AT&T
BELL LABORATORIES

TECHNICAL JOURNAL

October 1984
Vol. 63 No. 8 Part 1

A JOURNAL OF THE AT&T COMPANIES

QAM Binary Signaling

Digital Radio

Speech Processing

Packet/Circuit Switching

Traffic

AT&T BELL LABORATORIES TECHNICAL JOURNAL (ISSN0005-8580) is published ten times each year by AT&T, 550 Madison Avenue, New York, NY 10022; C. L. Brown, Chairman of the Board; T. O. Davis, Secretary. The Computing Science and Systems section and the special issues are included as they become available. Subscriptions: United States—1 year $35; 2 years $63; 3 years $84; foreign—1 year $45; 2 years $73; 3 years $94. 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 Technical Journal and sent to AT&T Bell Laboratories, Circulation Dept., Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, AT&T Bell Laboratories Technical Journal, Room 1H321, 101 J. F. Kennedy Pky, Short Hills, NJ 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 AT&T. 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, NJ 07078 and additional mailing offices. Postmaster: Send address changes to the AT&T Bell Laboratories Technical Journal, Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

AT&T Bell Laboratories

Technical Journal

VOL. 63 OCTOBER 1984 NO. 8, PART 1

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

Contrasting Performance of Faster Binary Signaling With QAM

By G. J. FOSCHINI*

(Manuscript received November 15, 1983)

In this paper we determine the performance of Faster Binary Signaling (FBS), an alternative method to Quadrature Amplitude Modulation (QAM) for achieving a high bit rate over an ideal, bandlimited, noisy channel. With this method, signaling is faster than the Nyquist rate. Consequently, there are fewer points in the signal constellation, resulting in a greater separation of the points when the average transmitter power is the same as for QAM. Thus, at the expense of introducing Intersymbol Interference (ISI), there is an apparent improvement in noise immunity. The ISI can be mitigated with maximum likelihood sequence detection. We explore the advisability of trading freedom from ISI for added noise immunity for the extreme case where the system with faster signaling uses a four-point constellation. The question of the efficacy of FBS has been difficult to approach, but FBS has loomed as a possibly strong competitor among alternatives to QAM. We show here how to analyze FBS, and we give examples involving FBS operating at up to five times the QAM rate. In the examples, FBS is revealed to be, at best, of marginal value even if one allows for implementation capabilities far beyond those of forthcoming processors.

I. INTRODUCTION

For data communication over channels such as voiceband analog telephone circuits, satellite links, or terrestrial digital radio hops, there is a search for practical techniques that are more efficient (in bits per cycle) than QAM. This search is intensifying because of the growing

* AT&T Bell Laboratories.

demand for data communication services over bandlimited channels and because of the continuing drop in the cost of high-speed processing required by advanced communication methods.

The bandlimited channel with additive white Gaussian noise (power $\mathcal{N}$), where the average transmitter power $\mathcal{P}$ ($\mathcal{P} \gg \mathcal{N}$) is constrained, serves as a proving ground for theoretical explorations of the relative efficacy of proposed techniques. Specific methods have been discussed as candidates for improving efficiency. Three candidates are Higher-Dimensional Constellations (HDCs),[1] Ungerbock's Trellis Coding (UTC),[2] and Faster Binary Signaling (FBS).[3] Both HDC and UTC have lent themselves to analysis and their significant value over the QAM method has been established. Moreover, we are beginning to understand the relative value of UTC over HDC.[4] On the other hand, the effectiveness of FBS has hitherto remained a mystery. In Ref. 5* some theoretical results on FBS for some special pulses are presented but the relative effectiveness issue is not settled.

To understand the FBS method, consider the elementary QAM method, 4-PSK (phase-shift keying), in a situation comfortably meeting a stringent probability of error (P_e) constraint. If the bit rate requirement increases it can be met without expanding bandwidth by increasing the number of points in the QAM constellation. FBS is a natural alternative means of transmitting at higher bit rates. In FBS, one fixes the constellation at four points and increases the symbol rate as much as necessary. Maximum Likelihood Sequence Detection (MLSD) is employed to overcome the consequent Intersymbol Interference (ISI) in the best way possible (see Refs. 6 through 8 for treatments of MLSD). The minimum separation between distinct points in the planar FBS constellation is greater than for the QAM constellation. One might say FBS trades freedom from ISI for added noise immunity and then MLSD is used to mitigate the ISI.

With the efficacy of FBS unknown, it looms as a possibly competitive technique. Here we help the process of evaluating the field of candidate methods for moving beyond the capabilities of QAM by showing how to analyze FBS. We show by examples that FBS is, at best, of marginal value relative to QAM, even if one allows for implementation capabilities far beyond those of forthcoming processors. Specifically, we allow for complexity of up to 10^8 states. Given the strides in processor technology in the last few decades and the imminent hardware advances, a number like 10^8 is chosen to avoid outdating of this paper for a long time. Such prudence is needed. Indeed, the analysis that follows leaves open the possibility that FBS

* The terminology "faster binary signaling" is not used in [A] and [M]. In [M] the term "faster than Nyquist signaling" is used.

would offer substantial improvement over QAM if complexity were not a consideration. Moreover, one must also consider that fast detectors could go far beyond conventional MLSD in processing efficiency.[9]

II. SYSTEM DESCRIPTION

2.1 Transmission medium and its use

The data transmission medium is represented here in the simplest idealized form as a lossless characteristic with additive white Gaussian noise. The channelization is shown in Fig. 1. On each channel the transmitted signal is subject to an average power constraint and it is assumed that, for all the systems that we discuss, the average transmitted power is much greater than the noise power. [The ramification of this high signal-to-noise ratio (s/n) assumption is discussed in Section III.] In the analysis that follows, we assume that the channels are isolated from each other in that it is not permitted to mitigate Adjacent Channel Interference (ACI) through some elaborate scheme requiring coordination among channels. It is required that R bits per second be transmitted over the channel. Whatever the form of modulation used, soft maximum likelihood sequence detection is employed at the receiver.[10]

2.2 Comments about the benchmark QAM method

It is because of the current prominence of QAM in applications that the work here is presented with QAM as the benchmark method. Since a flat channel transfer characteristic is assumed, it is trivial to relate results to the equivalent baseband channel representing a QAM rail. We use $M = m^2$ to denote the number of points in the QAM constellation. So, constellations with 16, 64, 256, and 1024 points correspond to FBS operating at 2, 3, 4, and 5 times the conventional rate. We elected to work with square QAM constellations even though certain departures from such constellations yield superior performance.[11] The reason for our choice is that we want to analyze FBS in isolation and the aforementioned departures can be viewed as the first step in using the HDC method. Finally, we employ the harmless expediency of dealing with M as if it is a continuous variable in our calculations and in some of our graphs. When M is a positive integer that is not a

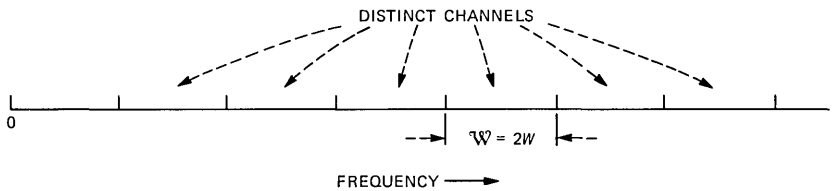


Fig. 1—Adjacent passband channels of bandwidth $\mathscr{W} = 2W$.

perfect square, one can find a two-dimensional constellation that realizes the situation covered by the analysis.

Throughout we will assume that the standard QAM method is meeting a P_e requirement (10^{-3} or less). When we compare alternatives to the QAM method, we will associate primed variables with parameters of the non-QAM system where needed to avoid ambiguity.

2.3 FBS

The generic system structure depicted in Fig. 2 is interpreted here for the special case of FBS. For FBS the binary data are blocked into successive 2-bit words. A pair of independent, synchronous, delta function streams are formed using the two bits to randomly sign the pair of delta functions that are input to the pair of baseband filters. The baseband filters nominally cut off at W cycles/second. On each rail the pulse rate is $r = R/2$. For convenience we use $T' = 1/r$ to denote the time interval between impulses. The filter outputs combine to form the in-phase and quadrature rails of a passband signal. Thus, it may seem that what we have is 4 PSK; but FBS is unusual in that its symbol rate, $1/T'$, is higher than the conventional $1/T \cong 2W$.

The higher rate does not increase bandwidth but it does cause ISI, which will be combatted with MLSD. The ISI is assumed to involve a

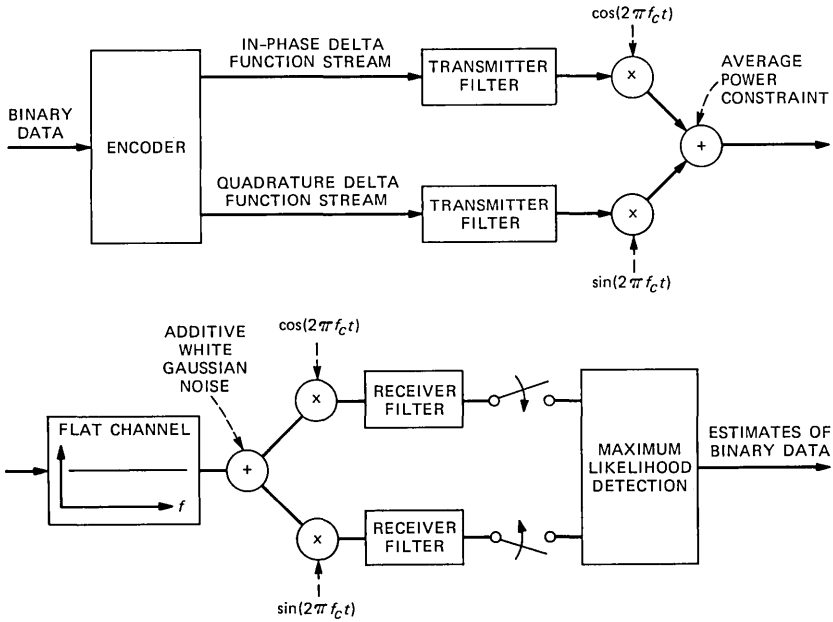


Fig. 2—Generic system structure. For QAM the transmit and receive filters are low-pass filters. For the FBS case they are a matched pair with finite memory in the sampled data domain. The nominal cutoff frequency of the low-pass filters is $W = \mathscr{W}/2$.

memory of ν, i.e., the bandlimited baseband filters are such that the impulse response sampled at times $\{n/T'\}_{n=-\infty}^{\infty}$ have the form $h = \overline{0}$, $h_0, h_1, \cdots h_\nu, \overline{0}$. Thus MLSD requires $2^{\nu+1}$ states (2^ν per rail).*

We would like to, if we could, make the following idealizations concerning $H(f)$, the Fourier transform of the impulse response of the baseband filter.

A. Each member of a bank of FBS passband systems is spectrally disjoint [in baseband, $H(f)$ vanishes outside $(-W, W)$].

B. $H(f)$ is a trigonometric polynomial of degree ν on $-W' \leqslant f \leqslant W'$ $[W' = 1/(2T') > W]$.

The first of the above is needed for spectral efficiency. Statement B is needed to be consistent with the assumption that MLSD involves a memory of ν. Mathematically it would seem impossible to meet A and B. After all, if $H(f)$ is a trigonometric polynomial vanishing on $[W, W']$, it vanishes everywhere. There is no real difficulty here. We will adhere to B with ν the degree of $H(f)$. While A will not strictly hold, one can get as close to ideal as desired as long as ν is large enough to meet out-of-band energy constraints. In MLSD a matched filter is used to initiate the detection process. The matched filter receiver serves to select the desired band $[-W, W]$. We have a dual view of what the frequency band is. From the point of view of where the signal power is concentrated, $[-W, W]$ is the band. From the point of view of MLSD, we are dealing with a sampled temporal response, which can only correspond to a transform that is a polynomial on $[-W', W']$. So long as the degree of the polynomial is large enough the two views of bandwidth can be reconciled.

There are several questions before us. Can we make the energy outside the $[-W, W]$ band so small that the interference between neighboring systems is negligible, yet the number of states involved in MLSD is reasonable? If we can accomplish this, does the FBS system perform better than the comparable QAM system? How much better does it perform and at what complexity?

We investigate these questions in the context of three kinds of discrete impulse responses. The first of these is the Nyquist responses ("brickwall spectra" on $[-W, W]$) truncated to memory ν. For these we shall see that the interference from neighboring systems is prohibitive for reasonable ν. The second set of impulse responses are the discrete prolate spheroidal wave functions, which, for fixed total energy and fixed ν, have the least interference from neighboring systems. We demonstrate that their performance is not good. Finally we explore optimally designed responses and find that for reasonable ν, even for the most favorable cases, the advantage over QAM is very modest.

* Notice that if the constellation were not the product of two one-dimensional constellations, as we have assumed, the complexity would grow as 4^ν instead of $2^{\nu+1}$.

III. PERFORMANCE CRITERION

3.1 Definition of gain

The probability of bit error, P_e, is an important performance measure for data communication systems. For QAM as well as the systems employing HDC, UTC, or FBS, if MLSD is used the probability of bit error decays exponentially as the noise spectral density is decreased (except for an algebraic multiplier). That is, an exponentially tight bound on P_e has the form

$$P_e \leq \kappa\sigma e^{-\frac{\xi}{2\sigma^2}} \qquad (\sigma \to 0),$$

where κ and ξ are independent of σ^2, the noise power spectral density on a single dimension. For FBS viewed in the MLSD context, the exponentially tight bound has the form

$$P_e \leq \kappa\sigma e^{\frac{-d^2}{2\sigma^2}} \qquad (\sigma \to 0).$$

The minimum distance is defined by

$$d^2_{\min} \triangleq \min_e ||h*e||^2, \tag{1}$$

where $*$ denotes convolution and the minimum is over all doubly infinite sequences e of the form

$$\overline{0}1\epsilon_1\epsilon_2 \cdots \epsilon_K\overline{0},$$

where ϵ_k belongs to $\{0, 1, -1\}$ and K can be any nonnegative integer.

Clearly, $d^2_{\min} \leq ||h||^2$. In cases where $d^2_{\min} = ||h||^2$ it is common to say that the matched filter bound is attained. What is meant is that the exponent of P_e is the same as if there were only a single data pulse to be detected (no ISI). The terminology stems from the fact that, when there is no ISI, MLSD employs simply a matched filter (along with a threshold comparator).[12]

We take the quantity ξ as a convenient indicator of performance in the high s/n realm. (We stress that, for models of specific systems, more refined computations estimating the actual error probability are often needed.) The "gain" of one system over another is expressed as

$$G = 10 \log_{10} \frac{\xi'}{\xi}.$$

We shall be concerned in this paper with estimating the gain that FBS exhibits over QAM. Both UTC and HDC exhibit substantial gain over QAM. For UTC, gains in the range of 3 to 6 dB have been reported and, for a 3-dB gain, the required complexity is extremely reasonable.[2]

For the conventional QAM system, $||h||^2$ denotes the energy per pulse prior to multiplication by a_i belonging to $[\pm 1, \pm 3, \cdots \pm$

$(L-1)]$, so the modulated pulse has average energy $(L^2-1)/3\,||h||^2$. If there is one symbol every T seconds, the average signal power is $[(L^2-1)/3]\,[||h||^2)/T]$. Since the information rate per rail is $(\log_2 L)/T$ b/s, FBS must operate at a rate of $(\log_2 L)/T$ pulses/s. Let h' be the impulse response for FBS. For the two systems to have identical signal power we must have

$$||h'||^2 = (L^2-1)||h||^2/(3\log_2 L).$$

For FBS, accounting for the wider bandwidth, the noise variance per sample is $\sigma^2(\log_2 L)/T$. Not necessarily all of $||h'||^2$ is realized in the error exponent.

The gain over the corresponding QAM system is expressed as

$$G = 10\log_{10}\frac{d_{\min}^2}{||h||^2}\frac{L^2-1}{3\log_2 L} = 10\log_{10}\frac{d_{\min}^2}{||h||^2}\frac{4^\rho-1}{3\rho}, \tag{2}$$

where $\rho = \log_2 L = W'/W = T/T'$. One could interpret $10\log_{10}[(4^\rho - 1)/3\rho]$ as a noise immunity gain and $10\log_{10}(d_{\min}^2/||h||^2)$ as the penalty for ISI.

Shortly it will prove useful to allow for replacing the noise power spectral density on the FBS system by a level greater than that on the slower system, say $(1+\beta)\sigma^2$ with $\beta > 0$ in place of σ^2. This will enable us to compensate for interference from adjacent channels. When, and if, the matched filter bound is attained the gain is expressed by $10\log_{10}(4^\rho - 1)/[3\rho(1+\beta)]$. Figure 3 depicts this function with β as a parameter. It is evident from the $\beta = 0$ curve that, depending on ρ, if the matched filter bound is attained, the gain can be considerable.

We consider now the interference in the band $(-W', W')$ that stems from those channels (other than the primary channel centered at zero) whose power spectral density is nonzero in $(-W', W')$. The determination of the additional power due to these interfering channels is straightforward. When measured for a single rail, at the output of the matched filter, $H^*(\omega)$, the power is the same as if σ^2 were replaced by

$$\sigma^2 + \int_{-W'}^{W'} |H'(f)|^2 \frac{\left[\frac{1}{T'}\sum_i |H'(f-i/T)|^2\right]}{\int_{-W'}^{W'} |H'(g)|^2 dg}\,df. \tag{3}$$

The sum is over all neighboring systems overlapping the $(-W', W')$ band. Because of its genesis, the term that adds to σ^2 in (3) is called the Adjacent-Channel Interference term or ACI. For $H'(f)$ with nearly all the energy in the $(-W, W)$ band, $|H'(f)|^2/\int_{-W'}^{W'} |H'(g)|^2 dg$ has a mean value of approximately T on $(-W, W)$. Since each channel is symmetrically disposed relative to its neighbors, the integral in the

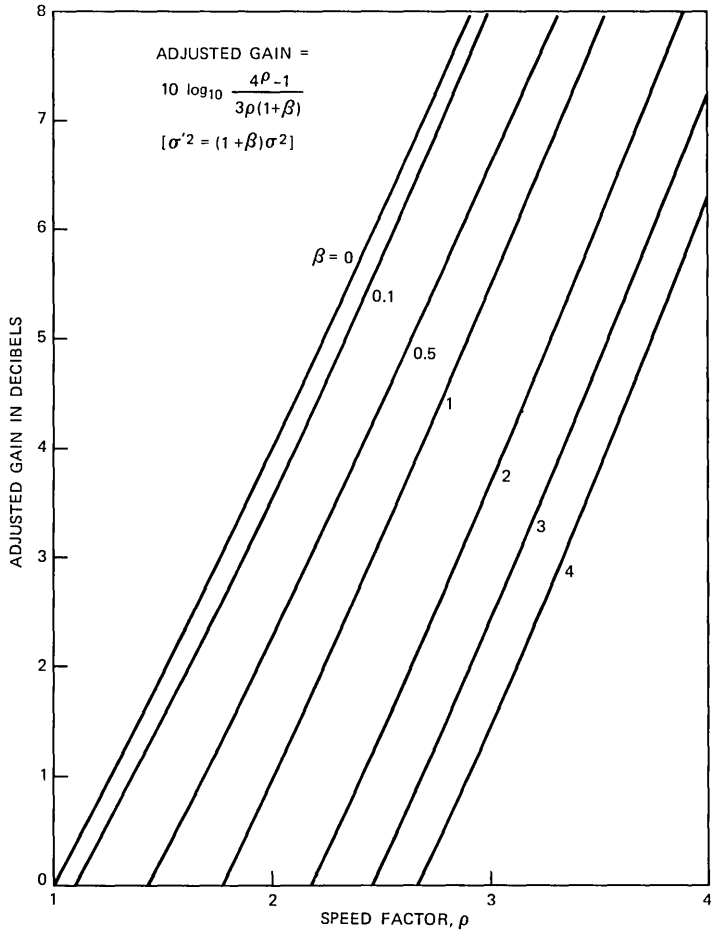


Fig. 3—Adjusted gain versus speed factor when matched filter bound is attained. Adjusted gain = $10 \log_{10} \frac{4^\rho - 1}{3\rho(1 + \beta)}$.

numerator of (3) can be replaced by $\int_{-W}^{W}$. It follows that the strength of the ACI term in (3) is roughly indicated by that energy in a pulse with transform $H'(f)$ that is out of the band $(-W, W)$. (We use OBE to denote out-of-band energy.) This approximation becomes more precise if $H'(f)$ is approximately flat in $(-W, W)$ (as is the case in Section IV).

If ACI is not negligible, it is reasonable to modify the gain by subtracting $10 \log_{10}(1 + OBE/\sigma^2)$ or, the more precise but more complex, $10 \log_{10}(1 + ACI/\sigma^2)$. No matter which gain expression is most appropriate, if we insist on some degree of spectral isolation it is unclear how much gain can be attained at specific levels of complexity

($2^{\nu+1}$ with $\nu \leq 26$). In the sequel we will find that the answer is not much gain. For the analysis in Sections IV and V we consider OBE in the range $[0, \sigma^2]$. As we note in Section VI there is no point in considering OBE outside this range.

3.2 Error events

We can see from the formula for minimum distance (1) that, if we slide a window of size ν along an error sequence e, a repeated state is forced to occur by 3^ν shifts. It follows that, to attain the minimum, one need not search over more than $3^{(3^\nu)}$ events. Since $\{3^{(3^\nu)} \mid \nu = 1, 2, 3, 4, \cdots\} = \{27, 19683, 7.63 \times 10^{12}, 4.4 \times 10^{38}, \cdots\}$. We see that, even for $\nu = 3$, a brute force search is extremely ambitious and for $\nu = 4$ it is completely out of the question. (Reference 13 discusses three other state symmetries as well as a repeat.)

For future reference, we borrow from Ref. 13 and list four useful representations and notations for error events in Appendix A.

3.3 Searching the tree of error events for $d^2_{\min}$

For each ν, it is useful to view a set of error events, one of which is guaranteed to achieve $d^2_{\min}$ as a tree. Construct a tree of sequences with three branches emanating out of each node and with the labeling illustrated in Fig. 4. The labels along each upward path represent the beginning string for the nonzero portion of an error event. Once a string of ν consecutive zeros is encountered, the growth out of such a node is pruned from the tree, since continuing the event with nonzero elements will correspond to creating labelings for beginnings of events with a greater $||h*e||^2$ than the all-zero continuation.

To envision a computer search for $d^2_{\min}$ for a specific h, one can think of climbing up the tree and to the left and at each node computing the accumulated

$$\sum \langle h^b, \Delta_j^+ \rangle^2 \quad \text{(notation in Appendix A)}$$

on the upward path to the node. There is one summand for each node in the upward path. Climb higher if the record low for a completed

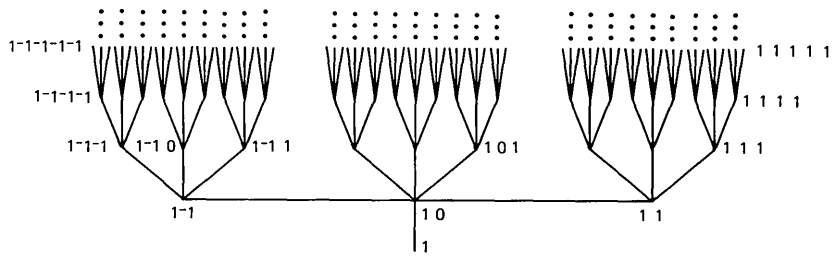


Fig. 4—Error event tree.

error event (a number $\leqslant ||h||^2$) has not been exceeded. Otherwise, climb down to the first node that offers an unclimbed branch and then climb that branch. Whenever a node with ν consecutive zeros is reached, and the old record has not been reached, record the new candidate event for achieving $d^2_{\min}$ and the new record before climbing down.

Some additional special search tools prove useful. Specifically, one can terminate an upward climb whenever any of the four symmetries $\Delta_i = \pm\Delta_j (i < j)$ or $\Delta_i = \pm\Delta_j^b (i \leqslant j)$ is detected (see Ref. 3). Two other search tools are very powerful, one for symmetrical h (see Appendix B) and one for nonsymmetrical h, which is discussed in Section 6.3.

Searching for $d^2_{\min}$ in the manner described will enable us to investigate the efficacy of FBS. In Sections IV, V, and VI, we consider three classes of FBS pulses.

IV. NYQUIST PULSES

4.1 Performance

One of the most elementary results in data communication theory is Nyquist's result that, for $WT = 1/2$, signals of the form $\sum_{-\infty}^{\infty} a_n h(t - nT)$ with $h(t) = \sin t/t$ are ISI-free. In FBS, we replace T by T/ρ with $\rho > 1$ and, as we have already mentioned, the signal bandwidth is invariant to ρ but ISI arises. A hypothetical FBS system based on such a pulse incorporates a level of idealization beyond the standard one associated with the abrupt cutoff. Namely, the system represents the limits of infinite decoding complexity as well as zero energy in the bands, $W < |f| \leqslant W'$.

For each system, assuming the pulse energy is normalized to 1 and W' is normalized to 1/2, we have

$$d^2_{\min} = \inf \frac{\rho}{2\pi} \int_{-\pi/\rho}^{\pi/\rho} \left| 1 - \sum_{1}^{K} \epsilon_k e^{jk\omega} \right|^2 d\omega. \tag{4}$$

The infimum is over all error events with ϵ_k belonging to $\{0, 1, -1\}$ and K ranging over the positive integers. Expression (4) is considered in Ref. 5 where it is demonstrated that $d^2_{\min} > 0$ for all $\rho \geqslant 1$. The intriguing question of whether, for such systems, a positive "gain" is available remains open.

Let

$$H_\rho^{NY}(\omega) \underline{\Delta} \begin{cases} \rho^{1/2} & |\omega| \leqslant \dfrac{\pi}{\rho} \\ 0 & \text{otherwise} \end{cases}$$

denote the FBS pulse transform normalized to unit energy. While expression (4) allows for infinite complexity, for any implementation, approximations of H_ρ^{NY} must be considered. The optimum least-mean-

square approximations, the Fourier series $\{H^{NY}_{\nu,\rho}(\omega) \triangleq \sum_0^\nu q_n e^{jn\omega}\}$, are natural responses to use to inquire whether, as ν increases, FBS performs better than QAM. Of course, if ν becomes too large the required detector becomes forbiddingly complex.

The tree search discussed in Section 3.3 was used to determine the "gain" for the least-mean-square approximations. The symmetry of the impulse responses allowed the addition of the test of Appendix B to significantly reduce the running time of the algorithm.

The results are shown in Fig. 5. The "apparent gains" are only meaningful if the Out-of-Band Energies (OBEs) are sufficiently small. Indeed, they are not sufficiently small as we now discuss. Figure 6 shows the out-of-band energy for a unit energy response for $\nu = 26$. Also shown are the noise levels for the benchmark QAM system providing the same information rate at a P_e of 10^{-3} and 10^{-6}. Of the four points $\rho = 2$, 3, 4, and 5, only $\rho = 2$ shows the out-of-band energy below the noise level. The margin for $P_e = 10^{-3}$ is slight ($\approx$4 dB) but, from Fig. 5, we see that for $\nu = 26$ the "gain" is negligible ($\approx$0.1 dB). For $\rho = 2$, if we reduce ν to increase the "gain", the attempt is undermined by the increase in out-of-band energy. The out-of-band energies for $\nu = 20$ and $\nu = 14$ are also shown in Fig. 6, for $\rho = 2$.

We conclude that, for the least-mean-square approximation of a Nyquist pulse, FBS signaling under the mild requirement $P_e = 10^{-3}$ does not offer any significant gain over QAM. In making the comparison, we have allowed FBS the extraordinary complexity of $2^{26} \approx 6 \times 10^7$ states per rail ($>10^8$ states total). If P_e were decreased, FBS would fare even worse.

Figure 7 illustrates approximate Nyquist spectra for $\nu = 26$ and minimizing error events. We note that for $\nu = 26$ the number of candidate error events exceed $10^{(10^{12})} < 3^{(3^{26})}$.

4.2 *Infinite complexity, complete spectral confinement asymptote*

Allowing for complexity not exceeding 10^8 states, FBS is not attractive relative to QAM for the examples considered thus far. As we mentioned in the last section, for the asymptote of infinite complexity ($\nu \to \infty$) and stringent out-of-band energy constraint, the limiting squared distance is expressed in (4). Although we cannot compute the gain G_p, we can find an upper bound using candidate error events. We used events revealed to be useful in the tree search for $\nu \leqslant 26$. The list of trigonometric polynomials below $\{E_\rho(\omega)\}_{\rho=2}^5$ was used to bound $d^2_{\min}$:

$$E_2(\omega) = (1 - e^{j\omega} + e^{3j\omega} - e^{4j\omega} + e^{6j\omega} - e^{7j\omega})$$

$$E_3(\omega) = (1 - e^{j\omega} + e^{4j\omega} - e^{5j\omega})$$

$$E_4(\omega) = (1 - e^{j\omega} - e^{2j\omega} + e^{3j\omega} + e^{4j\omega} - e^{5j\omega})(1 + e^{12j\omega})$$

$$E_5(\omega) = (1 - e^{j\omega} - e^{2j\omega} + e^{3j\omega} + e^{4j\omega} - e^{5j\omega}).$$

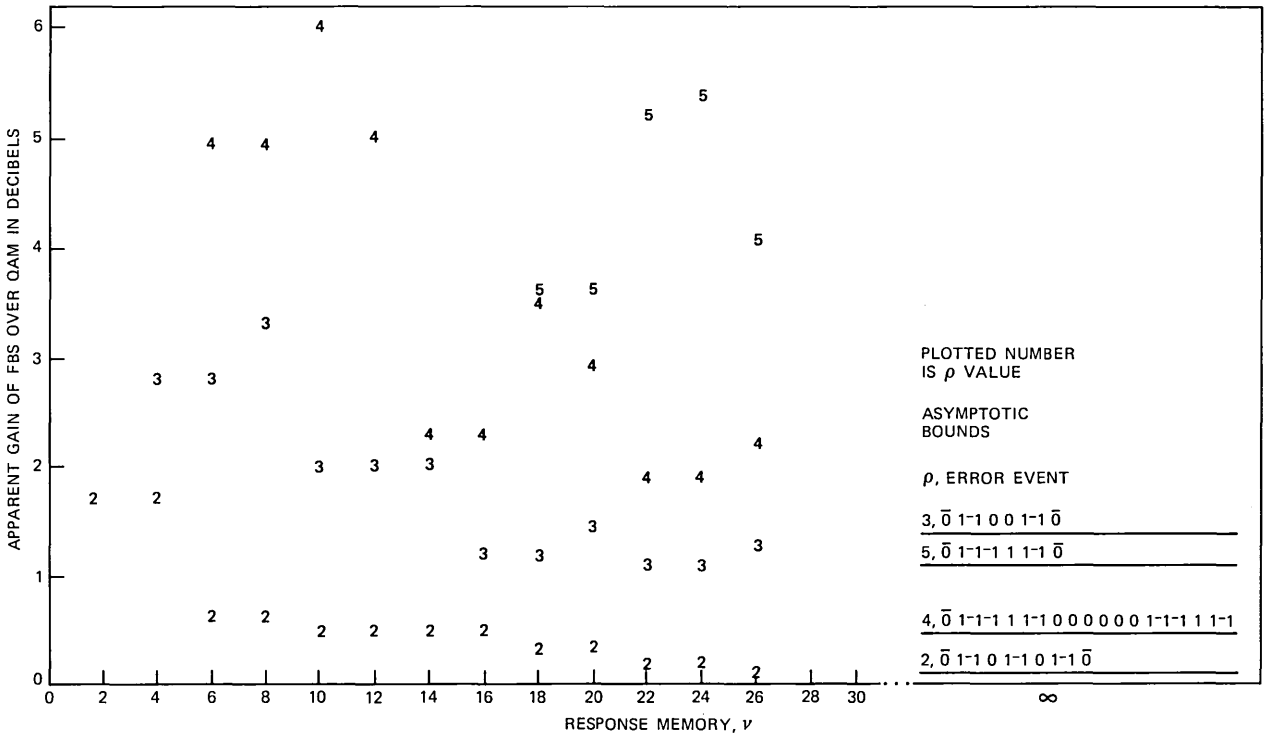


Fig. 5—Apparent gain of FBS over QAM (gains unachievable because of interference from adjacent bands).

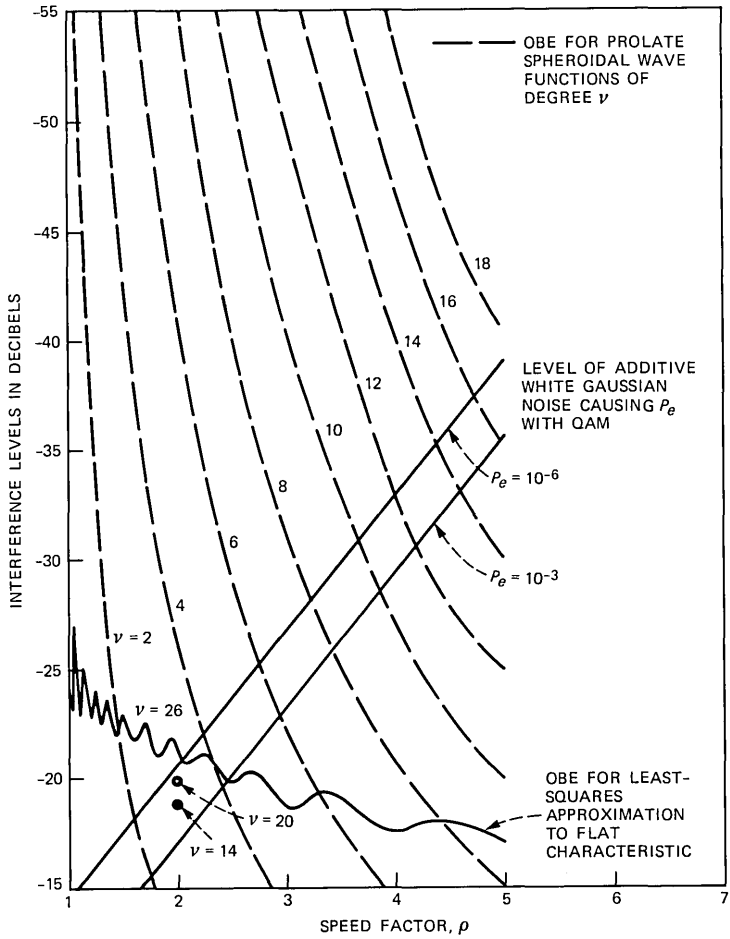


Fig. 6—Comparison of interference levels.

They yield $G_2 < 0.107$ dB, $G_3 < 1.4$ dB, $G_4 < 0.477$ dB, and $G_5 < 1.1$ dB. This shows that, even allowing for an arbitrarily large number of states, in the limit of stringent out-of-band energy requirements, the gains available using a Nyquist pulse are at best very modest. The $E_\rho(\omega)$ characteristics are illustrated in Fig. 8.

V. DISCRETE PROLATE SPHEROIDAL WAVE FUNCTIONS

In Section IV we investigated whether the approximations $H_{\nu,\rho}^{NY}$ have good distance properties for FBS with $\rho = 2$, 3, 4, and 5. We found that, for reasonable complexity and out-of-band energy constraints,

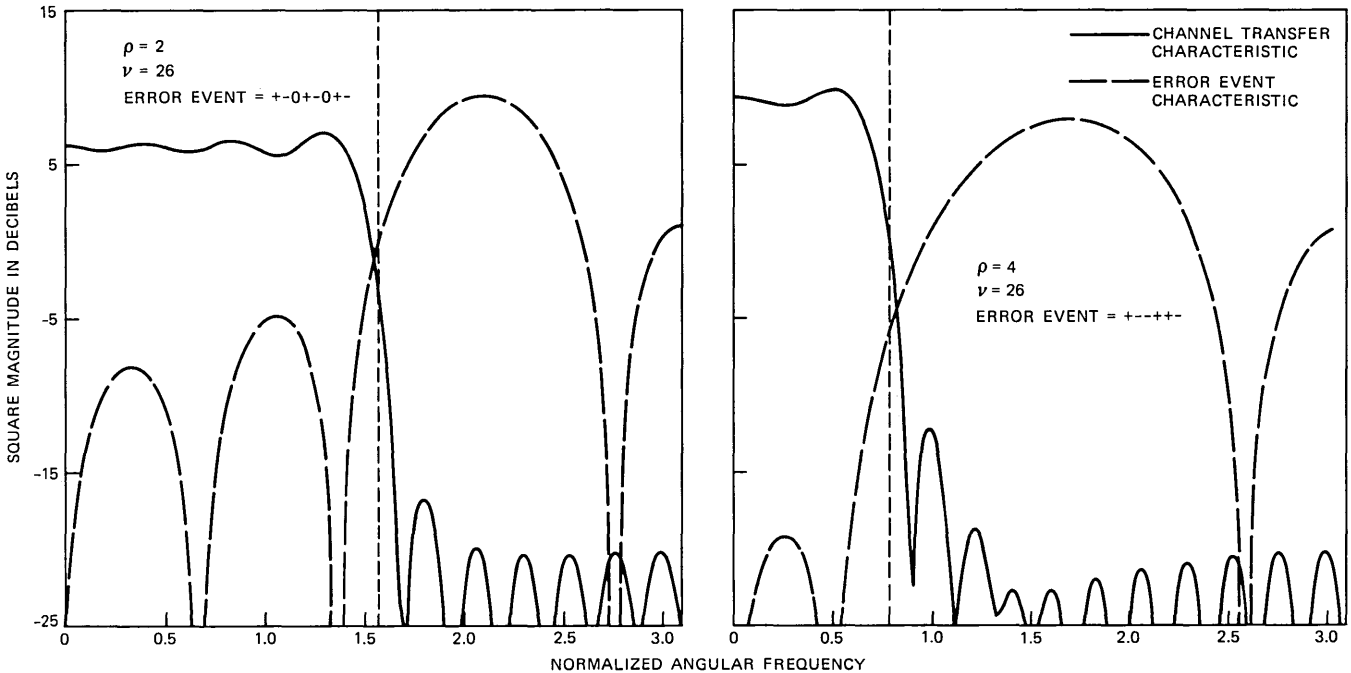


Fig. 7—Examples of $|H_{\nu,\rho}^{NY}|^2$ and extremal error event (+ and − mean +1 and −1). Vertical dotted line marks transition from in-band to out-of-band.

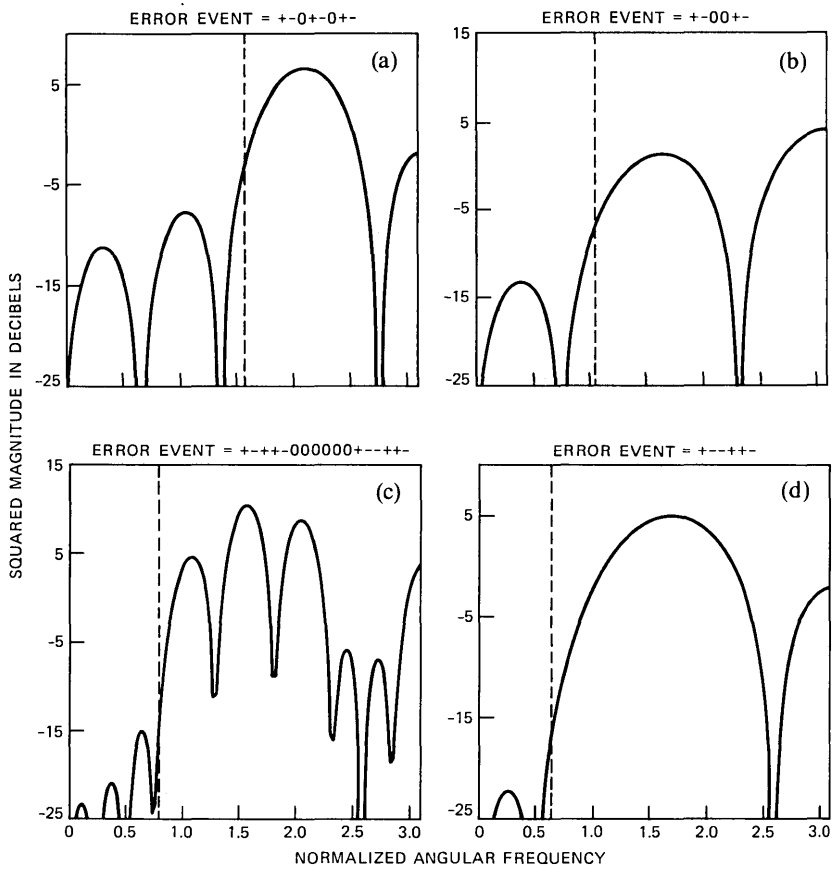


Fig. 8—Extremal error event for (a) $\rho = 2$, (b) $\rho = 3$, (c) $\rho = 4$, and (d) $\rho = 5$.

they do not. The $H_{\nu,\rho}^{NY}$ minimize $\int_{-\pi/\rho}^{\pi/\rho} |H_{\rho}^{NY} - \sum_0^{\nu} q_n e^{jn\omega}|^2 w(\omega) d\omega$ in the special case when the weight function $w(\omega) \equiv 1$. In light of the results of Section IV, we can reformulate the least-mean-square approximation using a $w(\omega)$ that is 0 on $(-\pi/\rho, \pi/\rho)$ and 1 otherwise. The weighting reflects the fact that it is essential to keep the out-of-band energy small but, having seen that the flat transform has no special distance properties, we have no motivation for keeping the transfer characteristic flat within $(-\pi/\rho, \pi/\rho)$.

The extremal responses so obtained are called the Discrete Prolate Spheroidal Wave Functions (DPSWF).[14] Their theory has been developed by Slepian.[15] Wyner has suggested their consideration for use in data communication systems for reasons other than those we are considering here.[16] Let $H_{\nu,\rho}^{PS}$ denote the transform of the discrete spheroidal wave function of memory ν corresponding to an FBS system

with parameter ρ. Since minimizing the out-of-band energy corresponds to maximizing the in-band energy, the coefficients of $H_{\nu,\rho}^{PS}$ are then components $(q_o, q_1, \cdots q_\nu)$ of the eigenvector of the matrix of the symmetric quadratic form

$$\frac{\rho}{2\pi} \int_{-\pi/\rho}^{\pi/\rho} |q_o + q_1 e^{j\omega} + \cdots q_\nu e^{j\nu\omega}|^2 d\omega$$

corresponding to the largest eigenvalue, $\lambda_{\nu,\rho}$. Since we normalize by constraining $H_{\nu,\rho}^{PS}$ to have unit energy, the quantity $1 - \lambda_{\nu,\rho}$ is the out-of-band energy.

In Fig. 6, $10 \log_{10}(1 - \lambda_{\nu,\rho})$ is plotted against ρ for various ν (see dashed curves). Unlike $H_{\nu,\rho}^{NY}$ we see that a significant portion of the loci for $H_{\nu,\rho}^{PS}$ are disposed well above curves for the noise levels for $P_e = 10^{-3}$ and $P_e = 10^{-6}$. Thus, there are spaces of systems of moderate complexity with small out-of-band energy, whose distance properties are of interest. What are the distance properties of $H_{\nu,\rho}^{PS}$ in the range $\rho = 2, 3, 4, 5$? They are not good. Use of the search algorithm of Section 3.2 demonstrated no gain for any $H_{\nu,\rho}^{PS}$ whose (ν, ρ) coordinate corresponded to an out-of-band energy below the level of the Gaussian noise for P_e of 10^{-3}. For example, for $\rho = 2$ at $\nu = 4$, the out-of-band energy is -26.3 dB, which is below the level of the additive noise. However, the minimum distance of $H_{2,2}^{PS}$ is poor, specifically $G = -1.33$ dB. For larger ν, the out-of-band energy drops precipitously but distance decreases as well. As ρ increases in the range 3, 4, and 5, the situation worsens: G values significantly below 0 dB occur with out-of-band energy prohibitively above the noise level. As ν is increased, the distance drops markedly.

At this point we have an interesting situation. The least-mean-square approximations to the Nyquist pulse are shown to have attractive "apparent gains" relative to QAM but the gains cannot be realized because the signal spectrum is not adequately confined. On the other hand, the results for DPSWF's show that great spectral confinement is possible, but these pulse shapes do not exhibit any gain over QAM. The question remains as to how much gain we can achieve under a spectral confinement constraint for a specific complexity. This is addressed in Section VI.

VI. OPTIMUM PULSE DESIGN

In this section we investigate the performance of optimally designed FBS responses of prescribed complexity (i.e., prescribed memory ν). By optimum we mean that the minimum distance is maximized. The transmitted power, which is proportional to $1/(2\pi) \int_{-\pi}^{\pi} |H(\omega)|^2 d\omega$, and the out-of-band energy, which is proportional to $1/\pi \int_{\pi\rho^{-1}}^{\pi} |H(\omega)|^2 d\omega$ are both constrained. Once the optimum $\mathbf{h} = h * h^b$ (equivalently $\mathbf{H} = |H(\omega)|^2$) is found, we factor $\mathbf{H}$ to determine an

optimal h. (Since $\mathbf{H}$ can have 2ν zeros disposed in inverse conjugate pairs there can be as many as 2^{ν} possible factors of $\mathbf{H}$ that have real coefficients.) The problem of finding optimal $\mathbf{h}$ is essentially a linear programming (LP) problem. The suggestion of viewing optimum MLSE system design as an LP problem appears in my paper with R. R. Anderson.[13] It turns out that, in most cases of interest, the number of constraints corresponding to the various error events is too large for the LP to be useful by itself. The LP is combined with the tree search algorithm that serves to eliminate most error events from consideration. The LP-tree search algorithm solves the design problem. We proceed now to describe the LP and show how it is integrated with a tree search algorithm. Then we present the performance results for optimally designed pulses.

6.1 Linear program

Recall that an LP problem is one of the following type: *Given a vector* $\mathbf{c}$, *find a vector* $\mathbf{y}$ *that maximizes* $\langle \mathbf{y}, \mathbf{c} \rangle$ *subject to a set of linear constraints of the form* $\langle \mathbf{y}, \mathbf{a}_i \rangle \geq b_i$, i *belonging to* $\mathcal{I}$, *a finite index set.* The $\mathbf{a}_i$ and b_i are given vectors and scalars, respectively. It is very useful that $\leq$ constraints can be converted into $\geq$ constraints by changing sign and so equality constraints can be represented by a pair of $\geq$ constraints.

In our application, $\mathcal{I}$ is infinite. Since we shall see that the feasible $\mathbf{y}$ exists in a bounded set we can, in principle, obtain a solution as close to optimum as desired by solving an LP with sufficiently many constraints.

6.2 Embedding systems in a $2\nu + 2$ space

Now $\mathbf{h} = \overline{0}\ \mathbf{h}_{-\nu}, \cdots, \mathbf{h}_o, \cdots, \mathbf{h}_\nu\ \overline{0}$. We will represent $\mathbf{h}$ in a $2\nu + 2$ dimensional space where the first $2\nu + 1$ coordinates are $(\mathbf{h}_{-\nu}, \cdots, \mathbf{h}_\nu)$. An additional coordinate augments the projection of $\mathbf{h}$ so that we have $(\mathbf{h}_{-\nu}, \cdots, \mathbf{h}_\nu, \mathbf{s})$. The augmented vector of $2\nu + 2$ components is denoted $\mathbf{y}$. The additional coordinate, $\mathbf{s}$, is a mathematical convenience that will facilitate maximization of the minimum squared distance, as we shall see.

To describe the linear constraints defining the set of admissible $\mathbf{h}$, we need to employ $\mathbf{1_k}$ to denote a vector that has all-zero coordinates except the kth coordinate, which is a one.

In $2\nu + 2$ space, we describe linear constraints defining the set of admissible $\mathbf{h}$.

