THE BELL SYSTEM TECHNICAL JOURNAL

SEPTEMBER 1983
VOL. 62, NO. 7, PART 1

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

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

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

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

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

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 62 September 1983 Number 7, Part 1

Stabilized Biasing of Semiconductor Lasers

By R. G. SWARTZ* and B. A. WOOLEY*

(Manuscript received December 10, 1982)

In this paper we analyze the design of biasing and control circuits for semiconductor lasers in a generalized context based on an idealized laser characteristic. In particular, we address three major design considerations: whether to bias the laser above or below threshold, how to stabilize the optical output levels independent of variation in the average output power, and to what degree the output levels can be stabilized relative to various circuit and device parameters. Results of our study indicate that to eliminate from the optical output any dependence on either variation in laser device characteristics or the dc average of the input signal, feedback control of both the prebias and modulation current is necessary.

I. INTRODUCTION

Within the past few years digital lightwave communication systems have become a practical reality. Several systems have been demonstrated for both interoffice trunk transmission and the subscriber plant.[1-4] In these applications optical fiber systems have the advantages of inherently large bandwidths and electrical isolation.

* Bell Laboratories.

High-bit-rate lightwave systems commonly use semiconductor laser diodes as the optical sources. The diodes are threshold devices whose characteristics depend on both age and operating temperature. As a consequence, large variations in the light output of the lasers will occur unless special measures are taken to properly bias and modulate these devices.

Several circuits for biasing and digitally modulating injection lasers have been reported. To ensure modulation of the laser output at the highest possible rates, these circuits typically dc bias the laser near its threshold. A modulation current is then superimposed on this bias to switch between the high and low light outputs. The circuits described to date commonly used negative feedback control of the bias current to stabilize the laser light output. In some early circuits the feedback stabilizes the average optical output of the laser. For this method to be successful, the digital input codes must exhibit a fixed on-off ratio (constant average value). More recent laser driver circuits employ balancing compensation of the modulation signal and purport to allow arbitrary on-off ratio digital codes.[5-8]

In this paper we consider the design of laser biasing and control circuits in a generalized context. Within this context we address three major design considerations: the choice of biasing the laser above or below threshold, how to stabilize the output independently of the nature of the laser modulation, and to what degree the laser output levels can be readily stabilized relative to various circuit and device parameters. Initially, we consider the stabilization obtainable by means of the approach adopted in a recently described monolithic laser driver, wherein feedback stabilization of the laser bias current is augmented by a simple balancing compensation of the modulation signal.[7] Following this analysis, we examine the benefits of using modulation current compensation.

II. FEEDBACK BIAS STABILIZATION

Figure 1 shows the luminosity versus current characteristic assumed for heterojunction lasers in this analysis. This relation can be characterized by three parameters: the threshold current, I_T, the subthreshold differential slope efficiency, η_1, and the above-threshold slope efficiency, η_2. (The variables used are defined at the back of this paper.) These parameters analytically approximate the characteristic of Fig. 1 by the piecewise linear relationships

$$L = \eta_1 I_L, \qquad I_L \leqslant I_T \tag{1}$$

and

$$L = \eta_1 I_T + \eta_2 (I_L - I_T), \qquad I_L \geqslant I_T, \tag{2}$$

where L is the luminosity (or light output intensity) of the laser.

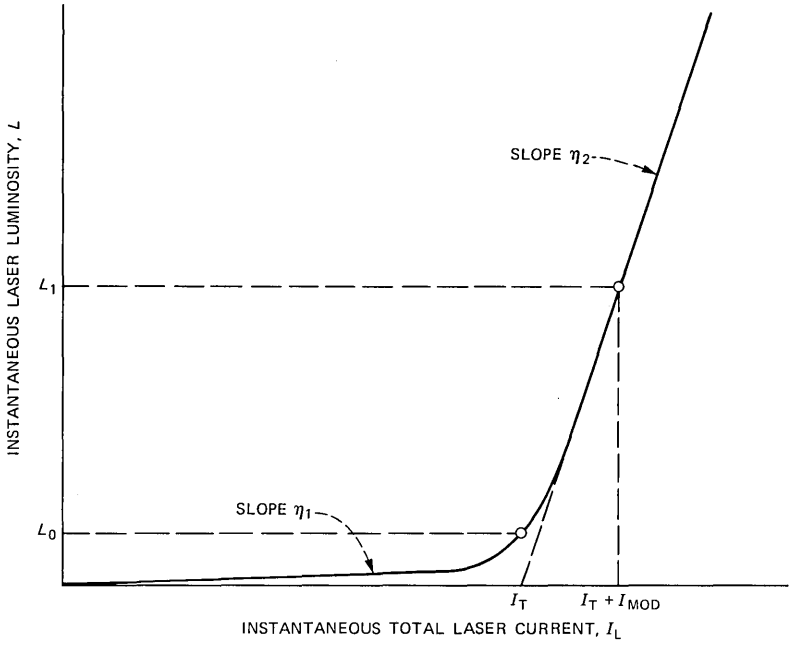


Fig. 1—The luminosity versus current relationship for an injection laser.

Figure 2 is a generalized circuit diagram for a recently reported integrated laser driver employing feedback stabilization.[7] In this circuit the laser is biased near its threshold by a prebias current I_A. Added to this bias is a modulation current, I_M, which switches the laser between its ZERO and ONE light output levels (L_0 and L_1). The prebias current is stabilized by a negative feedback loop comprising the laser, a photodetector, a reference current (I_B), a low-pass filter (C_A), and a current amplifier (A). The photodetector generates a current proportional to the optical output of the laser, typically by monitoring the light emitted from its rear face. The photodetector current (I_D) is compared to the reference current at the summing node, S, and the resulting current difference is then low-pass filtered and amplified to generate the prebias current. Because the modulation current, I_M, will alter the dc component of the laser output, the current I_X is added to node S to cancel this influence, as described below.

It is assumed throughout the analysis that:

1. The differential slope efficiency of the laser is much greater above threshold than below, i.e., $\eta_2 \gg \eta_1$.

2. As a consequence of the filtering provided by C_A, the response time of the feedback loop is long in comparison with the time constants of modulation-related parameters. It is also required, however, that the feedback loop response time be much shorter than the time

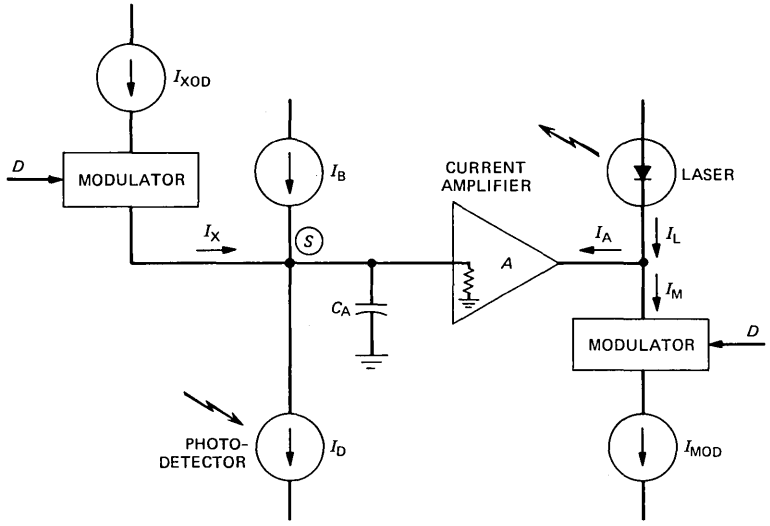


Fig. 2—A laser driver employing feedback stabilization of bias current.

constants of the laser parameter drift that occurs as a consequence of aging or changes in environment.

In the circuit of Fig. 2, the instantaneous laser current, I_L, is given by the sum of the prebias current and the instantaneous modulation current

$$I_L = I_A + I_M, \tag{3}$$

where

$$I_M = DI_{MOD} \tag{4}$$

and D is the binary data signal driving the modulator ($D = 0$ or 1). In Fig. 2, the capacitor, C_A, serves to average the summation of currents feeding the input to amplifier A. Thus,

$$I_A = A_I(I_B + \bar{I}_X - \bar{I}_D) \tag{5}$$

where A_I is the amplifier current gain, and $\bar{I}_X$ and $\bar{I}_D$ are the dc components (averages) of the balance and detector currents, respectively.

The output of the photodetector in Fig. 2 is assumed to be related to the laser light output by a proportionality factor, f. Thus, the average detector output current is given in terms of the average laser luminosity by

$$\bar{I}_D = f\bar{L}. \tag{6}$$

The laser light levels L_0 and L_1 are defined such that $L_0 \triangleq L$ when $D = 0$ and $L_1 \triangleq L$ when $D = 1$. It therefore follows that

$$\bar{L} = \bar{D}(L_1 - L_0) + L_0 \tag{7}$$

and then from (6) and (7) that

$$\bar{I}_D = f[\bar{D}(L_1 - L_0) + L_0]\,. \tag{8}$$

Implicit in this equation is the assumption that time delays in the responses of the laser and photodetector are negligible.

Inserting (8) into (5) and observing that $\bar{I}_X = \bar{D}I_{XOD}$ leads to the relationship

$$I_A = A_I\{I_B + \bar{D}I_{XOD} - f[\bar{D}(L_1 - L_0) + L_0]\}\,. \tag{9}$$

This result, together with the relationship between the laser luminosity (L) and current (I_L), as represented by (1) and (2), will next be used to determine the laser light output levels, L_0 and L_1. However, to proceed with this analysis we must first determine whether the laser is prebiased above threshold or below. This distinction, which seems minor at first glance, has important implications for the ultimate stability of the optical output. We first consider the above-threshold case.

2.1 Above-threshold prebiasing ($I_{L0} \geq I_T$)

From eq. (3) and the definition of L_0 and L_1 it follows that

$$I_{L0} = I_A \tag{10a}$$

and

$$I_{L1} = I_A + I_{MOD}. \tag{10b}$$

For the case where the laser is biased above threshold ($I_{L0} \geq I_T$), it thus follows from (2) that

$$L_0 = \eta_2(I_A - I_T) + \eta_1 I_T \tag{11a}$$

and

$$L_1 = \eta_2(I_A + I_{MOD} - I_T) + \eta_1 I_T\,. \tag{11b}$$

Substitution of (11) into (9) leads to the result

$$I_A = A_I\{I_B + \bar{D}I_{XOD} - f[\eta_2\bar{D}I_{MOD} + \eta_2(I_A - I_T) + \eta_1 I_T]\}\,. \tag{12}$$

This expression can be solved for I_A to obtain

$$I_A = \frac{A_I[I_B + (k_2 - k_1)I_T] + A_I\bar{D}(I_{XOD} - k_2 I_{MOD})}{1 + A_I k_2}, \tag{13}$$

where the laser-photodetector current efficiencies k_1 and k_2 are defined as $k_1 \triangleq f\eta_1$ and $k_2 \triangleq f\eta_2$.

Equations (11) and (13) can be now used to determine the light output levels for the laser:

$$L_0 = \frac{\eta_2 A_I I_B + (\eta_1 - \eta_2) I_T}{1 + A_I k_2} + \frac{A_I \bar{D} \eta_2 (I_{XOD} - k_2 I_{MOD})}{1 + A_I k_2} \tag{14a}$$

and

$$L_1 = L_0 + \eta_2 I_{MOD} . \tag{14b}$$

The quantity $A_I k_2$ represents the loop gain of the bias feedback loop. In a proper design the loop gain is necessarily very large ($A_I k_2 \gg 1$). Under this condition, together with the assumption that $\eta_2 \gg \eta_1$, we can simplify the expressions for the laser light output to

$$L_0 = \frac{1}{f}\left[I_B - \frac{I_T}{A_I}\right] + \frac{\bar{D}}{f}(I_{XOD} - k_2 I_{MOD}) \tag{15a}$$

$$L_1 = L_0 + \eta_2 I_{MOD} . \tag{15b}$$

The balance current, I_X, is incorporated in the circuit of Fig. 2 for the purpose of eliminating the dependence of the laser light levels on $\bar{D}$, the dc component of the data input signal. This is accomplished to first order by choosing $I_{XOD} = k_2 I_{MOD}$ so as to eliminate the second term in (15a). However, a dependence on $\bar{D}$ will reappear as a consequence of changes in the above-threshold conversion efficiency, k_2, that results from the drift of η_2 with time. In particular, if I_{XOD} is chosen to balance the circuit at some initial time when $\eta_2 = \eta_2^0$,

$$I_{XOD} = f\eta_2^0 I_{MOD} , \tag{16}$$

and if $\Delta\eta_2$ is defined to represent the subsequent drift in η_2,

$$\Delta\eta_2 \triangleq \eta_2 - \eta_2^0, \tag{17}$$

then it follows from eqs. (15) through (17) that the laser output levels can be expressed as

$$L_0 = \frac{1}{f}\left[I_B - \frac{I_T}{A_I}\right] - \bar{D} I_{MOD}(\Delta\eta_2) \tag{18a}$$

and

$$L_1 = L_0 + \eta_2 I_{MOD}. \tag{18b}$$

A principal function of the feedback loop in Fig. 2 is to eliminate the dependence of the optical output on the laser threshold current. However, there remain in (18) terms dependent on I_T, and we now consider their relative importance. Assume the laser is biased near threshold so that $I_{L0} \approx I_T$. Then, if the drift in η_2 is small so that $I_{XOD} = f\eta_2^0 I_{MOD} \approx f\eta_2 I_{MOD}$, it follows from (10a) and (13) that

$$I_{L0} \approx \frac{A_I[I_B + (k_2 - k_1) I_T]}{1 + A_I k_2} \approx I_T. \tag{19}$$

Thus,

$$\frac{I_T}{A_I} \approx \frac{I_B}{1 + A_I k_1}. \tag{20}$$

The quantity $A_I k_1$ is the *subthreshold* loop gain of the feedback loop, and if this gain is large ($A_I k_1 \gg 1$), then

$$\frac{I_T}{A_I} \ll I_B. \tag{21}$$

Therefore, the threshold current dependent terms in (18) are negligible. Thus, under the conditions that the feedback loop gain is large both above and below threshold ($A_I k_2 \gg A_I k_1 \gg 1$), the laser light output levels can be expressed simply as

$$L_0 \approx \frac{I_B}{f} - \bar{D} I_{MOD}(\Delta\eta_2) \tag{22a}$$

$$L_1 \approx \frac{I_B}{f} - \bar{D} I_{MOD}(\Delta\eta_2) + \eta_2 I_{MOD}. \tag{22b}$$

We can draw a number of conclusions with regard to above-threshold biasing from the results expressed in (22):

1. The laser light output levels, L_0 and L_1, are to first order independent of the subthreshold slope efficiency, η_1, and the threshold current, I_T.
2. If the above-threshold slope efficiency, η_2, is constant ($\Delta\eta_2 = 0$), the light output levels are independent of the data signal dc component, $\bar{D}$.
3. If η_2 drifts as a function of time or changes in environment, the light levels L_0 and L_1 will exhibit some dependence on $\bar{D}$.
4. In a proper design of the circuit represented by Fig. 2, the feedback loop gain must be large below, as well as above, the laser threshold.

2.2 Subthreshold prebiasing ($I_{L0} \leq I_T$)

When the laser is biased below threshold, it is necessarily the case that $I_{L0} = I_A \leq I_T$ and that $I_{L1} = I_A + I_{MOD} > I_T$; it therefore follows from (1), (2), and (10) that

$$L_0 = \eta_1 I_A \tag{23a}$$

and

$$L_1 = \eta_2(I_A + I_{MOD} - I_T) + \eta_1 I_T. \tag{23b}$$

Substitution of (23) into (9) then leads to the expression

$$I_A = A_I(I_B + \bar{D}I_{XOD} - f\{\eta_1 I_A + \bar{D}[\eta_2 I_{MOD} + (\eta_2 - \eta_1)(I_A - I_T)]\}) \quad (24)$$

and this result can be solved for I_A to obtain

$$I_A = \left[\frac{A_I}{1 + A_I k_1 + \bar{D}A_I(k_2 - k_1)}\right] \cdot \{I_B + \bar{D}[I_{XOD} - k_2 I_{MOD} + (k_2 - k_1)I_T]\}, \quad (25)$$

where $k_1 \triangleq f\eta_1$ and $k_2 \triangleq f\eta_2$.

As is the case for above-threshold biasing, the loop gain of the feedback circuit should be large both above and below the laser threshold ($A_I k_2 \gg 1$ and $A_I k_1 \gg 1$). Under these conditions, together with the assumption $\eta_2 \gg \eta_1$, (25) simplifies to

$$I_A \simeq \left(\frac{1}{k_1 + \bar{D}k_2}\right)\{I_B + \bar{D}[I_{XOD} - k_2(I_{MOD} - I_T)]\}. \quad (26)$$

To eliminate the dependence of the optical output levels on $\bar{D}$, I_{XOD} must be chosen so as to remove the dependence of I_A on $\bar{D}$. However, for subthreshold biasing a dependence on $\bar{D}$ will reappear in the event of drift in any of the laser parameters η_2, η_1, or I_T. If η_1^0, η_2^0 and I_T^0 denote the values of η_1, η_2, and I_T at the time when I_{XOD} is initially adjusted to cancel out the $\bar{D}$ dependence of I_A, then from (24) the appropriate value of I_{XOD} is

$$I_{XOD} \simeq k_2^0\left(\frac{I_B}{k_1^0} + I_{MOD} - I_T^0\right), \quad (27)$$

where $k_1^0 = f\eta_1^0$, $k_2^0 = f\eta_2^0$, and we have assumed that $\eta_2^0 \gg \eta_1^0$ and $A_I k_1 \gg 1$. For this choice of I_{XOD} the initial value of I_A is simply

$$I_A^0 \simeq \frac{I_B}{k_1^0}. \quad (28)$$

The expression for I_A at some time following the initialization of I_{XOD} is obtained by substituting (27) into (26), with the result

$$I_A = \left(\frac{1}{k_1 + \bar{D}k_2}\right)\left\{I_B\left(\frac{k_1^0 + \bar{D}k_2^0}{k_1^0}\right) + \bar{D}\,[k_2 I_T - k_2^0 I_T^0 - (k_2 - k_2^0)I_{MOD}]\right\}. \quad (29)$$

It then follows from (23) and (29) that for subthreshold prebiasing, and under the assumption of large feedback loop gain both above and below the laser threshold, the light output levels are given by

$$L_0 = \left(\frac{\eta_1}{\eta_1 + \bar{D}\eta_2}\right)\left\{\frac{I_B}{f}\left(\frac{\eta_1^0 + \bar{D}\eta_2^0}{\eta_1^0}\right)\right.$$
$$\left. + \bar{D}[\eta_2 I_T - \eta_2^0 I_T^0 - (\eta_2 - \eta_2^0)I_{MOD}]\right\} \tag{30a}$$

and

$$L_1 = \left(\frac{\eta_2}{\eta_1 + \bar{D}\eta_2}\right)\left\{\frac{I_B}{f}\left(\frac{\eta_1^0 + \bar{D}\eta_2^0}{\eta_1^0}\right)\right.$$
$$\left. + I_{MOD}(\eta_1 + \bar{D}\eta_2^0) - \eta_1 I_T - \bar{D}\eta_2^0 I_T^0\right\}. \tag{30b}$$

From these complex expressions, as compared to the simple results obtained in (22) for above-threshold biasing, we can draw the following conclusions with respect to subthreshold biasing:

1. The laser light output levels are not stabilized against individual variations in any of the parameters characterizing the laser (η_1, η_2, and I_T).

2. The light output levels will exhibit a dependence on the data signal average, $\bar{D}$, if changes occur in any one of the laser parameters.

III. MODULATION CURRENT COMPENSATION

As we demonstrated in Section 2.1, if the laser is prebiased above its threshold current, the circuit of Fig. 2 effectively stabilizes the optical output against variations in both the laser's subthreshold slope efficiency, η_1, and its threshold current, I_T. However, the light output levels remain sensitive to the above-threshold slope efficiency, η_2, and, as a consequence, to the average value of the input data, $\bar{D}$. One means of eliminating this remaining sensitivity is to compensate for changes in η_2 through control of the modulation source current, I_{MOD}. This approach has, in fact, already been proposed in specific designs.[6,8]

As in Section 2.1 we assume that the laser is prebiased above threshold ($I_{L0} \geq I_T$) so as to eliminate sensitivity of the optical output to η_1 and I_T. It is then apparent from (15) that the sensitivity of the light output levels to η_2 and $\bar{D}$ can be eliminated if the source current I_{MOD} can be continuously adjusted so as to hold the product $\eta_2 I_{MOD}$ constant. This can be accomplished by deriving a signal proportional to the difference between the ZERO and ONE light levels, and then using negative feedback to control I_{MOD} in a manner that stabilizes this difference. After a signal is obtained proportional to the difference $L_1 - L_0$, I_{MOD} is generated as

$$I_{MOD} = I_{REF} - \gamma(L_1 - L_0), \tag{31}$$

where γ is a constant characterizing the feedback loop controlling I_{MOD}, and I_{REF} is a modulation reference, or "baseline", current.

From (11) it follows that for above-threshold biasing

$$L_1 - L_0 = \eta_2 I_{MOD}. \tag{32}$$

Upon substituting this expression into (31) and then solving for I_{MOD}, we obtain

$$I_{MOD} = \frac{I_{REF}}{1 + \gamma\eta_2} \simeq \frac{I_{REF}}{\gamma\eta_2} \quad \text{for} \quad \gamma\eta_2 \gg 1. \tag{33}$$

In this equation the term $\gamma\eta_2$ represents the loop gain of the negative feedback loop controlling I_{MOD} and should necessarily be much greater than unity.

If we substitute (33) in (14) we obtain the following expressions for the laser light output levels:

$$L_0 = \frac{\eta_2 A_I I_B + (\eta_1 - \eta_2) I_T}{1 + A_I k_2} + \left(\frac{A_I \bar{D} \eta_2}{1 + A_I k_2}\right)\left(I_{XOD} - \frac{k_2 I_{REF}}{\gamma\eta_2}\right) \tag{34a}$$

and

$$L_1 = L_0 + \frac{I_{REF}}{\gamma}. \tag{34b}$$

If, as in Section 2.1, we assume that $\eta_2 \gg \eta_1$, $A_I k_2 \gg 1$ and $A_I k_1 \gg 1$, then, recognizing that $k_2 = f\eta_2$, it follows from (34) that

$$L_0 \approx \frac{I_B}{f} + \frac{\bar{D}}{f}\left(I_{XOD} - \frac{f I_{REF}}{\gamma}\right) \tag{35a}$$

and

$$L_1 = L_0 + \frac{I_{REF}}{\gamma}. \tag{35b}$$

It is apparent from (35) that the sensitivity of L_0 and L_1 to η_2 has been successfully eliminated. As in Section 2.1 the remaining dependence on $\bar{D}$ can be removed by the appropriate choice of I_{XOD}, namely,

$$I_{XOD} = \frac{f I_{REF}}{\gamma}. \tag{36}$$

The expressions for the ZERO and ONE light output levels then reduce to the very simple form

$$L_0 = \frac{I_B}{f} \tag{37a}$$

and

$$L_1 = \frac{I_B}{f} + \frac{I_{REF}}{\gamma}. \tag{37b}$$

Clearly these levels are now, to first order, independent of the laser parameters η_1, η_2, and I_T and the dc component of the data signal, $\bar{D}$.

Compensation of the modulation source current as described below can be implemented as illustrated in Fig. 3. Following the approach of Gruber, et al.,[6] the circuit of Fig. 2 is modified by the inclusion of high-speed buffers (B1 and B2), positive (B3) and negative (B4) peak detectors, and a summing amplifier (B5). The current I_{MOD} is developed at the output of the summing amplifier and is proportional to $L_1 - L_0$. The secondary negative feedback loop controlling I_{MOD} will act to hold $L_1 - L_0$ constant.

The modulation current feedback loop will have a characteristic response time. Consistent with our assumption that I_{MOD} is a parameter that changes slowly with respect to the response of the prebias

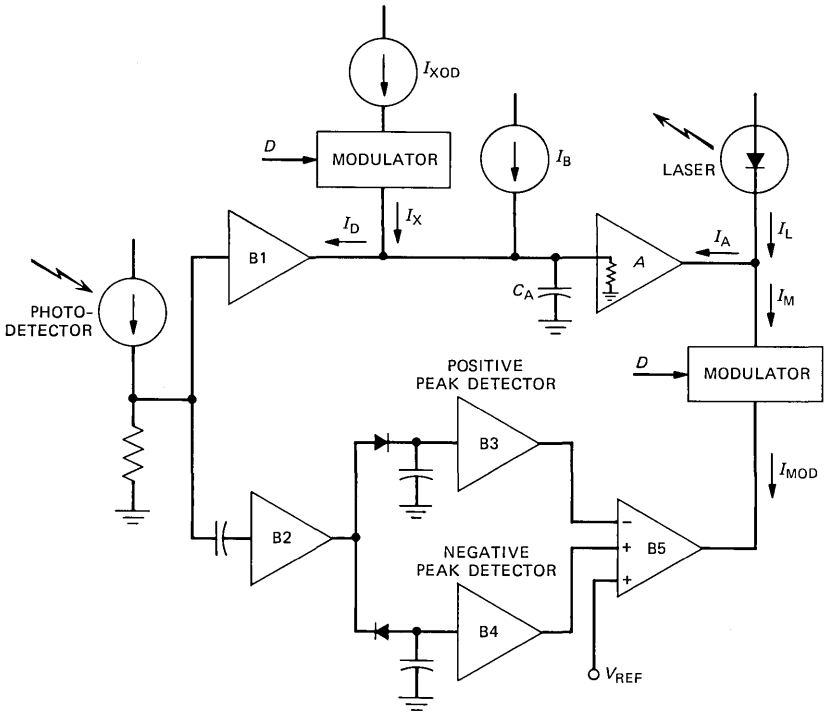


Fig. 3—An improved laser driver incorporating feedback control of both bias and modulation currents. Light output levels are set independently of laser parameters.

current feedback loop, the secondary feedback loop should respond slowly in comparison with the feedback controlling I_A.

IV. CONCLUSION

This paper has presented an analysis of a generalized method of negative feedback stabilized biasing and modulation of semiconductor lasers. Our objective was to evaluate the effectiveness of the stabilization and determine the critical feedback loop parameters. The analysis considered not only the direct influence of variations in η_1, η_2, and I_T, but also the effect of changes in the average modulation signal and the issue of biasing the laser above or below threshold.

For the more simple bias schemes reported to date, we showed that the laser light levels are susceptible to variation in any of the laser parameters when the laser is dc biased below threshold. When the laser is biased above threshold, only changes in η_2 affect the light output.

We also analyzed a method of stabilizing the difference $L_1 - L_0$ and thereby fixing laser light output independent of variations in any laser or modulation parameters. To maintain this independence the laser must, of course, be prebiased to remain above threshold under all expected conditions. Moreover, the optical output will still be sensitive to changes in the photodetector light-to-current conversion factor, f.

LIST OF VARIABLES

Device parameters

η_1	laser subthreshold differential slope efficiency
η_2	laser above-threshold differential slope efficiency
I_T	laser threshold current
f	photodetector light-to-current conversion factor
k_1	laser-photodetector subthreshold conversion efficiency $\triangleq f\eta_1$
k_2	laser-photodetector above-threshold conversion efficiency $\triangleq f\eta_2$

Modulation-related (rapidly changing) parameters

D	digital signal data (ONE or ZERO)
I_L	instantaneous total laser current
I_M	instantaneous modulation current: 0 or I_{MOD}
I_X	instantaneous balance current: 0 or I_{XOD}
I_D	instantaneous photodetector output current
L	instantaneous laser luminosity

Nominally DC (slowly changing) parameters

$\bar{D}$	average (dc) value of digital signal data
I_{L0}	logic ZERO laser current

I_{L1} logic ONE laser current
I_{MOD} modulation source current
I_{XOD} balance source current
L_0 logic ZERO laser luminosity
L_1 logic ONE laser luminosity
I_B bias current
I_A amplifier output current $\triangleq$ laser prebias current $= I_{L0}$
A_I amplifier current gain
I_{REF} modulation reference current
γ conversion efficiency of feedback loop controlling the modulation source current, I_{MOD}

Notation: For an arbitrary variable X, the bar notation $\bar{X}$ signifies the average or dc value of X.

REFERENCES

1. "Atlanta Fiber System Experiment," Special Issue, B.S.T.J., *57*, No. 6 (July–August 1978), pp. 1717–888.
2. M. I. Swartz, "Design and Performance of the FT3 Lightguide Transmission Medium," IEEE Int. Conf. Commun. Conf. Record, *1* (1981), pp. 6.1.1–8.
3. H. Buenning, H. W. Kreutzer, and F. Schmidt, "Subscriber Stations in Service Integrated Optical Broad Band Communications Systems," IEEE Trans. Commun., *COM-30*, No. 9 (September 1982), pp. 2163–71.
4. K. Asatani, R. Watanabe, K. Nosu, T. Matsumoto, and F. Nihei, "A Field Trial of Fiber Optic Subscriber Loop Systems Utilizing Wavelength Division Multiplexers," IEEE Trans. Commun., *COM-30*, No. 9 (September 1982), pp. 2172–84.
5. P. W. Shumate, Jr., F. S. Chen, and P. W. Dorman, "GaAlAs Laser Transmitter for Lightwave Transmission Systems," B.S.T.J., *57*, No. 6 (July–August 1978), pp. 1823–36.
6. J. Gruber, P. Marten, R. Petschacher, and P. Russer, "Electronic Circuits for High Bit Rate Digital Fiber Optic Communication Systems," IEEE Trans. Commun., *COM-26*, No. 7 (July 1978), pp. 1088–98.
7. R. G. Swartz, B. A. Wooley, A. M. Voshchenkov, V. D. Archer, and G. M. Chin, "An Integrated Circuit for Multiplexing and Driving Injection Lasers," IEEE J. Sol. State Circuits, *SC-17*, No. 4 (August 1982), pp. 753–60.
8. F. S. Chen, "Simultaneous Feedback Control of Bias and Modulation Currents for Injection Lasers," Electron. Letts., *16*, No. 1 (January 3, 1980), pp. 7–8.

AUTHORS

Robert G. Swartz, S.B.(EE), 1974, M.I.T.; M.S.(EE), 1975, Ph.D.(EE), 1979, Stanford University; Bell Laboratories, 1979—. Mr. Swartz has done research in a variety of circuit, device, and material areas including silicon molecular beam epitaxy, ultra-short gate MOSFETs, and piezoelectric transducers for medical imaging. Since arriving at Bell Laboratories, he has been particularly interested in circuits for laser control. Member, American Physical Society, IEEE, Sigma Xi; Secretary, 1981 International Solid-State Circuits Conference.

Bruce A. Wooley, B.S., 1966, M.S., 1968, and Ph.D., 1970 (Electrical Engineering), University of California, Berkeley; Bell Laboratories, 1970—. Since joining Bell Laboratories, where he is a member of the Microelectronics Research Department, Mr. Wooley's research interests have included monolithic broadband feedback amplifiers, integrated threshold logic circuits, analog-to-digital conversion, and integrated voiceband digital filters. Most re-

cently he has investigated integrated circuits and technology for tactile sensing and for very high speed communication systems. Fellow, IEEE; member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, IEEE Solid-State Circuits Council, and the Administrative Committee of the IEEE Circuits and Systems Society; Chairman, IEEE Solid-State Circuits and Technology Committee; Chairman, 1981 International Solid-State Circuits Conference; Guest Editor, IEEE Journal of Solid-State Circuits.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

Empirical Evaluation of Profile Variations in an MCVD Optical Waveguide Fiber Using Modal Structure Analysis

By A. CARNEVALE* and U. C. PAEK†

(Manuscript received November 15, 1982)

This paper presents a computer modeling evaluation of the effect on mode propagation of typical perturbations in the refractive index profile produced as a result of the MCVD process. Certain profile variations were eliminated, singly and/or collectively, and the resultant changes in modal structure analyzed. The most important factor in the deterioration of the bandwidth is the presence of a refractive index dip at the core center. Small shifts in the level of dopant were seen to broaden pulse shapes and thus reduce bandwidth. Much less effect on bandwidth is associated with the ripples in the profile, as is also true for dips in refractive index below the index of the cladding near the core-cladding interface. A perpendicular rise, or step, in index of refraction at the core-cladding interface reduces the number of altered modes near the cladding.

I. INTRODUCTION

In a previous paper[1] we reported on calculation of the modal structure as a function of wavelength (λ) of an optical waveguide produced by the Modified Chemical Vapor Deposition (MCVD) process using an exact numerical procedure. This technique[2] has been used extensively to investigate various theoretical aspects of optical waveguide propagation.[3-6] We observed excessive splitting and scattering in the

* Bell Laboratories. † Western Electric.

effective indices (N_e) and the group indices (N_g), which lead to a reduction of the bandwidth (BW) at any particular λ.[1] We concluded that the profile variations inherent in the MCVD process were responsible for the reduction of the BW. This paper analyzes in detail the same profile considered there.

The profile is shown in Fig. 1. The data, obtained by the laser beam refraction method,[7] indicates the change in refractive index (ΔN_r) as a function of the unnormalized radius (r). $\Delta N_r = 0$ corresponds to the cladding, in this case, SiO_2. The major deviations (imperfections) of this MCVD profile from a power law (α) profile are common to most profiles produced by the MCVD process. These include:

1. Region A in the plot, usually referred to as "BURNOFF" or "BURNOUT", which indicates a precipitous reduction in the refractive index at the core center. This reduction is believed to be due to the vaporization of dopant components, such as GeO_2, during the high-temperature collapse phase of the process.

2. Region B, which extends over most of the profile, containing what are generally called "RIPPLES", whose width and amplitude

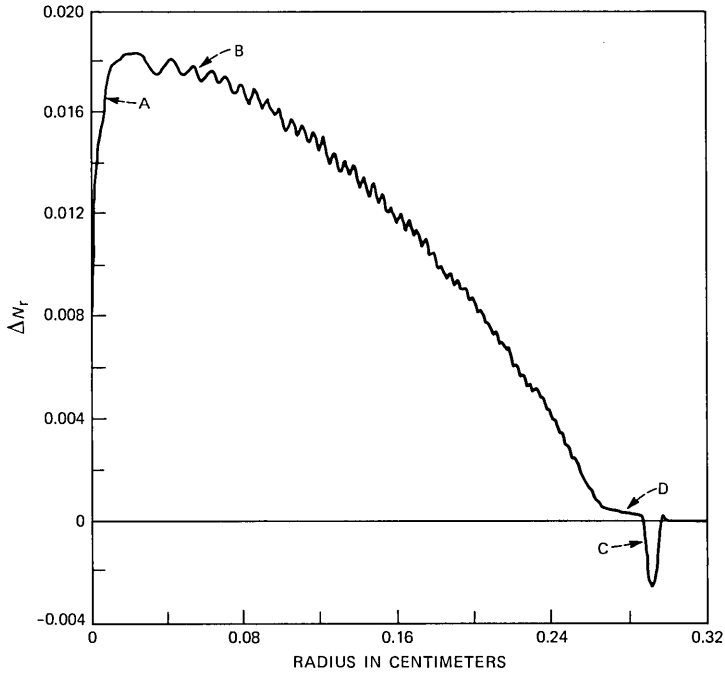


Fig. 1—MCVD profile. The region A denotes the burnout while the region B, which extends over most of the profile, shows the characteristic ripples referred to in the text. Regions C and D refer to the dip and the step, respectively. The radius data are not normalized.

increase toward the core center. These ripples are a direct result of the compositional gradient across the MCVD layer, burnoff of GeO_2 at the surface of the layer, and the finite number of discrete layers of changing composition (typically 50).

3. The sharp "DIP", C, near the core-cladding interface caused by the use of an index-reducing dopant in the deposited barrier layer, in this case, boron. In some special cases this "depressed cladding" feature can have a greater width than shown.

4. The "STEP", D, at the cladding.

The above deviations constitute the major obvious perturbations of the MCVD profile from an ideal profile. More subtle deviations, which are usually ignored, will be seen later in this report.

II. METHOD

It was deemed necessary and sufficient to find the effective indices and the group indices for the:

1. Meridional modes $TE_{0,q}$, and $TM_{0,q}$
2. Helical modes $HE_{1,q}$, and $EH_{1,q}$ over some wavelength region of interest.

The subscripts above denote the angular and radial mode number (m, q) respectively. Thus we conducted the modeling process at 11 specific wavelengths itemized below in Table I.

2.1 "As is" case

Initially, N_e and N_g of the modal groups at $m = 0$ and $m = 1$, were obtained for each of the λ's listed in Table I, operating on the MCVD profile as it exists; i.e., on an "as is" basis. Figure 2 shows examples of the distribution of N_g vs N_e for $\lambda = 0.82$ (a), 1.32 (b), and 1.55 (c) μm, respectively. We note the counterclockwise rotation of the data as a function of increasing λ; this rotation has been commented on in our early work.[4] The profile was shown to remain constant with λ, but the group index vs effective index changes in this manner owing to the dependence of bandwidth on α.

We further note the clear separation (splitting) of the *HE* modes (Δ) from all other modes (○, ●); this separation decreases as λ increases. The two groups appear to be moving roughly in parallel but

Table I—Wavelengths used for modeling process

Test No.	1	2	3	4	5	6	7	8	9	10	11
Wavelength (μm)	0.6328*	0.70	0.82*	0.90	1.00	1.10	1.20	1.23	1.32*	1.40†	1.55*

* Of current interest in engineering applications. All others listed are solely for the purpose of parameterizing the MCVD deviations from an ideal profile, as a function of wavelength, if possible.

† Water peak.

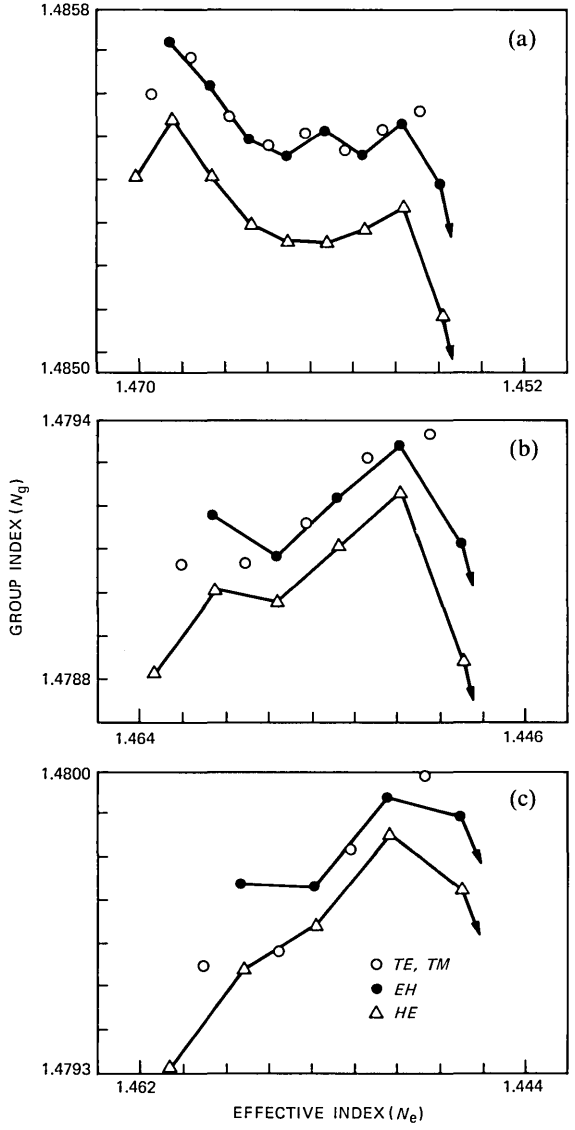


Fig. 2—N_g vs N_e—"as is" case. (a) $\lambda = 0.82$ μm. (b) $\lambda = 1.32$ μm. (c) $\lambda = 1.55$ μm.

displaced in time. This can result in two or more peaks, depending on any mode mixing, in the output pulse of the fiber, and a reduction in the BW at any λ. BW reduction would be more severe at short λ's, where the splitting is greater. Last, we observed the drop-off (lower N_g) of those high-order modes near the cladding. These modes have been altered because of their proximity to the cladding. The plots of

Fig. 2 can be compared to the ideal case of a power law profile using the computed α, ΔN, and r derived from the MCVD profile.[1] A plot of N_g at each λ results in a shape similar to those shown in the MCVD plot by (○, ●); i.e., other than the *HE* (Δ) modes.

2.2 "No ripple or dip" case

To analyze the effect of the MCVD imperfections on the modal structure, each was removed separately. Our first step was to remove the ripple in the section (B) between the burnout (A) and the step (D) in Fig. 1. Past experience in α curve fitting to MCVD profiles indicated that rarely can one fit section B with a single α. Because we did not want to destroy any short-term (local) variations in ΔN and/or α, we partitioned the B region into eight equal radial segments, and fitted each segment with a least squares parabola. We then used the fitted curve as the profile in each segment. The index of refraction data shown in Fig. 1 is collected at a number of discrete points and is generally not a continuous curve.

Next, the dip (C) was removed by connecting the end points with a two-point straight line, thus making the step (D) continuous. This correction caused negligible change in the spectrum and so for brevity we eliminated the ripple and dip in one operation.

One example, which is typical of all λ's tested, is given in Fig. 3 at $\lambda = 0.6328$ μm. Figure 3a is the modal display on an as is basis, while Fig. 3b is the result for no ripple or dip. The as is plot is consistent with those λ's shown in Fig. 2, i.e., the rotation of the data, the drop-off of the high-order modes, and, most important, the splitting of *HE* from all other modes. The relatively short λ of 0.6328 μm makes the N_e and N_g very sensitive to the short-term variations in the profile. We see the development of some jagged peaks and depressions in the display as we approach the cladding. Since it is the high-order modes that are most affected, imperfections in the profile near the cladding are most likely responsible for the development of these sharp features. It could be the step D, or as seems more likely, a shift in α or ΔN in the latter part of section B in the profile. Identification of the parameter responsible (ΔN or $\Delta\alpha$) is difficult. In any case, a change in either results in shifting N_g in the same direction. We can see two pronounced shifts at $r = 18$ and $r = 22$ in Fig. 1. If we examine the fine structure of the profile, we can detect other such shifts over the full range of the radius. Although these shifts appear to be small, they can cause large changes in the eigenvalues and, thereby, in N_g. We can therefore associate the ragged appearance of N_g in our plots with very small changes in ΔN_r.

Elimination of the ripple and the dip (Fig. 3b) reduces the scatter of the data and lessens the effect of the shifts in ΔN, probably because

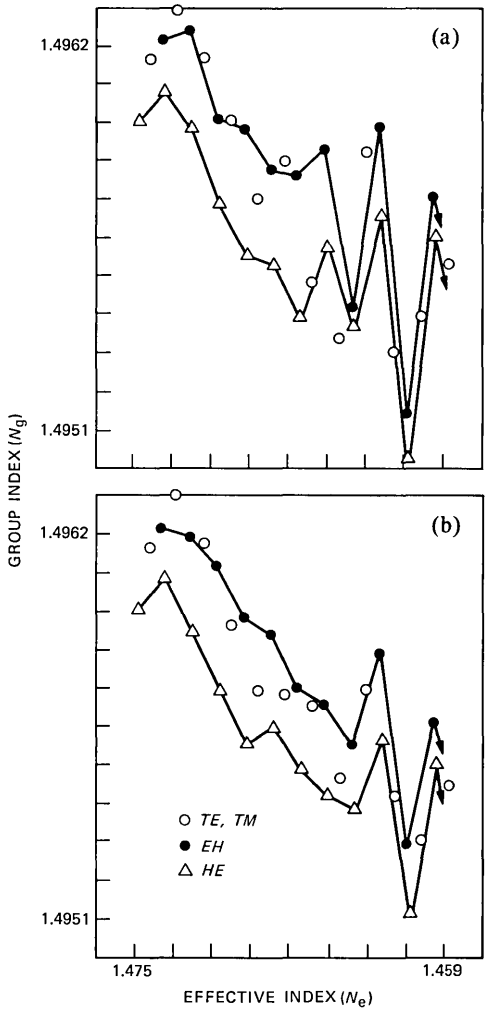


Fig. 3—Plot of N_g vs N_e for $\lambda = 0.6328$ μm. (a) "As is" case. (b) Ripple and dip removed.

it acts as a smoothing mechanism. However, it does not eliminate or reduce the separation of the *HE* modes and therefore is not the major cause of reduced BW resulting from this profile. The data in Figs. 3a and b appear to have approximately the same slope, but the slope in Fig. 3b is more clearly defined, probably as a result of the smoothing action.

2.3 "No burnout" case

The next MCVD deviation eliminated was the dip in the index of refraction at the core center, i.e., the burnout region (A) in Fig. 1. To

accomplish this, we found the maximum ΔN in the profile and extended it in a horizontal line to the core center. The rest of the profile was kept intact, thus preserving the original ripples, dip, and step. Figure 4 shows the results at $\lambda = 1.23$ μm. Figure 4a shows the results on an as is basis, while Fig. 4b shows the case without burnout.

The as is result at this λ is again consistent with our earlier results. At this longer λ, N_e and N_g are much less sensitive to the short-term variations in the profile, which is evident in the relatively smooth plot of N_g vs N_e as compared to the same plot (Fig. 3a) at $\lambda = 0.6328$ μm.

The elimination of the burnout (Fig. 4b) has removed the displacement in time of the *HE* modes and all modal groups are now traveling together. This remarkable result was observed at all λ's (see Table I) tested. Except for considerations of α optimum (α_{opt}), the burnout appears to be of paramount significance in the deterioration, or optim-

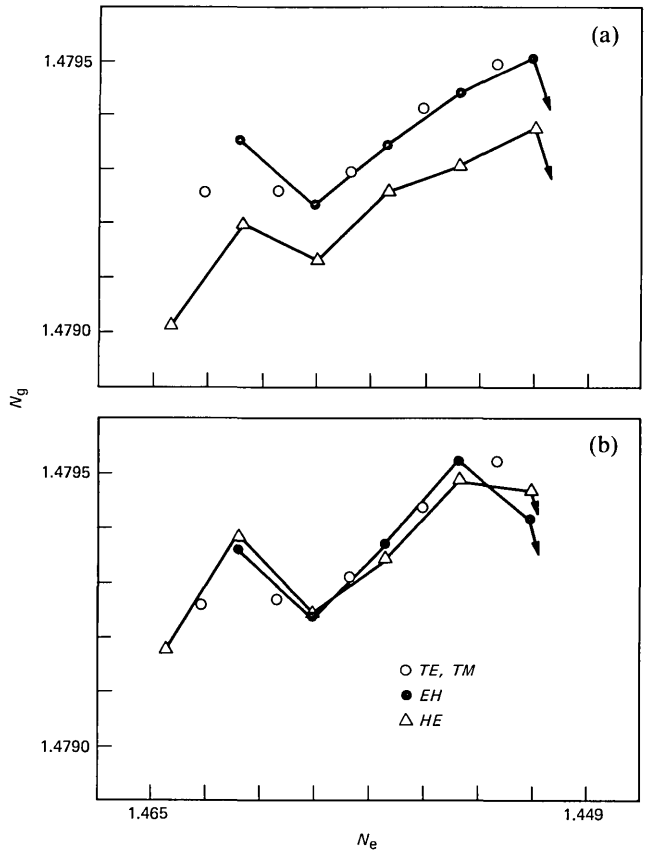


Fig. 4—N_g vs N_e for $\lambda = 1.23$ μm. (a) "As is" case. (b) "No burnout" case.

ization, of the BW. We should realize that the burnout region in this MCVD profile is much less than that seen in most MCVD profiles.

2.4 "Inverted burnout" case

Because of the previous results we decided to investigate a phenomenon sometimes seen in MCVD profiles. This defect can occur as a result of overcompensation of dopant in attempting to reduce or eliminate burnout. A study of some MCVD profiles[8] in which this defect was present indicated that we could simulate this characteristic by simply inverting the burnout. We found the difference of ΔN from the maximum ΔN and added this difference to the maximum, over the burnout region A in Fig. 1.

Figure 5 shows one result at $\lambda = 1.20$ μm, which was typical of all

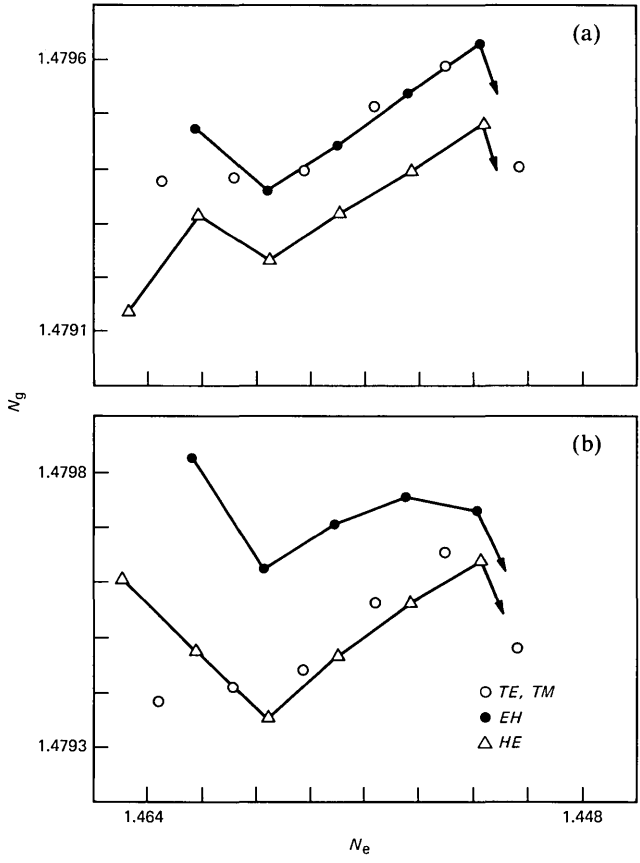


Fig. 5—N_g vs N_e for $\lambda = 1.20$ μm. (a) "As is" case. (b) "Inverted burnout" case.

λ's tested. Figure 5a is the modal display on an as is basis (which includes the burnout), and Fig. 5b is the result for the inverted burnout (overcompensation).

The as is result is consistent with our previous observations on similar spectra at other λ's.

Inverting the burnout (Fig. 5b) has displaced the *EH* modes from all the rest, and further, this displacement is much greater than the burnout effect, and thus is far more serious than burnout in the decrease of the BW. Our calculations indicate that if both burnout and overcompensation are present simultaneously (which has been observed at times), both splittings would also occur, i.e., the two effects would not be cancelled). Thus we would expect three or more peaks, depending on mode mixing, in the output pulse of the fiber.

2.5 "No step" case

One last perturbation in the MCVD profile remains to be examined, namely, the step (D) near the cladding. We used the least squares fitted curve for the ripple section of the profile nearest the step. We eliminated the step D (and also the dip C) by using the derived equation to extrapolate paired values of r, ΔN_r to the cladding. At $\Delta N_r = 0$, the radius is then renormalized to 25 μm. One comprehensive result is given in Fig. 6 at $\lambda = 0.70$ μm, where all plots are on identical scales and are called:

1. As is
2. No ripple or dip
3. No ripple, dip or burnout
4. No ripple, dip, burnout, or step.

The as is modal display (A) is consistent with our previous observations at other λ's on this basis. We see the separation of the *HE* modes due to the burnout and the increased sensitivity to the short-term variations in this profile at this relatively short λ. We note the development of the sharp features in the distribution as we approach the cladding; this is very similar to the comparable spectrum (Fig. 3a) at $\lambda = 0.6328$ μm.

Removing the ripple and dip (Fig. 6b) provides the same smoothing mechanism previously cited (see Fig. 3b for example).

Figure 6c is the modal distribution realized when the ripple, dip, and burnout have been removed, retaining the step. We see that all modes are now travelling together (no *HE* separation), and that we are now approaching the distribution to be realized for an ideal profile, except for the short-term shifts in ΔN_r in the profile.

Figure 6d is the modal distribution obtained when the ripple, dip, burnout, and step have been eliminated. This distribution represents the closest approach to an ideal profile. The severe drop-off in N_g of

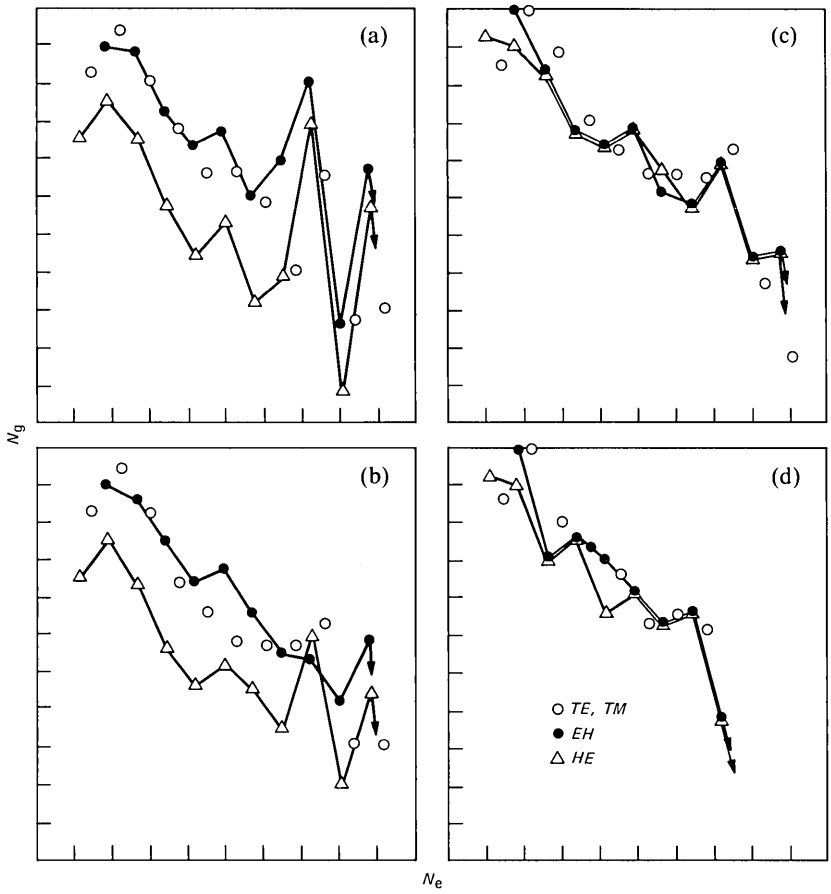


Fig. 6—N_g vs N_e for $\lambda = 0.70$ μm (Scales are identical in all sections). (a) "As is" case. (b) No ripple or dip case. (c) No ripple, dip, or burnout (step is in). (d) No ripple, dip, burnout, or step.

the high-order modes can lead to a reduction of the BW. Thus we see that the step is beneficial when we are attempting to optimize the BW. Note that there are many more modes in the distribution (Fig. 6c) than there are in our closest approach to an ideal profile (Fig. 6d). In this example, the step will introduce approximately 75 more modes to the distribution as compared to the ideal profile. Figure 7 uses data obtained at $\lambda = 0.6328$ μm (A) and $\lambda = 1.32$ μm (B) to show the steps beneficial effect. Each plot represents the results for the MCVD profile with the ripple, dip, burnout, and step removed. The solid line represents N_g values for an ideal profile using the computed values of α, ΔN, and r derived from the MCVD profile, at each λ. The dashed line indicates the path taken by the high-order modes, which have been

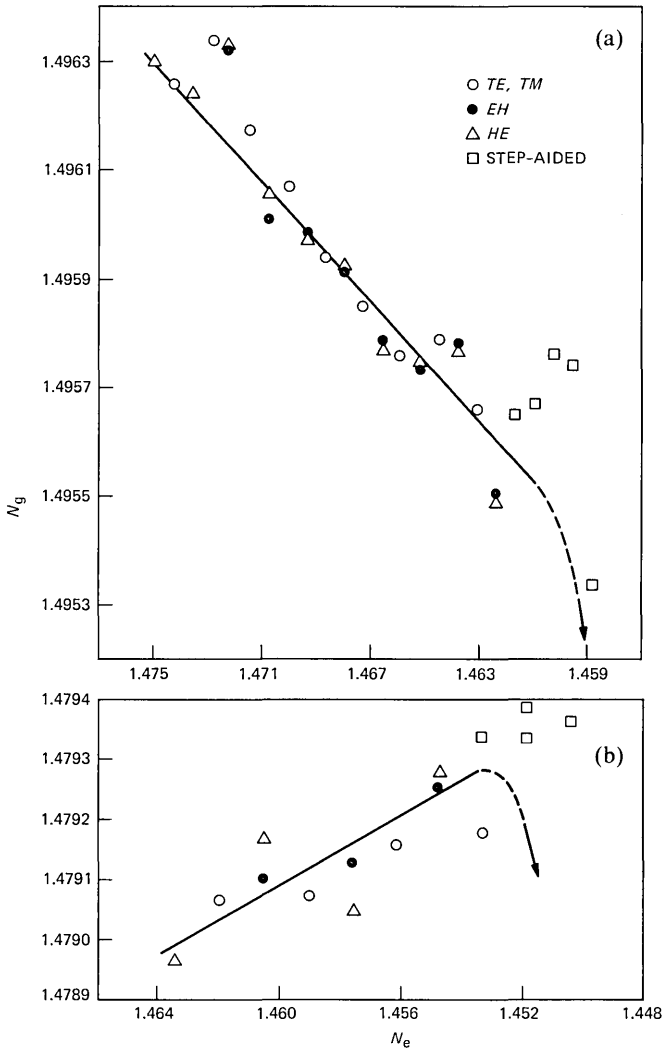


Fig. 7—N_g vs N_e. The solid line depicts an ideal profile using the pertinent parameters derived from the MCVD profile. (a) $\lambda = 0.6328$ μm. (b) $\lambda = 1.32$ μm.

altered by their proximity to the cladding. The filled squares in each plot show the N_g of some of these altered modes if the step is retained. Since this is the region of high modal density, the squares represent an addition of $\approx$100 modes to the modal distribution at $\lambda = 0.6328$ μm and $\approx$50 additional modes at $\lambda = 1.32$ μm. These additional modes can only be helpful when we are trying to optimize BW at any λ. We can now predict that the BW at these two λ's would be a function of the slope of the line.[4]

III. DISCUSSION

In our earlier work,[4] we have established that for a fixed profile that is a function of α, ΔN, and r, a least squares linear fit of N_g vs N_e, ignoring the high-order modes, will show an increasing slope for increasing λ. At a given λ, we have defined α_{opt} as that α which produces a slope of zero. A slope of zero will generate the minimum spread in N_g, in the group velocities (V_g), and of course, in the delay times. Using that criterion we can see[1] that this MCVD fiber should have an optimum operating wavelength (λ_0) of about 1.00 μm. Independent BW measurements[9] have established a BW performance of $\approx$1.5 GHz·km over the range

$$1.06\ \mu\text{m} \leq \lambda \leq 1.35\ \mu\text{m}$$

with a maximum of $\approx$2 GHz·km at $\lambda = 1.23\ \mu$m. We have reproduced the output pulse displays from this work in Fig. 8. In this figure the number in each plot is the λ in μm at which the distribution was obtained. A study of these distributions shows at least two peaks (broadening) at the short λ's, which appear to merge as λ increases. In all the pulses we see the common characteristic of a trailing tail. This tail is generated by the slowest modes in the modal distribution; at $\lambda = 0.7\ \mu$m (Fig. 6a) these would be the *low-order* modes, while at $\lambda = 1.55\ \mu$m (Fig. 2c) these would be the *high-order* modes. In Fig. 7, we have shown that the step at the core-cladding interface slows down the modes by raising the group indices. We have not attempted to derive the optimal step to optimize the BW for this profile. The reader should bear in mind that *all* modes will also be slowed down by increasing α.

