THE NOVEMBER 1981
VOL. 60, NO. 9

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published monthly, except for the May–June and July–August combined issues, by the American Telephone and Telegraph Company, C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; F. A. Hutson, Jr., Secretary. Editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-336, Crawfords Corner Road, Holmdel, N.J. 07733. Checks for subscriptions should be made payable to The Bell System Technical Journal and should be addressed to Bell Laboratories, Circulation Group, Bell Laboratories, 101 J. F. Kennedy Parkway, Short Hills, N.J. 07078. Subscriptions $20.00 per year; single copies $2.00 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A. Second-class postage paid at New Providence, New Jersey 07974 and additional mailing offices.

ISSN0005-8580

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-336, Crawfords Corner Road, Holmdel, N.J. 07733. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 60 November 1981 Number 11

Adaptive Cancellation of Intersymbol Interference for Data Transmission

By A. GERSHO and T. L. LIM

(Manuscript received April 22, 1981)

In this paper, we analyze a technique for accurately detecting transmitted data symbols contained in a modulated signal that has been degraded by a linear dispersive channel and additive Gaussian noise. The approach uses an adaptive equalizer which provides preliminary decisions to an adaptive canceller. The canceller output is used to remove the interference from an adaptive matching filter, resulting in the desired signal. Channel equalization attempts to invert the channel transfer function, while avoiding excessive noise enhancement. However, cancellation (as used in echo cancellers), attempts to generate a replica of the interfering signal and subtract it from the actual received signal containing the sum of the desired signal and interference. The cancellation approach, unlike equalization, offers the possibility of removing interference without enhancing the level of noise already present in the received waveform. Simulation results for transmission over practical channels show significant improvement of linear cancellation over both linear forward and decision-feedback equalization.

I. INTRODUCTION

For the past twenty years, engineers have been seeking new techniques to combat the intersymbol interference (ISI) in data transmission over band-limited channels. Adaptive equalization with the mean-square algorithm has been the major technique that allowed a substantial increase in attainable transmission rate.[1] If the channel has only phase distortion, then the linear fractionally spaced equalizer can eliminate virtually all of the ISI without enhancing the noise level.[2,3] However, when amplitude distortion is present in the channel, any adaptive linear equalizer (LE) must compromise between inverting the channel transfer function and avoiding excessive noise enhancement. Inevitably, some noise enhancement occurs. Decision-feedback equalization can offer somewhat improved performance when amplitude distortion is present.[4,5] By using the Viterbi algorithm,[6] maximum-likelihood receivers, in principle, offer the best performance possible but depend on adaptive estimators of the channel and require an impractically high complexity when the channel impulse response is long, as in the case of the typical telephone channel.

If there were no ISI, the probability of error in detecting the transmitted data level (i.e. ±1 or ±3) would be the same as if only one such pulse were transmitted in isolation. In that case, the optimal receiver (for Gaussian noise) would be a matched filter and would yield a certain error probability, P_0. When pulses are sent sequentially, the effect of ISI cannot be totally eliminated. The maximum likelihood estimator of the entire sequence of transmitted symbols is known to result in an error probability that is somewhat larger than P_0.

In this paper, we describe a cancellation technique designed to achieve isolated-pulse, matched-filter performance. Extensive simulation results confirm that there is a significant improvement over linear or decision-feedback equalization for severe amplitude-distorting channels of practical interest.

II. BACKGROUND AND MOTIVATION

The idea of cancellation was used for the echo problem in two-wire telephony (see Ref. 7), where the received signal contains an interference component that is a filtered and delayed version of an "originating" signal. In that application, the originating signal is actually available at the same location. However, in data transmission the originating signal—the transmitted data stream—is not directly observable at the receiver. The idea of using preliminary decisions to generate an intermediate estimate of the transmitted data signal was independently proposed by various investigators. All of the proposals included adapting the coefficients of a filter that forms a replica of the ISI. Adaptation is, of course, needed since the appropriate filtering operation is not known in advance and can vary with time. Hirsch and Proakis proposed the cancellation scheme with the essential structure in Figure 1.[8,9] Here the canceller attempts to remove the ISI directly from the received line signal. This approach does not achieve improved performance over linear equalization if there is phase distortion. In this paper, we describe a linear canceller (LC) structure where a

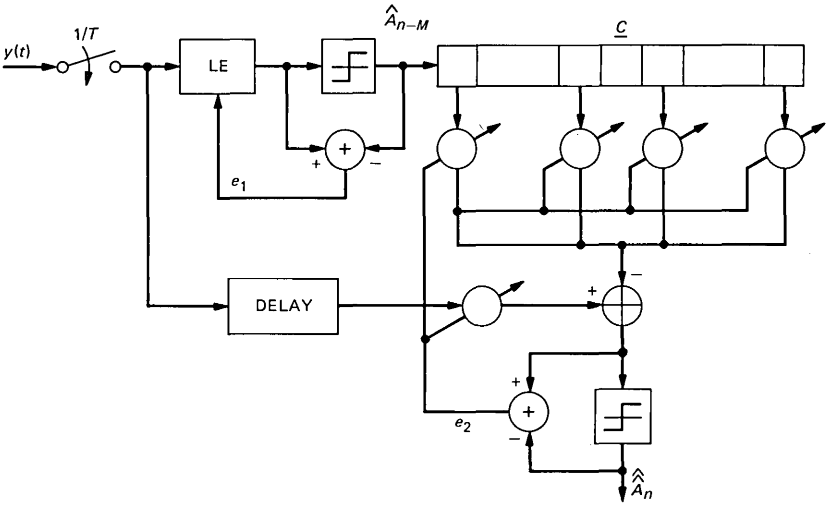


Fig. 1—First generation linear canceller.

transversal filter $\underline{W}$ is used instead of the delay in Fig. 1. Both $\underline{C}$ and $\underline{W}$ are adapted simultaneously with the error signal between the input to the final detector and the appropriate reference.

The motivation for this structure stems from the need to effectively detect high-speed data on channels that have both a high noise level and substantial amplitude (slope) and phase distortion. In certain conditions of practical interest, even if the equalizer were of infinite length, the noise enhancement of linear or decision feedback equalization makes it impossible to achieve the required error rate. Nevertheless, linear equalization is sufficient to obtain fairly moderate error rates so that the detected symbols may be adequate as preliminary decisions for a cancellation scheme.

In Section 3, we provide intuitive reasoning that the optimal choice for $\underline{W}$ is a matched filter, and the optimal $\underline{C}$ is a canceller whose tap weights are the samples of the channel autocorrelation function, except for the center tap, which has zero weight. We show in Section 4 that this is true under the assumption that the preliminary decisions $\hat{A}_k$ are correct. Section 5 covers adaptive operation and Section 6, simulation results.

III. FORMULATION

Let the transmitted data symbols be denoted A_1, A_2, A_3, $\cdots$ with each complex valued symbol having real and imaginary parts restricted to one of a finite set of values (i.e. ±1, ±3). A complex-valued pulse shape $p(t)$ is used to generate the baseband transmitted data signal

$$s(t) = \sum_k A_k p(t - kT). \tag{1}$$

The linear distortion of the channel results in the received waveform

$$X(t) = \sum_k A_k h(t - kT) + V(t), \tag{2}$$

where $V(t)$ is white noise and $h(t)$ is the overall channel impulse response.

Suppose initially that the receiver consists of a matched filter and a sampler (rate $1/T$) followed by a symbol detector as shown in Fig. 2. The matched filter has impulse response

$$W(t) = h^*(LT - t), \tag{3}$$

where * denotes complex conjugation and the integer L is chosen large enough so that the output is small for $t < 0$. The output of the sampler is then

$$U(mT) = \sum_k A_k r(mT - kT - LT) + V^1(mT), \tag{4}$$

where $V^1(t)$ is the colored noise at the matched filter output and $r(t)$ is the autocorrelation function of the pulse $h(t)$; that is, $r(t) = \int_{-\infty}^{\infty} h(s)g(s + t)ds$. Equation (4) can be written as

$$U(mT) = A_{m-L}r(0) + V^1(mT) + I_{m-L}, \tag{5}$$

where

$$I_{m-L} = \sum_{k \neq m-L} A_k r(mT - kT - LT). \tag{6}$$

Suppose that at time $t = mT$ the receiver must detect the currently observable symbol A_{m-L} and it knows all prior symbols A_{m-L-1}, $A_{m-L-2}, \ldots$ and all subsequent symbols A_{m-L+1}, $A_{m-L+2}, \ldots$, A_m that determine the value I_{m-L} of the total ISI. In this case, I_{m-L} is a known constant and can be subtracted from $U(mT)$. What remains is exactly the output value that the matched filter would produce if the transmitter sent a single isolated pulse $A_{m-L}p[t - (m - L)T]$. Hence, the ideal performance, with error probability P_0, would be achieved.

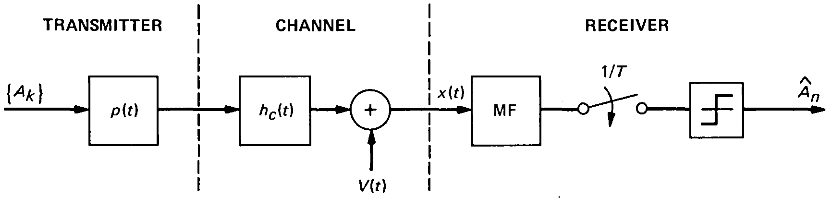


Fig. 2—Model of transmitter, channel and matched filter (MF).

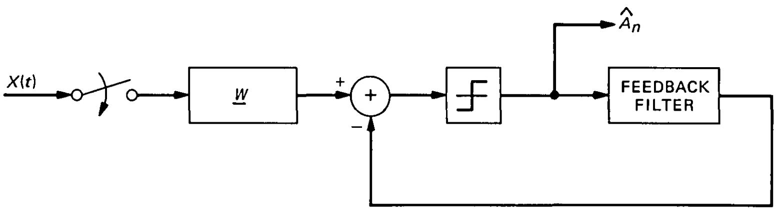


Fig. 3—Decision feedback equalizer.

The above reasoning suggests that we could approach the ideal—isolated symbol—performance with each symbol decision if we could generate a good estimate of the total ISI, I_{m-L}, at each sampling instant, $t = mT$. The decision feedback approach can be viewed as a partial step in this direction. This approach is based on the idea that we can estimate the prior symbols called precursors by storing and processing the outputs $\hat{A}_{m-L-1}$, $\hat{A}_{m-L-2}, \ldots$ already produced by the detector. The part of I_{m-L} determined by the precursors can then be constructed. By applying the decision $\hat{A}_i$ to a feedback filter, the output is subtracted from $U(mT)$ and the resultant signal is applied to the detector. The decision-feedback equalizer (DFE) is shown in Fig. 3. Since we have not removed all of the ISI, the resulting performance will be inferior to that of the isolated-pulse case. This discussion shows that the DFE technique can be regarded as a partial step towards the goal of totally removing ISI. We next examine how we go beyond the stage of postcursor cancellation to include precursor cancellation.

Suppose at time mT we could also eliminate the subsequent, postcursor, symbols A_{m-L+1}, $A_{m-L+2}, \ldots, A_m$ that also contribute to the total interference at time mT. Then, using eq. (6), an estimate of the total ISI, I_{m-L}, would be available at time instant mT. This is not possible using the output of the detector in Fig. 2. However, suppose a separate equalizer operates on the received data signal $y(t)$ as shown in (a) of Fig. 4. If optimally designed, it will have a modest error rate, and we can use its decision $\hat{A}_n$, as preliminary or tentative decisions for the purpose of constructing our estimate of I_{m-L}. Now there is no problem in obtaining both precursor and postcursor estimates needed to form I_{m-L}. By introducing a fixed time delay, D, to the received signal prior to the matched filter as shown in (a), the LE has a head-start in estimating data symbols. A practical implementation of an adaptive passband canceller would take the form shown in (b) of Fig. 4.

The delay D can actually be incorporated into the matched filter by choosing L suitably large. The cancellation filter $\underline{C}$ produces the estimate $\hat{I}_{m-L}$ of the actual interference

$$\hat{I}_{m-L} = \sum_{k \neq m-L} \tilde{A}_k r(mT - kT - LT), \tag{7}$$

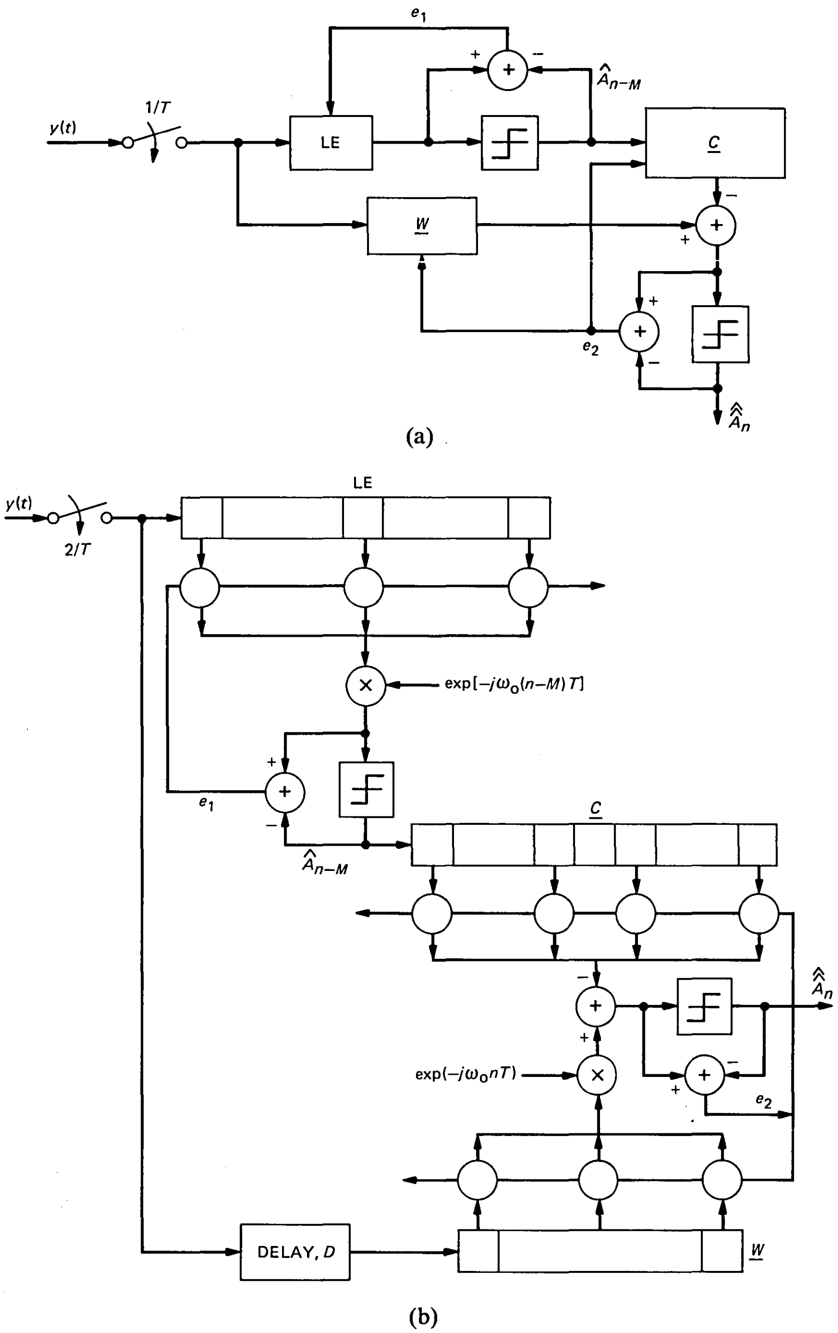


Fig. 4—(a) Linear canceller. (b) Passband linear canceller structure.

where $\tilde{A}_k$ is the sequence of preliminary decisions. Note that the transversal filter $\underline{C}$, which takes as input $\hat{A}_m$ and produces the output $\hat{I}_{m-L}$ at time mT, has an impulse response

$$c_i = \begin{cases} r[i - (m - L)T] & i \neq m - L \\ 0 & i = m - L \end{cases}. \tag{8}$$

We shall see in Section IV that, under the assumption of perfect preliminary decisions, this is indeed the optimum impulse response in the mean-square sense.

IV. OPTIMAL CANCELLATION

4.1 Derivation of optimal filter coefficients

To determine the optimal pair of filters $\underline{W}$ and $\underline{C}$ for the cancellation scheme, we make the simplifying assumption that the preliminary decisions available from the LE are correct. We focus on the structure shown in Fig. 5, where the filter $\underline{W}$, called the matching filter, is a $T/2$-spaced infinite length transversal filter preceded by a sampler operating at rate $2/T$ samples per second. The filter $\underline{C}$, called the canceller, is a T-shaped infinite length transversal filter.

The matching filter $\underline{W}$ has input samples, $y_l = y(lT/2)$, where $y(t)$ is the received line signal, and the output, U_n, of $\underline{W}$ is taken at time instants, $t = kT$, as indicated in Fig. 5 by the $1/T$ rate sampler. The input to canceller $\underline{C}$ is the true data sequence $\{A_n\}$ since we have assumed the tentative decisions are correct. Thus, Fig. 5 shows $\underline{C}$ being fed directly by the transmitted data symbols. It is not necessary to explicitly consider the time delay D since we are allowing infinite length filters. The output of the canceller V_n is subtracted from the matching-filter output. This difference producing $U_n - V_n$ to be applied to a slicer may be viewed as a linear estimator of the data symbol A_n. The goal is to determine the filters $\underline{W}$ and $\underline{C}$ that minimize

$$E_n = E(|\epsilon_n|^2), \tag{9}$$

the mean-square error (mse), where $\epsilon_n = U_n - V_n - A_n$.

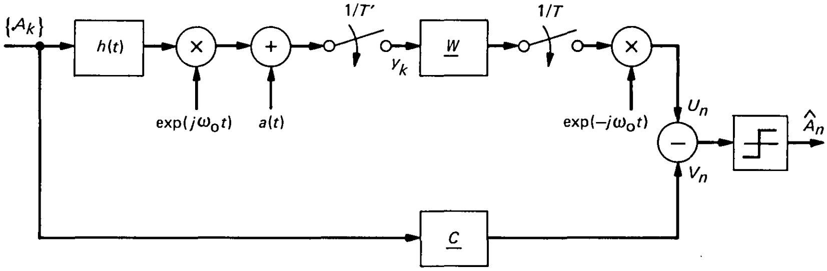


Fig. 5—Model of a linear canceller.

Let $\alpha(t)$ denote the additive receiver noise as shown in Fig. 5. Then, the input to the matching filter is given by

$$\begin{aligned} y_l &= \exp(j\omega_0 lT/2) \sum_v A_v h(lT/2 - vT) + \alpha(lT/2) \\ &= \exp(j\omega_0 lT/2) \left[\sum_v A_v h(lT/2 - vT) + \alpha_1(lT/2) \right], \end{aligned} \tag{10}$$

where $h(t)$ is the complex impulse response of the channel and includes the effect of the transmitter shaping filter $p(t)$ as treated in Section III. The term $\alpha_1(\cdot)$ is the complex baseband noise sample. All summations are over the integers from $-\infty$ to ∞, unless otherwise indicated. The output U_n of the matching filter is

$$U_n = \exp(-j\omega_0 nT) \sum_k W_k^* y_{2n-k}, \tag{11}$$

where W_i denotes the tap weights of $\underline{W}$. Let $h_l = h(lT/2)$ and define

$$b_l = \sum_v W_v^* h_{l-v}. \tag{12}$$

Then U_n may be written in the form

$$U_n = \sum_s A_s b_{2n-2s} + N_n, \tag{13}$$

where

$$N_n = \sum_s W_s^* \alpha_{2n-s} \tag{14}$$

is the noise at the output of the matching filter.

We shall let C_i denote the ith tap weight of the cancellation filter for each integer i, except $i = 0$. We make the constraint that the *center tap weight of the cancellation filter is zero.* This restricts the role of the canceller to removing ISI and prevents the canceller from making use of the current data symbol which must be estimated by the output signal from the matching filter. The canceller output V_n is given by

$$V_n = \sum_{i \neq 0} C_i^* A_{n-i}, \tag{15}$$

and we assume that

$$E|A_n|^2 = 1, \tag{16}$$

which is a convenient normalization of the data symbol power level.

To minimize the mse in eq. (9), we differentiate E_n with respect to the complex tap weights $\{C_k\}$ and $\{W_k\}$ and set the derivatives to zero. Using eq. (15), it can be shown to yield

$$E(\epsilon_n^* A_{n-m}) = 0 \qquad m \neq 0, \tag{17}$$

and using eq. (11),

$$\exp(-j\omega_0 nT) E(\epsilon_n^* y_{2n-m}) = 0 \qquad \text{all } m. \tag{18}$$

Thus, these optimality conditions require that the error signal ϵ be orthogonal to the observable inputs to the $\underline{C}$ and $\underline{W}$ filters, namely, $\{A_n\}$ and $\{y_n\}$.

We discuss two cases of interest. In Case 1, the channel is wideband. We assume that the $\underline{W}$ filter processes the line signal directly from the channel without prior band-limiting so that even when sampled at a rate of $2/T$, the noise samples are uncorrelated. In Case 2, which is more relevant to our situation of a channel band-limited to the voice frequency range, the noise samples at $T/2$ spacing are correlated.

Case 1. Uncorrelated noise samples

In Case 1,

$$E[\alpha(lT/2)\alpha^*(kT/2)] = \sigma^2 \delta_{lk},$$

where σ^2 is the noise variance. We also define a new term,

$$R_h(l) = \sum_j h^*(jT/2)h(jT/2 + lT/2), \tag{19}$$

which is the autocorrelation function of the $T/2$-sampled impulse response, $h(jT/2)$. Then, with

$$\begin{aligned} E_h &= \sum_j \left| h\left(j\frac{T}{2}\right) \right|^2 \\ &= R_h(0), \end{aligned} \tag{20}$$

we show in the Appendix that the matching filter has $T/2$-spaced tap weights

$$W_m = \exp(-j\omega_0 mT/2)\frac{h(-mT/2)}{E_h + \sigma^2}, \qquad \text{all } m, \tag{21}$$

which is clearly proportional to a matched-filter impulse response. The $\underline{C}$ taps are shown in the Appendix to be

$$C_m = \frac{1}{(E_h + \sigma^2)} R_h(2m), \qquad m \neq 0. \tag{22}$$

Thus, the canceller impulse response, for $m \neq 0$, is that of the overall T-spaced impulse response of the channel and matching filter.

Case 2. Correlated noise samples

As described earlier, Case 2 corresponds to the voiceband telephone channel, where the noise has approximately the same bandwidth as the signal so that noise samples at $T/2$ spacing are correlated.

The noise correlation is

$$E[(\alpha kT/2)\alpha^*(lT/2)] = R_n(k-l).$$

We define $W(\omega)$ as the Fourier transform of the $\underline{W}$ tap weights, $S_n(\omega)$ is the sampled noise spectrum, and $\tilde{H}(\omega)$ is the Fourier transform of the channel-sampled impulse response. Then, with

$$\xi = \frac{T}{4\pi}\int_{\frac{-2\pi}{T}}^{\frac{2\pi}{T}} \frac{|\tilde{H}(-\omega-\omega_0)|^2}{S_n(-\omega)}\,d\omega, \tag{23}$$

we show in the Appendix that

$$W(\omega) = \frac{\tilde{H}(-\omega-\omega_0)}{(1+\xi)S_n(-\omega)}. \tag{24}$$

The corresponding Fourier transform of the canceller is

$$C(\omega) = \frac{1}{1+\xi}\frac{|\tilde{H}(-\omega)|^2}{S_n(-\omega+\omega_0)}. \tag{25}$$

Note that since

$$R_h^*(-2m) = R_h(2m),$$

and $S_n(-\omega+\omega_0)$ is real, we see from eqs. (22) and (25) that the optimum, infinitely long canceller sampled impulse response is Hermitian symmetric about the center, i.e.

$$C_m = C_{-m}^*. \tag{26}$$

4.2 Derivation of mse

We now derive the mse achieved for Case 1, under the assumption that $\underline{W}$ and $\underline{C}$ have the optimal impulse responses given by eqs. (21) and (22), respectively. The matching filter output is given by

$$U_n = \frac{1}{E_h+\sigma^2}\sum_s A_s R_h(2n-2s) + N_n. \tag{27}$$

The noise, N_n, is the result of applying white noise with variance σ^2 to the matching filter eq. (65). Hence, using eqs. (66) and (67), we find that

$$E|N_n|^2 = \frac{\sigma^2 E_h}{(E_h+\sigma^2)^2}. \tag{28}$$

Also, the canceller output is given by

$$V_n = \frac{1}{E_h+\sigma^2}\sum_{s\neq n} A_s R_h(2n-2s). \tag{29}$$

Hence, the error signal, with $R_h(0) = E_h$, is

$$\epsilon_n = U_n - V_n - A_n = \frac{E_h}{E_h + \sigma_n^2} A_n + N_n - A_n$$

$$= -\left(\frac{\sigma^2}{E_h + \sigma^2}\right) A_n + N_n, \tag{30}$$

so that, using eq. (30), the mse is

$$E_n = E|\epsilon_n|^2 = \frac{\sigma^2}{E_h + \sigma^2}. \tag{31}$$

The s/n, ρ, at the output of the cancellation system is defined as the ratio

$$\rho = \frac{E|U_n - V_n - N_n|^2}{E|N_n|^2}. \tag{32}$$

Hence, we find that

$$\rho = \frac{E_h}{\sigma^2}. \tag{33}$$

4.3 On the property of invariance of mse to timing phase

The mse expression in eq. (31) can be written in terms of the channel-sampled power spectrum. If $H(\omega)$ is the Fourier transform of the overall impulse response $h(t)$, then the transform of $h(jT/2)$ is

$$\tilde{H}(\omega) = \frac{2}{T}\sum_k H\left[\omega + \frac{4\pi k}{T}\right], |\omega| < \frac{2\pi}{T}, \tag{34}$$

and

$$E_h = \sum_j \left|h\left(j\frac{T}{2}\right)\right|^2$$

$$= \frac{T}{4\pi}\int_{-\frac{2\pi}{T}}^{\frac{2\pi}{T}} |\tilde{H}(\omega)|^2 d\omega. \tag{35}$$

Therefore, from eq. (31) the mse of the optimum LC, normalized to unit signal power, is

$$E_{\text{LC}} = 1\bigg/\left[\frac{T}{4\pi\sigma^2}\int_{-\frac{2\pi}{T}}^{\frac{2\pi}{T}} |\tilde{H}(\omega)|^2 d\omega + 1\right] \tag{36}$$

If the matching filter were T-spaced, then we have

$$\tilde{H}(\omega) = \frac{1}{T}\sum_k H\left(\omega + \frac{2\pi k}{T}\right), \tag{37}$$

and

$$E_h = \frac{T}{2\pi}\int_{\frac{-\pi}{T}}^{\frac{\pi}{T}} |\tilde{H}(\omega)|^2 d\omega. \tag{38}$$

We can compare the results with similar expressions for the optimum LE and DFE in Ref. 5 which assume matched filters preceding the equalizers. They are

$$E_{\mathrm{LE}} = \frac{T}{2\pi}\int_{\frac{-\pi}{T}}^{\frac{\pi}{T}} \frac{1}{Y(\omega)+1}\, d\omega, \tag{39}$$

and

$$E_{\mathrm{DFE}} = \exp\left\{-\frac{T}{2\pi}\int_{\frac{-\pi}{T}}^{\frac{\pi}{T}} \ln[Y(\omega)+1]d\omega\right\}, \tag{40}$$

where

$$Y(\omega) = \frac{1}{\sigma^2}\sum_k \left|H\left(\omega+\frac{2\pi k}{T}\right)\right|^2. \tag{41}$$

The expression in eq. (39) is the same as that for an infinitely long fractionally spaced LE,[3] whereas the result for a symbol-spaced equalizer is

$$E_{\mathrm{LE}} = \frac{T}{2\pi}\int_{\frac{-\pi}{T}}^{\frac{\pi}{T}} \left[1\Big/\frac{1}{\sigma^2}\left|\sum_k H\left(\omega+\frac{2\pi k}{T}\right)\right|^2 + 1\right] d\omega. \tag{42}$$

Finally, we note that if the overall channel-impulse response has less than 100 percent rolloff and the matching filter is $T/2$-spaced, then

$$\tilde{H}(\omega) = \frac{1}{T}H(\omega),$$

and

$$E_h = \frac{T}{4\pi}\int_{\frac{-2\pi}{T}}^{\frac{2\pi}{T}} |H(\omega)|^2 d\omega.$$

Clearly, E_{LC} is independent of any phase characteristics and, hence, the canceller performance is insensitive to timing phase. This property is shared by the fractionally spaced LE[2,3] and the symbol-spaced LE and DFE which are preceded by matched filters, as exhibited by eqs. (39) and (40). On the other hand, by itself, the symbol-spaced LE is sensitive to timing phase, as seen in eq. (42), because the integrand is a function of the folded spectrum of the channel. This is because a symbol-spaced filter cannot properly synthesize a proper matched filter. Likewise, the LC with a T-spaced matching filter is sensitive to timing phase, although some simulations have indicated that the effect of a bad timing phase seems to be less on the LC than on the LE.

