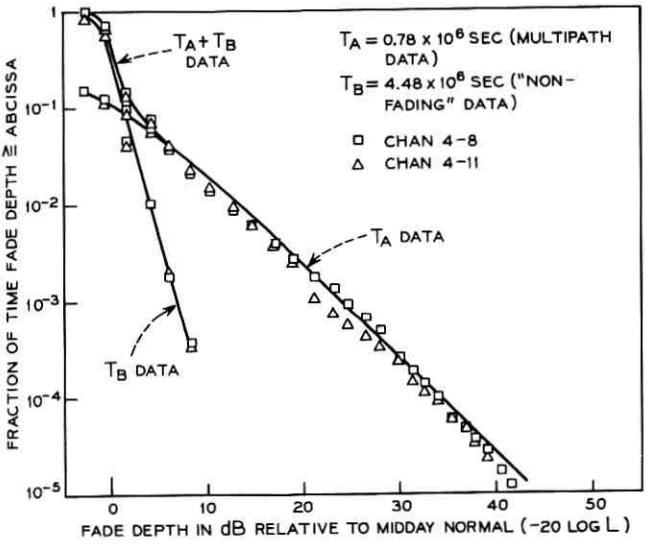


Fig. 24—4-GHz fade depth distribution for the entire test period, 1966 West Unity.

The lines on the figures are smoothed through the data, with the deep fade equations $0.25L^2$ and $0.53L^2$ (as given in Table II) used below -20 dB for 4 and 6 GHz respectively. As expected, the T_B data dominates the total distribution above the 10-percent point.

It should not be inferred from these results that there was zero probability of upfades above $+3$ dB. The equipment was designed to give this value whenever the signal level was in excess thereof.

The data for the fade depths less than 20 dB have been replotted on Figs. 26 and 27 on a normal probability scale where each set of data has been normalized to its own time base, e.g., the data for the multipath period are expressed as a fraction of 0.78×10^6 seconds (T_A). The data are given for only one of the channels in each band since the two channels have almost identical statistics in this fade depth range (see Figs. 24 and 25).

The plots show that neither the data for the total measurement period of 5.26×10^6 seconds ($T_A + T_B$) nor for the "nonfading" period of 4.48×10^6 seconds (T_B) are lognormal. During normal daytime periods of transmission on a single hop when the atmosphere is well mixed the envelope voltage scintillates and has a lognormal distribution with a standard deviation less than 1 dB. The T_B data is drawn from

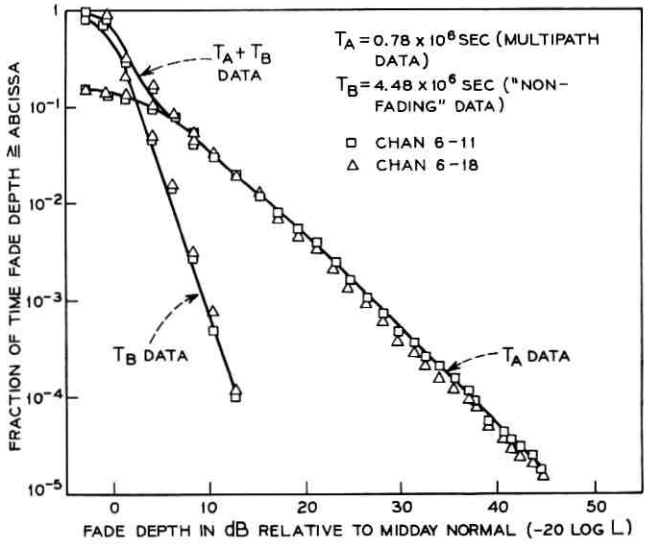


Fig. 25—6-GHz fade depth distribution for the entire test period, 1966 West Unity.

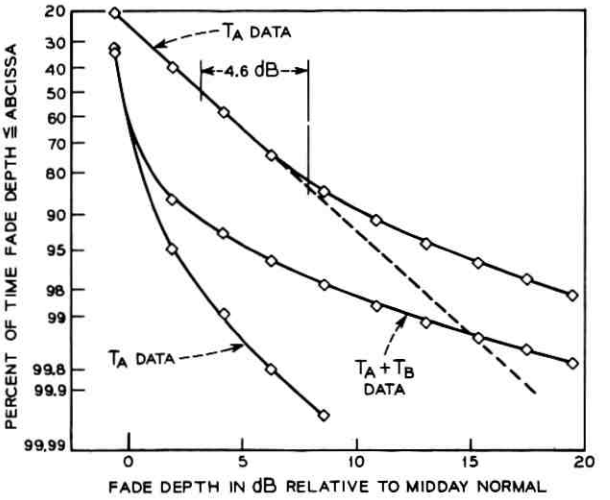


Fig. 26—4-GHz fade depth distribution, 1966 West Unity, probabilities for measurement intervals T_A (0.78×10^6 seconds), T_B (4.48×10^6 seconds), and $T_A + T_B$ (5.26×10^6 seconds).

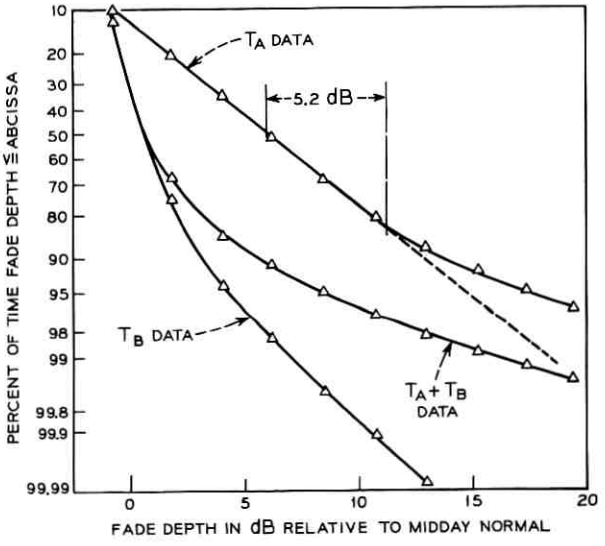


Fig. 27—6-GHz fade depth distribution, 1966 West Unity, probabilities for measurement intervals T_A (0.78×10^6 seconds), T_B (4.48×10^6 seconds), and $T_A + T_B$ (5.26×10^6 seconds).

a mixture of such periods and others with mild fading. This mixture, coupled with the approximately 2-dB quantizing intervals, makes it difficult to draw definitive conclusions from either the T_B or the $(T_B + T_A)$ data in the central part of the distribution.

The T_A data are approximately lognormal over the central 80 percent of the distribution, with the characteristics given in Table VIII. As can be seen from Figs. 26 and 27, the lognormal characteristic is useless for predicting the deep fade behavior. This seems to be a common finding; an observable which can be modeled as having multiplicative components is usually lognormal near its median. However, a more sophisticated model is needed for calculation of the tails of the distribution.[7]

TABLE VIII—CHARACTERISTICS OF SHALLOW FADES DURING PERIODS INCLUSIVE OF ALL DEEP MULTIPATH FADES

Characteristic	4 GHz	6 GHz
50% point	3.1 dB	6.0 dB
σ	4.6 dB	5.2 dB

8.3 *Number of Fades and Average Fade Durations*

Data on the number of fades and the average fade duration were also obtained for a 4-GHz and a 6-GHz radio channel, as shown on Figs. 28a–b and 29a–b respectively. The number of fades occurring during the deep fade total time (T_A) first increases and then decreases as the fade depth increases below 0 dB. The line through the deep fade region, $3670L$ for 4 GHz on Fig. 28a and $6410L$ for 6 GHz on Fig. 29a, are the least squares fitted lines to the data for all the channels in the separate bands.[4] The data for the balance of the measurement time (T_B) varies more rapidly as a function of fade depth, i.e., approximately a factor of 100 from 0 to −10 dB. Of course, the T_B data has many more fades at 0 dB fade than the T_A data. Note that the deep fade fitted line would overestimate the number of fades by a factor of 2 at a −10-dB fade depth but would be quite adequate for prediction at 0 dB fade depth.

The average fade duration at any fade depth is obtained from the ratio of the total time at or below the fade depth to the number of fades of this depth. Values for this variable have been obtained from the data for each of the three time bases—T_A, T_B, and $T_A + T_B$—as shown on Figs. 28b and 29b for 4 and 6 GHz respectively. The lines $408L$ (4 GHz) and $490L$ (6 GHz) have been obtained for the deep fade

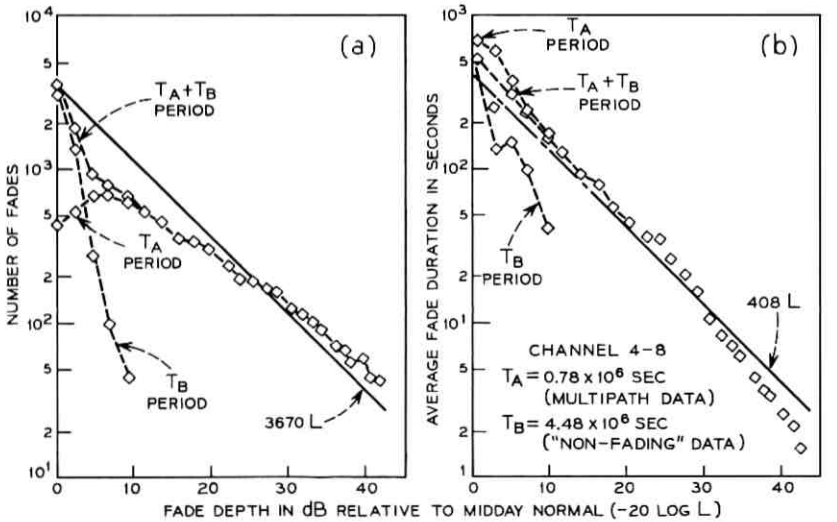


Fig. 28a—4-GHz number of fades for the entire test period.
Fig. 28b—4-GHz average fade duration for the entire test period.

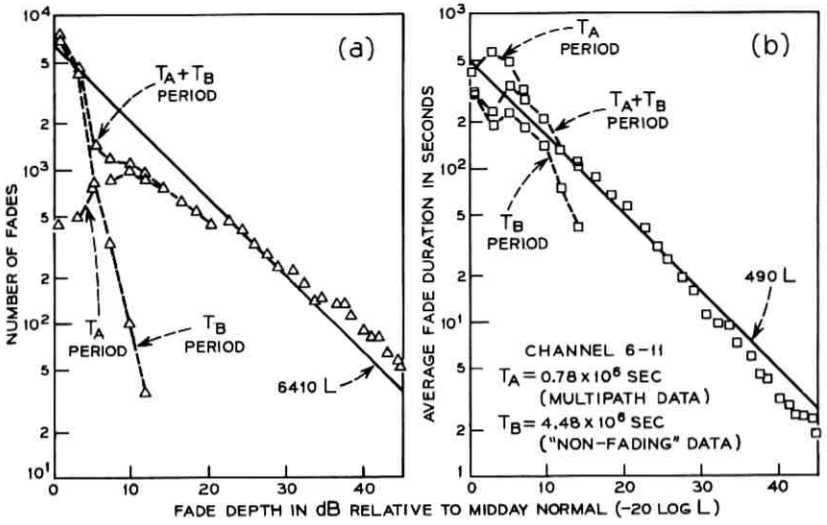


Fig. 29a—6-GHz number of fades for the entire test period.
Fig. 29b—6-GHz average fade duration for the entire test period.

data. However, these deep fade lines, extended to 0 dB, are a good representation of the data for the entire dB range. This is further evidence in support of Lin's finding that the average fade duration is less sensitive to the fading conditions than is either the number of fades or the fade depth distribution.

IX. ACKNOWLEDGMENTS

The author is indebted to many of his colleagues. The experimental data was obtained by MIDAS which was the creation of G. A. Zimmerman. The computer data processing capability was provided by C. H. Menzel. The data tabulation and plots were done by Miss E. J. Emer and Miss P. L. Russell. The interest and support of E. E. Muller and K. Bullington were invaluable.

REFERENCES

1. Vigants, A., "The Number of Fades in Space-Diversity Reception," B.S.T.J., *49*, No. 7 (September 1970), pp. 1513–1530.
2. Barnett, W. T., "Microwave Line-of-Sight Propagation With and Without Frequency Diversity," B.S.T.J., *49*, No. 8 (October 1970), pp. 1827–1871.
3. Chen, W. Y. S., "Estimated Outage in Long-Haul Radio Relay Systems with Protection Switching," B.S.T.J., *50*, No. 4 (April 1971), pp. 1455–1485.
4. Vigants, A., "Number and Duration of Fades at 6 and 4 GHz," B.S.T.J., *50*, No. 3 (March 1971), pp. 815–841.

5. Beckmann, P., and Spizzichino, A., *The Scattering of Electromagnetic Waves from Rough Surfaces*, New York: Pergamon Press, 1963, pp. 355–367.
6. Ruthroff, C. L., "Multiple-Path Fading on Line-of-Sight Microwave Radio Systems as a Function of Path Length and Frequency," B.S.T.J., *50*, No. 7 (September 1971), pp. 2375–2398.
7. Lin, S. H., "Statistical Behavior of a Fading Signal," B.S.T.J., *50*, No. 10 (December 1971), pp. 3211–3270.
8. Pearson, K. W., "Method for the Prediction of the Fading Performance of a Multisection Microwave Link," Proc. IEE, *112*, No. 7 (July 1965), pp. 1291–1300.
9. Yonezawa, S., and Tanaka, N., *Microwave Communication*, Tokyo: Maruzen Co., Ltd., 1965, pp. 25–60.
10. Wilk, M. B., and Gnanadesikan, R., "Probability Plotting Methods for the Analysis of Data," Biometrika, *55*, No. 1 (1968), pp. 1–17.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

A Fast Bipolar-IGFET Buffer-Driver

By G. MARR, G. T. CHENEY, E. F. KING, and E. G. PARKS

(Manuscript received August 24, 1971)

This paper discusses the performance and interface advantages of a self-isolating bipolar-IGFET (BIGFET) integrated structure as an output buffer-driver for IGFET integrated circuits. The low-capacitance, high-impedance input and low-impedance, high-current output characteristics make the BIGFET ideally suited to drive large output capacitances and to interface with bipolar logic circuits. It is shown that in a shift register application the operating speed is increased substantially when the BIGFET is used as output buffer and is essentially independent of output capacitance up to 100 pF. The application of BIGFET output circuits to 5-volt T^2L and 3-volt collector-diffusion-isolation (CDI) T^2L is also discussed.

I. INTRODUCTION

Due to the high output impedance normally associated with Insulated-Gate Field-Effect Transistors (IGFET) two problems often arise in digital IGFET integrated circuits: (*i*) Charging and discharging times for capacitances external to the integrated circuit are long compared to the corresponding times for internal circuit nodes. (*ii*) Interfacing with bipolar logic requires IGFETs to provide and/or sink currents which are larger than those normally available from IGFETs with typical integrated circuit geometries. Attempts to solve these problems usually involve large IGFET inverters or push-pull drivers as output stages. Since these types of output interface circuits employ large-geometry IGFETs and have higher input capacitances than those capacitances typically found at the nodes of the internal IGFET circuitry, the overall result is that circuit speed is degraded at the output interface.

This paper discusses the use of a self-isolating bipolar-IGFET (BIGFET) integrated structure in an output buffer-driver. Although this structure has been previously proposed,[1-3] there have been no reported experimental studies of improved circuit performance when the BIGFET is incorporated directly on a monolithic p-channel IGFET

integrated circuit. Since the BIGFET is capable of providing a low-capacitance, high-impedance input and a low-impedance, high-current output, it provides an almost ideal solution to the interface problems discussed above.

II. DEVICE STRUCTURE AND CHARACTERISTICS

A schematic and device cross section of a BIGFET are shown in Fig. 1. The structure is basically an IGFET and a vertical npn bipolar transistor in cascade. The collector of the npn transistor is common to the Silicon Integrated Circuit (SIC) substrate. A p-type diffusion performs the dual role of bipolar transistor base and p-channel IGFET drain. The emitter is formed by the same phosphorus diffusion that is used to make ohmic contact to the 6–9 Ω-cm n-type substrate.

The current-voltage characteristics for a typical BIGFET with $V_T = -1.0$ volt and $h_{FE} = 140$ at $I_c = 10$ mA are shown in Fig. 2. It may be seen in the figure that the output current is in the range of tens of milliamperes, although the IGFET gain factor, $\beta[(\mu\kappa\epsilon_o/t_{ox})W/L]$, for this structure is only 100 μmhos/volt. The overall effective gain factor is just the product of β and h_{FE} or, in this case, 14,000 μmhos/volt. Therefore, when using this structure for high-current output circuit applications, one may employ a small gain factor IGFET with correspondingly low input capacitance. Since this input capacitance need be no greater than that found at a typical internal node of an IGFET SIC, the delay through the BIGFET output-buffer, in turn, need be no greater than the intrinsic delays associated with the internal IGFET circuitry.

BIGFETs with the structure discussed above have been routinely

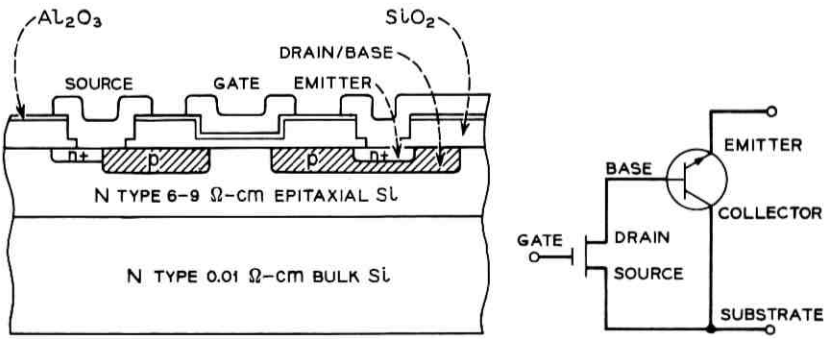


Fig. 1—BIGFET device schematic and structure.

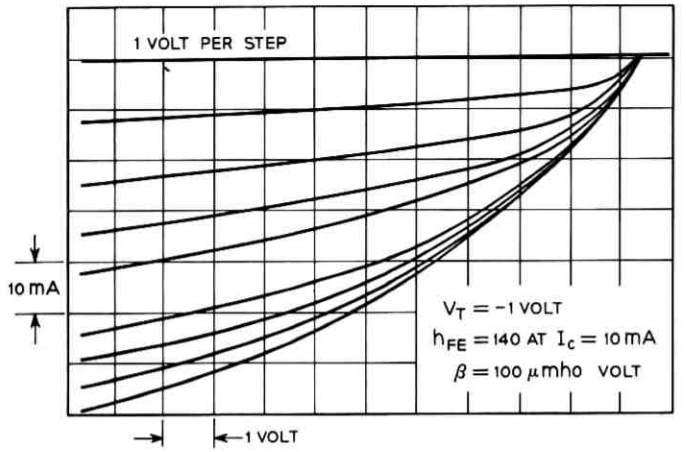


Fig. 2—BIGFET current-voltage characteristics.

fabricated with a minimum h_{FE} of 60 at $I_c = 1$ mA. Preliminary life-test data indicate that an end-of-life limit for h_{FE} of 50 is feasible for high-reliability applications. The temperature dependence of h_{FE} is $(dh_{FE}/dT)/h_{FE} \sim 1$ percent per degree from 0° to 80°C.

III. CIRCUIT PERFORMANCE

In order to assess empirically the circuit performance improvements achievable through the use of a BIGFET output driver, two four-bit static shift registers were designed, fabricated, and tested. One version of the shift register (SR1) has a large IGFET inverter ($\beta_{\text{driver}} = 60$ μmhos/volt) as the output stage. The βs of the IGFETs in the third and fourth bits are appropriately increased to achieve optimum design for maximum circuit speed. The second version (SR2) uses a BIGFET output driver which consists of a normal IGFET inverter ($\beta_{\text{driver}} = 20$ μmhos/volt) in cascade with a bipolar emitter follower. For the case of SR2, there was no increase in the gain factors of the IGFETs in the shift register bits just preceding the BIGFET buffer-driver. The two shift registers are shown schematically in Fig. 3.

To measure the maximum clocking frequency ($f_{\max}$) of the two shift registers, a 7-inverter cascade with a BIGFET output stage was used as signal discriminator. Signals from the shift register were acceptable only if they were capable of propagating through the seven-stage inverter cascade. Two voltage bias conditions were studied. In one case $V_{GG} = -3.0$ V and $V_{DD} = +5.0$ V while for the other $V_{GG} = 0$ V and $V_{DD} = +5.0$ V.

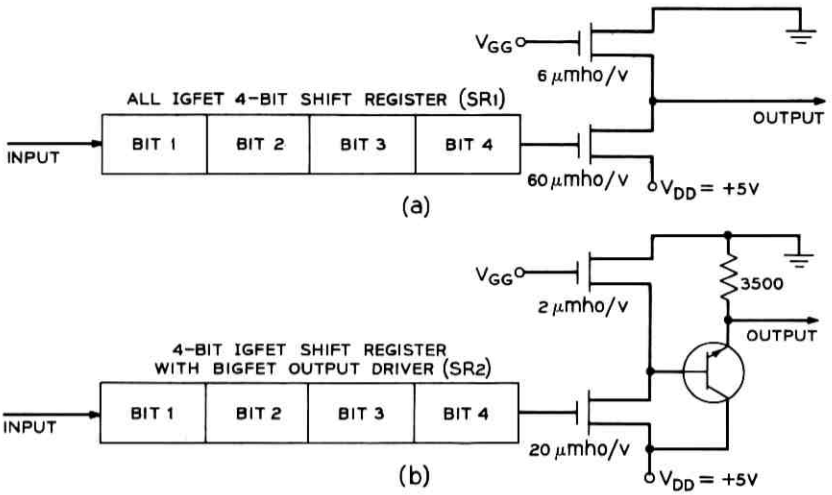


Fig. 3—Experimental circuits for comparative speed performance measurements.

The experimentally measured results for the two types of shift registers are summarized in Fig. 4. The maximum operating frequency is plotted as a function of the output capacitive load (C_o) for the two stated supply conditions. For SR1, f_{max} is twice as high at low values of C_o when two supplies are employed as when a single 5-volt supply is used. However, f_{max} decreases with increasing C_o at essentially the same rate regardless of the supplies used. On the other hand, SR2 is

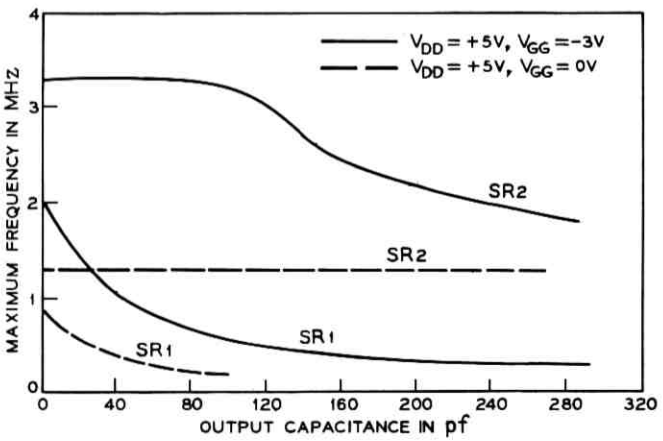


Fig. 4—Experimental performance results for all-IGFET (SR1) and BIGFET-output (SR2) shift registers.

capable of operating at 3.3 MHz for small values of C_o when two supplies are used and remains independent of C_o up to 100 pF. Beyond 100 pF the maximum operating frequency falls off in the same manner as SR1. In the single-supply case, f_{max} for SR2 is independent of C_o over the range investigated. Further comparison of the integrity of output waveforms with and without the BIGFET output buffer is demonstrated in Fig. 5. It can be seen that the output waveforms of SR1 with an IGFET output circuit are grossly degraded by the loading of 100 pF. The output of SR2 with the BIGFET is almost unaffected.

IV. CIRCUIT INTERFACE

In addition to its usefulness as an output driver, the BIGFET is also extremely versatile as a buffer to interface IGFET integrated circuits with bipolar logic. To interface with any bipolar logic, the primary design consideration is that the driver gate must furnish as well as sink currents required by the loading bipolar gate. The net result is that the value of the BIGFET emitter resistor R_E must be carefully chosen to reflect this requirement.

As an example, the choice of 1500 Ω for R_E allows a straightforward interface from BIGFET to low-power 5-volt T^2L logic. The circuit

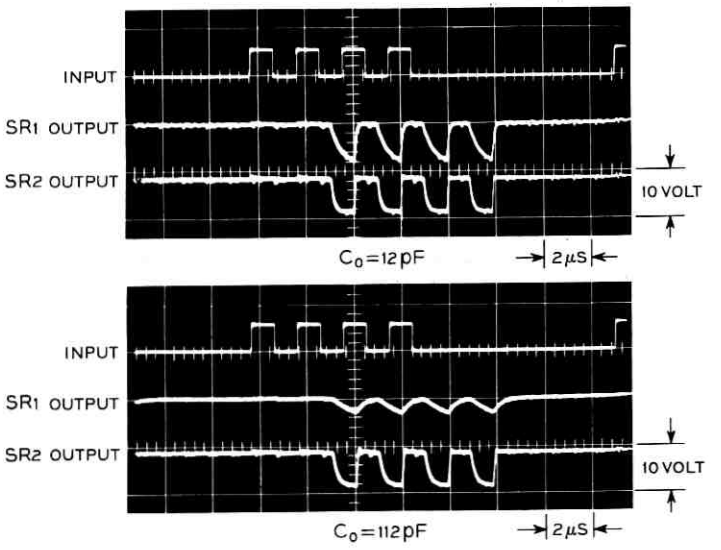


Fig. 5—Effects of capacitive loading on output waveforms for all-IGFET (SR1) and BIGFET-output (SR2) shift registers.

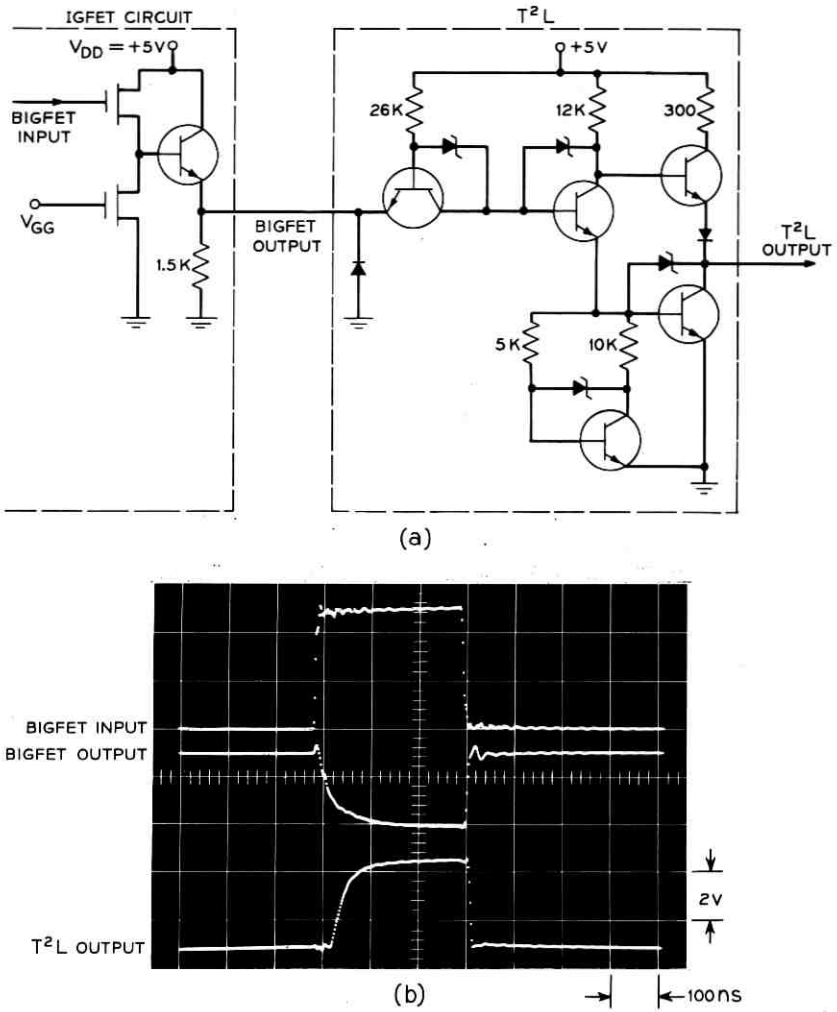


Fig. 6—(a) Circuit schematic of BIGFET-T²L interface. (b) Typical waveforms for BIGFET-T²L interface.

schematic is shown in Fig. 6a and the input and output waveforms of the circuit interface are shown in Fig. 6b. In like manner, a suitable choice of R_E allows the BIGFET to interface with RTL and DTL.

Interfacing with the 3-volt collector-diffusion-isolation (CDI)[4] T²L logic is less straightforward. If the same voltage biasing condition, i.e., R_E grounded, is used one finds that an R_E ladder of 900 Ω and 300 Ω

is needed to meet the current and voltage requirements of CDI-T²L. This is shown in Fig. 7a. The circuit shown requires a ±10-percent tolerance on the 300-Ω resistor which is not desirable for a high-yield, low-cost integrated circuit technology. Since a −3-volt supply is often available in low-threshold (V_T = −1 V) IGFET SIC applications, a higher-value and relaxed-tolerance R_E (~ 1500 Ω ± 20 percent) may be used if the emitter resistor is connected to the −3-volt supply. A schematic of this circuit configuration is shown in Fig. 7b.

Due to the voltage drop across the driver IGFET and the V_{BE} of the bipolar portion of the BIGFET, the output voltage level from the emitter follower may not be sufficient to provide adequate dc noise margin for low V_T IGFET SICs. However, this problem may be overcome by the introduction of a "pull-up" IGFET in parallel with the BIGFET output and using an IGFET as the active emitter load. This is shown in Fig. 8. The only requirement is that a gating signal

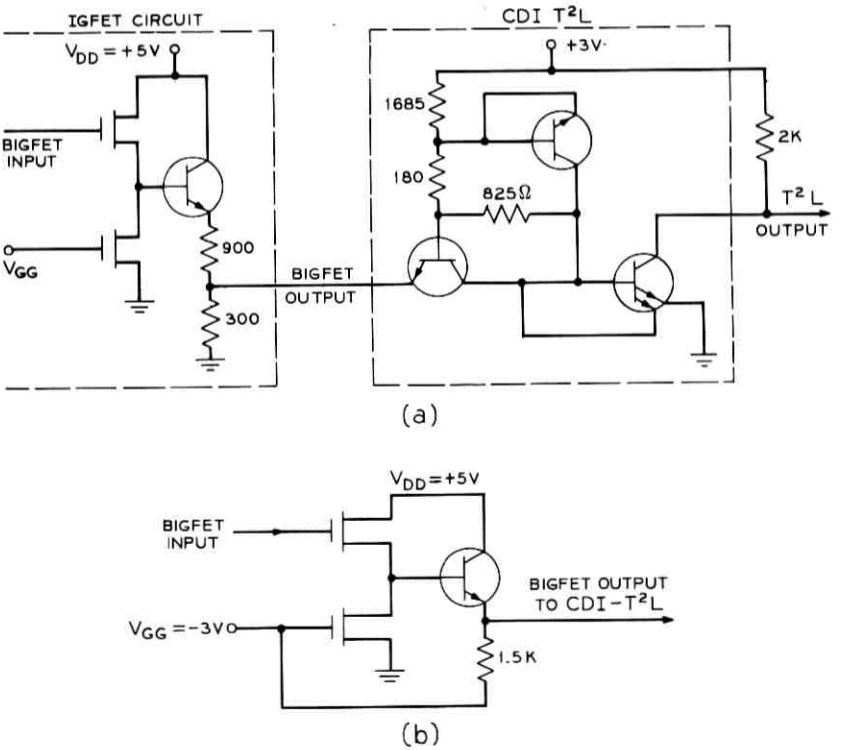


Fig. 7—Circuit schematic for BIGFET-T²L (CDI) interface: (a) resistor ladder output. (b) single emitter resistor.

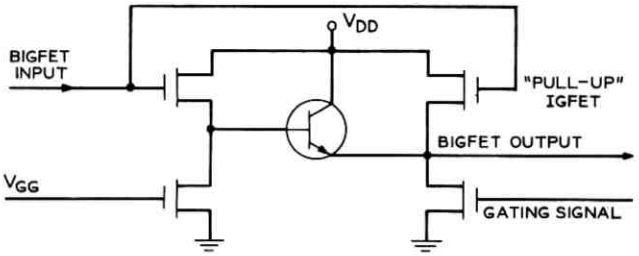


Fig. 8—IGFET "pull-up" circuit for BIGFET-IGFET interface.

must be applied to turn off the IGFET emitter load when the bipolar transistor and the associated "pull-up" IGFET are on. Such a gating signal is often conveniently available on circuits with timing signals, e.g., IGFET shift registers. An alternate solution is to provide a voltage level shifting buffer such as an IGFET source-follower at the input of the IGFET circuit to which the BIGFET interfaces.

V. CONCLUSION

This work demonstrates that there are significant advantages in using an integral bipolar-IGFET functional element as a fast interface buffer-driver. Specifically, the BIGFET driver

(*i*) requires no additional processing for isolation since the bipolar collector is common to the IGFET substrate,
(*ii*) significantly increases overall speed in multi-integrated circuit applications by reducing circuit-to-circuit propagation delays, and
(*iii*) allows direct interface with most forms of bipolar logic.

VI. ACKNOWLEDGMENTS

The authors would like to acknowledge helpful discussions with A. A. Mammele and B. H. Soloway concerning IGFET to CDI-T^2L interface.

REFERENCES

1. Price, J. E., U. S. Patent 3, 264, 493 (August 1966).
2. Lin, H. C., Ho, J. C., Ramachandran, R. I., and Kwong, K., "Complementary MOS-Bipolar Transistor Structure," IEEE Trans. Electron Devices, *ED-16*, November 1969, pp. 945–951.
3. Crawford, R. H., "Current Directions in MOS/Bipolar Interfacing," *1970 IEEE International Convention Digest*, March 1970, pp. 128–129.
4. Gary, P. A., Pedersen, R. A., Soloway, B. H., and Reed, R. A., "Designs of High Performance TTL Integrated Circuits Employing CDI Component Structures," *1970 ISSCC Digest*, February 1970, pp. 116–117.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

On Finding the Paths Through a Network

By N. J. A. SLOANE

(Manuscript received May 19, 1971)

Given a directed graph G, algorithms are discussed for finding (i) all paths through G with prescribed originating and terminating nodes, (ii) a subset of these paths containing all the edges, (iii) a subset containing all the edge-edge transitions, and (iv) a subset containing the most likely paths.

I. INTRODUCTION

Informally, a *directed graph* consists of a set of vertices or *nodes* together with a set of directed *edges* joining the nodes. (All of the figures below show directed graphs; for a formal definition see page 10 of Ref. 1. There may be more than one edge with the same originating and terminating nodes, and the originating and terminating nodes of an edge may coincide.)

Common examples of directed graphs are state diagrams of systems: the nodes represent states of the system and an edge directed from node N_i to node N_j means that it is possible for the system to go directly from state N_i to state N_j.

The following questions concerning the paths through a directed graph arose in testing for possible errors sections of the stored program of a No. 1 ESS electronic switching system.[2] However, these questions and the algorithms for their solution seem of sufficient general interest to warrant stating them independently of their origin.

Given a directed graph G, the questions are: (*i*) Find the set α of all paths through G with prescribed originating and terminating nodes. (A path is just what one would expect; a formal definition is given in Section II.) (*ii*) Find a small subset of α which contains every edge occurring in α. (*iii*) Find a small subset of α which contains all the edge-edge transitions occurring in any path in α. (*iv*) If a probability measure is associated with the edges of G, find the most probable paths in α.

These questions and algorithms for their solution are discussed in Sections III, V, VI, and VII, respectively. Section II is concerned with

the notation used to describe paths, and Section IV with an algorithm for partially solving a combinatorial problem encountered in Sections V and VI.

II. NOTATION FOR PATHS

Definition: A *path* from node N_1 to N_2 in a directed graph is a sequence of (not necessarily distinct) edges $e_1, e_2, \cdots, e_\ell$ with the property that there are nodes $N_1 = n_1, n_2, \cdots, n_{\ell+1} = N_2$ such that e_i is directed from n_i to n_{i+1} for $i = 1, 2, \cdots, \ell$. The *length* ℓ of a path is the number of edges it contains.

A path is specified by giving the ordered *string* $e_1e_2 \cdots e_\ell$ of its edges. (We are in fact describing paths by the notation used in automata theory to describe regular expressions, as given, for example, in Ref. 3 and chapter 5 of Ref. 4. However, the treatment given here is self-contained.)

It is convenient to include in the definition a path of zero length (whose endpoints N_1 and N_2 must coincide). This path is specified by the *empty string* Λ (not to be confused with the empty set ϕ).

A collection of paths is specified by the sum of the strings of the individual paths.

If S is a string, S^i denotes $SS \cdots S$ (S concatenated i times) and S^* denotes $\Lambda + S + S^2 + S^3 + \cdots$. For example, in Fig. 1 the collection of all paths from

N_2 to N_1 is ϕ,

N_1 to N_1 is Λ,

N_1 to N_2 is a,

N_4 to N_4 is $\Lambda + f + f^2 + f^3 + \cdots = f^*$,

N_2 to N_3 is $d + ce + cfe + cf^2e + \cdots = d + cf^*e$,

N_1 to N_3 is $ad + (ac + b)f^*e$.

Parentheses are used in the natural way. The following rules are easily verified. Here S is any sum of strings.

$$\phi + S = S, \phi S = S\phi = \phi$$

$$S^* = \Lambda + S + S^2 + S^3 + \cdots$$

$$\Lambda^* = \Lambda$$

$$\Lambda S = S$$

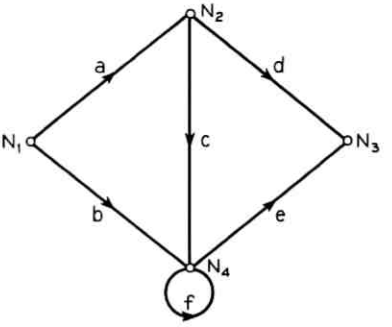


Fig. 1—An example.

$$(\Lambda + S)^* = \Lambda + SS^* = \Lambda + S^* = S^*$$

$$S_1 + S_2S_2^*S_1 = S_2^*S_1 \; ; \; S_1 + S_1S_2^*S_2 = S_1S_2^*$$

$$(S_1 + S_2)^* = (S_1^* + S_2^*)^*.$$

III. FINDING ALL PATHS THROUGH A GRAPH

Let G be a directed graph with n nodes labeled N_1 , N_2 , $\cdots$, N_n . Methods are given for finding all paths through G having prescribed starting node N_μ and (not necessarily distinct) terminating node N_ν . We first describe the McNaughton–Yamada Algorithm, which requires on the order of n^3 steps.

Definition: Let α_{ij}^k denote the set of paths which start at N_i , end at N_j , and do not pass through any intermediate node N_p with $p > k$, for $k = 0, 1, \cdots, n$, and $i, j = 1, 2, \cdots, n$.

The algorithm successively computes α_{ij}^0 for all i and j, then α_{ij}^1 for all i and j, $\cdots$, then α_{ij}^{n-1} for all i and j. The final step is to compute $\alpha_{\mu\nu}^n$, the set of all paths from N_μ to N_ν with no restriction on intermediate nodes, which is the desired result.

The inductive step proceeds as follows. Suppose $\alpha_{i,j}^{k-1}$ is known for all i, j, and we wish to obtain $\alpha_{i,j}^k$. Referring to Fig. 2, we see that the fundamental recurrence equation is

$$\alpha_{ij}^k = \alpha_{ij}^{k-1} + \alpha_{ik}^{k-1}(\alpha_{kk}^{k-1})^*\alpha_{kj}^{k-1}. \tag{1}$$

In words, this says that the paths from N_i to N_j containing intermediate nodes as high as k are made up of those containing intermediate nodes only as high as $k - 1$, α_{ij}^{k-1}, plus all possible paths containing N_k as an intermediate node, $\alpha_{ik}^{k-1}(\alpha_{kk}^{k-1})^*\alpha_{kj}^{k-1}$. When k is equal to either i or j,

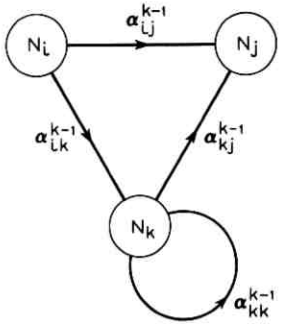


Fig. 2—The inductive step.

(1) may be simplified. We now give the complete statement of the algorithm.

THE MCNAUGHTON–YAMADA ALGORITHM[3]

1. *The Initial Step*

Define α_{ij}^0 for all $i, j = 1, \cdots, n$ by:

(1.1) if $i \neq j$,

$$\alpha_{ij}^0 = \begin{cases} \phi \text{ if there is no edge from } N_i \text{ to } N_j, \\ e_1 + e_2 + \cdots \text{ if edges labeled } e_1, e_2, \cdots \text{ join } N_i \text{ to } N_j; \end{cases}$$

(1.2) if $i = j$,

$$\alpha_{ii}^0 = \begin{cases} \Lambda \text{ if there is no edge from } N_i \text{ to itself,} \\ \Lambda + e_1 + e_2 + \cdots \text{ if edges labeled } e_1, e_2, \cdots \text{ join } N_i \text{ to itself.} \end{cases}$$

2. *The Inductive Step* (Refer to Fig. 2)

For $k = 1, 2, \cdots, n-1$ compute α_{ij}^k for all $i, j = 1, 2, \cdots, n$ from:

(2.1) if $k \neq i$, $k \neq j$ then

$$\alpha_{ij}^k = \alpha_{ij}^{k-1} + \alpha_{ik}^{k-1}(\alpha_{kk}^{k-1})^*\alpha_{kj}^{k-1};$$

(2.2) if $i \neq j$ and $k = i$,

$$\alpha_{ij}^i = (\alpha_{ii}^{i-1})^*\alpha_{ij}^{i-1};$$

(2.3) if $i \neq j$ and $k = j$,

$$\alpha_{ij}^i = \alpha_{ij}^{i-1}(\alpha_{jj}^{i-1})^*;$$

(2.4) if $i = j = k$,

$$\alpha_{ii}^i = (\alpha_{ii}^{i-1})^*.$$

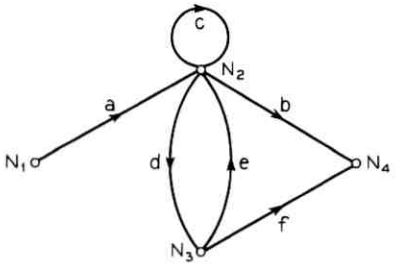


Fig. 3—An example.

3. *The Final Step*

Finally, use whichever of (2.1) to (2.4) is appropriate to calculate $\alpha^n_{\mu,\nu}$, the set of all paths from N_μ to N_ν.

Remark: In steps 2 and 3, after obtaining expressions of the form $\alpha_{\text{new}} = \cdots (\beta)^* \cdots$, it may be convenient to simplify $(\beta)^*$ by means of the rules given at the end of Section II.

An Example: We will use the McNaughton–Yamada algorithm to compute the set of all paths in Fig. 3 which start at N_1 and end at N_4, or, in other words, α^4_{14}.

Step 1.

		j			
	i	1	2	3	4
α^0_{ij}	1	Λ	a	ϕ	ϕ
	2	ϕ	$\Lambda + c$	d	b
	3	ϕ	e	Λ	f
	4	ϕ	ϕ	ϕ	Λ

Step 2. Since there are no paths into N_1, $\alpha^1_{ij} = \alpha^0_{ij}$ for all i, j.

		j			
	i	1	2	3	4
α^2_{ij}	1	Λ	$a(\Lambda + c)^*$	$a(\Lambda + c)^*d$	$a(\Lambda + c)^*d$
	2	ϕ	$(\Lambda + c)^*$	$(\Lambda + c)^*d$	$(\Lambda + c)^*b$
	3	ϕ	$e(\Lambda + c)^*$	$\Lambda + e(\Lambda + c)^*d$	$f + e(\Lambda + c)^*b$
	4	ϕ	ϕ	ϕ	Λ

where we have used the rules $\phi + S = S$ and $\phi S = S\phi = \phi$. This may be further simplified using the rules at the end of Section II as follows.

α_{ij}^2

	j			
i	1	2	3	4
1	Λ	ac^*	ac^*d	ac^*b
2	ϕ	c^*	c^*d	c^*b
3	ϕ	ec^*	$\Lambda + ec^*d$	$f + ec^*b$
4	ϕ	ϕ	ϕ	Λ

Since there are no paths out of N_4, $\alpha_{ij}^4 = \alpha_{ij}^3$ for all i, j. We can therefore go directly to Step 3:

$$\begin{aligned}\alpha_{14}^4 = \alpha_{14}^3 &= \alpha_{14}^2 + \alpha_{13}^2(\alpha_{33}^2)^*\alpha_{34}^2 \\ &= ac^*b + ac^*d(\Lambda + ec^*d)^*(f + ec^*b) \\ &= ac^*b + ac^*d(ec^*d)^*(f + ec^*b),\end{aligned}$$

which, if required, can be expanded to give

$$\begin{aligned}\alpha_{14}^4 = {} & ab + acb + ac^2b + \cdots \\ & + adf + adeb + adecb + adec^2b + \cdots \\ & + adedf + adedeb + adedecb + \cdots \\ & + adecdf + adecdeb + adecdecb + \cdots \\ & + acdf + acdeb + acdecb + acdec^2b + \cdots \\ & + acdedf + acdedeb + acdedecb + \cdots \\ & + \cdots .\end{aligned}$$

It may be verified that this includes all possible paths from N_1 to N_4.

Remarks: (*i*) When programmed in a computer language capable of handling strings, such as SNOBOL4,[5] this algorithm involves the calculation of n $n \times n$ matrices (requiring on the order of n^3 steps). Enough storage space is required to hold two $n \times n$ matrices (the current $[\alpha_{ij}^k]$, $i, j = 1, \cdots, n$, matrix and the previously calculated $[\alpha_{ij}^{k-1}]$, $i, j = 1, \cdots, n$, matrix) each entry of which is a string of letters, parentheses, $+$'s and $*$'s. (*ii*) With very little extra work Step 3 can be modified to give the paths between several pairs of nodes. This is valuable for analyzing large graphs, as we now show.

Analysis of Large Graphs by Partitioning

Since the time required for the McNaughton–Yamada algorithm grows as the cube of the number of states, large graphs cannot be handled directly. However, such graphs can usually be handled by partitioning them into smaller subgraphs, applying the algorithm to each subgraph separately, and then reapplying the algorithm to the network of subgraphs. The following simple example will illustrate the method.

Figure 4 shows a graph G partitioned into two subgraphs G_1 and G_2 which are interconnected at nodes N_2 and N_3. (Only edges between the subgraphs are shown.) Suppose we wish to find all paths from N_1 to N_4. If G_1, G_2 each contain 20 nodes, a direct application of the McNaughton–Yamada algorithm would require on the order of $40^3 = 64{,}000$ steps. This number is considerably reduced by the following technique.

Let $\beta_{ij}(G_\nu)$ denote the set of all paths starting at N_i, ending at N_j, and lying entirely in the subgraph G_ν.

We first apply the McNaughton–Yamada algorithm to G_1 and G_2 to obtain $\beta_{ij}(G_1)$, $i, j = 1, 2$, and $\beta_{ij}(G_2)$, $i, j = 3, 4$. That is, we first find all the paths between the interconnecting nodes that lie completely in one of the subgraphs. (This will take on the order of $2 \cdot 20^3 = 16{,}000$ steps.)

We now replace G by the condensed graph $\bar{G}$ of Fig. 5. $\bar{G}$ contains (*i*) nodes $\bar{N}_1$, $\bar{N}_4$ corresponding to the terminal nodes N_1, N_4, (*ii*) nodes $\bar{N}_2$, $\bar{N}_3$ corresponding to the interconnecting nodes N_2, N_3, (*iii*) edges a, b corresponding to the interconnecting edges a, b of G, and (*iv*) edges corresponding to all the paths $\beta_{ij}(G_1)$, $i, j = 1, 2$, and $\beta_{ij}(G_2)$, $i, j = 3, 4$, in G.

The McNaughton–Yamada algorithm is now used to obtain all paths from $\bar{N}_1$ to $\bar{N}_4$ in $\bar{G}$. (This takes on the order of $4^3 = 64$ steps.) It is clear that these paths are exactly all the paths from N_1 to N_4 in the original graph G. Partitioning into two equal subgraphs has thus reduced the number of steps by approximately a factor of four. (Parti-

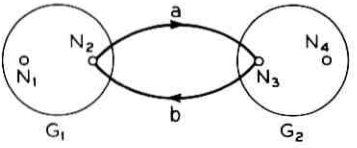


Fig. 4—A graph partitioned into subgraphs.

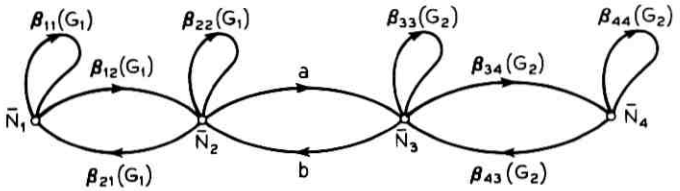


Fig. 5—The condensed graph corresponding to Fig. 4.

tioning into k equal subgraphs would reduce it by a factor of about k^2.)

The general method of analyzing a large graph by partitioning should now be clear.

If n_0 is the largest number of nodes that can be directly handled by the McNaughton–Yamada algorithm, then it is desirable to partition G in such a way that no subgraph has more than n_0 nodes, and that the total number of interconnecting nodes (which is the number of nodes in the condensed graph) is also less than n_0. (Of course the sugbraphs may themselves be partitioned.)

IV. THE COVERING PROBLEM

In Sections V and VI we will encounter a basic combinatorial problem, the covering problem, which may be stated as follows. Suppose a set $S = \{s_1, s_2, \cdots, s_n\}$ of n elements is given, together with a family $\mathfrak{F}$ of subsets of S,

$$\mathfrak{F} = \{X_1, X_2, \cdots, X_m\}, \qquad X_i \subseteq S.$$

The problem is to find a subfamily $\mathfrak{C} \subseteq \mathfrak{F}$, say

$$\mathfrak{C} = \{X_{i_1}, X_{i_2}, \cdots, X_{i_l}\},$$

where l is as small as possible, such that every element of S appearing in $\mathfrak{F}$ also appears in $\mathfrak{C}$, or formally, such that

$$X_1 \cup X_2 \cup \cdots \cup X_m = X_{i_1} \cup X_{i_2} \cup \cdots \cup X_{i_l}.$$

$\mathfrak{C}$ is called a *covering set* for $\mathfrak{F}$.

The family $\mathfrak{F}$ may be represented by an $m \times n$ (0, 1) matrix $\mathfrak{M} = (m_{ij})$, where

$$m_{ij} = 1 \quad \text{if} \quad s_j \in X_i,$$
$$0 \quad \text{if} \quad s_j \notin X_i.$$

The ith row of $\mathfrak{M}$, written $I(X_i)$, is called the *indicator vector* of X_i, since it indicates which elements of S belong to X_i.

The problem is to find a minimal set of rows which together contain a 1 in every nonzero column. Equivalently, if we relabel the matrix so that columns correspond to subsets and rows to elements, the problem is to find a minimal system of representatives for the subsets. This problem is known to be difficult (Ref. 6, page 521).

The direct attack is to look at the rows taken 1, 2, 3, $\cdots$, m at a time, until a covering set is found; this finds a minimal covering set, but may take up to $2^m - 1$ steps. Several methods[7-13] have been given which are faster than the direct attack, but are still impractical for large m. Roth's algorithm[14] finds a locally minimal cover which has a high probability of being the minimal cover, for quite large values of m (up to several hundred).

However, for our purposes, the following extremely simple (and appropriately named) algorithm is adequate. It finds a covering set in at most $\frac{1}{2}m^2$ steps, but may not find a minimal cover.

THE GREEDY ALGORITHM

The algorithm proceeds inductively, starting with $\mathfrak{C} = \phi$ and (greedily) *adding to* $\mathfrak{C}$, *each time that particular* X_i *which will contribute the greatest number of new elements.*

We keep track of the elements in $\mathfrak{C}$ at each step by means of the indicator vector

$$I(\mathfrak{C}) = I(\bigcup_{X \in \mathfrak{C}} X)$$

and stop when this is equal to

$$I(\mathfrak{F}) = I(\bigcup_{X \in \mathfrak{F}} X).$$

1. *The Initial Step*

Set $\mathfrak{C} = \phi$, $I(\mathfrak{C}) = (0, 0, \cdots, 0)$.

2. *The Inductive Step*

Search through all $X_i \in \mathfrak{F}$ that are not in $\mathfrak{C}$ and find an X_k which maximizes the number of elements of S which are in X_k but not in $\mathfrak{C}$, i.e., which maximizes weight $(I(X_k)$. AND. NOT. $I(\mathfrak{C}))$. (The weight of a vector is the number of its nonzero components, $(a_1, \cdots, a_n)$. AND. $(b_1, \cdots, b_n) = (a_1$ AND $b_1, \cdots, a_n$ AND $b_n)$, NOT. $(a_1, \cdots, a_n) =$ (NOT $a_1, \cdots,$ NOT $a_n)$, and .OR. is defined similarly.) Break ties in any way.

Add X_k to $\mathfrak{C}$, and calculate the new $I(\mathfrak{C})$ = old $I(\mathfrak{C})$. OR. $I(X_k)$. Repeat Step 2 until $I(\mathfrak{C}) = I(\mathfrak{F})$; then stop.

Remarks:

(*i*) The greedy algorithm often finds a covering set which is close to minimal, although it is possible to construct examples when the minimal covering set contains two subsets while the greedy algorithm uses more than N subsets, for any preassigned value of N. Are such examples rare? The behavior of the algorithm for a random family $\mathfrak{F}$ seems to be unknown.

(*ii*) Since the algorithm involves simple calculations with binary vectors it may be easily programmed on a computer.

Example 1: The set of all paths from N_1 to N_4 in Fig. 6 consists of

$$(a_1 + a_2)(b_1 + b_2)(c_1 + c_2) = a_1b_1c_1 + a_1b_1c_2 + a_1b_2c_1 + a_1b_2c_2 + a_2b_1c_1 + a_2b_1c_2 + a_2b_2c_1 + a_2b_2c_2 .$$

Suppose it is desired to find a minimal subset of these paths which contains all the edges $S = \{a_1 , a_2 , b_1 , b_2 , c_1 , c_2\}$. $\mathfrak{F}$ consists of the following eight subsets of S, shown together with their indicator vectors.

i	X_i	$I(X_i)$
1	$a_1b_1c_1$	1 0 1 0 1 0
2	$a_1b_1c_2$	1 0 1 0 0 1
3	$a_1b_2c_1$	1 0 0 1 1 0
4	$a_1b_2c_2$	1 0 0 1 0 1
5	$a_2b_1c_1$	0 1 1 0 1 0
6	$a_2b_1c_2$	0 1 1 0 0 1
7	$a_2b_2c_1$	0 1 0 1 1 0
8	$a_2b_2c_2$	0 1 0 1 0 1

The greedy algorithm then proceeds as follows.

Step 1. $\mathfrak{IC} = \phi$, $I(\mathfrak{IC}) = 000000$.

Step 2. Weight $(I(X_i)$. AND. $111111) = 3$ for all i, so we pick X_1 (any X_i will do) and add it to $\mathfrak{IC}$: $\mathfrak{IC} = \{X_1\}$, $I(\mathfrak{IC}) = 101010$.

Step 2 again. Weight $(I(X_i)$. AND. $010101)$ is maximized by $i = 8$.

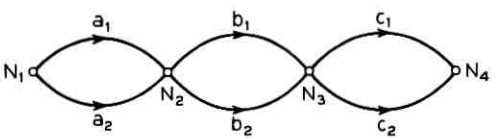


Fig. 6—An example.

Then $\mathfrak{C} = \{X_1 , X_8\}$, $I(\mathfrak{C}) = 101010.$ OR. $010101 = 111111.$ The algorithm terminates having found

$$\mathfrak{C} = \{a_1b_1c_1 , a_2b_2c_2\}$$

which is a correct solution.

Example 2: The greedy algorithm does not always find a minimal covering set, as the following example shows.

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$\mathfrak{F} = \{X_1 = \{1, 2, 3\}, X_2 = \{4, 5, 6\}, X_3 = \{1, 3, 4, 6\}.$$

The greedy algorithm finds $\mathfrak{C} = \{X_1 , X_2 , X_3\}$, while the minimal $\mathfrak{C}$ is $\{X_1 , X_2\}$.

V. FINDING A SMALL SET OF PATHS CONTAINING ALL EDGES

As before, let G be a directed graph with nodes $N_1 , N_2 , \cdots , N_n$. Let $\alpha_{\mu\nu}$ denote the set of paths from N_μ to N_ν.

Definition: A set $\beta_{\mu\nu}$ of paths from N_μ to N_ν is said to be a *spanning set* if every edge occurring in the set $\alpha_{\mu\nu}$ occurs in $\beta_{\mu\nu}$.

Example: In Fig. 7, the set of all paths from N_1 to N_2 is

$$\alpha_{12} = (a + b)(c + d) = ac + ad + bc + bd,$$

whereas an example of a spanning set is $\beta_{12} = ac + bd$.

The problem we consider in this section is to find a small spanning set $\beta_{\mu\nu}$. Finding a *minimal* spanning set appears difficult, and the only method we know is essentially an exhaustive search, as given in the next paragraph. The main algorithm of this section, algorithm B, gives a small spanning set $\beta_{\mu\nu}$ with a reasonable amount of computation.

Finding the Smallest Spanning Set $\beta_{\mu\nu}$ by Exhaustive Search

This may be accomplished by first applying the McNaughton–Yamada algorithm of Section III to produce a condensed list of all paths from N_μ to N_ν. Then truncate each expression S^* appearing in this list to $\Lambda + S$. (Since there is no need to go around a loop more than once in

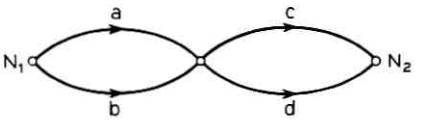


Fig. 7—An example.

succession, we can throw away the remaining terms of $S^* = \Lambda + S + S^2 + S^3 + \cdots$.) We now have a *finite* spanning set $\beta_{\mu\nu}$ and can use an exhaustive search to get a minimal set.

The difficulty with this method is that the number of terms obtained in the final list will be very large. To illustrate we apply the method to the four-node graph of Fig. 3. We found that the complete set of paths from N_1 to N_4 is

$$\alpha_{14} = ac^*b + ac^*d(ec^*d)^*(f + ec^*b).$$

Truncating each S^* to $\Lambda + S$, we obtain

$$a(\Lambda + c)b + a(\Lambda + c)d(\Lambda + e(\Lambda + c)d)(f + e(\Lambda + c)b),$$

which, when parentheses are removed, becomes

$$\begin{aligned} &ab + acb + adf + adeb + adecb + adedf + adedeb + adedecb \\ &\quad + adecdf + adecdeb + adecdecb + acdf + acdeb + acdecb + acdedf \\ &\quad + acdedeb + acdedecb + acdecdf + acdecdeb + acdecdecb. \end{aligned}$$

Then by inspection, or from the greedy algorithm of Section IV, we find that a minimal spanning set is for example

$$\beta_{14} = adf + adecb.$$

An Approximate Solution to the Problem–Algorithm B

We noticed in the above example that the difficulty was not in finding a minimal spanning set—indeed there are a large number of ways of choosing one—but rather in the very rapid increase in the number of terms to be handled. The algorithm to be described now keeps the lists involved small.

The basic idea is to follow the McNaughton–Yamada algorithm, but to use the greedy algorithm *twice at each step to reduce the complete path sets α_{ij}^k to small covering sets β_{ij}^k*.

Definition: Let β_{ij}^k be a set of paths from N_i to N_j containing no internal node N_p with $p > k$ and containing every edge appearing in α_{ij}^k.