In the "as is" section of this report (Section 2.1) we have seen that the burnout displaces the *HE* modes from all other modes and that the two modal groups will travel in parallel but be displaced in time. This will lead to at least two peaks in the output pulse shapes. Further, we have noted that the displacement in time will decrease as a function of increasing λ. This naturally suggests that the two peaks should merge (depending on instrumental resolution) in the output pulse of λ is increased (see Fig. 8).

IV. REDUCTION OF DATA

We now consider the separation of the various modal groups at several λ's (Table I) for the three conditions:

1. As is (includes burnout)
2. Inverted burnout (overcompensation)
3. No burnout,

and, for 2 and 3 above, the rest of the profile remains unaltered. Thus we are concerned here only with the index of refraction dip at the core

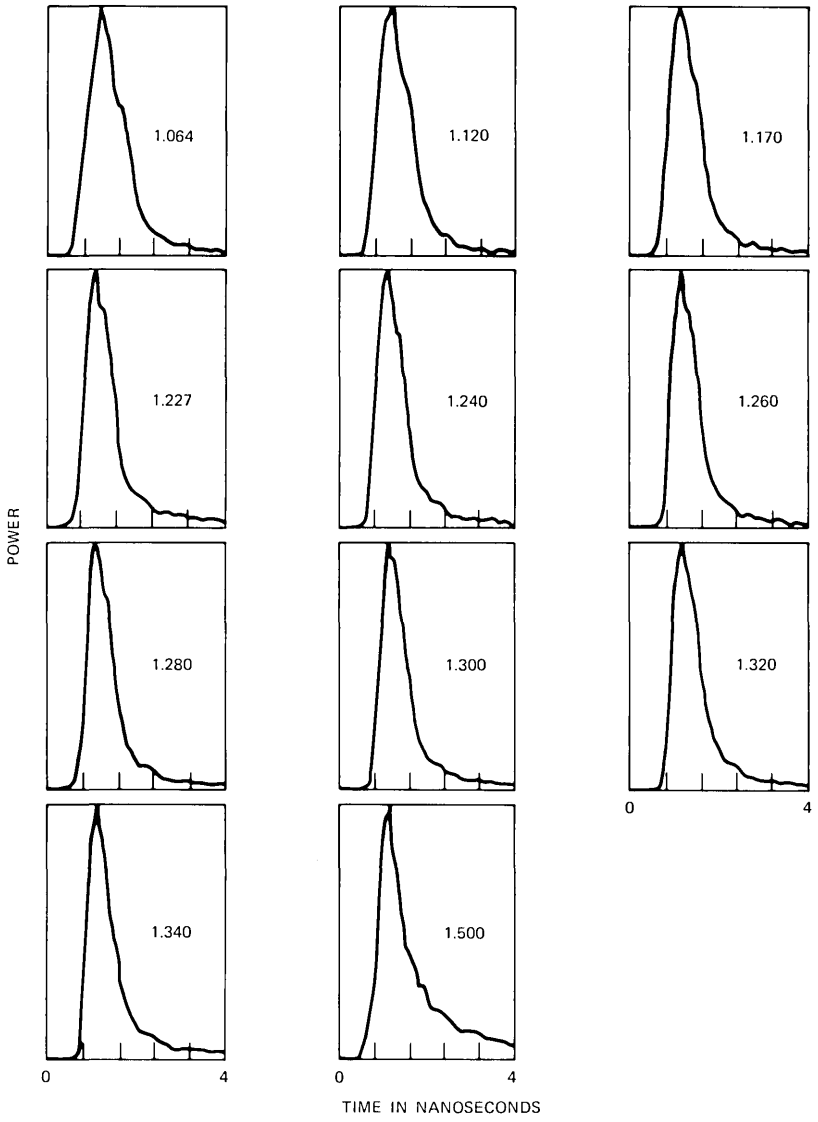


Fig. 8—Power vs time, as a function of the λ (μm) listed in each plot, for the specific MCVD fiber used in this study. Reprinted here, with permission, from Ref. 9.

center. We have previously stated that the separation of the modal groups appears to decrease for increasing λ, and this is true for 1 and 2 above. No such pattern was detected for the no burnout case. To codify these results over the range of λ's listed in Table I, we did the following.

At all λ's tested (Table I, Figs. 2a through 6a) we could generate at

least five distinct *EH-HE* splittings. We therefore found the average splitting ($\overline{\Delta N}_g$) over the first five such splittings at all λ's for each of the three cases listed above. These results are displayed in Fig. 9 as a function of λ. In the figure, the curves are defined as follows:

(A) Inverted burnout—*EH* modes separated from and moving slower than the comparable *HE* modes;

(B) As is—i.e., with burnout. *HE* modes separated from and moving faster than the comparable *EH* modes;

(C) No burnout—no real separation of modal groups—probably a random fluctuation about some relatively small average;

and 30 on the $\overline{\Delta N}_g$ axis is ≈1 ns.

Overcompensation (Curve A) is about 50 percent more deleterious in inducing splitting and is always worse in that respect than the burnout (Curve B). Further, if both phenomena are present, then the

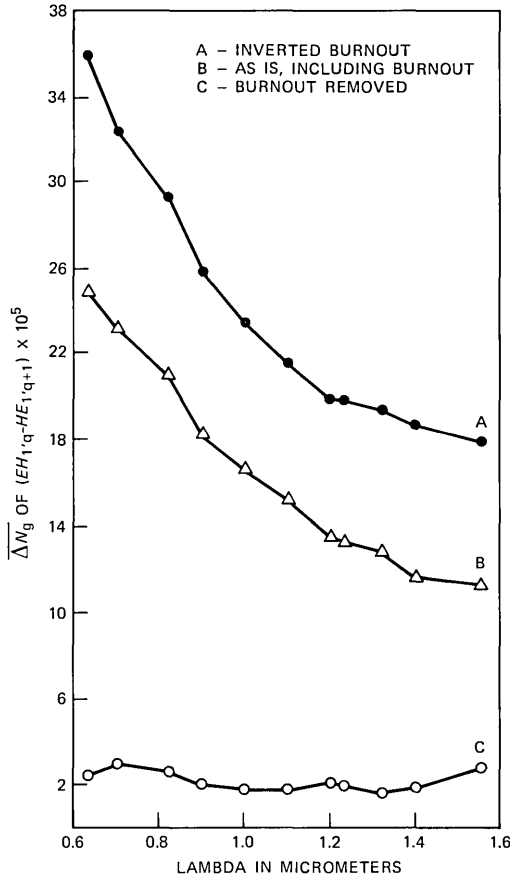


Fig. 9—The average splitting of N_g for the first five *HE-EH* modes, plotted as a function of λ, for the MCVD profile. The measurement 30 on the y axis is ≈1 ns.

total splitting is the sum of the two curves. Three distinct peaks (or more) can be produced in the output pulse of the fiber, and these modal groupings might be, among others:

1. *HE*
2. *EH*
3. *TE, TM*
4. MIXED (*HE, EH*).

In practice, preform makers[8] do not consider preforms that show significant overcompensation for fiber drawing. Experience has shown that this perturbation in the profile yields poor BW in drawn fibers.

Removal of the burnout (Curve C) has reduced the splitting of the *EH-HE* modes to a minimum that is constant over the full range of λ's investigated. This residual splitting can now be attributed to the perturbations still present in the profile, ripple, dip, and short-term ΔN_e, $\Delta\alpha$ variations, etc. If this condition could be realized in practice, i.e., burnout eliminated, the BW at each λ would be a function of α_{opt}.[4]

In our computational method,[2] we obtain the eigenvalues (N_e) at three closely spaced wavelengths, where

$$\lambda_1 < \lambda_2 < \lambda_3,$$

and λ_2 is the operating wavelength. The eigenvalues are inserted in the numerical derivative to obtain the group indices for λ_2 by

$$N_{g_2} = N_{e_2} - \lambda_2 \left| \frac{N_{e_1} - N_{e_3}}{\lambda_1 - \lambda_3} \right|.$$

Errors in the determination of N_e can be cancelled or amplified in the above formula. Thus we chose to obtain the average splitting in effective index ($\bar{\Delta}\bar{N}_e$) only at λ_2. The analogous display to Fig. 9 for $\bar{\Delta}\bar{N}_e$ is given in Fig. 10. In the figure, curves A, B, and C have the same connotation as before. We observe qualitatively the same trends for $\bar{\Delta}\bar{N}_e$ as for $\bar{\Delta}\bar{N}_g$ in Fig. 9. This information may be of some use to those computational methods[10] that calculate N_g directly from N_e.

V. CONCLUSION

We have examined the effect of various perturbations on the BW of an MCVD prepared fiber. We conclude that the primary reason for the deterioration of the BW in the MCVD as is profile is caused by the severe change in refractive index at the core center. This defect produces modal groupings that are displaced in time by as much as 0.5 ns, more or less, depending on the operating wavelength (λ_0). Short λ's are particularly sensitive to this defect, but even for longer λ's, the burnout will seriously affect attempts to maximize the BW. If burnout is minimized or eliminated in the MCVD process, we believe the next most important goal should be to optimize α at λ_0. Eliminating local

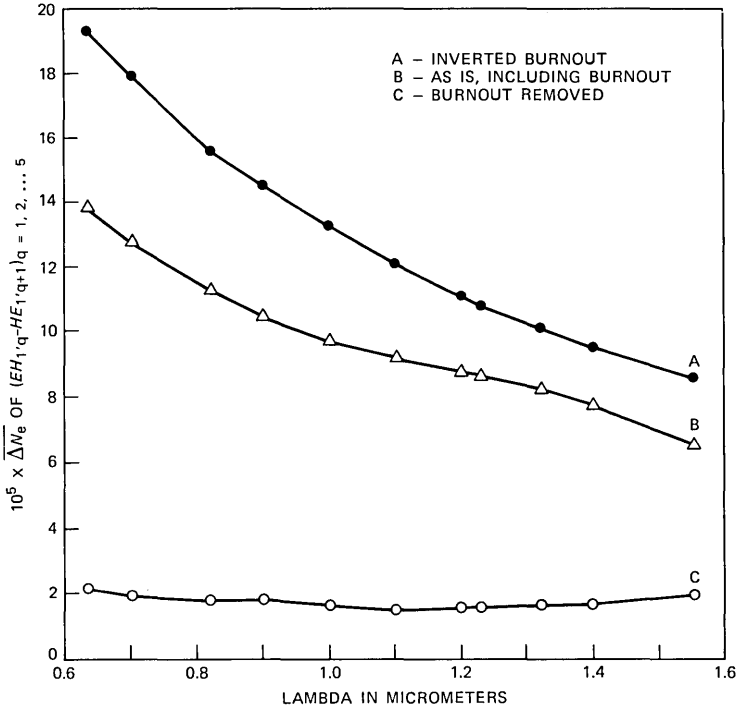


Fig. 10—Same as Fig. 9, except this is for the effective indices at the plotted λ.

shifts in ΔN would be the next priority, followed by reducing the amplitude of the ripples. This last is probably the least harmful in causing deterioration of BW. The step at the core-cladding interface appears to be beneficial in that it significantly *reduces* the number of high-order modes that have been altered by their proximity to the cladding. The works of Vassel[11] suggest that a perpendicular rise in the index of refraction at the core-cladding interface will have the same effect as the step (D) in Fig. 1.

In theory,[1,4] this MCVD fiber should have a maximum BW at $\lambda_0 \approx 1.00$ μm. The experimental λ_0 was determined[9] to be 1.23 μm, and this shift to the longer λ_0 was verified by the statistical study conducted in Ref. 1. In an ideal profile, the governing factor for the optimization of BW is the determination of α_{opt} for λ_0.[4] In the MCVD profile this criterion has been overridden by the excessive broadening of the output pulse induced by the burnout causing the theoretical λ_0 to shift longer λ's. Even at longer λ's the effect of the burnout on BW is significant and may preclude the optimization of BW beyond the present state of the art.

Optimization of BW, based on optimum parameters (α_{opt}, ΔN_{opt}, etc.) derived from theoretical (ideal) concepts, will have limited appli-

cation to MCVD fibers. This limited applicability can explain to a great extent why workers in the field endeavoring to optimize BW utilizing concepts based on the theory cannot reproduce these bandwidths.

The literature is replete with studies of the effects on BW caused by disturbances in the profile such as burnoff, ripple, and steps in the index of refraction at the core-cladding interface. Most of these studies (including the present work) describe a cause and effect relationship. We are not aware of any study correlating these effects in terms of local changes in the α value, which the burnoff and/or ripples cause, resulting in a deterioration of the BW.

In our most recent work, we have tried to determine the fundamental cause of mode splitting. In this project we believe we have been successful. The fundamental cause of mode splitting is easily applied to the width and depth of the burnoff, and to the width (variable or otherwise), amplitude, and frequency of the ripples. The project is incomplete at this time.

VI. ACKNOWLEDGMENTS

This work was supported by P. A. Fleury, M. I. Cohen, R. J. Klaiber, and L. S. Watkins. We are encouraged by the enthusiasm of the materials scientists engaged in the lightguide program at Bell Laboratories, Murray Hill and Atlanta, and at the Engineering Research Center (ERC), Western Electric Co., Princeton. We are indebted to the computer personnel at ERC, Western Electric Co., Princeton, without whose strong support and cooperation this task would not have been completed.

REFERENCES

1. A. Carnevale, and U. C. Paek, "Modal Structure of an MCVD Optical Waveguide Fiber," B.S.T.J., *62,* Part 1 (July-August 1983), pp. 1415–31.
2. G. E. Peterson, A. Carnevale, U. C. Paek, and D. W. Berreman, "An Exact Numerical Solution to Maxwell's Equation for Lightguides," B.S.T.J., *59,* No. 7 (September 1980), pp. 1175–96.
3. G. E. Peterson, A. Carnevale, and U. C. Paek, "Comparison of Vector and Scalar Modes in a Lightguide with a Hyperbolic Secant Index Distribution," B.S.T.J., *59,* No. 9 (November 1980), pp. 1681–91.
4. G. E. Peterson, A. Carnevale, U. C. Paek, and J. W. Fleming, "Numerical Calculation of Optimum α for a Germania-Doped Silica Lightguide," B.S.T.J., *60,* No. 4 (April 1981), pp. 455–70.
5. U. C. Paek, G. E. Peterson, and A. Carnevale, "Dispersionless Single-Mode Lightguides with α Index Profiles," B.S.T.J., *60,* No. 5 (May-June 1981), pp. 583–98.
6. U. C. Paek, G. E. Peterson, and A. Carnevale, "Electromagnetic Fields, Field Confinement, and Energy Flow in Dispersionless Single-Mode Lightguides With Graded-Index Profiles," B.S.T.J., *60,* No. 8 (October 1981), pp. 1727–43.
7. L. S. Watkins, "Laser beam refraction traversely through a graded-index preform to determine refractive index ratio and gradient profile," Applied Optics, *18* (July 1, 1979), pp. 2214–22.
8. S. R. Nagel, K. L. Walker, and S. G. Kosinski, private communication.
9. L. G. Cohen, and S. Jang, unpublished work.

10. T. Lenahan, "Calculation of Modes in an Optical Fiber Using the Finite Element Method and EISPACK," B.S.T.J., *62*, No. 9, Part 1 (November 1983).
11. M. O. Vassel, "Calculation of Propagating Modes in a Graded-Index Optical Fibre," Opto-Electron, *5* (1974), p. 271.

AUTHORS

Anthony Carnevale, B.S. (Physics), 1960, Fairleigh Dickinson University; Bell Laboratories, 1969—. At Bell Laboratories, Mr. Carnevale has been engaged in work on nuclear magnetic resonance, electron paramagnetic resonance, and computer software. For the last four years, his work has been devoted to fiber optics.

Un-Chul Paek, B.S. (Engineering), 1957, Korea Merchant Marine Academy, Korea; M.S., 1965, Ph.D., 1969, University of California, Berkeley; Western Electric, 1969—. At the Western Electric Engineering Research Center, Princeton, N. J., Mr. Paek has been engaged primarily in research on laser material interaction phenomena and fiber optics. Member, Optical Society of America, American Ceramic Society, Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

A Comparison of Line Difference Predictions for Time-Frequency Multiplexing of Television Signals

By R. L. SCHMIDT*

(Manuscript received January 27, 1983)

Studies are presented of a scheme utilizing time-frequency multiplexing (TFM) to multiplex two television signals on a microwave radio channel. This scheme transmits alternate lines of the television signal at full bandwidth, along with a differential signal for the remaining lines. The differential signal is derived by subtracting the true value of each picture element from a prediction for that value based on surrounding elements located in the transmitted lines. This differential signal is then placed on a carrier, and frequency multiplexed with the baseband signal. In this way, two television lines of information can be transmitted in one line time, allowing a second similarly constructed signal to be transmitted during the vacated line. Thus, the television signals will be time interleaved on a line-to-line basis. This paper describes the comparison of two different prediction algorithms used to construct the line-differential signal. Time and frequency domain analyses were carried out on the computer and then verified with hardware testing.

I. INTRODUCTION

Recently, a time-frequency multiplexing (TFM) system was described[1] as a method for multiplexing two National Television System Committee (NTSC) color television signals onto a single microwave radio channel. This system proposes to transmit alternate

* Bell Laboratories.

lines of a television signal at full bandwidth, and the remaining lines by means of a band-limited line-differential signal formed by using a prediction obtained from adjacent, full bandwidth lines. This differential signal is then placed on a carrier and frequency multiplexed with a full bandwidth television line. In this way two lines are sent in one line period, and, by time multiplexing, two pictures can be sent on a single microwave radio channel.

The reconstructed picture in the receiver is then a combination of alternate lines of full bandwidth video, and lines regenerated from the prediction and the band-limited line-differential signal. Consequently, a more accurate prediction would require less differential signal information. Stated another way, better picture quality can be attained for a given bandwidth limitation of the differential signal. The proposed differential signal band limiting for the TFM system is 3 MHz, and, within this constraint, an improved prediction would allow the transmission of more detailed pictures.

The originally proposed system had a prediction based on an average of adjacent picture elements of the same color subcarrier phase, from the previous and upcoming lines (Fig. 1). These elements were chosen because we did not have to take into account the color subcarrier, making this prediction the simplest to implement. However, since the prediction elements had to be of the same color subcarrier phase, they were not the elements closest to the element being predicted. This would create a larger differential signal on sharp vertical edges. This paper refers to this prediction as XP_1, and its associated differential signal as $XDIF_1$.

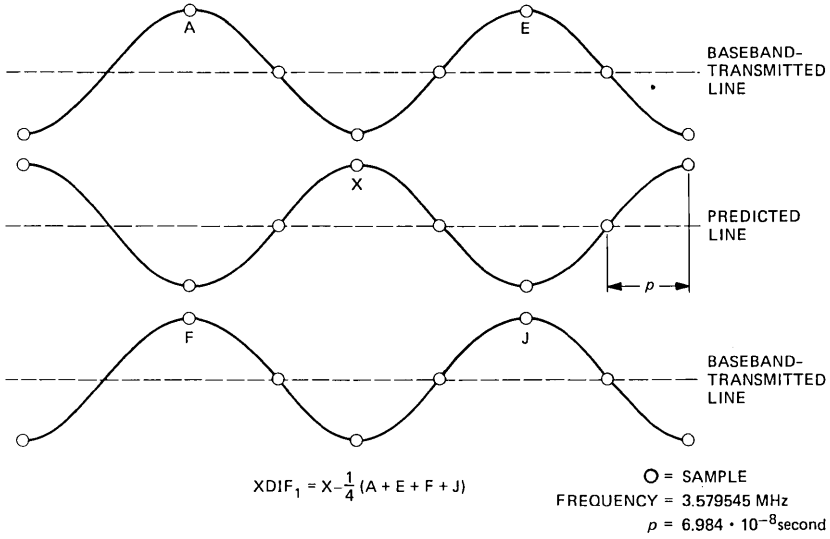


Fig. 1—Derivation of $XDIF_1$ from XP_1 prediction algorithm.

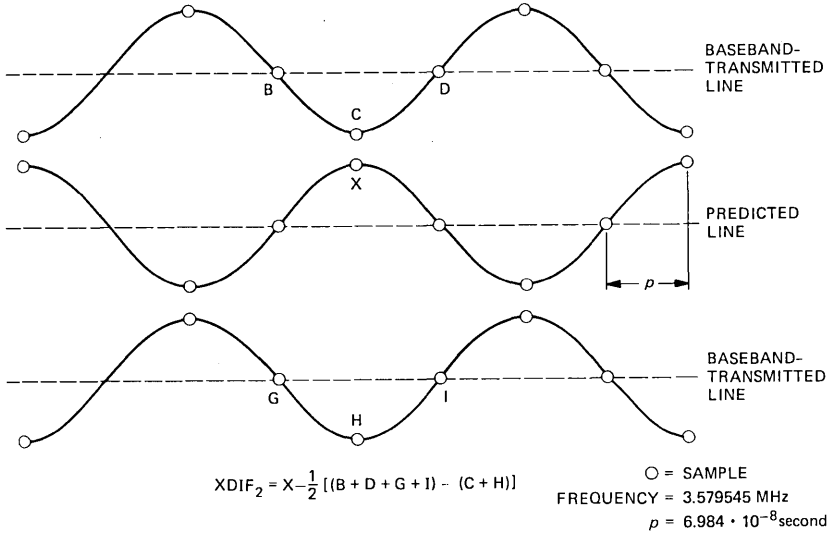


Fig. 2—Derivation of $XDIF_2$ from XP_2 prediction algorithm.

It was conjectured that picture quality would improve if we chose a line difference prediction that used the picture elements from the previous and upcoming lines closer to the coded element (Fig. 2), thus improving the resolution of the predicted value on vertical edges. Here we refer to this prediction as XP_2 and its associated differential signal as $XDIF_2$.

II. TIME DOMAIN ANALYSIS

The complexity of the baseband signal, and its quasi-periodic nature, would make analysis of the network response to an average video signal tedious and not very informative. However, since line-to-line correlation is high, some typical waveforms can be looked at and yield useful results.

Since there is a high content of color subcarrier in a typical television waveform, the differential signal should be small for that frequency. This criterion is met for both line difference predictions, and is shown in Appendix A to be zero for any flat field of constant color.

Appendix B shows that if there is color change from one line to the next the worst case differential signal will be the same for both predictions, with the amplitude being an average of the amplitudes of the two color sine waves for the adjacent lines. However, the worst case for each predictor does not occur at the same color subcarrier phase angle.

Along a horizontal line, color changes are almost always accompanied by luminance changes, and, since the bandwidth of the lumi-

nance is much higher than that of the chrominance, the effect of chrominance distortions will be secondary. With regard to luminance changes, three different waveforms are used to demonstrate the difference between the two prediction algorithms. The first waveform considered is the standard $2T$ pulse[2] shown in Fig. 3. This is a sine squared function, and is assumed repetitive from line-to-line. Calculations were done with a computer using the same general approach as shown in Appendices A and B. The results in Fig. 4 show that $XDIF_2$ (dotted line) has a reduced amplitude for both positive and negative directional peaks. In addition, the negative peaks are of shorter duration, resulting in less energy. Although this implies a slightly higher bandwidth, the magnitude of the band-limited signal for $XDIF_2$ is substantially less, indicating a more accurate reconstructed signal in the receiver, with less dependence on the differential signal.

The next waveform considered is the rising edge of the standard bar pulse[2] shown in Fig. 5. This function is represented a follows:

$$f(t) = \begin{cases} 0 & t < 0 \\ 6.4 \cdot 10^6 t & 0 \leqslant t \leqslant 156.25 \text{ ns} \\ 1 & t > 156.25 \text{ ns.} \end{cases}$$

Again, line-to-line repetition is assumed, and Fig. 6 shows the response of the two different algorithms. It can be seen that although the amplitudes of both differential signals are the same, the width is less for $XDIF_2$, resulting in less energy. Since the line-differential

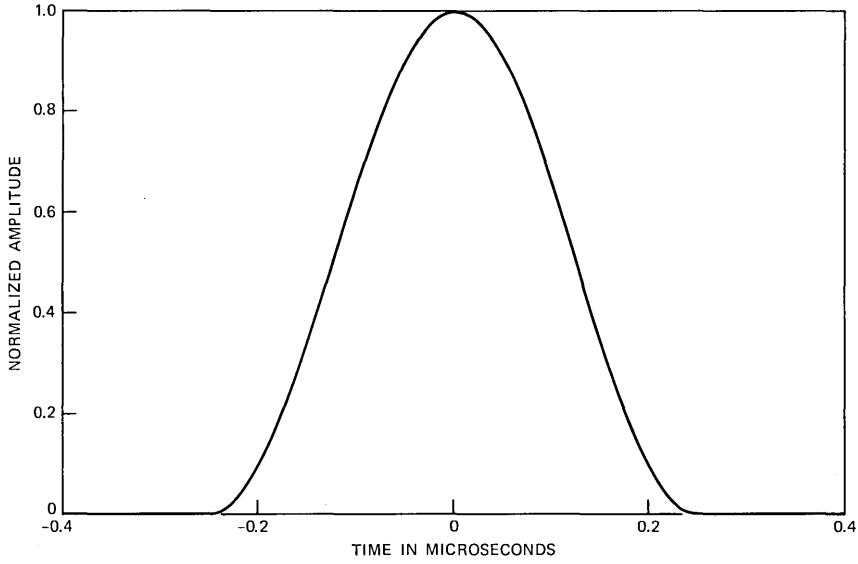


Fig. 3—Waveform of sine squared ($2T$) pulse from NTSC signal generator.

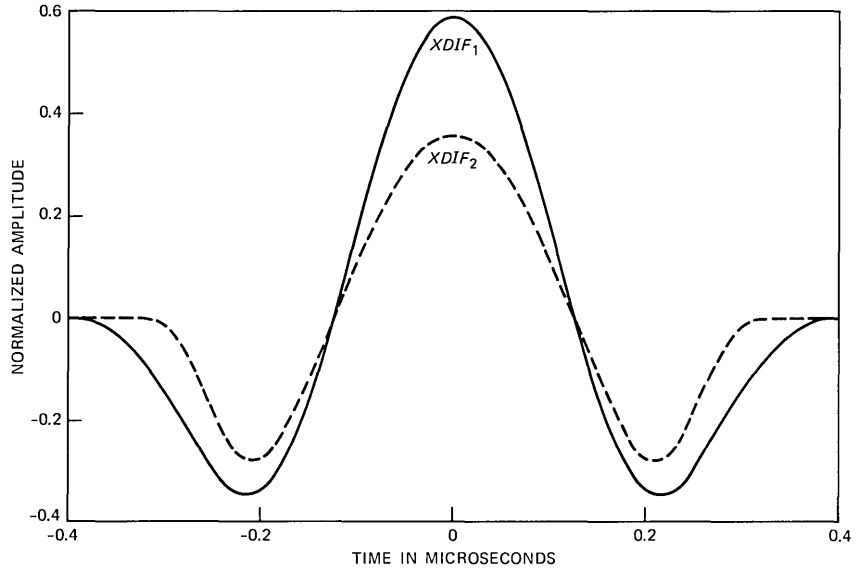


Fig. 4—Differential signal response of both algorithms to sine squared pulse.

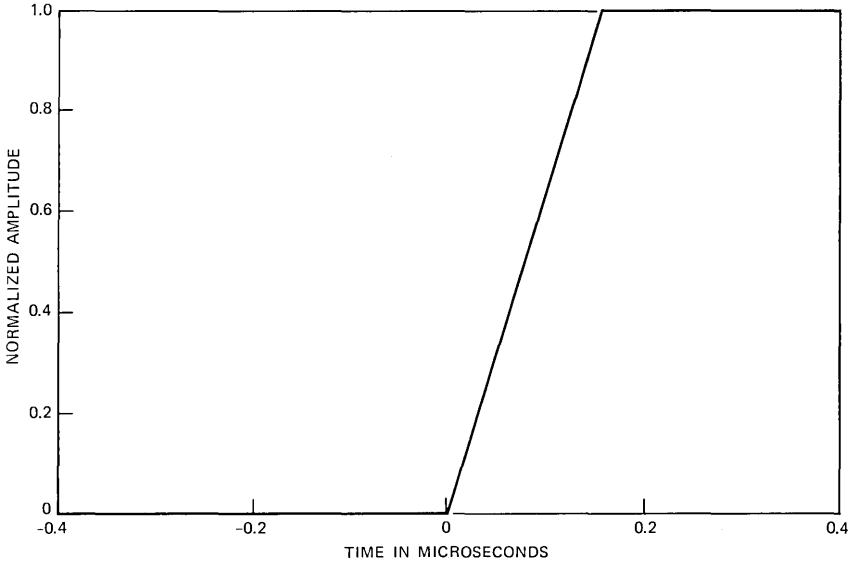


Fig. 5—Rising edge of bar pulse from NTSC signal generator.

signal in the proposed system is band limited to 3 MHz, this will again result in a smaller signal after filtering, with the same advantages as in the previous example.

The next and final waveform evaluated is a raised cosine function

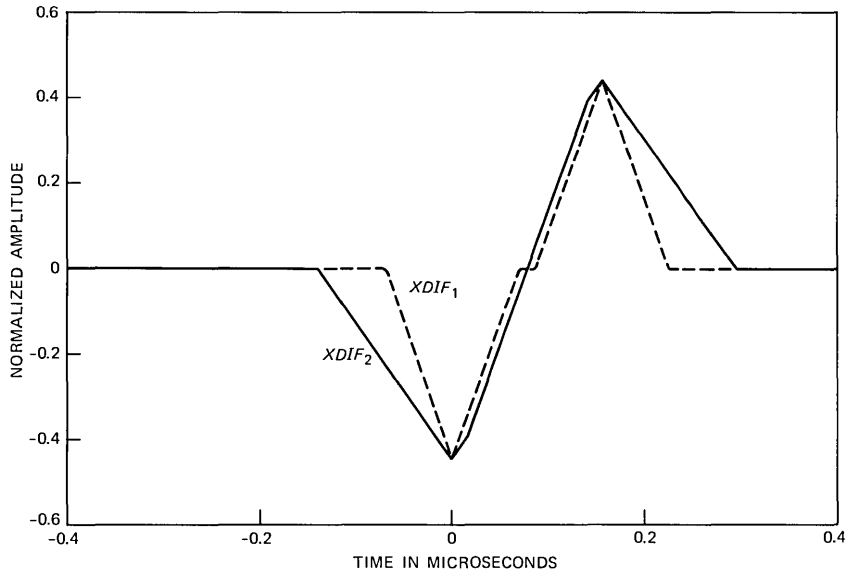


Fig. 6—Differential signal response of both algorithms to rising edge of bar pulse.

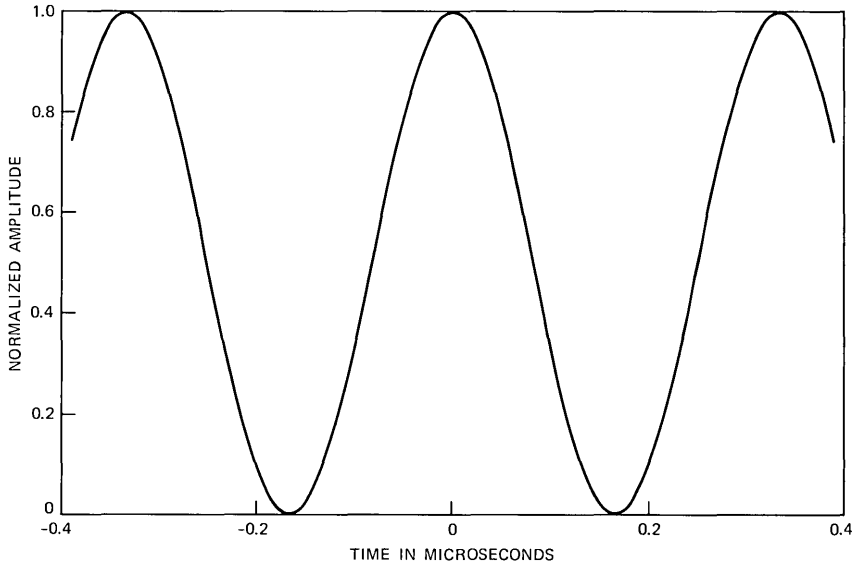


Fig. 7—Raised 3-MHz cosine demonstrating response of both algorithms to 3-MHz section of multiburst test signal.

at a frequency of 3 MHz, shown in Fig. 7. We chose this frequency because it is the intended maximum bandwidth of the differential signal. This would correspond to the fourth burst of the standard multiburst pattern[2] except that, for convenience, the amplitude is set to unity. Again, line-to-line repetition of the signal was assumed, and it can be seen in Fig. 8 that there is a 20-percent decrease in signal amplitude with $XDIF_2$. This again indicates a better reconstruction of the signal in the receiver with the available full bandwidth information.

III. FREQUENCY DOMAIN ANALYSIS

Before going into the analysis of the two predictors in the frequency domain, it will be useful to look at the frequency spectrum of a typical video signal.

It is well known[3] that the energy of a typical video luminance signal is concentrated on the harmonics of the horizontal or line sweep frequency F_H (≈15734 Hz). This means that most of the luminance energy of the video signal is distributed at multiples of F_H along the frequency axis. The chrominance information also peaks at line frequency intervals, and, by placing the color subcarrier F_c at an odd multiple of half the line frequency, i.e., $F_c = 455/2 \cdot F_H$, the chrominance information will be interleaved with luminance information peaks.

Chrominance information bandwidth is approximately 1.5 MHz, but, with the composite signal band-limited to 4.2 MHz, it results in

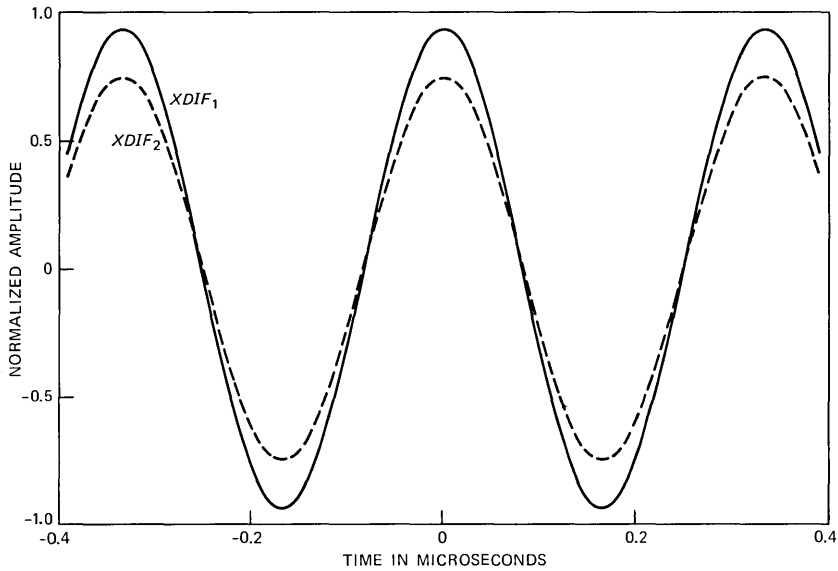


Fig. 8—Differential signal response of both algorithms to 3-MHz raised cosine.

a chrominance bandwidth of 1.5 MHz below the color subcarrier and 0.5 MHz above. Figure 9 shows the spectrum of a chrominance-modulated linear ramp test pattern. Figure 9a is the 0- to 4-MHz spectrum that shows the luminance information at low frequencies and the chrominance centered about $F_c = 3.579 \cdots$ MHz. Figure 9b shows the fundamental and the first nine harmonics of the line frequency, and Fig. 9c shows the color subcarrier and the line multiples on either side of it. It can be seen in this third photograph, also, that the luminance information (which peaks halfway between the chrominance peaks) is much lower than the chrominance at this frequency for this waveform.

The ideal frequency response of the linear system that produces the differential signal would then require that the differential signal be zero at even intervals of half the line frequency for the lower frequencies, and zero at odd intervals of half the line frequency for frequencies around the color subcarrier. Where the transition from odd to even intervals of half the line frequency should occur is not exactly defined, but it should not be lower than the lowest chrominance component, i.e., color subcarrier frequency minus the chrominance bandwidth, or approximately 2.08 MHz. In fact, the ideal value is picture dependent and, more exactly, would depend on the relationship between the luminance high-frequency content and the chrominance information content.

IV. FREQUENCY DOMAIN COMPARISON

It is well known from circuit theory that, in a linear system with input $x(t)$, output $y(t)$, and impulse response $h(t)$,

$$Y(j\omega) = H(j\omega) \cdot X(j\omega),$$

where uppercase letters designate Fourier Transforms. Since the time relationship of the differential predictors is known, the frequency domain representation of the network $H(j\omega)$ can be found easily, and this process is shown in Appendix C for both the original and the new predictors.

With the results of these simple calculations, a program was written to plot the network response $|H(j\omega)|$. Figures 10 and 11 show the 0- to 100-kHz response of the original and new differential signal networks, respectively. It can be seen that both signals are zero for the line frequency and its harmonics. Since this is where most of the incoming signal energy is concentrated, both of these algorithms should behave well in this regard.

Figures 12 and 13 show the response of the two algorithms in a 100-kHz interval centered around the color subcarrier. It can be seen that both algorithms are zero at the color subcarrier and are again periodic

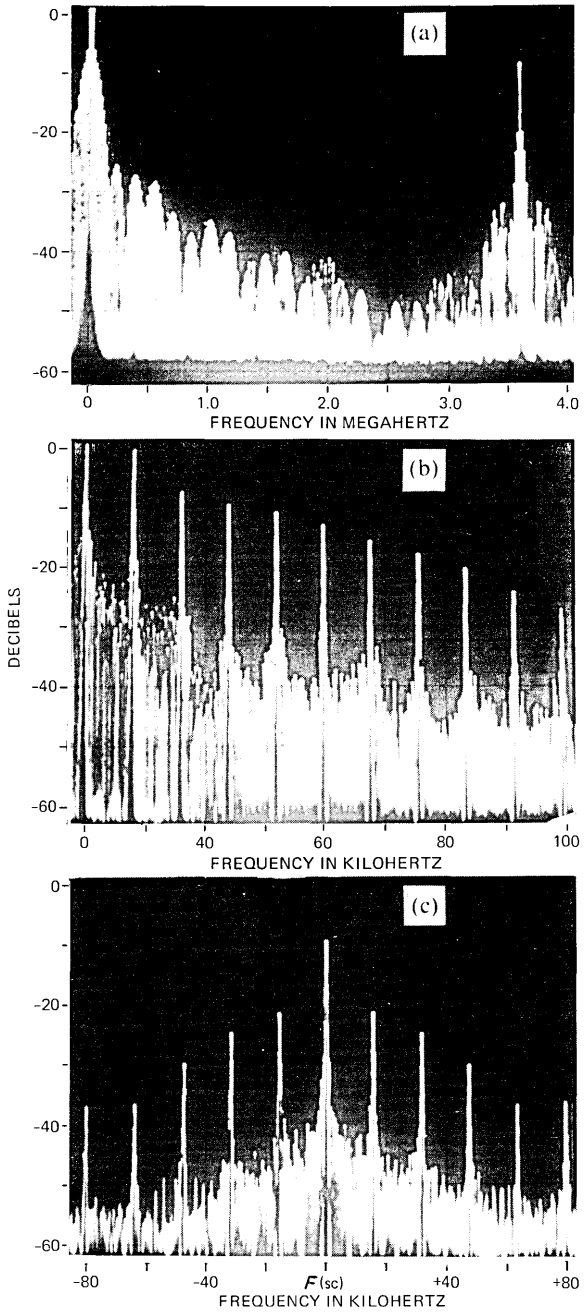


Fig. 9—Spectrum of chrominance-modulated linear ramp test pattern: (a) from 0 to 4.2 MHz; (b) from 0 to 100 kHz, with peaks at even multiples of half the line frequency; and (c) over a 100-kHz band centered around color subcarrier, with peaks at odd multiples of half the line frequency.

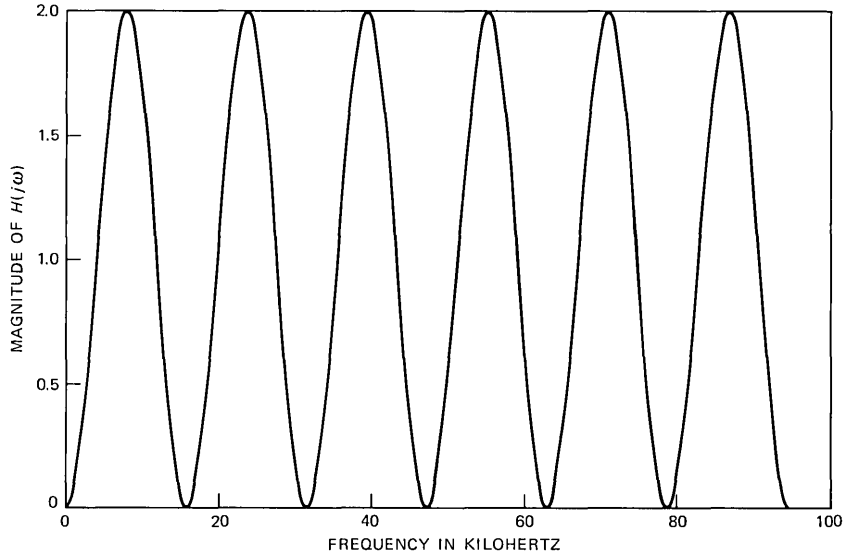


Fig. 10—Calculated differential signal spectrum of XP_1 algorithm from 0 to 100 kHz.

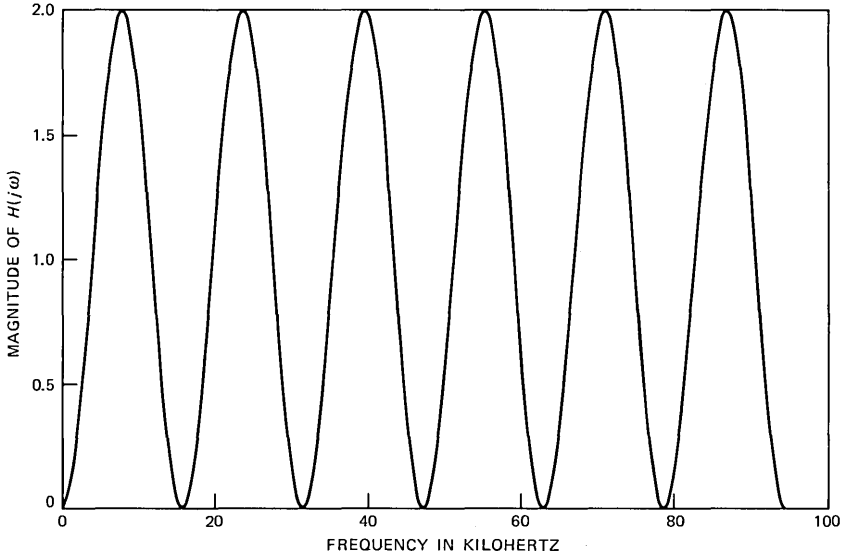


Fig. 11—Calculated differential signal spectrum of XP_2 algorithm from 0 to 100 kHz.

at a line rate. Since this corresponds to the high-energy chrominance portions of the incoming signal for this part of the spectrum, both algorithms should again behave well.

The difference in the response of the two algorithms can be seen,

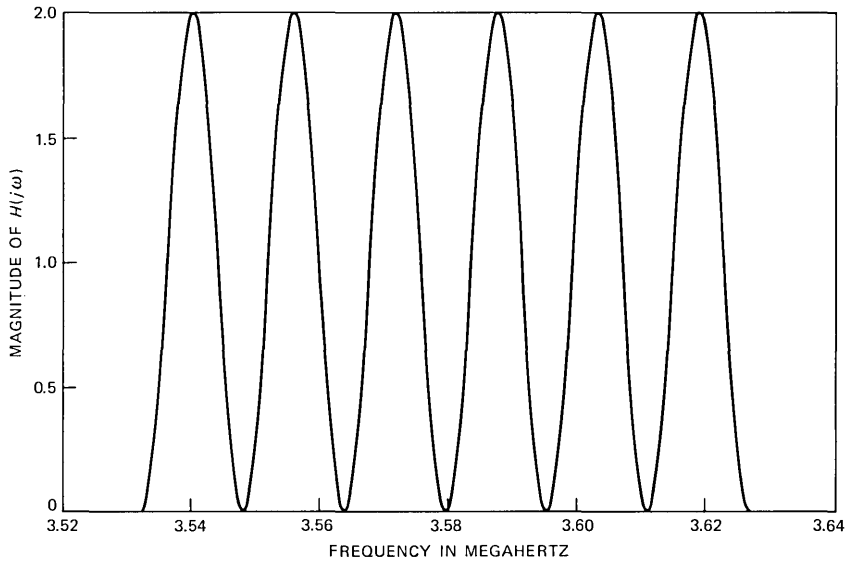


Fig. 12—Calculated differential signal spectrum of XP_1 algorithm for 100-kHz window centered around color subcarrier.

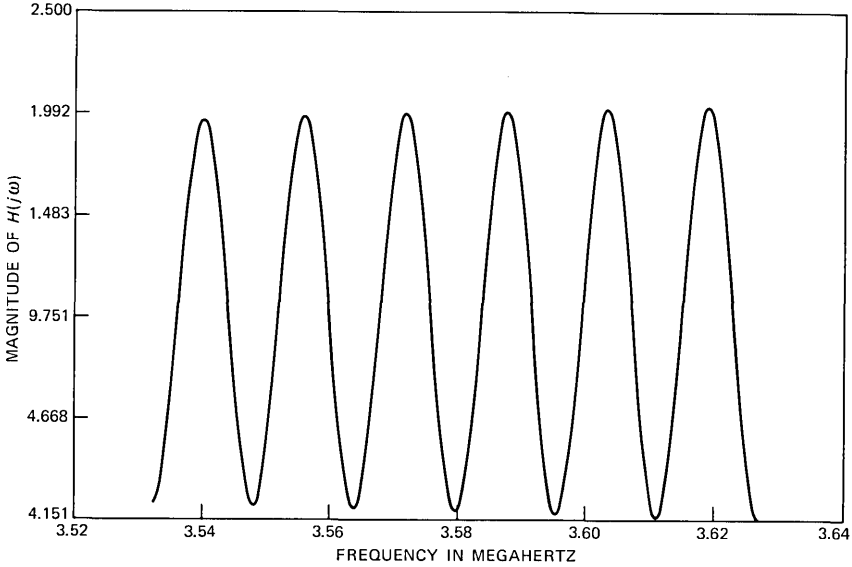


Fig. 13—Calculated differential signal spectrum of XP_2 algorithm for 100-kHz window centered around color subcarrier.

however, in Figs. 14 and 15. In these figures the differential signal response is plotted from 0 to 3 MHz, and the oscillatory response pattern is shown with the sine wave minimum values falling on even multiples of half the line frequency in the lower part of the spectrum, and at odd multiples of half the line frequency in the upper part.

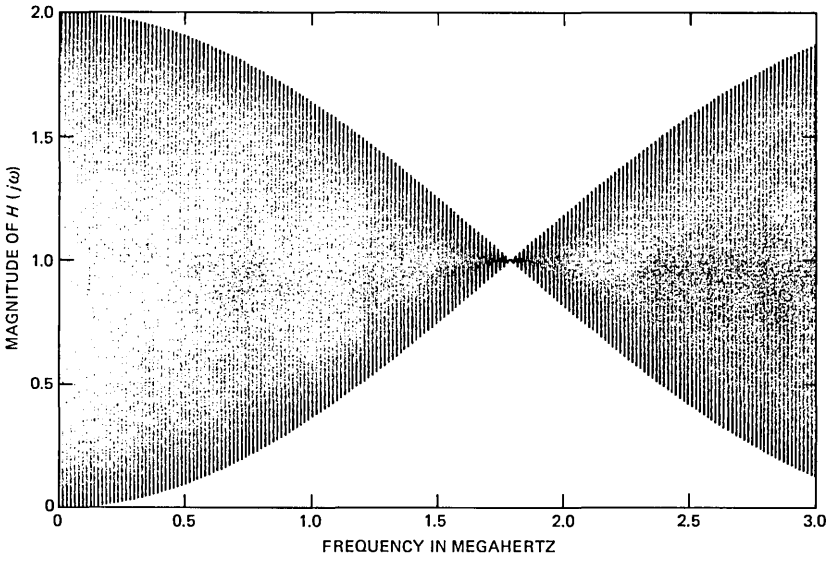


Fig. 14—Calculated differential signal spectrum of XP_1 algorithm from 0 to 3 MHz.

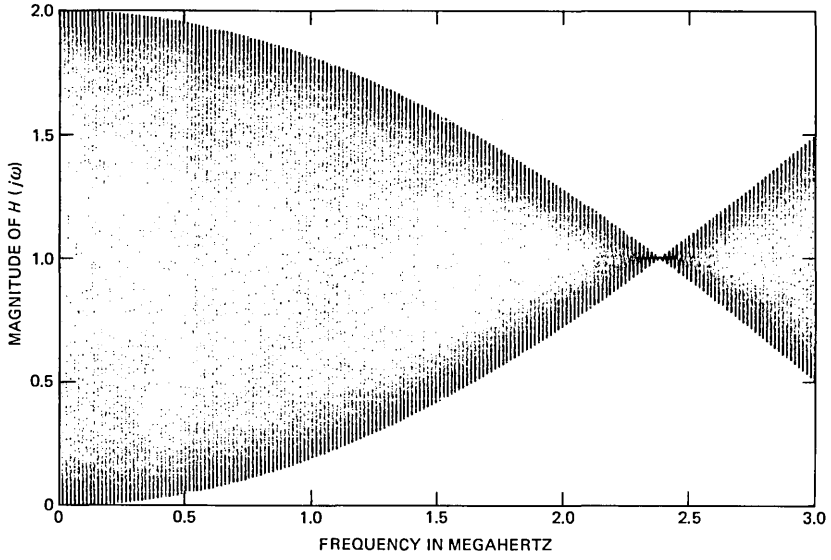


Fig. 15—Calculated differential signal spectrum of XP_2 algorithm from 0 to 3 MHz.

Because the lower limit of the chrominance information bandwidth is 2.08 MHz, we can see that the $XDIF_1$ transition is too soon (at about 1.8 MHz). This results in a greater than unity gain for the luminance information in this region and, therefore, a poorer prediction with no resulting savings in chrominance information because it is out of band. The $XDIF_2$ signal, however, is below unity for luminance response until 2.38 MHz, giving a better luminance prediction over an additional 500 kHz of bandwidth. This gives a poorer line difference prediction for the upper 300 kHz of the chrominance band, but since a fast chrominance change is almost always accompanied by a luminance change, the loss should be offset by an improvement in luminance response. Figure 16 plots the envelope for the luminance response and shows that the $XDIF_2$ signal will be lower throughout this whole region, indicating a better line difference prediction. These plots cross over at the color subcarrier, but that is of no concern since the differential signal will be band limited to 3 MHz in the proposed system.

V. HARDWARE VERIFICATION OF RESULTS

To verify these results, a Tektronix test signal generator was modified so that a sweep generator signal could be added to the flat field test pattern. In this way the system could be analyzed with a full field sine wave input. The differential signal was then put into a spectrum

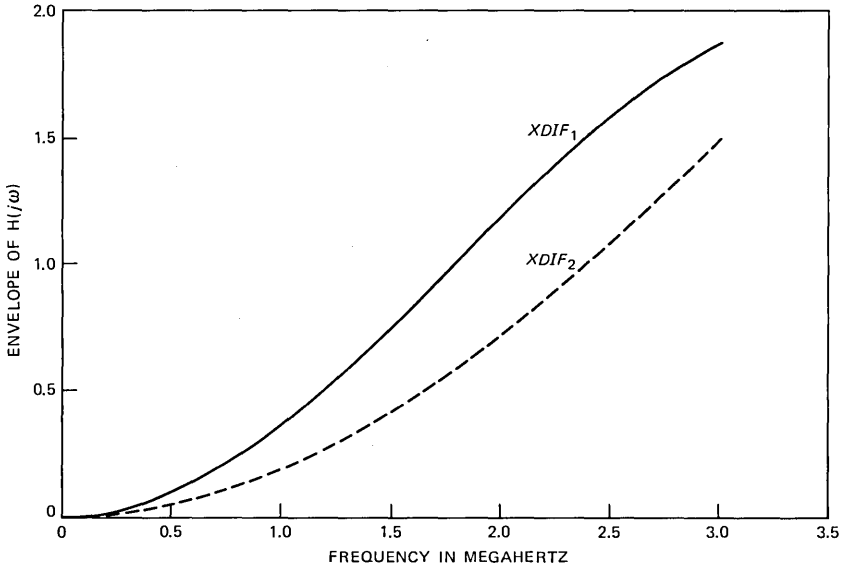


Fig. 16—Envelope of differential signal frequency response for XP_1 and XP_2, with samples taken at even multiples of half the line frequency.

analyzer. Figure 17 shows photographs of the 0- to 100-kHz spectrum and the 0- to 3-MHz spectrum. The response of the network is identical to the computed values shown in Figs. 10 through 16.

Time domain measurements were also made to verify the results of the analysis of this paper. The measurements were made with the full field pulse bar test pattern from the NTSC test signal generator. The differential signal output was band limited to 3 MHz, which resulted in some deviation from the calculated ideal results. Figures 18a and 18b show the response of the two algorithms to the sine squared pulse. A slight spreading of the signal caused by filtering can be seen, but otherwise the results are quite close to the ideal calculations. Figures 18c and 18d then show the response to the 156.25-ns ramp (bar pulse). Here again the high resolution of the ideal waveforms is lost because of filtering, but the predicted decrease in amplitude is clear.

Subjective tests were then carried out to determine if any degradation in picture quality was visible with either line difference prediction

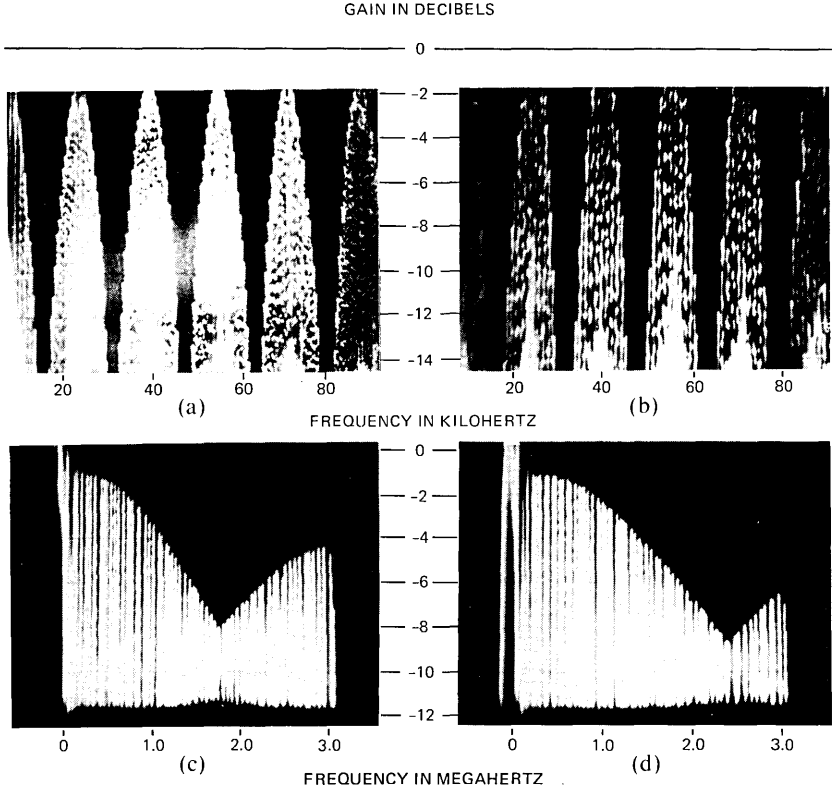


Fig. 17—Measured differential signal frequency responses of algorithms: (a) XP_1 from 0 to 100 kHz, (b) XP_2 from 0 to 100 kHz, (c) XP_1 from 0 to 3 MHz, and (d) XP_2 from 0 to 3 MHz.