1. As a convenient normalization, we assume that the energy in h cannot exceed 1, so $\mathbf{h}_o \leq 1$; therefore,

$$\langle \mathbf{y}, -1_{\nu+1} \rangle \geq -1.$$

2. $h*h^b = \mathbf{h}$, so for $i \leq \nu$, $\mathbf{h}_{-i} = \mathbf{h}_i$; therefore,

$$\langle \mathbf{y}, -\mathbf{1}_i + \mathbf{1}_{2\nu+2-i} \rangle = 0.$$

3. $\mathbf{H}(\omega)$ is nonnegative. $\mathbf{H}(\omega)$ is the function that has $\mathbf{h}$ for Fourier coefficients and the operation of Fourier series is a linear one. So the constraints $\mathbf{H}(\omega) \geq 0$, $0 \leq \omega \leq \pi$, can be put in the form $\langle \mathbf{y}, \mathbf{a}_\omega \rangle \geq 0$ by defining $\mathbf{a}_\omega$ appropriately. There is one constraint for each ω on $0 \leq \omega \leq \pi$. In our application we can use a discrete set of the form $\{\omega_n = (n\pi)/N\}_{n=0}^{N}$, with N sufficiently large to give adequate accuracy.

4. The out-of-band energy cannot exceed a prescribed amount θ, so $1/\pi \int_{\pi/\rho}^{\pi} \mathbf{H}(\omega)\, d\omega \leq \theta$. Let $\Theta(\omega)$ be defined to be the function that vanishes for $|\omega| < \pi/\rho$ and is 1 otherwise. Therefore, $1/(2\pi) \int_{-\pi}^{\pi} \mathbf{H}(\omega)\, \Theta(\omega)\, d\omega \leq \theta$. By the Parseval theorem, we can express this constraint as $\langle \mathbf{y}, \mathbf{a} \rangle \leq \theta$, where $\mathbf{a}$ is a $2\nu + 2$ vector with a zero in the last position and the first $2\nu + 1$ coordinates are Fourier coefficients of $\Theta(\omega)$ with index of absolute value $\leq \nu$.

5. The $2\nu + 2$ component, seemingly extraneous so far, now comes into play. Let $\{\mathbf{E}_j(\omega)\}_{j \in \mathcal{J}}$ be the error polynomials. Project them into a $2\nu + 2$ dimensional space using the successive Fourier coefficients with index of absolute value $\leq \nu$ to get the first $2\nu + 1$ components and use -1 for the last component. Call the resulting vectors $\{\mathbf{e}_j\}_{j \in \mathcal{J}}$. It will not bother us if some $\mathbf{E}_j(\omega)$ have nonzero Fourier coefficients with index exceeding ν. Taken together, the constraints $\{\langle \mathbf{y}, \mathbf{e}_j \rangle \geq 0\}_{j \in \mathcal{J}}$ amount to a statement that, for each admissible $\mathbf{h}$, the squared minimum distance is never larger than a candidate distance.

The optimal $\mathbf{h}$ is the one maximizing the minimum distance. So the constrained $\mathbf{y}$ attaining max $\langle \mathbf{y}, \mathbf{1}_{2\nu+2} \rangle$ has the optimum design for its first $2\nu + 1$ coordinates and the optimal exponent for the last coordinate.

In $2\nu + 2$ space, the set of all $\mathbf{y}$ meeting constraints is denoted $\mathbf{Y}$. $\mathbf{Y}$ is not empty. For example, it contains $\delta \mathbf{1}_{\nu+1}$, where δ is a number small enough that energy constraints are met. The optimization will not degenerate as $\mathbf{Y}$ is closed and bounded. $\mathbf{Y}$ is closed since it is expressible as the intersection of closed half-spaces. $\mathbf{Y}$ is bounded since each component of $\mathbf{y}$ is bounded by the pulse energy constraint. To see why, note that $\mathbf{e} = \mathbf{1}_{\nu+1}$ shows $y_{2\nu+2} \leq 1$. For the remaining bounds on the components of $\mathbf{y}$ we note $\mathbf{H}(\omega) = \sum_{-\nu}^{\nu} x_{m+1+\nu} e^{jm\omega} \geq 0$ and so factorization is possible, $\mathbf{H}(\omega) = H(\omega)\, H^*(\omega)$. Fourier coefficients $(x_1, x_2, \cdots, x_{2\nu+1})$ are sums of products and so, by the Schwarz inequality, $y_{\nu+1} = x_{\nu+1}$ bounds all the components of each $\mathbf{y}$ vector.

6.3 The optimization algorithm

The linear constraints include the error events and, in most exam-

ples of interest, there are too many error events. For example, for ν as small as 4, the estimate in Section 3.1 indicated that there are over 10^{38} error events about which we should be concerned. The difficulty of too many constraints may sometimes be handled by solving a problem with a manageable number of the constraints. If it can be verified that the optimum meets all constraints (not just the manageable ones), then the solution to the simplified problem is the same as the solution to the difficult problem. We design, via an LP, a response maximizing the minimum distance over some error events and then seek to verify, using a tree search, that the minimum distance is not reduced if one minimizes over all error events.

If the above procedure is unsuccessful, one can repeat it, enlarging $\mathscr{E}$ to include the minimizing error event revealed by the tree search. Eventually, the iteration process will converge. Prior to convergence, LP gives an upper bound while the tree search gives a lower bound to the $d^2_{\min}$ achievable by the optimum design.

The LP provides an $\mathbf{h}$, while the tree search requires an h. The minimum-phase deconvolution, $\hat{h}$, is suggested, since, among all h satisfying $h^b * h = \mathbf{h}$, $\hat{h}$ has the greatest $\sum_{i=0}^{k} h_i^2$ for each k.[17,18] This maximal frontal energy concentration expedites the tree search, which operates first on the leading coordinates of $\mathbf{h}$. (Orders of magnitude of difference in running time have been observed between minimum- and maximum-phase deconvolutions in the tree search algorithm.)

6.4 Performance

In estimating the performance of the optimum system employing an $\mathbf{h}$ of memory ν, power levels were set as follows: For the FBS system, as we mentioned in Section 6.2, the pulse energy was bounded above by one. The noise level was set to meet the required P_e in the benchmark system operating at maximum power. Finally, the out-of-band energy constraint was set to a fraction of the noise power.

The resulting gain versus ν curves are shown in Fig. 9 with ν as a parameter. The OBE is constrained to $\sigma^2/10$ so a penalty of $0.414 \approx 10 \log 11/10$ is included in the gain calculation. It is apparent from the curves that, even at extraordinary complexity (exceeding 10^8 states) and a P_e requirement of 10^{-3}, the resulting gains are very modest. This conclusion is not sensitive to the exact premises underlying the computation. Calculation shows that, for $\nu = 26$, if we allow θ to be larger than 1/10, the gains generally decrease because of the OBE penalty. There is little to achieve by making OBE smaller than $\sigma^2/10$ since the design is merely more constrained, and omitting the OBE penalty cannot add more to the gain than 0.414 dB. When the gain is positive, the ACI levels are generally within 0.25 dB of the OBE level so the gains based on ACI are not different in any important

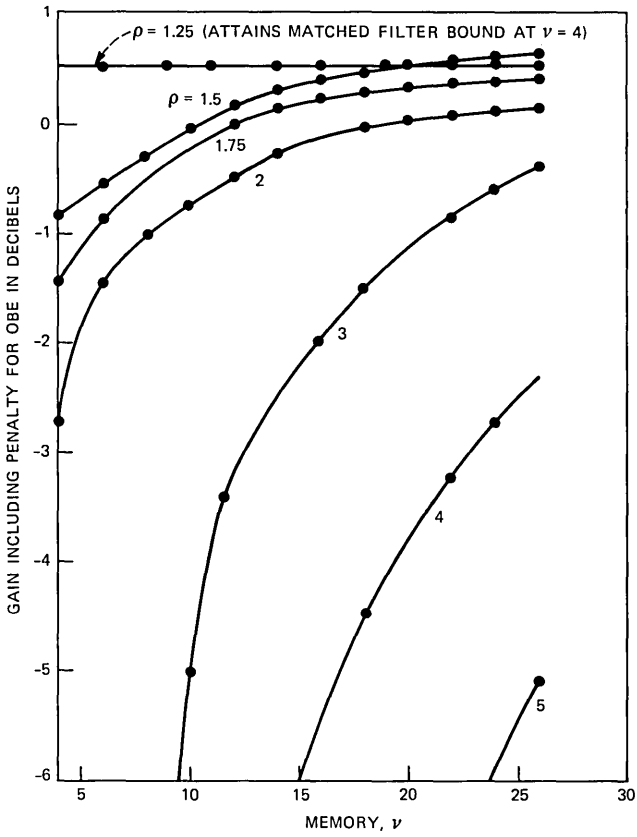


Fig. 9—Gain limit versus memory under spectral confinement constraint for $P_e = 10^{-3}$.

way from those based on OBE. There is no point in showing curves for $P_e < 10^{-3}$ in the benchmark system, as the gains can only decrease if the design is further constrained.

Figure 9 has enabled us to determine the relative merit of FBS for reasonable complexity. The possibility remains that, for extremely large ν, FBS could exhibit substantial gains and that these asymptotic gains could improve as ρ increases.

Figure 9 was derived using a list of 50 error events obtained by running the LP-tree search iteration for successive ν values. To conclude that FBS offers at best a very modest improvement over QAM, it is only necessary to present upper bounds in Fig. 9 rather than exact maximum gains. However, in preparing Fig. 9, we established that it is reasonable to exercise the LP-tree search algorithm to guarantee the precise optimum gain that can be attained for up to 10^5 states. It

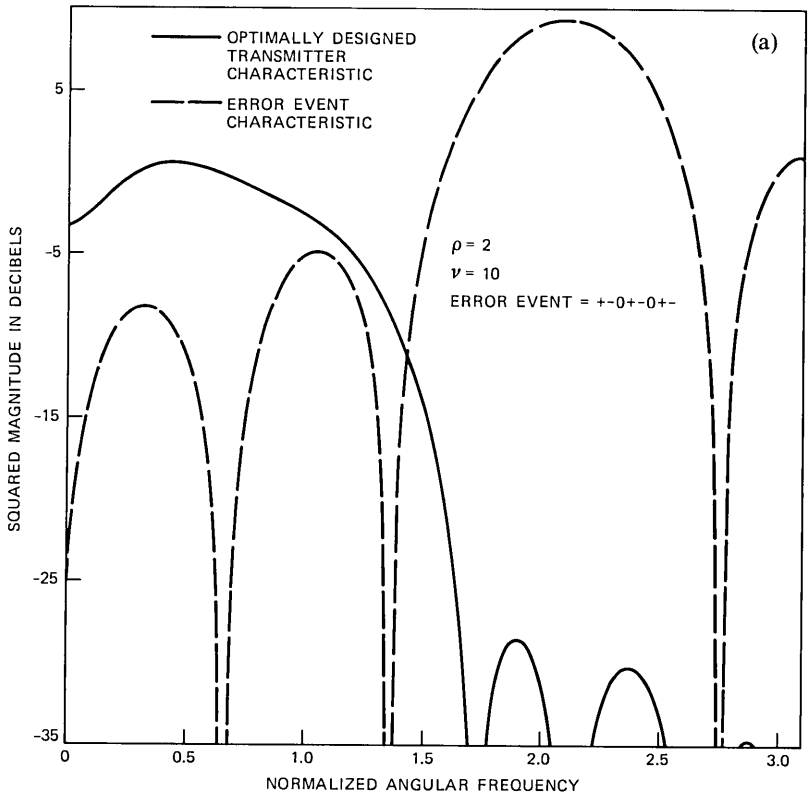


Fig. 10a—Example of an optimally designed transmitter characteristic and an error event characteristic. Vertical dotted line delineates the band edge.

is interesting to note that optimum system design can be accomplished for systems with such an enormous number of states.

Figure 10a illustrates an optimally designed spectrum and a corresponding minimizing error event. Figure 10b illustrates an interesting constrast between a pulse spectrum and an extremal error event.

For $\rho = 5/4$, the gain is only about 0.5 dB but a very interesting behavior is observed. Namely, with little complexity, the maximum distance possible is attained in the sense that the matched filter bound is obtained. From Section III, eq. (2), we can write the following expression for the gain (neglecting the OBE penalty):

$$G(\rho, \nu) = 10 \log_{10} \frac{(4^{\rho} - 1)}{3\rho} + 10 \log_{10} c(\rho, \nu),$$

where the function $c(\rho, \nu)$ gives the fraction of the matched filter energy attained. For fixed ρ, $c(\rho, \nu)$ is a nondecreasing function of ν. With ρ fixed, is $\lim_{\nu \to \infty} c(\rho, \nu) = 1$? In the case $\rho = 5/4$, we have seen

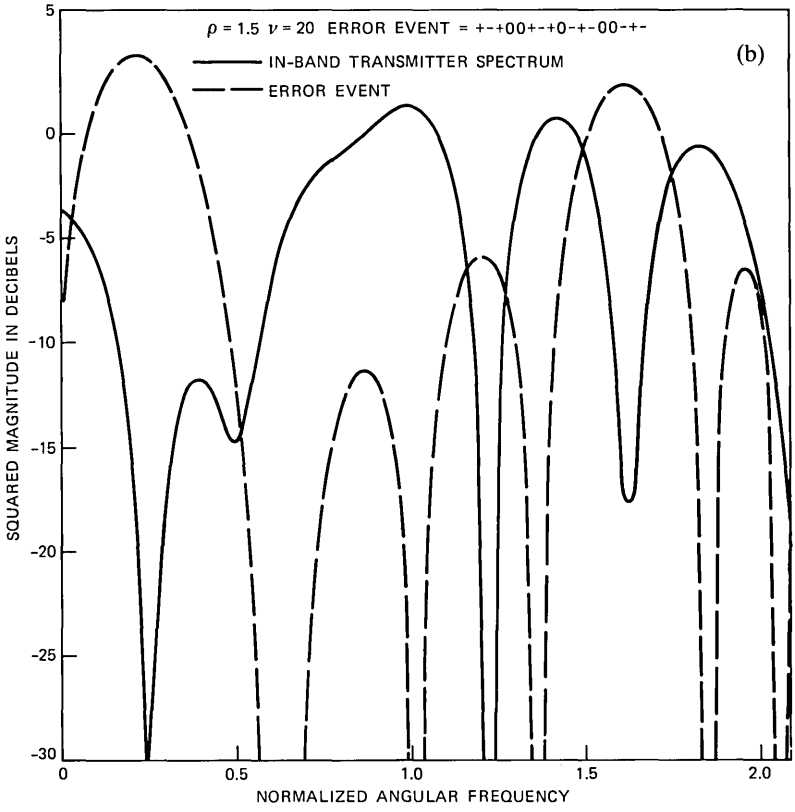


Fig. 10b—The in-band transmitter spectrum is optimized for a limited set of error events, one of which is shown. The extraordinary flexibility afforded by over one million states allows the optimum spectrum to have some peaks and valleys in opposition to those of the minimizing error event.

that the answer is yes. For $\rho > 3$ consideration of the error transform $|1 - e^{j\omega}|^2$ shows that the answer is no, as

$$\lim_{\nu \to \infty} c(\rho, \nu) \leq 4 \sin^2 \left(\frac{\pi}{2\rho} \right) < 1$$

in the limit of stringent out-of-band energy constraints. In light of the limited gains available with optimal FBS, it would be of only academic interest to pinpoint the largest ρ value for which the matched filter bound is attained as complexity is increased. Consequently, we shall not pursue this question further.

VII. DISCUSSION

At this point it is natural to question whether it is worthwhile to generalize and consider Faster Multilevel Signaling (FMS). Motiva-

tion for considering FBS comes from Ref. 5, where the first theoretical results on FBS were reported; from Ref. 3, where highly significant benefits of FBS were suggested but not established; and from discussions with J. Salz, who related that the idea of FBS has been around for many years and that it is important to settle the question of its merit. Since FBS proves to be unattractive, why should one consider FMS, especially when we know that increasing the number of levels toward that of the competing QAM system would seem to blur the distinction? Can one generally discount the competitiveness of FMS? We discuss why we cannot dismiss FMS and why, despite the findings on FBS, FMS systems may have some value.

7.1 Relieving the OBE constraint

FBS fared poorly. If we look back on our analysis of FBS it is obvious that it was the OBE constraint that drove the performance level of FBS. We noticed in Section IV suboptimal pulses exhibiting substantial gains that could not be realized because of prohibitive OBE. The stringency of the OBE constraint was necessitated by the substantial overlap of spectra between neighboring systems. As we move away from binary toward more levels, in the class of FMS systems, to compete with a fixed QAM system, the ratio $\rho = W'/W > 1$ decreases. The OBE constraint we need to impose is seen to be more relaxed.

Moreover, as we decrease ρ, systems are represented for which the ACI constraint is not of any direct importance. There is the interesting class of questions pertaining to transmitter filter smoothness considerations. For example, which performs better—a QAM system employing a square root raised cosine pulse with roll-off $\alpha = \rho - 1$, or an optimized FMS system with band-edge nulls of specified order and with system memory ν? The two systems are required to have the same power and information rate. The answer, of course, depends on M, α, ν, and the degree of the band-edge null. The band-edge null is useful for spectral confinement as well as for easing synthesizability. The imposition of nulls of specific order at specific frequencies lead to additional linear constraints and is easily handled by the LP-tree search program.

The simple partial response $\overline{0}$, 1, 1, $\overline{0}$ can be used to illustrate that there are situations where FMS can be very beneficial relative to QAM. Among all systems required to have a band-edge null, the system $\overline{0}$, 1, 1, $\overline{0}$ requires the least number of states, m per rail. It is easily shown that the response attains the matched filter bound independent of m. Figure 11 shows a gain versus roll-off plot, which speaks for itself concerning the substantial gains that are available in certain cases.

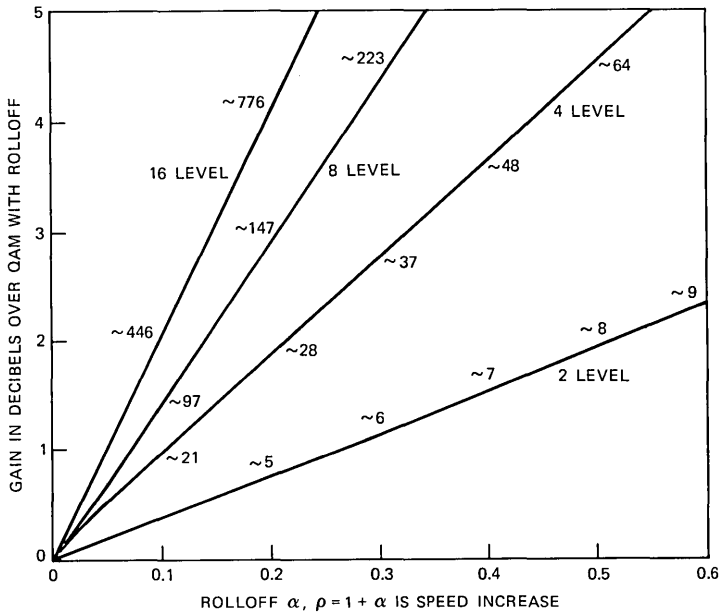


Fig. 11—Gain versus rolloff characteristics for partial response $2^{-1/2}$ ($\overline{0}$, 1, 1, $\overline{0}$), where the approximate number of constellation points are shown for competing QAM system, and the number of levels equals the number of states per rail.

A class of examples where ACI is not of direct concern occurs with the voiceband channel, which has severe band-edge attenuation. The channel shapes are irregular mounds and there is no obvious spectral support to assume. For a specific information rate, what is the best baud to use if the transmitter spectrum has a null of given order and zero rolloff? This is paraphrasal of the FMS issue coupled with a simpler question of where to center the signal spectrum. (The transmitter design must also account for the effect of nonlinearities and the fact that the exact modulus of the channel transfer characteristic is not known at the transmitter.)

7.2 The LP-tree search algorithm

Aside from the new information on FBS, a major finding of this report is that, for the class of MLSD systems considered, optimum designs can be accomplished involving numbers of states corresponding to the capabilities of forthcoming MLSD implementation technology (and far beyond). We have concentrated here on binary systems and a very special channel. However, the algorithm extends to apply to designing optimum m-ary systems of prescribed complexity operating over arbitrary linear dispersive channels. The astronomical number of error events is not an obstacle.

The extended algorithm, now being programmed in the course of joint work with G. Vannucci, will provide a basic tool for probing the fundamental relationship between attainable rates and system complexity for very general systems. Suppose one wants to achieve a certain information rate, under spectral confinement requirements and with a specific level of complexity. By exercising the LP-tree search algorithm for a sufficient number of ρ values, one can locate ρ_{opt}, the optimum $\rho \geq 1$, and the associated optimum gain over a corresponding QAM system with cosine rolloff spectral shaping.

VIII. ACKNOWLEDGMENT

It is a pleasure to acknowledge numerous valuable, stimulating discussions with G. Vannucci. These discussions involved both the theoretical and software aspects of this work. L. J. Domenico's assistance with the programming is greatly appreciated.

REFERENCES

1. A. Gersho and V. B. Lawrence, "Multidimensional Signal Design for Digital Transmission Over Bandlimited Channels," Proc. IEEE ICC, *1,* Amsterdam, May 1984, pp. 377–80.
2. G. Ungerboeck, "Channel Coding with Multilevel/Phase Signals," IEEE Trans. Inform. Theory, *IT-23,* No. 1 (January 1982), pp. 55–67.
3. A. S. Acampora, "Analysis of Maximum Likelihood Sequence Estimation Performance for Quadrature Amplitude Modulation," B.S.T.J., *60,* No. 6 (July–August 1981), pp. 865–85.
4. R. D. Gitlin and V. Lawrence, private communication.
5. J. E. Mazo, "Faster Than Nyquist Signaling," B.S.T.J., *54,* No. 8 (October 1976), pp. 1451–62.
6. G. D. Forney, Jr., "Lower Bounds on Error Probability in the Presence of Large Intersymbol Interference," IEEE Trans. Commun., *COM-20,* No. 1 (February 1972), pp. 76–7.
7. G. D. Forney, Jr., "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," IEEE Trans. Inform. Theory, *IT-18* (May 1972), pp. 363–78.
8. G. J. Foschini, "Performance Bound for Maximum Likelihood Reception of Digital Data," IEEE Trans. Inform. Theory, *IT-21* (January 1975), pp. 47–50.
9. G. J. Foschini, "A Reduced State Variant of MLSD Attaining Optimum Performance for High SNR," IEEE Trans. Inform. Theory, *IT-23,* No. 5 (September 1977), pp. 605–9.
10. A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding,* New York: McGraw Hill, 1979.
11. G. J. Foschini, R. D. Gitlin, and S. B. Weinstein, "Optimization of Two Dimensional Signal Constellations in the Presence of Gaussian Noise," IEEE Trans. Commun., *COM-22,* No. 1 (January 1974), pp. 28–38.
12. R. W. Lucky, J. Salz, and N. Weldon, *Principles of Data Communication,* New York: McGraw Hill, 1968.
13. R. R. Anderson and G. J. Foschini, "The Minimum Distance for MLSE Digital Data Systems of Limited Complexity," IEEE Trans. Inform. Theory, *IT-21,* No. 5 (September 1975), pp. 544–51.
14. A. Papoulis, *Signal Analysis,* New York: McGraw Hill, 1977, pp. 212–15.
15. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—V: The Discrete Case," B.S.T.J., *57,* No. 5 (May–June 1978), pp. 1371–430.
16. A. D. Wyner, "Signal Design for PAM Data Transmission to Minimize Excess Bandwidth," B.S.T.J., *57,* No. 9 (November 1978), pp. 3277–307.
17. E. A. Robinson, *Statistical Communication and Detection,* London: Griffin, 1967.

18. H. P. McKean and H. Dym, *Fourier Series and Integrals*, New York: Academic Press, 1972, pp. 175–6.

APPENDIX A

Error Event Representations

A.1 Sequence representation

$$\cdots 0 \cdots \overset{\nu}{0}0\epsilon_0\epsilon_1 \cdots \epsilon_K 0 0 \cdots \overset{\nu}{0} \cdots, \qquad \epsilon_0\epsilon_K \neq 0.$$

A.2 State representation

$\Delta_1, \Delta_2, \Delta_3, \cdots, \Delta_{K+\nu+1}$, where the states Δ_j are defined as the successive ν-tuples of the sequence representation, where the all-zero ν-tuples are omitted, except for the ν-tuple abutting ϵ_K.

$$(0, 0, \cdots, 0, \epsilon_0), (0, 0, \cdots, \epsilon_0, \epsilon_1),$$
$$\cdots, (\epsilon_K, 0, 0, \cdots 0), (0, 0, 0, \cdots 0).$$

A.3 Augmented state representation

$\Delta_1^+, \Delta_2^+, \Delta_3^+, \cdots \Delta_{K+\nu+1}^+$, where the augmented states Δ_j^+ are defined as the successive $(\nu + 1)$-tuples of the sequence representation

$$\underset{\nu+1}{(00 \cdots 0\epsilon_0)}, \underset{\nu+1}{(00 \cdots \epsilon_0\epsilon_1)} \cdots \underset{\nu+1}{(\epsilon_K 00 \cdots 0)}.$$

This representation derives its usefulness from the equality

$$\sum_{1}^{K+\nu+1} \langle h^b, \Delta_j^+ \rangle^2 = ||h*e||^2,$$

where $h^b \triangleq (h_\nu, h_{\nu-1}, \cdots, h_0)$ and the inner product is defined in the usual way. The b superscript is read "backward" and the b operation is also applied to error events in the memorandum.

A.4 Functional representation

The error sequence maps to the nonnegative cosine polynomial

$$\mathbf{E}(\omega) = |\epsilon_0 + \epsilon_1 e^{i\omega} + \cdots + \epsilon_K e^{iK\omega}|^2$$

on the interval $-\pi \leqslant \omega \leqslant \pi$.

We shall refer to each e as an error sequence, error event, or error pattern. Let $\mathbf{H}(\omega) \triangleq |h_0 + h_1 e^{i\omega} + \cdots + h_\nu e^{i\nu\omega}|^2$ and define $\langle \mathbf{E}, \mathbf{H} \rangle \triangleq 1/(2\pi) \int_{-\pi}^{p} \mathbf{E}(\omega)\mathbf{H}(\omega)\, d\omega$. Then by Parseval's theorem, $\langle \mathbf{E}, \mathbf{H} \rangle = ||e*h||^2$, so we have

$$d_{\min}^2(h) = \min_{\mathbf{E}} \langle \mathbf{E}, \mathbf{H} \rangle = \min_{\mathbf{E}} \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathbf{E}(\omega)\mathbf{H}(\omega) d\omega.$$

APPENDIX B

Expediting Tree Search When Response Is Symmetrical

We present some observations other than the four symmetry conditions that are useful for efficient calculation of $d^2_{\min}$ and a minimizing error event e for a symmetric transmitter impulse response.

Let $\{S_k\}_{-\infty}^{\infty}$ denote the resulting sequence of scalar products in $h*e$. Let K be the smallest integer satisfying $s_1 s_K \neq 0$ with $s_k = 0$ for $k \notin \{1, 2, \cdots K\}$. If, in the course of searching the tree, an error event breaking the current record, A, is encountered, then, for it to be a minimizing event, we must have

$$s_1^2 + s_2^2 + \cdots s_K^2 < A.$$

Let $[K/2]$ denote the largest integer less than or equal to $K/2$. For a record breaking error event or that error event in reverse (or both) we must have

$$s_1^2 + s_2^2 + \cdots s_{[K/2]}^2 < A/2. \tag{5}$$

Since $||h*e||^2 = ||h^*(\pm e^b)||_1^2$ error events for which it is established that the inequality (5) is reversed need not be explored further in the tree search for $d^2_{\min}$.

To expedite the process of seeking $d^2_{\min}$ among error sequences for which (5) holds, we discuss the calculation of the height at which the exploration of the growth of nodes terminates. Let L be the first integer for which the accumulated sum of $s_1^2 + s_2^2 + \cdots s_{L+1}^2 > A/2$ so that $s_1^2 + s_2 + \cdots s_L^2 \leqslant A/2$. Clearly, $L \geqslant [K/2]$ so $2L + 1 > K$. Once L is detected there is no need to explore any events involving $2L + 2$ scalar products. To put it another way, if $L' = L + 1$ is the first index for which $A/2$ is exceeded, then $s_{2L'} = 0$ and $2L' > K$.

It is not always necessary to search to height $2L'$ to terminate growth exploration. To see this note that from the height, ℓ', of occurrence of the last nonzero element in the event under exploration we have that $K \geqslant \ell' + \nu$. So once L' is determined exploration of growth is terminated if $\ell' + \nu \geqslant 2L'$.

AUTHOR

Gerard J. Foschini, B.S.E.E., 1961, Newark College of Engineering, Newark, NJ; M.E.E., 1963, New York University, New York; Ph.D., 1967 (Mathematics), Steven Institute of Technology, Hoboken, NJ; AT&T Bell Laboratories, 1961—. At AT&T Bell Laboratories Mr. Foschini initially worked on real-time program design. For many years he worked in the area of communication theory. In the spring of 1979 he taught at Princeton University. Mr. Foschini has supervised planning the architecture of data communications networks. Currently, he is involved with digital radio research. Member, Sigma Xi, Mathematical Association of America, IEEE, New York Academy of Sciences.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

Adaptive Transversal Equalization of Multipath Propagation for 16-QAM, 90-Mb/s Digital Radio

By G. L. FENDERSON,* J. W. PARKER,* P. D. QUIGLEY,* S. R. SHEPARD,* and C. A. SILLER, JR.*

(Manuscript received January 19, 1984)

Adaptive transversal equalization is an effective and relatively new countermeasure for dispersive multipath propagation in terrestrial digital radio networks. In this paper we describe the design and performance of a five-tap baseband analog equalizer developed for a family of 16-QAM, 90-Mb/s radio systems. Laboratory measurements and field evaluation during a five-month fading season in Palmetto, Georgia, indicate that the use of this adaptive transversal equalizer can significantly reduce the need for costly space-diversity equipment.

I. INTRODUCTION

The impairment of terrestrial digital microwave reliability due to multipath propagation is widely recognized.[1] Unlike FM radio systems, where system outage is predominantly determined by the thermal noise aspect of fading, digital radio is also affected by the dispersive character of multipath fading. This dispersion, engendered by significant amplitude and delay distortion across the channel bandwidth, causes considerable Intersymbol Interference (ISI) that degrades digital radio reliability well beyond that expected from the flat fade margin alone.[2] Multipath-induced distortion thus is considered the predominant cause of digital radio outage for frequencies under 12 GHz.

Presently, several methods are used to counter the impact of mul-

* AT&T Bell Laboratories.

tipath fading. These include frequency diversity,[3] space diversity,[4] and adaptive Intermediate Frequency (IF) equalization. Examples of the latter include slope equalizers[5] and notch or resonance equalizers.[6] However, frequency-selective fading corrupts both the amplitude and phase of a transmitted signal. While IF equalizers can be designed to condition a channel properly for minimum phase fading, they double the delay distortion during periods of nonminimum phase fading. (Minimum phase and nonminimum phase fading is clearly defined by Giger and Barnett[1] for a two-path statistical model of multipath propagation.) This effect naturally impacts the outage of those digital radio systems that rely solely on amplitude correction.[7]

Although adaptive transversal equalizers are a relatively new countermeasure to multipath fading in digital radio systems,[8] their prior application in mitigating the effects of linear distortion in other, lower-speed, digital communication networks is firmly established. Current practice is to use transversal equalizers in conjunction with IF equalizers. In a recent study, Foschini and Salz[9] considered the application of equalization techniques to digital data transmission over radio channels subject to frequency-selective fading. Their theoretical study of ideal transversal equalizers with an infinite number of taps clearly established the utility of linear equalization during multipath propagation. These equalizers are especially noteworthy in that they are capable of providing amplitude *and* delay equalization for minimum *and* nonminimum phase fades.

The baseband synchronous transversal equalizer briefly described here was designed for a family of 16-QAM (Quadrature Amplitude Modulation), 90-Mb/s radio systems. Designated DR 6-30 and DR 11-40, these digital systems provide 3-bit/Hz operation in the 6- and 11-GHz common carrier bands, respectively.[10] In this paper, we focus on the design and performance features of a high-speed (approximately 22.5-MHz) synchronous transversal equalizer. A theoretical development of equalization principles is specifically omitted since those points are amply discussed in the technical literature (for example, see Chapter 6 of Ref. 11).

II. GENERAL DESCRIPTION

2.1 DR 6/DR 11 radio system

Figure 1 functionally depicts the DR 6/DR 11 digital radio system. Two independent, 45-Mb/s random data streams are differentially encoded to form two rails, each with four-level amplitude states, and then modulated in quadrature to form a 16-QAM, 90-Mb/s IF signal at 70 MHz. The Radio-Frequency (RF) transmitter modulates the IF signal up to 6 or 11 GHz for transmission over a line-of-sight terrestrial path to the digital receiver. At the receiver the signal is down-converted

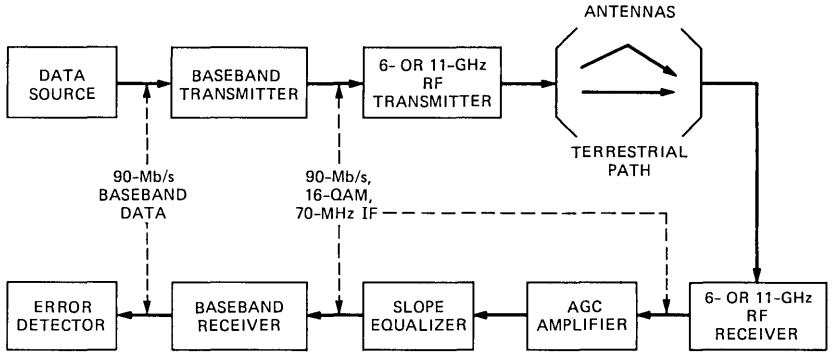


Fig. 1—DR 6/DR 11 digital radio system.

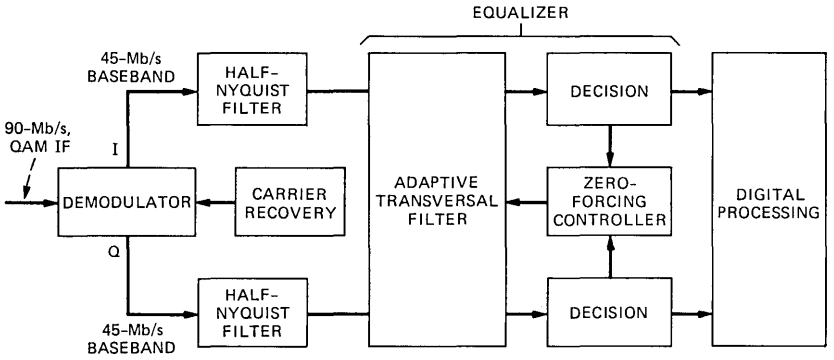


Fig. 2—Baseband receiver with adaptive transversal equalizer.

to IF, where it is processed by an Automatic Gain Control (AGC) amplifier and adaptive slope equalizer. The baseband receiver demodulates the IF signal into in-phase (I) and quadrature (Q) rails, where the baseband data states are detected and estimates of the original transmitted data are made. An error detector provides for system performance monitoring.

Figure 2 functionally illustrates the baseband receiver. As described above, the 90-Mb/s, QAM IF signal is demodulated into I and Q rails, each 45 Mb/s. After conventional half-Nyquist spectral shaping (using delay-equalized analog filters with raised-cosine shaping and 45-percent roll-off), the four-level signals enter the transversal equalizer for removal of the linear intersymbol interference previously generated by multipath propagation, imperfect Nyquist filtering, etc. After

baseband equalization, the transmitted symbol states are estimated at the decision point, and the decoded binary signals are used in subsequent digital processing.

2.2 Adaptive transversal equalizer

To remove in-rail and cross-rail intersymbol distortion, two adaptive transversal filters (each with five complex-valued tap weights) are configured for baseband equalization of QAM signals. The selection of five taps is based on theoretical studies of equalizer performance as a function of equalizer length. For example, Amitay and Greenstein[12] have investigated the multipath outage performance of digital radio receivers using finite-length adaptive equalizers. Using Rummler's statistical description of multipath channels,[13] equalizer performance for a broad ensemble of fading scenarios was simulated. Their study indicated that five synchronous taps considerably reduce ISI relative to performance attained with three taps and that equalizers with seven or more taps, while obviously further reducing ISI, exhibit a rapidly diminishing relative reduction in linear distortion. (Independently, Murase et al.,[14] and Takenaka et al.[15] have also selected five-tap filters for their transversal equalizer designs.)

The equalizer tapped-delay lines are fabricated using lumped-delay elements isolated with buffer amplifiers. The buffer amplifiers are Hybrid Integrated Circuits (HICs) and provide high isolation between the delay line and coefficient-weighting taps. Tap weighting is accomplished with variable gain amplifiers. These, too, are hybrid integrated circuits fabricated in single in-line packages, thereby permitting high-density electronics on each circuit board. Summing amplifiers (also HICs) then add the individual tap-weighted signals to form the filter output.

The coefficient control portion of the equalizer uses zero-forcing adaptation and is implemented with high-speed Emitter-Coupled Logic (ECL). The control circuit accepts error polarity and estimated-symbol polarity from the in-phase and quadrature decision circuits. Appropriately delayed versions of these bits are then correlated during each symbol period using exclusive OR gates. The time-averaged values of these correlations determine the weight of each tap in the two transversal filters. Time averaging is achieved using operational amplifier filters optimized for the trade-off between coefficient noise and dynamic multipath tracking ability.

The entire equalizer consists of three 1-inch plug-in circuit packs in a format compatible with the DR 6/DR 11 terminal or regenerative equipment. A photograph of these circuit packs appears in Fig. 3. Two of these packs are identical analog transversal equalizers, one for in-phase and cross-rail equalization of the I rail, the other for similar

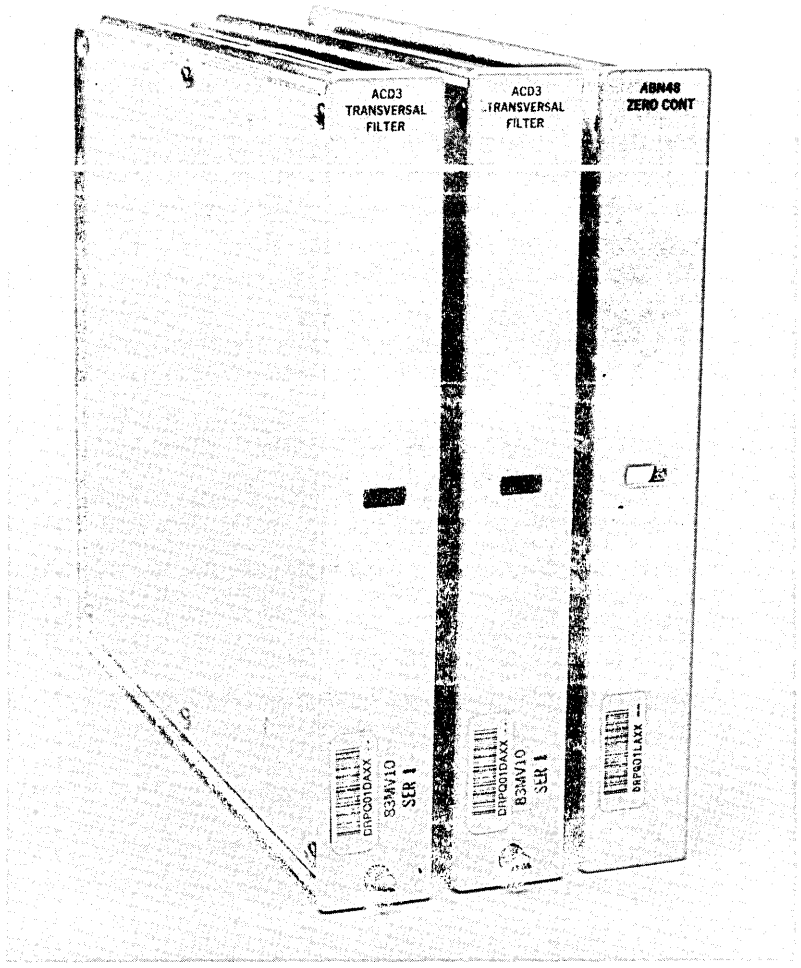


Fig. 3—Adaptive transversal equalizer consisting of two transversal filter circuit packs and one zero-forcing control circuit pack.

equalization of the Q rail. Equalizer coefficient control is generated in the third circuit package.

III. EQUALIZER PERFORMANCE

3.1 Theoretical performance

3.1.1 Reduction of peak distortion

As we noted above, five-tap synchronous transversal equalizers are theoretically capable of substantially reducing intersymbol interfer-

ence caused by frequency-selective fading. For zero-forcing coefficient adaptation, the peak distortion, D_p, of the corrupted digital signal is minimized.[11] (As used here, D_p is equivalent to Peak Eye Closure (PEC) for binary transmission.) Representative theoretical performance is illustrated in Fig. 4. In Fig. 4 we consider one digital rail (I or Q) and show the variation of peak distortion—with and without equalization—of a digital signal for a 20-dB fade notch depth as a function of notch position in a ±18-MHz channel. (Ideally, distortion in the other rail would be identical.) The ordinate on the right side of this figure provides the corresponding peak eye closure for a four-level signal, given by $PEC = D_p(L - 1)$, where L is the number of discrete transmitted symbol states on each rail. This illustrative fade grossly closes the digital eye with $D_p > 1$ over at least a portion of the channel

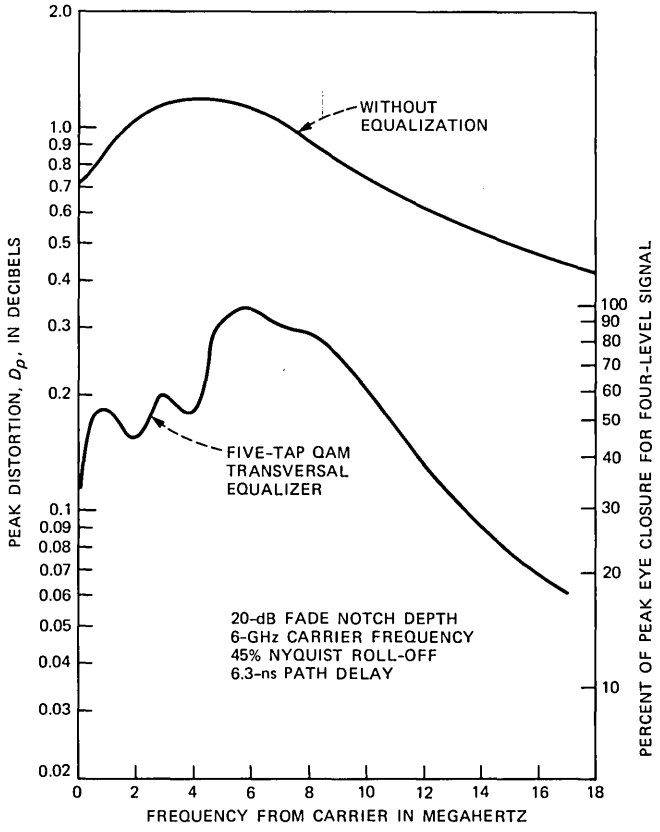


Fig. 4—Theoretical reduction in peak distortion, D_p, with a five-tap QAM transversal equalizer arrangement. Peak eye closure is noted for four-level transmission, that is, one rail of a 16-QAM system.

bandwidth. This latter condition highlights an analytic limitation of zero-forcing equalization: if $D_p > 1$, the coefficient set may be suboptimal.[11] In spite of this, other analysis (to be discussed shortly) and our own measured data show that adaptive transversal equalizers do, in fact, notably reduce intersymbol interference in just such an environment. Moreover, zero-forcing is known to assure a global minimum if $D_p < 1$, affords comparative ease of circuit realization, and minimizes Bit Error Rate (BER) in the high signal-to-noise ratio that typifies quiescent digital radio operation. The other dominant adaptation approach for automatic equalizers, Least-Mean-Square (LMS) algorithmic control, is more difficult to realize in high-speed circuits and has a proclivity for unsatisfactory local minima when used in the decision-directed mode.[16]

3.1.2 Equipment signatures

Equipment signatures[17,18] provide a particularly meaningful measure of equalization capability since they can be directly related to outage predictions for digital radio systems. The signatures are 10^{-3} BER contours: at each point on the contour, the fade notch depth corresponding to a 10^{-3} BER (defined as a digital radio outage) is specified as a function of notch position for a fixed-delay statistical model of multipath propagation. Figure 5 presents theoretical signatures computed by M. H. Meyers[19] for no equalization, slope equalization, and transversal equalization. Figures 4 and 5 confirm that five-tap transversal equalizers theoretically provide a significant reduction in linear distortion. Indeed, even the use of zero-forcing control for fades with $D_p > 1$ yields a degree of equalization that mitigates digital radio outage. The data of Fig. 5 indicate that a fade notch depth as shallow as 7 dB can cause an outage in the absence of countermeasures. When the radio receiver is equipped with a transversal equalizer, outages are not experienced until the notch depth reaches approximately 23 dB, which can occur for an unequalized $D_p > 1$, as shown in Fig. 4. Also observe from Fig. 5 that slope equalizer performance is influenced by the minimum or nonminimum phase character of the fade, as we mentioned earlier. This is not a limitation of transversal equalization.

3.2 Measured laboratory performance

3.2.1 Equipment signatures

The definition and significance of equipment signatures were previously mentioned. The laboratory measurement of these signatures is facilitated through the use of a new computer-controlled multipath fade simulator that continuously varies notch depth and notch frequency to achieve a 10^{-3} BER. The simulator is inserted in the IF path of the receiver just before the AGC amplifier (see Fig. 1).

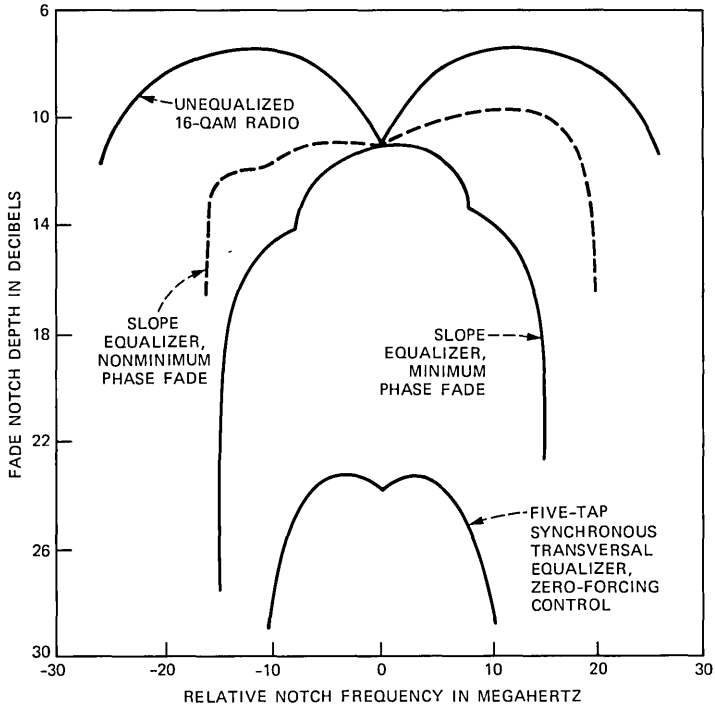


Fig. 5—DR 6 theoretical equipment signatures for 16-QAM digital radio. Performance for radio without equalization, with an adaptive slope equalizer, and with a five-tap synchronous transversal equalizer using zero-forcing control.