Then $\beta_{\mu\nu}^n = \beta_{\mu\nu}$ is an example of a spanning set of paths from N_μ to N_ν, which is what we are seeking.

The algorithm will form the β_{ij}^k by induction on k. At each step we will keep a record of the edges in β_{ij}^k by means of its indicator vector $I(\beta_{ij}^k)$.

The inductive step proceeds as follows. Suppose β_{ij}^{k-1} is known for all i, j, and we wish to obtain β_{ij}^k (see Fig. 8). We restrict ourselves here

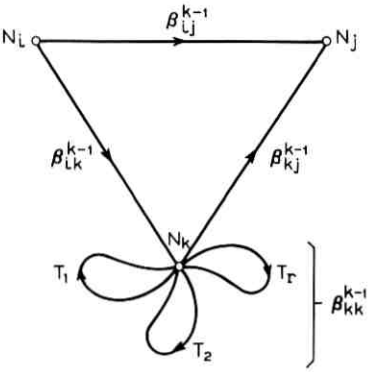


Fig. 8—The inductive step of algorithm B.

to the case when i, j and k are distinct, the other cases being left to the detailed statement of the algorithm.

Suppose $\beta_{kk}^{k-1} = T_1 + T_2 + \cdots + T_r$, where each T_a is a path from N_k to N_k. Then a possible choice for β_{ij}^{k} is

$$\beta_{ij}^{k-1} + \beta_{ik}^{k-1} T_{i_1} T_{i_2} \cdots T_{i_l} \beta_{kj}^{k-1} \tag{2}$$

where we have used just enough T_{i_r}'s to include all the edges in $T_1 + T_2 + \cdots + T_r$ that were not already contained in

$$\beta_{ij}^{k-1} + \beta_{ik}^{k-1} \beta_{kj}^{k-1}.$$

A better choice for β_{ij}^{k}, however, is to obtain (2) and then find a small spanning subset of (2) by the greedy algorithm.

We now give the algorithm.

ALGORITHM B

Each β_{ij}^{k} will have the form of a sum of strings of edges, without *'s or parentheses.

1. *The Initial Step*

Define β_{ij}^{0} for all $i, j = 1, \cdots, n$ by:

(1.1) if $i \neq j$,

$$\beta_{ij}^{0} = \begin{cases} \phi \text{ if there is no edge from } N_i \text{ to } N_j, \\ c_1 + c_2 + \cdots \text{ if edges labeled } c_1, c_2, \cdots \text{ join } N_i \text{ to } N_j; \end{cases}$$

(1.2) if $i = j$,

$$\beta_{ii}^{0} = \begin{cases} \Lambda \text{ if there is no edge from } N_i \text{ to itself,} \\ c_1 c_2 \cdots \text{ if edges labeled } c_1, c_2, \cdots \text{ join } N_i \text{ to itself.} \end{cases}$$

2. *The Inductive Step* (Refer to Fig. 8)

For $k = 1, 2, \cdots, n-1$, compute β_{ij}^{k} for all $i, j = 1, 2, \cdots, n$ as follows:

(2.1) If $k \neq i$, $k \neq j$:

(2.1.1) Let the terms of β_{kk}^{k-1} be

$$\beta_{kk}^{k-1} = T_1 + T_2 + \cdots + T_r .$$

(2.1.2) Form the indicator vector

$$I_1 = I(\beta_{ij}^{k-1}) .\text{OR}. I(\beta_{ik}^{k-1}) .\text{OR}. I(\beta_{kj}^{k-1}). \tag{2.1.3}$$

(This includes all the edges in the three sides of the triangle of Fig. 8.)
(2.1.4) Using the greedy algorithm, find a small subset of the T_a's in (2.1.1) which contains all the edges in

$$I(\beta_{kk}^{k-1}) .\text{AND. NOT.} I_1 ,$$

i.e., find a small subset of the terms T_a which includes all the new edges they contain. Let this subset be $T_{a_1} + T_{a_2} + \cdots + T_{a_m}$.

(2.1.5) Form the set

$$\beta_{ij}^{k-1} + \beta_{ik}^{k-1} T_{a_1} T_{a_2} \cdots T_{a_m} \beta_{kj}^{k-1} \cdots \tag{2.1.6}$$

(By construction, this now contains all the edges visible in Fig. 8.)
(2.1.7) Apply the greedy algorithm to the set (2.1.6) to find a small spanning subset. This is β_{ij}^{k}.

(2.2) If $i \neq j$ and $k = i$, replace (2.1.3) by $I_1 + I(\beta_{ij}^{i-1})$, and replace (2.1.6) by

$$T_{a_1} T_{a_2} \cdots T_{a_m} \beta_{ij}^{i-1}.$$

(2.3) If $i \neq j$ and $k = j$, replace (2.1.3) by $I_1 = I(\beta_{ij}^{j-1})$, and replace (2.1.6) by

$$\beta_{ij}^{j-1} T_{a_1} T_{a_2} \cdots T_{a_m} .$$

(2.4) If $i = j = k$:
Replace (2.1.3) by $I_1 = 0$ and replace steps (2.1.5) and (2.1.7) by

$$\beta_{kk}^{k} = T_{a_1} T_{a_2} \cdots T_{a_m} .$$

3. *The Final Step*

Use whichever of (2.1) to (2.4) is appropriate to calculate $\beta_{\mu\nu}^{n}$, the desired result.

Example: We use algorithm B to obtain a small spanning set β_{14} of paths from node 1 to node 4 in Fig. 3.

$\beta^0_{ij} = \beta^1_{ij}$

	j			
i	1	2	3	4
1	Λ	a	ϕ	ϕ
2	ϕ	c	d	b
3	ϕ	e	Λ	f
4	ϕ	ϕ	ϕ	Λ

β^2_{ij}

	j			
i	1	2	3	4
1	Λ	ac	acd	acb
2	ϕ	c	cd	cb
3	ϕ	ec	ecd	$f + ecb$
4	ϕ	ϕ	ϕ	Λ

The last step will be shown in detail.

(2.1.1) $\beta^2_{33} = ecd = T_1$.

(2.1.2) $I_1 = I(\beta^2_{14})$. OR. $I(\beta^2_{13})$. OR. $I(\beta^2_{34}) = 111000$. OR. 101100. OR. $011011 = 111111$.

(2.1.4) NOT. $I_1 = 000000$ so no T_a's need be used.

(2.1.5) $\beta^2_{14} + \beta^2_{13}\beta^2_{34} = acb + acd(f + ecb) = acb + acdf + acdecb$.

(2.1.7) From the greedy algorithm, $\beta^3_{14} = \beta^4_{14} = acdecb + acdf$, which is a minimal solution (although minimal solutions with shorter strings are possible, such as $acdeb + adf$).

Remarks: (*i*) If a fast version of the greedy algorithm is available, the computation time for algorithm B should not be much more than for the McNaughton–Yamada algorithm. (*ii*) An edge forming a loop of length one may be deleted from any sum of strings in which it appears more than once. If there are many such edges the algorithm should be modified to make a list of such edges and periodically delete duplicates from the β^k_{ij} . The modified algorithm would then give the improved solution $acdeb + adf$ to the above example. (*iii*) As in Section III, large networks may be handled by partitioning.

VI. FINDING A SMALL SET OF PATHS CONTAINING ALL EDGE-EDGE TRANSITIONS.

With $\alpha_{\mu\nu}$ defined as before, in this section we consider the problem of finding a small subset $\gamma_{\mu\nu}$ of $\alpha_{\mu\nu}$ with the property that every edge-edge transition appearing in any path from N_μ to N_ν appears in $\gamma_{\mu\nu}$.

For example, consider the graph of Fig. 6. Here the set of all paths from N_1 to N_4 is

$$\begin{aligned}\alpha_{14} &= (a_1 + a_2)(b_1 + b_2)(c_1 + c_2) \\ &= a_1b_1c_1 + a_1b_1c_2 + a_1b_2c_1 + a_1b_2c_2 \\ &\quad + a_2b_1c_1 + a_2b_1c_2 + a_2b_2c_1 + a_2b_2c_2 ,\end{aligned}$$

and an example of γ_{14} is

$$\gamma_{14} = a_1b_1c_1 + a_1b_2c_1 + a_2b_1c_2 + a_2b_2c_2 .$$

To check this we observe that α_{14} contains eight distinct edge-edge transitions:

$$a_1b_1 , a_1b_2 , a_2b_1 , a_2b_2 , b_1c_1 , b_1c_2 , b_2c_1 , b_2c_2 ,$$

and all of these appear in γ_{14} .

Of course γ_{14} is not unique, another example being

$$a_1b_1c_2 + a_1b_2c_2 + a_2b_1c_1 + a_2b_2c_1 .$$

The idea of the solution is to construct from G a new graph called the transition graph, G^T, which will have an edge for every edge-edge transition in G, and then to apply algorithm B to G^T.

Suppose then that G is given and it is desired to find $\gamma_{\mu\nu}$. First form the augmented graph $\bar{G}$ by adding to G a node N_o which is connected to N_μ by an edge z_1 , and a node N_{n+1} to which N_ν is connected by an edge z_2 (see Fig. 9).

From $\bar{G}$ we construct the transition graph G^T as follows. The nodes of G^T are (*i*) a node denoted (N_o), and (*ii*) nodes denoted $(e_{i1} , N_i) , \cdots ,$ (e_{ir_i} , N_i) if edges $e_{i1} , \cdots , e_{ir_i}$ enter N_i in $\bar{G}$, for $i = 1, 2, \cdots , n + 1$.

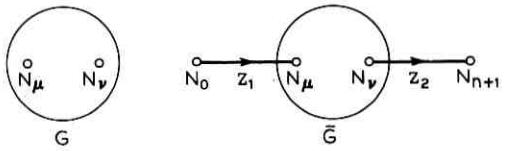
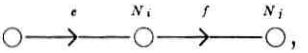
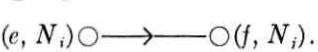


Fig. 9—Construction of augmented graph $\bar{G}$.

The edges of G^T are (*i*) an edge from (N_o) to $(z_1\ ,\ N_\mu)$, labeled (z_1); and (*ii*) for every edge-edge transition in $\bar{G}$,



there is a corresponding edge in G^T:



In general, we see that nodes of G^T have labels of the form (edge of $\bar{G}$, node of $\bar{G}$), and edges have labels of the form (edge-edge transition pair of $\bar{G}$).

By construction, apart from the edge (z_1) of G^T, there is a one-to-one correspondence between edge-edge transitions in $\bar{G}$ and edges of G^T.

To find $\gamma_{\mu\nu}$ we apply algorithm *B* to G^T. Each path through G^T from N_o to N_{n+1} will have the form

$$(z_1),\ (z_1e_{i_1}),\ (e_{i_1}\ ,\ e_{i_2}),\ (e_{i_2}\ ,\ e_{i_3}),\ \cdots\ ,\ (e_{i_r}z_2) \tag{3}$$

and this corresponds uniquely to the path

$$e_{i_1}\ ,\ e_{i_2}\ ,\ \cdots\ ,\ e_{i_r} \tag{4}$$

from N_μ to N_ν in G. The process of obtaining (4) from (3) will be called *contracting*.

We can now state the algorithm.

Algorithm C for Obtaining $\gamma_{\mu\nu}$

1. From G obtain $\bar{G}$ and then the transition graph G^T.

2. Apply algorithm B to find a small set of spanning paths from N_o to N_{n+1} in G^T.

3. Contract each of these paths to give a set of paths in G. This is $\gamma_{\mu\nu}$.

Example: Let G be the graph of Fig. 6. Then $\bar{G}$ and G^T are shown in Figs. 10–11.

Applying algorithm B, or in this case even by inspection, we see that a minimal spanning set for Fig. 11 is

$$\begin{aligned}
&(z_1)(z_1a_1)(a_1b_1)(b_1c_1)(c_1z_2)\\
+\ &(z_1)(z_1a_1)(a_1b_2)(b_2c_1)(c_1z_2)\\
+\ &(z_1)(z_1a_2)(a_2b_1)(b_1c_2)(c_2z_2)\\
+\ &(z_1)(z_1a_2)(a_2b_2)(b_2c_2)(c_2z_2),
\end{aligned}$$

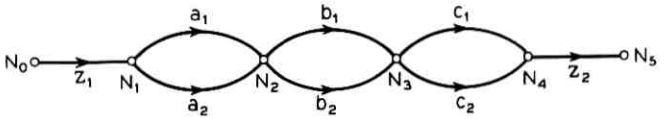


Fig. 10—Augmented graph $\bar{G}$ corresponding to Fig. 6.

which contracts to give the paths

$$a_1b_1c_1 + a_1b_2c_1 + a_2b_1c_2 + a_2b_2c_2 ,$$

the same solution as found before.

VII. FINDING THE MOST PROBABLE PATHS

A directed graph G is given with a conditional probability measure associated with the edges. More precisely G has nodes $N_1 , \cdots , N_n$, and associated with each edge e, directed say from N_i to N_j , is the conditional probability p_e that e will be traversed next, given that the last node reached was N_i .

We wish to find the most probable paths through the graph, starting at N_μ and ending at N_ν . The probability of a path is the product of the probabilities associated with the edges in the path.

In other words it is desired to find those paths P for which

$$\text{probability } (P) = \prod_{\substack{\text{all edges} \\ e \in P}} p_e$$

is the maximum, or is close to the maximum.

If we label each edge e of G with the "length"

$$q_e = -\log p_e$$

instead of with p_e , an equivalent problem is to find those paths P for which

$$\sum_{\text{all edges } e \, P} q_e$$

is the minimum, or is close to the minimum. In the new graph this corresponds to finding the shortest paths between N_μ and N_ν . This problem has been extensively studied and many good algorithms for its solution are available. We refer the reader to the recent survey by S. E. Dreyfus.[15] References 16 and 17 are earlier surveys covering a wide range of similar problems. The paper by H. Frank[18] is also relevant.

VIII. SUMMARY

Four questions which arise in testing a stored program for possible errors are stated quite generally in terms of listing the paths through

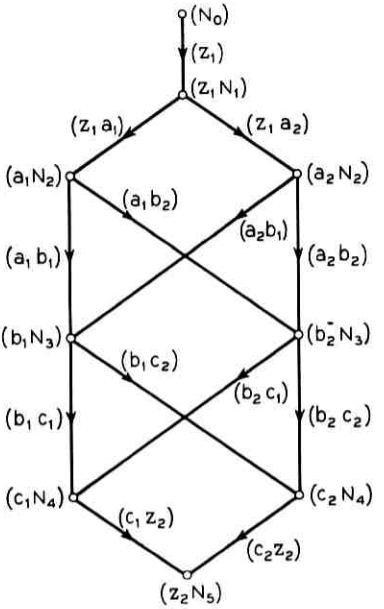


Fig. 11—Transition graph G^T corresponding to Fig. 10.

a directed graph. Question 1 may be answered for small graphs by the McNaughton–Yamada algorithm, and for large graphs by partitioning (Section III). Question 2 involves a difficult combinatorial problem, the minimal covering problem, a partial solution of which is given by the appropriately named greedy algorithm of Section IV. With the aid of the greedy algorithm, algorithm B solves question 2 (Section V). Question 3 is solved by the same method as question 2 (Section VI). Question 4 is shown to be equivalent to the widely studied "shortest-path problem," and references are given to the appropriate literature (Section VII).

IX. ACKNOWLEDGMENT

The author wishes to thank J. M. Scanlon for many interesting discussions and helpful suggestions.

REFERENCES

1. Harary, F., *Graph Theory*, Reading, Mass: Addison-Wesley, 1969.
2. Scanlon, J. M., personal communication.
3. McNaughton, R., and Yamada, H., "Regular Expressions and State Graphs for Automata," IRE Trans. Elec. Computers, *EC-9*, No. 1 (March 1960), pp. 39–47.

4. Hennie, F. C., *Finite-State Models for Logical Machines*, New York: Wiley, 1968.
5. Griswald, R. E., Poage, J. F., and Polonsky, I, P., *The SNOBOL4 Programming Language*, Englewood Cliffs, New Jersey: Prentice-Hall, 1968.
6. Mirsky, L., and Perfect, H., "Systems of Representatives," J. Math. Anal. and Appl., *15*, No. 3 (September 1966), pp. 520–568.
7. Breuer, M. A., "Simplification of the Covering Problem with Application to Boolean Expressions," J. Assoc. Comp. Mach., *17*, No. 1 (January 1970), pp. 166–181.
8. Cobham, A., Fridshal, R., and North, J. H., "An Application of Linear Programming to the Minimization of Boolean Functions," AIEE 2nd Annual Symposium on Switching Theory and Logical Design, 1961, pp. 3–10.
9. Geoffrion, A., "Integer Programming by Implicit Enumeration and Balas' Method," SIAM Review, *9*, No. 2 (April 1967), pp. 178–190.
10. House, R. W., Nelson, L. D., and Rado, T., "Computer Studies of a Certain Class of Linear Integer Problems," in *Recent Advances in Optimization Techniques*, edited by A. Lavi and T. Voge, New York: Wiley, 1966.
11. Lawler, E. L., "Covering Problems: Duality Relations and a New Method of Solution," J. SIAM Appl. Math., *14*, No. 5 (September 1966), pp. 1115–1132.
12. Mayoh, B. H., "On Finding Optimal Covers," Int. J. Comp. Math., *2*, No. 1 (January 1968), pp. 57–73.
13. McCluskey, E. J., Jr., "Minimization of Boolean Functions," B.S.T.J., *35*, No. 6 (November 1956), pp. 1417–1444.
14. Roth, R., "Computer Solutions to Minimum-Cover Problems," Operations Research, *17*, No. 3 (May–June 1969), pp. 455–465.
15. Dreyfus, S. E., "An Appraisal of Some Shortest-Path Algorithms," Operations Research, *17*, No. 3 (May–June 1969), pp. 395–412.
16. Fulkerson, D. R., "Flow Networks and Combinatorial Operations Research," Amer. Math. Monthly, *73*, No. 2 (February 1966), pp. 115–138.
17. Hu, T. C., "Recent Advances in Network Flows," SIAM Review, *10*, No. 3, (July 1968), pp. 354–359.
18. Frank, H., "Shortest Paths in Probabilistic Graphs," Operations Research, *17*, No. 4 (July–August 1969), pp. 583–599.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Wiring Telephone Apparatus from Computer-Generated Speech

By J. L. FLANAGAN, L. R. RABINER, R. W. SCHAFER, and J. D. DENMAN

(Manuscript received October 5, 1971)

Tape-recorded, spoken wiring instructions eliminate the need for a wireman to divert his eyes and hands from the equipment he is fabricating. A computer technique is described for automatically converting printed wire lists to synthetic speech. The technique was used to synthesize spoken wire lists for crossbar-4 equipment, and the result was tested informally on a production line at the Western Electric Company plant in Oklahoma City. No errors were made in wiring crossbar-4 circuitry from the computer-synthesized instructions.

I. INTRODUCTION

In many instances in fabricating and wiring telephone equipment, it is necessary for the wireman to use both hands and to visually "keep his place" in the equipment. Since it is inefficient and time consuming to divert either eyes or hands from the wiring task, a spoken presentation of the wire-list sequence is advantageous.

Tape-recorded, spoken wire lists have been used by Western Electric Company for switchgear wiring and cable forming at the Oklahoma City and Montgomery (Chicago) plants. The wire lists typically are read and recorded by a practiced announcer. The recordings are then checked and edited by another person in a separate listening operation. The final recording is then used in a cassette play-back whose start-stop control is wired to a footswitch. As the wireman needs items of the wiring sequence, he presses the footswitch for a time required to play back each item of the list. Because of the noisy environment he normally listens on an ear-insert earphone. The play-back normally is stopped while each connection is made. A typical wire list includes: lead length; color; beginning point; terminating point; and, sometimes, auxiliary instructions. Studies of the audio technique of wiring show accelerated training time and substantial improvements in quality and efficiency.

Fewer defects are found to occur and less time is needed to repair them.[1]

Wire lists for complex equipment are generally organized on computer cards. The audio technique therefore requires a listing of the card deck in a form convenient for the human announcer. The two human operations [(*i*) recording and (*ii*) editing] offer possibilities for errors to creep in. This sequence of operations is illustrated in the upper half of Fig. 1. Modifications in the wire list—made easily in the card deck and, generally, made often during the life of a typical list—require re-recording and re-editing of the audio tape. Consequently, there is considerable motivation to consider direct and automatic conversion of the card deck into a speech recording. One scheme for a direct and automatic generation of the spoken wire list uses synthetic speech and is illustrated in the lower half of Fig. 1.

II. COMPUTER-SYNTHESIZED INSTRUCTIONS

We recently have devised a computer technique for synthesizing speech from stored, low bit-rate data.[2,3] In its initial form the method has been applied to the synthesis of 7-digit telephone numbers, as might be used in an automatic intercept system. The system is implemented on one of the DDP-516 computers in the Acoustics Research

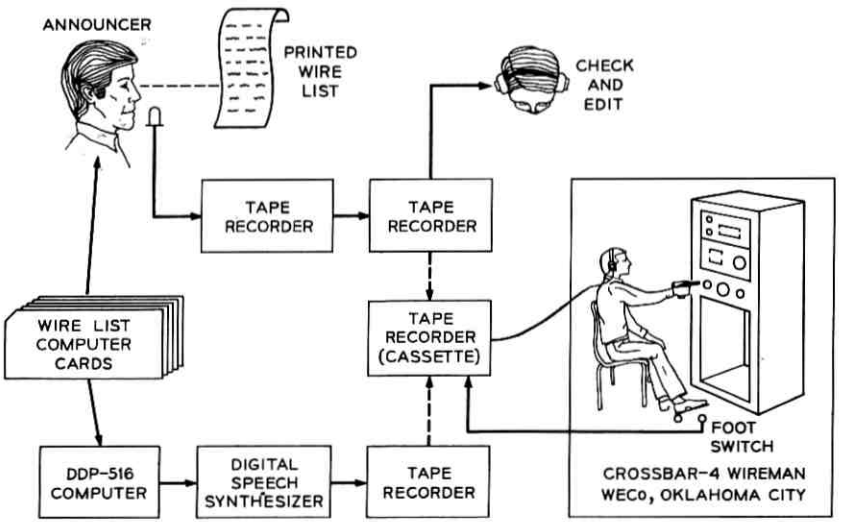


Fig. 1—Human and computer methods of preparing tape-recorded, spoken wiring instructions.

Department, and the major components of the system are shown in Fig. 2.

Individually-spoken words are analyzed in terms of their characteristic (formant) resonances, and the results described by a data rate of 530 bits per second.[2] These data are stored in the fast-access disc of the DDP-516 facility and constitute the vocabulary for the voice-response system. When a word-sequence is demanded by a control (answer-back) program, the formant data for the successive words are accessed from disc and are concatenated "head-to-tail." An analysis is made of the context in which the library words are to appear, and duration and voice pitch data are computed for each word by the synthesis program. The formant data at the boundaries between words are interpolated smoothly by a specially designed algorithm in the synthesis program. Finally, the formant and pitch data calculated for the required utterance are sent to a hardware digital filter whose resonances simulate those of the human vocal tract.[4,5] Digital-to-analog conversion of the filter output yields a synthetic speech signal.

We have used this voice-response system with simplified duration and pitch rules to synthesize wire lists for crossbar-4 switchgear. In this application the card deck comprising the wire list is simply put into the card reader of the DDP-516 and each wiring instruction is synthesized. A computer-controlled analog tape recorder records the output of the D/A converter, and this tape goes directly to the wireman's cassette. The items of the crossbar-4 wire list which were synthesized are shown in Table I. The synthesized list contained a total of 58 complete wire wrap instructions.

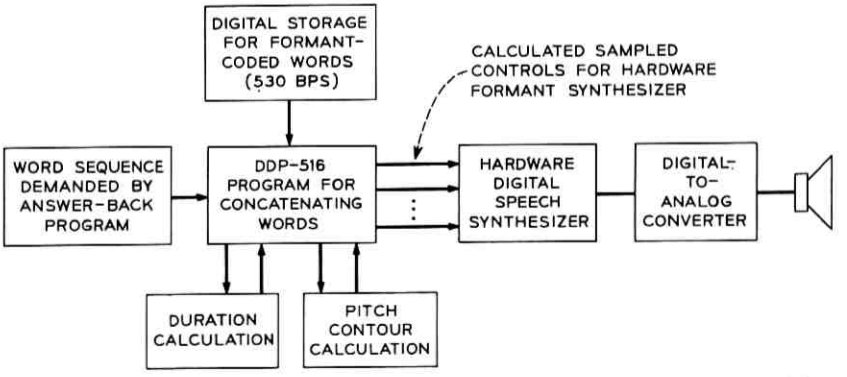


Fig. 2—DDP-516 computer system for automatic synthesis of spoken wiring instructions.

TABLE I—SYNTHESIZED WIRE LIST

SWJ99234T2
issue 7

List combination 1, 2, C, D
Apparatus not otherwise identified shall be considered to be a relay.
Wire colors not otherwise identified shall be green and the length will be in inches.
The sequence of operations will be: wire length, starting terminal and apparatus designation, ending terminal and apparatus designation.

	12	27A terminal strip	6R1
	3	28A terminal strip	2A tube socket
	11	37A terminal strip	4R1
	2	38A terminal strip	6A tube socket
	18	15A terminal strip	1 lower TP
	19	16A terminal strip	1 TP
	20	25A terminal strip	7 break TP
	4-1/2	26A terminal strip	1 top P
	20	35A terminal strip	7 TP
	16	36A terminal strip	2A repeat coil
	11	14A terminal strip	6 make R1
	4-1/2	24A terminal strip	2 top P
	11	33A terminal strip	4 make R1
	19	34A terminal strip	1 upper TP
	11	12A terminal strip	2 R1
	10-1/2	32A terminal strip	upper R1
	3-1/2	1A tube socket	8 top P
	4	5A tube socket	9 top P
	3-1/2	6A tube socket	4 top P
	3	8A tube socket	7 top P
	17	1 top P	6 make TP
	9	2 top P	10 break R1
	10	3 top P	8 make R1
	10	4 top P	top A capacitor
	2-1/2	5 top P	top D capacitor
	9	8 top P	bottom A capacitor
	2-1/2	bottom C2 capacitor	bottom C1 capacitor
	2-1/2	top C2 capacitor	top C1 capacitor
	5	bottom C1 capacitor	8 R1
	3-1/2	top C1 capacitor	top C resistor
	3-1/2	bottom C resistor	10 R1
	5	4 break R1	4A repeat coil
	5	6 break R1	7A repeat coil
	3	bottom A capacitor	8A repeat coil
	4	top A capacitor	3A repeat coil
	7	5A repeat coil	6 TP
	4	1 lower TP	right E capacitor
	4.5	1 upper TP	lower E capacitor
	2	2 lower TP	4 make TP
	3	2 upper TP	4 TP
	3	1A repeat coil	top A resistor
	3.5	6A repeat coil	bottom A resistor
	3.5	bottom A resistor	bottom B capacitor
	3.5	top A resistor	top B
	20	17A terminal strip	5 TP
	20	18A terminal strip	3 TP
	11.5	22A terminal strip	11 R1
	8	6 top P	top R1 upper terminal
	5.5	bottom D capacitor	bottom R1 lower terminal
	7	top R1 upper terminal	5A repeat coil
Red	2.5	11A terminal strip	4A tube socket
Black	3	31A terminal strip	7A tube socket
Black	9.5	7A tube socket	8 break R1
Black	2.5	8 break R1	2 make R1
Black	3	2 make R1	10 make R1
Black	10	10 make R1	top TP terminal

We have made informal experiments at the Western Electric Company plant in Oklahoma City, where we asked the wireman (or, rather, wiregirl) to use the synthetic speech recording to fabricate crossbar-4 equipment. A photograph of the wireman simultaneously wire-wrapping five identical chassis of crossbar-4 equipment is shown in Fig. 3a. The footswitch control of the synthetic speech tape on the cassette is shown in Fig. 3b.

Fig. 3a—Wireman on production line at Western Electric Company plant in Oklahoma City. The wireman is fabricating crossbar-4 equipment from the computer-spoken wire list.

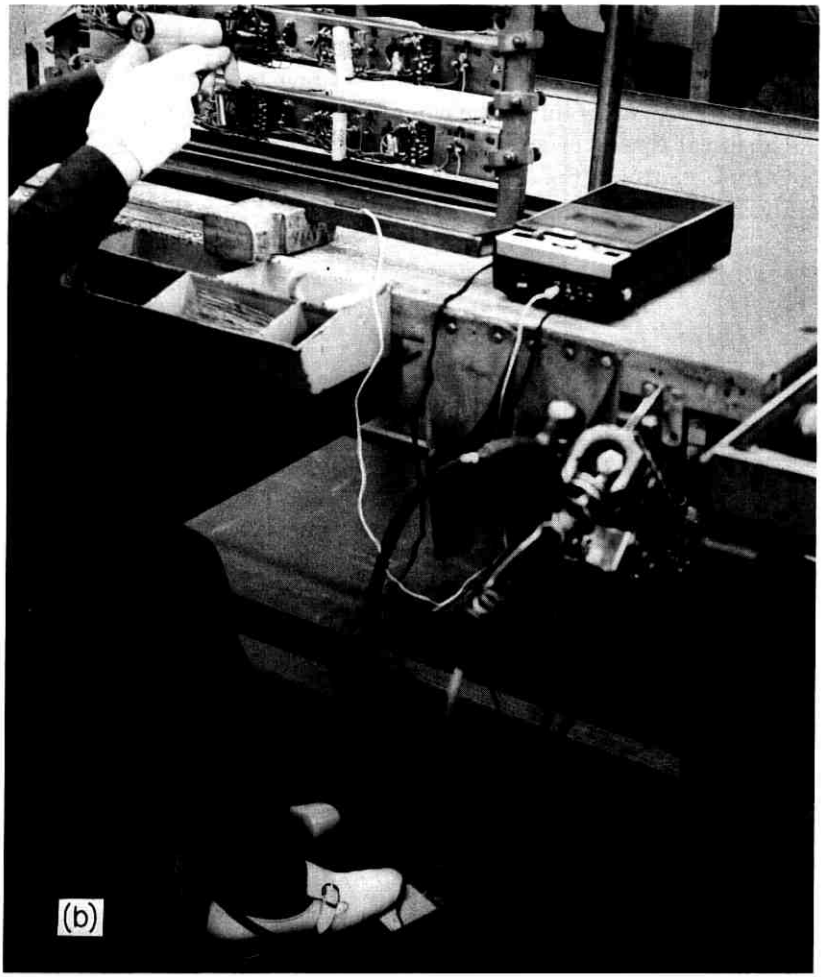


Fig. 3b—Wireman's footswitch to control the computer–synthesized speech tape.

While the quality of the synthetic speech is far from natural, the wireman (who had never heard synthetic speech) experienced no difficulty in using it immediately and, in fact, remarked that the "caricatured" nature of the synthetic signal seemed better for the noisy plant environment than natural speech. About 15 minutes/chassis are needed to wire the equipment shown in Fig. 3, and no wiring errors were made in the informal tests on the five chassis.

For speech material with as small a vocabulary and as rigid a con-

textual format as a wire list, the flexibility and storage economy of the synthesis system is not critically needed. In the case of brief lists, digital recordings of the naturally spoken vocabulary words may be made on the DDP-516 disc and these words can be concatenated automatically by the control program. This simpler approach does not, of course, permit smooth, natural joining of the words into a sentence, but utterances such as the components of a wire list can be rendered reasonably well virtually in isolation. This approach would be exceedingly economical in that no pre-analysis computation of formant data is required to establish the machine vocabulary, and all the advantages of automatic, computer-generation of the spoken instructions are retained.

One final comment may be in order about computer-generation of spoken wire lists. The human-pronounced list which had been in use for the crossbar-4 wiring had a very obvious pausal error throughout. (Look, for example, at the first item in Table I.) The girl announcer who recorded the tapes, and who apparently was unfamiliar with the wiring operation, consistently read the items as "Twelve (pause) Twenty-seven (pause) A terminal strip." The computer, although speaking with a machine accent, never makes this mistake.

REFERENCES

1. Kopack, J. A., and Mulligan, D. T., "Audio Aid Reduces Wiring Times," Assembly Eng., (October 1971), pp. 44–46.
2. Rabiner, L. R., Schafer, R. W., and Flanagan, J. L., "Computer Synthesis of Speech by Concatenation of Formant-Coded Words," B.S.T.J., *50*, No. 5 (May–June 1971), pp. 1541–1558.
3. Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schafer, R. W., and Umeda, N., "Synthetic Voices for Computers," IEEE Spectrum, *7*, (October 1970), pp. 22–45.
4. Rabiner, L. R., Jackson, L. B., Schafer, R. W., and Coker, C. H., "A Hardware Realization of a Digital Formant Speech Synthesizer," IEEE Trans. Commun. Tech., *COM-19*, No. 6 (December 1971), pp. 1016–1020.
5. Flanagan, J. L., "Focal Points in Speech Communication Research," IEEE Trans. Commun. Tech., *COM-19*, No. 6 (December 1971), pp. 1006–1015.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Man-Machine Interaction in Human-Face Identification

By A. J. GOLDSTEIN, L. D. HARMON, and A. B. LESK

(Manuscript received September 20, 1971)

How well can a computer identify a human face which is described by a person who is inspecting a photograph? We give an account of an interactive system that takes advantage both of the human's superiority in detecting and describing noteworthy features and of the machine's superiority in making decisions based on accurate knowledge of population statistics of stored face-features. Experiments using a population of 255 faces and 10 or fewer feature-descriptions showed that the population containing the described individual could be narrowed down to less than 4 percent in 99 percent of all trials.

I. INTRODUCTION

In a previous report[1] we described experiments in human-face recognition which were intended to establish a foundation for extended study. Those experiments provided a large body of reliable quantitative data based on 21 feature-descriptions of 255 human faces. These 21-dimensional vectors were shown to be sufficient for accurate individual identification, both by human and by computer search.*

The objective, then and now, is to explore new techniques for obtaining accurate recognition of vectors given imprecise component values. Our procedures involve searching through a population of vectors to retrieve one, a "target," whose components best match a searcher's imprecise specification.

There are two obvious kinds of such recognition and retrieval, just as in fingerprint-file search. One is that of finding the best match between an unidentified individual and a member of a file population. The other is that of assigning an individual to one of a number of

* Our population consisted of 255 white males aged 20–50 with no eyeglasses, facial hair, scars, or notable deformities. A panel of 10 observers independently evaluated 21 features (shown in Fig. 1) for each face. The average value of the observers' votes was used as the "official" description of each face-feature. Although individual feature-descriptions are restricted to *integral* values, averaging the panel's votes provides non-integral official descriptions. Reference 1 contains a detailed discussion of the features and population used.

predefined classes according to some systematic scheme. Ours is the first approach, matching, though the techniques developed could readily be used for the second, cataloging.

In our previous work, a subject was given a set of photographs of human faces and an official description of one of them. He was required to select that photograph which best matched the description. In the experiments reported here, the subject was shown a picture and was asked to describe it to a computer using features from a list given to him. The computer then searched a population of stored descriptions for best fit to the description furnished by the subject. In both studies we ran supplementary experiments employing computer simulation to establish theoretical limits of human performance under certain modeling assumptions.

In the earlier face-identification procedures, isolation was based on a binary-decision technique. At each step in the search, the population was progressively reduced by using a quantitative feature-description to determine which members of the remaining subset would be retained. On the average, eight feature-descriptions were required to isolate a face in a population of 255 males; about 50 percent correct identification was obtained. The binary process, however, obviously insured doom given just one error in the sequence.

A more lenient process is rank-ordering. If one ranks population members according to some goodness-of-fit criterion, any reasonably accurate description can be expected to place the target high on the rank-ordered list. Such a system can be quite useful in focusing attention on a small subset of the population that has high probability of containing the target. Population-reduction techniques like this are well-known to be useful in many tasks, from fingerprint-file search[2] and script recognition[3] to document retrieval.[4]

The present report deals with a real-time man-machine interactive system for human-face identification. The study has three main objectives:

(*i*) To develop a decision-making technique which replaces the earlier error-sensitive binary-decision selection process by a more forgiving rank-ordering process,

(*ii*) To design algorithms for optimizing the man-machine system so that we can take advantage of both the human's superiority in detecting noteworthy features and the machine's superiority in making decisions based on accurate knowledge of population statistics, and

(*iii*) To devise simple yet effective measures of performance.

II. SYSTEM DESIGN

The system design can be understood by considering our experimental procedure. A subject at a remote computer-terminal is given a photograph of one member of the population. He describes this target face to the computer using descriptive features chosen from a permitted set. The aim is to have the computer identify the target from the subject's description of it.

Subjects in our experiments used three-view photographs of target faces (two examples are shown in Fig. 6). The set of features from which descriptions were drawn is shown in Fig. 1.

In our experiments, features may be chosen by the subject or by the computer which uses an automatic feature-selection algorithm. There are three alternative modes of feature selection: the subject may choose all features, or he may choose some and then let the computer choose the rest, or the computer may choose all features.

After each feature description, the computer assigns a goodness-of-fit measure (a "weight") to each member of the population. This weight represents the similarity of the subject's description to the official description of each member of the population. At each feature-description step, the population is ranked by weight. After a predetermined number of steps, the process is terminated. We evaluate performance with respect to the target's rank and weight. An illustrative printout of one "portrait"* appears in Fig. 2.

Two aspects of system design are crucial: the weight-assignment algorithm and the feature-selection algorithm. They are described below. Following that, we discuss two critical experimental requirements, stopping criteria and measures of performance. The experiments reported in the succeeding section were designed to show how various modes of feature selection affected system performance.

2.1 *Weight Assignment*

The algorithm used to assign weights at each step must maintain a reasonable balance between penalizing descriptive errors so heavily that recovery from a mistake is impossible and penalizing these errors so lightly that no significant reduction of the population is achieved. The penalties assigned should distinguish between a minor descriptive error (e.g., medium-long vs long nose-length) from which recovery should be easy, and a major error (e.g., short vs long nose-length) from which recovery should be more difficult.

* A portrait is defined as a description consisting of a set of *integral* feature-values assigned by a subject; the subject is said to "portray" the target.

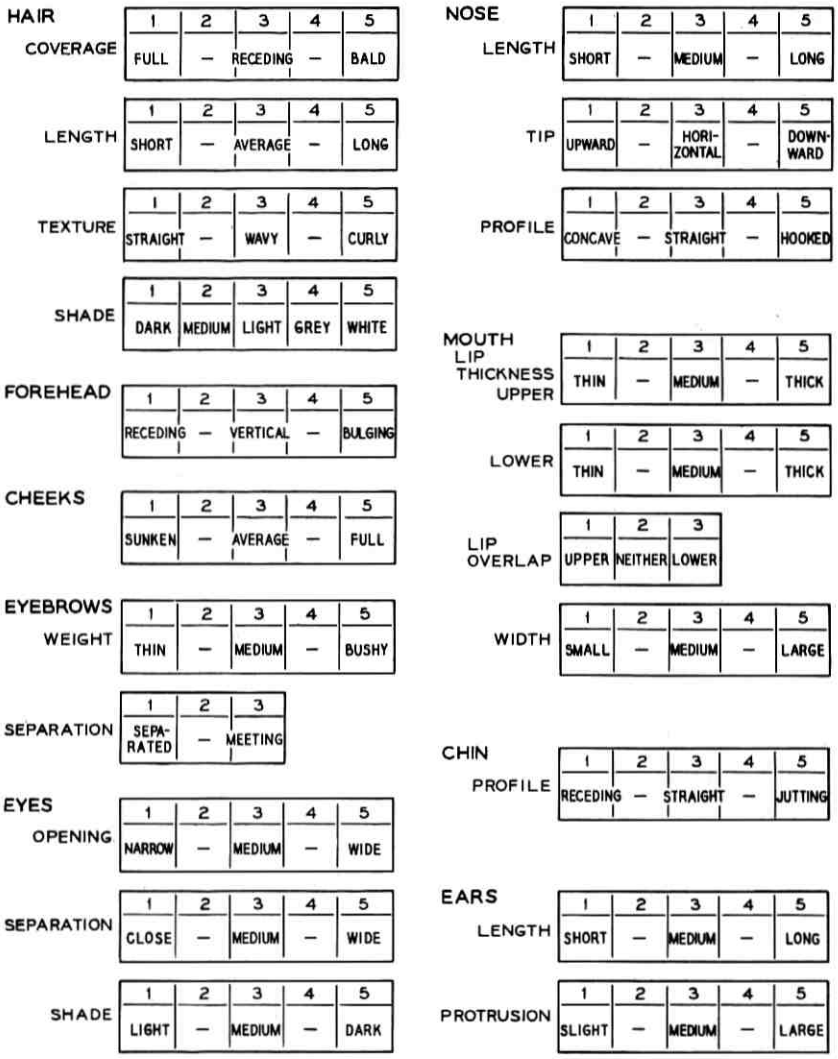


Fig. 1—Set of 21 face-features and their allowable values used for all experiments.

We chose

$$\sum_{i=1}^{s} |v_i - \hat{v}_i|^k = \sum_{i=1}^{s} \Delta_i^k$$

as the general form of an individual's weight at step s. For the feature described at step i, v_i is the individual's official value, $\hat{v}_i$ is the value

```
DESCRIBE NEXT PICTURE.

 FEATURE
      EYEBR0W WT.
 THIN        MEDIUM        BUSHY
   1     2     3     4     5
=1
   93  244  183  223  159
 1.00 1.00 1.00 1.00 0.82

 FEATURE
      EAR LENGTH
 SH0RT       MEDIUM       L0NG
   1     2     3     4     5
=1
   72  244  175   93   43
 1.00 1.00 0.82 0.67 0.66

 FEATURE
      LIP 0VERLAP
 UPPER        NEITHER      L0WER
   1            2            3
=1
   72  226  114  122   76
 1.00 0.73 0.66 0.61 0.60

 FEATURE
      HAIR TEXTURE
 STRAIGHT    WAVY         CURLY
   1     2     3     4     5
=4
   76  122   32  244   52
 1.00 0.74 0.56 0.55 0.50

 FEATURE

      AUT0MATIC FEATURE SELECTI0N
 ******EYE SHADE
 LIGHT        MEDIUM       DARK
   1     2     3     4     5
=3
   76   52   72  221  191
 1.00 0.56 0.45 0.38 0.36

 ******EYEBR0W SEP.
 SEPARATE    MEDIUM       MEETING
   1           2            3
=2
   76  147   52   84   72
 1.00 0.50 0.42 0.37 0.34

 ******EYE 0PENING
 NARR0W       MEDIUM       WIDE
   1     2     3     4     5
=2
   76   72  226   26  191
 1.00 0.51 0.40 0.38 0.36

 ******UPPER LIP
 THIN         MEDIUM       THICK
   1     2     3     4     5
=3
   76  191   72  221   52
 1.00 0.33 0.28 0.23 0.21

 ******HAIR SHADE
 DARK  MED.  LT.    GRAY  WHT.
   1     2     3     4     5
=2
   76  221   72  226  191
 1.00 0.34 0.34 0.33 0.25

 ******L0WER LIP
 THIN         MEDIUM       THICK
   1     2     3     4     5
=1
   76   72  221   84  191
 1.00 0.19 0.13 0.12 0.11

 PLEASE TYPE TARGET NUMBER.
=76

 0RDER FEATURE          DESCRIPTI0N      RANK
                        Y0U    AVG.    N0.    %
   1   EYEBR0W WT.       1     2.2     27   10.2
   2   EAR LENGTH        1     2.3      8    2.7
   3   LIP 0VERLAP       1     1.2      5    1.6
   4   HAIR TEXTURE      4     3.0      1    0.

   5   EYE SHADE         3     2.7      1    0.
   6   EYEBR0W SEP.      2     1.3      1    0.
   7   EYE 0PENING       2     2.6      1    0.
   8   UPPER LIP         3     2.9      1    0.
   9   HAIR SHADE        2     1.5      1    0.
  10   L0WER LIP         1     2.3      1    0.
```

Fig. 2—Printout of one interactive dialog. Computer requested feature; subject picked *Eyebrow Weight*. Computer printed allowable feature-values; subject voted *thin*. In next two lines computer displayed calculated weights of the top five individuals. First four faces, 93 · · · 223, tied with relative weights 1.00. Face no. 159 in fifth place was weighted 0.82 relatively. By step three the target (no. 76) was in fifth place, advancing to first rank by step four despite deliberately introduced errors on first two steps. Subject changed to AFS at step five, whereupon computer specified *Eye Shade*. Nearest neighbors were gradually separated; by step 10 the closest had relative weight of only 0.19. Portrait automatically terminated at step 10. Summary compares subject's assignments with official values ("AVG."); also displayed is target's rank at each step and percentage of population with higher rank.

assigned by the subject, and Δ_i is the magnitude of the difference between them.

A number of variants of this formulation were tested. In particular we found that $k = 1$ yielded results as good as or better than any other value of k. We also considered the effect of quantization error

arising from comparing the integral feature-values used by subjects to the non-integral official feature-values. For each feature description this error is at most 0.5. Alternative formulations of weight functions intended to minimize the effect of this error degraded performance. Our earliest formulations of weight used an exponential form. While presently unessential, the exponential has survived in our computational algorithms. Consequently, with $k = 1$ and with no compensation for quantizing effects, the weight assignment is

$$W = \exp\left(-\sum_{i=1}^{s} \Delta_i\right).$$

2.2 *Automatic Feature-Selection*

As noted above, features may be selected either by the subject or by the computer. The two methods have complementary advantages: The subject possesses exhaustive knowledge of the face he is portraying, but he knows very little about the characteristics of the population stored in the machine; conversely, the machine does not know who the target is, but it does possess the official descriptions of all population members and their goodnesses-of-fit to the target description.

We wish to find if the advantages of human and of computer feature-selection can be usefully combined, where the human can take advantage of *extreme* features, while the computer can utilize *discriminating* features.

An *extreme* feature of a target is a feature whose official value is near an extreme of that feature's range; e.g., long hair, short nose, small mouth. This classification does not depend on the target's other feature values or those of the population. It depends only on the feature's value and range.

Conversely, a *discriminating* feature is a purely relative concept, based on the population and the target description up to any given step. At each step, we refer to a feature as discriminating if its description will distinguish among those individuals whose official descriptions match the partial portrait well (i.e., the individuals who have large weights). Whether a feature is discriminating depends on the statistics of feature-value distribution over the population.

We wish to develop an automatic-feature-selection procedure that chooses the most discriminating feature available as the next one to be described in a portrait. How can we decide when a feature is discriminating?

Consider the two hypothetical distributions of official feature-values

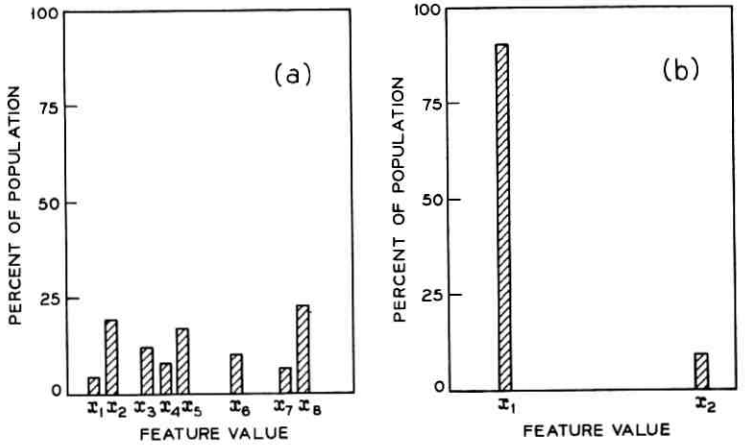


Fig. 3—Two types of distributions of official feature-values: (a) Relatively uniform distribution represents *discriminating* feature. (b) Relatively non-uniform distribution indicates a *nondiscriminating* feature.

shown in Fig. 3. If feature b were used, and if the target's value happened to be x_2, then the target would be well separated from the rest of the population. It is much more likely, however, that the target's value would be x_1, in which case the separation of population members would be poor. If feature a were used, one would always obtain some intermediate amount of population separation. In the extreme, if all members of the population had the same value of a particular feature, say very long ears, then the use of that feature would not lead to population separation. Conversely if the values were uniformly distributed over the population, maximum discrimination and most effective separation would be obtained.

In considering feature-value distributions, it is undesirable to utilize the official description of *every* member of the population for all unused features. Not only would this increase cost, but it would degrade performance. This can be seen from the following argument: The aim of automatic feature-selection (AFS) is to find a feature which will decrease the number of individuals who are described well by the portrait thus far. The distribution of feature values among those individuals may be completely different from the distribution in the whole population. If AFS considered all individuals, the distinguishing characteristics of the high-ranking individuals would be obscured by those of the overwhelming number of low-ranking individuals. To avoid waste of one's knowledge of the partial portrait, AFS considers the distribution of feature values only in the subset of the population

which the portrait describes well, although the feature chosen will be used to rerank the *entire* population. This subset should include the individuals who could easily attain first place in the rank-ordered population. In practice, we found that to consider all individuals with weight $\geqq 0.7$ times the weight of the first-ranked one, but at least 10 individuals, was effective.

As a result of the above arguments, we implemented an AFS procedure which chooses as the next feature that one for which the distribution of the feature-values of the high-ranking individuals is most nearly uniform. This will be the most-discriminating feature in the sense of efficient identification. Analytical details of the procedure are given in Appendix A.

2.3 *Stopping Criteria*

The portrait composition must continue for enough steps to insure accuracy. On the other hand, too many steps lead to subject fatigue and boredom. The rule which governs when portrait composition stops should satisfy both these requirements.

A stopping rule may be *dynamic* and depend on the ranks and/or the weights at each step, or the rule may be *static*, e.g., stop after a predetermined step. Our earlier experiments, employing a human binary-search process,[1] showed that, on the average, fewer than eight features were used when a target was successfully identified. One might conjecture that with 5-valued features some 2.3 bits of information could be available at each step, and so the present experiments should require fewer than 8 steps for isolation, and not less than $\log_2 255/\log_2 5 = 3.5$.

This argument, and information from trial runs indicating that fatigue and boredom commenced after the subject judged about ten features, were used to arrive at a static stopping-rule of ten steps. Experimental results have shown this to provide adequate accuracy. The data we obtained permitted us to formulate an efficient dynamic stopping-rule for future use; it is described in Section IV.

2.4 *Measures of Performance*

A binary search-procedure may be evaluated by whether and at what stage the target is ultimately isolated, or at what stage the target is rejected and the size of the smallest subset that contained the target. Meaningful measures of performance for a rank-ordering procedure are less obvious.

One useful measure, population reduction, can be transferred directly from binary search to rank ordering. We can consider the size of the

subset of the population with rank greater than that of the target, and how rapidly the population is reduced to that size. The concept of absolute isolation is thus replaced by one of relative identification.

We measure the population reduction at each step by the rank of the target. Since his rank usually changes from step to step, we use as an overall measure of performance the mean rank of the target from the sixth through tenth steps. The first five steps are not included because the target's rank then is usually large and changing rapidly.

Population reduction shows whether the target is separated from the rest of the population. It does not reveal, however, the *extent* of that separation. To do this, a "confidence" measure was introduced. It is based on the weights of the individuals in the ranked list, as follows: If the target is ranked first, his confidence is the ratio of his weight to that of the second-ranked individual; otherwise, the target's confidence is equal to the ratio of his weight to that of the first-ranked individual. A confidence value less than 1.0 denotes failure to place the target in first rank; confidence values greater than 1.0 correspond to varying degrees of success. Obviously, the magnitude of the confidence measure depends on the weighting function being used.

Confidence and rank are useful in evaluating a single portrait; their averages can be used to compare several sets of portraits. A third measure we find useful is the *rank cross-section*; this is meaningful only for comparing sets of portraits. For a set of portraits, the rank cross-section is the frequency with which targets reach or exceed a given threshold rank (e.g., first rank, or top 2 percent of population, etc.) at each step of a portrait. This indicates the average speed and extent of a target's rise in rank.

However, a target does not necessarily always rise in rank. A faulty feature-judgment may worsen his position. The weighting scheme is forgiving in that it permits recovery from a subject's error in feature judgment. Another way of viewing this is that once the target is entrenched in first place, i.e., has a large confidence, it takes a large error in judgment to displace him.

We can express this quantitatively as follows: Suppose the target is in first rank; let him have confidence c, and let the next feature judgment for him have an error Δ. Suppose that the error for the second-ranked individual is 0. Then with the weighting scheme that was adopted, we find that if $\Delta > \ln c$, the ranks will be reversed. Thus,

when confidence $c \leqq$	1.6	2.7	4.5	7.4	12.2	20.0
reversal occurs if $\Delta >$	0.5	1.0	1.5	2.0	2.5	3.0.

Data on subject error (see Section 3.1.1.1) show that 95 percent of the time $\Delta \leqq 1.0$. Thus a first-ranked target with confidence 2.7 or greater is rarely dislodged.

III. EXPERIMENTS

3.1 *Human Experiments*

An interactive experiment was run to evaluate the effectiveness of our overall system and to test the relative utility of three different modes of operation.* In one mode the subject selects *every* feature he describes to the computer, using first those he considers most extreme for the target. We shall refer to this mode as "NO AFS" (i.e., no automatic feature-selection). In another mode, termed "ALL AFS," the subject simply assigns feature values for each feature specified by the computer which is operating in the automatic-feature-selection mode described earlier. A third mode, termed "MIXED," requires a subject to select features until he decides there are no more he considers outstanding, then to invoke AFS.

We expected subject selection of extreme features to enhance separation, at least for the first few features, for many members of the population. When there are no extreme features to use, then computer selection of discriminating features should facilitate target separation. We expected that the mixed mode of operation, taking advantage of the best capability of both human and computer, would yield best results as measured by confidence and rank.

Fifteen subjects were used (13M, 2F). Twenty-one features were made available, as illustrated in Fig. 1. Each subject participated in three separate sessions, one in the NO-AFS mode, one in MIXED, and one in ALL AFS. Each of the 15 subjects, portraying 15 targets, provided us with 225 portraits. Five targets were portrayed in each session. Fifteen different targets were used; each subject thus portrayed all targets. The targets were individually selected at random from our population of 255; as an ensemble they were shown to preserve the feature distributions of the entire population. To minimize possible effects of learning, we randomized the order in which subjects used the three modes of feature selection and the order in which they portrayed the targets.

At the beginning of the experiment each subject was given 20–30 minutes of verbal instruction to familiarize him with the feature set. This used a collection of sample faces that were not employed in the

* The program which was used is described in Ref. 5.

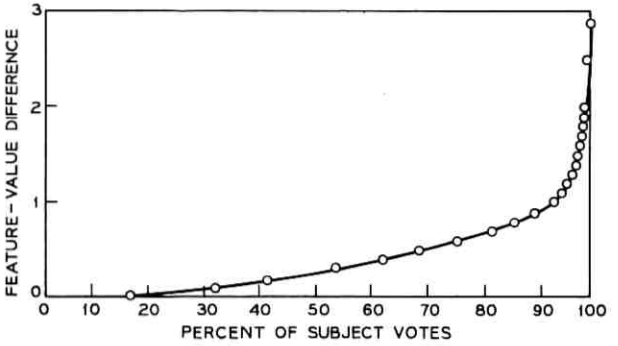


Fig. 4—Cumulative distribution of differences between subject votes and official values. The difference was never greater than 1.0 for 95 percent of the votes.

experiment. The subject then observed the experimenter portraying one target.

At the beginning of each session, the subject portrayed one practice target using the same mode of description (NO AFS, MIXED, or ALL AFS) to be employed in the experimental session. In all cases the subjects viewed the target's photograph while describing his features.

3.1.1 *Results*

3.1.1.1 *Feature-Judgment Reliability.* Our 15 subjects, making 2250 total judgments (15 subjects × 15 targets × 10 features), were in excellent agreement with the official feature-values. This can be seen in Fig. 4 which displays the cumulative distribution of magnitudes of the differences (Δ) between the subject judgments and the official values. In 95 percent of the 2250 judgments, the Δ was $\leqq 1.0$ (the maximum Δ is 4.0 for a 5-valued feature).* No judgments were as much as 3.0 off, only two were > 2.0 off, and only 24 of the 2250 judgments were different from the official values by more than 1.5.

Standard deviations were computed for the distributions of subject judgments, feature by feature. In both the ALL-AFS and the NO-AFS experiments, the standard deviation ranged from 0.42 to 1.1. The standard-deviation values for each feature are similar for ALL AFS and NO AFS, indicating no significant difference in subject accuracy as a function of whether feature selection is active or passive.

3.1.1.2 *Identification Accuracy.* The confidence and rank data,

* With the exception of two three-valued features. The data of Fig. 4, which include all 21 features, are not significantly changed by deleting the contributions of the two three-valued features.

averaged over all subjects and all targets, are shown in Fig. 5. For the combined 225 portraits, the mean confidence at step 10 was 5.65, and the mean rank over the sixth through tenth steps was 4.12. For 75 MIXED portraits, the mean confidence and rank were 6.79 and 2.75 respectively, while for 75 ALL-AFS portraits the corresponding figures were 4.41 and 6.71. The results of the 75 NO-AFS experiments were intermediate; mean confidence was 5.74, and mean rank was 2.91.

Subject performance varied considerably. Both the average confidence and the average rank had a range of 6:1 (from best to worst subjects). One subject's performance was consistently poor. When his scores are deleted, the average rank improves from 4.12 to 3.70, and the average confidence improves from 5.65 to 5.80.

To test for improved performance with practice during the course of the experiment, the data for each subject were examined according to their temporal sequence. No trends were observed.

The 15 targets received a rather wide range of performance indices. Number 99 had an average confidence measure of 20.3 (compared to the 15-target mean of 5.65), and his average rank was 1.39 (compared to the 15-target mean of 4.12). At the other extreme, no. 19 had a confidence measure of 0.88 and a rank of 9.16. These two individuals are depicted in Fig. 6.

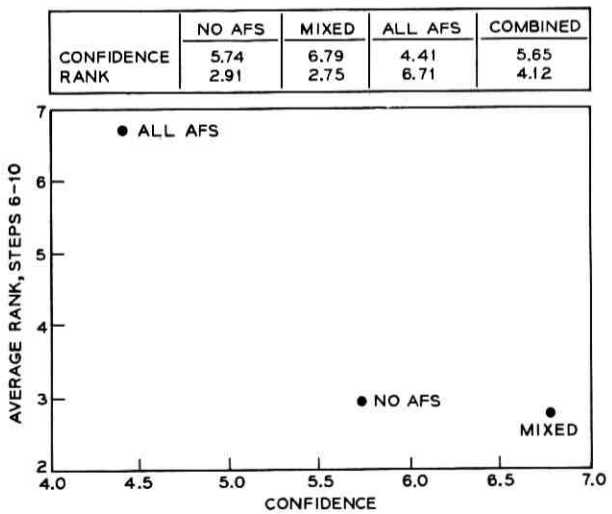


Fig. 5—Two measures of performance summarized for all subjects and targets. MIXED mode is clearly superior, while ALL AFS is markedly poorest. Combined results for all experimental data show that the average target, with a rank of 4.12, was in the upper 1.6 percent of the population over the sixth through tenth steps.

Fig. 6—Targets which produced two extremes of performance. No. 19 was difficult to retrieve, obtaining confidence 0.88 and rank 9.16, while no. 99 was outstandingly easy, obtaining confidence 20.3 and rank 1.39.

The reasons for the different success with the two targets are clear. In general, no. 19 is much closer to the population mean than is no. 99 who has a larger number of more extreme features than has no. 19. All ten subjects who portrayed no. 99 in either the MIXED or the NO-AFS mode started their portrait with hair texture; no. 99 has the curliest hair in the population. All ten also described his light hair-shade and thin upper lip, and all five NO-AFS portraits included his small-to-medium mouth width. By contrast, only one of no. 19's features received unanimous mention: his medium-to-wide eye opening.

3.1.1.3 *Performance Differences Among NO AFS, MIXED, and ALL AFS.* The differences in performance among the three modes of opera-

tion are clear and consistent. This can be seen first by noting the average rank of the target at each feature step. Figure 7 illustrates this by a plot of the percent of the population with better rank than the target at each step. Overall, the population reduction in early steps is quite rapid.