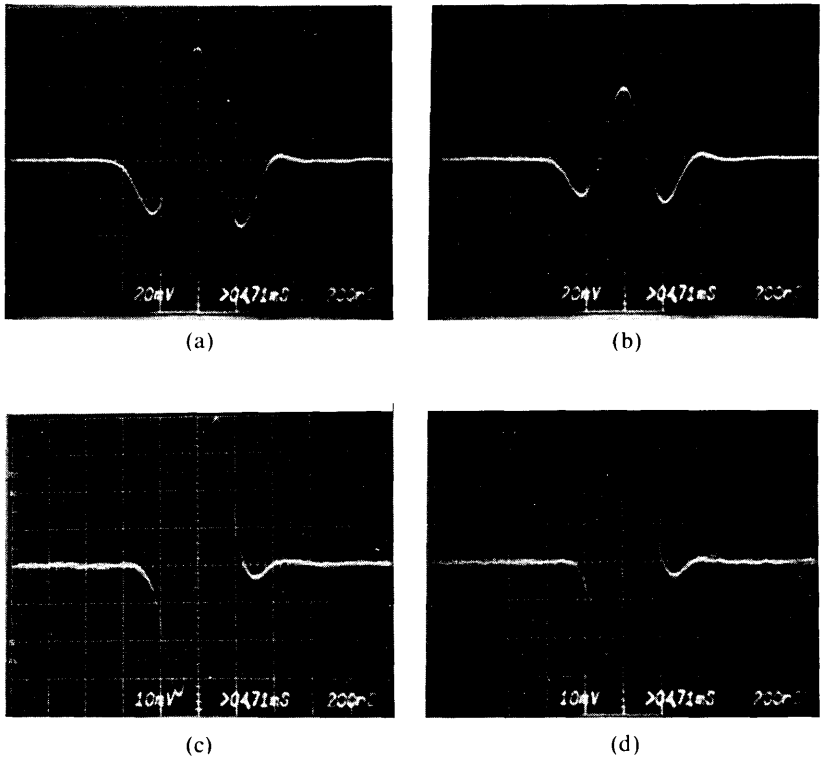
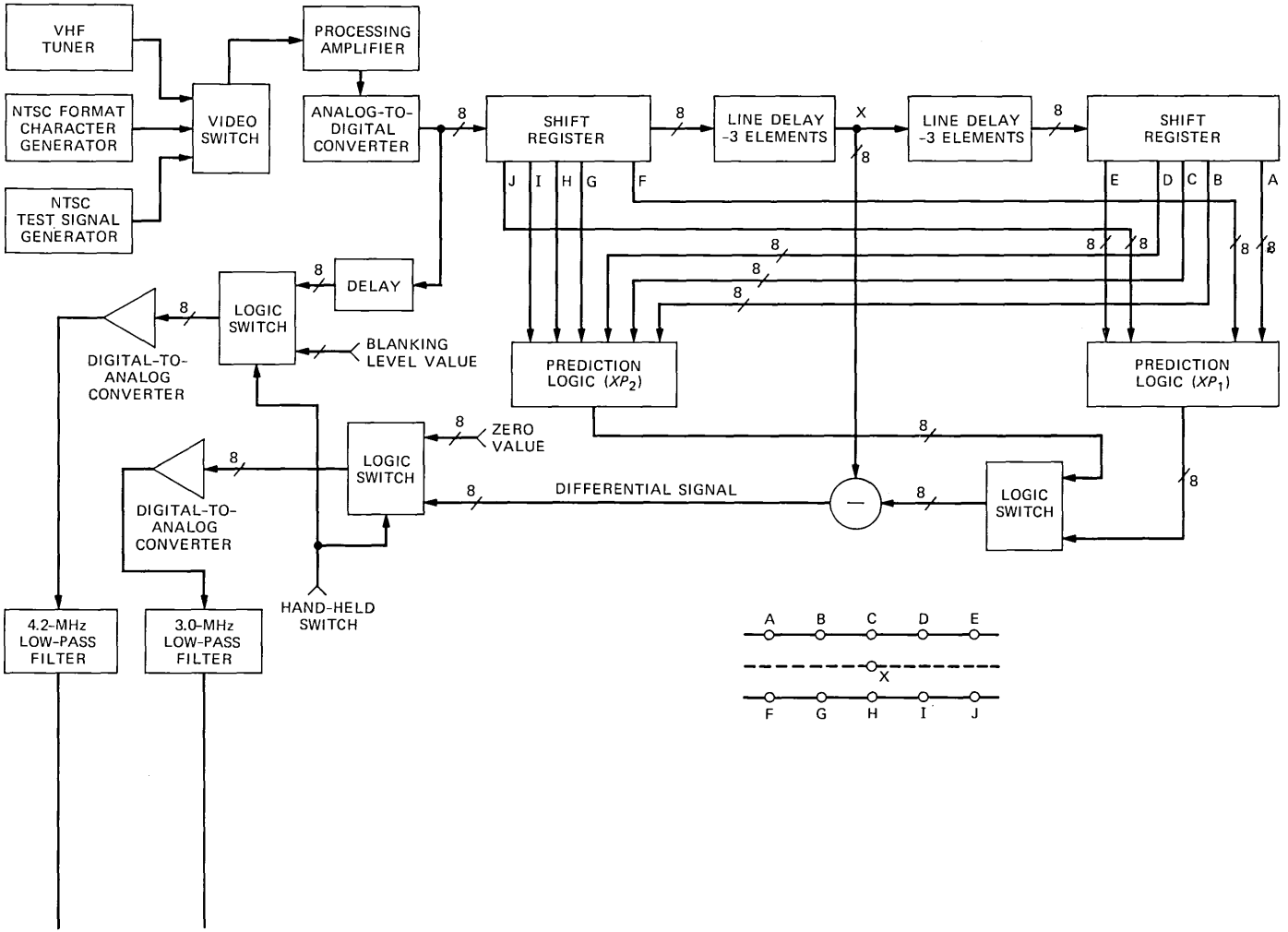
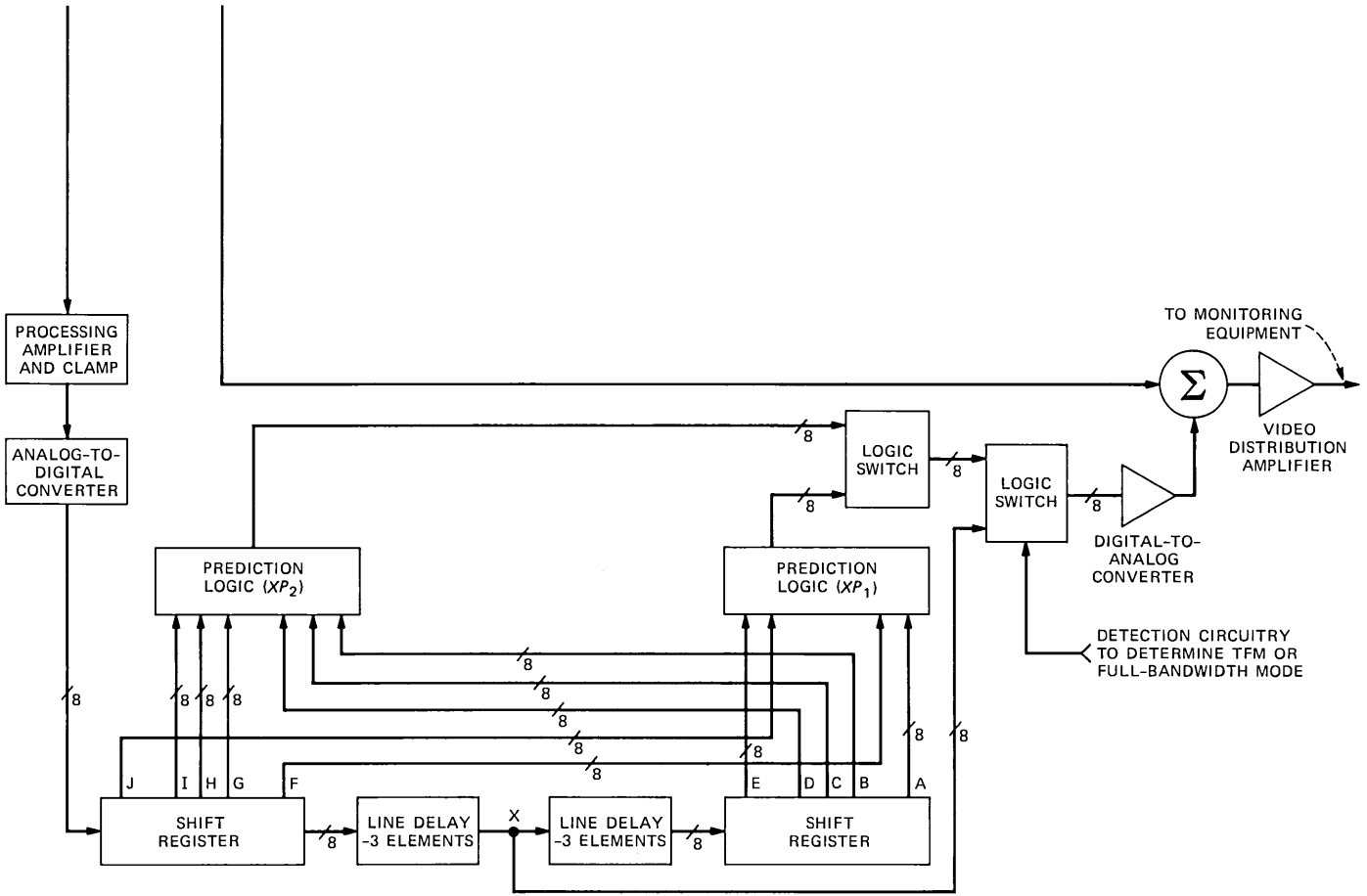


Fig. 18—Measured differential signal time domain responses of both algorithms to the sine for (a) XP_1 and (b) XP_2. Same responses to rising edge of the bar pulse for (c) XP_1 and (d) XP_2.

algorithm. To minimize distortion contributions from other hardware sources, so that contributions from the algorithm would be the primary cause of degradation, the TFM hardware was simplified. The band-limited differential signal and the time-multiplexed baseband signals from the transmitter were sent on individual coaxial cables to the receiver. This avoided signal degradations from the frequency multiplexing operation, and also transmission system degradations, leaving only the analog baseband amplifiers and two analog-to-digital-to-analog conversions as additional sources of degradation. Both the transmitter and the receiver use eight-bit pulse code modulation (PCM) for their digital processing, so that these additional contributions should be small in comparison with the differential signal band limiting. A block diagram of the test configuration is shown in Fig. 19.

Several video signals were evaluated using a stringent comparison between the reconstructed video signals and one which is unimpaired except for the two analog-to-digital conversions. The signals were switched on the same television monitor during the vertical blanking


VHF TUNER
NTSC FORMAT CHARACTER GENERATOR
NTSC TEST SIGNAL GENERATOR
VIDEO SWITCH
PROCESSING AMPLIFIER
ANALOG-TO-DIGITAL CONVERTER
SHIFT REGISTER
LINE DELAY -3 ELEMENTS
LINE DELAY -3 ELEMENTS
SHIFT REGISTER
J I H G F
E D C B A
DELAY
LOGIC SWITCH
BLANKING LEVEL VALUE
DIGITAL-TO-ANALOG CONVERTER
PREDICTION LOGIC (XP_2)
PREDICTION LOGIC (XP_1)
ZERO VALUE
LOGIC SWITCH
DIFFERENTIAL SIGNAL
LOGIC SWITCH
DIGITAL-TO-ANALOG CONVERTER
HAND-HELD SWITCH
4.2-MHz LOW-PASS FILTER
3.0-MHz LOW-PASS FILTER
A B C D E
X
F G H I J



NTSC – NATIONAL TELEVISION SYSTEM COMMITTEE
TFM – TIME FREQUENCY MULTIPLEXING

Fig. 19—Block diagram of test configuration used to elevate prediction algorithms.

intervals so that switching transitions were only visible when coding caused degradation. A toggle switch in both the transmitter and receiver allowed selection of either algorithm. The television monitor used for this evaluation did not contain a comb filter, so there is a possibility that some degradation caused by coding was masked by the luminance component band limiting. However, since both algorithms were evaluated with the same constraints, this did not appear to be a problem.

Results showed that with normal off-the-air video signals both algorithms performed well, and neither one produced any visible degradation. However, with electronically generated test patterns and text, the degradation of high-frequency signals, such as the $2T$ pulse, the multiburst pattern, and character edges, was less severe with the new algorithm.

VI. CONCLUSIONS

Subjective tests verified the results of this analysis. Thus, the quality of a picture using the band-limited differential signal is improved with the new algorithm. The implementation is somewhat more difficult than the original one and requires only the addition of two more arithmetic stages. It is, therefore, a practical and desirable change to make in the system.

REFERENCES

1. B. G. Haskell, "Time-Frequency Multiplexing (TFM) of Two NTSC Color TV Signals—Simulation Results," B.S.T.J., *60*, No. 5 (May–June 1981), pp. 643–660.
2. *NTC Report No. 7, Video Facility Testing Technical Performance Objectives,* June 1975, revised January 1976. Published by Public Broadcasting Service.
3. D. Fink, *Television Engineering Handbook,* New York: McGraw-Hill, 1957.

APPENDIX A

Comparison of Differential Signal Responses to a Constant Color, Flat Field Luminance Video Signal

Consider a uniform flat field of some color for the XP_1 algorithm shown in Fig. 1:

$$XDIF_1 = X - 1/4(A + E + F + J).$$

Let $X = U \cos(2\pi F_c t)$, where F_c = the color subcarrier frequency. Then, since it is a uniform flat field,

$$A = E = F = J = U \cos(2\pi F_c t) = X.$$

Therefore,

$$XDIF_1 = 0 \qquad \text{for all } t.$$

For the XP_2 prediction shown in Fig. 2 with the same X as above,

$$XDIF_2 = X - [1/2(B + D + G + I) - 1/2(C + H)].$$

Letting $\omega_c = 2\pi F_c$ we have

$$B = G = U \cos\left(\omega_c t + \frac{\pi}{2}\right),$$

$$D = I = U \cos\left(\omega_c t - \frac{\pi}{2}\right),$$

and

$$C = H = U \cos(\omega_c t + \pi);$$

therefore

$$\begin{aligned} XDIF_2 &= U \cos(\omega_c t) - \frac{U}{2}\left[2 \cos\left(\omega_c t + \frac{\pi}{2}\right)\right. \\ &\qquad \left. + 2 \cos\left(\omega_c t - \frac{\pi}{2}\right) + 2 \cos(\omega_c t)\right] \\ &= U \cos(\omega_c t) - U[-\sin(\omega_c t) + \sin(\omega_c t) + \cos(\omega_c t)] \\ &= U \cos(\omega_c t) - U \cos(\omega_c t) \\ &= 0 \qquad \text{for all } t. \end{aligned}$$

APPENDIX B

Comparison of Differential Signal Responses to an Adjacent Line Color Change, Flat Field Luminance Video Signal

Now consider a color change with no corresponding luminance change from one line to the next. The change will be considered between lines two and three of the prediction area, although calculations would be identical for a change from line one to two. Let U = magnitude of the first color, V = magnitude of the second color, and $\omega_c = 2\pi F_c$. Then for the XP_1 predictor,

$$\begin{aligned} XDIF_1 &= U \cos(\omega_c t) - \left[\frac{U}{2} \cos(\omega_c t) + \frac{V}{2} \cos(\omega_c t + \phi)\right] \\ &= \frac{U}{2} \cos(\omega_c t) - \frac{V}{2} \cos(\omega_c t + \phi), \end{aligned}$$

where the worst case would be for $\phi = \pi$, giving

$$XDIF_1 = \frac{U + V}{2} \cos(\omega_c t).$$

For the XP_2 predictor,

$$XDIF_2 = U\cos(\omega_c t) - \frac{1}{2}\left[U\cos\left(\omega_c t + \frac{\pi}{2}\right) + U\cos\left(\omega_c t + \frac{\pi}{2}\right) + V\cos(\omega_c t + \phi) + V\cos(\omega_c t + \pi + \phi) - U\cos(\omega_c t + \pi) - V\cos\left(\omega_c t + \frac{\pi}{2} + \phi\right)\right]$$

$$= U\cos(\omega_c t) - \frac{1}{2}[V\cos(\omega_c t + \phi) - V\cos(\omega_c t + \phi) + U\cos(\omega_c t) + V\sin(\omega_c t + \phi)]$$

$$= U\cos(\omega_c t) - \frac{U}{2}\cos(\omega_c t) - \frac{V}{2}\sin(\omega_c t + \phi)$$

$$= \frac{U}{2}\cos(\omega_c t) - \frac{V}{s}\sin(\omega_c t + \phi),$$

where the worst case will occur when $\sin(\omega_c t + \phi)$ equal $-\cos(\omega_c t)$ or at ϕ equal to $-\frac{\pi}{2}$, giving

$$XDIF_2 = \frac{U + V}{2}\cos(\omega_c t),$$

which is the same as for $XDIF_1$.

APPENDIX C

Transfer Function Derivation for Both Prediction Algorithms

Knowing the time relationship of the picture elements, we can find the Fourier Transform of the impulse responses of the differential signal algorithms. Let $p = 1$ picture element with a duration of $6.98412 \cdot 10^{-8}$ second. Then $910p = 1$ line $= 6.3555 \cdot 10^{-5}$ second.

For the XP_1 predictor:

$$XDIF_1(t) = X(t) - \frac{1}{4}[X(t + 908p) + X(t + 912p) + X(t - 908p) + X(t - 912p)],$$

$$XDIF_1(j\omega) = X(j\omega) - \frac{1}{4}[X(j\omega)e^{-j\omega 908p} + X(j\omega)e^{-j\omega 912p} + X(j\omega)e^{+j\omega 908p} + X(j\omega)e^{j\omega 912p}]$$

$$= X(j\omega)\left\{1 - \frac{1}{4}\left[e^{908pj\omega} + e^{-908pj\omega} + e^{912pj\omega} + e^{-912pj\omega}\right]\right\}$$

$$= X(j\omega)\left\{1 - \frac{1}{2}\left[\cos(908p\omega) + \cos(912p\omega)\right]\right\}$$

$$= X(j\omega)\cdot H(j\omega),$$

with

$$H(j\omega) = 1 - \tfrac{1}{2}[\cos(908p\omega) + \cos(912p\omega)].$$

Now for the XP_2 predictor:

$$XDIF_2(t) = X(t) - \{1/2[X(t-911p) + X(t-909p) + X(t+909p) + X(t+911p)] - 1/2[X(t-910p) + X(t+910p)]\},$$

$$\begin{aligned} XDIF_2(j\omega) &= X(j\omega) - \{1/2[X(j\omega)e^{+911pj\omega} + X(j\omega)e^{+909pj\omega} \\ &\quad + X(j\omega)e^{-909pj\omega} + X(j\omega)e^{-911pj\omega}] \\ &\quad - [X(j\omega)e^{910pj\omega} + X(j\omega)e^{-910pj\omega}]\} \\ &= X(j\omega)\{1 - [(e^{909pj\omega} + e^{-909pj\omega}/2) + (e^{911pj\omega} + e^{-911pj\omega}/2) \\ &\quad - (e^{910pj\omega} + e^{-910pj\omega}/2)]\} \\ &= X(j\omega)\{1 - [\cos(909pj\omega) + \cos(911pj\omega) - \cos(910pj\omega)]\}. \end{aligned}$$

Therefore,

$$H(j\omega) = 1 + \cos(910pj\omega) - \cos(909pj\omega) - \cos(911pj\omega).$$

AUTHOR

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

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

On the Recognition of Isolated Digits From a Large Telephone Customer Population

By J. G. WILPON* and L. R. RABINER*

(Manuscript received February 25, 1983)

A field study was initiated to learn about the effects of various telephone transmission and switching conditions on the algorithms currently used in the Bell Laboratories, Linear Predictive Coding (LPC)-based, isolated word recognizer. Digit recordings were obtained from customers over a variety of transmission facilities. During a 23-day recording period a total of 11,035 isolated digits were recorded. For each recording, statistics were recorded about the line condition, the background environment, and the customer's ability to speak his/her telephone number as a sequence of isolated digits. Also recorded was information about the ability of the automatic word endpoint detector to find each spoken digit and to accurately determine the correct endpoints. The results of several recognition tests are presented—one using a previously defined set of laboratory-created digit reference templates, and several others using new sets of reference templates from a subset of the recorded digits. The performance of the recognizer is poor (average digit accuracy of 77.4 percent) using the laboratory template set, but improves substantially (average digit accuracy of 93.1 percent) for a template set created from the field recordings. The reasons for this improvement in digit recognition accuracy are presented, along with their implications to future work in isolated word recognition.

I. INTRODUCTION

Research on the problems involved in speech recognition has been carried out at Bell Laboratories for close to a decade.[1-7] In all these

*Bell Laboratories.

studies the speech database consisted of utterances recorded under laboratory conditions (i.e., cooperative subjects, soundproof booths, and subject prompting) using dialed-up lines over a local Private Branch Exchange (PBX). Peak signal-to-noise ratios ranged from 40 to 60 dB under these conditions. The recognition systems previously studied involved either a user training phase (speaker-dependent systems) or no training phase (speaker-independent systems). The vocabulary sizes ranged from as few as 10 words,[2] to as many as 1109 words.[7] Our past studies of speaker-independent systems involved a relatively small number of subjects, typically 100 for training and 10 to 40 for test calculations. Our current recognition systems performed very well given these conditions.[8]

To test the viability of speaker-independent, isolated word recognition systems for large user populations, it was necessary to conduct an experiment under "real world" conditions. Such an experiment involves using noncooperative telephone customers speaking in an uncontrolled environment over a set of randomly dialed telephone lines. This paper presents such an experiment and its implications of future speech recognition work. During the course of the experiment, a speech database was collected over a 60-day period in a Bell System environment (i.e., recordings obtained directly at a Bell System switching office) from over 3100 subjects. The vocabulary chosen for this study was the 10 digits, zero through nine (the digit zero was generally pronounced "oh"). Since we wanted a system that could handle a large number of users with the least amount of burden to the user we chose to make the system speaker independent.

There are several very important recognition issues that need to be resolved, and only by using a very large speech database, such as the one we obtained, can these issues be addressed. The most important issues involve training the recognizer. Since we have required that the recognition system be speaker independent, several questions arise as to how one obtains the necessary training data. Should training tokens be used from only the "best" speakers over the cleanest telephone lines, or should all the training samples be randomized, i.e., from any talker over any quality transmission line? Another issue involves the number of training tokens needed to adequately represent an extremely large number of potential users. In past studies we have used at most 100 representations for each vocabulary item. Do we necessarily need more tokens? In this paper we investigate these issues among others.

Our results indicate a distinct number of "real world" problems that must be considered when implementing a speech recognition system with a widespread applicability. These include properly treating highly variable background conditions, devising procedures for handling au-

tomatic endpoint detector failures, and problems associated with obtaining isolated speech input. Methods of handling these problems will also be discussed in this paper.

In Section II we describe how we obtained the speech recordings. Section III presents an evaluation of our current speech recognition system, on the basis of a series of recognition experiments. A discussion of the overall results and their implications is given in Section IV.

II. RECORDING PROCEDURE

Figure 1 depicts the overall recording setup used in this study. Recordings were made at a Bell System switching office in Portland, Maine, and then transmitted back to the Murray Hill Laboratory for analysis. The sequence of operations to record a single user's speech was as follows:

1. A site observer (SO) issued a prerecorded spoken message (a prompt) requesting that the user speak his or her telephone number as a sequence of isolated digits. After the first three digits were spoken,* the observer then initiated recording (i.e., digitizing the spoken data) of the spoken number sequence. As each of the digits was spoken, the observer entered it on a keyboard. Four digits were nominally recorded (digitally at a 6.67-kHz rate) for each subject.

2. The observer determined if the digit sequence was not spoken in an isolated format (i.e., spoken without sufficient pauses). If so, the observer initiated another prerecorded spoken message (a reprompt) requesting the user to repeat the number with a longer pause between digits. After the subject completed speaking the number, he or she was given a prerecorded "Thank you" message.

During this first phase an observer at the Murray Hill Laboratory (MHO) had also been monitoring the call and now took over handling the call, performing the following operations:

1. The MHO also determined (based on listening) whether the digit sequence was spoken in an isolated manner. If the MHO decided that the speech was unacceptable (either because it was spoken in a connected manner or because of unacceptably bad telephone line conditions), a reject code was entered and the entire procedure was terminated for the current call.

2. If the call was acceptable, the MHO entered the sequence of spoken digits heard. This sequence was compared with that entered by the SO and any discrepancies (errors) were noted and fixed by listening to the recorded digits.

---

* For reasons of customer privacy we were not allowed to record all seven digits of the telephone number.

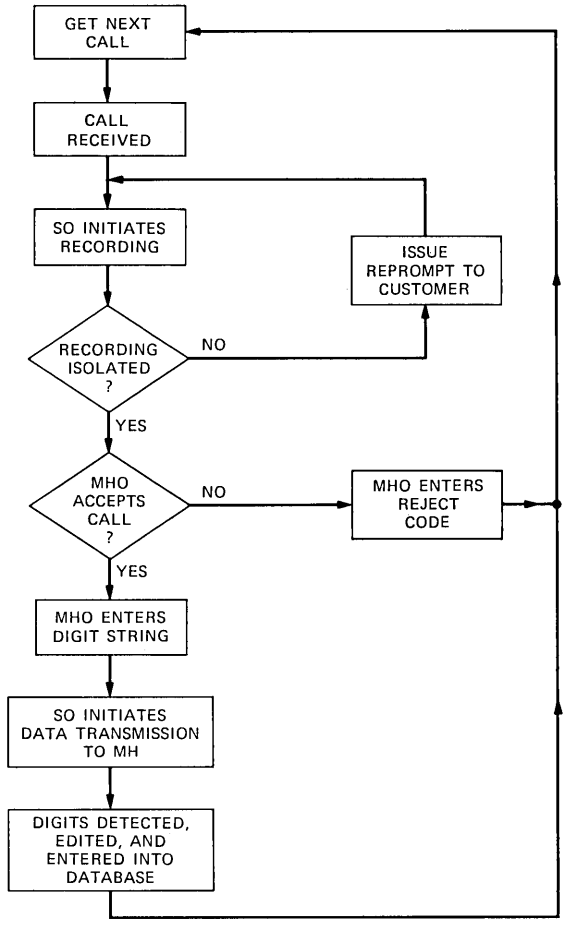


Fig. 1—The overall digit recording system.

3. At this point the MHO initiated digital transmission of the digitized speech from Portland to the MH laboratory.

Once the transfer was completed, an eighth-order Linear Predictive Coding (LPC) analysis, and automatic endpoint detection were performed.[9] The log energy of the waveform was displayed to the MHO, along with the automatically determined sets of endpoints indicating where in the recording interval the isolated words could be found. At this point the MHO had the option of modifying any or all sets of endpoints computed. The segmented speech was then entered into the database for later examination.

Using this procedure we recorded approximately 11,000 digits from

3100 subjects over a 23-day period. During the first 11 days no reprompting was used and recordings were taken for about 8 hours each day. Beginning wih the 12th day the reprompt procedure was instituted and we began recording for 12 hours each day. Figures 2 through 5 show the breakdown of the data recorded. Figure 2 shows a plot of number of digits recorded on a daily (session) basis, for males, females, and the total. For the first several days, only about 200 to 300 digits were recorded because both observers were learning. The dip in recordings around day 10 was due to a major snowstorm which curtailed recording. Figure 3 shows a histogram of the number of utterances recorded per digit for males, females, and total. The digits 2 and 3 had the highest number of occurrences, and the digits 9 and 0 had the fewest number of occurrences. This phenomenon is due to the fact that the digits 0,9 (and sometimes 1) are reserved in some positions for pay phones, businesses, etc. The general falloff in number of occurrences from 2 to 8 is due to the manner in which the telephone numbers were assigned in the Portland area.

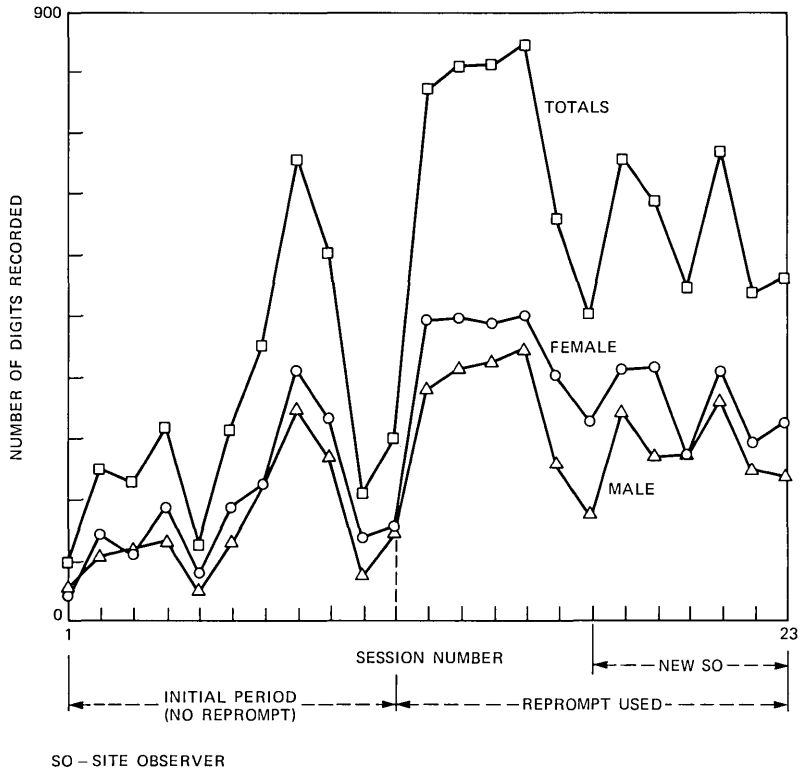


Fig. 2—The number of digits recorded as a function of session number for males, females, and combined (totals).

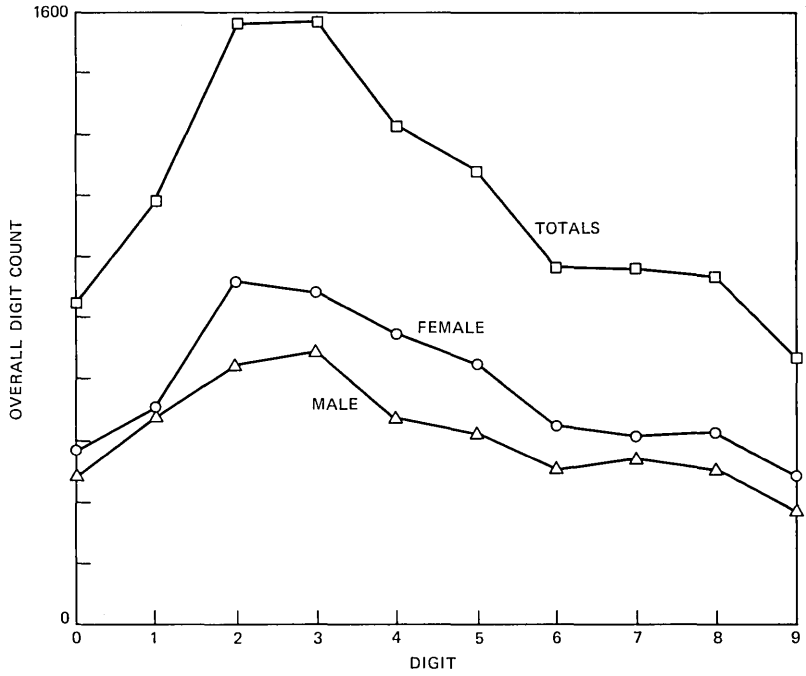


Fig. 3—Overall digit count as a function of the digit for males, females, and combined (totals).

There were several problems that occurred during the recording phase. These problems were classified as being in one of two groups. The first group contained problems associated with the telephone transmission conditions. Artifacts such as clicks, tones, and hum were often superimposed on the subject's speech. Resulting peak signal-to-noise ratios varied from as little as 10 dB to as much as 60 dB. The second group consisted of problems related to the talker and the environment in which he or she spoke. These included nonisolation of speech (i.e., the digits were connected) and the presence of extraneous background speech, such as people talking in the background or a television set being played at an audible level at the handset. Most of the user failures were severe enough to warrant elimination of the customer's speech from the database.

Figure 4 shows a plot of the percentage of calls accepted (i.e., at least 1 digit was extracted from the call) as a function of the session number. The rejected calls are a sum of both SO and MHO rejections. We can see that only 53 percent of calls yielded at least 1 digit. There were several reasons for such a low yield. The main reason for rejection was nonisolation of speech. This took one of two forms; either the subjects spoke in pairs of digits (e.g., 43 followed by 27), or they connected all four digits (e.g., 4327). Figure 5 shows a plot of the

percentage of these two types of rejections, as a function of total rejections, on a per-session basis. During the initial phase of recording, digit pair rejections accounted for about 30 percent of all rejections. Similarly, fully connected speech accounted for approximately 27

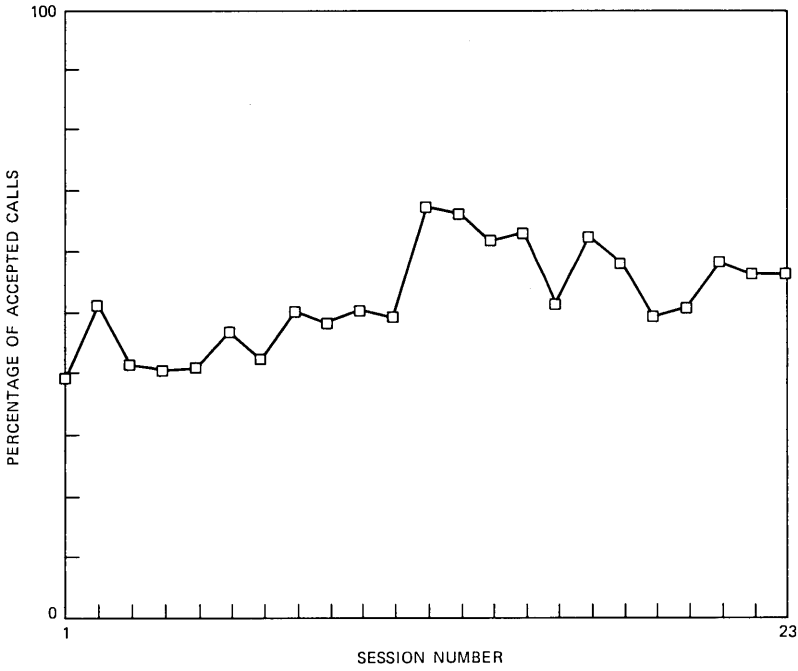


Fig. 4—The percentage of handled calls that were accepted as a function of session number.

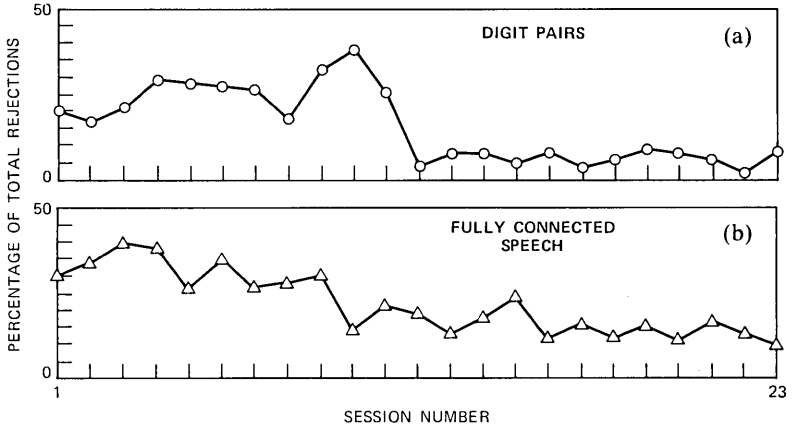


Fig. 5—Plots of percentage of total rejections because the number was spoken as (a) a set of digit pairs (b) as fully connected speech as a function of session number.

percent of total rejections. The plots show that a strong decrease in the number of these rejections occurred after the reprompting phase was initiated (Session 12), resulting in rejections rates of 9 and 15 percent for digit pairs and connected speech, respectively.

Another problem encountered during the recording process was the failure of the automatic endpoint detector to segment all the words properly. This algorithm[9] used the log energy contour of the speech and, after normalizing for background noise level, found as many digits as possible in the recorded string. For "ideal" transmission conditions and correct speaking of the number (i.e., as a sequence of isolated digits) the task of digit detection is relatively straightforward and essentially makes no errors.[9] Figure 6a illustrates such a case for the digit string /5946/. This figure shows the log energy contour of the recording. The dashed vertical lines indicate beginning and ending frames for automatically detected digits. For this example the peak-signal-to-average-background-noise ratio was about 54 dB (i.e., the average signal-to-noise ratio was close to 40 dB). Furthermore, the background noise was fairly stationary and at a low level.

Unfortunately, most of the recordings differed substantially from that of Fig. 6a. Figures 6b through 6e illustrate some problems that made automatic reliable detection of the digits very difficult. Figure 6b shows a log energy contour of the digit sequence /5282/ in which the signal level (during talking) was fairly low (peak-signal-to-background-noise ratio of 26 dB), and to further complicate matters, the background consisted of a mixture of noise and switching transients (clicks) generated within the telephone plant. For this sequence only three of the digits were automatically detected; the second digit was missed.

Figure 6c illustrates a case where the first three digits of the string /2383/ were spoken without a pause between the digits (i.e., in a connected format) and thus only one isolated digit could be used from this recording.

Figure 6d illustrates a case where the voice signal was corrupted by continuous signaling tones throughout the recording interval. The tones were set at a sufficiently high level so that the peak signal-to-tone ratio was only about 35 dB. Although, for this example, the locations of the major portions of each of the four spoken digits (4672) were properly detected, the parameterization of the signal (used later to recognize the digits) was greatly distorted by the tones present while the digits were being spoken. Furthermore, only the initial vowel portion of the digit six was located. The ending frication was lost in the tonal background.

Figure 6e shows an extreme case in which the background level of the line consisted of high-level noise and other extraneous sounds (i.e.,

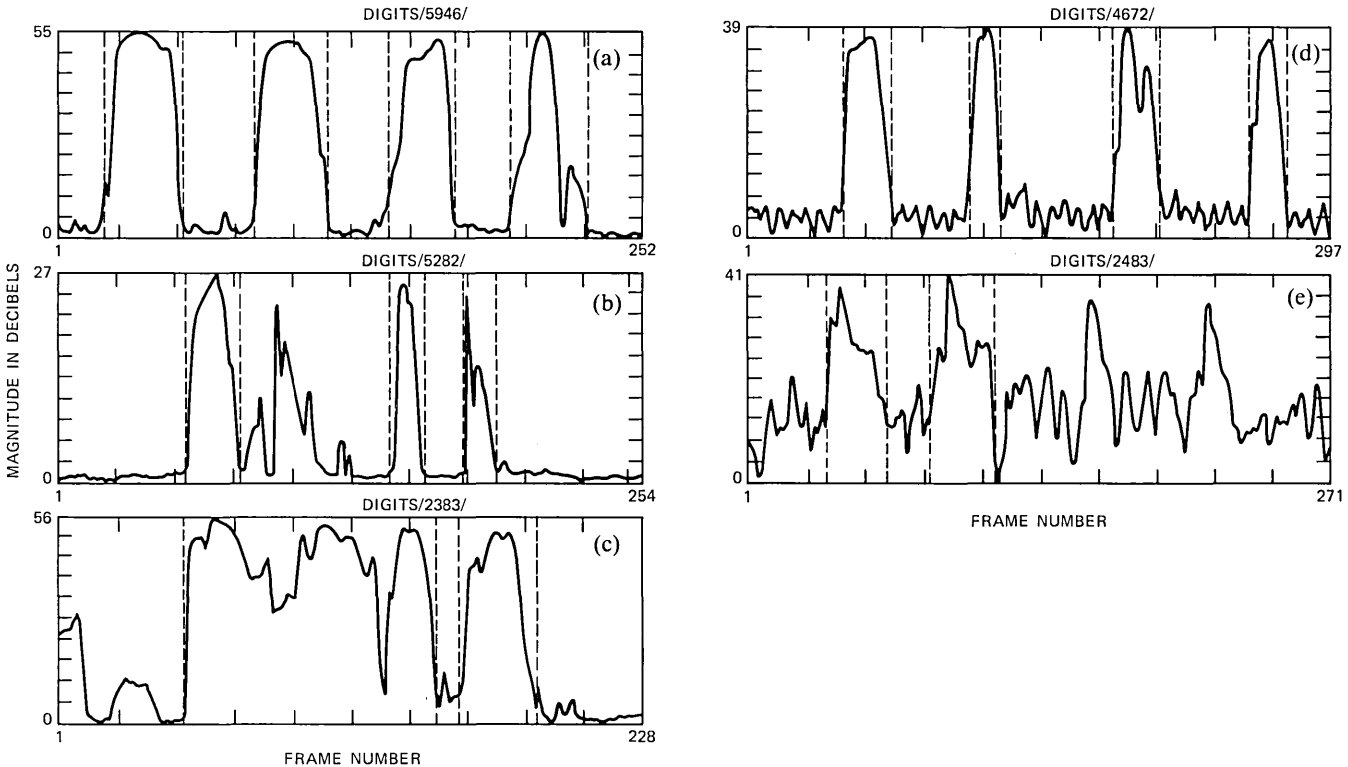


Fig. 6—Log energy contour for: (a) a high-quality recording of a 4-digit sequence; (b) a poor quality recording of a 4-digit sequence with low signal level and transmission clicks in the background; (c) a 4-digit sequence where the first three digits were not spoken in isolation; (d) a 4-digit sequence corrupted by continuous signaling tones throughout the recording interval; (e) a 4-digit sequence corrupted by high background noise and other extraneous sounds.

a very poor line). For this case it was impossible to detect accurate beginning and ending locations for any of the digits in the spoken string (/2483/).

The examples in Figs. 6d and 6e justify the importance of having the MHO check the accuracy of the automatic endpoint detector. In cases in which reliable endpoints are obtained automatically (no errors) the MHO allows the program to store the new digits in the database and update the relevant statistics. In all other cases the MHO can change endpoints or eliminate digits entirely. In this manner recordings with fewer than four isolated digits can provide one or more digits to the database and therefore the call is not entirely wasted. A discussion of the endpoint accuracy will be given in a later section.

An indication of how well the automatic endpoint detector performed was how often the MHO had to modify the endpoint sets. Figure 7 shows a plot of the percentage of calls requiring 0, 1, 2, 3, or 4 endpoint changes as a function of session number. This figure indicates that before using the reprompt, about 45 percent of the accepted calls needed no changes in word endpoints—i.e., no endpoint errors were made. After the reprompt was introduced the percentage rose to about 55 percent—i.e., a gain in endpoint accuracy of about 10 percent caused by the reprompt. This figure also shows that one set

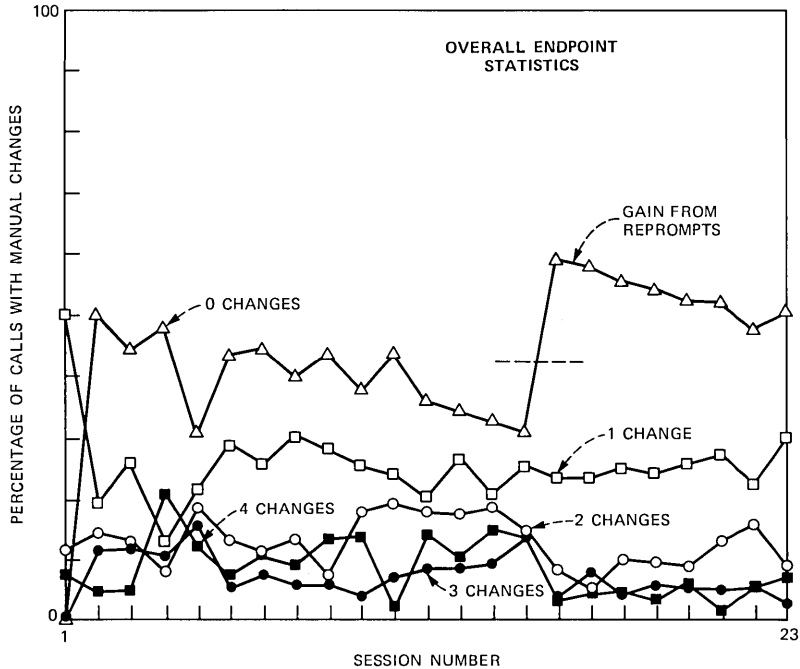


Fig. 7—Percentage of calls with from 0 to 4 sets of changed endpoints after manual corrections, as a function of session number.

of word endpoints needed to be modified about 25 percent of the time, and either three or four sets of endpoint modifications had to be made about 18 percent of the time. These results show clearly the necessity of improving the word endpoint detection. We are currently investigating into new methods of endpoint detection based on the types of problems encountered in this study.

III. ISOLATED DIGIT RECOGNITION EXPERIMENTS

3.1 Description of final database

After 23 days of recording over a two-month period, the final database contained 11,035 isolated digits taken from 3153 customers spoken during a wide range of telephone transmission conditions. Before any recognition experiments were performed, the full 11,035-digit database was listened to, and each digit was subjectively classified according to a set of background noise conditions to see if any one particular type of recording condition was harder for the recognition system to handle than another. Table I shows the individual types of conditions and the distribution of digits in the eight categories that were used. This table shows that 38.6 percent of the digits were classified as "acceptable". This meant that the background noise on the line was low (a signal-to-noise ratio in the range of 40 to 60 dB), and the given subject spoke in a clear, articulate voice, implying that the endpoint detector would have little problem with these utterances. These words were judged to be the "best" tokens, as close to laboratory data as was possible. The second category had similar signal-to-noise ratios, except here the callers did not speak in a normal fashion (presumably due to the novelty of being asked to speak in an isolated digit format). In such cases the customers dragged out words, or pronounced them in an abnormal way. This accounted for 10.3 percent of the digits. The remaining categories were used to describe the type of background noise present on the line. (If no background noise was present classes 1 and 2 were used.) About 12.5 percent of the digits came from strings that had a loud "crackling" noise superimposed on

Table I—Distribution of digits into categories based on background characteristics

Condition/Code	Count	Percent
Acceptable/1	4257	38.6
Customer-Related Problems/2	1133	10.3
Crackling Noise/3	1379	12.5
Pops, Clicks/4	1741	15.8
Tones, Whines/5	1386	12.6
Hum/6	752	6.8
Whirring/7	129	1.2
Background Speech/8	258	2.3
Totals	11035	100

the speech. Another 15.8 percent had pops and/or clicks throughout the recording. About 12.6 percent of the digits had loud tones present, mostly at 2600 Hz; and another 6.8 percent had loud "humming" noises superimposed with the speech from 200 to 400 Hz. We classified about 1.2 percent of digits as having a noise between "crackling" and "hum". (After we listened further we realized that this category could have been eliminated and its members classified as "crackling" noise.) The final 2.3 percent of the digits had background speech present—either other people's conversations, or television or radio sounds.

Table II shows the distribution of the 3153 calls, categorized by the number of digits (1, 2, 3, or 4) that were obtained from the calls. We see that for 68.8 percent of the calls four digits were actually obtained from the subject's spoken input. The average string length was 3.5 digits.

3.2 Description of recognition experiments

To determine how well we could recognize digits from this database, eight isolated word speech recognition experiments were performed. In the first experiment, the template set consisted of a set of 12 speaker-independent templates for each of the 10 digits plus the word "oh" (since the majority of talkers used "oh" instead of "zero" for the digit 0). The template set (MH templates) was obtained several years earlier from a clustering analysis of the speech of 100 talkers (50 male, 50 female) with recordings having been made over a local, dialed-up telephone line, providing from 40 to 60 dB peak-signal-to-noise ratios for all recordings.[2] The recognizer was the LPC-based recognizer, which has been in use in the Acoustics Research Department for several years.[1,2]

Figure 8 shows the results of this first recognition experiment. Figure 8a shows plots of the recognition accuracy for the top word candidate and the top two-word candidates as a function of the session number. The overall average word accuracy for the top candidate was 77.4 percent and was basically steady (to within statistical variations) with time.

Figure 8b shows a plot of the overall individual variations in digit accuracy. The digit 8 attained an accuracy of close to 95 percent, whereas the digit 4 was only 50 percent accurate. The major problem

Table II—Distribution of calls yielding 1, 2, 3, or 4 digits

	No. of Digits in String*				
	1	2	3	4	Total
Count	109	375	501	2168	3153
Percent	3.4	11.9	15.9	68.8	100

* Average string length −3.5 digits.

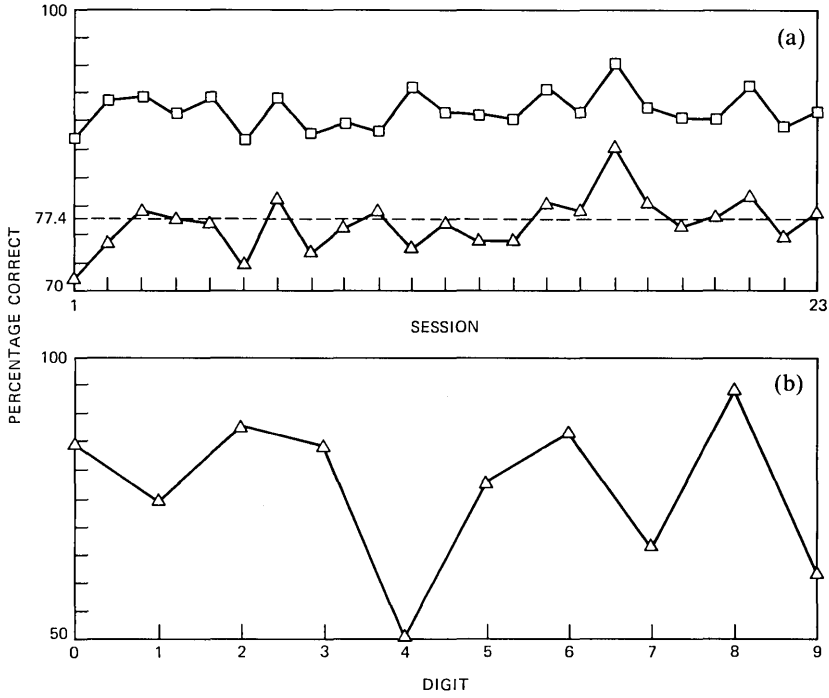


Fig. 8—Average digit recognition accuracy from: (a) MH templates for the top candidate and the top two candidates as a function of session number; (b) MH templates as a function of the spoken digit.

with the digit 4 was that about half the talkers pronounced it as /foe/, rather than /fore/, as represented in the template set. Hence for all such cases the digit 4 was recognized as 0 (i.e., the templates for "oh" provided the best match). Similarly, a modest number of confusions were found between the digits 5 and 9, and 1 and 9.

There was a basic problem with the training set of the first experiment. The pronunciations of the digits, as were prevalent in the Portland, Maine, area, were not well represented within the templates, and the difficult noise recording conditions led to large analysis degradations. Therefore, a second recognition experiment was run in which about 35 percent of the database was selected at random on a per-digit basis as a training set from which a new set of word reference templates [Portland (PO) random templates] was created. Reference template sets were generated using the UWA clustering algorithm of Rabiner and Wilpon,[6] yielding 12-, 20-, 25-, and 30-template-per-word sets, with the 30-template-per-word set yielding the best recognition results. The entire database of 11,035 digits was again used as a test set and the recognition results are given in Figs. 9 through 11 for the 30-template-per-word set. Figure 9a shows plots of the average word

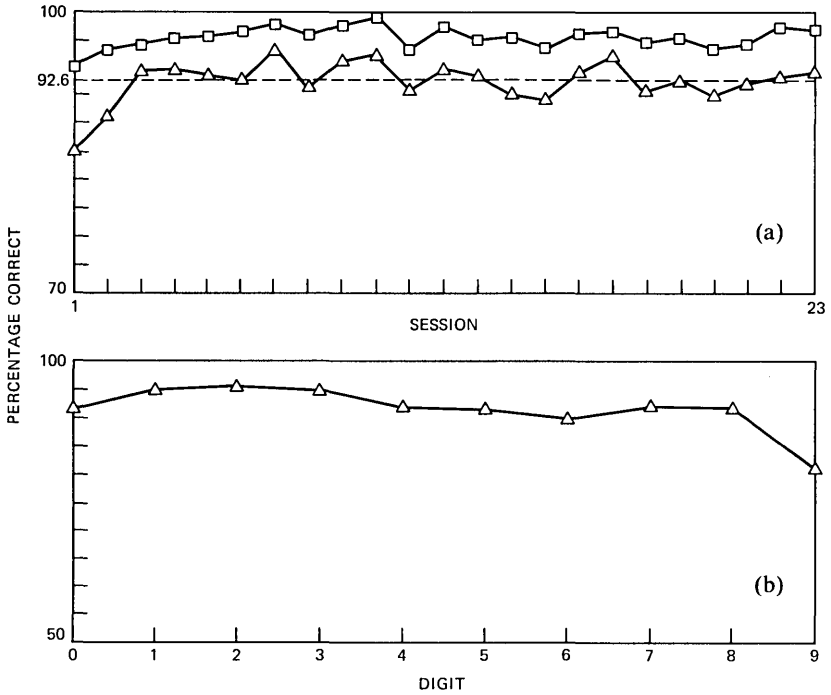


Fig. 9—Average digit recognition accuracy from: (a) PO random templates for the top candidate and the top two candidates as a function of session number; (b) PO random templates as a function of the spoken digit. (The same scale is used in Fig. 8b for comparison.)

recognition accuracy as a function of session number for both the top and the top two word candidates. The overall average word recognition accuracy for the top candidate was 92.6 percent and the individual session scores were fairly constant in time.

Figure 9b shows the overall recognition rates for the individual digits. We now see that all digits, except 9, were recognized with greater than 90 percent accuracy. Interestingly, the digit 9 was the one with the fewest tokens in the training set; hence improved recognition on 9 might result from a larger training set.

Figure 10 shows a plot of the overall word recognition accuracy as a function of the number of templates used per digit. The recognition rate is essentially flat for about 18 or more templates per word; hence only small reductions in accuracy would result from reducing the computation by almost one half.

Finally, Fig. 11 shows the individual accuracy scores for each digit as a function of the number of templates per digit. For some digits, e.g., 4, 6, 8, the recognition accuracy saturates for a small number of templates per digit, whereas for other digits, e.g., 0, 5, 7, 8, a large

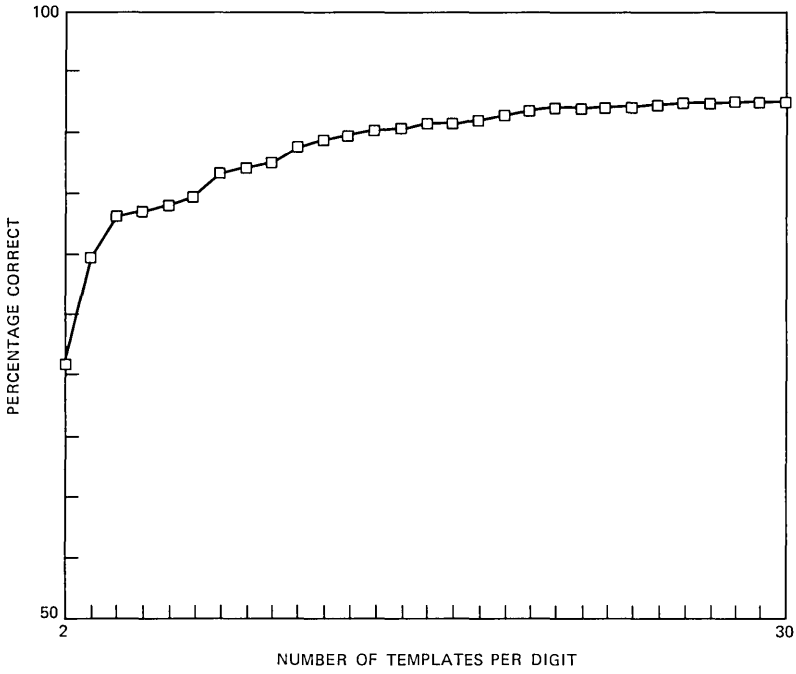


Fig. 10—Average digit recognition accuracy from PO random templates as a function of the number of templates used per digit.

number of templates per digit are needed. These results suggest that a reference set with a variable number of templates per digit could conceivably perform as well as the 30-template-per-digit set.

A problem in using a random set of utterances to train the system is the wide variability of transmission conditions present in the training tokens. The clustering procedure used to generate the templates tries to split the different spoken versions of a word into "similar" groups. If we now add an independent component to the speech, namely transmission variability, one would expect clusters also to be formed based on similarities in background conditions. Therefore we proposed the following recognition test. We used all the data that had been classified as "acceptable" (about 35 percent of the entire database) as a training set, from which we obtained a set of word reference tokens (PO clean templates). Since we were interested in a comparison with the PO random template experiment, a total of 30 reference templates were created for each digit. Again, the entire 11,035-word database was used as a testing set. The recognition results are summarized in Figs. 12 and 13. Figure 12a shows plots of recognition accuracy as a function of session number for the top and the top two word candidates. The average recognition accuracy over all days was

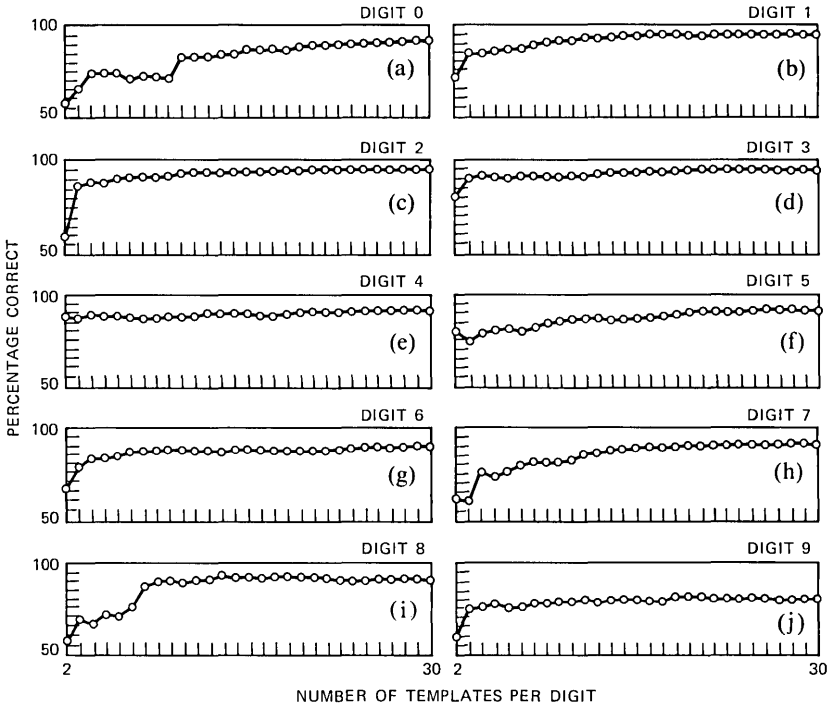


Fig. 11—Average recognition accuracy from PO random templates, for each individual digit, as a function of the number of templates used per digit.

93.1 percent. Compared with the PO random templates the difference in recognition accuracy was only 0.6 percent. Several inferences can be made from this result. First, it may not be necessary to go through the very time-consuming job of listening to several thousand spoken digit sequences and extracting only the "best" for use in training. Second, this result may indicate that the population of approximately 3900 training tokens was large enough to encompass all speaker variations (for most digits) and also all types of transmission system degradations.