V. ADAPTIVE CANCELLER STRUCTURE

In practice, we have finite length filters and the ensemble averages described above are not available to the receiver. As in the case of the adaptive LE, we adjust the complex $\{C_k\}$ taps of the canceller and $\{W_k\}$ taps of the matching filter to minimize the instantaneous squared error,

$$|\epsilon_n|^2 = |U_n - V_n - A_n|^2, \tag{43}$$

at each time instant nT. As shown in (b) of Fig. 4, U_n and V_n are the outputs of the matching filter and canceller, respectively, where

$$U_n = \exp[-j(\omega_0 nT + \hat{\theta}_n)] \sum_{-L}^{L-1} W_k^* y_{2n-k}, \tag{44}$$

and

$$V_n = \sum_{\substack{k=-M \\ \neq 0}}^{M} C_k^* \hat{A}_{n-k}, \tag{45}$$

where we have $2L$ complex $\underline{W}$ taps at $T/2$ spacing and $2M$ complex $\underline{C}$ taps at T spacing. Unlike the ideal case in Fig. 5, $\hat{A}_{n-k}$ are the tentative decisions obtained from the LE, and $\hat{\theta}\,O_n$ is the estimate of any phase jitter present. The value of A_n in eq. (43) is that of the ideal reference in the training mode and is the receiver's output decision in the decision-directed mode of operation. Note that ϵ_n in eq. (43) is a linear function of the tap weights $\{C_k, W_k\}$ so that it is possible in theory to jointly adapt the tap weights to achieve a unique mse.

It can be shown that the gradients of $|\epsilon_n|^2$ with respect to the tap weights are given by

$$\frac{\partial|\epsilon_n|^2}{\partial C_k} = -2\epsilon_n^* \hat{A}_{n-k}, \tag{46}$$

and

$$\frac{\partial|\epsilon_n|^2}{\partial W_k} = 2\exp[-j(\omega_0 nT + \hat{\theta}_n)]\epsilon_n^* y_{2n-k}. \tag{47}$$

Thus, the adjustment algorithms are

$$C_k(n+1) = C_k(n) + \beta\epsilon_n^* \hat{A}_{n-k} \qquad \begin{cases} k = -M, \cdots, M \\ k \neq 0, \end{cases} \tag{48}$$

and

$$W_k(n+1) = W_k(n) - \beta \exp[-j(\omega_0 nT + \hat{\theta}_n)]\epsilon_n^* y_{2n-k},$$
$$k = L, \cdots, L-1. \quad (49)$$

The step size β is chosen to obtain reasonably fast tap convergence without the algorithm becoming unstable. The problem is similar to that addressed in Ref. 10 for adaptive equalizers where it is proved that the step size needs to satisfy the constraint

$$0 < \beta < \frac{2}{L\,\langle \chi^2 \rangle},$$

where L is the total number of taps, $(2L + 2M)$ in our case, and $\langle \chi^2 \rangle$ is the average input signal power. The fastest initial convergence is achieved with

$$\beta = \frac{1}{L\,\langle \chi^2 \rangle}. \quad (50)$$

VI. COMPUTER SIMULATION

The performance of the system over three channel models is evaluated. Channel 1 has a flat amplitude response over the entire voiceband frequency with little delay distortion, except near the lower band edge. Channel 2 has moderate amplitude and delay distortion, which just meets the private-line conditioning, and Channel 3 has severe amplitude distortion. These characteristics are shown in Fig. 6. Pseudorandom digital data at levels ±1 and ±3 are used to modulate the quadrature amplitude modulation (QAM) transmitter at a rate of 9.6 kb/s and the resulting signal is sent over one of the channels. Additive Gaussian noise and, whenever desired, phase jitter are introduced into the channel. The input signal-to-channel noise ratio (s/n_i) will be taken to be either 30, 24, or 20 dB.

In the simulations, the LC filters $\{C_k\}$ and $\{W_k\}$ are allowed to start adapting after 1000 iterations to ensure that the LE has converged sufficiently and is able to provide good tentative decisions for the LC. The performance of the LC is measured by the output s/n (s/n_o). This is defined as the ratio of the average data symbol power to the output noise power, which is taken to be the time average of the squared error at the LC output. We exhibit plots of s/n_o as a function of time, expressed in number of data symbol intervals, for both the LE and LC. Scatter plots of the output signal constellation are also provided. In the absence of any impairments, the plot would show 16 points with the x and y coordinates at ±1, ±3.

In the simulations, both LC filters span 31 data symbols so that the matching filter has 62 $T/2$ taps, while the canceller has 31 taps. The

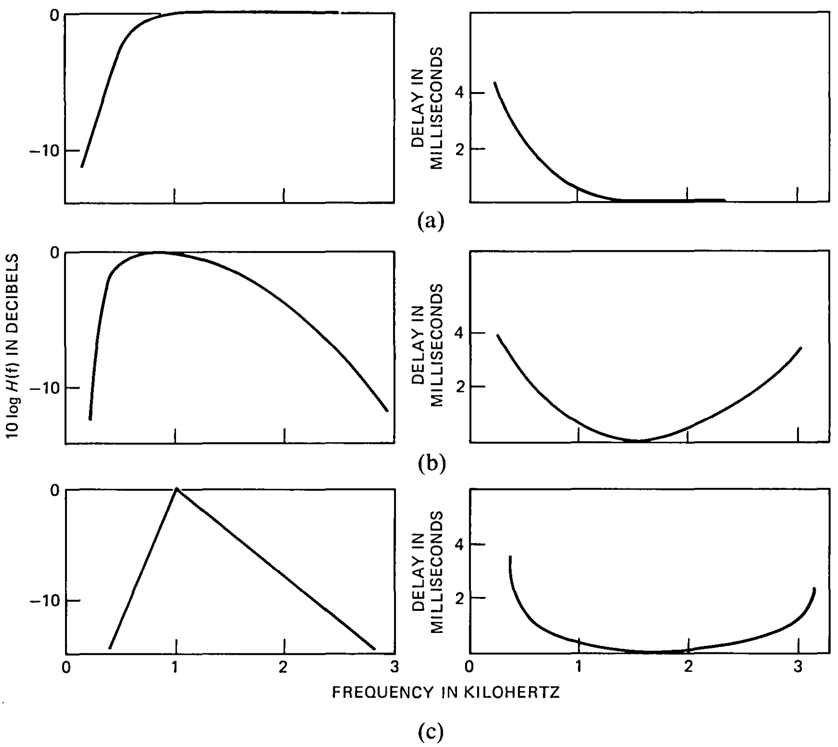


Fig. 6—Channel characteristics. (a) Channel 1. (b) Channel 2. (c) Channel 3.

LE has 64 $T/2$ taps to sufficiently span the channel-impulse response. An initial step size of 0.0005 is chosen for updating all the tap coefficients and it is reduced to 0.00005, when the taps have nearly converged, to get a small final mse.

We first compare the performance of the LC and LE over the three channels described earlier using an s/n_i of 24 dB. The receiver switches from the ideal reference to the decision directed mode of operation after 3000 iterations. Figures 7, 8, and 9 show the receiver s/n_o as a function of time for the three channels.

Note that in the absence of significant distortion, as in the case of the Channel 1, the LE and LC perform equally well, and we see that s/n_o is approximately s/n_i. However, the LE degrades more than the LC when the channel is more severely distorted in amplitude.

Table I summarizes the results of several simulation runs, for the LE, DFE, and LC, over Channels 2 and 3 and with various s/n_i. Results for Channel 1 are not presented because all three schemes do not suffer any significant degradation. These results are based on the use of an ideal carrier demodulating phase. Note that for Channel 2, the

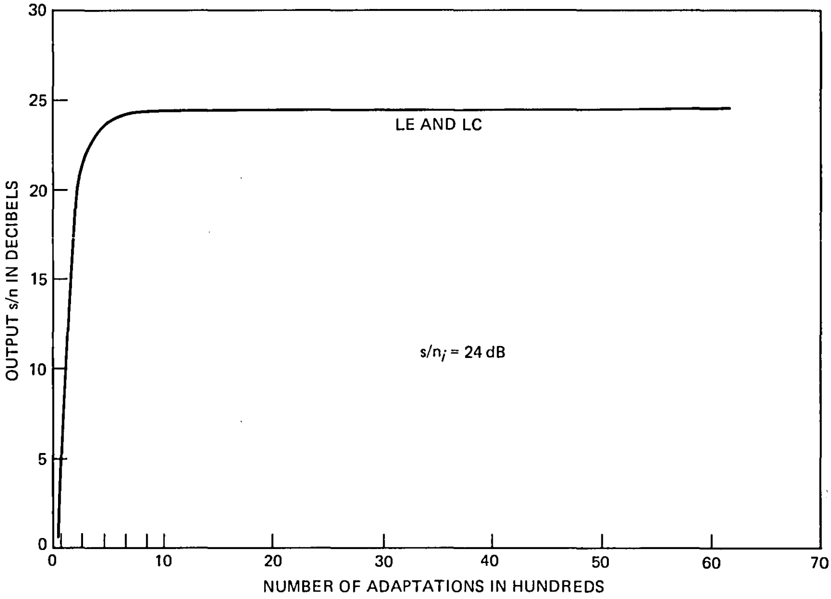


Fig. 7—Performance of LE and LC over Channel 1.

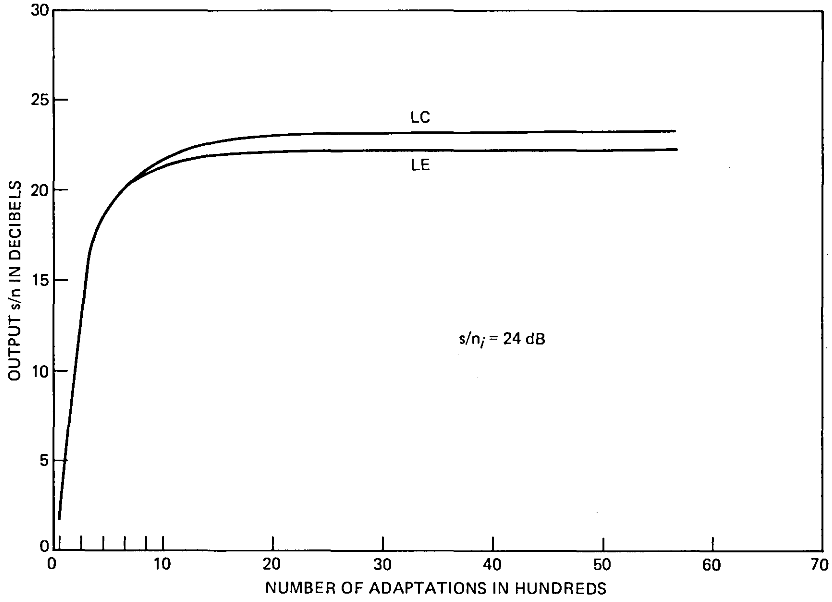


Fig. 8—Performance of LC and LE over Channel 2.

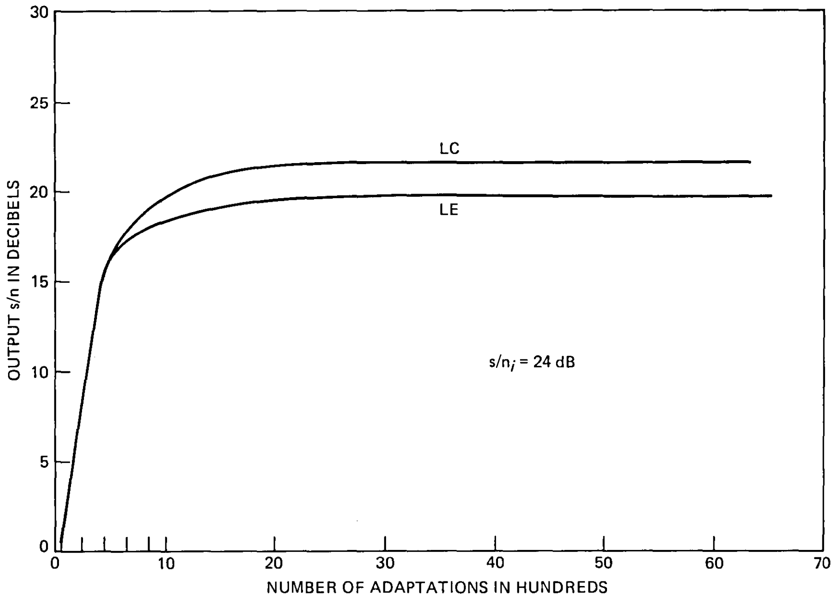


Fig. 9—Performance of LC and LE over Channel 3.

LC is able to attain s/n$_o$'s which are very close to s/n$_i$, for all three cases. The gain is about 2 dB over the LE and about 1 dB over the DFE. If we can use the rule of thumb that every 1-db gain corresponds to an order of magnitude reduction in error rate, then the LC would have an error rate two orders of magnitude less than the LE and one order less than the DFE. The results for Channel 3, again, show the improvement over both LE and DFE. We see that the LE performance degrades significantly when the channel has considerable slope distortion, and on the average, it suffers losses of about 4 dB from the input s/n. The LC shows improvements of about 3 db over the LE and 1 dB over the DFE in all three cases.

We conclude this section by comparing the scatter plots of the receiver outputs after processing by an LE and an LC. In this example, Channel 3 is used and the s/n$_i$ is 24 dB. Figure 10 is the result of linear

Table I—Comparison of linear equalizer, decision feedback equalizer, and canceller performances

	Channel 2			Channel 3		
s/n$_i$ (dB)	20	24	30	20	24	30
Linear equalizer s/n$_o$	18.0	22.1	26.8	15.4	20.4	25.5
Decision feedback equalizer s/n$_o$	19.1	23.0	27.9	17.3	22.0	27.1
Canceller s/n$_o$	19.7	23.8	29.2	18.5	23.1	28.7

equalization. There is a considerable point spread because of severe amplitude distortion. Figure 11 shows the result of cancellation, which considerably tightens the spread and provides a larger noise margin than equalization.

VII. CONCLUSION

Cancellation is a powerful alternative approach to linear equalization that strives to mitigate the effects of slope distortion. This technique is based on cancelling the sidelobes of the overall channel-impulse response, unlike equalization, which attempts to invert the overall channel.

Simulation results of QAM transmission at 9.6 kb/s have shown that the canceller performs impressively, especially for severely amplitude-

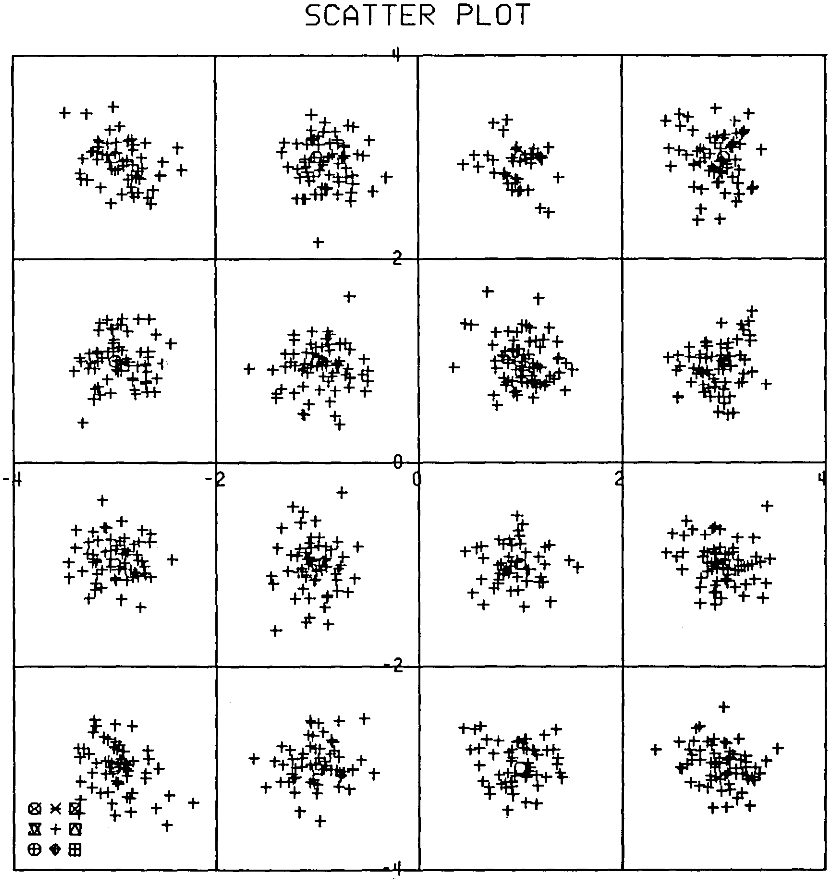


Fig. 10—Linear equalizer output signal constellation.

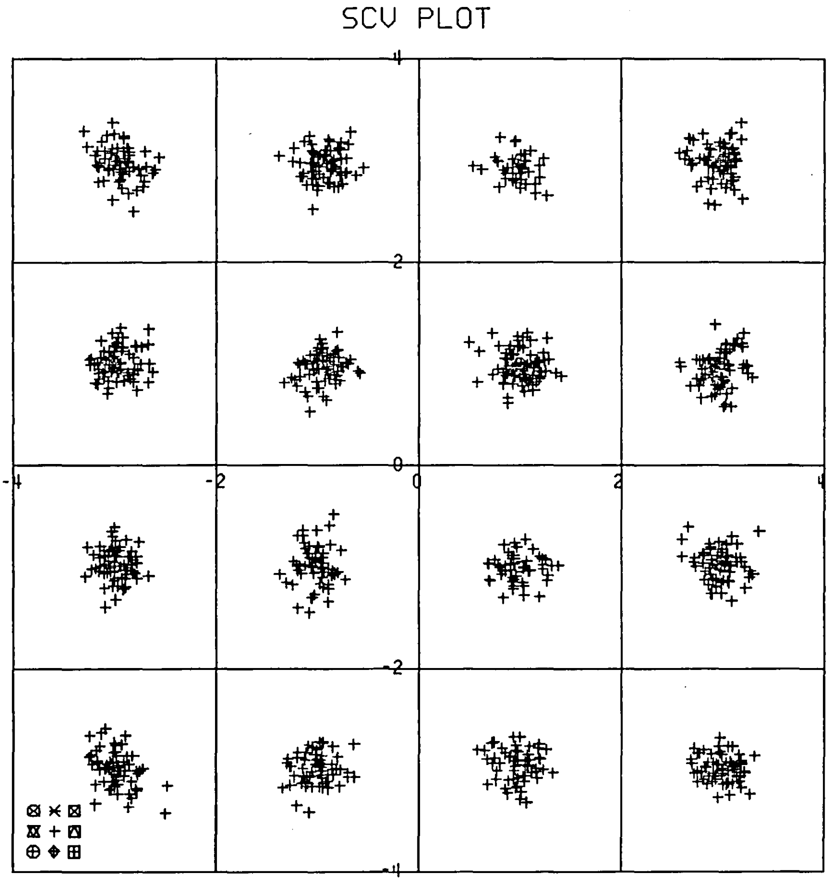


Fig. 11—Linear canceller output signal constellation.

distorted channels, where neither the LE nor the DFE is satisfactory. The LC needs an LE to provide it with reasonably good tentative decisions to perform the cancellation. It is then able to provide a final output which has a lower mse than the tentative output. In addition, the canceller is insensitive to phase distortion, provided the matching filter has an input sampling rate at least twice the data symbol rate. Hence, the LC accommodates the task of a fractionally spaced LE as well.

In summary, we have introduced a fundamentally new approach that performs better than either the linear or decision-feedback equalization, as confirmed by simulation results. Moreover, a recent theoretical study[11] applicable to this cancellation method further confirmed the effectiveness of our approach.

VIII. ACKNOWLEDGMENT

We gratefully acknowledge valuable discussions with R. D. Gitlin, E. Y. Ho, and V. B. Lawrence.

APPENDIX

We derive the optimum minimum mse LC filters $\underline{C}$ and $\underline{W}$ under the assumption of (*i*) uncorrelated noise samples, and (*ii*) correlated noise samples.

Case 1. Uncorrelated noise samples

In Case 1

$$E[\alpha(lT/2)\alpha^*(kT/2)] = \sigma^2\delta_{lk},$$

where σ^2 is the noise variance. Consequently, we have, from eqs. (10), (11), and (15),

$$E(U_n^* A_{n-m}) = \sum_k W_k \exp(j\omega_0 kT/2) \quad h^*[(2m-k)T/2], \tag{51}$$

$$E(V_n^* A_{n-m}) = C_m, \qquad m \neq 0, \tag{52}$$

$$\begin{aligned} E(y_{2n-k}^* y_{2n-m}) &= \exp[j\omega_0(k-m)T/2] \sum_p \sum_q E\{A_p^* h^*[(2n-k)T/2 \\ &\qquad -pT] + \alpha_1^*[(2n-k)T/2]\} \\ &\qquad \times \{A_q h[(2n-m)T/2 - qT] \\ &\qquad + \alpha_1[(2n-m)T/2]\} \\ &= \exp[j\omega_0(k-m)T/2] \sum_p h^*(pT - kT/2) \\ &\qquad \times h(pT - mT/2) + \sigma^2\delta_{km} \\ &= R_h^k(k-m) \qquad + \sigma^2\delta_{km}, \end{aligned} \tag{53}$$

where

$$R_h^l(\tau) = \begin{cases} \exp(j\omega_0\tau) \sum_k h^*(kT + T/2)h(kT + T/2 + \tau), & l \text{ odd} \\ \\ \exp(j\omega_0\tau) \sum_k h^*(kT)h(kT + \tau) & 1 \text{ even} \end{cases} \tag{54}$$

and we also define another term,

$$R_h(l) = \sum_j h^*(jT/2)h(jT/2 + lT/2), \tag{55}$$

which will be used later. Equation (55) is the autocorrelation function of the $T/2$-sampled impulse response $h(jT/2)$, whereas eq. (54) is T-spaced. So, from eq. (11)

$$\begin{aligned} E(U_n^*\, y_{2n-m}) &= \exp(j\omega_0 nT) \sum_k W_k E(y_{2n-k}^* y_{2n-m}) \\ &= \left[\sum_k W_k R_h^k(k-m) + \sigma^2 W_m\right] \exp(j\omega_0 nT). \end{aligned} \tag{56}$$

Also,

$$E(A^*_{n-k}y_{2n-m}) = \exp[j\omega_0(2n - m)T/2] \sum_p E(A^*_{n-k}A_p)$$

$$\times\, h[(2n - m)T/2 - pT]$$

$$= \exp[j\omega_0(2n - m)T/2]h[(2k - m)T/2]. \qquad (57)$$

Therefore,

$$E(V^*_n y_{2n-m}) = \exp[j\omega_0(2n - m)T/2] \sum_{i\neq 0} C_i h[(2i - m)T/2], \qquad (58)$$

and

$$E(A^*_k y_{2n-m}) = \exp[j\omega_0(2n - m)T/2]h(-mT/2). \qquad (59)$$

By substituting eqs. (51) to (59) into eqs. (17) and (18), we have

$$C_m = \sum_k W_k \exp(j\omega_0 kT/2)h^*[(2m - k)T/2], \qquad m \neq 0, \qquad (60)$$

and

$$\sum_k W_k R^k_h(k - m) + \sigma^2_n W_m = \exp(-j\omega_0 mT/2)$$

$$\times \sum_{i=0} C_i h[(2i - m)T/2] + \exp(-j\omega_0 mT/2)h(-mT/2), \quad \text{all } m. \qquad (61)$$

We next solve for $\{C_i\}$ and $\{W_i\}$ from this pair of equations. Substituting eq. (60) into eq. (61) gives

$$\sum_k W_k R^k_h(k - m) + \sigma^2_n W_m = \exp(-j\omega_0 mT/2) \sum_{i\neq 0} \sum_k W_k$$

$$\times \exp(j\omega_0 kT/2)h^*[(2i - k)T/2]$$

$$\times\, h[(2i - m)T/2]$$

$$+ \exp(-j\omega_0 mT/2)h(-mT/2)$$

$$= \sum_k W_k[\exp j\omega_0(k - m)T/2]$$

$$\times \sum_i h^*[(2i - k)T/2]$$

$$\times\, h[(2i - m)T/2] + \exp(-j\omega_0 mT/2)$$

$$\times\, h(-mT/2)[1 - \sum_k W_k \exp$$

$$\times\, (j\omega_0 kT/2)h^*(-kT/2)]$$

$$= \sum_k W_k R^k_h(k - m) + \exp(-j\omega_0 mT/2)$$

$$\times\, h(-mT/2)(1 - \beta), \qquad (62)$$

using eq. (55), and where

$$\beta = \sum_k W_k \exp(j\omega_0 kT/2) h^*(-kT/2). \tag{63}$$

Therefore, we obtain the equality

$$\sigma^2 W_m = \exp(-j\omega_0 T/2) h(-mT/2)(1-\beta), \qquad \text{all } m, \tag{64}$$

or

$$W_m = \exp(-j\omega_0 mT/2) h(-mT/2) \frac{(1-\beta)}{\sigma^2}. \tag{65}$$

On substituting eq. (65) into eq. (63), we can show that

$$\beta = \frac{E_h}{E_h + \sigma^2}, \tag{66}$$

where, using eq. (55),

$$\begin{aligned} E_h &= \sum_j \left| h\left(j\frac{T}{2}\right)\right|^2 \\ &= R_h(0). \end{aligned} \tag{67}$$

From eqs. (66) and (65), we see that the matching filter has $T/2$-spaced tap weights

$$W_m = \exp(-j\omega_0 mT/2) \frac{h(-mT/2)}{E_h + \sigma^2}, \qquad \text{all } m, \tag{68}$$

which is clearly proportional to a matched-filter impulse response.

Case 2. Correlated noise samples

As described in Section IV, Case 2 corresponds to the voiceband telephone channel where the noise has approximately the same bandwidth as the signal so that noise samples at $T/2$ spacing are correlated.

With noise correlation

$$E[(\alpha kT/2)\alpha^*(lT/2)] = R_n(k-l),$$

and eq. (53) becomes

$$E(y^*_{2n-k} y_{2n-m}) = R^k_h(k-m) + R_n(k-m). \tag{69}$$

Then, in place of eq. (56), we have

$$E(U^*_n \, y_{2n-m}) = \sum_k W_k[R^k_h(k-m) + R_n(k-m)]. \tag{70}$$

Consequently, eq. (62) becomes

$$\sum_k W_k R_n(k-m) = \exp(-j\omega_0 mT/2) h(-mT/2)(1-\beta), \qquad \text{all } m. \tag{71}$$

We have to solve a set of linear equations for the $\underline{W}$ taps, and the solution in the time domain is not obvious. Instead, we examine the results in the frequency domain. Transforming both sides of eq. (70) gives

$$W(\omega)S_n(-\omega) = (1-\beta)\tilde{H}(-\omega-\omega_0);$$

that is,

$$W(\omega) = (1-\beta)\tilde{H}(-\omega-\omega_0)/S_n(-\omega), \tag{72}$$

where $W(\omega)$ is the Fourier transform of the $\underline{W}$ tap weights, $S_n(\omega)$ is the sampled noise spectrum, and $\tilde{H}(\omega)$ is the Fourier transform of the channel-sampled impulse response. Again, $W(\omega)$ takes the form of a filter matched to the channel with additive band-limited noise.