It is clear that at any step the ALL-AFS mode places the target about twice as far down the ordered list as does either of the other two modes. This suggests that knowledge of the population statistics is not as effective as knowledge of a target's outstanding features. Both the MIXED and the NO-AFS modes are roughly equal and are superior to ALL AFS. From step seven on, with the MIXED and NO-AFS modes, the population having better rank than the target was reduced to 0.68 percent. We have seen (Fig. 5) that the confidence in the MIXED experiments is 18 percent higher than that in the NO-AFS experiments and 54 percent higher than that in the ALL-AFS experiments. Similarly, the rank results are superior for MIXED, being 11 percent ahead of NO AFS and 59 percent ahead of ALL AFS. Even for ALL AFS, however, the average rank was better than seventh place; i.e., 2.2 percent of the population had better rank than the target.

The plots of rank cross-section (see Section 2.4), displayed in Fig. 8, also make evident the relative inferiority of ALL AFS. The asymptotic levels of NO AFS and MIXED are virtually identical. For both MIXED and NO AFS, half the targets reach first place by step five, and by

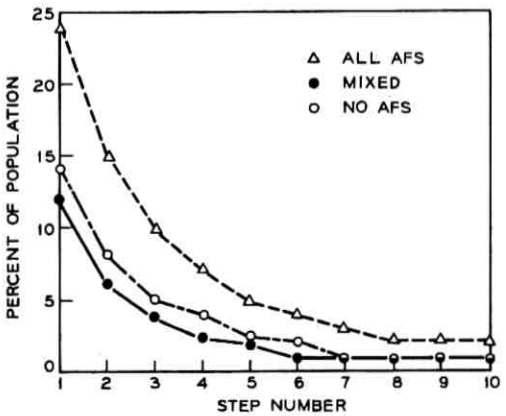


Fig. 7—Comparison of how three modes of system operation affect the percent of the population having better rank than the target. MIXED mode is clearly superior in early steps; with eight feature-steps ALL AFS reduces the population to 2 percent, and both other modes reduce it to 0.68 percent.

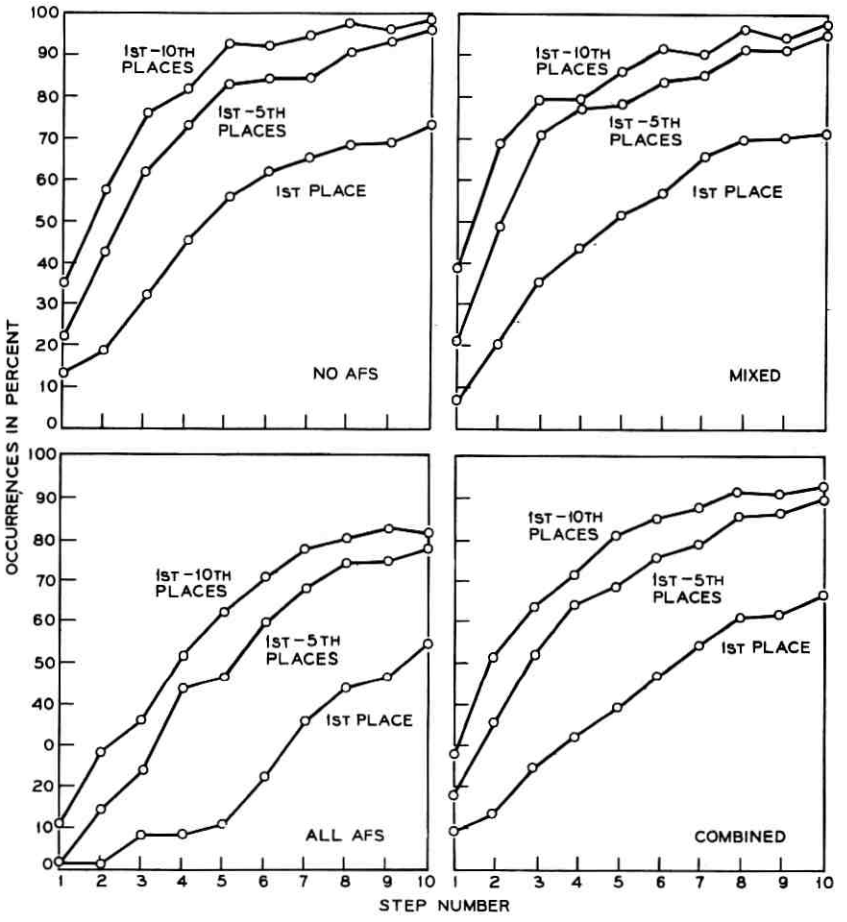


Fig. 8—Rank cross-section at each step. ALL-AFS mode is obviously inferior. Performances from NO AFS and MIXED are essentially alike. By step five, roughly half the targets reached the top in MIXED and NO AFS; by step 10, better than 70 percent reached first place.

step ten in both modes 99 percent of the targets are in no worse than tenth rank. And 96 percent are in no worse than fifth rank.

Although ALL AFS does not produce results comparable to those of the other modes, more than half the targets reach first place by the tenth step, and more than three-quarters of them reach fifth place or better.

The confidence measure (see Fig. 5) also indicates the relative inferiority of ALL AFS. Unlike the other measures discussed here, confidence

shows MIXED to be superior to NO AFS in separating the target from the rest of the population.

3.2 *Computer Experiments*

How does human performance compare with that of an "ideal" subject? The major variables in subject performance are the set of features selected, the accuracy with which they are judged, and the order in which they are described. Since the subject is constrained to use integral feature-values, the best judgment he can make on any feature is the nearest integer to the target's official description; we shall refer to this value as a "rounded" judgment. For each target there is a sequence of features which gives the largest confidence at step 10, and there is one which gives the best average rank. Either of these could be regarded as the optimal sequence chosen by an ideal subject. However, there is no easy way to find such optimal sequences; therefore the ideal subject was defined as follows:

For each target the sequence of features to be used by the ideal subject in a computer simulation was selected on the basis of feature "extremeness." The extremeness of an individual's feature is the magnitude of the difference between his official value and the feature's population mean. Our ideal subject, modeled on how our human subjects were instructed, was defined to be one who selected features in descending order of extremeness *and* used, for each feature's value, the rounded value of the official description.

This ideal subject was used to portray the 15 targets employed in the human experiments. The distribution of the step at which the target first achieved rank one and remained there through step 10 is

Step	1	2	3	4	5	6	7	8	9	10
Frequency	2	4	4	3	1	0	0	0	0	1.

For all targets, the average number of steps is 3.27, and the average rank (over steps 6 through 10) is 1.01 (i.e., virtually perfect). The confidence at step 10 ranged from 1.00 to 95.6 with an average of 21.5 and a median of 16.1.

These results are markedly superior to the results of the human experiments summarized in Fig. 5. Are the differences due to subject judgment-errors or to less-than-ideal feature selection owing to the fact that the subject does not know the population statistics?

To explore this question, three additional computer studies were performed with the same 15 targets used in the human experiments. The results of all four computer experiments are summarized in the tabulation below and are contrasted with the NO-AFS human experi-

ments. Experiment no. 1 is that described above, using the ideal subject. In the second experiment, the NO-AFS human experimental data were modified by replacing the subject judgments with rounded official-values. Third, the extreme features chosen by the ideal subject were used with human judgments. In the last computer experiment (no. 4), four random sequences of features were used with rounded feature-judgments. Finally, the results of the NO-AFS human experiments are shown.

Besides displaying confidence and mean rank (averaged over steps 6 through 10), the table shows the number of targets on which confidence was greater than, approximately equal to, and less than the confidence obtained by the ideal subject.

Exp.	Feature Selection	Judgment	Conf.	>	≅	<	Mean Rank
1	Extreme	Rounded	21.5	0	15	0	1.01
2	NO AFS	Rounded	10.6	4	6	5	1.23
3	Extreme	Human	8.25	1	2	12	1.68
4	Random	Rounded	4.08	0	2	13	1.76
5	NO AFS	Human	5.74	2	0	13	2.91

The confidence and mean rank show the performance of the ideal subject (exp. no. 1) to be better than that obtained in the experiment using NO AFS and rounded official-values. Notice, however, if one examines confidence for the ideal case and NO AFS rounded, *target by target*, then it is seen that NO AFS is better about as many times as it is worse. Since the only variable was feature selection, this indicates that the humans were almost as good as the ideal subject in their choice of features. The use of extreme features with human judgments (exp. no. 3) gives worse performance in rank and confidence than does NO AFS with rounded judgments. This shows that the advantage of extreme-feature selection was not sufficient to overcome human errors in judgment.

It might be argued that *any* feature sequence would produce good results. But the random experiment shows that perfect feature-judgments alone are not sufficient; feature selection is important.

In summary, humans are nearly ideal in feature selection while considerably less than ideal in feature-value assignment.

IV. EXTENSION TO LARGE POPULATIONS AND TO OTHER PROBLEMS

The procedures we have described for identification and retrieval are applicable to problems other than the face-recognition tasks we have

so far explored. Such searches as medical diagnosis and telephone-directory lookup also deal often with noisy data where probabilistic identification is made. With what generality can the procedures we have evolved be applied to tasks where descriptive components are imprecise and populations are large?

First, however, there are questions of economic feasibility. The storage and computing requirements in the present experiments are modest. For a population of 255, we require 1500 words* of disk and 14,400 words of core storage. Memory requirements grow at a rate of 7 words/face. The interactive computation process (slowed enormously by the human at a remote terminal) takes about 5–10 minutes real time (~5 seconds central-processor time) and costs $2.50 on the average. A key question for extended applications is: How do these numbers increase with population size?

In the earlier model of the binary-search identification process,[1] we showed a logarithmic growth of the number of steps (features) required to isolate a target. For a particular condition we found useful, the model predicts that an average of only 13.5 feature-descriptions will be required for a population of 4 million. If the actual growth of the number of steps required to isolate in the present rather different rank-ordering process is close to our model's prediction in the binary-search process, then a nonlinearity very important to economic treatment of large populations will be at hand. That this may indeed be so can be seen in Appendix B.

To investigate the effect of population size on the number of steps required for isolation, comparable runs were made with population sizes of 128, 255, and 510 individuals.† The first feature in all portraits was chosen at random, and all subsequent features were chosen by AFS. (Since the number of individuals used in the AFS computation is a function of each partial portrait, the cost varies from target to target.) The dynamic stopping-rule described at the end of this section was used. Feature judgments were drawn from the panel of observers whose averaged judgments comprise the official values. Randomly chosen observers supplied portraits. The data for each population size were averaged over five portraits of each of 15 randomly chosen individuals (75 portraits total). The results of this experiment are summarized below.

* The computer is a time-shared Honeywell-635 having 36-bit words.

† The 128-individual population is a randomly-chosen subset of the 255-face one. The 510-individual population is composed of the original 255 individuals plus 255 "new" pseudo-faces created by randomly shuffling the feature values of the old population.

	Population Size		
	128	255	510
Mean stopping step, std. dev.	9.5, 3.9	10.6, 4.2	11.7, 4.3
Relative total cost	1.00	1.98	4.56
Relative cost/step	1.00	1.77	3.76

While the mean stopping step appears to increase logarithmically with population size, P, the cost per step increases roughly in proportion to population size. That is,

$$\text{Total Cost/Step} \sim P$$

and the logarithmic growth of the mean stopping step with population size gives

$$\text{Total Cost} \sim P \ln P.$$

The mean stopping step increased very slowly with population size, from 9.5 to 11.7 for populations of 128 and 510. The final rank of the target rose on the average from 1.4 to only 2.5, less than a twofold increase for a fourfold increase in population size. Experience with MIXED and ALL AFS indicates that the corresponding figures for MIXED would be markedly better than those above, which were obtained with ALL AFS.

The cost of the AFS algorithm is linear with respect to the number of faces used to determine the next feature. Figure 9 shows that this number converges rapidly to a minimum. It is seen that, at most, less than 35 percent of the population is used in the AFS computation at step two and less than 15 percent at step three. From step four on (with but a slight exception at step five), only 3.9 percent is used; this is the minimum possible given our (arbitrary) convention of considering all individuals with relative weight $\geqq 0.7$, but at least 10 faces (10/255 = 3.9 percent).

Several kinds of algorithmic corner-cutting look attractive and are under consideration. The results displayed in Fig. 10 show that for a given performance level only some minimum proportion of the population need be considered at each step. For example, if flawless performance were required while operating in the MIXED mode, no more than half the population would need to be considered in steps three and four, and from step five on, at least 75 percent of the population could be ignored. In 95 percent of all trials, the target was in the top 10 percent of the population from the sixth step on. The computational

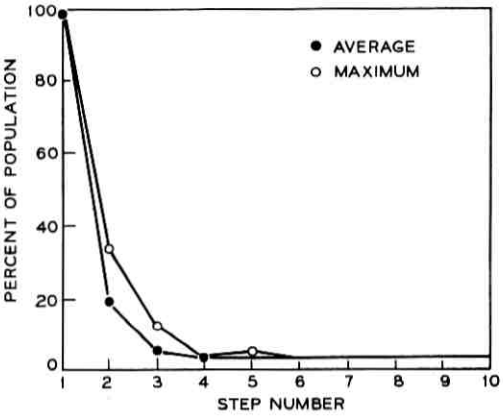


Fig. 9—Extent of AFS computation. For empirically determined rule of considering only those population members with weight equal to or greater than 0.7, or a minimum of 10, extent of computation drops rapidly. With but very slight exception, at and after step four no more than 10 individuals have weights above 0.7, indicating efficient separation of top members.

savings with such a limited-depth search would thus be considerable.

Another possible economy might be some form of individual or feature clustering. One could divide the population into small groups of "look-alikes" and create a "super-description" for each cluster whose official description was the mean of the individual descriptions. One could then order these clusters according to their resemblance to the target description and then search the clusters' members in that order to find a good individual match to the description. This scheme assumes that such a clustering can be achieved and that the cluster descriptions would be non-trivially different.

In a sense the 255 individuals we have dealt with comprise a cluster of the general population. Our 255-member subpopulation was deliberately chosen to be homogeneous (see footnote on page 399) to make isolation more difficult. Consequently, several highly reliable features (e.g., gender, race, age) could be added to our feature set for use with a more universal population. We might guess that the general population represented by the nonrepresentative subpopulation used in these studies is on the order of several thousand individuals.

An Empirical Dynamic Stopping Rule

An empirical dynamic stopping rule was developed using the data gathered from the 75 NO-AFS portraits. It is based on the concepts of confidence and rank and on tradeoff between the frequency and accuracy with which the rule stops portrait composition.

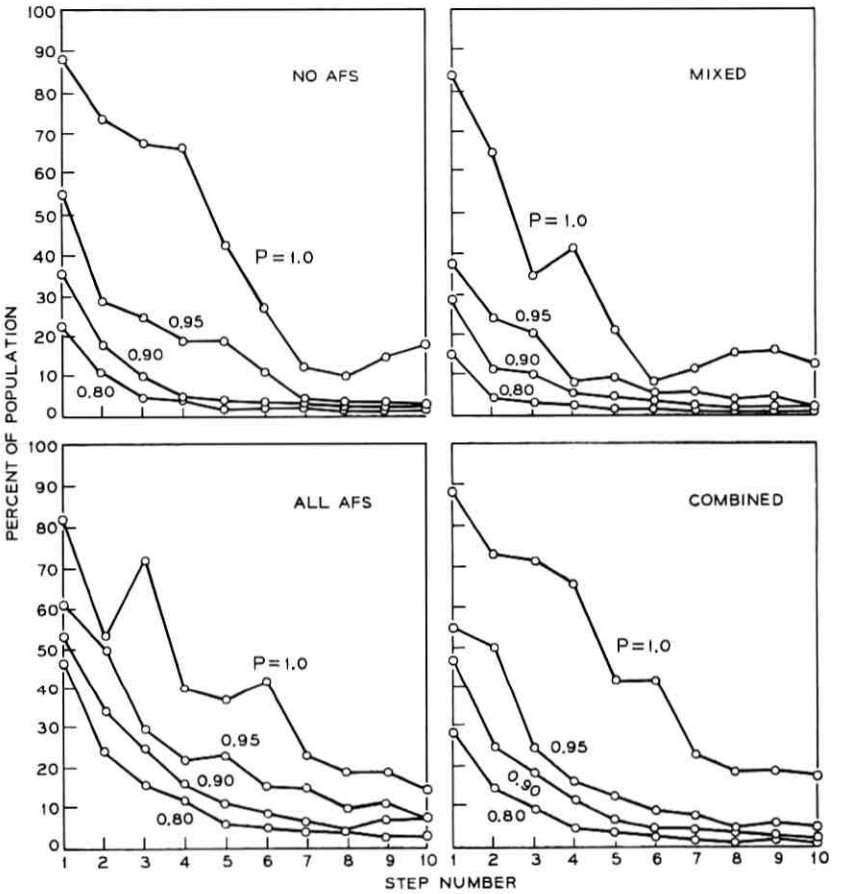


Fig. 10—Minimum envelope needed to capture the target with several probabilities at each step. $P = 1.0$ corresponds to the *worst* rank observed experimentally. After step five, target was in top 10 percent of the population for all cases except ALL AFS.

We consider first the *confidence*, which measures the degree of separation among population members. To formulate a stopping rule, we will use a variant "pseudo-confidence," the ratio of the weights of the first- and second-ranked individuals. (Note that this ratio is always $\geqq 1.0$). The experimental data show that when this ratio exceeded 3.5 at any step in the portrait, the target was then ranked first in 32 of the 34 cases, and the first-ranked individual was subsequently unseated in only two of 34 cases. We adopt this threshold as one component in our dynamic stopping rule: Whenever the pseudo-confidence exceeds 3.5, stop portrait composition.

Unfortunately, such a high pseudo-confidence occurs in fewer than half of the portraits. Another possible stopping criterion is an extended tenure-of-first-place by the same individual. Consequently, we adopt as the second component in our stopping rule: If the same individual has been in first place for the last six steps, stop regardless of the value of the pseudo-confidence. In only one of 37 cases did an individual change rank after holding first place for six or more consecutive steps.

We now have two criteria which would have terminated 80 percent of our experimental NO-AFS portraits. It was decided to use them as points on a linear stopping rule combining p, the pseudo-confidence, and s, the number of consecutive steps in which the same individual has been first-ranked: If $s + 2p > 8$, stop. This is the dynamic stopping rule used above to compare costs for various population sizes.

This empirical stopping rule was applied to the data from the rest of the experiment, and it provided another means of comparison (the mean stopping step) among the three types of portraits. The table below shows the results of applying the dynamic stopping rule to the NO-AFS, MIXED, and ALL-AFS runs.

	NO AFS	MIXED	ALL AFS
Decisions (Number of portraits terminated by stopping rule)	55	56	43
Correct decisions	49 (89%)	48 (86%)	30 (70%)
Mean stopping step, std. dev.			
Decisions only	6.8, 2.1	6.6, 2.2	7.7, 1.8
All portraits	7.7, 2.3	7.4, 2.4	8.7, 1.8
Mean rank of target			
Decisions only	1.4	1.6	3.6
All portraits	2.3	2.4	5.1

The number of decisions is the number of portraits (out of 75 in each case) which met the requirements of our stopping rule at or before the tenth step. A correct decision is one in which the target was in first place at the stopping step. The mean stopping step and its standard deviation are given for both the portraits which the stopping rule terminated ("Decisions only") and for all 75 portraits, considering the stopping step to be 10 for portraits in which no decision was made. The mean rank of the target at the stopping step is also given for both cases.

The data show the performance of MIXED and NO AFS to be almost identical. Both are superior in all respects to ALL AFS. The

mean stopping step and mean rank of the target are in the ranges one would expect from Fig. 7, which shows the progression of the average rank of the target. The stopping rule usually was satisfied soon after the position of the target had stabilized.

If this dynamic stopping rule had been used in our experiments, the average stopping step for a portrait would have been 7.9 instead of 10, a 21-percent saving with virtually no loss of accuracy in identification.

V. SUMMARY

An interactive system for the description and retrieval of multi-dimensional objects has been developed. This paper describes the system and its performance in face-identification experiments.

The system permits flexible description of target items using features chosen by either the user of the program or an automatic-feature-selection algorithm. At each step, AFS selects the feature which is most likely to be discriminating. It makes this choice on the basis of the partial portrait and the population statistics. Population members are ranked at each step on the basis of weights which reflect the match between the portrait description and each individual's official value. Performance is measured by two indices, confidence and rank.

The system was evaluated using 21 features, a population of 255 faces, and three modes of operation (NO AFS, MIXED, and ALL AFS). There were four principal results:

(*i*) The population was quickly and effectively reduced by all modes of operation. Over all trials, the population was reduced to less than 4 percent more than 93 percent of the time, and the target was successfully "isolated" (i.e., was in first place by portrait's end) 67 percent of the time (see Fig. 8). In 95 percent of all trials, the target remained in the top 10 percent of the population from the sixth step on.

(*ii*) The MIXED mode was the most effective in separating the target from the rest of the population as measured by confidence (see Fig. 5).

(*iii*) MIXED and NO AFS were equally effective with respect to population reduction, as measured by rank. The performance of these two modes was considerably superior to that of ALL AFS (see Figs. 5, 7). In the MIXED experiments, the population was reduced to less than 4 percent over 99 percent of the time, and the target was isolated 70 percent of the time (see Fig. 8).

(*iv*) The extent of the AFS computation drops rapidly with step number, reaching its minimum by step four (see Fig. 9).

These results can be summarized as follows: even in the worst case there is fair performance in singling out a target and good performance in narrowing down the population; and in the best case the population reduction is excellent.

This rapid population-reduction and the slow growth of the mean stopping step with population size (using the dynamic stopping rule) make the extension of these experiments to larger populations feasible. To process very large populations, say on the order of a million, new approaches would undoubtedly be needed. With the cost-cutting modifications we have described (dynamic stopping rule, limited-depth search), the present system could economically accommodate a population on the order of 5000.

VI. ACKNOWLEDGMENTS

We appreciate the skill and the stamina both of our subjects and of our early-draft critics W. S. Brown, Murray Eden, E. N. Gilbert, Newman Guttman, M. E. Harmon, S. C. Johnson, M. E. Lesk, R. C. Lummis, David Slepian, and Eric Wolman.

APPENDIX A

Automatic Feature-Selection

As discussed in the text (Section 2.2), the automatic-feature-selection algorithm selects, at each step, the most discriminating feature for the subject to describe next. The purpose of this Appendix is to formalize what is meant by a discriminating feature.

The AFS algorithm uses a subset of the population whose members are well-described by the subject's description of the target. In order to give greater importance to those members of the subset with high weight, each member's official feature-values were considered in proportion to his weight. The most discriminating feature, for that subset, thus is the one for which the distribution of the weighted feature-values is most uniform. Since the distribution of feature values may span different parts of the permissible feature ranges, distributions are shifted to facilitate equitable comparisons among features.

We shall define, for any shift, the deviation of the distribution of weighted feature-values from a uniform distribution. Formulae for the best shift and corresponding deviation are then derived.

Consider the subset of the population whose members are well-described by the subject's description of the target. Let the members of this subset have weights $W_1, \cdots, W_n$. The sum of the weights is W_T.

Let us concentrate on one feature. For convenience, scale its range to be from 0 to 1. Let the (scaled) official values corresponding to the above weights be $v_1, \cdots, v_n$.

Let

$$p_i = W_i/W_T .$$

We may interpret p_i as the probability that individual i is the target. When the sum of these probabilities in any interval is equal to the length of that interval, then the distribution of weighted feature-values is uniform. That is, if the interval is (x_1, x_2), then

$$\sum_{x_1 \leqq v_i \leqq x_2} p_i = x_2 - x_1 .$$

This is equivalent to

$$\sum_{0 \leqq v_i \leqq v} p_i = v, \qquad 0 \leqq v \leqq 1.$$

The deviation from uniformity can be measured by integrating the square of the difference between the left and right sides,

$$\int_0^1 \left(\sum_{0 \leqq v_i \leqq v} p_i - v \right)^2 dv.$$

If we define $F(v)$ by

$$F(v) = \sum_{0 \leqq v_i \leqq v} p_i ,$$

then the last formula becomes

$$\int_0^1 (F(v) - v)^2 \, dv.$$

Figure 11 gives a typical plot of $F(v)$ where $F(v) = 0$ for $v \leqq a$ and $F(v) = 1$ for $v \geqq b$. Now shifting $F(v)$ to the left or right [as long as neither a nor b is shifted out of the interval (0, 1)] does not change the essential shape of $F(v)$. It is reasonable to shift $F(v)$ to give the best approximation to v. We therefore define $E(s)$ to be the mean squared error when $F(v)$ is shifted by s; i.e.,

$$E(s) = \int_0^1 (F(v - s) - v)^2 \, dv \quad \text{for} \quad -a \leqq s \leqq 1 - b.$$

Then we redefine the deviation from uniformity by

$$E = \min_s E(s).$$

We derive the minimizing shift in the following lemma.

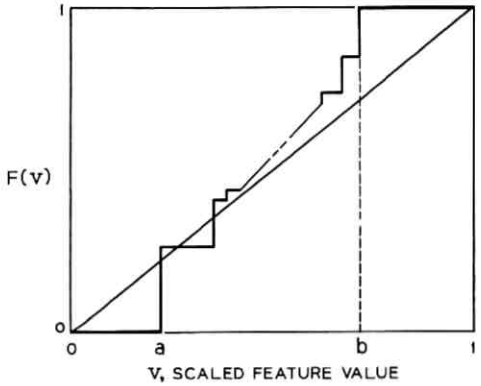


Fig. 11—A graph of a hypothetical $F(v)$, the cumulative distribution of P_i (the normalized weights) versus v (the scaled feature value). No individual has a scaled feature value less than a or greater than b.

Lemma: Let

$$E(s) = \int_0^1 (F(v - s) - v)^2 \, dv,$$

and let

$$e = \int_0^1 v \, dF(v).$$

Then for $-a < s < 1 - b$, $E(s)$ *is minimized for*

$$s = \begin{cases} -a & \text{if} \quad \tfrac{1}{2} - e \leqq -a \\ \tfrac{1}{2} - e & \text{if} \quad -a < \tfrac{1}{2} - e \leqq 1 - b \\ 1 - b & \text{if} \quad 1 - b < \tfrac{1}{2} - e. \end{cases}$$

Proof: To avoid needless mathematical complexity, let us suppose that $F(v)$ is a differentiable function. Then

$$E'(s) = -2 \int_0^1 (F(v - s) - v) F'(v - s) \, dv$$

$$= -F^2(1 - s) + F^2(-s) + 2 \int_0^1 v \, dF(v - s).$$

The first term is -1 since $b \leqq 1 - s$. The second term is 0 since $-s \leqq a$. The third term is easily shown to be $2(e + s)$ by using the substitution $u = v - s$ and the facts that $-s \leqq a$ and $b \leqq 1 - s$. Thus

$$E'(s) = 2(e + s - \tfrac{1}{2}).$$

$E(s)$ is minimized by having $e + s$ as close to $\frac{1}{2}$ as possible since $E'(s)$ is negative (positive) if $e + s$ lies to the left (right) of $\frac{1}{2}$.

APPENDIX B

Population Size and Identification Speed

We wish to show that the number of steps in the identification process grows at most logarithmically with P, the population size. More precisely, let r_k denote the rank of the target after the subject has given the kth feature value. It will be shown that under reasonable assumptions, given below, the expected value of r_k, for large k, satisfies

$$E(r_k) < P \cdot c^k = \exp\left(\ln P - k \ln \frac{1}{c}\right),$$

where $0 \leqq c < 1$, and c is a function of the distributions of the official values and the subject's errors in judgment. Thus, to achieve a given expected rank, the number of steps, k, need grow no faster than $\ln P$.

While we believe that these several assumptions lead to a reasonable model of our experiment, we expect them to provide only a qualitative indication of the growth of rank with population size. A quantitative analytical model is unobtainable at this point since the data we have are insufficient to extract the necessary statistical parameters. The assumptions are as follows:

Each of the P individuals in the population can be considered to be a vector $i = (i_1, i_2, \cdots)$ whose components are the official feature-values. We assume that these feature values are independent, identically distributed random variables and that the individuals are independent vectors. The subject describes the features of the target $t = (t_1, t_2, \cdots)$, and his judgments of the features are in error by $e_1, e_2, \cdots$. We assume that the errors are independent, identically distributed random variables.

By convention, the components of each vector are ordered in the sequence in which they are described by the subject.

Using the above notation and our definition of weight, the target has weight

$$\exp\left(-\sum_{j=1}^{k} | e_j |\right)$$

while an individual, i, has weight

$$\exp\left(-\sum_{j=1}^{k} | t_j + e_j - i_j |\right).$$

If we define

$$x_i(t, i) = | t_i + e_i - i_i | - | e_i | ,$$

then i's weight is larger than t's weight if

$$\sum_{j=1}^{k} x_j(t, i) < 0.$$

Let

$$s_k(t, i) = \begin{cases} 1 & \text{if } \sum_{j=1}^{k} x_j(t, i) < 0 \\ 0 & \text{otherwise.} \end{cases}$$

Define r_k, the rank of t, as the number of individuals with weight larger than t's weight. We then have

$$r_k = \sum_{i, i \neq t} s_k(t, i).$$

The expected value of r_k is

$$E(r_k) = E \sum_{i, i \neq t} s_k(t, i).$$

In $s_k(t, i)$, each summand $x_k(t, i)$ has, for fixed t, a distribution which clearly is a function of t. However, we are taking an expectation over all targets and populations. Thus, in this context, the $x_k(t, i)$ (for $t \neq i$) are independent, identically distributed variables since the t's, as well as the i's and e's, are independent, identically distributed variables. Hence

$$\begin{aligned} E(r_k) &= (P - 1)E(s_k(t, i)) = (P - 1) \Pr \{s_k(t, i) = 1\} \\ &= (P - 1) \Pr \left\{ \sum_{j=1}^{k} x_j(t, i) < 0 \right\}. \end{aligned}$$

Let the x_k's have common mean m and standard deviation σ. We apply the Central Limit Theorem to the last probability to obtain

$$\begin{aligned} \Pr \left\{ \sum_{j=1}^{k} x_j(t, i) < 0 \right\} &= \Pr \left\{ \frac{\sum_{j=1}^{k} x_j(t, i) - km}{\sqrt{k}\,\sigma} < -\sqrt{k}\,\frac{m}{\sigma} \right\} \\ &\sim \Phi(-\sqrt{k}\, m/\sigma) \end{aligned}$$

where Φ is the cumulative normal distribution. For large values of $\sqrt{k}\, m/\sigma$, the asymptotic formula* for Φ gives

* As $x \to \infty$, $\Phi(-x) \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x}$.

$$E(r_k) \sim (P - 1) \frac{1}{\sqrt{2\pi}} \frac{e^{-km^2/2\sigma^2}}{\sqrt{k}\, m/\sigma}$$

$$< P(e^{-m^2/2\sigma^2})^k = Pc^k$$

for k sufficiently large.

REFERENCES

1. Goldstein, A. J., Harmon, L. D., and Lesk, A. B., "Identification of Human Faces," Proc. IEEE, *59*, No. 5(1971), pp. 748–760.
2. Kingston, C. R., "Problems in Semi-Automated Fingerprint Classification," in *Proceedings of the First National Symposium on Law Enforcement Science and Technology*, Washington, D. C.: Thompson Book Co., 1967, pp. 449–457.
   Kingston, C. R., and Rudie, D. D., "Fingerprint Classification Methods Study—Second Status Report," N. Y. State Identification and Intelligence System/System Development Corp., October 1966.
3. Sitar, E. J., "A Handwriter Identification System," unpublished work.
4. Salton, G., and Lesk, M. E., "The SMART Automatic Document Retrieval System—an Illustration," Comm. ACM, *8*, No. 6 (1965), pp. 391–398.
5. Lesk, A. B., "An Interactive Program for Identification of Multi-Dimensional Objects," unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Power Distribution and Radiation Losses in Multimode Dielectric Slab Waveguides

By D. MARCUSE

(Manuscript received September 1, 1971)

Guided modes of multimode waveguides exchange power if the waveguide deviates in any way from its perfect geometry. The power exchange problem is studied for a multimode slab waveguide under the assumption that the power coupling is caused by irregularities of the core-cladding interfaces. The problem is treated by means of coupled power equations. The main result of this study is the realization that the power distribution versus mode number settles down to a steady state distribution if the waveguide is sufficiently long. The shape of the steady state distribution depends on the correlation length of the function describing the core-cladding interface irregularities. For very short correlation length only the lowest-order mode carries an appreciable amount of power while the power carried by all the other modes is orders of magnitude smaller. For very long correlation length, on the other hand, all guided modes carry equal amounts of power. The steady state distribution is achieved regardless of the way in which the power was distributed over all the modes at the beginning of the guide. However, the total power in the steady state mode distribution is dependent on the initial power distribution.

I. INTRODUCTION

Light communications systems using optical fibers as the guidance medium are presently being planned for two different modes of operation. High-capacity systems are likely to be used with a laser as the light source and should be operated in the fundamental HE_{11} mode in order to minimize delay distortion that accompanies multimode operation. For less ambitious, low-capacity systems excitation of the fiber with a light emitting diode appears more economical. However, the output of light emitting diodes cannot be used to excite a single fiber mode with high efficiency. A low-capacity fiber to be used with a light emitting diode must thus be designed to operate with many modes.

Multimode optical fibers are not as easily characterized as single-mode

fibers. The power loss of such a fiber is usually not simply expressed by an exponential decay law but depends in a complicated way on the distribution of the power over the many modes. The present study is an attempt to describe the loss behavior of multimode optical fibers. We use the TE modes of the simplified model of a slab waveguide with the added requirement that there is no field variation or change in the slab geometry in the y direction of the coordinate system. This model makes it possible to describe the multimode waveguide rather simply. Even though it cannot directly be used to predict the loss behavior of round multimode optical fibers, it provides insight into the operating principles of multimode waveguides that can be used to obtain an understanding of the properties of multimode fibers of different shape. Our treatment of the multimode dielectric slab waveguide is based on coupled power equations. It has been shown in an earlier paper[1] that the coupled wave equations of a multimode optical waveguide[2] can be used to derive much simpler coupled power equations provided that the coupling mechanism can be described by a stationary random process with Gaussian correlation function. The coupled power equations have the advantage that their coefficient matrix is constant, real, and symmetric. The system of coupled linear first-order differential equations can thus be solved by first finding eigensolutions with the common z dependence $\exp(-\alpha z)$. These can be used to express the general solution as a superposition of eigensolutions. This approach makes it clear that a steady state power distribution must exist. By allowing the field to travel far enough in the waveguide, so that all but the lowest-loss eigensolution has decayed to insignificant values, it is obvious that the distribution of power over the many modes assumes the shape of the lowest-order eigensolution regardless of the initial power distribution. The power loss of the steady state eigensolution obeys a simple exponential law and can thus be characterized by a single number, the lowest-order eigenvalue of the eigensolutions of the power rate equations.

The mechanism causing coupling between the many guided modes and of guided modes to the continuous spectrum of radiation modes will be assumed to consist of irregularities of the core-cladding interface. The coupling coefficients for this model have been evaluated in an earlier paper.[2] Any imperfection of the refractive index distribution and the slab geometry causes coupling between the modes. We choose the core-cladding interface irregularities because this coupling mechanism is of fundamental importance and because its properties are well understood. Mode coupling caused by irregularities of the refractive

index distribution will cause similar effects in some respects. However, there are differences that consist mostly in the dependence of the coupling process on the mode number of the coupled modes.

The coupling coefficient for the core-cladding irregularities can be expressed as a product of a term that is independent of the length coordinate z but depends on the mode number times a z-dependent function that describes the actual shape of the core-cladding interface. This function $f(z)$ is assumed to be a stationary random variable with a Gaussian correlation function that can be completely described by the rms deviation $\bar{\sigma}$ of the core-cladding interface from a perfect plane and by the correlation length D. The same process that couples the guided modes among each other also causes each mode to lose power to the continuous spectrum of radiation modes. The interplay between coupling among the guided modes and power loss to radiation is responsible for the shape of the steady state distribution as well as for the loss associated with that steady state distribution.

In order to spare readers not interested in the details of the theory the trouble of finding their way through the theoretical part of the paper, we present the results of the numerical analysis before the discussion of the details of the theory.

II. RESULTS OF THE NUMERICAL ANALYSIS

The theory has been evaluated for a slab waveguide with a core index of $n_1 = 1.5$ and a core-to-cladding-index ratio of $n_1/n_2 = 1.01$. Most numerical results hold for a slab waveguide supporting ten modes corresponding to the value $kd = 82$ (k = free-space propagation constant, d = slab half width). The only other case for which numerical values have been calculated corresponds to $kd = 165$ with twenty-one guided modes. It has been assumed throughout that the irregularities of the two core-cladding interfaces are statistically independent of each other but have the same rms deviation and the same correlation length.

Figure 1 is a plot of the steady state distribution of the ten-mode slab waveguide. The steady state mode power is plotted versus mode number. Actually, only integer values of the mode number have physical meaning. In order to be able to display the mode power distributions for several values of the correlation length on one graph, the power values at the integer mode numbers were connected by straight lines. The label $B_n^{(1)}$ of the vertical axis refers to the lowest-order eigenvector of the eigenvalue problem [see equation (62) of the theoretical part]. These values are proportional to the power in each mode. They are normalized so that the squares of the power values for all ten integer

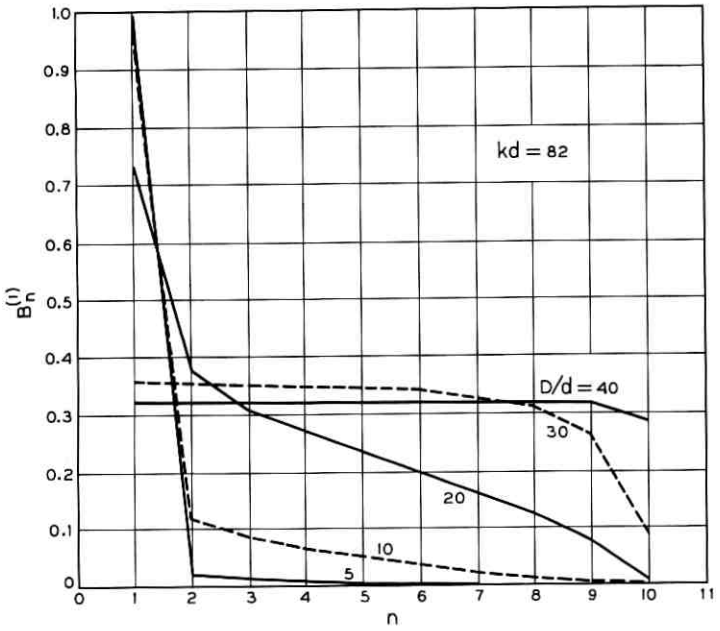


Fig. 1—Steady state power distributions for different values of the correlation length D. The multimode guide carries 10 modes.

mode numbers add up to unity. The most important aspect of Fig. 1 is the shape of the steady state power distribution for different values of D/d, the ratio of correlation length to slab half width. For very long correlation length each mode carries an equal amount of power regardless of the shape of the power versus mode distribution at the beginning of the guide. As D/d decreases more power is carried by the lower-order modes. For very small values of D/d (less than unity) essentially all the power is contained in the lowest-order mode. Figure 2 presents a similar graph for the case of twenty-one modes. The shape of the steady state distributions is essentially unchanged except that similarly shaped curves carry smaller D/d values showing that the number of modes does not affect the general behavior of the steady state distributions.

The shape of the steady state distributions can be explained as follows. For long correlation length only the high-order guided modes lose power directly to the radiation field while the guided modes couple in such a way that only next neighbors exchange power. It is thus understandable that the power tends to equalize among all the modes. For very short correlation length all guided modes couple directly to

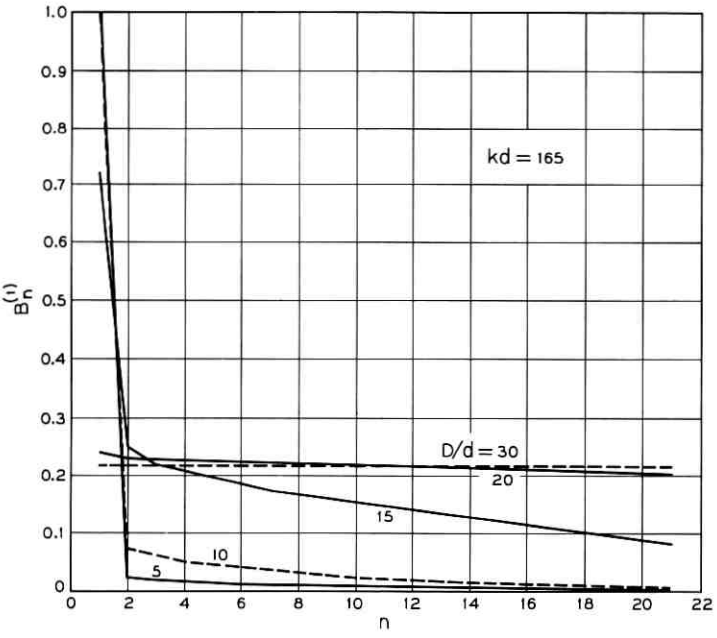


Fig. 2—Same as Fig. 1 for the 21-mode case.

radiation. Higher-order modes lose power by this mechanism at a higher rate than lower-order modes. In addition, each guided mode couples to all the other guided modes. Since the lowest-order mode loses the least power to radiation it is the one that "survives" after all the other modes have lost nearly all of their power. In general, the correlation length of the random core-cladding interface irregularities cannot be chosen at will. However, for multimode operation one would hope for a long correlation length which makes it possible to transmit power in all the modes. Coupling with short correlation length forces the multimode fiber into single-mode steady state operation.

Figure 3 shows the normalized steady state loss $\alpha d/(\tilde{\sigma}^2 k^2)$ of the slab waveguide (the lines labeled $i = 1$) as functions of D/d. The lines labeled $i = 2$ represent the second eigenvalue of the eigenvalue problem. The important feature of Fig. 3 is the existence of a maximum as a function of D/d and the separation between the curves of the first ($i = 1$) and second ($i = 2$) eigenvalues. With the help of these two curves it is possible to estimate the region where steady state operation has been achieved. The loss parameters $\alpha^{(1)}$ and $\alpha^{(2)}$ enter in the form $\exp(-\alpha^{(i)}z)$ as the first and second term of a series expansion [see

equation (62)]. When $\alpha^{(2)}z > 4.6$ we have exp $(-\alpha^{(2)}z) < 10^{-2}$ so that the second term of the series expansion is becoming insignificant and steady state is essentially achieved.

Figures 4, 5, and 6 show the way in which an initially uniform distribution of power settles down toward the steady state distribution. Three different values of correlation length were used. $D/d = 0.01$ is a sufficiently small value whose steady state distribution consists of only the lowest-order mode. The second mode carries only 10^{-4} of the power of the first mode at $z \to \infty$. The value $D/d = 20$ was chosen as an example for an intermediate correlation length. The steady state distribution in this case does not favor exclusively the lowest-order mode but assumes a shape in which higher-order modes carry decreasingly smaller amounts of power. The value of $D/d = 35$ is sufficiently large to produce an essentially uniform steady state distribution.

The next three figures, Figs. 7, 8, and 9, show how the steady state distribution establishes itself if initially all the power is launched in the first mode. The last three figures, Figs. 10, 11, and 12, show similar

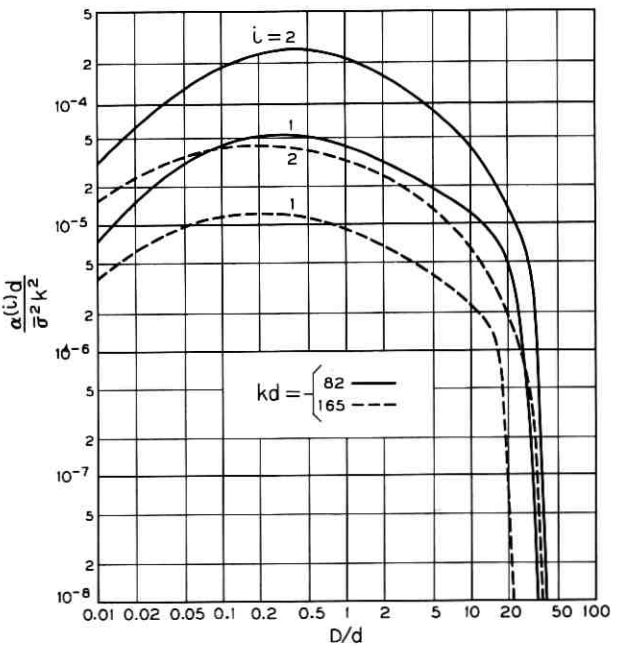


Fig. 3—The normalized first two eigenvalues $\alpha^{(i)}d/(\bar{\sigma}^2k^2)$ ($i = 1$ and $i = 2$) are shown as functions of D/d for the 10- and 21-mode case. The lowest-order eigenvalue, $i = 1$, is the steady state power loss of the multimode waveguide.

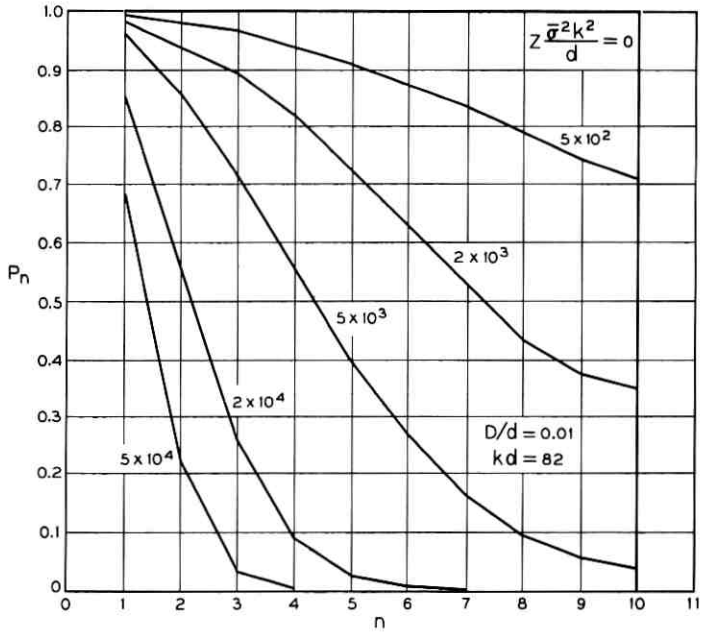


Fig. 4—Power distribution versus mode number for several values of normalized length along the guide for $D/d = 0.01$.

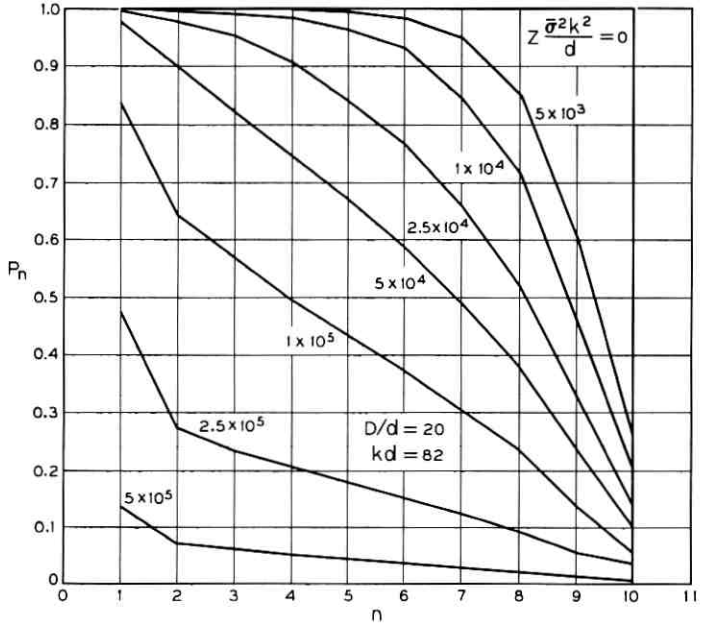


Fig. 5—Same as Fig. 4, $D/d = 20$.

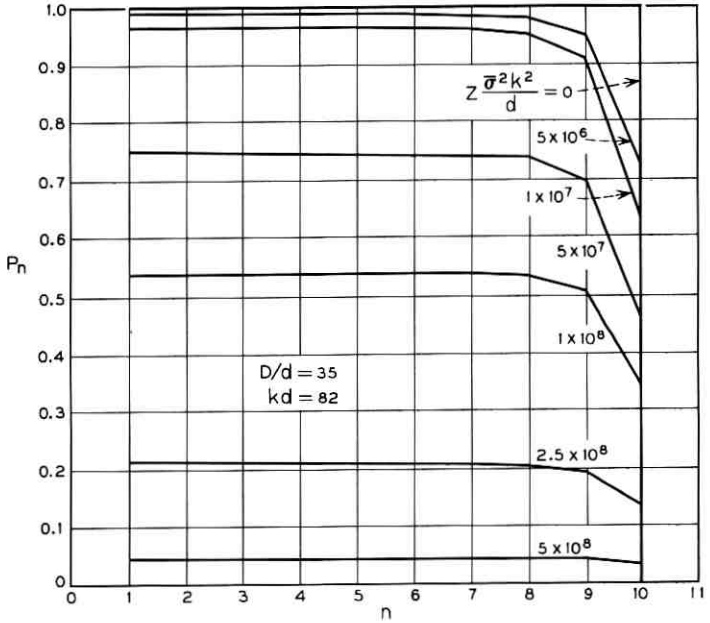


Fig. 6—Same as Fig. 4, $D/d = 35$.

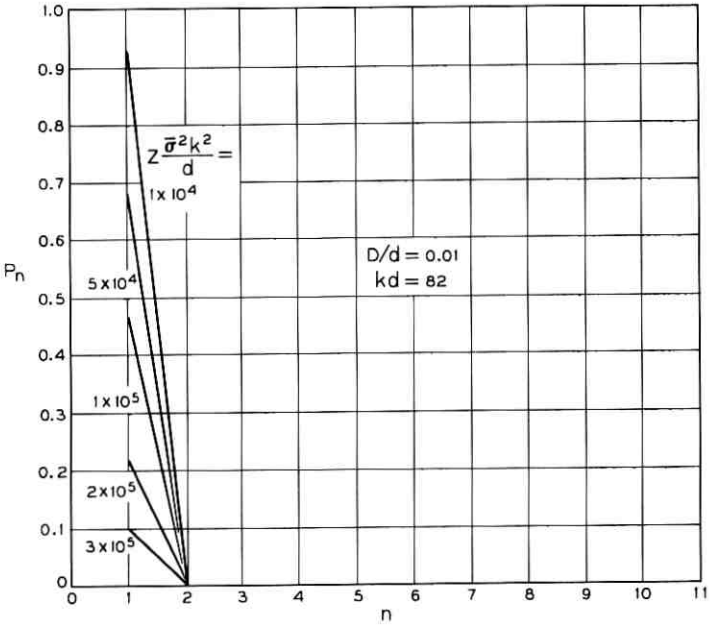


Fig. 7—Power distribution versus mode number for several values of the normalized length along the guide. Only mode 1 is excited at $z = 0$. $D/d = 0.01$.

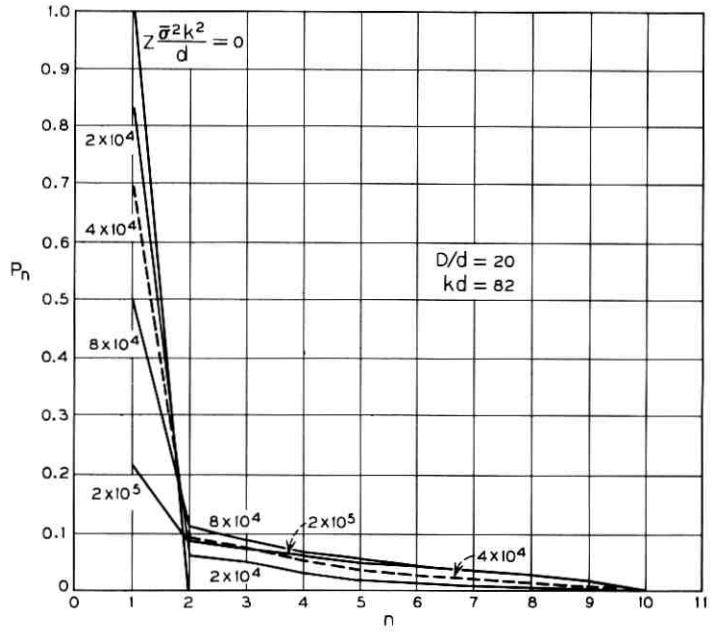


Fig. 8—Same as Fig. 7, $D/d = 20$.

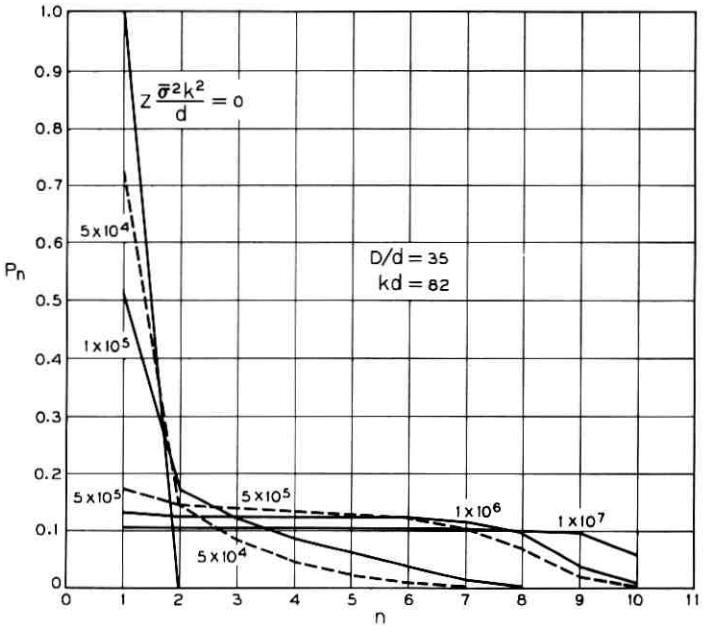


Fig. 9—Same as Fig. 7, $D/d = 35$.

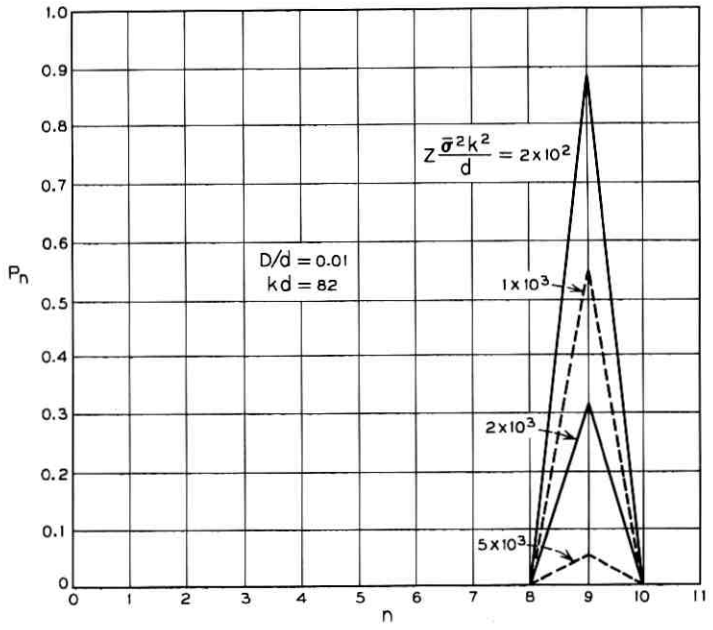


Fig. 10—Power distribution versus mode number for several values of the normalized distance along the guide. Only mode 9 is excited at $z = 0$. $D/d = 0.01$.

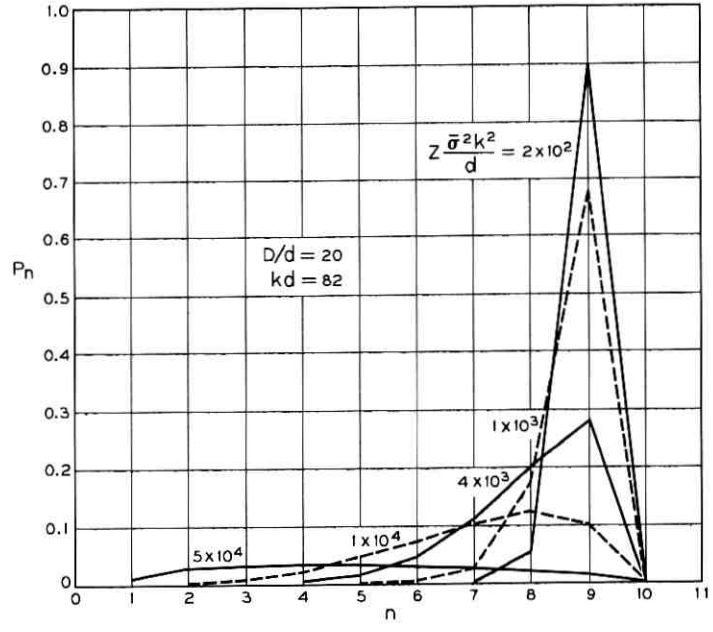


Fig. 11—Same as Fig. 10, $D/d = 20$.

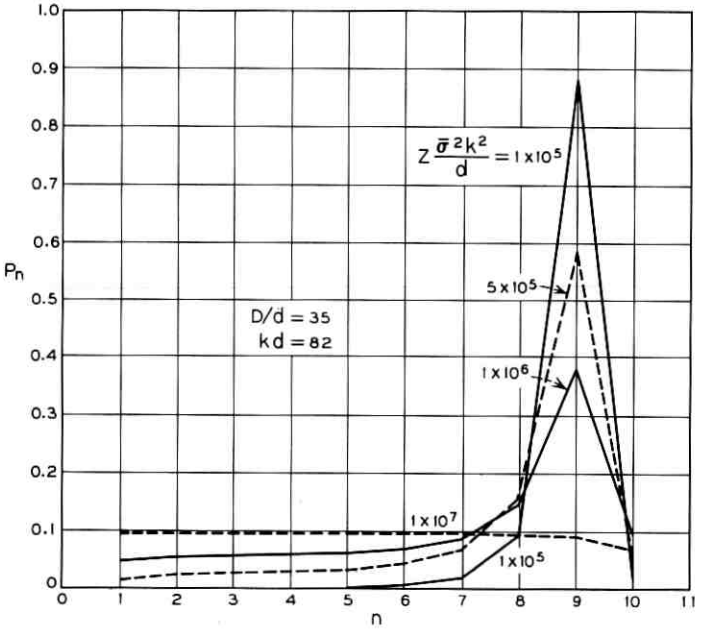


Fig. 12—Same as Fig. 10, $D/d = 35$.

plots for the case that all the power starts out in mode 9. The three values, $D/d = 0.01$, 20, and 35, have again been used.

It may be of interest to know the ratio of the power remaining in the steady state if initially only mode 1 or mode 9 were excited. We call this ratio $P(1, z)/P(9, z)$ and obtain $P(1, z)/P(9, z) = 1.55 \times 10^4$ for $D/d = 0.01$, $P(1, z)/P(9, z) = 9.75$ for $D/d = 20$, and $P(1, z)/P(9, z) = 1.08$ for $D/d = 35$.