Figure 12b shows a plot of the individual digit recognition scores. We see that the digit 9 still has a lower accuracy than the other digits. Figure 13a shows digit recognition accuracy as a function of condition code for the condition codes described previously. An accuracy of 96.6 percent was obtained when the testing set was also from the acceptable category (i.e., the same as the training data). The worst recognition results (86.7 percent) came from the digits categorized as having a very pronounced "humming" sound present during recording. Figure 13b shows string recognition accuracy as a function of condition code. The highest accuracy was 89.1 percent for "acceptable" data and the

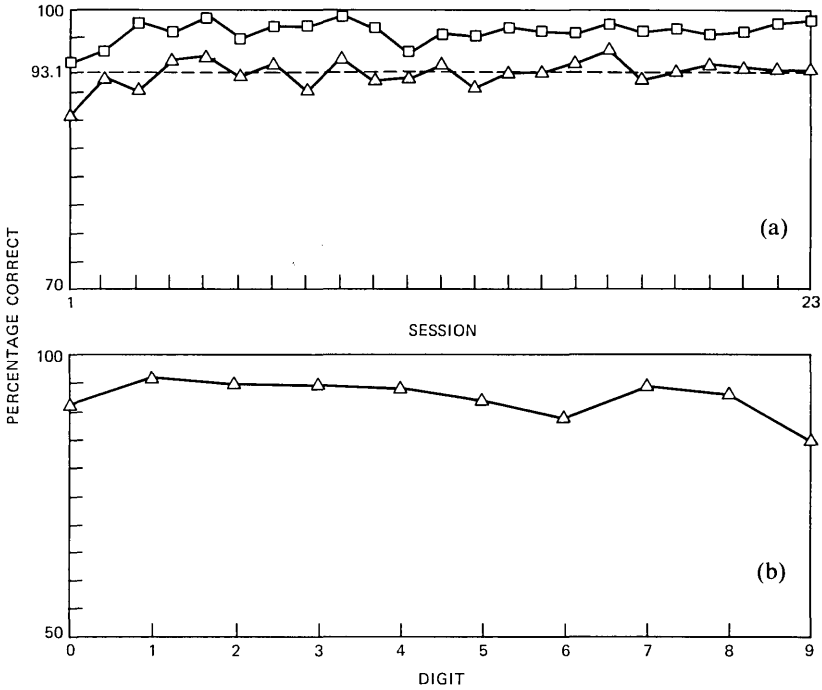


Fig. 12—Average digit recognition accuracy from the (a) PO clean template set for the top and top two candidates as a function of session number; (b) as a function of the spoken digit.

lowest accuracy was 64.3 percent for strings within the "hum" classification. The average string accuracy over all conditions was 80.8 percent. It was clear from these figures that the recognition errors were not distributed uniformly over all recording conditions. As a result, we were interested in determining whether digit errors were independent within each classification. Dependency might indicate that some sort of syntax could be applied to improve overall string accuracy.

To test the assumption that digit errors were independent within each string and classification category, a simple Bernoulli model was assumed in which the probability of error of a single digit was called α. Under these conditions, the probability, P, of four correct recognitions within a string of four digits is

$$\begin{aligned} P(\text{4 correct}) &= 1 - P(\text{single error}) + P(\text{double error}) \\ &\quad - P(\text{triple error}) + P(\text{quadruple error}) \\ &= 1 - 4\alpha + 6\alpha^2 - 4\alpha^3 + \alpha^4. \end{aligned} \tag{1}$$

For small α's eq. (1) can be approximated as

$$P(4 \text{ correct}) \approx 1 - 4\alpha. \quad (2)$$

Given that the average string length over all strings recorded was 3.5 digits

$$P(\text{string correct}) \approx 1 - 3.5\alpha. \quad (3)$$

In testing this hypothesis if we look at Fig. 13a, and multiply the digit error rates by 3.5 for each condition, we immediately see that eq. (3) above does indeed hold (at least to a reasonable level of accuracy). This indicates that within each classification the errors were probably distributed uniformly and randomly.

The next series of recognition experiments was carried out to determine whether we were using too many training tokens and therefore could reduce the amount of training data collected in future studies. Reference sets were created using one-third, one-half, and two-thirds of the data used in creating the 30-template-per-word PO clean template set (about 11.5, 17.2, and 23 percent of the entire database, respectively). The results are given in Tables III and IV. Table III shows the number of tokens in each training set or each digit, and Table IV shows the per-digit minimum, maximum, and

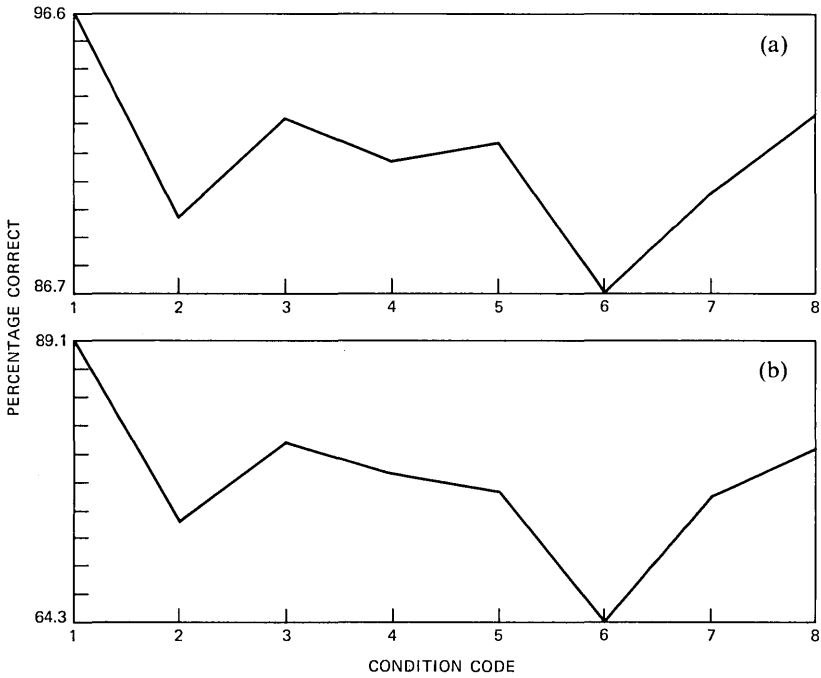


Fig. 13—Recognition accuracy as a function of PO clean template set for the: (a) acoustic classification on a per-digit basis; (b) acoustic classification on a string basis.

average accuracies for each experiment performed. The average per-digit recognition accuracy for the 12-template-per-word PO clean (one-third) template set was 89.7 percent for the 20-template-per-word PO clean (one-half) template set it was 91.4 percent; and for the 25-template-per-word PO clean (two-thirds) template set it was 92.2 percent.

The results show that the digit 9 consistently had the worst recognition scores. One possible explanation for this could be that in all cases the digit 9 had the smallest training set. The next recognition experiment was run to test this hypothesis. In this experiment we used a fixed number of training tokens per word. We were interested in seeing whether the 9 scores improved relative to the other digits. (Obviously we did not expect the overall accuracy to improve.) The results of this experiment are given in Table IV. We can see that the recognition accuracy for 9 has not changed (compared to the PO clean

Table III—Distribution of training tokens for individual words for each recognition experiment

	MH	PO Random	PO Clean	PO Clean (1/3)	PO Clean (1/2)	PO Clean (2/3)	PO Clean (Fixed)	PO Clean Variable
0	100	306	280	93	140	186	233	280
1	100	450	391	130	195	260	233	391
2	100	576	519	173	260	346	233	519
3	100	540	562	187	281	374	233	562
4	100	432	462	154	231	208	233	462
5	100	360	415	138	208	272	233	415
6	100	306	301	100	150	200	233	301
7	100	360	325	108	162	216	233	325
8	100	306	314	104	157	208	233	314
9	100	234	233	77	166	154	233	233

Table IV—Summary of recognition accuracies for the eight recognition experiments

	MH	PO Random	PO Clean	PO Clean (1/3)	PO Clean (1/2)	PO Clean (2/3)	PO Clean Fixed	PO Clean Variable
0	84.6	91.5	91.1	85.9	89.3	90.0	92.5	93.3
1	75.0	94.9	96.0	90.0	94.7	95.0	95.5	94.8
2	87.8	95.8	94.9	95.0	94.5	95.5	94.4	93.1
3	84.3	95.2	94.7	94.5	95.2	95.5	94.6	95.0
4	50.5	92.3	94.3	91.6	91.1	90.8	89.7	93.2
5	78.0	91.9	92.2	86.8	89.3	90.8	90.2	91.5
6	86.8	90.3	89.1	93.9	90.0	89.0	89.0	88.8
7	66.9	92.4	94.7	86.0	91.3	92.7	92.3	95.3
8	94.6	92.1	93.6	92.1	88.7	94.7	93.1	93.5
9	62.3	81.6	85.3	82.3	83.3	80.9	85.6	84.7
Average	77.4	92.6	93.1	89.7	91.4	92.2	92.1	92.6
Minimum (Word)	50.5(4)	81.6(9)	85.3(9)	82.3(9)	83.3(9)	80.9(9)	85.6(9)	84.7(9)
Maximum (Word)	94.6(8)	95.8(2)	96.0(1)	95.0(2)	95.2(3)	95.5(2)	95.5(1)	95.3(7)

template set), while the accuracies for the other digits have fallen. These results imply that some digits, e.g., 2, 3, need significantly fewer training tokens then do other digits, like 9.

In the final recognition experiment, we investigated the use of a variable number of templates per word. The number of templates per word was chosen on the basis of the curves of Fig. 11 at the point where the recognition accuracy leveled off. The training set for this experiment was the same as used in the PO clean template set, namely, approximately 35 percent of the entire database for each of the 10 digits. The average recognition accuracy using this new template set was 92.6 percent. This was only 0.5 percent worse than the recognition accuracy of the PO clean template set, but with 100 fewer templates (i.e., about two-thirds of the computation).

3.3 Imposing a rejection threshold

Figures 14 through 16 demonstrate the effects of imposing a threshold on the recognition task. A recognition distance score above the

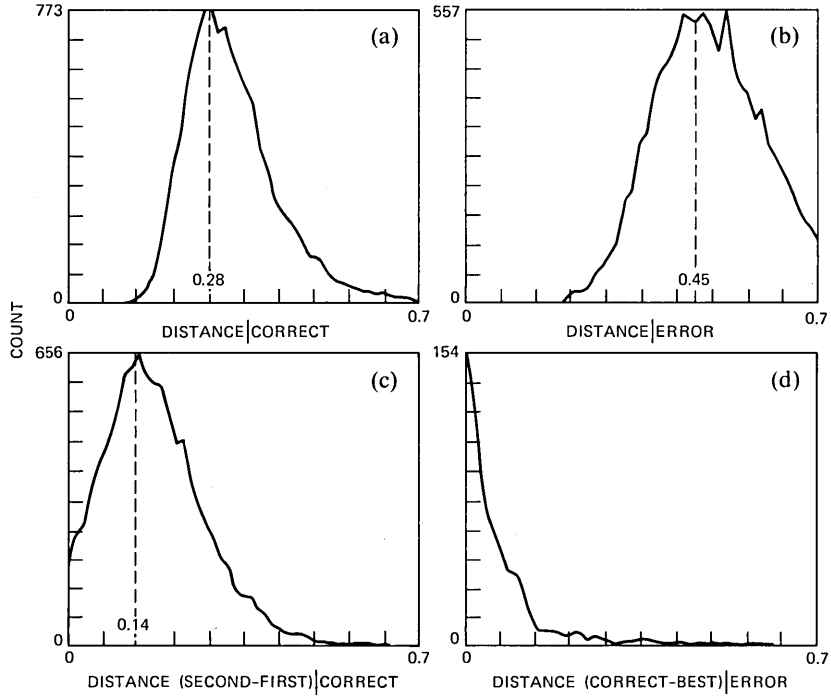


Fig. 14—Plot of histograms of LPC distances for (a) the correct word; (b) the closest incorrect word; (c) the difference between the best and second best choice, given the best choice is correct; (d) the difference between the correct word and the top recognition candidate, given the top candidate is not the correct word.

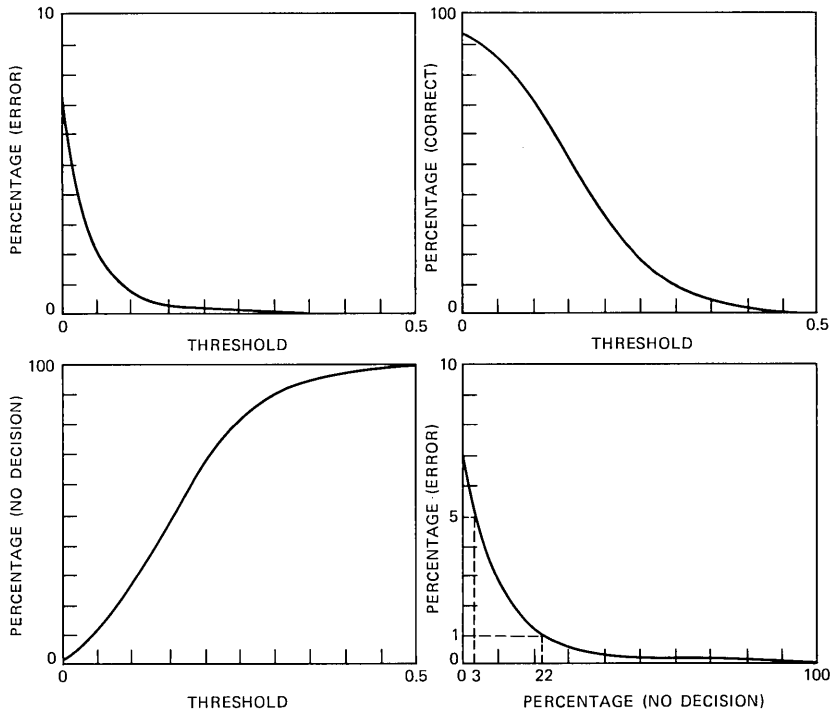


Fig. 15—Plot showing trade-off between no-decision rate and error rate on a per-digit basis, given a thresholding scheme imposed on the recognizer.

threshold would lead to a result of no decision by the recognizer. Figure 14a shows a histogram of LPC distances for the correct word (i.e., all 11,035 digits), with a mean correct distance of 0.28. Figure 14b shows a histogram of the scores for the closest incorrect word, with a mean distance of 0.45. Figure 14c shows a histogram of scores of the difference between the best choice and the second best choice, given the best choice is correct. Its mean of 0.14 indicates that when a word is recognized correctly the next closest word will on average have a distance score about 50 percent greater. Since LPC distances are on a log scale, this difference is a relatively large one. Figure 14d shows a histogram of the difference between the correct word and the top recognition candidate given the top candidate is not the correct word. This plot shows that when a word is misrecognized, the correct word has an LPC distance very close to that of the best choice. These results imply that a thresholding scheme can be applied to the recognition system and yield a good trade-off between accuracy and no-decision choices. Figure 15 shows the results of implementing such a thresholding technique. This plot shows the percent no decisions versus percent error rate, for the PO clean template set. We see that

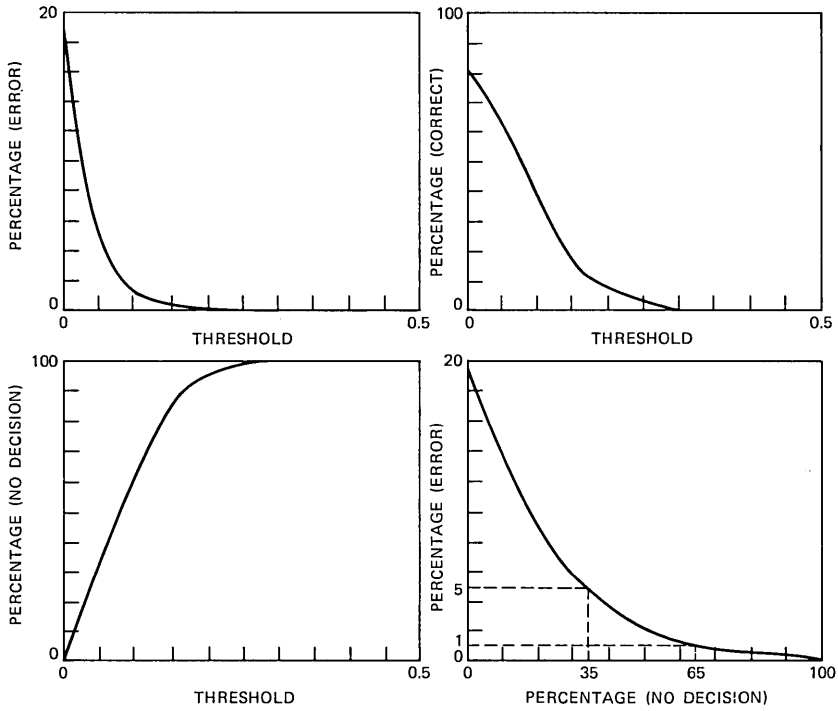


Fig. 16—Plot showing trade-off between no-decision rate and error rate on a string basis, given a thresholding scheme imposed on the recognizer.

if the task for which this recognition system is to be used can only tolerate a 1-percent per-digit error rate, a no-decision rate of 22 percent must also be accepted. However, a 5-percent probability of error yielded only 3 percent no decisions. Figure 16 shows a similar set of trade-off curves for full digit strings. For a digit string with an average of 3.5 digits, a 1-percent string error rate leads to a 65-percent no-decision rate and a 5-percent string error rate leads to a 35-percent no-decison rate. These results suggest that even though thresholding can be used to reduce error rate, other methods that do not lend themselves to such high no-decision rates should be investigated.

IV. DISCUSSION

The results presented in Sections III show that:

1. Significant problems exist in getting casual telephone customers to speak telephone numbers as a series of isolated digits. These are related to human factors issues (people don't normally speak in isolated word sequences) and problems induced by a wide variety of transmission and switching conditions.

2. The ability to detect words automatically in noisy or nonideal

environments is not adequate for about 50 percent of the recordings. There are some obvious ways of improving the current algorithm for finding words; and the database we collected in this test is currently being used to test various modifications of the algorithm.

3. The recognition results given in Table IV show that the accuracy of the recognizer, even when trained on a subset of the test data, is at best marginally acceptable with a maximum accuracy of 93.1 percent. However, some digits had recognition accuracies well over 93.1 percent, while others (i.e., the digit 9) had scores around 85 percent. More detailed analyses of the type of errors and their causes must be undertaken and some improvements in the training and recognition procedures must be made to approach the accuracies obtained for laboratory recordings (close to 98 percent).

Each of the above problems will be investigated carefully to improve all aspects of automatic recording, detection, and recognition of isolated words.

Another data collection exercise will begin soon at another site. Such issues as the effects of regional dialect on the reference templates will be studied. In addition, further testing of an improved endpoint detector will be performed.

V. SUMMARY

In this paper we have presented recognition results obtained from a speech database, consisting of 11,035 isolated digits, collected in an actual telephone environment from 3100 nonsolicited subjects. Several different reference template sets were used with a maximum recognition accuracy reported at 93.1 percent.

VI. ACKNOWLEDGMENTS

The authors would like to acknowledge the many individuals involved in collecting the data described in this paper. In particular, Eddie Youngs, Cindy Karhan, Marty Viets and Helen Hollinka provided the interface to the Bell System and were responsible for direct recording of the spoken digit sequences. Frank Pirz and Keith Bauer designed, built, and programmed the special-purpose recording hardware. Finally, Don Bock and Sandy MacNeil aided in the collection, processing, and analysis of the customer data at the Murray Hill laboratory. To each of these individuals we express our thanks for their help.

REFERENCES

1. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 67–72.
2. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker

Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 4 (August 1979), pp. 336–49.
3. L. R. Rabiner and J. G. Wilpon, "Speaker Independent, Isolated Word Recognition for a Moderate Size (54 word) Vocabulary," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 6 (December 1979), pp. 583–7.
4. J. G. Wilpon, L. R. Rabiner, and A. F. Bergh, "Speaker Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary," J. Acoust. Soc. Amer., *72*, No. 2 (August 1982), pp. 390–6.
5. L. R. Rabiner and J. G. Wilpon, "A Simplified, Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68*, No. 5 (October 1980), pp. 1069–70.
6. L. R. Rabiner and J. G. Wilpon, "Considerations in Applying Clustering Techniques to Speaker Independent Word Recognition," J. Acoust. Soc. Amer., *66*, No. 3 (September 1979), pp. 663–73.
7. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated Word Recognition for Large Vocabularies," B.S.T.J., *61*, No. 10, Part 1, (December 1982), pp. 2989–3005.
8. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Recognition—Theory and Selected Applications," IEEE Trans. Commun., *29*, No. 5 (May 1981), pp. 621–59.
9. L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-29*, (August 1981), pp. 777–85.

AUTHORS

Jay G. Wilpon, B.S., A.B. (cum laude) in Mathematics and Economics, respectively, 1977, Lafayette College, Easton, Pa; M.S. (Electrical Engineering/Computer Science), 1982, Stevens Institute of Technology, Hoboken, N.J.; Bell Laboratories, 1977—. Since June 1977 Mr. Wilpon has been with the Acoustics Research Department at Bell Laboratories, Murray Hill, N.J., where he is a Member of the Technical Staff. He has been engaged in speech communications research and is presently concentrating on problems of speech recognition.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D., 1967 (Electrical Engineering), The Massachusetts Institute of Technology; Bell Laboratories, 1962—. From 1962 through 1964, Mr. Rabiner participated in the cooperative plan in electrical engineering at Bell Laboratories. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983). Former President, IEEE, ASSP Society; former Associate Editor, ASSP Transactions; former member, Technical Committee on Speech Communication of the Acoustical Society, ASSP Technical Committee on Speech Communication; Member, IEEE Proceedings Editorial Board, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

Comparing Batch Delays and Customer Delays

By W. WHITT*

(Manuscript received January 25, 1983)

For a large class of queueing systems in which customers arrive in batches, Halfin (1983) showed that the delay distribution of the last customer in a batch to enter service coincides with the delay distribution of an arbitrary customer when the batch-size distribution is geometric. Halfin's result can be applied to study the performance of complicated communication systems in which messages are divided into packets for transmission. Then packets are customers and the delay of a message is the delay of the last customer in a batch to enter service. If the assumptions are satisfied and if packet delays are easier to analyze, then packet delays can be used to calculate message delays. In this paper, we show that these two delay distributions are stochastically ordered when the batch-size distribution is NBUE or NWUE (new better or worse than used in expectation). The delays of arbitrary customers tend to be less (more) when the batch-size distribution is NBUE (NWUE). In addition to the bounds provided by the stochastic ordering, we also suggest an approximation for the relation between the two expected delays based on known results for the $M^B/G/1$ queue having a batch-Poisson arrival process.

I. INTRODUCTION

Halfin[1] recently showed that for a large class of queueing systems in which customers arrive in batches the delay distribution of the last customer in a batch to enter service equals the delay distribution of an arbitrary customer when the batch size has a geometric distribution. Halfin's result is important when we study the performance of communication systems in which messages are divided into packets for

* Bell Laboratories.

transmission. Then packets are customers and the delay of a message is the delay of the last customer in a batch to enter service. Halfin's result is useful to approximately describe message delays in complicated communication systems, e.g., with contention-resolving schemes such as Carrier Sense Multiple Access (CSMA).[1] In many of these systems message delay is more important but more difficult to analyze than packet delay. Halfin provides conditions under which packet delay results directly yield message delay results.

In this paper we identify conditions on the batch-size distribution under which the delay distribution of the last customer to enter service is stochastically greater than or equal to the delay distribution of an arbitrary customer. We show that it suffices for the batch-size distribution to be NBUE (new better than used in expectation). The ordering is reversed for batch-size distributions that are NWUE (new worse than used in expectation). The batch size B has an NBUE distribution if its mean is greater than all conditional means given tail events, i.e., if $EB \geq E(B - n \mid B \geq n)$ for all n. A distribution is NBUE (NWUE) if it has increasing (decreasing) failure rate. The NBUE and NWUE properties are now quite standard for stochastic comparisons. For further discussion, see Barlow and Proschan,[2] and Whitt.[3]

As Halfin observed, his result about batch arrivals can be viewed as a special case of a discrete analog of Poisson Arrivals See Time Averages (PASTA).[4] In the same way, our stochastic comparison results parallel those for customer-stationary and time-stationary characteristics of queues.[5-7] Random quantities associated with batches (e.g., delays of the last customer in a batch) constitute an embedded sequence in the sequence of random quantities associated with all customers (e.g., the delays of arbitrary customers), just as customer arrival points or departure points constitute embedded sequences in continuous time.

Our approach is similar to that of the East German School (Franken et al.) because we begin in Section II with a stationary version, what we call an "equilibrium batch." In this setting, we easily obtain a representation of the expected average associated with an arbitrary customer that makes the desired comparisons and Halfin's result transparent. The representation involves the stationary-excess distribution of the batch-size distribution. (From the theory of stationary point processes, as contained in Franken et al., this can also be viewed as a consequence of the Palm theory.)

The relation between what an arbitrary customer sees and what the last customer in a batch sees can be explained in part as follows. One customer in each batch is the last customer in the batch, but there are more customers in big batches than in small batches. Hence, the distribution of the batch size containing a last customer is the ordinary

batch-size distribution, but the distribution of the position in a batch of an arbitrary customer is the batch-size stationary-excess distribution (see Section III.2 of Cohen,[8] Burke,[9] and Section 5.10 of Cooper[10]). Hence, the relation between the two kinds of delays reduces to the relation between the batch-size distribution and its associated stationary-excess distribution.

These stochastic comparison results are only qualitative. With the methods we use, we are unable to obtain related quantitative results. Moreover, the difference between the expected delays obviously will depend on the context, whereas the qualitative results here generally hold true. Following Halfin's example, we make very few assumptions for our stochastic comparisons. However, in Section III we examine quantitative results by considering the special case of the $M^B/G/1$ queueing model, which has a batch-Poisson arrival process. We get some idea about what to expect in more complicated situations by examining existing quantitative results for this special case. We use these results to obtain an approximate quantitative relationship between expected batch delays and expected customer delays as a function of the first two moments of the batch-size distribution.

In Section IV we briefly show how our equilibrium batch can be viewed as a time-average limit. Throughout we talk about batch arrivals to a queue, but as in Wolff[4] the results apply more generally. The model need not be a queue, and for a queueing model the process need not be an arrival process; for an arrival process the customers in a batch need not arrive together. The batch is simply a label assigned to random variables, as we explain in Section II.

II. THE EQUILIBRIUM BATCH

We consider a batch in equilibrium, defined in terms of a positive-integer-valued random variable B and a sequence of random variables $\{X_k, k \geq 1\}$, all defined on a common probability space. The variable B represents the size of the batch and the variable X_j is a random quantity associated with the jth customer in the batch. For example, X_j might be a function of the delay of the jth customer to enter service. We are only interested in X_j for $j \leq B$. In fact, we can consider X_j defined only on the subset $\{B \geq j\}$. Hence, we can speak of the basic data as the random vector $(B, X_1, \cdots, X_B)$.

The variables $X_1, X_2, \cdots$ are typically dependent. Our *basic assumption* is that the events $\{B = j\}$ for $j \geq k$ are independent of the random vector $(X_1, \cdots, X_k)$ for $k \geq 1$. This corresponds to the Lack of Anticipation Assumption (LAA) in Wolff.[4]

Let the probability mass function (pmf) of B be defined as

$$p_n = P(B = n), \qquad n = 1, 2, \cdots, \tag{1}$$

and let p_n^* be the *stationary-excess* pmf associated with p_n, defined by

$$p_n^* = \bar{p}_n \Big/ \sum_{k=1}^{\infty} \bar{p}_k, \qquad n \geq 1, \tag{2}$$

where

$$\bar{p}_n = \sum_{k=n}^{\infty} p_k, \qquad n \geq 1, \tag{3}$$

and

$$\sum_{k=1}^{\infty} \bar{p}_k = \sum_{k=1}^{\infty} k p_k = EB < \infty. \tag{4}$$

We are interested in the relationship between the random quantities associated with an *arbitrary* customer and the *last* customer in a batch so we can compare the expected values in these two cases. We interpret the expected value associated with an arbitrary customer as the expected value for all customers in the batch divided by the expected number of customers in a batch. (A way to justify this interpretation is described in Section IV). Because of our basic assumptions, these quantities are

$$A = \frac{\sum_{n=1}^{\infty} p_n(EX_1 + \cdots + EX_n)}{\sum_{n=1}^{\infty} np_n} \tag{5}$$

and

$$L = EX_B = \sum_{n=1}^{\infty} p_n EX_n. \tag{6}$$

The desired relationships between A and L are derived from the following alternate representation for A.

Theorem 1: $A = \sum_{n=1}^{\infty} p_n^* EX_n$.

Proof: Change the order of summation in (5) and apply (2) to obtain

$$A = \sum_{n=1}^{\infty} EX_n \left(\frac{\sum_{k=n}^{\infty} p_k}{\sum_{k=1}^{\infty} kp_k} \right) = \sum_{n=1}^{\infty} EX_n \left(\frac{\bar{p}_n}{\sum_{k=1}^{\infty} \bar{p}_k} \right) = \sum_{n=1}^{\infty} EX_n p_n^*.$$

Corollary 1: (*Halfin*[1]) *$A = L$ for all $\{EX_n\}$ if and only if $p_n = p_n^*$ for all n or, equivalently, if p_n is geometric, i.e.,*

$$p_n = (1 - p)p^{n-1}, \qquad n = 1, 2, \cdots \tag{7}$$

for some p, $0 \leq p \leq 1$.

Proof: Sufficiency is immediate. It is well known that p_n is geometric if and only if $p_n = p_n^*$ for all n (e.g., see Corollary 3.3 of Whitt[3]). Necessity is almost as easy: Choose different sequences $\{EX_n\}$, e.g., $EX_n = 1$ if $n = k$, and 0 otherwise.

Remark: Theorem 1 and its corollaries apply immediately to distributions as well as means. For example, our original sequence X_n can be replaced by $f(X_n)$ where $f(x) = I_{(-\infty,x)}$, so that A is the expected proportion of customers for which $X_n \le x$ and $L = P(X_B \le x)$.

We now establish inequalities between A and L as a function of the shape of the batch-size pmf p_n. These follow immediately from known stochastic-order relations between p_n and p_n^*.

For two pmf's p_n^1 and p_n^2 on the positive integers, we define stochastic-order relations $p_n^1 \le_{st} p_n^2$ ($p_n^1 \le_{ic} p_n^2$) to hold if

$$\sum_{k=1}^{\infty} f(k)p_k^1 \le \sum_{k=1}^{\infty} f(k)p_k^2 \tag{8}$$

for all nondecreasing (nondecreasing and convex) real-valued functions f on the positive integers for which the sums converge.

A pmf p_n is NBUE (new better than used in expectation) if

$$EB = \sum_{k=1}^{\infty} \bar{p}_k \ge \sum_{k=n}^{\infty} \bar{p}_k/\bar{p}_n = E(B - n \mid B \ge n), \qquad n \ge 1, \tag{9}$$

and NWUE with the inequality in (9) reversed.

From Theorem 1, we obtain:

Corollary 2: $A \le L$ for all nondecreasing sequences $\{EX_n\}$ if and only if $p_n^* \le_{st} p_n$ or, equivalently, if p_n is NBUE.

Proof: From (6), (8), and Theorem 1, $A \le L$ for all nondecreasing $\{EX_n\}$ if and only if $p_n^* \le_{st} p_n$. It is well known that $p_n^* \le_{st} p_n$ if and only if p_n is NBUE [e.g., see Theorem 3.2 (iii) of Whitt[3]]. For necessity, choose $EX_n = 1$ for $n \ge k$, and 0 otherwise.

Remarks: (1) Corollary 2 remains valid with $p_n^* \ge_{st} p_n$ instead of $\le_{st}$, which is equivalent to p being NWUE, if either $A \ge L$ or $\{EX_n\}$ is nonincreasing (but not both). (2) For Halfin's problem involving delays, note that the delays of the successive customers in any batch to begin service are nondecreasing with probability one.

We obtain another corollary using the stochastic-order relation $\le_{ic}$ defined in (8).

Corollary 3: $A \le L$ for all nondecreasing convex sequences $\{EX_n\}$ if and only if $p_n^* \le_{ic} p_n$ or, equivalently, if p_n new is better than p_n^* used in expectation, i.e., if

$$\sum_{k=1}^{\infty} \bar{p}_k \ge \sum_{k=n}^{\infty} \bar{p}_k^*/\bar{p}_n^*, \qquad n \ge 1.$$

Proof: Follow the proof of Corollary 2 using the convexity to treat $\leq_{ic}$. The characterization of $p_n^* \leq_{ic} p_n$ follows easily from the fact that $p_n^1 \leq_{ic} p_n^2$ if and only if $\sum_{k=n}^{\infty} \bar{p}_k^1 \leq \sum_{k=n}^{\infty} \bar{p}_k^2$; see Definition 3.1(iv) and Theorem 3.2(iv) of Whitt.[3]

III. The $M^B/G/1$ QUEUE

To illustrate the qualitative results and obtain some related quantitative results, we now consider the special case of the $M^B/G/1$ queue having a batch-Poisson arrival process (see Section 5.10 of Cooper[10]). Let W_F, W_A, and W_L be the equilibrium delay (waiting time before beginning service) of the first customer in a batch, an arbitrary customer, and the last customer in a batch. Let τ and σ^2 be the mean and variance of the service time; let m and $\hat{\sigma}^2$ be the mean and variance of the batch size B; and let B^* have the batch-size stationary-excess distribution p_n^* in (2), which has mean

$$EB^* = \sum_{n=1}^{\infty} np_n^* = (\hat{\sigma}^2 + m^2 + m)/2m. \tag{10}$$

Let λ be the rate of the Poisson process and, for stability, assume that $\lambda m\tau < 1$.

From Cooper, we obtain

$$E(W_F) = \frac{\lambda m^2\tau^2}{2(1 - \lambda m\tau)}\left(1 + \frac{m\sigma^2 + \tau^2\hat{\sigma}^2}{m^2\tau^2}\right). \tag{11}$$

From Theorem 1 and (10), we obtain

$$\begin{aligned} E(W_A) &= E(W_F) + \sum_{n=1}^{\infty} p_n^*(n-1)\tau \\ &= E(W_F) + [E(B^*) - 1]\tau \\ &= E(W_F) + \frac{(m-1)\tau}{2} + \frac{\hat{\sigma}^2\tau}{2m}. \end{aligned} \tag{12}$$

From (6), we obtain

$$\begin{aligned} E(W_L) &= E(W_F) + (EB - 1)\tau \\ &= E(W_F) + (m-1)\tau. \end{aligned} \tag{13}$$

Let c_B^2 be the squared coefficient of variation of the batch-size B, i.e., the variance of B divided by the square of the mean. We can rewrite (12) as

$$E(W_A) = E(W_F) + [m(c_B^2 + 1) - 1]\tau/2, \tag{14}$$

so that $EW_A \leq EW_L$ if and only if

$$mc_B^2 + 1 \leq m, \tag{15}$$

which is consistent with Corollaries 1 through 3 in Section II because $m = 1/p$ and $c_B^2 = 1 - p$ for the geometric distribution in (7).

In applying Theorem 1, we should observe that the lack of anticipation (LAA) assumption holds for the $M^B/G/1$ queue. Moreover, the Poisson property is only used to get (11); (12) and (13) remain valid provided that the LAA assumption holds. For example, the LAA assumption holds if the intervals between batch arrivals are a stationary sequence independent of the successive batch sizes.

Formulas (11) through (14) suggest an approximation for more general systems:

$$\frac{E(W_A) - E(W_F)}{E(W_L) - E(W_F)} = \frac{m(c_B^2 + 1) - 1}{2(m - 1)}. \tag{16}$$

IV. LIMITING AVERAGES

Instead of using the framework defined in Section II, we could also begin with a sequence of random vectors $\{(B_k, X_{k1}, \cdots, X_{kB_k}), k \geq 1\}$, where B_k represents the size of the kth batch, indexed in order of arrival, and X_{kj} represents the random quantity of interest (delay, etc.) associated with the jth customer to enter service in the kth batch. Let B_k take values on the positive integers and let X_{kj} be nonnegative for each k and j. Let Y_k be the random quantity associated with the kth customer indexed, first according to the batch and then according to the order of entering service, defined by

$$Y_k = X_{jm}, \qquad B_1 + \cdots + B_{j-1} + m = k \leq B_1 + \cdots + B_j$$

for $k \geq 1$.

The limiting average value of X_{kj} over *all* customers is naturally defined by

$$\bar{A} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} Y_k}{n}. \tag{17}$$

We shall work with the related quantity $\hat{A}$ defined by

$$\hat{A} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \sum_{j=1}^{B_k} X_{kj}}{\sum_{k=1}^{n} B_k}. \tag{18}$$

In most situations, the limits $\bar{A}$ and A exist and are equal.

We are interested in the relation between A and the limiting average value of X_{kj} over the *last* customer in each batch, defined by

$$\hat{L} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} X_{kB_k}}{n}. \tag{19}$$

We assume that there exists a random vector $(B, X_1, \cdots, X_B)$ such that

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} B_k}{n} = EB < \infty \tag{20}$$

and

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \sum_{j=1}^{B_k} X_{kj}}{n} = E \sum_{j=1}^{B} X_j < \infty, \tag{21}$$

so that

$$\hat{A} = \frac{E \sum_{j=1}^{B} X_j}{EB}. \tag{22}$$

We also assume that

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} X_{kj} I_{\{B_k = j\}}}{n} = P(B = j)EX_j. \tag{23}$$

To obtain (23) it is natural to assume that the basic independence assumption holds for each k, i.e., $\{B_k = j\}$ is independent of $(X_{k1}, \cdots, X_{kn})$ for all $j, j \geq n$. From (23), we have

$$\hat{L} = EX_B. \tag{24}$$

For eqs. (20) through (24) to be valid, in addition to the basic independence assumption, it suffices for the basic sequence $\{(B_k, X_{k1}, \cdots, X_{kB_k}), k \geq 1\}$ to be stationary and ergodic.

With (22) and (24) we obtain the framework defined in Section II.

V. ACKNOWLEDGMENTS

I am grateful to Shlomo Halfin for introducing me to this problem and to Moshe Segal for making helpful comments.

REFERENCES

1. S. Halfin, "Batch Delays Versus Customer Delays," B.S.T.J., this issue.
2. R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, New York: Holt, Rinehart and Winston, 1975.

3. W. Whitt, unpublished work.
4. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30*, No. 2 (March–April 1982), pp. 223–31.
5. D. König and V. Schmidt, "Stochastic Inequalities Between Customer-Stationary and Time-Stationary Characteristics of Queueing Systems With Point Processes," J. Appl. Prob., *17*, No. 3 (September 1980), pp. 768–77.
6. P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes*, Berlin: Akademie-Verlag, 1981.
7. S. Niu, unpublished work.
8. J. W. Cohen, *The Single Server Queue*, Amsterdam: North Holland, 1969.
9. P. J. Burke, "Delays in Single-Server Queues With Batch Input," Oper. Res., *23*, No. 4 (July–August 1975), pp. 830–3.
10. R. B. Cooper, *Introduction to Queueing Theory*, Second Edition, New York: North Holland, 1981.

AUTHOR

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; Bell Laboratories, 1977—. At Yale University, from 1973 through 1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At Bell Laboratories he is in the Operations Research Department. His work focuses on stochastic processes and stochastic models in operations research.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

Batch Delays Versus Customer Delays

By S. HALFIN*

(Manuscript received January 6, 1982)

For a large class of contention schemes with messages transmitted by subdividing them into packets, we show that the delay distribution of a message is the same as that for an individual packet. We conclude this from analyzing queueing systems with batch arrivals, where batch sizes have a geometric distribution and the queue discipline is indifferent to batch sizes and service times. There we prove that the customer (packet) delay distribution is the same as the batch (message) delay distribution, where delay is defined to be the delay of the *last served* customer in the batch. The proof is based on the discrete-time analog of the Poisson Arrivals See Time Averages (PASTA) theorem. We conclude that, in many cases, we can obtain message delays by calculating or measuring the packet delays, which is usually an easier task.

I. INTRODUCTION

Consider a queueing system with batch arrivals. We define the delay of batch to be the maximal delay of the customers in the batch. For example, consider the case where data messages arrive at a node of a network and await transmission. Each message is subdivided into packets that may be transmitted individually. Here the customers correspond to the packets and the batch to a message, and it is natural to say that the message is delayed as long as at least one of its packets is delayed.

Next, assume that the number of customers in a batch has a geometric distribution:

$$\Pr(\text{batch size} = n) = p(1-p)^{n-1}, \qquad n = 1, 2, \cdots.$$

* Bell Laboratories.

In our example, if the message lengths (in bits) are exponentially distributed, and each message m is "chopped" into $i(m)$ packets whose lengths (in bits) are independent and identically distributed and independent of the message length, then $i(m)$ has a geometric distribution. [Note that the $i(m)$th packet usually will not be full.]

We prove that for a wide class of such systems the batch delay in equilibrium has the same distribution as the individual customer delay. Here delay is the time from arrival to start of service. In Section II we discuss the various queue disciplines for which this result holds. A discete-time analog of the Poisson Arrivals See Time Averages (PASTA) theorem is introduced in Section III. The main result is stated and proved in Section IV. Section V contains some additional comments, and conclusions are stated in Section VI.

II. QUEUE DISCIPLINES

The stated result does not hold for all queue disciplines. Consider the case where customers are selected for service independent of their service times. Then it is well known that the expected delay of a customer is independent of the queue discipline.[1] On the other hand, the expected batch delay may depend on the discipline. For instance, if one always chooses the next customer from a batch with the smallest number of remaining customers, the expected batch delay will be smaller than if the next batch is chosen randomly, or in a preassigned order. This can be proved by arguments similar to those showing that giving preferential treatment to customers with small service time reduces the expected waiting time.[2] Thus, to prove equality of the delay distributions, such disciplines must be excluded.

Next, we characterize disciplines for which the result holds:

Definition: A queueing discipline will be called *impartial* if it selects customers independently of their services times, and selects batches independently of their sizes.

The following are examples of impartial disciplines:

- First in, first out (FIFO) for batches and for customers within batches.
- Last in, first out (LIFO) for batches and for customers within batches.
- Random choice of a batch, and then a random choice of a customer from that batch (but not a random choice among all waiting customers because this will favor large batches).
- Random choice of a batch, and then serving all the customers of that batch in FIFO, LIFO, or random order.
- Contention schemes: each batch chooses a candidate customer for next service (independent of its service time) and the contention

between the candidates is resolved independent of the sizes of their respective parent batches.

The last family includes disciplines that are applicable to the case of transmitting messages by subdividing them into packets. Here the candidate packet for transmission in each message is the first packet not yet transmitted. Examples of relevant contention-resolving methods are (1) the round robin (token) scheme, and (2) Carrier Sense Multiple Access (CSMA) scheme.[3-5]

III. DISCRETE-TIME ANALOG OF PASTA

The following result will be needed later. Let X_n, $n = 1, 2, \cdots$ be a sequence of random variables, and let B be a set in the value space of X. Let U_n be the indicator function of the event $X_n \in B$, for all n. Let A_n, $n = 1, 2, \cdots$ be a sequence of i.i.d. Bernoulli random variables defined on the same probability space as the X_n's.

$$\text{Let} \quad V_n = n^{-1} \sum_{i=1}^{n} U_i, \quad Y_n = \sum_{i=1}^{n} U_i A_i, \quad \text{and} \quad Z_n = Y_n \left(\sum_{i=1}^{n} A_n \right)^{-1}.$$

Then the following hold:

Theorem 1: If for every n the set of random variables $X_1, X_2, \cdots, X_n$ is independent of the set $A_n, A_{n+1}, \cdots$ then: $V_n \to V$ w.p. 1 if and only if $Z_n \to V$ w.p. 1, as $n \to \infty$.

Remark: This theorem is a discrete-time analog of the PASTA theorem,[6] which is stated in terms of a continuous time stochastic process $X(t)$, and a Poisson process $A(t)$. The assumption in the theorem is called by Wolff 'Lack of Anticipation Assumption' (LAA). The name PASTA comes from the special case where $A(t)$ is an arrival process to a queue, and $X(t)$ is the number of customers in the system at time t. In that case the theorem states that the long-term proportion of time for which $X(t) = m$ (arbitrary nonnegative integer) is equal to the long-term proportion of arriving customers that find the system in state m.

Proof: Wolff's Lemma 1, Lemma 2, and Theorem 1[6] can be easily stated and verified in the discrete-time case, and thus the conclusion holds.

IV. THE RESULT AND ITS PROOF

Theorem 2: Given a queueing system with batch arrivals where:

1. *The batch sizes are i.i.d. with a geometric distribution, and independent of the arrival times process*
2. *The service times are independent of the arrival times and batch sizes*

3. *The queue discipline is impartial, then the customer delay distribution is equal to the batch delay distribution.*

Remark: We interpret the conclusion of the theorem in a time average sense. Thus, for any $x \geq 0$ we compare the proportions of customers and batches that are delayed x time units or less. We prove that if one of these proportions has a long-term limit, so does the other, and they are equal. In the common case, when the queueing system is ergodic, the conclusion can be also interpreted in terms of delay distributions of individual customers and batches.

Proof: Let $X_1, X_2, \cdots$ be the sequence of delays of customers arranged in the order in which they go into service. Let $x \geq 0$ be fixed, and let $B = [0, x]$. Let $A_n = 1$, if the nth customer in the above order is the last to go to service in its batch, and $A_n = 0$ otherwise. Next, observe that the LAA assumption holds for this setup because for any n, $X_1, \cdots, X_n$ are determined by the arrival process, service times of the first $n - 1$ served customers, and queue discipline, all of which are independent of the batch size of the nth customer. (Note that if the discipline is not blind to service time, the above service times could depend on the batch sizes. Also note that the delays do depend in general on the *number* of present batches, and thus A_n and X_{n+1} would typically be dependent. For instance, if $A_n = 1$, the probability that no customer remains waiting, implying $X_{n+1} = 0$, increases.) The proof is now completed by applying Theorem 1.

V. FURTHER COMMENTS

The assumptions of Theorem 2 are quite weak. The theorem is true for a queue with many servers, even with different service rates, as long as the assignment of customers to servers is again independent of the present batch sizes.

We do not require that a service begins immediately when a server becomes free, even if customers are waiting. There may be a "dead period", as is the case in some contention schemes. However, the length of this dead period has to be independent of the present batch sizes.

If we want to derive a similar theorem for *time in system* (rather than delay), then it will not hold in general in the multiserver case. For example, it will not be true for an infinite number of servers, and nonconstant service time. (Ward Whitt[7] provided this example.)

Although the last customer to enter service in a given batch has the same distribution of time in system as a typical customer, its time in system may be different from the batch's time in system. This is so because, in the multiserver case, when that customer completes service, other customers of its batch may still be in service. If we apply

Theorem 2 to the setup where we order the customers by the time they leave the system and let the X_n's be the corresponding times in system, the application will fail because the LAA assumption will no longer be valid.

VI. CONCLUSION

We have shown that for many queueing systems with batch arrivals having geometrically distributed sizes, the individual customers experience the same delay distribution as the batches themselves. This seems to affect the case of transmitting messages by packets, since in many cases it is possible to obtain packet delays, while from a performance point of view one is more interested in the message delays.

VII. ACKNOWLEDGMENT

I would like to thank Ward Whitt for his useful comments and suggestions.

REFERENCES

1. L. Kleinrock, *Queueing Systems*, Vol. 2, New York: Wiley, 1976, p. 113.
2. R. W. Conway, W. L. Maxwell, and L. W. Miller, *Theory of Scheduling*, New York: Addison-Wesley, 1967.
3. L. Kleinrock and F. A. Tobagi, "Packet Switching in Radio Channels: Part I—Carrier Sense Multiple Access Modes and Their Throughput—Delay Characteristics," IEEE Trans. Commun., *23* (1975), pp. 1400–16.
4. F. A. Tobagi and V. B. Hunt, "Performance Analysis of Carrier Sense Multiple Access with Collision Detection," Proc. LACN Symp., May 1979, pp. 217–44.
5. D. P. Heyman, "An Analysis of the Carrier Sense Multiple Access Protocol," B.S.T.J., *61*, No. 8 (October 1982), pp. 2023–51.
6. R. W. Wolff, "Poisson Arrivals See Time Averages," Op. Res., *30*, No. 2 (1982), pp. 223–31.
7. W. Whitt, private communication.

AUTHOR

Shlomo Halfin, M.Sc., 1958, and Ph.D., 1962 (Mathematics), The Hebrew University of Jerusalem; Bell Laboratories, 1968—. Mr. Halfin's areas of interest are stochastic processes and mathematical programming. He currently works on applications of these methods to communication networks. Member, SIAM, ORSA.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

Combined Source and Channel Coding for Variable-Bit-Rate Speech Transmission

By D. J. GOODMAN* and C.-E. SUNDBERG†

(Manuscript received January 19, 1983)

Motivated by potential applications to mobile radio, we studied variable-bit-rate speech communication through Gaussian-noise and Rayleigh-fading channels. For convenience we used a constant signaling rate of 32 kb/s and adjusted the source-coding and channel-coding rates in response to changing transmission quality. When the channel quality was good enough, we used all 32 kb/s for speech transmission. When the channel quality was lower, we reduced the source rate to 24 or 16 kb/s and introduced channel coding to control distortion due to transmission errors. We concentrated on specific source and channel codes that could be implemented with hardware of modest complexity. The source code was embedded differential pulse code modulation, which is amenable to variable-bit-rate operation and economical to implement. For error control we introduced punctured convolutional codes and a Viterbi decoder with only 16 states. Although the source/channel codec was simple, it offered good performance. Speech quality was at the level of normal telephony when the channel was good; the error-correcting codes extended by up to 13 dB the range of channel signal-to-noise ratios that support adequate quality. Our performance estimates were based on a new analysis of transmission errors in embedded differential pulse code modulation and on computer simulations of speech transmission through fixed and fading channels.

I. INTRODUCTION

1.1 Motivation

Every communication channel has an optimum rate for digital transmission of speech. If the actual rate is below the optimum, speech

* Bell Laboratories. † University of Lund, Sweden.

quality is impaired by unnecessary quantizing distortion. If the rate is higher than optimum, the distortion due to transmission errors is excessive. These observations are embodied in rate-distortion theory, which provides theoretical bounds on performance. Implicitly, they influence practical systems in which design decisions are compromises between goals of quality, bandwidth efficiency, and equipment economy.

When the properties of the communication channel are predictable and invariant, one may explore this compromise and arrive at an acceptable design. However, in many instances (for example, in switched telephony) prior knowledge of the channel is only statistical, and in others (fading microwaves) the channel changes drastically with time. With these channels, performance is characterized statistically by the fraction of users experiencing each possible level of quality or by the fraction of time a single user experiences each quality level.

Both spatial and temporal channel fluctuations are inherent in mobile telephony. Because channel characteristics depend on the position of the mobile unit, all users have, at any time, different channels and, because mobile units are moving, their channels change with time. The conventional approach to this statistical nature of the channel is to set design goals of acceptable quality for a large fraction (say 90 percent) of users or to set "outage" limits, i.e., small fractions of time (say 10 percent) when quality is allowed to fall below a certain specification. In meeting these statistical requirements with a fixed coding and modulation scheme, we have a system in which performance is worse than necessary for most users most of the time. The common deficiency is too much quantizing noise, because 90 percent of the channels are better than the marginal one for which the system is designed. The other 10 percent of the users have too much distortion because of transmission errors.

In this paper we investigate variable-bit-rate transmission, with the information rate adjusted according to the properties of the channel. To facilitate implementation, we maintain a constant signaling rate through the channel and reciprocally adjust the information rate of the source and the amount of channel coding for forward error correction. The result, relative to fixed-rate transmission, is improved grade of service or, for the same grade of service, the ability to serve more users within each geographical area.[1]

We limited the content of this paper to transmission components of a variable-bit-rate system: the source and channel codecs and their performance in two types of channel. Control components, which involve measurements of channel quality and two-way communication between user pairs, require further study.

1.2 Source and channel codes

The most general variable-bit-rate scheme requires control information, an adjustable voice coder/decoder (codec), an adjustable channel codec that provides forward error correction to combat channel impairments, and perhaps an adjustable modem (modulator/demodulator), as indicated in Fig. 1a. In our studies, we have focused on specific source and channel codes that are simple to implement and are especially well suited to variable-bit-rate operation. We do not claim that these codes are optimum, even within a constraint of limited hardware complexity. However, they are strong candidates for practical implementation and this study of their performance provides valuable insights into general principles of variable-bit-rate transmission.

The source codes are embedded Differential Pulse Code Modulation (DPCM),[2] which can be implemented with fixed analog-to-digital and digital-to-analog converters and simple digital control of the number of bits transmitted. The channel codes are convolutional codes, with rates 1/2, 2/3, and 3/4, that provide significant coding gains with maximum likelihood decoding. At rates 2/3 and 3/4, "punctured" code[3] realizations simplify the decoder and lend themselves to variable-bit-

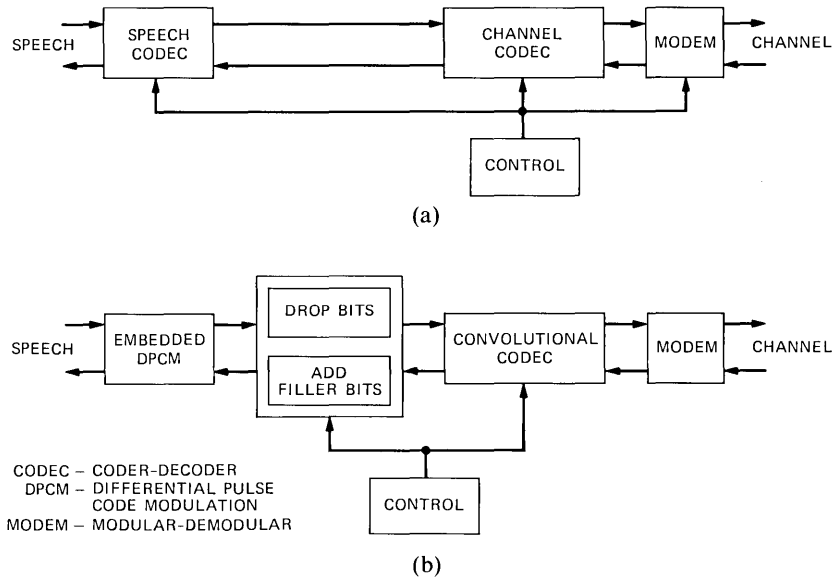


Fig. 1—Elements of a variable-bit-rate system. (a) Adjustable source codec, channel codec, and modem. (b) Fixed embedded source codec and modem. The control of the source rate is simple.

rate operation because the encoder and decoder structures are the same as for the rate 1/2 code. In fact, the punctured codes are rate 1/2 codes with a fraction (1 out of 4 or 2 out of 6) of the channel bits deleted.

In the combined source/channel codec the channel signaling rate is constant; when the channel deteriorates, the least significant speech bits are deleted and the most significant speech bits are protected by a convolutional code. Thus the variable-rate encoding and decoding can be performed, as indicated in Fig. 1b, with a fixed speech codec, a fixed modem, and modest additional hardware relative to fixed-rate operation. Furthermore, the performance characteristics of the embedded source code and the punctured channel code are virtually as good as their conventional counterparts, which require complicated adjustments in a variable-bit-rate environment.

To explore the principle of combining embedded DPCM and punctured convolutional codes we have studied channels with signaling rates of 32 kb/s. We consider four speech transmission formats:

1. All 32 kb/s used for speech transmission
2. 24 kb/s speech transmission, all speech bits protected by a rate 3/4 code
3. 24 kb/s speech transmission, the most significant 2 bits of every 3-bit code word protected by a rate 1/2 code
4. 16 kb/s speech transmission, all speech bits protected by a rate 1/2 code.

These transmission formats, listed in order of increasing resistance to channel impairments, are summarized in Table I, which also presents properties of the convolutional codes discussed in Section III.

Table I—Source and channel code formats, convolutional code properties

	Format 1	Format 2	Format 3	Format 4
1. Source code				
a) bits/sample	4	3	3	2
b) bits/s	32k	24k	24k	16k
c) bits/sample protected	0	3	2	2
2. Channel code				
a) rate	no code	3/4	2/3	1/2
b) $U_0 - U_4$		10101	11001	11101
c) $D_0 - D_4$	encoder, Fig. 4	11111	11011	01011
d) switching pattern		UDDD	UDU	UD
e) free distance, d		4	5	7
f) weight w_d		22	25	4
w_{d+1}	error	0	112	12
w_{d+2}	properties	1687	357	20
w_{d+3}		0	1858	72
w_{d+4}		66964	8406	225

1.3 An example of performance

Figure 2 pertains to phase-shift-keying modulation in a Rayleigh-fading channel. The curves show speech quality, expressed as segmental signal-to-noise ratio (s/n) (Section 2.4.2), as a function of channel s/n. This is defined as the ratio of average energy per channel symbol to noise power per Hertz. Format 1 (32 kb/s speech transmission) is the best choice when the channel is very good (s/n > 22 dB), but it is also the most vulnerable to transmission impairments. By contrast Format 4 (16 kb/s speech), with substantially more quantizing distortion, can pass through very poor channels with no added degradation. Adaptive DPCM at 16 kb/s is highly intelligible, though somewhat fuzzy. Assuming that the threshold of "adequate" performance is 0.5 dB less than the s/n of 16 kb/s error-free transmission, we see in Fig. 2 that Format 4 extends the useful range of channel quality by 13 dB relative to Format 1.