The constant β can be written as

$$\beta = \frac{T}{4\pi}\int_{\frac{-2\pi}{T}}^{\frac{2\pi}{T}} \tilde{H}^*(-\omega-\omega_0)W(\omega)d\omega \tag{73}$$

On substituting eq. (72) into eq. (73), we can solve for β as

$$\beta = \frac{\xi}{1+\xi},$$

where

$$\xi = \frac{T}{4\pi}\int_{\frac{-2\pi}{T}}^{\frac{2\pi}{T}} \frac{|\tilde{H}(-\omega-\omega_0)|^2}{S_n(-\omega)}d\omega. \tag{74}$$

Consequently, from eq. (72),

$$W(\omega) = \frac{\tilde{H}(-\omega-\omega_0)}{(1+\xi)S_n(-\omega)}. \tag{75}$$

Even though the $\underline{C}$ and $\underline{W}$ filters are jointly adapting, the $\underline{C}$ taps are, in fact, slaved to the $\underline{W}$ taps. On substituting eq. (68) into eq. (60), we obtain, using eq. (55),

$$\begin{aligned} C_m &= \frac{1}{(E_h+\sigma^2)}\sum_k h(-kT/2)h^*[(2m-k)T/2] \\ &= \frac{1}{(E_h+\sigma^2)}R_h(2m), \qquad m \neq 0. \end{aligned} \tag{76}$$

Thus, the canceller-impulse response (for $m \neq 0$) is that of the overall T-spaced impulse response of the channel and matching filter; that is, the canceller recreates the entire ISI component present in the matching filter output signal as long as the correct data symbols are applied to the canceller.

The C taps corresponding to eq. (75) can similarly be obtained. From eq. (71), we have

$$h[(2m-k)T/2]$$
$$=\frac{1}{(1-\beta)}\exp[j\omega_0(k-2m)T/2]\sum_l W_l R_n(l-k+2m). \quad (77)$$

On substituting into eq. (60), we have

$$C_m=\frac{\exp(j\omega_0 mT)}{(1-\beta)}\sum_{k,l} W_k W_l^* R_n^*(2m+l-k), \qquad m\neq 0. \quad (78)$$

If, in addition, we define

$$C_0=\frac{1}{(1-\beta)}\sum_{k,l} W_k W_l^* R_n^*(l-k),$$

we can transform the sequence $\{C_m\}$ to obtain

$$\begin{aligned} C(\omega) &= \sum_m C_m \exp(-j\omega mT) \\ &= \frac{1}{(1-\beta)}\sum_k W_k \exp[-j(\omega-\omega_0)kT/2] \\ &\quad \times \sum_l W_l^* \exp[j(\omega-\omega_0)lT/2] \\ &\quad \times \sum_m R_n^*(2m+l-k) \\ &\quad \times \exp[-j(\omega-\omega_0)(2m+l-k)T/2] \end{aligned} \quad (79)$$

Since both indices k and l run from $-\infty$ to ∞, we can replace the summation over m as

$$\begin{aligned} \sum_p R_n^*(p)\exp[-j(\omega-\omega_0)pT/2] &= S_n^*(-\omega+\omega_0) \\ &= S_n(-\omega+\omega_0). \end{aligned}$$

Therefore, eq. (79) can be simplified to

$$C(\omega)=\frac{1}{(1-\beta)}\,|W(\omega-\omega_0)|^2 S_n(-\omega+\omega_0). \quad (80)$$

Substituting for $W(\omega)$ from eq. (72) gives

$$C(\omega)=\frac{(1-\beta)\,|\tilde{H}(-\omega)|^2}{S_n(-\omega+\omega_0)}, \quad (81)$$

or, from eq. (74),

$$C(\omega)=\frac{1}{1+\xi}\,\frac{|\tilde{H}(-\omega)|^2}{S_n(-\omega+\omega_0)}. \quad (82)$$

REFERENCES

1. A. Gersho, "Adaptive Equalization of Highly Dispersive Channels for Data Transmission," B.S.T.J., *48,* No. 1 (January 1969), pp. 55–70.
2. G. Ungerboeck, "Fractional Tap-Spacing Equalizer and Consequences for Clock Recovery in Data Modems," IEEE Trans. Communication, *COM-24,* No. 8 (August 1976), pp. 856–64.
3. R. D. Gitlin and S. B. Weinstein, "Fractionally-Spaced Equalization: An Improved Digital Transversal Equalizer," B.S.T.J., *60,* No. 2 (February 1981), pp. 275–96.
4. M. E. Austin, "Decision-Feedback Equalization for Digital Communication over Dispersive Channels," Lincoln Laboratory Report No. 437 (1967).
5. J. Salz, "Optimum Mean-Square Decision Feedback Equalization," B.S.T.J., *52,* No. 8 (October 1973), pp. 1341–73.
6. G. D. Forney, "The Viterbi Algorithm," Proc. IEEE, *61* (March 1973), pp. 268–78.
7. S. B. Weinstein, "Echo Cancellation in the Telephone Network," IEEE Communication Magazine, *15* (January 1977), pp. 9–15.
8. D. Hirsch, private communication.
9. J. Proakis, "Adaptive Nonlinear Filtering Techniques for Data Transmission," IEEE Symposium on Adaptive Processes, Decision and Control (1970), pp. XV.2.1-5.
10. G. Ungerboeck, "Theory on the Speed of Convergence on Adaptive Equalizers for Digital Communication," IBM J. Res. and Dev., No. 6 (November 1972), pp. 546–55.
11. M. S. Mueller and J. Salz, "A Unified Theory of Data-Aided Equalization," B.S.T.J., this issue.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

A Unified Theory of Data-Aided Equalization

By M. S. MUELLER and J. SALZ

(Manuscript received March 17, 1981)

A unified theory is presented for data-aided equalization of digital data signals passed through noisy linear dispersive channels. The theory assumes that some past and/or future transmitted data symbols are perfectly detected. We use this hypothesis to derive the minimum mean-square error receiver. The optimum structure consists of a matched filter in cascade with a transversal filter combined with a linear intersymbol interference canceler which uses the ideally detected data symbols. The main result is an expression for the optimized mean-square error as a function of the number and location of the canceler coefficients, the s/n, and the channel transfer function. When the number of canceler coefficients is zero, we get the well-known result for linear equalization. When the causal or postcursor canceler approaches infinite length, we obtain the well-known decision feedback result. When both the precursor and postcursor cancelers become infinite, we obtain the very best result possible, namely, the matched-filter bound dictated from fundamental theoretical considerations. Neither the decision feedback nor the matched-filter results can be achieved in practice since their implementation requires infinite memory and storage. Our theory can be used to calculate the rate of approach to these ideals with finite cancelers.

I. INTRODUCTION

The theory of linear and decision feedback equalization to mitigate the effects of intersymbol interference (ISI) and noise in digital data transmission is well known.[1-4] In this paper, the problem of equalization is cast in a general framework of an ISI canceler aided by past and/or future data values. This general structure is suggested from optimal detection theory and is shown in Fig. 1. The optimal detector of digital data in the presence of additive Gaussian noise and ISI is comprised of a matched filter and an ISI estimator which is used to cancel the

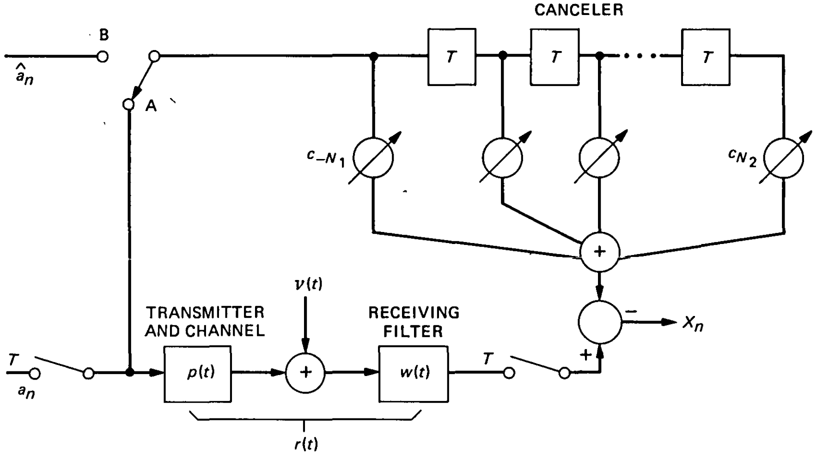


Fig. 1—Block diagram of data-aided equalization.

interference.[5,6] The implementation of this structure is often impractical because of its complexity.[7,8]

In our theory we postulate that some portion of the ISI can be perfectly synthesized and, therefore, subtracted from the incoming signal. In other words, we replace the optimal estimator with a practical one. The effect of the remaining interference is then minimized by a linear filter or a conventional linear equalizer. In practical systems, however, perfect estimation cannot be achieved; therefore, our results serve as ideal limits. The inclusion of occasional errors in our theory has proved mathematically intractable so far.

In Section II, we determine the minimal mean-square error (mse) when an arbitrary set of data symbols is known to the receiver. In Section III, the optimal receiving filter is derived and analyzed. The performance of the infinite linear equalizer, the decision feedback equalizer and the infinite canceler are obtained as special cases of the general result. Section IV covers a discussion of numerical results.

II. MINIMUM MSE FOR DATA-AIDED EQUALIZATION

In Fig. 1, the transmitter generates the data sequence $\{a_n\}$ whose elements are assumed to be independent identically distributed (i.i.d.) discrete random variables. These discrete amplitudes sequentially modulate the pulse $p(t)$ at a rate $1/T$ to produce the transmitted signal. The pulse shape, $p(t)$, can be viewed as the overall impulse response of the transmitting filter and the transmission channel. White noise, $\nu(t)$, is added to the received signal which is then applied to the linear receiving filter, $w(t)$. The output signal is sampled at the symbol rate $1/T$ and combined with the output of the canceler. The linear canceler is modeled as a transversal filter with coefficients, $\{c_n\}$, where $n \in S$, and where the set of integers S denotes the range of the canceler's taps.

For most applications, and for the special cases investigated later in this Section, the range will contain the neighboring taps of the reference location but never the reference location itself, and consequently, $S = \{-N_1, \cdots, -1, 1, \cdots, N_2\}$. The canceler operates on the past received symbols $a_{n-1}, \cdots, a_{n-N_2}$ and on the future received symbols $a_{n+1}, \cdots, a_{n+N_1}$, which are assumed to be known to the receiver. Clearly, to realize an operation on the future data symbols, a time delay of at least N_1T seconds has to be introduced.

For a general set S the output signal $x_n = x(nT)$ can, thus, be written as

$$x_n = \sum_{k=-\infty}^{\infty} r_k a_{n-k} - \sum_{k\in S} c_k a_{n-k} + \xi_n . \tag{1}$$

Where $r_k = r(kT)$ is the overall impulse response evaluated at $t = kT$,

$$r_k = T \int_{-\infty}^{\infty} w(\tau)p(kT - \tau)d\tau , \tag{2}$$

and where $\xi_k = \xi(kT)$, i.e.

$$\xi_k = \int_{-\infty}^{\infty} w(\tau)\nu(kT - \tau)d\tau, \quad \text{for} \quad k = -\infty, \cdots, -1, 0, 1, \cdots, \infty. \tag{3}$$

To facilitate modeling of various types of linear modulation schemes, the data sequence, the noise, and all impulse responses are assumed to be complex valued. In general, $p(t)$ will be the preenvelope of the passband transmission system with respect to a carrier frequency. This notation has become extremely useful and economical in this field.[9] Specifically, it permits a unified presentation of baseband and passband systems.

The output signal, x_n, after slicing or quantizing is usually taken to be an estimate of the transmitted data symbol a_n. Our goal now is to obtain a receiving filter, $w(t)$, and canceler taps, $\{c_n\}$, so that the mse,

$$\epsilon = E\{|x_n - a_n|^2\} , \tag{4}$$

is a minimum. To determine the optimal canceller coefficients, $\{c_n\}$, we differentiate eq. (4) with respect to c_n, $n \in S$, and set the result to zero

$$\frac{\partial \epsilon}{\partial c_n} = \sigma_a^2[c_n^* - r_n^* + c_n - r_n] = 0, \quad \text{for} \quad n \in S , \tag{5}$$

where

$$E\{a_n a_k^*\} = \sigma_a^2 \delta_{n,k} , \tag{6}$$

and $\delta_{n,k}$ is the Kronecker delta. The immediate conclusion from eq. (5) is that for $n \in S$

$$c_n = r_n . \tag{7}$$

Inserting this into eq. (1), we get

$$x_n = \sum_{k \notin S} r_k a_{n-k} + \int_{-\infty}^{\infty} w(\tau)\nu(nT - \tau)d\tau . \tag{8}$$

Thus far, our approach is perfectly obvious. By knowing the data symbols for all integers $k \in S$, it is possible to synthesize the resulting ISI associated with these symbols and subtract it from the current signal sample x_n. If the set S contains all integers $k < n$, we use all the already-decided-upon data symbols (available at the receiver without delay) to synthesize the postcursor ISI. This is precisely what is done in decision feedback equalization. If the set S contains all the integers, except the one associated with the present instant, n, all ISI is eliminated. But this, of course, requires infinite delay. In practice, the set S will be finite and our main concern will be to determine how it influences the mse.

We now proceed to optimize the receiving filter, $w(t)$, for a given set S. Inserting eq. (8) into eq. (4) and using eq. (6), the resulting mse can be expressed as

$$\epsilon = \sigma_a^2 \left[\sum_{k \notin S} |r_k|^2 - r_0 - r_0^* + 1 + \sigma^2 \right], \tag{9}$$

where

$$E\{\nu(t)\nu(t+\tau)^*\} = \sigma_\nu^2 T \delta(\tau) , \tag{10}$$

and where

$$\sigma^2 = T \frac{\sigma_\nu^2}{\sigma_a^2} \int_{-\infty}^{\infty} |w(t)|^2 dt . \tag{11}$$

We remark that more general noise covariances can be included, but the calculations become more cumbersome without yielding additional insights.

To obtain the optimum $w(t)$, let

$$w(t) = w_0(t) + \lambda\mu(t) \tag{12}$$

and define

$$U_k = T \int_{-\infty}^{\infty} w_0(\tau) p(kT - \tau) d\tau , \tag{13}$$

where $w_0(t)$ is the optimum impulse response of the receiving filter and where the U_k are the samples of the optimized overall impulse response. It follows immediately from eq. (2) that

$$r_k = U_k + \lambda T \int_{-\infty}^{\infty} \mu(\tau) p(kT - \tau) d\tau . \qquad (14)$$

When

$$\left. \frac{\partial \epsilon}{\partial \lambda} \right|_{\lambda=0} = 0 \qquad (15)$$

is calculated from eq. (9), we obtain an equation for the optimum receiving filter:

$$w_0(t) N_0 = p(-t)^* - \sum_{k \notin S} U_k p(kT - t)^* , \qquad (16)$$

where

$$N_0 = \frac{\sigma_\nu^2}{\sigma_a^2} . \qquad (17)$$

The interpretation of eq. (16) is standard: the optimum receiving filter is comprised of a matched filter $p(-t)^*$ in cascade with a transversal filter having taps only at those locations where the canceler has none. This structure is shown in Fig. 2.

To obtain our central result, an expression for the optimal mse, we multiply eq. (16) by $Tw_0(t)^*$ and integrate from $-\infty$ to $+\infty$. This yields with the aid of eqs. (9), (11), and (13) the result

$$\epsilon_{\text{opt}} = \sigma_a^2 (1 - U_0) . \qquad (18)$$

The explicit determination of U_0 is the subject of the next section.

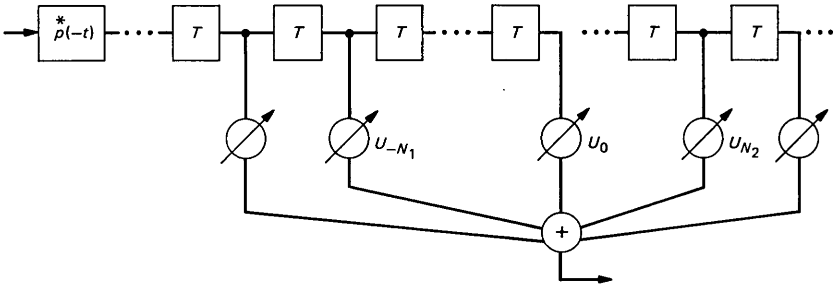


Fig. 2—Optimal receiving filter, $w_0(t)$.

III. THE OPTIMAL FILTER

To determine the sample values, $\{U_m\}$, of the optimal overall impulse response, define the autocorrelation function of the channel impulse response as

$$R_m = T \int_{-\infty}^{\infty} p(mT - \tau) p(-\tau)^* d\tau \tag{19}$$

and its Fourier transform as

$$R(\omega) = \sum_{m=-\infty}^{\infty} R_m e^{-jm\omega T}. \tag{20}$$

After multiplying eq. (16) by $Tp(nT - \tau)$ and integrating from $-\infty$ to $+\infty$, we get the following system of linear equations in $\{U_m\}$,

$$U_m N_0 = R_m - \sum_{k \notin S} U_k R_{m-k} \quad \text{for} \quad \text{all}\, m. \tag{21}$$

To determine the optimal receiving filter, we only need to know U_m for $m \notin S$. From eq. (21) we extract the equations necessary to determine U_0 and partition them as follows

$$U_0(N_0 + R_0) = R_0 - \sum_{k \notin J} U_k R_{-k} \quad \text{for} \quad m = 0\,, \tag{22}$$

$$\sum_{k \notin J} U_k M_{m-k} = (1 - U_0) R_m \quad \text{for} \quad m \notin J\,, \tag{23}$$

where we defined

$$M_k = R_k + N_0 \delta_{k,0}\,, \tag{24}$$

and where J is the set S augmented by $m = 0$.

Note that the indices of the unknowns and the indices of the right-hand sides of eq. (23) have gaps of the same size and at the same locations. (See the definition of J above.) Thus, the set of equations in (23) is not in a standard form and the solution technique is not obvious. In Appendix A we develop a technique to solve this infinite set of equations with finite gaps. It involves the solution of a special infinite set of equations without gaps. To compensate for the gaps, we augment the original set of equations. Specifically, we add for $m \in J$ a finite number of equations to the infinite set such that the solution vanishes for $m \in J$. From Appendix A, we determine that the optimum mse becomes

$$\epsilon_{\text{opt}} = \sigma_a^2 \frac{N_0}{N_0 + H_0}\,, \tag{25}$$

where H_0 is determined from the following set of equations:

$$\sum_{m \in J} M_{k-m}^{-1} H_m = \delta_{k,0} - N_0 M_k^{-}, \quad \text{for} \quad k \in J \tag{26}$$

and where M_k^{-1} is the inverted sequence of M_k, i.e., it satisfies

$$\sum_n M_n M_{k-n}^{-1} = \delta_{k,0} \quad \text{for} \quad \text{all } k\,. \tag{27}$$

In the following section, we investigate the minimal mse for some special cases. As mentioned initially, we use the realistic assumption that the set S contains the neighboring locations $\{-N_1, \cdots, -1, 1, \cdots, N_2\}$. Then $J = \{-N_1, \cdots, N_2\}$ and the coefficient matrix in eq. (26) is a finite Toeplitz matrix. The solution of eq. (26) and, thus, H_0 is unique and is guaranteed to exist when $R(\omega) + N_0$ is bounded away from zero and infinity.[10] These conditions are very mild and are satisfied in most cases of practical interest.

3.1 Infinite length equalizer

For $N_1 = N_2 = 0$ the set J includes only the zero integer and all canceler coefficients vanish. Consequently, eq. (26) degenerates to a single equation

$$H_0 M_0^{-1} = 1 - N_0 M_0^{-1}. \tag{28}$$

Solving this for H_0 and inserting it into eq. (25), we obtain the standard result for the optimum linear equalizer,[2]

$$\epsilon_{\text{opt}} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{N_0 d\omega}{R(\omega) + N_0}, \tag{29}$$

where we expressed M_0^{-1} in terms of its Fourier transform

$$M_0^{-1} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{d\omega}{R(\omega) + N_0}. \tag{30}$$

3.2 Matched-filter

When $N_1 = N_2 = \infty$ the set J is infinite; i.e., the canceler subtracts all the ISI. Equation (26) in this case yields

$$\sum_{m=-\infty}^{\infty} M_{k-m}^{-1}(H_m + N_0 \delta_{m,0}) = \delta_{k,0} \quad \text{for} \quad \text{all } k\,. \tag{31}$$

Comparing this with eq. (27) gives the result

$$M_m = H_m + N_0 \delta_{m,0}\,. \tag{32}$$

From eqs.(32) and (23) it follows that $H_m = R_m$ for all m. Thus, the optimum mse for this case is

$$\epsilon_{\text{opt}} = \sigma_a^2 \frac{N_0}{N_0 + R_0}. \tag{33}$$

This is recognized as the matched-filter bound for the optimal detection of a known signal in noise and in the absence of ISI.

3.3 One-sided canceler of infinite length

For $N_1 = 0$ and $N_2 = \infty$ the canceler performs as an ideal decision feedback equalizer of infinite length. In Appendix B the mse is derived for the more general case $N_1 \neq 0$, i.e. for a decision feedback equalizer with a limited number of noncausal taps. The result is

$$\epsilon_{\text{opt}} = \sigma_a^2 \frac{N_0}{\sum_{k=0}^{N_1} |M_k^+|^2}, \tag{34}$$

where the coefficientsM_k^+ are determined from the following equation

$$\sum_{k=0}^{\infty} M_k^+ M_{m-k}^- = M_m \quad \text{for} \quad \text{all}\, m. \tag{35}$$

Here, $\{M_k^+\}$ is the causal "root" of the two sided sequence $\{M_k\}$. It satisfies $M_k^+ = 0$ for $k < 0$ and $M_k^+ = (M_{-k}^-)^*$ for $k \geq 0$. It is shown in Ref. 3 that

$$|M_0^+|^2 = N_0 \exp\left[\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \ln\left(\frac{R(\omega)}{N_0} + 1\right) d\omega\right], \tag{36}$$

and when this formula is inserted into eq. (34) we get the well-known result for the decision feedback equalizer,

$$\epsilon_{\text{opt}} = \sigma_a^2 \exp\left[-\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \ln\left(\frac{R(\omega)}{N_0} + 1\right) d\omega\right]. \tag{37}$$

Unfortunately, there is no similar simple expression for $|M_k^+|^2$, for $k \neq 0$; therefore, we are forced to numerically factor the two-sided sequence $\{M_k\}$ into its causal and anticausal root.

IV. DISCUSSION OF NUMERICAL RESULTS

In this section, the minimal mse of data-aided equalization is evaluated numerically for certain channels and for various sets of canceler taps. We will exhibit and discuss the behavior of the mse, $\epsilon_{\text{opt}}(N_1, N_2)$, as a function of N_1 and N_2 for typical telephone channels. As a point of reference, note the following easily proved inequalities:

$$\epsilon_{\text{opt}}(0, 0) \geq \epsilon_{\text{opt}}(0, \infty) \geq \epsilon_{\text{opt}}(\infty, \infty).$$

In the following, we examine three types of cancelers:

(i) Starting from the infinite length linear equalizer whose mse is $\epsilon_{\text{opt}}(0, 0)$, we increase the number of causal canceler taps; i.e., we examine $\epsilon_{\text{opt}}(0, N_2)$ for $N_2 = Q = 0, \cdots, 15$.

(ii) Starting from the infinite length linear equalizer, we increase the number of known data symbols alternating between causal and noncausal ones; i.e., we examine $\epsilon_{\text{opt}}(N_1, N_2)$ for $Q = N_1 + N_2 = 0, \cdots, 15$ and where $N_1 = N_2$ for Q even, $N_2 = N_1 + 1$ for Q odd.

(iii) Starting from the infinite length decision feedback equalizer whose mse is $\epsilon_{\text{opt}}(0, \infty)$, we examine the behavior when noncausal taps are added; i.e., $\epsilon_{\text{opt}}(N_1, \infty)$ for $N_1 = Q = 0, \cdots, 15$.

Equation (25) is used to determine the mse for cases (i) and (ii), where H_0 is obtained as the solution of eq. (26). To determine the sequence $\{M_k^{-1}\}$ for all k, we observe that the Fourier transform of the sequence $\{R_k\}$ is related to the overall transfer function of the channel as follows:

$$R(\omega) = \sum_{k=-\infty}^{\infty} \left| P\left(\omega - k\frac{2\pi}{T}\right)\right|^2 . \tag{38}$$

Clearly, $R(\omega)$ is periodic with period $2\pi/T$, and it is only dependent on the magnitude of the overall channel transfer function, $P(\omega)$. Therefore, phase distortion in the channel has no effect on the mse. This is reflected in the well-known fact that phase distortion can be perfectly equalized without noise enhancement. Therefore, the sequence, $\{M_k^{-1}\}$, is obtained as follows

$$M_k^{-1} = \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{e^{jk\omega T}d\omega}{R(\omega) + N_0}, \tag{39}$$

where $j = \sqrt{-1}$. Fast Fourier transform techniques are used to evaluate eq. (39). Numerical tests show that it suffices to take 64–128 samples of $R(\omega) + N_0$ in the interval $[-\pi/T, \pi/T]$. The fact that the coefficient matrix in eq. (26) is positive definite and Toeplitz, makes it possible to obtain the desired solution, H_0, recursively. This is done with the Levinson algorithm.[11–14]

For case (iii), we evaluate eq. (34). The sequence, $\{M_k^+\}$, is obtained from the following approach.[3] First determine $\{F_k\}$ for all k such that

$$\ln(M(\omega)) = \sum_{k=-\infty}^{\infty} F_k e^{-jk\omega T}. \tag{40}$$

Then it follows that

$$M(\omega)^+ = \exp\left\{\sum_{k=0}^{\infty} F_k e^{-jk\omega T}\right\}, \tag{41}$$

and

$$M_k^+ = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} M(\omega)^+ e^{jk\omega T} d\omega . \tag{42}$$

Fast Fourier transforms are used to obtain consecutively $\{F_k\}$, $M(\omega)$, and M_k^+. The overall channel power transfer function, $|P(\omega)|^2$, is assumed to consist of a raised cosine shaped transmitting filter with relative excess bandwidth, $\alpha = 0.15$,

$$|T(\omega)|^2 = \begin{cases} 1 & \text{for } |\omega| < (1-\alpha)\pi/T \\ 0.5\left[1 - \sin\left(\omega - \dfrac{\pi}{T}\right)\dfrac{T}{2\alpha}\right] & \text{for } (1-\alpha)\dfrac{\pi}{T} < \omega < (1+\alpha)\dfrac{\pi}{T} \\ 0.5\left[1 + \sin\left(\omega + \dfrac{\pi}{T}\right)\dfrac{T}{2\alpha}\right] & \text{for } -(1+\alpha)\dfrac{\pi}{T} < \omega < -(1-\alpha)\dfrac{\pi}{T} \\ 0 & \text{elsewhere} \end{cases}$$

and cascaded with the channel power transfer function $|G(\omega)|^2$.

Figure 3 shows two different channel power transfer functions, $|G(\omega)|^2$, which are used to derive the subsequent numerical results. In (a) we show the equivalent baseband transfer function for the worst channel meeting the basic conditions of private lines (BASICBAD).[15] Part (b) shows a transfer function (CABLE) with linearly increasing attenuation. The parameters P_1 and P_2 indicate the attenuation at $\omega = -\pi/T$ and $\omega = \pi/T$. A model for a baseband cable channel is obtained when $P_1 = P_2$.