Tables I, II, and III show the ten eigenvalues of the steady state mode distributions together with the first eigenvector for the same three values of D/d. The lowest-order and the second eigenvalue appear also in Fig. 3. The other eigenvalues are given in the tables for the sake of completeness. It should be noted that the integer values in the left-hand column of these tables have different meaning for the eigenvalues and the eigenvector. The eigenvalues $\alpha^{(i)}$ are ordered in increasing value and originate as the ten solutions of the eigenvalue problem [equation (60)] of a symmetric 10 by 10 matrix. The first eigenvector $B_\nu^{(1)}$ belongs to the lowest eigenvalue $\alpha^{(1)}$. The subscript ν is a mode label in this case. The eigenvector $B_\nu^{(1)}$ is proportional to the steady state power distribution. Table I shows clearly that mode 1 carries the overwhelming

TABLE I—EIGENVALUES AND THE FIRST EIGENVECTOR FOR $D/d = 0.01$

i or ν	$\frac{d}{\bar{\sigma}^2 k^2}\alpha^{(i)}$	$B_\nu^{(1)}$
1	7.624×10^{-6}	9.999×10^{-1}
2	3.044×10^{-5}	8.486×10^{-5}
3	6.845×10^{-5}	7.160×10^{-5}
4	1.214×10^{-4}	6.789×10^{-5}
5	1.891×10^{-4}	6.630×10^{-5}
6	2.711×10^{-4}	6.546×10^{-5}
7	3.666×10^{-4}	6.497×10^{-5}
8	4.743×10^{-4}	6.466×10^{-5}
9	5.909×10^{-4}	6.444×10^{-5}
10	7.049×10^{-4}	6.424×10^{-5}

TABLE II—EIGENVALUES AND THE FIRST EIGENVECTOR FOR $D/d = 20$

i or ν	$\frac{d}{\bar{\sigma}^2 k^2}\alpha^{(i)}$	$B_\nu^{(1)}$
1	5.175×10^{-6}	7.284×10^{-1}
2	1.402×10^{-5}	3.799×10^{-1}
3	4.576×10^{-5}	3.212×10^{-1}
4	1.008×10^{-4}	2.770×10^{-1}
5	1.951×10^{-4}	2.376×10^{-1}
6	3.329×10^{-4}	2.011×10^{-1}
7	5.217×10^{-4}	1.648×10^{-1}
8	7.596×10^{-4}	1.248×10^{-1}
9	1.046×10^{-3}	7.475×10^{-2}
10	4.345×10^{-3}	3.129×10^{-3}

TABLE III—EIGENVALUES AND THE FIRST EIGENVECTOR FOR $D/d = 35$

i or ν	$\frac{d}{\bar{\sigma}^2 k^2}\alpha^{(i)}$	$B_\nu^{(1)}$
1	6.415×10^{-9}	3.296×10^{-1}
2	2.552×10^{-7}	3.294×10^{-1}
3	1.112×10^{-6}	3.293×10^{-1}
4	4.251×10^{-6}	3.291×10^{-1}
5	7.789×10^{-6}	3.290×10^{-1}
6	1.608×10^{-5}	3.287×10^{-1}
7	3.149×10^{-5}	3.279×10^{-1}
8	5.837×10^{-5}	3.248×10^{-1}
9	1.036×10^{-4}	3.094×10^{-1}
10	1.765×10^{-4}	2.067×10^{-1}

amount of power in the steady state if $D/d = 0.01$. The distribution of the first eigenvector for $D/d = 20$ appears also in Fig. 1.

In order to obtain a feeling for the amount of irregularity of the core-cladding interface that will cause a certain amount of loss, we consider the case $kd = 82$ (10 modes). For $\lambda = 1$ μm we then obtain $d = 13$ μm for the slab half width. We now ask the question: What value of the rms deviation $\bar{\sigma}$ causes 10 dB/km steady state radiation loss? The result is obtained from Fig. 3 or from Tables I, II, and III and is shown in Table IV. The tolerance requirements, arising from the need

TABLE IV—RMS DEVIATION OF CORE-CLADDING INTERFACE CAUSING 10 dB/km LOSS FOR $kd = 82$, $\lambda = 1$ μm

D/d	$\bar{\sigma}/d$	$\bar{\sigma}(\mu m)$
0.01	7.65×10^{-4}	9.94×10^{-3}
0.3	3.26×10^{-4}	4.25×10^{-3}
20.0	9.30×10^{-4}	1.21×10^{-2}
35.0	2.63×10^{-2}	3.42×10^{-1}

for keeping the steady state radiation losses low, are thus very stringent since the rms deviation of the core-cladding interface must be kept within fractions of micrometers.

III. APPLICATION TO DELAY DISTORTION*

Multimode waveguides suffer from delay distortion that occurs because the modes contributing to the power transmission travel with different group velocities. Modes with a higher group velocity arrive at the receiver earlier than modes with a slower group velocity. A pulse, whose power is shared in some way by many modes, is thus distorted and lengthened by this effect. If the modes exchange power rapidly among each other this pulse lengthening effect of multimode waveguides can be substantially reduced. S. D. Personick[4] first pointed out the beneficial effect of tight mode coupling for the reduction of pulse delay distortion. For a two-mode waveguide Personick's results have been confirmed by a rigorous analysis by H. E. Rowe and D. T. Young.[5] Our present work has some applications to the reduction of delay distortion by mode mixing. It is clear that if the coupling between the modes is strong, as would be desirable for delay distortion reduction,

* A more rigorous discussion of pulse distortion in multimode waveguides will be published in a later issue of B.S.T.J.[3]

the steady state power distribution is reached sooner. From Fig. 1 we see that only if the correlation length is large do many modes contribute to energy transport in the waveguide. As far as delay distortion is concerned it might appear advantageous to operate with a short correlation length forcing the multimode waveguide into essentially single-mode steady state operation. However, this method has the disadvantage that most of the power that is initially launched into higher-order modes is lost by radiation so that the waveguide suffers high transient losses. If a light emitting diode is to be used as the transmitter, single-mode operation is most undesirable. That leaves us only with the choice of a long correlation length (if indeed we have a choice) to reduce the power loss from high-order modes. In the limit of very long correlation length all the modes are excited equally strongly in the steady state distribution. If we can still provide strong coupling between the modes there is a chance that the power will be exchanged among all the modes making possible the reduction of delay distortion by mode mixing. Mode mixing takes place via coupling between nearest neighbors in case of long correlation length. The diffusion of power from mode 1 to the highest-order mode and vice versa is thus likely to be slow.

We can get a rough idea of the "speed" with which the power travels from mode 1 to mode 9 (or from mode 9 to mode 1) from Figs. 9 and 12. It is apparent from both figures that it takes approximately $z\bar{\sigma}^2k^2/d = 10^{-6}$ to 10^{-7} before the mode at the other end of the mode spectrum has received an appreciable amount of power from the mode that is initially excited. The same power diffusion must, of course, take place for any other excitation of the modes. But the effect becomes observable when we launch all the power in one mode and watch how it redistributes itself over the other modes. This redistribution of power is part of the transient behavior that results in the steady state distribution. It is thus possible to estimate the distance that is required for one transit of power from mode 1 to mode 9 (or from mode 9 to mode 1) by looking at the second eigenvalue. We know that the steady state is reached as soon as the second term in the series expansion (of power in terms of steady states) becomes negligible compared to the leading first term [see equation (62)]. The second term is quite small when $\alpha^{(2)}z = 2.3$. We thus define a diffusion length L_d by the relation

$$L_d = \frac{2.3}{\alpha^{(2)}}. \tag{1}$$

L_d is the distance along the waveguide that is required for the power in one of the modes at the end of the mode spectrum to transfer an appre-

ciable amount of power (which is a somewhat undefined quantity) across to the mode at the other end of the mode spectrum. For $D/d = 35$ we obtain from Table III and equation (1) $L_d\bar{\sigma}^2k^2/d = 9 \times 10^6$. From Figs. 9 and 12 we see that this is indeed a reasonable estimate for the distance required for a power exchange between mode 1 and 9. The definition (1) allows us immediately to determine the steady state power loss that accompanies this power diffusion among the guided modes. The steady state power loss that occurs over a distance $z = L_d$ is given by

$$\alpha^{(1)}L_d = 2.3\frac{\alpha^{(1)}}{\alpha^{(2)}}. \tag{2}$$

Table V shows a number of values for $\alpha^{(1)}L_d$ for various correlation lengths. The penalty in radiation loss that must be paid for this mode mixing process is relatively high but it improves with increasing correlation length. Mode mixing via the next neighbor power exchange is not likely to be very effective in reducing delay distortion since only a small fraction of power traveling initially in one mode is transferred to the mode at the other end of the mode spectrum in the distance L_d. One might expect that many such diffusion distances would have to fit into the overall length of the guide before delay distortion reduction by mode mixing becomes appreciable. Table V shows that a large correlation length to slab half width ratio is required in order to keep the loss per distance L_d small. Also shown in the table is the normalized exchange length L_d. The numbers were computed for $kd = 82$, the ten-mode case.

For delay distortion equalization it appears desirable to make L_d much shorter than the total guide length L. If we choose $L/L_d = 100$, for example, we compute from the last column of Table V for $D/d = 40$ with $L = 1$ km, $kd = 82$, $\lambda = 1$ μm, and $d = 13$ μm the value $\bar{\sigma} = 3.14$ μm for the required rms deviation of the core-cladding interface irregularities. This value is much larger than accidental irregularities need

TABLE V—LOSS PENALTY $\alpha^{(1)}L_d$ AND NORMALIZED POWER EXCHANGE LENGTH L_d FOR $kd = 82$

D/d	$\alpha^{(1)}L_d$(dB)	$\frac{\bar{\sigma}^2k^2}{d}L_d$
20	3.7	1.64×10^5
30	0.33	4.41×10^5
35	0.25	9.02×10^6
40	0.064	3.01×10^8

to be. It is thus conceivable that an optical fiber could be designed with an intentional core-cladding interface irregularity with long correlation length for the purpose of reducing pulse delay distortion.

IV. THE COUPLING COEFFICIENTS

In Ref. 1 coupled power equations were derived from the coupled wave equations. The coupled wave equations have the form

$$\frac{dA_\nu}{dz} = \sum_{\mu \neq \nu}^{N} c_{\nu\mu} A_\mu e^{i(\beta_\nu - \beta_\mu)z}. \tag{3}$$

With the coupling coefficient written as

$$c_{\nu\mu} = K_{\nu\mu} f(z) \tag{4}$$

and with the assumption that the correlation function of $f(z)$ is Gaussian,

$$\langle f(z) f(z - u) \rangle = \bar{\sigma}^2 e^{-(u/D)^2} \tag{5}$$

($\langle\ \rangle$ indicates an ensemble average), the coupled equations for the average power assume the form

$$\frac{dP_\nu}{dz} = -\alpha_\nu P_\nu + \sqrt{\pi}\, \bar{\sigma}^2 D \sum_{\mu=1}^{N} |K_{\nu\mu}|^2 e^{-[(D/2)(\beta_\nu - \beta_\mu)]^2} (P_\mu - P_\nu). \tag{6}$$

The term $-\alpha_\nu P_\nu$ was added to account for the radiation losses of the modes. Coupling coefficients describing the coupling between the guided modes of a slab waveguide caused by core-cladding interface irregularities were derived in Ref. 2. To obtain the coupling coefficient $c_{\nu\mu}$ from our earlier work, we observe that equations (53) and (60) of Ref. 2 correspond to a solution by perturbation theory of equation (3) for the special case that only the lowest-order even guided TE mode of the slab waveguide is excited. Comparison between the corresponding perturbation solution of (3) and equations (53) and (60) of Ref. 2 allows us to find

$$c_{\nu\mu} = \frac{(n_1^2 - n_2^2)\mathrm{k}^2 a_\nu a_\mu (\gamma_\nu \gamma_\mu)^{\frac{1}{2}}}{2i[|\beta_\nu \beta_\mu| (1 + \gamma_\nu d)(1 + \gamma_\mu d)]^{\frac{1}{2}}} [f(z) - (-1)^{\nu+\mu} h(z)]. \tag{7}$$

The symbols appearing in (7) have the following meaning:

d = core half thickness
n_1 = index of refraction of core material
n_2 = index of refraction of cladding material
$\mathrm{k} = 2\pi/\lambda$ = free-space propagation constant
β_ν = propagation constant of mode ν

$$\gamma_\nu = (\beta_\nu^2 - n_2^2 k^2)^{\frac{1}{2}} \tag{8}$$

$$\kappa_\nu = (n_1^2 k^2 - \beta_\nu^2)^{\frac{1}{2}} \tag{9}$$

$$a_\nu = \begin{cases} \cos \kappa_\nu d & \text{for } \nu = 0, 2, 4 \cdots \\ \sin \kappa_\nu d & \text{for } \nu = 1, 3, 5 \end{cases} \tag{10}$$

$f(z)$ = distortion function of upper core-cladding interface ($f(z) = 0$ indicates a perfect interface at $x = d$)

$h(z)$ = distortion function of lower core-cladding interface ($h(z) = 0$ indicates a perfect interface at $x = -d$).

The propagation constants of the even and odd guided TE modes are obtained with the help of (8) and (9) from the eigenvalue equations. We have for even modes

$$\tan \kappa_\nu d = \frac{\gamma_\nu}{\kappa_\nu} \qquad \nu = 0, 2, 4, \cdots \tag{11}$$

and for odd modes

$$\tan \kappa_\nu d = -\frac{\kappa_\nu}{\gamma_\nu} \qquad \nu = 1, 3, 5, \cdots . \tag{12}$$

With the help of the eigenvalue equations, we can express (10) in the following form:

$$a_\nu = \begin{Bmatrix} (-1)^{\nu/2} \\ (-1)^{(\nu-1)/2} \end{Bmatrix} \frac{\kappa_\nu}{(n_1^2 - n_2^2)^{\frac{1}{2}} k} \quad \begin{matrix} \text{for } \nu = 0, 2, 4 \cdots \\ \text{for } \nu = 1, 3, 5 \cdots . \end{matrix} \tag{13}$$

It is convenient to describe the guided modes in terms of a mode angle θ_ν. We can introduce this angle by the equations

$$\kappa_\nu = n_1 k \sin \theta_\nu , \tag{14}$$

$$\beta_\nu = n_1 k \cos \theta_\nu . \tag{15}$$

Equations (14) and (15) represent the transverse and longitudinal components of the propagation vector of a plane wave in the core and are clearly compatible with (9). The guided mode can be represented as a superposition of two plane waves traveling inside of the core of the slab waveguide. $\pm\theta_\nu$ is the angle that these plane waves form with the waveguide axis.

The function $f(z)$ appearing in (4) is replaced with the sum of $f(z)$ and $h(z)$ in (7). Assuming that the two functions are uncorrelated, and assuming further that they have the same correlation function, we find

$$\langle [f(z) - (-1)^{\nu+\mu} h(z)][f(z-u) - (-1)^{\nu+\mu} h(z-u)] \rangle = 2\bar{\sigma}^2 e^{-(u/D)^2}. \tag{16}$$

By collecting all our results we can finally express the absolute square value of $K_{\nu\mu}$ defined by (4) in the form

$$|K_{\nu\mu}|^2 = \frac{n_1^2 k^2 \sin^2\theta_\nu \sin^2\theta_\mu}{2d^2\left(1+\frac{1}{\gamma_\nu d}\right)\left(1+\frac{1}{\gamma_\mu d}\right)\cos\theta_\nu \cos\theta_\mu}. \tag{17}$$

For small values of $(n_1/n_2 - 1)$ we have $\theta_\nu \ll 1$. For modes far from cutoff we have, in addition, $\gamma_\nu d \gg 1$. Under these conditions, (17) simplifies to the expression

$$|K_{\nu\mu}|^2 = \frac{n_1^2 k^2 \theta_\nu^2 \theta_\mu^2}{2d^2}. \tag{18}$$

Far from cutoff we can approximate the solution of the eigenvalue equations (11) and (12) as follows:

$$\kappa_\nu d = (\nu + 1)\frac{\pi}{2} \quad \text{for} \quad \nu = 0, 1, 2, 3, 4 \cdots . \tag{19}$$

The mode angles can then be expressed as

$$\sin\theta_\nu = \frac{\pi(\nu+1)}{2n_1 kd}. \tag{20}$$

Finally, it is important to know the largest mode angle that can occur. Mode guidance ceases to exist when the angle, that the plane-wave components of the guided mode form with the core-cladding interface, exceeds the total internal reflection angle θ_c defined by

$$n_1 \cos\theta_c = n_2 . \tag{21}$$

For $(n_1/n_2 - 1) \ll 1$ we obtain approximately

$$\theta_c = \left[2\left(1-\frac{n_2}{n_1}\right)\right]^{\frac{1}{2}}. \tag{22}$$

Combining (20) and (22) allows us to find an approximate value for the number of modes N that the slab waveguide can support:

$$N = \frac{2}{\pi} n_1 kd\left[2\left(1-\frac{n_2}{n_1}\right)\right]^{\frac{1}{2}} - 1. \tag{23}$$

V. RADIATION LOSSES

In order to be able to evaluate the coupled power equations (6) we need convenient approximations for the radiation losses α_ν of the guided modes.

The radiation loss problem has been solved in a general way in Ref. 6. Equation (14) of Ref. 6 gives the radiation losses of the even and odd guide modes.

$$\alpha_\nu = \int_{-n_2 k}^{n_2 k} \langle |\, F(\beta_\nu - \beta)\, |^2 \rangle I_\nu(\beta)\, d\beta \tag{24}$$

with

$$I_\nu(\beta) = \frac{(n_1^2 - n_2^2)k^3 n_1 \sin^2 \theta_\nu}{2\pi d \cos \theta_\nu \left(1 + \dfrac{1}{\gamma_\nu d}\right)} \cdot \left[\frac{\rho \cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} + \frac{\rho \sin^2 \sigma d}{\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d}\right] \tag{25}$$

and with the Fourier coefficient of the core-cladding interface function

$$F(\beta_\nu - \beta) = \frac{1}{\sqrt{L}} \int_0^L f(z) e^{-i(\beta_\nu - \beta)z}\, dz. \tag{26}$$

Equations (10), (13), (14), and (15) have been used to express (25) in this form. In addition, (25) has been multiplied with a factor 2 to account for the fact that both core-cladding interfaces have irregularities (contrary to the assumption in Ref. 6) that are statistically independent of each other but have the same correlation function. There are two new parameters in (25):

$$\rho = (n_2^2 k^2 - \beta^2)^{\frac{1}{2}} \tag{27}$$

and

$$\sigma = (n_1^2 k^2 - \beta^2)^{\frac{1}{2}}. \tag{28}$$

The parameter β is the propagation constant (in z direction) of the radiation modes. Using (5) and assuming that $L \gg D$ we obtain from (26)

$$\langle |\, F(\beta_\nu - \beta)\, |^2 \rangle = \sqrt{\pi}\, \tilde{\sigma}^2 D e^{-[(D/2)(\beta_\nu - \beta)]^2}. \tag{29}$$

The loss expression (24) must be simplified before it can be used for our purposes. We are interested only in multimode waveguides with $kd \gg 1$. The functions $\sin \sigma d$ and $\cos \sigma d$ thus vary rapidly as functions of β. It is impossible to obtain an approximation for all values of D/d. We begin by assuming $D/d \ll 1$. In this case, we can replace the exponential function in (29) by unity and obtain

$$\alpha_\nu = \frac{(n_1^2 - n_2^2)k^3 n_1 \sin^2 \theta_\nu}{2\sqrt{\pi} \cos \theta_\nu \left(1 + \frac{1}{\gamma_\nu d}\right)} \bar{\sigma}^2 \frac{D}{d}$$

$$\cdot \int_{-n_2 k}^{n_2 k} \left[\frac{\rho \cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} + \frac{\rho \sin^2 \sigma d}{\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d} \right] d\beta. \quad (30)$$

Consider the terms in the integrand. The first term can be written

$$G_1 = \frac{\rho \cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} = \frac{\rho \cos^2 \sigma d}{\rho^2 + (\sigma^2 - \rho^2) \sin^2 \sigma d}. \quad (31)$$

The sine and cosine functions oscillate rapidly while ρ is only a slowly varying function. The contribution of the second term in the denominator is slight since this term vanishes when the cosine term in the numerator assumes its maximum. On the other hand, when the second term in the denominator is largest, the numerator is zero so that the value of the denominator does not matter. We can thus write to a crude approximation

$$G_1 \approx \frac{\cos^2 \sigma d}{\rho}. \quad (32)$$

The average value of the cosine square function is 1/2 so that we approximate further

$$G_1 \approx \frac{1}{2\rho}. \quad (33)$$

It appears that this approximation may be very poor at $\rho = 0$. However, by converting the integration variable from ρ to β, we see that

$$d\beta = -\frac{\rho}{\beta} d\rho \quad (34)$$

showing that there is no pole at $\rho = 0$. By an analogous argument we find that the second term in the integrand can also be approximated as

$$G_2 \approx \frac{1}{2\rho}. \quad (35)$$

The integral in (30) thus assumes the value

$$\int_{-n_2 k}^{n_2 k} (G_1 + G_2)\, d\beta \approx \int_{-n_2 k}^{n_2 k} \frac{1}{\rho} d\beta = \pi. \quad (36)$$

For $D/d < (1/2n_2 kd)$ we obtain the following approximation for the radiation loss of the νth guided mode of the slab waveguide:

$$\alpha_\nu = \frac{\sqrt{\pi}\,(n_1^2 - n_2^2) n_1 k^3 \sin^2 \theta_\nu}{2 \cos \theta_\nu \left(1 + \frac{1}{\gamma_\nu d}\right)} \bar{\sigma}^2 \frac{D}{d}. \tag{37}$$

Next we try to obtain an approximation for large values of the correlation length, $D/d \gg 1$. The general expression for the radiation loss of the νth guided mode follows from (24), (25), and (29):

$$\alpha_\nu = \frac{(n_1^2 - n_2^2) n_1 k^3 \sin^2 \theta_\nu}{2\sqrt{\pi}\left(1 + \frac{1}{\gamma_\nu d}\right) \cos \theta_\nu} \bar{\sigma}^2 \frac{D}{d}$$

$$\cdot \int_{-n_2 k}^{n_2 k} e^{-[(D/2)(\beta_\nu - \beta)]^2}$$

$$\cdot \left[\frac{\rho \cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} + \frac{\rho \sin^2 \sigma d}{\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d}\right] d\beta. \tag{38}$$

The exponential function under the integral sign decreases very rapidly with increasing values of $\beta_\nu - \beta$ for $D/d \gg 1$. Since the largest value that β can assume is $\beta = n_2 k$ only the immediate vicinity of the upper limit of the integration range contributes to the integral. In this region we have $\rho \ll n_2 k$. In order to be able to work out approximations for the case of large correlation length we must consider two more subdivisions, the case that D/d is small enough so that the exponential factor under the integral sign in (38) varies slowly compared to the rapid oscillations of the sine and cosine functions and the opposite case where the exponential function decays to insignificant values within one cycle of the oscillations of the oscillatory functions.

In the first case, slowly varying exponential function, we can consider ρ and σ approximately constant over one cycle of oscillation except for the σd term appearing in the argument of the oscillatory functions and consider the average of the integral over one period

$$\frac{1}{\beta_2 - \beta_1} \int_{\beta_1}^{\beta_2} \frac{\rho \cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} d\beta$$

$$\approx \frac{\rho}{2\pi} \int_0^{2\pi} \frac{\cos^2 x}{\rho^2 \cos^2 x + \sigma^2 \sin^2 x} dx \approx \frac{1}{\sigma + \rho}. \tag{39}$$

And, similarly, for the second term of the integrand we find

$$\frac{1}{\beta_2 - \beta_1} \int_{\beta_1}^{\beta_2} \frac{\rho \sin^2 \sigma d}{\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d} d\beta \approx \frac{1}{\sigma + \rho}. \tag{40}$$

Since only small values of ρ can contribute to the integral, because of the rapid decay of the exponential function, we use the approximations

$$\frac{1}{\sigma + \rho} \approx \frac{1}{\sqrt{n_1^2 - n_2^2 \mathrm{k}}} \tag{41}$$

and

$$\beta = n_2 \mathrm{k} - \frac{\rho^2}{2n_2 \mathrm{k}}. \tag{42}$$

Using all these approximations and the change of integration variable

$$d\beta = -\frac{\rho}{\beta} d\rho \approx -\frac{\rho}{n_2 \mathrm{k}} d\rho, \tag{43}$$

the integral in (38) can be approximated by the following expression:

$$\begin{aligned}\int_{-n_2\mathrm{k}}^{n_2\mathrm{k}} \frac{2}{\sigma + \rho} e^{-[(D/2)(\beta_\nu - \beta)]^2} d\beta &\approx \frac{2}{n_2 \mathrm{k}^2 \sqrt{n_1^2 - n_2^2}} \\ &\cdot \int_0^\infty \rho \exp\left\{-\left[\frac{D}{2}\left(\beta_\nu - n_2\mathrm{k} + \frac{\rho^2}{2n_2\mathrm{k}}\right)\right]^2\right\} d\rho \\ &= \frac{4}{\mathrm{k}\sqrt{n_1^2 - n_2^2}\, D} \int_{[(D/2)(\beta_\nu - n_2\mathrm{k})]}^\infty e^{-u^2} du \\ &= \frac{2\sqrt{\pi}}{\mathrm{k}D\sqrt{n_1^2 - n_2^2}} \left\{1 - \operatorname{erf}\left[\frac{D}{2}(\beta_\nu - n_2\mathrm{k})\right]\right\}. \end{aligned} \tag{44}$$

We have now finally obtained the result that the radiation loss of the νth mode can be approximated by

$$\alpha_\nu = \frac{n_1 \mathrm{k}^2 \sqrt{n_1^2 - n_2^2} \sin^2 \theta_\nu}{d\left(1 + \frac{1}{\gamma_\nu d}\right) \cos \theta_\nu} \bar{\sigma}^2 \left\{1 - \operatorname{erf}\left[\frac{D}{2}(\beta_\nu - n_2\mathrm{k})\right]\right\}. \tag{45}$$

The range of applicability of (45) is obtained by considering that we must require the exponential function in (44) to change only slightly over the range $\Delta\rho$ corresponding to $\Delta(\sigma d) = 2\pi$. This condition can be expressed as

$$\frac{\rho \Delta\rho D^2}{2n_2\mathrm{k}} \left(\beta_\nu - n_2\mathrm{k} + \frac{\rho^2}{2n_2\mathrm{k}}\right) \ll 1. \tag{46}$$

The increment $\Delta\rho$ is obtained from

$$\Delta\rho = 2\pi \frac{\sigma}{\rho} \frac{1}{d}. \tag{47}$$

The condition (46) must hold primarily for small values of ρ. We use, therefore,

$$\rho = \eta \, \Delta\rho. \tag{48}$$

With η being a small number such as 2 or 5 and obtained from (46), (47), and (48) with $\sigma \approx (n_1^2 - n_2^2)^{\frac{1}{2}}\mathrm{k}$,

$$\frac{2}{(\beta_\nu - n_2\mathrm{k})d + \dfrac{(n_1^2 - n_2^2)\mathrm{k}d}{2n_2}} < \frac{D}{d} < \left\{ \frac{n_2}{\pi\left[(\beta_\nu - n_2\mathrm{k})d + \eta\pi \dfrac{(n_1^2 - n_2^2)^{\frac{1}{2}}}{n_2}\right](n_1^2 - n_2^2)^{\frac{1}{2}}} \right\}^{\frac{1}{2}}. \tag{49}$$

The left-hand side of this inequality follows from the requirement that the exponential function in (44) must drop to small values as ρ grows from 0 to approximately $(n_1^2 - n_2^2)^{\frac{1}{2}}\mathrm{k}$.

Finally, we obtain an approximation for very large values of D/d if we assume that the exponential factor in (38) decreases very appreciably over an interval corresponding to one oscillation period of the oscillatory functions. We can now use the approximation

$$\sigma = (n_1^2 - n_2^2)^{\frac{1}{2}}\mathrm{k} \tag{50}$$

treating σ as independent of ρ. Expanding the factor that multiplies the exponential function in (38) in a power series in terms of ρ at $\rho = 0$, keeping only the first none-vanishing term, results in

$$\begin{aligned}
&\int_0^\infty \frac{\rho}{\beta} e^{-[(D/2)(\beta_\nu - \beta)]^2} \\
&\quad \cdot \left[\frac{\rho \cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} + \frac{\rho \sin^2 \sigma d}{\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d} \right] d\rho \\
&\approx \frac{1}{n_2(n_1^2 - n_2^2)\mathrm{k}^3} (\cot^2 \sigma d + \tan^2 \sigma d) \\
&\quad \cdot \int_0^\infty \rho^2 \exp\left\{ -\left[\frac{D}{2}\left(\beta_\nu - n_2\mathrm{k} + \frac{\rho^2}{2n_2\mathrm{k}}\right)\right]^2 \right\} d\rho \\
&\approx \frac{2(\pi n_2\mathrm{k})^{\frac{1}{2}}}{\mathrm{k}^2(n_1^2 - n_2^2)(\beta_\nu - n_2\mathrm{k})^{\frac{1}{2}} D^3} (\cot^2 \sigma d + \tan^2 \sigma d) e^{-[(D/2)(\beta_\nu - n_2\mathrm{k})]^2}.
\end{aligned} \tag{51}$$

If $\beta_\nu = n_2\mathrm{k}$, (51) becomes infinitely large. The approximation leading to the solution of the integral is violated in this case and the result

becomes meaningless. However, this violation of the applicability of the approximate solution of (51) can happen only for the highest-order mode and only if it happens to be directly at its cutoff frequency. Our approximation, if applied to this case, gives a loss value that is too large. Using too large a radiation loss for the highest-order mode affects the power distribution in all the other modes only slightly. The radiation loss of the highest-order mode is large in any case. Power coupled from the neighboring guided modes to this mode is lost rapidly. Using too large a loss value for this mode makes little difference to any of the other modes. We thus use (51) for all values of β_ν .

A more serious violation of the applicability of (51) occurs if either the function cot σd or the function tan σd should become infinite or at least very large. In both of these cases the integral assumes the form

$$\int_0^\infty \frac{1}{\beta} e^{-[(D/2)(\beta_\nu-\beta)]^2}\, d\rho \approx \frac{1}{n_2 \mathrm{k}} \int_0^\infty e^{-[(D/2)(\beta_\nu - n_2\mathrm{k} + (\rho^2/2n_2\mathrm{k}))]^2}\, d\rho$$

$$\approx \frac{\pi^{\frac{1}{2}}}{D[n_2\mathrm{k}(\beta_\nu - n_2\mathrm{k})]^{\frac{1}{2}}} e^{-[(D/2)(\beta_\nu - n_2\mathrm{k})]^2}. \tag{52}$$

For very large D/d the radiation loss approximation is

$$\alpha_\nu = \bar{\sigma}^2 \frac{n_1 \mathrm{k}^3 \sin^2 \theta_\nu e^{-[(D/2)(\beta_\nu - n_2\mathrm{k})]^2}}{2d[n_2\mathrm{k}(\beta_\nu - n_2\mathrm{k})]^{\frac{1}{2}}\left(1 + \dfrac{1}{\gamma_\nu d}\right) \cos \theta_\nu}$$

$$\cdot \begin{cases} \dfrac{2n_2}{\mathrm{k}D^2(\beta_\nu - n_2\mathrm{k})} (\cot^2 \sigma d + \tan^2 \sigma d) & \text{for } \tan \sigma d \neq 0 \\ & \text{and } \cot \sigma d \neq 0 \\ (n_1^2 - n_2^2) & \text{for } \tan \sigma d = 0 \\ & \text{or } \cot \sigma d = 0. \end{cases} \tag{53}$$

Equation (53) holds for values of D/d that are much larger than the D/d values in the range indicated in (49).

VI. THE EIGENVALUE PROBLEM

Knowing the coupling coefficients and the radiation losses allows us to determine the power distribution in the multimode dielectric slab waveguide as a function of the distance z along the guide. Introducing the abbreviations

$$h_{\nu\mu} = \sqrt{\pi}\, \bar{\sigma}^2 D \mid K_{\nu\mu} \mid^2 e^{-[(D/2)(\beta_\nu - \beta_\mu)]^2} \tag{54}$$

and

$$b_\nu = \sum_{\mu=1}^{N} h_{\nu\mu} , \tag{55}$$

we can write the coupled power equations (6) in the form

$$\frac{dP_\nu}{dz} = -(\alpha_\nu + b_\nu)P_\nu + \sum_{\mu=1}^{N} h_{\nu\mu}P_\mu . \tag{56}$$

The trial solution

$$P_\nu = B_\nu e^{-\alpha z} \tag{57}$$

converts (56) into an eigenvalue problem

$$\sum_{\mu=1}^{N} [h_{\nu\mu} - (\alpha_\mu + b_\mu - \alpha)\delta_{\nu\mu}]B_\nu = 0. \tag{58}$$

The coefficient matrix of this problem is real and symmetric as can be seen from (54) and the condition (11) of Ref. 1. This latter condition can be expressed in the form

$$| K_{\nu\mu} |^2 = | K_{\mu\nu} |^2. \tag{59}$$

The symmetry condition (59) follows also directly from (17). The eigenvalue α is obtained from the eigenvalue equation

$$| h_{\nu\mu} - (\alpha_\mu + b_\mu - \alpha)\ \delta_{\nu\mu} | = 0. \tag{60}$$

The vertical lines in (60) indicate that the determinant of the matrix, whose $\nu\mu$ element appears explicitly, must be formed. The eigenvalue equation is an algebraic equation of order N providing N different solutions for the eigenvalues $\alpha^{(i)}$. The eigenvectors, whose elements are $B_\nu^{(i)}$, are mutually orthogonal and will be assumed to be normalized,

$$\sum_{\nu=1}^{N} B_\nu^{(i)}B_\nu^{(j)} = \delta_{ij} . \tag{61}$$

The general solution of (56) can now be expressed as a linear superposition of the N eigensolutions,

$$P_\nu(z) = \sum_{i=1}^{N} c_i B_\nu^{(i)} e^{-\alpha^{(i)}z}. \tag{62}$$

The expansion coefficients c_i must be determined from the given power distribution at $z = 0$. With the help of (61) we obtain from (62)

$$c_i = \sum_{\nu=1}^{N} B_\nu^{(i)}P_\nu(0). \tag{63}$$

VII. CONCLUSIONS

In this paper we have shown that coupling between the guided modes of a multimode waveguide causes the power versus mode number distribution to settle down to a steady state provided the signal is allowed to travel far enough in the waveguide. This steady state applies, of course, only to the CW case. For very long correlation length of the core-cladding interface irregularities the steady state distribution contains equal power in all the modes. For very short correlation length, on the other hand, only the lowest-order mode carries an appreciable amount of power, forcing the fiber into single-mode steady state operation.

The results of this paper have some application to delay distortion equalization. If the power carried by the guided modes is exchanged rapidly among them, the pulse distortion caused by the different group velocities of the modes is partially compensated. Coupling among the modes is of necessity accompanied by radiation losses. Effective pulse delay distortion equalization has a chance of working only if the correlation length of the core-cladding irregularities is long since the penalty paid in radiation loss becomes high for short correlation length. In addition, only for long correlation length do all the modes carry power in the steady state distribution.

A detailed discussion of the numerical results and the properties of multimode waveguides is to be found at the beginning of the paper.

REFERENCES

1. Marcuse, D., "Derivation of Coupled Power Equations," B.S.T.J., *51*, No. 1 (January 1972), pp. 229–237.
2. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3187–3215.
3. Marcuse, D., "Pulse Propagation in Multimode Dielectric Waveguides," to be published in B.S.T.J.
4. Personick, S. D., "Time Dispersion in Dielectric Waveguides," B.S.T.J., *50*, No. 3 (March 1971), pp. 843–859.
5. Rowe, H. E., and Young, D. T., "Transmission Distortion in Multimode Random Waveguides," to be published in IEEE Trans. MTT.
6. Marcuse, D., "Radiation Losses of Dielectric Waveguides in Terms of the Power Spectrum of the Wall Distortion Function," B.S.T.J., *48*, No. 10 (December 1969), pp. 3233–3242.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Hot Carrier Effects in the Integral Charge-Control Model for Bipolar Transistors

By G. PERSKY

(Manuscript received August 17, 1971)

The integral charge-control model for bipolar transistors is rederived with the purpose of elucidating hot carrier effects. In its original derivation the model contained an additive hot carrier contribution to the base charge of possible significance in narrow-base transistors. Inclusion of this term is shown to be unnecessary. However, careful examination of the potentials appearing in the formalism has disclosed other hot carrier effects. These could lower the transconductance of a transistor operating in or near saturation, particularly if the base has a low number of impurities per unit area, but would otherwise be unobservable.

I. INTRODUCTION

The integral charge-control model (ICM) provides an elegant and compact description of the one-dimensional transport physics of transistors by relating collector current to the junction voltages and total base majority carrier charge.[1,2] The original derivation of the model indicates a possible need for supplementing the base charge in the ICM relation with a term inversely proportional to the minority carrier saturation velocity when base widths are very small ($\sim$1000 Å).[1] It is shown herein that this term is an artifact arising from inappropriate treatment of the diffusion current contribution to the transport equation. There are, however, additional hot carrier modifications of the charge-control relation that have not been included in previous treatments. These originate in the heating of a reverse current by the built-in field in a junction *not* supporting a large reverse bias, and should be manifest only in saturated or near-saturated transistor operation. With this exception, the standard ICM relation [equation (15) of Ref. 1] remains valid to the same extent as the macroscopic current transport equation, even for very narrow base widths.

II. DERIVATION

Considering a pnp transistor, we integrate the one-dimensional macroscopic equation for hole transport to obtain the integral charge-control relation. Our derivation largely parallels that of H. K. Gummel.[1] The essential differences are representation of diffusive transport by $-q\nabla(Dp)$ rather than $-qD\nabla p$, and a more detailed treatment of the potentials. Both diffusion expressions are, of course, identical if D is coordinate independent. When coordinate dependencies arise from local carrier heating, the former can be more readily justified by integration of the Boltzmann equation, and is therefore to be preferred.[3] Thus, as a starting equation we take

$$j_p = q\left(\frac{qE}{kT_d}\eta - \frac{d\eta}{dx}\right), \tag{1}$$

where j_p is the hole current density, E is the electric field, and kT_d and η are given by

$$kT_d \equiv qD/\mu, \tag{2}$$

$$\eta \equiv Dp. \tag{3}$$

In relation (2), T_d is the hole "diffusion temperature," which is *defined* from the local diffusion coefficient and mobility by the Einstein relation. The variable η is the product of the local diffusion coefficient and hole density.

The full solution to (1) is the sum of the homogeneous solution for $j_p = 0$, and the particular solution. From the homogeneous equation we obtain

$$\eta_h = e^{-\psi(x)}, \tag{4}$$

where

$$\frac{d\psi}{dx} = -\frac{qE}{kT_d}. \tag{5}$$

Note that ψ is a potential normalized to the local value of kT_d, and is nonconservative in regions where T_d varies. The particular solution to (1) is

$$\eta_p = -\frac{1}{q}e^{-\psi(x)}\int^x j_p(x')e^{\psi(x')}\,dx'. \tag{6}$$

Thus

$$\eta = e^{-\psi(x)} - \frac{1}{q}e^{-\psi(x)}\int^x j_p(x')e^{\psi(x')}\,dx'. \tag{7}$$

Equation (7) is now evaluated at x_E, the outer edge of the emitter junction, and x_C, the outer edge of the collector junction. This procedure yields

$$\eta(x_E)e^{\psi(x_E)} - \eta(x_C)e^{\psi(x_C)} = \frac{1}{q}\int_{x_E}^{x_C} j_p(x')e^{\psi(x')}\,dx'. \tag{8}$$

Following Gummel,[1] we may account in a crude way for recombination through the introduction of a quantity $\bar{a}$ defined by

$$\int_{x_E}^{x_C} j_p(x)e^{\psi(x)}\,dx = \bar{a}j_c\int_{x_E}^{x_C} e^{\psi(x)}\,dx, \tag{9}$$

where $j_c = j_p(x_C)$ is the collector hole current density. Consequently, $\bar{a} \geqq 1$, and assumes the value unity in the absence of recombination. Upon substitution of (9) into (8), the resulting equation may be solved for j_c.

$$j_c = \frac{q}{\bar{a}}\frac{\eta(x_E)e^{\psi(x_E)} - \eta(x_C)e^{\psi(x_C)}}{\int_{x_E}^{x_C} e^{\psi(x)}\,dx}. \tag{10}$$

There remains evaluation of the contributions to (10). At coordinates x_E and x_C in the undepleted bulk material of the emitter and collector there is no carrier heating and the diffusion coefficient has its zero field value D_o. Hence, assuming the emitter and collector have the same low field mobility,

$$\eta(x_E) = D_o p(x_E) \tag{11a}$$

$$\eta(x_C) = D_o p(x_C) \tag{11b}$$

so that (10) may be rewritten

$$j_c = \frac{qD_o}{\bar{a}}\frac{p(x_E)e^{\psi(x_E)} - p(x_C)e^{\psi(x_C)}}{\int_{x_E}^{x_C} e^{\psi(x)}\,dx}. \tag{12}$$

Since the normalized potentials in (12) are, in general, nonconservative, it is convenient to introduce a conservative electrical potential $\hat{\psi}(x)$ which is everywhere normalized to the lattice temperature T_o. Then in regions where the holes are not heated, their concentration is given by

$$p(x) = n_i e^{\varphi_p(x)-\hat{\psi}(x)}, \tag{13}$$

where n_i is the intrinsic carrier concentration and $\varphi_p(x)$ is the hole

quasi-Fermi level normalized to T_o. Equation (13) can be invoked at x_E and x_C, yielding

$$j_c = \frac{qn_iD_o}{\bar{a}} \cdot \frac{e^{\psi(x_E)-\hat{\psi}(x_E)} \cdot e^{\varphi_p(x_E)} - e^{\psi(x_C)-\hat{\psi}(x_C)} \cdot e^{\varphi_p(x_C)}}{\int_{x_E}^{x_C} e^{\psi(x)-\hat{\psi}(x)} \cdot e^{\hat{\psi}(x)}\, dx}. \tag{14}$$

The relationship between $\psi(x)$ and $\hat{\psi}(x)$ is arbitrary to within a constant, permitting a choice of the coordinate at which $\psi(x)$ and $\hat{\psi}(x)$ coincide. Although (14) is implicitly "gauge invariant," its explicit form will depend on the choice made. The most symmetrical appearance is obtained if one relates $\hat{\psi}(x)$ to $\psi(x)$ by

$$\hat{\psi}(x) = -\frac{q}{kT_o}\int_{x_B}^{x} E\, dx + \psi(x_B), \tag{15}$$

where x_B is any coordinate in the base. Then (14) becomes

$$j_c = \frac{qn_iD_o}{\bar{a}} \frac{\gamma(x_E)e^{\varphi_p(x_E)} - \gamma(x_C)e^{\varphi_p(x_C)}}{\int_{x_E}^{x_C} \gamma(x)e^{\hat{\psi}(x)}\, dx}, \tag{16}$$

where

$$\gamma(x) \equiv e^{\psi(x)-\hat{\psi}(x)} = \exp\int_{x_B}^{x} \frac{qE(T_d - T_o)}{kT_dT_o}\, dx. \tag{17}$$

The function $\gamma(x)$ provides a uniform treatment of the hot carrier effects in (16), which all arise when hole current is drifted in the direction of the field, and power absorption from the field raises the hole diffusion temperature T_d above T_o.

The ICM relation follows from (16) if the quasi-Fermi level for the electrons in the base may be regarded as constant. This implies the absence of substantial dc base majority carrier current, such as would arise if there were both high-level injection and poor current gain.[4] For a constant electron quasi-Fermi level φ_{nb} one has

$$n(x) = n_i e^{\hat{\psi}(x)-\varphi_{nb}} \tag{18}$$

and

$$V_{eb} = \frac{kT_o}{q}(\varphi_p(x_E) - \varphi_{nb}), \tag{19a}$$

$$V_{cb} = \frac{kT_o}{q}(\varphi_p(x_C) - \varphi_{nb}), \tag{19b}$$

where V_{eb} and V_{cb} are respectively the applied emitter-base and collector-base voltages exclusive of ohmic drops. Insertion of (18) and (19) into (16) results in

$$j_c = \frac{n_i^2 q D_a}{\bar{a}} \frac{\gamma(x_E)e^{qV_{eb}/kT_a} - \gamma(x_C)e^{qV_{cb}/kT_a}}{\int_{x_E}^{x_C} \gamma(x)n(x)\,dx}. \tag{20}$$

Letting A denote the active cross-sectional area of the transistor, and defining an effective base majority carrier charge by

$$Q_b^* = qA \int_{x_E}^{x_C} \gamma(x)n(x)\,dx, \tag{21}$$

one arrives at the ICM relation for the collector current I_c.

$$I_c = -\frac{(qn_iA)^2 D_a}{\bar{a}} \frac{\gamma(x_E)e^{qV_{eb}/kT_a} - \gamma(x_C)e^{qV_{cb}/kT_a}}{Q_b^*}. \tag{22}$$

If one neglects carrier heating, $\gamma(x) = 1$ for all x and Q_b^* reduces to Q_b, the total majority charge (within the active region) that communicates with the base terminal. Equation (22) then becomes identical to the integral charge-control relation derived by Gummel.[1]

III. DISCUSSION AND CONCLUSION

We have shown that inclusion of the diffusion coefficient within the gradient operation in the current transport equation automatically eliminates additive contributions to the defining integral for the base charge in the ICM. However, careful examination of the nonconservative potentials appearing in the formalism discloses other hot carrier contributions that have not been previously considered. In equation (22) these are embodied in $\gamma(x_E)$, $\gamma(x_C)$, and Q_b^*. For forward operation of the transistor, $\gamma(x_E) = 1$ because the holes do not absorb power from the emitter junction field. On the other hand, carrier heating can occur in the collector junction and, in accordance with (17), result in $\gamma(x_C) > 1$. However, this effect would be discernible only for reasonably large values of exp (qV_{cb}/kT), requiring the transistor to be in or near saturation. Heating must then be produced by the built-in field. Similar considerations apply to the effective charge defined by (21). Reverse bias of the collector causes the Boltzmann tail of $n(x)$ to fall off very fast within the collector junction and make little contribution to the integral. Saturated or near-saturated operation of the transistor is therefore required for carrier heating to affect Q_b^*. Further-

more, the number of impurities per unit area of the base must be low for the Boltzmann tails within the junctions to make any significant contributions to the total base charge. Since $\gamma(x) > 1$ within the collector junction, Q_b^* will exceed Q_b. Therefore, by increasing $\gamma(x_C)$ and the effective base charge, carrier heating in the collector junction tends to decrease the collector current for a given set of applied voltages. The diminution in I is plausible in view of the decreased effectiveness of the collector junction as a sink for the minority holes diffusing across the base when their mobility within the junction is lowered by carrier heating.

A number of important effects, such as impact ionization and base crowding, have not been included in this treatment. High current gain has been assumed. The question of the ultimate validity of the macroscopic transport equation in inhomogeneous high-field regions has not been addressed.

IV. ACKNOWLEDGMENTS

The author wishes to thank R. Edwards, H. K. Gummel, B. T. Murphy, and D. L. Scharfetter for useful conversations and critically reading the manuscript. D. L. Scharfetter was especially helpful in suggesting a key improvement in the treatment.

REFERENCES

1. Gummel, H. K., "A Charge Control Relation for Bipolar Transistors," B.S.T.J., *49*, No. 1 (January 1970), pp. 115–120.
2. Gummel, H. K., and Poon, H. C., "An Integral Charge Control Model of Bipolar Transistors," B.S.T.J., *49*, No. 5 (May–June 1970), pp. 827–852.
3. Persky, G., and Bartelink, D. J., Phys. Rev., *B1*, No. 4, (1970), pp. 1614 ff.
4. Gradients in the quasi-Fermi level can also be produced by large transient base charging currents. Ohmic drops across the inactive base can account for quasi-Fermi level gradients therein. In the active base the problem is essentially two-dimensional and beyond the scope of this treatment.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

The Design and Embodiment of Magnetic Domain Encoders and Single-Error Correcting Decoders for Cyclic Block Codes

By S. V. AHAMED

(Manuscript received July 19, 1971)

This paper explores the possibilities of accomplishing the functional requirements of encoders and single-error correcting decoders for cyclic block codes using the inherent properties of magnetic domains. Typical designs and embodiments of such encoders and decoders are presented with field access propagations for moving the magnetic domains.

I. INTRODUCTION

The properties of magnetic domains have been studied by A. H. Bobeck,[1-3] by A. J. Perneski,[4] by U. F. Gianola,[3] and by A. A. Thiele.[5,6] The applications of such magnetic domains for storage and logic are described by A. H. Bobeck, R. F. Fischer, and A. J. Perneski,[7,8] and by P. I. Bonyhard, et al.[9] This paper proposes the possible applications of these results to the construction of encoders and single-error correcting decoders for cyclic block codes.

Single-error correcting codes were introduced by R. W. Hamming in 1950.[10] In 1960, R. C. Bose and D. K. Ray-Chaudhuri[11] formulated a class of multiple-error correcting cyclic codes. W. W. Peterson[12] has presented a variety of logic circuits for encoders and decoders. These circuits conventionally employ semiconductor logic elements. Such circuits are discussed in some detail by R. W. Lucky, J. Salz, and E. J. Weldon, Jr.,[13] and by E. R. Berlekamp.[14]

The encoders proposed in this paper are for cyclic block codes and the decoders are limited to single-error correcting decoders for such codes. In general, these codes constitute a set of BCH codes named after Bose, Chaudhuri, and A. Hocquenghem.[15] The paper is divided into four parts. Part A provides an insight into the fundamentals necessary for a qualitative understanding of magnetic domain functions. Part B deals with encoders for cyclic block codes and Part C deals

with single-error correcting decoders for such codes. Part D discusses the various types of magnetic materials suitable for embodiments and their characteristics. Each of the two parts B and C is divided into two sections; section 1 introduces the fundamentals of encoding and decoding while section 2 leads into the conversion of conventional encoding and decoding to serial encoding and decoding, and describes these functions with magnetic domains. The expert in magnetic domain devices may skip Part A and the expert in coding theory may skip the first sections of Parts B and C.

(PART A)

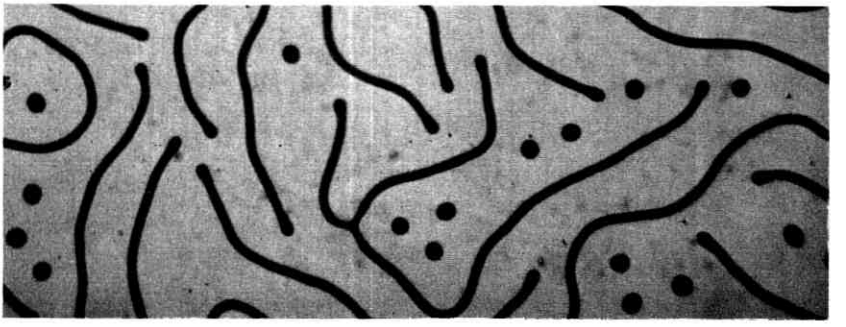
II. MAGNETIC DOMAINS AND THEIR FUNCTIONS

2.1 *Introduction to Orthoferrites and Domains*

Rare earth orthoferrites have a formula $RFeO_3$ where R is a rare earth. Very thin platelets of the orthoferrite crystals are prepared so that the appropriate crystalline axis (generally the 001 or *c* axis) is normal to the surface of the platelet. Magnetic domains with their direction or magnetization normal to the surface of such platelets may be observed by Faraday effect. Such domains may also be observed (Fig. 1a) in very thin epitaxial garnet films deposited on suitable substrates. When these domains are subjected to a bias field opposing the magnetic moment enclosed within them, they shrink (Fig. 1b) to microscopic and submicroscopic sizes and are cylindrical in shape. Such cylindrical magnetic domains, sometimes called bubbles, generally

(a)

Fig. 1a—Magnetic domains in a typical epitaxial film 5 to 8 microns deep, deposited on Gadolinium–Gallium–Garnet (GGG) substrate 20 to 40 mils thick. Magnification is 340.



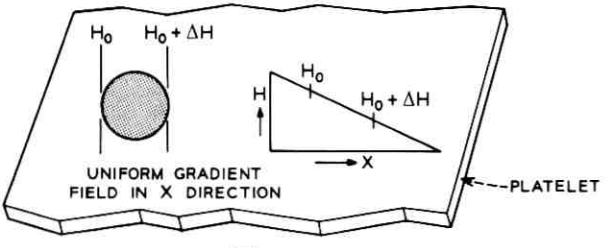
(b)

Fig. 1b—Formation of "bubbles" from magnetic domains at a bias field of 30 Oe in the same material used in Fig. 1a. Magnification is 340.

are a few microns in diameter and are stable under proper bias field conditions. Bubbles may be used to store information and to carry out certain elementary logical functions.

2.2 *Propagation of Bubbles*

Bubbles respond to bias field gradients in the plane of the platelet (or film) hosting them by moving in a direction which tends to minimize the net energy. A bubble of diameter d located in a uniform gradient field would tend towards the position of reduced bias (Fig. 2a). Bubble velocities yielding a bit rate of over two or three megacycles have been achieved in selected magnetic materials. There are two basic methods of providing such an inplane field gradient. The first method depends on a current in a conductor loop which produces a field to attract the neighboring bubble directly beneath a loop formed by a conductor (Fig. 2b). This method is called "conductor propagation" since a



(a)

Fig. 2a—A cylindrical domain located in a uniform gradient field.

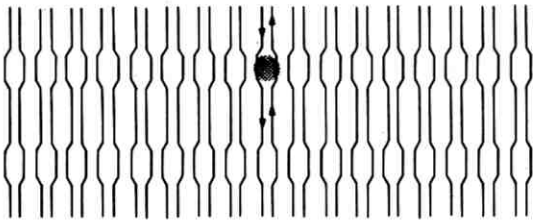


Fig. 2b—Conductor propagation of bubbles.

sequence of bubble positions may be propagated by exciting a series of conductor loops wired to carry current pulses. The second method depends on the alternating magnetization of a patterned soft magnetic overlay embedded on the surface of the platelet (Fig. 2c). The magnetization is imposed by a rotating inplane magnetic field generated by a set of two coils carrying an alternating current and surrounding the platelet with their axes in its plane. This method is called the "field access propagation" and each of the bubbles is propagated to the next pattern in the overlay during one cycle of the exciting current in the surrounding coils.

Field access propagation is more suitable for constructing magnetic domain encoders and decoders, even though it is possible to construct these devices with conductor propagation. Storage, propagation, and the synchronization of incoming data with the outgoing data may all be accomplished by one clock driven at one frequency which is a multiple of the transmission rate. For this reason, only the embodiments of encoders and decoders with field access propagation will be discussed in this paper.

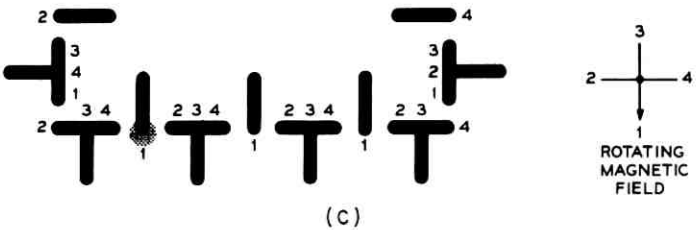


Fig. 2c—Field access propagation of bubbles.

2.3 *Bubble Functions*

2.3.1 *Generation of Bubbles*

Bubbles are generated from an original source bubble. The source bubble rotates around the periphery of a disk of soft magnetic material and when subjected to the localized field of a properly placed current loop it is duplicated (Fig. 3). One section is led away from the generator

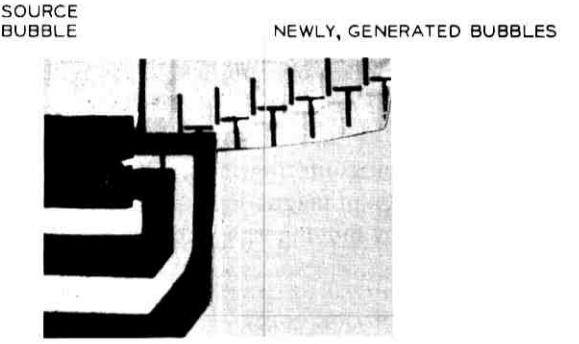


Fig. 3—Generation of bubbles in field access propagation.

and the other section keeps rotating. In most cases the bubbles are generated only when subjected to the field of a current, which is generally controlled by the information bits, or by readings of sensors in a circuit.

2.3.2 *The exclusive-or operation*

This function is accomplished by the mutual repulsion of two bubbles when they are brought in close proximity. In Fig. 4 any one of the two input bubbles A or B finds its way to the output in the absence of a repulsive force due to the other input bubble. Two input bubbles mutually repel themselves into two annihilators. Such an annihilator operates by merging the incoming bubble with a bubble of its own and the diameter of the bubble in the annihilator remains as it was before the merging of the incoming bubbles.

(PART B)

Overview

The basic vehicle chosen for introducing the principles of encoding is the single-error correcting (7,4)* cyclic block Hamming code. The

* The notation is explained in Section 3.1.

principles are then extended to a (31,21) cyclic block code. The code chosen to demonstrate the feasibility of the designs and the embodiments of magnetic domain encoders is (30,20) shortened block code. It is derived from the original (31,21) code. This choice, even though it is inherently a double-error correcting code, facilitates the presentation of serial encoding with field access propagation. The generality of the embodiment for another code is also presented.

In the design of encoders and decoders, time plays a critical part and it becomes necessary to choose a unit of time for any given code. In a (n,k) code, if the incoming information is received at the rate of k bits every P seconds, then the outgoing information is relayed at the rate of n bits every P seconds. If t is defined as $P/(n \times k)$ seconds, then the average interval between the incoming information is $(n \times t)$ and the interval between the outgoing information is $(k \times t)$ seconds. As it will become evident in the design of magnetic domain encoders and decoders, t plays a dominant role in moving the bubbles from one location to the next.

III. ENCODERS FOR CYCLIC BLOCK CODES

(PART B, SECTION 1)

3.1 *Cyclic Block Codes and their Construction**

Block codes constitute a set of codes in which a binary block of k

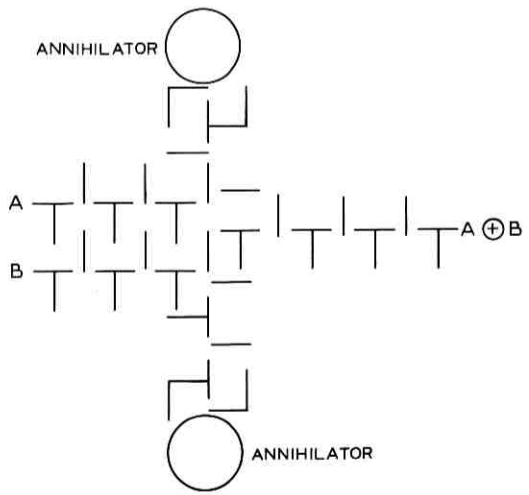


Fig. 4—Exclusive-or operation in field access propagation.

* This topic is discussed in Refs. 12, 13, and 14.

information bits has a binary block of $(n - k)$ parity bits appended to it, thus constituting a (n,k) block code. The n-bit binary cyclic block is represented as a polynomial $c(X)$ as follows.

Let the n-bit binary string be 1010001. The polynomial representation would be

$$c(X) = X^6 + X^4 + X^0 = X^6 + X^4 + 1 \tag{1}$$

corresponding to unity in the first, third, and seventh binary positions. Cyclic block codes have the attractive property that if coefficients of $c(X)$ are cyclically shifted, the new polynomial also represents a code word. For instance, cyclically shifting the coefficients of $c(X)$ once, yields $(X^5 + X + 1)$ which represents another code word.

Consider a new polynomial,

$$g(X) = X^3 + X^2 + 1, \tag{2}$$

which is four binary bits long. If $c(X)$ is divided by $g(X)$ as

$$\begin{array}{r} X^3 + X^2 + 0 + 1 \\ \hline X^3 + X^2 + 0 + 1 \,|\, X^6 + 0 + X^4 + 0 + 0 + 1 \\ \underline{X^6 + X^5 + 0 + X^3} \\ X^5 + X^4 + X^3 + 0 \\ \underline{X^5 + X^4 + 0 + X^2} \\ 0 + X^3 + X^2 + 0 \\ \underline{0 + 0 \;\; + 0 \;\; + 0} \\ X^3 + X^2 + 0 + 1 \\ \underline{X^3 + X^2 + 0 + 1} \\ 0 \;\; + 0 + 0, \end{array}$$

the remainder is three binary zeros. When polynomials obtained by cyclically shifting the coefficients of $c(X)$ once, twice, etc., are divided by $g(X)$, the three-bit remainders obtained are always zero. For each cyclic code there exists such a polynomial $g(X)$ which divides every codeword. This polynomial is called the generator of the code.