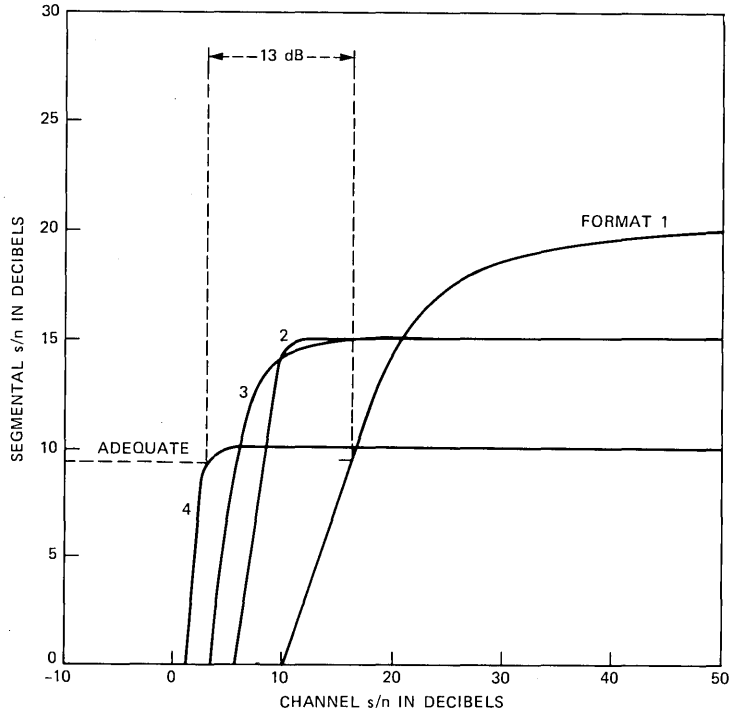


Fig. 2—Performance of the four code formats as a function of the s/n of a Rayleigh-fading channel. If a segmental s/n of 9.5 dB represents "adequate" performance, the variable-rate mechanism extends the threshold of usable channel s/n's from 16 dB (Format 1) to 3 dB (Format 4).

1.4 Organization of this paper

In the following three sections we describe the source and channel codes and present theoretical performance bounds and measures obtained from computer simulations of speech communication over idealized channels. The theory of the effects of transmission errors in embedded DPCM will be published in a separate paper,[4] which extends to embedded DPCM Sundberg and Rydbeck's analysis[5–7] of Pulse Code Modulation (PCM) s/n in noisy channels. Our approach to combined source coding and channel coding is similar in spirit to the work of Modestino and Daut on image transmission.[8–10] Section V lists several issues to be addressed in evaluating applications of these techniques to mobile radio communication.

II. EMBEDDED DIFFERENTIAL PULSE CODE MODULATION

2.1 Encoder and decoder

Within the bit stream of an embedded code is a slower bit stream that represents the analog signal source with reasonable quality. Embedded coding simplifies variable-bit-rate communication. The analog-to-digital and the digital-to-analog converters always operate at the same high rate. If necessary, the transmission system deletes bits from the source code and adds filler bits prior to decoding. Although PCM is an embedded code, differential PCM, which is more efficient for speech transmission, is not. On the other hand, minor modifications of the conventional DPCM encoder and decoder do produce an embedded code. In the embedded codec, the integrators in the encoder and decoder process a low-bit-rate representation of the error signal. Then, when bits are deleted in transmission, the same signal arrives at both integrators and large errors are avoided.

Figure 3 models the embedded DPCM encoder and decoder as the combination of two separate codecs: a "minimal" DPCM codec operating with 2 bits/sample, and a "supplemental" PCM codec, which transmits the quantization error of the minimal encoder. The supplemental codec operates at 2 bits/sample and the combination of minimal and supplemental quantizers can be viewed as a two-stage, successive-approximation realization of a 4-bit/sample quantizer. The 2 bits/sample of the minimal codec are always transmitted, while one or both bits of the supplemental codec can be deleted to reduce the transmission rate from 32 kb/s (4 bits/sample) to 24 or 16 kb/s (3 or 2 bits/sample). The codec functions properly because the encoder and decoder predictors operate on the same signal regardless of the number of bits deleted prior to transmission.

Relative to conventional DPCM, the feedback loop of the embedded DPCM encoder operates with reduced resolution when more than 2

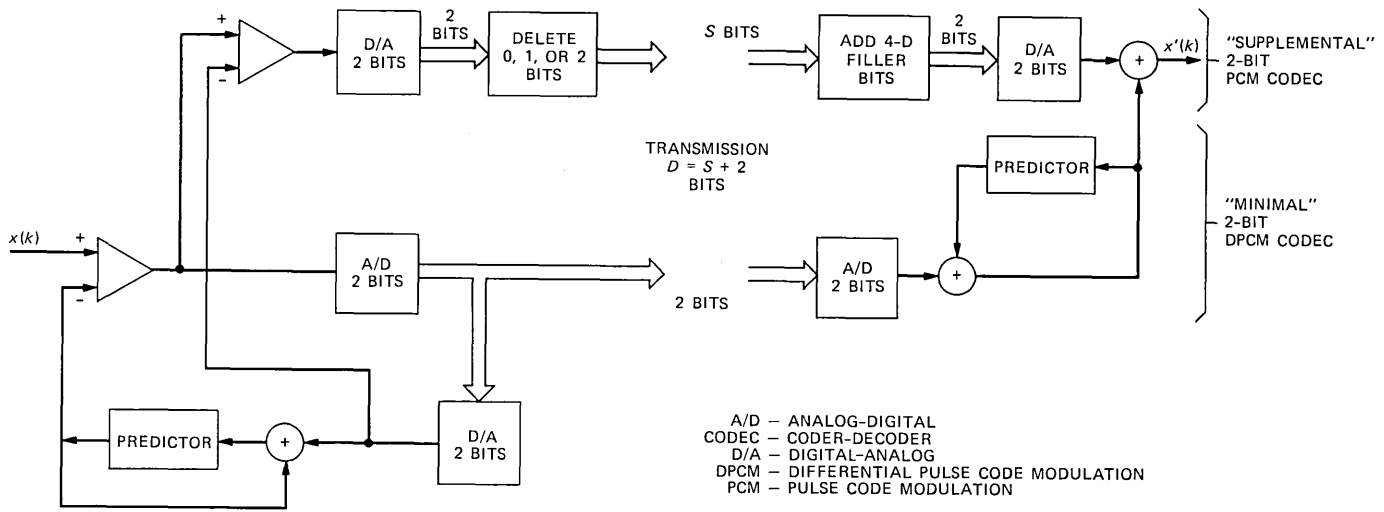


Fig. 3—The embedded DPCM codec modeled as the combination of a "minimal" DPCM codec and a "supplemental" PCM codec. Bits from the supplemental codec can be deleted without severely degrading performance because this deletion has no effect on the ability of the predictor at the decoder to follow the predictor at the encoder.

bits/sample are transmitted. Reference 2 shows that the impact on performance due to this reduced resolution is small, causing no more than a 0.7 dB decrease in s/n in the single-integration codec studied here. (On the other hand, conventional DPCM is unsuited to variable-bit-rate operation. If the transmission system deletes 1 or 2 bits/sample there is a 5.5-dB s/n penalty.)

2.2 Effects of transmission errors

Reference 4 presents a detailed analysis of transmission errors in embedded DPCM. Specialized to a single-integration codec with prediction coefficient a, a key result of that analysis is the formula for the error in the kth sample:

$$x'(k) - x(k) = n_q(k) + e_A(k) + \sum_{i=1}^{\infty} a^i e_M(k - i), \tag{1}$$

in which $x(k)$ and $x'(k)$ are the encoder input and decoder output, respectively. On the right side of (1), $n_q(k)$ is the quantizing distortion of the 2, 3, or 4 transmitted bits/sample, $e_A(k)$ is the noise due to binary errors in all of the transmitted bits, and $e_M(k)$ is the noise due to binary errors in the 2 bits/sample of the minimal quantizer. Equation (1) shows that the errors in these two bits are amplified by the integrator in the decoder.

2.3 Audio signal-to-noise ratio

Our analytic tools[4] enable us to compute the mean-square value of (1) under the condition that the quantizer is in a granular state (output and input differ by no more than half a quantizing step). The derived signal-to-noise ratio, therefore, fails to reflect overload distortion. This apparent deficiency does not limit the value of the granular mean-square error as a predictor of speech quality. Subjective tests show that granular noise and overload are quite different perceptually and that adding their squares is often misleading.

A similar issue is raised by the addition in (1) of a granular-noise term (n_q) to e_A and e_M, which result from transmission errors. Is the mean-square sum a meaningful measure of signal quality? We are optimistic that it is because, contrary to overload, which is signal-dependent, granular noise and noise due to transmission errors are essentially uncorrelated with the input. With this rationale we have derived signal-to-noise ratio formulas of the form[4]

$$\text{s/n} = \frac{E[x(k)]^2}{E[x'(k) - x(k)]^2} = \frac{C}{\sigma_q^2 + \sigma_{qt}^2}, \tag{2}$$

in which C depends on the codec configuration and the source statistics, σ_q^2 is the quantizing noise power, and σ_{qt}^2 is the distortion due to

transmission errors, including the effects of correlation between quantizing noise and noise due to transmission errors.

With single-integration DPCM and a 2-bit minimal quantizer

$$C = \frac{1 - a^2L^2/48}{L^2(1 - 2a\rho + a^2)} \tag{3}$$

and

$$\sigma_q^2 \approx 2^{-2D}/3, \tag{4}$$

where D (2, 3, or 4) is the number of bits transmitted, a is the predictor coefficient of the DPCM feedback loop, ρ is the adjacent-sample autocorrelation coefficient of the encoder input signal, $x(k)$, and L is the quantizer load factor: the ratio of overload point to rms quantizer input.

The other quantity in (2), σ_{qt}^2, depends on the binary representation of quantizer output signals and on the transmission conditions including the modulator and demodulator, the channel, and the encoder and decoder (if they are included) for forward error correction. Table II contains the formulas for σ_{qt}^2 for the four transmission formats studied in this paper. The formulas pertain to a sign-magnitude binary representation and a quantizer with "3.16-sigma loading", i.e., the quantizer overload point is $\sqrt{10}$ times the rms value of the quantizer input.

Table II—Formulas for computing audio signal-to-noise ratio

General formula

$$s/n = \frac{(1 - a^2L^2/48)}{L^2(1 - 2a\rho + a^2)} \cdot \frac{1}{(\sigma_{qt}^2 + 2^{-2D}/3)}$$

Coder parameters

Predictor coefficient $a = 0.85$

Quantizer load factor $L = \sqrt{10}$

Input signal

Correlation coefficient $\rho = 0.85$

Transmission effects, sign-magnitude representation

P: binary error probability of unprotected bits

P_e: binary error probability of convolutional code

Format 1: $D = 4$ (no channel coding)

$$\sigma_{qt}^2 = 0.723P + 0.271P^2 + \frac{a^2}{1 - a^2}(0.725P + 0.275P^2)$$

Format 2: $D = 3$ (rate 3/4 code)

$$\sigma_{qt}^2 = 0.25P_e\left(3.37 + \frac{a^2}{1 - a^2}1.73\right)$$

Format 3: $D = 3$ (rate 2/3 code)

$$\sigma_{qt}^2 = 0.5P_e\left(1.52 + \frac{a^2}{1 - a^2}1.73\right) + 0.0625P$$

Format 4: $D = 2$ (rate 1/2 code)

$$\sigma_{qt}^2 = P_e\left(1.56 + \frac{a^2}{1 - a^2}1.73\right)$$

2.4 Speech transmission

2.4.1 Adaptive DPCM

Practical DPCM codecs have time-varying step sizes to accommodate the wide dynamic range of speech sounds. In this study, we have used a robust adaptive quantizer[11] in which the step size of the encoder changes with each sample according to the rule

$$\delta_{k+1} = M\delta_k^{\beta}, \tag{5}$$

where δ_k is the step size used to encode the kth sample, M is a multiplicative factor that depends on the most recent encoder output, and β is a leakage factor that helps the codec work properly in the presence of transmission errors.

At the decoder the step size is δ_k' with

$$\delta_{k+1}' = M'(\delta_k')^{\beta}, \tag{6}$$

where M' depends on the received code word. It can differ from M if there has been an error in transmitting the kth code word.

To make the adaptive quantizer work with embedded DPCM, we restrict M to one of two possible values:

$$M = M_1 < 1$$

if the input is in the lower half of the quantizer range, and

$$M = M_2 > 1$$

otherwise. Then with four-bit code words represented in a sign-magnitude format, M and M' depend only on the most significant magnitude bit. As a consequence the decoder step size can track the encoder step size when one or both of the two least significant bits are deleted.

In the simulation studies reported in this paper the adaptive quantizer constants were $M_1 = 0.85$, $M_2 = 1.5$, and $\beta = 0.98$. We selected these values because they offer a good compromise between performance over an ideal channel and tolerance of transmission errors.

2.4.2 Segmental signal-to-noise ratio

In addition to the limitations discussed in Section 2.3, the s/n formulas of Table II are of limited value in evaluating adaptive DPCM coding of speech because they do not account for errors due to different step sizes at the transmitter and receiver. On the other hand, segmental s/n is a quality measure that is reasonably well correlated with listener reports of the quality of speech transmitted by adaptive DPCM over noisy channels. It is defined as the average of the s/n measured in decibels over short segments of the signal, and it can be calculated in a straightforward manner in computer simulations of speech trans-

mission.[12] In our simulations we measure s/n over 16-ms segments (128 samples) and exclude from the average segments detected as silent (power at least 55 dB below the saturation level of the quantizer). We also limit the s/n of each segment to the range −10 to 80 dB.

III. CONVOLUTIONAL CODING AND DECODING, PSK MODULATION

Our design, analysis, and implementation of the rate 3/4, 2/3, and 1/2 coders and decoders for the most part follow Chapter 6 of Ref. 13. The punctured codes[3] (rate 1/2 codes with a fraction of the channel symbols deleted) are especially well suited to variable-bit-rate operation because the encoder and decoder retain their basic structures as the code format changes; the things that do change are a set of constants (code generators) and the switching patterns that govern the output of channel symbols at the encoder and the input of channel symbols at the decoder.

3.1 Code configuration

In our calculations and simulations, we use codes generated according to Fig. 4. With a five-stage shift register at the encoder, the codes have 4-bit memory (16 unique nodes in the code trellis). This number provides an attractive compromise between performance (enhanced by long memory) and decoder simplicity (short memory). When the rate 1/2 code is employed, the DPCM signal arrives at 16 kb/s, and for each input bit, the switch at the encoder gets one output bit from the top branch of Fig. 4 and one output bit from the bottom branch. When the rate is 2/3, the two most significant bits in each 3-bit DPCM code word stimulate three output bits from Fig. 4. The first of the two DPCM bits results in two output bits—one from the upper branch of Fig. 4 and one from the lower branch, as in the rate 1/2 code. When the second DPCM bit arrives, the encoder releases one bit from the upper branch. Along with these three bits from the convolutional encoder the least significant DPCM bit is transmitted without protection. When the rate is 3/4, the first member of every block of three input bits results in two output bits (from upper and lower branch of Fig. 4); the remaining two input bits in the block stimulate one output bit each (both from the lower branch). The code generators and switching patterns are listed in Table I.

The receiver is a maximum-likelihood decoder, implementing the Viterbi Algorithm. A signal processor produces a measurement (between 0 and 1) for each received bit. This measurement is near 0 if there is a high likelihood that the bit was transmitted as zero and near 1 if there is a high likelihood that one was transmitted. With each block of two (rate 1/2), three (rate 2/3), or four (rate 3/4) received bits, the Viterbi decoder accepts combined measurements of the first

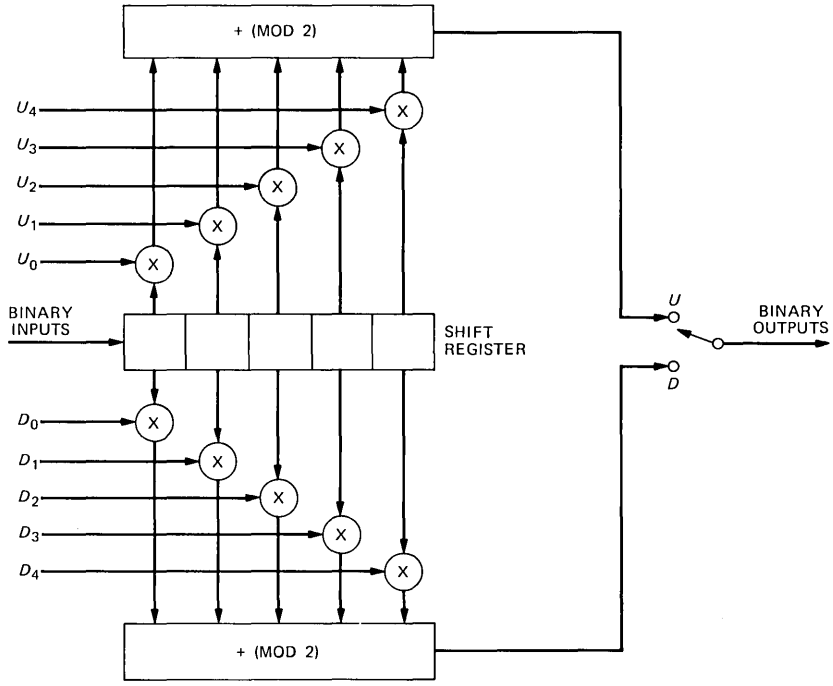


Fig. 4—Structure of the convolutional encoder. U_i and D_i are binary coefficients. For rate 1/2 code there are two output bits (U and D) for each input bit. For rate 2/3 there are two outputs (U and D) for one input and one output (U) for the next input. For rate 3/4 there are two outputs (U and D) for the first input and one output (D) for each of the next two input bits.

two bits and updates the metrics and path memories associated with the 16 possible decoder states. The decoder then releases one bit corresponding to an input bit delayed by the length of the decoder path memory. (In the simulations reported here the path memory was 30 bits for all codes.) If the rate is 2/3 or 3/4, the measurement of each remaining bit in the block causes the decoder to update all metrics and path memories and to release one additional bit.

3.2 PSK modulation, likelihood measurements

To study the performance of variable-bit-rate transmission with embedded DPCM and convolutional codes we have confined our attention to Binary Phase-Shift Keying (BPSK) modulation with coherent detection. In a Gaussian-noise channel the demodulator output for each channel symbol is either

$$r = -A + n \quad \text{or} \quad r = A + n,$$

depending on whether "zero" or "one" was transmitted. The constant,

A, is the received sinewave amplitude and the noise, n, is a sample of a zero-mean normal random variable with variance σ^2. The symbol s/n is $\rho = A^2/2\sigma^2$. In our computer simulations, the likelihood measurement for each symbol is $m = 0.5 + r/\text{const}$, where the constant is large enough to ensure $0 < m < 1$. (With m effectively unquantized, as in our simulations, the constant is otherwise arbitrary. With m quantized to a few bits the constant is a compromise between dynamic range and resolution.)

To update the state metrics of the Viterbi decoder we add m or $1 - m$, depending on whether the relevant branch of the code tree is associated with zero or one transmitted. For a constant-amplitude, white-Gaussian-noise channel, this procedure produces a precise likelihood metric. In a fading channel, which is of principal interest in mobile radio, this metric is only an approximation to the likelihood function.[14] A true likelihood metric involves elaborate computations that would probably be prohibitively complicated in practice.

3.3 Probability of error

The theory of convolutional coding provides upper (union) bounds on bit-error probability as a function of s/n. These bounds are infinite series that are truncated for the purpose of numerical computation. In our calculations we have used sums of five terms to estimate the error probabilities of the convolutional codes:

$$P_e = \frac{1}{b} \sum_{n=d}^{d+4} w_n f_{n,C}(\rho), \tag{7}$$

where d is the free distance of the code; b is the number of source bits per code word (1, 2, 3 for rate 1/2, 2/3, 3/4); and $f_{n,C}(\rho)$ is the probability that an incorrect path, n bits removed from the correct path, has a lower metric than the correct path. This probability is a function of the type of channel, ($C = g$ for a Gaussian channel, $C = R$ for Rayleigh fading), the modulation technique, and the s/n ρ.

Table I contains the free distances, d, and the weights, w_n, of the three convolutional codes we have studied. For BPSK with coherent detection we have the precise formula for a nonfading channel,

$$f_{n,g}(\rho) = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2n\rho}}^{\infty} \exp\left[-\frac{x^2}{2}\right] dx. \tag{8}$$

For independent Rayleigh fading signals in white Gaussian noise, we do not have a precise formula for $f_{n,R}(\rho)$. Instead we use the upper bound[14]

$$f_{n,R}(\rho) < \left[\exp(\beta/\rho)\, \frac{\sqrt{1 + 2/\beta} - 1}{\sqrt{1 + 2/\beta} + 1}\right]^n, \tag{9}$$

$$\beta = \sqrt{\rho^2 + 1} - 1. \tag{10}$$

Figures 5 and 6 show estimates of binary error rates computed from (7) through (10) and the results of computer simulations of random data transmitted through a convolutional coder, a random channel, and a Viterbi decoder with path memory 30 bits. In Fig. 5 the simulations and calculations for the nonfading channel agree very closely. On the other hand Fig. 6 suggests that the bound in (9) is quite loose over the range of conditions of interest to us here; the estimated channel s/n corresponding to a given error rate is 3 to 4 dB lower in the simulations than in the computed curves. Figures 5 and

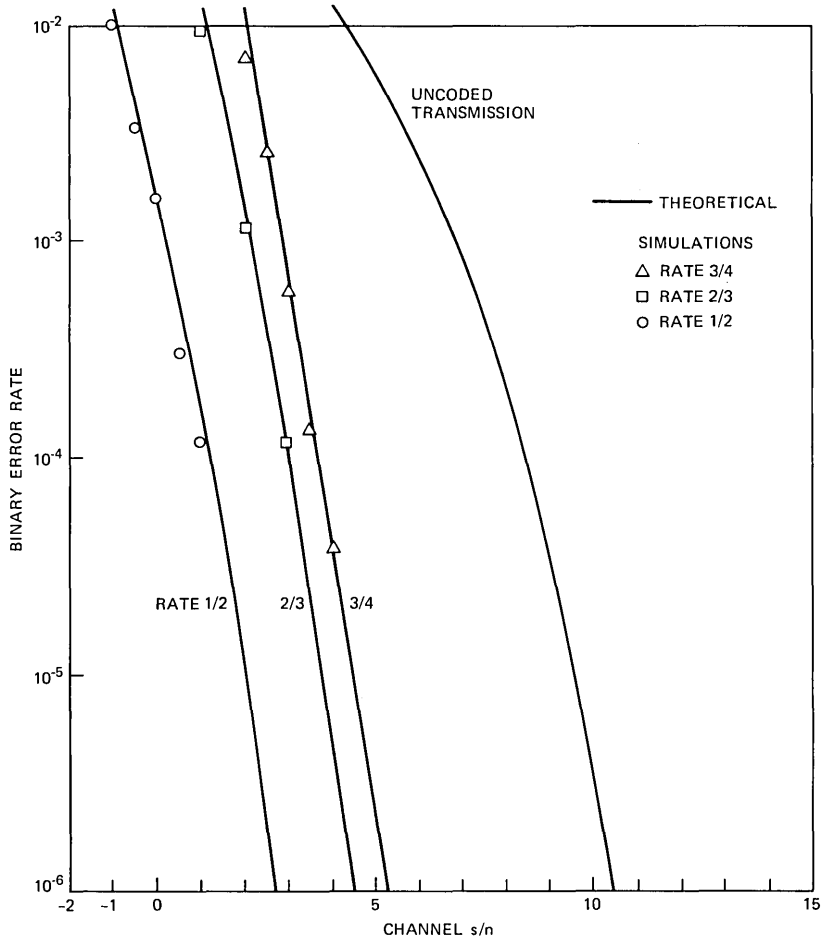


Fig. 5—Binary error rates of convolutional codes and uncoded Phase-Shift Keying (PSK) in a nonfading channel. The theoretical curves and the results of computer simulations are in close agreement.

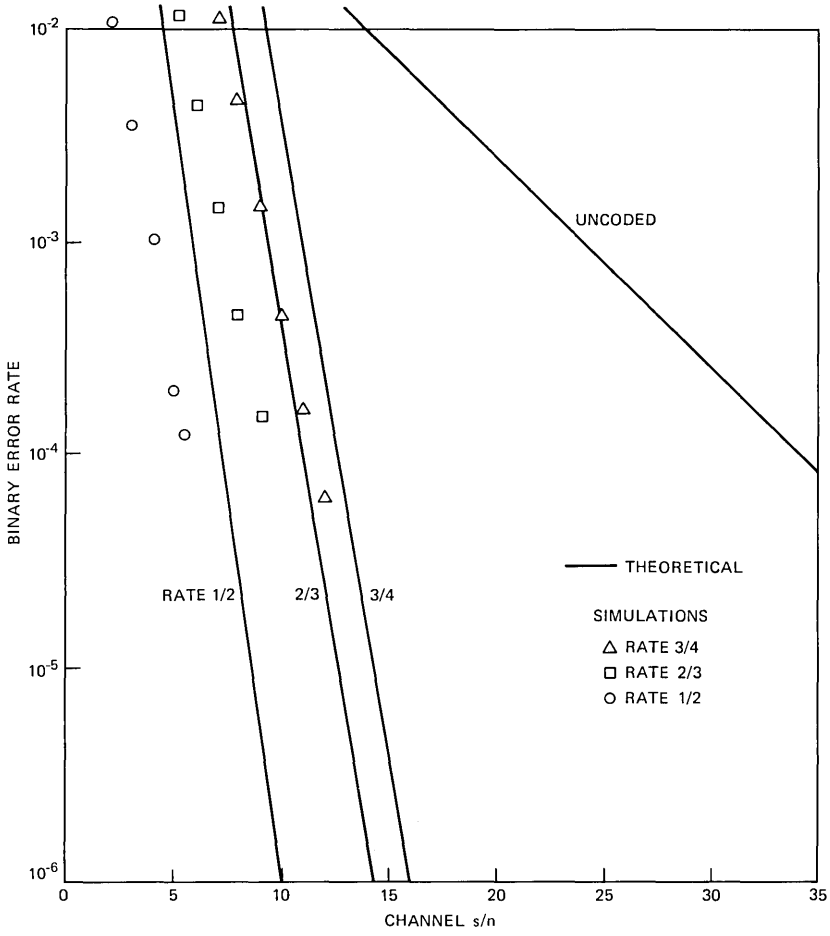


Fig. 6—Binary error rates of convolutional codes and uncoded PSK in a Rayleigh-fading channel. The theoretical curves, based on upper bounds on path error probabilities, are loose over the range of error rates (10^{-4} to 10^{-2}) of interest to us here.

6 also show P, the binary error rates for uncoded transmission, which are $P = f_{1,g}(\rho)$ in (8) for the nonfading channel and

$$P = \frac{1}{2}\left[1 - \sqrt{\frac{\rho}{1+\rho}}\right] \tag{11}$$

with Rayleigh fading.[15]

Note that the independent variable in Figs. 5 and 6 is the s/n of *channel symbols* rather the s/n of *information bits*, which is often plotted in comparisons of coding schemes for digital data transmission.

For our purpose the channel s/n is an appropriate measure because it remains constant as we change code formats.

IV. PERFORMANCE

To evaluate the potential effectiveness of variable-bit-rate transmission we have computed audio s/n as a function of channel s/n using the formulas in Table II and (7) through (10). Figure 7 applies to a nonfading Gaussian channel and Fig. 8 applies to independent Rayleigh fading signals (the s/n's of all channel symbols are mutually independent) in Gaussian noise. All curves pertain to a sign-magnitude representation of source symbols, which is generally more tolerant of transmission errors than the natural-binary representation.

We assume that an audio s/n within 0.5 dB of the s/n of error-free 16 kb/s transmission provides "adequate" voice quality; therefore, in Fig. 7 we estimate that relative to conventional 32 kb/s transmission, variable-bit-rate operation extends the range of useful channel conditions by 4.3 dB in a nonfading channel. (Without channel coding

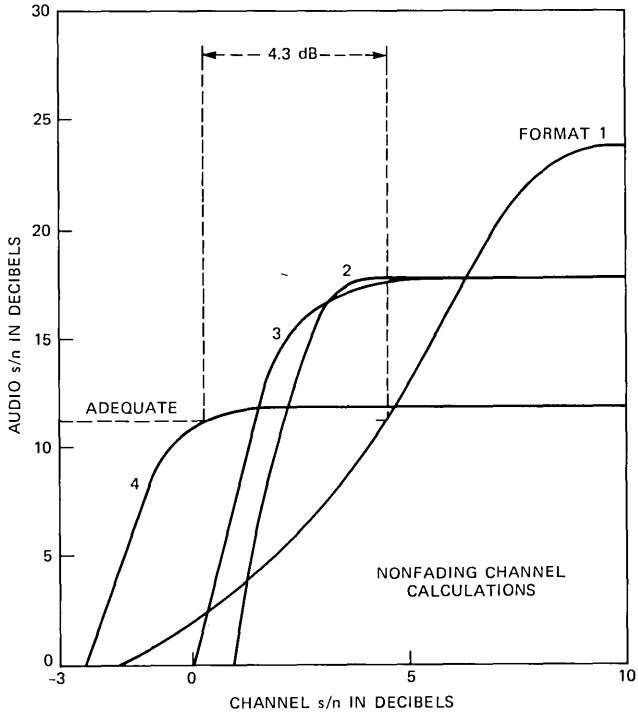


Fig. 7—Calculated performance of the four code formats as a function of the s/n of a nonfading channel. Convolutional coding extends the range of channel s/n's that offer adequate communication by 4.3 dB.

voice quality is inadequate when the channel s/n is less than 4.5 dB. With coding this threshold is 0.2 dB.) Figure 8 suggests that with Rayleigh fading the coding gain is 9.2 dB.

The theoretical formulas, though simple, rely on many bounds, unverified assumptions, and idealizations of a practical communications environment. To obtain more realistic estimates of performance we have resorted to computer simulations of adaptive DPCM transmission of a 2.5-second speech sample. The simulated Viterbi decoder has a path memory of 30 bits. Performance measurements, with speech quality measured as segmental s/n (see Section 2.4.2), are shown in Fig. 9 for a nonfading channel and in Fig. 2 for the Rayleigh-fading channel.

We see that the approximations in Figs. 7 and 8 provide the same qualitative information as the speech simulations in Figs. 9 and 2. For the nonfading channel, simulations and theory lead to nearly equal estimates (4.3 dB and 4.7 dB) of coding gain. In the case of Rayleigh fading the theory underestimates this gain (by 3.8 dB) relative to

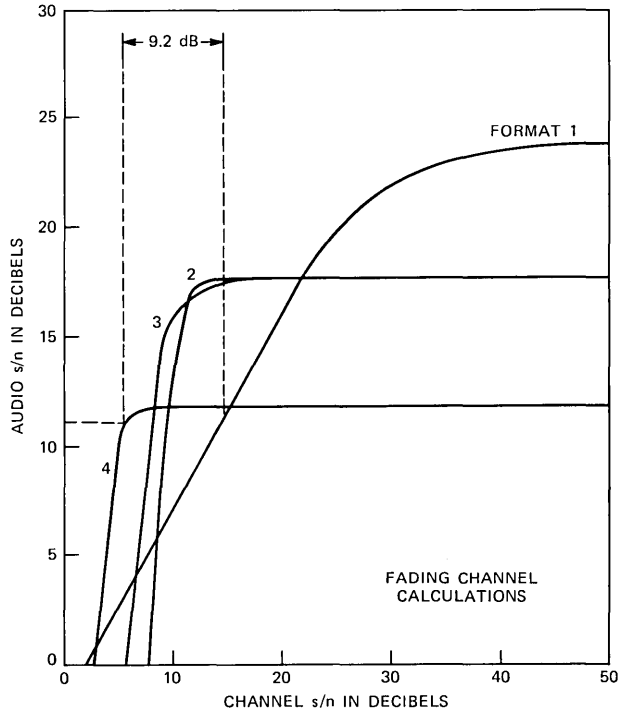


Fig. 8—Calculated performance of the four code formats in a fading channel. The estimate of 9.2-dB coding gain is a lower bound because the computed error probabilities of the Viterbi decoder are loose upper bounds.

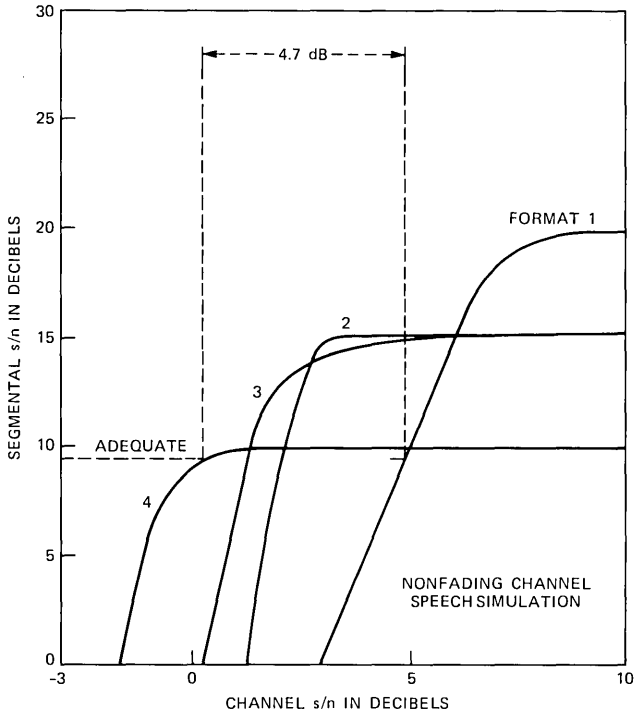


Fig. 9—Performance of the four code formats in simulated speech transmission over a nonfading channel. Coding lowers by 4.7 dB the channel s/n required for adequate quality.

simulation. This discrepancy is largely due to the loose upper bound on the binary error rate of the Viterbi decoder in a fading channel. (See Fig. 6 and Section 3.3.)

All of the performance curves show that transmission Format 2 (24 kb/s speech, all bits protected) is of very limited value. When the coded bits are essentially error free, Format 2 is only slightly better than Format 3 (24 kb/s, 2 of 3 bits/sample protected). In this condition, the transmission noise of Format 3 is due to errors in the third, uncoded, bit, which have a small effect on the quality of embedded DPCM because the errors are not enhanced by the receiver integrator. (In nonembedded DPCM Format 2 would be more effective than Format 3 over a wide range of channel conditions.[4]) In a practical application, therefore, Format 2 would be omitted and the system would switch among Formats 1, 3, and 4.

V. CONCLUSIONS: APPLICATIONS TO MOBILE RADIO

We have described the key elements of a combined source/channel

codec and evaluated its ability to communicate speech over idealized channels. While this approach has other possible applications, we were motivated to study its potential for enhancing mobile radiotelephony. A detailed study of the codec in a mobile-radio environment is currently under investigation. We conclude this paper by reviewing the features of the scheme and listing important issues to be addressed in assessing its value in a mobile-radio context.

A principal advantage of the source codes and channel codes is their simplicity. The embedded DPCM coder and decoder can be realized on a single-chip microcomputer[16] and the Viterbi decoder, with only 16 states, is within the state-of-the-art of special-purpose integrated circuits.[17] Because the channel signaling rate is constant, no special modem is required. Notwithstanding this simplicity, the speech quality for good channels is comparable to that of conventional telephony and the extension of the useful range of channel conditions (4.5 dB in a nonfading channel, 13 dB with independent Rayleigh fading) is substantial.

To extend this idealized study to the context of variable-bit-rate communication over mobile-radio channels, we are led to study the:

1. Control mechanism for altering code formats
2. Channel characteristics, including the temporal nature of the Rayleigh fading and the geographical distribution of signal-to-interference ratio in a mobile-radio service area
3. Alternative modulation techniques to binary-phase-shift keying
4. Possibility of diversity reception.

REFERENCES

1. D. J. Goodman and C.-E. Sundberg, "Quality of Service and Bandwidth Efficiency of Cellular Mobile Radio with Variable-Bit-Rate Speech Transmission," Trans. Veh. Technol., *VT-32*, No. 3 (August 1983).
2. D. J. Goodman, "Embedded DPCM for Variable Bit Rate Transmission," IEEE Trans. Commun., *COM-28*, No. 7 (July 1980), pp. 1040–6.
3. J. B. Cain, G. C. Clark, and J. M. Geist, "Punctured Convolutional Codes of Rate $(n-1)/n$ and Simplified Maximum Likelihood Decoding," IEEE Trans. Inf. Theory, *IT-25* (January 1979), pp. 97–100.
4. D. J. Goodman and C.-E. Sundberg, "Transmission Errors and Forward Error Correction in Embedded Differential Pulse Code Modulation," B.S.T.J., *62*, No. 9, Part 1 (November 1983).
5. N. Rydbeck and C.-E. Sundberg, "Analysis of Digital Errors in Non-Linear PCM Systems," IEEE Trans. Commun., *COM-24*, No. 1 (January 1976), pp. 59–65.
6. C.-E. Sundberg and N. Rydbeck, "Pulse Code Modulation with Error-Correcting Codes for TDMA Satellite Communication Systems," Ericsson Technics, *32*, No. 1 (1976), pp. 1–56.
7. N. Rydbeck and C.-E. Sundberg, "PCM/TDMA Satellite Communication Systems with Error Correcting and Error Detecting Codes," Ericsson Technics, *32*, No. 3 (1976), pp. 195–247.
8. J. W. Modestino and D. G. Daut, "Combined Source and Channel Coding of Images," IEEE Trans. Commun., *COM-27*, No. 11 (November 1979), pp. 1644–59.
9. J. W. Modestino, D. G. Daut, and A. L. Vickers, "Combined Source-Channel Coding of Images Using the Block Cosine Transform," IEEE Trans. Commun., *COM-29*, No. 9 (September 1981), pp. 1261–74.

10. D. G. Daut and J. W. Modestino, "Two-Dimensional DPCM Image Transmission Over Fading Channels," IEEE Trans. Commun., *62*, No. 9, Part 1 (November 1983).
11. D. J. Goodman and R. W. Wilkinson, "A Robust Adaptive Quantizer," IEEE Trans. Commun., *COM-23*, No. 11 (November 1975), pp. 1362–5.
12. C. Scagliola, "Evaluation of Adaptive Speech Coders Under Noisy Channel Conditions," B.S.T.J., *58*, No. 6, Part 2 (July–August 1979), pp. 1369–94.
13. G. C. Clark and J. B. Cain, *Error-Correction Coding for Digital Communications*, New York: Plenum Press, 1981, Chapter 6, pp. 227–66.
14. J. Hagenauer, "Viterbi Decoding of Convolutional Codes for Fading and Burst Channels," Proc. 1980 Int. Zurich Seminar (March 1980) pp. G2.1–7.
15. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw Hill, 1966, p. 407.
16. J. R. Boddie, et al., "Adaptive Differential Pulse-Code-Modulation Coding," B.S.T.J., *60*, No. 7, Part 2 (September 1981), pp. 1547–61.
17. R. M. Orndorff, et al., "Viterbi Decoder VLSI Integrated Circuit for Bit Error Correction," Nat. Telecommun. Conf., New Orleans, Louisiana, December 1981, Conference Record, pp. E1.7.1–4.

AUTHORS

David J. Goodman, B.E.E., 1960, Rensselaer Polytechnic Institute; M.E.E., 1962, New York University; Ph.D. (Electrical Engineering), 1967, Imperial College, London; Bell Laboratories, 1967—. Mr. Goodman has studied various aspects of digital communications including analog-to-digital conversion, digital signal processing, subjective assessment of voiceband codecs, and spread spectrum modulation for mobile radio. He is Head, Communications Methods Research Department. In 1974 and 1975 he was a Senior Research Fellow and in 1983 a Visiting Professor at Imperial College, London, England. Member, IEEE.

Carl-Erik W. Sundberg, M.S.E.E., 1966, and Dr. Techn., 1975, Lund Institute of Technology, University of Lund, Sweden; Bell Laboratories, 1981–1982. Mr. Sundberg is an Associate Professor in the Department of Telecommunication Theory, University of Lund, and a consultant in his field. He is Director of the consulting company SUNCOM, Lund. During 1976 he was with the European Space Research and Technology Centre (ESTEC), Noordwijk, The Netherlands, as an ESA Research Fellow. He has been a Consulting Scientist at LM Ericsson and SAAB-SCANIA, Sweden, and at Bell Laboratories. His research interests include source coding, channel coding (especially decoding techniques), digital modulation methods, fault-tolerant systems, digital mobile radio systems, spread spectrum systems, and digital satellite communication systems. He has published a large number of papers in these areas during the last few years. Senior Member, IEEE; member, SER, Sveriges Elektroingenjörers Riksförening.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

Alternative Cell Configurations for Digital Mobile Radio Systems

By C.-E. SUNDBERG*

(Manuscript received January 19, 1982)

This paper introduces a novel class of antenna configurations and applies it to cellular digital mobile radio systems with frequency reuse. Directional antennas are used extensively. Cooperation between more base stations than are in the conventional three-corner directional antenna scheme is required for the alternative cells. The paper studies the relationship between signal-to-cochannel interference and trunking efficiency (availability of channels) and compares conventional systems with the same base-station locations. We concluded that the new antenna configurations can significantly improve trunking efficiency. Time-division retransmission systems with space diversity are considered for some cases. Furthermore, we show how transmitter power weighting (i.e., certain transmitters have higher output power than others) can improve the signal-to-cochannel-interference ratio for the novel cellular systems.

I. INTRODUCTION

Good spectral efficiency in digital mobile radio systems is obtained through frequency reuse. That is, each cell in the cellular system is assigned a number of channels in a frequency-division system, and each channel (frequency) is reused at a cell further away. The closer this second cell is, the higher the system capacity. On the other hand, cochannel interference increases when the interfering cells are too close. Omnidirectional and directional antenna arrangements have

* Bell Laboratories. Currently at the University of Lund, Sweden.

been considered in cochannel interference reduction. The concept of centrally located base stations with omnidirectional antennas is one possibility. Another suggested scheme is cooperating base stations with 120-degree directional antennas in three corners in a hexagonal cell. In this paper we generalize these antenna configuration ideas. Cooperation between more base stations and extensive use of directional antennas reduce cochannel interference. For example, all channels allocated to three cells in a conventional three-corner scheme can be made available in one "supercell" with the area three times the initial cell. The local availability of channels is now significantly improved. This leads to improved trunking efficiency.

As we will see, working with cells can sometimes be confusing. However, all comparisons below will be made with the same number of base stations per unit area. As a matter of fact, the base-station locations are always the same.

Choice of modulation scheme also affects spectral efficiency. In this paper we will, however, only deal with the frequency-reuse issue for a given modulation scheme, e.g., Quadriphase Shift Keying (QPSK).

This paper will discuss two main ideas. The first is the design of supercells with cooperation between a large number of directional base-station transmitters. The second idea is base-station transmitter power weighting. Since cochannel interference is affected differently by different transmitters, it can be reduced by choosing proper power levels for the base-station transmitters in cellular systems composed by supercells.

The rest of this paper is organized as follows. Section II contains background material on conventional cellular systems, time-division retransmission, and propagation and interference models used in the analysis. Section III contains the new antenna configurations, and Section 3.1 the calculations of signal-to-cochannel-interference ratios for cellular systems with these configurations. Section 3.2 presents the transmitter power weighting analysis. Section IV contains a discussion and conclusions.

II. BACKGROUND MATERIAL

Before we discuss cellular systems in any detail, we will give some brief background information about conventional cellular arrangement, the time-division retransmission method, and signal propagation and interference models for the fading land mobile radio channel.

2.1 Conventional cellular arrangement

We will discuss frequency reuse in cellular systems for digital mobile radio systems. Figure 1 shows an example of a cellular system with three hexagonal cells ($N = 3$) per cluster. Frequencies are assigned as

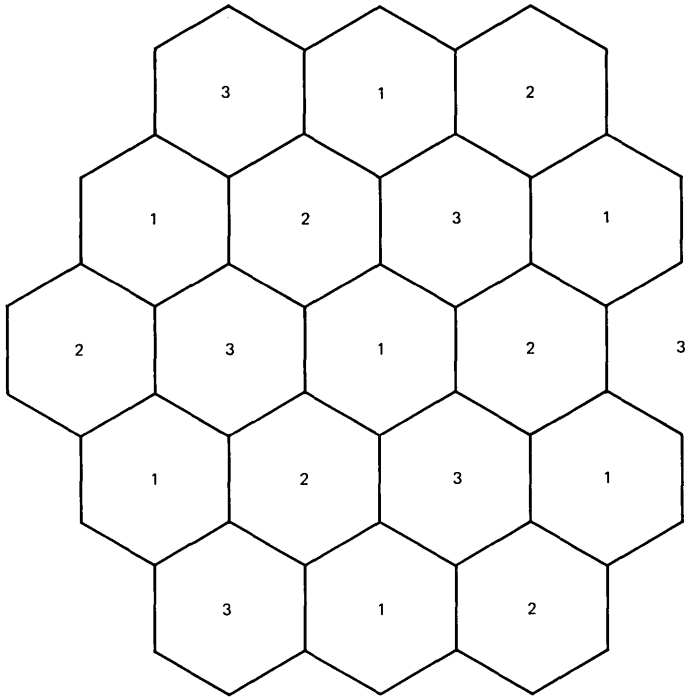


Fig. 1—Example of a cellular system. Each cell is a hexagon of equal area. The frequencies are reused. Cells marked with the same digit use the same frequency channel set. In a cluster of $N = 3$ cells, all channels are used.

in Fig. 2. The channel bandwidth is denoted B_c and the total system bandwidth is denoted B_T. Note that for a fixed total number of channels, a fixed antenna configuration, and a fixed cell size, a large frequency reuse factor, N, gives good cochannel interference (the interfering cells get further away) but a lower number of channels per base station (a low number of channels available at any particular location). This number is inversely proportional to N.

The antennas in each cell in Fig. 1 might be omnidirectional and then located in the center of the cell or 120-degree directional located in each of three corners.[1,2]

2.2 Time-division retransmission

The time-division retransmission (TDR) concept is described in Refs. 3 and 4. The basic ideas are the following. The fading channel changes "slowly" (during one burst or package). Communication in both directions takes place in packages transmitted in the same frequency band: first mobile-to-base and then base-to-mobile. During mobile-to-base transmission, the channel is estimated for maximal-

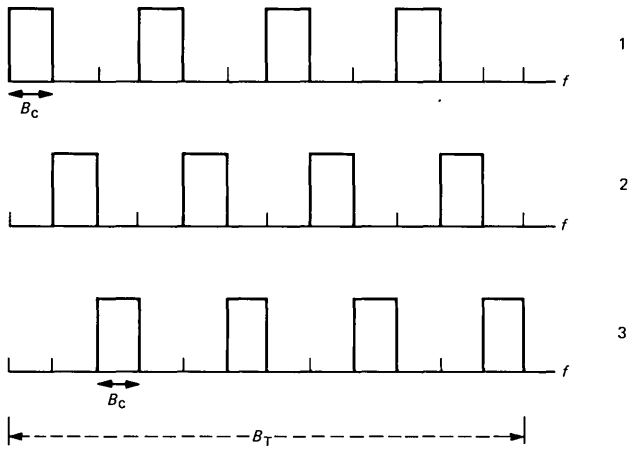


Fig. 2—Frequency plan for the cellular system with $N = 3$, shown in Fig. 6. The channel set marked 1 is used in each cell marked 1, etc. By channel we will consistently mean a two-way channel, including overhead for synchronization (when required).

ratio, space-diversity combining in both directions. Several schemes[3] have been proposed for performing the required co-phasing, for transmission from and reception at the base station. Because all diversity combining takes place at the base stations, the mobile equipment is relatively simple. As a consequence, it is feasible to use more than two branches of diversity with TDR.

We mention the concept of TDR because some of the cellular systems might require diversity because of short-frequency reuse distance and thus high cochannel interference.

2.3 Signal propagation and interference

It is assumed that mobile radio reception in an urban environment is characterized by

$$P(\bar{r}) = |\bar{r}|^{-\alpha} S(\bar{r}) \cdot R^2(\bar{r}), \tag{1}$$

where $P(\bar{r})$ is the received signal power at location $\bar{r}$ (position vector relative to a transmitter).[1,3,4] The first factor $|\bar{r}|^{-\alpha}$ is a reduction factor due to the distance between the mobile unit and the transmitter and α is the propagation constant. It is normally assumed that α is in the range three to four in the urban environment.[1,3] In free space, $\alpha = 2$.

The second factor, $S(\bar{r})$, represents shadow fading,[1,3,4] and the third factor, $R^2(\bar{r})$, represents Rayleigh fading.[4,5] R is the envelope of the received signal. It is modeled as a random variable with the density function

$$p(R) = 2Re^{-R^2}, \tag{2}$$

with $E\{R^2\} = 1$ (see Refs. 4 and 5). In general, R varies with vehicle location and signal frequency.

We will consider propagation and interference in cellular systems with frequency reuse, and use the same basic assumptions as in Refs. 1 and 3.

It is assumed that the cochannel interference is formed by the incoherent sum of contributions from many interfering sources. This sum is assumed to be equivalent to stationary Gaussian noise.[1,3] It is assumed that the shadow and Rayleigh fading of the total interference is negligible compared to the fading of the signal.[1,3,4]

It is also assumed that cochannel interference is the main source of additive signal degradation. The additive thermal background noise is assumed to be negliglible compared to the cochannel interference. Thus, the transmitter power and the cell sizes are assumed to be such that alien background noise from sources other than mobile units and transmitters in the cellular system is negligble compared to cochannel and adjacent-channel interference.

Below we will use the same technique as in Refs. 1 and 3 for calculating cochannel interference. The signal-to-interference ratio is defined as the ratio of the signal power to the total interference power, based on the $|\bar{r}|^{-\alpha}$ propagation law and averaged over shadow and Rayleigh fading. It is assumed that the fading is flat over the band of each channel.

III. NOVEL ANTENNA CONFIGURATIONS

In this section we will present some new methods of organizing antennas covering the hexagons in a cellular system. Analysis of signal-to-cochannel interference for some of these schemes is presented in Section 3.1. Adjacent channel interference is calculated in Ref. 6. These calculations are carried out based on the propagation assumptions made in the introduction and in Refs. 1 and 3.

Figure 3 illustrates the key idea in this paper. Assume base stations in three of the corners in Fig. 1. Form "supercells" by grouping the three cells with C/3 different channels in each into a new "cell" and let the total number of channels, i.e., C, be available throughout the cell. A new cellular system is formed, where the building block now is a cell with a centerbase station and six 120-degree corner stations.

Consequently, the cochannel interference properties of the cell in Fig. 3 are merely fair. By simply rotating the 120-degree corner antennas 30 degrees, one obtains a new larger hexagon (see Fig. 4). The radius of this cell is $r\sqrt{3}$, where r is the radius of the basic hexagonal cell in the three-corner system. The large hexagon is shown solid in Fig. 4.

It is clear from Fig. 4 that the area of the large hexagon ("super

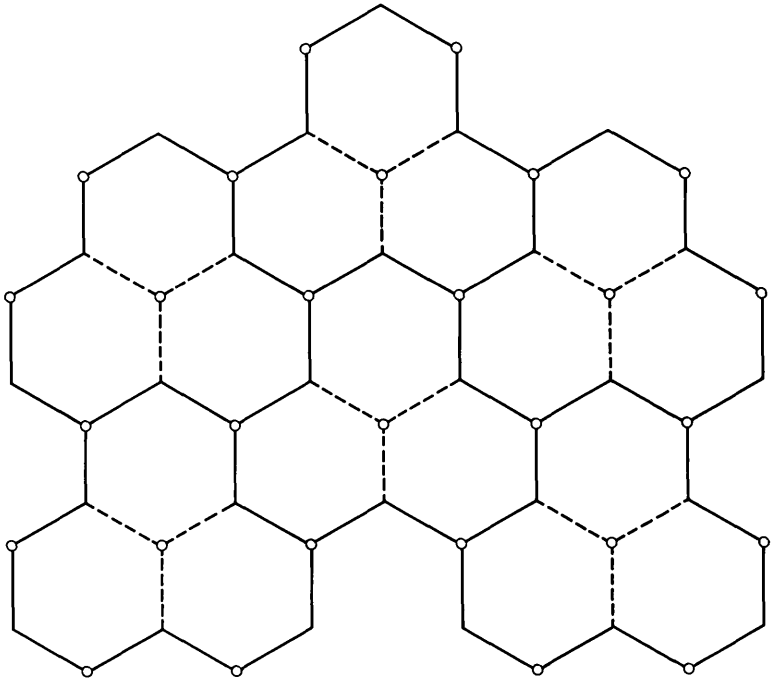


Fig. 3—A cellular system where "supercells" are formed by merging three conventional cells.

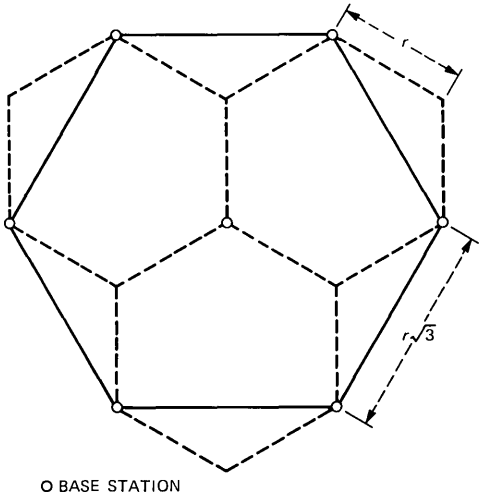


Fig. 4—Comparison between three conventional hexagonal cells (dashed) and a "superhexagon" (solid).

hexagon") is the same as that of three basic hexagonal cells. It is also clear that the cellular system based on the superhexagon has its base stations at exactly the same locations as the basic system with which we started.

Below, we will consider systems based on the superhexagon (and variations thereof). Above, we said that C/3 channels from each of the three basic cells were combined into C channels available throughout the supercell. This will mean not only increased channel availability but also somewhat increased cochannel interference, as we will see. The supercell will also be used with C/3 different channels available throughout the cell. This will mean decreased channel availability and decreased cochannel interference.

Figure 5 shows a number of ways to organize the antennas to cover a cell (hexagon). All the cells in Fig. 5 are shown with equal base-station locations. The shortest distance between base stations is $r\sqrt{3}$, where r is the radius of the basic cell a. It is assumed that the cellular system consists of a number of cells with equal antenna arrangements in each cell. Different cellular systems can be constructed based on the cells shown in Fig. 5.

By a cell we will mean the basic building block in a cellular system, like those in Fig. 5. Sometimes we will refer to a cell as a supercell. For these cases, this cell can be thought of as a combination of a group of smaller cells. Throughout the paper, comparisons will only be made between cellular systems where the base stations have identical locations.

Other cell types (triangular, square, etc.) are, of course, also conceivable (see e.g., Ref. 7). Groups of such basic cells can also be combined into "supercells," much the same way as we have done here for hexagons. The discussion in this paper will, however, be confined to the cells in Fig. 5.

Cases a, b, and c in Fig. 5 are considered in Refs. 1 and 3. In Fig. 5a the antenna is centrally located in each cell and is omnidirectional. In Fig. 5b the antenna is also centrally located in each cell, but the base station now consists of three 120-degree antennas. One of them serves a particular mobile unit at a particular moment. Because of the directivity of the antenna, interference is reduced during mobile-to-base transmission. Further improvements are obtained by placing 120-degree antennas in three cell corners (see Fig. 5c). This arrangement improves signal-to-interference ratios in both directions. The reason is that the distance from a base station to the desired mobile unit is improved relative to the distance to interference sources [see eq. (1) above]. The maximum distance from the desired mobile unit to the closest base station is the same in cells a, b, and c. In cell c it is assumed that the corner base station with the best signal-to-interfer-

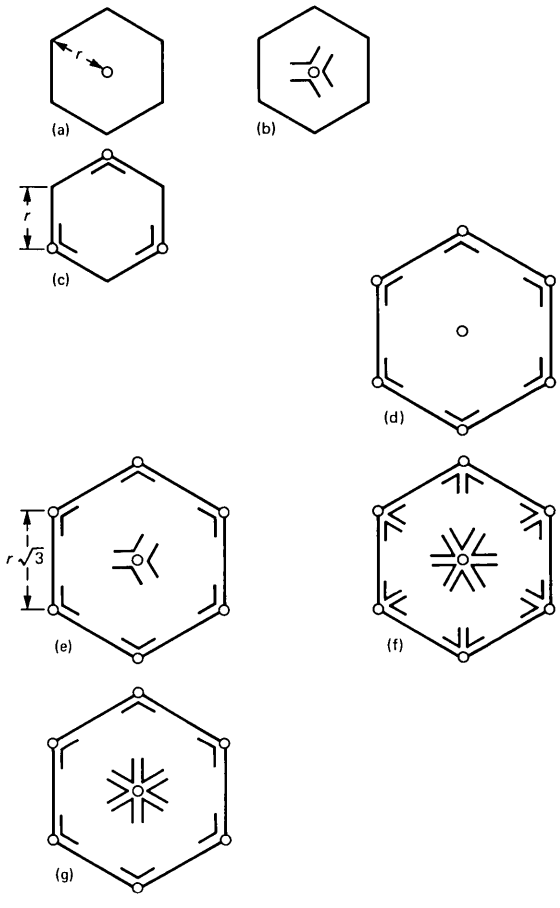


Fig. 5—Different ways of arranging antennas in a cell. Cell a has an omnidirectional antenna in the center of the cell. Cell b has 120-degree antennas in the center of the cell. Cell c has 120-degree antennas in three corners. Cells d, e, f, and g have antennas in all corners and in the center of the cell. Various combinations of 120-degree antennas, 60-degree antennas, and omnidirectional antennas are considered. The cells above will be referred to as cell a to cell g in the text. The cell radius is denoted r for the cells a to c. The radius is $r\sqrt{3}$ for cells d to g. The base-station locations are the same for all cells.

ence ratio serves the mobile unit. The three base stations require coordination so that the desired mobile unit is served by the base station with the best signal-to-interference ratio.

Figures 5d to 5g show arrangements with antennas placed in each cell corner and in the center of the cell. Figure 5e (cell type e) is the one we arrived at in the example above in Fig. 4. Figure 5d shows 120-degree antennas in the corners and an omnidirectional antenna in the center. Figure 5c shows 120-degree antennas in all places. Figure 5f shows the corresponding case with 60-degree antennas. Finally, Fig.

5g shows a hybrid between e and f. Below, we will concentrate on case e and f. Analysis will also be presented for some cases with cell d. For simplicity, the cells will be referred to as cells a to g in the following.

It is evident from Fig. 5 that the maximum distance from a desired mobile unit to the closest base station, d_0, is unchanged in cells d, e, f, and g, compared to cells a, b, and c. Figures 6 and 7 show cells e and f in more detail. It is easily seen that the maximum distance to the closest serving base station is $d_0 = r$ for all cells in Fig. 5.