Figure 4 shows the mse as a function of the number of canceler taps for the various channel transfer functions and for s/n of 20 dB at the receiver input. The dotted line represents type (*i*) canceler; the dashed line, type (*ii*); and the solid line, type (*iii*). The curves for types (*i*) and (*ii*) start at the minimal mse for the infinite length equalizer and the curve for type (*iii*) starts at the minimal mse of the infinite decision feedback equalizer.

As can be observed, all curves converge very rapidly to their asymptotes. The curve for type (*i*) indicates that only 3 causal coefficients suffice to closely approximate the performance of an infinite decision feedback equalizer. The curve for type (*ii*) suggests that a total of 6 coefficients (3 causal and 3 anticausal) results in a performance which is very close to the optimal (the matched-filter bound). The curve for type (*iii*) reaches very close to the mse obtained from the matched-filter bound with only 3 noncausal coefficients, in addition to an infinite decision feedback equalizer. These results are virtually independent of the channel involved.

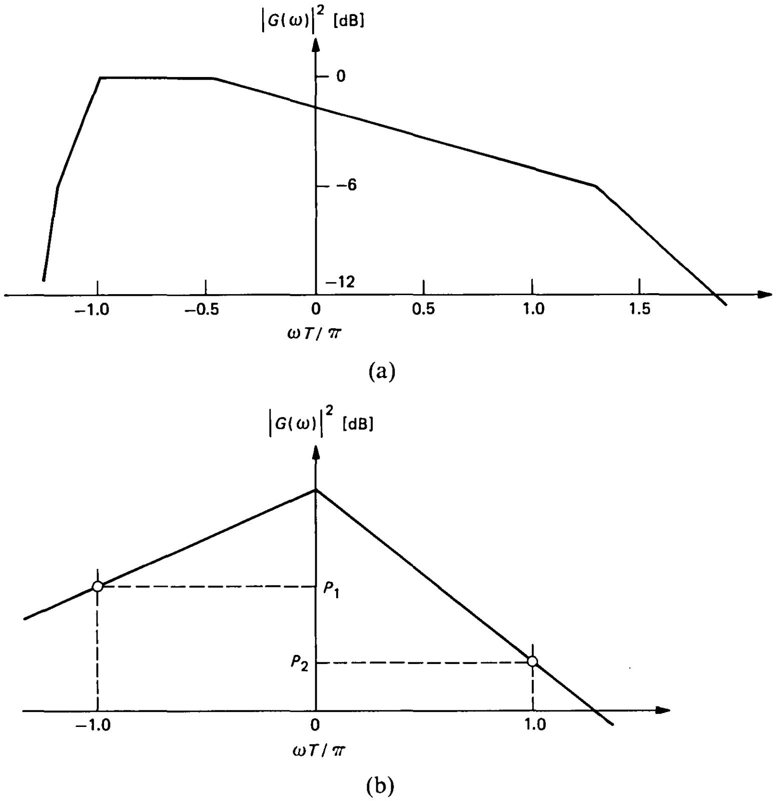


Fig. 3—Power transfer functions. (a) BASICBAD channel; (b) CABLE channel.

The channel, however, influences the best mse which is obtainable with the infinite equalizer, i.e. $\epsilon_{\text{opt}}(0, 0)$, and with the infinite decision feedback equalizer, i.e. $\epsilon_{\text{opt}}(0, \infty)$. Table I shows these figures for the various channels.

The minimal mse obtained from the matched-filter bound is −20.04 dB. Therefore, 1.8 to 4.7 dB can be gained for the channels considered if a canceler of three causal and three noncausal coefficients is included.

APPENDIX A

Solution of an Infinite Set of Equations with Finite Gaps

Let

$$M_k = R_k + N_0\delta_{k,0}\,, \tag{43}$$

and consider

$$\sum_{k \notin J} U_k M_{m-k} = (1 - U_0)R_m \quad \text{for} \quad m \notin J, \tag{44}$$

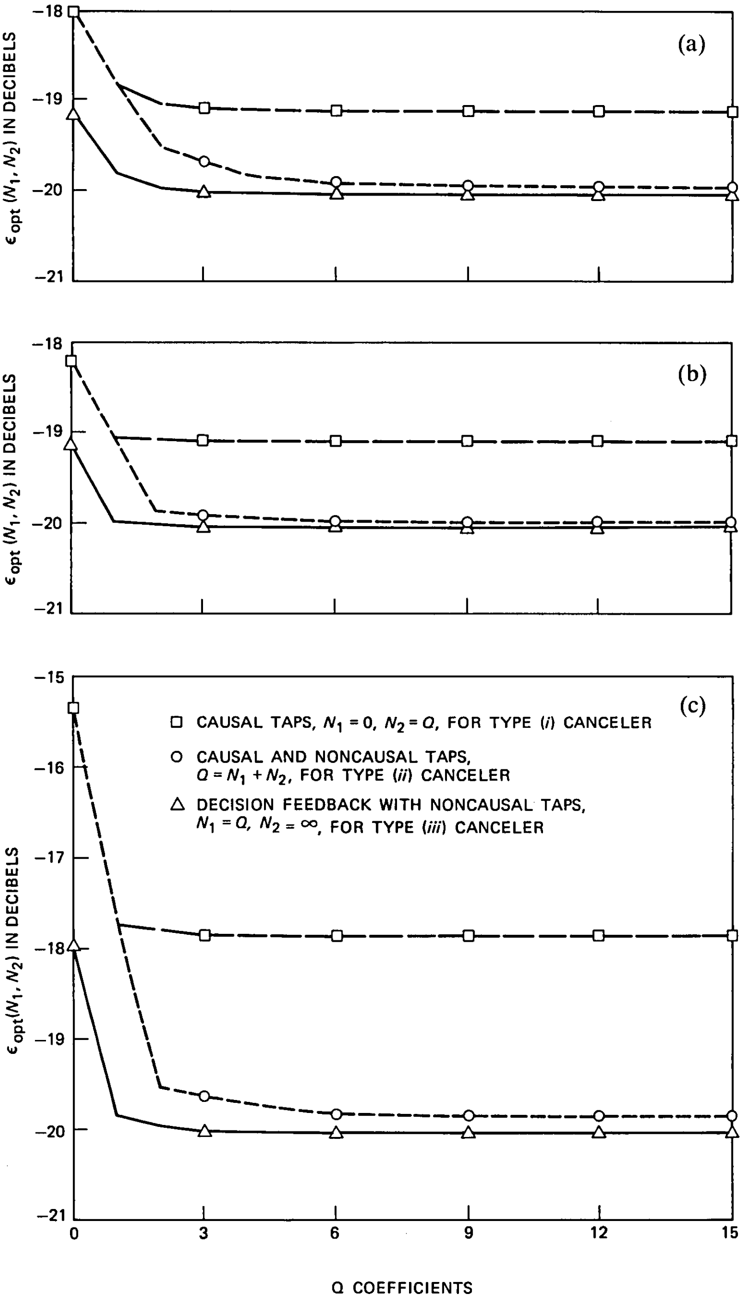


Fig. 4—Mean-square error for data-aided equalizations. (a) BASICBAD channel; (b) CABLE channel, $P_1 = P_2 = -10$ dB; (c) CABLE channel, $P_1 = -10$ dB, $P_2 = -20$ dB.

Table I—Minimal mean-square error

Mean-Square Error Channel	∞ Equalizer ϵ_{opt} (0, 0) in dB	∞ Decision Feedback Equalizer ϵ_{opt} (0, ∞) in dB
BASICBAD	−18.05	−19.16
CABLE		
$P_1 = 10, P_2 = 10$	−18.22	−19.13
$P_1 = 10, P_2 = 20$	−15.35	−17.97

where J is a finite set. It always contains the number zero but is otherwise arbitrary. Equation (44) is an infinite set of equations with a finite gap in both the indices of the unknowns U_k and the right-hand sides R_m. Notice that eq. (44) reduces to a discrete convolution and, therefore, is easy to solve if k and m are allowed to take on all the integers, or if the gap could be removed somehow.

Now consider instead of eq. (44) the following set of equations

$$\sum_{k=-\infty}^{\infty} V_k M_{m-k} = (1 - U_0)(R_m - H_m) \quad \text{for} \quad \text{all}\, m, \tag{45}$$

which is a discrete convolution. In order that eq. (45) conform to eq. (44), the auxiliary sequence $\{H_m\}$ must satisfy

$$H_m = 0 \quad \text{for} \quad m \notin J, \tag{46}$$

and

$$V_m = 0 \quad \text{for} \quad m \in J. \tag{47}$$

To accomplish this, the values of H_m for $m \in J$ are determined such that these constraints are forced to be satisfied. This is always possible since there are $N_1 + N_2 + 1$ free parameters, H_m, and the same number of conditions on V_m. From the above, it follows that

$$U_k = V_k \quad \text{for} \quad k \notin J. \tag{48}$$

This is easily proved by subtracting eq. (45) for $m \notin J$ from eq. (44). Now define the sequence $\{M_k^{-1}\}$ such that

$$\sum_{k=-\infty}^{\infty} M_k M_{m-k}^{-1} = \delta_{m,0}. \tag{49}$$

It can then be shown that

$$V_k = (1 - U_0) \sum_{m=-\infty}^{\infty} (R_m - H_m) M_{k-m}^{-1} \quad \text{for} \quad \text{all}\, k. \tag{50}$$

Since $V_k = 0$ for $k \in J$, we conclude from eq. (50) that

$$\sum_{m=-\infty}^{\infty} H_m M^{-1}_{k-m} = \sum_{m=-\infty}^{\infty} R_m M^{-1}_{k-m} \quad \text{for} \quad k \in J\,,$$

and since $H_m = 0$ for $m \notin J$, it follows that

$$\sum_{m \in J} H_m M^{-1}_{k-m} = \delta_{k,0} - N_0 M_k^{-1} \quad \text{for} \quad k \in J\,, \tag{51}$$

where we used eq. (43) and eq. (49) for the right-hand side. Eq. (51) can be solved if the Toeplitz matrix generated by the sequence $\{M_k^{-1}\}$ is not singular. This is always the case when $M(\omega) = R(\omega) + N_0$ is bounded away from zero and infinity, i.e. for all systems of practical interest.[10]

For evaluation of the mse, we need

$$\sum_{k \notin J} U_k R_{-k} = \sum_{k} V_k R_{-k}\,, \tag{52}$$

where the equation holds because $V_k = 0$ for $k \in J$. Since J always contains the number zero, we conclude from eq. (43) that R_{-k} can be replaced by M_{-k}. This yields together with eq. (52) and eq. (45)

$$\sum_{k \notin J} U_k R_{-k} = (1 - U_0)(R_0 - H_0)\,. \tag{53}$$

We use eq. (53) together with eq. (22) to obtain

$$U_0 = \frac{H_0}{N_0 + H_0}\,, \tag{54}$$

which finally leads to

$$\epsilon_{\text{opt}} = \sigma_a^2 \frac{N_0}{N_0 + H_0}\,, \tag{55}$$

the desired main result.

APPENDIX B

Analysis of the One-Sided Canceler of Infinite Length

For N_1 finite and $N_2 = \infty$, the set of equations (24) which determines U_k reads as follows:

$$\sum_{k<-N_1} U_k M_{m-k} = (1 - U_0) R_m \quad \text{for} \quad m < -N_1\,. \tag{56}$$

To solve this one-sided convolution, we factor the sequence $\{M_m\}$ into a causal part $\{M_m^+\}$ and an anticausal part $\{M_m^-\}$; i.e.,

$$M_m = \sum_{n=0}^{\infty} M_n^+ M_{m-n}^-\,, \tag{57}$$

where

$$M_n^+ = 0 \quad \text{for} \quad n < 0 \tag{58a}$$

$$M_n^- = 0 \quad \text{for} \quad n > 0\,. \tag{58b}$$

We now define a sequence Y_n such that

$$\sum_{m=0}^{\infty} M_n^+ Y_{m-n} = (1 - U_0) R_m \quad \text{for all} \quad m\,. \tag{59}$$

Now insert eq. (57) into eq. (56) to obtain

$$\sum_{k<-N_1} U_k \sum_{n=0}^{\infty} M_n^+ M_{m-k-n}^- = (1 - U_0) R_m \quad \text{for} \quad m < -N_1\,. \tag{60}$$

In addition, substract eq. (60) from eq. (59) for $m < -N_1$ and obtain the following set of equations

$$Y_m = \sum_{k<-N_1} U_k M_{m-k} \quad \text{for} \quad m < -N_1\,. \tag{61}$$

Multiply eq. (61) by M_m^+, sum over all $m < -N_1$, and use eq. (57) on the right-hand side to obtain

$$\sum_{k<-N_1} Y_m M_{-m}^+ = \sum_{k<-N_1} U_k M_{-k}\,. \tag{62}$$

Recall that

$$M_k = R_k + N_0 \delta_{k,0}\,, \tag{63}$$

and compare eqs. (57) and (59) which determines Y_m as

$$Y_m = (1 - U_0) M_m^- \quad \text{for} \quad m \neq 0\,. \tag{64}$$

Now insert eq. (64) into eq. (62) and make use of eq. (63) once more to obtain the one-sided sum which is required in eq. (53):

$$\sum_{k<-N_1} R_{-k} U_k = (1 - U_0) \sum_{m<-N_1} M_m^- M_{-m}^+\,. \tag{65}$$

Since $M_m^{-*} = M_{-m}^+$, it can be shown that

$$\begin{aligned} \sum_{k<-N_1} R_{-k} U_k &= (1 - U_0) \sum_{m<-N_1} |M_m^-|^2 \\ &= (1 - U_0) \sum_{m>N_1} |M_m^+|^2. \end{aligned} \tag{66}$$

Also, insert eq. (66) into eq. (53) and use eq. (63) to find

$$N_0 + H_0 = M_0 - \sum_{m>N_1} |M_m^+|^2. \tag{67}$$

With eq. (57) evaluated for $m = 0$, we finally obtain

$$N_0 + H_0 = \sum_{m=0}^{N_1} |M_m^+|^2, \tag{68}$$

and with eq. (55) we get our desired result,

$$\epsilon_{\text{opt}} = \sigma_a^2 \frac{N_0}{\sum_{m=0}^{N_1} |M_m^+|^2}. \tag{69}$$

REFERENCES

1. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968.
2. T. Berger and D. W. Tufts, "Optimal Pulse Amplitude Modulation, Part I. Transmitter-Receiver Design and Bounds from Information Theory," IEEE Trans. Information Theory, *IT-13*, No. 2 (April 1967), pp. 196–208.
3. J. Salz, "Optimum Mean-Square Decision Feedback Equalization," B.S.T.J., *52*, No. 8 (October 1973), pp. 1341–73.
4. D. D. Falconer and G. J. Foschini, "Theory of Minimum Mean-Square-Error QAM Systems Employing Decision Feedback Equalization," B.S.T.J., *52*, No. 10 (December 1973), pp. 1821–49.
5. T. M. Cover, J. F. Hayes, and J. Riera, unpublished work.
6. T. Kailath, "A General Likelihood-Ratio Formula for Random Signals in Gaussian Noise," IEEE Trans. Information Theory, *IT-15*, No. 3 (May 1969), pp. 350–61.
7. G. D. Forney, Jr., "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," IEEE Trans. Information Theory, *IT-18* (May 1972), pp. 363–78.
8. D. D. Falconer and F. R. Magee, Jr., "Adaptive Channel Memory Truncation for Maximum Likelihood Sequence Estimation," B.S.T.J., *52*, No. 9 (November 1973), pp. 1541–62.
9. D. D. Falconer, "Jointly Adaptive Equalization and Carrier Recovery in Two-Dimensional Digital Communication Systems," B.S.T.J., *55*, No. 3 (March 1976), pp. 317–34.
10. U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, Berkeley: University of California Press, 1958.
11. B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Englewood Cliffs: Prentice-Hall, 1979.
12. N. Levinson, "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction," J. Math. and Phys., *25* (1946), pp. 261–78.
13. W. F. Trench, "Weighting Coefficients for the Prediction of Stationary Time Series from the Finite Past," SIAM J. Appl. Math., *15*, No. 6 (1967), pp. 1502–10.
14. S. Zohar, "The Solution of a Toeplitz Set of Linear Equations," J. Assoc. Computing Machinery, *21*, No. 2 (April 1974), pp. 272–6.
15. Bell System Technical Reference, "Data Communications Using Voiceband Private Line Channels," October 1973, Table 3, p. 13.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9 November 1981
Printed in U.S.A.

Adaptive Equalization and Phase Tracking for Simultaneous Analog/Digital Data Transmission

By T. L. LIM and M. S. MUELLER

(Manuscript received May 14, 1981)

The general problem of equalization for data transmission where one of the two data sources produces continuous amplitude data samples is studied. There are various ways to configure a modem for such a transmission scheme, and we describe how a standard quadrature amplitude modulation structure can be modified to operate in this mode. This solution can be specialized to include various linear modulation schemes, such as single sideband and vestigial sideband. Theoretical analysis shows that adaptive equalization and adaptive phase tracking can be achieved with similar quality as in the familiar digital-only modem. We provide extensive computer simulation results which confirm the validity of our theory.

I. INTRODUCTION

Recently, considerable interest arose in finding ways to transmit and receive digital and analog data simultaneously over the 2-wire voiceband telephone channel. We investigate a system using quadrature amplitude modulation (QAM) where digital and analog data modulate the two quadrature channels. The same scheme was independently proposed in Ref. 1. The effects of various channel impairments, as well as an imbalance of signal powers between the digital and analog signals, on the error probability of the system has been studied in Ref. 2.

In this paper, we analyze an adaptive equalizer for this type of hybrid modulation, using a cross-coupled transversal filter as described in detail by Falconer in Refs. 3 and 4. The difference here is that one of the two quadrature channels transmits analog, i.e. continuous amplitude, data. If the communication channel is time invariant, it is possible to train the equalizer with digital data on both quadrature channels and freeze the equalizer taps after convergence. Analog data can then be sent. However, possible changes in channel characteristics warrant the use of an adaptive equalizer to update the taps continually. Since references for the analog data are not available, especially when the receiver is in a decision-directed mode, the update algorithms reported in Refs. 3, 4 are not applicable in this case. We propose a modification which only minimizes the mean-square error (mse) in the digital data path. Thus, only the error signal from this path is required for adaption. This analysis is similar to that in Refs. 5 and 6.

Results from computer simulations are given to verify our analytical results. We observed that because the analog data does not aid in the equalization, it actually acts as an interferer. As such, it would seem advantageous to reduce the analog signal power, thus, unbalancing the system. On the other hand, this would degrade the analog s/n. Therefore, there is a trade off in allocating different power levels to the two channels depending on which signal is more important.

In the scheme described here, we assume that data symbols are sent every T seconds. Thus, the analog signal has to be limited in bandwidth to $1/(2T)$ in order to avoid aliasing. Alternatively, the two quadrature channels can be used to transmit primarily analog data with occasional digital data for purposes of equalizer updating. Then we could have one high-rate analog channel and a low data rate digital channel.

II. MATHEMATICAL MODEL

The general QAM transmission scheme of Fig. 1 is considered. Two data sequences, $\{a_n\}$ and $\{b_n\}$, are applied to the in-phase and quadrature inputs of the cross-coupled transmission filters with impulse responses $g_p(t)$ and $g_q(t)$. Their output signals are modulated by sine and cosine waves of carrier frequency ω_0 to form the passband signal

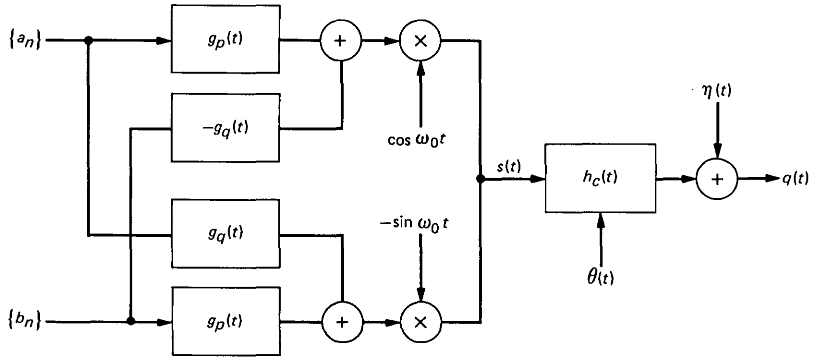


Fig. 1—Model of transmitter and channel.

$$s(t) = \text{Re}\left[\sum_n D_n G(t - nT)\exp(j\omega_0 t)\right], \tag{1}$$

where T is the symbol interval,

$$D_n = a_n + jb_n \tag{2}$$

and

$$G(t) = g_p(t) + jg_q(t). \tag{3}$$

In the above equations, $\{D_n\}$ is the complex sequence of data symbols and $G(t)$ is the complex impulse response of the transmission filter. These parameters can be specialized to represent any linear modulation scheme,[7] e.g., amplitude/phase modulation, single sideband (SSB), vestigial sideband (VSB), and QAM.

Throughout this paper it will be assumed that at least one data sequence, for example, $\{a_n\}$ is digital. In particular, we will report results for a system with $G(t)$ real and where $\{a_n\}$ and $\{b_n\}$ are digital and analog data sequences, respectively. These sequences are assumed to have zero-mean and the following correlation properties:

$$E\{a_n a_m\} = P_a \delta_{nm} \tag{4a}$$

$$E\{b_n b_m\} = P_b \delta_{nm} \tag{4b}$$

$$E\{a_n b_m\} = 0 \tag{4c}$$

$$\delta_{nm} = \begin{cases} 1, & n = m \\ 0, & \text{otherwise} \end{cases}.$$

After passing through a noisy, dispersive, and phase-jittered channel, the signal at the input to the receiver can be expressed as

$$q(t) = \text{Re}\left(\exp\{j[\omega_0 t + \theta(t)]\} \sum_k D_k U(t - kT)\right) + \eta(t),$$

where $U(t)$ is the combined complex impulse response of the transmitting filter and the baseband component of the pass-band channel $h_c(t)$ with respect to the carrier frequency ω_0. The noise process $\eta(t)$ is independent of the data sequences, while the phase shift $\theta(t)$ caused by the channel is assumed to vary much more slowly than the channel-impulse response, $h_c(t)$, and is typically about 10 degrees peak-to-peak. It is mutually independent of the additive noise process, as well as of the data symbols.

The general QAM receiver is shown in Fig. 2. The received signal $q(t)$ is bandpass-filtered by the phase splitter pair with impulse response $F(t) = f(t) + j\check{f}(t)$ where $\check{f}(t)$ denotes the Hilbert transform of $f(t)$.

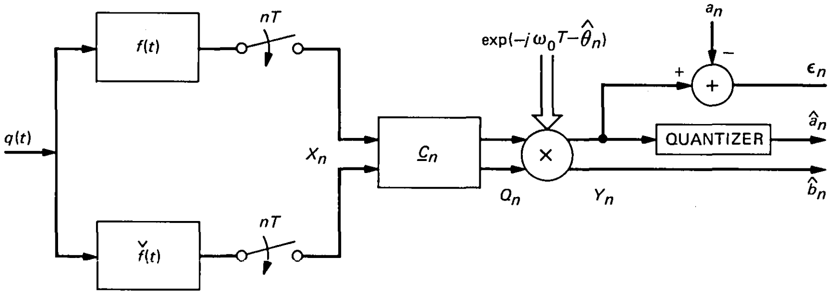


Fig. 2—Receiver structure.

The pair of outputs of the phase splitter at time kT is written as the complex signal sample

$$\begin{aligned} X_k &= x_k + j\check{x}_k \\ &= \exp[j(\omega_0 kT + \theta_k)] \sum_l D_l H_{k-l} + N_k, \end{aligned} \tag{5}$$

where $\{H_k = H(\tau + kT)\}$ are the samples of the overall complex baseband equivalent impulse response, and $\{N_k\}$ are the complex samples of the filtered-noise process. The latter process is uncorrelated with the signal and has an autocorrelation R_{NN}. Thus, we have

$$E\{N_n N_m^*\} = R_{NN}[(n-m)T] \tag{6a}$$

$$E\{N_n N_m\} = 0 \tag{6b}$$

$$E\{a_n N_m\} = 0 \tag{6c}$$

$$E\{b_n N_m\} = 0 \tag{6d}$$

for all integers n and m.

The complex signal sequence $\{X_k\}$ is passed through a cross-coupled passband equalizer with $2M + 1$ complex taps, the output of which at time nT is given by

$$\begin{aligned} Q_n &= \sum_{k=-M}^{M} C_k^* X_{n-k} \\ &= \underline{C}^* \underline{X}_n, \end{aligned} \tag{7}$$

where we write the complex tap as

$$C_k = c_k + j\check{c}_k, \qquad c_k \text{ and } \check{c}_k \text{ real}$$

and define the vectors

$$\underline{C}^t = [C_{-M}, \cdots, C_M] \tag{8}$$

$$\underline{X}_n^t = [X_{n+M}, \cdots, X_{n-M}]. \tag{9}$$

We use the * to denote conjugation for scalars and conjugate transposition for vectors and matrices. The symbol t denotes transposition.

The signal Q_n is demodulated to baseband by multiplication with $\exp(-j\hat{\theta}_n - j\omega_0 nT)$, where $\hat{\theta}_n$ is the estimated phase offset (or jitter) at time nT. The resulting signal Y_n can then be written as

$$Y_n = Q_n \exp(-j\hat{\theta}_n - j\omega_0 nT) \tag{10a}$$

$$= y_n + j\check{y}_n. \tag{10b}$$

The demodulated outputs y_n and $\check{y}_n$ are estimates of the transmitted data samples. In the following section, an optimum equalizer tap vector is derived which minimizes the mse of an appropriate cost function.

III. OPTIMIZATION OF EQUALIZER COEFFICIENTS

3.1 Analysis of the minimum mse criterion for the in-phase branch

In Refs. 3 and 4 an optimum equalizer that minimizes an mse criterion was derived. The mse was defined as

$$\begin{aligned} E[|Y_n - D_n|^2] &= E[\mathrm{Re}^2(Y_n - D_n) + \mathrm{Im}^2(Y_n - D_n)] \\ &= E[(y_n - a_n)^2] + E[(\check{y}_n - b_n)^2], \end{aligned} \tag{11}$$

which is the sum of the mse's in both branches of the equalizer output. It was found that the optimum equalizer coefficients can be calculated adaptively provided the complex output error $Y_n - D_n$ is available to the receiver. While this is the case for a transmission system with digital data in both branches where references can be estimated easily, it is not for the system considered here. In this application only one reference sequence is assumed to be available. Consequently, only the error signal in this branch can be made available for updating purposes.