Now consider a new polynomial $d(X)$ which corresponds to the first four bit positions of $c(X)$ yielding

$$d(X) = X^6 + X^4. \tag{3}$$

If $d(X)$ is divided by $g(X)$ the remainder corresponds to the polynomial

$$r(X) = 0 \cdot X^2 + 0 \cdot X + 1 = 1, \tag{4}$$

corresponding to the last three bits of the polynomial $c(X)$, since $g(X)$ divides $c(X)$ completely. If the first four bits of $c(X)$ were to denote information bits of a code, then, the last three bits may be thought of as the parity bits, and are in general obtained by dividing a data polynomial $d(X)$ by the generator $g(X)$ and calculating the remainder.

The paper by Bose and Chaudhuri[11] has proved that a large number of codes may be generated by various choices of n and k, provided a generator polynomial $g(X)$ exists for the particular combination of n and k. The value of n is initially limited to $(2^m - 1)$ where m is an integer number. The series of polynomials $g(X)$ for each value of n are readily available in any standard textbook in coding theory (see Ref. 12 or 13). One such value of n is 31 (i.e., $2^5 - 1$) and one of the polynomials for $g(X)$ is

$$g(X) = X^{10} + X^9 + X^8 + X^6 + X^5 + X^3 + 1. \tag{5}$$

The highest degree of the remainder $r(X)$ in the division of a polynomial $d(X)$ by $g(X)$ is always one less than the degree of divisor $g(X)$. In this case, the highest degree of $r(X)$ is 9 and is 10 bits long. Hence the cyclic block code constructed with $n = 31$ and the prechosen value of $g(X)$ has 10 parity bits leading to a (31,21) code. It is however possible to reduce both n and k by a selected number and obtain shortened block codes. For example, if the first bit of an original (31,21) code is eliminated by considering it as being always zero, then a (30,20) code is obtained yielding 10 parity bits for every 20 information bits and the rate corresponding to the ratio of k to n is 2/3.

3.2 *The Function of Encoding for Cyclic Block Codes*

The encoder receives information $d(X)$ in blocks from the data source and yields the code word $c(X)$ in blocks. The two subfunctions are

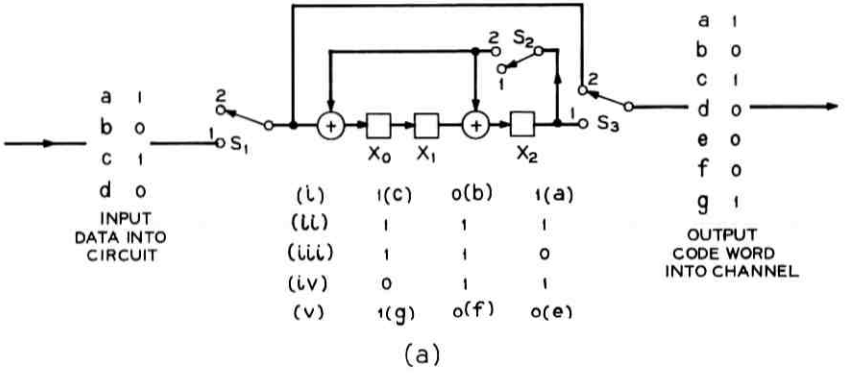
(*i*) Divide the incoming data string $d(X)$ by the generator function $g(X)$, and
(*ii*) Append the remainder after the division to the incoming data string.

These subfunctions are commonly accomplished by semiconductor electronic circuitry in conventional encoders. The division in Section 3.1 has four steps. During the first step of the division cycle the nonzero terms of $g(X)$ are added (by an exclusive-or operation) to the appropriate terms of the data polynomial $d(X)$. At the successive steps of the division cycle, the partial remainder from the earlier step is treated the same way, and the nonzero terms of $g(X)$ are added (by the exclusive-or

function) to the appropriate terms, provided the highest order of the partial remainder is a nonzero quantity. When a zero quotient is encountered in the highest order (such as the third step of the division), then a set of zeros is added (by an exclusive-or function) to the partial remainder.

In the electronic circuitry these functions may be explicitly accomplished. The steps in the well-known (but not frequently used) encoder,* shown in Fig. 5a, are as follows.

Step 1:	Switch	S_1	S_2	S_3
	Position	1	1	1



(a)

Fig. 5a—Encoding for a (7,4) block cyclic code with generator function $g(X) = 1 + X^2 + X^3$.

The first three data positions 101 of $d(X)$ in (3) are shifted into the encoder and transmitted into the channel [line (i) in Fig. 5a].

Step 2:	Switch	S_1	S_2	S_3
	Position	1	2	2

The last data bit, i.e., 0, is shifted into the register and transmitted into the line. The contents of the shift register are shown in line (ii). (Also see the partial remainder after the first step of the division cycle in Section 3.1.)

Step 3:	Switch	S_1	S_2	S_3
	Position	2	2	2

* It will be seen that this type of encoder will present certain operating advantages with magnetic domain configurations in which storage is quite inexpensive as compared to semiconductor configurations.

The contents of the shift register are shifted three times (corresponding to the three remaining steps of the four-step division cycle). The contents of the shift register are shown by (iii), (iv), and (v).

Step 4:	Switch	S_1	S_2	S_3
	Position	2	1	1

The contents of the shift register are emptied into the channel and these correspond to the parity bits $r(X)$ in (4).

A more commonly used configuration of the encoder arrangement is shown in Fig. 5b. Four data bits are shifted with switches S_1, S_2, and S_3 in position 2. The switches are moved down to position 1 and the contents of the shift register are emptied into the channel. The lines a, b, c, and d in Fig. 5b indicate the contents of the register as the data bits corresponding to $d(X)$ in (3) are received.

It is to be observed here that the arrangement in Fig. 5b necessitates that the two exclusive-or functions be done serially between the arrival of data bits, whereas the arrangement in Fig. 5a requires that the two exclusive-or functions be accomplished simultaneously. In the magnetic domain technology this consideration makes the configuration of Fig. 5a more favorable for the implementation.

Encoders for various codes are similarly constructed. The location of the exclusive-or gates is determined by the nonzero terms in the generator polynomial $g(X)$ exclusive of the highest-degree term. Figures 5c and 5d indicate the conventional encoder arrangements for the (30,20) shortened block code discussed earlier with $g(X)$ in (5). The complete encoder also adjusts for the difference of rate between the

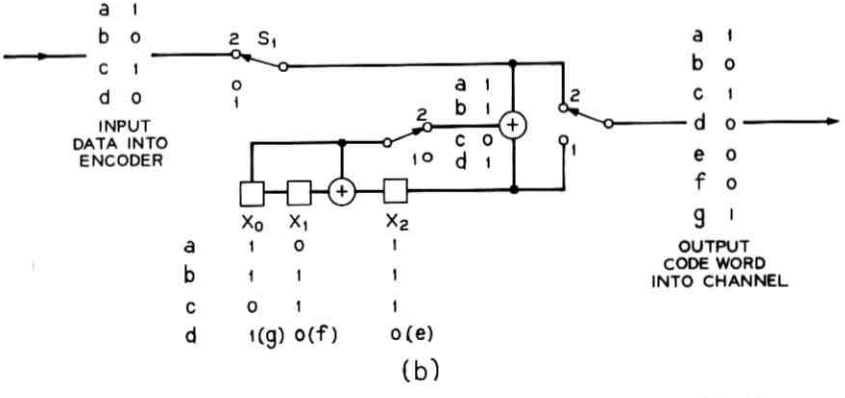


Fig. 5b—Conventional encoder for the (7,4) block cyclic code with the same generator used in Fig. 5a.

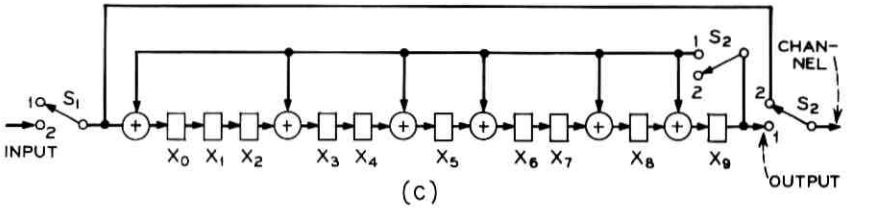


Fig. 5c—An encoding arrangement for (30,20) cyclic block code with generator function

$$g(X) = 1 + X^3 + X^5 + X^6 + X^8 + X^9 + X^{10}.$$

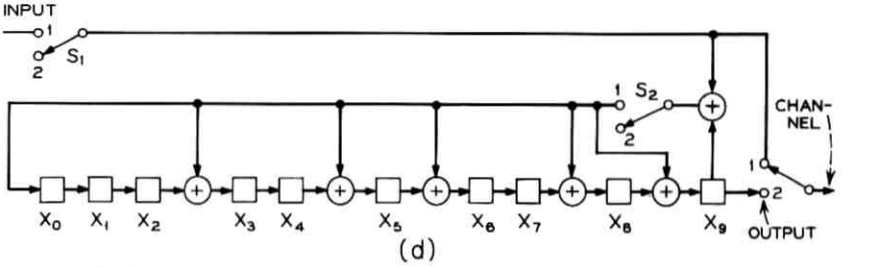


Fig. 5d—Conventional encoder configuration for the (30,20) code with same generator used in Fig. 5c.

arrival rate at the input of the encoder and its output into the channel. In this case the input rate in the encoder is two-thirds the output rate.

The generator $g(X)$ in (5) has seven terms. During each step of the division cycle the highest order term is eliminated from the partial remainder in the shift register. This leads to an unconditional zero coefficient for the term to which X^{10} is added. This fact may be used to limit the number of terms in $g(X)$ to six terms (excluding the highest-order term), provided the highest-order term is eliminated from the partial remainder after sensing its value.* Under such conditions the six remaining terms may be written as

$$g'(X) = 1 + X^3 + X^5 + X^6 + X^8 + X^9. \tag{6}$$

(PART B, SECTION 2)

3.2.1 *Serial Arrangement of Encoders for Cyclic Block Codes*

In the conventional configurations (Figs. 5c and 5d), the output of the rightmost stage feeds back into six different exclusive-or gates

* The presence of one in the highest-order position requires that the other six terms be added (by exclusive-or function) to the corresponding terms in the partial remainder.

corresponding to the nonzero terms of $g(X)$. Alternatively, the information may be fed back at one location with one exclusive-or gate but at six different instants of time. Each step of the 20*-step division cycle is effectively performed by circulating the partial remainder through this gate. The input to the gate is dictated by the nonzero terms of $g'(X)$. A configuration incorporating such a serial feedback arrangement is shown in Fig. 6a. The switch S_a is designed to respond to the contents of x_{10}, closing only if the content is one. The contents of the shift register $g'(X)$ are initialized to 1101101001 corresponding to the X^9, X^8, X^6, X^5, X^3, 1 terms of the function $g'(X)$ in (6). The contents of the shift register $g'(X)$ are circulated in synchronism with the main shift register SR. The circulation time of both registers is the time between the arrival of bits in the incoming data stream.

The operation of this type of encoder after emptying its contents is as follows:

Step 1:	Switch	S_1	S_2	S_2'	S_3
	Position	1	2	2	1

The first 10 bits of a data block are shifted into the main shift register SR.

Step 2:	Switch	S_1	S_2	S_2'	S_3
	Position	1	1	2	1

The shift register is shifted once more so that the highest-order bit is in x_{10} and the 11th bit of the data block enters position x_0 simultaneously.

Step 3:	Switch	S_1	S_2	S_2'	S_3
	Position	1	2	1	1

The shift register is completely circulated once.

Step 4:	Switch	S_1	S_2	S_2'	S_3
	Position	1	1	2	1

The highest-order bit is entered in x_{10} as in step 2 and the 12th bit of data enters position x_0. The process in steps 3 and 4 is repeated 10 times (i.e., 8 more times). After the 20th data bit enters the shift register, the switch S_1 is moved to position 2 and the shift register is circulated 10 more times as in steps 2 and 3. The division is now complete.

Step 5:	Switch	S_1	S_2	S_2'	S_3
	Position	2	2	2	2

* i.e., 30–10 or k, the number of information bits in the block.

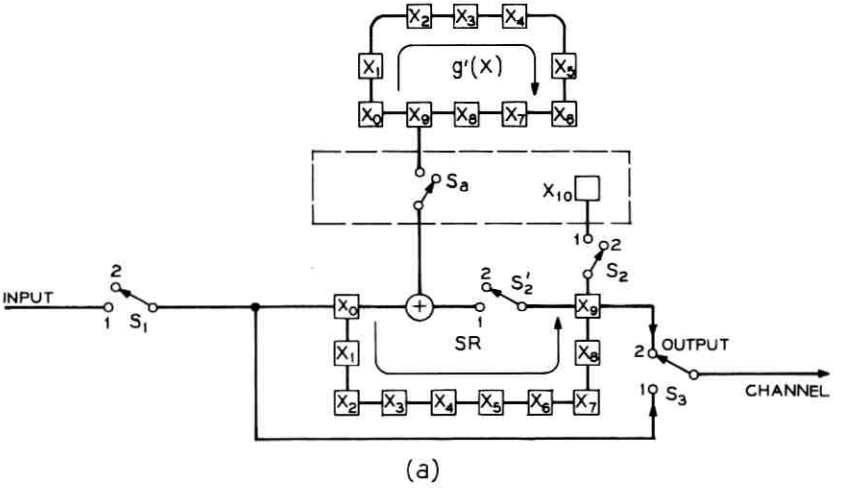


Fig. 6a—Serial encoding for the (30,20) code.

The parity bits are emptied into the channel. The process is repeated for the next data block by returning to step 1.

3.2.2 *Complete Encoder with Serial Arrangement for Cyclic Block Code*

Figure 6b shows a complete encoder. The incoming data ($k = 20$ information bits) arrives uniformly at the encoder and coded information ($n = 30$ bits consisting of 20 information and 10 parity bits) is recovered uniformly. The operation of the switch S_a is explained earlier. Only one of three poles of switch S_b is closed at any given instant of time. Coded information is emptied out of d′, d″, or d_p one bit every $20t$ seconds. The data-stores d′ and d″ store the first and second 10 bits, and d_p stores the parity bit. The data-store d_i holds the first 10 bits of any block on an interim basis while the main shift register SR is calculating the parity bits of the previous data block.

When the register is full, the contents of d_i are moved into both the main shift register SR and d′ within the $30t$ seconds preceding the arrival of the next data bit. The shift is synchronized with moving the parity bits from SR to d_p with S_5 in position 2. The arrival of the 11th bit is synchronized with the movement of the first bit into x_{10}, thus emptying the location x_0 in SR for this 11th bit. The second 10 bits arrive at location x_0 of SR via switch S_1 and are also entered in d″. The circulation SR and division with $g(X)$ continues 20 times. The data-store d_i would then have emptied the first 10 bits of the next data

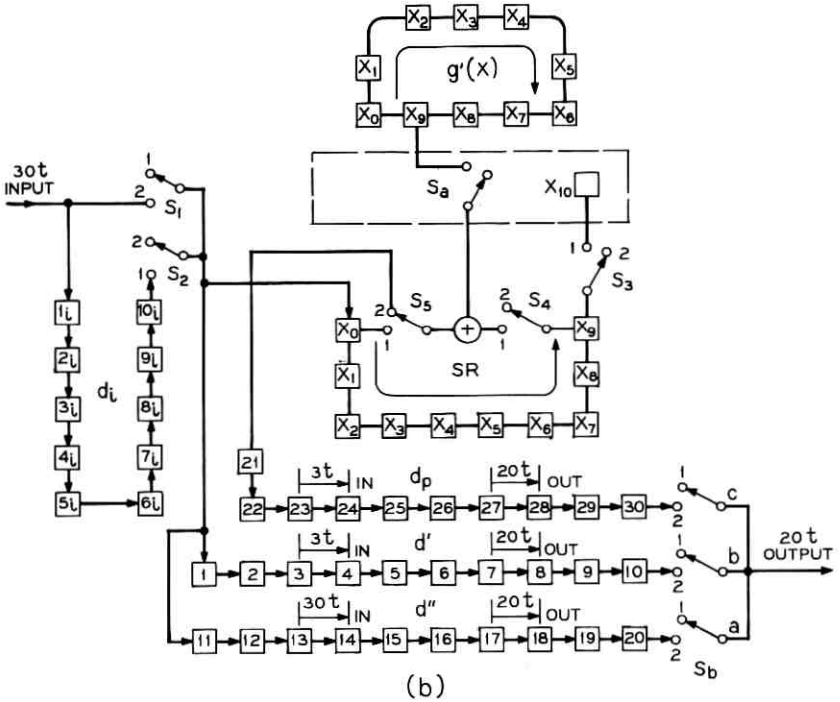


Fig. 6b—A complete serial encoder for the (30,20) code.

block into SR and d_p would have the parity bits for the data just processed. The cycle can be repeated indefinitely provided the stored d', d'', and d_p have been emptied into the transmission channel at appropriate times. The operation of switches a, b, and c meet this requirement. The timing diagram of the encoder is shown in Fig. 6c. The incoming data is shown in line (i) and the outgoing information is shown on line (iii). The data-stores d' and d_p shift in during a $30t$-second interval and shift out into the channel one bit every $20t$ seconds. The data-store d'' shifts in one bit every $30t$ seconds and uniformly shifts out one bit every $20t$ seconds.

3.3 *Magnetic Domain Encoders for Cyclic Block Codes with Field Access Propagation*

In field access propagation all the bubbles in the region are propagated by one pitch (or period) during one cycle time of the rotating magnetic field. It is advantageous to equate the cycle time of this magnetic field with t defined earlier.

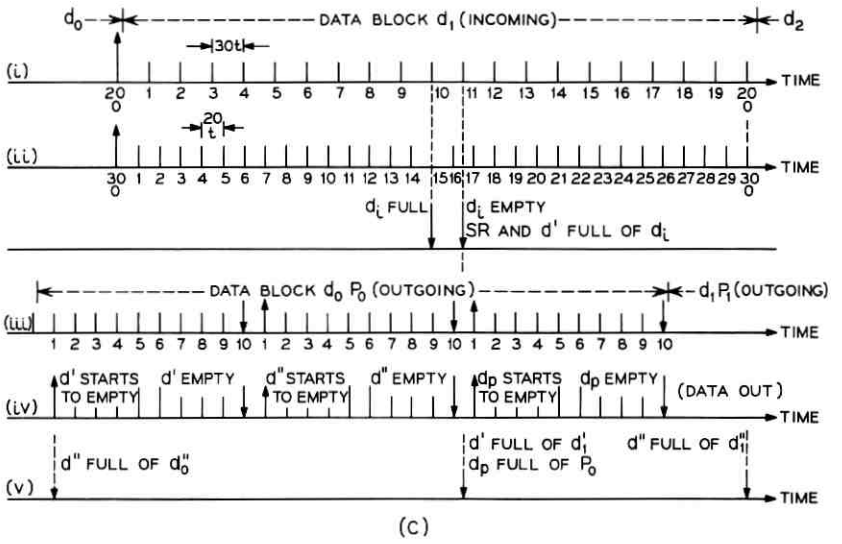


Fig. 6c—Timing diagram for encoder shown in Fig. 6b.

The encoder arrangement is shown in Fig. 7. The incoming data pulses generate bubbles at the information bubble generator. These are accumulated in loop 1 at consecutive periods since there are 29 periods and the incoming data arrives every $30t$ seconds. The channeling gate g_1 opens every 20 circulations* to permit a sequence of 20 bubble positions to enter the duplicator D. The data is circulated in loop 3. Loop 2 performs one step of the 20-step division cycle every circulation. The sensor S_g reads the leading bubble position every $30t$ seconds and controls the generator G_g to inject a series of bubbles corresponding to the nonzero terms of the generator $g(X)$ in the exclusive-or gate. A string of bubbles 11101101001, corresponding to X^{10}, X^9, X^8, X^6, X^5, X^3, 1 terms in $g(X)$, is generated if S_g has sensed a bubble. The distances between the sensor S_g, the exclusive-or gate, and generator G_g are adjusted so that the bubble corresponding to the X^{10} position of $g(X)$ arrives into the exclusive-or gate in synchronism with the leading bubble position that S_g sensed. After 20 circulations, the 10 parity bits are left in loop 2. The parity and data bits are channeled into loops 4 and 5 by the action of the channeling gates (Ref. 9) g_2 and g_3 respectively.

The code word (data and parity) is retrieved and transmitted in two sections. The data is read by the sensor S_d in loop 5 every $20t$ seconds. The parity is read by the sensor S_p in loop 4 every $20t$ seconds. The

* A circulation corresponds to the contents of the loop going around once.

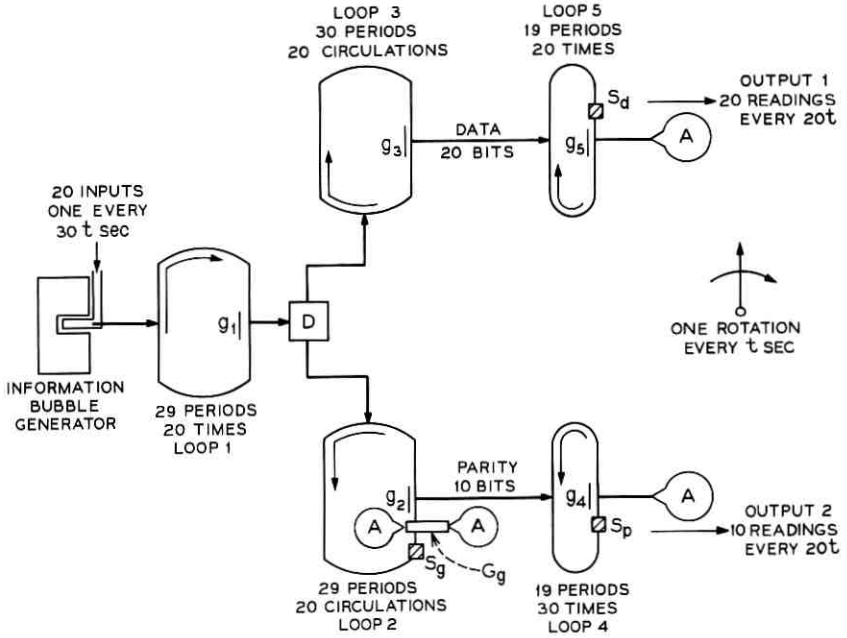


Fig. 7—Serial encoder with domains under field access propagation for the (30,20) code with $t = P/30 \times 20$.

diverting gates g_4 and g_5 function identically. Every time the sensor S_d or S_g is read, the diverting gate g_5 or g_4 diverts the bubble position read into annihilator A.

The generality of this embodiment is exemplified by another serial encoder shown in Fig. 8 for (31,26) cyclic block code. The generator function for this code is

$$g(X) = 1 + X^2 + X^5. \tag{7}$$

This encoder operates along the same principles described earlier. Such encoders cannot be constructed when the loops 1 through 5 become extremely small, and thus codes with very short block lengths cannot be easily implemented. Generally block codes with block lengths of 30 or over are well suited for such embodiment.

(PART C)

Overview

The basic vehicle chosen for introducing the principles of decoding is the single-error correcting (7,4) cyclic block Hamming code discussed earlier. The general concepts of decoding and single-error correcting

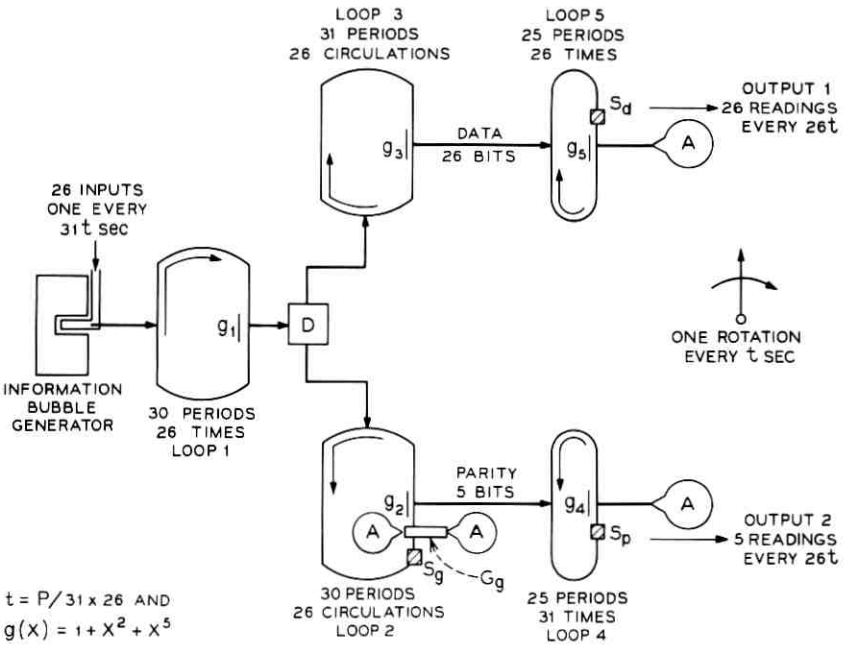


Fig. 8—Serial encoder for the (31,26) code with domains under field access propagation with $t = P/31 \times 26$.

are then extended to cyclic block codes. The particular code chosen to demonstrate the embodiment of single-error correcting decoders for cyclic block codes is the (30,20) shortened cyclic code.

For a (n,k) code, the decoders receive information from the channel at an interval of one bit every $(k \times t)$ seconds and recover the original information at an interval of one bit every $(n \times t)$ seconds. Further, the decoders discussed here detect and correct single errors in the received information. Multiple-error correcting decoders are not discussed in this paper.

IV. SINGLE-ERROR CORRECTING DECODERS FOR CYCLIC BLOCK CODES

(PART C, SECTION 1)

4.1 *Decoding of Cyclic Block Codes**

Coded information in the form of $c(X)$ in (1) is received from the channel into the decoder. The decoder recovers the original information

* This topic is discussed in Refs. 16 and 17.

bits [polynomial $d(X)$] from $c(X)$ even if one of the bits of $c(X)$ was received in error at the decoder. The function of detecting and correcting single errors for the (7,4) Hamming code is explained as follows. Let $S_1(X)$, $S_2(X)$, $\cdots$, $S_7(X)$ be the remainders obtained by dividing X, X^2, $\cdots$ X^7, by $g(X)$ in (2). These polynomials may be calculated as:

$$s_1(X) = X; \qquad s_2(X) = X^2 \tag{8a; b}$$

$$s_3(X) = X^2 + 1; \qquad s_4(X) = X^2 + X + 1 \tag{8c; d}$$

$$s_5(X) = X + 1; \qquad s_6(X) = X^2 + X, \tag{8e; f}$$

and finally

$$s_7(X) = \Gamma(X) = 0 \cdot X^2 + 0 \cdot X + 1 = 1. \tag{8g}$$

Now if the received word has a single error in the second location,* then the received word $R(X)$ will differ from $c(X)$ as follows:

$$R(X) = c(X) + X^5 = X^6 + X^5 + X^4 + 1 \tag{9}$$

and the remainder [also known as the syndrome $s(X)$] obtained by dividing $R(X)$ by $g(X)$ is

$$s(X) = X + 1. \tag{10}$$

This polynomial is seen to be $s_5(X)$ from (8e) indicating that a single error in the ith location yields a syndrome corresponding to $s_{7-i}(X)$. Next consider the polynomial obtained by shifting $s(X)$ two (i.e., $7 - 5$) times,

$$X^2 \cdot s(X) = X^3 + X^2; \tag{11}$$

and the remainder, denoted by $\rho(X)$, obtained by dividing the shifted polynomial by $g(X)$ is

$$\rho(X) = 0 \cdot X^2 + 0 \cdot X + 1 = 1. \tag{12}$$

This value of $\rho(X)$ corresponds to $s_7(X)$ or $\Gamma(X)$ in (8g), since

$$X^i \cdot s(X) = X^i \cdot s_{7-i}(X), \tag{13}$$

and the remainder obtained by dividing the right side of (13) by $g(X)$ does in fact represent the remainder obtained by dividing $(X^i \cdot X^{7-i})$ or X^7 by $g(X)$ and is indeed $\Gamma(X)$. This leads to the conclusion that if the remainder obtained by dividing $s(X)$ shifted i times by $g(X)$ corresponds to $\Gamma(X)$ then the ith bit is in error. Correction is accomplished

* It should be noted that the error in the ith bit corresponds to adding the X^{7-i} term to $c(X)$.

by complementing this bit. In this case, the corrected data corresponding to the first four bits of $R(X)$ is 1010 originally represented as $d(X)$ in (3).

This reasoning may be extended to general cyclic codes and in particular to the (30,20) code. For this code, $\Gamma(X)$ is the remainder obtained by dividing X^{30} by $g(X)$ in (5) and it can be calculated as

$$\Gamma(X) = X^4 + X^5 + X^6 + X^7. \tag{14}$$

In semiconductor circuitry the division by $g(X)$ is accomplished by the top section of the decoder shown in Fig. 9 and the comparison of the contents of the register with $\Gamma(X)$ is accomplished by the AND gate. In the complete shift register two shift registers are used with one performing the comparison while the other is calculating the syndrome of the next data block.

(PART C, SECTION 2)

4.2 *Serial Decoding of Block Cyclic Codes*

In serial decoding the division is carried out by one exclusive-or gate as in serial encoding discussed in Section 3.2.1. Further, the comparison of the content of the shift register is also done serially bit by bit in contrast to the simultaneous evaluation and comparison of all the bits by the AND gate used in conventional decoders (Fig. 9).

An exclusive-or gate is used for serial comparison instead of the AND gate. In the (30,20) code the comparison cycle lasts for 10 bits

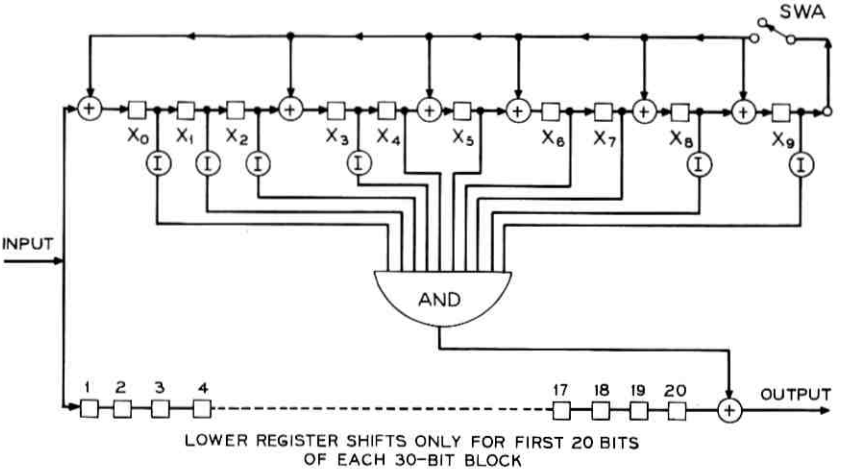


Fig. 9—Conventional single-error correcting decoder with (30,20) code.

(i.e., the number of bits in the remainder). Bits duplicated from the shift register are serially fed into an exclusive-or gate together with bits corresponding to $\Gamma(X)$. A perfect match between the two inputs yields a zero output from the exclusive-or gate for the entire interval of comparison. One or more outputs from the exclusive-or gate during the interval indicates a mismatch. This principle is used in the magnetic domain decoders discussed next.

4.3 *Single-Error Correcting Magnetic Domain Decoder for Cyclic Block Codes with Field Access Propagation*

Figure 10 shows a decoder for the (30,20) code. The operation of the decoder closely resembles the operation of the encoder shown in Fig. 7. The incoming data generates a series of bubbles at G_i . This data is led

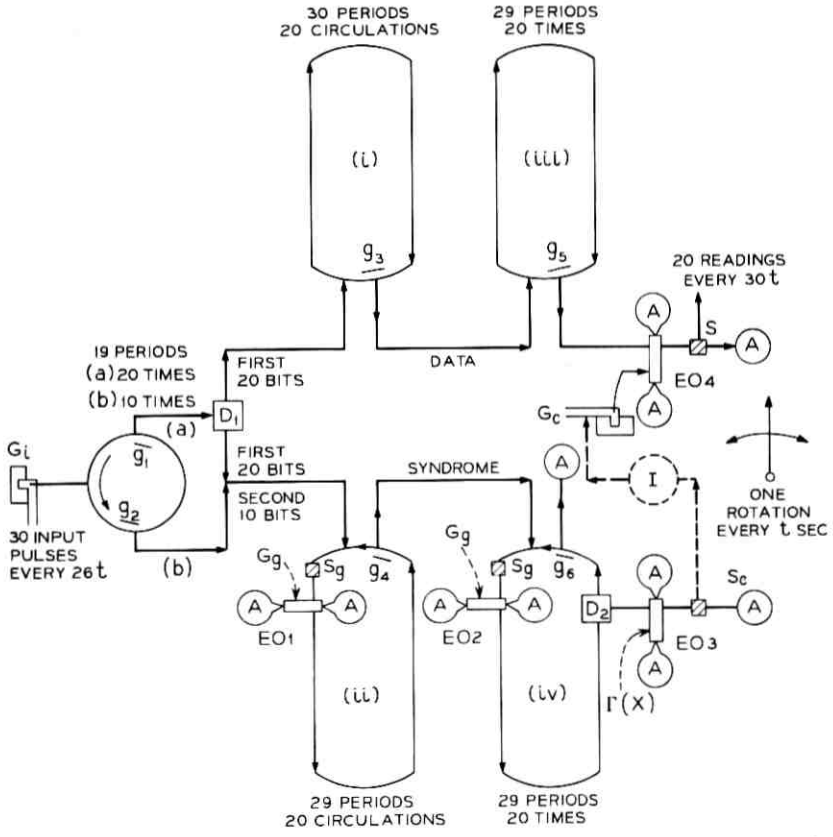


Fig. 10—Serial arrangement of single-error correcting decoder with magnetic domains under field access propagation for the (30,20) code with $t = P/30 \times 20$.

into a loop with 19 periods and two exit gates g_1 and g_2. A string of bubbles are formed in adjoining periods in the loop since the main magnetic field rotates every t seconds and the bubbles arrive every $20t$ seconds. When the first 19 bubble positions have arrived in the loop, the gate g_1 empties the bubble stream into channel (a). This stream is duplicated at D_1 and it enters loops (i) and (ii). This data is allowed to circulate in (i) while the division in (ii) progresses. The sensor S_g in (ii) senses a bubble every $30t$ seconds and directs G_g to generate a stream of bubbles into the exclusive-or gate E01 only if a bubble is sensed at the leading end of the data stream. The generator bubble stream is 11101101001 and is consistent with the nonzero coefficients of $g(X)$. A leading bubble corresponding to the 10th power of X is necessary since there is no special arrangement to eliminate the X^{10} bubble as in the conductor pattern propagation (see Fig. 6b). After the parity bits are accumulated in the first loop, they are channeled into (ii) via gate g_2. To ascertain that the last 10 bits arrive at the correct location in (ii), it is necessary to adjust the number of periods between the exit points g_1 and g_2 in the first loop.

When the division is complete [i.e., 20 circulations of (ii), each circulation accomplishing one step of the 20-step division cycle] the data and syndrome may be transferred out of (i) and (ii) by gates g_3 and g_4 respectively. Such gates have been designed and implemented for other applications (Ref. 9 and Ref. 18). In loop (iv) the syndrome is again divided by the generator function $g(X)$. The exclusive-or gate E02 performs this function. The remainder after this division is duplicated at D_2. One section circulates in (iv) and the other is compared with the remainder function $\Gamma(X)$ by the exclusive-or gate E03. A perfect match results in a zero reading of sensor S_c during the entire comparison time which lasts for $10t$ seconds. Meanwhile, the data in (iii) is also being circulated. The gate g_5 permits one leading bubble to be channelized out of the loop once every circulation into the exclusive-or gate E04. This gate receives a complementing bubble only if S_c has not sensed a bubble after comparing the contents of (iv) with the remainder function $\Gamma(X)$.

The generality of the embodiment is exemplified by another serial decoder shown in Fig. 11 for (31,26) cyclic block codes. The encoder for this code is shown in Fig. 8 and the generator function is given by equation (7). This decoder operates on a principle of serial decoding discussed earlier and the value of $\Gamma(X)$, the remainder obtained by dividing X^{31} by $g(X)$ in (7), is

$$\Gamma(X) = 0 \cdot X^5 + 0 \cdot X^4 + 0 \cdot X^3 + 0 \cdot X^2 + 0 \cdot X + 1 = 1. \quad (15)$$

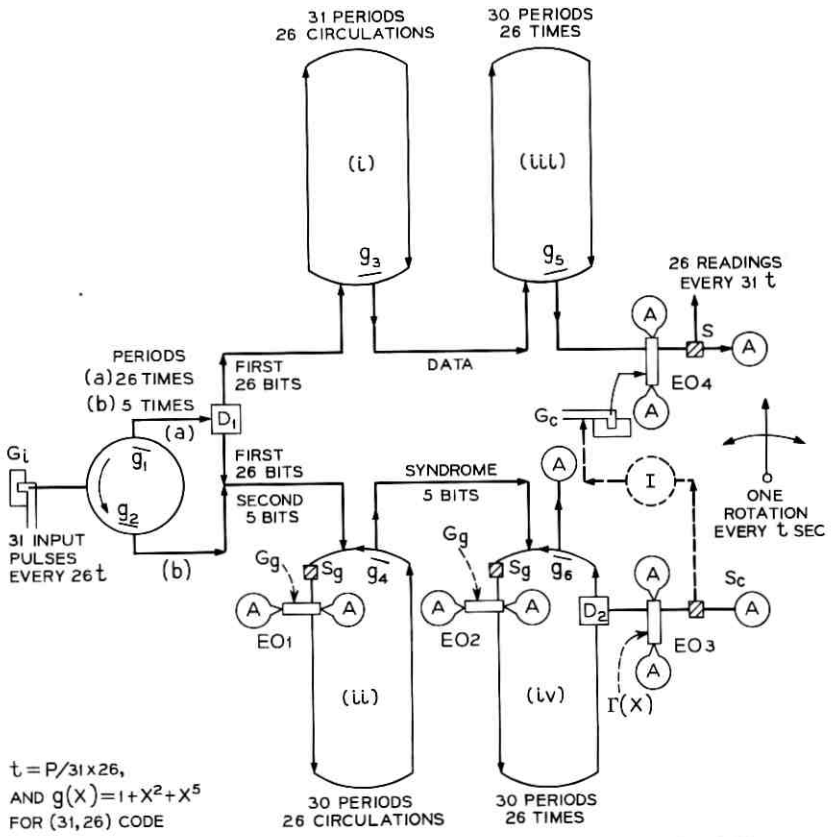


Fig. 11—Serial decoder for the (31,26) code with domains under field access propagation with $t = P/31 \times 26$.

(PART D)

V. DISCUSSION OF MAGNETIC DOMAIN ENCODERS AND DECODERS

Magnetic domain technology offers inexpensive storage but requires time for every operation (generation, propagation, exclusive-or operation, sensing, annihilation, etc.,) in contrast to the instantaneous operation of semiconductor circuitry. Such a constraint induces the variations in the design of encoders and decoders from conventional devices in semiconductor technology.

The basic functions utilized in the domain devices in this paper are: controlled generation of information, propagation, storage, duplication, gating of bubble streams, exclusive-or operation, sensing, and annihilation. Most of these functions have been successfully accomplished

within the Bell System. Some of the functions are discussed in Part A and the others are reported in Refs. 8, 9, and 18. Serial arrangements of these encoders and decoders utilize fewer exclusive-or operations than the nonserial arrangements.

The packing density (which ultimately influences the active chip area in the devices) depends upon the nature of the uniaxial magnetic material chosen. Typical orthoferrites[19] ($YbFeO_3$, $YFeO_3$, and $Sm_{0.55}Tb_{0.45}FEO_3$) can hold bubbles of 40 to 50 micron diameters at 200 micron spacing yielding about 1.6×10^5 bits per square inch. Typical garnets ($Er_2Tb_1Al_{1.1}Fe_{3.9}O_{12}$ and $Gd_{2.3}Tb_{0.7}Fe_5O_{12}$) can support bubbles of 4 to 8 micron diameters at 25 micron spacing yielding about 10^6 bits per square inch. Newer epitaxial garnet films have yielded up to 1.6×10^6 bits per square inch of storage.[20] The hexagonal ferrites ($PbAl_4Fe_8O_{19}$) support bubbles of 4 to 8 microns in diameter.

The domain velocity (which ultimately influences the speed of devices) depends on the field difference across the bubble diameter and the magnetic material used. A nominal value of 20 Oe can be generated in field access propagation with a T-bar type of overlay. Orthoferrites require the lowest time to move a bubble from one position to the next data position approximately four diameters away, thus yielding a data rate of about one megacycle of 20 Oe field difference. The highest rate achieved is about three megacycles. Some of the earlier garnets have lower mobilities and a data rate of 140 kHz has been achieved with field access propagation. Some of the newer garnet films have yielded data rates of up to one megacycle.[20] Hexagonal orthoferrites have the lowest mobilities and are suitable for 10 to 60 kHz application. The data rates thus far attained in orthoferrites and garnets are sufficient to construct encoders and decoders at normal data transmission rates. For instance a transmission rate of 4800 bits per second would demand a data rate of about 125 kHz.

One of the differences between the conventional semiconductor devices and the serial type of bubble devices is the ease of converting one generator polynomial to another generator of the same degree without changing the control or propagating circuitry. If it is desired to change the generator, then it is necessary only to change the sequence of bubbles injected by G_g in Fig. 7 for the encoder and Fig. 11 for the decoder together with the generator $\Gamma(X)$, without altering the rest of the circuitry. Further, the embodiments presented indicate that the serial encoders and decoders with field access propagation yield flexible designs for block codes whose block length is about thirty bits or more.

VI. CONCLUSIONS

Magnetic domains may be used to construct encoders and single-error correcting decoders for cyclic block codes. The magnetic material chosen to host the bubbles depends on the transmission rate, and the generator of the code may be changed from one polynomial to another of the same order without altering the embodiment or the control circuitry in the serial type of devices.

In the field access propagation only one clock frequency is utilized to accomplish storage, division, and synchronizing the input and the output. The same clock excites the main propagating magnetic field once during an interval calculated as $(P/n \times k)$ seconds, where P is the time required to transmit one block of data through the transmission channel, n is the total number of bits in the block, and k is the number of information bits in the block.

VII. ACKNOWLEDGMENTS

The author thanks D. D. Sullivan, A. H. Bobeck, P. I. Bonyhard, A. J. Perneski, and E. R. Berlekamp at Bell Laboratories for the discussions during the development of the possible configurations of the magnetic domain encoders and decoders.

REFERENCES

1. Bobeck, A. H., "Properties and Device Applications of Magnetic Domains in Orthoferrites," B.S.T.J., *46*, No. 8 (October 1967), pp. 1901–1925.
2. Bobeck, A. H., "Properties of Cylindrical Magnetic Domains in Orthoferrites," IEEE Trans. Magnetics, *4*, (1968), pp. 450 ff.
3. Bobeck, A. H., and Gianola, U. F., "Magnetic Domains," Science and Technology, No. 86 (1969).
4. Perneski, A. J., "Propagation of Cylindrical Magnetic Domains in Orthoferrites," IEEE Trans. Magnetics, *5*, No. 3 (September 1969), pp. 554–557.
5. Thiele, A. A., "The Theory of Circular Magnetic Domains," B.S.T.J., *48*, No. 10 (December 1969), pp. 3287–3335.
6. Thiele, A. A., "Theory of Static Stability of Cylindrical Domains in Uniaxial Platelets," J. Appl. Phys., *41*, No. 3 (March 1970), pp. 1139–1145.
7. Bobeck, A. H., Fischer, R. F., Perneski, A. J., Remeika, J. P., and Van Uitert, L. G., "Application of Orthoferrites to Domain Wall Devices," IEEE Trans. Magnetics, *5*, No. 3 (September 1969), pp. 544–553.
8. Bobeck, A. H., Fischer, R. F., and Perneski, A. J., "A New Approach to Memory and Logic Cylindrical Domain Devices," Proc. FJCC, (1969), pp. 489–498.
9. Bonyhard, P. I., Danylchuk, I., Kish, D. E., and Smith, J. L., "Application of Bubble Devices," IEEE Trans. Magnetics, *6*, No. 3 (September 1970), pp. 447–451.
10. Hamming, R. W., "Error Detecting and Error Correcting Codes," B.S.T.J., *29*, No. 1 (January 1950), pp. 147–160.
11. Bose, R. C., and Ray-Chaudhuri, D. K., "On Class of Error Correcting Binary Group Codes," Information Control, *3*, pp. 68–79.
12. Peterson, W. W., *Error-Correcting Codes*, Cambridge, Massachusetts: MIT Press, 1961.

13. Lucky, R. W., Salz, J., and Weldon, E. J., *Principles of Data Communication*, New York: McGraw-Hill Book Company, 1968.
14. Berlekamp, E. R., *Algebraic Coding Theory*, New York: McGraw-Hill Book Company, 1968.
15. Hocquenghem, A., "Codes Correcteurs d'erreurs," Chieffres, *2*, (1959), pp. 147–156.
16. Peterson, W. W., *loc. cit.*, pp. 201–202.
17. Berlekamp, E. R., *loc. cit.*, p. 123.
18. Bonyhard, P. I., private communication.
19. Bobeck, A. H., Danylchuk, I., Remeika, J. P., Van Uitert, L. G., and Walters, E. M., "Dynamic Properties of Bubble Domains," presented at the International Conference on Ferrites, July 6–9, 1970, Kyoto, Japan.
20. Fischer, R. F., private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Delay Distortion in Weakly Guiding Optical Fibers Due to Elliptic Deformation of the Boundary

By W. O. SCHLOSSER

(Manuscript received September 13, 1971)

The delay distortion in glass fiber optical waveguides due to small elliptical deformations of the cross section is calculated. Simple approximations are given for the case of small differences in index of refraction between core and cladding (weak guidance). Since the delay distortion is quadratically dependent on the index difference it is found that it is generally possible to keep it negligible by judiciously choosing the guide parameters.

I. INTRODUCTION

Recent results[1] indicate that glass fibers are a potential transmission medium for optical communication. For high-capacity systems (in excess of 100 MBaud) dispersion and the associated delay distortion is an important factor to be considered. For such an application the fiber cannot be used in a frequency range where more than one mode propagates, since the difference in group velocity between the various modes causes excessive delay distortion.[2] In the single-mode range the two best known sources of delay distortion are the material and waveguide dispersion.[2] In this paper we will treat a different source of dispersion. If the fiber is elliptically deformed two different polarizations with different group velocities are possible. We will calculate the delay distortion due to these elliptical deformations and establish simple relationships between the allowable delay distortion and the tolerances with which the fiber has to be manufactured.

II. DEFORMATION OF A SQUARE DIELECTRIC WAVEGUIDE

Let us consider the dielectric waveguide of square cross section first, since its properties are very similar to the round waveguide[3] and the mode structure is quite simple. This will allow us to explain the effects

of deformation more easily than for the round waveguide, where the necessary formalism clouds the physics somewhat. For a rectangular dielectric waveguide the HE_{11} dominant mode can be thought of as being composed of the E mode on a slab of height $2a$ and the H mode on a slab of height $2b$ (Fig. 1). The propagation constant is approximated by

$$\beta^2 = n_c^2 k_0^2 - (\beta_H^2 + \beta_E^2) \tag{1}$$

where β_H and β_E can be determined from the characteristic equations for the slab modes:

$$[(n_c^2 - n^2)k_0^2 - \beta_H^2]^{1/2} = \beta_H \tan \beta_H a \qquad \text{H mode,} \tag{2a}$$

$$\frac{n_c^2}{n^2}[(n_c^2 - n^2)k_0^2 - \beta_E^2]^{1/2} = \beta_E \tan \beta_E b \qquad \text{E mode.} \tag{2b}$$

We postulate that the deformation of a quadratic cross section into a rectangular one corresponds to the elliptic deformation of a circular cross section. The height $2a$ of the square is increased by $2\Delta a$ and the width is decreased by $2\Delta a$. The change in propagation constant β due to this deformation can be calculated from

$$\Delta\beta = \left(\frac{\partial\beta}{\partial a} - \frac{\partial\beta}{\partial b}\right)\Bigg|_{a=b=a_0} \cdot \Delta a = \frac{1}{\beta}\left(\beta_E \frac{\partial\beta_E}{\partial b} - \beta_H \frac{\partial\beta_H}{\partial a}\right)\Bigg|_{a=b=a_0} \cdot \Delta a. \tag{3}$$

This shows that the effects from the widening of one dimension and narrowing of the other tend to cancel each other. From the well known properties of β we can make predictions about $\Delta\beta$. For small differences of refractive indices $n_c \approx n$ and β is given by $\beta = n \cdot k_0 + \Delta \cdot f(a, b, k_0)$ where $\Delta = (n_c - n)/n$. The derivative of β with respect to the dimensions a, b will therefore have a factor Δ and $\Delta\beta$ must hence possess this factor. We observe furthermore from the equations (2a) and (2b)

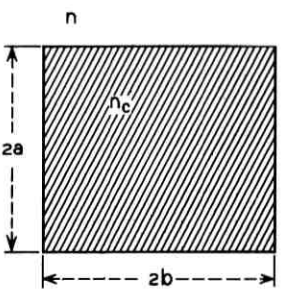


Fig. 1—Cross section of rectangular waveguide.

that β_E and β_H obey the same characteristic equation except for a factor $n_c^2/n^2 = 1 + 2\Delta$. The difference,

$$\beta_H \frac{\partial \beta_H}{\partial a} - \beta_E \frac{\partial \beta_E}{\partial b}\bigg|_{a=b=a_0},$$

will therefore be proportional to Δ^2. We can thus expect $\Delta\beta$ to be quadratic in Δ and the difference in group delay between the two polarizations will have the form

$$\Delta\tau = \Delta^2 \cdot \frac{\Delta a}{a} f(a, k_0, \Delta). \tag{4}$$

Recently, a letter by R. B. Dyott and J. R. Stern[4] has been published, approximating the phase difference between the two polarizations by the phase difference of two modes on circular guides of different diameters. This results in an approximation which overestimates the delay distortion considerably. We can see that easily by calculating the variation in propagation constant due to an increase of height and width by $2\Delta a$,

$$\Delta\beta = \frac{\partial \beta}{\partial a}\Delta a = -\left(\frac{1}{\beta_H}\frac{\partial \beta_H}{\partial a} + \frac{1}{\beta_E}\frac{\partial \beta_E}{\partial b}\right)\bigg|_{a=b=a_0} \cdot \Delta a.$$

Since the two derivatives are added, $\Delta\beta$ will only be of order Δ and not Δ^2 as in equation (4). In practical cases Δ is in the order of 10^{-2} and therefore this approximation gives a considerably bigger value than (4).

III. DEFORMATION OF A CIRCULAR DIELECTRIC WAVEGUIDE

The function $f(a, k, \Delta)$ in equation (4) has to be determined for the circular fiber. This will now be done by first determining $\Delta\beta$ and then differentiating with respect to ω. We make use of the fact that the cross section of the guide is only slightly elliptic (Fig. 2). In this case the propagation constant β is approximated by the first two terms of a Taylor series

$$\beta = \beta_0 + \left(\frac{e}{d}\right)^2 \partial\beta/\partial\left(\frac{e}{d}\right)^2\bigg|_{e=0} \tag{5}$$

where β_0 is the propagation constant on the circular guide and e the small excentricity. The propagation constant β is generally determined from the zeros of the characteristic equation:

$$F(d, e^2, \beta, k_0, \Delta) = 0. \tag{6}$$

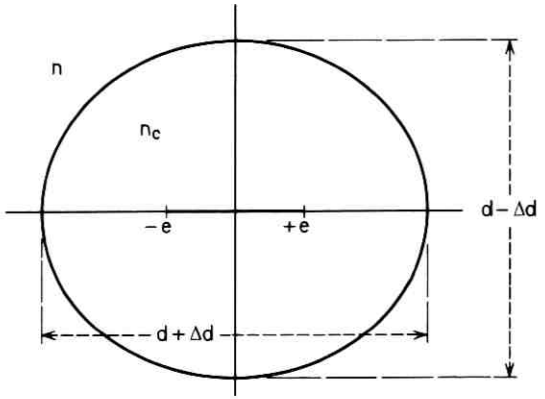


Fig. 2—Cross section of elliptically deformed fiber.

The derivative $\partial\beta/\partial(\frac{e}{d})^2$ can thus be expressed by the derivatives of F,

$$\partial\beta/\partial\left(\frac{e}{d}\right)^2\Big|_{e=0} = \frac{\partial F/\partial\left(\frac{e}{d}\right)^2}{\partial F/\partial\beta}\Big|_{e=0}. \tag{7}$$

The problem is now reduced to finding F. This has been done for small ellipticities in Ref. 5. The reduction of the results of Ref. 5 to the case of small difference in refractive indices finally yields $\Delta\beta$:

$$\Delta\beta = \Delta^2 2\,\frac{\Delta d}{d}\cdot n\cdot\left[\frac{\left(\frac{K_0(w)}{wK_1(w)} + \frac{1}{w^2}\right)u^4w^4}{J_1^2(u)v^6}\right], \tag{8}$$

where

$$u = \frac{d}{2}k_0\sqrt{n_c^2 - (\beta/k_0)^2}$$

$$w = \frac{d}{2}k_0\sqrt{(\beta/k_0)^2 - n^2}$$

$$v = \frac{d}{2}k_0\sqrt{n_c^2 - n^2}.$$

$\Delta\beta$ has the Δ dependence predicted from the quadratic case. It must be noted that the part in square brackets is only dependent on the parameter v, the normalized frequency. As shown in Ref. 6 the magnitude of v determines uniquely the properties of the mode independent

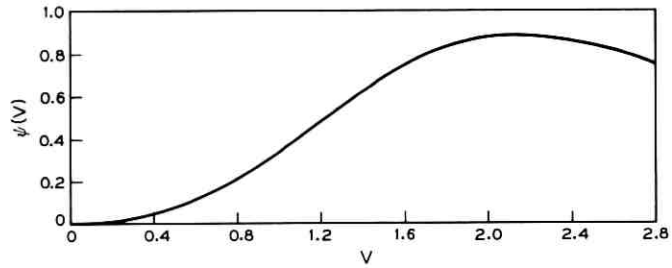


Fig. 3—The normalized perturbation $\psi(v)$ of the propagation constant.

of the guide parameters (only for small Δ of course). We can therefore express $\Delta\beta$ as the product of the guide parameters and a function $\psi(v)$ depending only on the binding properties of the mode:

$$\Delta\beta = \Delta^2 \cdot 2 \frac{\Delta d}{d} n\psi(v). \tag{9}$$

$\psi(v)$ is plotted in Fig. 3. The group delay difference between the two polarizations is given by

$$\Delta\tau = \frac{L \cdot n}{c} \Delta^2 2 \frac{\Delta d}{d} \frac{d}{dv} (v \cdot \psi(v)). \tag{10}$$

The v dependent part is plotted in Fig. 4. Since it never exceeds 1.6, we can use the upper limit

$$\boxed{\Delta\tau_{\max} = 3.2 \frac{L \cdot n}{c} \frac{\Delta d}{d} \Delta^2} \tag{11}$$

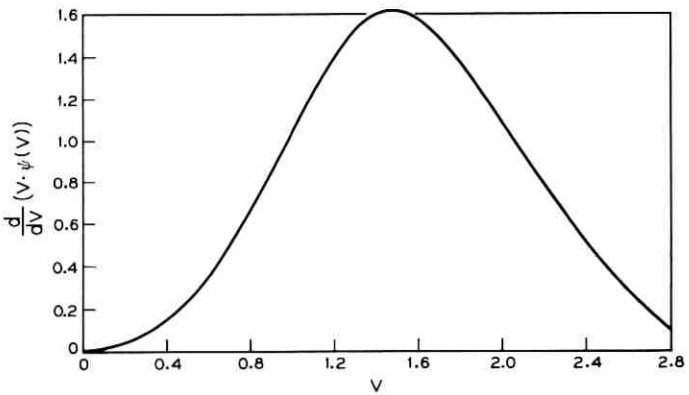


Fig. 4—The normalized delay distortion $d\ (v \cdot \psi(v))/dv$.

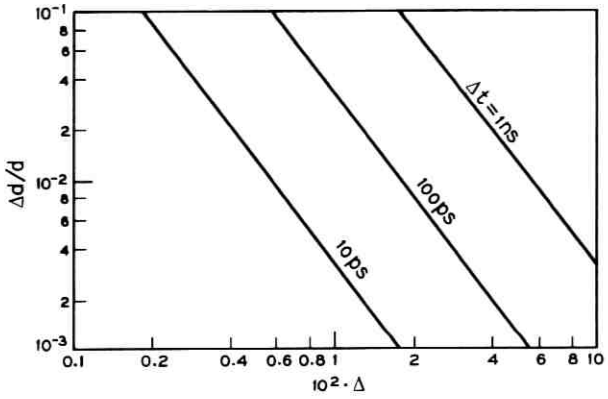


Fig. 5—The maximum diameter variation as a function of Δ.

to establish tolerances for $(\Delta d)/d$ as a function of Δ, assuming that the core diameter is chosen to get appropriate binding properties. Figure 5 shows the upper limit of $(\Delta d)/d$ as a function of Δ corresponding to various delay distortions. The length L of the guide is 3 km and $n = 1.5$. For 1 ns delay distortion and $\Delta = 1.8 \times 10^{-2}$ the diameter variation must be smaller than 10 percent which is not difficult to achieve. The requirements for a 10-ps delay distortion however ($\Delta < 1.8 \times 10^{-3}$ for $(\Delta d)/d < 10$ percent) are not so trivial anymore. We can conclude from these data that for a system in the several 100-Mb/s range the delay distortion due to elliptical deformation does not cause any difficult tolerance problems. At much higher speeds, however, the effect must be seriously considered.

REFERENCES

1. Kapron, F. P., Keck, D. B., and Maurer, R. D., "Radiation Loss in Glass Optical Waveguides," Appl. Phys. Letters, *17*, (1970), pp. 423–425.
2. Gloge, D., "Dispersion in Weakly Guiding Fibers," Appl. Opt., *10*, (1971), pp. 2442–2445.
3. Schlosser, W. O., and Unger, H. G., "Partially Filled Waveguides and Surface Waveguides of Rectangular Cross Section," Advances in Microwaves, *1*, (1966), pp. 319–387.
4. Dyott, R. B., and Stern, J. R., "Group Delay in Glass Fiber Waveguide," Elec. Letters, *7*, (1971), pp. 82–84.
5. Schlosser, W. O., "Die Stoerung der Eigenwerte des runden dielektrischen Drahtes bei schwacher elliptscher Deformation der Randkontur," AEU, *19*, (1965), pp. 1–8.
6. Gloge, D., "Weakly Guiding Fibers," Appl. Opt., *10*, (1971), pp. 2252–2258.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Optimal Reception for Binary Partial Response Channels

By M. J. FERGUSON

(Manuscript received September 3, 1971)

This paper describes an exceptionally simple scheme for binary partial response signal formats of the form $a_k \pm a_{k-l}$ *(for* $l \geq 1$*, and* $a_k = \pm 1$*). The receiver implements the maximum likelihood detector of the sequence* a_k *assuming additive white Gaussian noise as the channel impairment. It is simpler and more efficient than the scheme recently described by G. D. Forney.*[1] *It is, however, not generalizable to multilevel signaling while still retaining its simplicity.*

I. INTRODUCTION

There has recently been considerable interest in using the inherent redundancy of the partial response signal formal to approach the error rate versus signal-to-noise-ratio performance equivalent to binary antipodal signaling. Forney[1] at the 1970 International Symposium on Information Theory discussed a simple decoding scheme which he shows to be asymptotically optimal for high signal-to-noise ratio for channels with white additive Gaussian noise.

This paper describes a receiver for binary partial response signaling which is optimal for white additive Gaussian noise. This demodulator is much simpler than the equivalent two-level scheme of Forney. However, the extension to four or more levels seems to result in a scheme of much greater complexity than Forney's. In the first part of the paper we briefly review binary Class IV partial response signaling. Then we derive the optimal detection scheme for binary signaling which has a particularly simple implementation. A simple analysis of the memory requirements of the implementation follows. Finally we discuss some of the problems of extensions to multilevel signaling.