Figure 6 shows the service areas for the different antennas for the cell in Fig. 5e, assuming all base stations are transmitting at equal power level. It is clear that the worst-case locations (in terms of

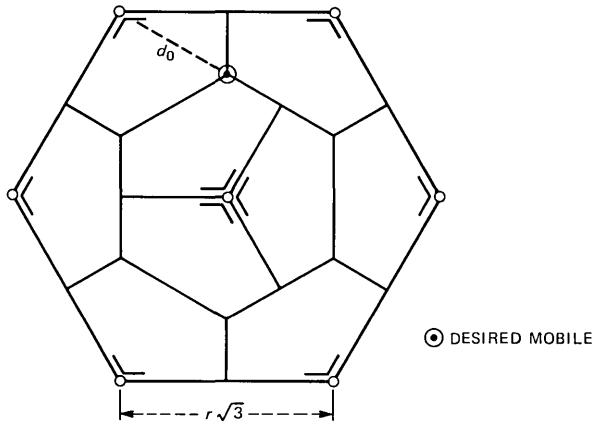


Fig. 6—Coverage areas for the 120-degree antennas in cell e. The example shows the location of a desired mobile unit with maximum distance to a base station.

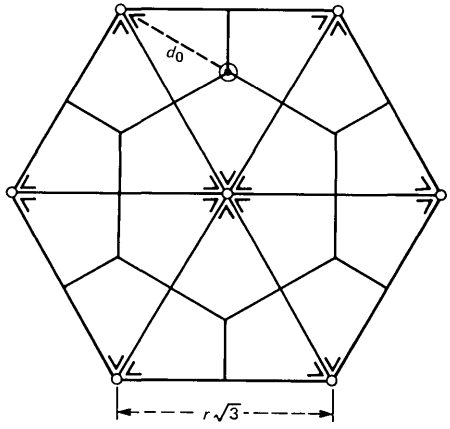


Fig. 7—Same as Fig. 6 for cell f.

maximum distance to the closest base station) for the desired mobile unit is on distance $d_0 = r$ from three base stations. Figure 6 shows that seven different antenna locations have to be coordinated so that a mobile unit within a particular cell is served by the base station with the best signal-to-interference ratio. Typically, only three stations are involved in this comparison for one particular location.

Figure 7 shows the service areas of the 60-degree antennas for the cell in Fig. 5f. Cell f has large similarities to cell e. The signal-to-interference ratio is improved for some cases, though (see tables below). This is, of course, due to the increased directivity of the base stations.

3.1 Cochannel-interference analysis for novel antenna configurations

Tables I to III summarize the results of some of the cochannel-interference calculations based on the assumptions presented above. For background, see Refs. 1 and 3. The cells are defined in Fig. 5. Column 1 also gives the number of base stations per cell (3/3 indicates 3 120-degree stations, 12/6 indicates 12 60-degree stations, etc.). The total number of base stations per unit area and the base-station locations are the same for all schemes (cells). For details of the calculations for cell e, i.e., 120-degree base stations in each of the six corners of the hexagon and in the center (see the appendix). Note that the signal-to-cochannel-interference ratio is not dependent on a particular modulation scheme or on a particular cell size, r. It is purely given by the relationships of distances from the desired base station to the desired mobile unit, and from the interfering base stations or mobile units.

It is assumed in the calculations that the background noise from other sources is negligible. The calculated signal-to-interference ratio determines which modulation method can be used (in terms of required detection efficiency) and how many branches of diversity (M) are required.[1,8]

Table I gives the results for various cell types for a cluster of one cell ($N = 1$), i.e., all frequency channels are used in each cell. The closest (and largest) cochannel interference comes from adjacent cells. The propagation exponent is assumed to be $\alpha = 3$. Some results for $N = 1$, $\alpha = 4$ are summarized in Table II.

Table I gives both average and worst-case signal-to-interference ratios for both mobile-to-base transmission and base-to-mobile transmission. For all cases, it is assumed that the desired mobile unit is in the least favorable location, with respect to the base stations in the cell where it is served. Worst-case mobile-to-base interference means that all mobile units in the interfering cells are in such positions that their contribution to the total interference is maximum. Average

Table I—Signal-to-cochannel interference ratio, $N = 1$, $\alpha = 3$

Base-Station Configuration		Mobile-to-Base						Base-to-Mobile			
		Worst Case		Average				Worst Case		Average	
Cell	No./Type of Stations per Cell	Center	Corner	Center,	M	Corner,	M		M		M
a	1	−10		−6.9	24			−4.8	20	−4.8	20
c	3/3		−4.3			−1.8	12	−1.8	12	>−1.8	12
d	6/3+1	−2.8	2.9	0.3	8	5.4	4	4.4	4	>4.4	
e	6/3+3/3	2.0	2.9	5.1	4	5.4	4	4.4	4	10.6	3
f	12/6+6/6	5.0	5.9	8.1	3	8.4	3	4.4	4	>10.6	

* (Cell type, see Fig. 5.)

Table II—Signal-to-cochannel interference ratio, $N = 1$, $\alpha = 4$

Base-Station Configuration		Mobile-to-Base						Base-to-Mobile			
		Worst Case		Average				Worst Case		Average	
Cell	No./Type of Stations per Cell	Center	Corner	Center,	M	Corner,	M		M		M
a	1	−10.5						−3.8		−3.8	
c	3/3		−3.6					−0.8		>−0.8	
d	6/3+1	−1.5	6.1	4.2				8.3			
e	6/3+3/3			9.0	3			8.3	3	>8.3	
f	12/6+6/6			12.0	2			8.3	3	>8.3	

* (Cell type, see Fig. 5.)

mobile-to-base interference means that the desired mobile unit still is in its worst possible location, but the interfering mobile units are at equally probable locations within their respective cells. The average is formed as described in Ref. 1.

The base-to-mobile signal-to-interference ratio for the worst case occurs when the desired mobile unit is in its least favorable position with respect to any serving base station in the cell, and when the interfering base stations are those which are as close as possible to the desired mobile unit. With one omnidirectional centrally located base station, there is only one case—worst case and average case are the same. When several 120-degree antennas serve the mobile units from various directions, as in cell e, for example, the worst case is a very pessimistic assumption (see the appendix for details).

Tables I to III also contain columns for the number of diversity branches (M) required with space diversity and ideal maximal-ratio

Table III—Signal-to-cochannel interference ratio, $N = 3$, $\alpha = 3$

Base-Station Configuration		Mobile-to-Base						Base-to-Mobile			
		Worst Case		Average				Worst Case		Average	
Cell	No./Type of Stations per Cell	Center	Corner	Center,	M	Corner,	M		M		M
a	1	0.6		2.8							
c	3/3					7.5	3	8.0	3		
d	6/3+1			10.0		14.7		12.6			
e	6/3+3/3			14.8	2	14.7	2	12.6	2	17.9	2
f	12/6+6/6			17.8	2	17.7	2	12.6	2	>17.9	2

* (Cell type, see Fig. 5.)

combining and coherent ideal Binary Phase Shift Keying (BPSK) modulation to achieve the bit error probability of 10^{-3}. This is to give an idea of what the calculated signal-to-interference ratios mean in a cellular system.

For details about the calculations of the signal-to-interference ratios in Tables I to III, see the appendix. Cells a and c have been considered before in Refs. 1 and 3. Some of the signal-to-interference ratios from Tables I to III are from these references. From the signal-to-interference results in Table I, we note that the cells a and c give extremely low values. Many branches of space diversity are required. With cells e and f, more "reasonable" numbers of M are required. Note that cell d is not attractive, since the center base station is very sensitive to interference during mobile-to-base transmission.

Table III gives the signal-to-interference results for clusters consisting of $N = 3$ cells and for a propagation constant of $\alpha = 3$. The results for cells a and c are from Refs. 1 and 3. Note the significant improvements by using cells e and f. For this case, two branches of diversity ($M = 2$) are sufficient for several modulation schemes.

Table III contains results for $N = 3$, $\alpha = 3$. The corresponding results for cell c for $\alpha = 4$ are given in Refs. 1 and 3. For cells d, e, and f, these $\alpha = 4$ results will, of course, be better than the corresponding $\alpha = 3$ results in Table III. They can easily be obtained using the same technique as for the $N = 1$, $\alpha = 4$ case and for the $N = 3$, $\alpha = 3$ case.

Figure 8 shows a detailed comparison of cells e and c. The comparison in Fig. 8 is as before at equal d_0 and at equal base-station locations.

The base-station transmitter power is also equal for the two schemes in Fig. 8. Thus, when comparing the schemes in Fig. 8, we observe the following:

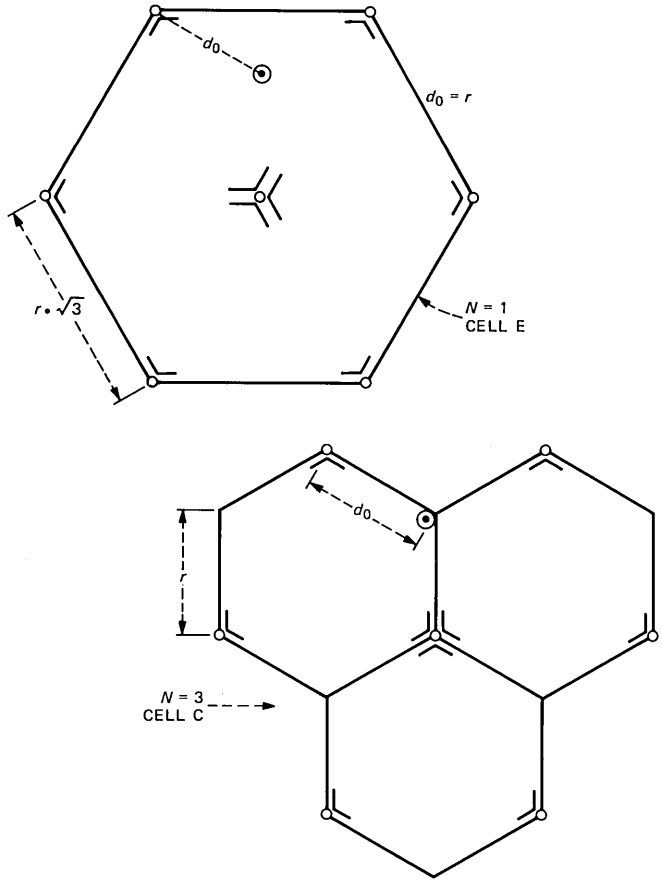


Fig. 8—Comparisons between $N = 3$, cell c radius r and $N = 1$, cell e, radius $r\sqrt{3}$.

1. The number of base stations per unit area is *the same.*
2. The locations of the base stations are the same.
3. The average cochannel interference is comparable (see Table I).
4. The total number of channels served per cell ($N = 1$, cell e or f) and per cluster of cells ($N = 3$, cell c) is the same.
5. The number of channels available at any point of location is three times higher in cells e and f than in cell c. This leads to improved trunking efficiency (see Section 3.3).
6. The number of channels per 120-degree transmitter is three times larger in cells e and f than in cell c.
7. The adjacent-channel-interference problem is worse in cells e and f than in cell c.

It should also be pointed out that, at the worst-case location for

mobile-to-base transmission, there are two base stations within the same distance from the mobile unit for cell c, and sometimes three for cell e (see Fig. 8).

Note that the comparison made in Fig. 8 for $N = 1$, cells e and f and $N = 3$, cell c, can be extended to other N's with similar conclusions. For example, the case with $N = 3$, cells e, and f should be compared to $N = 9$, cell c for the same number of base stations per unit area and with an increase of a factor of three of the number of locally available channels at any particular point of location (see Fig. 9). The adjacent-channel-interference problem is now less serious.

We can also compare two systems based on different cells with the same number of cells (N) per cluster. As an example, consider $N = 3$, cell c and e (or f) with the same number of base stations per unit area and the same base-station locations. Table III gives cochannel-interference results. For this situation we can conclude that:

1. The number of base stations per unit area is the same (same locations).
2. The average cochannel interference is better with cells e and f than with c.
3. The total number of channels served per cell is the same. The total number of channels per three cells of type c is three times that of the number of channels available in one cell of type e, f.
4. The total number of channels available at any specific location is the same in the two cases.
5. The number of channels per 120-degree transmitter is the same.
6. The adjacent-channel-interference problems are worse for cell c than e, f.

3.2 Transmitter power weighting

All the transmitters in systems based on cells a, b, and c (see Fig. 5) have the same properties. Except for the edges in the whole cellular system, the interference situation is the same around each transmitter. All transmitters above are assumed to be transmitting at the same power levels.

The situation is different for cells of type e, f. The center transmitters clearly have a different environment than the corner transmitters. It is not immediately clear why these transmitter power levels should be the same, as we assumed in the analysis above. On the contrary, there are possible improvements in base-to-mobile cochannel interference through reduction of the transmitter power levels for the center stations compared to the corner stations. The consequences of this are that the worst-case location for base-to-mobile and mobile-to-base transmission cochannel interference might be different and that the

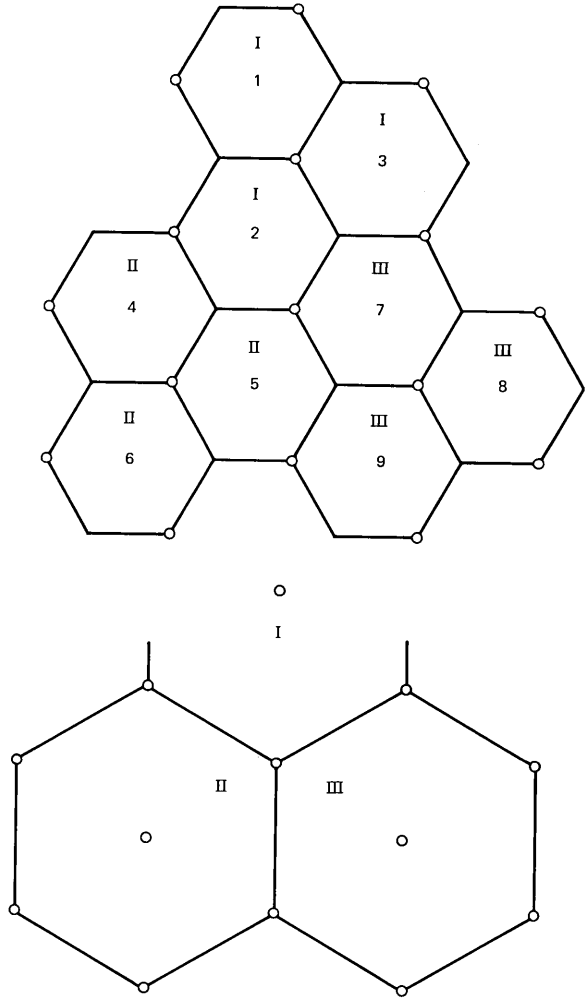


Fig. 9—Comparison between a conventional cellular system with $N = 9$ cells in a cluster and a system based on cell e with $N = 3$ cells in a cluster.

base-to-mobile interference can be improved, for both worst case and average. Mobile-to-base transmission is not affected.

We will now give two examples of what the weighting should be for improving the worst-case components among all the contributions to cochannel interference for the base-to-mobile transmission.

Figure 10 shows the $N = 1$ case with cell e. Assume for simplicity that the radius is one. Assume that the mobile unit is in position M. Furthermore, assume that the worst-case cochannel-interference components will occur somewhere along a straight line from the center

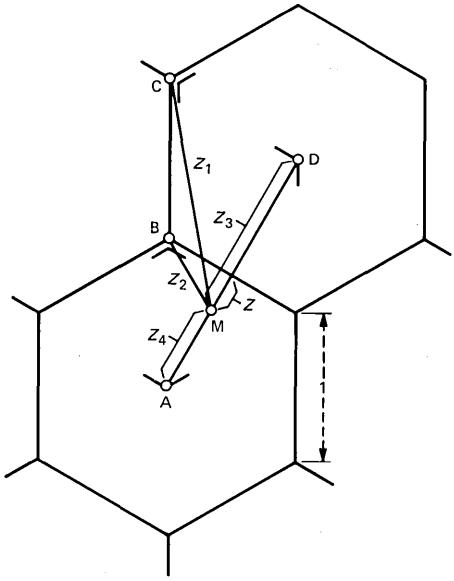


Fig. 10—Notations for analysis of worst-case cochannel interference for cell e, $N = 1$ with transmitter-power weighting.

stations A and D (due to symmetry). The two closest interfering stations are D and C and the two closest base stations serving the mobile unit M are A and B.

Assume that the transmitter power of the corner transmitters is P. The center station has the transmitter power βP, where β is the weighting factor. The dominating components in the cochannel-interference ratio are:

1. For a mobile unit served from corner station B, the signal-to-cochannel-interference ratio is

$$\left(\frac{Z_1}{Z_2}\right)^{\alpha} = \left[\frac{1 + \left(\frac{\sqrt{3}}{2} + Z\right)^2}{\frac{1}{4} + Z^2}\right]^{\alpha/2}, \tag{3}$$

when the interferer is station C, and is

$$\frac{1}{\beta}\left(\frac{Z_3}{Z_2}\right)^{\alpha} = \frac{1}{\beta}\left[\frac{\left(\frac{\sqrt{3}}{2} + Z\right)^2}{\frac{1}{4} + Z^2}\right]^{\alpha/2}, \tag{4}$$

when the interferer is station D. Z is the distance from the mobile unit to the cell boundary (see Fig. 10).

2. For a mobile unit served from the center station A, the signal-to-cochannel-interference ratio is

$$\beta\left(\frac{Z_1}{Z_4}\right)^{\alpha} = \beta\left[\frac{1+\left(\frac{\sqrt{3}}{2}+Z\right)^2}{\left(\frac{\sqrt{3}}{2}-Z\right)^2}\right]^{\alpha/2}, \tag{5}$$

when the interferer is station C, and is

$$\left(\frac{Z_3}{Z_4}\right)^{\alpha} = \left[\frac{\left(\frac{\sqrt{3}}{2}+Z\right)^2}{\left(\frac{\sqrt{3}}{2}-Z\right)^2}\right]^{\alpha/2}, \tag{6}$$

when the interferer is station D.

What is the best choice for β? Select β so that the worst-case location (M) has as large a signal-to-interference ratio as possible. This optimization depends on the propagation exponent, α, i.e., β is a function of α. We found that the best weighting parameter is $\beta = 0.23$, with $Z = 0.416$ for $\alpha = 4$.

This should be compared with the nonweighted case where $\beta = 1$. Here the worst-case term for the cochannel-interference occurs for $Z = 0$. Thus, the improvement of the worst-case cochannel-interference component with weighting is 6.4 dB for $\alpha = 4$.

It can also be expected that the average base-to-mobile cochannel interference is improved significantly by means of weighting. The worst contributors to the noise are the center base stations, and they are now reduced in power. The optimum β with this criterion might differ from the optimum β derived above.

More terms than the worst one, of course, have to be taken into account in a complete analysis. There seems to be room for signficant improvements by means of weighting, however.

Transmitter-power weighting can also be employed for cell e, Fig. 5, with $N = 3$ cells in a cluster. The definitions of locations and parameters are equivalent to the above (see Fig. 11). We assume that the worst-case mobile location for the base-to-mobile cochannel-interference contribution occurs on the straight line from transmitter location A to E (see Fig. 11).

Using the same technique as in the appendix and above, we have the following worst-case cochannel-interference components:

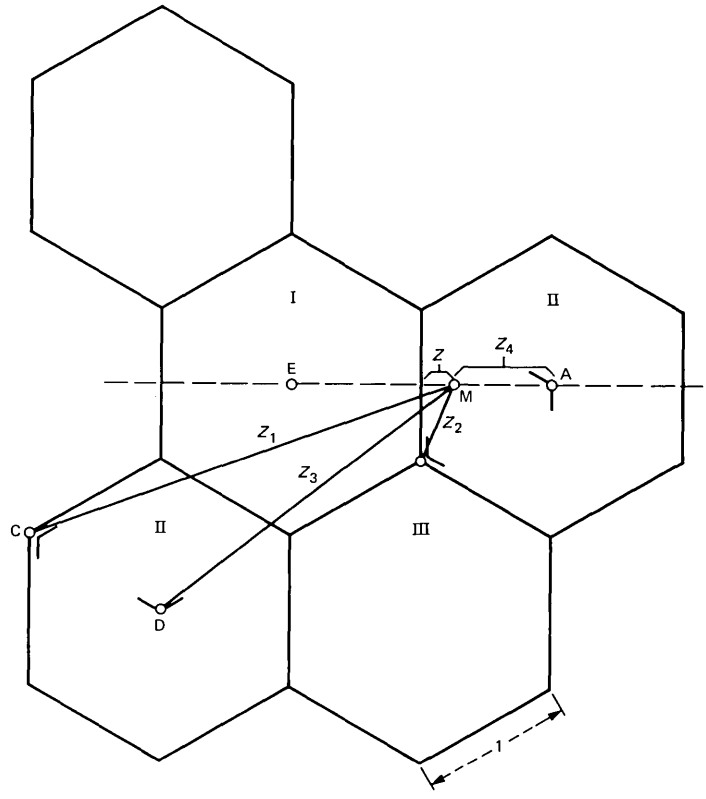


Fig. 11—Same as Fig. 10 with $N = 3$.

1. For a mobile unit served from corner station B, the signal-to-cochannel-interference ratio is

$$\left(\frac{Z_1}{Z_2}\right)^{\alpha} = \left[\frac{1 + \left(\frac{3\sqrt{3}}{2} + Z\right)^2}{\frac{1}{4} + Z^2}\right]^{\alpha/2}, \tag{7}$$

when the interferer is station C, and is

$$\frac{1}{\beta}\left(\frac{Z_3}{Z_2}\right)^{\alpha} = \left[\frac{\left(\frac{3}{2}\right)^2 + (\sqrt{3} + Z)^2}{\frac{1}{4} + Z^2}\right]^{\alpha/2}, \tag{8}$$

when the interferer is station D.

2. For a mobile unit served from the center station A, the signal-to-cochannel-interference ratio is

$$\beta\left(\frac{Z_1}{Z_4}\right)^{\alpha} = \left[\frac{1+\left(\frac{3\sqrt{3}}{2}+Z\right)^2}{\left(\frac{\sqrt{3}}{2}-Z\right)^2}\right]^{\alpha/2}, \tag{9}$$

when the interferer is station C, and is

$$\left(\frac{Z_3}{Z_4}\right)^{\alpha} = \left[\frac{\left(\frac{3}{2}\right)^2+(\sqrt{3}+Z)^2}{\left(\frac{\sqrt{3}}{2}-Z\right)^2}\right]^{\alpha/2}, \tag{10}$$

when the interferer is station D. As in the previous example, the transmitter power of all center stations is βP, where P is the transmitter power of the corner stations. The parameter β is the weighting.

The minimum worst-case cochannel-interference components [eqs. (7) to (10)] occur for the parameters in Table IV. This table shows optimum weighting for minimizing the worst-case contribution to the signal-to-cochannel interference ratio for $N = 3$, cell e.

The gain in the table above is the improvement of the worst-case cochannel-interference contribution with optimum weighting compared with no weighting. For the no-weighting case ($\beta = 1$), the worst case is $Z = 1/2\sqrt{3} \approx 0.289$.

The average cochannel interference will also be improved for the $N = 3$ case with proper weighting. The optimum β has to be found, however. The optimization above was only carried out for the worst-case components.

The extreme case of center base-station power weighting is $\beta = 0$. For this case, there are no center stations at all. This cell is shown in Fig. 12. One third of the base stations can be saved. However, the cochannel interference increases compared to cell e, Fig. 5, particularly mobile-to-base. The mobile transmitter power must also be increased

Table IV—Optimum weighting

α	β	Z	Gain
4	0.276	0.401	5.6 dB
3.7	0.303	0.401	5.2 dB
3	0.381	0.401	4.2 dB

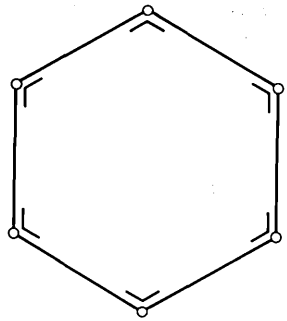


Fig. 12—A cell with one-third fewer base stations than cell e, Fig. 5.

somewhat. The distance to the nearest transmitter is significantly increased for the cell in Fig. 12 compared with cell e.

3.3 Trunking efficiency

We saw above in Section 3.1 that the local availability of channels with cell e is three times that with a cluster of three conventional cells of type c. This three-fold increase in availability of channels leads to improved trunking efficiency with cell e. To illustrate this gain, we give the following example.

Compare a system based on the "superhexagon" cell e with a total of 45 available channels and a system based on the conventional cell c. In the latter case, we assume that 15 channels are available at each station. Thus, in a cluster of 3 cells, type c, there are a total of 45 channels, as with cell e. However, with cell e, the 45 channels are available at all locations throughout the cell e, while with the cluster of 3 cells of type c, 15 channels are available in each cell.

The traffic behavior of Mobile Radio-Telephone Systems might be modeled by Erlang C tables. (The calls are delayed, but eventually placed. The calls are not rerouted over alternate facilities, nor do they go away.)

Assume a blocking probability of two percent using Erlang C tables; the traffic carried by the system in cell e is 32.03 erlangs, while the system with cell c carries 8.03 erlangs in each cell, or 24.1 erlangs in a cluster of three cells. Thus, the system based on cell e carries 7.93 erlangs more per area corresponding to one cell e (3 cells of type c) than the conventional system.

Instead, using Erlang B tables, the traffic carried by the system with cell e is 35.61 erlangs, while with cell c, the traffic in each cell is 9.01 erlangs and in a cluster of three cells about 27.0 erlangs. Thus, the system based on cell e carries 8.6 erlangs more per area corresponding to one cell e (3 cells of type c) than the conventional system.

The best model is perhaps a modified Erlang B—something between Erlang B and C. Some percentage of the calls may never be made or may be rerouted. In any case, the basic conclusion is the same: A significant improvement of the trunking efficiency is achieved with cell e.

3.4 Relationship with dynamic channel assignment schemes

In the example above, it is evident that the trunking efficiency is improved by using cell e instead of a cluster of three cells of type c. The increased local availability of channels should also improve the capability of matching nonuniform geographic traffic patterns.

With conventional cells, this problem is dealt with by using dynamic channel assignment schemes.[4,9] The novel cell types, e.g., cell e, are somewhat like this. The control algorithm is fixed with cell e, however. Further work, especially simulations, is required for evaluating the above relationship.

IV. DISCUSSION AND CONCLUSIONS

Novel antenna configurations for cellular digital radio systems are introduced and analyzed in this paper. Two key ideas are presented. The first is designing new cells by keeping the base-station locations in conventional cellular systems and rearranging the antenna patterns. Basically, merging conventional cells into larger, novel cells improves trunking efficiency. The local availability of channels is increased. The second idea is transmitter-power weighting. The base-station transmitters in the novel cells have different roles depending on the exact location. The center stations and the corner stations contribute differently to the cochannel interference. Thus, the signal-to-cochannel interference ratio can be improved by proper weighting of the transmitter-power levels.

Many unsolved problems remain. The analysis above (as that in Refs. 1 and 3) is based on idealized assumptions about the selection of the serving base station and on very simple and idealized channel and interference models. More refined analysis and simulations will be necessary. The analysis above was carried out under the idealized assumptions of flat fading. Uniform transmission conditions were assumed for all cells. No delay spread was considered. Perfect timing and synchronization was assumed with coherent detection and ideal maximal-ratio combining. It was furthermore assumed that perfect synchronization for the time-division retransmission scheme was established. The analysis was confined to local-mean values of signal and interference at isolated points. Consequently, the effects of shadow fading were not taken into account, and no results were obtained for overall signal-to-interference statistics throughout entire cellular

areas. The effect of channel occupancy (the fraction of time that a channel is in use) on interference was not considered. All of the above problems and others have to be taken into account in a refined system analysis.

Some of the aspects of choosing a modulation scheme for digital cellular mobile radio systems are dealt with in Refs. 6 and 8. Adjacent-channel interference in cellular systems is calculated in Ref. 6 for some bandwidth-efficient constant amplitude modulation schemes.

Further work should be devoted to the traffic-carrying aspects of cellular systems based on the novel cells. Detailed comparisons with dynamic channel assignment schemes with conventional (and novel) cells should be made.

V. ACKNOWLEDGMENTS

Thanks are due to the reviewers, who pointed out the trunking efficiency aspects of the novel cells and the relationship with dynamic channel assignment.

REFERENCES

1. Y. S. Yeh and D. O. Reudink, "Efficient Spectrum Utilization for Mobile Radio Systems Using Space Diversity," IEEE Trans. Commun., *COM-30*, No. 3 (March 1982), pp. 447–55.
2. V. H. MacDonald, "The Cellular Concept," B.S.T.J., *58*, No. 1 (January 1979), pp. 15–41.
3. P. S. Henry and B. S. Glance, "A New Approach to High Capacity Digital Mobile Radio," B.S.T.J., *60*, No. 8 (October 1981), pp. 1891–904.
4. W. C. Jakes, Jr., *Microwave Mobile Communications*, New York: Wiley, 1974.
5. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw-Hill, 1966.
6. C.-E. Sundberg, "On Continuous Phase Modulation in Cellular Digital Mobile Radio Systems," B.S.T.J., this issue.
7. D. C. Cox, "Cochannel Interference Considerations in Frequency Reuse Small-Coverage-Area Radio Systems," IEEE Trans. Commun., *COM-30*, No. 1 (January 1982), pp. 135–42.
8. C.-E. Sundberg, "Error Probability of Partial Response Continuous Phase Modulation with Coherent MSK-type Receiver Diversity and Slow Rayleigh Fading in Gaussian Noise," B.S.T.J., *61*, No. 8 (October 1982), pp. 1933–63.
9. T. J. Kahwa and N. D. Georganas, "A Hybrid Channel Assignment Scheme in Large-Scale, Cellular-Structured Mobile Communication Systems," IEEE Trans. Commun., *COM-26*, No. 4 (April 1978), pp. 432–8.

APPENDIX

Details of Cochannel-Interference Calculations

This appendix contains some details about the calculations of the signal-to-cochannel-interference ratios given in Tables I to III in Section 3.1 above.

Figure 13 shows the worst-case location for the desired mobile unit for the case of transmission from base to mobile with $N = 1$, $\alpha = 3$, and 120-degree base stations in all corners and in the center (cell e).

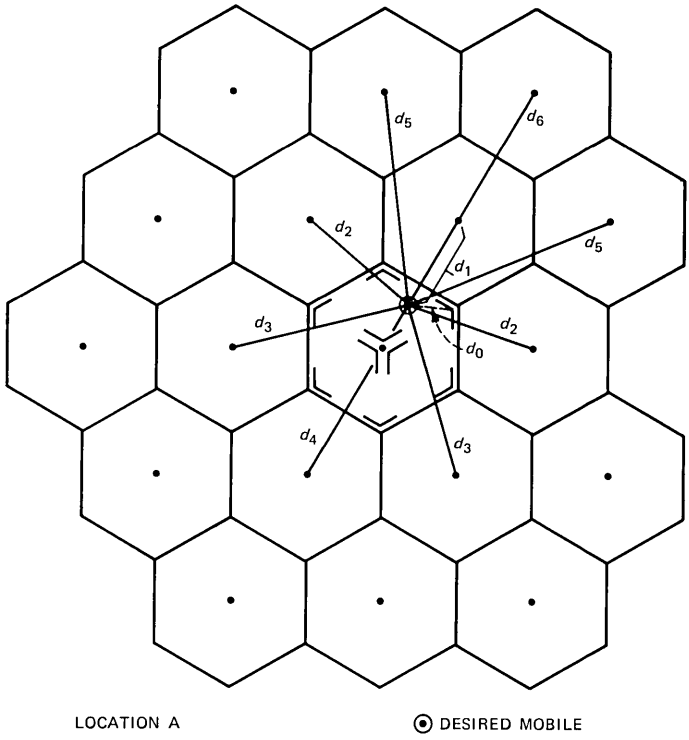


Fig. 13—Worst-case location for the desired mobile unit for worst-case base-to-mobile transmission $N = 1$, $\alpha = 3$, cell e. This location of the desired mobile unit is referred to as location A.

Assume that the base stations in nearby cells are serving mobile units in interfering frequency slots in such a way that the base station in a particular cell is as close as possible to the desired mobile unit, thus contributing maximally to the cochannel interference. Thus, the worst-case interference power to signal power is given by the distance relationships[1,3]

$$\frac{P_I}{P_S} = \left(\frac{d_0}{d_1}\right)^3 + 2\left(\frac{d_0}{d_2}\right)^3 + 2\left(\frac{d_0}{d_3}\right)^3 + 2\left(\frac{d_0}{d_5}\right)^3 + \left(\frac{d_0}{d_4}\right)^3 + \left(\frac{d_0}{d_6}\right)^3 + \text{terms from cells further away.} \quad (11)$$

The worst-case position above is given by the location where the mobile unit is as far away as possible from the nearest base station in the cell. Thus $d_0 = r_0\sqrt{3}$, where r_0 is the cell radius (of cell e) (see Fig. 8).

The dominating term in eq. (11) is

$$\left(\frac{d_0}{d_1}\right)^3 = \left(\frac{r_0/\sqrt{3}}{2r_0/\sqrt{3}}\right)^3 = \left(\frac{1}{2}\right)^3. \tag{12}$$

Compare the dominating term in the $N = 1$ schemes with centrally located base stations or corner stations only in Refs. 1 and 3. In this case, this term is 1, independent of α. For the above case, the cochannel-interference noise suppression improves with the propagation constant α.

Taking into account contributing terms from the 23 interfering cells closest to the cell with the desired mobile unit we have (see Table I)

$$\frac{P_S}{P_I} \approx 4.4 \text{ dB}. \tag{13}$$

This is the signal-to-interference ratio for base-to-mobile transmission with the worst-case location of the desired mobile unit. All the interfering base stations are assumed to be in a worst-case mode of operation too. The propagation exponent is $\alpha = 3$, and the frequency plan is such that all cells use the total number of frequency slots, $N = 1$. We have assumed that each cell has 120-degree base stations in each corner and in the center, cell e.

Figure 14 shows the location of the desired mobile unit in the cell where the worst-case relationship is between the distance to the nearest transmitter in the cell and the distance to the nearest interfering base station in the adjacent cells. In this case, the interference-to-signal ratio is

$$\frac{P_I}{P_S} = \left(\frac{d_0}{d_1}\right)^3 + 2\left(\frac{d_0}{d_2}\right)^3 + 4\left(\frac{d_0}{d_3}\right)^3 + 2\left(\frac{d_0}{d_4}\right)^3 + \text{terms from interfering cells further away.} \tag{14}$$

The dominating term is now given by $d_0 = r_0/2$, $d_1 = r_0\sqrt{3}/2$, thus

$$\left(\frac{d_0}{d_1}\right)^3 = \left(\frac{1}{\sqrt{3}}\right)^3. \tag{15}$$

This term is, of course, larger than eq. (12). However, the other terms in eq. (14) are smaller than their counterparts in eq. (11) because d_0 is smaller in Fig. 14 than in Fig. 13. Thus, it is not immediately evident which location, A or B, is generally the worst-case location for the desired mobile unit.

Carrying out the calculations in eq. (14) with the same number of terms as in eq. (11), we have the signal-to-distortion ratio

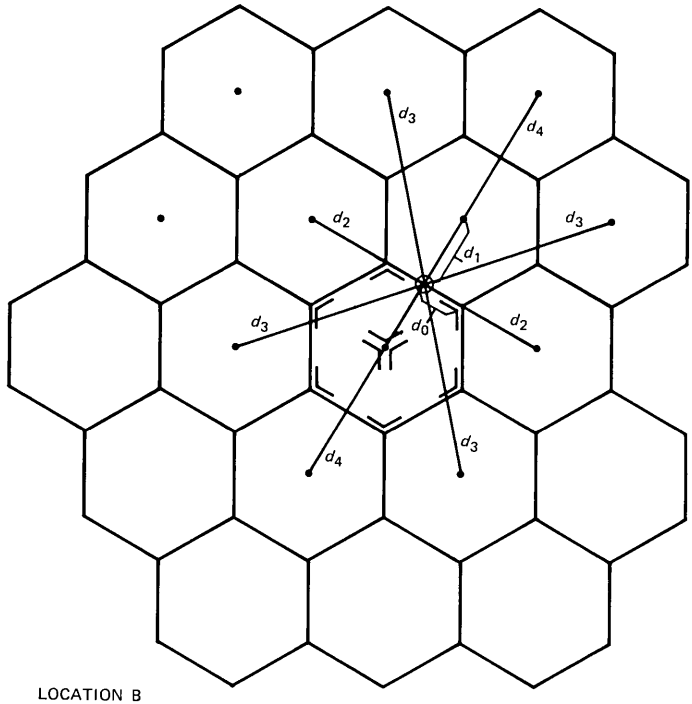


Fig. 14—Location of the desired mobile unit where (d_0/d_1) is maximum. Otherwise, same case as Fig. 13. This location of the desired mobile unit is referred to as location B.

$$\frac{P_S}{P_I} \cong 4.5 \text{ dB} \tag{16}$$

for the worst case in Fig. 14, for $\alpha = 3$. Thus, the location in Fig. 13 is slightly worse for the $N = 1$, $\alpha = 3$ case.

It is even conceivable that some location between that in Figs. 13 and 14 is the worst case for $\alpha = 3$. We do not expect more than a very small (if any) deviation from eq. (13), however.

We have also carried out the calculations of the base-to-mobile worst-case signal-to-cochannel-interference ratio for the cases in Figs. 13 and 14 for the propagation exponent $\alpha = 4$. For this case, the dominating term for location A is

$$\left(\frac{d_0}{d_1}\right)^4 = \left(\frac{1}{2}\right)^4, \tag{17}$$

and for location B,

$$\left(\frac{d_0}{d_1}\right)^4 = \left(\frac{1}{\sqrt{3}}\right)^4. \tag{18}$$

Carrying out the calculation of eqs. (11) and (14) for $\alpha = 4$ for the locations A and B we have, for $\alpha = 4$,

$$\frac{P_S}{P_I} \approx 9.2 \text{ dB location A,}$$

$$\frac{P_S}{P_I} = 8.3 \text{ dB location B.}$$

Thus, in this case ($\alpha = 4$), location B is the worst case. This is what might be expected from the formulas above. For large α's, the first term [in generalizations of eqs. (11) and (14)] will play an increasing role.

Above, we have calculated the signal-to-interference ratio for the worst-case base-to-mobile transmission where we assumed worst-case conditions, i.e., that the desired mobile unit is in the worst location and that the mobile units in neighboring cells are all in such positions that the interference from all cells is maximum. This is, of course, a very pessimistic assumption. Next, we will calculate the average interference power when it is assumed that all positions of a particular mobile unit in a cell are equally probable. Thus, the probability that a mobile unit is served by a particular 120-degree base station is 1/5 in configuration e (see Fig. 5e). Note that it is still assumed that the desired mobile unit is in its worst location. Other positions for the desired mobile unit will yield better signal-to-interference ratios. With the assumptions above, we have the contributions to the average interference-to-signal power ratio for the closest cell,

$$\frac{P_I}{P_S} = \frac{1}{9}\left(\frac{d_0}{d_1}\right)^3 + \text{other smaller terms within the cell}$$

$$+ \text{terms from other interfering cells.} \tag{19}$$

Note that the contribution from each base station is scaled with either 1/9, when it is pointed towards the desired mobile unit, or 0, when it is pointed in a different direction. The worst-case term $(d_0/d_1)^3$ is now scaled down with a factor of 1/9. Continuing the calculation in eq. (19) and including all the cells included in the calculation of eq. (11), we have the approximate average signal-to-cochannel-distortion ratio with the desired mobile unit in the worst location A for the $\alpha = 3$, $N = 1$ case with 120-degree base stations in all corners and in the center of the cell,

$$\frac{P_S}{P_I} \approx 10.6 \text{ dB.} \tag{20}$$

This is significantly better than the worst case [eq. (13)] of 4.4 dB for the worst-case location for the desired mobile unit. The large improvement is, of course, due to the fact that the surrounding interfering cells only use the interfering antennas that are closest to the desired mobile unit part of the time. Sometimes, there is no interference from a particular cell, because the mobile unit served on the same frequency channel in that particular cell is served by an antenna that is not pointed in the direction of the desired mobile unit.

Figure 15 shows the worst-case mobile-to-base cochannel interference for the 120-degree corner base station in a cell with 120-degree base stations in all corners and in the center. The desired mobile unit is assumed to be in its worst position, A, on distance $d_0 = r\sqrt{3}$ from the serving base station (see Fig. 13). Mobile units on interfering frequency channels are all assumed to be in the least favorable position in their cells (see Fig. 15). With straightforward calculations like those earlier in this appendix, we arrive at the signal-to-interference ratio of approximately 2.9 dB for $\alpha = 3$ (see Table I).

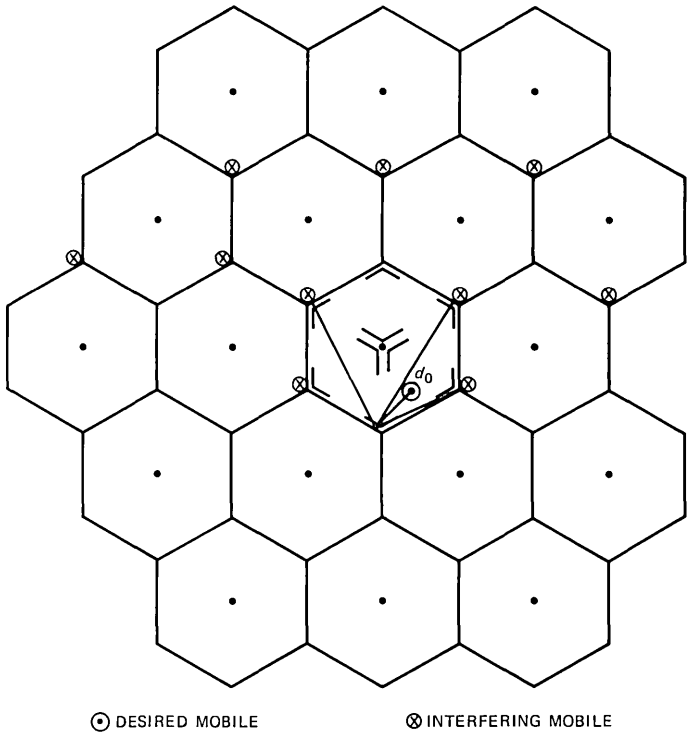


Fig. 15—$N = 1$, $\alpha = 3$, cell e, desired mobile unit in location A. Worst-case mobile-to-base signal-to-interference ratio for a corner station.

The center base station is more sensitive to worst-case mobile-to-base cochannel interference that the corner base stations in Fig. 15. Figure 16 shows this. The worst locations for interfering mobile units are closer to the base station serving the desired mobile unit in this case. The signal-to-interference ratio for this worst case is approximately 2.0 dB.

It is clear from Fig. 16 that an ominidirectional center antenna is very sensitive to mobile-to-base cochannel interference. The lack of directivity forces it to receive interference from all directions, while the desired mobile unit is in a particular direction. Thus, it is advantageous to use 120-degree antennas rather than omnidirectional antennas. Further improvements are obtained by using 60-degree antennas (see Fig. 7 and Table I).

The worst-case base-to-mobile cochannel interference is the same for omnidirectional, 120- or 60-degree center antennas (see Table I).

The assumption that all interfering mobile units are in their worst positions in their respective cells is, of course, a very pessimistic one. It is more realistic to calculate an average interference, where the mobile units are assumed, with equal probability, to be in any position

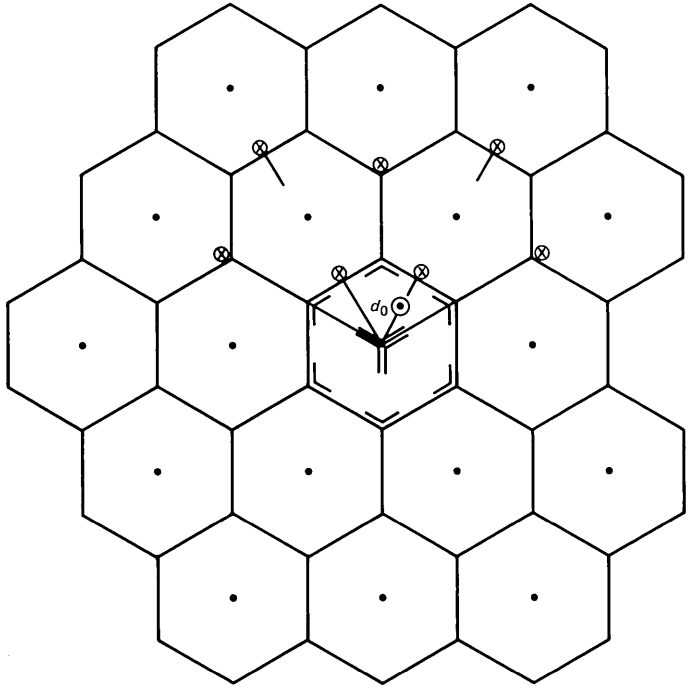


Fig. 16—Same as Fig. 15 for a center station.

within a cell. Using the averaging technique, described in detail in Ref. 1, we obtain the average signal-to-interference ratios for mobile-to-base transmission in Tables I to III in Section 3.1.

By using 60-degree antennas, the interference above is reduced by a factor of two. Since the center base station is the most sensitive one, it is, of course, conceivable with a hybrid scheme with 120-degree antennas in all the corners and 60-degree antennas in the center locations (see Fig. 5g).

The cochannel-interference calculations behind the remaining signal-to-cochannel-interferences ratios in Tables I to III above are carried out with the same techniques as above. The averaging technique for the mobile-to-base cochannel interference in Ref. 1 is easily extended to the propagation exponent $\alpha = 4$. Location A is the worst case for $N = 3$, $\alpha = 3$ with cell e.

AUTHOR

Carl-Erik W. Sundberg, M.S.E.E and Dr. Technology, Lund Institute of Technology, University of Lund, Lund, Sweden, in 1966 and 1975, respectively. Currently, Mr. Sundberg is a Research Professor in the Department of Telecommunication Theory, University of Lund. He is also Director of the consulting company SUNCOM, Lund. During 1976 he was with the European Space Research and Technology Centre (ESTEC), Noordwijk, The Netherlands, as an ESA Research Fellow. He has held positions as Consulting Scientist at LM Ericsson and SAAB-SCANIA, Sweden, and at Bell Telephone Laboratories, Crawford Hill, N. J. His research interests include source coding, channel coding, especially decoding techniques, digital modulation methods, fault-tolerant systems, digital mobile radio systems, spread spectrum systems, and digital satellite communication systems. He has published a large number of papers in these areas during the last couple of years. Dr. Sundberg holds several U.S. and European patents. Senior Member, IEEE; member, SER (Sveriges Elektroingenjörers Riksförening).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

On Continuous Phase Modulation in Cellular Digital Mobile Radio Systems

By C.-E. SUNDBERG*

(Manuscript received January 19, 1982)

This paper considers the application of constant-amplitude, partial-response, continuous-phase modulation with simple near-optimum receivers to cellular digital mobile radio systems. These smoothed modulation schemes have low spectral sidelobes and narrow main lobes. A combiner for time-division retransmission systems with space diversity is given for continuous-phase modulation. Cochannel interference and adjacent-channel interference are calculated for continuous-phase modulation in cellular systems with frequency reuse. The efficiency of 3RC and 4RC modulations with space diversity is evaluated for conventional cellular systems.

I. INTRODUCTION

This paper considers a class of constant-amplitude digital modulation schemes with smoothed partial-response, continuous-phase modulation. This gives a bandwidth-efficient digital modulation scheme. With the baseband quadrature combiner described below, these modulation schemes can also be used with time-division retransmission. The concept of time-division retransmission is briefly described below. This scheme makes it possible to use simple equipment at the mobile unit, and yet use several branches of space diversity, assuring a reasonable error rate on the fading channel. Most of the signal processing and all diversity combining is performed at the base stations.

To obtain good spectral efficiency in digital mobile radio systems,

* Bell Laboratories. Currently at the University of Lund, Sweden.

frequency reuse is employed.[1-4] Each cell in the cellular system is assigned a number of channels in a frequency division system. Each channel (frequency) is reused at a cell further away. The closer this cell is, the larger the number of available channels. On the other hand, the signal-to-cochannel interference increases when the interfering cells are too close. The interference is also affected by the way the base-station antennas are arranged. Omnidirectional and directional antennas have been considered. Furthermore, the location of the base station in the cell is important for the signal-to-interference ratio. In this paper we consider two schemes: Centrally located base stations with omnidirectional antennas, and cooperating base stations with 120-degree directional antennas in three alternate corners in a hexagonal cell. Below, we will see how cochannel and adjacent-channel interference is calculated for continuous-phase modulation in a cellular environment.

The rest of this paper is organized as follows. Section II contains background material on modulation methods, conventional cellular systems, time-division retransmission, and propagation and interference models used in the analysis. Section III presents a baseband maximal-ratio combiner for time-division retransmission with constant-amplitude modulations. Section 4.1 contains the calculations of signal-to-cochannel interference ratios for cellular systems with continuous-phase modulation. Section 4.2 gives some results on adjacent-channel interference. Section V contains a discussion and conclusions.

II. BACKGROUND MATERIAL

Before the cellular systems are discussed in any detail, we will give some brief background information about the class of constant-amplitude modulation systems considered in this paper, conventional cellular arrangement, the time-division retransmission method, and signal propagation and interference models for the fading land mobile radio channel.

2.1 Constant-amplitude modulations

Constant-amplitude modulation schemes are considered for application to cellular digital mobile radio systems. The transmitted signal is of the form

$$s(t) = \sqrt{\frac{2E}{T}} \cos[\omega_c t + \phi(t)], \tag{1}$$

where the information-carrying phase is

$$\phi(t) = 2\pi h \sum_i \alpha_i q(t - iT), \tag{2}$$

with

$$q(t) = \int_0^t g(\tau)d\tau. \tag{3}$$

The pulse shape $g(t)$ (instantaneous frequency) determines the spectral behavior of the modulation scheme, while h is the modulation index, and the α_i's are data symbols. For detailed background on these schemes, see Refs. 5–11. In this paper, we will mainly concern ourselves with schemes having modulation index $h = 1/2$ and with $g(t)$ in the form of raised cosine pulses of length LT, where T is the bit time. This pulse shape is defined by

$$g(t) = \begin{cases} \dfrac{1}{2LT}\left[1 - \cos\left(\dfrac{2\pi t}{T}\right)\right] & 0 \le t \le LT \\ 0 & t \le 0,\ t \ge LT. \end{cases} \tag{4}$$

For $L > 1$, the pulses are overlapping in time, thus creating controlled intersymbol interference. In this paper, only binary transmission $\alpha_i = \pm 1$ will be considered. The reason is that binary schemes with modulation index $h = 1/2$ have simple, near-optimum receivers, i.e., "minimum shift keying (MSK)-type" receivers (see Refs. 7–9).

Figures 1 to 3 show examples of power spectra of (unfiltered)

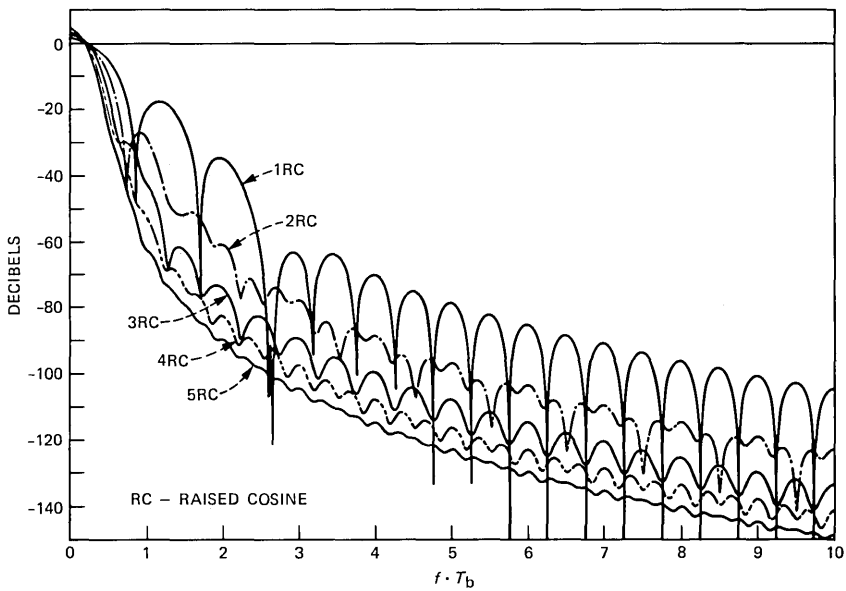


Fig. 1—Power spectra for binary RC schemes when $h = 1/2$, $L = 1$ to 5. For details, see Refs. 5, 6, and 11. T_b is bit time (symbol time). (Spectra are shown one-sided normalized with total power.)

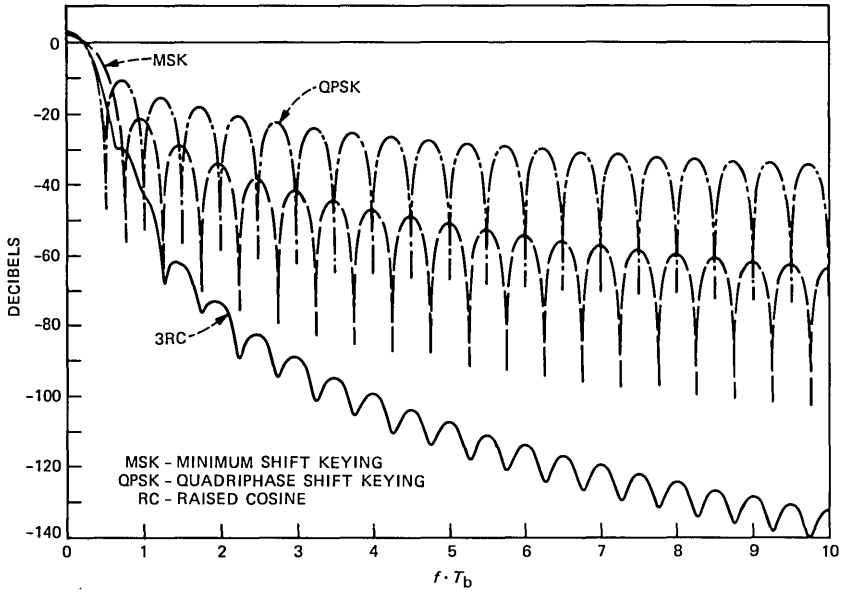


Fig. 2—Power spectra for binary 3RC with $h = 1/2$, QPSK, and MSK.

constant-amplitude modulation schemes in the raised cosine family. A scheme with a pulse shape $g(t)$ of a raised cosine shape with length $3T$ is denoted by 3RC (raised cosine). Minimum shift keying,[9] tamed frequency modulation (TFM),[8] and Gaussian MSK (GMSK)[10] are obtained from eqs. (1) and (2) by selecting proper pulses $g(t)$ (see Refs. 5 to 7 for details).

Even better spectral tail behavior is obtained with spectral raised cosine (SRC) pulse shapes instead of raised cosine pulse shapes.[5,6] These schemes are more difficult to analyze exactly, however. A 3SRC pulse $g(t)$ has a length of $3T$ between the first nulls in time around the peak value. The Fourier Transform is a raised cosine. Thus, the time function is infinitely long. In practice, these pulses are truncated.

Figures 4 and 5 show examples of error probability curves for coherent binary phase shift keying (BPSK), quadriphase shift keying (QPSK), 3RC and 4RC for slow Rayleigh fading, space diversity with M branches, and ideal maximal-ratio combining. These are examples of calculations in Refs. 7 and 12. The analysis for the Gaussian channel is presented in Ref. 7, using analytical tools from Refs. 13 and 14. The formulas for the Gaussian channel are generalized to the fading channel in Ref. 12. Note that the error probability curves in Figs. 4 and 5 are shown without differential encoding/decoding. In practice, differential encoding is employed as the simplest way of resolving phase ambiguity in the coherent MSK-type receiver for $h = 1/2$

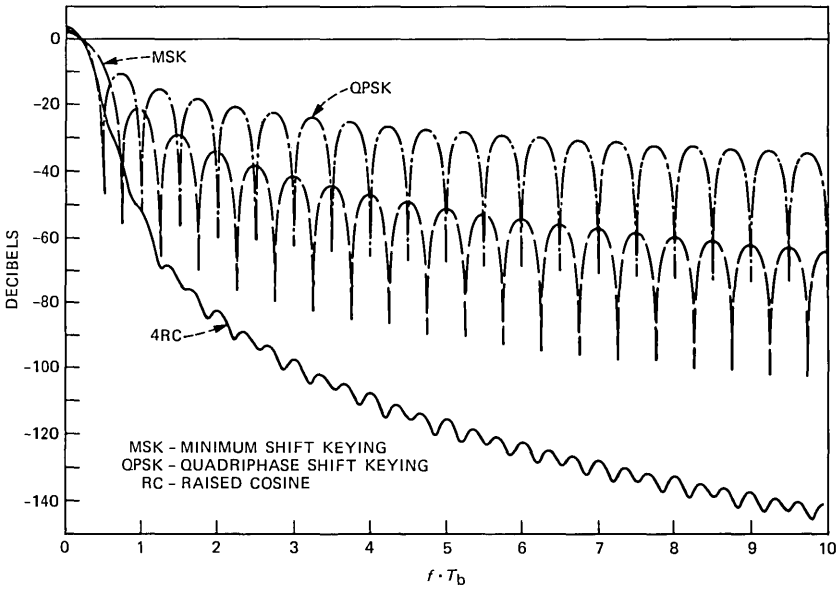


Fig. 3—Power spectra for binary 4RC with $h = 1/2$, QPSK, and MSK.

schemes.[7,8,9,15] For this case, all error probability values in Figs. 4 and 5 are multiplied by a factor of two.

The longer the pulse $g(t)$, the more narrow the power spectrum.[6] However, the detection efficiency is reduced with increasing L for fixed $h = 1/2$, particularly for suboptimum receivers.[7] The optimum receiver (the Viterbi detector) is more robust.[6] Thus, as we will see, choosing the length of the pulse shape $g(t)$ plays an important part in the overall system optimization.

Power amplifiers, both in the mobile unit and at the base stations, sometimes contain nonlinearities. For such cases, it is advantageous to use a constant envelope modulation scheme. Such a scheme will not be subject to widened spectra after the nonlinearity. This is the case for nonconstant amplitude modulations. A QPSK signal passed through a bandpass filter can be quite narrowband. The signal amplitude varies, however, and after a nonlinearity, the sidelobes are restored to a level that is unacceptable from an adjacent-channel interference point of view.