Signatures were measured using two DR 6 receivers, the first equipped with an adaptive slope equalizer (the standard arrangement) and the second equipped with both the adaptive slope equalizer and a five-tap adaptive transversal equalizer. The 10^{-3} BER minimum phase and nonminimum phase equipment signatures appear in Fig. 6. As the data reveal, the adaptive slope equalizer performs best when used for minimum phase fades, with a performance deterioration experienced for nonminimum phase fades. We commented earlier that IF equalizers typically double delay distortion during nonminimum phase fading, and this effect can degrade equipment signature performance. The same effect naturally occurs when the adaptive slope and synchronous transversal equalizers are used together. Comparing both sets of curves, however, we also observe the significant improvement in equipment signature performance that can be ascribed to the transversal equalizer alone.

The relative reduction in digital radio outage time is estimated using a prescription described by Meyers,[21] wherein the areas under equipment signature contours, with and without transversal equalization,

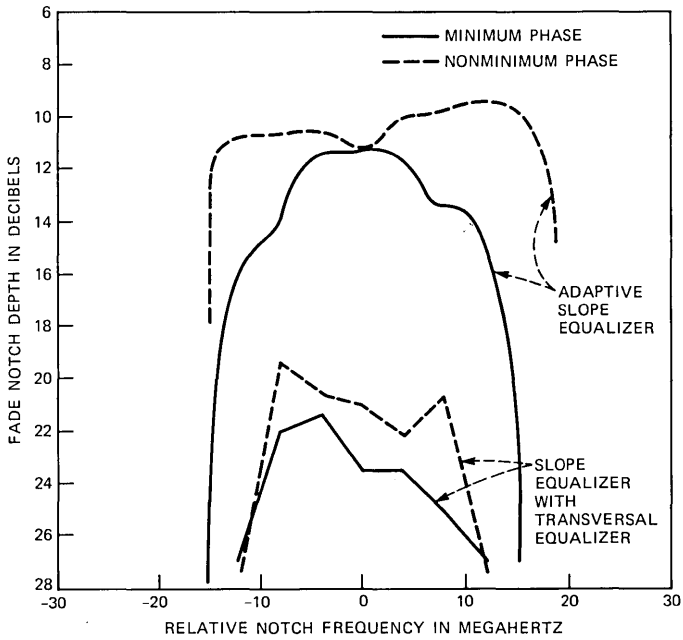


Fig. 6—DR 6 measured equipment signatures for 16-QAM digital radio. Performance for radio with adaptive slope equalization and adaptive slope and transversal equalization for 6.3-ns path delay. (Adapted from Ref. 20.)

are compared. The predicted relative outage reduction factor, derived from theoretical equipment signatures for combined ideal slope and transversal equalization (see Fig. 5) is 5. This assumes equally probable minimum and nonminimum phase fading. The predicted relative outage reduction factor for the measured equipment signatures (see Fig. 6) is 4.5, again assuming equally probable minimum and nonminimum phase fading. The relative reduction factors for other ratios of minimum to nonminimum phase fading range from 4 to 5. The measurements in Fig. 6 attest to the quality of the transversal equalizer circuit design itself. Regarding this point, baseband implementation of the equalizer permits integration of substantial portions of the circuitry, thus simplifying design and manufacture. The development of new carrier and timing recovery circuits also helps place laboratory performance close to the theoretical limit shown in Fig. 5.

3.2.2 Simulation of dynamic fading

An important aspect of multipath propagation is its rapid temporal variation. To assure optimal equipment performance in the field, dynamic (time-varying) tests were performed during the development phase. Dynamic multipath fading is produced in the laboratory by

controlling the continuously variable fade simulator with a microcomputer. Realistic time sequences of multipath behavior were programmed into the simulator. Equalizer performance was monitored during the simulation of these dynamic fades, thereby permitting optimization of the equalizer timing-recovery, carrier-recovery, and coefficient-updating loop parameters.

Several aspects of an equalizer's response to dynamic multipath propagation are exercised with the following test sequence (schematically depicted in Fig. 7): starting with a shallow fade notch depth d_1 at a particular notch frequency f_1, the notch depth increases at a rate s_1 until a notch depth d_2 is reached. The notch then sweeps across a band of frequencies from f_1 to f_2 at a rate s_2. At the notch frequency f_2, the notch depth decreases from d_2 back to d_1 at a rate s_3. This fading trajectory retraces itself and is repeated several times for statistical averaging of the receiver's error performance. A test sequence like this tests the receiver's ability to track notch depth and notch frequency dynamics. For trajectory parameters of $d_1 = 6$ db, $d_2 = 15$ dB, $s_1 = s_3 = 9$ dB/s, $f_1 = -12$ MHz (12 MHz below the IF frequency), $f_2 = +12$ MHz, and $s_2 = 24$ MHz/s, the transversal equalizer consistently operates error free. Those test velocities are also faster than 90 percent of all observed notch depth and notch position rates of change reported by Sakagami et al.[22]

3.3 Field evaluation

The adaptive transversal equalizer was installed in a DR 6-30 field test facility at Palmetto, Georgia, on June 4, 1982. This modified

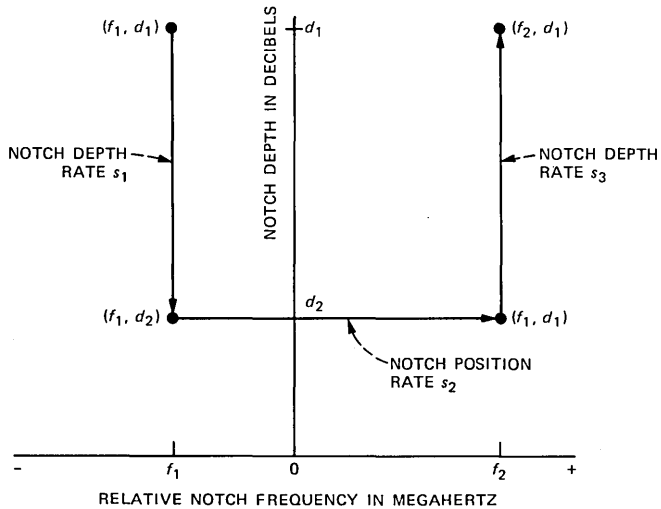


Fig. 7—Test sequence for dynamic simulation of multipath propagation.

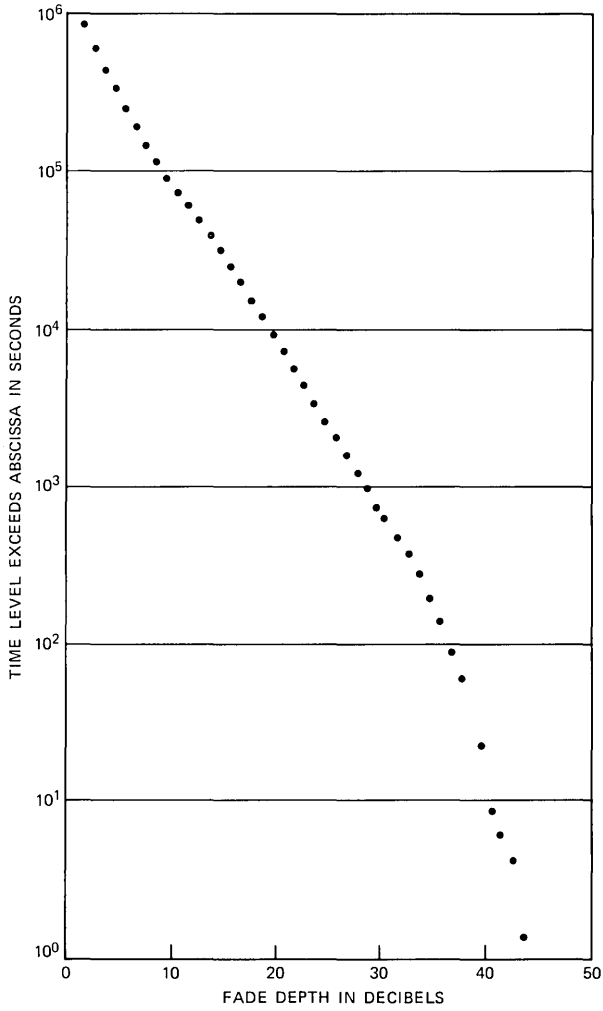


Fig. 8—Time-below-level propagation data for Palmetto, Georgia, in 1982.

baseband receiver was compared with a standard DR 6 receiver (equipped with an adaptive slope equalizer) during a multipath season from June 6 to November 6, 1982. Propagation data collected during the field evaluation period are shown in Fig. 8.[23] The abscissa of this figure reports fade notch depth; the ordinate indicates time faded below the respective abscissa value. A considerable amount of fading exhibits notch depths in excess of 10 dB, which, from Fig. 5, could correspond to an outage in the absence of suitable countermeasures. The two baseband receivers shared the same RF and IF front ends. Field measurements, monitored by AT&T Bell Laboratories personnel

from Merrimack Valley, were grouped into 11 two-week intervals. In Fig. 9, the number of seconds for which BER > 10^{-6} is presented for both receivers for each of the two-week intervals. Also presented is the ratio of these two time measurements, representing a composite improvement factor attributable to the transversal equalizer, alone. Figure 10 presents similar data for a BER > 10^{-4}. In Fig. 11 we show the incidence of frame loss with and without the equalizer, as well as the corresponding reduction in loss of frame.

For the 22 weeks represented in Figs. 9 through 11, the adaptive

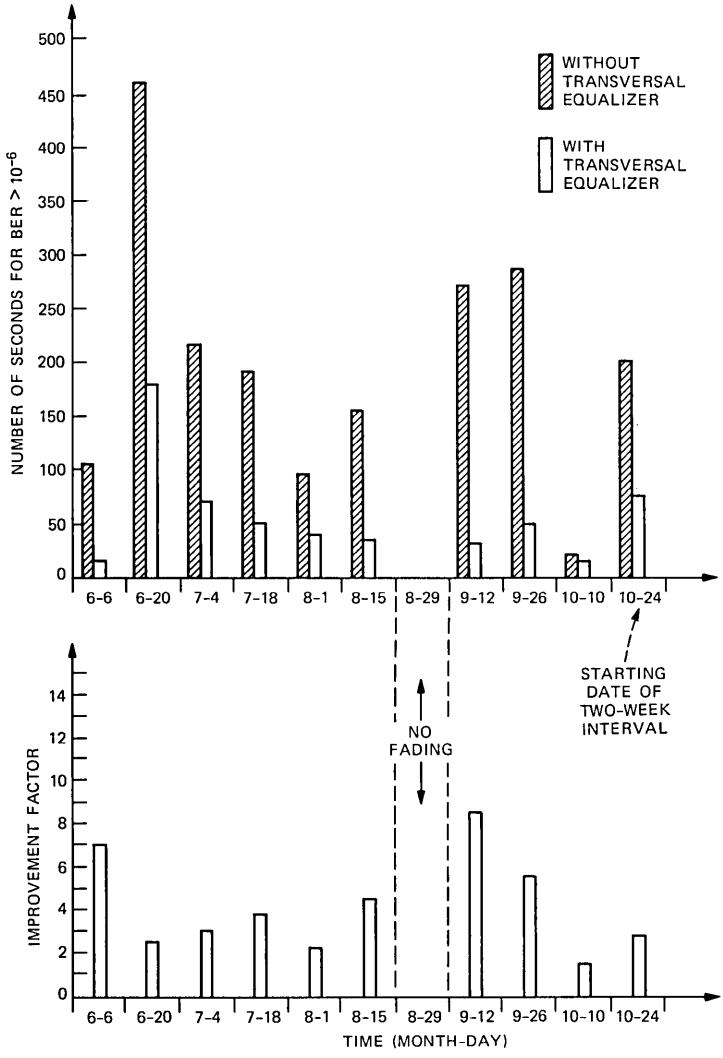


Fig. 9—Field performance for BER > 10^{-6}.

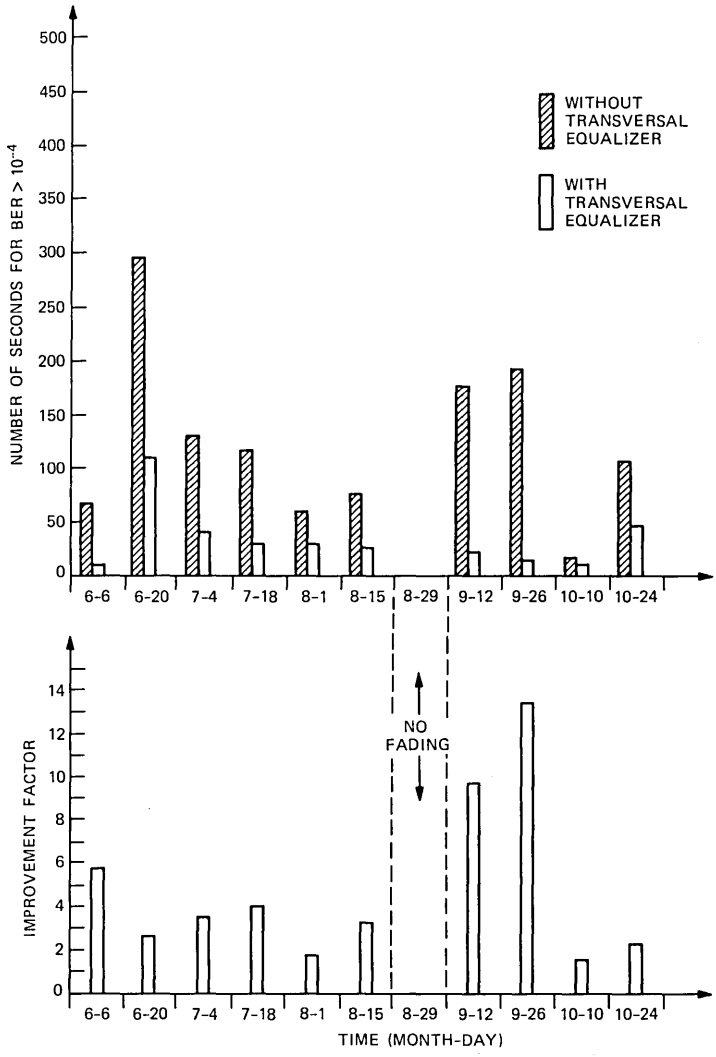


Fig. 10—Field performance for BER > 10^{-4}.

transversal equalizer provided composite improvement factors of 3.6 for BER > 10^{-6}, 3.7 for BER > 10^{-4}, and 2.9 for frame loss. The 10^{-4} BER improvement factor of 3.7 is only 20 percent below the predicted improvement factor of 4.5, based on laboratory-measured equipment signatures.

IV. CONCLUSIONS

Because of their ability to adaptively equalize multipath-induced amplitude and delay distortion for minimum and nonminimum phase

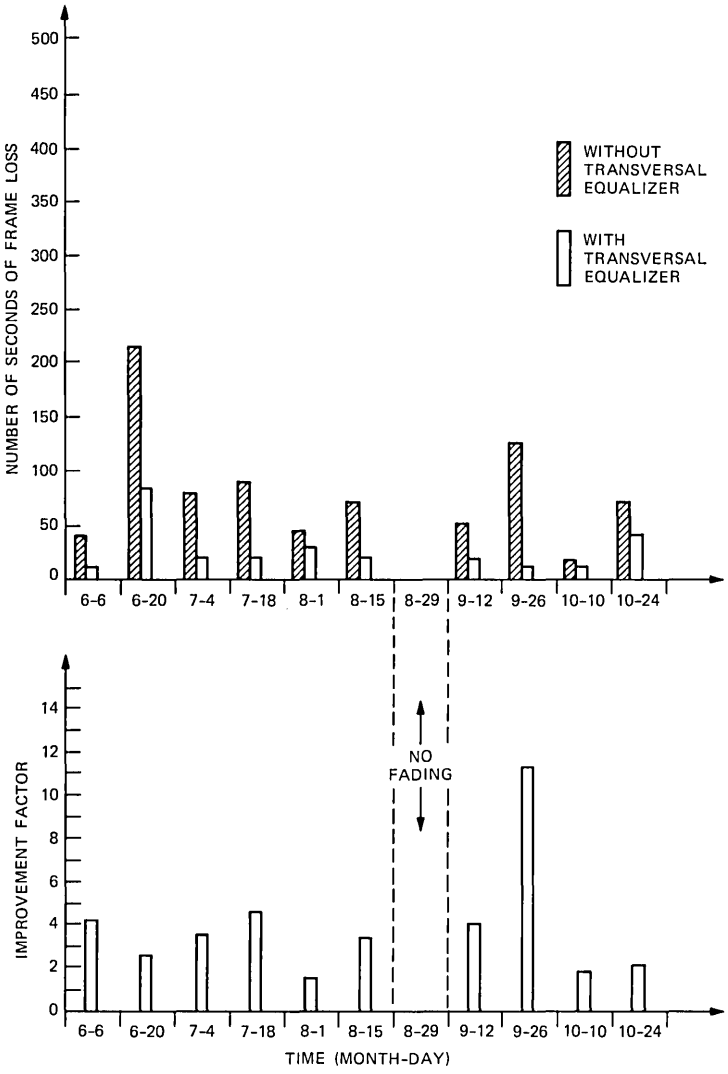


Fig. 11—Field performance for frame loss.

fading, synchronous transversal equalizers promise to play an important role as a multipath countermeasure for terrestrial digital microwave networks.

In this paper we summarize the major design and performance features of a five-tap analog transversal equalizer for the baseband receivers of two 16-QAM, 90-Mb/s digital radio systems. The equalizers heavily rely on HIC technology for their tapped-delay line buffer amplifiers, tap-weighting coefficients, and summing circuitry. The

zero-forcing adaptation portion of the equalizer is realized with high-speed ECL logic. The entire equalizer is packaged in three 1-inch plug-in circuit packs.

During design, the equalizer was tested for its static equipment signature performance and dynamic tracking capability. The latter evaluation was facilitated with a special-purpose, computer-controlled multipath fade simulator. During a 22-week field trial evaluation in Palmetto, Georgia, the equalizer reduced the overall incidence of DR 6-30 radio outage by more than a factor of 3. System estimates indicate that this improvement factor could eliminate the need for space-diversity reception on more than 50 percent of the short-haul digital radio hops that currently use it. Use of the baseband adaptive transversal equalizer thus can provide considerable cost savings.

V. ACKNOWLEDGMENTS

Completion of this project required the synergistic efforts of many individuals. The authors are pleased to acknowledge the following contributors: J. S. Bitler, for designing many of the analog integrated circuits; G. L. Frazer, for special insights into carrier and timing recovery; M. H. Meyers, for theoretical BER and equipment signature calculations; K. L. Seastrand, for initially proposing transversal equalization in our digital radio systems; M. A. Skinner, for overseeing the project during its development phase; and R. B. Ward, for contributions to the carrier recovery design.

REFERENCES

1. A. J. Giger and W. T. Barnett, "Effects of Multipath Propagation on Digital Radio," IEEE Trans. Commun., *COM-29*, No. 9 (September 1981), pp. 1345–52.
2. C. W. Anderson et al., "The Effect of Selective Fading on Digital Radio," IEEE Trans. Commun., *COM-27*, No. 12 (December 1979), pp. 1870–5.
3. A. Vigants and M. V. Pursley, "Transmission Unavailability of Frequency-Diversity Protected Microwave FM Radio Systems Caused by Multipath Fading," B.S.T.J., *58*, No. 8 (October 1979), pp. 1779–96.
4. A. Vigants, "Space Diversity Engineering," B.S.T.J., *54*, No. 1 (January 1975), pp. 103–42.
5. W. J. Schwarz, Jr., et al., "Radio Repeater Design for 16 QAM," ICC '79 (June 1979), pp. 48.2.1–6.
6. E. R. Johnson, "An Adaptive IF Equalizer for Digital Transmission," ICC '81 (June 1981), pp. 13.6.1–4.
7. S. Komaki et al., "Characteristics of a High Capacity 16 QAM Digital Radio System in Multipath Fading," IEEE Trans. Commun., *COM-27*, No. 12 (December 1979), pp. 1854–61.
8. C. A. Siller, Jr., "Multipath Propagation: Its Associated Countermeasures in Digital Microwave Radio," IEEE Commun. Mag., *22*, No. 2 (February 1984), pp. 6–15.
9. G. J. Foschini and J. Salz, "Digital Communications Over Fading Radio Channels," B.S.T.J., *62*, No. 2 (February 1983), pp. 429–56.
10. J. J. Kenny, "Digital Radio for 90-Mb/s, 16-QAM Transmission at 6 and 11 GHz," Microw. J. (August 1982), pp. 71–80.
11. R. W. Lucky et al., *Principles of Data Communication*, New York: McGraw-Hill, 1968.
12. N. Amitay and L. J. Greenstein, "Multipath Outage Performance of Digital Radio Receivers Using Finite-Tap Adaptive Equalizers," NATO/AGARD 33rd Sympo-

sium of the Electromagnetic Wave Propagation Panel, Spatind, Norway, October 4–7, 1983.
13. W. D. Rummler, "A New Selective Fading Model: Application to Propagation Data," B.S.T.J., *59*, No. 5 (May–June 1979), pp. 1037–71.
14. T. Murase et al., "200 Mb/s 16-QAM Digital Radio System With New Countermeasure Techniques for Multipath Fading," ICC '81 (June 1981), pp. 46.1.1–5.
15. S. Takenaka et al., "A Transversal Fading Equalizer for a 16-QAM Microwave Digital Radio," ICC '81 (June 1981), pp. 46.2.1–5.
16. J. E. Mazo, "Analysis of Decision-Directed Equalizer Convergence," B.S.T.J., *59*, No. 10 (December 1980), pp. 1857–76.
17. M. Emshwiller, "Characterization of the Performance of PSK Digital Radio Transmission in the Presence of Multipath Fading," ICC'78 (June 1978), pp. 47.3.2–6.
18. C. W. Lundgren and W. D. Rummler, "Digital Radio Outage Due to Selective Fading—Observation vs. Prediction From Laboratory Simulation," B.S.T.J., *58*, No. 5 (May–June 1979), pp. 1073–100.
19. M. H. Meyers, personal communication.
20. C. P. Bates and M. A. Skinner, "Impact of Technology on High-Capacity Digital Radio Systems," ICC'83 (June 1983), pp. F2.3.1–5.
21. M. H. Meyers, unpublished work.
22. S. Sakagami et al., "Inband Amplitude Dispersion Characteristics During Multipath Fading on Microwave Links," Rev. Elec. Commun. Lab., *29* (November–December 1981), pp. 1295–303.
23. A. Ranade, personal communication.

AUTHORS

Gerald L. Fenderson, B.S.E.E., 1960, University of Maine; M.S.E.E., 1963, Northeastern University; AT&T Bell Laboratories, 1960—. Since joining AT&T Bell Laboratories, Mr. Fenderson has participated in the development of FM and digital microwave radio systems. His most recent responsibility has been in the design of digital modems, including adaptive transversal equalization. Mr. Fenderson received an AT&T Bell Laboratories Distinguished Technical Staff Award in December 1982. Member, Tau Beta Pi, Eta Kappa Nu.

James W. Parker, A.S.M.E.T., 1978, Vermont Technical College; AT&T Bell Laboratories, 1978—. Mr. Parker has worked on a variety of physical design projects for digital radio systems and is currently involved in the design and development of advanced and international radio bays. Member, Tau Alpha Pi.

Patrick D. Quigley, A.S.E.E.T., 1978, S.U.N.Y. at Alfred; B.S.E.E.T. 1983, University of Lowell, Lowell, MA; AT&T Bell Laboratories, 1978—. Mr. Quigley has worked on analog, digital, and firmware development projects for the AR6 and DR6 radio systems. He is presently a Member of Technical Staff in the Microwave Radio Systems department, working on the development of a digital monitoring receiver for advanced digital radio systems.

Scott R. Shepard, B.S. (Electrical Engineering) and B.S. (Physics), 1979, Kansas State University; M.S. (Electrical Engineering), 1981, The Massachusetts Institute of Technology; AT&T Bell Laboratories, 1979—. While at K.S.U., Mr. Shepard was employed as a research assistant for various problems in electromagnetic theory, including the design and construction of optical logic devices and the focusing of charged particles in a linear accelerator. His master's research involved the analysis of a nonlinear optical pulse-shaping device by means of symbolic manipulation software at M.I.T.'s artificial

intelligence laboratory. Mr. Shepard's major responsibilities at AT&T Bell Laboratories have been the design of the adaptive transversal equalizer and the design of the dynamic multipath fade simulator. Currently, in addition to channel modeling and channel conditioning, he is involved in high-speed, high-accuracy, analog-to-digital and digital-to-analog signal conversion. Member, IEEE, Eta Kappa Nu, Sigma Pi Sigma, Tau Beta Pi, Phi Kappa Phi, and Sigma Xi.

Curtis A. Siller, Jr., B.S.E.E., 1966, M.S. (Plasma Physics), 1967, Ph.D. (Electrical Engineering), 1969, The University of Tennessee, at Knoxville; AT&T Bell Laboratories, 1969–1978, 1979—. Mr. Siller's earliest experience at AT&T Bell Laboratories was in the analysis and design of reflector antennas for terrestrial microwave communications. He subsequently initiated an exploratory investigation of adaptive transversal equalization for advanced digital radio systems. His more recent research interests were in digital signal processing, particularly equalizer control algorithms and new techniques for digital filtering. Mr. Siller is presently involved in system engineering of future digital transmission systems. Mr. Siller is the recipient of an AT&T Bell Laboratories Distinguished Technical Staff Award. Member, Phi Eta Sigma, Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi, Sigma Xi, American Association for the Advancement of Science, and the IEEE, where he is a Senior Member and serves on the Signal Processing and Communication Electronics Technical Committee.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

Enhancement of ADPCM Speech by Adaptive Postfiltering

By V. RAMAMOORTHY* and N. S. JAYANT†

(Manuscript received February 8, 1984)

Adaptive Differential Pulse Code Modulation (ADPCM) systems can provide high-quality digitizations of telephone-bandwidth speech at a bit rate of 32 kb/s. At a lower bit rate such as 24 kb/s, the quality of the speech is limited by an easily perceptible level of quantization noise. This paper proposes an adaptive postfiltering procedure that can significantly enhance the quality of lower bit rate ADPCM. The coefficients of the postfilter are easily derivable from the predictor coefficients in the ADPCM decoder. In a subjective test involving 18 listeners and two sentence-length test inputs, the enhanced 24-kb/s speech with an optimized postfilter design ranks very close to conventional 32-kb/s speech. A suggested application of the postfiltering procedure is in packet voice or mobile radio systems where substandard bit rates such as 24 kb/s or 16 kb/s are sometimes necessary. The postfiltering algorithm has also been successfully tested in non-DPCM situations, such as in the enhancement of speech degraded by additive white Gaussian noise.

I. INTRODUCTION

Recent algorithms for adaptive prediction[1] and adaptive quantization[2] have led to the realization of high-quality ADPCM systems at 32 kb/s. This bit rate is the result of 8-kHz sampling and quantization using 4 bits/sample. The quality of 24-kb/s speech using the same prediction algorithm and 3-bits/sample coding is limited by a clearly perceptible level of quantization noise. This paper proposes a very simply implemented postfiltering algorithm, which provides a significant enhancement of 24-kb/s quality. In a subjective test to be

* University of Linköping, Sweden. † AT&T Bell Laboratories.

described later in this paper, the enhanced 24-kb/s ADPCM system was ranked very close to conventional 32-kb/s ADPCM.

A natural application of the postfiltering procedure would be in variable bit rate ADPCM systems such as packet networks or mobile radio where substandard bit rates such as 3 or 2 bits/sample are occasionally encountered. The postfiltering technique described in this paper is particularly effective at the bit rate of 3 bits/sample. It is also effective in non-DPCM situations such as in the enhancement of speech degraded by additive white Gaussian noise. When the signal-to-noise ratio (s/n) at the input to the postfilter is too low (as in 2-bit ADPCM or with white Gaussian noise at a relative noise level exceeding approximately −3 dB), noise suppression can only be achieved at the expense of severe distortion of the speech signal itself. When the 24-kb/s ADPCM is enhanced, the introduction of speech distortion is perceptible, but the effect of noise reduction is by far the more dominant phenomenon.

II. A SEMIQUANTITATIVE EXPLANATION OF THE POSTFILTERING TECHNIQUE

The philosophy of the postfiltering technique is represented in Fig. 1. Part (a) of the figure shows a signal spectrum with two narrowband components in the frequency regions W_1 and W_2, and a flat noise spectrum that is 15 dB below the first signal component but 5 dB above the second signal component. An ideal postfilter for this situation would have a gain of unity (0 dB) in the regions W_1 and W_2 and a gain of zero ($-\infty$ dB) in the rest of the frequency range. In real speech applications, implementation of such all-or-none responses is impractical except in the special cases where the stopband regions of the postfilter are merged into a single contiguous frequency region as in a low-pass or high-pass postfilter.[3,4]

A more practical approach, proposed in this paper, is the use of a postfilter frequency response that peaks in the regions W_1 and W_2, but is significantly lower in the rest of the frequency range. Figure 1b illustrates an extreme example of this approach. Here, the transfer function of the postfilter is chosen to be identical to the input signal spectrum in Fig. 1a. The resulting spectra of postfiltered signal and postfiltered noise preserve the original signal-to-noise ratios of 15 dB and −5 dB in the regions W_1 and W_2, respectively. However, the noise in the rest of the illustrated frequency range is now much lower, relative to the signal levels, than in part (a) of the figure. Specifically, the signal-to-background-noise ratios for regions W_1 and W_2 are now 45 dB and 10 dB, in place of 15 dB and −5 dB in the absence of postfiltering. This suppression of background noise also implies that

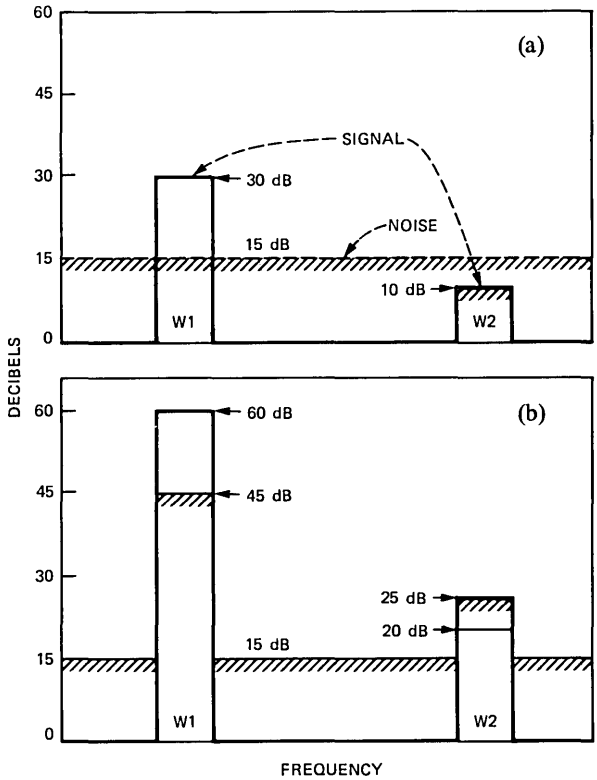


Fig. 1—An idealized explanation of the effects of postfiltering, assuming a signal with two narrowband components and a noise spectrum that is white. (a) Signal and noise spectra at the input to the postfilter, showing signal-to-noise ratios of 15 dB and −5 dB in signal frequency bands W_1 and W_2. (b) Spectra of postfiltered signal and postfiltered noise, assuming a postfilter transfer function identical to the signal spectrum in (a). Regions W_1 and W_2 continue to have local signal-to-noise ratios of 15 dB and −5 dB as in (a), but the signals are now 45 dB and 10 dB above the out-of-band noise level. In (a) the corresponding numbers are only 15 dB and −5 dB. The overall effect is a reduction of perceived noise, but the price paid is a change in the relative strengths of the signal components in W_1 and W_2.

the residual noise spectrum after postfiltering is very similar to the input signal spectrum itself. In speech applications, noise that is shaped in this manner tends to be perceived as speech.

A postfiltering operation such as that in Fig. 1b provides a significant amount of signal enhancement for the reasons just described. It should be noted, however, that such a postfilter also distorts the signal. For example, the difference in signal levels in the regions W_1 and W_2 has been distorted, from 20 dB in Fig. 1a to 35 dB in Fig. 1b. The postfiltering technique to be described in the next section provides a controlled exchange between signal distortion and noise suppression.

In the applications discussed, the technique realizes a broad range of postfilter design over which the phenomenon of noise suppression dominates the phenomenon of signal distortion.

Although speech enhancement[5] is an "ancient" art, we believe that the adaptive postfiltering technique discussed in the next section is novel. It can be used as a very general technique for speech enhancement. It can also be used very effectively in the specific context of ADPCM noise. The coefficients of the proposed postfilter are inspired by the coefficients of the adaptive predictor in ADPCM coding, and are in fact very closely related to these coefficients.

III. POSTFILTERING OF ADPCM SPEECH

Figures 2 and 3 provide block diagram descriptions of ADPCM with adaptive postfiltering.

Figure 2 shows the decoder part of the system. Broken lines in the figure refer to parts of the system that compute the coefficients of the adaptive predictor and the adaptive postfilter. The coefficients used in the postfilter are differently scaled versions of the coefficients used in the adaptive predictor. These coefficients are already available in conventional ADPCM. In the case of a system with Backward-Adaptive Prediction (APB), the predictor coefficients are updated in gradient-search algorithms driven by a recent history of the input and output of the ADPCM decoder.

A more complete block diagram of the ADPCM system appears in Fig. 3. The quantizer and predictor assumed in this paper are both

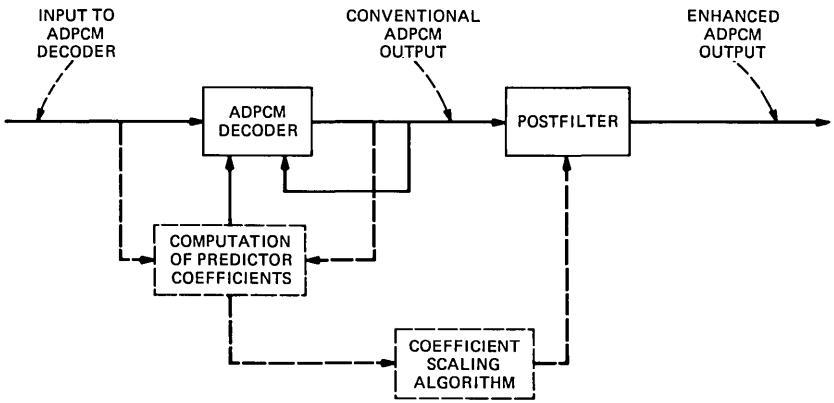


Fig. 2—Adaptive postfiltering of the output of an ADPCM decoder. The coefficients of the postfilter are scaled versions of the coefficients of the adaptive predictor in DPCM. In DPCM-APB, the predictor coefficients are obtained on the basis of observations of a recent history of decoder input and decoder output. The parts of the circuit that determine coefficient values are shown by broken lines.

backward-adaptive devices, implying that no special side information needs to be explicitly transmitted to the ADPCM decoder to enable adaptations of quantizer step size and predictor coefficients.

The adaptive quantizer assumed in this paper is one based on the use of a one-word memory,[6] but the results of this paper are fully expected to extend to a system that may use the more generalized version of this quantizer, as described in Ref. 2. In the quantizer used in this paper, the ratio of maximum step size to minimum step size is 512, and the minimum step size is in the order of 2^{-12} times the peak-to-peak value of the speech input $x(n)$.

This adaptive predictor we assumed is a pole-zero predictor, similar to that in Ref. 1. As Fig. (3) shows, the predicted value $\hat{x}(n)$ of input $x(n)$ is a combination of two components, the outputs $\hat{x}_z(n)$ and $\hat{x}_p(n)$ of an all-zero predictor $B(z)$ and an all-pole predictor $A(z)$. Formally,

$$\hat{x}(n) = \hat{x}_z(n) + \hat{x}_p(n)$$

$$\hat{x}_p(n) = \sum_{j=1}^{2} a_j(n)y(n-j)$$

$$\hat{x}_z(n) = \sum_{j=1}^{6} b_j(n)u(n-j),$$

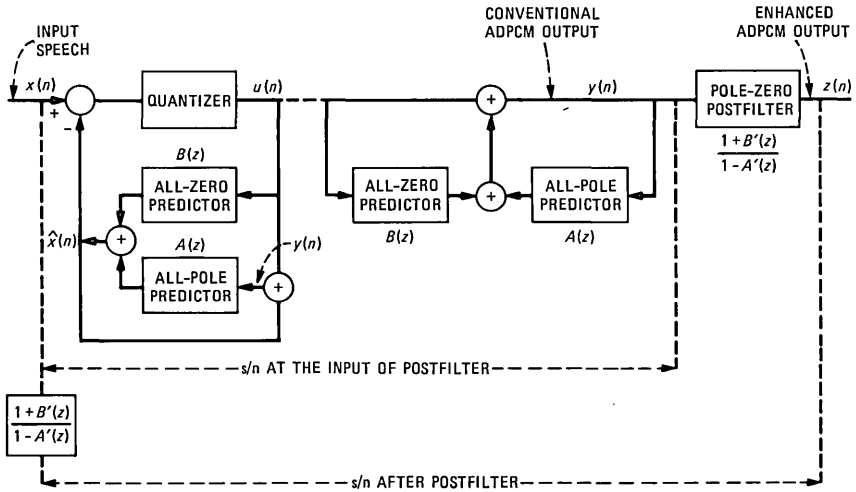


Fig. 3—Complete block diagram of an ADPCM system with a pole-zero predictor [defined by all-zero and all-pole components $A(z)$ and $B(z)$] and a pole-zero postfilter [defined by components $A'(z)$ and $B'(z)$ that are derived from $A(z)$ and $B(z)$]. The extreme case of $A'(z) = B'(z) = 0$ results in conventional ADPCM without postfiltering. The case of $A'(z) = A(z)$ and $B'(z) = B(z)$ results in a postfilter transfer function that is identical to the input signal spectrum, as in Fig. 1b.

where $u(n)$ is the quantized version $Q[\cdot]$ of the prediction error and $y(n)$ is the reconstructed output:

$$u(n) = Q[x(n) - \hat{x}(n)]$$
$$y(n) = \hat{x}(n) + u(n).$$

Adaptation of the predictor coefficients a_j and b_j follow the updating algorithms[7]

$$a_j(n) = \lambda_j a_j(n-1) + \mu_j \text{sgn}[u(n-1)]\text{sgn}[y(n-1-j)]$$
$$j = 1, 2; \quad \lambda_1 = 511/512; \quad \lambda_2 = 255/256; \quad \mu_1 = \mu_2 = 0.008$$

and

$$b_j(n) = \lambda_j' b_j(n-1) + \mu_j' \text{sgn}[u(n-1)]\text{sgn}[u(n-1-j)]$$
$$j = 1 \text{ to } 6; \quad \lambda_j' = 255/256 \quad \text{and} \quad \mu_j' = 0.008 \quad \text{for all } j.$$

The coefficients of the all-pole predictor are further controlled, for stability reasons, by the following constraints:

$$-0.75 \leq a_2 \leq 0.97$$
$$|a_{1,\max}| = 0.97 - a_2; \quad |a_1| = \min\{|a_1|, |a_{1,\max}|\}$$
$$a_1 = |a_1| \text{sgn } a_1.$$

3.1 Coefficients of the postfilter

A good starting point for designing the postfilter is the frequency response of the inverse predictor. This is the system whose input and output are the innovations $u(n)$ and the reconstruction $y(n)$. Its transfer function, derivable from linear equations that relate $u(n)$, $\hat{x}(n)$, and $y(n)$, is

$$\frac{Y(z)}{U(z)} = \frac{1 + B(z)}{1 - A(z)},$$

where

$$A(z) = \sum_{j=1}^{2} a_j z^{-j}; \quad B(z) = \sum_{j=1}^{6} b_j z^{-j}.$$

The speech-like transfer function of Fig. 1b is approximated if the postfilter response is identical to the function $[Y(z)]/[U(z)]$. This is because the spectrum of the quantized innovations $u(n)$ is approximately white and that of the reconstruction $y(n)$ is hopefully an approximation to that of the input $x(n)$. More generally, as in Fig. 3, we propose a postfilter transfer function

$$F(z) = \frac{1 + B'(z)}{1 - A'(z)},$$

where

$$A'(z) = \sum_{j=1}^{2} a_j \alpha^j z^{-j}; \qquad B'(z) = \sum_{j=1}^{6} b_j \beta^j z^{-j}$$

$$0 \leq \alpha \leq 1; \quad \text{and} \quad 0 \leq \beta \leq 1.$$

The extreme situation of Fig. 1b is approximated when $\alpha = \beta = 1$. In practice, this approximation can be quite poor because of the effects of nonideal predictor adaptation, usually resulting in an inverse predictor transfer function that is a flattened version of the input speech spectrum, with poles and zeros that may also be significantly shifted from their original locations. The case of $\alpha = \beta = 0$ corresponds to conventional ADPCM without any postfiltering. As we discuss in the next section, intermediate designs provide different mixes of noise suppression and speech distortion.

Figure 4 shows an illustrative spectrum of input speech and compares it with the transfer functions $F(z)$ for ($\alpha = 0.2$; $\beta = 1.0$) and ($\alpha = 1.0$; $\beta = 1.0$). The latter condition simply corresponds to the transfer function of the inverse predictor.

IV. EXPERIMENTAL RESULTS WITH ADPCM SPEECH

The speech inputs used in the experiment were the sentence-length

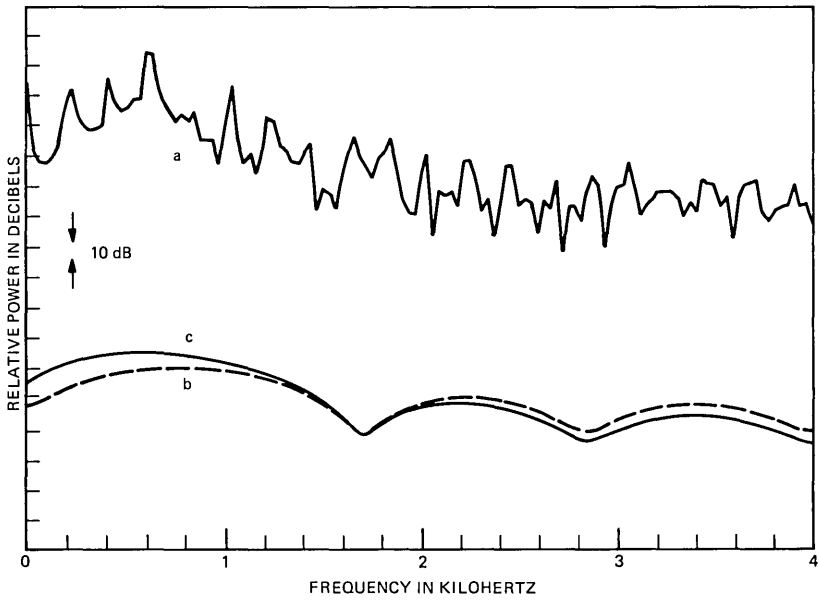


Fig. 4—(a) Input speech spectrum; and power transfer functions of postfilter with scaling coefficients for (b) $\alpha = 0.2$, $\beta = 1.0$, and for (c) $\alpha = 1.0$, $\beta = 1.0$. The plot (c) is merely the transfer function of the inverse predictor in the DPCM-APB system. [The 0-dB line is the same for (b) and (c) but different for (a)].

utterances "The Lathe is a big tool" and "The chairman cast three votes," bandlimited to 3.2 kHz in each case and sampled at 8 kHz. These inputs will be referred to as L8 and C8, respectively.

4.1 Signal-to-noise ratio results

Figures 1a and 1b indicate that postfiltering can result in significant improvements in s/n. Table I further demonstrates this for the examples of 3-bit and 2-bit DPCM. The results tabulated are the values of the s/n at the input of the postfilter and the s/n after postfiltering. (See Fig. 3.) Table I also shows corresponding values of the segmental s/n. In the ranges $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ for the coefficient scaling factors, the greatest gains in the s/n are obtained when $\alpha = \beta = 1$. These gains are seen to be as high as 8.9 dB for both L8 and C8. The gains of the s/n at the input of the postfilter are always lower for the design $\alpha = 0.2$ and $\beta = 1.0$. But we presently note that these settings of α and β provide a subjectively desirable design.

4.2 Subjective results

Tables II and III provide the results of a subjective test involving 14 listeners, including 9 from the AT&T Bell Laboratories Acoustics Research department and 5 listeners who had no prior exposure to speech coding experimentation or testing. A total of eight stimuli were included in the test. These included 4-bit ADPCM without postfiltering, 3-bit ADPCM with six postfiltering conditions (including the no-postfiltering case of $\alpha = \beta = 0$), and 4-bit ADPCM with 6-kHz sampling and a substandard speech bandwith of $W = 2.6$ kHz. This last condition was included to provide a 4-bit, 24-kb/s alternative to the 3-bit ADPCM stimuli, all of which also had a bit rate of 24 kb/s. The values of α and β used in the test were selected on the basis of a pilot test

Table I—Values of s/n at input of postfilter and after postfiltering (see Fig. 3). Numbers in parentheses are corresponding values of segmental s/n ratio

Input	R (bits/ sample)	s/n at Input of Postfiltering	s/n After Postfiltering (dB)	
			$\alpha = 1.0$ $\beta = 1.0$	$\alpha = 0.2$ $\beta = 1.0$
L8	3	21.7 (23.6)	28.9 (32.5)	27.0 (28.1)
	2	15.2 (17.1)	20.9 (25.3)	19.5 (21.0)
C8	3	18.9 (20.7)	27.2 (29.6)	23.9 (24.6)
	2	12.7 (14.9)	20.0 (22.9)	17.1 (18.2)

Table II—Number of wins in a round-robin tournament involving eight coding conditions and four listeners, where the maximum possible score is 196 for any given coding condition

Bits/ Sample	4	3						4 (W = 2.6 kHz)
α, β	0, 0	0, 0	0.2, 1.0	0.4, 0.8	0.4, 0.6	0.6, 0.6	0.6, 0.4	0, 0
L8	127	75	118	116	104	113	106	25
C8	111	80	129	120	106	117	102	19

Table III—Rank ordering of coding conditions by the group of 14 listeners and by a subgroup of 9 listeners from the Acoustics Research Department

Bits/ Sample	4	3						4 (W = 2.6 kHz)
α, β	0, 0	0, 0	0.2, 1.0	0.4, 0.8	0.4, 0.6	0.6, 0.6	0.6, 0.4	0, 0
L8 (G14)*	1	7	2	3	6	4	5	8
L8 (G9)†	1	7	3	2	6	4	5	8
C8 (G14)	4	7	1	2	5	3	6	8
C8 (G9)	2	7	1	3	5	4	6	8

* G14 group of 14 listeners.
† G9 group of 9 listeners.

that identified the interesting ranges of these parameters from the point of view of perceived mixes of noise suppression and speech distortion.

In general, use of postfiltering results in an amplification of the speech signal as suggested in Fig. 1b. The postfiltered speech stimuli were therefore appropriately scaled down to mitigate differences in stimulus loudness.

The subjective test involved an exhaustive pairwise comparison of all possible stimulus pairs, with each pair appearing at random places in the test once in each possible order of presentation. The total number of AB comparisons was therefore 768 ($8 \times 7 = 56$ possible stimulus pairs for each of 14 listeners).

Table II shows, separately for inputs L8 and C8, the total number of wins of each stimulus, with a maximum possible score of 196 for each stimulus [a maximum score of $2 \cdot (8 - 1)$ for each of 14 listeners]. It is seen that the worst two coding conditions stand apart from the rest. These conditions are 3-bit ADPCM with no prefiltering and 4-bit ADPCM with 2.6-kHz bandwidth speech input (and output). This latter condition gets a particularly low total score. Table II also shows that the above results are not very different for the inputs L8 and C8.