II. A PARTIAL RESPONSE SYSTEM*

The motivation for binary partial response signaling schemes is to

* This section is almost entirely due to D. D. Falconer.[2] E. R. Kretzmer[3] and A. Lender[4] did the original work in this area and Lucky, Salz, and Weldon[5] have a good survey and summary.

allow transmission of two bits per cycle of bandwidth without requiring ideal boxcar filters. The train of signal waveforms is shaped so that inherent intersymbol interference does not affect decisions made by the receiver.

Figure 1 shows a basic partial response signaling scheme transmitting $1/T$ bits per second. Information bits (a_k) are represented by $+1$s and -1s. Signal shaping is done by the filter whose transfer function is $X(\omega)$. A "Class IV" partial response function $X(\omega)$ and its associated sampled impulse response are shown in Fig. 2. This particular scheme is useful since it has no transmitted dc component. It is used in several existing and proposed partial response modems.

The transmitted Class IV partial response signal $s(t)$ can be represented in terms of the sequence of samples (x_k) spaced at Nyquist intervals (T seconds) as

$$s(t) = A \sum_k x_k \operatorname{sinc}\left(\frac{\pi t}{T} - k\pi\right) \tag{1}$$

where

$$\operatorname{sinc}(x) = \frac{\sin x}{x}$$

and

$$x_k = a_k - a_{k-2}, \qquad k = 1, 2, \cdots. \tag{2}$$

When the information symbols a_k take on values ± 1, then the samples (x_k) have three possible levels: 0, $+2$, or -2. Thus the scheme would be expected to be more sensitive to noise than is a comparable binary antipodal scheme in which $x_k = \pm 1$, and in which the transmitting filter's transfer function is a "boxcar." In fact, when independent hard decisions are made on each bit, it can be shown to require 3 dB higher signal-to-noise ratio in order to achieve the same error rate as the comparable binary antipodal scheme. A more efficient conventional partial response configuration which is 2.1 dB worse than binary antipodal[5] is to provide a matched filter at the receiver by replacing

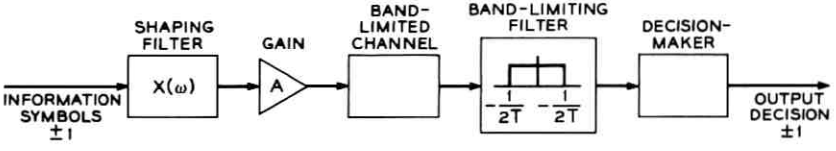


Fig. 1—Partial response system.

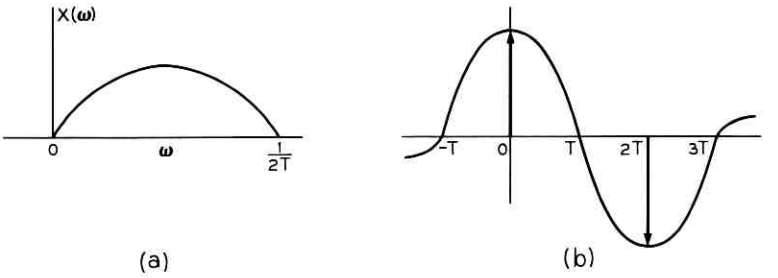


Fig. 2—Class IV partial response: (a) spectrum; (b) sampled impulse response.

$X(\omega)$ at the transmitter by $X(\omega)^{\frac{1}{2}}$ at the transmitter and $X(\omega)^{\frac{1}{2}}$ at the receiver. Other classes of binary partial response systems are worse than the ideal binary antipodal scheme by various amounts (see Table 4-2, page 91, in Ref. 5).

With $a_k = \pm 1$, $x_k = a_k - a_{k-2}$ is a sequence of three-level signals. However not all the sequences are possible! For example, if $a_k = +1$, $x_k = +2$ or 0 and if $a_k = -1$, $x_k = 0$ or -2. All the schemes described use this inherent redundancy to win back the 2.1-dB loss alluded to previously. Finally we note that all partial responses of the form

$$x_k = a_k - a_{k-l}, \qquad l \geqq 1,$$

produce l noninteracting streams of x_ks. For $l = 2$, the even x_ks and the odd x_ks are entirely independent. A scheme for $l = 1$ can be used for any $l \geqq 1$ by time sharing its operation with the other independent streams of x_ks. This observation allowed Forney to assert the applicability of his scheme for all l. It also allows us to consider only $l = 1$.

III. DERIVATION OF OPTIMAL RECEIVER

The receiver that we develop is to be optimal for additive white Gaussian noise and the signaling format

$$x_k = a_k - a_{k-1} \tag{3}$$

with

$$a_{-1} = 1.$$

A simple way to describe the sequence of x_ks resulting from a sequence of a_ks is given by the trellis in Fig. 3. The branches of the trellis are the x_k values and the nodes are the a_k values. The upper nodes are $+1$

and the lower values are -1. We trace a particular sequence of x_ks by following the branches joining the nodes for the appropriate a_k. For instance if we have the sequence starred (*) in Fig. 3, namely $\cdots$,

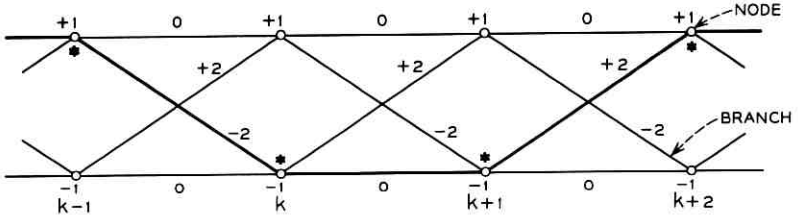


Fig. 3—Signal trellis.

$a_{k-1} = 1$, $a_k = -1$, $a_{k+1} = -1$, $a_{k+2} = +1$, $\cdots$, then we have the output sequence $\cdots$, $x_k = -2$, $x_{k+1} = 0$, $x_{k+2} = +2$, $\cdots$ by following the appropriate branches. Notice that we have the capability of describing any possible sequence of x_ks using Fig. 3. Further we note that any node has two branches leading to it and away from it.

The channel is assumed to add white Gaussian noise n_k, with density $N(0, \sigma^2)$* giving a received signal $y_k = x_k + n_k$. It is well known that the maximum likelihood receiver chooses the infinite sequence of $\hat{a}_k$s which maximize

$$\frac{1}{2}\sum_{k=0}^{\infty} y_k(\hat{a}_k + \hat{a}_{k-1}) - \frac{1}{4}\sum_{k=0}^{\infty}(\hat{a}_k - \hat{a}_{k-1})^2 \tag{4}$$

$$\hat{a}_{-1} = 1,$$

for a given sequence of y_ks. The $\hat{a}_k$s are the estimates of the transmitted sequence $\{a_k\}$. While it is clearly impossible to maximize (4) directly, it is possible to maximize (4) sequentially. We note that we can represent all possible sequences of $\hat{a}_k$ by paths in the signal trellis in Fig. 3. We also note that we can represent all possible sums in (4) as the result of paths through a trellis. We then obtain the trellis in Fig. 4. When $\hat{a}_k$ and $\hat{a}_{k+1}$ are of the same sign, the branch contributes 0 to the sum in (4) but when $\hat{a}_k = +1$, $\hat{a}_{k-1} = -1$, it contributes $y_k - 1$ to the sum. Similarly when $\hat{a}_k = -1$, $\hat{a}_{k-1} = +1$, the branch contributes $-y_k - 1$ to the sum. We say that a specific sequence of $\hat{a}_k$s through the trellis describes a *path*. All paths must have $\hat{a}_k = +1$ or $\hat{a}_k = -1$.

* $N(a, b)$ is the Gaussian density with mean a and variance b.

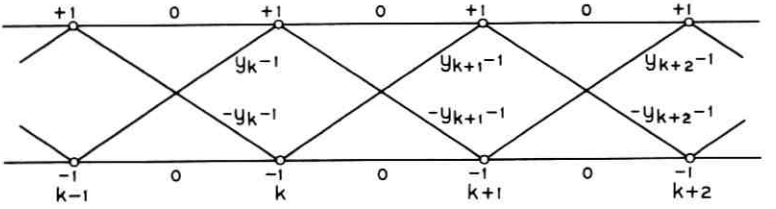


Fig. 4—Received signal trellis.

For all those paths with $\hat{a}_k = +1$ we can write (4) as

$$\left\{\frac{1}{2}\sum_{l=0}^{k} y_l(\hat{a}_l - \hat{a}_{l-1}) - \frac{1}{4}\sum_{l=0}^{k}(\hat{a}_l - \hat{a}_{l-1})^2\right\} + \left\{\sum_{l=k+1}^{\infty} y_l(\hat{a}_l - \hat{a}_{l-1}) - \frac{1}{4}\sum_{l=k+1}^{\infty}(\hat{a}_l - \hat{a}_{l-1})^2\right\} \tag{5}$$

where

$$\hat{a}_k = 1.$$

Thus it is *necessary* that any path with $\hat{a}_k = 1$ and which maximizes (5) also maximizes the first bracketed sum in (5). But this first bracketed sum in (5) depends only on $\{\hat{a}_0, \cdots, \hat{a}_{k-1}\}$ and this portion of the path can be chosen *independently* of the rest of the path. Define

$$f_k^+ \triangleq \max_{\substack{\text{all paths} \\ \text{with } \hat{a}_k = 1}} \left\{\frac{1}{2}\sum_{l=0}^{k} y_l(\hat{a}_l - \hat{a}_{l-1}) - \frac{1}{4}\sum_{l=0}^{k}(\hat{a}_l - \hat{a}_{l-1})^2\right\}. \tag{6a}$$

We similarly define, for the best path leading to $\hat{a}_k = -1$,

$$f_k^- = \max_{\substack{\text{all paths} \\ \text{with } \hat{a}_k = -1}} \left\{\frac{1}{2}\sum_{k=0}^{k} y_l(\hat{a}_l - \hat{a}_{l-1}) - \frac{1}{4}\sum_{l=0}^{k}(\hat{a}_l - \hat{a}_{l-1})^2\right\}. \tag{6b}$$

Finally we see that there are only four branches from the kth to the $(k+1)$st node. Hence, if we have the best path to $\hat{a}_k = \pm 1$, then at $\hat{a}_{k+1} = +1$ we must choose between only two paths, the one coming from $\hat{a}_k = +1$ having a value f_k^+ and the one from $\hat{a}_k = -1$ having a value $f_k^- + y_{k+1} - 1$. The best path is obviously the one with the largest value. Thus we have

$$f_{k+1}^+ = \max \begin{cases} f_k^+ & (+ \text{ PATH}) \\ f_k^- + y_{k+1} - 1 & (- \text{ PATH}). \end{cases} \tag{7a}$$

Similarly we have the best path to $\hat{a}_{k+1} = -1$ as the solution of

$$f_{k+1}^{-} = \max \begin{cases} f_k^{+} - y_{k+1} - 1 & (+ \text{ PATH}) \\ f_i^{-} & (- \text{ PATH}). \end{cases} \tag{7b}$$

Thus at any point in time we have two paths, one of which must be the beginning of the one which optimizes (4).* We say we are not *merged* at $(k-1)$ if we still have two paths left at k. Figure 5 shows

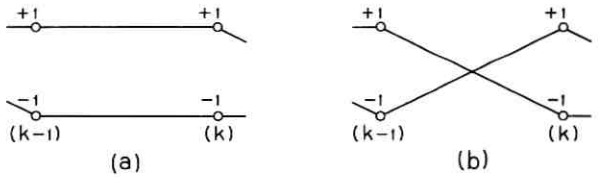


Fig. 5—Possible nonmerge paths.

the only two possibilities to remain unmerged. For the $(+)$ path to go to $\hat{a}_{k+1} = +1$ and the $(-)$ path to go to $\hat{a}_{k+1} = -1$ (Fig. 5a) we need both

$$f_k^{+} - f_k^{-} - y_{k+1} + 1 \geqq 0 \qquad \text{from (7a)}$$

and

$$-(f_k^{+} - f_k^{-}) + y_{k+1} + 1 \geqq 0. \qquad \text{from (7b)}$$

Thus we require

$$-1 \leqq f_k^{+} - f_k^{-} - y_{k+1} \leqq 1. \tag{8}$$

For us to remain unmerged on the "crossover" path of Fig. 5b we require

$$\begin{cases} \text{and} & (f_k^{+} + f_k^{-}) - y_{k+1} > 1 \\ & (f_k^{+} - f_k^{-}) - y_{k+1} < -1 \end{cases}.$$

This is clearly impossible. Thus we remain unmerged if and only if (8) is true. Hence when (8) is true, we are unmerged and the most likely path *from* the $+$ node leads *to* the $+$ node, and the most likely

* The formulation and solution of this problem is a simple example of Dynamic Programming and an application of Bellman's Principle of Optimality.[6] This may also be considered as a simple example of the Viterbi Algorithm which was shown by J. K. Omura[7] to be equivalent to Dynamic Programming. Finally, the identical formulation and solution to this problem was also obtained independently by H. Kobayashi[8] and M. Segal (unpublished).

path *from* the $-$ node leads *to* the $-$ node. We also note that

$$f_k^+ - f_k^- - y_{k+1} > 1 \tag{9}$$

implies both best paths came from $\hat{a}_{k-1} = +1$ [a $(+)$ merge] and

$$f_k^+ - f_k^- - y_{k+1} < 1 \tag{10}$$

implies both best paths came from $\hat{a}_{k-1} = -1$ [a $(-)$ merge]. Finally we see that all decisions as to merge or not are based on $f_k^+ - f_k^-$ and not on either separately. We then define

$$\Delta_k \triangleq f_k^+ - f_k^-.$$

Subtracting (7b) from (7a) and noting (8), (9), and (10) gives

$$\Delta_{k+1} = \begin{cases} y_{k+1} + 1, & \Delta_k - y_{k+1} > 1 & (+ \text{ MERGE}) \text{ at } k \\ \Delta_k, & -1 < \Delta_k - y_{k+1} < 1 & (\text{NO MERGE}). \\ y_{k+1} - 1, & \Delta_k - y_{k+1} < -1 & (- \text{ MERGE}) \end{cases} \tag{11}$$

The optimal receiver implements (11). We see that while unmerged, Δ_k remains the same. Only the testing to see if we have finally merged depends on the incoming data. The value of Δ_k while unmerged is just that resulting from the two paths leading from the most recent merge. Thus if the most recent merge was $(+)$ at node $l - 1$ then $\Delta_k = y_l + 1$ for $k \geqq l$ and no merge. Between merges, only two sequences are possible, either $\{1, 1, 1, \cdots\}$ or $\{-1, -1, -1, \cdots\}$. Hence in our implementation all we have to do is save Δ_l and the location of the most recent merge. Since we will be placing our data in a storage register prior to outputing it, we must decide which of the two between-merge sequences should we store. Obviously, the best is the most likely of the two.

We remember that after a $(+)$ merge at the $(k - 1)$ node

$$\Delta_k = y_k + 1.$$

If the $(+)$ merge is *correct* then

$$\begin{aligned} y_k &= a_k - a_{k-1} + n_k \\ &= -1 + a_k + n_k \end{aligned}$$

giving

$$\Delta_k = a_k + n_k .$$

Since $a_k = \pm 1$, then $\Delta_k = 1 + n_k$ if the transmitted path leads to the $+$ node and $\Delta_k = -1 + n_k$ if the transmitted path leads to the

— node. Hence the determination of the most likely path leading *from* the $(k-1)$ node is a binary hypothesis testing problem. The solution is to say the most likely path leads to the + node if $\Delta_k > 0$ and the — node if $\Delta_k < 0$.

For a correct — merge, the test is identical. The most likely path initially then is the one leading to $\text{sgn}(\Delta_k)$. If when we finally merge, the sign of Δ_l at the merge point is the same as the merge, the most likely path is the same as the most likely path initially chosen on the basis of $\Delta_k \gtrless 0$.

IV. IMPLEMENTATION

An implementation is suggested by the flow diagram of Fig. 6. We suppress the subscripts. The newly received signal is y and the previously stored difference is Δ. The decoded data are stored in a

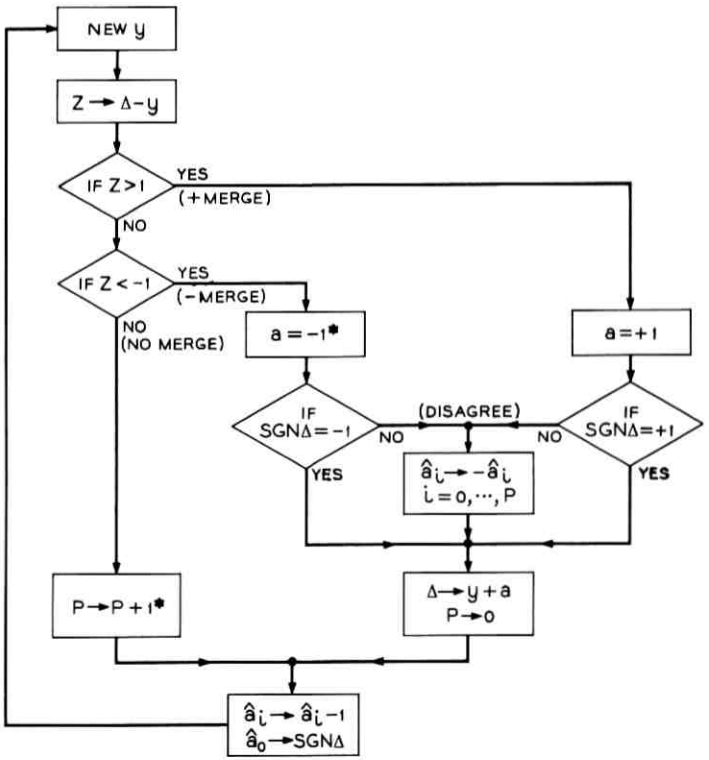


Fig. 6—Flow diagram for $x_k = a_k - a_{k-1}$.

register of length $N + 1$. $\hat{a}_0$ is the most recent decision and $\hat{a}_N$ is the bit about to be outputed. There is also a pointer, p, which indicates where the first data bit *after* the most recent merge point is located in the data register. The equations implemented are those of (11). The values of $\hat{a}_0, \cdots, \hat{a}_p$ are those of the initially most likely path as described in Section III. Referring to Fig. 6; when a new value y is obtained, we subtract it from the stored difference Δ calling this sum z. We then check to see if a merge has occurred according to (11). If $z > 1$ then we have a $+$ merge, if $z < -1$ we have a $-$ merge, and if $-1 \leqq z \leqq 1$ we have no merge. If a merge has occurred, then we will eventually replace Δ by $y + 1$ for a $+$ merge and $y - 1$ for a $-$ merge. We thus let $a \triangleq \pm 1$ for a $\pm$ merge. If the most likely path is actually the one we have been saving, then they must agree at the merge. We check this by finding out whether sgn Δ is the same as the merge value ± 1. If it is, then we have saved the most likely path. If it does not agree, the most likely path is the complement of the one we saved up to the most recent merge point p. We then complement $\hat{a}_0, \cdots, \hat{a}_p$. After we have our data set up, we replace Δ by its new value $y + a$, and reset the pointer p to 0. At this point we shift the register and place $\hat{a}_0 = \text{sgn } \Delta$.

If there is no merge, then life is simpler; Δ is the same and the pointer is advanced by 1; the register is then shifted and $\hat{a}_0 = \text{sgn } \Delta$. We are now ready for a new piece of data. Figure 7 shows a possible implementation of the above flow diagram.

V. BUFFER OVERFLOW (P > N) STRATEGY

Since we are saving the most likely sequence, we just output the buffer and keep $p = N$. If when we merge, we do indeed have the most likely sequence, then all is fine. If the most likely sequence is not actually held, then we complement the entire register. Although we have sent some suboptimally detected bits, this appears to be the best strategy. We could save ourselves most of the problem if we differentially encoded ("PRECODED"[5]) the data. This means we would let

$$x_k = \tilde{a}_k - \tilde{a}_{k-1}$$

where

$$\tilde{a}_k \tilde{a}_{k-1} = a_k .$$

Under these circumstances, a single suboptimal decision in the decoded

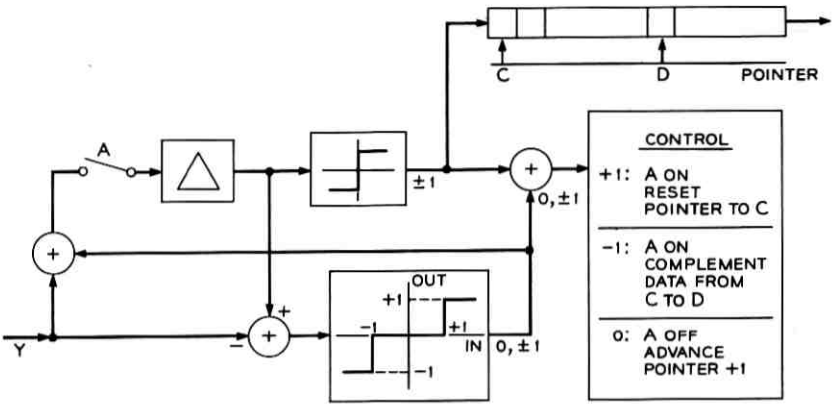


Fig. 7—Possible implementation for $x_k = a_k - a_{k-1}$.

$\tilde{a}_k$ path results in two errors in the a_k path. However, if we hold the complement $\tilde{a}_k$ path after a merge point rather than the most likely path, we only make a *single* suboptimal decision in decoding the a_ks rather than a possibly long burst. If we now complement the entire register, then we make an additional single error for the data bit affected by both $\hat{a}_N$ and $\hat{a}_{N+1}$ because $\hat{a}_N$ was complemented and $\hat{a}_{N+1}$ was not. If we do not complement the entire register, then the same phenomenon occurs at the next merge point. In both cases we make only two single suboptimal decisions whenever the buffer overflows. This obviously makes differential encoding of the data advisable.

VI. ANALYSIS OF BUFFER SIZE

In this section we determine the approximate probability of overflow of a buffer of length $(N + 1)$. If we have a $(+)$ merge then

$$\Delta = y_0 + 1 \tag{12}$$

and for a $(-)$ merge

$$\Delta = y_0 - 1 \tag{13}$$

where the "0" subscript refers to the merge position. Because y_0 is $N(-2, \sigma^2)$ for a $+-$ transition, $N(0, \sigma^2)$ for a $++$ or $--$ transition, and $N(+2, \sigma^2)$ for a $-+$ transition (see Fig. 3) we have, for a correct decision, substituting into (12) and (13), that Δ is $N(1, \sigma^2)$ for a transition *leading to* a $+$ node and Δ is $N(-1, \sigma^2)$ for a transition *leading to* a $-$ node for both types *of* previous merges. We know we remain

unmerged for N transitions if

$$-1 \leqq \Delta - y_l \leqq 1 \qquad l = 1, \cdots, N. \tag{14}$$

Since the $++$, $+-$, $-+$, $--$ transitions are all equally likely, we have $\Delta - y_l$ with the densities

$$\Delta - y_l \sim N(+3, \sigma^2) \text{ with probability } 1/4 \tag{15a}$$

$$\Delta - y_l \sim N(+1, \sigma^2) \text{ with probability } 1/4 \tag{15b}$$

$$\Delta - y_l \sim N(-1, \sigma^2) \text{ with probability } 1/4 \tag{15c}$$

$$\Delta - y_l \sim N(-3, \sigma^2) \text{ with probability } 1/4. \tag{15d}$$

For small σ^2 (large signal-to-noise ratio) both (15a) and (15d) lead to (14) *not* being satisfied with a very high probability. We can ignore these two events. (15b) and (15c) both correspond to $y_l \sim N(0, \sigma^2)$ and occur together with probability 1/2. Because of the symmetry of (14), (15b), (15c), and y_l we can write the probability of no merge for at least N nodes as

$$\begin{aligned} P(N) &\cong Pr(-1 \leqq \Delta - y_l \leqq 1, l = 1, \cdots, N) \\ &\cong \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp - \frac{(\Delta - 1)^2}{2\sigma^2} \left[\frac{1}{2} \int_{-1-\Delta}^{1-\Delta} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-y^2/2\sigma^2}\, dy\right]^N d\Delta. \end{aligned}$$

Now if σ^2 is small then Δ is concentrated about 1 and the limit $-1 - \Delta$ can be replaced by $-\infty$. We then have

$$P(N) \cong \left(\frac{1}{2}\right)^N \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x^2/2\sigma^2)} \left[\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-y^2/2\sigma^2}\, dy\right]^N dx. \tag{16}$$

Letting

$$Q(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-y^2/2\sigma^2}$$

and noting

$$dQ(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2/2\sigma^2},$$

we write (16) as

$$\begin{aligned} P(N) &= \left(\frac{1}{2}\right)^N \int_{-\infty}^{\infty} [Q(x)]^N\, dQ(x) \\ &= \frac{1}{N+1} \left(\frac{1}{2}\right)^N. \end{aligned} \tag{17}$$

To obtain (17) we have really used only the fact that the distributions of Δ and y are translates of each other and that they are symmetric about their mean value. Equation (17) was also derived independently by Kobayashi.[8]

The Forney scheme[1] has a probability of buffer overflow of 2^{-N}. Equation (17) indicates a factor of N improvement in this case. For $N = 20$, (17) gives $P(N) = 4.5 \times 10^{-8}$.

VII. GENERALIZATIONS

The most obvious generalization we would like is to four-level signaling. We can obtain a signal trellis in the same way as in the binary case but now we have four nodes at each time instant. Again we can write, using the same arguments as before, equations equivalent to (7). However, all the special structure which led to the exceptionally simple results of (11) seems to be missing. Instead of only one set of no-merge paths as indicated by Fig. 5a, we have many. Instead of only one possible way for either a + or − merge to occur, we have several. It also appears that all four possible paths through the trellis must be kept. In short, for a four-level signaling, Forney's scheme seems to be the simplest but for binary signaling, the one described here is best.

VIII. CONCLUSIONS

The system here is applicable to partial response signaling with binary data of the form

$$x_k = a_k \pm a_{k-l} \qquad \text{for all } l \geqq 1.$$

Differential encoding of the data is helpful to reduce the effects of buffer overflow. The extension to multilevel signaling destroys the beauty and simplicity of the binary scheme.

IX. ACKNOWLEDGMENTS

The author wishes to thank R. Gitlin and J. Salz for stimulating discussion.

REFERENCES

1. Forney, G. D., "Error Correction for Partial Response Modems," talk presented at the 1970 International Symposium on Information Theory, Noordwijk, the Netherlands, June 1970.
2. Falconer, D. D., "Forney's Error Correction Scheme for Partial Response Signals," unpublished work.

3. Kretzmer, E. R., "Generalization of a Technique for Binary Data Communication," IEEE Trans. Commun. Technology, *COM-14*, February 1966.
4. Lender, A., "The Duobinary Technique for High-Speed Data Transmission," 1963 IEEE Winter General Meeting.
5. Lucky, R. W., Salz, J., and Weldon, E. J., Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968, pp. 83–92.
6. Bellman, R., *Dynamic Programming*, Princeton, N. J.: Princeton University Press, 1957.
7. Omura, J. K., "On Optimum Receivers for Channels with Intersymbol Interference," talk presented at the 1970 International Symposium on Information Theory, Noordwijk, the Netherlands, June 1970.
8. Kobayashi, H., "Correlative Level Coding and Maximum-Likelihood Decoding," IEEE Trans. Inf. Theory, *IT-17*, No. 5 (September 1971), pp. 586–594.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Controllability and Observability in Linear Time-Variable Networks With Arbitrary Symmetry Groups

By H. RUBIN and H. E. MEADOWS

(Manuscript received May 21, 1971)

This paper presents a unified treatment of linear time-variable networks displaying arbitrary geometrical symmetries by incorporating group theory into an analysis scheme. Symmetric networks have their elements arranged so that certain permutations of the network edges result in a configuration which is identical with the original. These permutations lead to a group of monomial matrices which are shown to commute with the network A-matrix and the state transition matrix of the normal form equation. The representation theory of groups facilitates the study of those network properties which are determined solely by symmetry. By using group theory, a simple arithmetic condition is derived which, when satisfied, implies that the network is noncontrollable or nonobservable because of symmetry alone. The results allow the determination by inspection of linear combinations of the original state variables which result in noncontrollable variables. It is shown that networks displaying axial point group symmetry are generally only weakly controllable.

I. INTRODUCTION

In the past two decades, engineers and applied mathematicians have devoted a great deal of attention to diverse aspects of linear time-variable networks and systems. However, one problem that has not been treated in depth is that of analyzing time-varying networks displaying arbitrary geometrical symmetries. A symmetric network may be regarded as a set of identical subnetworks connected in a symmetric pattern. Such a circuit may be more easily implemented in an integrated form than is a nonsymmetric network, especially when the circuit is time-variable and the construction and synchronization of the variable elements are major technical problems. Since the trend in integrated circuit technology is towards large-scale integration, it may soon become

practically important to consider large networks displaying arbitrary geometrical symmetries. The present research was undertaken partly as a possible first step toward developing a modular approach to linear network design.

While it has long been known that network symmetries can be used to facilitate analysis, previous work on symmetric networks dealt mainly with bisection techniques for networks with mirror-plane symmetry and has not incorporated general types of symmetries into an analysis scheme. The present work treats arbitrary symmetries by utilizing the mathematics of group theory, a natural tool for studying symmetry.

Network controllability and observability are important concepts in analysis and synthesis, and group theory may be employed in determining symmetry-constrained noncontrollability and nonobservability of the network. Furthermore, the determination of these properties may be made by inspection without writing network equations. The group-theoretic approach enables us to prove several theorems concerning controllability and observability of a wide class of symmetric networks. The theorems would be difficult or impossible to prove, or to state precisely, without the use of group theory.

The reader who is unfamiliar with the results and notation used in both the abstract and representation theories of groups can find this material in Appendixes A and B in a form consistent with that used in the remainder of the text. The reader may wish to study the appendixes before continuing to Section II.

II. GROUP THEORY AND NETWORK EQUATIONS

Symmetric networks have their elements arranged so that certain permutations of the network edges result in a configuration which is identical with the original. For example, the geometrically symmetric network shown in Fig. 1 is invariant under permutations of the network edges which result from a rotation of the network structure by π radians about an axis perpendicular to the plane of the paper or from reflections in the planes σ_1 and σ_2.

Definition 1: A *covering operation* or *symmetry operation* is a transformation (rotation, reflection, etc.) which will bring the symmetric object (network) into a form indistinguishable from the original one. The following is well-known and shown in Ref. 1.

Theorem 1: The set of symmetry operations of an object constitutes a group.[‡]

[‡] See Appendix A for definitions of pertinent group theoretic terms.

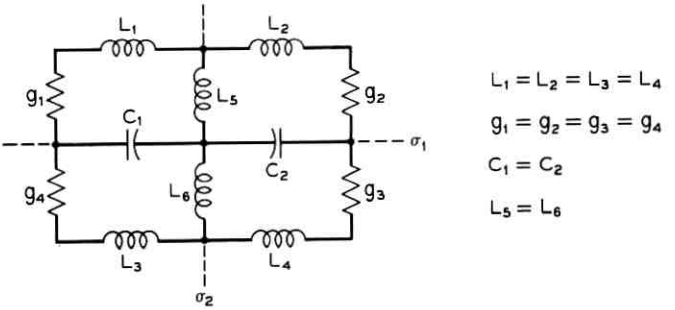


Fig. 1—Network with C_{2v} symmetry.

The effect of each symmetry operation is to permute the network edges. Specifically, only resistive edges are permuted among themselves, capacitive edges are permuted among themselves, etc., i.e., only edges of like type and equal element value or variation may be permuted. The letter R is used to denote the general symmetry operation of the symmetry group. Thus, R denotes either E, C_2, σ_1, or σ_2 for the network of Fig. 1, where E denotes the identity, C_2 denotes rotation by π radians, and σ denotes reflection in the plane σ. Thus, the operations $\{R\}$ form a group G_S which describes the symmetry of the network structure. The following exposition shows how group theory may be incorporated into the network analysis scheme.

Analysis in the time-domain can proceed from the normal form equation

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{1}$$

where $x(t)$ is an n-vector of state variables, $u(t)$ is a k-vector of inputs, and $A(t)$ and $B(t)$ are time-variable matrices conformable with x and u. In the context of our analysis, it is sufficient to consider $A(t)$, T. R. Bashkow's A-matrix.[2] The A-matrix contains some information about the network topology and also determines the natural response of the network. B. K. Kinariwala[3] showed that the A-matrix description is valid for time-varying as well as for fixed networks. The explicit form of the A-matrix given by P. R. Bryant[4] may be found in textbooks such as Refs. 5 and 6.

The A-matrix is derived with respect to a normal tree, i.e., a tree containing a maximum number of capacitive edges and a minimum number of inductive edges. It is assumed that the reader is familiar with the procedures needed to obtain the A-matrix, and the form of the network equilibrium equations is therefore given below:‡

‡ A superscript t appearing with a matrix denotes the transpose of that matrix; a superscript * denotes the complex conjugate of a scalar quantity.

$$\begin{bmatrix} pC_t & 0 & 0 & H_0 & H_1 & H_2 \\ 0 & G_t & 0 & 0 & H_3 & H_4 \\ 0 & 0 & \Gamma_t p^{-1} & 0 & 0 & H_5 \\ -H_0^t & 0 & 0 & D_c p^{-1} & 0 & 0 \\ -H_1^t & -H_3^t & 0 & 0 & R_c & 0 \\ -H_2^t & -H_4^t & -H_5^t & 0 & 0 & pL_c \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ i_3 \\ i_2 \\ i_1 \end{bmatrix} = \begin{bmatrix} j_1 \\ j_2 \\ j_3 \\ e_3 \\ e_2 \\ e_1 \end{bmatrix}. \tag{2}$$

In (2), the *subscript* t denotes elements in the tree, the *subscript* c denotes elements in the cotree, and p is differentiation with respect to time. Submatrices H_0 through H_5 express the topological relationship between elements in the tree and elements in the cotree. The letters e and j refer to voltage and current sources, respectively, while V and i denote branch variables. In order to obtain the A-matrix from equation (2), the nondynamic vector variables V_2 , V_3 , i_2 , i_3 are eliminated algebraically, thus yielding the equation

$$\begin{bmatrix} pC + G & T \\ -T^t & pL + R \end{bmatrix} \begin{bmatrix} V \\ i \end{bmatrix} = \begin{bmatrix} j \\ e \end{bmatrix}, \tag{3}$$

where j and e are regarded as inputs. For a complete interpretation of submatrices in the above equations, see Ref. 5. Equation (3) may be put into the form of equation (1) by choosing capacitor charges ($q = CV$) and inductor fluxes ($\phi = Li$) as state variables and writing

$$\frac{d}{dt}\begin{bmatrix} q \\ \phi \end{bmatrix} = -\begin{bmatrix} G & T \\ -T^t & R \end{bmatrix} \begin{bmatrix} C^{-1} & 0 \\ 0 & L^{-1} \end{bmatrix} \begin{bmatrix} q \\ \phi \end{bmatrix} + \begin{bmatrix} j \\ e \end{bmatrix}. \tag{4}$$

With reference to equation (2), observe that tree edges are used to define basic cutsets[7] of the network graph and hence current-law equations, while cotree edges are used to define basic loopsets[7] of the network graph and hence voltage-law equations. Thus, if a symmetry operation of the network structure permutes an edge in the tree with one of the cotree, the equilibrium equation (2) will be in a form different from that of the original equations; such an operation is not a symmetry operation of the network equilibrium equations. Those covering operations of the network structure which do not permute tree edges with cotree edges form a subgroup G_N of the group G_S (if two operations R_1 and R_2 do not permute tree edges with cotree edges, then the compound operation R_1R_2 also possesses that property), and the group G_N thus contains the symmetry operations of the equilibrium equations. Since the network equilibrium equations are being considered, the transformations of edge

voltages and currents are of importance rather than merely the permutations of network edges. The operations R of the group G_N may transform a voltage (or current) into the negative of another voltage (or current). If the network contains b edges, a b-dimensional monomial matrix[‡] $\hat{D}(R)$ may be formed which represents the transformation of the b voltages and currents under the symmetry operation R. The rows and columns of $\hat{D}(R)$ correspond to edge voltages and currents, and the matrix entries show how these quantities transform under the symmetry operation. Matrices $\hat{D}(R)$ form a reducible representation of the group G_N.

2.1 *Commutativity Relations*

In (2), denote the column vector of edge currents and voltages by $\hat{f}$, the column vector of current sources and voltage sources by $\hat{g}$, and the coefficient matrix by $\hat{N}$. Thus, equation (2) becomes

$$\hat{N}\hat{f} = \hat{g}. \tag{5}$$

Consider the new arrangement of sources and edges obtained by operating on the network with symmetry operation R, i.e., consider the equilibrium equations for the case

$$\begin{aligned} \hat{\hat{g}} &= \hat{D}(R)\hat{g} \\ \hat{\hat{f}} &= \hat{D}(R)\hat{f}. \end{aligned} \tag{6}$$

Since the operation R yields a network configuration, including the choice of tree, which is identical to the original one, it must be that

$$\hat{N}\hat{\hat{f}} = \hat{\hat{g}}. \tag{7}$$

Hence,

$$\hat{N}\hat{D}(R)\hat{f} = \hat{D}(R)\hat{g} = \hat{D}(R)\hat{N}\hat{f}, \tag{8}$$

where use is made of equations (5) and (6). For a network of b edges, b linearly independent vectors $\hat{f}$ may be specified such that the current-law and voltage-law equations are satisfied. Furthermore, b values of $\hat{g}$ are then obtained such that for each $\hat{f}$ chosen, the terminal relations of the network elements are satisfied. Thus, equality of the first and last members of (8) implies that

$$\hat{N}\hat{D}(R) = \hat{D}(R)\hat{N},$$

or

$$\hat{D}^{-1}(R)\hat{N}\hat{D}(R) = \hat{N}. \tag{9}$$

Thus the monomial representation $\hat{D}(R)$ commutes with $\hat{N}$.

[‡] A monomial matrix has only one nonzero entry in any row or column. The nonzero entry is restricted here to the values $+1$ or -1.

Equation (3) shows how elimination of nondynamic variables in equation (2) reduces $\hat{N}$ to a new matrix

$$N = \begin{bmatrix} pC + G & T \\ -T^t & pL + R \end{bmatrix},$$

and reduces $\hat{f}$ and $\hat{g}$ to the vectors

$$f = \begin{bmatrix} V \\ i \end{bmatrix} \quad \text{and} \quad g = \begin{bmatrix} j \\ e \end{bmatrix}.$$

The algebraic operations have eliminated from $\hat{N}$ the rows and columns corresponding to submatrices G_t , R_c , Γ_t , and D_c in equation (2). The elimination of rows and columns of $\hat{D}(R)$ which correspond to G_t , R_c , Γ_t , and D_c results in a group of matrices $D(R)$ which show only how tree capacitive voltages and cotree inductive currents are transformed under the operation R. By expanding equation (9) in terms of the submatrices of equation (2), it is possible to show[1] that $D(R)$ satisfies the following commutativity relation:

$$D^{-1}(R)ND(R) = N,$$

or (10)

$$D^{-1}(R)\begin{bmatrix} pC + G & T \\ -T^t & pL + R \end{bmatrix} D(R) = \begin{bmatrix} pC + G & T \\ -T^t & pL + R \end{bmatrix},$$

where $D(R)$ commutes with $\left[\begin{smallmatrix} C & 0 \\ 0 & L \end{smallmatrix}\right]$ and with $\left[\begin{smallmatrix} G & T \\ -T^t & R \end{smallmatrix}\right]$. Thus, the following may be stated.

Theorem 2: For a symmetric network, construct the monomial representation $D(R)$ of the symmetry group G_N , where $D(R)$ shows how the tree capacitive voltages and cotree inductive currents are transformed under the symmetry operation R. The monomial representation $D(R)$ commutes with the network A-matrix based either on voltages and currents or fluxes and charges as state variables. That is,

$$D^{-1}(R)A(t)D(R) = A(t), \quad \textit{for all} \quad R \in G_N .$$

The commutativity relation given in Theorem 2 establishes a basic connection between group theory and the network analysis problem, and allows group theoretic methods to be applied to linear networks displaying arbitrary geometrical symmetries.

The state transition matrix $\Phi(t, \tau)$ is the matric solution to the homogeneous part of equation (1) which satisfies

$$\Phi(\tau, \tau) = I, \tag{11}$$

where I is the unit matrix of appropriate order. $\Phi(t, \tau)$ is given in series form[8] as

$$\Phi(t, \tau) = \sum_{k=0}^{\infty} \Phi_k(t, \tau), \tag{12}$$

where

$$\Phi_{k+1}(t, \tau) = \int_{\tau}^{t} A(\rho)\Phi_k(\rho, \tau)\, d\rho \tag{13}$$

$$\Phi_0 = I.$$

Theorem 3: For a symmetric network, the monomial representation $D(R)$ of the symmetry group G_N commutes with $\Phi(t, \tau)$, i.e.,

$$D^{-1}(R)\Phi(t, \tau)D(R) = \Phi(t, \tau).$$

Proof: From (12) and (13),

$$D^{-1}(R)\Phi(t, \tau)\, D(R) = \sum_{k=0}^{\infty} D^{-1}(R)\Phi_k(t, \tau)\, D(R).$$

An induction procedure shows that $D(R)$ commutes with each term $\Phi_k(t, \tau)$ in the above sum.

$$D^{-1}(R)\Phi_1(t, \tau)\, D(R) = \int_{\tau}^{t} [D^{-1}(R)A(\rho)\, D(R)]I\, d\rho$$

$$= \int_{\tau}^{t} A(\rho)\, d\rho = \Phi_1(t, \tau).$$

Assume that $D(R)$ commutes with $\Phi_k(t, \tau)$. Hence,

$$D^{-1}(R)\Phi_{k+1}(t, \tau)\, D(R) = \int_{\tau}^{t} [D^{-1}(R)A(\rho)\, D(R)][D^{-1}(R)\Phi_k(\rho, \tau)\, D(R)]\, d\rho$$

$$= \int_{\tau}^{t} A(\rho)\Phi_k(\rho, \tau)\, d\rho = \Phi_{k+1}(t, \tau).$$

Thus, the theorem is proved.

III. Explicit Form of Transformation α to Reduce $A(t)$

In Appendix B, a procedure is given for the construction of a unitary matrix α from the irreducible representations of symmetry group G_N and representation $D(R)$. The important property possessed by the transformation α is that it transforms the state space to a new basis

in which $D(R)$ appears in block diagonal form and in which $A(t)$ appears in block diagonal form.[9] For the remainder of this paper, it is important to determine the positions of zero elements in the matrix α. Thus, the characterization of α in an explicit form is undertaken at this point. The following definition is adapted from group theory in a way useful to network analysis.

Definition 2: A symmetric network is said to be *transitive* if there is at least one group operation which transforms a given state variable into any other state variable (with plus or minus sign). The network is *intransitive* if it is not transitive.

Since an inductor and a capacitor cannot be permuted by any symmetry operation, general *RLC* symmetric networks are intransitive. The state variables can be partitioned into sets such that the group operations permute among themselves only those variables in the same set. Hence, each set is transitive, and the state variables are said to be partitioned into transitive sets.

Theorem 4: For a symmetric network, the number of transitive sets into which the state variables may be partitioned is equal to the number of times the totally symmetric irreducible representation [i.e., $D^{(1)}(R)$ having all group operations represented by unity] appears in $D_p(R)$, the permutation representation obtained from $D(R)$ by replacing each -1 entry in $D(R)$ by $+1$.

Proof: This result follows from a theorem given by W. Burnside (Ref. 10, p. 191) which states that

$$gs = \sum_{r=1}^{n} r\nu_r , \tag{14}$$

where g is the order of the group, n the number of symbols (state variables) operated on by the group, s the number of transitive sets in which the n symbols are permuted, and ν_r the number of group operators which leave exactly r symbols unchanged.

Let c_1^p denote the number of times that $D^{(1)}(R)$ appears in $D_p(R)$ and χ the trace of D. From (52) in Appendix A,

$$\begin{aligned} c_1^p &= \frac{1}{g} \sum_R \chi^{(1)}(R)^* \chi_p(R) \\ &= \frac{1}{g} \sum_R \chi_p(R) \end{aligned} \tag{15}$$

since $\chi^{(1)}(R) = 1$ for all R. Because $D_p(R)$ is a permutation representa-

tion, $\chi_p(R)$ is precisely the number of state variables left unchanged by operation R, and hence is an integer from zero to n. The group operators can be partitioned such that all operations in a given set leave unchanged the same number of state variables. It is now evident that (14) and (15) are identical sums, and c_1^p is equal to s.

The column vectors $\alpha_{p\pi a}$, $a = 1, \cdots, c_p$, of the matrix α are given in (56) in Appendix B and repeated here for convenience. They are c_p linearly independent (and orthonormal) columns of

$$G_{\pi}^{(p)} = \sum_R D^{(p)}(R)^*_{\pi\pi} D(R)I, \tag{16}$$

where I is the unit matrix and $p\pi a$ are indices defined as follows. The index p denotes one of the distinct irreducible representations of the symmetry group, the index π runs from 1 to l_p and denotes a row of the matrix $D^{(p)}(R)$ [so that the dimension of $D^{(p)}(R)$ is l_p], and the index a denotes one of the c_p appearances of $D^{(p)}(R)$ in $D(R)$. Thus, α has the form

$$\alpha = [\alpha_{111}, \cdots, \alpha_{\mu 11}, \cdots, \alpha_{\mu 1 c_\mu}, \cdots, \alpha_{\mu\pi 1}, \cdots, \alpha_{\mu\pi c_\mu}, \cdots]. \tag{17}$$

We consider a typical column vector $\alpha_{\mu\pi a}$ corresponding to the πth row of $D^{(\mu)}(R)$. Let e_m be the vector containing all zeros except for unity in the mth row. From (16), $\alpha_{\mu\pi m}$ may be considered to result from (we delete the normalization factor)

$$\alpha_{\mu\pi m} = \sum_R D^{(\mu)}(R)^*_{\pi\pi} D(R)e_m . \tag{18}$$

If c_μ is less than or equal to c_1^p, the number of transitive sets into which the state variables may be partitioned (Theorem 4), the c_μ values of the index m can be chosen such that each vector e_m corresponds to a different transitive set. The operation $D(R)e_m$ results in a new vector e_k where m and k are in the same transitive set. Thus, the c_μ vectors $\alpha_{\mu\pi m}$ chosen above are necessarily linearly independent if they are not zero. If any choice of m yields a zero result in (18), merely choose a value of m corresponding to a different transitive set; c_μ linearly independent $\alpha_{\mu\pi m}$ must be obtained in this way since the matrix $G_{\pi}^{(\mu)}$ has rank c_μ.[11]

Lemma 1: If the vectors $\alpha_{\mu\pi m}$ are chosen as outlined above, only one of the vectors having indices μ and π can possibly have a nonzero result in row r, namely, $\alpha_{\mu\pi\rho}$ where r and ρ are in the same transitive set.

The rth component of $\alpha_{\mu\pi m}$ is denoted by $\alpha^r_{\mu\pi m}$. The group operators may be partitioned into the set $\{R_\rho^r\}$ and its complement $\{\bar{R}_\rho^r\}$, where $\{R_\rho^r\}$ consists of all group operations which take the ρth state variable

into the rth state variable. Thus,

$$\begin{aligned}\alpha_{\mu\pi\rho} &= \sum_{R} D^{(\mu)}(R)^{*}_{\pi\pi}\, D(R)e_{\rho} \\ &= \sum_{R_\rho{}^r} D^{(\mu)}(R_\rho^r)^{*}_{\pi\pi}\, D(R_\rho^r)e_{\rho} + \sum_{\bar{R}_\rho{}^r} D^{(\mu)}(\bar{R}_\rho^r)^{*}_{\pi\pi}\, D(\bar{R}_\rho^r)e_{\rho} \\ &= \sum_{R_\rho{}^r} s_\rho^r\, D^{(\mu)}(R_\rho^r)^{*}_{\pi\pi}e_r + \sum_{\bar{R}_\rho{}^r} D^{(\mu)}(\bar{R}_\rho^r)^{*}_{\pi\pi}\, D(\bar{R}_\rho^r)e_{\rho}\,, \end{aligned} \tag{19}$$

where s_ρ^r is $+1$ or -1 as R_ρ^r transforms state variable x_ρ into state variable x_r with positive or negative sign, respectively. Hence, except for a scale factor,

$$\alpha^r_{\mu\pi\rho} = \sum_{R_\rho{}^r} s_\rho^r\, D^{(\mu)}(R_\rho^r)^{*}_{\pi\pi}\,. \tag{20}$$

In determining whether the rth component of vectors $\alpha_{\mu\pi\rho}$ is zero, there may be some ambiguity in choosing the index ρ in the same transitive set as r. The following lemma eliminates any ambiguity in this choice.

Lemma 2: A necessary and sufficient condition for $\alpha^r_{\mu\pi\rho}$ to be zero for all ρ in the same transitive set as r is that $\alpha^r_{\mu\pi r}$ be equal to zero, i.e., that

$$\sum_{R_r{}^r} s_r^r\, D^{(\mu)}(R_r^r)^{*}_{\pi\pi} = 0. \tag{21}$$

Proof: Consider the subgroup $\mathfrak{K}$ of the group G_N, where $\mathfrak{K}$ consists of those group operations which transform the rth state variable into itself with either positive or negative sign. The subset H of $\mathfrak{K}$ which transform the rth state variable into itself with positive sign forms a subgroup of index two in $\mathfrak{K}$.[12] Thus, $\mathfrak{K}$ may be partitioned into cosets with respect to H as

$$\mathfrak{K} = H, PH, \tag{22}$$

where P is an operation of $\mathfrak{K}$ not contained in H, and thus transforms the rth state variable into itself with minus sign. The group G_N may be partitioned into cosets with respect to $\mathfrak{K}$ as

$$G_N = H, PH, R_r^{\rho}H, R_r^{\rho}PH, \cdots, R_r^{\eta}H, R_r^{\eta}PH,$$

where R_r^i denotes a group operation which transforms the rth state variable into the ith with plus sign. It is clear from the above that if any group operations transform any symbols (state variables) with negative sign, there must exist an equal number of group operations which transform the symbols with positive sign. Hence, for each transitive set of symbols (state variables) operated on by G_N, the subset of group operations which permutes the symbols with positive sign forms

a subgroup of index two. Therefore, there exists a one-dimensional irreducible representation $D^{(\bar{\mu})}(R)$ of G_N in which each group operation which transforms the symbols with plus sign is represented by $+1$, while each operation which transforms the symbols with minus sign is represented by -1.[12] Thus, in equation (20),

$$s_\rho^r = D^{(\bar{\mu})}(R_\rho^r). \tag{23}$$

The orthogonality relation (51) for irreducible representations requires that

$$\sum_R D^{(\bar{\mu})}(R)\, D^{(\mu)}(R)^*_{\pi\pi} = \frac{g}{l_{\bar{\mu}}}\, \delta_{\bar{\mu}\mu} = 0 \tag{24}$$

if $\bar{\mu} \neq \mu$. Notice that if $D(R)$ is a permutation representation, then $\bar{\mu} = 1$, and $D^{(1)}(R)$ is the totally symmetric irreducible representation.

Let the transitive set to which r belongs be denoted by M_r consisting of $\{r, \rho, \cdots, \eta\}$. Hence, the set of group operations may be partitioned into $\{R_r^r\}, \{R_r^\rho\}, \cdots, \{R_r^\eta\}$, and from equations (23) and (24),

$$\sum_R D^{(\bar{\mu})}(R)\, D^{(\mu)}(R)^*_{\pi\pi} = \sum_{R_r^r} s_r^r\, D^{(\mu)}(R_r^r)^*_{\pi\pi} + \sum_{R_r^\rho} s_r^\rho\, D^{(\mu)}(R_r^\rho)^*_{\pi\pi}$$

$$+ \cdots + \sum_{R_r^\eta} s_r^\eta\, D^{(\mu)}(R_r^\eta)^*_{\pi\pi} = 0. \tag{25}$$

Now, the matrices $D^{(\mu)}(R)$ are unitary, and $\{R_r^\rho\} = \{[R_\rho^r]^{-1}\}, \cdots, \{R_r^\eta\} = \{[R_\eta^r]^{-1}\}$. Hence, if $\alpha^r_{\mu\pi\rho} = \cdots = \alpha^r_{\mu\pi\eta} = 0$, then by virtue of (20), equation (25) implies that $\alpha^r_{\mu\pi r} = 0$ as well. This proves necessity.

There are two cases to consider in proving sufficiency, namely, $c_\mu \leqq c_1^p$ and $c_\mu > c_1^p$.

Case (a) $c_\mu \leqq c_1^p$:

In this case c_μ linearly independent vectors $\alpha_{\mu\pi a}$ may be obtained by choosing vectors e_m in (18) so that each m corresponds to a different transitive set. Suppose e_r is chosen corresponding to the set M_r. The addition of $\alpha_{\mu\pi\rho}$ to the set thus results in a dependent set. Furthermore, by construction, all vectors except $\alpha_{\mu\pi r}$ are zero in positions where $\alpha_{\mu\pi\rho}$ is nonzero. Thus, $\alpha_{\mu\pi r}$ and $\alpha_{\mu\pi\rho}$ are proportional, i.e., if $\alpha^r_{\mu\pi r} = 0$, then $\alpha^r_{\mu\pi\rho} = 0$. The last result holds true for all $\rho \in M_r$, and sufficiency is established for Case (a).

Case (b) $c_\mu > c_1^p$:

In this case, it may be possible to choose more than one index in the transitive set M_r such that the $\alpha_{\mu\pi a}$ vectors obtained from equation (18) are linearly independent. A direct argument shows that a contradiction results if $\alpha^r_{\mu\pi r} = 0$ while $\alpha^r_{\mu\pi\rho} \neq 0$; namely, more linearly independent

vectors than is actually possible can be obtained from equation (18) by using indices m corresponding to the transitive set M_r . Thus, sufficiency for Case (b) is proved.

Hence, it has been established that for any indices μ and π, the vanishing of the rth component of $\alpha_{\mu\tau\sigma}$ is completely determined by the matrices $D^{(\mu)}(R_r^r)$, where the only group operations involved are those which transform the rth state variable into itself. This result will be used in the next section.

IV. CONTROLLABILITY OF SYMMETRIC NETWORKS

The concept of controllability relates to the degree to which the state of a system is affected by the application of some input. The following definition may be found in Ref. 13.

Definition 3: The system (1) is *completely controllable on an interval* (t_o , t_1) if for any state x_o at t_o and any desired final state x_1 at t_1 , there exists an input $u(t)$ defined on (t_o , t_1) such that $x(t_1) = x_1$.

The system (1) is *totally controllable on an interval* (t_o , t_1) if it is completely controllable on every subinterval of (t_o , t_1).

For networks with sufficiently smooth time-variations, controllability of the linear time-varying system (1) may be characterized by the controllability matrix[13]

$$\left.\begin{aligned} Q_c &= [P_0 P_1 \cdots P_{n-1}], \\ P_k &= -A(t)P_{k-1} + \dot{P}_{k-1} , \; P_0 = B \end{aligned}\right\} \quad (26)$$

where

and n is the order of the system.

The following theorem is a paraphrase of Theorem 4 of Ref. 13.

Theorem 5: For the system (1) assume that $A(t)$ and $B(t)$ together with their first $n - 2$ and $n - 1$ derivatives, respectively, are continuous functions. System (1) is totally controllable on the interval (t_o , t_1) if and only if Q_c does not have rank less than n on any subinterval of (t_o , t_1).

Lemma 3: The system described in partitioned form by

$$\frac{d}{dt}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u(t) \quad (27)$$

is noncontrollable (i.e., not completely controllable).

A sufficient condition for noncontrollability is given in the above lemma. This section is concerned with determining conditions in which symmetry alone is sufficient to cause the network to be noncontrollable.

Definition 4: A symmetric network is said to be NCS (possess the NCS property) if it is noncontrollable because of symmetry alone.

In (4), if only k of the inputs $[{}^{i}_{\bullet}]$ are nonzero, the equation can be rewritten using the k-vector of inputs, $u(t)$. Thus,

$$\dot{x} = A(t)x + Bu(t) \tag{28}$$

where B is an $n \times k$ constant matrix and $x = [{}^{q}_{\phi}]$. By making the unitary change of variable[‡] $Z = \alpha^{\dagger}x$, we arrive at the block diagonal system of equations

$$\frac{d}{dt}\begin{bmatrix} Z_1^1 \\ \vdots \\ Z_\pi^\mu \\ \vdots \\ Z_{l_\beta}^\beta \end{bmatrix} = \begin{bmatrix} \bar{A}_1^1 & & & & \\ & \ddots & & & \\ & & \bar{A}_\pi^\mu & & \\ & & & \ddots & \\ & & & & \bar{A}_{l_\beta}^\beta \end{bmatrix} \begin{bmatrix} Z_1^1 \\ \vdots \\ Z_\pi^\mu \\ \vdots \\ Z_{l_\beta}^\beta \end{bmatrix} + \alpha^{\dagger}Bu(t), \tag{29}$$

where Z is shown partitioned according to the submatrices $\bar{A}_\pi^\mu$. Controllability of the network reduces to that of all of the subsystems corresponding to the $\bar{A}_\pi^\mu$. From Lemma 3, the network is noncontrollable if a submatrix of $\alpha^{\dagger}B$ corresponding in its partition location to one of the $\bar{A}_\pi^\mu$ is zero. This occurrence is due solely to symmetry; we now investigate this condition more closely.

First consider the case where a single input is coupled only to the rth state variable, i.e., in (28), $B = e_r$ and $u(t)$ is a scalar. From α given in (17), it is evident that the submatrix of $\alpha^{\dagger}e_r$ corresponding to $\bar{A}_\pi^\mu$ is simply the $c_\mu \times 1$ partition consisting of the rth components of the vectors $\alpha_{\mu\pi 1}, \cdots, \alpha_{\mu\pi c_\mu}$. As mentioned in Lemma 1, at most one of these vectors can have nonzero entry in row r. Using the notation developed previously and Lemma 2, the following theorem has therefore been established.

Theorem 6: A symmetric time-varying network having a single input coupled only to the rth state variable is noncontrollable by virtue of its symmetry (i.e., is NCS) if and only if there is a μ such that $D^{(\mu)}(R)$ appears in $D(R)$ and

$$\sum_{R_r{}^r} s_r^r \, D^{(\mu)}(R_r^r)^*_{\pi\pi} = 0$$

for some value of π.

It is clear that if a table of irreducible representations is available, the arithmetic computation involved in the above theorem is quite simple. For any μ, all values of $\pi = 1, \cdots, l_\mu$ should be checked to determine

[‡] The complex conjugate transpose of the matrix α is denoted by $\alpha^{\dagger}$.

which state variables are uncontrollable. For most cases of interest, $l_\mu = 1$, 2, or 3; the point group of the regular icosahedron has irreducible representations of order five.[12]

If the irreducible representation in (21) is one-dimensional, the quantities $D^{(\mu)}(R)_{\pi\pi}$ are unambiguous. However, for irreducible representations whose dimension exceeds unity, any representation which is equivalent to $D^{(\mu)}(R)$ may be used to form the matrix α which reduces the system of (28) to that of (29). Clearly, although the block diagonal form of (29) will be essentially the same under transformations produced from equivalent irreducible representations, the matrix α will be different depending on which irreducible representation is used to construct it. Hence, for multidimensional irreducible representations, it is possible for $\alpha^\dagger B$ in equation (29) to have a zero submatrix if $D^{(\mu)}(R)$ is used to construct α, whereas nonzero submatrices may result if a representation equivalent to $D^{(\mu)}(R)$ is used to construct the transformation α. The above discussion shows that for multidimensional irreducible representations,

$$\sum_{R_r^r} s_r^r \, D^{(\mu)}(R_r^r)^*_{\pi\pi} \neq 0$$

is not sufficient to conclude that the network is *not* NCS. The inequality to zero of the sum in (21) must be shown for all representations equivalent to $D^{(\mu)}(R)$. In most cases of interest, the set $\{R_r^r\}$ consists of very few elements, and it may be quite easy to determine an irreducible representation which satisfies (21). The points mentioned in the above discussion will be illustrated by example in the sequel.