In the following discussion, we define an optimum tap vector which minimizes the mse in that branch where a reference signal is available or can be estimated easily. The resulting tap vector will be compared with the result for the case where both references are available. Assuming we have a reference for $\{a_n\}$, we define the mse in that branch as the cost function to be minimized

$$\epsilon_n^2 = E[(y_n - a_n)^2] \tag{12}$$

with

$$y_n = \mathrm{Re}[\underline{C}^* \underline{X}_n \exp(-j\hat{\theta}_n - j\omega_0 nT)]. \tag{13}$$

It is convenient to express y_n in vector-matrix notation as follows:

$$y_n = \mathbf{C}^t T(\Delta\theta_n)\mathbf{X}_n, \tag{14}$$

where we partitioned the complex vectors $\underline{C}$ of the passband equalizer coefficients and $\underline{X}_n \exp(-j\theta_n - j\omega_0 nT)$ of the ideally demodulated received signal to get real vectors $\mathbf{C}$ and $\mathbf{X}_n$ with twice the original dimension

$$\mathbf{C}^t = [\mathrm{Re}(\underline{C}^t) \,|\, \mathrm{Im}(\underline{C}^t)] \tag{15}$$

$$\begin{aligned} \mathbf{X}_n^t = \{&\mathrm{Re}[\underline{X}_n^t \exp(-j\theta_n - j\omega_0 nT)] \\ &= |\,\mathrm{Im}[\underline{X}_n^t \exp(-j\theta_n - j\omega_0 nT)]\} \end{aligned} \tag{16}$$

$$\Delta\theta_n = \theta_n - \hat{\theta}_n. \tag{17}$$

In eq. (14), $T(\Delta\theta)$ is a transformation matrix expressing the effect of the demodulating phase error $\Delta\theta_n$ and is defined as

$$T(\alpha) = \left[\begin{array}{cc|cc} \cos\alpha & 0 & -\sin\alpha & 0 \\ & \ddots & & \ddots \\ 0 & \cos\alpha & 0 & -\sin\alpha \\ \hline \sin\alpha & 0 & \cos\alpha & 0 \\ & \ddots & & \ddots \\ 0 & \sin\alpha & 0 & \cos\alpha \end{array}\right]. \tag{18}$$

Note that this matrix is orthogonal; that is,

$$T(\alpha) \times T^t(\alpha) = T^t(\alpha) \times T(\alpha) = I. \tag{18a}$$

Furthermore,

$$T(-\alpha) = T^t(\alpha) \tag{18b}$$

and

$$T(\alpha + \beta) = T(\alpha) \times T(\beta) = T(\beta) \times T(\alpha). \tag{18c}$$

In order to get the mse eq. (12) as an explicit function of the coefficient vector $\mathbf{C}$, we introduce the autocorrelation matrix $\mathbf{A}$ of the demodulated received signal

$$\mathbf{A} = E\{\mathbf{X}_n \mathbf{X}_n^t\} \tag{19}$$

and the cross-correlation vector $\mathbf{V}$ between the demodulated received signal and the reference

$$\mathbf{V} = E(\mathbf{X}_n a_n). \tag{20}$$

With eqs. (14), (19), and (20), we can express eq. (12) as follows:

$$\epsilon_n^2 = \mathbf{C}^t T(\Delta\theta_n) \mathbf{A} T(-\Delta\theta_n) \mathbf{C} - 2\mathbf{C}^t T(\Delta\theta_n) \mathbf{V} + E(a_n^2). \tag{21}$$

Setting the partial derivatives with respect to $\mathbf{C}$ to zero yields the vector equation for the optimum tap vector $\mathbf{C}_{\mathrm{opt}}$

$$\mathbf{A} T(-\Delta\theta_n) \mathbf{C}_{\mathrm{opt}} = \mathbf{V}. \tag{22}$$

The definition of **A** in eq. (19) ensures that it is positive definite. Consequently, the equation has a unique solution.

$$\mathbf{C}_{\text{opt}} = T(\Delta\theta_n)\mathbf{A}^{-1}\mathbf{V}. \tag{23}$$

Inserting eq. (23) into eq. (21) yields for the minimum mse

$$\epsilon_{\text{opt}}^2 = E(a_n^2) - \mathbf{C}_{\text{opt}}^t\, T(\Delta\theta_n)\mathbf{V} = E(a_n^2) - \mathbf{V}\mathbf{A}^{-1}\mathbf{V}. \tag{24}$$

In Appendix A, we show that, for stationary data sequences $\{a_n\}$ and $\{b_n\}$ uncorrelated with the noise, the autocorrelation matrix **A** and the cross correlation vector **V** are independent of the time instant n. It is important to note that the minimum mse is independent of the constant demodulation phase error $\Delta\theta_n$. This is a consequence of eq. (23) and indicates that even by minimizing the mse in one of the two equalizer outputs, the optimum coefficient vectors can take care of any phase error, in the same manner as in a cross-coupled equalizer which minimizes the total mse at the output.

These facts have been reported in Refs. 5 and 6 for VSB- and SSB-modulated pulse amplitude modulation signals. Our analysis shows, however, that this holds in general for stationary sequences $\{a_n\}$ and $\{b_n\}$. Thus, the independence of the minimum mse on the demodulation phase is a property of the cross-coupled equalizer which is not adversely affected by the particular selection of the cost function nor by the nature of the sequence $\{b_n\}$.

When the power of the two data sequences is balanced, i.e., $P_a = P_b$, it can be seen from eq. (80) that the resulting equation for the optimum tap vector coincides exactly with the equation resulting from minimizing the total mse in both branches. In this case, both methods will give the same optimal coefficient vector and the same total mse.

In all our analysis we have assumed uncorrelated data as described by eqs. (4a) through (4c). If, instead of eq. (4b), we have

$$E(b_n b_m) = R_b(n - m),$$

then the expressions for $\mathbf{A}_1(k, l)$ and $\mathbf{A}_2(k, l)$ in Appendix A would be more complicated but they would remain stationary matrices. Then, assuming correlated data, eq. (24) is in general still valid, except that **V** and **A** are more complicated than the expressions derived in Appendix A.

Although we have based our analysis on a symbol spaced equalizer, we can also handle fractional spacing and derive similar results. As an example, we can view the $T/2$ equalizer as two parallel symbol-spaced equalizers, where the first of the two data samples during each baud is processed by one of the equalizers while the second is processed by the other equalizer. Let us denote the two equalizer tap vectors as$\underline{C}_1$ and

$\underline{C}_2$ and the input vectors as $\underline{X}_n$ and $\underline{X}_{n+1/2}$. Then we can define

$$\begin{aligned}
\mathbf{C}^t &= [\mathrm{Re}(C_1^t)\,|\,\mathrm{Im}(\underline{C}_1^t)\,|\,\mathrm{Re}(\underline{C}_2^2)\,|\,\mathrm{Im}(\underline{C}_2^t)] \\
\mathbf{X}_n^t &= \{\mathrm{Re}[\underline{X}_n^t \exp(-j\theta_n - j\omega_0 nT)]\,|\,\mathrm{Im}[\underline{X}_n^t \exp(-j\theta_n - j\omega_0 nT)] \\
&\quad |\,\mathrm{Re}[\underline{X}_{n+1/2}^t \exp(-j\theta_n - j\omega_0 nT)] \\
&\quad |\,\mathrm{Im}[\underline{X}_{n+1/2}^t \exp(-j\theta_n - j\omega_0 nT)]\} \\
T(\alpha) &= \left[\begin{array}{c|c} T_1(\alpha) & 0 \\ \hline 0 & T_1(\alpha) \end{array}\right],
\end{aligned}$$

where $T_1(\alpha)$ is given in eq. (18). With these definitions, all the previous results for $\mathbf{C}_{\mathrm{opt}}$ and $\epsilon_{\mathrm{opt}}^2$ in eqs. (23) and (24) follow.

3.2 Mean square error in the quadrature branch

We now analyzed the mse in the second branch of the equalizer. The output of this branch is

$$\check{y}_n = \mathrm{Im}[\underline{C}^* \underline{X}_n \exp(-j\hat{\theta}_n - j\omega_0 nT)]. \tag{25}$$

Since

$$\mathrm{Im}\{\underline{Z}\} = \mathrm{Re}\left[\underline{Z} \exp\left(-j\frac{\pi}{2}\right)\right], \tag{26}$$

we are able to express $\check{y}_n$ in terms of $\mathbf{C}$ and $\mathbf{X}$ defined in eqs. (15) and (16) using the transformation properties in eq. (18)

$$\check{y}_n = \mathbf{C}^t T\left(\Delta\theta_n - \frac{\pi}{2}\right) \mathbf{X}_n. \tag{27}$$

Therefore, the mse in the second branch can be expressed in vector-matrix notation as follows:

$$\check{\epsilon}_n^2 = E[(\check{y}_n - b_n)^2] \tag{28}$$

$$\begin{aligned}
&= \mathbf{C}^t T\left(\Delta\theta_n - \frac{\pi}{2}\right) \mathbf{A} T\left(-\Delta\theta_n + \frac{\pi}{2}\right) \mathbf{C} \\
&\quad -2\mathbf{C}^t T\left(\Delta\theta - \frac{\pi}{2}\right) E(\mathbf{X}_n b_n) + E(b_n^2).
\end{aligned} \tag{29}$$

Using eq. (18) and an approach similar to eq. (66) through eq. (70) in Appendix A, we show that

$$T\left(-\frac{\pi}{2}\right) E\{\mathbf{X}_n b_n\} = \mathbf{V}(P_b/P_a). \tag{30}$$

On substituting eq. (30) into eq. (29) we obtain

$$\check{\epsilon}_n^2 = \mathbf{C}^t T\left(\Delta\theta_n - \frac{\pi}{2}\right) \mathbf{A} T\left(-\Delta\theta_n + \frac{\pi}{2}\right) \mathbf{C} - 2\mathbf{C}^t T(\Delta\theta_n)\mathbf{V}\frac{P_b}{P_a} + P_b. \tag{31}$$

For the balanced power case, i.e., $P_a = P_b$, Appendix A shows that

$$T\left(-\frac{\pi}{2}\right) \mathbf{A} T\left(\frac{\pi}{2}\right) = \mathbf{A}, \tag{32}$$

and it follows that

$$\check{\epsilon}^2 = \mathbf{C}^t T(\Delta\theta_n)\mathbf{A}T(-\Delta\theta_n)\mathbf{C} - 2\mathbf{C}^t T(\Delta\theta_n)\mathbf{V} + P_a. \tag{33}$$

This is exactly the same expression as for ϵ^2 in eq. (21). Consequently,

$$\check{\epsilon}^2 = \epsilon^2. \tag{34}$$

Thus, we conclude that the mse's in both branches are equal. In this special case, minimizing the mse in one branch also minimizes the mse in the other branch.

3.3 Analysis of an infinitely long equalizer

While the formal solution eq. (24) for the minimum mse already shows the independence of a constant-phase error, it does not reveal anything about the influence of channel parameters (amplitude, or phase distortion) or of the sampling instant. To obtain further insight into this dependence, we analyze an infinite length equalizer.

We show in Appendix B that the resulting minimum mse for an infinite tap equalizer can be written as

$$\epsilon_{\text{opt}}^2 = \frac{TP_a}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{P_b[Z(\omega) + Z(-\omega)] + 1}{4P_aP_bZ(\omega)Z(-\omega) + (P_a + P_b)[Z(\omega) + Z(-\omega)] + 1}\, d\omega, \tag{35}$$

where

$$Z(\omega) = \left|\frac{H_{\text{eq}}(\omega)}{N_{\text{eq}}(\omega)}\right|^2. \tag{36}$$

In eq. (36) $H_{\text{eq}}(\omega)$ is the Fourier transform of the sampled impulse response $H(\tau + kT)$, where τ indicates the sampling instant. It is related to the transfer function $H(\omega)$ as follows:

$$H_{\text{eq}}(\omega) = \frac{\exp(j\omega\tau)}{T}\sum_{k=-\infty}^{\infty} H\left(\omega + k\frac{2\omega}{T}\right)\exp\left(j2\pi k\frac{\tau}{T}\right), \tag{37}$$

and $|N_{\text{eq}}(\omega)|^2$ is the baseband component of the noise power spectrum

$$|N_{\text{eq}}(\omega)|^2 = \sum_{k=-\infty}^{\infty} R_{NN}(kT) \exp[-j(\omega + \omega_0)kT]. \tag{38}$$

The formula given in eq. (35) can be evaluated for all the different modulation schemes which can be modeled by a linear transmission system. The only frequency-dependent term appearing in eq. (35) is $Z(\omega)$, the s/n of the sampled received signal. According to eq. (37) this will, in general, depend on the sampling instant, τ, and the phase characteristic of the overall channel transfer function, $H(\omega)$. If the sampling theorem is satisfied, i.e., if $H(\omega) = 0$ for all ω outside the interval $[\omega_1, \omega_1 + 2\pi/T]$, where ω_1 is arbitrary, combining eqs. (37) and (36) shows that the minimum mse is only dependent on the amplitude characteristic of the channel transfer function and of the noise power density spectrum. A QAM transmission system with no excess bandwidth is an example; a system transmitting only one data sequence and using VSB-modulation with less than 100 percent excess bandwidth is another, more realistic example. In case of balanced power in the transmitted data sequences, i.e., $P_a = P_b$ or for $Z(\omega)$ symmetric around ω_1, i.e., $Z(\omega + \omega_1) = Z(-\omega + \omega_1)$, the mse is given by,

$$\epsilon^2_{\text{opt}}|_{P_a=P_b} = \frac{TP_a}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{d\omega}{2P_aZ(\omega) + 1}. \tag{39}$$

In eq. (100) of Appendix B we show that the partial derivative of ϵ^2_{opt} with respect to P_b is nonnegative. Therefore, an increase in the analog signal power P_b causes an increase in ϵ^2_{opt}. So the analog signal acts as an interferer to the digital signal.

IV. ANALYSIS OF THE GRADIENT ALGORITHM FOR JOINT EQUALIZATION AND PHASE TRACKING

In an adaptive receiver, the equalizer is assumed to know the reference data for startup and to operate in a decision-directed mode when random data is being sent. In either case, the tap weight vector is being updated continuously. Similarly, the phase offset, or jitter, would be continuously tracked in order to remain in synchronism. As Falconer did, in Refs. 3 and 4, we assume gradient algorithms are used in these updatings as follows (C is now time-varying),

$$\mathbf{C}_{n+1} = \mathbf{C}_n - \frac{\beta}{2} \nabla\mathbf{C}\epsilon^2_n \tag{40a}$$

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{\alpha}{2} \frac{\partial}{\partial\hat{\theta}_n} \epsilon^2_n, \tag{40b}$$

where $\nabla\mathbf{C}\epsilon_n^2$ denotes the gradient of ϵ_n with respect to $\mathbf{C}$.

It can be shown from eq. (21) that

$$\nabla\mathbf{C}\epsilon_n^2 = 2T(\Delta\theta_n)[\mathbf{A}T^t(\Delta\theta_n)\mathbf{C}_n - \mathbf{V}] \tag{41a}$$

and

$$\frac{\partial}{\partial\hat{\theta}_n}\,\epsilon_n^2 = 2\mathbf{C}_n^t\,\frac{\partial T}{\partial\hat{\theta}_n}\,(\Delta\theta_n)[\mathbf{A}T^t(\Delta\theta_n)\mathbf{C}_n - \mathbf{V}]. \tag{41b}$$

Using the definitions of $\mathbf{A}$ in eq. (19) and $\mathbf{V}$ in eq. (20) and substituting eq. (14) we can recast eqs. (41a) and (41b) into

$$\nabla\mathbf{C}_n\epsilon_n^2 = 2E[T(\Delta\theta_n)\mathbf{X}_n(y_n - a_n)] \tag{42a}$$

$$\frac{\partial\epsilon_n^2}{\partial\hat{\theta}_n} = 2E\left[\mathbf{C}_n^t\,\frac{\partial T(\Delta\theta_n)}{\partial\hat{\theta}_n}\,\mathbf{X}_n(y_n - a_n)\right]. \tag{42b}$$

Finally, we use another property of the transformation matrix $T(\alpha)$, namely,

$$\frac{\partial T(\alpha + \beta)}{\partial\alpha} = T\left(\alpha + \beta + \frac{\pi}{2}\right),$$

together with eq. (27) to obtain

$$\frac{\partial\epsilon_n^2}{\partial\hat{\theta}_n} = 2E[\check{y}_n(y_n - a_n)]. \tag{42c}$$

Equations (42a) and (42c) can be used to update the equalizer according to eq. (40). Note that all the signals required to update the equalizer are readily available at the receiver. Also note that $y_n - a_n = e_n$ is the error in the branch where a reference is available and $\check{y}_n$ is the output of the other branch.

By taking expected values as shown in eqs. (42a and b), we obtain the estimated gradient algorithm. In practice, a stochastic gradient approach is used to avoid the long delay involved in estimating the averages. The update equations are obtained by omitting the expectations in eq. (42) and making small corrections in the direction of the instantaneous values instead. Hereafter, we shall discuss only the stochastic gradient method.

Inserting eq. (42a) into eq. (40a) yields the update equation for the coefficient vector

$$\mathbf{C}_{n+1} = \mathbf{C}_n - \beta'(y_n - a_n)T(\Delta\theta_n)\mathbf{X}_n. \tag{43}$$

The corresponding equation for the update of the phase jitter corrector is

$$\check{\theta}_{n+1} = \check{\theta}_n - \frac{\alpha'}{a_n^2}\,\check{y}_n(y_n - a_n). \tag{44}$$

Division by a_n^2 has been included in eq. (44) to give the corrections a smaller weight, if the nominal data value has a larger absolute value. Note that because of the special nature of $T(\Delta\theta_n)$ defined in eq. (18), the matrix multiplication in eq. (43) requires only $2(2M + 1)$ multiplication. Equation (43) can be reexpressed in terms of the complex equalizer $\underline{C}_n$ and input $\underline{X}_n$ as

$$\underline{C}_{n+1} = \underline{C}_n - \beta' \underline{X}_n[(y_n - a_n)\exp(jn\omega_0 T + j\theta_n)]^*. \qquad (45)$$

The equalizer coefficient and phase adjustment algorithms represented by eqs. (44) and (45) are similar to the ones published in Ref. 3. The main difference is that here, only the error in the digital data path appears in the adjustment algorithms.

An important parameter in the evaluation of the dynamic behavior of the control loop is the rate of convergence (ROC). For the case of updating the equalizer only, the ROC can be analyzed using, for example, Ungerboeck's method,[8] since the stochastic recurrence equations for the excess mse can be cast in the form analyzed in Ref. 8. Combining eq. (43) with eq. (14), it is seen that for the stochastic gradient case

$$T(-\Delta\theta_n)\mathbf{C}_{n+1} = [I - \beta'\mathbf{X}_n\mathbf{X}_n^t]T(-\Delta\theta_n)\mathbf{C}_n + \beta' a_n\mathbf{X}_n. \qquad (46)$$

From eqs. (21) and (24) the excess mse is given by

$$\epsilon_n^2 - \epsilon_{\mathrm{opt}}^2 = (\mathbf{C}_n - \mathbf{C}_{\mathrm{opt}})^t T(\Delta\theta_n)\mathbf{A}T(-\Delta\theta_n)(\mathbf{C}_n - \mathbf{C}_{\mathrm{opt}}). \qquad (47)$$

Therefore,

$$\epsilon_{n+1}^2 - \epsilon_{\mathrm{opt}}^2 = \mathbf{D}_n^t\mathbf{A}\mathbf{D}_n, \qquad (48)$$

with

$$\mathbf{D}_n = (I - \beta'\mathbf{X}_n\mathbf{X}_n^t)T(-\Delta\theta_n)(\mathbf{C}_n - \mathbf{C}_{\mathrm{opt}}) - \beta'\mathbf{X}_n[\mathbf{X}_n^t T(-\Delta\theta_n)\mathbf{C}_{\mathrm{opt}} - a_n]. \qquad (49)$$

The convergence of the excess mse eq. (48) is analyzed in Ref. 8. There it is shown further that for stability the control loop constant has to satisfy

$$0 \le \beta' \le \frac{2}{E(\mathbf{X}_n^t\mathbf{X}_n)}, \qquad (50)$$

with

$$\beta_0' = \frac{1}{E(\mathbf{X}_n^t\mathbf{X}_n)}. \qquad (51)$$

The control loop constant, β_0', is chosen to give the fastest initial convergence.

In steady-state the excess mse is given by

$$\epsilon_\infty^2 - \epsilon_{\rm opt}^2 = \frac{\beta' E(\mathbf{X}^t\mathbf{X})\epsilon_{\rm opt}^2}{2 - B'E(\mathbf{X}^t\mathbf{X})}. \tag{52}$$

From eq. (52) it can be seen that the excess mse can be reduced to an arbitrarily small value by selecting β' small enough.

It is interesting to note that eq. (50) specifies a stability region for β' equal to that for an equalizer using error signals from both branches $[E(\mathbf{X}^t\mathbf{X}) = E(\underline{X}^*\underline{X})]$. From eq. (52) it follows that for a particular β' the ratio of the excess mse to $\epsilon_{\rm opt}^2$ is the same as for an equalizer using error signals from both branches.

We have not been able to analyze the ROC for the joint operation of equalizer and phase jitter loops. In contrast to the equalizer update equation, the transformation matrix $T(\Delta\theta_n)$ in the joint case is involved in a nonlinear manner. We, therefore, resort to computer simulation of the loop behavior. These results are reported in the next section.

V. SIMULATION RESULTS

Here, we present simulation results for the hybrid modulation scheme described earlier. Pseudorandom digital data selected from the $\{\pm 1, \pm 3\}$ alphabet is used to modulate the in-phase channel, while a set of pseudorandom numbers with a Gaussian distribution $N(O, P_b)$ modulates the quadrature channel, where P_b is the analog signal power. Additive Gaussian noise at −30 dB below the average signal level is introduced in the channel and, whenever desired, phase jitter is introduced. The latter process is modeled by a 60-Hz sinusoid of 10 degrees peak-to-peak. The channels used in the simulation are (*i*) a good channel with a flat amplitude frequency response within most of the frequency band of interest and small delay distortion, except near the lower band edge, and (*ii*) a channel just violating the requirements for basic conditioning with both moderate amplitude and delay distortion. These characteristics are shown in Fig. 3.

Although it is not intended to start up an equalizer with a reference signal only in one branch, we report results of such simulations. This gives good insight into the dynamic behavior of the adaptive equalizer update loop and facilitates the comparison with a conventional passband equalizer.

The first run described is for balanced power ($P_a = P_b = 5$), where the receiver has a 64-tap $T/2$ complex equalizer and the error signal is derived from the in-phase digital channel alone. The basic channel described above is used. Figure 4 is a sample simulation run displaying the s/n as a function of time where s/n is defined as the ratio of the digital signal power to the digital mse. The latter is taken to be a weighted sum of past and present instantaneous squared errors. The

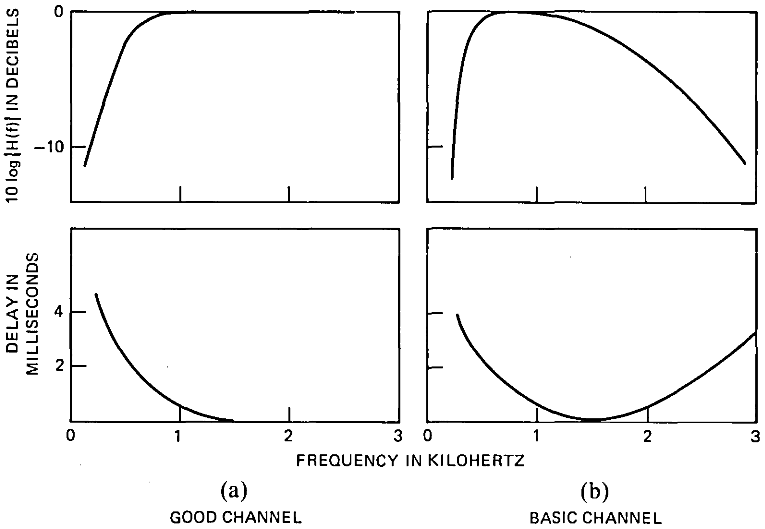


Fig. 3—Channel frequency responses.

two curves are for two different timing phases in the receiver. As shown, the equalizer converges in about 2000 iterations for a step size of $\beta = 0.0005$. The step size giving the fastest convergence, according to eq. (53), would be $\beta_0 = 0.0015$. A lower value is used in order to re-

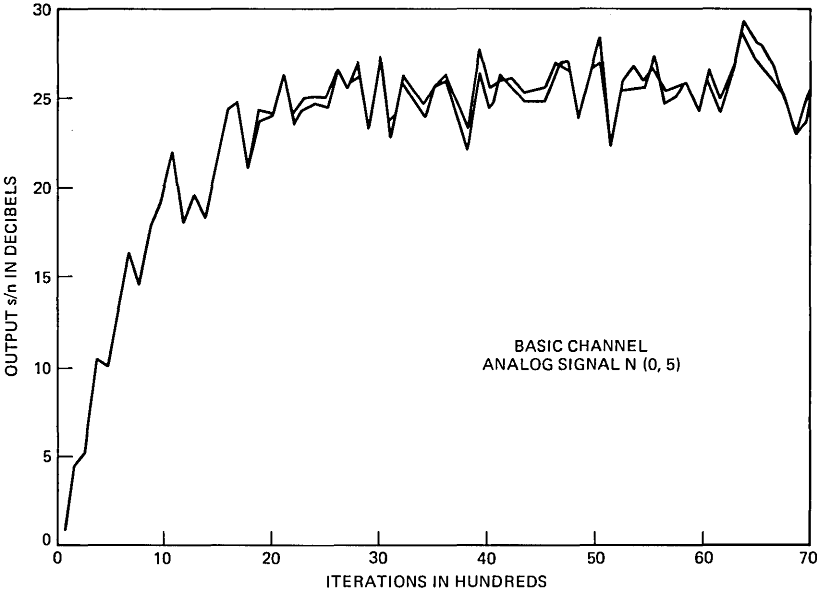


Fig. 4—Convergence of hybrid equalizer.

duce the mse in steady state. The receiver's digital and analog outputs after equalization and demodulation to baseband are displayed in Fig. 5 as a scatter plot. The vertical and horizontal axes represent analog and digital values respectively and, ideally, the data points would be on vertical lines passing through the X-axis at ±1, ±3. Owing to channel noise, there is both lateral and vertical displacement from the true values. The decision thresholds for the digital data are the vertical lines with abscissae 0 and ±2. In this example, and all the others to be described, the input s/n is about 30 dB so that the equalizer has done a reasonable job in removing intersymbol interference (ISI) caused by the basic channel.

We next exhibit the results for the regular QAM in Figs. 6 and 7, with digital data on both branches and using exactly the same channel and receiver parameters as before. The ideal constellation in the scatter plot would be 16 points with coordinates ±1, ±3 (in the absence of ISI

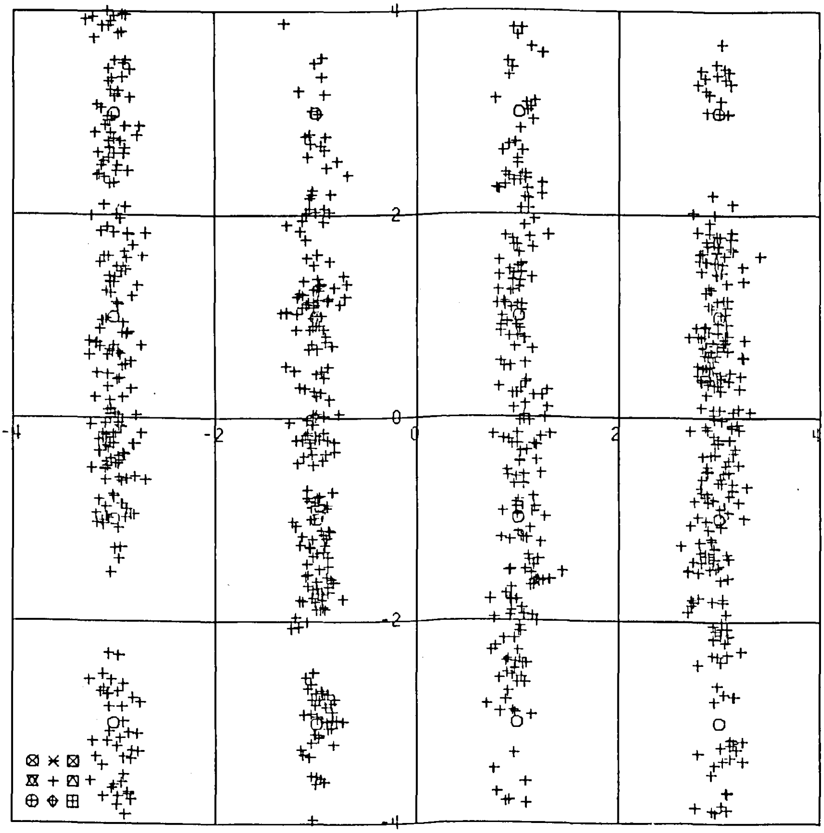


Fig. 5—Hybrid receiver output constellation.