Table III shows, separately for inputs L8 and C8, the rank ordering

of the eight coding conditions in the subjective test. Results are shown separately for the total group of 14 listeners and the group of 9 listeners from the Acoustics Research department. It is seen that the rankings are not significantly different for the two populations. The best setting of the coefficient scaling parameters is defined by

$$\alpha = 0.2; \qquad \beta = 1.0$$

in each case, and for both L8 and C8. With input L8, 3-bit ADPCM postfiltered as above is ranked a close second to 4-bit ADPCM speech of equal bandwidth. In fact, the 4-bit ADPCM coder is ranked only fourth when the results of all 14 listeners are pooled together. The second and third ranks in this category belong to postfilters with the design ($\alpha = 0.4$; $\beta = 0.8$) and ($\alpha = 0.6$; $\beta = 0.6$). The preference for the design ($\alpha = 0.2$; $\beta = 1.0$) has a simple interpretation. It suggests a postfilter transfer function that mimics the approximate speech spectrum (the inverse predictor function) very closely at the zeros of that spectrum ($\beta = 1.0$), but very loosely at the poles ($\alpha = 0.2$). This suggests a condition that seeks to maximize background noise suppression and minimized perceived speech distortion. In the case of voiced speech segments, the poles tend to correspond to formant frequencies and the value of $\alpha = 0.2$ prevents an undue emphasis of the higher-amplitude spectral peaks, a situation that was indeed encountered in the example of Fig. 1b.

4.3 Enhancement of 2-bit ADPCM

As we see in Table I, the s/n gains due to postfiltering are equally significant for both 3-bit ADPCM and 2-bit ADPCM. Perceptually, however, the general noise level in 2-bit ADPCM speech is such that a useful degree of noise suppression requires the design of ($\alpha = 1.0$; $\beta = 1.0$). With this design, the speech distortion introduced by the postfilter is also substantial. For this reason, the case of 2-bit ADPCM is not considered to be of sufficient practical importance to pursue formal subjective testing. Informal testing shows, however, that the design of ($\alpha = 1.0$; $\beta = 1.0$) is again preferable to conventional 2-bit ADPCM ($\alpha = 0$; $\beta = 0$).

V. ENHANCEMENT OF SPEECH DEGRADED BY ADDITIVE WHITE GAUSSIAN NOISE

The specific postfiltering algorithm of Fig. 1b ($\alpha = 1.0$; $\beta = 1.0$) was also applied to speech degraded by additive white Gaussian noise. The input speech was L8, the speech-to-noise ratios ranged from −3 dB to 17 dB, and all coefficients were obtained simply by simulating the easily available case of 5-bit ADPCM, a bit rate high enough to introduce very little quantization noise in comparison with the levels

of Gaussian noise being studied. We find the postfiltering algorithm provides a very useful enhancement of noisy speech if the input to the postfilter had a s/n of at least +3 dB. For lower values of s/n, postfiltering provides noise suppression, but at the cost of substantial distortion of the speech itself.

REFERENCES

1. T. Nishitani et al., "A 32 kb/s Toll Quality ADPCM Codec Using a Single Chip Signal Processor," Proc. ICASSP (April 1982), pp. 960–3.
2. D. W. Petr, "32 kb/s ADPCM-DLQ Coding for Network Applications," Proc. IEEE Globecom Conf., December 1982. pp. A8.3.1–A8.3.5.
3. N. S. Jayant, "Adaptive Postfiltering of ADPCM Speech," B.S.T.J., *60*, No. 5 (May–June 1981), pp. 707–17.
4. J. O. Smith and J. B. Allen, "Variable Bandwidth Adaptive Delta Modulation," B.S.T.J., *60*, No. 5 (May–June 1981), pp. 719–37.
5. J. S. Lim (ed.), *Speech Enhancement*, Englewood Cliffs, NJ: Prentice Hall, 1983.
6. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–18.
7. D. W. Petr, unpublished work.

AUTHORS

Nuggehally S. Jayant, B.Sc. (Physics and Mathematics), Mysore University, 1962, B.E., 1965, and Ph.D. (Electrical Communication Engineering), 1970, Indian Institute of Science, Bangalore; Research Associate, Stanford University, 1967–1978; AT&T Bell Laboratories, 1968—. Mr. Jayant was a visiting scientist at the Indian Institute of Science in 1972 and 1975 and a Visiting Professor at the University of California, Santa Barbara, in 1983. Mr. Jayant has worked in the field of digital coding and transmission of waveforms, with special reference to robust speech communications. Editor, IEEE Reprint Book, *Waveform Quantization and Coding* and co-author of *Digital Coding of Waveforms: Principles and Applications to Speech and Video* (Prentice Hall, 1984).

Venkatasubbarao Ramamoorthy, B.E. (Electrical Engineering), 1970, The Regional Engineering College at Tiruchirappalli, India; M.Tech., 1972, The Indian Institute of Technology, Madras, India; Tekn.Dr., 1981, University of Linköping, Linköping, Sweden. Mr. Ramamoorthy was with the Indian Space Research Organisation at Bangalore, India, prior to his joining as a staff member at the department of Electrical Engineering, the University of Linköping, Sweden, in 1974. He visited AT&T Bell Laboratories during the summer of 1983. His current research interests include speech processing in mobile and packet radio environments, channel and source coding, digital modulation techniques, and development of handicap aids for children with speech problems.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

On Using the Itakura-Saito Measures for Speech Coder Performance Evaluation

By B.-H. JUANG*

(Manuscript received October 14, 1983)

The purpose of this paper is to discuss theoretical, as well as psychophysical, aspects of using the Itakura-Saito type of measures for evaluating the quality of coded speech. We present psychoacoustic interpretations of the measures and identify their effectiveness as well as limitations within the theoretical framework of a generalized waveform coder distortion model. The discussions then point out some specific issues to be resolved through psychoacoustic research effort.

I. INTRODUCTION

A "good" speech quality measure is central to progress in the research and development of speech processing systems. In speech coding, for example, we need a quality measure to provide insight into different distortions that are present in a coder output. If such a measure existed, it would help speech researchers identify how various kinds of distortions could be traded in order to improve the perceptual performance of the speech coder. In an engineering context, a measure that indicates the perceptual quality is a criterion to be optimized in speech coder design. Without such a measure, tuning coding schemes to achieve optimal quality is not a trivial task and the performance cannot be conveniently evaluated.

Speech quality assessment, however, involves subjective, psychological attributes of human perception, an area in which mathematicians

* AT&T Bell Laboratories.

and engineers are usually not well versed. Thus, speech quality evaluation has never been established satisfactorily in mathematical terms. The conventional signal-to-noise ratio (s/n), widely used in characterizing signal transmission/reception environments, is an ineffective measure of speech quality. Several other measurement methods and parameters, such as the isopreference method[1] and the subjective s/n,[2] have been proposed during the last two decades. General surveys of classical approaches can be found in Refs. 1 and 3. Reference 2 and its references also provide a summary of past efforts. Among these approaches, one particular class of measures based upon the Itakura-Saito measure has attracted engineers and scientists taking an analytical approach toward the problem. The Itakura-Saito measure and its variations, such as the Itakura or log likelihood ratio measure[4] and the likelihood ratio measure,[5] have been employed in noise studies by Sambur and Jayant;[6] in vocoder designs by Juang et al.[7] and Wong et al.;[8] in automatic speech recognition by Itakura[9] and Rabiner;[10] and as quality measures by Goodman et al.,[11] Crochiere et al.,[12] and Barnwell et al.[13]

Although successful applications of this class of measure are widespread in speech processing, none of them comes close to being justified as *the* speech quality measure. This paper attempts to identify the effectiveness as well as limitations of using this class of measure for speech quality within the theoretical framework of a generalized waveform coder distortion model.[14,15] We will further point out that such limitations also exist in current automatic speech recognizers that rely upon spectral matching. We then present some considerations relating to psychoacoustic studies, aiming at a better understanding of the fundamental concepts of speech quality in the presence of spectral distortion. These considerations will help direct future relevant psychoacoustic experiments for studying the dynamics of speech perception.

II. PRELIMINARIES

Let $s(i)$ and $s'(i)$ be two sampled speech signals, and let $x_n(i)$ and $x'_n(i)$ be two windowed segments, or frames, of $s(i)$ and $s'(i)$, respectively. Segments $x_n(i)$ and $x'_n(i)$ are obtained by applying a window function $w(i)$, with $w(i) = 0$ for $i < 0$ and $i \geq N$, to the speech signals at instance n; in particular,

$$x_n(i) = w(i)s(i + n) \tag{1}$$

and

$$x'_n(i) = w(i)s'(i + n). \tag{2}$$

The windowing operation greatly facilitates using spectral represen-

tations for speech analysis because speech is considered as a quasi-stationary signal. We denote the z-transform of $x_n(i)$ and $x'_n(i)$ by $X_n(z)$ and $X'_n(z)$, respectively. The Fourier transform is obtained by evaluating the z-transform on the unit circle, i.e., $z = e^{j\omega}$, and thus the notations $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$ are used to designate the Fourier transform of two windowed signals, respectively. For every such pair of spectral representations, $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$, a spectral distortion $\rho[X_n, X'_n]$ can be defined to measure the dissimilarity between $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$. In speech analysis, one particularly interesting distortion measure is the Itakura-Saito measure, which is defined as

$$\rho_{IS}[X_n, X'_n] \triangleq \int_{-\pi}^{\pi} [e^{\Lambda(\omega)} - \Lambda(\omega) - 1] \frac{d\omega}{2\pi}, \tag{3}$$

where

$$\Lambda(\omega) = \log |X_n(e^{j\omega})|^2 - \log |X'_n(e^{j\omega})|^2. \tag{4}$$

This mathematically tractable distortion measure has been successfully employed in vocoder designs.[7] Detailed analytical properties of the measure can be found in Refs. 4 and 5.

It has been shown in short-time Fourier analysis that a signal can be reconstructed from a properly time-sampled sequence of short-time Fourier transforms.[16] We can, thus, further represent the two signal sequences, $s(i)$ and $s'(i)$, by their corresponding short-time spectral sequences. Using $\oplus$ to denote the reconstruction process,

$$\{s(i)\} = .. \oplus X_{(n-1)l}(z) \oplus X_{nl}(z) \oplus X_{(n+1)l}(z) \oplus .. = \bigoplus_n X_n(z), \tag{5}$$

and

$$\{s'(i)\} = .. \oplus X'_{(n-1)l}(z) \oplus X'_{nl}(z) \oplus X'_{(n+1)l}(z) \oplus .. = \bigoplus_n X'_n(z). \tag{6}$$

In the above l is the underlying interval for short-time Fourier analysis and has been dropped in the final expressions without ambiguity. Such a representation allows us to characterize the dissimilarity between $s(i)$ and $s'(i)$ in terms of distortion measures obtained from short-time spectral representations. A distortion sequence between two speech signals is then defined as

$$\rho[s(i), s'(i)] = \{\rho_n\}, \tag{7}$$

where n is, as in (1) and (2), the frame index designating the window location, and

$$\rho_n = \rho[X_n, X'_n].$$

We will call ρ_n spectral distortion and $\{\rho_n\}$ a distortion sequence.

Extending the definition (3) to (7), then, we have a sequence of Itakura-Saito distortions.

The Itakura-Saito distortion measure defined by (3) and (4) is in fact *the* distortion measure for all-pole signal modeling; it was originally introduced as an error-matching function in maximum likelihood estimation of autoregressive spectral models.[17] Therefore, we shall confine ourselves to the analysis of Mth-order all-pole signal models despite the fact that a distortion measure could be more general. Several important results of the measure related to all-pole signal modeling are:

1. $$\rho_{IS}[X_n, \sigma_n/A_n] = (\alpha_n/\sigma_n^2) + \log \sigma_n^2 - \log \alpha_{n,\infty} - 1, \tag{8}$$

where

$$\alpha_n \triangleq \int_{-\pi}^{\pi} |X_n(e^{j\omega}) \cdot A_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \tag{9}$$

$$\alpha_{n,\infty} \triangleq \exp\left\{\int_{-\pi}^{\pi} \log |X_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}\right\}, \tag{10}$$

σ_n is a scalar, called the gain term, and

$$A_n(z) = 1 + \sum_{i=1}^{M} a_{i,n} z^{-i}. \tag{11}$$

2. $$\rho_{IS}[\sigma_n/A_n, \sigma_n'/A_n'] = \frac{\sigma_n^2}{\sigma_n'^2} \int_{-\pi}^{\pi} \frac{|A_n'(e^{j\omega})|^2}{|A_n(e^{j\omega})|^2} \frac{d\omega}{2\pi} + \log \sigma_n'^2 - \log \sigma_n^2 - 1, \tag{12}$$

which reduces to

$$\rho_{IS}[\sigma_n/A_n, \sigma_n/A_n'] = \int_{-\pi}^{\pi} \frac{|A_n'(e^{j\omega})|^2}{|A_n(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1 = \rho_{IS}[1/A_n, 1/A_n'] \tag{13}$$

when the gain terms are identical. $A_n'(z)$ takes the same form as $A_n(z)$ in (11). In the above expressions, we have assumed that $A_n(z)$ and $A_n'(z)$ have all their roots within the unit circle. Therefore,[18]

$$\int_{-\pi}^{\pi} \log |A_n(e^{j\omega})|^2 \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} \log |A_n'(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 0.$$

For clarity, we further define the likelihood ratio measure and the log likelihood ratio (or Itakura) measure as follows:

1. Likelihood ratio measure[7] $\rho_{LR}[X_n, X_n']$,

$$\rho_{LR}[X_n, X_n'] \triangleq \rho_{IS}\left[\frac{1}{A_n}, \frac{1}{A_n'}\right]; \tag{14}$$

2. Log likelihood ratio (Itakura) measure,[4]

$$\rho_I[X_n, X'_n] \triangleq \rho_{IS}\left[\frac{\sqrt{\underline{\alpha}_n}}{\underline{A}_n}, \frac{\sqrt{\hat{\alpha}_n}}{\underline{A}'_n}\right]$$

$$= \log\left(\frac{\hat{\alpha}_n}{\underline{\alpha}_n}\right). \tag{15}$$

In defining the above two measures, $\underline{A}_n(z)$ and $\underline{A}'_n(z)$ are the *optimal* Mth-order inverse filters of $X_n(z)$ and $X'_n(z)$, respectively.[18] Furthermore,

$$\hat{\alpha}_n = \int_{-\pi}^{\pi} |X_n(e^{j\omega})\underline{A}'_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \tag{16}$$

and

$$\underline{\alpha}_n = \int_{-\pi}^{\pi} |X_n(e^{j\omega})\underline{A}_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}. \tag{17}$$

Note that $\underline{\alpha}_n$ is the *minimum* Mth-order prediction residual energy pertaining to signal $X_n(z)$.

The Itakura-Saito distortion between the input and output signals of a linear system $H(e^{j\omega})$ can be easily calculated. Denoting the input power spectrum as $|X_n(e^{j\omega})|^2$, we have the output power spectrum $|X'_n(e^{j\omega})|^2 = |X(e^{j\omega})H(e^{j\omega})|^2$. Therefore,

$$\Lambda(\omega) = \log|X_n(e^{j\omega})|^2 - \log|X_n(e^{j\omega})H(e^{j\omega})|^2$$

$$= -\log|H(e^{j\omega})|^2, \tag{18}$$

and hence,

$$\rho_{IS}[X_n, X'_n] = \int_{-\pi}^{\pi}\left[\frac{1}{|H(e^{j\omega})|^2} + \log|H(e^{j\omega})|^2 - 1\right]\frac{d\omega}{2\pi}. \tag{19}$$

Of particular interest here is a class of $H(e^{j\omega})$ of the form

$$H(e^{j\omega}) = \frac{\underline{A}_n(e^{j\omega})}{B_n(e^{j\omega})}, \tag{20}$$

where $\underline{A}_n(z)$, as defined above, is the optimal Mth-order inverse filter of $X_n(z)$ and $B_n(z)$ is another Mth-order Finite Impulse Response (FIR) filter, taking the same form as (11). We also assume that $\underline{A}_n(z)$ and $B_n(z)$ both have all their roots within the unit circle. The input/output relationship of the system is illustrated in Fig. 1. Since $\underline{A}_n(z)$ is the optimal Mth-order inverse filter of $X_n(z)$, $E_n(z)$ is then the residual signal. $X'_n(z)$ is obtained by driving another all-pole filter $1/B_n(z)$ with such a residual signal. The distortion between $X_n(z)$ and

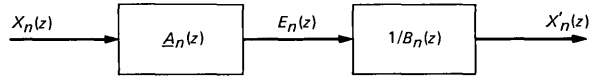


Fig. 1—A particular class of linear system in which $\underline{A}_n(z)$ is the optimal Mth-order inverse filter of $X_n(z)$.

$X'_n(z)$ under this condition is thus

$$\rho_{IS}[X_n, X'_n] = \int_{-\pi}^{\pi} \frac{|B_n(e^{j\omega})|^2}{|\underline{A}_n(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1$$

$$= \rho_{IS}\left(\frac{1}{\underline{A}_n}, \frac{1}{B_n}\right), \tag{21}$$

which is determined by the two all-pole filters, and has the same expression as the likelihood ratio measure of (14). This result gives us a convenient means of modifying a signal in order to achieve a prescribed distortion level from the original signal. Detailed discussions in Section IV are based upon this concept. It is, however, important to note that in eq. (21), $B_n(z)$ is not unique, and is not necessarily the optimal Mth-order inverse filter of the output signal $X'_n(z)$. It is simply stated that within the Mth-order autoregressive model framework, a prescribed Itakura-Saito spectral distortion can be obtained from a given signal through proper filtering operations, which will be convenient to realize.

III. A WAVEFORM CODER MODEL

Figure 2 shows a block diagram of the waveform coder distortion model used by Crochiere et al. for an interpretation of the log likelihood ratio measure.[15] This coder distortion model is composed of a time-varying linear filter $h(i)$, to model the "linearly correlated" distortions, and an additive noise source $q(i)$, to account for the nonlinear, uncorrelated distortions in the coder. Since the model attempts to split the components of distortion, it was expected that distinctively different

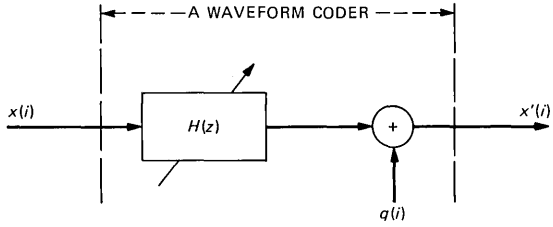


Fig. 2—Waveform coder distortion model.

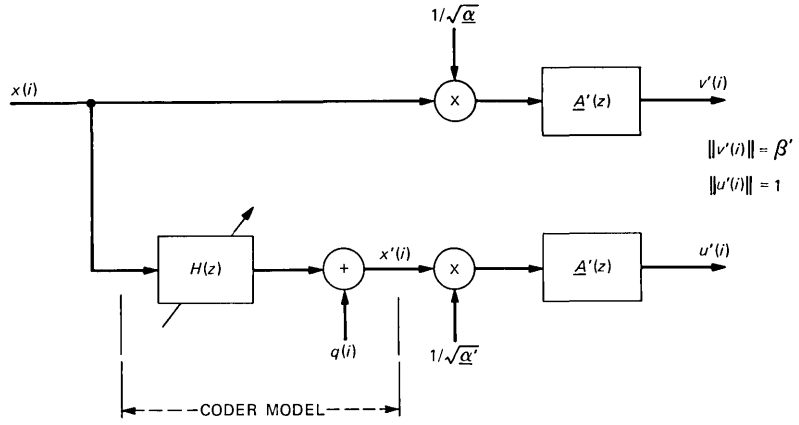


Fig. 3—Measuring coder performance with the likelihood ratio in a forward manner.

perceptual effects could be meaningfully studied separately with such a model.

Measurement of the coder performance with the likelihood ratio measure is shown in Fig. 3, which introduces the notion of inverse filtering. We use the likelihood ratio measure, rather than the Itakura-Saito measure, because we try to avoid, in the following discussions, extra complications in speech quality measurement due to amplification or attenuation. We follow the notation of Section II, except that the subscript indicating the frame index has been dropped, since, for most of the subsequent expressions, signal stationarity is assumed. We shall reinstate the frame index wherever necessary. The two parameters, α and α', are defined as in (17) by

$$\underline{\alpha} = \int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi} \tag{22}$$

and

$$\underline{\alpha}' = \int_{-\pi}^{\pi} |X'(e^{j\omega})\underline{A}'(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \tag{23}$$

where $\underline{A}(z)$ and $\underline{A}'(z)$ are the optimal Mth-order inverse filters of $X(z)$ and $X'(z)$, respectively. In other words, $\underline{\alpha}$ and $\underline{\alpha}'$ are the minimum Mth-order prediction residual energies corresponding to $x(i)$ and $x'(i)$ sequences, respectively. The energy of $v'(i)$, denoted by β', is then

$$\beta' = \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}'(e^{j\omega})|^2 \frac{d\omega}{2\pi}. \tag{24}$$

The energy of $u'(i)$, on the other hand, is unity due to the normalization factor $1/\sqrt{\underline{\alpha}'}$ and eq. (23). The energy ratio of the two filtered

signals, $v'(i)$ and $u'(i)$, is then equal to β'. By substituting (22) into (24), we have

$$\beta' = \frac{\int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}'(e^{j\omega})|^2 \frac{d\omega}{2\pi}}{\int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi}}. \tag{25}$$

The right-hand side of eq. (25) is the so-called likelihood ratio, and it can be reduced to

$$\beta' = \int_{-\pi}^{\pi} \frac{|\underline{A}'(e^{j\omega})|^2}{|\underline{A}(e^{j\omega})|^2} \frac{d\omega}{2\pi} \tag{26}$$

since both $\underline{A}(z)$ and $\underline{A}'(z)$ are Mth-order FIR filters, and the first $M + 1$ autocorrelation coefficients of the $\{[x(i)]/\sqrt{\alpha}\}$ sequence are equal to those of the impulse response of $1/\underline{A}(z)$. Therefore, the likelihood ratio measure of (14) can be expressed in terms of the energy ratio of the two filtered outputs, $v'(i)$ and $u'(i)$, and

$$\rho_{LR}(X, X') \triangleq \rho_{IS}\left[\frac{1}{\underline{A}}, \frac{1}{\underline{A}'}\right] = \int_{-\pi}^{\pi} \frac{|\underline{A}'(e^{j\omega})|^2}{|\underline{A}(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1$$

$$= \beta' - 1. \tag{27}$$

Alternatively, we may replace the filter $\underline{A}'(z)$ by $\underline{A}(z)$, the inverse filter of the $x(i)$ sequence, as shown in Fig. 4. In such a case, the energy of $u(i)$, denoted by γ, is the likelihood ratio, and the distortion

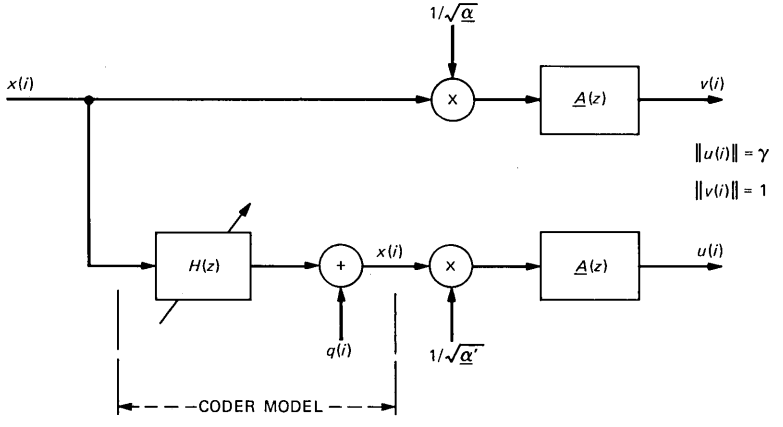


Fig. 4—Measuring coder performance with the likelihood ratio in a backward manner.

measurement is accomplished in a reversed direction, i.e.,

$$\rho_{LR}(X', X) \triangleq \rho_{IS}\left[\frac{1}{\underline{A}'}, \frac{1}{\underline{A}}\right] = \int_{-\pi}^{\pi} \frac{|\underline{A}(e^{j\omega})|^2}{|\underline{A}'(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1$$

$$= \gamma - 1. \tag{28}$$

The interpretation of the log likelihood ratio as a coder performance measure by Crochiere et al. follows the comparison order of eq. (28).[17] More specifically, the measure they discussed was log γ, instead of $\gamma - 1$. The difference between the log likelihood ratio measure and the likelihood ratio measure may be insignificant in terms of measurement. However, the likelihood ratio measure of eq. (14) appears to correspond more closely to the Itakura-Saito measure in representing the distortion relationship between the input and output signals of a particular class of linear systems. This was shown in eq. (21).

We now express the measures within the coder model. Referring to Fig. 2 and denoting the Fourier transforms of $h(i)$ and $q(i)$ by $H(e^{j\omega})$ and $Q(e^{j\omega})$, respectively, we have

$$X'(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega}) + Q(e^{j\omega}). \tag{29}$$

Furthermore, since $x(i)$ and $q(i)$ are uncorrelated,

$$|X'(e^{j\omega})|^2 = |X(e^{j\omega})H(e^{j\omega})|^2 + |Q(e^{j\omega})|^2. \tag{30}$$

For simplicity we assume that $H(z)$ does not have poles and zeros *on* the unit circle. From (24), the likelihood ratio distortion measured from $\{x(i)\}$ to $\{x'(i)\}$ is thus

$$\rho_{LR}[X, X'] = \beta' - 1$$

$$= \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} \frac{|\underline{A}'(e^{j\omega})|^2}{|H(e^{j\omega})|^2} \{|X'(e^{j\omega})|^2$$

$$- |Q(e^{j\omega})|^2\} \frac{d\omega}{2\pi} - 1. \tag{31}$$

On the other hand, the distortion measured from $\{x'(i)\}$ to $\{x(i)\}$ is

$$\rho_{LR}[X', X] = \gamma - 1$$

$$= \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} \{|X(e^{j\omega})H(e^{j\omega})|^2 + |Q(e^{j\omega})|^2\}$$

$$\cdot |\underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi} - 1. \tag{32}$$

Note that $Q(e^{-j\omega})$ is the complex conjugate of $Q(e^{j\omega})$ since $q(i)$ is real.

Different distortion components of the coder may be decoupled in the following way:

1. Additive noise distortion, ρ_a, is defined when there is no correlated spectral distortion, i.e.,

$$\begin{aligned}\rho_a^{(f)} &= \rho_{LR}[X, X'] \mid_{|H(e^{j\omega})|=1} \\ &= \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} |\underline{A}'(e^{j\omega})|^2 \{|X'(e^{j\omega})|^2 - |Q(e^{j\omega})|^2\} \frac{d\omega}{2\pi} - 1 \\ &= \frac{\underline{\alpha}'}{\underline{\alpha}} - 1 - \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} |\underline{A}'(e^{j\omega}) Q(e^{j\omega})|^2 \frac{d\omega}{2\pi} \end{aligned} \tag{33}$$

and

$$\begin{aligned}\rho_a^{(b)} &= \rho_{LR}[X', X] \mid_{|H(e^{j\omega})|=1} \\ &= \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} |\underline{A}(e^{j\omega})|^2 \{|X(e^{j\omega})|^2 + |Q(e^{j\omega})|^2\} \frac{d\omega}{2\pi} - 1 \\ &= \frac{\underline{\alpha}}{\underline{\alpha}'} - 1 + \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} |\underline{A}(e^{j\omega}) Q(e^{j\omega})|^2 \frac{d\omega}{2\pi}. \end{aligned} \tag{34}$$

In the above, the superscripts, f and b, denote the forward and backward measurements, respectively.

2. Correlated spectral distortion, ρ_c, is defined when the additive noise component vanishes, i.e.,

$$\begin{aligned}\rho_c^{(f)} &\triangleq \rho_{LR}[X, X'] \mid_{Q(e^{j\omega})=0} \\ &= \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} \frac{|X'(e^{j\omega}) \underline{A}'(e^{j\omega})|^2}{|H(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1 \end{aligned} \tag{35}$$

and

$$\begin{aligned}\rho_c^{(b)} &\triangleq \rho_{LR}[X', X] \mid_{Q(e^{j\omega})=0} \\ &= \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} |X(e^{j\omega}) H(e^{j\omega}) \underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi} - 1. \end{aligned} \tag{36}$$

The above decomposition of the measure into additive noise and correlated spectral distortions provides a helpful means in cross-verification between the measure and many known perceptual attributes. In the following we shall discuss the merits as well as limitations of the above measure in measuring the perceptual quality of waveform-coded speech signals. Such discussions point to some necessary psychophysical experiments for a closer link between objective and subjective measures.

3.1 Additive noise distortion

The key contribution of the uncorrelated additive noise, $q(i)$, appears in the integral terms in (33) and (34). Let us consider (34), where the integrand involves the inverse filter $\underline{A}(z)$ for the input speech signal.

The integral

$$\int_{-\pi}^{\pi} |\underline{A}(e^{j\omega})Q(e^{j\omega})|^2 \frac{d\omega}{2\pi}$$

is minimized subject to the constraint

$$\int_{-\pi}^{\pi} |Q(e^{j\omega})|^2 \frac{d\omega}{2\pi} = P_q, \tag{37}$$

where P_q is a constant, when $\underline{A}(z)$ is the optimal (Mth-order) inverse filter of the $q(i)$ sequence. In other words, for a given noise power, the integral is minimized if *the noise has the same spectral shape as the input speech*, within the Mth-order autoregressive signal modeling framework. This appears to be in very good agreement with the results of auditory masking that has been proposed as a method for improving the perceived quality of digitally encoded speech.[19,20] The same observation can also be made on (33), where the integrand involves $\underline{A}'(z)$ instead of $\underline{A}(z)$. $\underline{A}'(z)$ is the optimal Mth-order inverse filter of the *encoded output sequence* $x'(i)$. If $q(i)$ is truly uncorrelated with $x(i)$ (recall that $|H(e^{j\omega})| = 1$ here) and has the same spectral shape as $x(i)$, then $\underline{A}'(z)$ is, in fact, identical to $\underline{A}(z)$. However, when exact shaping of noise spectra is not achievable (as in most practical coder systems), (33) and (34) lead to significantly different distortion measurements since $\underline{\alpha}'$ involves $\underline{A}'(e^{j\omega})$, which demonstrates attributes of $Q(e^{j\omega})$. The following example illustrates the difference between the forward and the backward measurements.

Consider two signals, one being tonelike and the other being white noise. These two signals are represented in terms of second-order all-pole models as $1/A_t(z)$ and $1/A_w(z)$, where

$$A_t(z) = 1 - 1.2726\, z^{-1} + 0.81\, z^{-2} \tag{38}$$

and

$$A_w(z) = 1. \tag{39}$$

The two roots of $A_t(z)$ are $0.9\, e^{\pm j\pi/4}$, which indicate a resonance at $\pi/4$ normalized frequency or at 1000 Hz when the sampling frequency is 8000 Hz. These two all-pole models have corresponding reflection coefficient vectors $\underline{k}_t$ and $\underline{k}_w$:[18]

$$\underline{k}_t^t = [k_{t1}\, k_{t2}] = [-0.7\ 0.81] \tag{40}$$

and

$$\underline{k}_w^t = [k_{w1}\, k_{w2}] = [0\ 0]. \tag{41}$$

Using eq. (7) of Ref. 7,

$$\rho_{LR}\left[\frac{1}{A_t}, \frac{1}{A_w}\right] = \frac{(k_{w1} - k_{t1})^2(1 + k_{w2})^2}{(1 - k_{t1}^2)(1 - k_{t2}^2)} + \frac{(k_{w2} - k_{t2})^2}{I - k_{t2}^2}, \tag{42}$$

we can easily calculate the distortion in each direction and obtain

$$\rho_{LR}\left[\frac{1}{A_t}, \frac{1}{A_w}\right] = 4.7 \tag{43}$$

and

$$\rho_{LR}\left[\frac{1}{A_w}, \frac{1}{A_t}\right] = 2.26. \tag{44}$$

Clearly, if measured in the forward direction, when an input tonelike signal is being distorted into white noise, the distortion is higher than vice versa. The result is reversed if the distortion is measured in the backward direction; that is, distorting an input noise signal into a tonelike signal will result in a more serious objective distortion measurement than distorting a tone-like signal into white noise. Previous studies in auditory masking demonstrated a similar asymmetry of masking between tone and noise.[21–23] In particular, it has been reported that noise masks a tone more effectively than a tone masks noise. A 1-kHz tone masked by noise that is one critical band wide typically is inaudible at a signal-to-masker ratio of −4 dB, while the corresponding ratio for noise signal masked by tone is approximately −24 dB.[24] In other words, it is easier to perceive noise in a tone than it is to perceive a tone in noise. For an objective measure to consistently predict the perceived quality, we thus would require that such a measure show higher distortion when the input tone is corrupted by noise and that it show lower distortion when input noise is distorted by an additive tone signal. Despite the slight difference between masking and distortion, forward measurements of (33) thus appear to be more justifiable. More rigorous psychoacoustic studies are obviously very important in carefully resolving this measurement direction issue.

3.2 Correlated spectral distortion

Compared to additive noise, correlated spectral distortion has not been as well studied in the past, but it is a key factor affecting the

perceived quality. One well-known example is that "telephone speech", which is essentially bandlimited to the range of 200 to 3200 Hz, is considered to be of poorer quality and of lower intelligibility than the unfiltered original speech. Since correlated spectral distortion can be a result of the filtering operation, we shall discuss it using linear filtering concepts.

Linear systems can be categorized into time-invariant and time-variant systems. Accordingly, correlated spectral distortion can be time-invariant or time-variant as demonstrated in eq. (19), where the correspondence between the filtering operation and the distortion measure was established. The above-mentioned bandpass filtered speech signals, such as telephone speech, have essentially a time-invariant spectral distortion (here we are not considering tone noise, clicks, or channel variations, etc.), while Linear Predictive Coding (LPC) vocoders involve many time-variant spectral distortions, as will be discussed shortly.

The use of the Itakura-Saito type of measure for *time-invariant spectral distortions*, such as (3), (14) and (15), appears to be justifiable, at least within the short-time frame boundary where stationarity is reasonably assumed. This can be seen from the application of the likelihood ratio measure in vector quantization for voice coding.[7,8] In fact, the code words designed for vector quantization using (14) are substantially consistent with the vowel triangle of Peterson and Barney from an acoustic-phonetic point of view.[8] It also has been shown that the log likelihood ratio measure usually leads to a better recognition rate in speech recognition schemes.[10,25] (Note that the log likelihood ratio and the likelihood ratio measure make no significant difference in most speech recognition applications. The only theoretical difference is in template generation where minimization of some criterion, such as the average distortion or maximum distortion, is required.) For interests in psychoacoustic studies, however, it may be desirable to further translate the measurement into a perceptual scale that better interprets the relative perceived quality. (The complication here is the possible sound dependence on a perceptual scale. Consider the following example. Suppose X has been distorted, resulting in Y and Z. We can confidently say Y sound is closer to X sound than Z sound is, if $\rho[X, Y] < \rho[X, Z]$. However, we are not sure that Y is perceptually closer to X than Z is to W, even if $\rho[X, Y] < \rho[Z, W]$.)

Beyond the short-time stationary segment level, the time-variant distortion is a more important and complicated factor to consider in speech processing. Spectral distortion measures are defined for every pair of spectral representations. A natural extension of the distortion measure for measuring dissimilarity between time-varying signals is thus the distortion sequence is expressed by (7). Previous experiments

and several reported results that help illustrate the effect of time-variant spectral distortions upon speech quality are in order.

Voice coding results in time-variant spectral distortion. The key contribution to the time variation of distortion in voice coding such as LPC is a result of parameter quantization, although the parameter analysis procedure itself may also introduce some time-variant distortion because of frame alignment, change in excitation, etc. The effect of such distortion thus can be best explained in performance comparison of different parameter quantization schemes.

The experiment in Ref. 7 that compared the distortion performance of vector and scalar quantization for LPC voice coding provides important insights in this regard. In order to conduct the so-called equal average distortion comparison in the experiment, speech signals were vocoded at a lower bit rate with vector quantization and at a higher bit rate with scalar quantization. Subjective comparison of these two sets of synthesized signals of equal average distortion showed that the vector quantization synthesis samples sounded smoother and more pleasant, and were considered of better quality. Substantial background warble was perceived in the scalar quantization samples. Differences in spectral continuity, distortion contour, and some statistics of the distortion process $\{\rho_n\}$ between the two sets of synthesis samples were then reported to explain the difference in the perceived synthesis quality. It was concluded that a coder that preserves more spectral continuity, achieves smoother distortion contour, and produces less divergent distortion statistics is better than a coder with otherwise different distortion performance, even though they yield the same average distortion. Vector quantizers appear to produce "better" distortion sequences than do scalar quantizers in LPC voice coding. The importance of considering the distortion as a *process* or *sequence* (instead of just an average distortion) and of looking into the spectral continuity (a mathematical definition of which has yet to be obtained) was thus highlighted.

The concepts of time-variant distortions and spectral continuity also raise a possible explanation for the experimental results of Tribolet et al.[26] Here, performances of four different types of waveform coders at three different bit rates were compared. An average noise-to-signal measure [eq. (2) of Ref. 26], ℓ_m, which was derived through the concepts of log likelihood ratios, was used as an objective measure to predict the subject performance. As seen from Figs. 5 and 6 (duplicated from Figs. 7 and 9a of Ref. 26), the main failures of the likelihood-ratio-derived measure are in predicting the performance of all coders, at 9.6 kb/s [in particular, Sub-band Coder (SBC) at 9.6 kb/s] and Adaptive Differential PCM (ADPCM) coder with a fixed predictor at 24 kb/s. At 9.6 kb/s all coders perform subjectively worse than objec-

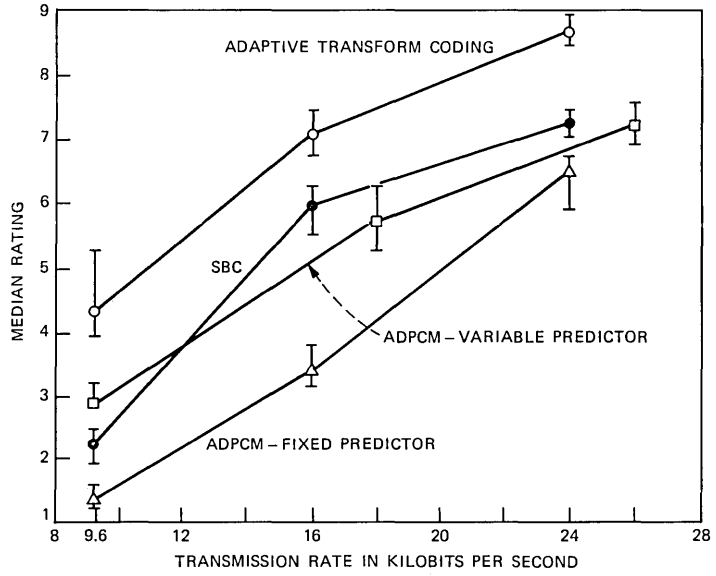


Fig. 5—Quality median rating of 12 coders (65 listeners by 4 talkers).

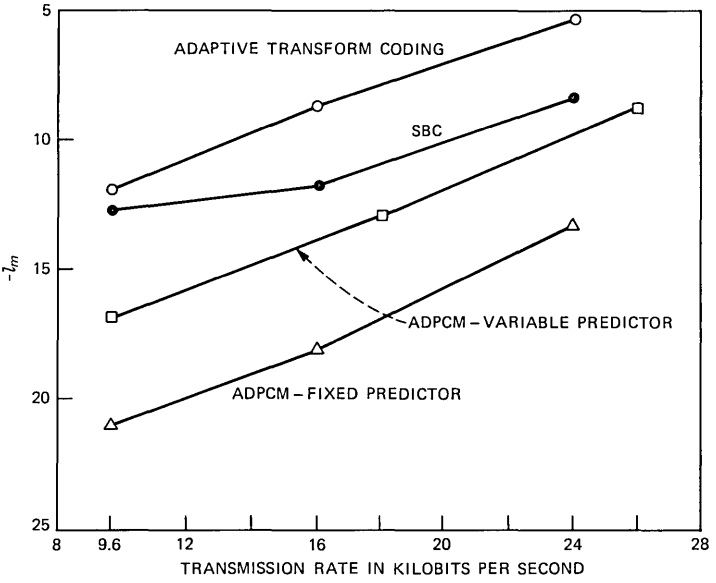
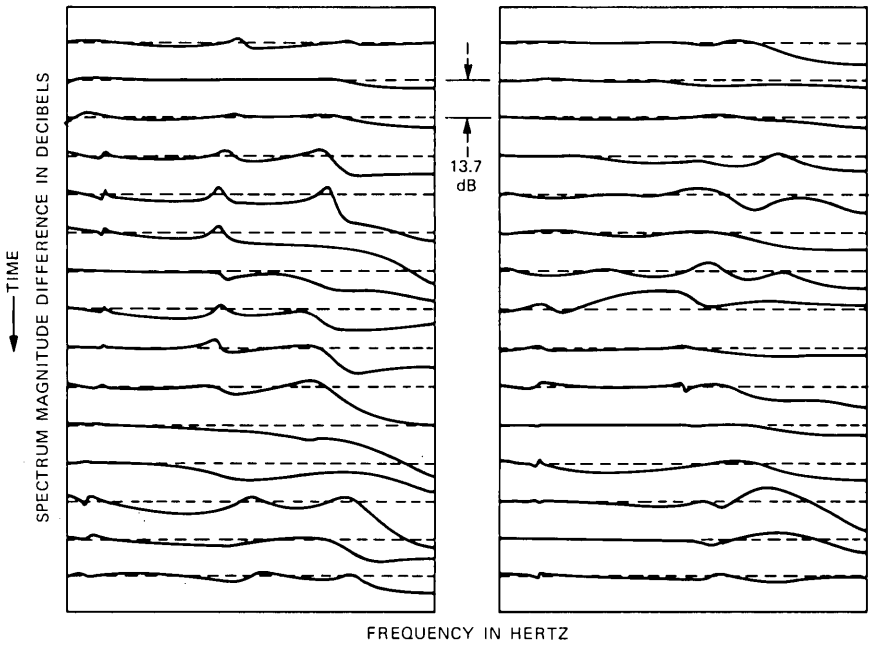


Fig. 6—Objective noise-to-signal measure, ℓ_m, averaged over 16 articulation bands for the 12 coders in Fig. 6.

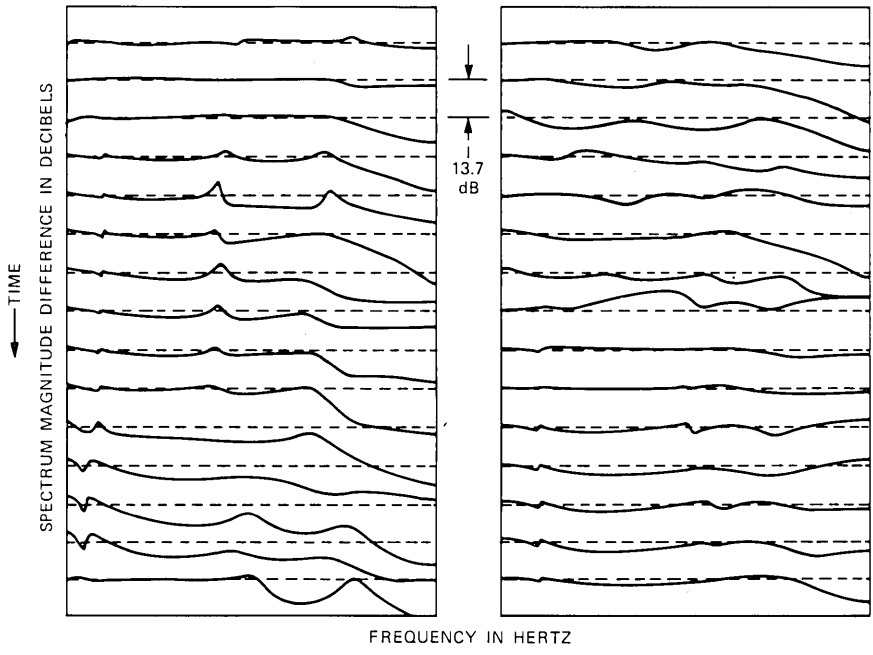
tively predicted. At 9.6 kb/s, SBC is objectively very close to the Adaptive Transform Coder (ATC) but turns out to be subjectively even worse than the ADPCM with a variable predictor. At 24 kb/s, ADPCM with fixed predictor is objectively much worse than ADPCM with variable predictor, but they in fact are subjectively very close. These failures can be attributed to the fact that ℓ_m does not correctly consider the correlated spectral distortion, and more importantly, it is only an average over the entire speech sample, revealing no information on possible perceptual degradation due to time-variant spectral distortions. The outcome that all coders perform subjectively worse than objectively predicted at lower bit rates is probably a result of increased sporadic spectral distortions and reduced spectral continuity along the time axis. Sub-band coding schemes inherently preserve less spectral continuity at lower bit rate, and thus it is possible that relatively more quality degradation is perceived at 9.6 kb/s with SBC. Finally, the ADPCM coder with adaptive, variable predictor potentially introduces more spectral discontinuity, due to quantization of the predictor parameters, than does the ADPCM with a fixed, unquantized predictor.

To illustrate this, plots of the log spectral (eighth-order all-pole) difference between the original and the reconstructed speech signals are shown in Fig. 7. Coders used in Fig. 7 are ADPCM with fixed predictors and adaptive predictors, respectively. More spectral discontinuity is observed in the adaptive predictor case, particularly in the low frequency region. Therefore, even though adaptive predictors yield higher prediction gain than fixed predictors,[27] this objective advantage has been subjectively offset by the perceptual sensitivity to time-variant distortions, particularly at higher bit rates, where the effect of additive noise becomes relatively less significant. As a result, the subjective performance gap between the two coders is substantially reduced.

Similar limitations apply to automatic speech recognition schemes that use one single average or accumulative figure to represent the dissimilarity between the spectral sequences of the input speech and the stored reference template. In parallel with the concept of measuring speech quality with the segmental s/n, recognition schemes usually resort to segmentation and time warping in order to obtain better distortion or distance measurements for more accurate recognition decisions. Nevertheless, segmentation schemes produce hard segmental boundaries, instead of natural, soft transitions, and are never completely reliable. The original problem of measuring the dissimilarity between time-varying signals thus has never been entirely solved.



(a)



(b)

Fig. 7—Log spectral difference between the original and reconstructed signals: (a) with a fixed, unquantized predictor; (b) with an adaptive, quantized predictor.

The above considerations clearly point out the necessity of psychophysical experiments for developing a better speech quality measure. Specifically, with regard to using the Itakura-Saito type of measures, issues to be further studied are: the measurement direction, the feasibility of characterizing subjective quality by distortion sequences, and the incorporation of some transitive functions into the distortion measure to account for spectral continuity. In light of the analytical features of the Itakura-Saito type of measures, research on these issues appears to be vitally important to an analytical speech perception model.

IV. SOME CONSIDERATIONS IN FUTURE PSYCHOPHYSICAL STUDIES

It is beyond the scope of this paper to propose and discuss in detail the psychophysical experiment procedures necessary to answer all the questions above. It is, however, appropriate to address one of the difficulties in psychoacoustic experiment designs here. In addition, we shall propose to consider a class of transitive functions to be used in defining the spectral continuity measure.

4.1 Inverse filtering as a tool

One of the fundamental difficulties in designing psychoacoustic experiments is the control of test stimuli. How to characterize and control the test signals is obviously not a simple matter when the stimuli are real running speech signals. In parallel with this problem is the difficulty in defining a refined speech production model that, at least, adequately describes the real speech production mechanism.