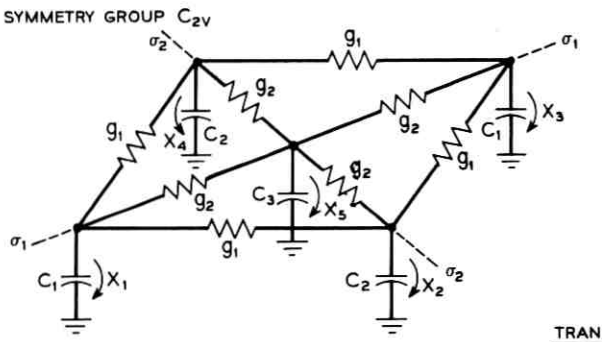
As an example illustrating the use of Theorem 6, consider the network of Fig. 2. The network has C_{2v} symmetry, and a table of irreducible representations of the group is given in Fig. 2. By utilizing (52), it is determined that the monomial‡ representation $D(R)$ contains $D^{(1)}(R)$ three times, $D^{(2)}(R)$ zero times, and $D^{(3)}(R)$ and $D^{(4)}(R)$ each one time. If a current source $I(t)$ is placed in parallel with the capacitor associated with state variable x_1, the following calculations can be made with regard to Theorem 6. The group operations which leave x_1 invariant are E and σ_1. Thus,

$$\sum_{R_r^r} D^{(1)}(R_r^r)^*_{\pi\pi} = 1 + 1 \neq 0$$

$$\sum_{R_r^r} D^{(3)}(R_r^r)^*_{\pi\pi} = 1 + 1 \neq 0$$

$$\sum_{R_r^r} D^{(4)}(R_r^r)^*_{\pi\pi} = 1 - 1 = 0.$$

‡ $D(R)$ is a permutation representation in this case, so that $s_r^r = +1$.



SYMMETRY OPERATIONS

$E = (1)(2)(3)(4)(5)$

$C_2 = (13)(24)$

$\sigma_1 = (24)$

$\sigma_2 = (13)$

IRREDUCIBLE REPRESENTATIONS

R	E	C_2	σ_1	σ_2
$D^{(1)}(R)$	1	1	1	1
$D^{(2)}(R)$	1	1	−1	−1
$D^{(3)}(R)$	1	−1	1	−1
$D^{(4)}(R)$	1	−1	−1	1

TRANSFORMATION

$$\alpha = \frac{1}{\sqrt{2}} \left[\begin{array}{ccc|c|c} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & \sqrt{2} & 0 & 0 \end{array}\right]$$

$D^{(1)}(R)$ $D^{(3)}(R)$ $D^{(4)}(R)$

Fig. 2—Network with C_{2v} symmetry, including symmetry operations, irreducible representations, and transformation matrix α.

Hence, if the excitation is coupled solely to state variable x_1, the basis function corresponding to $D^{(4)}(R)$ will be uncontrollable. Indeed, the block-diagonal system has the form

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix} = \begin{bmatrix} a & b & c & 0 & 0 \\ d & e & f & 0 & 0 \\ g & h & i & 0 & 0 \\ 0 & 0 & 0 & j & 0 \\ 0 & 0 & 0 & 0 & k \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix} + \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} I(t). \tag{30}$$

It is seen from the matrix α in Fig. 2 that $z_5 = x_2 - x_4$, and it is this linear combination of the original state variables which is uncontrollable in the present example. Note that since $D^{(2)}(R)$ does not appear in $D(R)$ in the above example, no basis functions are associated with it, and hence a computation is not made for this irreducible representation.

The next example serves to illustrate some complications that arise when the symmetry group possesses irreducible representations of dimension greater than unity. The network shown in Fig. 3 possesses symmetry C_{3v}. Two equivalent two-dimensional irreducible representations are given, and two transformation matrices α_1 and α_2 are shown

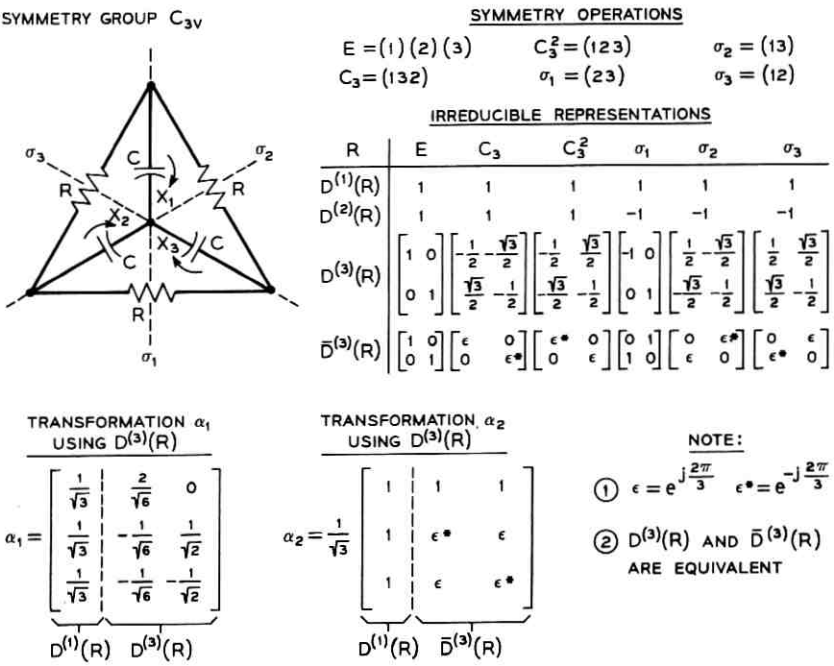


Fig. 3—Network with C_{3v} symmetry, including symmetry operations, irreducible representations, and transformation matrix α.

which will transform the differential equations to block diagonal form. It is determined that the irreducible representations $D^{(1)}(R)$ and $D^{(2)}(R)$ are contained one time and zero times, respectively, in the permutation representation $D(R)$, while the two-dimensional irreducible representation is contained once in $D(R)$. The transformation matrix α_1 is constructed using the real two-dimensional representation while α_2 is constructed using the complex two-dimensional representation. Both α_1 and α_2 are given in Fig. 3.

If a current source $I(t)$ is placed in parallel with the capacitor associated with state variable x_1, the block-diagonal system has the form (using the transformation α_1)

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & b \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 1/(3)^{\frac{1}{2}} \\ 2/(6)^{\frac{1}{2}} \\ 0 \end{bmatrix} I(t),$$

while if α_2 is used as the transformation matrix, the block-diagonal

system has the form

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & b \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} I(t).$$

The network is uncontrollable as shown with α_1 above. Uncontrollability of the network may be determined by inspection by using the real two-dimensional representation in Theorem 6. The set $\{R_r^r\}$ consists of $\{E, \sigma_1\}$, where $r = 1$.

The following corollary results from a trivial application of Theorem 6, but is by no means obvious without the use of the theorem.

Corollary 1: Given the assumptions of Theorem 6, if there is just one group operation that leaves the rth state variable invariant, then the network cannot be NCS.

Proof: The lone group operation must be the identity, and $D^{(\mu)}(E)_{\pi\pi} = 1$ for all μ and π.

An interesting and important result of Corollary 1 is that a network whose only symmetry operations are rotations (i.e., C_n groups) cannot be NCS except in the special case treated in Corollary 2 which follows.

Corollary 2: If the symmetric network contains a state variable which is invariant under all the group operations, and if the single input is coupled solely to this state variable, the network is NCS.

Proof: Since $\{R_r^r\}$ is the entire group, equation (51) yields

$$\sum_{R,r} s_r^r\, D^{(\mu)}(R_r^r)^*_{\pi\pi} = \sum_R D^{(\bar{\mu})}(R)\, D^{(\mu)}(R)^*_{\pi\pi} = \frac{g}{l_{\bar{\mu}}}\,\delta_{\mu\bar{\mu}} = 0, \qquad \mu \neq \bar{\mu}.$$

The network of Fig. 2 may also serve to illustrate Corollary 2. State variable x_5 is invariant under all the group operations. If an excitation $I(t)$ is coupled only to x_5, the block-diagonal system has the form ($\bar{\mu} = 1$ since $D(R)$ is a permutation representation)

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix} = \begin{bmatrix} a & b & c & 0 & 0 \\ d & e & f & 0 & 0 \\ g & h & i & 0 & 0 \\ 0 & 0 & 0 & j & 0 \\ 0 & 0 & 0 & 0 & k \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} I(t). \tag{31}$$

Corollary 3: The state variables associated with $D^{(\bar{\mu})}(R)$ represent an always excitable portion of the network, i.e., these variables are never NCS. If $\bar{\mu} \neq 1$, basis functions corresponding to $D^{(1)}(R)$ are always NCS.

Proof: From (22), the subset of group operations which transform the rth state variable into itself with plus sign forms a subgroup of index two in the group $\mathcal{K}$ of operations which transforms the rth state variable into itself with either plus or minus sign. Hence, the quantities s_r^r in equation (21) form an irreducible representation $D^{(\bar{\mu})}(R)$ of the group $\mathcal{K}$. Thus, for the basis functions corresponding to $D^{(\bar{\mu})}(R)$, equation (21) becomes

$$\sum_{R_r{}^r} s_r^r \, D^{(\bar{\mu})}(R_r^r)^* = \sum_{R_r{}^r} D^{(\bar{\mu})}(R) \, D^{(\bar{\mu})}(R)^* = k \neq 0,$$

where k denotes the order of the group $\mathcal{K}$. Likewise, $D^{(1)}(R)$ forms an irreducible representation of $\mathcal{K}$ since the totally symmetric representation is an irreducible representation of any abstract group. Thus, for the basis functions corresponding to $D^{(1)}(R)$, equation (21) becomes

$$\sum_{R_r{}^r} s_r^r \, D^{(1)}(R_r^r)^* = \sum_{R_r{}^r} D^{(\bar{\mu})} R_r^r \, D^{(1)}(R_r^r)^* = 0.$$

Thus, the corollary is proved.

Corollary 3 is illustrated in (30) and (31) where the excitation is coupled to the state variables associated with $D^{(1)}(R)$. (Matrix $D(R)$ is a permutation representation for the example of Fig. 2, so that $\bar{\mu} = 1$.)

The applicability of Theorem 6 is extended somewhat by considering the case where the single input is coupled to more than one state variable. For the moment, it is assumed that only two state variables are coupled to the input so that in equation (28)

$$B(t) = h_1(t)e_r + h_2(t)e_i \, , \tag{32}$$

where $h_1(t)$ and $h_2(t)$ are scalars.

Corollary 4: A symmetric network having a single input coupled only to the rth and jth state variables is NCS if and only if a proper value[‡] of μ exists such that for some value of π,

$$\sum_{R_r{}^r} s_r^r \, D^{(\mu)}(R_r^r)^*_{\pi\pi} = 0 \quad \textit{and} \quad \sum_{R_j{}^i} s_i^i \, D^{(\mu)}(R_i^i)^*_{\pi\pi} = 0. \tag{33}$$

Proof: For r and j in the same transitive set and $c_\mu \leqq c_1^p$, the partition of $\alpha^\dagger B$ in equation (29) corresponding to $\bar{A}^{\,\mu}_{\pi}$ can have nonzero terms only from[§]

[‡] A proper value of μ is one for which $D^{(\mu)}(R)$ is contained in $D(R)$.

[§] A normalization factor is not included in the vector $\alpha_{\mu\pi r}$.

$$(\alpha_{\mu\pi r})^{\dagger} B = h_1(t) \sum_{R_r^r} s_r^r D^{(\mu)}(R_r^r)^*_{\pi\pi} + h_2(t) \sum_{R_r^j} s_r^j D^{(\mu)}(R_r^j)^*_{\pi\pi} . \qquad (34)$$

Provided that $h_1(t)$ and $h_2(t)$ are not specially chosen to cause (34) to vanish, Lemma 2 implies that (34) vanishes if and only if

$$\sum_{R_r^r} s_r^r D^{(\mu)}(R_r^r)^*_{\pi\pi} = 0$$

(only one of the sums in (33) need be computed in this case).

For the case where $c_\mu > c_1^n$, the possibility exists that $\alpha_{\mu\pi r}$ and $\alpha_{\mu\pi j}$ are linearly independent. This linear independence also occurs when r and j are in different transitive sets. Thus, for these cases, the possible nonzero terms in the partition of $\alpha^{\dagger}B$ corresponding to $\bar{A}_{\pi}^{\mu}$ in (29) arise from $(\alpha_{\mu\pi r})^{\dagger} h_1(t) e_r$ and from $(\alpha_{\mu\pi j})^{\dagger} h_2(t) e_j$. From Lemma 2, $(\alpha_{\mu\pi k})^{\dagger} e_k$ is zero if and only if

$$\sum_{R_k^k} s_k^k D^{(\mu)}(R_k^k)^*_{\pi\pi} = 0, \qquad k = r, j.$$

Thus, the proof is complete.

The above method may be extended in a fairly obvious manner to treat the case where any number of state variables are coupled to the single input. A separate statement is required for each set of variables in a given transitive set.

At this point, we consider the problem of determining general conditions which guarantee that (21) will or will not be satisfied for some proper value of μ. Thus, the summations for all values of μ need not be computed. A partial solution to this problem is offered in Theorems 7, 8, and 9 below. It is assumed that the single input is coupled only to the rth state variable; the results can be extended to the case of multiple couplings by utilizing the reasoning in Corollary 4 above.

Use is made of the following well-known properties of finite groups[10].

(*i*) The order of a group G which is transitive on k symbols is mk where m is an integer giving the number of group operations which leave any given symbol unchanged.

(*ii*) If a group G is intransitive on k symbols, the symbols may be partitioned into transitive sets M_r , M_s , $\cdots$. If the operations of G are allowed to operate only on symbols in the transitive set M_r (i.e., permutations of symbols not in M_r are simply ignored), G reduces to a new group G_r . The result is that

$$g = g_r g_{\bar{r}} , \qquad (35)$$

where the lower-case letters indicate the orders of the appropriate groups, and $G_{\bar{r}}$ is the invariant subgroup leaving fixed all symbols in M_r .

A general intransitive symmetric network is considered in which the state variables are partitioned into the transitive sets M_r , M_s , $\cdots$. The set M_r contains the rth state variable, x_r , and the number of state variables in M_r is denoted by k_r . The following theorem is a simple application of property (i) above; it guarantees that the rth state variable is left unchanged by only one operation of the network symmetry group, G.

Theorem 7: If G is isomorphic with G_r and if k_r is equal to the order of G, the network is not NCS.

Proof: Since G_r is necessarily transitive on the k_r symbols of M_r , property (i) above implies that $g_r = mk_r$. However, $g_r = g$ since G and G_r are isomorphic. Thus, $g = mk_r$. By hypothesis, k_r equals the order of G; m must be unity. Hence, only one group operation of G leaves the rth state variable invariant, and the NCS property is impossible as shown in Corollary 1 to Theorem 6.

Theorem 8: Let G be an axial point group and let G_r be a proper subgroup of G. The symmetric network with symmetry group G is NCS.

Proof: Since G_r is a proper subgroup of G, equation (35) implies that $g_{\bar{r}}$ is greater than unity. For the axial point groups excluding D_{nh} groups, only the identity and a reflection plane σ_v (a rotation C_2 about a two-fold axis perpendicular to the principal axis may be included instead of a reflection plane) can have an invariant effect on any given state variable. For D_{nh} groups, in addition to E and σ_v , a C_2 operation and a σ_h operation can have an invariant effect on a given state variable. Furthermore, at most only one symmetry plane σ_v (rotation C_2) can leave a given state variable unchanged. Hence $g_{\bar{r}} = 2$, or possibly $g_{\bar{r}} = 4$, for a D_{nh} group. Thus, the subgroup $G_{\bar{r}}$ in property (ii) above is either $\{E, \sigma_v\}$, $\{E, C_2\}$, or $\{E, \sigma_v , C_2 , \sigma_h\}$, and these operations leave invariant all variables in the transitive set M_r .

To show that the networks considered in this theorem are NCS, a proper value of μ is determined for use in (21); we compute c_2 , the number of times $D^{(2)}(R)$ is contained in $D(R)$, using (52). To facilitate the computation of c_2 , the n state variables are partitioned into h_2 transitive sets of two variables each, h_3 transitive sets of three variables each, $\cdots$, h_{k_r} transitive sets of k_r variables each, etc. Thus,

$$n = 2h_2 + 3h_3 + \cdots + h_{k_r}k_r + \cdots . \tag{36}$$

In $D^{(2)}(R)$, all σ_v and C_2 operations are represented by -1 while E and σ_h are represented by $+1$.[12] Hence, c_2 is at least

$$c_2 = \frac{1}{g}\sum_R \chi^{(2)}(R)^*\chi_p(R)$$

$$= \begin{cases} \frac{1}{g}[n - h_{k_r}k_r], & \text{for } C_{nv}, D_n, D_{nd} \text{ symmetry} \\ \text{or} \\ \frac{1}{g}[n - h_{k_r}k_r + h_{k_r}k_r - h_{k_r}k_r], & \text{for } D_{nh} \text{ symmetry.} \end{cases} \tag{37}$$

In general, c_2 is not zero, and (21) is satisfied for $\mu = 2$ since $\{R_r^r\} = \{E, \sigma_v\}$ or $\{E, C_2\}$, or $\{E, \sigma_v, C_2, \sigma_h\}$, and

$$\sum_{R_r{}^r} s_r^r D^{(2)}(R_r^r)^*_{\pi\pi} = 1 - 1 = 0, \quad \text{for } C_{nv}, D_n, \text{ or } D_{nd} \text{ groups}$$

and

$$\sum_{R_r{}^r} s_r^r D^{(2)}(R_r^r)^*_{\pi\pi} = 1 - 1 + 1 - 1 = 0, \quad \text{for } D_{nh} \text{ groups.}$$

Thus, these networks are NCS.

As an example illustrating the use of Theorems 7 and 8, consider the network of Fig. 4 which includes the operations of the symmetry group C_{2v} for this case. A table of irreducible representations of C_{2v} may be found in Fig. 2. There are two transitive sets, namely, $M_1 = \{x_1, x_4\}$ and $M_2 = \{x_2, x_3, x_5, x_6\}$. By allowing the permutations of G to operate only on state variables in M_1, G reduces to

$$G_1 = \{E, \sigma\}$$

where

$$E = (1)(4)$$
$$\sigma = (14).$$

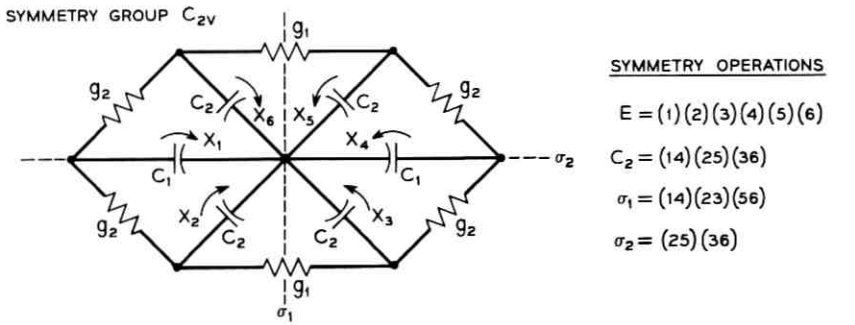


Fig. 4—Network with C_{2v} symmetry and symmetry operations.

By allowing the permutations of G to operate only on state variables in M_2, G reduces to $G_2 \sim G$ (i.e., G_2 and G are isomorphic). Furthermore, the number of state variables in M_2 is equal to the order of G. Thus, by Theorem 7, no symmetry constraints are placed on controllability if the input is coupled to one of the variables in the set M_2. However, from Theorem 8, if the input is coupled either to x_1 or to x_4, the network has the NCS property. The above statements are verified by computing the sum in equation (21) for each case.

Theorem 9: Let G be an axial point group having at least one irreducible representation of dimension two. A network possessing symmetry group G is NCS.

Proof: Let $D^{(\mu)}(R)$ be an irreducible representation of G of dimension two, and let c_μ be the number of times that $D^{(\mu)}(R)$ appears in the monomial representation $D(R)$. Since the character $\chi^{(\mu)}(R)$ of all group operators, excluding E, that can possibly have an invariant effect on any state variable is zero (see tables of irreducible representations in Ref. 12),

$$c_\mu = \frac{1}{g} \sum_R \chi^{(\mu)}(R)^* \chi(R) = \frac{1}{g} \chi^{(\mu)}(E)^* \chi(E) = \frac{1}{g} 2 \cdot n.$$

Hence, $D^{(\mu)}(R)$ is contained in $D(R)$. From the proof of the previous theorem, $\{R_r^r\} = \{E, \sigma_v\}$ or $\{E, C_2\}$ or $\{E, \sigma_v, C_2, \sigma_h\}$. A table of irreducible representations[12] shows that $D^{(\mu)}(R)$ is equivalent to a representation in which

$$D^{(\mu)}(E) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$D^{(\mu)}(\sigma_v) \quad \text{or} \quad D^{(\mu)}(C_2) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$D^{(\mu)}(\sigma_h) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Therefore, equation (21) is satisfied for $\pi = 1$ (or $\pi = 2$ if $s_r^r = -1$), and the network is NCS. The theorem is proved.

An example illustrating Theorem 9 has already been given in the discussion following Theorem 6. The example concerned the network displaying C_{3v} symmetry shown in Fig. 3. Since all C_{nv}, D_n, D_{nd}, and D_{nh} groups of complexity C_{3v} or greater possess two-dimensional irreducible representations, Theorem 9 shows that the general symmetric network displaying axial point group symmetry is NCS.

The single-input case, considered extensively here, assumes an added significance in view of the following definition found in Ref. 14.

Definition 5: A k-input system is said to be *strongly controllable* if it is controllable by each input separately while all others are zero; otherwise it is *weakly controllable.*

From the discussion of the present section, it is observed that a symmetric network is generally only weakly controllable. Thus, several inputs are required to control the state of a symmetric network in general. The results of the present section can be used to determine the number and placement of inputs to insure that the network is not NCS.

The previous results of this section may be applied to the multiple-input case by means of the following.

Theorem 10: The k-input system of equation (28) *with symmetry group* G_N *is NCS if and only if a proper value of* μ *exists such that for some value of* π,

$$\sum_{R_\phi^\phi} s_\phi^\phi \, D^{(\mu)}(R_\phi^\phi)_{\pi\pi}^* = 0, \qquad \phi = r, \cdots, j,$$

where ϕ *is an index denoting all nonzero couplings of the inputs to the state variables in the k columns of* B.

Proof: It follows from Theorem 6 and its Corollary 4 that if the above conditions hold, the submatrix of $\alpha^\dagger B$ in (29) corresponding to $\bar{A}_\pi^\mu$ is identically zero. Thus, the network is noncontrollable due to symmetry constraints. From Theorem 6, the above conditions are also necessary for the NCS property.

The discussion just concluded shows that simple arithmetic computations involving the irreducible representations of the network symmetry group can be used to detect noncontrollability which is due solely to symmetry. For an input coupled to a given state variable, the NCS property is determined completely by those group operations that leave the given state variable unchanged. The interpretation of (21) is obtained from (52), in which the generating matrix $G_r^{(\mu)}$ is obtained by analogy with the projection operation $P_r^{(\mu)}$. If the input is coupled to the rth state variable, then, by Lemma 2, condition (21) is equivalent to the statement that the projection of the input onto the invariant subspace associated with the πth row of $D^{(\mu)}(R)$ is zero.

In the application of Theorem 6, if equation (21) is not satisfied, the network *may* be controllable. The nonsatisfaction of equation (21) amounts to a necessary condition for controllability of a symmetric network. The transformation α obtained using group theory then enables

us to test controllability of several smaller subsystems [equation (29)] rather than that of system (1).

V. OBSERVABILITY OF SYMMETRIC NETWORKS

The concept of observability relates to the degree to which the past state of a system may be determined from knowledge of the system outputs. The following definition may be found in Ref. 13.

Definition 6: The system (1) is said to be *completely observable on an interval* (t_o , t_1) if any initial state x_o at t_o can be determined from knowledge of the system output over (t_o , t_1).

The system (1) is said to be *totally observable on an interval* (t_o , t_1) if it is completely observable on every subinterval of (t_o , t_1).

For networks with sufficiently smooth time variations, observability of the linear time-varying system

$$\begin{aligned} \dot{x} &= A(t)x + B(t)u(t) \\ y(t) &= C(t)x \end{aligned} \tag{38}$$

may be characterized by the observability matrix [13]

$$Q_0 = [S_0 S_1 \cdots S_{n-1}] \tag{39}$$

where

$$S_k = A^t S_{k-1} + \dot{S}_{k-1} , \qquad S_0 = C^t$$

and n is the order of the system.

The following theorem is a paraphrase of Theorem 5 of Ref. 13.

Theorem 11: For the system (38), *assume that* $A(t)$ *and* $C(t)$ *and their first* $n - 2$ *and* $n - 1$ *derivatives, respectively, are continuous functions. System* (38) *is totally observable on the interval* (t_o , t_1) *if and only if* Q_0 *does not have rank less than* n *on any subinterval of* (t_o , t_1).

The results of this section are completely analogous to those obtained for controllability. Hence, only some theorems will be presented; their proofs follow exactly from their counterparts in the previous sections.

Lemma 4: The system described in partitioned form by

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} &= \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + Bu(t) \\ y(t) &= [C_1 \quad 0]\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \end{aligned} \tag{40}$$

is unobservable (i.e., not completely observable).

Definition 7: A symmetric network is said to be NOS (possess the NOS property) if it is nonobservable because of symmetry alone.

To examine the NOS property, we reduce $A(t)$ to block-diagonal form; for $Z = \alpha^{\dagger}x$, equation (38) becomes

$$\frac{d}{dt}\begin{bmatrix} Z_1^1 \\ \vdots \\ Z_\pi^\mu \\ \vdots \\ Z_{l_\beta}^\beta \end{bmatrix} = \begin{bmatrix} \bar{A}_1^1 & & & & \\ & \ddots & & 0 & \\ & & \bar{A}_\pi^\mu & & \\ & 0 & & \ddots & \\ & & & & \bar{A}_{l_\beta}^\beta \end{bmatrix} \begin{bmatrix} Z_1^1 \\ \vdots \\ Z_\pi^\mu \\ \vdots \\ Z_{l_\beta}^\beta \end{bmatrix} + \alpha^{\dagger}Bu(t) \tag{41}$$

$$y(t) = C(t)\alpha \begin{bmatrix} Z_1^1 \\ \vdots \\ Z_\pi^\mu \\ \vdots \\ Z_{l_\beta}^\beta \end{bmatrix}.$$

Hence, if a submatrix of $C(t)\alpha$ corresponding to Z_π^μ is zero, these variables will not be observed in the output. In analogy to Section IV, first consider the output in (38) to be a function of a single state variable, x_r. Thus, $C(t)$ is $[e_r]^t$, and $C(t)\alpha$ in (39) is the rth row of α.

Theorem 12: A symmetric time-varying network whose output is a function of the rth state variable only is NOS if and only if there exists a proper value of μ and a value of π such that

$$\sum_{R_r^r} s_r^r \, D^{(\mu)}(R_r^r)^* = 0. \tag{42}$$

Corollaries 1–4 of Theorem 6 carry through directly for this case with slight and obvious changes of wording (i.e., "NOS" replaces "NCS", etc.), and they are not repeated here. Of course the other results of Section IV follow for observability with slight modification of the wording.

VI. APPEARANCE OF BASIS FUNCTIONS IN $\Phi(t, \tau)$

It may happen that one or more basis functions of the normal form differential equation do not appear in the expression for component $\phi_{ij}(t, \tau)$ of the state-transition matrix $\Phi(t, \tau)$. In the case of fixed systems, this condition corresponds to one in which certain modes are cancelled.

The present section investigates the use of symmetry in predicting such cancellation of basis functions in a symmetric network.

In Section II, it was shown that the monomial representation $D(R)$ of the group G_N commutes with $\Phi(t, \tau)$ for a symmetric network. Hence, $\Phi(t, \tau)$ is reduced to block-diagonal form by the same transformation α which reduces $D(R)$ and $A(t)$, and we can write

$$\Phi(t, \tau) = \alpha\bar{\Phi}(t, \tau)\alpha^{\dagger}$$

$$= \alpha \begin{bmatrix} \bar{\phi}_1^1 & & & & & & \\ & \ddots & & & & & \\ & & \bar{\phi}_1^{\mu} & & & 0 & \\ & & & \ddots & & & \\ & & & & \bar{\phi}_{\pi}^{\mu} & & \\ & & & & & \ddots & \\ & & 0 & & & \bar{\phi}_{l_\mu}^{\mu} & \\ & & & & & & \ddots \\ & & & & & & \bar{\phi}_{l_\beta}^{\beta} \end{bmatrix} \alpha^{\dagger}, \tag{43}$$

where $\bar{\phi}_{\pi}^{\mu}$ is a $c_\mu \times c_\mu$ matrix which does not depend on π.‡ Hence,

$$\phi_{ij}(t, \tau) = [\alpha_{111}^{i}, \cdots, \alpha_{\mu\pi 1}^{i}, \cdots, \alpha_{\mu\pi c_\mu}^{i}, \cdots, \alpha_{\beta l_\beta c_\beta}^{i}]$$

$$\times \begin{bmatrix} \bar{\phi}_1^1 & & & & & & \\ & \ddots & & & & & \\ & & \bar{\phi}_1^{\mu} & & & 0 & \\ & & & \ddots & & & \\ & & & & \bar{\phi}_{\pi}^{\mu} & & \\ & & & & & \ddots & \\ & & 0 & & & \bar{\phi}_{l_\mu}^{\mu} & \\ & & & & & & \ddots \\ & & & & & & \bar{\phi}_{l_\beta}^{\beta} \end{bmatrix} \begin{bmatrix} \alpha_{111}^{j} \\ \vdots \\ \alpha_{\mu\pi 1}^{j} \\ \vdots \\ \alpha_{\mu\pi c_\mu}^{j} \\ \vdots \\ \alpha_{\beta l_\beta c_\beta}^{j} \end{bmatrix}^{*} \tag{44}$$

It is evident from (44) that basis functions associated with $D^{(\mu)}(R)$

‡ In Ref. 2 it is shown that submatrices $\bar{A}_{\pi}{}^{\mu}$ of the A-matrix of (29) do not depend on π. Hence, submatrices $\bar{\phi}_{\pi}{}^{\mu}$ given above are independent of π.

will appear in $\phi_{ij}(t, \tau)$ if

$$K_\mu = [\alpha^i_{\mu 11}, \cdots, \alpha^i_{\mu 1 c_\mu}, \cdots, \alpha^i_{\mu \pi 1}, \cdots, \alpha^i_{\mu \pi c_\mu}, \cdots, \alpha^i_{\mu l_\mu 1}, \cdots, \alpha^i_{\mu l_\mu c_\mu}]$$

$$\times \begin{bmatrix} \bar{\phi}^\mu_1 & & & \\ & \ddots & & \\ & & \bar{\phi}^\mu_\pi & \\ & & & \ddots \\ & & & \bar{\phi}^\mu_{l_\mu} \end{bmatrix} \begin{bmatrix} \alpha^j_{\mu 11} \\ \vdots \\ \alpha^j_{\mu 1 c_\mu} \\ \vdots \\ \alpha^j_{\mu \pi 1} \\ \vdots \\ \alpha^j_{\mu \pi c_\mu} \\ \vdots \\ \alpha^j_{\mu l_\mu 1} \\ \vdots \\ \alpha^j_{\mu l_\mu c_\mu} \end{bmatrix}^* \not\equiv 0, \qquad (45)$$

where K_μ is used to denote all terms corresponding to $D^{(\mu)}(R)$ in (44). From Lemma 1, only one of the c_μ terms $\alpha^i_{\mu \pi a}$ (or $\alpha^j_{\mu \pi a}$) can possibly be nonzero, $a = 1, \cdots, c_\mu$; $\pi = 1, \cdots, l_\mu$. Thus, using the results of Section II and excluding a scale factor, we have

$$K_\mu = (\bar{\phi}^\mu_\pi)_{dd} \sum_{\pi=1}^{l_\mu} [\sum_{R_i{}^i} s^i_i \, D^{(\mu)}(R^i_i)^*_{\pi\pi}][\sum_{R_i{}^j} s^j_i \, D^{(\mu)}(R^j_i)_{\pi\pi}] \qquad (46a)$$

if i and j are in the same transitive set, or

$$K_\mu = (\bar{\phi}^\mu_\pi)_{de} \sum_{\pi=1}^{l_\mu} [\sum_{R_i{}^i} s^i_i \, D^{(\mu)}(R^i_i)^*_{\pi\pi}][\sum_{R_j{}^j} s^j_i \, D^{(\mu)}(R^j_i)_{\pi\pi}] \qquad (46b)$$

if i and j are in different transitive sets.‡ In (46), $(\bar{\phi}^\mu_\pi)_{dd}$ and $(\bar{\phi}^\mu_\pi)_{de}$ are components of the matrix $\bar{\phi}^\mu_\pi$, and are removed from the π-summation because they are independent of π. Thus, the following theorem has been established.

Theorem 13: Basis functions corresponding to $D^{(\mu)}(R)$ will not appear in $\phi_{ij}(t, \tau)$ as a result of the symmetry if and only if

$$\sum_\pi [\sum_{R_i{}^i} s^i_i \, D^{(\mu)}(R^i_i)^*_{\pi\pi}][\sum_{R_i{}^j} s^j_i \, D^{(\mu)}(R^j_i)_{\pi\pi}] = 0. \qquad (47)$$

‡ If $c_\mu > c_1{}^p$, equation (46b) may apply even if i and j are in the same transitive set. See Lemma 2.

The following corollaries to the above theorem may be established; some of the corollaries are similar to those following Theorem 6.

Corollary 1: All basis functions corresponding to $D^{(\bar{\mu})}(R)$ appear in every $\phi_{ij}(t, \tau)$ (if all the s_i^j equal unity, $\bar{\mu} = 1$).

Corollary 2: If there is just one group operation taking the ith state variable into the jth state variable, basis functions corresponding to $D^{(\mu)}(R)$ appear in $\phi_{ij}(t, \tau)$ if $\chi^{(\mu)}(R_i^j) \neq 0$, where $\chi^{(\mu)}(R)$ denotes the trace of $D^{(\mu)}(R)$.

Proof: The hypothesis requires that only one group operator leave the ith state variable invariant.[10] This operator must be the identity, and $D^{(\mu)}(E)_{\pi\pi} = 1$ for all values of μ. Hence, equation (46a) becomes (s_τ^r equals unity for the identity operation)

$$K_\mu = (\bar{\phi}_\tau^\mu)_{dd} \sum_\pi D^{(\mu)}(R_i^j)_{\pi\pi} = (\bar{\phi}_\pi^\mu)_{dd} \chi^{(\mu)}(R_i^j). \tag{48}$$

Thus, basis functions corresponding to $D^{(\mu)}(R)$ appear in $\phi_{ij}(t, \tau)$ if $\chi^{(\mu)}(R_i^j) \neq 0$.

Corollary 3: If no group operation transforms the ith state variable into the jth state variable, the types of basis functions which appear in $\phi_{ij}(t, \tau)$ are those which are common to $\phi_{ii}(t, \tau)$ and $\phi_{jj}(t, \tau)$.

The proof of Corollary 3 is a straightforward application of Theorem 13.

Corollary 4: If the kth state variable is invariant under all the group operators, the only basis functions appearing in $\phi_{ik}(t, \tau)$ are those which correspond to $D^{(\bar{\mu})}(R)$.

With regard to Section IV (Section V) the following statement can be made about noncontrollability (nonobservability) due to symmetry.

Theorem 14: A symmetric time-variable linear network with a single input coupled only to the rth state variable (output proportional only to the rth state variable) is NCS (NOS) if there exists a proper value of μ such that basis functions corresponding to $D^{(\mu)}(R)$ do not appear in $\phi_{rr}(t, \tau)$.

Proof: If the basis functions corresponding to $D^{(\mu)}(R)$ do not appear in $\phi_{rr}(t, \tau)$, Theorem 13 shows that

$$\sum_\pi [\sum_{R_r^r} s_r^r D^{(\mu)}(R_r^r)_{\pi\pi}^*][\sum_{R_r^r} s_r^r D^{(\mu)}(R_r^r)_{\pi\pi}] = 0. \tag{49}$$

Since the squared magnitude of the bracketed term appears in the

above equation, it is necessary that

$$\sum_{R_r{}^r} s_r^r \, D^{(\mu)}(R_r^r)^*_{\pi\pi} = 0. \tag{50}$$

Under the conditions of the present theorem, the network is NCS (NOS) by Theorem 6 (Theorem 12).

VII. CONCLUSION

We have presented a unified treatment of linear time-variable networks displaying arbitrary geometrical symmetries by incorporating group theory into the analysis scheme. Symmetric networks have their elements arranged so that certain permutations of the network edges result in a configuration identical with the original. The complete set of such permutations constitutes a group G_S, the symmetry group of the network structure. A group G_N of monomial matrices may then be determined, and it was shown that these matrices commute with the A-matrix and the state transition matrix of the normal form equation. The commutativity result establishes a basic connection between group theory and the network analysis problem and allows group theoretic methods to be employed in the study of networks with arbitrary symmetries. The group G_N may be a proper subgroup of G_S, since G_N contains those operations of G_S which do not permute edges in the tree with those in the cotree.

Group representation theory makes it possible to obtain information about properties of the network differential equations without writing or solving them. For the case of a network with a single input coupled to only one of the state variables, an extremely simple arithmetic condition is derived which determines whether symmetry alone causes the network to be noncontrollable. The condition involves only those group operators which transform the state variable in question into itself. It is equivalent to the algebraic statement that the projection of the input vector onto a subspace associated with an irreducible representation of the group be zero. The results allow a determination by inspection of linear combinations of the original state variables which result in noncontrollable variables. It was demonstrated that a network with axial point group symmetry is always noncontrollable if its symmetry group possesses an irreducible matrix representation of dimension two. This result agrees with intuition in that the network will be noncontrollable if the symmetry is high enough. Thus, networks with axial point group symmetry are generally noncontrollable because of symmetry alone. The case where the input is coupled to more than

one state variable and the multiple input case were also treated. Furthermore, dual results were stated for network observability.

By utilizing the symmetry, a transformation may be constructed which transforms the A-matrix into block-diagonal form. The original differential equation is thereby resolved into several differential equations of relatively low order. Hence, there results an appreciable economy of effort in obtaining solutions for symmetric networks.

APPENDIX A

This appendix provides some basic definitions and results from the abstract theory of finite groups and the corresponding representation theory. A more detailed treatment of concepts mentioned here may be found in Refs. 10 and 15, 16.

Definition 8: A set of elements $G = \{A_1, A_2, A_3, \cdots\}$ is a *group* if

(*i*) for $A_p, A_q \in G$, $A_pA_q \in G$ (closure)
(*ii*) for $A_p, A_q, A_r \in G$, $(A_pA_q)A_r = A_p(A_qA_r)$ (associativity)
(*iii*) there exists $E \in G$ such that $A_pE = EA_p = A_p$ (identity element)
(*iv*) there exists $A_p^{-1} \in G$ such that $A_p^{-1}A_p = A_pA_p^{-1} = E$ (inverse element).

If the number of distinct elements of the group is finite, the group is said to be a *finite group*; the number of distinct elements in a finite group is called its *order*.

Definition 9: Two groups G and G' are said to be *isomorphic* if there exists a one-to-one correspondence (denoted $\sim$) between their elements such that products correspond to products, i.e., if $A \sim A'$ and $B \sim B'$, then $AB \sim A'B'$.

Definition 10: If among the elements of a group G there exists a subset H of elements satisfying the definition of a group, then H is said to be subgroup of the group G.

Consider a subgroup H of G, where the order of H is h while that of G is g. Now consider any element x_1 of G which is not contained in H, and form the product x_1H. That is, multiply every element of H on the left by x_1. Since x_1 is not in H, the resulting set of elements is different from H (H contains the identity, E, and hence x_1H contains x_1). The set of elements x_1H is called a *left coset* of G with respect to the subgroup H. A coset is not a subgroup since it does not contain the identity (H does not contain x_1^{-1}). If there are any elements of G not contained in H or x_1H, choose one of these elements, x_2 say, and form the coset x_2H.

Continue in this manner until all elements of G are exhausted. Thus, a partition has been effected of the group G into left cosets with respect to the subgroup H.

$$G = H, x_1H, x_2H, \cdots, x_{l-1}H.$$

A similar partition could be effected using right cosets, defined analogously. The quantity $l = g/h$ is an integer[10] called the *index* of H in G.

Definition 11: If H is a subgroup of G and $x \in G$, then $x^{-1}Hx$ is called a *conjugate subgroup* of H in G. If H coincides with all its conjugates (i.e., $x^{-1}Hx = H$, for all $x \in G$), then H is said to be an *invariant subgroup.*

Consider the set of n symbols $a_1, a_2, \cdots, a_n$. A rearrangement of these same symbols into the order $b_1, b_2, \cdots, b_n$ is called a *permutation.* Here, the symbol a_1 is replaced by b_1, a_2 by b_2, etc. One way of indicating this permutation is

$$\begin{pmatrix} a_1 a_2 \cdots a_n \\ b_1 b_2 \cdots b_n \end{pmatrix},$$

so that each symbol on the upper line is replaced by the symbol appearing below it. A more convenient notation is to write the permutation as a set of cycles. To do so, begin by choosing any symbol on the top line, say q, writing it followed by the symbol r on the bottom line which replaces it. Now find where r appears on the upper line, obtain the symbol which replaces r and write that. This procedure is continued until we arrive at the symbol which is replaced by q, the first symbol in the cycle. This step completes a cycle. If any symbols remain unused, a new cycle is written by choosing as the leading symbol any one of those which did not appear in the first cycle. This procedure is continued until all symbols are exhausted. The cycles are enclosed in parentheses. Thus,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 3 & 5 & 1 & 2 & 6 \end{pmatrix} = (14)(235)(6) = (14)(235),$$

where 1 is replaced by 4, 4 is replaced by 1, etc. Cycles composed of a single symbol, such as (6), need not be written. In examples of symmetric networks, the cycle notation may be used to make easy the identification of matrices $D(R)$ (Section II) by inspection.

Some important results in group representation theory are presented next.

Definition 12: A group of matrices $D(\cdot)$ is said to form a *representation* of a group $G = \{E, A, \cdots, R, \cdots\}$ if there exists a correspondence

(denoted $\sim$) between the matrices and the group elements such that products correspond to products, i.e., if $R_1 \sim D(R_1)$ and $R_2 \sim D(R_2)$, then $(R_1R_2) \sim D(R_1)D(R_2) = D(R_1R_2)$.
An example of a representation is the so-called *totally symmetric* representation in which each group element is represented by the scalar quantity unity.

Definition 13: A representation is said to be *reducible* if it can be converted to block-diagonal form via a similarity transformation; i.e.,

$$D(R) = \begin{bmatrix} D_1(R) & 0 \\ 0 & D_2(R) \end{bmatrix}$$

is reducible. Otherwise, it is said to be irreducible. For a finite group, there can be only a finite number of distinct irreducible representations, and the irreducible representations may generally be specified to within a similarity transformation. The irreducible representations of a finite group satisfy the following important orthogonality relation.[15]

$$\sum_R D^{(i)}(R)^*_{\alpha\beta} D^{(j)}(R)_{\alpha'\beta'} = \frac{g}{l_i} \delta_{ij} \delta_{\alpha\alpha'} \delta_{\beta\beta'}, \tag{51}$$

where $D^{(k)}(R)_{\alpha\beta}$ denotes the $\alpha\beta$-element of irreducible representation $D^{(k)}(R)$, l_i denotes the dimension of $D^{(i)}(R)$, g denotes the order of the group, $\delta_{\mu\nu}$ is Kronecker's delta, and asterisk denotes the complex conjugate.

If a reducible representation $D(R)$ is reduced to block-diagonal form, the nonzero submatrices along the diagonal will be the irreducible representations of the group.[15] Some irreducible representations may appear more than once (i.e., several nonzero blocks may be identical) in $D(R)$ while others may not appear at all. The number of times that $D^{(k)}(R)$ appears in $D(R)$ is denoted by c_k and is given by[15]

$$c_k = \frac{1}{g} \sum_R \chi^{(k)}(R)^* \chi(R), \tag{52}$$

where $\chi(R)$ is the trace of $D(R)$ and $\chi^{(k)}(R)$ is the trace of $D^{(k)}(R)$.

A very brief account is now given of so-called axial point groups. Some important statements regarding networks with axial point group symmetry may be found in Theorems 7, 8, and 9. For a more complete treatment of these groups, see Ref. 17.

A point group is one whose symmetry operations leave fixed a point at the center of symmetry. Some symmetry operations contained in these groups are described in the following five definitions.

Definition 14: The *identity* is the trivial operation which does not transform the object at all. It is denoted by the letter E.

Definition 15: A rotation operation by $2\pi/n$ radians about an axis is denoted by C_n where $2\pi/n$ is the smallest angle for which the object may be rotated invariantly about this axis. The axis is said to be an *n-fold rotation axis.*

Definition 16: A reflection operation in a plane of symmetry is labelled σ. If the plane of symmetry is perpendicular to the principal rotation axis of symmetry, it is labelled σ_h ; if it contains the principal axis, it is labelled either σ_v or σ_d .

Definition 17: The *rotation-reflection* operation S_n is a compound operation consisting of a rotation by $2\pi/n$ radians about an axis followed by a reflection in a plane perpendicular to the axis. Thus, $S_n = \sigma_h C_n$.

Definition 18: The *inversion*, denoted by i, is a reflection in the center of symmetry.

The distinguishing characteristic of axial point groups is their single n-fold axis of symmetry ($n > 2$), called the principal symmetry axis. A diagram, called an equivalent point diagram, often used to visualize the operations of an axial point group, is described below. The number of points in the diagram is equal to the order of the group;[12] the points transform into one another under the group operations. In the equivalent point diagram, a plus, +, and circle, ○, denote points above and below the plane of the paper, respectively. Reflection planes not in the plane of the paper are indicated by dotted lines while rotation axes are indicated by solid lines marked with one of the symbols ○, △, □, etc. to indicate a two-fold, three-fold, four-fold axis, etc. Reflection in the plane of the paper is indicated as a solid circle in the point diagram. In the equivalent point diagram, the principal symmetry axis is assumed to be perpendicular to the plane of the paper so that reflection in that plane is σ_h . See Fig. 5 for equivalent point diagrams of all the axial point groups mentioned below.

C_n *groups* have one n-fold rotation axis. The group operations consist of the rotations C_n^r of the object by $(2\pi r)/n$ radians ($r = 1, 2, \cdots, n$). These groups are cyclic.

S_n *groups* have one n-fold rotation-reflection axis.

C_{nv} *groups* have a symmetry axis C_n and n symmetry planes σ_v .

C_{nh} *groups* have a symmetry axis C_n and one symmetry plane σ_h perpendicular to it.

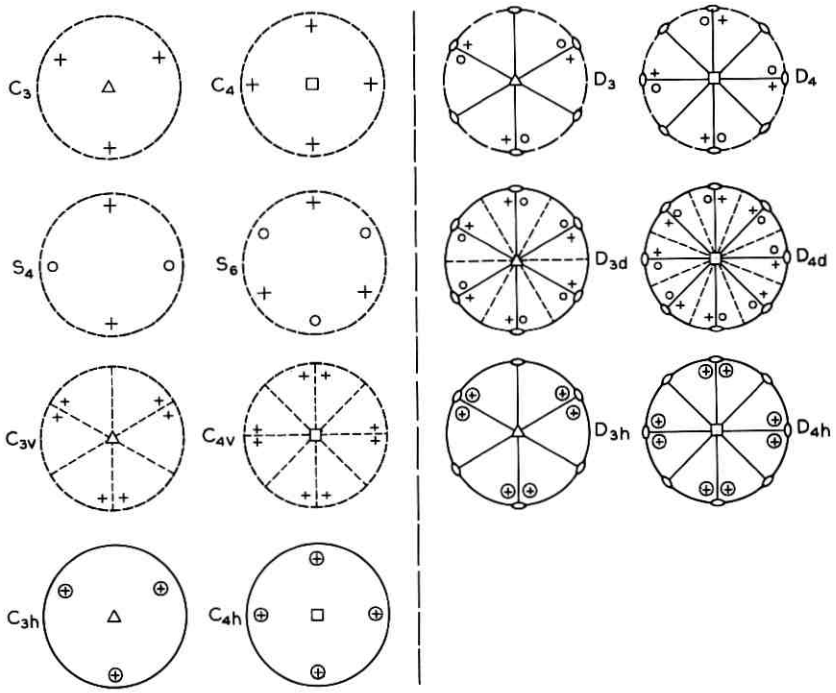


Fig. 5—Some axial point groups and their equivalent point diagrams.

D_n *groups* have an n-fold rotation axis and n two-fold rotation axes perpendicular to the principal axis. The angle between two adjacent two-fold axes is π/n radians.

D_{nd} *groups* contain all the symmetries of D_n and in addition contain n vertical symmetry planes σ_d which contain the principal axis and bisect the angles between the two-fold axes.

D_{nh} *groups* contain all the symmetries of D_n and in addition contain the symmetry plane σ_h perpendicular to the principal axis. These symmetries imply the existence of n symmetry planes σ_v containing both the principal axis and a two-fold rotation axis.

APPENDIX B

For a given reducible representation $D(R)$ of a group G_N, it is possible to construct a unitary matrix α such that the transformed representation[‡] $\alpha^\dagger D(R)\alpha$ is in block-diagonal form. It is next shown how α is constructed.

[‡] The complex conjugate transpose of matrix α is denoted by $\alpha^\dagger$.

Let $D^{(i)}(R)$ be an irreducible representation contained in the reducible representation $D(R)$ of the group G_N .

Definition 19: A set of k vectors $v_1^{(i)}, v_2^{(i)}, \cdots, v_k^{(i)}$ is said to form a *basis for an irreducible representation* $D^{(i)}(R)$ of dimension k if the effect of all group operators on these vectors is to produce a vector which is a linear combination of those already in the set. The set of vectors is said to *transform according to* $D^{(i)}(R)$.

Definition 20: The vector $v_x^{(i)}$ is said to *belong to* (or transform according to) the xth row of the irreducible representation $D^{(i)}(R)$ if it satisfies

$$\sum_R D^{(i)}(R)^*_{xx} D(R)v_x^{(i)} = \frac{g_N}{k} v_x^{(i)}, \tag{53}$$

where g_N is the order of G_N . The other vectors in the basis belong to other rows of $D^{(i)}(R)$ and are called *partners* of $v_x^{(i)}$. They satisfy

$$\frac{g_N}{k} v_m^{(i)} = \sum_R D^{(i)}(R)^*_{mx} D(R)v_x^{(i)}. \tag{54}$$

Let P_R be the operator which denotes the effect of operating on a vector with group operation R. Form the operator

$$P_k^{(j)} = \sum_R D^{(j)}(R)^*_{kk} P_R . \tag{55}$$

$P_k^{(j)}$ has the important property that its effect on any arbitrary vector v is to produce the component vector (which may be zero) which belongs to the kth row of $D^{(j)}(R)$.[17] Hence, $P_k^{(j)}$ is a projection operation.

The transformation α which places a reducible representation $D(R)$ in a block diagonal form may now be constructed. Let c_p be the number of times the irreducible representation $D^{(p)}(R)$ of dimension l_p appears in $D(R)$. Form the $n \times n$ generating matrix $G_\pi^{(p)}$ by analogy with the projection operator of equation (55), so that

$$G_\pi^{(p)} = \sum_R D^{(p)}(R)^*_{\pi\pi} D(R)I, \tag{56}$$

where I is the unit matrix. Some of the column vectors of $G_\pi^{(p)}$ may be zero and several may be identical; the number of linearly independent vectors among the columns of $G_\pi^{(p)}$ is c_p ,[11] and each of these c_p vectors belongs to the πth row of $D^{(p)}(R)$. They are orthogonal and may be normalized to unity. Following Kerns' notation,[11] these c_p column vectors are labelled $\alpha_{p\pi 1}, \cdots, \alpha_{p\pi a}, \cdots, \alpha_{p\pi c_p}$, and are used as c_p column vectors of the matrix α. For every one of the vectors $\alpha_{p\pi a}$, $l_p - 1$ partner vectors must be constructed. The partner vectors are denoted by $\alpha_{p\mu a}$ where $\mu = 1, \cdots \pi - 1, \pi + 1, \cdots, l_p$ and $a = 1, \cdots, c_p$,

and may be calculated as [using equation (54)]

$$\alpha_{p\mu a} = [\sum_{R} D^{(p)}(R)^{*}_{\mu\pi} D(R)]\alpha_{p\pi a} . \tag{57}$$

The index p runs over all distinct irreducible representations of the symmetry group G_N . Thus, if a table of irreducible representations is available, the matrix α may be computed relatively easily, and has the form

$$\alpha = [\underbrace{\alpha_{111}, \cdots, \alpha_{11c_1}}_{D^{(1)}(R)}, \underbrace{\alpha_{211}, \cdots, \alpha_{2l_2c_2}}_{D^{(2)}(R)}, \cdots], \tag{58}$$

where the columns of α are shown associated with the appropriate irreducible representation in the above equation.

REFERENCES

1. Rubin, H., "Symmetric Linear Time-Variable Networks and the Theory of Finite Groups," Tech. Report No. 117, Systems Research Group, Dept. of Electrical Engineering, Columbia University, June 1970.
2. Bashkow, T. R., "The *A* Matrix, New Network Description," IRE Trans. Circuit Theory, *CT-4*, (September 1957), pp. 117–119.
3. Kinariwala, B. K., "Analysis of Time-Varying Networks," IRE Conv. Rec., *9*, Part IV, 1961, pp. 268–276.
4. Bryant, P. R., "The Explicit Form of Bashkow's *A* Matrix," IRE Trans. Circuit Theory, *CT-9*, (September 1962), pp. 303–306.
5. Stern, T. E., *Theory of Nonlinear Networks and Systems*, Reading, Massachusetts: Addison Wesley Publishing Co., Inc., 1965.
6. Kim, W. H., and Meadows, H. E., *Modern Network Analysis*, New York: John Wiley and Sons, 1971.
7. Kim, W. H., and Chien, R. T. W., *Topological Analysis and Synthesis of Communication Networks*, New York: Columbia University Press, 1962.
8. Erugin, N. P., *Linear Systems of Ordinary Differential Equations*, New York: Academic Press, 1966.
9. Rubin, H., "Symmetric Basis Functions for Linear Time-Variable Networks With Arbitrary Symmetry Groups," IEEE Trans. Circuit Theory, *CT-18*, No. 5 (September 1971), pp. 547–549.
10. Burnside, W., *Theory of Groups of Finite Order*, New York: Dover Publications, Inc., 1955.
11. Kerns, D. M., "The Analysis of Symmetrical Waveguide Junctions," Journal of Research of the National Bureau of Standards, *46*, (April 1951), pp. 267–282.
12. Auld, B. A., "Applications of Group Theory in the Study of Symmetrical Waveguide Junctions," Stanford University, Stanford, California, MLR-157, March 1952.
13. Silverman, L. M., and Meadows, H. E., "Controllability and Observability in Time-Variable Linear Systems," J. SIAM Control, *5*, No. 1 (February 1967), pp. 64–73.
14. Kreindler, E., and Sarachik, P. E., "On the Concepts of Controllability and Observability of Linear Systems," IEEE Trans. Aut. Control, *AC-9*, (April 1964), pp. 129–136.
15. Tinkham, M., *Group Theory and Quantum Mechanics*, New York: McGraw-Hill Co., Inc., 1964.
16. Carmichael, R. D., *Introduction to the Theory of Groups of Finite Order*, New York: Dover Publications, Inc., 1956.
17. Shonland, D. S., *Molecular Symmetry*, London: D. Van Nostrand Co., Ltd., 1965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Controlled Response of a Ceramic Microphone

By R. E. NICKELL and D. C. STICKLER

(Manuscript received May 17, 1971)

One of the electromechanical transducer candidates for the electronic telephone is a bilamellar piezoelectric ceramic. In order to meet the design template for transducer response in the acoustic band, 0.3 kHz–3.0 kHz, a controlled resonant condition must be introduced at the upper end of the spectrum.*

An analytical program, consisting of three complementary parts, was carried out in order to understand the phenomenology of the transducer/support system response to acoustic loading. The three parts are: (i) a simple direct variational model, used to generate parametric design information; (ii) an exact solution with a lumped mechanical model of the support structure, used to evaluate the effect of using different rubber materials in relation to the design goal; and (iii) a finite element modal survey of the system, used to determine the necessary design modifications and to expose deficiencies in the previous models.

I. INTRODUCTION

Several alternative designs are under consideration as transducer elements for the electronic telephone.[1] One of the leading candidates is a bilamellar piezoelectric ceramic plate consisting of two thin circular ceramic wafers that are electroded on both surfaces and cemented together.[2] The disks are joined with opposing polarity so that the flexural response of the assembly to applied acoustic loading results in additive voltage output.

The design objectives for the transducer response as a function of driving frequency are shown in Fig. 1.[3] The cross-hatched areas indicate the template within which the response should fall. Important characteristics of this template are: (*i*) the electrical output rolls off below 300 Hz in order to exclude low-frequency room noise; (*ii*) output is

* It has since been decided to eliminate rubber from the design and to replace the bilamellar transducer by a metal/ceramic combination.

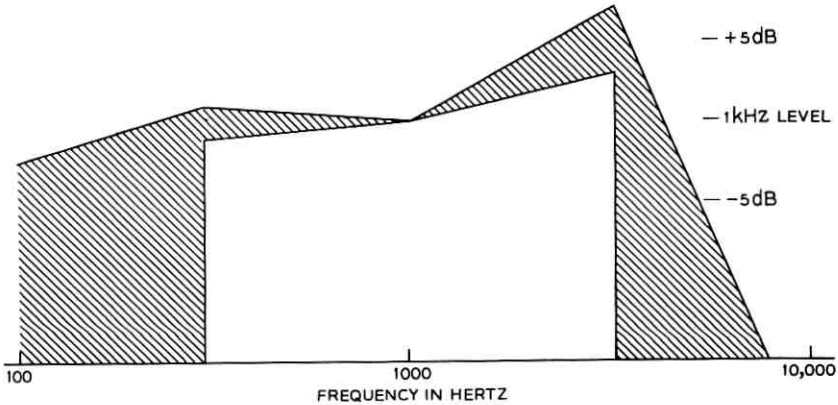


Fig. 1—Microphone response design objective.

relatively flat between 300 and 1000 Hz; (*iii*) output increases by approximately 7 dB between 1000 and 3000 Hz in order to compensate for transmission loop losses and to improve speech recognition; and (*iv*) output rolls off rapidly at frequencies above 3000 Hz in order to eliminate crosstalk and other high-frequency noise. The object of this investigation is to determine the extent to which transducer support damping can be used to achieve these characteristics while maximizing the microphone sensitivity.

The basic approach is to design a bilamellar structure with a fundamental flexural resonance near 3000 Hz—this will guarantee flat response up to frequencies just below the resonance—while providing a support configuration which will permit shaping the response around the resonant peak. In addition, this shaping should include the suppression of resonant response at higher frequencies. The shaping of the response curve at frequencies below 300 Hz is considered to be a manageable problem.

One support configuration that has been tried is shown in Fig. 2. The ceramic disks are mounted between soft rubber "O"-rings that are held in place by a relatively rigid housing. Such a design has one serious drawback—the lack of stability of the transducer response with respect to slight changes in rubber precompression. This sensitivity is due to the large contact area increase, and corresponding increase in support stiffness, as a function of slight changes in precompression. Large excursions in support stiffness will affect the location of the fundamental resonance and, thereby, distort the response.