2.2 Conventional cellular arrangement

This paper will consider frequency reuse in cellular systems for digital mobile radio systems. Figure 6a shows an example of a cellular system with $N = 3$ hexagonal cells per cluster. Frequencies are assigned as in Fig. 6b. Note that for a fixed total number of channels, a fixed

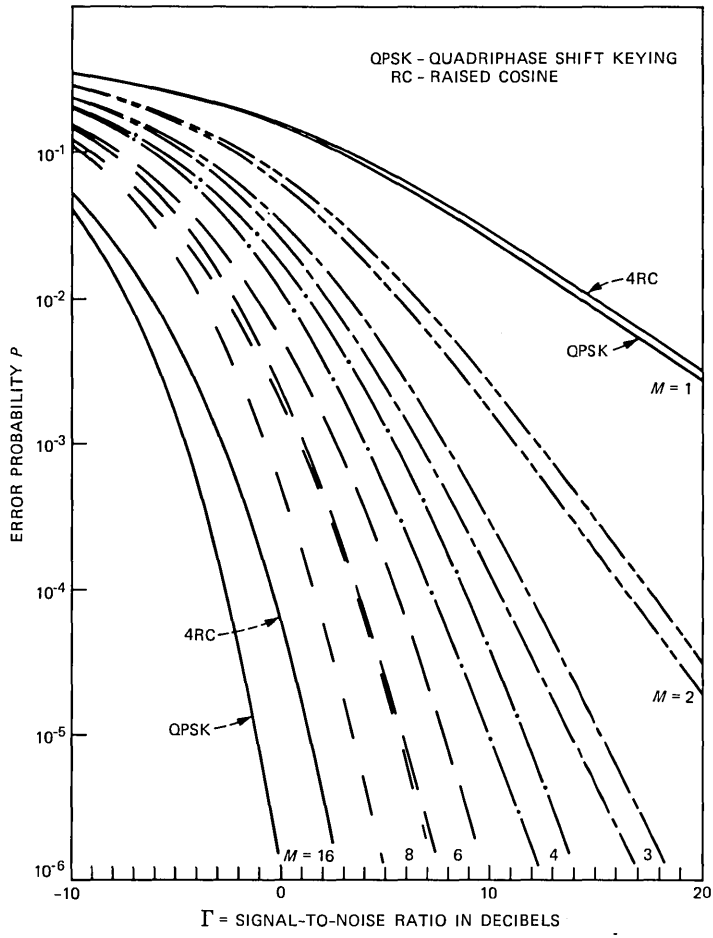


Fig. 4—Error probability, P, versus average per diversity branch (and per bit) signal-to-noise ratio Γ for a coherent MSK-type receiver with Selected Pulse Amplitude Modulation (SPAM) receiver. Maximal-ratio combining and modulation scheme 4RC are used. Curves shown without differential encoding/decoding.

antenna configuration, and a fixed cell size, a large frequency reuse factor, N, gives good cochannel interference (the interfering cells get further away) but low system capacity, since the number of available channels in one cell is inversely proportional to N.[1-3] In this paper, we will basically confine the discussion to one type of cell, namely the type with three 120-degree corner antennas, as shown in Fig. 7. The maximum distance from any base station d_0 is equal to the cell radius r. For details about cellular systems based on this cell, see Refs. 1 and 3. Other cellular concepts are considered in Ref. 16.

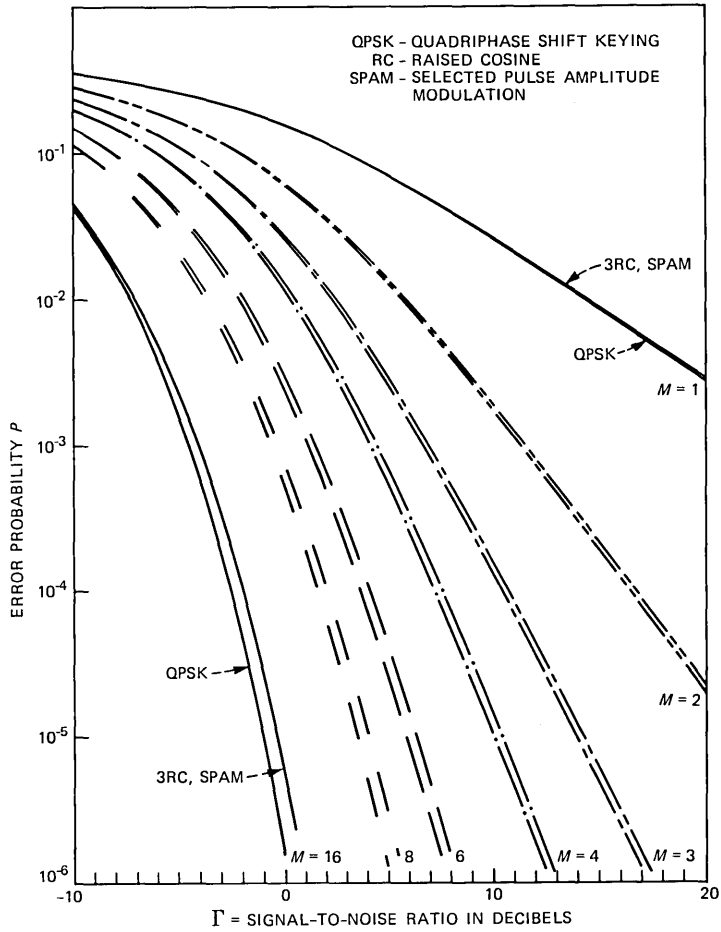


Fig. 5—Same as Fig. 4 with 3RC and SPAM-filter M branch diversity with maximal-ratio combining.

2.3 Time-division retransmission

The time-division retransmission (TDR) concept is described in Refs. 3 and 4. The basic ideas are as follows. The fading channel changes "slowly" (during one burst or package). Communication in both directions takes place in packages transmitted in the same frequency band: first mobile-to-base and then base-to-mobile. During mobile-to-base transmission, the channel is estimated for maximal-ratio, space-diversity combining in both directions. Several schemes[3] have been proposed for performing the required co-phasing for transmission from and reception at the base station. Because all diversity combining takes place at the base stations, the mobile equipment is

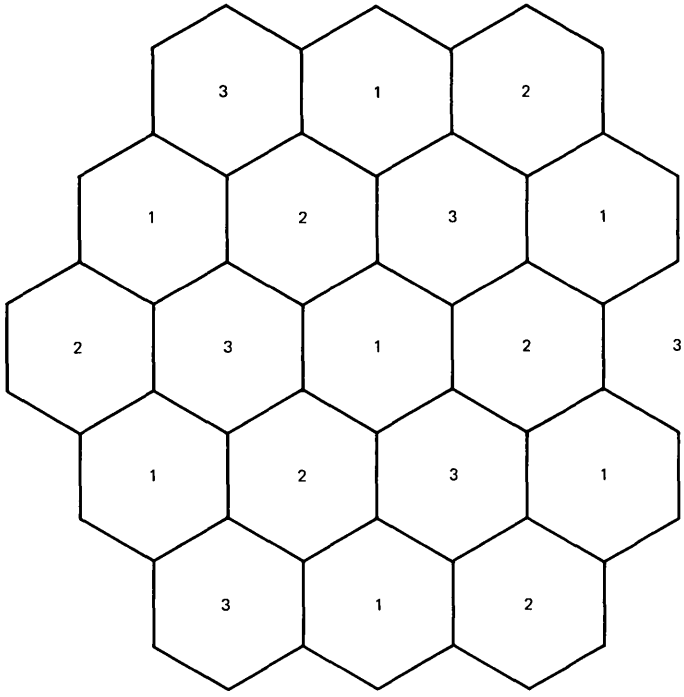


Fig. 6a—Example of cellular system. Each cell is a hexagon of equal area, and frequencies are reused. Cells marked with same digit use same frequency channel set (see Fig. 6b). In cluster of $N = 3$ cells, all channels are used.

relatively simple. As a consequence, it is feasible to use more than two branches of diversity with TDR.

2.4 Signal propagation and interference

It is assumed that mobile radio reception in an urban environment is characterized by

$$P(\bar{r}) = |\bar{r}|^{-\alpha} S(\bar{r}) \cdot R^2(\bar{r}), \tag{5}$$

where $P(\bar{r})$ is the received signal power at location $\bar{r}$ (position vector relative to a transmitter).[1,3,4] The first factor, $|\bar{r}|^{-\alpha}$, is a reduction factor due to the distance between the mobile unit and the transmitter, and α is the propagation constant. It is normally assumed that α is in the range three to four in the urban environment.[1,3] In free space, $\alpha = 2$.

The second factor, $S(\bar{r})$, represents shadow fading,[1,3,4] and the third factor, $R^2(\bar{r})$, represents Rayleigh fading.[4,17] R is the envelope of the received signal and is modeled as a random variable with the density function

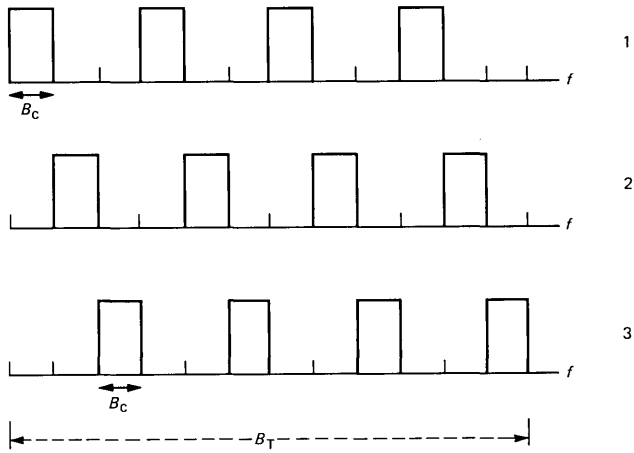


Fig. 6b—Frequency plan for cellular system with $N = 3$, shown in Fig. 6a. Channel set marked 1 used in each cell marked 1, etc. Channel bandwidth is denoted B_c, and total system bandwidth is B_T. Channel means two-way channels, including overhead for synchronization (when required).

$$p(R) = 2Re^{-R^2}, \tag{6}$$

with $E\{R^2\} = 1$ (see Refs. 4 and 17). In general, R varies with vehicle location and signal frequency.

This paper also considers propagation and interference in cellular systems with frequency reuse. We will make use of the same basic assumptions as in Refs. 1 and 3.

It is assumed that the cochannel interference and adjacent-channel interference from cells other than those containing the desired mobile unit are formed by the incoherent sum of contributions from many interfering sources. This sum is assumed to be equivalent to stationary Gaussian noise.[1,3] It is also assumed that the shadow and Rayleigh fading of the total interference is negligible compared to the fading of the signal.[1,3,4]

It is also assumed that cochannel interference is the main source of additive signal degradation. Adjacent-channel interference from other cells will be considered to some extent in a few cases, where it is assumed that the adjacent-channel interference is the incoherent sum of many sources forming a stationary additive Gaussian adjacent-channel noise, which is added to the cochannel interference. The additive thermal background noise is assumed to be negligible compared to the cochannel and the adjacent-channel interference. Thus, the transmitter power and the cell sizes are assumed to be such that alien background noise from sources other than mobile units and transmitters in the cellular system is negligible compared to cochannel and adjacent-channel interference.

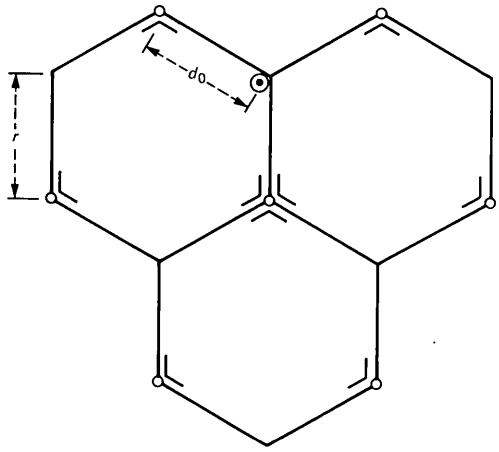


Fig. 7—Details for conventional three-corner, 120-degree antenna cell. Desired mobile shown in its worst-case locations.

Below, we will use the same technique as in Refs. 1 and 3 for calculating adjacent-channel interference and cochannel interference for continuous-phase modulation schemes. The signal-to-interference ratio is defined as the ratio of the signal power to the total noise power, based on the $|\bar{r}|^{-\alpha}$ propagation law and averaged over shadow and Rayleigh fading. It is assumed that the fading is flat over the band of each channel.

III. A BASEBAND COMBINER FOR TIME-DIVISION RETRANSMISSION WITH CONSTANT-AMPLITUDE MODULATION

The time-division retransmission (TDR) concept has so far been applied only to binary modulation schemes with nonconstant amplitude.[3] For constant-amplitude modulation schemes, such as continuous-phase modulation, the combiner in Ref. 3 needs to be slightly generalized. That is the subject of this section. We have not yet analyzed the impact of the constant-amplitude format on synchronization. As a first rough estimate, it is assumed that the same scheme as that in Ref. 3 can be used also for constant-amplitude modulation. Further work in considering other schemes[3] for synchronization and establishing phase references is required for constant-amplitude modulation schemes.

The baseband combiner for space diversity with maximal-ratio combining used by P. S. Henry and B. Glance[3] for binary phase shift keying (BPSK) is easily generalizable to quadrature constant-amplitude modulation. The base-station signal processing equipment for one branch of diversity is shown in Fig. 8 for time-division retransmission with quadrature constant-amplitude modulation. During the

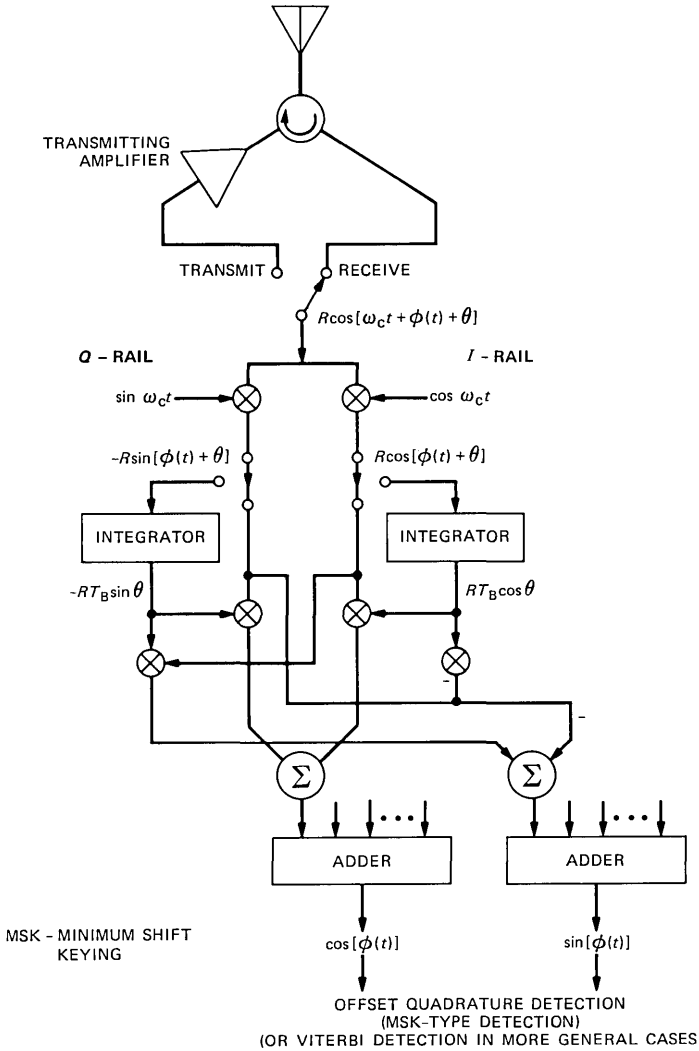


Fig. 8—Baseband combiner for constant-amplitude modulation and time-division retransmission. Compare this to Fig. 5 in Ref. 3.

reference transmission interval, the carrier burst of duration T_B is received, having the form

$$R\cos(\omega_c t + \theta), \tag{7}$$

where R is the Rayleigh amplitude, and θ is the unknown phase. It is assumed that the R and θ are essentially constant until the next reference burst is transmitted.[3] Then the stored R and θ are updated.

After down-conversion and integration, the reference coefficients $T_B R \cos\theta$ and $T_B R \sin\theta$ are produced and stored for message demodulation. During information transmission, the received signal can be written

$$R\cos[\omega_c t + \phi(t) + \theta] + I_c \cos\omega_c t + I_s \sin\omega_c t, \tag{8}$$

where $\phi(t)$ is the information-carrying phase (see the introduction). I_c and I_s are Gaussian interference with zero mean and variance S^2 (see Ref. 3). After down-conversion and multiplication with the reference coefficients and combinations following the network in Fig. 8, we have in each quadrature arm the signals

$$\begin{aligned} S_I = {} & T_B R^2 \cos[\phi(t) + \theta]\cdot\cos\theta \\ & + T_B R^2 \sin[\phi(t) + \theta]\cdot\sin\theta \\ & + I_c T_B R \cos\theta - I_s T_B R \sin\theta \end{aligned} \tag{9}$$

in the cosine branch (I-rail), and

$$\begin{aligned} S_Q = {} & T_B R^2 \sin[(\phi(t) + \theta]\cdot\cos\theta \\ & - T_B R^2 \cos[\theta(t) + \theta]\cdot\sin\theta \\ & - I_c T_B R \sin\theta + I_s T_B R \cos\theta \end{aligned} \tag{10}$$

in the sine branch (Q-rail). Using simple trigonometric formulas we have

$$S_I = T_B R^2 \cos[\phi(t)] + I_c T_B R \cos\theta - I_s T_B R \sin\theta, \tag{11}$$

$$S_Q = T_B R^2 \sin[\phi(t)] - I_c T_B R \sin\theta + I_s T_B R \cos\theta. \tag{12}$$

The special case of BPSK in Ref. 3 is obtained by letting $\phi(t)$ be

$$\phi(t) = \begin{cases} 0 & \text{for data } +1 \\ \pi & \text{for data } -1. \end{cases} \tag{13}$$

Thus $\cos\phi(t) = \pm 1$, and $\sin\{\phi(t)\} = 0$. Only one output is required in this case. Maximal-ratio combining at the output is obtained by adding the components from the M branches.[3] Coherent reception is obtained and the data can be demodulated by processing $\cos\phi(t)$ and $\sin\phi(t)$. This can, for example, be done by means of the "MSK-type" receiver (offset quadrature receiver)[7–9] or, for the more general case, even a Viterbi detector.[6,13]

The combiner in Fig. 8 followed by an "MSK-type receiver" supplies the motivation for considering the family of $h = 1/2$ partial response FM or continuous-phase modulation (CPM) schemes[6,7] for the case of slow Rayleigh fading with M-branch space diversity with maximal-ratio combining.[4,7]

For transmission from the base station to the mobile unit, all stored phase references in Fig. 8 are changed from θ to $-\theta$. Thus, through the time-division retransmission procedure, maximal-ratio combining is approximately obtained at the mobile unit with only one antenna and one receiver.[3] All signal processing is performed at the base-station transmitter.

IV. COCHANNEL AND ADJACENT-CHANNEL INTERFERENCE ANALYSIS FOR CPM

4.1 Cochannel interference analysis

The technique for the cochannel interference analysis of CPM is basically the same as that for BPSK QPSK in Refs. 1, 3, and 16. The signal-to-cochannel interference ratios are given by the geometry of the cells and antennas and of the propagation exponent α [in the range three (pessimistic) to four]. From Refs. 1 and 3, we have the average signal-to-cochannel interference ratio with a three-corner cell system with three cells per cluster

$$P_S/P_I = 7.5 \text{ dB} \tag{14}$$

for mobile-to-base transmission and

$$P_S/P_I = 8.0 \text{ dB} \tag{15}$$

for base-to-mobile transmission. Figure 4 shows the theoretical number of diversity branches required in a space diversity (TDR) 4RC and QPSK scheme. Figure 5 shows the corresponding curves for 3RC.

With a larger number of cells per cluster (large channel sets), the required signal-to-cochannel interference ratio is larger, and thus fewer diversity branches are required.

It is assumed in the calculations that the background noise from other sources is negligible. The calculated signal-to-interference ratio determines which modulation method can be used (in terms of required detection efficiency) and how many branches of diversity are required.

Note that the signal-to-cochannel interference ratio is not dependent on a particular modulation scheme. It is given by the relationships of distances from the desired base station to the desired mobile unit and from the interfering base stations or mobile units. See Refs. 1, 3, and 16 for details. Thus, the cell geometry gives the available signal-to-cochannel interference ratio. This number plus the bit error probability requirement determines the number of diversity branches for a given modulation scheme by using results like those in Ref. 12 (see Figs. 4 and 5).

Reference 16 gives cochannel interference results for a generalized class of cells.

4.2 *Adjacent-channel interference analysis*

The effect of adjacent-channel interference in cellular mobile radio systems is briefly considered for some cases with constant-amplitude modulation schemes.

Figure 9 shows the so-called fractional out-of-band power curves for the binary-raised cosine family for modulation index $h = 1/2$.[6,7] These curves show the fraction of the spectral power relative to the total power outside the band $-f + f_0$, $f + f_0$ for the data rate $f_b = 1/T_b$. Thus, the bandwidth of the 4RC with the portion 10^{-3} of the total power outside the band is approximately $0.95/T_b$. Note that the figure shows half the bandwidth of one channel. Figures 10 and 11 show 3SRC and 4SRC. Similar curves for GMSK are published in Ref. 10.

Curves like those in Fig. 9 are useful for calculating the adjacent-

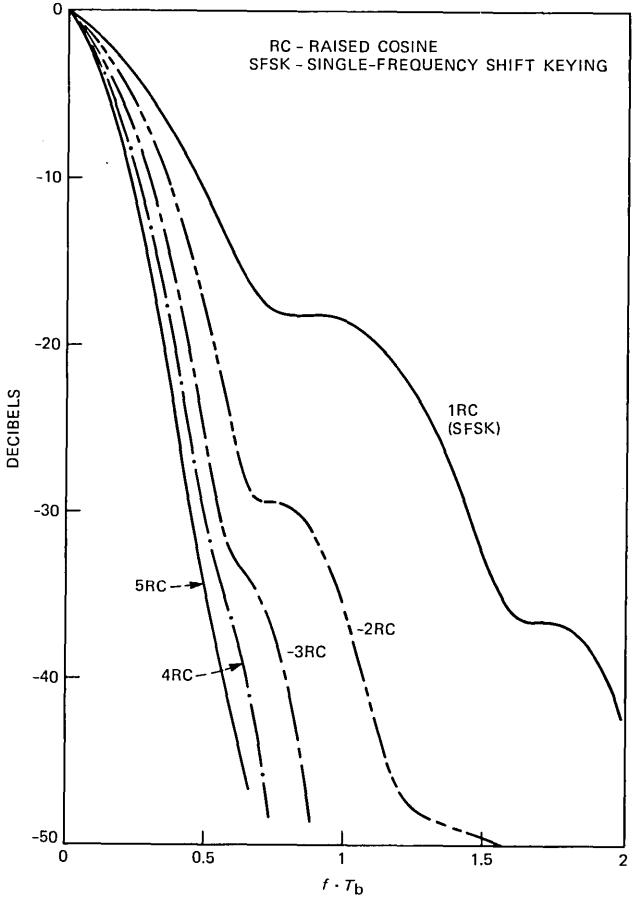


Fig. 9—Fractional out-of-band power curves for the binary modulation schemes 1RC, 2RC, 3RC, 4RC, and 5RC, $h = 1/2$.

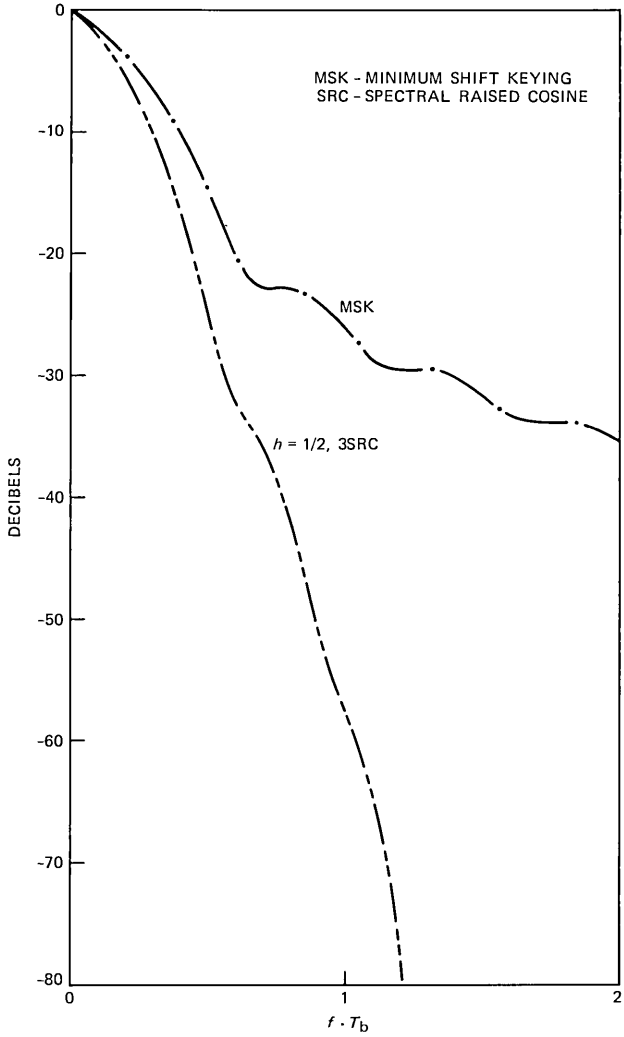


Fig. 10—Fractional out-of-band power curves for 3SRC, $h = 1/2$, and MSK.

channel interference. From the curves in Fig. 9, we obtain an upper bound on the interference power in the two adjacent channels on each side. Since the power spectra fall off rapidly with f for increasing frequencies, this upper bound is a good estimate of the adjacent-channel interference power level.

From the results in the appendix and from the cochannel interference results above and in Refs. 1, 3, and 16, it is evident that with "wide" channels compared to the bit rate, average adjacent-channel interference is very small compared to cochannel interference. In the

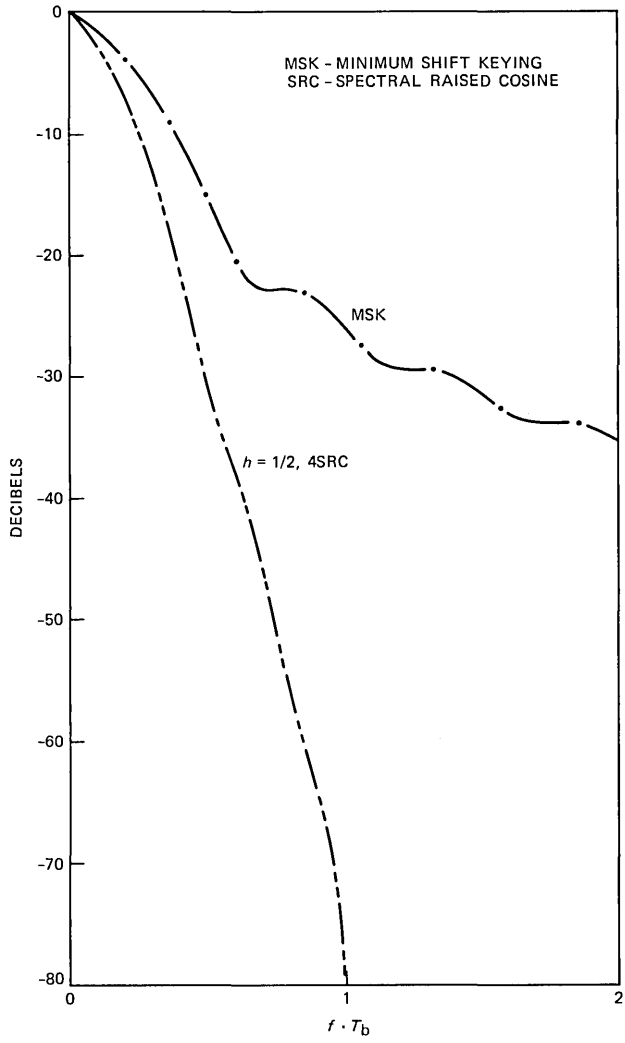


Fig. 11—Fractional out-of-band power curves for 4SRC, $h = 1/2$, and MSK.

appendix, a parameter, A, is introduced, which is a measure of the adjacent-channel interference suppression with a given modulation scheme and a given relationship between the bit rate and the channel bandwidth. Even with A as large as 10^{-2}, the average adjacent-channel interference is almost negligible compared to the cochannel interference. Much smaller A's are quoted in the literature (10^{-6}–10^{-7}).[8,10,18] With A as small as 10^{-2}, the detection efficiency of the modulated signal will be affected by the filtering. In the system discussion below, we will choose "small" A's.

For schemes with $N \geq 3$, the adjacent-channel interference with one particular cell is small with modulation schemes like 3RC-4RC (see Figs. 1, 6, and 7). The average adjacent-channel interference from adjacent cells is small compared to the cochannel interference. The same is true for $N = 1$ for the adjacent-channel interference from adjacent cells. However, for $N = 1$, the adjacent-channel interference within one cell is a more severe problem. This will restrain the value of bit rate f_b, relative to the channel bandwidth B_c. Cellular systems with $N = 1$ are considered in Refs. 1 and 16.

V. DISCUSSION AND CONCLUSIONS

We can now put the various pieces of information together and evaluate the number of available channels for cellular schemes with different modulation schemes. Time-division retransmission is used in several cases but not all. For the cases without time-division retransmission, it is expected that the number of space diversity branches cannot be larger than $M = 2$. It is inconvenient to have more than two antennas on the mobile unit.

Let B_T be the total system bandwidth, and let B_c be the bandwidth of one two-way TDR channel (or two one-way channels without TDR). Let the bit rate over one channel be $f_b = 1/T_b$ bits/s. The total number of available channels is B_T/B_c. Let N be the number of cells in a cluster. The number of channels, C, available in each cell in the cluster is $B_T/B_c \cdot 1/N$. The total system bandwidth will be kept constant for comparison of various systems. It is evident that C can be improved by decreasing N or decreasing B_c (at a given B_T and a given transmitted bit rate f_b). Decreasing N has been discussed in Ref. 16. Decreasing B_c can be done to a certain degree by selecting narrowband modulation schemes. However, the detection efficiency decreases for schemes with simple receivers. Furthermore, the adjacent-channel interference increases. There seems to be a preferred modulation scheme with an intermediate degree of smoothing (see Ref. 7 for details). These we have chosen below.

The total system bandwidth is assumed to be 40 MHz, and the carrier frequency $f_c = 850$ MHz. The propagation constant is $\alpha = 3$, and the error probability is 10^{-3}. It is assumed that digitized speech is transmitted with 32 kb/s. If 16 kb/s is used (and all other system parameters are kept unchanged), the total system capacity in terms of channels is of course doubled for all schemes. When comparisons are made to other schemes,[3,15] these will be scaled to "our" system bandwidth 40 MHz and data rate $f_b = 32$ kb/s.

The calculations of number of diversity branches given below assume ideal conditions all the way through, i.e., no degradation due to

filtering, nonlinearities, timing, synchronization, combining, etc. Sometimes there is a margin, though.

Below we will calculate the number of channels per cell (antenna) for a few different cases.

1. $N = 3$, i.e., a cell with 120-degree base stations in three corners. Time-division retransmission. BPSK with 81 kb/s data rate (the overhead over 64 kb/s is due to the co-phasing procedure for the TDR scheme). Channel bandwidth $B_c = 100$ kHz (see Ref. 3). With this channel bandwidth, the transmitted BPSK signal must be filtered quite hard, resulting in a nonconstant amplitude. $M = 3$ branch diversity. This scheme gives about 133 two-way channels.

2. $N = 3$ and the same cells as in 1. TDR with 81 kb/s in a channel bandwidth of 81 kHz, 4RC (or better 4SRC[6]) modulation (see Figs. 9 to 11 for adjacent-channel interference estimates). Mild filtering of the 4RC might be necessary, causing small fluctuations of the envelope (see Fig. 4 for error probability considerations). TDR and $M = 4$ branches of diversity. The number of channels is 164.

3. $N \cong 11$, a cell with centrally located, omnidirectional antenna. $M = 2$ branches of diversity, no TDR. GMSK modulation with $B_bT \cong 0.25$ (see Refs. 10 and 15 for parameter definitions and other details). This corresponds roughly to 3RC (3SRC). Adjacent-channel interference is about -70 dB. 64 kb/s in 107 kHz. The results in Ref. 15 were scaled to the total bandwidth of 40 MHz and the data rate of 32 kb/s in the one-way speech channel. About 33 channels are available.

4. Same as 3 above with harder filtering. 64 kb/s in 58 kHz. Adjacent-channel interference is about -20 dB.[15] This leads to 62 channels.

The reasons that schemes 3 and 4 above yield such a low number of channels compared to the others are twofold. Time-division retransmission is not considered. Therefore an $M = 2$ is the highest degree of diversity considered. This leads to the high N value. The other important reason is that only centrally located omnidirectional antennas are considered. Thus, even with a good modulation scheme, the number of channels remains low.

A class of constant-amplitude modulation schemes is considered for the cellular digital radio system. Such features as power spectral density, fractional out-of-band power, and detection efficiency with simple near-optimum detectors for the slow Rayleigh fading channel with space diversity and maximal-ratio combining are reported.

From the analysis and discussions above, we have seen that a large number of channels can be provided in a digital cellular system by the proper combination of antenna configuration, modulation scheme, and diversity scheme. We have seen that constant-amplitude digital mod-

ulation schemes can give high-capacity systems. It should be pointed out that higher system-capacity levels can be expected if nonconstant-amplitude modulation schemes are acceptable.

A large number of unsolved problems remain. The analysis above (as indeed that in Refs. 1, 3, and 15) is based on very simple and idealized channel and interference models. More refined analysis and simulations will be necessary. The analysis above was carried out under the idealized assumptions of flat fading and of negligible effects of filtering and nonlinearities on the modulation scheme. Uniform transmission conditions were assumed for all cells. No delay spread was considered. Perfect timing and synchronization were assumed with coherent detection and ideal maximal-ratio combining. It was furthermore assumed that perfect synchronization for the time-division retransmission scheme was established. All of the above problems and others have to be taken into account in a refined system analysis.

At least the same number of channels seems to be within reach with the space-diversity and time-division retransmission schemes as with the spread spectrum multiple access approach to digital mobile radio (see Refs. 3, 19, 20, and 21).

VI. ACKNOWLEDGMENT

Thanks are due to Tor Aulin, who calculated the fractional out-of-band power curves in Figs. 9 to 11 and to Arne Svensson, who calculated the error probability curves in Figs. 4 and 5 (see Ref. 12).

REFERENCES

1. Y. S. Yeh and D. O. Reudink, "Efficient Spectrum Utilization for Mobile Radio Systems Using Space Diversity," IEEE Trans. Commun., *COM-30*, No. 3 (March 1982), pp. 447–55.
2. V. H. MacDonald, "The Cellular Concept," B.S.T.J., *58*, No. 1 (January 1979), pp. 15–41.
3. P. S. Henry and B. S. Glance, "A New Approach to High Capacity Digital Mobile Radio," B.S.T.J., *60*, No. 8 (October 1981), pp. 1891–904.
4. W. C. Jakes, Jr., *Microwave Mobile Communications*, New York: Wiley, 1974.
5. T. Aulin and C.-E. Sundberg, "Continuous Phase Modulation—Part I: Full Response Signaling," IEEE Trans. Commun., *COM-29*, No. 3 (March 1981), pp. 196–209.
6. T. Aulin, N. Rydbeck, and C.-E. Sundberg, "Continuous Phase Modulation—Part II: Partial Response Signaling," IEEE Trans. Commun., *COM-29*, No. 3 (March 1981), pp. 210–25.
7. T. Aulin, C.-E. Sundberg, and A. Svensson, "MSK-Type Receivers for Partial Response Continuous Phase Modulation," Int. Conf. Commun., Philadelphia, Pennsylvania, June 1982, pp. 6F3.1–6.
8. F. deJager and C. B. Dekker, "Tamed Frequency Modulation, A Novel Method to Achieve Spectrum Economy in Digital Transmission," IEEE Trans. Commun., *COM-26*, No. 5 (May 1978), pp. 534–42.
9. R. deBuda, "Coherent Demodulation of Frequency-Shift Keying with Low Deviation Ratio," IEEE Trans. Commun., *COM-20* (June 1972), pp. 429–35.
10. K. Murota and K. Hirade, "GMSK Modulation for Digital Mobile Radio Telephony," IEEE Trans. Commun., *COM-29*, No. 7 (July 1981), pp. 1044–50.
11. T. Aulin and C.-E. Sundberg, "Numerical Calculation of Spectra for Digital FM

Signals," Nat. Telecommun. Conf., New Orleans, Louisiana, Conference Record, December 1981, pp. D.8.3.1–7.
12. C.-E. Sundberg, "Error Probability of Partial Response Continuous Phase Modulation With Coherent MSK-type Receiver Diversity and Slow Rayleigh Fading in Gaussian Noise," B.S.T.J., *61*, No. 8 (October 1982), pp. 1933–63.
13. T. Aulin, C.-E. Sundberg, and A. Svensson, "Viterbi Detectors with Reduced Complexity for Partial Response Continuous Phase Modulation," Nat. Telecommun. Conf., New Orleans, Louisiana, Conference Record, December 1981, pp. A.7.6.1–7.
14. W. Hirt and S. Pasupathy, "Suboptimum Reception of Binary CPSK Signals," Proc. IEE, Part F, *128*, No. 3 (June 1981), pp. 125–34.
15. K. Murota, K. Konoshita, and K. Hirade, "Spectrum Efficiency of GMSK Land Mobile Radio," Int. Conf. Commun., Denver, Colorado, Conference Record, June 1981, pp. 23.8.1–5.
16. C.-E. Sundberg, "Alternative Cell Configurations for Digital Mobile Radio Systems," B.S.T.J., this issue.
17. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw-Hill, 1966.
18. D. Muilwijk, "Tamed Frequency Modulation—A Bandwidth-Saving Digital Modulation Method Suited for Land Mobile Radio," Philips Telecommun. Rev., *37*, No. 1 (March 1979), pp. 35–49.
19. G. R. Cooper and R. W. Nettleton, "A Spread-Spectrum Technique for High-Capacity Mobile Communications," IEEE Trans. Veh. Technol., *VT-27* (November 1978), pp. 264–75.
20. P. S. Henry, "Spectrum Efficiency of a Frequency-Hopped DPSK Spread-Spectrum Mobile Radio System," IEEE Trans. Veh. Technol., *VT-28* (November 1979), pp. 327–32.
21. D. J. Goodman, P. S. Henry, and V. K. Prabhu, "Frequency-Hopped Multilevel FSK for Mobile Radio," B.S.T.J., *59*, No. 7 (September 1980), pp. 1257–75.

APPENDIX

Details of adjacent-channel interference calculations

In this appendix, we will demonstrate how the method for analyzing cochannel interference in Refs. 1 and 3, with minor modifications, can be used for calculation of adjacent-channel interference from adjacent (and further away) cells.

The average adjacent-channel interference from adjacent cells is calculated with basically the same method as the cochannel interference. It is assumed that the total adjacent-channel interference is formed by incoherent additions of several interference sources forming a stationary additive Gaussian noise, which is added to the cochannel interference noise. We will see that with "reasonable" spectral shape of the modulation scheme, this intereference source is, in most cases, small compared to the cochannel interference.

Let A be the relationship between the adjacent-channel interference power on one side of the channel to the total power in this channel. This number is upper bounded by 1/2 of the fractional out-of-band power level at $f = B_C$ (see Figs. 9 to 11 and Ref. 5). It is assumed that the receiver filter consists of an ideal bandpass filter of width B_C in these idealized calculations (the same assumptions are used in Ref. 15).

The base stations and the mobile units in nearby cells will cause adjacent-channel interference because of nonideal, nonband-limited

signals. We will consider the constant-amplitude raised cosine schemes (see Section II) transmitted without bandpass filtering. The spectral tails will cause adjacent-channel interference.

First, we will consider adjacent-channel interference for the case $N = 3$, $\alpha = 3$ with centrally located omnidirectional base stations (see Fig. 12). It is immediately clear from the frequency plan (Figs. 6a and 6b) that the adjacent-channel interference from sources within the same cell is extremely small with any spectrally efficient modulation scheme because each adjacent channel within one cell is three channels apart. The significant contributions to the total adjacent-channel interference come from adjacent channels in adjacent cells.

Once the major sources of interference are identified, it is straightforward to use the same methods as in Refs. 1, 3, and 16 for the calculation of the interference. The average adjacent-channel interference is suppressed by a factor of A compared to the cochannel inter-

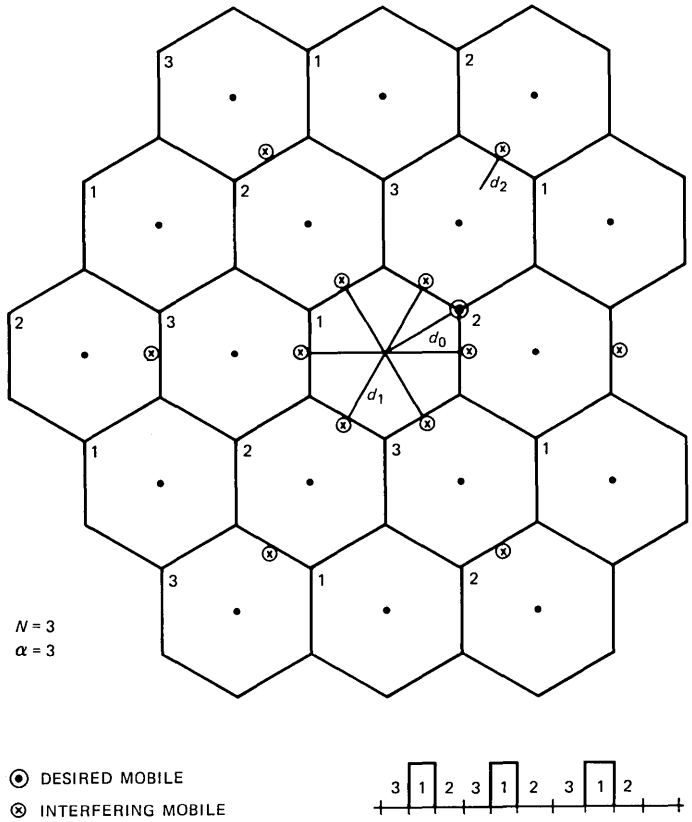


Fig. 12—Worst-case adjacent-channel interference cases for mobile-to-base transmission with centrally located antennas.

ference from the same sources. Space diversity helps against average adjacent-channel interference just as well as cochannel interference.[1,3]

Figure 12 shows the calculation of the worst-case (noncoherent addition) adjacent-channel interference for the $N = 3$, $\alpha = 3$ case with centrally located base stations and mobile-to-base transmission. Adjacent-channel interference comes from all cells marked 2 and 3. (Cochannel interference comes from cells marked 1; this is considered in Refs. 1, 3, and 16.) The distance to the desired mobile unit serviced from the omnidirectional antenna in the center of the cell is d_0. The distance to the nearest (worst-case) six interfering mobile units is d_1 (see Fig. 12). The distance to the next six interfering mobile units is d_2. Thus, the worst-case adjacent-channel interference to signal ratio for mobile-to-base transmission is

$$\frac{P_I}{P_S} = A\left[6\left(\frac{d_0}{d_1}\right)^3 + 6\left(\frac{d_0}{d_2}\right)^3 + \text{further terms}\right]. \tag{16}$$

For the case in Fig. 12,

$$P_I/P_S \cong 9.5 \cdot A. \tag{17}$$

Assuming that the interfering mobile units can be in any equally probable location in their cells, we found that the average adjacent-channel interference to signal ratio for mobile-to-base is

$$P_I/P_S \cong 3.6 \cdot A. \tag{18}$$

It is straightforward to obtain similar results for base-to-mobile transmission for the case in Fig. 12. We have the adjacent-channel interference to signal ratio

$$P_I/P_S \cong 2.5 \cdot A. \tag{19}$$

It is also straightforward to calculate the formulas for a cell with three 120-degree corner stations in each cell (see Fig. 7). For $N = 3$, $\alpha = 3$, we have for this case:
base-to-mobile,

$$P_I/P_S \cong 1.2A, \tag{20}$$

mobile-to-base average,

$$P_I/P_S \cong 1.5A. \tag{21}$$

In principle, it is straightforward to calculate the corresponding adjacent-channel interference formulas for the $N = 3$, $\alpha = 3$ case for other cells.[16] However, as a rough estimate one can use A times the cochannel-interference to signal ratio for the corresponding $N = 1$ case.[16]

For all calculations above, it was assumed that the desired mobile unit is in the least favorable position in the cell, much the same way as in Refs. 1 and 16.

Note that in the approximate calculations above, we only considered averages. It is possible that the desired mobile unit and the interfering mobile unit in the adjacent cell occupying a frequency channel immediately adjacent to that of the desired mobile unit are almost at the same location near the cell dividing line. For such cases, the adjacent-channel interference might be such that the space diversity does not suppress this interference (the combiner in the desired mobile unit adds the interference and treats this interference just as the desired signal). For this case, a sufficiently small A is required. It is, of course, also conceivable to have a frequency channel change for one of the mobile units in that case.

For the $N = 1$ case, the adjacent-channel interference problem becomes larger than for the $N = 3$ case. The background average adjacent-channel interference from sources in adjacent cells is still there as before. On top of that, there is interference from (primarily two) adjacent channels within the cell of the desired mobile unit. A small A is required to suppress this interference.

AUTHOR

Carl-Erik W. Sundberg, M.S.E.E and Dr. Technology, Lund Institute of Technology, University of Lund, Lund, Sweden, in 1966 and 1975, respectively. Currently, Mr. Sundberg is a Research Professor in the Department of Telecommunication Theory, University of Lund. He is also Director of the consulting company SUNCOM, Lund. During 1976 he was with the European Space Research and Technology Centre (ESTEC), Noordwijk, The Netherlands, as an ESA Research Fellow. He has held positions as Consulting Scientist at LM Ericsson and SAAB-SCANIA, Sweden, and at Bell Laboratories, Crawford Hill, N. J. His research interests include source coding, channel coding, especially decoding techniques, digital modulation methods, fault-tolerant systems, digital mobile radio systems, spread spectrum systems, and digital satellite communication systems. He has published a large number of papers in these areas during the last couple of years. Dr. Sundberg holds several U.S. and European patents. Senior Member, IEEE; member, SER (Sveriges Elektroingenjörers Riksförening).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

A Compatible High-Definition Television System

By T. S. RZESZEWSKI*

(Manuscript received March 11, 1983)

A compatible HDTV system with the potential of having significantly better picture quality than the present NTSC color TV system is proposed. It realizes an increase in horizontal and vertical resolution and has considerably less crosstalk between the composite signal components compared to the NTSC signal. The increased resolution will allow a display as large as present home projection televisions with a sharper-looking and more detailed image than is possible with the present NTSC system. Also, the elimination of crosstalk adds to picture quality. A large screen size together with improved image quality should provide the user with a feeling of realism and involvement. This system also allows the use of more detailed graphics and more text per page for new services such as teletext.

I. INTRODUCTION

A compatible high-definition television (HDTV) system capable of producing an image quality significantly better than NTSC is proposed. Viewing the pictures produced by this system on a large screen should result in an increased sense of realism over the present NTSC television. This new system uses a split-luminance and split-chrominance (SLSC) type of transmission. The areas where the primary benefits are expected from this HDTV system over the National Television System Committee (NTSC) system are:

1. Increased horizontal resolution,
2. Increased vertical resolution, and
3. Less crosstalk between the components of the signal.

* Bell Laboratories.

Several different approaches to HDTV can offer excellent quality. However, there are quality-independent attributes that will aid the acceptance of this type of service. In particular, compatibility with present NTSC receivers and a bandwidth of no more than twice the present 6-MHz channel for broadcast are critical.

Many different forms of compatibility are discussed today. However, as used here, compatibility means receiver compatibility. The signal must be able to feed an HDTV and an NTSC TV simultaneously and be received on the NTSC receiver with substantially the same quality picture that those sets presently realize, while the HDTV receiver realizes all the benefits, such as increased resolution.

Spectrum usage is another problem with some HDTV approaches. The NHK (Japanese) system uses 30 MHz of baseband.[1] This bandwidth is so large that this type of service could not be broadcast in the same manner as the present broadcast service, and would leave this service with fewer delivery systems. Of course, video tape or cable are still possible delivery systems. Direct broadcast by satellite (DBS) is also possible; however, the prime allocations that are not affected by the weather are likely to be used for DBS of NTSC in the near future.

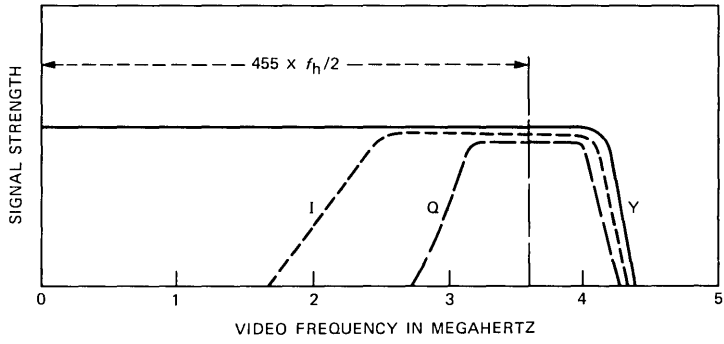
Some other proposed systems have preserved some aspect of compatibility.[2,3] They retain the scanning format, but change the encoding such that a present receiver will not operate correctly.

A format is proposed here that uses a 10-MHz baseband composite signal that can be transmitted as a vestigial sideband, amplitude-modulated signal in a bandwidth of 12 MHz. Also, an NTSC receiver will operate with the same quality as at present when receiving this signal. The NTSC receiver must be tuned to the lower 6-MHz portion of the 12-MHz spectrum. The price that is paid for this compatibility is a reduction in the number of broadcast or CATV channels to approximately one half the present NTSC allocations. However, VHF stations are normally spaced at least every second channel allocation apart in a given location for broadcast, and UHF channels are spaced even further apart. Therefore, the impact on the present broadcast service is likely to be small. Also, new systems with bandwidth requirements like the NHK system would produce at least a six-to-one reduction in the number of channels compared to an NTSC service.

II. A COMPATIBLE APPROACH

2.1 System description

This new system is built around the NTSC baseband spectrum shown in Fig. 1. The NTSC baseband signal is modulated for broadcast as a vestigial-sideband, amplitude-modulated signal as shown in Fig. 2. The composite HDTV baseband is illustrated in Fig. 3. Note that there is approximately 0.75 MHz of spectrum left over for other



Y – THE LUMINANCE SIGNAL
I – I COMPONENT OF CHROMINANCE
Q – Q COMPONENT OF CHROMINANCE

Fig. 1—NTSC luminance and chrominance bandwidth allocation.

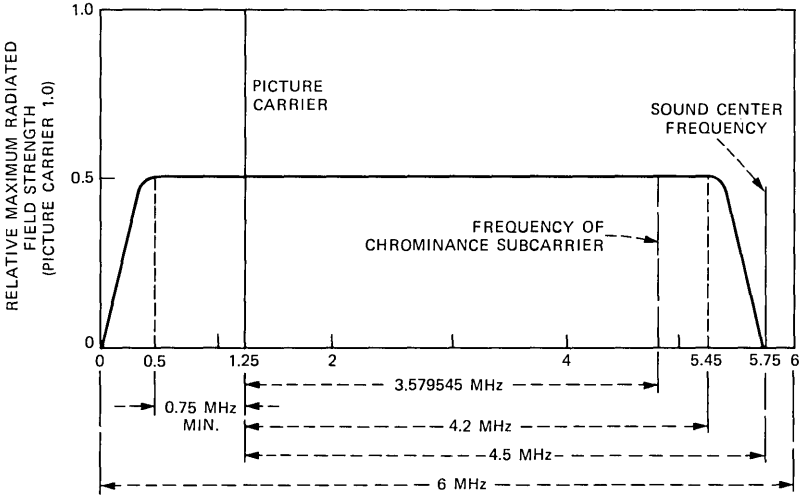


Fig. 2—Idealized picture transmission amplitude characteristic for NTSC.

services such as teletext and/or multichannel sound. The composite HDTV baseband signal could be modulated as a vestigial-sideband, amplitude-modulated signal for broadcast, as illustrated in Fig. 4. The selectivity of the NTSC receiver will reject the additional high-frequency portion of this signal at least as well if not better than an adjacent NTSC station.

The composite signal of Fig. 3 is obtained by starting with a 1050-line progressive scan source of wide-bandwidth red, green, blue (R, G, B) signals. The technique for improved vertical resolution is described

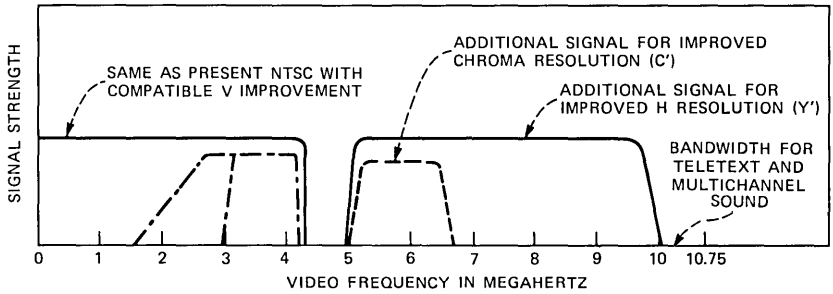


Fig. 3—SLSC HDTV baseband spectrum.

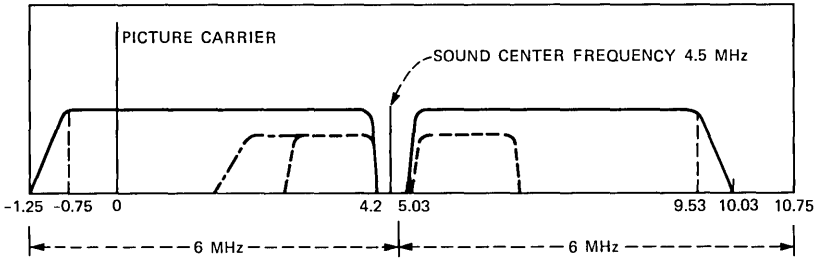


Fig. 4—Idealized broadcast picture transmission amplitude characteristic for SLSC HDTV.

in the literature.[4] That technique will be described in greater depth later. For now, it is sufficient to say that the wideband R, G, B signals are filtered and converted to a 525-line signal by a scan conversion that deletes every second line to obtain a 525-line signal suitable for transmission. Next, the wideband R, G, B signals are matrixed into a Y, I, Q format. This process is illustrated in Fig. 5. The NTSC encoder provides the appropriate selectivity and processing to create a standard NTSC signal that occupies the lower 4.2 MHz of the baseband spectrum in Fig. 3. The additional signal for improved horizontal resolution (Y′) that occupies the frequency region of approximately 5 to 10 MHz in the baseband is processed by the High Frequency Luminance Encoder. The High Frequency Chrominance Encoder creates the signal C′.* These signals are added to the signal obtained from the NTSC encoder to produce the composite HDTV baseband signal of Fig. 3. A detailed block diagram that elaborates on the encoding of these extra signals will be given later.

* A method of recreating the high-frequency components of the composite NTSC signal that were filtered out at the transmitter has recently been developed.[5] If this technique is proven to be acceptable for this application, the additional signal for extra chrominance resolution (C′) is not needed.

The system description in this paper utilizes the anti-aliasing filtering of the R, G, B signals in the encoder, as shown in Fig. 5. Consequently, the interpolation filtering in the decoder will also be performed on R, G, B as described later. An alternative approach that may have some advantages is to transform the R, G, B signals into a Y, I, Q format before performing the anti-aliasing filtering. In that case the I and Q signals will be filtered to a smaller bandwidth before the anti-aliasing filter. This approach simplifies two of the anti-aliasing and interpolation-filter designs and requires less frame-store memory. It will be considered further in the section on system alternatives.

Figure 6 contains the general block diagram of the decoder. Not shown are the tuner, IF, and video detector that are required to select and demodulate the HDTV signal to a baseband signal if the composite signal is modulated for broadcast; these functions will be described later. In the decoder, the three portions of the HDTV signal (the low-frequency signals obtained from the NTSC composite signal, the

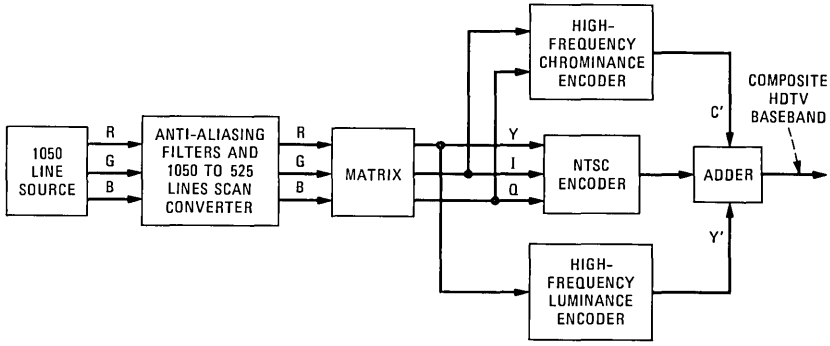


Fig. 5—SLSC HDTV encoder.

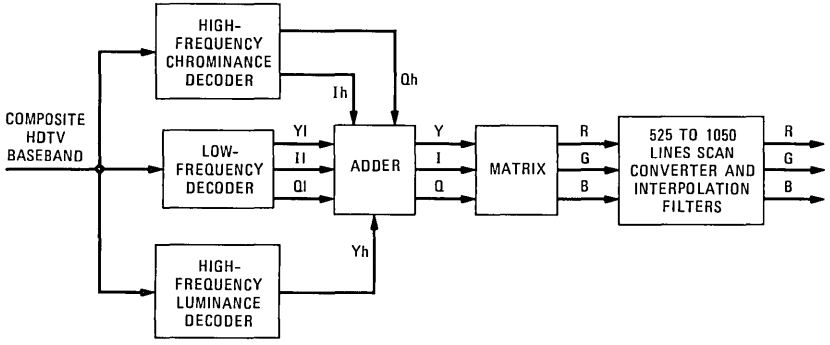


Fig. 6—SLSC HDTV decoder.

additional high-frequency chrominance signal, and the additional high-frequency luminance signal) are selected and processed in parallel. A detailed block diagram of these portions will be given later. The additional chrominance signals (Ih and Qh) are added to the chrominance produced by the low-frequency decoder to get I and Q signals that have a 2-MHz bandwidth each. These signals are then matrixed to produce 2-MHz color-difference signals. They are then added to the full 7.5-MHz-bandwidth luminance signal to produce R, G, B signals. The wideband luminance, Y, results from the addition of Yl and Yh.