In studying perceptual responses to various spectral distortions in order to better analytically and dynamically characterize speech quality with the Itakura-Saito measure, the difficulty fortunately can be greatly alleviated. In particular, the result of (21) allows us to modify a speech signal conveniently to meet the prescribed distortion requirements. Clearly, if

$$\begin{aligned} \{s'(i)\} &= \bigoplus_n X'_n(z) \\ &= \bigoplus_n \frac{E_n(z)}{B_n(z)} \\ &= \bigoplus_n X_n(z)\,\frac{A_n(z)}{B_n(z)}, \end{aligned} \tag{45}$$

then

$$\rho[s(i),\, s'(i)] = \left\{\rho_{IS}\left[\frac{1}{A_n}, \frac{1}{B_n}\right]\right\}_n. \tag{46}$$

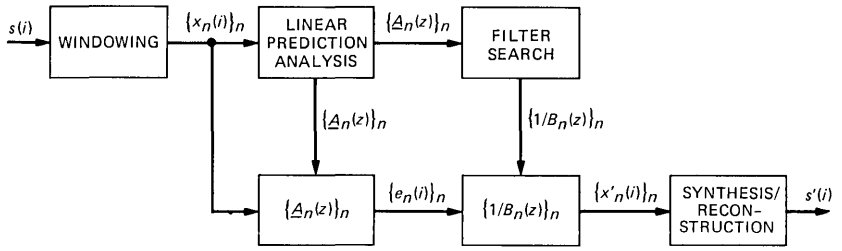


Fig. 8—Signal modification procedures to achieve prescribed distortion characteristics.

Figure 8 illustrates the modification procedure. The speech signal is first inverse filtered by $\underline{A}_n(z)$ to obtain the residual $E_n(z)$, which then drives the chosen filter $1/B_n(z)$ to form the desired signal.

Choosing $1/B_n(z)$ such that

$$\rho_{IS}\left[\frac{1}{\underline{A}_n}, \frac{1}{B_n}\right] \cong \rho,$$

a prescribed value can be made simple if we have a good-sized vector code book, as designed in vector quantization.[7] The search for $1/B_n(z)$ is then *quantum-selectively finite*, although there are theoretically infinite number of all-pole filters. Also, the test stimuli designed according to (45) are free from excitation variations, such as fundamental frequency changes, that are better considered separately.

4.2 Spectral continuity

As discussed above, spectral continuity is an important factor affecting the perceived quality of speech signals. Speech signals carry distinctive time-frequency or spectral transition patterns. Phonetic manifestation in articulated speech signals could be very fast, like /str/ in "strange", or sustainingly slow, like /i/ in "eat". To avoid complications due to such an inherent nonuniform spectral change, Ref. 7 used the model error spectral sequence $\{\Delta_n(\omega)\}$, defined by

$$\Delta_n(\omega) = \log \frac{1}{|\underline{A}_n(e^{j\omega})|^2} - \log \frac{1}{|\hat{A}_n(e^{j\omega})|^2}, \tag{47}$$

where $\hat{A}_n(z)$ is a quantized version of $\underline{A}_n(z)$, to illustrate the difference of the ability of various quantization schemes in preserving spectral continuity. The rationale was based upon the fact that the ultimate spectral continuity to be retained is the inherent spectral transition pattern, and that if a coder produces spectral distortion that is independent of time, that is,

$$\Delta_n(\omega) = \Delta(\omega) \quad \text{for all } n, \tag{48}$$

then the time-variant spectral distortion is completely eliminated. While $\{\Delta_n(\omega)\}$ adequately explained the spectral continuity differences,

more rigorous alternatives are necessary for, at least, the following reasons: (1) the variation in $\Delta_n(\omega)$ along the frequency axis, ω, as well as the time axis, n, is often so substantial that it is difficult to use only (47) to define a spectral continuity measure; (2) it was never concluded that the change in $\Delta_n(\omega)$ along the time axis, if regarded as an indication of spectral smoothness, is indeed perceptually independent of the spectral transition pattern of the speech signal.

Before we can completely characterize the spectral continuity along both the frequency and time axis, we would like to propose to tentatively consider two transitive functions that indicate the spectral changes in a speech signal as a function of time. The notion of eq. (21), measuring the distortion between two all-pole spectra, is emphasized in defining such transitive functions. Denoted by $\phi_f(k)$, the forward transitive function is defined by

$$\phi_f(k) \triangleq \sum_{n=0}^{\infty} e^{-n\lambda_f} \rho_{IS}\left[\frac{1}{\underline{A}_k}, \frac{1}{\underline{A}_{k-n}}\right], \tag{49}$$

where λ_f is a time constant and, $\underline{A}_k(z)$ and $\underline{A}_{k-n}(z)$ are the optimal Mth-order inverse filters of $X_k(z)$ and $X_{k-n}(z)$, respectively. $\phi_f(k)$ measures the all-pole spectral change in the speech signal in a forward manner, i.e., it measures the distortion resulting from replacing the current spectral envelope with previous spectral envelops. Characteristic changes in excitation, such as the pitch inflection, are not actively considered in $\phi_f(k)$, although they may affect the estimation of all-pole spectral models. One interpretation of measuring the transition in speech by the distortion between all-pole models instead of speech spectra is that we try to keep the current excitation signal unchanged, as if it were present in the previous segments as implied by eq. (21). We also assume that the time constant λ_f, accounting for short-time auditory memory,[28] is independent of the particular sound that is articulated and perceived.

Similarly, we define the backward transitive function $\phi_b(k)$ as

$$\phi_b(k) \triangleq \sum_{n=0}^{\infty} e^{-n\lambda_b} \rho_{IS}\left[\frac{1}{\underline{A}_{k-n}}, \frac{1}{\underline{A}_k}\right]. \tag{50}$$

Note that if the distortion measure were symmetrical and if $\lambda_f = \lambda_b$, the two transitive functions would be identical. The appropriateness of these functions remains to be studied.

The transitive functions are to be regarded as part of the speech signal. When a speech signal is distorted because of processing or encoding, the corresponding transitive functions are distorted also. The distortion, or noise, in the transitive functions thus provides a measure of the time-variant spectral distortion that affects the spectral continuity in the original signal. Further research effort, of course, is

necessary to verify the suitability of these functions or to develop a better spectral continuity measure. We feel that the concept in (49) and (50) provides a good starting point.

V. CONCLUSION

While the Itakura-Saito distortion measure and its variations have been widely employed and are considered promising in characterizing speech quality,[15] limitations in such measures still exist and have been identified within the theoretical framework of a generalized waveform coder distortion model in the above discussion. This type of measure is inherently nonsymmetric and therefore, in measuring distortions, a proper measurement direction needs to be determined. Subjective quality evaluation involves perceptual response to various degrees of distortion that has to be considered as a time function or a stochastic process. The feasibility of describing the subjective quality by finite-order statistics of the distortion process is to be studied. Furthermore, evidence shows that speech spectral continuity is also a key, if not the most important, factor affecting the subjective quality and thus, the speech spectral transition pattern should be regarded as a vital part of the speech signal. An even more fundamental and difficult task is, then, the incorporation of the spectral transition patterns into the rather static measurements of the Itakura-Saito distortion. Psychoacoustic studies are necessary to resolve these issues.

REFERENCES

1. W. A. Munson and J. E. Karlin, "Isopreference Method for Evaluating Speech-Transmission Circuits," J. Acoust. Soc. Amer., *34* (1962), pp. 762–74.
2. M. Nakatsui and P. Mermelstein, "Subjective Speech-to-Noise Ratio as a Measure of Speech Quality for Digital Waveform Coders," J. Acoust. Soc. Amer., *72*, No. 4 (1982), pp. 1136–44.
3. M. H. L. Hecker and N. Guttman, "Survey of Methods for Measuring Speech Quality," J. Aud. Eng. Soc., *15* (1976) pp. 400–3.
4. A. H. Gray, Jr. and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-24* (October 1976), pp. 380–91.
5. R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion Measures for Speech Processing," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-28* (August 1980), pp. 367–376.
6. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis From Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-24* (December 1976), pp. 448–94.
7. B.-H. Juang, D. Y. Wong, and A. H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-30* (April 1982), pp. 294–304.
8. D. Y. Wong, B.-H. Juang, and A. H. Gray, Jr., "An 800 Bits/s Vector Quantization LPC Vocoder," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-30* (October 1982), pp. 770–80.
9. F. Itakura, "Minimum Predication Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-23* (February 1975), pp. 67–72.
10. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Acoustics, Speech, Signal Processing, *ASSP-26* (February 1978), pp. 34–42.
11. D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner, and J. Goodman,

"Objective and Subjective Performance of Tandem Connections of Waveform Coders With an LPC Vocoder," B.S.T.J., *58*, No. 3 (March 1979), pp. 601–29.
12. R. E. Crochiere, L. R. Rabiner, N. S. Jayant, and J. M. Tribolet, "A Study of Objective Measures for Speech Waveform Coders," in Proc. 1978 Zurich Seminar on Digital Commun., pp. H1.1–7.
13. T. P. Barnwell, III, A. M. Bush, R. M. Mersereau, and R. W. Schafer, "Speech Quality Measurement," Georgia Inst. Technol., Atlanta, Tech. Rep. E21-655-77-TB-1, June 1977.
14. M. R. Aaron, J. S. Fleischman, R. W. McDonald, and E. N. Protonatarios, "Response of Delta Modulation to Gaussian Signals," B.S.T.J., *48*, No. 5 (May–June 1969), pp. 1165–95.
15. R. E. Crochiere, J. M. Tribolet, and L. R. Rabiner, "An Interpretation of the Log Likelihood Ratio as a Measure of Waveform Coder Performance," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-28* (June 1980), pp. 318–23.
16. J. B. Allen, "Short-Term Spectral Analysis and Synthesis and Modification by Discrete Fourier Transform," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-25* (June 1977), pp. 235–8.
17. F. Itakura, "Speech Analysis and Synthesis Systems Based on Statistical Method," Doctor of Engineering Dissertation (Department of Engineering, Nagoya University, Japan, 1972). (In Japanese).
18. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, New York: Springer-Verlag, 1976.
19. M. R. Schroeder, B. A. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," J. Acoust. Soc. Amer., *66* (1979), pp. 1647–52.
20. B. J. McDermott and C. Scagliola, unpublished work.
21. J. L. Hall, unpublished work.
22. J. L. Hall and M. R. Schroeder, "Loudness of Noise in the Presence of Tones: Measurements and Non-linear Model Results," in *Psychophysical, Physiological, and Behavioral Studies in Hearing*, G. van den Brink and F. A. Bilsen, Eds., Delft, The Netherlands: Delft University Press, 1980, pp. 329–32.
23. R. P. Hellman, "Asymmetry of Masking Between Tones and Noise," Percept. Psychophys., *11* (1972), pp. 241–6.
24. J. Zwislocki, "Analysis of Some Auditory Characteristics," in *Handbook of Mathematical Psychology*, Vol. 3, R. D. Luce, R. R. Bush, and E. Galanter, Eds, New York: Wiley, pp. 1–97.
25. B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," J. Acoust. Soc. Amer., Supplement 1, *72* (Fall 1982), p. 531.
26. J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A Comparison of the Performance of Four Low-Bit-Rate Speech Waveform Coders," B.S.T.J., *58*, No. 3 (March 1979), pp. 699–712.
27. P. Noll, "A Comparative Study of Various Schemes for Speech Encoding," B.S.T.J., *54*, No. 9 (November 1975), pp. 1597–1614.
28. G. A. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits in Our Capacity for Processing Information," Psychol. Rev., *63* (1956), pp. 81–97.

AUTHOR

Biing-Hwang Juang, B.Sc. (Electrical Engineering), 1973, National Taiwan University, Republic of China; M.Sc. and Ph.D. (Electrical and Computer Engineering), University of California, Santa Barbara, 1979 and 1981, respectively; Speech Communications Research Laboratory (SCRL), 1978; Signal Technology, Inc., 1979–1982; AT&T Bell Laboratories, 1982; AT&T Information Systems Laboratories, 1983; AT&T Bell Laboratories, 1983—. Before joining AT&T Bell Laboratories, Mr. Juang worked on vocal tract modeling at Speech Communications Research Laboratory, and on speech coding and interference suppression at Signal Technology, Inc. Presently, he is a member of the Acoustics Research department, where he is researching speech communications techniques and stochastic modeling of speech signals.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

A Packet/Circuit Switch

By Z. L. BUDRIKIS* and A. N. NETRAVALI†

(Manuscript received July 29, 1983)

We propose a switch, suitable for an integrated local communications network, that will support packet switching and circuit switching, with a wide range of bit rates. Key components are two serial memories; a multiplicity of access units, each capable of writing and reading uniformly formatted, addressed information; and a programmed controller. Circuit switching is achieved when the controller repeatedly allocates memory slots, following call setup. Data communications can proceed concurrently without setup, competing for unused slots. We give an example of a 10,000-telephone-line switch carrying a similar load of other traffic. The switch would delay voice by less than 5 ms and could be interfaced to the existing telephone system. We indicate a method of fault detection and isolation that will limit the impact of a failure on a serial memory to an arbitrarily small group of connected lines. We define an index for measuring failure impact and use it to derive most-favorable fault-isolating partitions.

I. INTRODUCTION

The telephone system is by far the world's largest communications network. It was primarily designed for voice, but its role widens continuously, as it adapts to new requirements. Presently it is changing to accommodate data communications.

Already the network extensively caters to data communications, but not yet as well as it might. Although internally the telephone system

* AT&T Bell Laboratories. On leave from the Department of Electrical and Electronic Engineering, University of Western Australia, Nedlands. † AT&T Bell Laboratories.

is rapidly becoming a vast interconnected computer system, proffered data are still largely carried as analog signals externally. That will change, however, as special provisions for data come on-line. As more of the plant, including switches, becomes digital, it will be possible to offer, on a selective basis, switched digital telephone channels usable for 56-kb/s data throughput. Also, packet-switched data services will widen in scope and access. Packet-switched data services are overlay networks that use the digital transmission facilities of the telephone network but bypass its switches, of which many are still analog. The packet networks eventually may become totally interconnected, just as the voice network, and also may become integrated with it.

In-house, or proprietary, telephone networks can benefit from the changing character of the overall network more immediately. Already available are switches and other components that permit an all-digital network that will accommodate on one facility both voice and data. As good as this already is, we are proposing a switch that could make the private network even better. Eventually it might even influence the entire system.

Currently available switches provide only circuit-switched connections. This gives fixed-capacity channels on a continuous basis, whereas much of data comes in bursts. Thus, computer communications are characterized by very long call durations with only low average, but in many instances very high, peak rates. Given the option, direct memory transfers could proceed in some instances at rates of many megabits per second. This is far too high for a switched and continuously held circuit.

It is true that the needs of bursty traffic can be catered to by what already is available, namely by some packet-switched networks. But that introduces a separate communications network for data, with the consequences of proliferating wiring plans, divided responsibilities, and probable long-term dyseconomies. It is better for one facility to serve all communications, and to do so without imposing mismatches.

We propose a switch and, more generally, a new switch architecture that support within one switching fabric both circuit- and packet-switched connections. This would largely avoid mismatches in respect to bursty data traffic, while preserving unity in communications.

The cardinal components of the switch (see Fig. 1) are a pair of Serial Memories (SMs), a Central Controller (CC), and Accessing Units (AUs). The memories do not recirculate and both ends (head and tail) of each terminate on the central controller. The AUs are connected to read-and-write taps along the SMs, an AU having one connecting tap to each memory. The two taps of an AU form a symmetrical pair: the tap to the second memory is as many places from the tail end as that to the first is from the head. Thus, each AU

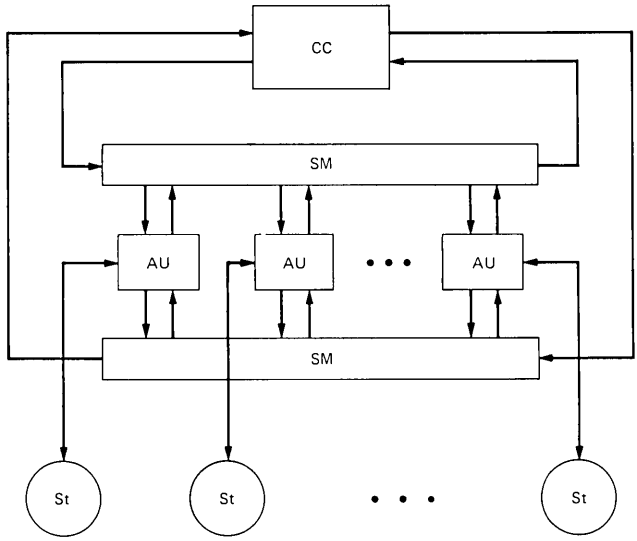


Fig. 1—Block schematic of switch. Access Units (AUs) communicate via two Serial Memories (SMs) on behalf of client Stations (Sts). Central Controller (CC) reserves slots for circuit-switched communications.

can reach every other AU by either one, or the other, memory. It can reach, and be reached by, the central controller by either memory. All writing is logical OR.

An AU acts as an agent of a client station (St) (e.g., telephone, facsimile terminal, computer) and mediates communications between it and other stations by way of corresponding AUs. Communications are carried on by write-and-reads in memory/time slots of uniform length and format. Each slot consists of a data field and several control fields. Collectively, the control fields provide synchronization, "Slot busy" indication, source and destination addressing, and slot pleading. A circuit-switched communication is carried on in regularly recurring slots, which are appropriately premarked by the central controller. For a packet-switched communication an AU simply uses the next available slot.

The capacity of a connected circuit can vary over a wide range, from a small fraction of a single (64-kb/s) telephone channel up to a large multiple of that capacity. It is settled by negotiation with the CC at the time of circuit setup, and need not be the same on different occasions. The capacity available to a packet-switched communication depends on the prevailing competition and can be any portion of the total switch capacity. The latter is a function of size and would be just several megabits per second for a 100-line switch and several hundred megabits per second for a switch that supports 10,000 lines.

A simple realization of the serial memories would be by clocked shift registers. The shift registers can be bit-paralleled to any degree needed to keep the clock rate low. The memories and all access units can be located centrally at the controller, with all connections to the switch then forming a single star. But it is also possible to segment the memories and form the network in clusters. The segments of memory would be connected serially to each other and to the central controller by transmission lines, forming two contrary rings, and the clusters again would form star topologies.

All elements of our proposal are well established and tried. Central, or stored program, control in circuit switching is over twenty years old.[1] The idea of switching by time slot interchanges is even older,[2] followed shortly by its realization through read-and-writes in computer memory.[3,4] Packet switching is more recent,[5] but is also well established both in local and wide-area networks.[6–9]

In essence, our scheme is an adaptation of seemingly diverse procedures, so that they may coexist. Time division slots are enlarged from what is usual in circuit switching, so that they can carry the control information essential to packet switching. Unlike normal packet-switched schemes, packets are of a single fixed length so that they can also be circuit-switched. Instead of separate time and space division stages, common in current telephone switches, we have a combined space/time fabric, abstracted from ring and bus networks, with a particular debt to Fasnet.[9] This makes packet switching possible without controller intervention. Finally, the controller maintains circuit connections by repetitive slot allocations, which is only marginally different from what takes place in a time division stage of a standard switch.

Also, our proposal is not first in its suggestion that voice and data be integrated on a common network.[10–14] But it appears to be first in suggesting a common switch for circuits and packets as the basis for that integration. With few exceptions,[11] prior suggestions have been to treat voice as data and to packet-switch it both in local and wide-area networks. However, these proposals have attendant delays that have to be addressed.

The point is important since, in the global telephone system, transmission delays can limit the quality of many possible connections. In the case of our switch, the delay of voice signals can be kept to less than 5 ms. It depends only on the clock rate, the size of switch, and the size of slots. Since delay considerations have an overriding sway on system choices, we discuss them in Section II. In Section III we give further details of our proposal.

In Section IV we address the question of reliability in our switch. We do this because our proposal may be seen as being particularly

vulnerable, since all its communications are to take place via two serial memories to which all AUs have writing privileges. We introduce a scheme for sectional detection and isolation of faults applicable to our switch. We show that this would limit the impacts of faults in our case to those that would prevail in switches that have much more dispersed and/or redundant architectures.

II. DELAY CONSTRAINTS AND RATE REQUIREMENTS

A communication system is expected to deliver messages to the destination in a timely fashion. The permitted delay is different in character for data and for real-time signals. We review the two cases separately.

2.1 Data transmission

Within limits, the exact times of arrival at the destination of the different parts in a data stream generally are unimportant. It is usually required that the sequence in the stream be preserved and that the average delay does not exceed some specified value. When data are presented for transmission at a fluctuating rate and there is not sufficient transmission capacity to cope with the peaks, the flow is smoothed by buffering. Waiting times in buffer stores are the predominant cause of delay.[15]

2.2 Real-time signals

In the transmission of real-time signals, the delay should be a constant and not greater than a specified value. Given fluctuation in transmission rate, there will be a time-varying delay $W_s(t)$ in a buffer at the sending end. A further delay $W_r(t)$ must be deliberately introduced in a buffer at the receiver,[16] so that the total delay could stay constant:

$$W_s + W_r = D. \tag{1}$$

D is the fixed buffer delay with which the system has been designed.

If, at some time, W_s exceeds D, then, at the same time, the buffer at the receiver will become empty and there will be a break in the received signal. Hence, there is no point in storing more at the transmitter than the amount of data that represents the total designed delay. If the rate λ of the real-time data is constant, as in Pulse Code Modulation (PCM) voice, then waiting times are directly related to amounts of stored data. The buffer-store capacities, N_s and N_r, that need to be provided at the two ends are equal, given by

$$N_r = N_s = \lambda D. \tag{2}$$

If λ is not constant, then the required capacities are still equal and are found by substituting the maximum value of λ in eq. (2).

It is important to note that, given fluctuations in data and/or transmission rate(s) and buffer stores to smooth them, the relevant delay for real-time signals is the maximum, i.e., designed, value, not the statistical average. How much larger that designed delay is to be than the average depends on the actual fluctuations in rate(s) and the relative tolerance to lost quality by signal discontinuities and by delay.

III. DESCRIPTION OF SWITCH

We now detail several aspects of the proposed switch. We give an indication of architectural options, describe protocols, and suggest suitable parameters for a 10,000-line switch.

3.1 Architecture

The basic configuration of the switch was shown in Fig. 1 and outlined in the Introduction. The functions of the AUs and the CC will be defined in more detail when we discuss protocols in the next subsection. It will be seen that there are considerable differences in the tasks of an AU that is mediating a circuit-switched, as compared to packet-switched, communication. Further differences in speed and buffer requirements may be identified between, and within, those two categories.

Clearly, there is a choice between designing a number of special-purpose AUs and designing a single universal AU. Further choices concern sharing of, and multitasking by, access units. Should AUs be placed in a common pool and shared by a larger group of stations? That would entail further switching outside the main switch to mediate connections between AUs and stations. Should an AU be multitasked, serving simultaneously different stations? That would make the AU a more complex device. Figure 2 illustrates a switch that incorporates both sharing and multitasking.

Our inclination is towards universal AUs, one to each station, and towards neither sharing nor multitasking. True, this calls for the largest number of AUs, and not the least complex, at that. But it has the advantage of uniformity and, in the light of technology trends, of likely overall economy.

The next choice concerns the serial memories. They may be active, made in semiconductor, or also, reverting to earlier technologies, passive, e.g., acoustic or electromagnetic delay lines. Passive components are attractive because they promise more reliability. However, our purpose would be better served by clocked shift register memories in an arrangement as, say, shown in Fig. 3. This makes for easier synchronization and permits bit-paralleling to hold down clock rates.

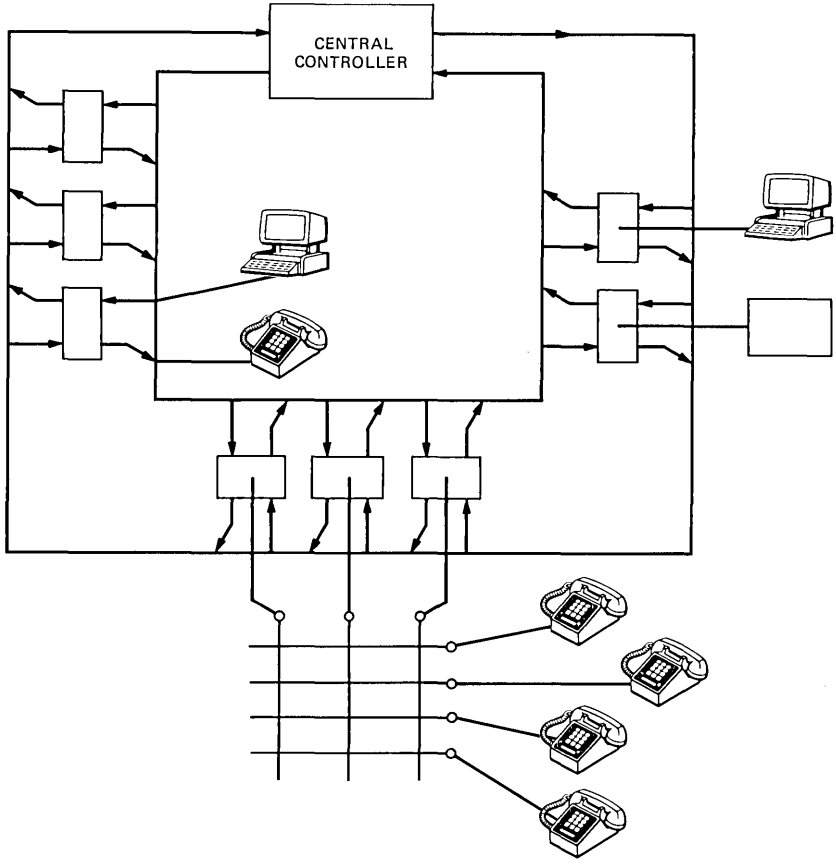


Fig. 2—Different options in AU tasking. All AUs could be of one type, each serving a single station (right); AUs could be shared by a larger group of stations requiring selector switching outside the main switch (center); or an AU could be multitasked, serving more than one type of station (left).

Reliability is a matter of overall design and implementation. In Section IV we discuss an architecture-related aspect of reliability, namely isolation of faults to limited sections.

Finally, we have the question of overall network topology. Three different arrangements are shown in Fig. 4. Figure 4a shows the traditional topology of a central switch and star network. In Fig. 4b a completely distributed arrangement is shown in which the serial memories wend their way past every station. This would make it similar to a local area ring network and would be possible only with passive lines as the memories. A compromise between the above two, and an interesting topology for a PBX that has to serve an extended area, is shown in Fig. 4c. The serial memories are cut into sections, and each section is placed close to the group of stations that it serves.

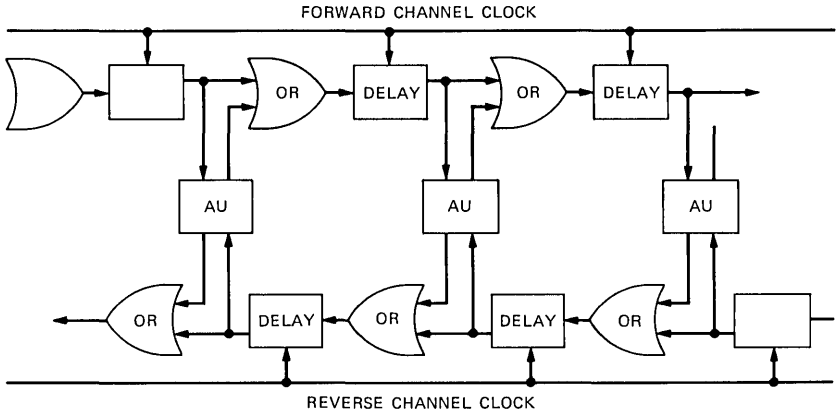


Fig. 3—Shift register realization of serial memories. Access points are at the inputs of clocked unit delays; the writing is through OR gates.

The lines connecting the individual sections and the central controller could be optic fibers, which would carry the total information streams serially even in a large switch.

3.2 Protocols

In the context of our proposal, a packet used in packet-switched communication is made up of five control fields and data, as shown in Fig. 5. The same format could be used in circuit-switched packets. But for these, at least one of the two address fields is unnecessary. Its space may either be added to the data field, or it could be used as a separate channel, a companion to the main channel.

The six fields marked in Fig. 5 are:

1. BUSY—a single bit to indicate slot occupancy
2. RQST—a single bit, common channel used for slot pleading
3. SNDR—address or password of AU sending packet
4. RCVR—address or password of AU intended to receive packet
5. DATA—data field
6. SYNC—synchronization field.

The roles of all the fields, except RQST and SYNC, are self-evident. RQST is used by packet-switching AUs and we will see its function presently when we discuss data communications. The SYNC field is written by the central controller to ensure slot and frame synchronization. Although both synchronizations could be achieved with just one bit per slot, a field of two bits will make them more secure. Altogether, the following numbers would be of the right order: BUSY and RQST one bit each, SYNC two bits, the addresses 14 bits each, and DATA 192 bits, for a total packet of 224 bits, or 28 bytes.

Corresponding to a packet, one may think of a time slot and of a

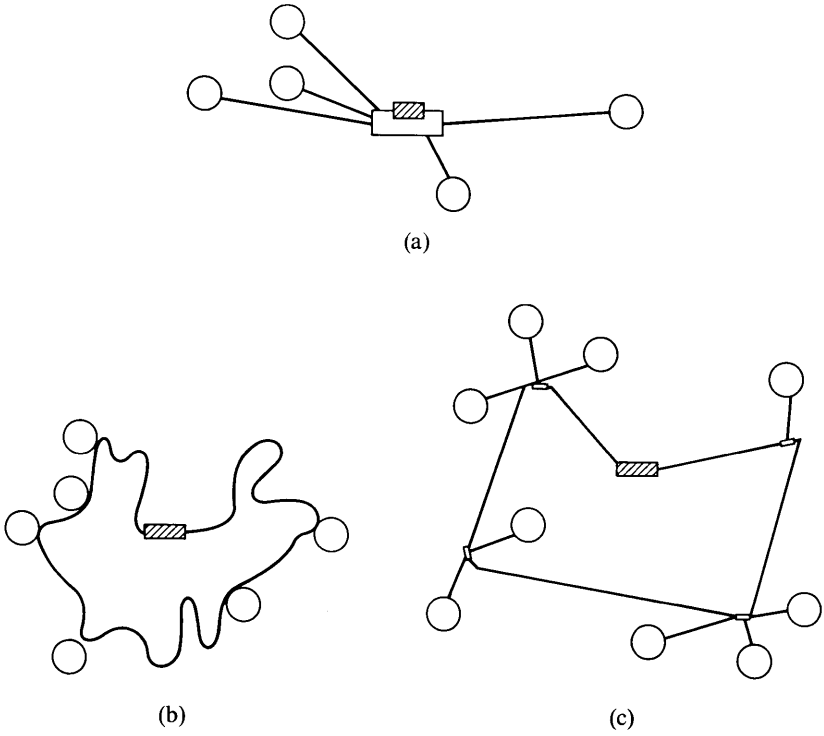
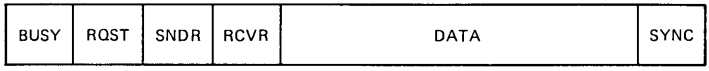


Fig. 4—Network topology options: (a) central switch, stations connected by lines; (b) switch completely distributed with AUs at individual stations and connected by the serial memories realized as buses; (c) switch distributed in clusters, the seral memories within clusters realized by shift registers and between clusters by transmission lines.

propagated memory block as consisting of the same number of bits and divided among respective fields. Note, however, that it is not necessary that the different parts of a packet be placed into a single slot or block. There may be interleaving of packet parts to any extent that is desirable.

Thus, it is conceivable that in order to alleviate the pressure of time for the signaling from receiver to transmitter within the AU, the



BUSY – SLOT BUSY FIELD
RQST – SLOT REQUEST FIELD
SNDR – SENDER ADDRESS
RCVR – RECEIVER ADDRESS
DATA – DATA FIELD
SYNC – SYNCHRONIZATION FIELD

Fig. 5—Slot format. Typically, BUSY, RQST and SYNC would be one-bit fields, the address fields could be two bytes each and the data field 24 bytes.

BUSY field could refer to the state of occupancy—not of the slot that it is riding in, but of the following slot. Similarly, other fields could be advanced or retarded, and not necessarily by only one slot, nor, indeed, just as a complete field. Thus, the DATA field could be broken into single bytes or even bits, and the fragments made to follow the header as an arbitrarily dispersed tail, provided only that all packets are fragmented and dispersed identically.

The extreme fragmentation of packets, as just alluded to, may seem an attractive way of restoring smoothness to data flow for circuit-switched communications. Indeed, almost complete smoothness is possible for any one chosen rate. But it would be at the expense of considerable complication for all other communications, particularly the packet-switched and the circuit-switched that have higher rates than the one singled out for favorable treatment. We will dismiss it from further consideration and turn to describing procedures.

3.2.1 Data communication

Assume that AU addresses are in numerical order along the two memories, ascending in the direction of propagation along one and descending along the other. We will call the memory with ascending addresses the forward channel, and hence the other the reverse channel.

Suppose that an AU has to communicate to another AU of higher address. It must send a message, or packet(s), on the forward channel. To do so, the dispatch processor of the AU will follow the data dispatch routine of Fig. 6. This can be understood more easily with the help of the state diagram of Fig. 7. For the sake of description, this diagram relates to an exclusive forward channel dispatcher, although in practice a single dispatcher would service both directions.

When idle, the dispatcher is normally in the "Go" state and monitors the sending buffer (for the forward channel), checking whether it contains a packet for transmission. If it does, it reads the BUSY field of the next block on the forward channel and at the same time writes a "ONE" in that field so as to seize the slot, should it be available. If it is not, i.e., BUSY was already "ONE," then it will write "ONE" in the next RQST field on the reverse channel and wait for the next BUSY field on the forward channel. It will repeat reading and writing of BUSY on the forward channel and sending RQSTs on the reverse channel until a "ZERO" BUSY occurs. It will then write in the related SNDR, RCVR, and DATA fields, so dispatching a packet.

Having sent a packet, the dispatcher moves to the 'One packet sent' state. If the sending buffer has at that moment one or more further packets for dispatch, then the dispatcher will behave exactly as in the "Go" state and send off the next packet, thereby moving to the "Two

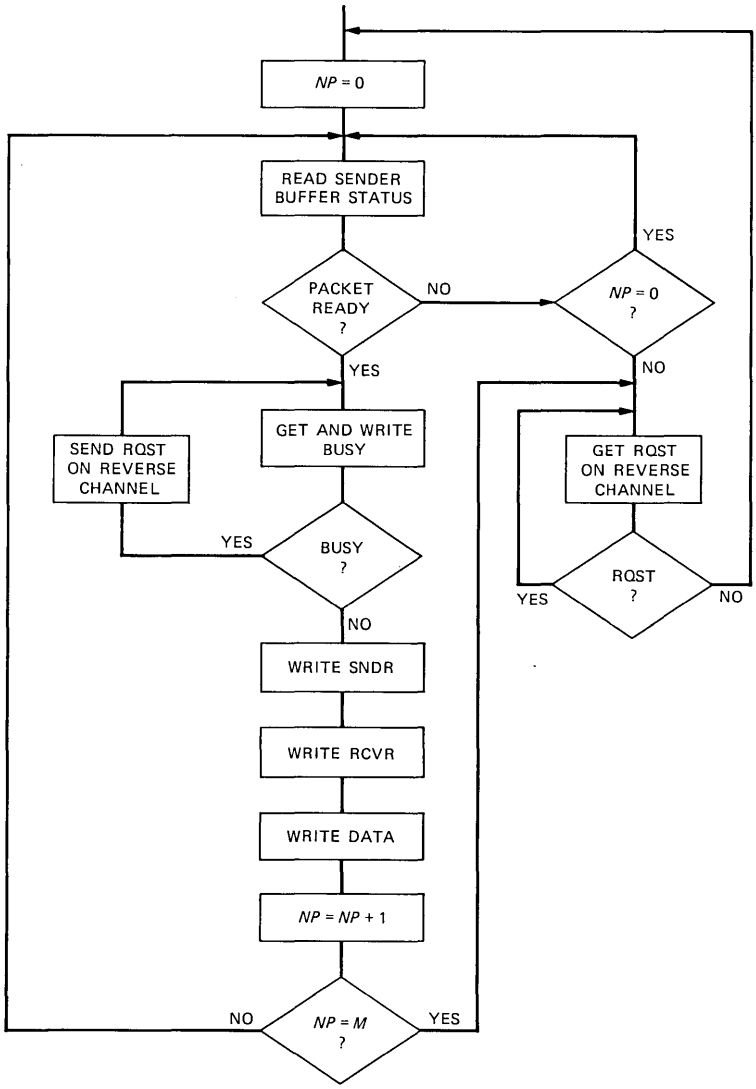


Fig. 6—Flowchart of forward channel data dispatch routine.

packets sent" state. But if there is no packet in the sending buffer on entry to the "One packet sent" state, then the dispatcher will proceed to the "Halt" state. It will remain there until the next "ZERO" is written in the RQST fields on the reverse channel, whereupon it will revert to the "Go" state. Similar conditions apply on entry to the "Two packets sent" and further states, until the dispatcher has sent in a contiguous sequence M packets and entered the "M packets sent" state. From this it must proceed unconditionally to "Halt."

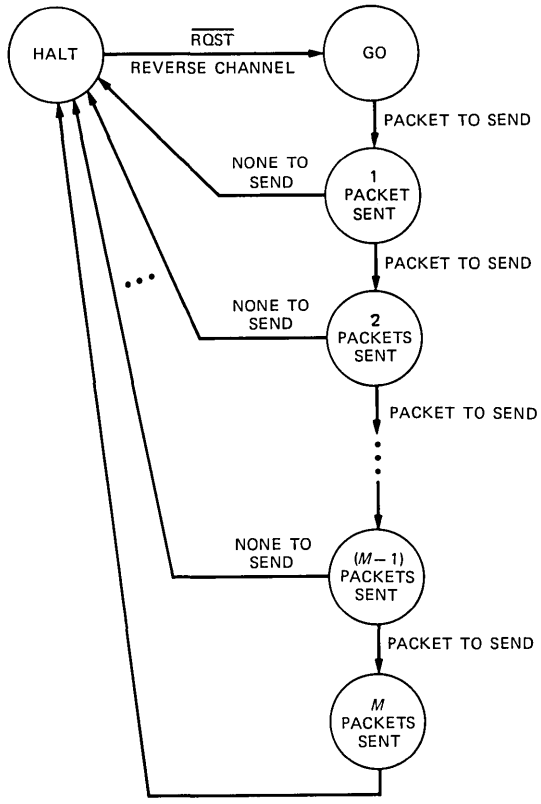


Fig. 7—State diagram of data dispatcher. The dispatcher goes temporarily into HALT state whenever it has no more packets to send or has already sent M packets since the last HALT. It goes from HALT to GO as soon as the RQST bit on the reverse channel is ZERO.

M is a parameter that may vary with AU. It represents priority standing: The larger its value, the less sensitive the AU is to pleadings for slots by other AUs that are downstream from it. It is normally set in relation to the rate of the station that the AU serves.

The task of receiving is less involved but no less time consuming, and an AU will have a separate processor for it. A routine that it could follow is given in Fig. 8. This is set out on the assumption that the SNDR and RCVR fields of a packet would precede the DATA field by one slot.

3.2.2 Real-time signal transmission

An AU serving a real-time device has to act in two distinct modes, one in setting up or tearing down a circuit and the other in transmitting and receiving the real-time signals when the circuit is set up. We

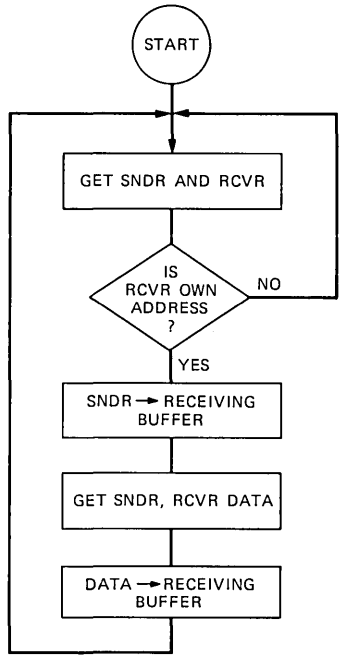


Fig. 8—Flowchart of data receiver routine.

outline the procedure, limiting our attention to telephony. Other devices requiring circuit connections would be served similarly.

Looked at from the telphone, the AU would appear as the line selector of the standard switch. When the telephone is taken off hook, the AU would supply dial tone. As the number is dialed, it would be stored by the AU, which, on completion, would assemble a packet for transmission to the CC. The DATA field of that packet would disclose the fact that a telephone link is being sought, and the numbers of the calling and called stations. The sending procedure for the packet could follow the routine of Fig. 6, even though a simpler routine is possible since no "Halt" state is necessary.

The CC would process received requests using a routine that could be as in Fig. 9. First, the CC would check the total switch capacity already committed to circuit traffic, and from this it would decide whether the setting up of the further circuit is permitted. If it is not permitted, then the CC would inform the originating AU, and that would terminate the processing. If setting up the circuit is permitted, then the CC would determine which AU serves the called station and check whether it is engaged. If it is engaged, then the CC would inform the originating AU accordingly. If it is not engaged, then the CC would tag both AUs as engaged and send messages to both AUs and inform

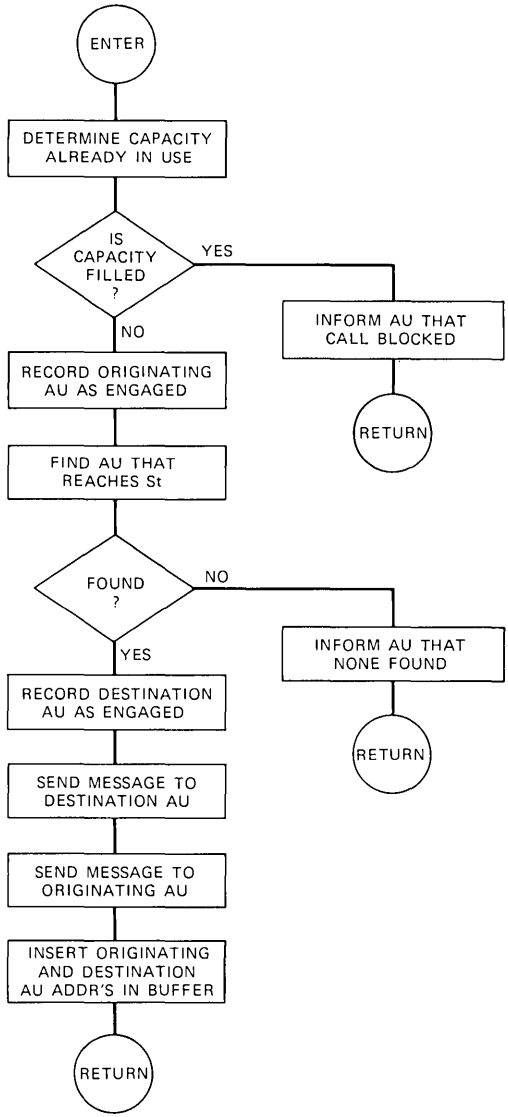
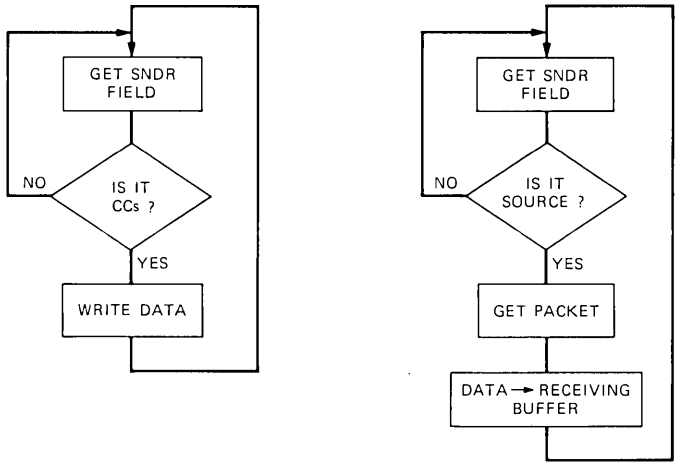


Fig. 9—Flowchart of circuit setup routine in central controller.

them of each other's addresses. The two addresses would also be inserted at appropriate places in ring buffers to cause the necessary premarking of slots by writing of BUSY and SNDR on the correct channels at the right frequencies. This would complete the setting up of the two-way circuit. Given a setup circuit, the dispatch and reception of the real-time data would follow the routines of Fig. 10.

Note that only the SNDR address is used in circuit-switched trans-



(a) (b)

Fig. 10—Flowchart of (a) real-time signal dispatch, and (b) real-time signal reception.

missions: The receiver is given the sender's address and recognizes it for the duration of the call. Apart from saving one address field for other use, there is a further bonus in that more than one receiver can be given the same SNDR address and simultaneously receive the same real-time signal. This leads to the possibility of a simple arrangement for broadcasting to designated outlets for, say, a public address system. If, furthermore, the AU's receiver capability was enlarged to noting several SNDR addresses and taking in packets with those markings, then a telephone conference facility, with voice signal summation at each receiver, would be possible.

The setting up, tearing down, and maintaining of calls to subscribers outside the switch's own area would have to interwork with equipment in other offices. But there is no particular problem about this. The AU serving a trunk would interface with the outside system, sending and responding to signals in conformity with existing specifications. But, in other respects, it would not be different from an AU serving a local subscriber. Data out of, and into, the local switch area could also be carried by circuit-switched trunks, with suitable interfacing to a wider-area data network. The role of an AU providing that interfacing would then amount to that of a gateway processor.

3.3 Packet size and clock rates

The bit rate required along the SMs is related to the total peak load for which the system is designed, multiplied by a factor that accounts for efficiency. Assuming a telephone voice signal sampled at 8 kHz,

represented by 8 bits per sample and an allowed delay due to packetization of 3 ms, a packet may contain 24 samples or 192 bits of data. The overheads are mainly in the addresses: Assuming a 10,000-voice-line switch and a total number of AUs not exceeding 16K, the SNDR and RCVR fields could be 14 bits each. As we already noted, BUSY and RQST need be only 1 bit each, and SYNC 2 bits. The total overhead will then be 32 bits and the packet length will be 224 bits.

Another criterion by which the overall size of a packet can be decided is efficiency. Since the allowable delay for voice is binding, the best size indicated for maximum efficiency will be of interest only if it is smaller than that already decided.

It is a reasonable simplification to suppose that all offered traffic divides into two categories: very short bursts, and prolonged streams. Furthermore, it is reasonable to assume that the number of packets-per-second from the very short burst will be independent of packet size. Thus, such very short bursts would be produced by single ASCII characters from, and echoed to, computer terminals, when carried in individual packets. On the other hand, circuit-switched traffic and data file transfers are examples of streamed flow.

Consider the total bit rate, R, that results from traffic consisting of b, short bursts per second, and an aggregated stream flow of S bits per second. If the packet has h bits of header and x bits of DATA, then

$$R = [b(h + x)] + [(S/x)(h + x)]. \tag{3}$$

This will be a minimum when

$$x = \sqrt{(S \cdot h)/b}. \tag{4}$$

In a system serving a business, one might provide for a busy-hour voice traffic of 10 ccs (hundred call seconds) per telephone. In the switch, this will divide equally between the two memories. With 10,000 telephones, the aggregate stream S_v on each SM due to voice would then be

$$S_v = 10{,}000 \cdot 5 \cdot 64{,}000/36 = 89 \text{ Mb/s}.$$

A reasonable assumption for the present is that all other traffic would amount to 20 percent of the total, or in our example it would be a further 22 Mb/s.