A more dimensionally stable configuration is shown in Fig. 3. Conical

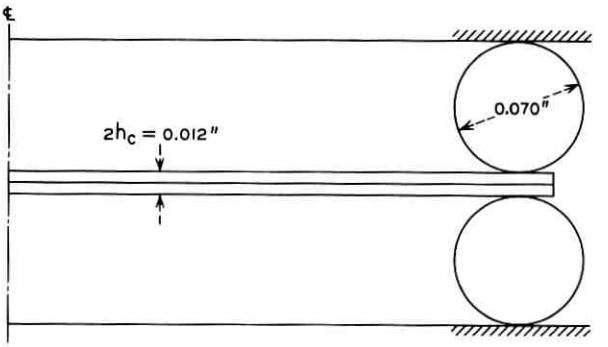


Fig. 2—"O"-ring support configuration.

rubber washers are placed above and below the disk and have flat surfaces in contact with the ceramic, insuring relatively constant stiffness with respect to precompression. The stiffness of the washer is primarily dependent on the thickness and height of the cross section, with a secondary dependence on the cone angle, shown as 25 degrees in Fig. 3. This design can be easily modified in order to achieve the response objectives by making suitable adjustments in these parameters.

While the concern here is with controlled response through support damping, other concepts, such as the addition of acoustic elements to the design (acoustic mass, compliance, and resistance) or constrained layer damping, could be considered.[3] Economic constraints are paramount in deciding the most feasible concept, however, so that manufacturability, material availability, and unit cost are vital ingredients.

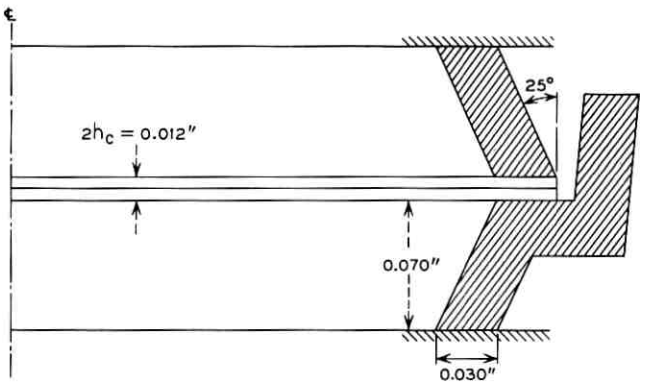


Fig. 3—Conical washer configuration.

In the following sections, an analytical effort that consists of three mutually complementary elements is described. First, a direct variational procedure is used to calculate, approximately, the first two resonant frequencies of the ceramic and its support system. The attraction of these approximate methods is their simplicity of expression and the consequent ease in assessing system trade-offs. The second phase deals with the washer as a lumped mechanical network that acts at an "effective support radius" and is accounted for by a boundary condition on the shear resultant at this location. The exact solution for the forced response of the transducer, as a function of driving frequency, is then found. The third phase consists of a resonant frequency and modal shape survey of the complete structure by using an elastic, axisymmetric, dynamic finite element code. These finite element results provide a check on previous calculations and give more detailed information on the motion executed by the support system.

From this description the role of analysis is seen to have several facets: (*i*) as a design guide (to identify parameters and to check initial experiments); (*ii*) as a key to understanding the phenomenology; (*iii*) to assess hard designs through detailed analysis; and (*iv*) to provide guidelines for future designs.

II. RUBBER CHARACTERISTICS

Before proceeding with the analytical details, a few comments on the thermomechanical properties of viscoelastic support materials are in order. From the transducer response template in Fig. 1, the primary information needed is the complex viscoelastic moduli over the frequency range 100 Hz–10,000 Hz. In addition, since the transducer must have stable response characteristics with respect to temperature changes down to about −40°C and up to about +50°C, the effect of temperature on these moduli must be known.

For these analyses, the candidate polymers were assumed to be isotropic, thus reducing the number of moduli about which knowledge is required down to two (e.g., the shear and bulk moduli). Also, it was assumed that the materials were nearly incompressible over the frequency and temperature ranges of interest (the bulk modulus much larger than the shear modulus), reducing the number down to one.[4] For example, if the complex extensional modulus is known, the complex shear modulus is found by dividing by three. It suffices, therefore, to know the storage modulus, in either extension or shear, and the loss tangent over the acoustic frequency range at temperatures in the environmental range.

An additional simplification is possible by assuming the polymers to be thermorheologically simple, so that the theory of reduced variables[5] applies (e.g., frequency and temperature are interrelated). Thus, if the complex extensional modulus as a function of circular frequency ω at a temperature T_o is given by ($i = \sqrt{-1}$)

$$E(\omega, T_o) = E'(\omega, T_o) + iE''(\omega, T_o), \quad (1)$$

where E' and E'' are the storage and loss moduli, respectively, then the extensional modulus at a temperature T is given by $E(\omega a_T, T_o)$. The multiplier a_T is referred to as the time-temperature shift function. For a typical polymer, such as polybutadiene, a plot of extensional modulus versus frequency (see Fig. 4) at room temperature can be used to generate data at other temperatures provided the shift function has been experimentally determined and provided that the room temperature data extends over a sufficient frequency range.

Characterizing each polymer to this extent is prohibitive, however, and the usual approach is to generate data at a fixed frequency while varying the temperature. Such data is shown in Fig. 5 for two polymers of interest: (*i*) a blend of 50 percent cis-4 polybutadiene and 50 percent styrene-butadiene rubber (called PBD/SBR) and (*ii*) a blend of 75 percent cis-4 polybutadiene and 25 percent chlorobutyl rubber (called PBD/CBT). A fixed frequency of 110 Hz (data obtained with a Vibron Viscoelastometer marketed by Imass, Inc., Accord, Mass.) was used and the temperature was varied sufficiently to capture the transition regions of interest.[6] Note that the loss tangent, defined by

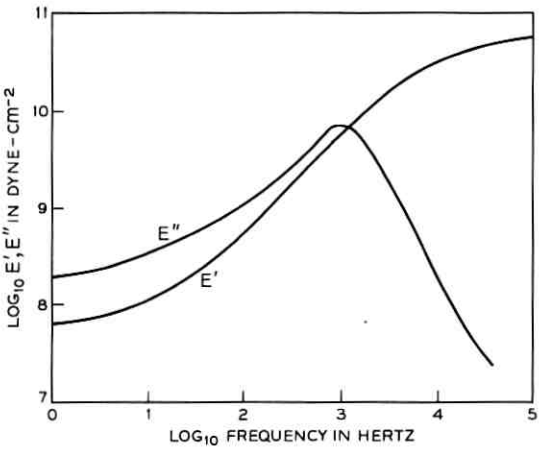


Fig. 4—Complex extensional modulus, polybutadiene, $T = 20°C$.

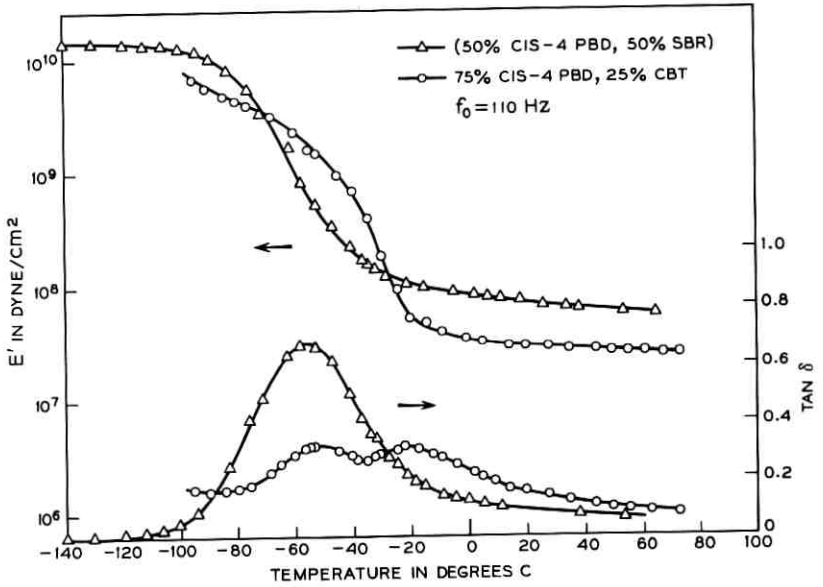


Fig. 5—E' and tan δ vs temperature.

$$\tan \delta = E''/E', \tag{2}$$

has been shown in lieu of the imaginary component of the modulus.

In order to construct modulus versus frequency from these data, the shift function for each polymer must be known. Ordinarily, a_T would be determined experimentally from several fixed frequency runs and their graphical superposition. It is often convenient, however, to assume a form for the shift function that is found to fit a wide variety of polymers and is called the WLF equation:[7]

$$\log_{10} a_T = -c_1(T - T_R)/(c_2 + T - T_R), \tag{3}$$

where c_1 and c_2 are constants and T_R is a reference temperature. Common practice is to use $c_1 = 8.86$ and $c_2 = 101.5$ as the constants and a reference temperature in the middle of the transition region. For the analytical work described here, T_R for the PBD/SBR was selected as -60°C and for the PBD/CBT was chosen as -40°C. Then, with the help of (3), the data of Fig. 5 was converted to the form of Fig. 4 for specific temperatures.

These modulus values can be approximately converted to effective stiffness by using simple strength of materials considerations. If the cone angle is neglected and the washer is assumed to be in a state

of plane stress, then the factor which converts modulus to stiffness (per radian) can be written

$$\kappa = \frac{2\bar{r}t}{l(1 - \nu_R^2)}, \tag{4}$$

where $\bar{r}$ is an average radius for the washer, t is its thickness, l is the height, ν_R is the Poisson's ratio for the rubber, and the multiplier indicates that both washers are being taken into account. The cantilever frustum, which does not provide support stiffness in its unconstrained configuration, is neglected. If ν_R is assumed to be 0.5 and the dimensions of Fig. 3 are used, $\kappa \doteq 0.8$, indicating that the effective rubber stiffness per radian is eight-tenths of the extensional modulus. This conversion factor will be used in the next section in order to help generate parametric design information.

III. SIMPLE VARIATIONAL SOLUTION

As a first step in the rational design process, a procedure for estimating the two lowest resonant frequencies of the transducer, as a function of geometric and material parameters, is developed. A Rayleigh–Ritz procedure is used for deriving these design equations. First, a functional is written which represents the strain energy and kinetic energy of the plate and its deformable supports, less the work done by the acoustic loading.[8] Classical infinitesimal plate theory is used (rotatory inertia and shear deformation are neglected) and piezoelectric stiffening effects are ignored. Then

$$\begin{aligned} F(w) = {} & \frac{1}{2}\int_0^{r_e} D(r)\left\{\left(\frac{\partial^2 w}{\partial r^2} + \frac{1}{r}\frac{\partial w}{\partial r}\right)^2 - 2(1-\nu)\left(\frac{\partial^2 w}{\partial r^2}\right)\left(\frac{1}{r}\frac{\partial w}{\partial r}\right)\right\} rdr \\ & - \tfrac{1}{2}\omega^2 \int_0^{r_e} \rho(r)h(r)\{w(r,t)\}^2 rdr - \int_0^{r_e} p(r,t)w(r,t)rdr \\ & + \tfrac{1}{2}r_s k_s\{w(r_s,t)\}^2 - \tfrac{1}{2}\omega^2 r_s M_s\{w(r_s,t)\}^2, \end{aligned} \tag{5}$$

where r is the radial coordinate of the circular plate, t is the time, and $w(r, t)$ is transverse deflection. The flexural stiffness, density, thickness, Poisson's ratio, and radius of the plate are $D(r)$, $\rho(r)$, $h(r)$, $\nu(r)$, and r_e, respectively. The effective mass, effective stiffness, and effective support radius for the rubber are denoted by M_s, k_s, and r_s, respectively. The circular frequency is ω and the applied acoustic pressure is $p(r, t)$.

Next, an approximate deflected shape is assumed, in terms of one

or more undetermined parameters, and substituted into (5). This shape function should satisfy the geometric boundary conditions for the plate (i.e., those on deflection and slope) identically, but may also satisfy natural boundary conditions (i.e., on shear and bending moment). After substitution, the spatial integration is carried out; then, the stationary value of the functional is found through the first variation and subsequent solution of simultaneous equations for the undetermined parameters. Eigenvalues are found from the homogeneous system. The procedure has been used for clamped and simply supported plates[9] and is usually found to be within a few percent of exact solutions.

The trial function for the microphone is taken to be

$$w(r) = \alpha_o + \alpha_1\left[1 - \left(\frac{6 + 2\nu}{5 + \nu}\right)\left(\frac{r}{r_c}\right)^2 + \left(\frac{1 + \nu}{5 + \nu}\right)\left(\frac{r}{r_c}\right)^4\right] \tag{6}$$

where α_o and α_1 are the undetermined parameters (the harmonic time dependence has been suppressed). This trial function has the properties that

$$w(0) = \alpha_o + \alpha_1 , \tag{7a}$$

$$w(r_c) = \alpha_o , \tag{7b}$$

and

$$M_r(r_c) = 0. \tag{7c}$$

This implies that the generalized coordinate α_o represents the motion at the outside edge of the plate and that the generalized coordinate α_1 represents motion of the center of the plate relative to edge motion. The boundary condition on shear at the outside edge is not, and need not be, satisfied by the trial function; the boundary condition or radial bending moment at the outside edge, which also need not be satisfied by the trial function, is explicitly satisfied, as indicated by (7c). Note that (6) has the value $w(r_s) = \bar{w}_s$ at the effective support radius.

Carrying out the steps previously indicated yields the matrix equation governing the system:

$$[[K] - \omega^2[M]]\begin{Bmatrix}\alpha_o \\ \alpha_1\end{Bmatrix} = \{F\}, \tag{8}$$

where the stiffness matrix, $[K]$; the mass matrix, $[M]$; and the load vector, $\{F\}$, are given by

$$[K] = \frac{D}{r_c^2}\begin{bmatrix} \lambda & \lambda\bar{w}_s \\ \lambda\bar{w}_s & \lambda\bar{w}_s^2 + \dfrac{32(1 + \nu)(7 + \nu)}{3(5 + \nu)^2} \end{bmatrix}, \tag{9a}$$

$$[M] = \rho h r_c^2 \begin{bmatrix} \frac{1}{2} + m & \frac{1}{6}\left(\frac{7+\nu}{5+\nu}\right) + m\bar{w}_s \\ \frac{1}{6}\left(\frac{7+\nu}{5+\nu}\right) + m\bar{w}_s & \frac{(3\nu^2 + 36\nu + 113)}{30(5+\nu)^2} + m\bar{w}_s^2 \end{bmatrix}, \tag{9b}$$

and

$$\{F\} = \tfrac{1}{2} p_0 r_l^2 \begin{Bmatrix} 1 \\ \left[1 - \left(\frac{3+\nu}{5+\nu}\right)\left(\frac{r_l}{r_c}\right)^2 + \frac{1}{3}\left(\frac{1+\nu}{5+\nu}\right)\left(\frac{r_l}{r_c}\right)^4\right] \end{Bmatrix}, \tag{9c}$$

respectively. The dimensionless parameters λ and m are defined by

$$\lambda = \frac{r_c^2 r_s k_s}{D}; \qquad m = \frac{r_s M_s}{\rho h r_c^2}; \tag{10}$$

and the variables p_0 and r_l are the uniform acoustic pressure and the loading radius ($0 < r_l < r_s$).

As an illustration of the Rayleigh–Ritz procedure, consider a transducer composed of two PZT-5A disks, each 0.006 inch thick and 0.590 inch in diameter. The bonding layer is assumed to have negligible thickness. With an in-plane extensional modulus of 6.1×10^{11} dynes/cm^2 and a Poisson's ratio of 0.35, the flexural stiffness for the plate is $D = 1.644 \times 10^6$ dyne-cm. The effective radius of the rubber support (the centerline of contact with the conical washer) is taken to be 0.708 cm, $r_c = 0.75$ cm, the density for PZT-5A is 7.8 g/cm^3, and the total thickness of the plate is 0.0305 cm.

Using these data, approximate values for the first two resonant frequencies can be found as a function of the stiffness ratio, λ, and the mass ratio, m. Figure 6 shows these two resonances plotted parametrically with respect to λ and m. From this plot, the primary effect of the rubber effective mass is to lower both resonances (the second much more markedly than the first).

The Rayleigh–Ritz results are summarized in Table I. Effective translational inertia is found from Ref. 10, where the effective mass of a rubber block bonded between two plates was shown to be slightly larger than one-third of the total mass. Since the total rubber volume per radian is 0.0312 cm^3 and $r_s = 0.708$ cm, then

$$M_s = 0.0146\ \rho_R \text{ g/cm/radian}, \tag{11}$$

where ρ_R is the density of the rubber ($\rho_R \doteq 1.2$ for PBD/SBR and 1.0 for PBD/CBT). The extensional modulus values are obtained by fitting the WLF shifted data of Fig. 5 by collocation;[11] the effective

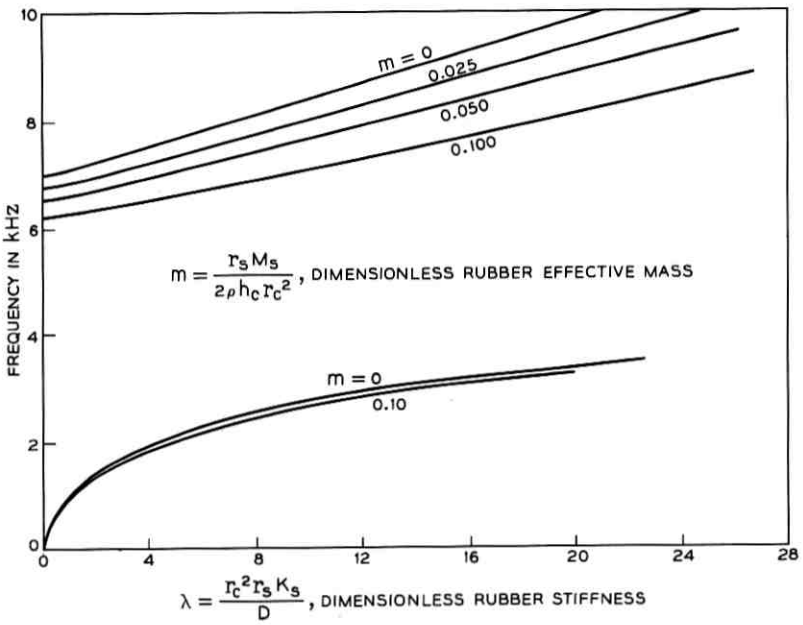


Fig. 6—First and second resonances vs rubber mass and stiffness.

stiffness is then computed using the conversion factor, $\kappa = 0.8$, found previously. The frequencies shown in Table I will be seen later to be in excellent agreement with measured results.

The direct variational calculations can be extended to include forced response and complex rubber properties. Rather than rely on (8) entirely, however, a more exact representation is formulated in the next section.

TABLE I—RAYLEIGH–RITZ RESULTS

	PBD/SBR		PBD/CBT	
	3 kHz	10 kHz	3 kHz	10 kHz
M_s	0.0176	0.0176	0.0146	0.0146
m	0.093	0.093	0.078	0.078
E	116×10^6	148×10^6	48×10^6	59×10^6
k_s	92×10^6	118×10^6	38×10^6	47×10^6
λ	22.5	28.5	9.5	11.5
f_1	3.3 kHz		2.7 kHz	
f_2	9.2 kHz		7.3 kHz	

IV. EXACT SOLUTION–LUMPED SUPPORT PARAMETERS

In this section, the effect of the support stiffness and damping is treated through a boundary condition on the shear resultant. This resultant is assumed to act at the mean radius of the conical washer, implying that the contact area of the rubber is small in comparison to the area of the ceramic (see Fig. 3). Then, the forced harmonic response of the plate can be found from the solution to

$$\nabla^4 w - k^4 w = \begin{cases} -\dfrac{p_o}{D} e^{i\omega t}, & 0 < r < r_l \\ 0, & r_l < r < r_c \end{cases} \tag{12}$$

where the wave number, k, is defined by

$$k^4 = \frac{\rho h \omega^2}{D}. \tag{13}$$

Due to the assumed cylindrical symmetry of the pressure, only axisymmetric solutions of (12) are sought. The plate is then divided into three regions: (*i*) $0 < r < r_l$; (*ii*) $r_l < r < r_s$; and (*iii*) $r_s < r < r_c$. Solutions over these regions are pieced together by satisfying continuity (boundary) conditions on the transverse displacement, the slope, the radial bending moment, and the shear force at the radii r_l and r_s ; in addition, the homogeneous boundary conditions on radial bending moment and shear at the free outer edge of the plate are satisfied. At the effective support radius,

$$Q(r_{s-}) - Q(r_{s+}) = [\omega^2 M_s - k_s(\omega)]w(r_s); \tag{14}$$

i.e., the net shear is opposed by a complex impedance that is proportional to the transverse displacement at that point. The impedance is composed of an inertia term, represented by $\omega^2 M_s$, and a complex stiffness, written as a generalized Maxwell model[12] in the form

$$k_s(\omega) = k_s'(\omega) + ik_s''(\omega), \tag{15}$$

where

$$k_s'(\omega) = k_R + \sum_{n=1}^{N} \frac{k_n \omega^2 \tau_n^2}{1 + \omega^2 \tau_n^2}, \tag{16a}$$

$$k_s''(\omega) = \sum_{n=1}^{N} \frac{k_n \omega \tau_n}{1 + \omega^2 \tau_n^2}, \tag{16b}$$

and k_R , k_n , and τ_n are the equilibrium (rubbery) stiffness, an incre-

mental stiffness, and a relaxation time associated with an incremental stiffness, respectively.

The technique that is used to solve this system, subject to the above stated boundary conditions, does not require the determination of the eigenfunctions of the problem. This procedure can be avoided since a particular solution is known: namely,

$$w_p = \begin{cases} \frac{p_o}{k^4 D}, & 0 < r < r_l \\ 0, & r_l < r < r_e . \end{cases} \tag{17}$$

To this solution it is necessary only to add properly weighted homogeneous solutions to satisfy the boundary conditions—in this case, ordinary and modified Bessel functions of zero order. The matrix inversion required to find the proper weights is carried out on a digital computer and the solution to (12) can then be determined as a function of the acoustic driving frequency. The voltage output is then found from the expression derived in the Appendix.

Two numerical examples are solved in order to illustrate the procedure and to compare the exact (lumped parameter) results with experiment. The transducer design is identical in both cases; the only difference is the conical washer material—in the first case, the PBD/SBR blend; in the second case, the PBD/CBT blend. Effective mass and stiffness are computed by procedures that were described previously. The comparison to experiment for the PBD/SBR blend is shown in Fig. 7 and a similar comparison for the PBD/CBT blend is shown in Fig. 8.

The response comparison for both examples is favorable up to frequencies slightly above the first resonance; then, in both cases, an intermediate response peak is not captured by this model and the response peak at the next resonance is predicted to be much lower than shown by experiment. The location of this latter peak is, however, quite favorable. Perhaps the most disturbing feature of the comparison is the encouraging proximity of the analytical results to the design goal (see Fig. 1)—encouragement that is not borne out by the actual transducer performance. It seems apparent that other deformation mechanisms, not represented adequately by the lumped mechanical model of the conical washer support, are dominating the response at the higher frequencies. For this reason, a more exact model of the support structure is in order.

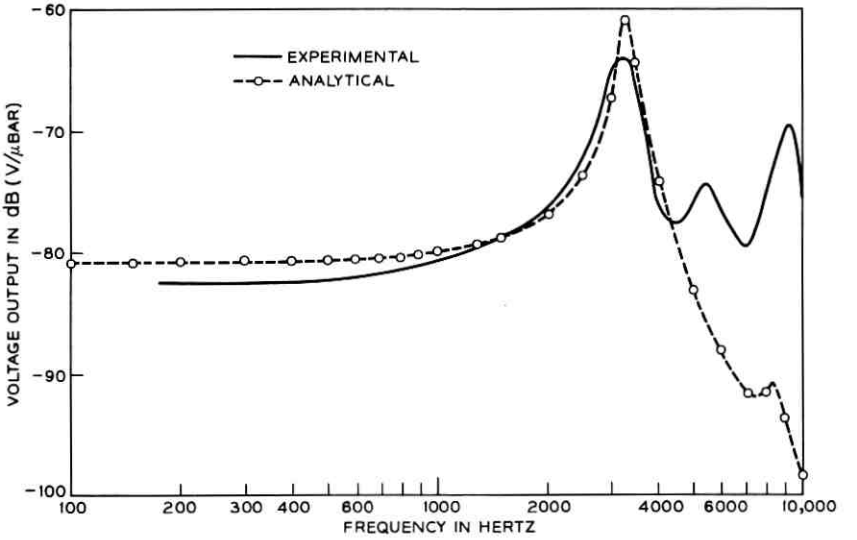


Fig. 7—Voltage output vs frequency, rubber: 50 percent cis-4 PBD, 50 percent SBR.

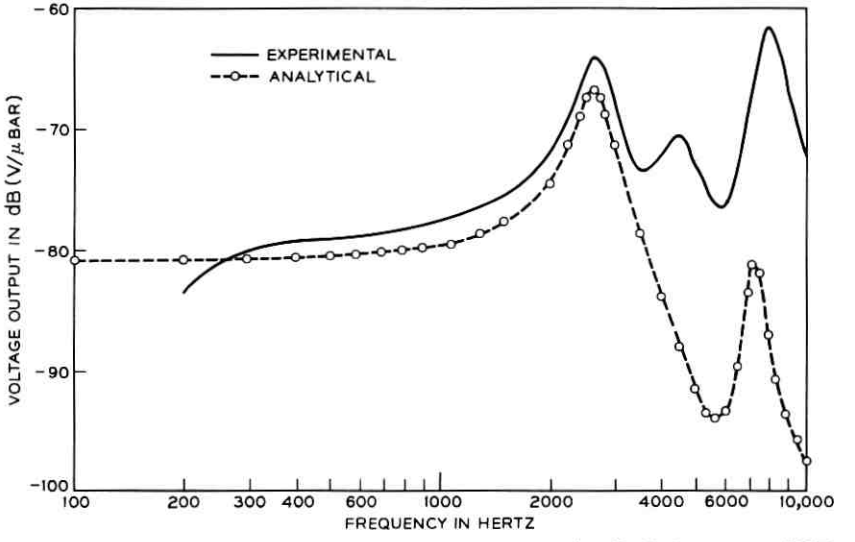


Fig. 8—Voltage output vs frequency, rubber: 75 percent cis-4 PBD, 25 percent CBT.

V. FINITE ELEMENT ANALYSIS

In order to understand the limitations of the lumped mechanical model of the rubber support, a finite element code[13] was exercised. The code was designed to dynamically analyze axisymmetric elastic solids subjected to arbitrary time-dependent loads and includes, as an option, the frequency and mode shape calculations for the solid. For this application, the ceramic was discretized into eighteen plate bending elements and the conical washers were discretized into three successively finer grids, with the most dense grid containing 92 quadrilateral continuum elements. All materials were treated as elastic—the absolute value of the complex extensional modulus of the rubber was used—and Poisson's ratio was chosen to be either 0.45 or 0.49.

A modal survey was then conducted for varying values of rubber extensional modulus. The four lowest resonant frequencies and their corresponding mode shapes were calculated for each modulus value. Typical results are shown in Figs. 9a–9d. These figures portray the influence of the rotatory inertia of the cantilevered frustum, which vibrates either in-phase or out-of-phase with the outer edge rotation of the ceramic. Note that the out-of-phase modes, Figs. 9a and 9c, are not strongly piezoelectrically active, whereas the in-phase modes, Figures 9b and 9d indicate substantial edge rotation with reference to central deflection.

A composite plot of all the results obtained from the modal survey is shown in Fig. 10. This plot correlates well with the experimental results of Figs. 7 and 8. Note that the results are only slightly dependent on the value of Poisson's ratio and on the discretization.

VI. CONCLUSIONS

With the knowledge gained from these three phases of analysis, a coherent set of design conclusions can be drawn. These recommendations fall into two categories: (*i*) rubber material selection and (*ii*) conical washer design modification. In Figs. 11 and 12 the response variation of the ceramic transducer and its rubber supports is shown, as a function of environmental temperature, for the two different rubbers.[14,15] Clearly, the rubber modulus is increasing too rapidly and the loss tangent is not holding an adequate value at the lower temperatures. In addition, the rotatory inertia of the unconstrained rubber is creating intolerable amplitude levels at the higher frequencies. In a recent investigation,[16] block copolymers cast from different solvents seemed to yield dynamic mechanical properties with desirable damping

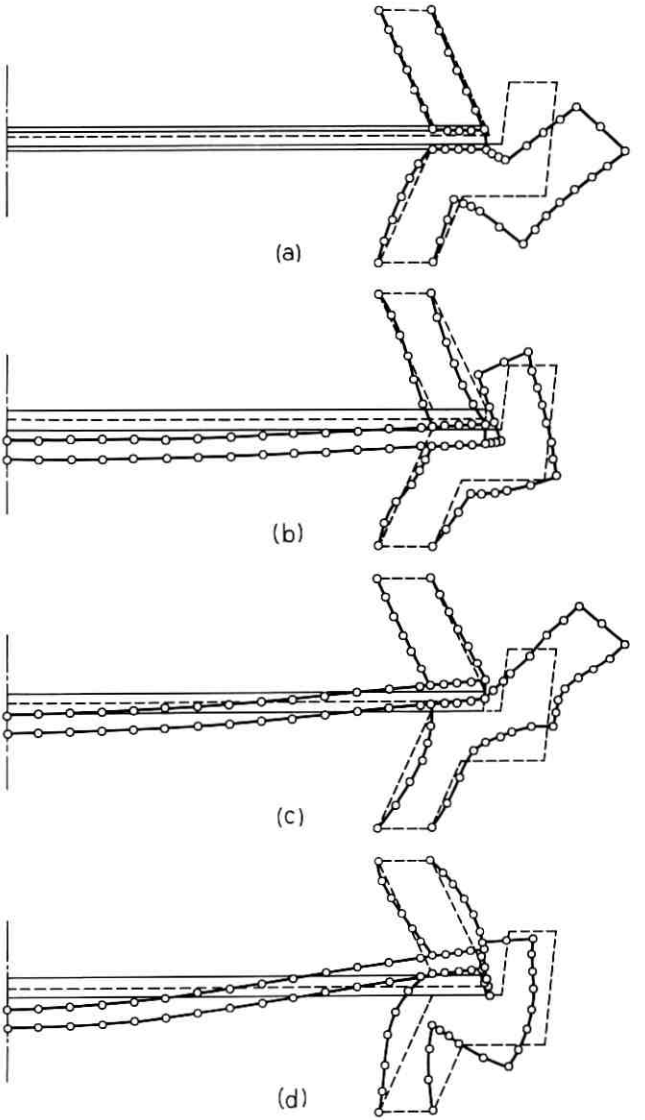


Fig. 9a—Resonant mode shape, $f_0 = 2250$ Hertz, $E_R = 40 \times 10^6$ dynes/cm^2.

Fig. 9b—Resonant mode shape, $f_0 = 2700$ Hertz, $E_R = 40 \times 10^6$ dynes/cm^2.

Fig. 9c—Resonant mode shape, $f_1 = 5400$ Hertz, $E_R = 60 \times 10^6$ dynes/cm^2.

Fig. 9d—Resonant mode shape, $f_2 = 8500$ Hertz, $E_R = 80 \times 10^6$ dynes/cm^2.

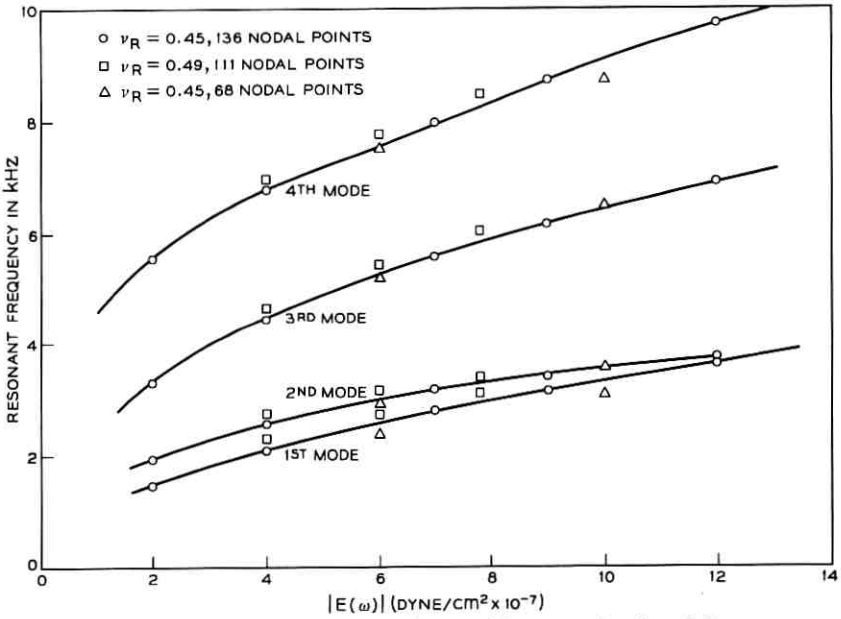


Fig. 10—Finite element frequencies vs rubber extensional modulus.

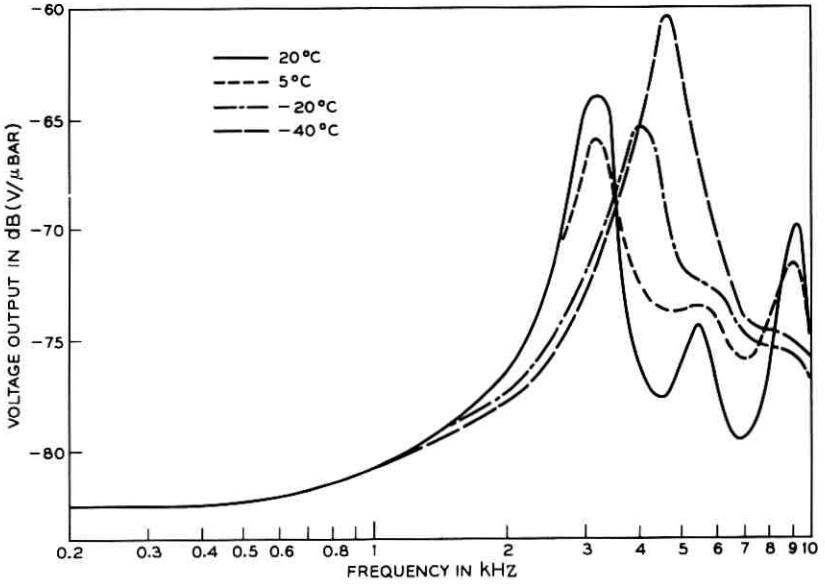


Fig. 11—Frequency response vs temperature change, rubber: 50 percent PBD, 50 percent SBR.

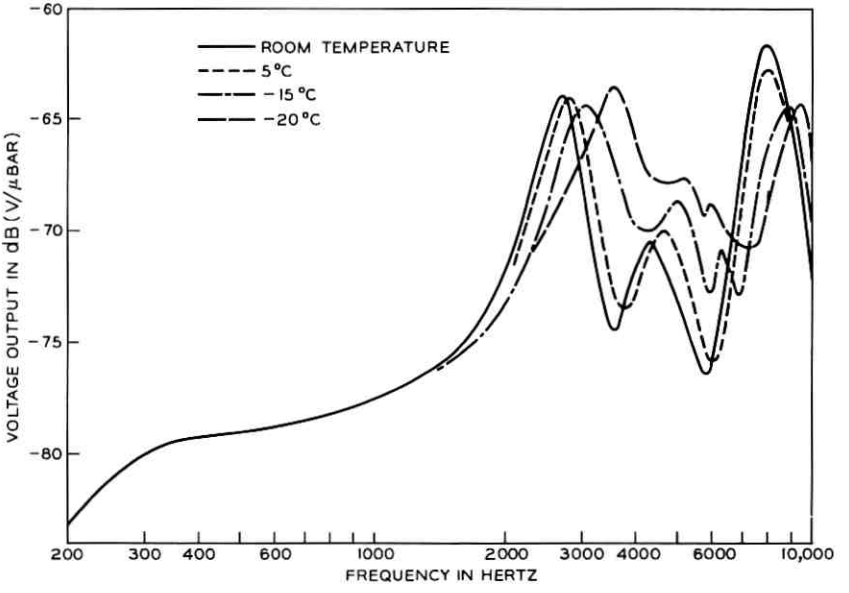


Fig. 12—Frequency response vs temperature change, rubber: 75 percent PBD, 25 percent CBT.

characteristics and relatively stable stiffness. Because of the dual transition (see Fig. 13), a styrene-butadiene-styrene block copolymer, obtained from solutions in carbon tetrachloride (c), toluene (T), ethyl acetate (E), and methyl ethyl ketone (M), has a sufficiently high loss tangent over a 200°C temperature range and also has a relatively constant modulus over a 130°C range. If the modulus is too high over this range, the washer design can be modified—thinner and taller cross section—to achieve nominal stiffness. A material tailoring program might produce a rubber which will help the transducer meet the design template.

In addition to the improvement of rubber mechanical properties, the washer design should be altered in order to decrease, substantially, the rotatory inertia of the unrestrained cantilever section. One possibility is shown in Fig. 14. An inverted vee-shaped design is depicted for the bottom washer that has three salient features: (*i*) a thinner cross section in order to maintain the modulus/stiffness ratio; (*ii*) restraint of the cantilevered section; and (*iii*) a seating lip to aid in fabrication of the transducer.

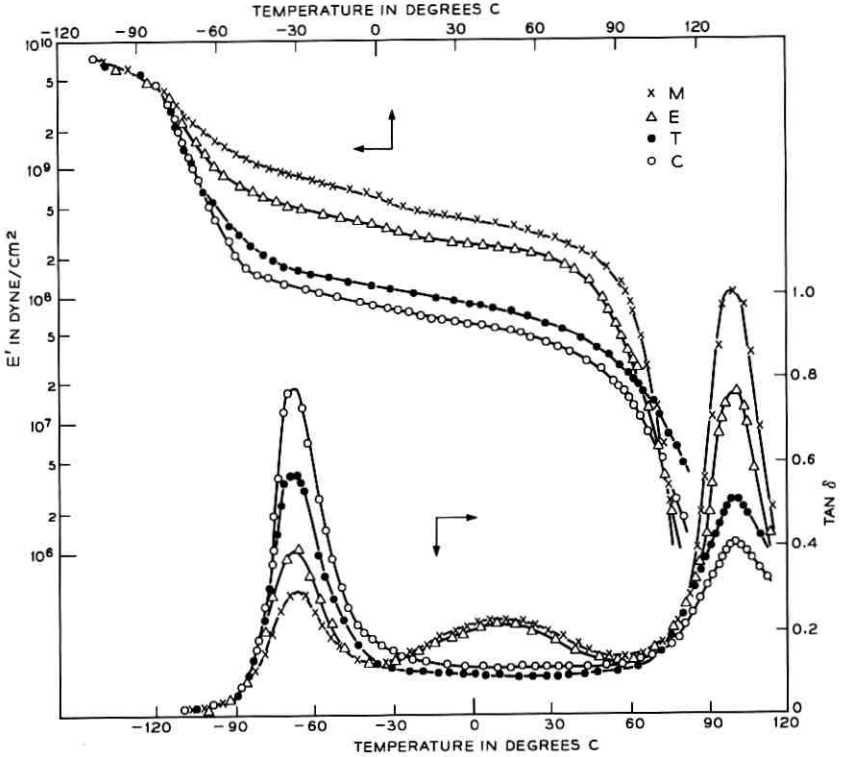


Fig. 13—E' and tan δ vs temperature.

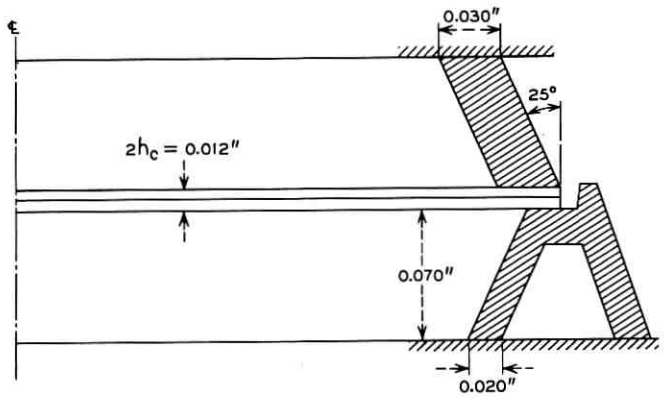


Fig. 14—Modified conical washer design.

VII. ACKNOWLEDGMENTS

The authors would like to acknowledge the design work and measurements of H. W. Bryant and T. C. Austin. They would also like to acknowledge the efforts of D. L. Loan and W. F. Moore for their work on the rubber materials.

In particular, they would like to thank W. D. Goodale for his very professional coordination of the project.

APPENDIX

Voltage Output Calculation

The voltage output of the bimorph ceramic microphone is calculated from the strain field under the following assumptions.

The significant contributions to the component of the electric field, $E_z(r)$, perpendicular to the midplane of the plate are generated by the two strain components ϵ_{rr} and $\epsilon_{\theta\theta}$. This assumes that the electric displacement field is negligible. Hence

$$\begin{aligned} \mathbf{Z}\cdot\mathbf{E}(r) = E_z(r) &= \mathbf{Z}\cdot(-\mathsf{h}\cdot\mathbf{S} + \boldsymbol{\beta}\cdot\mathbf{D}) \\ &\doteq -h_{12}\epsilon_{12} - h_{13}\epsilon_{13}\,, \end{aligned} \tag{18}$$

where $\mathbf{Z}$ is the unit normal vector to the plate, h the piezoelectric tensor relating strain, $\mathbf{S}$, to electric field, and $\boldsymbol{\beta}$ the electric displacement field, $\mathbf{D}$, to electric field.[17] However, for a ceramic poled in the $\mathbf{Z}$ direction, $h_{13} = h_{12}$; hence,

$$\begin{aligned} E_z(r) &= -h_{13}(\epsilon_{12} + \epsilon_{13}) \\ &= -h_{13}(\epsilon_{rr} + \epsilon_{\theta\theta}). \end{aligned} \tag{19}$$

From thin-plate theory, the strain term is given in terms of the midplane displacement, $w(r)$, by the expression

$$\epsilon_{rr} + \epsilon_{\theta\theta} = -z\left[\frac{1}{r}\frac{d}{dr}\left(r\frac{dw}{dr}\right)\right]; \tag{20}$$

thus

$$E_z(r) = +zh_{13}\left[\frac{1}{r}\frac{d}{dr}\left(r\frac{dw}{dr}\right)\right]. \tag{21}$$

The average electric field over the plated area is given by

$$E_{ave} = \frac{1}{\pi r_p^2}\int_0^{r_p}\int_0^{2\pi} E_z(r) r\, dr\, d\theta, \tag{22}$$

where r_p is the electroded radius.

Substitution of (21) in (22) yields

$$E_{ave} = \frac{2zh_{13}}{r_p^2}\left(r\frac{dw}{dr}\right)_{r=r_p}. \tag{23}$$

The potential difference across one plate is therefore given by the line integral $\int_0^{h_c} E_{ave}\,dz$, but, since the two plates are series connected, the total voltage is given by

$$V(f) = \frac{2h_{13}}{r_p^2}\,h_c^2 r\left(\frac{dw}{dr}\right)_{r=r_p}, \tag{24}$$

where h_c is the thickness of one ceramic disk, and $h = 2h_c$ the total thickness.

REFERENCES

1. Bryant, H. W., "Telephone Transmitter Designs Having Ceramic Transducers," unpublished work.
2. O'Bryan, H. M., Jr., and Bryant, H. W., "Ceramic Assemblies for Microphones," unpublished work.
3. Austin, T. C., and Herson, R. J., "Response Shaping of Transducer by Optimization of Housing Geometry," unpublished work.
4. Ferry, J. D., *Viscoelastic Properties of Polymers*, Second Edition, New York: John Wiley and Sons, Inc., 1970, p. 145.
5. Ibid., pp. 292–351.
6. Daane, J. H., private communication.
7. Ferry, loc. cit., p. 307.
8. Timoshenko, S., and Woinowsky-Krieger, S., *Theory of Plates and Shells*, Second Edition, New York: McGraw-Hill Book Company, 1959, pp. 344–346.
9. Rajappa, N. R., "Free Vibration of Rectangular and Circular Orthotropic Plates," AIAA J., *1*, May 1963, pp. 1194–1195.
10. Hilyard, N. C., "Effective Mass of Bonded Rubber Blocks," J. Acoust. Soc. Amer., *47*, 1969, pp. 1463–1465.
11. Schapery, R. A., "Approximate Methods of Transform Inversion for Viscoelastic Stress Analysis," Proc. Fourth U. S. Nat. Conf. Appl. Mech., June 1961, pp. 1075–1085.
12. Fung, Y. C., *Foundations of Solid Mechanics*, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1965, pp. 25–28.
13. Ghosh, S., and Wilson, E. L., "Dynamic Stress Analysis of Axisymmetric Structures Under Arbitrary Loading," Report No. *EERC 69–10*, Earthquake Engineering Research Center, University of California, Berkeley, September 1969.
14. Austin, T. C., private communication.
15. Bryant, H. W., private communication.
16. Miyamoto, T., Kodama, K., and Shibayama, K., "Structure and Properties of a Styrene-Butadiene-Styrene Block Copolymer," J. Polymer Sci. A-2, *8*, 1970, pp. 2095–2103.
17. Berlincourt, D. A., Curran, D. R., and Jaffe, H., "Piezoelectric and Piezomagnetic Materials and Their Function in Transducers," in *Physical Acoustics*, Vol. IA, edited by W. P. Mason, New York: Academic Press, 1964.

Contributors to This Issue

Syed V. Ahamed, B.E., 1957, University of Mysore, India; M.E., 1958, Indian Institute of Science; Ph.D., 1962, University of Manchester, U. K.; Post-Doctoral Research Fellow, 1963, University of Delaware; Assistant Professor, 1964, University of Colorado; Bell Laboratories, 1966—. Mr. Ahamed was working in Computer Aided Engineering Analysis and Software Design at Whippany. Presently, he is investigating the applications of Algebraic Techniques for Domain Circuits.

William T. Barnett, B.S.E.E., 1958, Illinois Institute of Technology; M.E.E., 1960, New York University; Western Electric, 1953–1958; Bell Laboratories, 1958—. Mr. Barnett has worked on problems related to microwave radio relay systems. Since 1966, he has supervised a group concerned with propagation problems. Member, IEEE.

Glen T. Cheney, A. B., 1960, and M. S. (Physics), 1964, San Diego State College; Convair Astronautics, 1957–1960; General Atomic, 1960–1964; Bell Laboratories, 1964—. Before joining Bell Laboratories, Mr. Cheney was involved in applied research activities in the areas of thermal and thermionic properties of reactor materials and radiation-induced conductivity in plastic films. At Bell Laboratories, he has been concerned with the development of device and circuit technology for both bipolar and unipolar silicon integrated circuits. His current assignment is Supervisor, Digital IGFET Circuits Group.

J. D. Denman, Central State University; Western Electric Company, 1960—. Mr. Denman is a Planning Engineer at the Oklahoma City Works and has worked in quality assurance, industrial engineering, and product engineering. He is presently in Crossbar Equipment Assembly and Wiring Engineering where his responsibilities include solderless wrap control and cost reduction coordination. Member, IEEE, Society of Manufacturing Engineers, Institute for Certification of Engineering Technicians, Society for the Advancement of Management, American Radio Relay League, Oklahoma Industrial Arts Association.

MICHAEL J. FERGUSON, B.A.Sc., 1962, University of Toronto; M.S., 1963, California Institute of Technology; Ph.D., 1966, Stanford University. Mr. Ferguson is an Assistant Professor of Electrical Engineering at McGill University, Montreal, Canada. He spent the summer of 1970 in the Data Theory Department at Bell Laboratories where he did work in data transmission, coding, and communication networks.

JAMES L. FLANAGAN, B.S., 1948, Mississippi State University; S.M., 1950, and Sc.D., 1955, Massachusetts Institute of Technology; Faculty of Electrical Engineering, Mississippi State University, 1950–1952; Air Force Cambridge Research Center, 1954–1957; Bell Laboratories, 1957—. Mr. Flanagan has worked in speech and hearing research, computer simulation and digital encoding, and acoustics research. He is Head, Acoustics Research Department. Fellow, IEEE; Fellow, Acoustical Society of America; Tau Beta Pi; Sigma Xi; member of several government and professional society boards, including committees of the National Academy of Sciences and the National Academy of Engineering.

A. JAY GOLDSTEIN, B.S. (Physics), 1948, and M.A. (Mathematics), 1951, Pennsylvania State University; Ph.D. (Mathematics), 1955, Massachusetts Institute of Technology; mathematics faculty of Polytechnic Institute of Brooklyn, 1954–1957; Bell Laboratories, 1957—. Mr. Goldstein has worked on network analysis and synthesis, computer-oriented combinatoric algorithms, and interactive computing systems. He is now supervisor of the Mathematical Techniques Group.

LEON D. HARMON, B.S.E.E., 1956, New York University; Institute for Advanced Study, Princeton, N. J., 1950–1956; Bell Laboratories, 1956—. At the Institute for Advanced Study, Mr. Harmon was engaged in the research and development of high-speed digital computing systems. At Bell Laboratories, his work has included studies in visual pattern recognition by machines, sensory psychophysics, and information processing in the nervous system. His present research in the Systems Theory Research Department includes the analysis of neurophysiological systems using electronic neural analogs, and studies of automatic machine processing of visual and auditory patterns. Member, AAAS, IEEE, Society for Neuroscience.

ERNAM F. KING, E. E. T., 1963, Capital Radio Engineering Institute; Bell Laboratories, 1957—. Mr King has worked on the design and development of traveling-wave tubes, bipolar transistors, and IGFET integrated circuits. Member, IEEE.

ANN B. LESK, B.A. (Chemistry), 1968, Radcliffe College; Research Associate, Arthur D. Little, Inc., Cambridge, Mass., 1968–1969; Bell Laboratories, 1969—. Ms. Lesk has worked in the fields of computational linguistics and cancer chemotherapy. At Bell Laboratories, she is working on face recognition and digital picture-processing.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Laboratories at the University of Utah where he wrote a book on quantum electronics. He is presently working on the transmission aspect of a light communications system. Member, IEEE, Optical Society of America.

GEORGE MARR, B. A., 1963, Hope College; M. S., 1965, Miami University; Ph.D. (Physics), 1968, The Ohio State University; Bell Laboratories, 1968—. Mr. Marr has worked on the design and development of IGFET and BIGFET integrated circuits. His recent work includes the improvement of IGFET integrated circuit performance by the incorporation of both bipolar transistors and ion-implanted, depletion-mode IGFETs in the same integrated circuit. Member, American Physical Society, IEEE.

HENRY E. MEADOWS, B.E.E., 1952, M.S., 1953, and Ph.D., 1959, Georgia Institute of Technology; Instructor in Electrical Engineering, Georgia Institute of Technology, 1955–1958; Bell Laboratories, 1959–1962; Professor of Electrical Engineering in the Department of Electrical Engineering and Computer Science, Columbia University, 1962—. During 1968–1969, Mr. Meadows was a Ford Foundation Resident in Engineering Practice at the Philco-Ford Corporation, Palo Alto, California. His main interests lie in systems engineering including controls and simulation. Member, IEEE, Sigma Xi.

ROBERT E. NICKELL, B.S., 1963, M.S., 1964, and Ph.D., 1967, (all in Engineering Science), University of California at Berkeley; Bell Laboratories, 1968—. Mr. Nickell has been assigned to both military and Bell System projects, involving such aspects as structural dynamics of ballistic missile interceptors, thermal shock of ceramic nozzles for continuous copper casting, and vibration-enhanced soil penetration. He is currently a supervisor in the Ocean System Planning Department and is on assigned teaching to Brown University, Providence, R. I., for the academic year 1971–1972. Member, Tau Beta Pi, Chi Epsilon, Sigma Xi, AAAS, Society of Rheology, AIAA, ASME, ASCE. He is currently serving on technical committees of ASME and ASCE.

EUGENE G. PARKS, A. E. E., 1958, Pennsylvania State University; Bell Laboratories, 1958—. Mr. Parks has been concerned with the development of process technologies for many types of silicon semiconductor devices. Recently, he has been responsible for the operation of the device development laboratory at the Allentown branch laboratory.

G. PERSKY, B.S.E.E., 1959, Rensselaer Polytechnic Institute; M.S.E.E., 1961, and Ph.D. (Physics), 1968, Polytechnic Institute of Brooklyn; Bell Laboratories, 1967—. Since 1967, Mr. Persky has worked on problems of high-field transport in semiconductor devices. Member, IEEE, American Physical Society, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

LAWRENCE R. RABINER, S.B. and S.M., 1964, and Ph.D. (E.E.), 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962–1964, 1967—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Since 1967, he has been engaged in research on speech communication, signal analysis, digital filtering, and techniques for waveform processing. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, IEEE; Fellow, Acoustical Society of America. He is chairman of the IEEE G-AE Technical Committee on Digital Signal Processing, and member of the technical committees on speech communication of both the IEEE and the Acoustical Society.

HARVEY RUBIN, B.S., 1965, M.S.E.E., 1966, and Eng.Sc.D., 1970, Columbia University; Bell Laboratories, 1965–1968, 1970—. Mr. Rubin has participated in the development of test equipment for use in evaluating the performance of telephone communications systems. From 1966 to 1968 he was involved in software design for the TSPS system. Presently, as a member of the Exploratory Integrated Electronics Group, he is developing filters and equalizers which meet critical specifications. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

RONALD W. SCHAFER, B.S. (E.E.), 1961, and M.S. (E.E.), 1962, University of Nebraska; Ph.D., 1968, Massachusetts Institute of Technology; Bell Laboratories, 1968—. Mr. Schafer has been engaged in research on digital waveform processing techniques and speech communication. Member, Phi Eta Sigma, Eta Kappa Nu, Sigma Xi, IEEE, Acoustical Society of America, and the IEEE G-AE Technical Committees on Digital Signal Processing and Speech Communication.

WOLFGANG O. SCHLOSSER, Dr. Ing., 1964, Technische Hochschule, Darmstadt, Germany; Research Associate, Technische Hochschule, Braunschweig, Germany, 1963–1966; Bell Laboratories, 1966—. Mr. Schlosser's work has included the design of microwave IMPATT oscillators and the design of mm-wave phase switches and PIN diodes. He is now working on optical communication subsystems. Member, IEEE.

NEIL J. A. SLOANE, B.E.E., 1959, and B.A. (Hons.), 1960, University of Melbourne, Australia; Postmaster General's Department, Commonwealth of Australia, 1956–1961; M.S., 1964, and Ph.D., 1967, Cornell University; assistant professor of electrical engineering, Cornell University, 1967–1969; Bell Laboratories, 1969—. Mr. Sloane is engaged in research in coding theory, communication theory, and combinatorial mathematics. Member, IEEE, American Mathematical Society, Mathematical Association of America.

D. C. STICKLER, B.Sc., 1956, M.Sc., 1959, and Ph.D., 1964, The Ohio State University; Bell Laboratories, 1965—. Since joining Bell Laboratories, Mr. Stickler has studied propagation in underwater acoustics. The microphone analysis covered in his paper was completed while on an internship. Currently he is in the Ocean Physics Research Department. Member, Sigma Xi.

B. S. T. J. BRIEF

Effect of Ambient Temperature on Infrared Transmission Through a Glass Fiber

by R. W. DAWSON

(Manuscript received December 2, 1971)

Recent progress in reducing the losses in optical fibers[1] has increased the possibility that such fibers might be used as dielectric waveguides in future optical communication systems.[2] The loss properties of such fibers are therefore of interest. This note concerns the measured change in transmission loss of a glass fiber for an ambient temperature variation of −196°C to +200°C. The results indicate that the loss in glass fibers varies only slightly with temperature.

The attenuation measurements were made on single fibers taken from a Corning type 5900 optical fiber bundle. These fibers were approximately 60 μm in diameter and had a very thin cladding with a refractive index about 10 percent below that of the core. The elevated temperature measurement was performed on a 39-meter length which was made up of three 10- to 20-meter lengths joined into a single piece by a low-loss fusing process.[3] The reduced-temperature tests were made with a 12-meter segment of this same fiber. The light source consisted of a 50-μm-diameter gallium arsenide light emitter diode (GaAs LED)[4] which was used to supply about 0.05 mW of power into the fiber at a wavelength of 0.9 μm. Detection was accomplished with a silicon PIN photodetector. The high-temperature test was carried out with the major portion of the fiber on a reel in an oven; for the low-temperature test, the fiber was coiled in a Dewar flask filled with liquid nitrogen. In both cases, the source and the detector were outside the test chamber in a room-temperature ambient.

Calibration and stability tests showed that no measurable change occurred in the LED output when the input current was maintained within ±0.05 percent; that the optical power from the LED decreased 0.24 percent for a one-degree rise in the ambient temperature (near 25°C); and that there was no measurable change in the detector sen-

sitivity for an ambient temperature variation of ±2°C. The recording voltmeter drifted less than 0.001 mV (the estimated reading accuracy) in a 5-hour period. The time of an individual run was less than one hour.

Results of the measurements are summarized in Table I. The absolute attenuation of the fiber at room temperature was determined by measuring the received power, first through the entire length, and then (after breaking the fiber near the source) through a very short length. The tabulated results include a correction for a 0.144-percent reduction in input power due to a 0.6°C increase in the ambient temperature near the LED during the course of the heating run. The ambient temperature near the detector changed less than one degree during both runs, and thus no correction was required. The difference in loss per unit length of the two sections of fiber listed in the table is typical, in our experience, of the variation in the infrared loss of different individual fibers from the same bundle.

On the assumption of a thermal expansion coefficient for the glass of approximately 1×10^{-5}/°C, the change in fiber length over these large temperature ranges is significant. The estimated increase was 7 cm for a temperature increase of 175°C, for example. While this expansion slightly changes the value of the fiber loss per unit length, it does not affect consideration of the total change in loss between two terminals connected by a specific fiber.

The small variation in transmission loss measured for rather extreme temperature changes indicates that glass or glass-like dielectric waveguides should have transmission characteristics essentially unaffected

TABLE I—MEASURED CHANGE IN OVERALL TRANSMITTED POWER THROUGH CORNING 5900 FIBER

Fiber Length @25°C (m)	Ambient Temp. (°C)	Received Power (dBm)	Fiber Loss (dB)	Change in Effective Attenuation Constant (%)
39.20	+25	−44.83	31.82†	
	+200	−44.78*	31.76‡	−0.19*
12.02	+25	−26.29	13.28†	
	−196	−26.41	13.42‡	+1.05

* Corrected for change in source output due to ambient-temperature change of source

† Measured

‡ Derived from change in received power

by ordinary ambient temperature changes. The measured value for this fiber, of the order of 0.001 percent per degree Centigrade above room temperature, would correspond to a variation of the order of 10^{-3} dB/°C between repeaters separated by a transmission line with an overall attenuation of 50–60 dB. Thus the complex active temperature compensation required by coaxial cable systems[5] probably would not be necessary in a system employing fiber transmission lines.

These results qualitatively confirm the expectation that the absorption characteristics of wide band-gap materials should change very little as a result of normal temperature fluctuations. Furthermore, the direction of the small but detectable changes measured here is consistent with an increase in the lower-energy-state population of the glass at lower temperatures.

REFERENCES

1. Kapron, F. P., Keck, D. B., and Maurer, R. D., Appl. Phys. Letters, *17*, (1970), p. 423.
2. Miller, S. E., "Integrated Optics: An Introduction," B.S.T.J., *48*, No. 7 (September 1969), pp. 2059–2069.
3. Bisbee, D. L., "Optical Fiber Joining Technique," B.S.T.J., *50*, No. 10 (December 1971), pp. 3153–3158.
4. Burrus, C. A., and Dawson, R. W., "Small-Area High-Current-Density GaAs Electroluminescent Diodes and a Method of Operation for Improved Degradation," Appl. Phys. Letters, *17*, No. 3 (August 1971).
5. Members of the Technical Staff—Bell Telephone Laboratories, *Transmission Systems for Communications*, Third Edition, Winston-Salem, N.C.: Western Electric Company, 1964, Chapter 14.