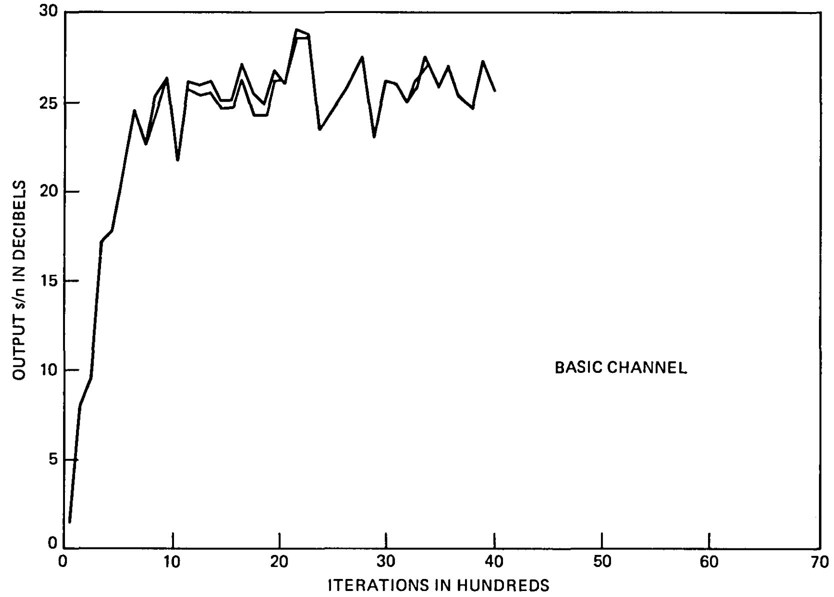


Fig. 6—Convergence of regular QAM equalizer.

and channel noise). As seen in Fig. 6, the convergence is faster than for the hybrid modulation since here we used a complex error function, leading to a tap weight vector $\underline{C}_n$ that fluctuates less than when using a one-sided error signal only.

We mentioned in Section 3.1 that when the powers in the digital and analog branches are balanced, the optimum equalizer and mse are the same as that of a regular QAM receiver. Indeed, we see by comparing Figs. 4 and 6 that the digital mse's reach the same levels. It was also observed that the significant tap weights for both receiver timing phases differ by less than 5 percent.

Now we exhibit the effect on the mse of unbalancing the analog signal power from $P_b = 5$ to some other value. The simulated receiver has an AGC that scales the received signal to an average power of "ten" and, hence, the equalizer tap adjustment step size is kept the same in this series of runs. The analog mse is obtained by averaging all the past instantaneous mse's over time. Instead of presenting a series of curves, we summarize the results in Table I. The digital signal power is fixed at "five," but the analog power is varied from "one" to "nine." The output mse normalized to the power of the corresponding signal is presented in Table I for the good channel. Thus, we see that, as the analog power is increased, the digital mse increases, confirming the analytical results in Section 3.3 for the infinite length equalizer. It was

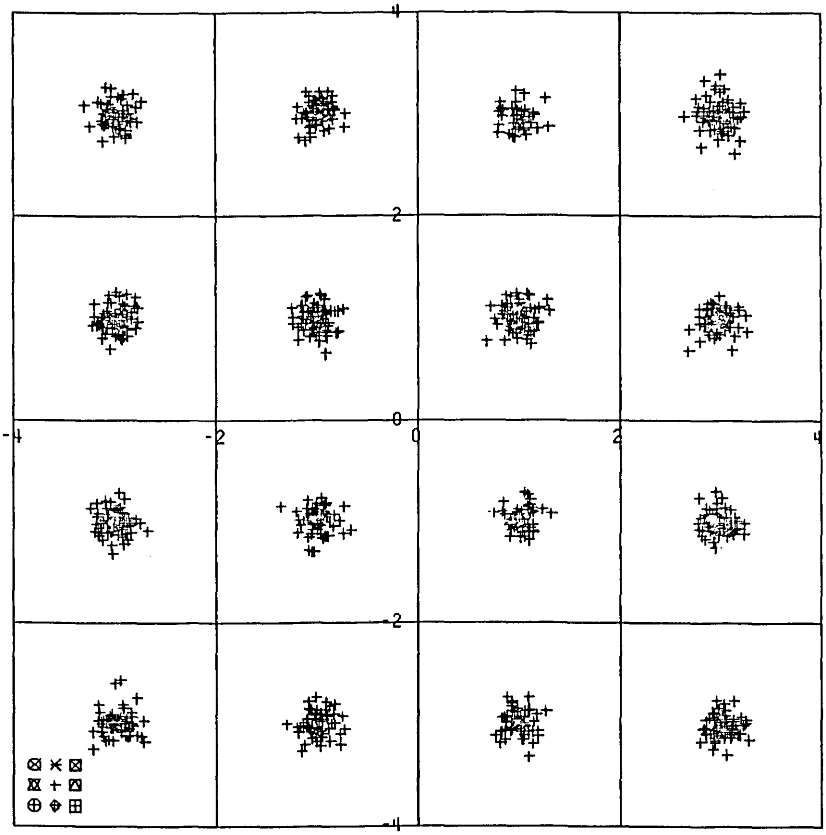


Fig. 7—Regular QAM receiver output constellation.

also shown in Ref. 2 that the error probability upper bound is increased when a power imbalance exists in favor of the analog signal. As expected, we also see that the normalized analog mse is smaller with a larger analog signal power. We also note that the digital and analog

Table I—Comparison of analog and digital mse for different P_b (good channel)

Analog Power P_b	Normalized mse	
	Digital	Analog
1	0.0007	0.0040
3	0.0010	0.0017
5	0.0013	0.0013
7	0.0016	0.0011
9	0.0018	0.0009

mse's are almost the same for balanced power, as theoretically predicted in Section 3.2.

Simulations were also performed to study the hybrid equalizer performance in the presence of sinusoidal phase jitter of 60 Hz, where the phase tracker modeled by eq. (44) is used to estimate the jitter process. Figure 8 shows a sample run where, after allowing the equalizer to reach steady-state, 60-Hz jitter with 10 degrees peak-to-peak amplitude is introduced causing a degradation in performance. The plot in Fig. 9 shows the same run, except that the phase tracker is turned on shortly after the phase jitter is introduced. The lower the jitter frequency and amplitude, the more effective we can expect the phase tracker to be.[4]

VI. CONCLUSION

A data transmission system capable of transmitting and receiving analog and digital data simultaneously has been studied in detail. We found that it is possible to perform adaptive equalization of the channel even when only one of the two quadrature channels is carrying digital data. Moreover, the minimum mse and tap-weight vector are unchanged from that of the regular QAM as long as the analog and digital

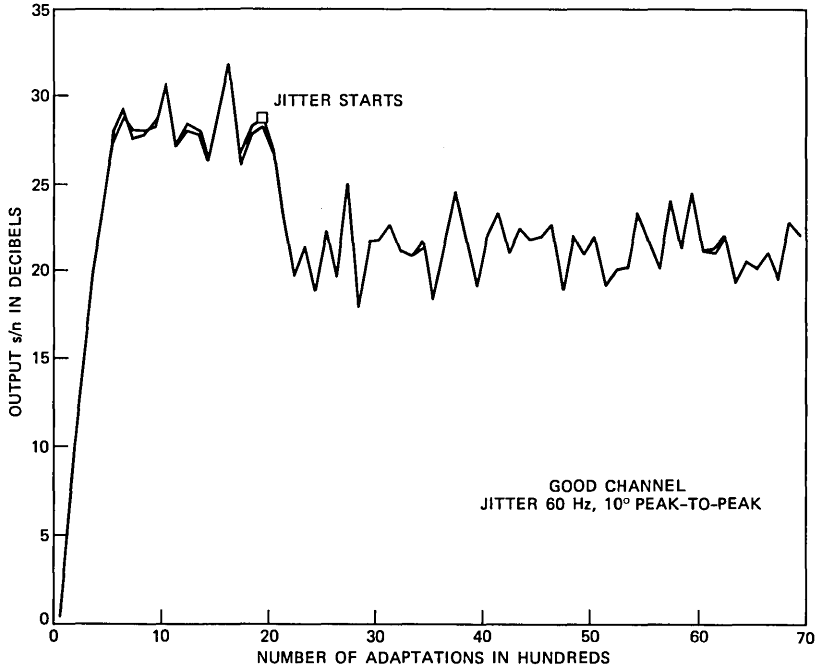


Fig. 8—Hybrid receiver performance in presence of phase jitter (no carrier recovery).

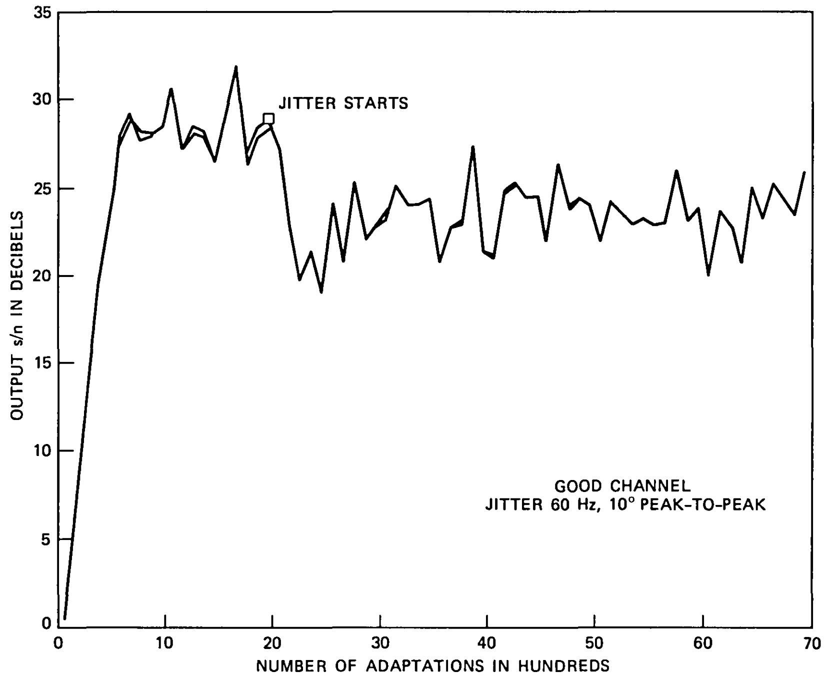


Fig. 9—Hybrid receiver performance in presence of phase jitter (carrier recovery, $\alpha = 0.5$).

signal powers are equal. However, start-up with simultaneous analog and digital data is slower by approximately a factor of two compared to the case of a conventional QAM system. An efficient start-up procedure might be to train the receiver with digital data on both channels and then switch one channel to analog data upon convergence of the equalizer, since in the case of balanced power, the equalizer taps are the same. We also found that the scheme can tolerate moderate amounts of phase jitter.

APPENDIX A

Analysis of A *and* V

We define,

$$\underline{X}_{p,n} = \mathrm{Re}[\underline{X}_n \exp(-j\theta_n - j\omega nT)] \tag{53a}$$

$$\underline{X}_{q,n} = \mathrm{Im}[\underline{X}_n \exp(-j\theta_n - j\omega_0 nT)]. \tag{53b}$$

According to eq. (16), it follows that

$$\mathbf{X}_n^t = (\underline{X}_{p,n}^t | \underline{X}_{q,n}^t). \tag{54}$$

Inserting eq. (54) into the definition eq. (19) of **A** yields

$$\mathbf{A} = E\left(\begin{array}{c|c} \underline{X}_{p,n}\underline{X}^{t}_{p,n} & \underline{X}_{p,n}\underline{X}^{t}_{q,n} \\ \hline \underline{X}_{q,n}\underline{X}^{t}_{p,n} & \underline{X}_{q,n}\underline{X}^{t}_{q,n} \end{array}\right). \tag{55}$$

This can be expressed as

$$\mathbf{A} = \frac{1}{2}\left(\begin{array}{c|c} \mathrm{Re}(\mathbf{A}_1 + \mathbf{A}_2) & -\mathrm{Im}(\mathbf{A}_1 - \mathbf{A}_2) \\ \hline \mathrm{Im}(\mathbf{A}_1 + \mathbf{A}_2) & \mathrm{Re}(\mathbf{A}_1 - \mathbf{A}_2) \end{array}\right), \tag{56}$$

where

$$\mathbf{A}_1 = E[(\underline{X}_{p,n} + j\underline{X}_{q,n})(\underline{X}_{p,n} - j\underline{X}_{q,n})^{t}] = E(\underline{X}_n\underline{X}_n^{*}) \tag{57}$$

$$\begin{aligned} \mathbf{A}_2 &= E[(\underline{X}_{p,n} + j\underline{X}_{q,n})(\underline{X}_{p,n} + j\underline{X}_{q,n})^{t}] \\ &= E[\underline{X}_n\underline{X}_n^{t} \exp(-2j\theta - 2j\omega_0 nT)]. \end{aligned} \tag{58}$$

We note that $\mathbf{A}_2$ is symmetric and $\mathbf{A}_1$ is Hermitian. For uncorrelated data and noise sequences the k, lth entry in the matrix $\mathbf{A}_2$ is computed with eqs. (5), (6), (9), and (58) as follows

$$\begin{aligned} \mathbf{A}_2(k, l) &= E[X_{n-k}X_{n-l}\exp(-2j\theta_n - 2j\omega_0 nT)] \\ &= \sum_{\nu}\sum_{\mu} E(D_\nu D_\mu)H_{n-k-\nu}H_{n-l-\mu} \\ &\quad \exp[j(\theta_{n-k} + \theta_{n-l} - 2\theta_n) - j\omega_0(k + l)T] \\ &\quad + E[N_{\mu-k}N_{n-l}\exp(-2j\theta_n - 2j\omega_0 nT)]. \end{aligned} \tag{59}$$

From eqs. (2) and (4) we get

$$E(D_\nu D_\mu) = (P_a - P_b)\delta_{\nu\mu}. \tag{60}$$

With the assumption that the phase jitter is quasi-stationary over the equalizer length, i.e. $\theta_{n-k} = \theta_n$ for all $k \in [-M, M]$, we obtain

$$\mathbf{A}_2(k, l) = (P_a - P_b)\sum_{\nu} H_\nu H_{\nu+k-l}\exp[-j\omega_0(k + l)T]. \tag{61}$$

Note that $\mathbf{A}_2(k, l)$ is zero, if the powers of the sequences $\{a_n\}$ and $\{b_n\}$ are equal, i.e., $P_a = P_b$.

Similarly the k, lth entry in $\mathbf{A}_1$ is

$$\begin{aligned} \mathbf{A}_1(k, l) &= E(X_{n-k}X^{*}_{n-l}) \\ &= \sum_{\nu}\sum_{\mu} E(D_\nu D^{*}_\mu)H_{n-k-\nu}H^{*}_{n-l-\mu} \\ &\quad \times \exp[j(\theta_{n-k} - \theta_{n-l}) + j\omega_0(l - k)T] \\ &\quad + E[N_{n-k}N^{*}_{n-l}], \end{aligned} \tag{62}$$

with

$$E(D_\nu D^{*}_\mu) = (P_a + P_b)\delta_{\nu\mu}, \tag{63}$$

$$E(N_\nu N_\mu^*) = R_{NN}[(\mu - \nu)T], \tag{64}$$

and the assumption of quasi-stationary phase jitter, this finally yields

$$\mathbf{A}_1(k, l) = (P_a + P_b) \sum_\nu H_\nu H_{\nu+k-l}^* \times \exp[j\omega_0(l - k)T] + R_{NN}[(k - l)T]. \tag{65}$$

With the definition of **V** in eqs. (20) and (54), we find

$$\mathbf{V} = E\begin{pmatrix} \underline{X}_{p,n}a_n \\ \underline{X}_{q,n}a_n \end{pmatrix} \tag{66}$$

or

$$\mathbf{V} = \begin{pmatrix} \underline{V}_1 \\ \underline{V}_2 \end{pmatrix}, \tag{67}$$

with

$$\underline{V}_1 = E[\underline{V}_{p,n}a_n] \tag{68a}$$

and

$$\underline{V}_2 = E[\underline{X}_{q,n}a_n]. \tag{68b}$$

The kth entry in $\underline{V}_1$ is calculated by inserting eqs. (5), (9), and (53a) into (68b)

$$\begin{aligned} V_1(k) &= E\{a_n\mathrm{Re}[X_{n-k}\exp(-j\theta_n - j\omega_0 nT)]\} \\ &= E\{a_n\mathrm{Re}[\exp(j\theta_{n-k} - j\theta_n - j\omega_0 kT) \sum_l D_l H_{n-k-l}]\}. \end{aligned} \tag{69}$$

Again, under the assumption of quasi-stationary phase jitter and with eqs. (2) and (4), we have

$$V_1(k) = P_a\mathrm{Re}[H_{-k}\exp(-j\omega_0 kT)]. \tag{70}$$

Following the same lines, one obtains for $V_2(k)$,

$$V_2(k) = P_a\mathrm{Im}[H_{-k}\exp(-j\omega_0 kT)]. \tag{71}$$

A transformed version of **A** is needed for the analysis of the mse in the second branch. According to eq. (18), the transformation matrix is

$$T\left(-\frac{\pi}{2}\right) = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ & \ddots & & \ddots \\ 0 & 0 & 0 & 1 \\ \hline -1 & 0 & 0 & 0 \\ & \ddots & & \ddots \\ 0 & -1 & 0 & 0 \end{array}\right]. \tag{72}$$

Inserting this into eq. (56) yields

$$T\left(-\frac{\pi}{2}\right)\mathbf{A}T\left(\frac{\pi}{2}\right)=\frac{1}{2}\left[\begin{array}{c|c}\mathrm{Re}(\mathbf{A}_1-\mathbf{A}_2) & -\mathrm{Im}(\mathbf{A}_1+\mathbf{A}_2)\\ \hline \mathrm{Im}(\mathbf{A}_1-\mathbf{A}_2) & \mathrm{Re}(\mathbf{A}_1+\mathbf{A}_2)\end{array}\right]. \tag{73}$$

By comparing eq. (73) with eq. (56), it is found that

$$T\left(-\frac{\pi}{2}\right)\mathbf{A}T\left(\frac{\pi}{2}\right)=\mathbf{A}+\mathbf{A}' \tag{74}$$

with

$$\mathbf{A}'=\left[\begin{array}{c|c}-\mathrm{Re}\mathbf{A}_2 & -\mathrm{Im}\mathbf{A}_2\\ \hline -\mathrm{Im}\mathbf{A}_2 & \mathrm{Re}\mathbf{A}_2\end{array}\right]. \tag{75}$$

In eq. 75, $\mathbf{A}'$ is symmetric but not positive definite. Then, referring to eq. (61), we see that for the balanced power case, i.e., $P_a = P_b$,

$$\mathbf{A}'=0 \quad \text{and} \quad T\left(-\frac{\pi}{2}\right)\mathbf{A}T\left(\frac{\pi}{2}\right)=\mathbf{A}.$$

APPENDIX B

Optimum Tap Weight and mse of the Infinitely Long Equalizer

Define

$$\mathbf{C}^t_{\mathrm{opt}}=(\underline{C}^t_1 \mid \underline{C}^t_2) \tag{76}$$

and use eqs. (56) and (67). Then, the equation for the optimum coefficient vector eq. (22) can be expressed as follows

$$\begin{aligned}\mathrm{Re}(\mathbf{A}_1+\mathbf{A}_2)\underline{C}_1-\mathrm{Im}(\mathbf{A}_1-\mathbf{A}_2)\underline{C}_2&=2\underline{V}_1\\ \mathrm{Im}(\mathbf{A}_1+\mathbf{A}_2)\underline{C}_1+\mathrm{Re}(\mathbf{A}_1-\mathbf{A}_2)\underline{C}_2&=2\underline{V}_2,\end{aligned} \tag{77}$$

where it is assumed that the demodulation phase error $\Delta\theta_n$ equals zero. (It has been stated that the minimum mmse is independent of $\Delta\theta_n$, at least for a constant phase offset. Therefore, no loss of generality results because of this simplification).

Let

$$\underline{C}=(\underline{C}_1+j\underline{C}_2) \tag{78}$$

and

$$\underline{V}=(\underline{V}_1+j\underline{V}_2). \tag{79}$$

This allows us to write eq. (77) in complex notation

$$\mathbf{A}_1\underline{C}+\mathbf{A}_2\underline{C}^*=2\underline{V}. \tag{80}$$

Equation (80) can be expressed in the components of the vector $\underline{V}$ and matrices $\mathbf{A}_1$ and $\mathbf{A}_2$

$$\sum_{l=-M}^{M} [\mathbf{A}_1(k, l)C(l) + \mathbf{A}_2(k, l)C^*(l)] = 2V(k). \tag{81}$$

Note from eq. (61) that since $\mathbf{A}_2$ is not a Toeplitz matrix, the sum in eq. (81) is not a convolution even if the dimension M approaches infinity. Define

$$\tilde{C}(l) = C(l)\exp(j\omega_0 lT), \tag{82}$$

$$\tilde{V}(l) = V(l)\exp(j\omega_0 lT), \tag{83}$$

$$\begin{aligned}\tilde{\mathbf{A}}_1(k-l) &= \mathbf{A}_1(k, l)\exp[-j\omega_0(l-k)T],\\ &= (P_a + P_b)\sum_{\nu} H_\nu H^*_{\nu+k-l}\\ &\quad + R_{NN}(k-l)\exp[-j\omega_0(l-k)T],\end{aligned} \tag{84}$$

$$\begin{aligned}\tilde{\mathbf{A}}_2(k-l) &= \mathbf{A}_2(k, l)\exp[j\omega_0(k-l)T],\\ &= (P_a - P_b)\sum H_\nu H_{\nu+k-l}.\end{aligned} \tag{85}$$

Equations (61) and (65) were used to get eqs. (84) and (85). With these definitions, we can transform the problem to baseband. Inserting eqs. (82) to (85) into eq. (81) yields

$$\sum_{l=-M}^{M} [\tilde{\mathbf{A}}_1(k-l)\tilde{C}(l) + \tilde{\mathbf{A}}_2(k-l)\tilde{C}^*(l)] = 2\tilde{V}(k). \tag{86}$$

Here the sums will be convolutions if $M \rightarrow \infty$. Consequently, the equation can be expressed in the domain of Fourier transform as

$$\tilde{\mathbf{A}}_1(\omega)\tilde{C}(\omega) + \tilde{\mathbf{A}}_2(\omega)\tilde{C}^*(-\omega) = 2\tilde{V}(\omega). \tag{87a}$$

Together with the transform of the conjugate complex of eq. (86), which is

$$\tilde{\mathbf{A}}_1^*(-\omega)C^*(-\omega) + \tilde{\mathbf{A}}_2^*(-\omega)\tilde{C}(\omega) = 2\tilde{V}^*(-\omega), \tag{87b}$$

it can be used to find the formal solution

$$\tilde{C}(\omega) = \frac{2\tilde{V}(\omega)\tilde{\mathbf{A}}_1^*(-\omega) - 2\tilde{V}^*(-\omega)\tilde{\mathbf{A}}_2(\omega)}{\tilde{\mathbf{A}}_1(\omega)\tilde{\mathbf{A}}_1^*(-\omega) - \tilde{\mathbf{A}}_2(\omega)\tilde{\mathbf{A}}_2^*(-\omega)}. \tag{88}$$

The Fourier transform of the sampled impulse response $H(\tau + kT)$, where τ indicates the sampling instant, is defined as follows

$$H_{\mathrm{eq}}(\omega) = \sum_{K=-\infty}^{\infty} H(\tau + kT)\exp(-j\omega kT), \tag{89a}$$

where

$$H(\tau + kT) = \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} H_{\text{eq}}(\omega)\exp(j\omega kT)d\omega. \tag{89b}$$

Assuming the impulse response $H(t)$ has the Fourier transform $H(\omega)$, Poisson's sum formula can be used to get the following relation between the spectra of the sampled and the continuous functions

$$H_{\text{eq}}(\omega) = \frac{\exp(j\omega\tau)}{T}\sum_{k=-\infty}^{\infty} H\left(\omega + 2\pi\frac{k}{T}\right)\exp\left(j2\pi k\frac{\tau}{T}\right). \tag{90}$$

Using the spectral density of the filtered noise in baseband

$$|N_{\text{eq}}(\omega)|^2 = \sum_{k=-\infty}^{\infty} R_{NN}(kT)\exp[-j(\omega + \omega_0)(kT)] \tag{91}$$

and eqs. (89a) and (89b), the transforms of $\tilde{\mathbf{A}}_1$, $\tilde{\mathbf{A}}_2$, and $\tilde{V}$ can be shown to be

$$\tilde{\mathbf{A}}_1(\omega) = (P_a + P_b)|H_{\text{eq}}(-\omega)|^2 + |N_{\text{eq}}(-\omega)|^2, \tag{92a}$$

$$\tilde{\mathbf{A}}_2(\omega) = (P_a - P_b)H_{\text{eq}}(\omega)H_{\text{eq}}(-\omega), \tag{92b}$$

$$\tilde{V}(\omega) = P_a H_{\text{eq}}(-\omega). \tag{92c}$$

From eq. (24) and with eqs. (67), (76), and (78) the minimum mse can be expressed as

$$\epsilon_{\text{opt}}^2 = P_a - \mathbf{C}_{\text{opt}}^t\mathbf{V} = P_a - R_e(\underline{C}^*\underline{V}). \tag{93}$$

In eq. 93, $\underline{C}^*\underline{V}$ can be viewed as the zeroth term of the convolution of the sequences $(\underline{V}_n)$ represented by $\underline{V}$ and (C_{-n}^*) represented by $\underline{C}^*$. Multiplying the spectra and transforming back to time domain yields

$$\text{Re}(\underline{C}^*\underline{V}) = \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T}\frac{1}{2}[C^*(\omega)V(\omega) + C(-\omega)V^*(-\omega)]d\omega. \tag{94}$$

Taking into account the modulation in eqs. (82) and (83), this also can be expressed in terms of $\tilde{C}(\omega)$ and $\tilde{V}(\omega)$, which are defined above.

$$\text{Re}(\underline{C}^*\underline{V}) = \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T}\frac{1}{2}[\tilde{C}(\omega)\tilde{V}(\omega) + \tilde{C}(-\omega)\tilde{V}(-\omega)]d\omega. \tag{95}$$

Combining eqs. (89a), (89b), (92a), (92b), (93), and (95) finally yields the desired expression of the minimum mse

$$\epsilon_{\text{opt}}^2 = \frac{TP_a}{2}\int_{-\pi/T}^{\pi/T}\frac{P_b[Z(\omega) + Z(-\omega)] + 1}{\begin{gathered}4P_aP_bZ(\omega) \times Z(\omega)\\ + (P_a + P_b) \times [Z(\omega) + Z(\omega)] + 1\end{gathered}}d\omega, \tag{96}$$

where

$$Z(\omega) = \left|\frac{H_{\rm eq}(\omega)}{N_{\rm eq}(\omega)}\right|^2. \tag{97}$$

In the balanced power case $P_a = P_b$, we can simplify further to yield

$$\epsilon_{\rm opt}^2 = \frac{TP_a}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{1}{2[2P_aZ(\omega) + 1]} + \frac{1}{2[2P_aZ(-\omega) + 1]}\,d\omega. \tag{98}$$

Since both terms are integrated over one full period this yields

$$\epsilon_{\rm opt}^2 = \frac{TP_a}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{d\omega}{2P_aZ(\omega) + 1}. \tag{99}$$

It can be shown that the derivative of the minimum mse eq. (96) with respect to the power in the second branch is

$$\frac{\partial\epsilon_{\rm opt}^2}{\partial P_b} = \frac{TP_a^2}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{[Z(\omega) - Z(-\omega)]^2 d\omega}{D^2}, \tag{100}$$

where D is the denominator of the integrand in eq. (96). This means the derivative is positive unless $Z(\omega)$ is symmetric in which case the former is zero. In general, a decrease in the power of the second branch will decrease the minimum mse in the first branch.