For the sake of illustration, assume that the very short burst rate, b, is 20,000 packets per second. If each of these carries only one 8-bit byte, then the net traffic from them is 160 kb/s, a negligible amount within the assumed 22 M/bs. But the gross traffic may be much larger, depending on packet size. Hence, the decision for best size of DATA field, which, with the numbers already invoked, follows from eq. (4):

$$x = \sqrt{(111{,}000{,}000 \cdot 32)/20{,}000} = 421 \text{ bits}.$$

For x_{opt} to be less than 192 bits, decided by delay considerations, the very short burst traffic would have to be 4.8 times larger than was assumed. But the assumed rate is already large, and therefore it is unlikely that efficiency considerations would indicate a smaller packet than given from delay.

Given packets of 224 bits and the numbers cited above, the rate, R, in each memory follows from eq. (3) as 134 Mb/s. If 8-bit bytes are propagated in parallel, then the required clock rate is 16.75 MHz, a none too demanding frequency for present technology. The packet rate, which is of greater relevance to AU and CC speeds, would be 598 kHz.

A switch would be designed for a given ultimate size and given an appropriate clock rate from the start. But it would not be necessary to give it immediately the full complement of AUs, nor, indeed, full lengths of memories. AUs could be added without any disruption and memory sections with only a minor pause.

IV. RELIABILITY

Availability of communications services is extremely important and has prompted switch designers to adopt the very highest standards of reliability.[18,19] Thus, it is accepted practice to have two identical central controllers, one being a "hot" standby that can take over at any instant. This and other common practices would also apply to our switch. The features by which our switch is rendered most vulnerable in respect to reliability are its serial memories, which carry all messages and are accessed by all AUs. Below we consider the general question of disruptive impact by failures and suggest a measure for it. Then we introduce a fault detection and isolation scheme that would make the robustness of switching by serial memories with multiple read-and-write taps comparable to that of much more redundant architectures.

4.1 Failure impact

In switching equipment, including ours, failures are unequal in likelihood and in disruptive consequence. We introduce the notion of expected failure impact. Let π_{ik} be the probability that component C_i will fail during the course of one year; let the expected repair time for it be τ_{ik}; and the number of potential communication connections that are unavailable while C_i is in the failed state be ν_{ik}. We define U_i, the expected per annum failure impact (EPAFI) of C_i, as

$$U_i = \sum_k \pi_{ik} \cdot \tau_{ik} \cdot \nu_{ik}. \tag{5}$$

We assume that failures are statistically independent and disregard the probability of another component failing during the repair time of

an existing failure. EPAFI values then are additive, and U, with respect to an assembly of N components, is

$$U = \sum_{i=1}^{n} U_i. \tag{6}$$

We consider the expected failure of the SMs and all the AUs connected to them. Suppose that there are altogether N AUs, each with (1) a per annum rate π_1 of failing in a way that affects only one subscriber and takes time τ_1 to repair, and (2) a rate π_2, which disrupts communications on the memory past the failed AU and takes τ_2 to repair. Also, let the memories have a rate π_3 of failing at each of the $2N$ connecting points, with expected repair times τ_3.

With N mutually communicating AUs in the system, the number of potential two-way communication links is $N(N-1)/2$. If an AU failure of the first kind occurs, then $\nu_1 = (N-1)$ of these are disrupted. When the failure of the AU is of the second kind, or when a memory fails, then the number of disrupted links is much larger and depends on the actual location of the failure. One can calculate an average number on the assumption that all potential failure locations are equally likely. It is found to be approximately $(N^2)/3$. Hence the total expected impact due to failures of AUs and memory links is

$$U = [N\cdot(N-1)\cdot\mu_1] + [(N/3)\cdot N^2\cdot\mu_2] + [2\cdot(N/3)\cdot N^2\cdot\mu_3], \tag{7}$$

where $\pi_i\nu_i$ has been contracted to μ_i.

4.2 Failure detection and isolation

We propose to divide the memories into sections and have a fault detector at the end of each section. Further, each section would have a bypass and, in case of a detected failure, a switch would be actuated to pass on to the next section the data stream at the output of the bypass (Fig. 11). Thus, effectively, the consequence of the fault is isolated to one section. A possible realization of the switch is shown in Fig. 12.

A Fault Detector (FD) would compare the data streams at the outputs of the memory section and the bypass, and it would decide that failure has occurred when the evident modification to the stream in passage through the memory section violates existing constraints. The particular constraint of the several that exist in our case and which we use is the following: There may never be a change of any field that is already nonzero. Detecting any such illegal changes will catch failures both in AUs and the memories. The detection will, of course, rely on the output from the bypass being a flawless replica of what entered that section. If necessary, redundancies and error control could be implemented on the bypasses to make that more sure.

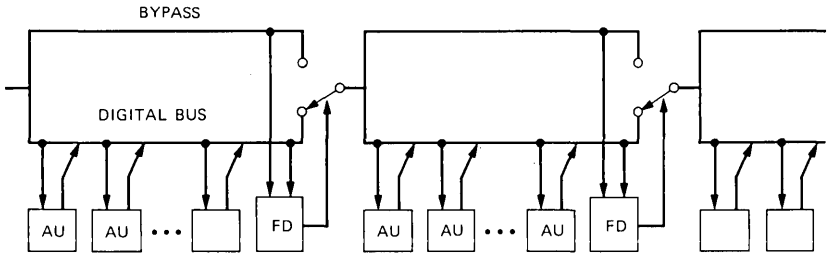


Fig. 11—Fault isolation using bypasses. Whenever the Fault Detector (FD) detects constraint violations, it switches input to next section to output of the bypass from the previous section.

With the memories divided into sections and with fault detection and isolation in place, failure impacts will be reduced. Unless the fault is in the fault detector itself, then if each memory is divided into m sections, only the N/m stations within a section will be affected by a fault. The disruption depends on which section and which point within the section is involved. Again assuming equal likelihoods for locations, we can derive the average number of potential connections that are disrupted and find this, to a good approximation, amounting to 0.75 (N^2/m).

Suppose that fault detectors have their own failure rate π_4 for failures that produce an open line and π_5 for switching off a section when it should not be. Further, suppose that these have expected repair times τ_4 and τ_5, or $\mu_4 = \pi_4\tau_4$ and $\mu_5 = \pi_5\tau_5$. On average, these events disrupt, respectively, $N^2/3$ and $N^2/2m$ potential connections.

The total EPAFI value, with division into m sections and fault detection and isolation, is then

$$U_m = [N\cdot(N-1)\cdot\mu_1] + [0.75\cdot N^3\cdot\mu_2/m] + [1.5 \cdot N^3\cdot\mu_3/m] + [2\cdot m\cdot N^2\cdot\mu_4/3] + [N^2\cdot\mu_5]. \quad (8)$$

The first and last terms in eq. (8) are much smaller than the others and may be neglected. The value of m that results in minimum U_m is

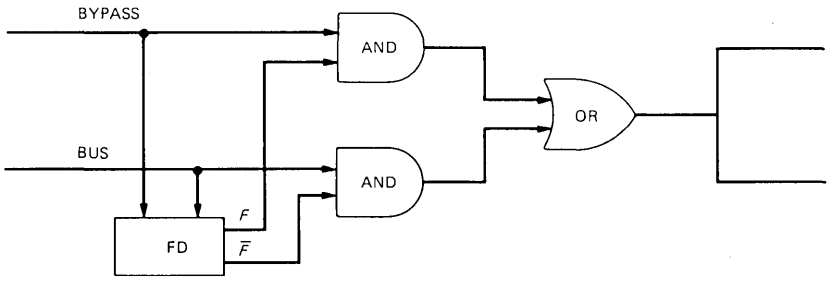


Fig. 12—Realization of two-way switch for fault isolation. FD outputs F and $\overline{F}$ are complementary, signifying FAULT and NO FAULT.

$$m = (3/2)\cdot\sqrt{N\cdot(\mu_2 + 2\cdot\mu_3)/2\cdot\mu_4}, \quad (9)$$

and the failure impact, with the first and last terms in eq. (8) neglected, comes to

$$U_{\text{opt}} = (4/3)\cdot N^{5/2}\cdot\sqrt{\mu_4\cdot(\mu_2 + 2\cdot\mu_3)/2}. \quad (10)$$

This should be compared with the expected failure impact without sectional detection and isolation, as found in eq. (7). The improvement ratio is

$$U/U_{\text{opt}} = (1/3)\cdot\sqrt{N\cdot(\mu_2 + 2\cdot\mu_3)/2\cdot\mu_4}. \quad (11)$$

Further improvement is possible by instituting super sections by which a number of consecutive sections would be bypassed and again fault-tested and isolated, as shown in Fig. 13. Indeed, one can take the hierarchy of protection to any number of levels.

With just one level of protection and, say, N = 10,000, and the different failure rates and repair times comparable to each other, the optimum number of sections would be around 185, or 55 AUs to one section. The improvement over no protection would be by a factor of 40. A second level of protection would increase the improvement by a further factor of around 5. Asymptotically, as the hierarchy of protection is taken to higher levels, the functional dependence of EPAFI on N becomes quadratic, which is the relationship that applies when the effect of a failure is confined to a single AU.

V. CONCLUSIONS

We have proposed a switch architecture that supports circuit- and packet-switched communications. Both kinds of communications can proceed at widely varying rates: Circuits can be set up with different capacities, selectable as a binary fraction or multiple of a basic capacity, while packet-switched communications share in the pool of the total switch capacity that is not in use at any given time. Thus, the proposed switch could cater efficiently in mediating real-time signals

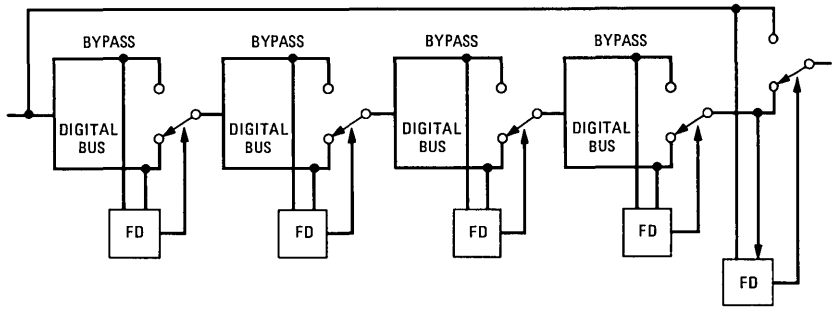


Fig. 13—Fault detection and isolation in super sections.

and data. Specifically, it could be a PBX that, apart from voice, could provide other circuit- and packet-switched services.

The possibility for the two modes is brought about by having enlarged time slots that can include addressing information, and then by making these of fixed length so that they can be made available regularly. Further, the switching is performed by access units that write and read on serial memories on which synchronization can be maintained without interruptions and information transfers can occur without collisions.

We have proposed that data packets be fixed at 192 bits, or 24 samples of pulse-code-modulated voice. This limits the delay due to packetization to 3 ms. The total delay, which includes propagation along the memories, will then be less than 4 ms, even in a very large switch.

It is recognized that a switch used in telephony should conform to very high standards of reliability. We have proposed a scheme of fault detection and isolation applicable to memories as in our switch. This would substantially overcome any added vulnerability due to the serial nature of the signal paths. However, other issues (e.g., overall system reliability) not addressed by us remain to be resolved. In summary, our proposed switch offers the possibility of integrating voice and data in a way that would preserve the quality and reliability of voice communications and therefore, in turn, could be integrated with the telephone system at large. We believe that, provided no compromises need to be made, very real benefits flow from having all communications mediated by a common facility. It is possible that our proposed switch could meet such objectives.

VI. ACKNOWLEDGMENT

The authors would like to thank M. G. Hluchyj for helpful discussions during early phases of this work.

REFERENCES

1. J. Bellamy, *Digital Telephony*, New York: Wiley, 1982.
2. H. E. Vaughan, "Research Model for Time-Separation Integrated Communication," B.S.T.J., *38* (July 1959), pp. 909–32.
3. H. Inose et al., "A Time-Slot Interchange System in Time-Division Electronic Exchanges," IEEE Trans. Commun. Syst., *CS-11* (September 1963), pp. 336–45.
4. Special Issue on No. 4 ESS, B.S.T.J., *56* (September 1977).
5. L. G. Roberts, "The Evolution of Packet Switching," Proc. IEEE, *66* (November 1978), pp. 1307–13.
6. A. G. Fraser, "Datakit—A Modular Network for Synchronous and Asynchronous Traffic," Proc. ICC (June 1979).
7. R. M. Metcalfe and D. R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Comm. ACM, *19* (July 1976), pp. 395–404.
8. G. T. Hopkins and P. E. Wagner, "Multiple Access Digital Communications Systems," U.S. Patent 4,210,780, issued July 1, 1980.
9. J. O. Limb and C. Flores, "Description of FASNET—A Unidirectional Local-Area Communications Network," B.S.T.J., *61* (September 1982), pp. 1413–40.

10. T. W. Forgie and A. G. Nemeth, "An Efficient Packetized Voice/Data Network Using Statistical Flow Control," IEEE Commun. Conf. III, 1977, pp. 38.2.44–48.
11. N. F. Maxemchuck, "A Variation on CSMA/CD That Yields Movable TDM Slots in Integrated Voice/Data Local Networks," B.S.T.J., *61* (September 1982), pp. 1527–50.
12. G. J. Coviello et al., "System Design Implications of Packetized Voice," IEEE Commun. Conf. III, 1977, pp. 38.3.48–53.
13. T. Bially et al., "Voice Communication in Integrated Digital Voice and Data Networks," IEEE Trans. Commun., *COM-28* (September 1980), pp. 1478–90.
14. D. H. Johnson and G. C. O'Leary, "A Local Access Network for Packetized Digital Voice Communication," IEEE Trans. Commun., *COM-29* (May 1981), pp. 679–88.
15. L. Kleinrock, *Queuing Systems, Vol. 1: Theory*, New York: Wiley-Interscience, 1975, *Vol 2: Computer Applications*, New York: Wiley-Interscience, 1976.
16. Z. L. Budrikis, J. L. Hullet, and D. Q. Phiet, "Transient-Mode Buffer Stores for Nonuniform Code TV," IEEE Trans. Commun. Technol., *COM-19* (December 1971), pp. 913–22.
17. E. T. Klemmer, "Subjective Evaluation of Transmission Delay in Telephone Conversations," B.S.T.J., *46* (July–August 1967), pp. 1141–7.
18. R. W. Downing, J. S. Nowak, and L. S. Tuomenoksa, "No. 1 ESS: Maintenance Plan," B.S.T.J., *43* (September 1964), pp. 1961–2019.
19. W. P. Karas, "Reliability and Maintainability Improvements Through Distributed Controls in Communication Systems," NTC Record 1981, pp A4.4.1–7.

AUTHORS

Zigmantas L. Budrikis, B.Sc., 1955, and B.E. (Hons I, Electrical Engineering), 1957, University of Sydney; Ph.D., 1970, University of Western Australia; P.M.G. (now Telecom Australia) Research Laboratories, 1958–1960; Aeronautical Research Laboratories, Fishermen's Bend, 1961; Electrical Engineering Faculty at University of Western Australia, 1962—. Mr. Budrikis has had a number of visiting appointments: University of California at Berkeley, 1968; AT&T Bell Laboratories, 1972, 1973, 1981, 1983, 1984; TU Munich, 1977. He is interested in problems in communications, man-machine interfaces, and foundations of electromagnetism. Fellow, IE Australia; member, IEEE, Optical Society of America, New York Academy of Science.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, Ph.D. (Electrical Engineering), 1970, Rice University; Optimal Data Corporation, 1970–1972; AT&T Bell Laboratories, 1972—. Mr. Netravali has worked on problems related to filtering, guidance, and control for the space shuttle. At AT&T Bell Laboratories, he has worked on various aspects of digital processing and computing. He was a Visiting Professor in the Department of Electrical Engineering at Rutgers University and the City College. He is presently Director of the Computer Technology Research Laboratory. Mr. Netravali holds over 20 patents and has had more than 60 papers published. He was the recipient of the Donald Fink Prize Award for the best review paper published in the Proceedings of the IEEE and the journal award for the best paper from the SMPTE. Editorial board, Proceedings of the IEEE; Editor, IEEE Transactions on Communications; senior member, IEEE; member, Tau Beta Pi, Sigma Xi.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

An Approximate Analysis of Sojourn Times in the M/G/1 Queue With Round-Robin Service Discipline

By P. J. FLEMING*

(Manuscript received February 24, 1984)

In most time-shared computer systems a program is processed by the central processing unit for, at most, a fixed period of time called a time slice, or quantum. If the program requires more processing after it has received its quantum, it is placed at the end of a run queue. This procedure is repeated until the program has finished executing. To the user who submitted the program the two most important performance measures of such a system are the mean and variance of the program's total elapsed time of execution. This total elapsed time is often referred to as the "response time". In this paper we investigate the effect of the quantum size on the mean and variance of the response time.

I. INTRODUCTION

The round-robin queue has been studied by several authors as a model of time-shared computer systems. In a time-shared system, the arrivals of requests for service as well as the service times may be thought of as random variables. From the user's point of view, the two most important measures of performance in such a system are the mean and variance of the response time. The round-robin discipline implicitly favors jobs with shorter service times, in the sense that the mean response time is approximately a linear function of the service time.[1] Thus far, however, the variance has proved to be intractable in

* AT&T Bell Laboratories.

the case of a general service-time distribution. In the case of exponential service times, Muntz has found the Laplace transform of the waiting-time distribution.[2]

The round-robin model can be described as follows: New arrivals join the end of the queue, and all jobs in the queue are served on a first-come first-served basis until they have completed their service requirement or have received one quantum of service. When a job has completed service, it leaves the system, and the next job in the queue begins service immediately. If a job requires more service after receiving its quantum, it rejoins the queue. In this paper we assume that the arrival process is Poisson but the service times are governed by a general distribution. Overhead due to switching between jobs can be included by adjusting the service requirement. For simplicity of exposition we assume that the quantum is constant. Variable quantum sizes also yield to our method of analysis.

In this paper we address the following questions:

1. What value of the quantum minimizes the mean sojourn time of a given class of jobs?

2. For a given class of jobs what is the variance of the sojourn time and what quantum minimizes the variance in the sojourn time?

Here "sojourn time" refers to the total amount of time that a job is in the queueing system, both in the queue and in service. To answer the second question, we use a new light traffic–heavy traffic interpolation (which we will call the RS interpolation) developed by M. Reiman and B. Simon, which makes use of a "light traffic derivative".[3]

In Section II we describe exactly how one calculates the mean waiting time in a round-robin queue. In Section III the results on minimizing the response time and a simple method for finding the optimal quantum are presented. The RS interpolation is described in Section IV, and in Section V we present some numerical examples.

II. THE MEAN WAITING TIME

The mean waiting time in a round-robin queue has been studied by several authors.[4,5] By "waiting" time we mean the total amount of time that the job spends in the queue (but not in service). In this section we describe the authors' analysis and set the notation that will be used throughout the paper. As mentioned above, we assume that the arrival process is Poisson with rate λ and that the service times of newly arriving jobs are independent, identically distributed random variables with distribution function F and density f. Let q denote the quantum size. We say that a job in the system is type j if its service time requirement as a newly arriving job is between $(j-1)q$ and jq, and we say that it is type (i, j) if it is type j and has received between $(i-1)q$ and iq units of service.

The following are additional notations:

m is the mean service time.

ρ is the traffic intensity, λm.

p_j is the probability a newly arriving job is type j.

m_{ij} is the mean amount of time that a type (i, j) job in the queue will occupy the server the next time it receives service.

M_{ij} is the second moment of the amount of time that a type (i, j) job in the queue will occupy the server the next time it receives service.

R_{ij} is the mean forward recurrence time of a type (i, j) job.

Q_{ij} is the mean number of type (i, j) jobs in the queue.

ω_i is the mean amount of time that a job must wait in the queue in the ith time in the queue.

W is the mean amount of time that an arbitrary job must wait in the queue before it completes its total service-time requirement.

2.1 Linear equations for the ω_i

A derivation of the following equations is contained in Ref. 5.

$$\omega_1 = \sum_{\substack{j\geq 1 \\ 1\leq i\leq j}} m_{ij}Q_{ij} + \lambda p_j m_{ij}R_{ij}, \tag{1}$$

and for $n \geq 2$

$$\omega_n = \sum_{\substack{j\geq n \\ 1\leq i\leq j-n+1}} m_{ij}Q_{ij} + \lambda p_j m_{ij}m_{n+i-1,j} + \sum_{i=1}^{n-1}\left(\sum_{k\geq i}\lambda p_k m_{ik}\right)(\omega_{n-i} + q). \tag{2}$$

Using Little's Law, which takes the form $Q_{ij} = \lambda p_j \omega_i$ in this case, we can eliminate Q_{ij} from (1) and (2). One can easily verify that the matrix form of the equations for the ω_i is

$$\omega = \rho M\omega + \rho b, \tag{3}$$

where ω is the vector whose ith component is ω_i, and M and b are a matrix and a vector, respectively, that are independent of ρ. Finally, the mean waiting time, W, is given by

$$W = \sum_{i=1}^{\infty} p_i\omega_i. \tag{4}$$

The following fact can be used to simplify the expression for b:

$$b_1 = 2m^{-1}\sum_{\substack{j\geq 1 \\ 1\leq i\leq j}} p_j M_{ij},$$

and for $n \geq 2$

$$b_n = q. \tag{5}$$

2.2 Favoring a class of jobs

Using the above results we can compute the mean waiting time W^1 of a particular class of jobs if we are given the service-time distribution F_1 of arrivals of that class:

$$W^1 = \sum_{j=1}^{\infty} (1 - F_1((j-1)q))\omega_j.$$

Figure 1 suggests that, at least in some fairly typical cases, the quantum size that minimizes the mean waiting time of a preferred class of jobs is neither very big [which is, in essence, First-In First-Out (FIFO)] nor very small (processor-sharing), but is just big enough to let all the preferred jobs complete service without having to feed back.

In this example, F_1 is uniform between one and three, F_2 is uniform between four and eight, and $F = 0.75\ F_1 + 0.25\ F_2$. In addition, $\lambda = 0.1$, so $\rho = 0.3$. Note that $q = 3$ minimizes W^1 over all q, and in addition, the mean waiting time of the complementary class decreases with increasing q.

2.3 A simple method for approximating the optimal quantum size

Since the expression for W^1 is complicated, it is quite difficult to find the optimal quantum q_0 without significant computational effort. The following observation (from numerical experiments) yields a simple method for finding q_0.

Observation: The value of q that minimizes $[dW^1(0)]/(d\rho)$ is close to the value of q that minimizes W^1 (for arbitrary values of ρ, $0 \leq \rho < 1$).

An easy calculation using (3) implies that

$$\frac{dW^1(0)}{d\rho} = \sum_{j=1}^{\infty} \{1 - F_1[(j-1)q]\}b_j.$$

Using (5) we see that

$$\frac{dW^1(0)}{d\rho} = b_1 + q \sum_{j=1}^{\infty} [1 - F_1(jq)].$$

III. THE RS INTERPOLATION

3.1 Description

We describe the interpolation as it is applied to the steady-state waiting time distribution in the round-robin queue with Poisson arrivals. The interested reader is directed to Ref. 3 for a general

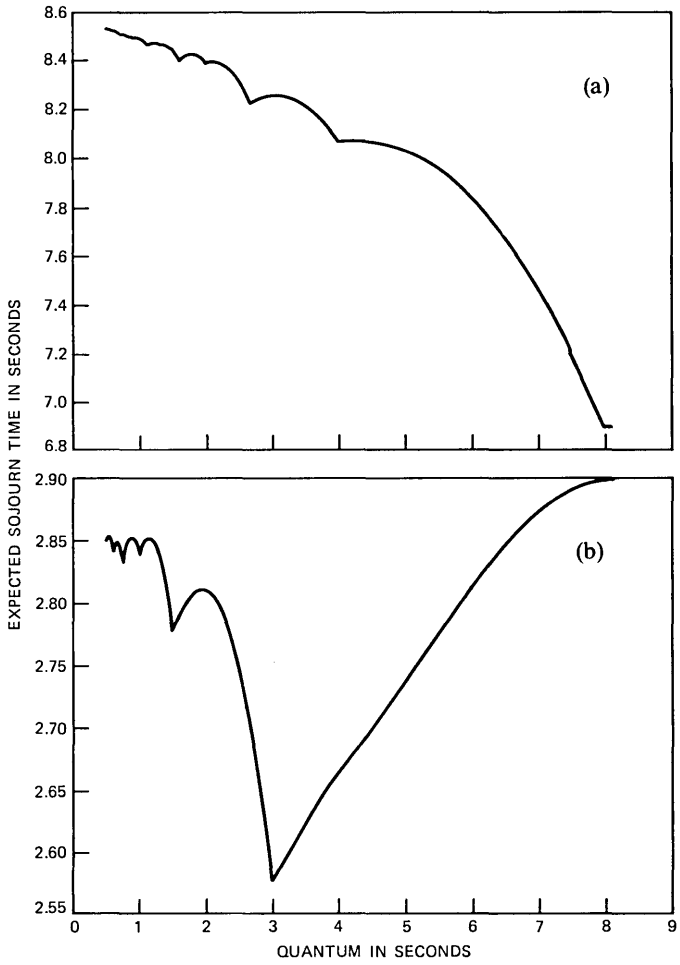


Fig. 1—Expected sojourn times as a function of the quantum for (a) type II jobs and (b) type I jobs.

treatment of the method. Let $W_n(\rho)$, $0 \leq \rho \leq 1$, be the nth moment of the steady-state waiting time distribution as a function of ρ, the traffic intensity. Let $\bar{W}_n(\rho) = (1 - \rho)^n W_n(\rho)$ when $0 \leq \rho < 1$, and let $\bar{W}_n(1) = \lim_{\rho \to 1}(1 - \rho)^n W_n(\rho)$. $\bar{W}_n(1)$ is well-defined and finite by the argument in Ref. 6.

Note that $\bar{W}_n(0) = W_n(0)$ is zero and $\bar{W}_n(1)$ can be calculated as the heavy traffic limit using a diffusion process. One can interpolate between light and heavy traffic to get the approximation formula:

$$W_n(\rho) \approx \frac{\bar{W}_n(1)\rho + \bar{W}_n(0)}{(1 - \rho)^n} = \frac{\bar{W}_n(1)\rho}{(1 - \rho)^n}.$$

This idea is, of course, well known. The novelty of the RS interpolation is that it makes use of the derivative of $W_n(\rho)$ at $\rho = 0$, which we will denote by $W_n'(0)$. It is clear that $W_n'(0) = \bar{W}_n'(0)$. The RS interpolation is

$$W_n(\rho) \approx \frac{(\bar{W}_n(1) - W_n'(0))\rho^2 + W_n'(0)\rho}{(1-\rho)^n}.$$

3.2 The light traffic derivative

Let q be a fixed quantum, and let $\sigma(x, a, b)$ be the total waiting time of a tagged job, J_1, that requires a units of service given that in the entire history of the system exactly one other job, J_2, arrives x units of time after J_1 and requires b units of service. Here, a and b are nonnegative real numbers, and x is any real number with negative x, implying that J_2 arrived before J_1. Figure 2 describes σ.
Let

$$\bar{\sigma}_n(a, b) = \int_{-\infty}^{\infty} \sigma(x, a, b)^n dx$$

and let F_i be the service-time distribution of J_i, $i = 1, 2$. The following theorem allows us to calculate $W_n'(0)$.
Theorem 1:

$$W_n'(0) = E(F_2)^{-1} \int_0^{\infty} \int_0^{\infty} \bar{\sigma}_n(a, b) dF_1(a) dF_2(b).$$

The proof of a more general version of this result is contained in Ref. 3. We now present a formula for $W_n'(0)$ using the above theorem. A straightforward calculation yields

$$\begin{aligned}\bar{\sigma}_n(a, b) = {} & \{b - \max[0, (\lfloor b/q \rfloor - \lfloor a/q \rfloor)q]\}^{n+1}/(n+1) \\ & + \max(0, \lfloor b/q \rfloor - \lfloor a/q \rfloor)[(\lfloor a/q \rfloor + 1)^{n+1} \\ & \quad - \lfloor a/q \rfloor^{n+1}]q^{n+1}/(n+1) \\ & + \max(0, \lfloor a/q \rfloor - \lfloor b/q \rfloor)qb^n + q^{n+1} \sum_{k=0}^{\min} k^n,\end{aligned}$$

where min = min $(\lfloor a/q \rfloor, \lfloor b/q \rfloor)$ and $\lfloor x \rfloor$ is the greatest integer less than x.

Hence,

$$\begin{aligned}W_n'(0) &= E(F_2)^{-1} \sum_{i,j \geq 0} \int_{jq}^{(j+1)q} \int_{iq}^{(i+1)q} \bar{\sigma}_n(a, b) dF_1(a) dF_2(b) \\ &= E(F_2)^{-1} \sum_{i,j \geq 0} (F_1((i+1)q) - F_1(iq)) \int_{jq}^{(j+1)q} \bar{\sigma}_n(i, b) dF_2(b).\end{aligned}$$

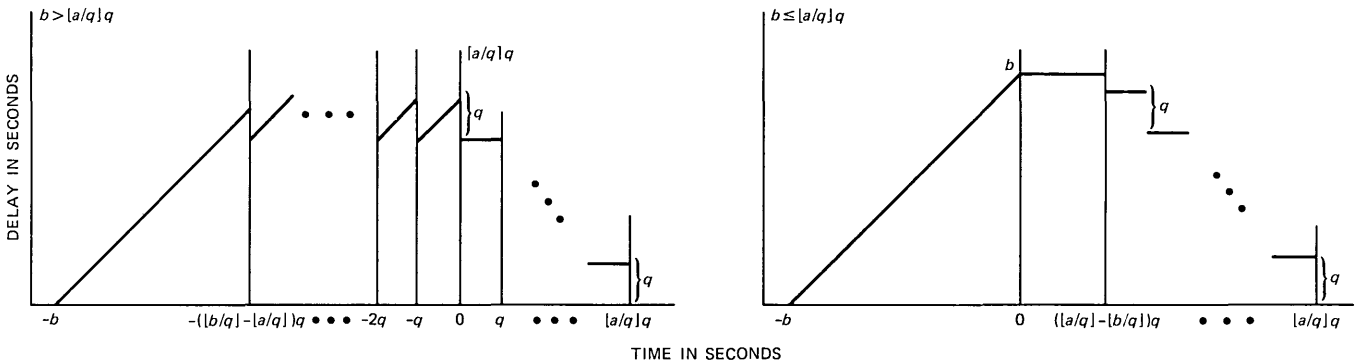


Fig. 2—A description of σ.

And

$$\int_{jq}^{(j+1)q} \bar{\sigma}_n(i, b)dF_2(b) = M^2_{j,n+1}[\max(0, j - i)q]/(n + 1)$$
$$+ M^2_{j,n}(0)\max(0, i - j)q$$
$$+ M^2_{j,0}\left[q^{n+1}\sum_{k=0}^{\min(i,j)} k^n + \max(0, j - i)\right.$$
$$\left.\cdot[(i + 1)^{n+1} - i^{n+1}]q^{n+1}/(n + 1)\right].$$

Here $M^2_{j,k}(x) = \int_{jq}^{(j+1)q} (b - x)^k dF_2(b)$.

3.3 The heavy traffic limit

The queueing system under consideration in this paper is a special case of the multiclass feedback queue analyzed in Ref. 6. Here we follow the development in Ref. 7 for the readers' convenience.

Let $U(t)$ denote the unfinished work process. Formally,

$$U(t) = V(t) - \inf_{0<s<t}\{V(s)\},$$

where $V(t) = L(t) - t$ and $L(t)$ is the total amount of work entering the system in $[0, t]$. (We assume the system is empty at $t = 0$.) Define a sequence of systems whose parameters, and queue-length and sojourn-time processes are indexed by $n \geq 1$, and consider the normalized processes

$$\hat{U}_{(n)}(t) = \frac{U_{(n)}(nt)}{\sqrt{n}}, \qquad n \geq 1$$

over some finite interval, which we normalize to $[0, 1]$ for convenience. For a heavy traffic limit we assume $\lambda^{(n)} \to m^{-1}$ as $n \to \infty$. We have the result of D. C. Igelhart and W. Whitt (1971).

Theorem 2: If

$$\lim_{n\to\infty}(\rho^{(n)} - 1)\sqrt{n} = c, \qquad -\infty < c < \infty,$$

then as $n \to \infty$

$$\hat{U}^{(n)} \Rightarrow \hat{U} = \text{RBM}[c, m^{-1}(s + m^2)],$$

where $\Rightarrow$ *denotes weak convergence.*

Here s is the variance in the service times, and a is the variance in the interarrival times. RBM(d, σ^2) is one-dimensional reflected Brownian motion with drift d and infinitesimal variance σ^2.

Now consider the normalized queue-length process

$$\hat{Q}_{ij}^{(n)}(t) = \frac{Q_{ij}^{(n)}(nt)}{\sqrt{n}}, \qquad 0 \le t \le 1,\ n \ge 1,\ j \ge 1,\ 1 \le i \le j.$$

We have

$$U^{(n)}(t) \approx \sum_{j\ge 1} \sum_{i=1}^{j} \hat{Q}_{ij}^{(n)}(t) \sum_{k=i}^{j} m_{kj}.$$

Here $\approx$ means

$$\lim_{n\to\infty} \sup_{0\le t\le 1} \left| \hat{U}^{(n)}(t) - \sum_{j\ge 1} \sum_{i=1}^{j} \hat{Q}_{ij}^{(n)}(t) \sum_{k=i}^{j} m_{kj} \right| = 0$$

in probability.

Theorem 3: Let

$$Y^{(n)}(t) = \sup |\, \lambda p_{j'} \hat{Q}_{ij}(t) - \lambda p_j \hat{Q}_{i'j'}^{(n)}(t) \,|,$$

where the supremum is taken first over all $1 \le i \le j$ *and* $1 \le i' \le j'$ *and then over all* $j, j' \ge 1$. *If* $\lambda^{(n)} \to m^{-1}$, *then as* $n \to \infty$, $\sup_{0\le t\le 1} Y^{(n)}(t)$ *converges in probability to zero.*

Using these results one can prove the following.

Theorem 4: Under the conditions of Theorem 2,

$$\hat{Q}_{ij}^{(n)} \Rightarrow \lambda p_j \Gamma^{-1} \hat{U}$$

as $n \to \infty$, *where*

$$\Gamma = \sum_{j\ge 1} \lambda p_j \sum_{i=1}^{j} i m_{ij}.$$

Next we consider sojourn times. Again we use the normalization

$$\hat{\omega}_i^{(n)}(t) = \frac{\omega_i^{(n)}(nt)}{\sqrt{n}}, \qquad 0 \le t \le 1,\ n, i \ge 1.$$

For single-pass jobs (i.e., $i = 1$),

$$\hat{\omega}_1^{(n)}(t) \approx \sum_{\substack{j\ge 1 \\ 1\le i\le j}} m_{ij} \hat{Q}_{ij}^{(n)}(t).$$

So from Theorem 4 we see that

$$\hat{\omega}_1^{(n)}(t) \approx \Gamma^{-1} \hat{U}^{(n)}(t) \sum_{\substack{j\ge 1 \\ 1\le i\le j}} \lambda p_j m_{ij}.$$

Thus, from $\rho^{(n)} \to 1$ and Theorem 3 one can prove Theorem 5.

Theorem 5: Under the conditions of Theorem 2,

$$\hat{\omega}_i^{(n)} \Rightarrow i\Gamma^{-1} \hat{U}$$

as $n \to \infty$.

Let $\hat{\omega}_1(t) = \Gamma\hat{U}(t)$ so that

$$\hat{\omega}_1(t) = \text{RBM}[-\Gamma^{-1}, \Gamma^{-2}(\lambda s + \lambda^{-1})].$$

Since we are interested only in the stationary behavior of the queueing system, we observe that the stationary density of RBM (α, β), $(\alpha < 0)$ is

$$2 \mid \alpha \mid \beta^{-1}\exp(2\alpha\beta^{-1}x).$$

So the kth moment of the stationary distribution of RBM(α, β) is

$$k!\,(2 \mid \alpha \mid \beta^{-1})^{-k}$$

for $k = 1, 2, \cdots$.

Let $\omega_i^k(\rho)$, $0 \le \rho < 1$ be the kth moment of the amount of time a job waits in the queue the ith time in the queue, and let

$$\tilde{\omega}_i^k(\rho) = \begin{cases} \omega_i^k(\rho)(1-\rho)^k & 0 \le \rho < 1 \\ \lim\limits_{\rho \to 1} \omega_i^k(\rho)(1-\rho)^k & \rho = 1. \end{cases}$$

The above results yield Theorem 6.

Theorem 6: For all i,

$$\tilde{\omega}_i^k(1) = \omega_1^k(1) = k![2\Gamma/(\lambda s + \lambda^{-1})]^{-k}.$$

Furthermore, $\bar{W}_n(1)$, is given by

$$\bar{W}_k(1) = \tilde{\omega}_1^k(1) \sum_{j \ge 1} j^k p_j.$$

V. SOME NUMERICAL EXAMPLES

In this section, we present numerical examples of two sorts. Figures 3 and 4 demonstrate the accuracy of the interpolation when applied to the mean sojourn time. Here we are comparing the interpolation with the exact formula given in Section II. In addition, these examples tell us that for some typical load situations, the quantum that minimizes the sojourn time for the type I jobs does not seriously degrade sojourn time for the type II jobs.

Figures 5 through 10 present the approximation to the variance in the waiting time for some typical choices of service times using the RS interpolation to approximate the second moment $W_2(\rho)$ of the waiting time. An interesting thing here is the similarity between the mean and variance regardless of the service-time distribution. Notice that the qualitative properties of the mean and variance of the sojourn time as a function of the quantum size are quite sensitive to the service-time distribution. In the case of exponential service times it appears that the waiting time gets smaller as the quantum gets smaller,

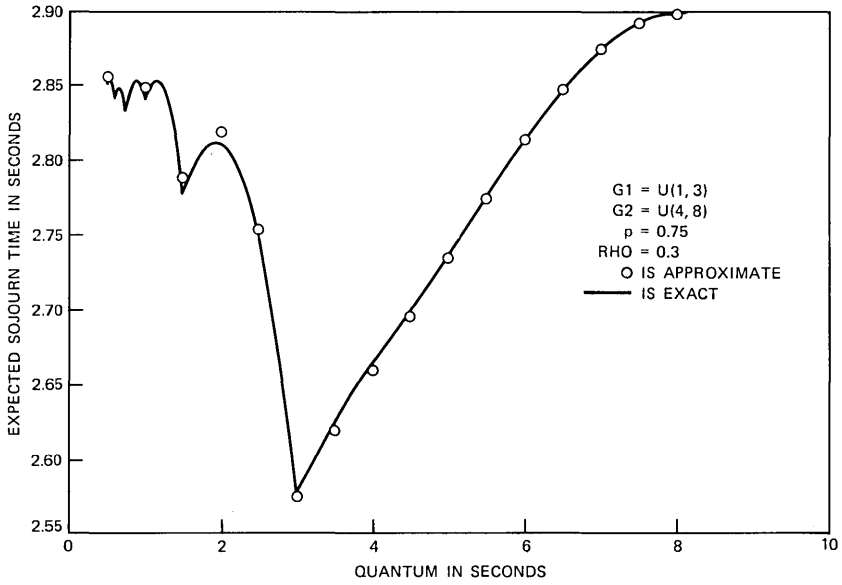


Fig. 3—Expected sojourn time for type I jobs.

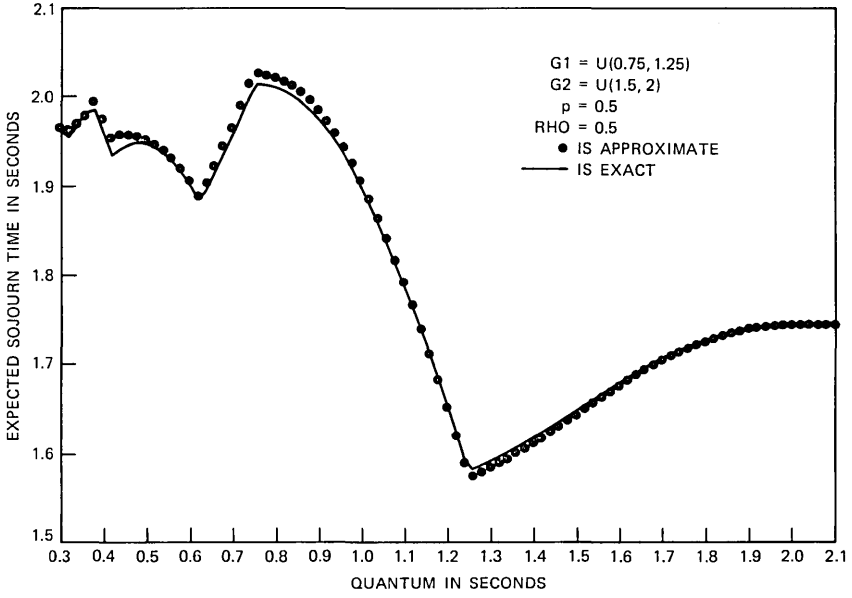


Fig. 4—Expected sojourn time for type I jobs.

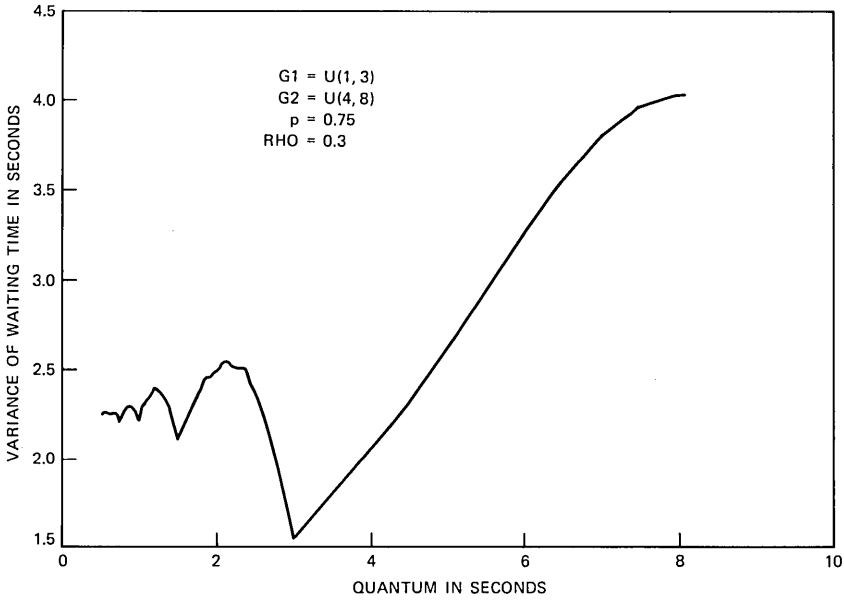


Fig. 5—Variance of waiting time for type I jobs.

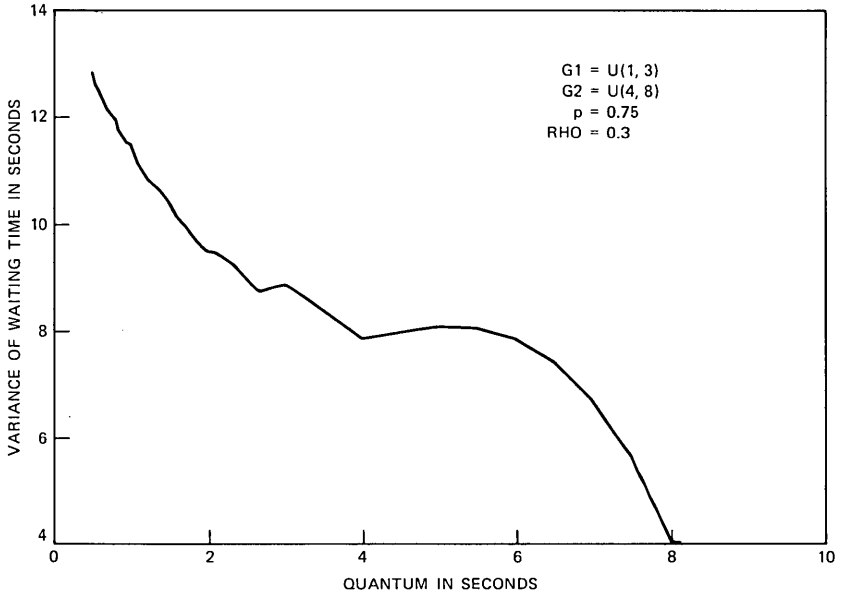


Fig. 6—Variance of waiting time for type II jobs.

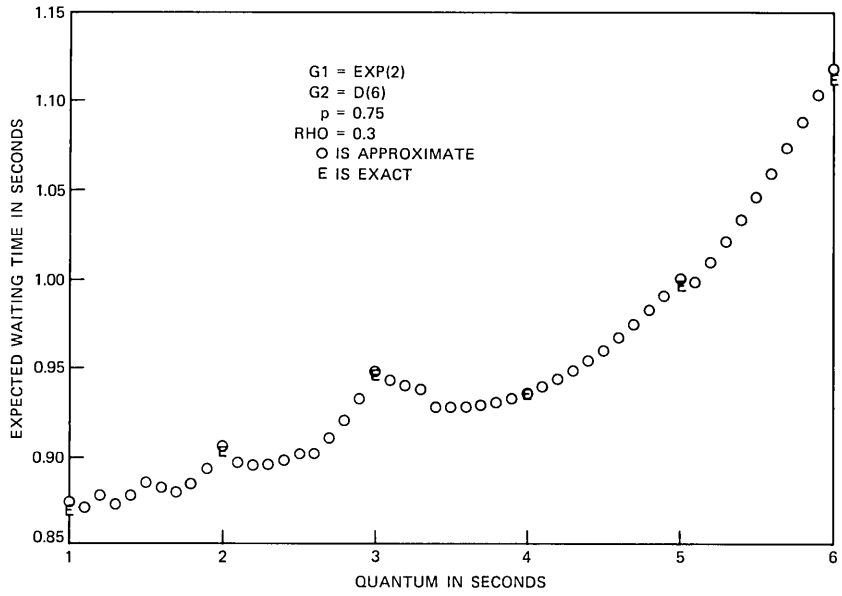


Fig. 7—Expected waiting time for type I jobs.

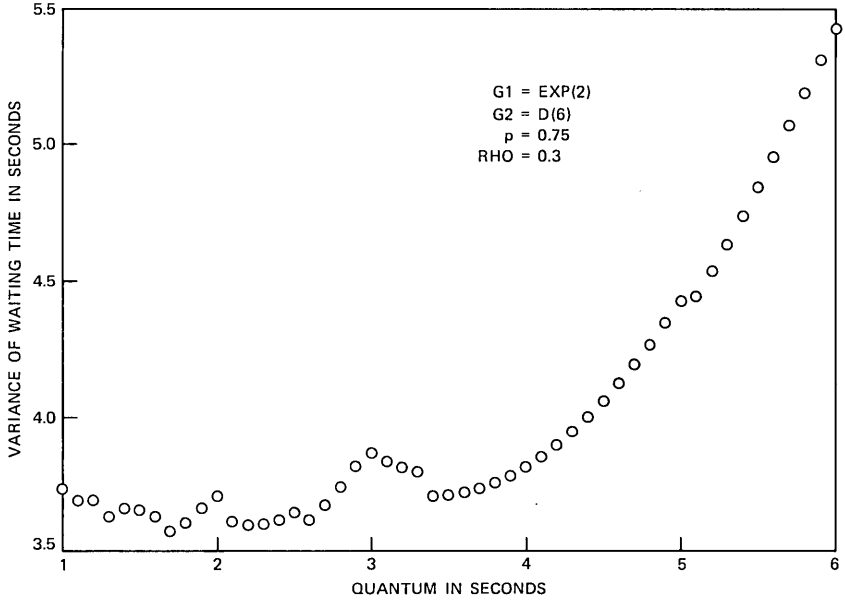


Fig. 8—Variance of waiting time for type I jobs.