However, before being fed to the display, the signals must be processed to realize the increased vertical resolution. The process consists of a scan conversion back to 1050 lines from the interlaced 525-line transmission format. The monitor may be either a 1050-line progressive scan or interlaced scan at the discretion of the manufacturer. However, progressive scan is needed to realize the full potential of this system. Interpolation filtering is performed in the scan-conversion process, and the high-definition R, G, B signals are fed to the display device.

2.2 Vertical resolution

An analysis of the resolution capability of the NTSC signal is helpful to assess the improvements that HDTV will realize. Resolution is expressed in terms of vertical and horizontal equivalent TV lines. The vertical resolution tells the number of horizontal lines alternating between black and white that can be resolved in the TV image. It is tempting to equate this to the total number of scan lines minus the lines in the vertical interval that are not used for display. Unfortunately, the scanning process that changes the image into an electrical signal in the camera and then reassembles the image on the display is really a sampling process. It is well known that sampled signals must first be bandlimited or aliasing will occur. It is the aliasing and the replicated spectra that further reduce the vertical resolution. Equation 1 below expresses the actual vertical resolution in a TV picture by the addition of a Kell factor (k) that takes this extra loss into account.

$$Rv = (Nt - 2Nv)k. \tag{1}$$

Rv is the vertical resolution; Nt is the total number of scan lines (525 for NTSC); Nv is the number of lines in the vertical interval (21 for NTSC); and k is the Kell factor. The Kell factor normally ranges between 0.6 and 0.7 for TV systems and results in an Rv range between 290 and 338 lines for NTSC TV. The method of vertical resolution improvement employed for the SLSC HDTV system allows the vertical

resolution to approach the full 483 lines of the active video; the Kell factor approaches unity.[4]

The modulation transfer function (MTF) in the camera and the display are analogous to the frequency response in linear system theory. It can be adjusted by shaping the electron beam in the camera and a CRT display. The contour of the scanning spot can be thought of as a two-dimensional impulse response commonly called the point-spread function. A narrow scanning spot in the vertical direction means a wide vertical spatial frequency spectrum and aliasing, and a wide scanning spot means overlapping of adjacent lines and low-pass filtering in the vertical direction (defocusing). In NTSC, the scanning spot is adjusted to compromise between aliasing and defocusing. Anti-aliasing (prefiltering) on a 1050-line progressive-scan source signal and interpolation (postfiltering) that eliminates replicated spectra aid the compromise. These filters can be used together with more lines at the camera in a compatible fashion to increase the vertical resolution.[4]

2.3 Horizontal resolution

The horizontal resolution is also expressed in terms of lines (equivalent vertical lines) that are the same width as the horizontal lines used to determine the vertical resolution above. There are two lines per cycle of video bandwidth. In other words, in the time—and therefore the horizontal space—it takes to display a cycle of the highest-frequency signal that will pass through the system bandwidth, the system will produce two lines on the display, one white and one black. The width of the lines is the same as for the vertical resolution, and the 4-to-3 aspect ratio is taken into account. Equation 2 below can be used to determine the horizontal resolution of a television system per unit of video bandwidth (Rh').

$$Rh' = 2Ta/AR. \tag{2}$$

Ta is the total active time for a horizontal line, and AR is the aspect ratio (these are 53.5 microseconds and 4/3 respectively for the NTSC system). This results in approximately 80 lines/MHz. Most NTSC receivers have at least 3 MHz of bandwidth, yielding a minimum of 240 lines of horizontal resolution. However, the total system luminance bandwidth can be 4.2 MHz if a comb filter is used in the NTSC receiver. This results in 336 lines.

The horizontal resolution is increased for the HDTV system proposed here by adding extra bandwidth to the baseband luminance signal. Figure 7a shows the 7.5-MHz bandwidth that is allocated to luminance in this compatible system. The same scanning format for transmission as NTSC is used; therefore, eq. (2) above indicates that the system will produce 80 lines/MHz. The 7.5-MHz bandwidth will

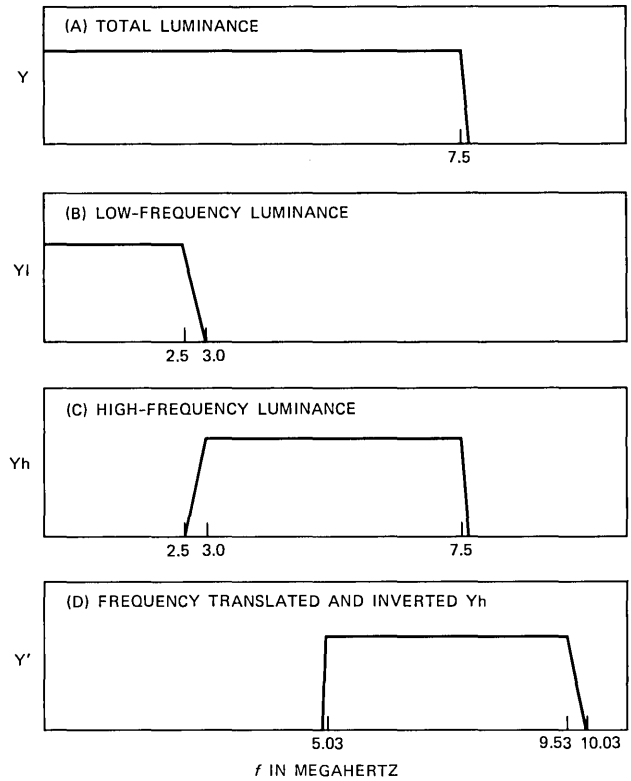


Fig. 7—Luminance (Y) processing in SLSC HDTV.

result in 600 lines of horizontal resolution.* This 7.5-MHz-bandwidth luminance spectrum feeds (1) an NTSC encoder that only uses the lower 4.2 MHz, and (2) a circuit that processes the high-frequency luminance detail (2.5 to 7.5 MHz) in a separate parallel channel.

Figures 7b through 7d illustrate the high-frequency luminance processing. Luminance (Y) is split up into a low-frequency portion (Yl) and high-frequency portion (Yh), each with a controlled roll off between 2.5 and 3 MHz (see Fig. 7c). If these two signals are added together, they will result in the original spectrum (Y) with a 7.5-MHz bandwidth. The high-frequency luminance will be processed so that it can be multiplexed with the NTSC baseband signal into a new, compatible HDTV baseband signal. This is accomplished by reversing

* This will result in 24 percent more horizontal resolution than vertical resolution for this system. There is the possibility of limiting the horizontal resolution to equal the vertical and simplifying the system. The amount of benefit produced by allowing greater horizontal resolution than vertical can be investigated when the system is implemented.

the frequency sense of Yh and translating it up in frequency to the location illustrated in Fig. 7d as Y′.

Note that the spectrum for Yh is cut off at a frequency below 2.5 MHz. Only the lower portion of the NTSC signal, which is substantially free from chrominance, is used for low-frequency luminance information (Yl).* Use of this lower cutoff frequency reduces the cross luminance problems that trouble the present NTSC system. An optimum cutoff frequency should be determined based on experiment since the higher the cutoff frequency the greater the horizontal resolution. The value indicated in Fig. 7 is a likely choice.

Figure 3 shows the complete baseband selectivity that defines the HDTV spectrum. Notice the additional signal for improved chrominance resolution (C′). Additional luminance resolution requires extra associated chrominance resolution and therefore additional chrominance bandwidth to appreciate the full improvement. The additional chrominance signal, C′, is time multiplexed between the Ih signal and the Qh signal components on alternating scan lines. However, unlike SECAM, the C′ signals are modulated as a single-sideband signal. The frequency of the carrier is selected to be a multiple of the horizontal frequency to minimize the crosstalk between this additional signal and the high-frequency luminance information, Y′. The purpose is to interleave the luminance (Y′) and chrominance (C′). The next section includes an explanation of the interleaving.

2.4 Detailed block diagrams

Figure 8 is a detailed block diagram of the encoder shown in Fig. 5. The method for increasing the vertical resolution mentioned previously is used,[4] so that a 1050-line progressive-scan source (camera) feeds an anti-aliasing filter. The R, G, B signals are all processed in the same fashion, each one is passed through a scan converter that converts from 1050 lines in a progressive-scan format to 525 lines in an interlaced format for compatible transmission. These signals are then matrixed into a Y, I, Q format.

The additional high-frequency signal for improved chrominance is time multiplexed to carry the Ih and Qh signals on alternate horizontal lines. However, it is not frequency modulated as in SECAM; rather, it is single-sideband amplitude modulated. This gives a spectrum that tends to cluster at multiples of the horizontal frequency and odd

---

* Yl will be completely free from any cross-luminance produced by the Q component of the color signal. If the I component is allowed the present specification of 1.5-MHz bandwidth, it can cause some cross-luminance. However, there is some question of whether that bandwidth should be transmitted since there are no consumer receivers that make use of this extra bandwidth; consequently, the extra bandwidth of the I signal can only cause cross-luminance problems in present NTSC receivers.

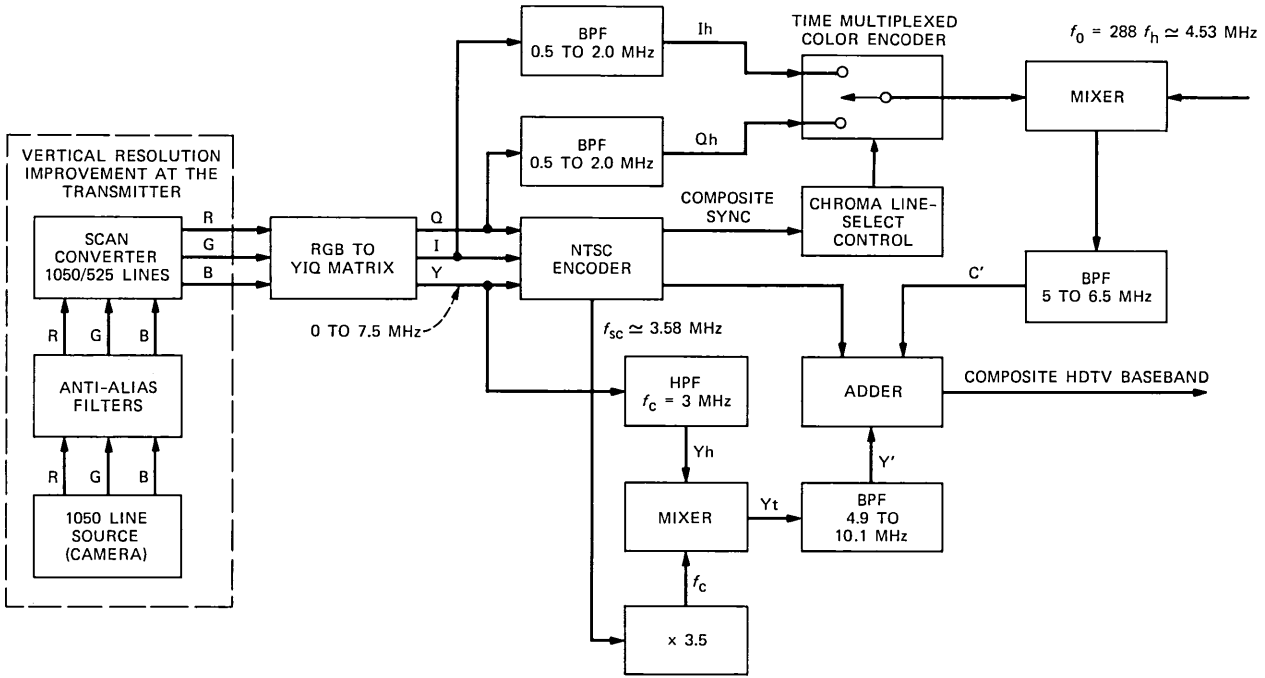


Fig. 8—Detailed block diagram of SLSC HDTV encoder.

multiples of half the horizontal frequency. The outputs of the two bandpass filters that select the Ih and Qh signals and the composite synchronization signal from the NTSC encoder feed the time-multiplexed color encoder (switch) that provides the additional signal for improved chrominance resolution, C′. The switch connects the Ih signal and then the Qh signal to the mixer on alternate scan lines. A carrier frequency, f_0, is the second input.*

$$f_0 = 288\, f_h \sim 4.53 \text{ MHz}. \tag{3}$$

A tone burst of frequency f_0 could be inserted into the vertical interval for phase reference at the receiver. The bandpass filter at the output of the modulator selects only the sum signal, C′, that is in the frequency range of 5 to 6.5 MHz.

The NTSC encoder functions as it would normally. It provides the composite synchronization signal to the line-select control for the high-frequency chrominance processing described above, and the color subcarrier for the high-frequency luminance processing.

The translated and frequency-inverted high-frequency luminance signal is formed by first high-pass filtering the 7.5-MHz luminance to produce Yh. When Yh is applied to the mixer illustrated in Fig. 8, a double-sideband suppressed carrier signal is created. The carrier input to the modulator is f_c.

$$f_c = 3.5\, f_{sc} = 3185(f_h/4) \sim 12.53 \text{ MHz}. \tag{4}$$

The NTSC color subcarrier, f_{sc}, is available from the NTSC encoder and can be used to derive f_c as indicated by eq. (4). However, a tone burst of f_c would be inserted into the vertical interval for phase reference at the receiver. The bandpass filter on the output only allows the frequency-inverted lower sideband, Y′, to pass. Y′ has a quarter-line frequency offset as can be seen from eq. (4). Consequently, f_0 is made equal to an exact multiple of f_h in eq. (3) so that the time-multiplexed signal C′ will interleave with Y′. This signal, Y′, is added to the output of the NTSC encoder and to the output of the high-frequency chrominance encoder, C′, to produce the composite HDTV signal of Fig. 3. Audio information is added just prior to transmission; this is the conventional manner of adding audio information to the NTSC broadcast signal. Additional subcarriers could be added to the baseband of the composite HDTV signal for multichannel sound or teletext as indicated in Fig. 3.

A detailed block diagram of the decoder is shown in Fig. 9. The decoder input comes from the video detector(s); more will be said

* This is just one of several strategies for selecting f_0. f_0 could be made 300 $F_h \sim 4.72$ MHz; then a low-level tone of this frequency can be inserted in the SLSC baseband.

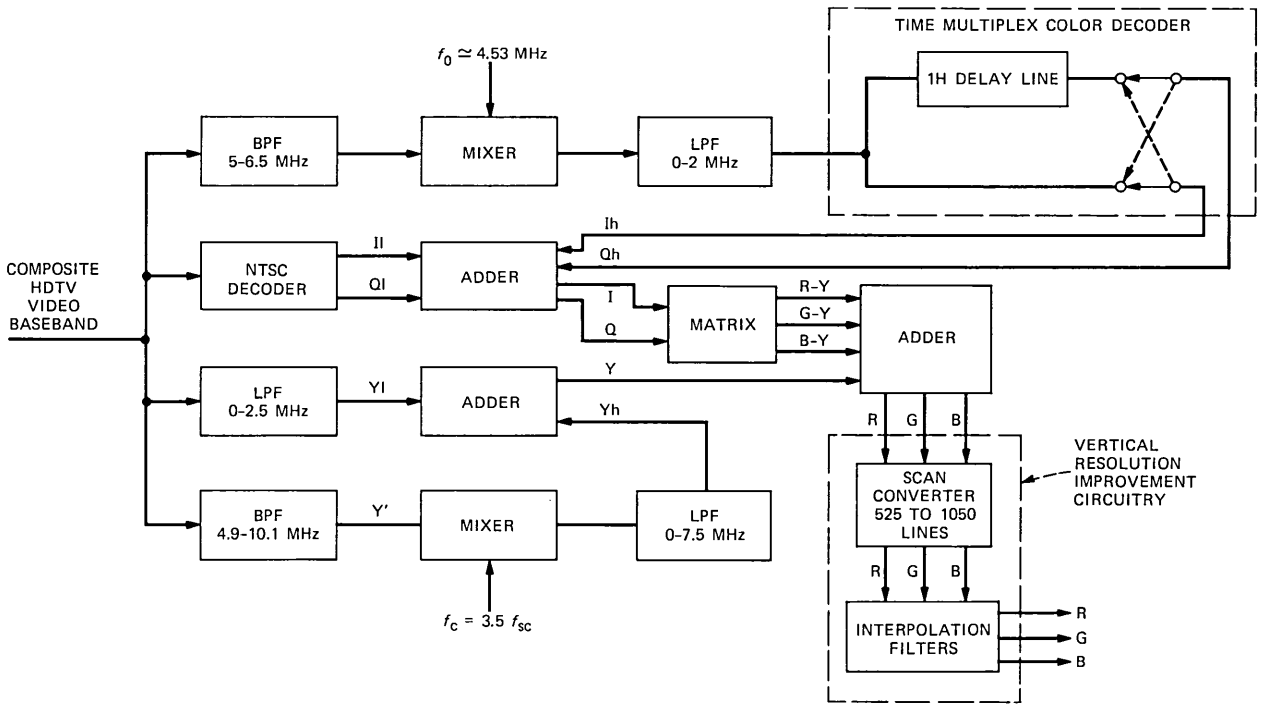


Fig. 9—Detailed block diagram of SLSC HDTV decoder.

about the receiver circuitry between the antenna and the decoder in Section 3.1. Since the composite HDTV signal was made up of three separate parts, each of these parts must be decoded. The high-frequency chrominance is decoded by first selecting it from the composite signal via the 5- to 6.5-MHz bandpass filter. This signal feeds a single-sideband demodulator that consists of a mixer and a bandpass filter that extracts the 0.5- to 2-MHz high-frequency color signal. However, this signal is still time multiplexed at the single-sideband demodulator output, so this output is fed into a time-multiplexed decoder to obtain simultaneous I and Q signals. The time-multiplexed decoder consists of a delay line that provides the storage of one horizontal line of color information, and the appropriate switches indicated in Fig. 9. The time-multiplexed decoder functions such that the present Ih (Qh) signal and the previous line Qh (Ih) signal are both present on the outputs providing the simultaneous Ih and Qh signals as outputs.

The NTSC decoder provides its normal I and Q color signals, called Il and Ql to emphasize that they convey the low-frequency portion (0 to 0.5 MHz) of the HDTV color signal. (Virtually all NTSC decoders produce I and Q signals of 0.5 MHz bandwidth in spite of the fact that the I channel should have extra bandwidth.) Il and Ql are added to Ih and Qh to form the complete 2-MHz I and Q signals. The high-frequency luminance, Y′, is passed by the 4.9- to 10.1-MHz bandpass filter and fed to the mixer. The carrier input is 3.5 f_{sc}, derived from the color subcarrier provided by the NTSC decoder with phase coherence obtained from a tone of f_c in the vertical interval. The demodulator output is filtered by a low-pass filter with a 7.5-MHz cutoff. The resultant signal, Yh′, which occupies a 2.5- to 7.5-MHz spectrum, is added to the 0- to 2.5-MHz low-frequency luminance signal obtained by passing the composite HDTV spectrum through a low-pass filter with a 2.5-MHz cutoff. The adder outputs the luminance signal, Y. Y, I, and Q are matrixed to provide color-difference signals and then added to Y to output R, G, B signals to the circuit that provides the vertical-resolution improvement. This circuit consists of a scan converter that inserts extra lines, converting the interlaced 525-line transmission standard to a progressive or interlaced 1050-line format along with interpolation filtering for each of the R, G, B signals. The resulting R, G, B signals feed circuitry that drives the picture display.

2.5 System alternatives

This system is flexible in that several variations are possible depending upon the needs and possibilities that surface during testing. The vertical resolution is limited to a maximum of 483 lines because the technique for vertical improvement can only promise a maximum resolution equal to the maximum number of active scan lines in the

transmission standard. The actual resolution could be slightly less than this because the anti-aliasing filter and the interpolation filter may reduce the vertical frequency response somewhat. The maximum horizontal resolution that corresponds to 7.5 MHz of bandwidth is 600 lines. This resolution can be achieved by using a comb filter at the receiver (covering the 5- to 6.5-MHz portion of the baseband signal shown in Fig. 3), or by not transmitting C′ and using the inferred highs processing to restore wideband chrominance signals at the receiver.[5] Both approaches result in less vertical resolution (483 lines) than horizontal resolution (600 lines).

Alternatively, a horizontal resolution equal to the vertical resolution could be chosen, thus simplifying the system. If 480 lines of horizontal resolution were chosen, only 6 MHz of luminance bandwidth is required. The high-frequency luminance of Fig. 3 need not overlap the high-frequency chrominance. Therefore, a comb filter is not needed to realize the 480 lines of horizontal resolution while using a C′ signal. An additional simplification is possible; the carrier, f_c, used to translate the high-frequency luminance spectrum in Fig. 8 can be three times the color subcarrier, 3 f_{sc}, rather than three and one half times that frequency.

Less extensive changes are possible in the interest of optimizing the system. For example, the 2.5-MHz cutoff of the low-pass filter that forms Y1 (illustrated in Figs. 6 and 9) could be reduced to 2 or even 1.5 MHz in order to completely eliminate any possible crosstalk between low-frequency chrominance and luminance. Alternatively, the cutoff could be increased to 3 MHz if it is shown that the extra resolution is a more important benefit than the penalty of a small amount of extra crosstalk. The 2.5-MHz cutoff and many other parameters given are reasonable choices that may change somewhat once the system is tested. Another possibility is to bandlimit the I channel in the NTSC encoder to 0.5 MHz since virtually no NTSC decoders use more than 0.5 MHz at this time and the extra I channel bandwidth can only cause crosstalk in NTSC consumer receivers. With equal-bandwidth (0.5 MHz) I and Q channels in the NTSC encoder, the 2.5-MHz cutoff illustrated in Fig. 7 will result in no cross luminance from Ih and Qh.

A change in the encoder and decoder (illustrated in Figs. 5, 6, 8, and 9) is possible to simplify the system by taking the output of the 1050-line source in Figs. 5 and 8 and transforming to a Y, I, Q signal format immediately. Then the I and Q signals could be bandlimited to 4 MHz (twice the 2-MHz transmission bandwidth) rather than 15 MHz for R, G, B signals (twice the 7.5-MHz Y bandwidth) for a source with a 30-Hz frame rate. Anti-aliasing and interpolation filter implementation in these channels is simplified because of the lower-frequency

operation. Framestore memory is also minimized because of the smaller bandwidth of I and Q.

III. DELIVERY SYSTEMS

3.1 AM broadcast

The front end of the HDTV set consists of a tuner, IF amplifier, and video detector. These functions can be handled in three basic ways, as illustrated in Fig. 10. Figures 10a and b represent the case where the two 6-MHz channels are adjacent to each other, as illustrated in Fig. 4.

Figure 10a illustrates a single wideband tuner and IF amplifier that is wide enough to pass the entire HDTV spectrum. The IF amplifier may be centered around the same IF frequency presently used, or a different center frequency may be chosen. Figure 10b uses a wideband tuner, but splits the spectrum up into two parts: an NTSC IF to receive that portion of the spectrum that is the same as the NTSC signal, paralleled by a second IF specially designed to receive the extra portion of the HDTV signal. After video detection in each channel, the two baseband spectra are added together to produce the baseband spectrum of Fig. 3.

A more versatile but complicated approach is illustrated in Fig. 10c; shown are two separate tuners for each portion of the spectrum, the NTSC portion and the extra portion. After processing by the respective IF amplifiers and detectors in each channel, the two portions of the baseband spectrum are added back together to obtain the spectrum of Fig. 3. If the arrangement of Fig. 10c is used, the extra 6-MHz spectrum for HDTV need not be located at the upper 6-MHz band shown in Fig. 4. It could be a totally separated channel of 6-MHz bandwidth. If the additional HDTV information is carried in the 6-MHz channel directly above the NTSC channel, 40.75 to 35.75 MHz is a likely IF frequency range for the additional IF. However, if the additional HDTV information is one or several channels away, another frequency may be chosen.

3.2 Cable

Modern CATV systems usually use a tree structure that has the capability of handling up to 50 or more NTSC TV channels per cable. Since present channel allocation is 6 MHz, any additional bandwidth requirement results in a minimum of two channels per HDTV signal. This two-to-one reduction in stations would be more acceptable to the CATV industry than HDTV systems requiring more than two 6-MHz channels per signal. It is almost certain that a higher-bandwidth system would cause the CATV industry large problems, since it would

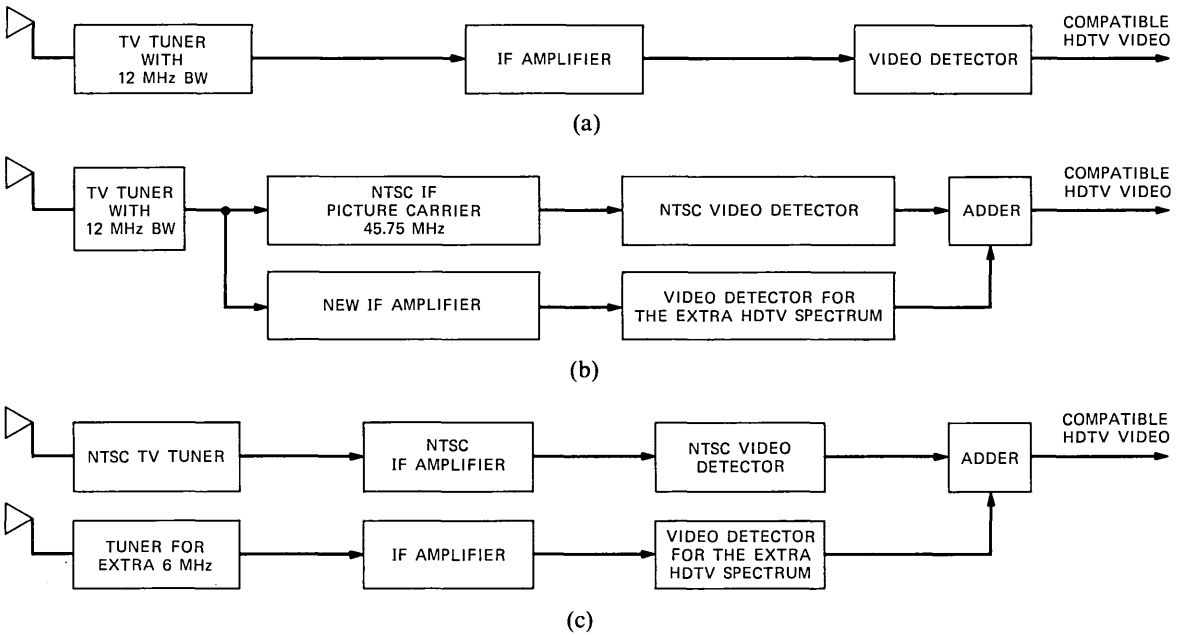


Fig. 10—Front-end alternatives for SLSC HDTV. (a) Single IF amplifier approach. (b) Single-tuner multiple IF amplifier approach. (c) Multiple-tuner approach.

severely limit the amount of programming that could be carried by a CATV system.

Switched video cable systems that have the ability to provide video on demand are just beginning to surface. It is very difficult to switch extremely wideband video; therefore, a reasonably compact HDTV spectrum is needed for this new type of cable system.

This compatible HDTV system has been carefully designed to occupy a 10-MHz baseband bandwidth. It will fully utilize two 6-MHz channels for transmission if desired. Therefore, it should have a minimum negative impact on present and future cable systems.

3.3 Satellite

Direct satellite broadcast to the home is a delivery system that is just about to become important. The modulation format and the frequency allocation of the service will eliminate compatibility with the present NTSC receivers. NTSC compatibility may lose some importance for this delivery system. However, bandwidth is always at a premium for satellite systems; thus the compact baseband as shown in Fig. 3 is very important. Also, the ability to modulate the HDTV signal into two NTSC channels could be advantageous.

3.4 Prerecorded

Even when the prerecorded media such as video tape and video disc are used as a delivery system, bandwidth is still important. It is much easier to record a 10-MHz baseband signal than a 30-MHz signal. Consequently, a tape recorder for the 10-MHz signal should be more economical. There may be a considerable problem with making a video disc unit that handles a 30-MHz baseband signal and still has sufficient playing time per disc.

IV. COMPARISONS

4.1 Vertical resolution comparisons

The vertical resolution, Rv, of a TV system is given by eq. (1). For the current NTSC system, the following is obtained:

$$Rv = (525 - 2x21)0.65 = 313 \text{ lines.} \tag{5}$$

This assumes a nominal Kell factor of 0.65. If the same Kell factor and 21-line vertical interval is applied to the NHK system, the result is 704 lines. For the Dortmund system using USA scanning standards and the proposed split-luminance, split-chrominance (SLSC) compatible system, the result should be the same as the NTSC with a Kell factor approaching unity.[4] Therefore, the vertical resolution should approach 483 lines. These comparisons together with modifications of the BBC and IBA systems for the USA scanning standards are shown

in Table I.[2,3] (Note that there are actually two Dortmund systems described in the literature.[4] The system considered here is the diagonal sampling approach that provides both increased vertical and horizontal resolution over the present standard. The other system does not provide increased horizontal resolution.)

4.2 Horizontal resolution comparisons

Table I also contains a comparison of some HDTV systems with respect to horizontal resolution, crosstalk, and bandwidth requirements adapted to the NTSC environment where appropriate. The NTSC system is used as a reference. The horizontal resolution of the NTSC system ranges from 240 lines to 336 lines. The NHK system should be capable of approximately 30 lines/MHz based on eq. (2). This predicts a horizontal resolution, Rh, of approximately:

$$Rh = (30 \text{ lines/MHz})(20 \text{ MHz}) = 600 \text{ lines}. \quad (6)$$

The Dortmund system approach applied to the NTSC system should have the ability to reproduce the same horizontal resolution as vertical. Further, the vertical resolution should approximate the number of active scan lines. Therefore, it should result in 480 lines.[4]

4.3 Crosstalk

Crosstalk is a very important factor in the image quality of a video system. There are three types of component crosstalk in present television standards. They are cross luminance (crosstalk of chrominance into luminance), cross color (luminance crosstalking into chrominance), and chrominance to chrominance crosstalk. Cross luminance, also called dot crawl, normally shows up in the NTSC picture at the edges of color areas. The high-frequency subcarrier signal appears in the luminance channel as a dot pattern crawling up the edges of color areas. Cross color is most obvious when the scene contains a detailed pattern such as a striped shirt or tweed suit. It is

Table I—HDTV comparison chart

System	*V* Resolution (k = 0.65)	*H* Resolution	Min. No. NTSC Channels When Broadcast	Compatible	Crosstalk
NTSC	313	240–336	1	Yes	Bad–Small
NHK	704	600	—	No	Small
Dortmund*	483	480	1	—	—
BBC*	313†	336	2	Can Be	None
IBA*	313†	336	2	No	None
SLSC	483	600	2	Yes	Small–None

* The values are adapted to the USA scanning standard.
† This value can be made 483 by applying the vertical improvement technique.

seen as a color pattern over the area of detail that obviously does not belong, and destroys the ability to see the detail in that area. Chrominance-to-chrominance crosstalk is less obvious than the other two because it is correlated with the scene. It results in color distortions.

The first two types of crosstalk mentioned above can produce significant picture degradation in the NTSC system unless the luminance bandwidth is sacrificed or a comb filter is used. However, a comb filter tends to produce some luminance degradation of its own. It reduces vertical resolution by averaging two or more lines at a time and rejects diagonal lines in the picture. This situation can be improved for luminance by only comb filtering the high-frequency luminance where the actual crosstalk frequency components occur. Further, a comb filter will produce a dot crawl on horizontal edges of saturated colors in a picture that is not present without it.

As shown in Fig. 7, the low-frequency luminance (Yl) used by the SLSC HDTV system is the lower 2.5 MHz of the NTSC luminance. This portion of the NTSC luminance is not interleaved with any of the Q component of the NTSC color signal. There is some interleaving of the I component of the NTSC signal and the low-frequency luminance, Yl, of the HDTV signal. However, there is some question of whether the full frequency range of the I signal should be transmitted for NTSC receivers since there are no consumer receivers that make use of it, and it can only cause crosstalk in the present NTSC receivers.

Cross luminance in the high-frequency luminance (Yh) can also be minimized. The upper 1.5 MHz of the luminance—the 6-MHz to 7.5-MHz region of the original luminance signal—interleaves with the high-frequency chrominance. This upper end of the luminance ends up occupying the 5- to 6.5-MHz region of the frequency inverted and translated spectrum Y′ in Fig. 3. This region can be comb filtered—rolled off at 6 MHz (480 lines of resolution)—or simply allowed to pass with the small amount of cross luminance mentioned above.

The last alternative is possible with only a small amount of degradation because of the high-frequency nature of the cross luminance and the fact that the chrominance that is talking into the luminance will be down in level compared to the signal that is producing cross luminance in the NTSC system (it is only the 0.5- to 2-MHz region of the chrominance that will contribute to crosstalk here). Since this relatively low-level chrominance will also be producing a much smaller dot pattern in the luminance than the NTSC color subcarrier would cause in the NTSC system, the crosstalk should be much less obvious.

Cross color can be reduced in several ways. A comb filter could be used to remove that portion of the luminance that interferes with the low-frequency chrominance. Alternatively, use can be made of the fact that a portion of the luminance that represents 2.5 to 4.2 MHz of the

NTSC luminance is repeated in the high-frequency luminance portions of the composite HDTV signal. This portion of the high-frequency luminance is free of any interleaved chrominance. However, the corresponding portion of the NTSC part of the signal contains chrominance interleaved with the luminance. These two corresponding portions of luminance can be subtracted, leaving an uncontaminated chrominance signal. Also, the luminance above 3 MHz could be rolled off in the NTSC part of the baseband as another way to eliminate cross color without any degradation to the vast majority of NTSC receivers that do not use a comb filter. Cross color into C′ can be reduced with a comb filter.

4.4 Encoding errors

There are a number of color deficiencies in the NTSC color standard that degrade the quality of the reproduced image.[6-8] The crosstalk problems mentioned in the previous section are a part of these deficiencies. Encoding errors are another facet of these problems. The study of encoding errors or distortions in the NTSC system is an involved topic that will only be briefly mentioned here. The problems center around the fact that part of the luminance is carried by the chrominance when color is broadcasted, and it is aggravated by the nonlinear gamma characteristics of the system.[6] The transmitter corrects for a nominal gamma of 2.2 by raising the signal to the 1/2.2 power so that the overall response is linear. The result is that saturated (vivid) colors lose details. Also, there may be substantial errors in the transients between certain colors (usually complementary colors produce the worst transients). With wideband color signals, there will be fewer errors. The luminance is still carried partly by the chrominance, but the chrominance is now wideband and it degrades the end result less.

V. SUMMARY

A split-luminance, split-chrominance (SLSC) HDTV system has been described that is NTSC compatible and uses a 10-MHz baseband signal. This baseband signal can be modulated to produce an amplitude-modulated, vestigial-sideband signal in a 12-MHz bandwidth for broadcast. Alternatively, this compatible signal can also occupy two separate 6-MHz channels. Compatibility and bandwidth conservation are two of the most important attributes of this HDTV system. A compatible system is likely to penetrate the market place much more rapidly than a non-compatible system because of the huge investment in NTSC equipment. The two most common delivery systems are broadcast and CATV; both are very sensitive to the bandwidth requirements of a new system.

The present NTSC channels are 6 MHz. Conventional channel allocation forces additional bandwidth to come in increments of 6 MHz. A two-channel, 12-MHz requirement is the largest bandwidth that these systems can reasonably accommodate for this new HDTV service. Therefore, this new compatible system has been designed to occupy a 12-MHz bandwidth when using the present broadcast format of vestigial-sideband amplitude modulation.

The important parameters of this new compatible HDTV system are a horizontal resolution potential up to 600 lines, a vertical resolution potential up to 483 lines, and less crosstalk between the individual components of the color signal compared to the NTSC system. Further minimization of color-encoding errors may be possible by using recently developed processing techniques.[5]

VI. ACKNOWLEDGMENTS

The author gratefully acknowledges Joseph LoCicero for many useful discussions that aided his understanding of much of the HDTV area. The author expresses his gratitude to Bruce Briley, Kristin Kocan, and Harvey Lehman for their review of this paper.

REFERENCES

1. T. Fujio, "High-Definition Wide-Screen Television System for the Future-Present State of the Study of HD-TV Systems in Japan," IEEE Trans. Broadcasting, *BC-26* (December 1980), pp. 113–24.
2. C. Sandbank and M. Moffat, "High-Definition Television and Compatibility with Existing Standards," *Tomorrow's Television*, Scarsdale, New York: SMPTE, 1982, pp. 170–85.
3. T. Robson, "A Compatible High Fidelity TV Standard for Satellite Broadcasting," *Tomorrow's Television*, Scarsdale, New York: SMPTE, 1982, pp. 218–36.
4. B. Wendland, "High Definition Television Studies on Compatible Basis with Present Standards," *Television Technology in the 80's*, Scarsdale, New York: SMPTE, 1981, pp. 151–65.
5. D. Richman, "Color Television Receiving System Utilizing Inferred High Frequency Signal Components to Reduce Color Infidelities in Regions of Color Transitions," U.S. Patent 4,181,917, Jan. 1, 1980; "Color Television Receiving System Utilizing Multimode Inferred Highs Correction to Reduce Color Infidelities," U.S. Patent 4,183,051, Jan. 8, 1980; and "Color Television Receiving System Utilizing Inferred High Frequency Signal Components to Reduce Color Infidelities in Regions of High Color Saturation." U.S. Patent 4,245,239, Jan. 13, 1981.
6. K. McIlain and C. Dean, *Principles of Color Television*, New York: John Wiley, 1956, Chapter 11.
7. W. Gibson and A. Schroeder, "Color Television Luminance Detail Rendition," Proc. I.R.E., *43* (August 1955), pp. 918–23.
8. D. Livingston, "Reproduction of Luminance Detail by NTSC Color Television Systems," Proc. I.R.E., *42* (January 1954), pp. 228–34.

AUTHOR

Theodore S. Rzeszewski, B.S.E.E., 1964, M.S.E.E., 1965, University of Illinois; E.E., 1971, D.E. (honorary), 1973, Midwest College of Engineering; Bell Laboratories, 1982—. Mr. Rzeszewski is a member of the Exploratory Wideband Telecommunications Department. His current areas of interest include high-definition television and digital television. Before joining Bell Laboratories, he was Manager of Technical Operations of Matsushita Applied Research.

LETTER TO THE EDITOR

Comments on "Three-Stage Multiconnection Networks Which Are Nonblocking in the Wide Sense," by F. K. Hwang*

Two theorems presented in this paper are incorrect. Theorem 2 stated by Hwang can be reformulated as follows.

Theorem 2: $v(m, n_1, r_1, n_2, r_2)$ *is nonblocking as a* (q_1, q_2) *multiconnection network under Strategy 2, for* $r_1 \geq q_1q_2n_1$ *and* $r_2 \geq q_1q_2n_2$, *if and only if*

$$m \geq q_1q_2(n_2 - 1) + q_1 + n_1 - 1.$$

Proof: Sufficiency. Consider the connection of the pair (x, Y). The input switch that contains x can be connected already to at most $n_1 - 1$ distinct middle switches under Strategy 2. Each output switch in Y can be connected already to at most $n_2q_1 - 1$ distinct middle switches under Strategy 2. Since $|Y| \leq q_2$, we need q_2 sets of $n_2q_1 - 1$ distinct middle switches, if the sets are disjoint. However, under Strategy 2 these sets are not disjoint and the number of middle switches must be replaced by

$$q_2(n_2q_1 - 1) - (q_1 - 1)(q_2 - 1).$$

Then the total number of middle switches, including one switch that must be available to connect the pair (x, Y), is

$$(n_1 - 1) + q_2(n_2q_1 - 1) - (q_1 - 1)(q_2 - 1) + 1.$$

After rearrangement we obtain

$$m \geq q_1q_2(n_2 - 1) + q_1 + n_1 - 1.$$

The necessity can be proved with ease by presenting the network with

$$m < q_1q_2(n_2 - 1) + q_1 + n_1 - 1,$$

in which a new call is blocked.

Similarly, Theorem 3 can be reformulated as follows.

Theorem 3: $v(m, n_1, r_1, n_2, r_2)$ *is nonblocking as a* (q_1, q_2) *multiconnec-*

* B.S.T.J., *58*, No. 10 (December 1979), pp. 2183–87.

tion network under Strategy 3, *for* $r_1 \geqslant q_1q_2n_1$ *and* $r_2 \geqslant q_1q_2n_2$, *if and only if*

$$m \geqslant q_1q_2(n_1 - 1) + q_2 + n_2 - 1.$$

Proof: Analogous to the proof of Theorem 2.

Andrzej Jajszczyk
Technical University of Poznań
Institute of Electronics
ul. Piotrowo 3a, 60-965 Poznań
Poland

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Barnes E. S., Sloane N. J. A., **New Lattice Packings of Spheres.** Can J Math 34(1):117–130, 1983.

Barnes E. S., Sloane N. J. A., **The Optimal Lattice Quantizer in Three Dimensions.** Siam J Alg 4(1):30–41, 1983.

Bazelow A. R., Raamot J., **On the Microprocessor Solution of Ordinary Differential-Equations Using Integer Arithmetic (Letter).** IEEE Comput 32(2):204–207, 1983.

Buckingham M. A., Golumbic M. C., **Partitionable Graphs, Circle Graphs, and the Berge Strong Perfect Graph Conjecture.** Discr Math 44(1):45–54, 1983.

Fleckenstein W. O., **Challenges in Software-Development.** Computer 16(3):60–64, 1983.

Garey M. R., Johnson D. S., Tarjan R. E., Yannakakis, M., **Scheduling Opposing Forests.** Siam J Alg 4(1):72–93, 1983.

Grambsch P., **Sequential Sampling Based on the Observed Fisher Information to Guarantee the Accuracy of the Maximum-Likelihood Estimator.** Ann Statist 11(1):68–77, 1983.

Hwang F. K., **Fast Solutions for Consecutive-K-OUT-OF-N-F-System.** IEEE Reliab 31(5):447–448, 1982.

Lin G. K., Menon P. R., **Totally Preset Checking Experiments for Sequential-Machines.** IEEE Comput 32(2):101–108, 1983.

Logan B. F., **An Integral-Equation Connected with the Jacobi-Polynomials.** Siam J Math 14(2):269–322, 1983.

Logan B. F., **Extremal Problems for Positive-Definite Bandlimited Functions. 1. Eventually Positive Functions with Zero Integral.** Siam J Math 14(2):249–252, 1983.

Logan B. F., **Extremal Problems for Positive-Definite Bandlimited Functions. 2. Eventually Negative Functions.** Siam J Math 14(2):253–257, 1983.

Logan B. F., **Hilbert Transform of a Function Having a Bounded Integral and a Bounded Derivative.** Siam J Math 14(2):247–248, 1983.

Mallows C. L., **Minimax Aspects of Bounded-Influence Regression Comment.** J Am Stat A 78(381):77, 1983.

Stroustrup B., **Adding Classes to the C-Language—An Exercise in Language Evolution.** Software 13(2):139–161, 1983.

ENGINEERING

Alley G. D., **High-Voltage Two-Dimensional Simulations of Permeable Base Transistors.** IEEE Device 30(1):52–60, 1983.

Azbel M. Y., **Resonance Tunneling and Localization Spectroscopy.** Sol St Comm 45(7):527–530, 1983.

Capasso F., **Avalanche Photo-Diodes with Enhanced Ionization Rates Ratio—Towards a Solid-State Photomultiplier.** IEEE Nucl S 30(1):424–428, 1983.

Chemla D. S., Maruani A., **Non-Linear Optical Effects Associated with Excitonic-Molecules in Large Gap Semiconductors (Review or Bibliog.).** Prog Q Elec 8(1):1–77, 1982.

Joy D. C., **Practical Quantitation for Energy-Loss Spectra.** Scan Elec M1982(P2):505–515, 1982.

Levinson M., Temkin H., Bonner W. A., **Electron-Bombardment Induced Defect States in P-InP.** J Elec Mat 12(2):423–432, 1983.

Meyer W. G., Fair R. B., **Dynamic Behavior of the Buildup of Fixed Charge and Interface States During Hot-Carrier Injection in Encapsulated MOSFETs.** IEEE Device 30(2):96–103, 1983.

Salem A. A. M., Cox D. C., **Improving the Power-Added Efficiency of FET Amplifiers Operating with Varying-Envelope Signals.** IEEE Micr T 31(1):51–56, 1983.

Shang H. T., Stone J., Burrus C. A., **Low-OH MCVD Fibers Without a Barrier Layer Using OH→OD-Exchanged Substrate Tubes.** Electr Lett 19(3):95–96, 1983.

MANAGEMENT/ECONOMICS

Gordon R. H., **Interest-Rates, Inflation, and Corporate Financial Policy.** Brookings P1982(2):461–488, 1982.

PHYSICAL SCIENCES

Abrahams S. C., et al., **Detwinning Cell for High-Transition-Temperature Atmosphere-Sensitive Ferroelastic Crystals.** J Appl Crys 16(FEB):96–98, 1983.

Alavi K., Temkin H., Wagner W. R., Cho A. Y., **Optically Pumped 1.55-μm Double Heterostructure $Ga_xAl_yIn_{1-x-y}As/Al_uIn_{1-u}As$ Lasers Grown by Molecular-Beam Epitaxy.** Appl Phys L 42(3):254–256, 983.

Alpern M., Kitahara K., Krantz D. H., **Perception of Color in Unilateral Tritanopia.** J Physl Lon 335(FEB):683–697, 1983.

Bally J., Predmore R., **Ionized-Gas in Active Molecular Cloud Cores.** Astrophys J 265(2):778–790, 1983.

Bally J., Lada C. J., **The High-Velocity Molecular Flows Near Young Stellar Objects.** Astrophys J 265(2):824–847, 1983.

Baraff G. A., Schluter M., Allan G., **Calculation of Si and Al Interstitials in Silicon Using the Cluster-Extended Greens-Function Technique.** Physica B&C 116(1–3):76–78, 1983.

Baraff G. A., Schluter M., Allan G., **Theory of Enhanced Migration of Interstitial Aluminum in Silicon.** Phys Rev L 50(10):739–742, 1983.

Bayman A., Hansma P. K., Kaska W. C., Dubois L. H., **Inelastic Electron-Tunneling Spectroscopic Study of Acetylene Chemisorbed on Alumina Supported Palladium Particles.** Appl Surf S 14(2):194–208, 1983.

Bean J. C., Mcafee S. R., **Silicon Molecular-Beam Epitaxy—A Comprehensive Bibliography 1962–82.** J Physique 43(NC-5):153–172, 1982.

Beck S. M., Brus L. E., **Transient Spontaneous Raman Study of Photo-Ionization Kinetics at the Hydrocarbon Water Interface in Micellar Solutions.** J Am Chem S 105(5):1106–1111, 1983.

Benton J. L., Kimerling L. C., Stavola M., **The Oxygen Related Donor Effect in Silicon.** Physica B&C 116(1–3):271–275, 1983.

Besomi P., Wilson R. B., Wagner W. R., Nelson R. J., **Enhanced Indium-Phosphide Substrate Protection for Liquid-Phase Epitaxy Growth of Indium-Gallium-Arsenide-Phosphide Double Heterostructure Lasers.** J Appl Phys 54(2):535–539, 1983.

Bridges T. J., Patel C. K. N., Strnad A. R., Wood O. R., Brewer E. S., Karlin D. B., **Syneresis of Vitreous by Carbon-Dioxide Laser-Radiation.** Science 219(4589:1217–1219, 1983.

Cooper J. A., Nelson D. F., **High-Field Drift Velocity of Electrons at the $SiSiO_2$ Interface as Determined by a Time-of-Flight Technique.** J Appl Phys 54(3):1445–1456, 1983.

Crane P., Tyson J. A., Saslaw W. C., **Optical-Emission in the Radio Lobes of Radio Galaxies. 2. New Observations of 21 Radio Lobes.** Astrophys J 265(2):681+, 1983.

Cross M. C., Daniels P. G., Hohenber P. C., Siggia E. D., **Phase-Winding Solutions in a Finite Container Above the Convective Threshold.** J Fluid Mec 127(Feb):155–183, 1983.

Cummins S. E., Green S., Thaddeus P., Linke R. A., **The Kinetic Temperature and Density of the Sagittarius-B2 Molecular Cloud From Observations of Methyl Cyanide.** Astrophys J 266(1):331–338, 1983.

Dutta N. K., **Calculation of Auger Rates in a Quantum Well Structure and its Application to InGaAsP Quantum Well Lasers.** J Appl Phys 54(3):1236–1245, 1983.

Edgerton M. E., Handler F. A., Wu C. W., **Methods for the Analysis of the Kinetics of Cooperative Binding of Ligands to a Two-Site Lattice.** Biopolymers 22(2):787–805, 1983.

Eibschutz M., Salomon D., Disalvo F. J., **Mossbauer Studies of the 6.2 keV Gamma-Rays of ^{181}Ta in Ta-Dichalcogenides.** Phys Lett A 93(5):259–261, 1983.

Eilenberger D. J., Isaacs E. D., Aumiller G. D., **Near-Infrared, Tunable, Oxazine-750 Perchlorate, Synchronously-Pumped Picosecond Ring Dye-Laser.** Opt Commun 44(5):350–352, 1983.

Furukawa T., Date M., Johnson G. E., **Polarization Reversal Associated with Rotation of Chain Molecules in Beta-Phase Polyvinylidene Fluoride.** J Appl Phys 54(3):1540–1546, 1983.

Gallagher P. K., Sinclair W. R., **Oxidation of Polycrystalline Niobium Nitride.** ISR J Chem 22(3):222–226, 1982.

Gurvitch M., Washington M. A., Huggins H. A., **High-Quality Refractory Josephson Tunnel-Junctions Utilizing Thin Aluminum Layers.** Appl Phys L 42(5):472–474, 1983.

Hasegawa A., Wakatani M., **Plasma Edge Turbulence.** Phys Rev L 50(9):682–686, 1983.

Heaven M., Dimauro L., Miller T. A., **Laser-Induced Fluorescence-Spectra of Free-Jet Cooled Organic Free-Radicals—Vinoxy, Cyclopentadienyl, and Benzyl.** Chem p Lett 95(4–5):347–351, 1983.

Helfand E., **On Inversion of the Williams-Watts Function for Large Relaxation-Times.** J Chem Phys 78(4):1931–1934, 1983.

Hertel G., Bishop D. J., Spencer E. G., Rowell J. M., Dynes R. C., **Tunneling and Transport Measurements at the Metal-Insulator-Transition of Amorphous Nb-Si.** Phys Rev L 50(10):743–746, 1983.

Hutson A. R., **Charge Transport by Dislocations in II-VI Compounds.** Physica B&C 116(1–3):650–656, 1983.

Julian M. M., Stillinger F. H., Festa R. R., **The 3rd Law of Thermodynamics and the Residual Entropy of Ice: "Stillwater" or $\Delta S^{H_2O}_{f,T=0} = 0$.** J Chem Educ 60(1):65–66, 1983.

Khanarian G., Schilling F. C., Cais R. E., **Kerr Effect and Dielectric Study of Polyvinyl-Chloride and its Oligomers.** Macromolec 16(2):287–291, 1983.

Kim S., et al., **Resonance Raman-Spectra and Electronic-Structure of the μ-Disulfur-Bridged Diruthenium Complex $[(H_3N)_5RuSSRu(NH_3)_5]^{4+}$ and Related Systems.** J Am Chem S 105(3):336–339, 1983.

Kimerling L. C., **Defects in Semiconductors—1982 (Editorial).** Physica B&C 116(1–3):1–3, 1983.

Kimerling L. C., Benton J. L., **Electronically Controlled Reactions of Interstitial Iron in Silicon.** Physica B&C 116(1–3):297–300, 1983.

Lamola A. A., et al., **Spectral Study of the Photochemistry of Dipyrrole Models for Bilirubin Bound to Human-Serum Albumin.** Photochem P 37(3):263–270, 1983.

Le H., Rosenbluh M., Lax B., Miller T. A., **Observation of ^{1}P-^{1}D, ^{1}P-^{1,3}F, and ^{1}P-^{1,3}H Motional Stark-Effect-Induced Anticrossings in ^{4}He: Determination of Zero-Field Level Separations.** Phys Rev A 27(2):895–905, 1983.

Lin W., Hill D. W., **Oxygen Segregation in Czochralski Silicon Growth.** J Appl Phys 54(2):1082–1085, 1983.

Lines M. E., Eibschutz M., **Correlation-Effects in the Mossbauer-Spectra of Amorphous Metallic Magnetic-Materials.** Sol St Comm 45(5):435–439, 1983.

Macrander A. T., Johnston W. D., **Deep Levels in InGaAsP/InP p+n Diodes Grown by Vapor-Phase Epitaxy.** J Appl Phys 54(2):806–813, 1983.

Manzione L. T., Osinski J. S., **Predicting Reactive Fluid-Flow in RIM and RTM Systems.** Mod Plast 60(2):56–57, 1983.

Mills A. P., **Measurement of the Decay-Rate of the Positronium Negative-Ion.** Phys Rev L 50(9):671–674, 1983.

Muhlhausen C. W., Serri J. A., Tully J. C., Becker G. E., Cardillo M. J., **Scattering of N-2 From Ag(001).** ISR J Chem 22(4):315–320, 1982.

Pearsall T. P., Carles R., Portal J. C., **Single Longitudinal-Mode Optical Phonon-Scattering in $Ga_{0.47}In_{0.53}As$.** Appl Phys L 42(5):436–438, 1983.

Pearsall T. P., Eaves L., Portal J. C., **Photo-Luminescence and Impurity Concentration in $Ga_xIn_{1-x}As_yP_{1-y}$ Alloys Lattice-Matched to InP.** J Appl Phys 54(2):1037–1047, 1983.

Pelleg J., Murarka S. P., **Phosphorus Distribution in $TaSi_2$ Films by Diffusion From a Polycrystalline Silicon Layer.** J Appl Phys 54(3):1337–1345, 1983.

Shank C. V., Greene B. I., **Femtosecond Spectroscopy and Chemistry.** J Phys Chem 87(5):732–734, 1983.

Silfvast W. T., Wood O. R., **Gain Scaling of Short-Wavelength Plasma-Recombination Lasers.** Optics Lett 8(3):169–171, 1983.

Sternheim M., Vangelder W., Hartman A. W., **A Laser Interferometer System to Monitor Dry Etching of Patterned Silicon.** J Elchem So 130(3):655–658, 1983.

Stolen R. H., Mollenauer L. F., Tomlinson W. J., **Observation of Pulse Restoration at the Soliton Period in Optical Fibers.** Optics Lett 8(3):186–188, 1983.

Sturge M. D., Cohen E., Logan R. A., **Dynamics of Intrinsic and Nitrogen-Induced Exciton Emission in Indirect-Gap $Ga_{1-x}Al_xAs$.** Phys Rev B 27(4):2362–2373, 1983.

Tewksbury S. K., **Formation of High-Density Electron-Hole Plasma in Silicon Metal-Oxide-Semiconductor Transistors Below 25-Degrees-K.** J Appl Phys 54(2):868–875, 1983.

Thompson M. O., Mayer J. W., Cullis A. G., Webber H. C., Chew N. G., Poate J. M., Jacobson D. C., **Silicon Melt, Regrowth, and Amorphization Velocities During Pulsed Laser Irradiation.** Phys Rev L 50(12):896–899, 1983.

Wagner W. R., Schumaker N. E., Zilku J. L., Anthony P. J., Swaminathan V., **Lattice-Parameter Changes in $Al_{0.39}Ga_{0.61}As$ Due to O, Ge, Si, and S Doping.** J Elchem So 130(3):670–675, 1983.

Watson E. A., Gibbs H. M., Mattar F. P., Cormier M., Claude Y., Mccall S. L., Feld M. S., **Quantum Fluctuations and Transverse Effects in Superfluorescence.** Phys Rev A 27(3):1427–1434, 1983.

Wood T. H., White J. C., Thacker B. A., **Ultraviolet Photo-Decomposition for Metal-Deposition—Gas Versus Surface Phase Processes.** Appl Phys L 42(5):408–410, 1983.

Yan M. F., Cannon R. M., Bowen H. K., **Space-Charge Distributions Near Interfaces During Kinetic Processes.** J Appl Phys 54(2):779–791, 1983.

Yan M. F., Cannon R. M., Bowen H. K., **Space-Charge, Elastic Field, and Dipole Contributions to Equilibrium Solute Segregation at Interfaces.** J Appl Phys 54(2):764–778, 1983.

SOCIAL AND LIFE SCIENCES

Breitmeyer B. G., Kropfl W., Julesz B., **The Existence and Role of Retinotopic and Spatiotopic Forms of Visual Persistence.** Act Psychol 52(3):175–196, 1982.

Eisinger J., Flores J., **Cytosol-Membrane Interface of Human-Erythrocytes—A Resonance Energy-Transfer Study.** Biophys J 41(3):367–379, 1983.

Graedel T. E., Mcrae J. E., **Total Organic-Component Data—A Study of Urban Atmospheric Patterns and Trends—Reply (Discussion).** Atmos Envir 17(3):670–671, 1983.

Nelson W. L., **Physical Principles for Economies of Skilled Movements.** Biol Cybern 46(2):135–147, 1983.

Reed A. V., **Parallel Models of Associative Memory—Hinton, G. E., Anderson, J. A. (Book Review).** Cont Psycho 28(2):137–139, 1983.

Smith E.R., Miller F. D., **Mediation Among Attributional Inferences and Comprehension Processes—Initial Findings and a General-Method.** J Pers Soc 44(3):492–505, 1983.

SPEECH/ACOUSTICS

Sondhi M. M., Resnick J. R., **The Inverse Problem for the Vocal-Tract Numerical-Methods, Acoustical Experiments, and Speech Synthesis.** J Acoust So 73(3):985–1002, 1983.

CONTENTS, OCTOBER 1983

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/ Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System