REFERENCES

1. F. Akashi, Y. Sato, and M. Eguchi, unpublished work.
2. T. L. Lim, unpublished work.
3. D. D. Falconer, "Jointly Adaptive Equalization and Carrier Recovery in 2-Dimensional Digital Communication Systems," B.S.T.J., *55* (March 1976), pp. 317–34.
4. D. D. Falconer, "Analysis of a Gradient Algorithm for Simultaneous Passband Equalization and Carrier Phase Recovery," B.S.T.J., *55* (April 1976), pp. 409–28.
5. M. S. Mueller, "Complex Valued Analysis of Cross Coupled Equalizers in Synchronous Data Receivers," (in German) Swiss Federal Inst. of Technology, Ph.D. Thesis 5711, 1976.
6. G. Sandegren, "QAM and SSB Modulation for Data Transmission over Telephone Lines Including Equalizer," Ericsson Technics, *33* (1977), pp. 247–98.
7. H. B. Voelcker, "Toward a Unified Theory of Modulation-Part I: Phase Envelope Relationships," Proc. IEEE, *54* (March, 1966), pp. 340–53.
8. G. Ungerboeck, "Theory on the Speed on Convergence in Adaptive Equalizers for Digital Communications," IBM J. Res. and Dev. (November, 1972), pp. 546–55.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

High-Speed Measurement and Control of Fiber-Coating Concentricity

By D. H. SMITHGALL and R. E. FRAZEE

(Manuscript received May 19, 1981)

A technique has been implemented to measure and control the eccentricity of lightguide fiber in transparent polymer coating materials. It is based upon a model which describes the characteristics of a forward-scattered light pattern generated by transversely illuminating coated fiber with a laser beam. The model predicts the behavior of the principal characteristics of the pattern as a function of fiber eccentricity within the coating. The implementation automatically detects and controls the position of the dominant pattern feature to maintain an average fiber-coating concentricity within 2 μm over multikilometer lengths of fiber.

I. INTRODUCTION

The concentricity of an optical fiber in a plastic coating affects the fiber strength, transmission loss, and the connectorization process. Self-centering coating techniques, such as the flexible tip applicator,[1] or hydrodynamically designed applicators,[2] have not been completely effective in maintaining the fiber centered in the coating. An alternative to such passive fiber centering is to measure and control the location of the fiber in the coating.

Eichenbaum has described a technique in which the coated fiber is illuminated by a laser beam (Fig. 1) and the symmetry of certain features of the forward scattering pattern are used to determine coating concentricity.[3] His model analyzed coatings whose refractive index is greater than that of the fiber, and eccentricities are normal to the direction of the laser beam. The effects of refractive index and relative fiber and coating diameters for the concentric case are also described. The model is limited, however, to fiber offset normal to the direction of incidence and does not consider the effect of the fiber core upon the forward-scattered pattern.

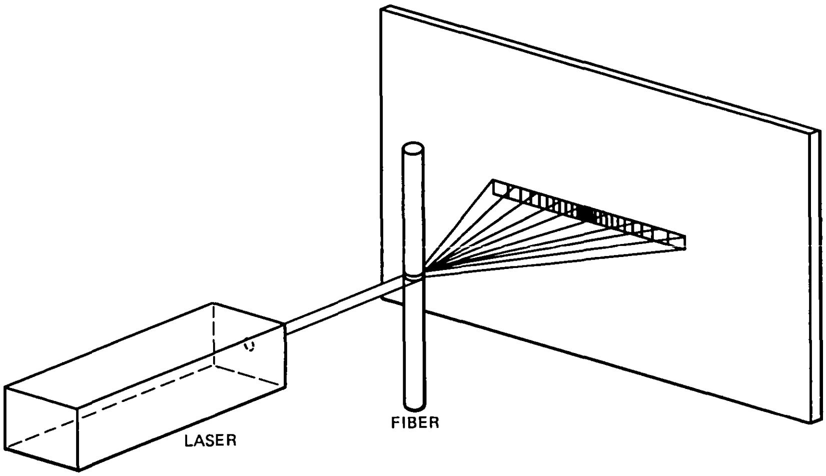


Fig. 1—Forward-scattering pattern generated from laser-illuminated fiber.

An alternate approach, first examined by Presby and subsequently described by Marcuse for silicone coatings, examines the backward-scattered pattern from a coated fiber illuminated by a laser beam.[4,5] The structure of the pattern is more complex than a forward-scattered pattern because the scattered light passes through the coated-fiber structure twice.

The technique of measuring fiber properties by forward-light scattering has been shown to be very powerful, and a similar approach could prove useful to automatically measure the fiber concentricity during the coating operation, and control the position of the fiber within the coating material.[6-9] Of the two scattering techniques described, the forward-scattering technique has the advantages that the pattern has a simpler structure, and as a practical matter, the forward-scattered pattern contains more light energy than the backward-scattered pattern, thus, requiring a smaller laser to provide sufficient information to a detector. In developing a centering control based upon forward-scattered light, the effects of both the fiber core and an arbitrary fiber eccentricity upon the structure of the scattering pattern must be considered.

II. CONCENTRIC FIBER RAY-SCATTERING MODEL

The smoothed envelope of the intensity of the forward-scattered pattern from a graded-index, multimode fiber coated with an epoxy acrylate material is shown in Fig. 2. The pattern is typical for fiber with a 50-μm core with 0.23 NA (numerical aperture), 125-μm clad

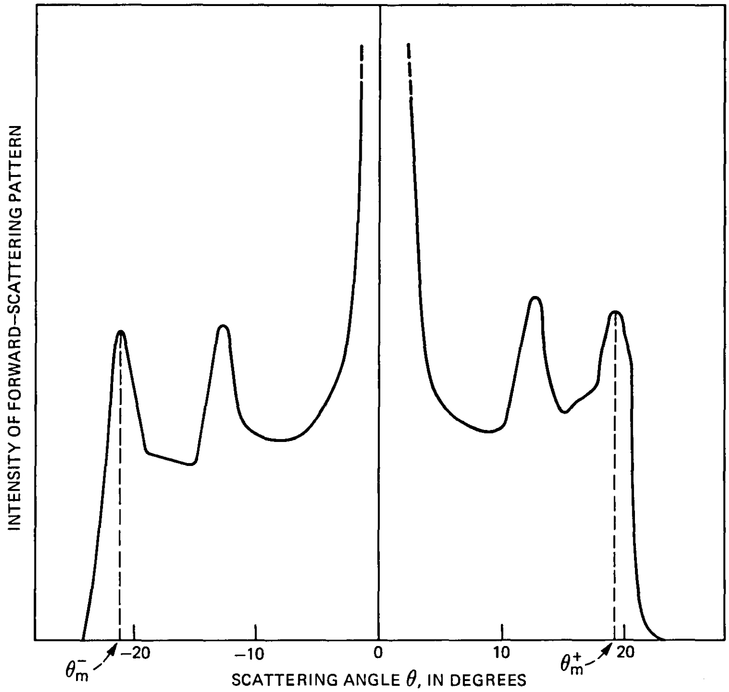


Fig. 2—Smoothed intensity of forward-scattered pattern for graded-index fiber in high-index coating.

diameter, and 250-μm coating diameter. The refractive indices of the cladding and coating are 1.457 and 1.539, respectively, at 0.6328 μm.

The structure of this pattern can be analyzed by considering the three-level concentric model shown in Fig. 3, where the radii are

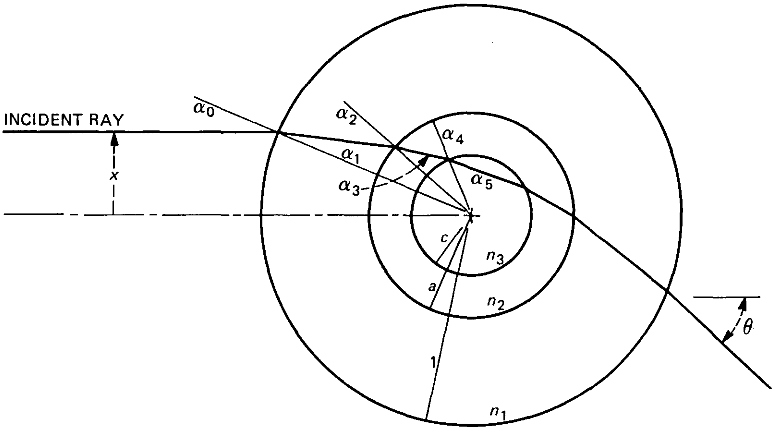


Fig. 3—Three-level concentric model.

normalized to the coating radius, and the refractive indices are assumed constant. The scattering angle θ for a ray striking the coating a distance x from the optical axis can be determined from the formula of Bouger,[10] when the ray passes through the core, as $x \leq n_2\, c$, and

$$\theta = 2\ (\alpha_0 - \alpha_1 + \alpha_2 - \alpha_3 + \alpha_4 - \alpha_5), \tag{1}$$

where

$$\begin{aligned} x &= \sin \alpha_0 \\ &= n_1 \sin \alpha_1 = n_1\, a \sin \alpha_2 \\ &= n_2\, a \sin \alpha_3 = n_2\, c \sin \alpha_4 \\ &= n_3\, c \sin \alpha_5. \end{aligned} \tag{2}$$

The exit angle, θ, for rays refracted through the cladding, $n_2\, c < x \leq n_1\, a$, is given by

$$\theta = 2\ (\alpha_0 - \alpha_1 + \alpha_2 - \alpha_3), \tag{3}$$

where the angles are defined in eq. (2).

In the case of a high-index coating, eq. (3) holds over the range $n_2\, c \leq x < n_2\, a$. In the range $n_2\, a \leq x < n_1\, a$, rays are reflected from the surface of the fiber, and the exit angle is given by

$$\theta = 2\ (\alpha_0 - \alpha_1 + \alpha_2 - \pi/2). \tag{4}$$

The exit angle for rays refracted through the coating only is

$$\theta = 2\ (\alpha_0 - \alpha_1). \tag{5}$$

For the parameters given above, the relationship between the scattering angle and the normalized incident ray position is shown in Fig. 4 for x between 0 and 1. For x in the range 0 to -1, $\theta(-x) = -\theta(x)$. Ray positions less than x_c pass through the core of the fiber; rays between x_c and x_2 pass through the fiber clad and coating. Ray positions greater than x_2 pass through the coating only.

Over the angular range θ_c to θ_u, there is an interference pattern generated by rays passing through both the core and cladding of the fiber. This interference is observed as an intensity maxima near ±10° in Fig. 2. The structure of the interference pattern depends upon the core index gradient,[11] and for typical near-parabolic index gradients, there exists a single intensity maximum whose location is nearer θ_c than θ_u. The angle θ_c is independent of core structure, but depends upon core diameter. For single-mode fibers, $\theta_c \rightarrow 0$, and there will be no measurable intensity peak in the scattered-light pattern corresponding to the core-clad interface.

The intensity maxima near ±20° in Fig. 2 result from the focusing

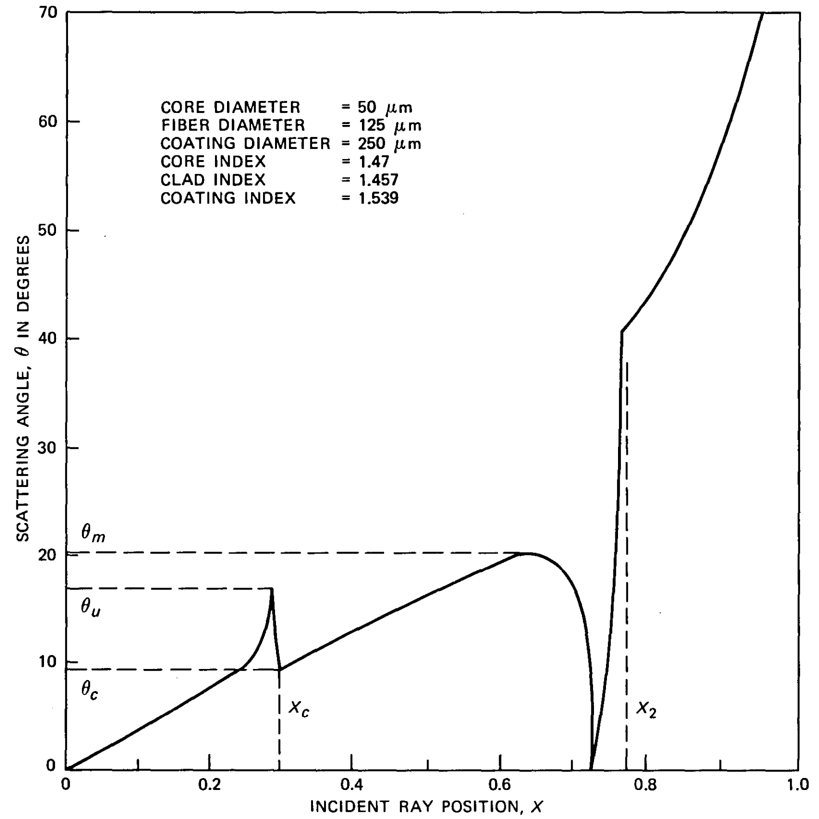


Fig. 4—Ray-scattering pattern for step-index, multimode fiber with high-index coating.

of rays passing near the clad-coating interface. The location of these caustics in Fig. 4 is the angle θ_m, where $d\theta/dx = 0$. If the slope of the inverse scattering curve $dx/d\theta$ is interpreted as the relative energy density of the scattered light, then $dx/d\theta \to \infty$ is a region of focused energy which will result in an intensity maximum. Conversely, the energy density of the forward-scattered pattern for angles greater than θ_m is very low. This is observed in Fig. 2 as an abrupt reduction in detected energy near the peak intensity at θ_m.

III. ECCENTRIC FIBER RAY-SCATTERING MODEL

The influence of fiber eccentricity within the coating upon the forward-scattered pattern may be determined by examining its effect upon the position of the intensity maximum at θ_m. The maxima at θ_c may or may not exist, depending upon the structure of the fiber. The results in the preceding section show that for typical single or multimode fibers, the effect of the core and the effect of the fiber/coating relationship upon the scattering pattern are distinct and separate. Therefore, in the subsequent development, a homogeneous fiber structure will be assumed with no loss of generality.

Consider the general nonconcentric structure of Fig. 5 in which the center of the fiber is displaced from the center of the coating by a distance d at an angle ϕ from incidence. The radii of the fiber and coating have been normalized to "a" and 1, respectively. The refractive indices are n_2 and n_1, respectively. Using Snell's Law and the geometric construction of Ref. 5, expressions for the angles describing the ray path can be determined by modifying the equations describing backscattered rays to describe the refraction of forward-scattered rays.

$$\alpha_0 = \sin^{-1} x. \tag{6}$$

$$\alpha_1 = \sin^{-1}\left(\frac{1}{n_1}\sin\alpha_0\right). \tag{7}$$

$$\gamma_1 = \frac{1}{a}\sin\alpha_1 - \frac{d}{a}\sin(\alpha_0 - \alpha_1 + \phi). \tag{8}$$

$$\alpha_2 = \sin^{-1}\gamma_1. \tag{9}$$

$$\gamma_2 = \frac{n_1}{n_2}\sin\alpha_2. \tag{10}$$

$$\alpha_3 = \sin^{-1}\gamma_2. \tag{11}$$

$$\alpha_4 = \alpha_3. \tag{12}$$

$$\alpha_5 = \alpha_2. \tag{13}$$

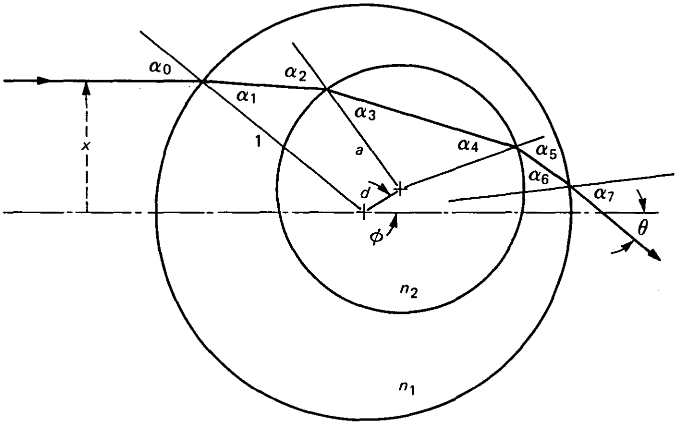


Fig. 5—Refracted ray in two-level structure with offset core.

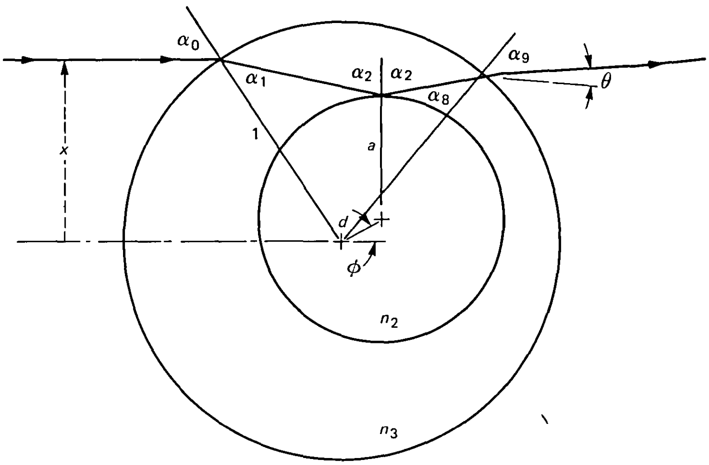


Fig. 6—Internal reflected ray in two-level structure with offset core.

$$\alpha_6 = \sin^{-1} \{a \sin \alpha_2 + d \sin[\alpha_0 - \alpha_1 + 2(\alpha_2 - \alpha_3) + \phi]\}. \tag{14}$$

$$\alpha_7 = \sin^{-1} (n_1 \sin \alpha_6). \tag{15}$$

These equations represent an arbitrary ray passing through the coated fiber. Certain conditions, however, yield internally reflected rays or rays passing only through the coating material. Hence, special cases of eqs. (6) to (15) need to be considered in determining the exit angle $\theta(x)$.

For a ray which passes through the fiber and coating,

$$\gamma_1, \gamma_2 < 1, \quad \text{and} \tag{16}$$

$$\theta = (\alpha_0 - \alpha_1) + 2\,(\alpha_2 - \alpha_3) + (\alpha_7 - \alpha_6). \tag{17}$$

For a ray which passes through the coating only,

$$\gamma_1 > 1, \quad \text{and} \tag{18}$$

$$\theta = 2\,(\alpha_0 - \alpha_1). \tag{19}$$

For a ray which passes through the coating and is reflected from the fiber surface (Fig. 6),

$$\gamma_1 < 1.$$

$$\gamma_2 > 1. \tag{20}$$

$$\gamma_3 = \alpha_0 - \alpha_1 + \phi + 2\alpha_2 - \pi/2. \tag{21}$$

$$\alpha_8 = \sin^{-1} (a \sin \alpha_2 + d \sin \gamma_3). \tag{22}$$

$$\alpha_9 = \sin^{-1} (n_1 \sin \alpha_8). \tag{23}$$

$$\theta = \alpha_0 + 2\alpha_2 + \alpha_9 - \alpha_1 - \alpha_8 - \pi. \tag{24}$$

There is one additional case occurring when $n_1 \sin \alpha_6 > 1$ or $n_1 \sin \alpha_8 > 1$, which represents total internal reflection at the point where the ray should exit the coating.

Equations (6) to (24) can be used to generate ray-scattering angles $\theta(x)$. By appropriate choices of d and ϕ, scattering angles for arbitrary fiber eccentricity may be examined.

IV. RESULTS FOR TWO-LEVEL MODEL

For the purpose of investigating the effect of fiber eccentricity, consider the model wherein the ratio of core and outer diameter $a = 0.5$, and the refractive indices of the core and outer layers are 1.457 and 1.539, respectively. These index values correspond to those of fused silica and an epoxy acrylate coating material. The effects of changing the material indices or the fiber/coating diameter ratio were examined in Ref. 3 for a concentric structure.

The effect of offset "d" along the direction of incidence ($\phi = 0$), and normal to the direction of incidence ($\phi = \pi/2$*) upon θ_m is shown in Fig. 7. For each case, there is a θ_m corresponding to refraction of the light rays to both sides of the axis of the incident light. The angle θ_m^+ results from incident rays striking the fiber at normalized positions in the range 0 to +1, and the angle θ_m^- results from incident rays striking the fiber for x between 0 and −1. There is a symmetry such that θ_m^+ and θ_m^- exchange roles for $\phi = \phi + \pi$. This is illustrated in Fig. 7 by showing the dependence of θ_m upon negative d.

For the case $\phi = 0$, the loci of θ_m^+ and θ_m^- are symmetric about the vertical axis with increasing separation as the fiber is displaced away from the laser source. For d varying from −0.2 ($d = 0.2$ @ $\phi = \pi$) to $d = 0.2$, θ_m varies from 18 to 26 degrees.

For the case $\phi = \pi/2$, the maxima θ_m^+ and θ_m^- are symmetric about the axis only for $d = 0$. For increasing d at $\phi = \pi/2$, θ_m^+ increases and θ_m^- decreases. Furthermore, the separation increases as d increases, from 41 degrees at $d = 0$ to 54 degrees at $d = 0.2$. For $\phi = 3\pi/2$, the roles of θ_m^+ and θ_m^- are reversed.

At a fiber displacement of approximately $d = 0.2$, the ray generating θ_m^+ is totally internally reflected at the point where it would exit the coating material, and the scattering-pattern feature corresponding to θ_m^+ does not exist.

For the more general case of an arbitrary ϕ, define the separation of

* In the subsequent discussion, the angle of the fiber offset is given in radians and the ray-scattering angle is given in degrees.

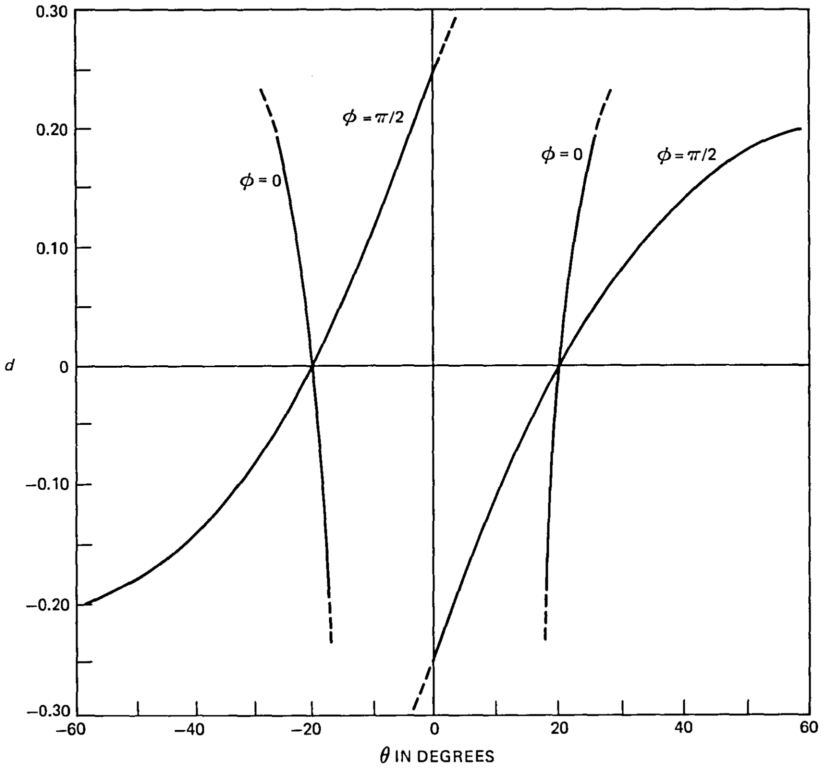


Fig. 7—Locus of intensity maxima θ_m as a function of spatial offset for two angles of offset.

the scattering-pattern intensity maxima as

$$\Delta = \theta_m^+ - \theta_m^-. \tag{25}$$

It has been illustrated in the previous discussion that Δ as a function of both d and ϕ will provide an indication of the composite behavior of the scattering pattern. In addition, Δ is a factor in the measured eccentricity.

$$E = \frac{\theta_m^+ + \theta_m^-}{\Delta}. \tag{26}$$

Figure 8 shows a plot of Δ as a function of the offset orientation, ϕ, with the offset magnitude, d, as a parameter. The pattern is cyclic in ϕ as expected. For $d > 0.15$, $\Delta \rightarrow \infty$ near the orientation $\phi = \pm\pi/4$. The intensity maxima θ_m^+ does not exist for $d > 0.15$ at this angular orientation. The fiber is sufficiently close to the surface of the coating that the cumulative refraction of the light rays remains monotonic with ray height, x, as in the case of the homogeneous medium.

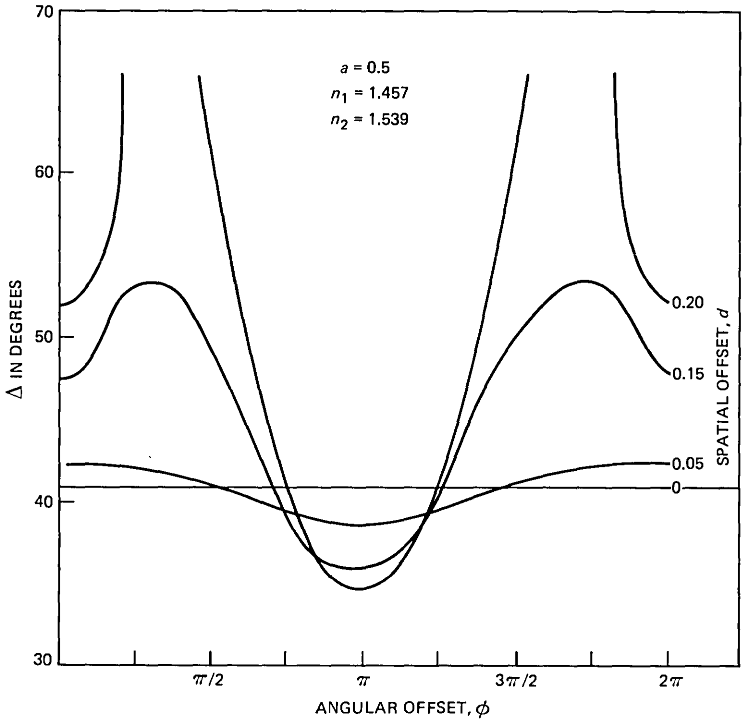


Fig. 8—Separation (Δ) of intensity maxima as a function of spatial and angular offset with respect to incident laser beam.

Therefore, there is no region in which scattered energy is concentrated for eccentricities greater than 15 percent of the coating radius. This condition limits the range over which a high-intensity feature could be used to detect fiber eccentricity. Subsequently, it will be shown that this is a minor restriction.

V. OPTICAL DESIGN FOR COATING CONCENTRICITY MONITOR

The model developed in Section III has shown that if the intensity maxima at θ_m^+ and θ_m^-, or, correspondingly, the abrupt change in intensity at the edges of the pattern, could be detected, their position would provide a monotonic measure of the eccentricity of the fiber within the coating. By illuminating the coated fiber in two orthogonal directions, the degree of fiber eccentricity is uniquely established. The fiber may be centered within the coating by establishing symmetry within each of the two orthogonal patterns.

The optical layout for the automatic detection system is shown in Fig. 9. The beam for a 1-mW HeNe laser is attenuated by a neutral density filter and split into two equal-intensity beams which are each

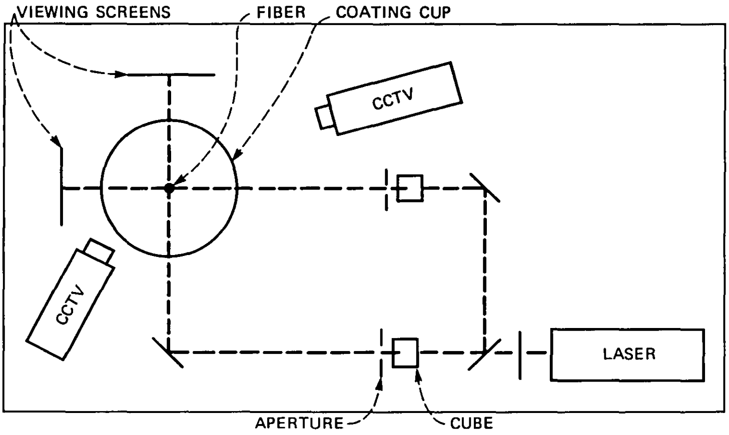


Fig. 9—Optics layout for automatic centering unit.

reflected to intersect at the fiber. Located approximately four centimeters behind the fiber is a viewing screen of white bond paper. The paper provides a high contrast background, and partially diffuses the scattering pattern, eliminating the fine structure corresponding to interference of refracted and reflected rays. The scattering pattern, as viewed on the screen, appears as a bright bar on either side of the very bright central spot (Fig. 10). At the end of each bar is a spot of slightly greater intensity corresponding to the outer peaks of Fig. 2. For multimode fiber, a second bright spot may be discerned between the central peak and the edge of the bright region.