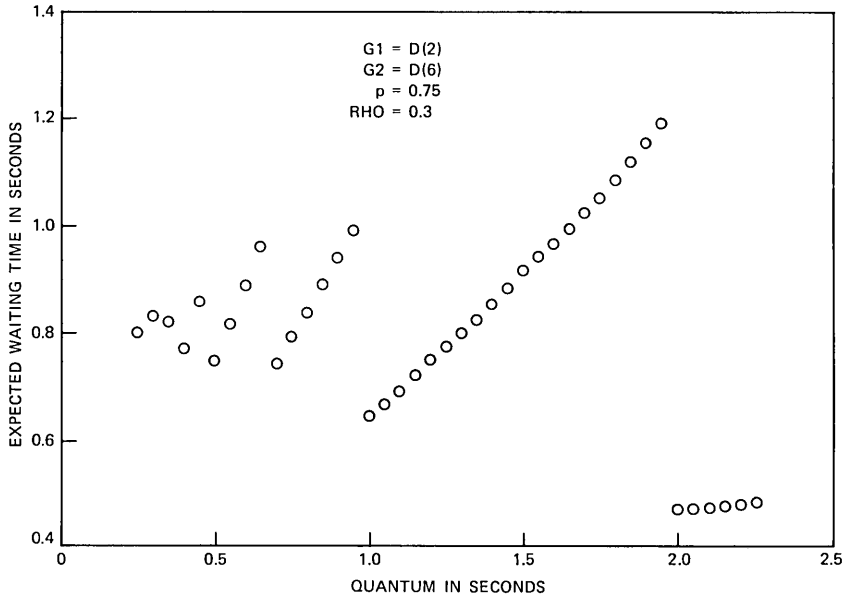


Fig. 9—Expected waiting time for type I jobs.

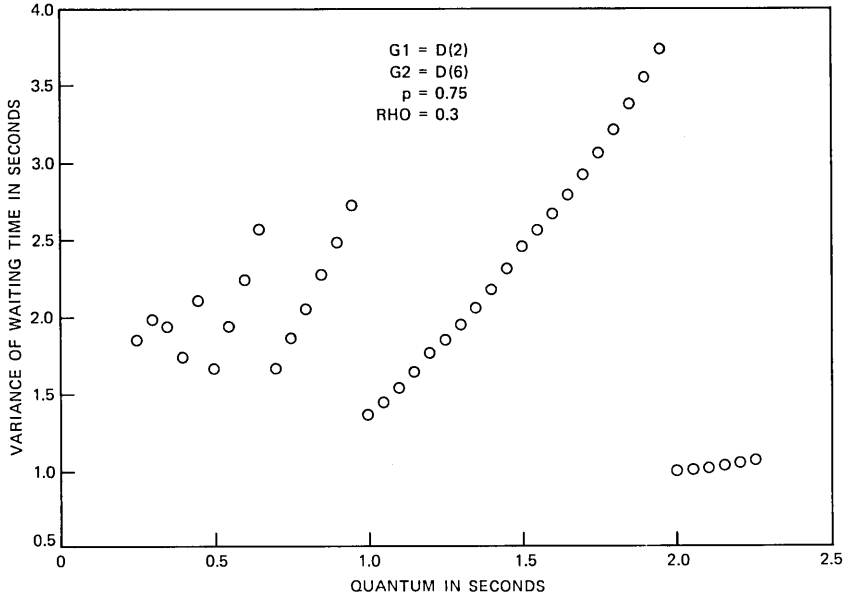


Fig. 10—Variance of waiting time for type I jobs.

whereas in the deterministic and uniform cases it is best to allow all the favored jobs to finish in one quantum.

The following notations are used in Figs. 3 through 10.

rho	is the overall traffic intensity of the example queue.
p	is the proportion of jobs that are type I.
G1	is the service-time distribution of the type I jobs.
G2	is the service-time distribution of the type II jobs.
EXP(X)	means an exponential distribution with mean X.
D	means a deterministic distribution with mean X.
U(X,Y)	means a distribution that is uniform between X and Y.

VI. ACKNOWLEDGMENT

I would like to thank Marty Reiman and Burt Simon for many useful suggestions.

REFERENCES

1. L. Kleinrock, *Queueing Systems*, Vol. 2, New York: Wiley, 1976, pp. 166–70.
2. R. Muntz, "Waiting Time Distribution for Round-Robin Queueing Systems," Symp. on Computer-Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn (April 4–6, 1972), pp. 429–39.
3. M. I. Reiman, and B. Simon, unpublished work.
4. M. Sakata, S. Noguchi, and J. Oizumi, "An Analysis of the M/G/1 Queue Under Round-Robin Scheduling," Oper. Res., *19*, No. 1 (March–April 1971), pp. 317–85.
5. R. W. Wolff, "Time Sharing With Priorities," SIAM J. Appl. Math., *19*, No. 3 (November 1970), pp. 566–74.
6. M. I. Reiman, unpublished work.
7. E. G. Coffman, Jr. and M. I. Reiman, "Diffusion Approximations for Computer/Communication Systems," in *Mathematical Computer Performance and Reliability*, G. Iazeolla, T. J. Courtois, and A. Hortijk, eds., New York: North Holland, pp. 33–53.
8. D. W. Igelhart and W. Whitt, "Multiple Channel Queues in Heavy Traffic," Advance. Appl. Probab., *2* (1970), pp. 150–77, 355–64.

AUTHOR

Philip J. Fleming, B.A. (Mathematics), 1974, Wayne State University; M.A. and Ph.D. (Mathematics), University of Michigan, Ann Arbor, Michigan, 1977 and 1981, respectively; AT&T Bell Laboratories, 1982—. Before joining AT&T Bell Laboratories, Mr. Fleming was Assistant Professor of Mathematics and Statistics at Case Western Reserve University in Cleveland, Ohio, from 1981 through 1982. Mr. Fleming joined AT&T Bell Laboratories as a Member of Technical Staff in the Systems Design and Exploratory Development department. His work has been in the area of computer performance modeling and analysis, queueing theory, and cryptography as it relates to computer security. He is currently a member of the Operating System Architecture department. Member, AMS and ORSA.

AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

Analysis of a TDMA Network With Voice and Data Traffic

By M. L. HONIG*

(Manuscript received May 16, 1983)

An analysis of an integrated voice-data network with Demand Assignment Time Division Multiple Access (TDMA) is presented using the following model: (1) voice calls that cannot be serviced are blocked, whereas requests to transmit data messages are queued; (2) no traffic boundaries are assumed, i.e., any new traffic arrival may be assigned to any unassigned time slot; (3) message lengths are exponentially distributed with the mean voice message length assumed to be much larger than the mean data message length; (4) traffic requests are generated according to two independent Poisson processes; and (5) time slot assignments are made instantaneously and no priorities are assumed. Such a model applies to a single-channel TDMA network in which voice and data traffic arrivals are serviced on a first-come first-served basis. An approximate analysis, based upon physical insight, is presented that yields the blocking probability for voice messages, the mean number of queued data requests, and the mean value of the peaks of the data queue process. Comparisons with simulation results indicate that the analytical results are very accurate. Performance curves are presented and compared with analogous results for TDMA networks that handle only one traffic type.

I. INTRODUCTION

The popularity of integrated voice-data networks has motivated numerous analyses of associated network queueing models.[1-7] This paper presents an analysis of a voice-data network using Demand Assignment Time Division Multiple Access (DA/TDMA). This work

* AT&T Bell Laboratories; present affiliation Bell Communications Research, Inc.

evolved from a study of a multichannel DA/TDMA protocol that handles different traffic types, and that has received much attention in the context of satellite communications.[8,9] These networks typically have a few hundred megabits of capacity and can be used to carry real-time digital voice traffic in addition to data traffic. Analysis in Ref. 10 indicates that a large multichannel TDMA network (i.e., a network with more than 100 communicating traffic nodes) behaves much like a single-channel TDMA network with an equivalent number of time slots per frame. We therefore concentrate on the simpler single-channel network and attempt to characterize its performance when used to handle both real-time voice and data traffic. The results in this paper carry over to the multichannel case when the number of traffic nodes in the network is large.[10]

A TDMA protocol divides the broadcast channel into a series of time slots of identical width. A prespecified number of time slots forms a TDMA frame that continually repeats itself. A demand assignment protocol assumes that when a traffic source has a message to transmit, it must first send a message to a central controller indicating that it wishes to transmit a message to a specified destination address. The central controller assigns specific time slots to each received request on a noninterfering basis. Only one time slot per frame is assigned to each request. Once a time slot is assigned, it remains assigned to the same traffic source for the duration of the message. We therefore assume that each data message consists of a variable number of packets, where the length of a packet is the number of bits per time slot. Voice calls are assigned a dedicated time slot for the duration of the call. (Full-duplex voice traffic actually requires two time slots per frame, one for each direction.) Figure 1 shows an example in which there are four time slots per frame. The numbers in each slot specify the source and destination addresses. The controller has assigned slot 1 to the source-destination pair 1-2. Since the message generated by source 1 requires more than one time slot, source 1 also uses the first time slot in the succeeding frame. The number of time slots per frame and number of bits per time slot are design parameters that can vary from system to system. Notice, however, that an additional constraint might be that the number of frames per second and number of bits

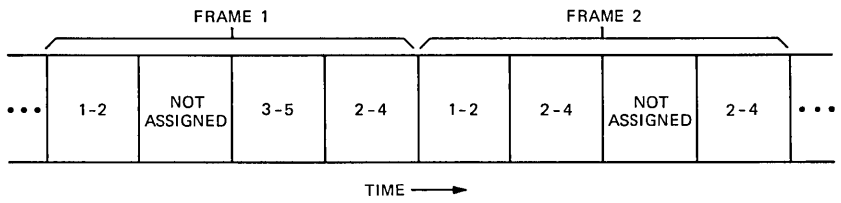


Fig. 1—Assignment of source-destination pairs to time slots.

per time slot must be selected so that each time slot represents a 64-kb/s circuit, which is necessary to provide toll-quality PCM voice transmission.

As the source traffic intensity increases, so does the probability of not being able to assign a new voice or data request because all slots have already been assigned. New voice requests that cannot be assigned are blocked, whereas unassigned data requests enter a queue. As slots become free, requests from the queue are then assigned. The network model analyzed here is different from the models used in Refs. 1 through 7 in one or more of the following respects: (1) Each channel time slot can be assigned to either traffic type. In particular, no moving boundaries[1,6,7] are assumed that partition the time slots in each frame into a section reserved for voice traffic and a section reserved for data traffic. (2) The duration of each message measured in number of time slots, or equivalently, the number of TDMA frames, is an exponentially distributed random variable. Furthermore, the mean voice message length is assumed to be at least an order of magnitude larger than the mean data message length. (3) No priorities are assumed, so that traffic is serviced on a first-come first-served basis.

To simplify the analysis, approximations based upon physical insight are used. Results are expressions for voice blocking probability, mean number of queued data requests, and the mean value of the peaks of the data queue process as functions of the number of time slots and traffic parameters. Comparisons with simulation results indicate that these analytical results are quite accurate.

The next section describes the network queueing model in detail. Section III presents the analytical results, and Section IV presents performance curves and compares them with analogous curves for systems having only one input traffic type.

II. NETWORK QUEUEING MODEL

The TDMA network is modeled as a c-server queueing system, where c is the number of time slots per TDMA frame. We assume voice and data traffic requests to be generated according to two independent Poisson processes with respective arrival rates λ_v and λ_d. We therefore implicitly assume an infinite source model. Service times (or message lengths) in both cases are assumed to be exponentially distributed, with service rates μ_v for voice messages and μ_d for data messages. In particular, the mean number of time slots required for a voice message is $1/\mu_v$. Notice that in practice the service distribution must be discrete, since messages can only last an integral number of time slots. The continuous distribution assumed here is a good approximation to a "discrete" exponential distribution as long as the

typical message length is a relatively large number of time slots (i.e., > 10).

Each voice or data arrival demands one time slot per frame for the duration of the message. The rates (i.e., bits per second) at which both voice and data messages access the channel are therefore identical. Voice messages that cannot be assigned a time slot immediately are blocked (i.e., disappear), whereas unassigned data requests enter a queue. To simplify the analysis, we assume that time slot assignments occur instantaneously, rather than at the beginning of the next frame. In particular, as soon as a request is generated, it is immediately assigned, assuming unassigned slots exist. This approximation is reasonable as long as the average message lengths are much greater than one time slot. Given that there are queued data requests, they are assigned as soon as time slots are relinquished by traffic already assigned. Voice messages that arrive while data requests are queued are therefore blocked. As either the voice or data traffic intensity increases, the data traffic tends to grab enough time slots to empty any queue that may appear, thereby depriving voice traffic of available time slots. One therefore expects that under high traffic intensities, voice blocking probability is quite high, whereas the mean data queue length (i.e., number of queued requests) is moderate.

An exact analysis of the model just described can be performed by extracting the associated embedded Markov chain. In this case, the two-dimensional state defining the embedded Markov chain is (d, v), where d and v represent the number of data and voice messages, respectively, in the system. Transition probabilities are easily determined in terms of the traffic parameters and number of time slots, thus a solution for the steady-state probability distribution $p(d, v)$ can be theoretically obtained. If we assume that the data message queue can be arbitrarily large, however, the number of states in the Markov chain becomes infinite. As the dimension of the problem increases, the amount of numerical computation required to obtain $p(d, v)$ increases, which in turn causes further propagation of round-off errors. The exact analysis just outlined was attempted.[11] However, it was not successful due to finite word-length effects. The approximate analysis in the next section is therefore proposed as a simple alternative.

III. ANALYTICAL RESULTS

We start by deriving an approximate expression for voice blocking probability. In steady state, the minimum number of time slots per frame needed to ensure that the system remains stable (i.e., the number of queued data requests does not become infinite with probability one) is

$$\left\lfloor \frac{\lambda_d}{\mu_d} \right\rfloor + 1,$$

where $\lfloor x \rfloor$ denotes the largest integer less than x. If the number of slots is less than this amount, then the system would be unstable even with no additional voice traffic. If the number of slots is greater than this amount, then the system must be stable since unassigned voice requests disappear, and because voice requests cannot preempt queued data requests. All queued data requests must therefore be assigned before any voice messages can be assigned.

The number of time slots per frame available for data messages is a random process that varies according to how many time slots are assigned to voice traffic. Assuming that the mean service time for voice messages ($1/\mu_v$) is orders of magnitude greater than the mean service time for data messages ($1/\mu_d$), the voice "state", i.e., number of voice-occupied slots per frame, varies much more slowly than the data state, which is the total number of data messages present in the system. It is therefore a good approximation to assume that the time spent in each voice state is much longer than the time it takes the number of data requests present in the system to reach steady-state behavior. This "steady-state approximation" is the basis for the analysis that follows. Using this approximation, it follows that if the number of voice messages in the system is greater than or equal to

$$v_0 \equiv c - \left\lfloor \frac{\lambda_d}{\mu_d} \right\rfloor, \tag{1}$$

the normalized data traffic intensity conditioned on the number of voice-occupied slots, $\lambda_d/[(c - v_0)]\mu_d$, is greater than one, and data requests become queued with probability one. Since the number of queued data requests is assumed to reach steady-state behavior, this queue cannot be emptied until a voice message relinquishes a time slot. The steady-state approximation therefore implies that the number of voice messages in the system is never greater than v_0. Computer simulations of the queueing model considered have verified that the probability of the voice state v becoming greater than v_0 is indeed very small when μ_v is much less than μ_d. The "competition" of voice and data traffic for available time slots can therefore be expected to produce intermittent queue "spikes," representing times at which the voice state $v = v_0$. During this time, the data queue process experiences a "transient instability."

Given that v time slots are assigned to voice traffic, the probability that an incoming voice message is blocked is equal to the probability that the number of data requests in the system, d, is greater than or equal to $c - v$. Using the steady-state approximation, this is simply

the steady-state probability that a queue exists in an $M/M/c-v$ queueing system and is given by[12]

$$p(d \geq c - v \mid v) = \frac{p_0}{(c-v)!}\left(\frac{\lambda_d}{\mu_d}\right)^{c-v} \frac{1}{1-(\lambda_d/\mu_d)}, \tag{2}$$

where

$$p_0 = \left[\sum_{k=0}^{c-v-1} \frac{1}{k!}\left(\frac{\lambda_v}{\mu_v}\right)^k + \frac{1}{(c-v)!}\left(\frac{\lambda_v}{\mu_v}\right)^{c-v} \frac{1}{1-(\lambda_v/\mu_v)}\right]^{-1}. \tag{3}$$

The blocking probability for voice traffic is therefore

$$P_B = \sum_{v=0}^{v_0} p(d \geq c - v \mid v)p(v), \tag{4}$$

where $p(v)$ is the probability that v time slots are assigned to voice traffic.

Consider now a blocking system with v_0 servers and one Poisson input. Given $v < v_0$, let ϕ_v denote the probability that a new arrival can be served (i.e., even though all servers are not busy, a newly arriving request is blocked with probability $1 - \phi_v$). Assuming exponential service times, the probability that v servers are busy is known to be[12]

$$p(v) = \frac{\frac{1}{v!}\left(\frac{\lambda_v}{\mu_v}\right)^v \prod_{j=0}^{v-1} \phi_j}{\sum_{i=0}^{v_0}\left[\frac{1}{i!}\left(\frac{\lambda_v}{\mu_v}\right)^i \prod_{j=0}^{i-1} \phi_j\right]}. \tag{5}$$

This exactly describes the voice-data system considered, where the "entrance" probability,

$$\phi_v = p(d \geq c - v \mid v), \tag{6}$$

and is given by (2). Substituting (2) and (5) into (4) therefore gives the desired result. Notice that because the arrival processes are Poisson, the blocking probability P_B is equal to the probability of being in a blocking state (i.e., all time slots are busy), which is equal to the probability that data requests are queued.

An analogous argument can be applied to compute the mean number of queued data requests. Denoting this queue length as q, we have

$$E(q) = \sum_{v=0}^{v_0} p(v)E(q \mid v), \tag{7}$$

where $E(q \mid v)$ is the mean number of queued data requests given v assigned voice messages. Assuming $\mu_v \ll \mu_d$ implies that $E(q \mid v)$ is

approximately equal to the mean queue length for an $M/M/c-v$ queueing system,[12] i.e.,

$$E(q \mid v) \approx \frac{p_0}{c-v}\left(\frac{\lambda_v}{\mu_v}\right)^{c-v} \frac{(c-v)\lambda_v\mu_v}{[(c-v)\mu_v - \lambda_v]^2}, \qquad 0 \le v < v_0, \quad (8)$$

where p_0 is given by (3). If $v = v_0$, this expression no longer applies, however, since the system becomes unstable. To approximate $E(q \mid v_0)$, note that when $v = v_0$, the mean data queue length increases approximately at rate $\lambda_d - (c - v_0)\mu_d$ until a voice message relinquishes its time slot. At this point the queue starts to empty at rate $(c - v_0 + 1)\mu_d - \lambda_d$. As the queue empties, more time slots may be relinquished by voice messages, causing the queue to empty at a faster rate. Suppose that we assume

$$E(q \mid v_0) \approx \frac{1}{t_0}\int_0^{t_0} E[q(t)]dt, \quad (9)$$

where t_0 is the duration of the queue spike and $q(t)$, $0 \le t \le t_0$, is the queue length as a function of time (given that v increased from $v_0 - 1$ to v_0 at $t = 0$). Furthermore, we assume that $E[q(t)]$ is piecewise linear (fluid flow approximation).[4] Then it is shown in Appendix A that

$$E(q \mid v_0) \approx \frac{1}{t_0}\left\{\frac{1}{2}\frac{\lambda_d - \mu_d}{(v_0\mu_v)^2} + \frac{\bar{q}_{i_0}^2}{2[(c - v_0 + i_0 + 1)\mu_d - \lambda_d]} + \sum_{j=1}^{i_0}\frac{1}{2(v_0 - j)\mu_v}[\bar{q}_{j-1} + \bar{q}_j]\right\}, \quad (10)$$

where

$$t_0 = \frac{\bar{q}_i}{(c - v_0 + i_0 + 1)\mu_d - \lambda_d} + \sum_{j=0}^{i_0}\frac{1}{(v_0 - j)\mu_v}, \quad (11)$$

$$\bar{q}_i = \frac{\lambda_d - \mu_d}{v_0\mu_v} - \sum_{j=1}^{i}[(c - v_0 + j)\mu_d - \lambda_d]\frac{1}{(v_0 - j)\mu_v}, \quad (12)$$

and

$$i_0 = \max\{i \mid \bar{q}_i > 0\}. \quad (13)$$

Substituting (5), (8), and (10) into (7) therefore gives the approximate mean queue length.

The expressions for voice blocking probability and mean data queue length presented thus far have been found to be quite accurate when compared with simulation results. To gain further insight into the behavior of the system, however, we now attempt to characterize the transient instabilities, or queue spikes, which occur when $v = v_0$.

3.1 Analysis of transient instabilities

Figure 2 illustrates the problem under consideration. The queue spikes appear whenever the voice state is equal to v_0. Not shown are queues that occur when the voice state v is less than v_0. The height of each spike is denoted as d_M, the duration of each spike is denoted as τ_w, and the time between spikes is denoted as T. Figure 2 is not meant to indicate a typical sample function for the data message queue process. Significant queues appear when $v < v_0$; however, the purpose of the following analysis is to determine whether the transient instabilities shown in Fig. 2 cause serious performance degradation.

We begin by computing the distribution of the peak value of each spike. Let $p(\tilde{d}, t)$ denote the probability that $\tilde{d}$ data messages are *queued* at time t, given that $v = v_0$. At $t = 0$ we assume $\tilde{d} = 0$. The following equations can be derived in a straightforward manner,[12]

$$\frac{d}{dt}p(\tilde{d}, t) = \lambda_d p(\tilde{d} - 1, t) - [(c - v_0)\mu_d + \lambda_d]p(\tilde{d}\ t) + (c - v_0)\mu_d p(\tilde{d} + 1, t) \quad \text{for} \quad \tilde{d} > 0 \tag{14a}$$

and

$$\frac{d}{dt}p(0, t) = -\lambda_d p(0, t) + (c - v_0)\mu_d p(1, t). \tag{14b}$$

Solving (14) gives the probability that the maximum queue length is equal to $\tilde{d}$ given the time until the first voice departure is t. (Recall that as soon as v decreases from v_0 to $v_0 - 1$, the mean queue length decreases.) We know, however, that the time until the first voice departure is exponentially distributed with parameter $v_0\mu_v$, so that

$$\begin{aligned} q_M(d_M) &\equiv \Pr\{\text{maximum queue length} = d_M\} \\ &= \int_0^\infty v_0\mu_v e^{-v_0\mu_v t}p(d_M, t)dt \\ &= v_0\mu_v Q_M(v_0\mu_v, d_M), \end{aligned} \tag{15}$$

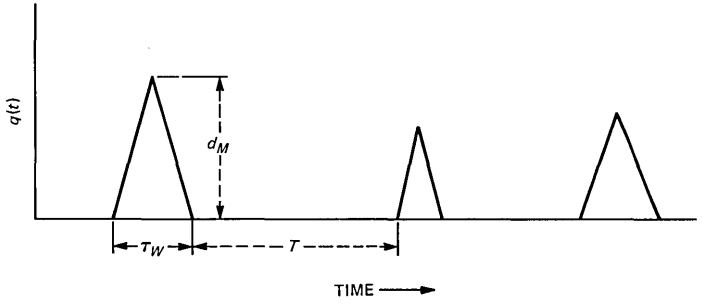


Fig. 2—Transient instabilities in the data queue process.

where

$$Q_M(s, d_M) = \int_0^\infty e^{-st} p(d_M, t)dt \tag{16}$$

is the Laplace transform of $p(d_M, t)$. Equations (14a) and (14b) are standard "birth-death" equations.[12] The Laplace transform, $Q_M(s, \tilde{d})$, can be computed directly from (14) and is given by

$$Q_M(s, d_M) = \frac{r_2^{d_M}(s)}{(c - v_0)\mu_d[r_1(s) - 1]}, \tag{17}$$

where

$$r_1(s) = \frac{1}{2(c - v_0)\mu_d}\{s + \lambda_d + (c - v_0)\mu_d + \sqrt{[s + \lambda_d + (c - v_0)\mu_d]^2 - 4(c - v_0)\mu_d\lambda_d}\} \tag{18a}$$

and

$$r_2(s) = \frac{1}{2(c - v_0)\mu_d}\{s + \lambda_d + (c - v_0)\mu_d - \sqrt{[s + \lambda_d + (c - v_0)\mu_d]^2 - 4(c - v_0)\mu_d\lambda_d}\}. \tag{18b}$$

We therefore have

$$q_M(d_M) = q_M(0)r_2^{d_M}(v_0\mu_v), \tag{19}$$

where

$$q_M(0) = \frac{v_0\mu_v}{(c - v_0)\mu_d[r_1(v_0\mu_v) - 1]}. \tag{20}$$

Since

$$\sum_{d_M=0}^{\infty} q_M(d_M) = 1, \tag{21}$$

it follows that

$$q_M(0) = 1 - r_2(v_0\mu_v). \tag{22}$$

The distribution of the maximum of the queue spike is therefore geometric with parameter $r_2(v_0\mu_v)$. Consequently,

$$E(d_M) = \frac{r_2(v_0\mu_v)}{1 - r_2(v_0\mu_v)} \tag{23}$$

and

$$E[d_M - E(d_M)]^2 = \frac{r_2(v_0\mu_v)}{[1 - r_2(v_0\mu_v)]^2}. \tag{24}$$

Notice that as μ_v decreases relative to μ_d, $E(d_M)$ increases.

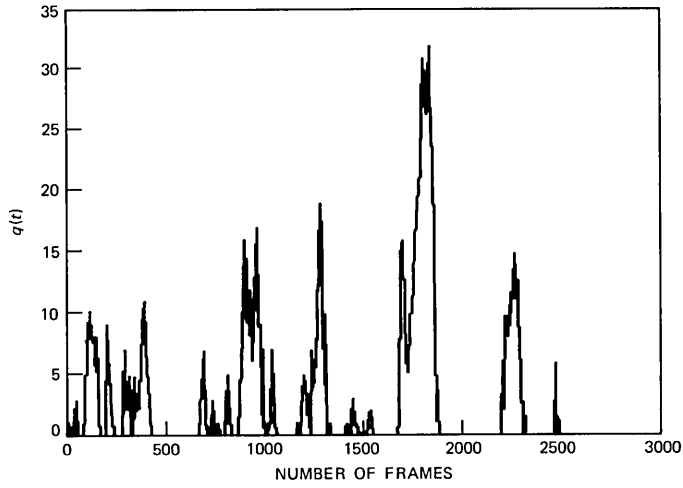


Fig. 3—Sample path for queue process. Every tenth sample is shown.

A sample path of the number of queued data requests versus time is shown in Fig. 3. (Every 10th sample is plotted.) Queues periodically build and empty, which suggests that as an approximation, the mean value of the peaks of these buildups is given by (23). This approximation becomes more accurate as μ_v decreases relative to μ_d. Note, however, that (23) gives the mean value of the peaks of the queue process with initial conditions $v = v_0$ and $\tilde{d} = 0$. It is possible that the mean value of the peaks of the queue process, assuming $v < v_0$, is larger than that predicted by (23). [In some cases, the conditional mean queue length given by (8) for $v = v_0 - 1$ is in fact larger than $E(d_M)$.] A better approximation to the mean value of the peaks of the queue process can be obtained by computing the conditional means assuming $v = 0, 1, \cdots, v_0$, and then using the distribution $p(v)$ given by (5) to form a weighted average. As the present analysis is concerned with evaluating the performance degradation caused by transient instabilites, this computation was not performed.

We now compute the mean duration of the queue spike. Denoting this quantity as $\bar{\tau}_w$, it is apparent that

$$\bar{\tau}_w = \bar{\tau}_{w,1} + \bar{\tau}_{w,2}, \tag{25}$$

where $\bar{\tau}_{w,1}$ is the mean time it takes the queue to reach its maximum value given $v = v_0$, and $\bar{\tau}_{w,2}$ is the mean time it takes the queue to empty. From the previous discussion,

$$\bar{\tau}_{w,1} = \frac{1}{v_0 \mu_v}. \tag{26}$$

Let $\bar{\tau}_{\tilde{d},v}$ denote the mean time it takes to reach state $\tilde{d} = 0$ (i.e., no data message queue) given an initial state of $\tilde{d}$ queued data messages and v voice-occupied time slots. We have the following transition equation,

$$\bar{\tau}_{\tilde{d},v} = \frac{1}{\sigma(v)} + p_v\bar{\tau}_{\tilde{d}-1,v} + q_v\bar{\tau}_{\tilde{d}+1,v} + w_v\bar{\tau}_{\tilde{d},v-1} \tag{27a}$$

with initial condition

$$\bar{\tau}_{0,v} = 0, \tag{27b}$$

where

$$\sigma(v) \equiv v\mu_v + (c - v)\mu_d + \lambda_d \tag{28}$$

and $1/[\sigma(v)]$ is the mean amount of time spent in state $(\tilde{d}, v)$ before a state transition occurs, and

$$p_v = \frac{(c - v)\mu_d}{\sigma(v)}, \tag{29a}$$

$$q_v = \frac{\lambda_d}{\sigma(v)}, \tag{29b}$$

and

$$w_v = \frac{v\mu_v}{\sigma(v)} \tag{29c}$$

are, respectively, the probabilities of going to states $(\tilde{d} - 1, v)$, $(\tilde{d} + 1, v)$, and $(\tilde{d}, v - 1)$ from state $(\tilde{d}, v)$. Equation (27) is a two-dimensional difference equation that is nonlinear in v. Notice, however, that we desire

$$\begin{aligned}\bar{\tau}_{w,2} &= \sum_{\tilde{d}=0}^{\infty} q_M(\tilde{d})\bar{\tau}_{\tilde{d},v_0-1} \\ &= [1 - r_2(v_0\mu_v)] \sum_{\tilde{d}=0}^{\infty} r_2^{\tilde{d}}(v_0\mu_v)\bar{\tau}_{\tilde{d},v_0-1} \\ &= [1 - r_2(v_0\mu_v)]D\left[\frac{1}{r_2(v_0\mu_v)}, v_0 - 1\right],\end{aligned} \tag{30}$$

where

$$D(z, v) = \sum_{\tilde{d}=0}^{\infty} z^{-\tilde{d}}\bar{\tau}_{\tilde{d},v} \tag{31}$$

is the partial z-transform of $\bar{\tau}_{\tilde{d},v}$. An iterative method for computing $D(1/[r_2(v_0\mu_v)], v_0 - 1)$ is discussed in Appendix B.

To compute the mean time between unstable periods, we first define $\bar{T}_v$ as the expected value of the first passage time it takes to go from

an initial voice state $v < v_0$ to voice state v_0. The mean time between unstable periods is then

$$\bar{T} = \sum_{v=0}^{v_0-1} p^*(v)\bar{T}_v, \tag{32}$$

where $p^*(v)$ is the probability that the voice state is v at the end of a queue spike (i.e., when $\tilde{d}$ returns to zero). The computation of $p^*(v)$ is similar to the computation of $\bar{\tau}_w$. In particular, let $p(v \mid v_1, \tilde{d})$ denote the probability of v voice-occupied time slots at the end of an unstable period given an initial state $(v_1, \tilde{d})$. The following transition equation is easily obtained,

$$p(v \mid v_1, \tilde{d}) = q_v p(v \mid v_1, \tilde{d} + 1) + p_v p(v \mid v_1, \tilde{d} - 1) + w_v p(v \mid v_1 - 1, \tilde{d}), \tag{33a}$$

where q_v, p_v, and w_v are defined by (29). The initial conditions are

$$p(v \mid v - 1, \tilde{d}) = 0 \tag{33b}$$

and

$$p(v \mid v_1, 0) = \begin{cases} 1 & \text{if } v = v_1 \\ 0 & \text{otherwise.} \end{cases} \tag{33c}$$

In analogy with (27), (33) is a two-dimensional difference equation that is nonlinear in v. As before, we desire

$$\begin{aligned} p^*(v) &= [1 - r_2(v_0)] \sum_{\tilde{d}=0}^{\infty} r_2^{\tilde{d}}(v_0\mu_v) p(v \mid v_0 - 1, \tilde{d}) \\ &= [1 - r_2(v_0)] D\left[\frac{1}{r_2(v_0\mu_v)}, v \mid v_0 - 1\right], \end{aligned} \tag{34}$$

where

$$D(z, v \mid v_1) = \sum_{\tilde{d}=0}^{\infty} z^{-\tilde{d}} p(v \mid v_1, \tilde{d}). \tag{35}$$

An iterative method for computing $D(z, v \mid v_1)$ is discussed in Appendix B.

The mean time between unstable periods, given the initial starting state, can be approximated by again assuming voice traffic service times are very long relative to data traffic service times. For each voice state, we assume that the data traffic exhibits steady-state behavior. This leads to the following difference equation,

$$\bar{T}_v = t_v + r_v \bar{T}_{v-1} + s_v \bar{T}_{v+1}, \tag{36}$$

where

$$t_v = \frac{1}{\lambda_v \phi_v + v\mu_v} \tag{37a}$$

is the mean time spent in voice state v before going to state $v - 1$ or $v + 1$, ϕ_v is the "entrance" probability for an incoming voice message and is equal to the probability that a data message queue exists, and

$$r_v = \frac{v\mu_v}{\lambda_v \phi_v + v\mu_v} \tag{37b}$$

and

$$s_v = \frac{\lambda_v \phi_v}{\lambda_v \phi_v + v\mu_v} \tag{37c}$$

are, respectively, the probabilities of going from voice state v to state $v - 1$ and from voice state v to state $v + 1$. In Appendix C we show

$$\bar{T}_v = \frac{1}{\lambda_v} \sum_{j=v}^{v_0-1} \left\{ \left(\frac{\mu_v}{\lambda_v}\right)^j \frac{j!}{\gamma_j} \left[1 + \sum_{m=0}^{j-1} \left(\frac{\mu_v}{\lambda_v}\right)^{m-j} \frac{\gamma^{j-m-1}}{(j-m)!} \right] \right\}, \tag{38}$$

where

$$\gamma_j = \prod_{m=0}^{j} \phi_m. \tag{39}$$

Therefore, computation of $\bar{T}_v$ and $p^*(v)$ by way of (38) and the method given in Appendix B, respectively, yields the mean time between unstable periods. The relative frequency, or probability, that the system is in an unstable state ($v = v_0$) is approximated by

$$p_u = \frac{\bar{\tau}_w}{\bar{\tau}_w + \bar{T}}, \tag{40}$$

where $\bar{\tau}_w$ and $\bar{T}$ are given by (25) and (32), respectively. Notice that this expression should be approximately equal to the value of $p(v_0)$ obtained using (5).

This completes the presentation of analytical results. These results are used in the next section to evaluate the performance of an integrated voice-data TDMA network, and to demonstrate the improvement over analogous TDMA networks that handle only one traffic type.

IV. PERFORMANCE RESULTS

The objective of this section is to demonstrate how the integrated voice-data network described in Sections I and II performs as a function of (1) input traffic intensity, (2) traffic blend (ratio of voice traffic intensity to total traffic intensity), and (3) system size, as

measured by the number of time slots. In all cases, half-duplex voice traffic is assumed. We expect the full-duplex case, where each voice message requests two time slots, to exhibit similar behavior. Denoting the voice traffic intensity as $\rho_v = \lambda_v/(c\mu_v)$, and the data traffic intensity as $\rho_d = \lambda_d/(c\mu_d)$, where c is the number of time slots, the total normalized offered load is defined as

$$\rho = \rho_v + \rho_d, \tag{41}$$

and the traffic blend is

$$r = \frac{\rho_v}{\rho_v + \rho_d}. \tag{42}$$

Initial traffic parameters are selected to produce a preselected value of r, and the input traffic intensity ρ is varied between zero and one by multiplying both λ_v and λ_d by a constant. The service rates used are $\mu_v = 0.001$ and $\mu_d = 0.025$, which corresponds to a mean voice message length of 1000 time slots and a mean data message length of 40 time slots. In practice, the mean voice message is much greater than 1000 time slots; however, the analytical results in the last section become more accurate as μ_v decreases relative to μ_d.

Figures 4a, 4b, and 4c show voice message blocking probabilities computed by means of (4) versus the offered load for systems with 20, 100, and 500 time slots, respectively. In each case, curves are shown for three different traffic blends. A few randomly selected points from Figs. 4a, 4b, and 4c were compared with results obtained by a computer simulation of the queueing model under consideration. In each case the approximate result and the computer-simulated result were nearly identical. Also shown are plots of blocking probabilities versus load produced by systems that handle only (half-duplex) voice traffic and which have rc time slots, where c is the number of time slots in the voice-data network. These curves are computed directly from (5), where $\phi_j = 1$ for $j < rc$.

At high traffic intensities, blocking probabilities for the integrated systems are consistently higher than those for the analogous single traffic systems. In this region, data queues often form, causing data messages to grab time slots relinquished by voice messages. In contrast, at lower traffic intensities, data message queues are less likely, so that voice messages often have access to additional time slots. At low traffic intensities, integrated systems therefore always exhibit superior performance when compared with analogous voice-only systems. In each comparison, there appears to be a unique traffic intensity, ρ^*, where both systems give the same blocking probability. Notice that as the traffic blend r decreases, ρ^* increases. This is due to the fact that the probability of a data message queue existing at a fixed, normalized traffic intensity decreases as the number of time slots allocated for

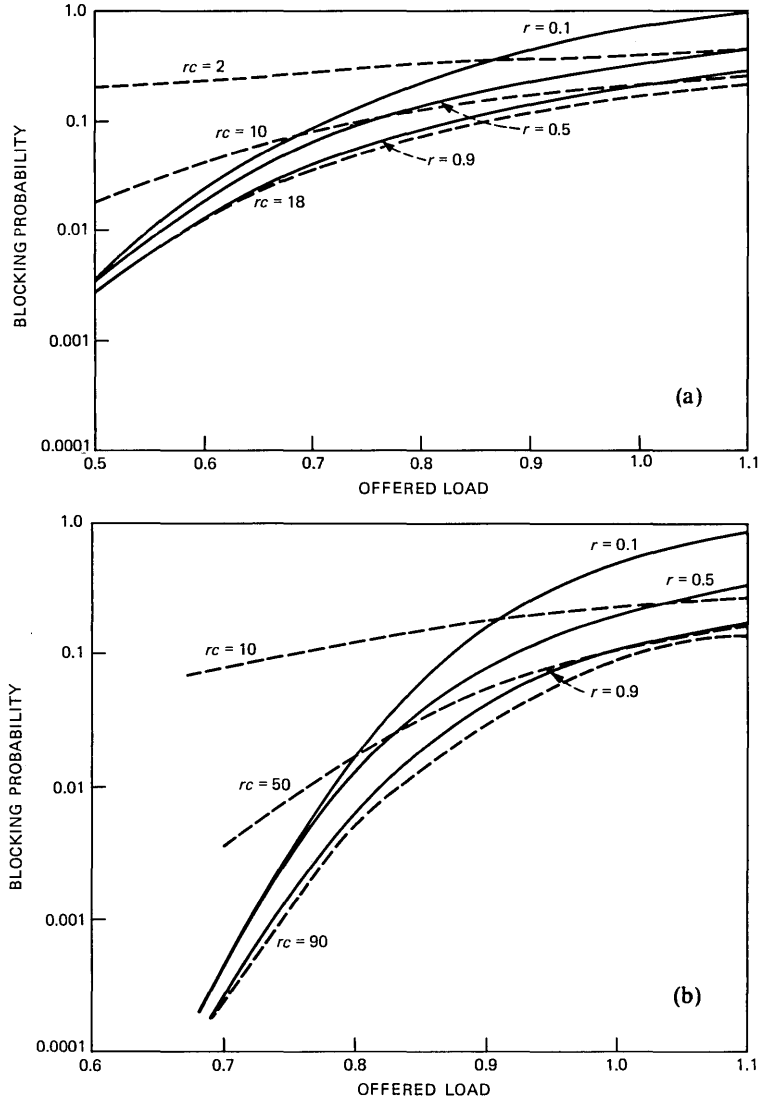


Fig. 4—Voice message blocking probability versus offered load for integrated systems with: (a) 20 slots and single traffic systems with 2, 10, and 18 slots; (b) 100 slots and single traffic systems with 10, 50, and 90 slots.

data traffic increases. As r decreases, voice messages therefore often have access to additional time slots not used by data traffic. At a fixed traffic intensity, as r decreases, the blocking probability produced by the integrated system should therefore decrease, relative to the blocking probability produced by the analogous voice-only system. A final observation is that as the traffic intensity decreases, blocking proba-

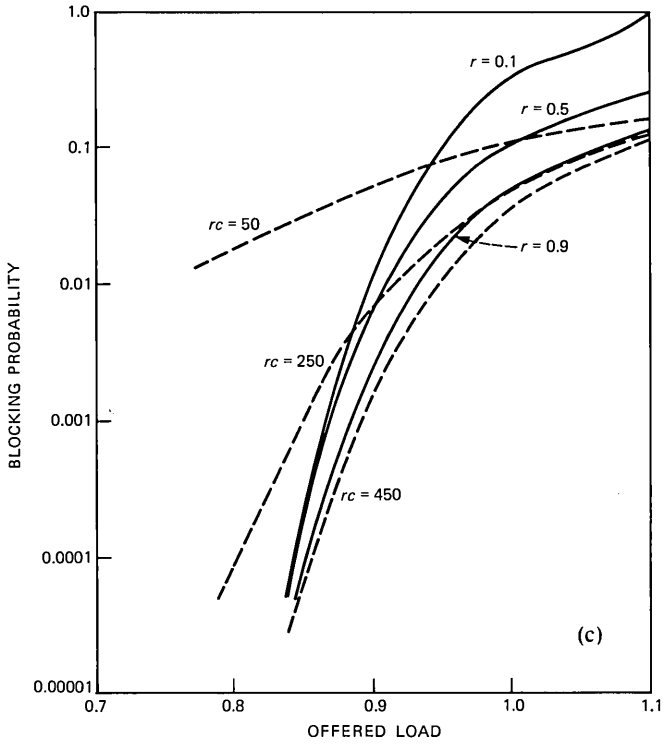


Fig. 4(c)—Voice message blocking probability versus offered load for integrated systems with 500 slots and single traffic systems with 50, 250, and 450 slots.

bilities obtained using the integrated systems become insensitive to the traffic blend. This is in contrast to the analogous voice-only systems, which result in a much wider variation in blocking probabilities as the number of time slots is varied.

Figures 5a, 5b, and 5c show plots of mean data message queue length (number of queued data requests), computed by means of (7), (8), and (10), versus normalized offered load for systems with 20, 100, and 500 time slots, respectively. Curves are again shown for three different traffic blends. Also plotted is the mean number of queued data requests produced by a system handling data traffic only with c time slots. Close agreement was again found between randomly selected points from these figures and computer simulation results. At high traffic intensities, the variation between curves is caused by the different traffic loads, at which the mean data queue length approaches infinity. In particular, the single traffic curve has its asymptote at $\rho = \rho_d = 1$. In contrast, because queued data messages can grab relinquished voice-occupied time slots, the integrated traffic curves have asymptotes at $\rho_d = 1$, which corresponds to $\rho = 1.1$, $\rho = 2$, and $\rho = 10$ for $r = 0.1$,

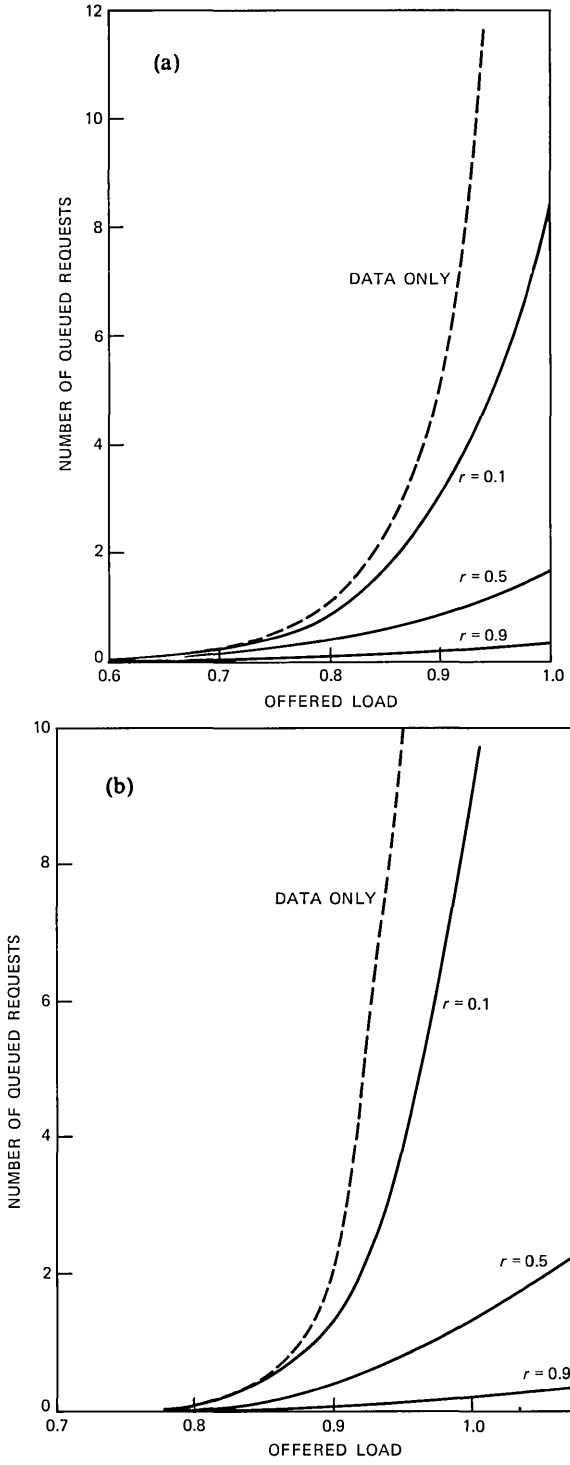


Fig. 5—Mean data message queue length versus offered load for integrated systems and a single traffic system with: (a) 20 slots, and (b) 100 slots.

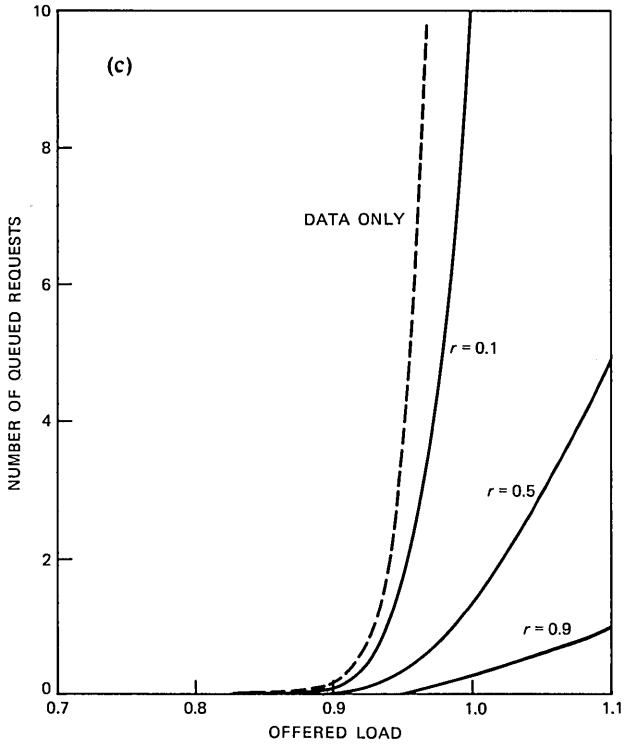


Fig. 5(c)—Mean data message queue length versus offered load for integrated systems and a single traffic system with 500 slots.

$r = 0.5$, and $r = 0.9$, respectively. Mean queue length in these high traffic intensity regions is not of interest, however, since the corresponding voice blocking probability is near unity.

Figures 6a, 6b, and 6c show plots of $E(d_M)$ given by (23) versus offered load, which indicates the mean value of the maximum of queue buildups. The discontinuities in Fig. 6a are caused by discontinuous changes in the voice instability state v_0 as a function of traffic load. As an example, for the case $c = 20$ and $r = 0.5$ shown in Fig. 6a, v_0 changes from 14 to 13 as the traffic intensity ρ increases from $0.7 - \epsilon$ to 0.7, where ϵ is small. Discontinuities were observed in all curves shown in Fig. 6; however, in most cases these discontinuities were hardly noticeable. In particular, as the number of slots c increases, $E(d_m)$ becomes less sensitive to changes in v_0. A comparison of the results in Figs. 6a, 6b, and 6c with simulated sample paths of the data-message queue process indicates that the results presented here are typically about 10 to 25 percent smaller than the actual peaks observed, indicating that the peaks that occur when $v < v_0$ are often greater than those that occur when $v = v_0$. As μ_v decreases relative to μ_d,

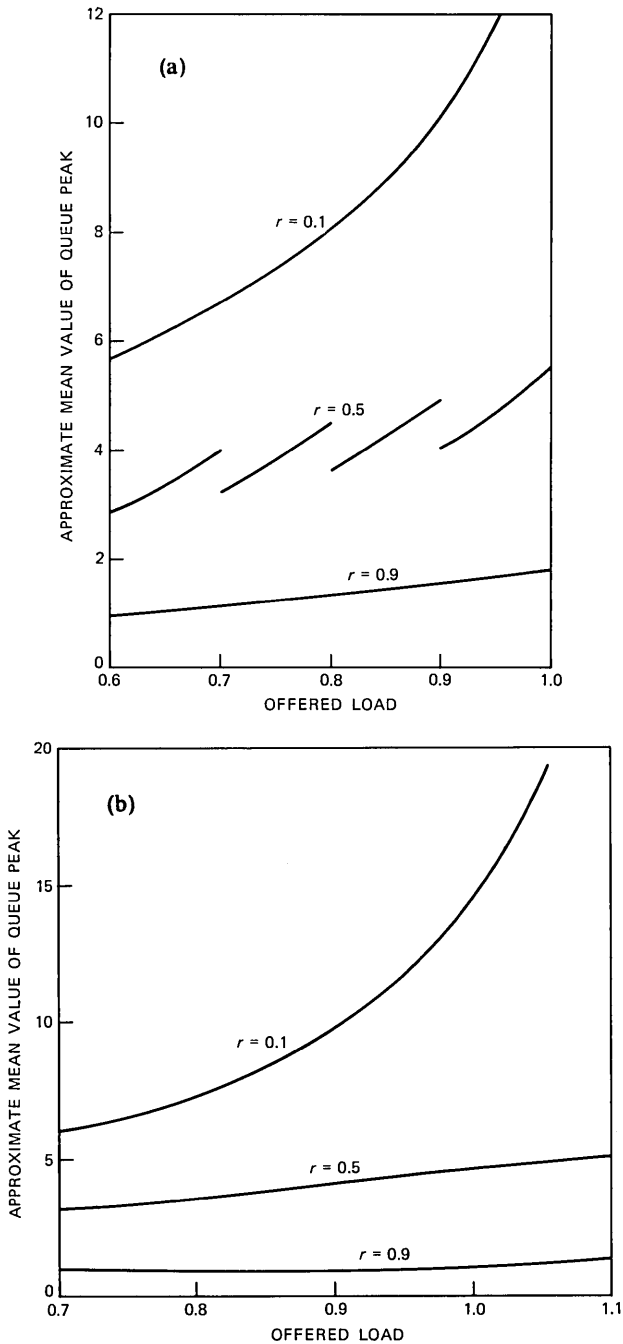


Fig. 6—Approximate mean value of queue peaks versus offered load from eq. (23) for integrated systems with: (a) 20 slots and (b) 100 slots.

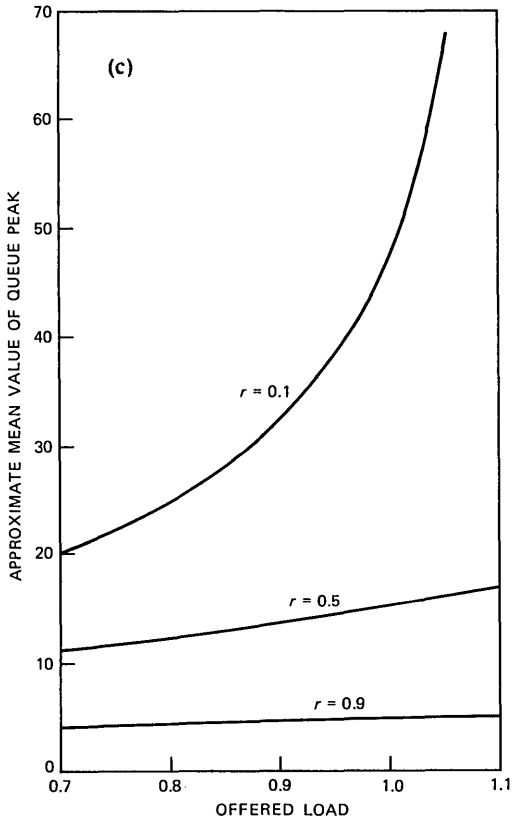


Fig. 6(c)—Approximate mean value of queue peaks versus offered load from eq. (23) for integrated systems with 500 slots.

however, $E(d_M)$ given by (23) should give a more accurate indication of the mean value of these peaks.

For each point computed in Figs. 4 through 6, the probability of being in an "unstable" state $p(v_0)$, the mean duration of the unstable state ($\bar{\tau}_w$), and the mean time between unstable states ($\bar{T}$) were calculated by means of the results in Section 3.1. A few representative points are listed in Table I. The probability of being in an unstable state and the resulting queues that form are typically too small to cause significant degradation in system performance.

V. CONCLUSIONS

Results obtained from the analysis of an integrated voice-data TDMA protocol indicate that voice message blocking probability in the integrated system is insensitive to the blend of traffic at low input traffic intensities. As the traffic intensity increases, the blocking

Table I—Mean duration of unstable state ($\bar{\tau}_w$ in number of frames), mean time between unstable states ($\bar{T}$ in number of frames), and the probability of being in an unstable state [$p(v_0)$] for a few representative cases (assuming $r = 0.5$)

Time Slots	Load (ρ)	$\bar{\tau}_w$	$\bar{T}$	$p(v_0)$
20	0.6	129.2	$2.05 \cdot 10^5$	$6.3 \cdot 10^{-4}$
20	1.0	191.8	2898	0.062
100	0.7	44.8	$1.85 \cdot 10^9$	$2.4 \cdot 10^{-8}$
100	1.0	60.8	$4.15 \cdot 10^4$	0.0015
500	0.85	$6.66 \cdot 10^6$	$4.12 \cdot 10^{19}$	$1.6 \cdot 10^{-13}$
500	1.0	$2.88 \cdot 10^7$	$5.87 \cdot 10^{13}$	$4.9 \cdot 10^{-7}$

probability of the integrated system increases, relative to the blocking probability of the analogous voice-only system. The traffic intensity at which the two blocking probability curves intersect is a function of the traffic blend. Mean queue length in the integrated system displays a wide variation with traffic blend at high traffic intensities, due to the variation in traffic intensity at which instability occurs. At offered loads of 0.7 to 0.8, very good performance can be achieved (i.e., a blocking probability <0.01 and a mean queue length near zero) with moderately sized systems (~100 slots/frame). Finally, the data message queues that form during the unstable transients are moderate for the cases examined, and the frequency at which these transients occur is in most cases quite small. As the number of time slots per TDMA frame increases, results presented here show significant improvements in system performance. This is an important observation since most practical networks are much larger than those considered here (i.e., greater than 1000 time slots per frame).

VI. ACKNOWLEDGMENT

The author thanks S. M. Barta and B. E. Simon for many helpful conversations concerning this work.

REFERENCES

1. M. J. Fischer and T. C. Harris, "A Model for Evaluating the Performance of an Integrated Circuit- and Packet-Switched Multiplex Structure," IEEE Trans. Commun., *COM-24* (February 1976), pp. 195–202.
2. C. J. Weinstein, M. L. Malpass, and M. J. Fisher, "Data Traffic Performance of an Integrated Circuit- and Packet-Switched Multiplex Structure," IEEE Trans. Commun., *COM-28* (June 1980), pp. 873–8.
3. M. Schwartz and B. Kraimeche, "Comparison of Channel Assignment Techniques for Hybrid Switching," Proc. 1982 IEEE ICC, Philadelphia, PA, June 1982, pp. 2F.3.1–5.
4. D. P. Gaver and J. P. Lehoczky, "Channels That Cooperatively Service a Data Stream and Voice Messages," IEEE Trans. Commun., *COM-30* (May 1982), pp. 1153–62.
5. M. J. Ross and O. A. Mowafi, "Performance Analysis of Hybrid Switching Concepts for Integrated Voice/Data Communications," IEEE Trans. Commun., *COM-30* (May 1982), pp. 1073–87.

6. N. Janakiraman, B. Pagurek, and J. E. Neilson, "Performance Analysis of an Integrated Switch with Fixed or Variable Frame Rate and Movable Voice/Data Boundary," IEEE Trans. Commun., *COM-32,* No. 1 (January 1984), pp. 34–9.
7. A. G. Konheim and R. L. Pickholtz, "Analysis of Integrated Voice/Data Multiplexing," IEEE Trans. Commun., *COM-32,* No. 2 (February 1984), pp. 140–7.
8. R. Cooperman and W. G. Schmidt, "A Satellite Switched SDMA/TDMA System for Wideband Multibeam Satellites," ICC Conf. Rec., Seattle, Washington, June 1973.
9. D. O. Reudink and Y. S. Yeh, "A Scanning Spot Beam Satellite System," B.S.T.J., *56,* No. 8 (October 1977), pp. 1549–60.
10. S. M. Barta and M. L. Honig, "Analysis of a Demand Assignment TDMA Blocking System," AT&T Bell Lab. Tech. J., *63,* No. 1 (January 1984), pp. 89–114.
11. B. E. Simon, unpublished work.
12. T. L. Saaty, *Elements of Queuing Theory with Application,* New York: McGraw-Hill, 1961.

APPENDIX A

We wish to show (10) using the approximation (9). In addition, we assume that $E[q(t)]$ is piecewise linear. (This would be true, for instance, if the data message queue length $q(t)$ were allowed to assume negative values.) At time $t = 0$ we assume that the state of the system is $(v_0, c - v_0)$. Let $t_v(i)$ denote the mean time it takes $i + 1$ voice messages to relinquish their time slots. Then

$$t_v(i) = \sum_{j=0}^{i} \frac{1}{(v_0 - j)\mu_v}, \tag{43}$$

and

$$E[q(t)] \approx \begin{cases} q_M v_0 \mu_v t & 0 < t < t_v(0) \\ E\{q[t_v(i)]\} - [(c - v_0 + i + 1)\mu_d - \lambda_d][t - t_v(i)], & t_v(i) < t < t_v(i+1) \end{cases}, \tag{44}$$

where q_M is the peak value of $E[q(t)]$ and is given by

$$q_M = \frac{\lambda_d - (c - v_0)\mu_d}{v_0 \mu_v}. \tag{45}$$

Notice that for $t > t_v(0)$,

$$\begin{aligned} E\{q[t_v(i)]\} \equiv \bar{q}_i &= \bar{q}_{i-1} \\ &\quad - [(c - v_0 + i)\mu_d - \lambda_d][t_v(i) - t_v(i-1)] \\ &= q_M - \sum_{j=1}^{i} [(c - v_0 + j)\mu_d - \lambda_d] \frac{1}{(v_0 - j)\mu_v}. \end{aligned} \tag{46}$$

To calculate the area under $E[q(t)]$, we must first compute

$$t_0 \equiv \inf\{t \mid E[q(t)] = 0 \quad \text{and} \quad t > 0\}. \tag{47}$$

Letting

$$i_0 = \max\{i \mid \bar{q}_i > 0\}, \tag{48}$$

it follows that

$$t_v(i_0) < t_0 < t_v(i_0 + 1). \tag{49}$$

For $t_v(i_0) \le t \le t_0$,

$$E[q(t)] = \bar{q}_{i_0} - [(c - v_0 + i_0 + 1)\mu_d - \lambda_d][t - t_v(i_0)], \tag{50}$$

and setting $\bar{q}_{i_0} = 0$ gives

$$t_0 = \frac{\bar{q}_{i_0}}{(c - v_0 + i_0 + 1)\mu_d - \lambda_d} + t_v(i_0). \tag{51}$$

A plot of $E[q(t)]$, $0 \le t \le t_0$, assuming three voice departures ($i_0 = 2$) is shown in Fig. 7. If we use (9), it follows that

$$E(q \mid v_0) \approx \frac{1}{t_0} \sum_{j=0}^{i_0+1} A_j, \tag{52}$$

where A_j is the area of region R_j. It is apparent that

$$A_0 = \frac{1}{2} \frac{q_M}{t_v(0)} = \frac{1}{2} \frac{\lambda_d - (c - v_0)\mu_d}{(v_0\mu_v)^2} \tag{53}$$

and that

$$\begin{aligned} A_{i_0+1} &= \frac{1}{2}[t_0 - t_v(i_0)]\bar{q}_{i_0} \\ &= \frac{\bar{q}_{i_0}^2}{2[(c - v_0 + i_0 + 1)\mu_d - \lambda_d]}. \end{aligned} \tag{54}$$

Finally, regions $R_1, \cdots, R_{i_0}$ are trapezoids with upper-boundary $E[q(t)]$, so that for $1 \le j \le i_0$,

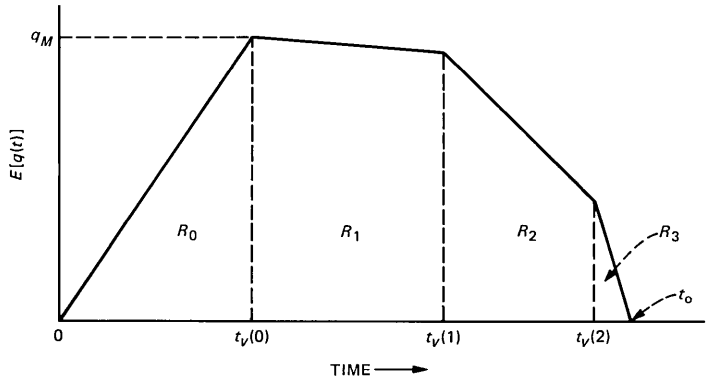


Fig. 7—Mean queue length versus time during transient instability.

$$A_j = \frac{1}{2}[t_v(j) - t_v(j-1)](\bar{q}_{j-1} + \bar{q}_j)$$
$$= \frac{1}{2(v_0 - j)\mu_v}(\bar{q}_{j-1} + \bar{q}_j). \quad (55)$$

Substituting (51) and (53) through (55) into (52), and using (43) and (46) gives (10).

APPENDIX B

We are interested in computing $D\{1/[r_2(v_0\mu_v)], v_0 - 1\}$ and $D\{1/[r_2(v_0\mu_v)], v \mid v_1\}$, where $D(z, v)$ and $D(z, v \mid v_1)$ are given, respectively, by (31) and (35) and $r_2(v_0\mu_v)$ is given by (18b). Multiplying both sides of (27a) by $z^{-\tilde{d}}$ and summing from $\tilde{d} = 1$ to infinity gives, after algebraic manipulation,

$$D(z, v) = \frac{(z-1)[v\mu_v D(z, v-1) - \lambda_d \bar{\tau}_{1,v}] + 1}{z(1-z)s(v, z^{-1})}, \quad (56)$$

where

$$s(v, z) = \mu_d(c - v)z^2 - [(c - v)\mu_d + \lambda_d + v\mu_v]z + \lambda_d, \quad (57)$$

which has real roots

$$r_k(v) = \frac{1}{2(c-v)\mu_d}\{\sigma(v) \pm \sqrt{\sigma^2(v) - 4(c-v)\mu_d\lambda_d}\}, \quad (58)$$

where $k = 1$ corresponds to "+", $k = 2$ corresponds to "−", and $\sigma(v)$ is given by (28). Notice that the roots $r_k(v_0)$ given by (58) are identical to the roots $r_k(v_0\mu_v)$ given by (18). For convenience, we therefore refer to the roots $r_k(v_0\mu_v)$ as $r_k(v_0)$, for $k = 1, 2$. From Rouche's Theorem,[12] it follows that $r_2(v) < 1$ and $r_1(v) \geq 1$. The z-transform, $D(z, v)$, must be analytic outside the unit circle, and hence $\bar{\tau}_{1,v}$ in (56) must be selected to cancel the pole at $1/[r_2(v)]$. In particular, $s(v, z^{-1})$ has roots $1/[r_1(v)] < 1$ and $1/[r_2(v)] > 1$, so that

$$\bar{\tau}_{1,v} = \frac{v\mu_v}{\lambda_d} D\left[\frac{1}{r_2(v)}, v - 1\right] + \frac{r_2(v)}{[1 - r_2(v)]\lambda_d}. \quad (59)$$

As an example, suppose that we assume the data message queue empties with probability one before $k + 1$ voice messages relinquish their time slots. Once the voice state becomes $v = v_0 - k - 1$, the voice service rate $\mu_v = 0$. Substituting these values for μ_v and v in (56) gives

$$D(z, v_0 - k - 1)$$
$$= \frac{z[1 - \lambda_d(z-1)\bar{\tau}_{1,v_0-k-1}]}{(z-1)[1 - r(v - k - 1)z][1 - r(v - k - 1)z]}, \quad (60)$$

where from (58),

$$r_1(v_0 - k - 1) = 1 \quad \text{and} \quad r_2(v_0 - k - 1) = \frac{\lambda_d}{(c - v_0 + k + 1)\mu_d}. \tag{61}$$

Selecting $\tilde{\tau}_{1,v_0-k-1}$ to cancel the pole at $1/[r_2(v_0 - k - 1)]$ and simplifying gives

$$D(z, v_0 - k - 1) = \frac{z}{(z-1)^2} \frac{1}{\mu_d(c - v_0 + k + 1) - \lambda_d}, \tag{62}$$

which has inverse transform

$$\tilde{\tau}_{\tilde{d},v_0-k-1} = \frac{\tilde{d}}{\mu_d(c - v_0 + k + 1) - \lambda_d}. \tag{63}$$

Equation (62) constitutes an initial condition for (56), which can be iterated numerically using (59). In particular, assuming no more than k voice messages can relinquish their time slots after the queue begins to empty, $D(z, v_0 - k - 1)$ for $z = 1/[r_2(v_0)]$ and $z = 1/[r_2(v_0 - k)]$ is calculated from (62). The value of $\tilde{\tau}_{1,v_0-k}$ is subsequently computed from (59) and is used in (56) to compute $D(z, v_0 - k)$ for $z = 1/[r_2(v_0)]$ and $z = 1/[r_2(v_0 - k + 1)]$. Equation (59) is subsequently used to compute $\tilde{\tau}_{1,v_0-k+1}$, which is used to compute $D(z, v_0 - k + 1)$, and so forth until $D\{1/[r_2(v_0)], v_0 - 1\}$ is computed.

To compute $D(z, v \mid v_1)$ given by (35) at $z = 1/[r_2(v_0)]$, we multiply both sides of (33a) by $z^{-\tilde{d}}$ and sum from $\tilde{d} = 1$ to infinity to get

$$D(z, v \mid v_1 + 1) = \frac{\begin{array}{r} -v\mu_v D(z, v \mid v_1) + \lambda_d p(v \mid v_1 + 1, 1) \\ + v\mu_v \delta_{v,v_1} - [\sigma(v) - \lambda_d z]\delta_{v,v_1+1} \end{array}}{zs(v, z^{-1})}, \tag{64}$$

where δ_{ij} is the Kronecker delta. Using the condition (33b) gives the boundary condition

$$D(z, v_1 \mid v_1 - 1) = 0. \tag{65}$$

Because $D(z, v \mid v_1)$ must be analytic outside the unit circle, $p(v \mid v_1, 1)$ is selected to cancel the pole at $z = 1/[r_2(v_0)]$. This implies that

$$p(v \mid v_1, 1) = \frac{v\mu_v}{\lambda_d} \left\{ D\left[\frac{1}{r_2(v)}, v \mid v_1 - 1\right] - \delta_{v,v_1-1} \right\} + \frac{\sigma(v)r_2(v) - \lambda_d}{\lambda_d r_2(v)} \delta_{v,v_1}. \tag{66}$$

To compute $D(z, v \mid v_0 - 1)$ at $z = 1/[r_2(v_0)]$ for $v = v_0 - 1, v_0 - 2, \cdots, v_0 - k - 1$, where k is the maximum number of voice departures allowed, the boundary condition (65) is first used in (66) to get

$$p(v_0 - j \mid v_0 - j, 1) = \frac{\sigma(v_0 - j)r_2(v_0 - j) - \lambda_d}{\lambda_d r_2(v_0 - j)}, \tag{67}$$

which is used in (64) to compute

$$D(z, v_0 - j \mid v_0 - j) = \frac{\lambda_d p(v_0 - j \mid v_0 - j, 1) - [\sigma(v_0 - j) - \lambda_d z]}{zs(v_0 - j, z^{-1})},$$

where j is initially one and ranges from one to $k + 1$. This expression is evaluated at $z = 1/[r_2(v_0 - j)]$ and substituted into (66) to obtain $p(v_0 - j \mid v_0 - j + 1, 1)$, which is used in (64) to compute $D(z, v_0 - j \mid v_0 - j + 1)$ at the appropriate values of z. This procedure continues until the value of $D\{1/[r_2(v_0)], v_0 - j \mid v_0 - 1\}$ is obtained, whereupon j is incremented and the procedure starts over again. In this way the values $D\{1/[r_2(v_0)], v_0 - j \mid v_0 - 1\}$ for $j = 1, 2, \cdots, k + 1$ are generated systematically.

APPENDIX C

We wish to show that (38) is the solution to (36). We first rewrite (36) as

$$\bar{T}_{v+1} = \left(1 + \frac{v\mu_v}{\lambda_v \phi_v}\right) \bar{T}_v - \frac{v\mu_v}{\lambda_v \phi_v} \bar{T}_{v-1} - \frac{1}{\lambda_v \phi_v}. \tag{68}$$

Letting

$$y_v = \bar{T}_v - \bar{T}_{v-1}, \tag{69}$$

(68) can be rewritten as

$$y_{v+1} = \frac{v\mu_v}{\lambda_v \phi_v} y_v - \frac{1}{\lambda_v \phi_v}, \quad \text{for} \quad 0 \le v \le v_0, \tag{70}$$

which can be iterated to give

$$y_{v+1} = \left(\frac{\mu_v}{\lambda_v}\right)^{k+1} \frac{v!\gamma_{v-k-1}}{(v - k - 1)!\gamma_v} y_{v-k} - \frac{1}{\lambda_v} \sum_{m=0}^{k} \left(\frac{\mu_v}{\lambda_v}\right)^m \frac{v!\gamma_{v-m}}{(v - m)!\gamma_v}, \tag{71}$$

where γ_v is given by (39). Using the initial condition,

$$y_1 = \bar{T}_1 - \bar{T}_0 = -\frac{1}{\lambda_v \phi_0}, \tag{72}$$

and substituting $k = v - 1$ in (71) gives

$$y_{v+1} = -\frac{1}{\lambda_v} \left(\frac{\mu_v}{\lambda_v}\right)^v \frac{v!}{\gamma_v} \left[1 + \sum_{m=0}^{v-1} \left(\frac{\mu_v}{\lambda_v}\right)^{m-v} \frac{\gamma_{v-m-1}}{(v - m)!}\right]. \tag{73}$$

From (69),

$$\bar{T}_v = \sum_{j=1}^{v} y_j + \bar{T}_0, \tag{74}$$

which has boundary condition

$$\bar{T}_{v_0} = \sum_{j=1}^{v_0} y_j + \bar{T}_0 = 0. \tag{75}$$

This implies that

$$\bar{T}_0 = -\sum_{j=1}^{v_0} y_j, \tag{76}$$

so that

$$\bar{T}_v = -\sum_{j=v+1}^{v_0} y_j. \tag{77}$$

Combining (73) and (77) gives (38).

AUTHOR

Michael L. Honig, B.S. (Electrical Engineering), 1977, Stanford University; M.S. and Ph.D. (Electrical Engineering), 1978 and 1981, respectively, University of California, Berkeley; Bell Laboratories, 1981–1982; AT&T Information Systems, 1983. Present affiliation Bell Communications Research, Inc. At AT&T Bell Laboratories and AT&T Information Systems Mr. Honig worked on modulation, coding, and echo cancellation of voiceband data signals, performance analysis of local area networks, and office information systems. He is now a member of the Communications Principles Research group in the Applied Research Area of Bell Communications Research, Inc. Member, IEEE, Tau Beta Pi, Phi Beta Kappa.

PAPERS BY AT&T BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Graham M. H., Yannakakis M., **Independent Database Schemas.** J Comput Sy 28(1): 121–141, 1984.

Heyman D. P., Whitt W., **The Asymptotic Behavior of Queues With Time-Varying Arrival Rates.** J Appl Prob 21(1): 143–156, 1984.

Johnson D. S., Klug A., **Testing Containment of Conjunctive Queries Under Functional and Inclusion Dependencies.** J Comput Sy 28(1): 167–189, 1984.

Johnson D. S., **The NP-Completeness Column—an Ongoing Guide.** J Algorithm 5(1): 147–160, 1984.

Landwehr J. M., Pregibon D., Shoemaker A. C., **Graphical Methods for Assessing Logistic-Regression Models.** J Am Stat A 79(385): 61–71, 1984.

Massey W. A., **Open Networks of Queues—Their Algebraic Structure and Estimating Their Transient Behavior.** Adv Appl P 16(1): 176–201, 1984.

Ott T. J., **On the M/G/1 Queue With Additional Inputs.** J Appl Prob 21(1): 129–142, 1984.

Sethi R., **Preprocessing Embedded Actions.** Software 14(3): 291–297, 1984.

Witsenhausen H. S., Wyner A. D., **On Storage Media With Aftereffects.** Inf Contr 56(3): 199–211, 1983.

Zaniolo C., **Database Relations With Null Values.** J Comput Sy 28(1): 142–166, 1984.

ENGINEERING

Aumiller G. D., **Infrared Dye Laser in the 685–880-nm Range (Letter).** Appl Optics 23(5): 651, 1984.

Avanessians A. A., Beck E. C., Corcoran G. T., Eldumiati I. I., Elward J. P., Glaser A. B., Irving R. H., Spiwak R. R., Wiederhold R. P., **A VLSI Link-Level Controller.** ISSCC Diges 27: 196–197, 1984.

Barness Y. et al., **Cross-Coupled Phase-Locked Loop With Closed-Loop Amplitude Control (Letter).** IEEE Commun 32(2): 195–199, 1984.

Bonyhard P. I., Ekholm D. T., Hagedorn F. B., Muehlner D. J., Nelson T. J., Roman B. J., **Characterization of 8-μm Period Half-Megabit Ion-Implanted Memory Chip Designs.** IEEE Magnet 20(1): 129–134, 1984.

Capasso F., Alavi K., Cho A. Y., Hutchinson A. L., **Electroabsorption $Al_{0.48}In_{0.52}As$ p-i-n Avalanche Photodiodes Grown by Molecular-Beam Epitaxy.** IEEE Elec D 5(1): 16–17, 1984.

Chang T. L., Fisher P. D., **High-Speed Normalization and Rounding Circuits for Pipelined Floating-Point Processors.** IEEE Acoust 31(6): 1403–1408, 1983.

Geary J. M., Vella-Coliero G. P., **Cryogenic Wafer Prober for Josephson Devices.** IEEE Magnet 19(3): 1190–1192, 1983.

Geballe T. H., **The Science of Useful Superconductors—and Beyond.** IEEE Magnet 19(3): 1300–1307, 1983.

Gilbert J. A. et al., **Ultra Low-Frequency Holographic Interferometry Using Fiber Optics.** Opt Laser E 5(1): 29–40, 1984.

Hong M. et al., **Multifilamentary Nb-Nb_3Sn Composite by Liquid Infiltration Method—Superconducting, Metallurgical, and Mechanical Properties.** IEEE Magnet 19(3): 912–916, 1983.

Jackson S. A., Fulton T. A., **The Effect of Quantum-Mechanical Uncertainty on Punchthrough Probability in a Josephson Junction.** IEEE Magnet 19(3): 1143–1146, 1983.

Kelkar S. S., Lee F. C., **Adaptive Input Filter Compensation for Switching Regulators.** IEEE Aer El 20(1): 57–66, 1984.

Kelkar S. S., Lee F. C., **Stability Analysis of a Buck Regulator Employing Input Filter Compensation.** IEEE Aer El 20(1): 67–77, 1984.

Kohl P. A., D'Asaro L. A., Wolowodiuk C., Ostermayer F. W., **Photoelectrochemical Plating of Via GaAs-FET's.** IEEE Elec D 5(1): 7–9, 1984.

Korotky S. K., Alferness R. C., Joyner C. H., Buhl L. L., **14 Gbits/s Optical Signal Encoding for $\lambda = 1.32$ μm With Double Pulse Drive of a Ti: $LiNbO_3$ Waveguide Modulator.** Electr Lett 20(3): 132–133, 1984.

Levinson S. E., **Some Experiments With a Linguistic Processor for Continuous Speech Recognition.** IEEE Acoust 31(6): 1549–1556, 1983.

Marcuse D., Lee T. P., **Rate-Equation Model of a Coupled-Cavity Laser.** IEEE J Q El 20(2): 166–176, 1984.

Pei S. S., Nakahara S., Schreiber H., Gates J. V., **Microstructures of Lead Alloy Josephson Junction Electrode Materials: PbInAu and PbSb.** IEEE Magnet 19(3): 972–975, 1983.

Picinbon B., Bouvet M., Kadota T., **(Fr) Detection of a Deterministic Signal After Random Reflections.** Ann Telecom 38(7–8): 287–296, 1983.

Ross D. G., Paski R. M., Ehrenberg D. G., Eckton W. H., Moyer S. F., **A Regenerator Chip Set for High-Speed Digital Transmission.** ISSCC Diges 27: 240+, 1984.

Rutledge J. E., Dynes R. C., Narayanamurti V., **Superconducting Tunneling in a Pair-Breaking Microwave Field.** J L Temp Ph 54(5–6): 547–554, 1984.

Sabnis A. G., **Characterization of Annealing of Co^{60} Gamma-Ray Damage at the Si/SiO_2 Interface.** IEEE Nucl S 30(6): 4094–4099, 1983.

Swartz R. G., Chin G. M., Voshchenkov A. M., Ko P., Wooley B. A., Finegan S. N., Bosworth R. H., **Digital NMOS Test Circuits Fabricated in Silicon MBE.** IEEE Elec D 5(2): 29–31, 1984.

Swartzlander E. E., Young W. K. W., Joseph S. J., **A VLSI Delay Commutator for FFT Implementation.** ISSCC Diges 27: 266+, 1984.

Tsang W. T., Olsson N. A., Linke R. A., Logan R. A., **1.5 μm Wavelength GaInAsP C^3 Lasers: Single-Frequency Operation and Wideband Frequency Tuning.** Electr Lett 19(11): 415–417, 1983.

Vandover R. B., Bacon D. D., **Properties of NbN/Pb Josephson Tunnel Junctions.** IEEE Magnet 19(3): 951–953, 1983.

Lucky R. W., **The Office of the Future.** Chemtech US 14(3): 135–137, 1984.

MANAGEMENT/ECONOMICS

Lucky R. W., **The Office of the Future.** Chemtech US 14(3): 135–137, 1984.

PHYSICAL SCIENCES

Aeppli G. et al., **Spin Correlations Near the Ferromagnetic-to-Spin-Glass Crossover.** J Appl Phys 55(6): 1628–1633, 1984.

Boring J. W., Johnson R. E., Reimann C. T., Garret J. W., Brown W. L., Marcantonio K. J., **Ion-Induced Chemistry in Condensed Gas Solids.** Nucl Instru 218(1–3): 707–711, 1983.

Boutique J. P., Riga J., Verbist J. J., Delhalle J., Fripiat J. G., Haddon R. C., Kaplan M. L., **Electronic Structure of 3,7-Diphenyl- and 3.7-Bis(dimethylamino)-1,5-dithia-2,4,6,8-tetrazocines: Ab Initio Calculations and Photoelectron Spectra.** J Am Chem S 106(2): 312–318, 1984.

Brand H. R., Cladis P. E., **Smectic-X—the First Truly Ferroelectric Liquid Crystal.** J Phys. Lett 45(5): L217–L222, 1984.

Brand H. R., Pleiner H., **Macroscopic Dynamics of Chiral Smectic-C.** J Physique 45(3): 563–573, 1984.

Buck T. M., Wheatley G. H., Jackson D. P., **Quantitative Analysis of First and Second Surface Layers by LEIS (TOF).** Nucl Instru 218(1–3): 257–265, 1983.

Carlson N. W., Geschwind S., Devlin G. E., Batlogg B., Dillon J. F., Rupp L. W., **Spin-Flip Raman-Scattering in $Eu_{0.54}Sr_{0.46}S$.** J Appl Phys 55(6): 1679–1681, 1984.

Cava R. J., Fleming R. M., Rietman E. A., **Diffuse X-Ray Scattering Study of Single-Crystal Alpha-AgI.** Sol St Ion 9–10(Dec): 1347–1351, 1983.

Chen C. H., **Electron-Diffraction Study of the Charge-Density Wave Superlattice in 2H-$NbSe_2$.** Sol St Comm 49(7): 645–647, 1984.

Chen H. S., Sherwood R. C., Jin S., Chi G. C., Inoue A., Masumoto T., Hagiwara H.,

Mechanical Properties and Magnetic Behavior of Deformed Metal Glass Wires. J Appl Phys 55(6): 1796–1798, 1984.

Chin A. K., Chen F. S., Ermanis F., **Failure Mode Analysis of Planar Zinc-Diffused $In_{0.53}Ga_{0.47}As$ p-i-n Photodiodes.** J Appl Phys 55(6): 1596–1606, 1984.

Dillon J. F., Albiston S. D., Batlogg B., Schreiber H., **Domain Observation in Ferromagnetic and Reentrant $Eu_xSr_{1-x}S$.** J Appl Phys 55(6): 1673–1675, 1984.

Downey P. M., Schwartz B., **Picosecond Photoresponse in $^3He^+$ Bombarded InP Photoconductors.** Appl Phys L 44(2): 207–209, 1984.

Dubois L. H., Schwartz G. P., Camley R. E., Mills D. L., **Inelastic Scattering of Electrons From Ionic Crystals With a Highly Conducting Overlayer.** Phys Rev B 29(6): 3208–3216, 1984.

Fischer-Colbrie A., Fuoss P. H., **X-Ray RDF Analysis of 1500 Å Thick Amorphous Films.** J Non-Cryst59-6(Dec): 859–862, 1983.

Fisk Z., Thompson J. D., Lawrence J. M., Smith J. L., Batlogg B., **Phase Diagram and Critical Points of Ce Alloys.** J Appl Phys 55(6): 1921–1924, 1984.

Forrest S. R., Kaplan M. L., Schmidt P. H., **Organic-on-Inorganic Semiconductor Contact Barrier Diodes. 1. Theory With Applications to Organic Thin Films and Prototype Devices.** J Appl Phys 55(6): 1492–1507, 1984.

Fratello V. J., Wolfe R., Blank S. L., Nelson T. J., **High Curie-Temperature Drive Layer Materials for Ion-Implanted Magnetic-Bubble Devices.** J Appl Phys 55(6): 2554–2556, 1984.

Garcia Iniguez L., Powers L., Chance B., Sellin S., Mannervik B., Milovan A. S., **X-Ray Absorption Studies of the Zn^{2+} Site of Glyoxalase I.** Biochem 23(4): 685–689, 1984.

Gerhardmulthaupt R. et al., **Investigation of Piezoelectricity Distributions in Poly(vinylidene Fluoride) by Means of Quartz-Generated or Laser-Generated Pressure Pulses.** J Appl Phys 55(7): 2769–2775, 1984.

Geschwind S., Devlin G. E., Dillon J. F., Batlogg B., Maletta H., **Elastic Light-Scattering From the Reentrant Spin-Glass $Eu_xSr_{1-x}s$.** J Appl Phys 55(6): 1676–1678, 1984.

Gibson J. M., McDonald M. L., **A Simple Liquid-He-Cooled Specimen Stage for an Ultrahigh Resolution Transmission Electron Microscope.** Ultramicros 12(3): 219–222, 1984.

Gottscho R. A., Burton R. H., Flamm D. L., Donnelly V. M., Davis G. P., **Ion Dynamics of RF Plasmas and Plasma Sheaths—a Time-Resolved Spectroscopic Study.** J Appl Phys 55(7): 2707–2714, 1984.

Heaven M., Miller T. A., Bondybey V. E., **Chemical Formation and Spectroscopy of S_2 in a Free Jet Expansion.** J Chem Phys 80(1): 51–56, 1984.

Heinekamp S., Pelcovits R. A., Fontes E., Chen E. Y., Pindak R., Meyer R. B., **Smectic-*C** to Smectic-*A* Transition in Variable-Thickness Liquid-Crystal Films: Order-Parameter Measurements and Theory.** Phys Rev L 52(12): 1017–1020, 1984.

Hong M., Gyorgy E. M., Bacon D. D., **DC Getter Sputtered Amorphous GdCo Films—Magnetic-Anisotropy and In-Depth Chemical Composition.** Appl Phys L 44(7): 706–708, 1984.

Inoue A., Chen H. S., Krause J. T., Masumoto T., **The Effects of Quench Rate and Cold Drawing on the Structural Relaxation and Young Modulus of an Amorphous $Pd_{77.5}Cu_6Si_{16.5}$ Wire.** J Non-Cryst 61-2(Jan): 949–954, 1984.

Jayaraman A., **The Diamond-Anvil High-Pressure Cell.** Sci An 250(4): 544+, 1984.

Jin S., Sherwood R. C., Chin G. Y., Wernick J. H., Bordelon C. M., **Soft Magnetic Properties of a Ferritic Fe-Ni-Cr Alloy.** J Appl Phys 55(6): 2139–2141, 1984.

Jin S., Vandover R. B., Sherwood R. C., Tiefel T. H., **Magnetic Sensors Using Fe-Cr-Ni Alloys With Square Hysteresis Loops.** J Appl Phys 55(6): 2620–2622, 1984.

Kammlott G. W., Franey J. P., Graedel T. E., **Atmospheric Sulfidation of Copper Alloys. 1. Brasses and Bronzes.** J Elchem So 131(3): 505–511, 1984.

Kammlott G. W., Franey J. P., Graedel T. E., **Atmospheric Sulfidation of Copper Alloys. 2. Alloys With Nickel and Tin.** J Elchem So 131(3): 511–515, 1984.

Klauder J. R. et al., **Quantum-Mechanical Path Integrals With Wiener Measures for All Polynomial Hamiltonians.** Phys Rev L 52(14): 1161–1164, 1984.

Lam E. et al., **Spectroscopic Characterization of Nitrated Purple Membranes.** Biochem Int 8(2): 217–224, 1984.

Langer W. D. et al., **Carbon and Oxygen Isotope Fractionation in Dense Interstellar Clouds.** Astrophys J 277(2): 5814, 1984.

Lanzerotti L. J., Medford L. V., **Local Night, Impulsive (P12-Type) Hydromagnetic Wave Polarization at Low Latitudes.** Planet Spac 32(2): 135–142, 1984.

Larson R. G., Monroe K., **The BKZ as an Alternative to the Wagner Model for Fitting Shear and Elongational Flow Data of an LDPE Melt.** Rheol Act 23(1): 10–13, 1984.

Lax M., Odagaki T., **Hopping Conduction From Multiple-Scattering Theory and Continuous-Time Random Walk to the Coherent Medium Approximation.** AIP Conf Pr (109): 133–154, 1984.

Levine B. F., Bethea C. G., **Error Rate Measurement for Single Photon Detection at 1.3 μm.** Appl Phys L 44(7): 649–650, 1984.

Lien S. C., Huang C. C., Goodby J. W., **Heat-Capacity Studies Near the Smectic-A–Smectic-C (-Smectic-C*) Transition in a Racemic (Chiral) Smectic Liquid-Crystal.** Phys Rev A 29(3): 1371–1374, 1984.

Liu P. L., Heritage J. P., Martinez O. E., **Temperature Dependence of the Threshold Current of an InGaAsP Laser Under 130-ps Electrical Pulse Pumping.** Appl Phys L 44(4): 370–372, 1984.

Lucchese R. R., Tully J. C., **Trajectory Studies of Rainbow Scattering From the Reconstructed Si(100) Surface.** Surf Sci 137(2–3): 570–594, 1984.

MacDonald J. R., Feldman L. C., Silverman P. J., Davies J. A., Griffith K., Jackman T. E., Norton P. R., Unertl W. N., **Auger Electron Emission Induced by MeV H^+ and He^+ Ions.** Nucl Instru 218(1–3): 765–770, 1983.

Maki A. H., Weers J. G., Hilinski E. F., Milton S. V., Rentzepis P. M., **Time-Resolved Spectroscopy of Intramolecular Energy Transfer in a Rigid Spiran.** J Chem Phys 80(6): 2288–2297, 1984.

Marcuse D., **Microdeformation Losses of Single-Mode Fibers.** Appl Optics 23(7): 1082–1091, 1984.

Ocko B. M., Kortan A. R., Birgeneau R. J., Goodby J. W., **A High-Resolution X-Ray Scattering Study of the Phases and Phase Transitions in N-(4-n-Butyloxybenzylicene)-4-n-Heptylaniline (40.7).** J Physique 45(1): 113–128, 1984.

Pei S. S., Vandover R. B., **Ion-Beam Oxidation for Josephson Circuit Applications.** Appl Phys L 44(7): 703–705, 1984.

Pfeiffer L., Kovacs T., Di Salvo F. J., **Amplitude of the Charge-Density Waves in $1T$-$TaSe_2$ and $2H$-$TaSe_2$.** Phys Rev L 52(8): 687–690, 1984.

Rabinovich E. M., **Sol-Gel Preparation of Transparent Silica Glass.** J Non-Cryst 63(1-2): 155–161, 1984.

Sacharoff A. C., Westervelt R. M., Bevk J., **Magnetoresistance of Disordered Ultrathin Pt Wires.** Phys Rev B 29(4): 1647–1652, 1984.

Senft D. C., Boyd G. D., Thurston R. N., **Multiplexing the Bistable Boundary-Layer Liquid-Crystal Display.** Appl Phys L 44(7): 655–657, 1984.

Slusky S. E. G., Ballintine J. E., **Effective Mobility Limit in Small-Bubble, Low-Damping-Bubble Materials.** J Appl Phys 55(6): 2548–2550, 1984.

Smith P. W., Ashkin A., Bjorkholm J. E., Eilenberger D. J., **Studies of Self-Focusing Bistable Devices Using Liquid Suspensions of Dielectric Particles.** Optics Lett 9(4): 131–133, 1984.

Stephens P. W. et al., **High-Resolution X-Ray-Scattering Study of the Commensurate-Incommensurate Transition of Monolayer-Kr on Graphite.** Phys Rev B 29(6): 3512–3532, 1984.

Teo B. K., Snyder-Robinson P. A., **Metal Tetrathiolenes. 8. Molecular Structures of Two Isostructural Two-Electron Systems: $(Ph_3P)_2(CO)XIr(C_{10}Cl_4S_4)$ (X = Cl.H). The First Member of a Novel Series of Metal Tetrathiolene Complexes.** Inorg Chem 23(1): 32–39, 1984.

Tiefel T. H., Jin S., **Microduplex Fe-Ni-Mo Semihard Magnet Alloys.** J Appl Phys 55(6): 2112–2114, 1984.

Tu C. W., Sheng T. T., Macrander A. T., Phillips J. M., Guggenheim H. J., **Lattice-Matched Single-Crystalline Dielectric Films ($Ba_xSr_{1-x}F_2$) on InP (001) Grown by Molecular-Beam Epitaxy.** J Vac Sci B 2(1): 24–26, 1984.

Vanuiter L. G., **An Empirical Relation Fitting the Position in Energy of the Lower D-Band Edge for Eu^{2+} or Ce^{3+} in Various Compounds.** J Luminesc 29(1): 1–9, 1984.

Vonseggern H., West J. E., **Stabilization of Positive Charge in Fluorinated Ethylene Propylene Copolymer.** J Appl Phys 55(7): 2754–2757, 1984.

Wakatani M., Hasegawa A., **A Collisional Drift Wave Description of Plasma Edge Turbulence.** Phys Fluids 27(3): 611–618, 1984.

Walton C. R., Goodby J. W., **Esters That Exhibit Smectic-F to Isotropic Liquid Phase Transitions.** Molec Cryst 92(9–10): 263–269, 1984.

Weber T. A., Stillinger F. H., **The Effect of Density on the Inherent Structure in Liquids.** J Chem Phys 80(6): 2742–2746, 1984.

Weiss M. A., Karplus M., Patel D. J., Sauer R. T., **Solution NMR Studies of Intact Lambda Repressor.** J Bio Struc 1(1): 151–157, 1983.

Weiss M. A., Patel D. J., Sauer R. T., Karplus M., **Two-Dimensional ^{1}HNMR Study of the λ Operator Site O_L1: A Sequential Assignment Strategy and Its Application.** Proc Natl Acad Sci USA 81(1): 130–134, 1984.

White J. C., Craighead H. G., Howard R. E., Jackel L. D., Behringer R. E., Epworth R. W., Henderson D., Sweeney J. E., **Submicron, Vacuum Ultraviolet Contact Lithography With an F_2 Excimer Laser.** Appl Phys L 44(1): 22–24, 1984.

Wood T. H., Burrus C. A., Miller D. A. B., Chemla D. S., Damen T. C., Gossard A. C., Wiegmann W., **High-Speed Optical Modulation With GaAs/GaAlAs Quantum Wells in a *p-i-n* Diode Structure.** Appl Phys L 44(1): 16–18, 1984.

Yafet Y., Vier D. C., Schultz S., **Conduction Electron-Spin Resonance and Relaxation in the Superconducting State.** J Appl Phys 55(6): 2022–2024, 1984.

Yurke B., Denker J. S., **Quantum Network Theory.** Phys Rev A 29(3): 1419–1437, 1984.

SOCIAL AND LIFE SCIENCES

Julesz B., **A Brief Outline of the Texton Theory of Human Vision.** Trends Neur 7(2): 41–45, 1984.

SPEECH/ACOUSTICS

Buschvishniak I. J., **Response of an Edge-Supported Circular Membrane Electret Earphone. 1. Theory.** J Acoust So 75(3): 977–989, 1984.

Buschvishniak I. J., **Response of an Edge-Supported Circular Membrane Electret Earphone. 2. Experimental Results.** J Acoust So 75(3): 990–995, 1984.

Nelson W. L., Perkell J. S., Westbury J. R., **Mandible Movements During Increasingly Rapid Articulations of Single Syllables—Preliminary Observations.** J Acout So 75(3): 945–951, 1984.

Roberts L. A., Mathews M. V., **Intonation Sensitivity for Traditional and Nontraditional Chords.** J Acoust So 75(3): 952–959, 1984.

CONTENTS, NOVEMBER 1984

AT&T BELL LABORATORIES 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.

AT&T
Bell Laboratories