The scattering pattern is viewed by a closed circuit television (CCTV) camera through a 0.633-μm interference filter. The filter permits the device to operate in normal room lighting conditions, while only the scattering pattern is observed. The camera is mounted such that the scattering pattern is crossed by a multiplicity of scan lines. Thus, along each scan line the intensity of the pattern is sampled once. By extracting this information from the CCTV output, the scattering pattern can

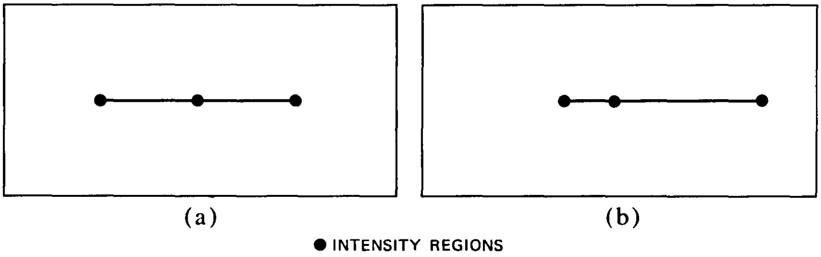


Fig. 10—Forward-scattering patterns as seen on viewing screen. (a) Fiber centered in coating. (b) Fiber eccentric in coating.

be reconstructed as in Fig. 2. Furthermore, the location of important features within the pattern can be determined by counting the number of samples, or scan lines, between features. With the full-field view of CCTV, the detector is insensitive to tilt in the scattering pattern resulting from an angularly-misaligned camera or fiber.

To determine the centered position of the fiber within the coating, the forward-light-scattered pattern is made symmetric with respect to a central intensity maximum of the pattern, generated by the incident laser beam. If the coated fiber is not centered in the beam, a slight offset in the pattern with respect to the central peak exists, and when corrected by moving the coating applicator, results in an off-center coating. In addition, centering the coated fiber in the laser beam maximizes the energy transferred from the beam into the scattering pattern, and ensures that the two wings of the scattering pattern will have equal energy, maximizing the sensitivity of the measurement system.

To align the laser beams to the coated fiber, a rotatable cube, mounted on the shaft of a small servo motor is located in each optical path. An aperture placed between the cube and the viewing screen eliminates spurious scattering effects from the corners of the cube.

The optical components, screens, and CCTVs are mounted on a platform with the coating cup. The platform is adjustable with respect to a fixed base plate which, in turn, is attached to the fiber draw tower. Thus, the coated fiber structure is fixed with respect to the illuminating laser beams, and the fiber moves within the coating, limited by the orifice in the tip of the applicator.

VI. VIDEO SIGNAL PROCESSING

Two CCTV cameras are mounted such that each views one of the orthogonally mounted viewing screens and cuts the projected forward-scattering pattern with a plurality of horizontal scan lines. The camera outputs are treated independently of each other but via identical circuitry (Fig. 11). For each, the composite video signal is separated into the video and synchronization components.

The video portion of the signal is input to two separate integrator circuits. One integrator sums the collective value of all the video pulses contained in one complete vertical field. Therefore, the output voltage is proportional to the total energy contained in the forward-scattered pattern. This signal, input to the microprocessor through an analog-to-digital (A/D) converter, is used to center the laser beam on the coated fiber. The update rate of the signal strength measurement is equal to the vertical field scan rate (60 Hz).

The second integrator sums the signal levels contained in each horizontal scan line. The output, controlled by line synchronization

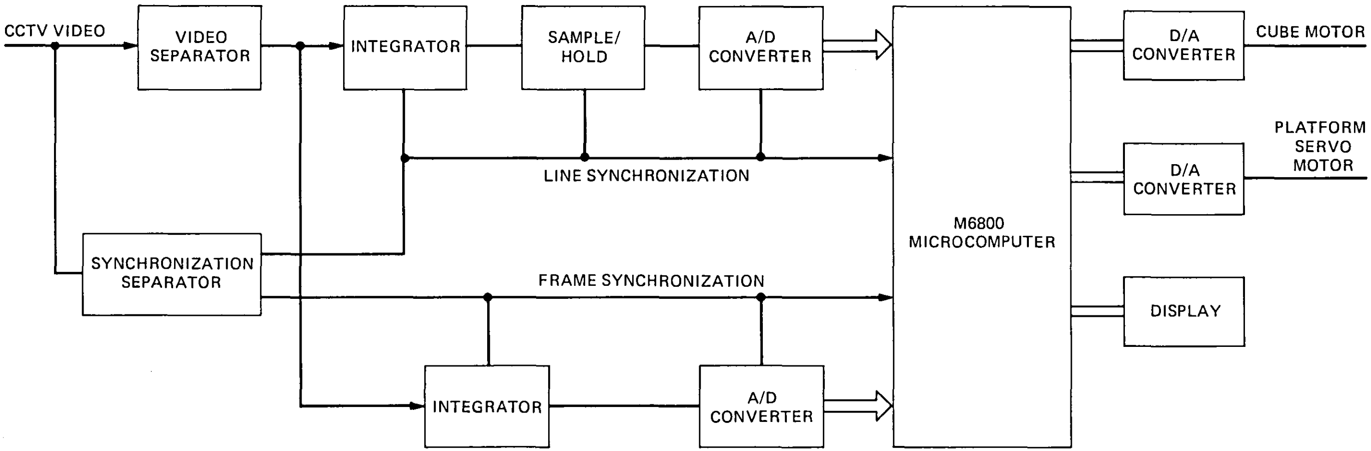


Fig. 11—Video signal processing circuit.

pulses to form a serialized boxcar representation of the envelope, is input to a high-speed A/D converter which converts each scan line level to eight bits of binary data, which are input to the microprocessor. The video scan line signal is conditioned at the horizontal scan rate of approximately 63 μs.

VII. MEASURE OF FIBER ECCENTRICITY

Once the laser beam has been centered on the coated fiber structure the envelope of the scattering pattern is examined to determine fiber eccentricity. Establishing a reference on the frame synchronization pulse, and subsequently, on each of the line synchronization pulses, 230 consecutive samples are read from the A/D converter. Each sample represents the intensity of the scattering pattern detected during line scan.

The data is scanned to locate the intensity changes which correspond to the edges of the laser beam and the forward-scattered pattern. From the scattering model it is known that an intensity maximum must occur near the locations at which the intensity goes to zero.

Let $\{X_i,\ i = 1, \cdots, n\}$ be the intensities of the samples $\{i\}$. A maxima is defined as:

$$\begin{aligned} M_i\colon X_i - X_{i+1} &\geq 0 \\ X_i - X_{i-1} &> 0. \end{aligned} \tag{27}$$

The two maxima corresponding the θ_m^+ and θ_m^- may be determined by searching the data near the locations of the edges of the pattern. From the relative locations of the central intensity maximum corresponding to the laser beam and the two maxima, the angles θ_m^+ and θ_m^- can be determined and the eccentricity measure

$$e = \theta_m^+ + \theta_m^- \tag{28}$$

defined. Equation (28) represents a monotonic measure of fiber eccentricity, since it can take on both positive and negative values. Furthermore, the function can be used directly to generate a corrective control signal to center the fiber in the coating.

VIII. CONTROL OF FIBER POSITION

The coating material, with a viscosity of ~50 poise at room temperature, presents a strong dampening effect on the motion of the fiber. Consequently, when the cut is moved with respect to the fiber, a settling time of several seconds is observed. Conversely, once the fiber is centered within the coating cup it will tend to remain centered such that readjustment is only occasionally required. It has been observed that fifteen minutes may be required to reach this equilibrium state, with frequent position adjustments necessary prior to reaching equilibrium.

The algorithm implemented for fiber positioning is a deadband control, where for e given by eq. (28), the control u is:

$$\begin{aligned} u &= 0 && 1 \leq e \leq 1 \\ u &= u_0 && e > 1 \\ u &= -u_0 && e < -1. \end{aligned} \tag{29}$$

The value of u_0 is selected large enough to overcome friction and hysteresis in the servo motors and platform transport mechanism, yet, small enough to account for the slow dynamic response of the fiber moving through the coating material. The deadband about the point of pattern symmetry prevents the servo motor from constantly responding to small disturbances resulting from short-term fluctuations in eccentricity, or single quantum steps in the computation of the eccentricity function [eq. (25)]. The normal system condition for a process in equilibrium is one of quiescence.

For the typical case of an epoxy acrylate-coated silica fiber, with 125-μm-fiber diameter and 250-μm-outer diameter, a scattering pattern with the intensity maxima at ±20 degree is generated. A pattern nearly 4 cm wide projected onto the viewing screen is intersected by approximately 60 horizontal scan lines on the CCTV. With the algorithm previously described, centering is controlled within ±1 part in 30, or ±2/3 degree. From Fig. 7, the sensitivity to fiber offset normal to the incident laser beam is 1 percent of the coating radius per degree shift in concentricity. Therefore, the control of the average concentricity is ±1 μm.

Localized fluctuations may exceed this value. Experience has shown that long lengths of fiber may be coated with coating concentricities within 1.5 percent of the coating diameter.

IX. CONCLUSION

A multilevel model for a lightguide fiber coated with material with refractive index greater than that of the fiber, has been used to determine characteristics of the forward-scattered pattern with respect to the coating concentricity. The results of the model have been used as a basis for a system which detects and controls the average position of the fiber in the coating within 2 μm. Experimentally, the fiber-coating concentricity has been maintained within 1.5 percent over multikilometer lengths of fiber. With this control, the predicted limitation in detectable eccentricity of 7.5 percent of coating diameter is not a significant factor.

REFERENCES

1. A. C. Hart and R. V. Albarino, "An Improved Fabrication Technique For Applying Coatings to Optical Fiber Waveguides," Technical Digest of Topical Meeting on Optical Fiber Transmission II, Williamsburg, Virginia, February 22–4, 1977, Paper TuB2.
2. P. W. France, P. L. Dunn, and M. H. Reeve, "Plastic Coating of Glass Fibers and Its Influence on Strength," Fiber and Integrated Optics, *2,* No. 3–4, pp. 267-86.
3. B. R. Eichenbaum, "The Centering of Optical Fiber Coatings by Monitoring Forward Scattering Patterns—Theory and Practice," B.S.T.J., *59,* No. 3 (March 1980), pp. 313–32.
4. H. M. Presby, "Geometrical Uniformity of Plastic Coatings on Optical Fibers," B.S.T.J., *55,* No. 10 (December 1976), pp. 1525–37.
5. D. Marcuse and H. M. Presby, "Optical Fiber Coating Concentricity: Measurement and Analysis," Appl. Optics, *16,* No. 9 (September 1977), pp. 2383–90.
6. D. H. Smithgall, L. S. Watkins, and R. E. Frazee, "High Speed Non-contact Fiber Diameter Measurement Using Forward Light Scattering," Appl. Optics, *16,* No. 9 (September 1977), pp. 2395–402.
7. L. S. Watkins, "Scattering From Side—Illuminated Clad Glass Fibers for Determination of Fiber Parameters," J. Opt. Soc. of Am., *64* (June 1974), pp. 767–72.
8. M. A. G. Abushagur and N. George, "Measurement of Optical Fiber Diameter Using Fast Fourier Transform," Appl. Opt., *19,* No. 12 (June 15, 1980), pp. 2031–3.
9. L. S. Watkins and R. E. Frazee, "High Speed Measurement of Core Diameter of a Step Index Optical Fiber," Appl. Opt., *19,* No. 22 (November 15, 1980), pp. 3756–62.
10. M. Born and E. Wolf, *Principles of Optics*, New York: Pergamon Press, 1964.
11. L. S. Watkins, "Laser Beam Refraction Traversely Through a Graded Index Preform to Determine Refractive Index Ratio and Gradient Profile," Appl. Opt., *18,* No. 13 (July 1, 1979), pp. 2214–22.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

Simultaneous Transmission of Speech and Data Using Code-Breaking Techniques

By R. STEELE and D. VITELLO

(Manuscript received April 19, 1981)

A system whereby speech is used as a data carrier is proposed. The speech, sampled at 8 kHz, is divided into blocks of N samples, and provided the correlation coefficient and mean square value of the samples exceed system thresholds, data is allowed to be transmitted. If the data is a logical 0, the samples are sent without modification; however, if a logical 1 is present, frequency inversion scrambling of the samples occurs. The receiver performs the inverse process to recover both the speech and data. Data rates of 700 b/s were achieved without data errors or speech distortion via an ideal channel. The effects of additive background and channel noise were investigated, and the system was shown to operate at 126 b/s with no data errors when the additive noise was as high as 10 dB below the mean square value of the speech signal.

I. INTRODUCTION

There are numerous schemes[1,2] for analog scrambling of speech signals, but they all require a scrambling key. For example, we may sample the speech at a rate in excess of its Nyquist rate, parcel the samples into blocks, and rearrange the blocks prior to transmission. This rearrangement of the blocks breaks up the rhythm in the speech making it difficult for an eavesdropper to comprehend the conversation. The shuffling of the block positions is done under the auspices of the scrambling-key, and provided the receiver knows this key and, hence, the descrambling key, the blocks of speech can be correctly repositioned and made intelligible to the desired recipient.

It is not our purpose to describe the numerous scrambling techniques, but rather to suggest a method whereby speech and data can be transmitted simultaneously over the channel by using scrambling strategies. The principle is very simple. The scrambling key becomes the data to be transmitted. The receiver adopts the role of codebreaker. Every time the receiver correctly guesses the key and breaks the code, it recovers both the speech and the data. For the scheme to have any significance, the receiver must break the code successfully at nearly every attempt. Therefore, we must select scrambling keys which are easy to break, and this means that we are not aiming for speech privacy (although a degree of privacy may be achieved as a by-product). The scrambling process is, therefore, a catalyst which enables the data to be transmitted.

At first sight, it might appear that we are getting something for nothing. With care we can arrange for the data to be transmitted at negligible error rate, the speech faithfully recovered, and a small bandwidth expansion of the transmitted signal compared to the original speech. These rewards are derived from the inherent redundancy in the speech signal. Indeed, we emphasize that the method will work with any signal that has correlative features, such as speech, television, facsimile, and analog-plant control signals, like pressure and temperature variations, etc.

II. SIMULTANEOUS SPEECH AND DATA TRANSMISSION USING FREQUENCY INVERSION SCRAMBLING

As a demonstration of the concept, we describe the transmission of data using the simplest of scrambling methods, frequency inversion. In this method, speech, band-limited to 3.4 kHz, is sampled at 8 kHz and N samples are processed at a time. Let us represent these samples as

$$S_1 = x_0, x_1, x_2, \cdots, x_{N-1}\,. \tag{1}$$

To invert the frequency components associated with these N samples, all we need to do is to alter the polarity of every other sample,[3,4] namely,

$$S_2 = x_0, -x_1, x_2, -x_3, \cdots, -x_{N-1} \tag{2}$$

N even.

In frequency-inversion scrambling (FIS), sequence S_2 would always be transmitted, but in our scheme, sequence S_2 is used when we decide to transmit data and, further, the data is a logical 1. Observe only one bit per N speech samples is transmitted.

To minimize the number of bits received in error, we proceed as follows. The calculation

$$\rho = \frac{\sum_{i=0}^{N-2} x_i x_{i+1}}{\sum_{i=0}^{N-1} x_i^2} \tag{3}$$

is made on the original speech sequence S_1 and called here the correlation coefficient, and the mean square value

$$\sigma_x^2 = \frac{1}{N}\sum_{i=0}^{N-1} x_i^2 \tag{4}$$

in the block of speech samples is also found. Notice that the correlation coefficient ρ_s of the scrambled sequence S_2 is $-\rho$. Figure 1 shows the block diagram of the system. Mean square value σ_x^2 and correlation coefficient ρ are compared with system threshold parameters T and K in comparators COMP 1 and COMP 2, respectively. Parameters T and K may be selected such that $\sigma_x^2 > T$ and $\rho > K$ generally implies the absence of unvoiced speech and silence, assuming there is no additive background noise. This strategy aids in reducing the number of received bit errors when transmitting through noisy channels. Later we will give details of how T and K are selected.

Data is only transmitted when the Boolean equation

$$y = C_1C_2 \tag{5}$$

is a logical 1, where

$$C_1 = \begin{cases} \text{logical 1 } \textit{if } \sigma_x^2 \geq T \\ \text{logical 0 } \textit{if } \sigma_x^2 < T \end{cases}$$

and

$$C_2 = \begin{cases} \text{logical 1 } \textit{if } \rho \geq K \\ \text{logical 0 } \textit{if } \rho < K \end{cases}.$$

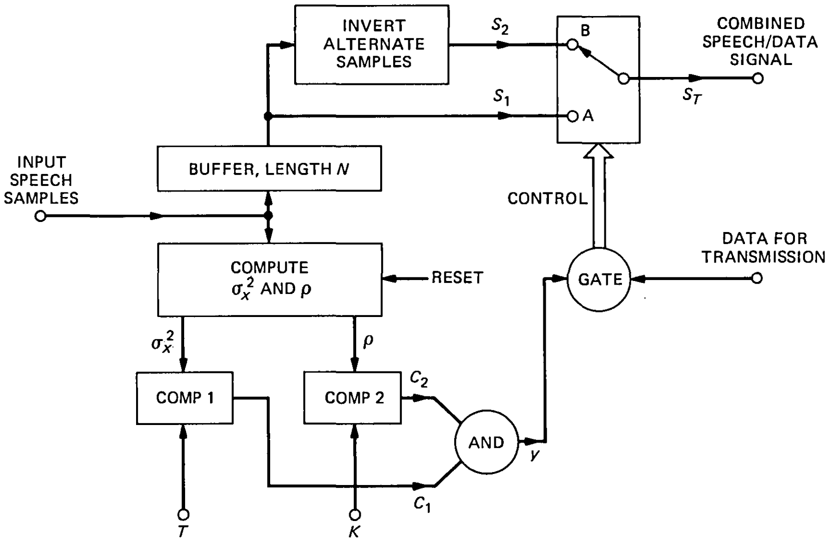


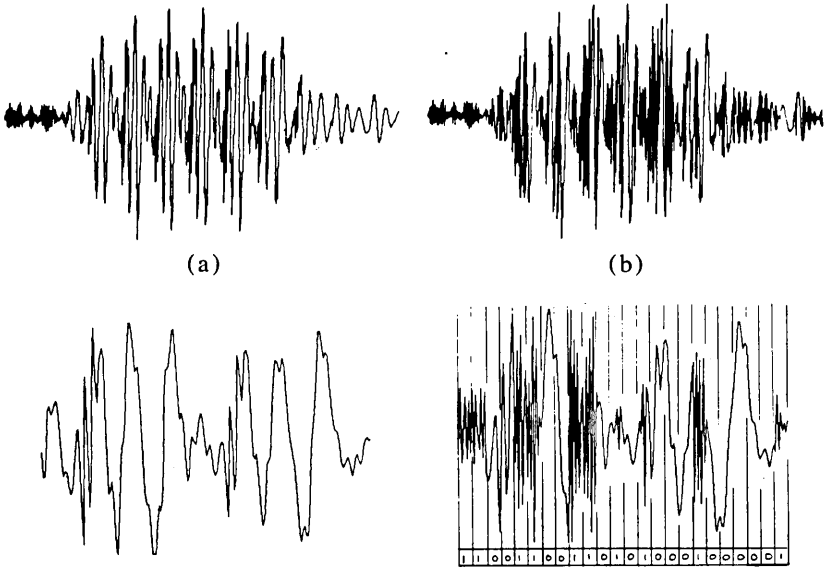
Fig. 1—Block diagram of the SSDT/FIS system at the transmitting end for the simultaneous transmission of speech and data.

The data sequence is allowed to select S_1 or S_2 if eq. (5) is satisfied. Thus, if $y = 1$, the switch in Fig. 1 is set in position A or B if the data is logical 0 or 1, respectively, i.e., a sequence S_T is generated according to

$$S_T = \begin{cases} S_1, \text{ data} = \text{logical } 0 \\ S_2, \text{ data} = \text{logical } 1 \end{cases}. \tag{6}$$

Whenever $y = 0$, $S_T = S_1$, the unscrambled speech. The sequence S_T is appropriately filtered and transmitted as the combined speech and data signal.

To illustrate the effect of the imposition of data on the speech signal, we show the waveforms in Fig. 2. In (a) and (b) of Figure 2 an arbitrary segment of speech and the corresponding transmitted signal containing data for 120 blocks are shown, respectively. The envelope of the signal is barely changed, and blocks conveying zeros are not scrambled. Hence, the transmitted signal is perceived as a distorted version of the input speech—intelligible but tiresome to a listener. A smaller segment of the original speech signal, and the resulting transmitted signal for the logical values of the data shown, are displayed, respectively, in (c) and (d) of Fig. 2. There are now only 24 blocks, and the frequency inversions are apparent when the data is a logical 1.



(a)

(b)

(c)

(d)

Fig. 2—Arbitrary segments of speech are shown in (a) and (c), and the corresponding transmitted signals are displayed in (b) and (d), respectively, $N = 8$, $T = 0$, $K = 0.6$. The logical values of the data signal are shown for the transmitted signal (d).

The signal emerging from the transmission channel is sampled at 8 kHz to give $\hat{S}_T$, where a caret (^) above the symbol signifies its presence at the receiver. In the absence of channel impairments $\hat{S}_T = S_T$, the power $\hat{\sigma}_x^2$ and correlation coefficient $\hat{\rho}$ of the sequence $\hat{S}_T$ in the block of N samples is computed according to eqs. (3) and (4). The operations associated with eq. (5) are implemented, and the following processes are performed until a decision is reached.

(*i*) If $\hat{y}$ is a logical 1, data is assumed to be transmitted of value logical 0, and $\hat{S}_T = \hat{S}_1$ is the recovered speech sequence.

(*ii*) If $\hat{y}$ is a logical 0, data may or may not be present. To determine whether data is present, every other sample in $\hat{S}_T$ is inverted and the scrambled correlation coefficient $\hat{\rho}_s$ is computed. Then,

(a) if $\hat{y}$ remains a logical 0, it is decided that no data was sent. The recovered speech sequence is, therefore, the original received sequence $\hat{S}_T$.

(b) if $\hat{y}$ becomes a logical 1, it is decided that data is present of value logical 1, and the recovered speech sequence is the scrambled $\hat{S}_T$ sequence.

Observe that if the conditions are not correct for the conveyance of data, or if a logical 0 is transmitted, the speech is dispatched without being scrambled. Only when a logical 1 is transmitted is scrambling performed, and this is done twice, once at the transmitter and once at the receiver. Should a data error occur, the speech at the output of the receiver may be erroneously scrambled. The resulting error samples in the block of length N have a rate of 4 kHz, and magnitudes double that of the original speech samples.

III. PERFORMANCE PARAMETERS FOR DATA TRANSMISSION

From a data transmission point of view we are interested in the transmitted bit rate (TBR) and the total bit error rate (TBER). Data will only be transmitted when y of eq. (5) is a logical 1, and the efficiency η of the system to transmit data is given by

$$\eta = \frac{\text{actual data rate}}{\text{possible data rate}} \tag{7}$$

from which

$$\text{TBR} = \frac{\eta f_s}{N}, \tag{8}$$

where f_s is the sampling rate of the speech signal. Error bits are those bits generated incorrectly at the receiver, and the number of bit errors per second is the TBER. Let the measure of the deficiency of the system that results in erroneous data at the output of the receiver be known as the data transmission deficiency,

$$\lambda = \frac{\text{data error rate}}{\text{possible data rate}}. \tag{9}$$

Then

$$\text{TBER} = \frac{\lambda f_s}{N} \tag{10}$$

or

$$\text{TBER} = \text{BER} + \text{FBR}, \tag{11}$$

where BER is the conventional bit-error rate that relates to those bits transmitted that were erroneously received. The term FBR is the false-bit rate that is associated with the generation of bits at the receiver when none were actually transmitted, and the declaration at the receiver that no bits were transmitted when they really were. Representing the states when the transmitter does not transmit data, and when the receiver deems that no data was transmitted, by the symbol −1, and using the logical data symbols of 1 and 0, we are able to construct Table I, which shows all the possible data-error conditions.

Let us consider the case of no additive noise to the speech input signal, and an ideal channel. In this case, the false bits are always a logical 1 and occur when no data (−1) was transmitted. This is state A in Table I. These errors occur when the power in the block is above the threshold, $\sigma_x^2 \geq T$, and the correlation ρ is below its threshold, $\rho < K$, prohibiting transmission of data. Now K is a positive number, and the bit error will occur if ρ is negative having a magnitude K_1, say, that is greater than K. At the receiver, $\hat{S}_T = \hat{S}_1$, $\hat{y}$ = logical 0 and, hence, the received sequence is scrambled. Because the correlation

Table I—Data error table and output speech status

	Data Status		Recovered Speech Status
State	TX	RX	at RX
A	−1	1	I
B	−1	0	C
C	1	0	I
D	1	−1	I
E	0	1	I
F	0	−1	C

Note: Logical states of the data are represented by 1 and 0. When no data is sent, or no data received, −1 is used. When the output speech at the receiver for the block of N samples is correct, the symbol C is used; when it is scrambled, i.e., frequency inverted, I is used.

coefficient of the scrambled sequence is $\hat{\rho}_s = -\hat{\rho} = +K_1$, and $K_1 > K$, $\hat{y}$ is now a logical 1, and data is deemed to be present having a value logical 1. Thus, the probability of a false bit being generated is very low, being the joint probability that $\sigma_x^2 \geq T$ and $\rho < -K$.

When the speech signal is in a noisy environment, the symbols $x_{(.)}$ representing speech in eqs. (1) to (4) are replaced by $x'_{(.)} = x_{(.)} + n_{(.)}$, where $n_{(.)}$ is the noise component and the $'$ above the symbols means noise contamination. The effect of the noise is to increase $\sigma_{x'}^2$ and decrease ρ', and as both $\sigma_{x'}^2$ and ρ' must exceed their thresholds [see eq. (5)] for data to be transmitted, the TBR decreases. Provided the channel is ideal, the TBER will depend on the correlative properties of the received speech, and the only source of errors derives from state A, i.e., TBER = FBR.

When clean speech is used and the channel is noisy, the TBR is unaffected. However, the TBER increases with channel noise power because the noise decorrelates the received signal, causing the receiver to sometimes erroneously presume that no data was transmitted. Thus, states D and F apply for this condition, and as the existence of other states occurs with a much lower probability, the received bit rate is approximately the difference between TBR and FBR.

Dispersive channels alter both the power and correlation of the recovered signal. The most common state is D which occurs when $\hat{\rho} < |K|$. State C occurs when the scrambled speech arrives with a correlation $\hat{\rho} \geq K$, causing 1 to be interpreted as a 0. State F occurs when $\hat{\rho} < |K|$, or $\hat{\sigma}_x^2 < T$, or when both $\hat{\rho} < |K|$ and $\hat{\sigma}_x^2 < T$. The other states were found to rarely happen.

IV. DATA TRANSMISSION PERFORMANCE

The simultaneous speech and data transmission using frequency inversion scrambling, SSDT/FIS, described here, was investigated using the sentences: "Live wires should be kept covered," and "To reach the end he needs much courage"—spoken by a male and female, respectively. The speech signal was sampled at 8 kHz to yield 38,912 samples, a number sufficiently large to give a good indication of the system's performance. The amplitude of the speech samples was confined to the range extending from −6000 to +6000 arbitrary units, and the mean square value of the samples averaged over both sentences was $MS_x = 1.09 \times 10^6$ or 60.4 dB relative to a mean square value of unity. The time waveforms for these two sentences and an expanded version of the magnitude of these speech samples to give the time variation of the low-level sounds are shown in Fig. 3.

Our objectives were to determine how to select K, T, and N for high TBR and low or negligible TBER, and to study how the performance

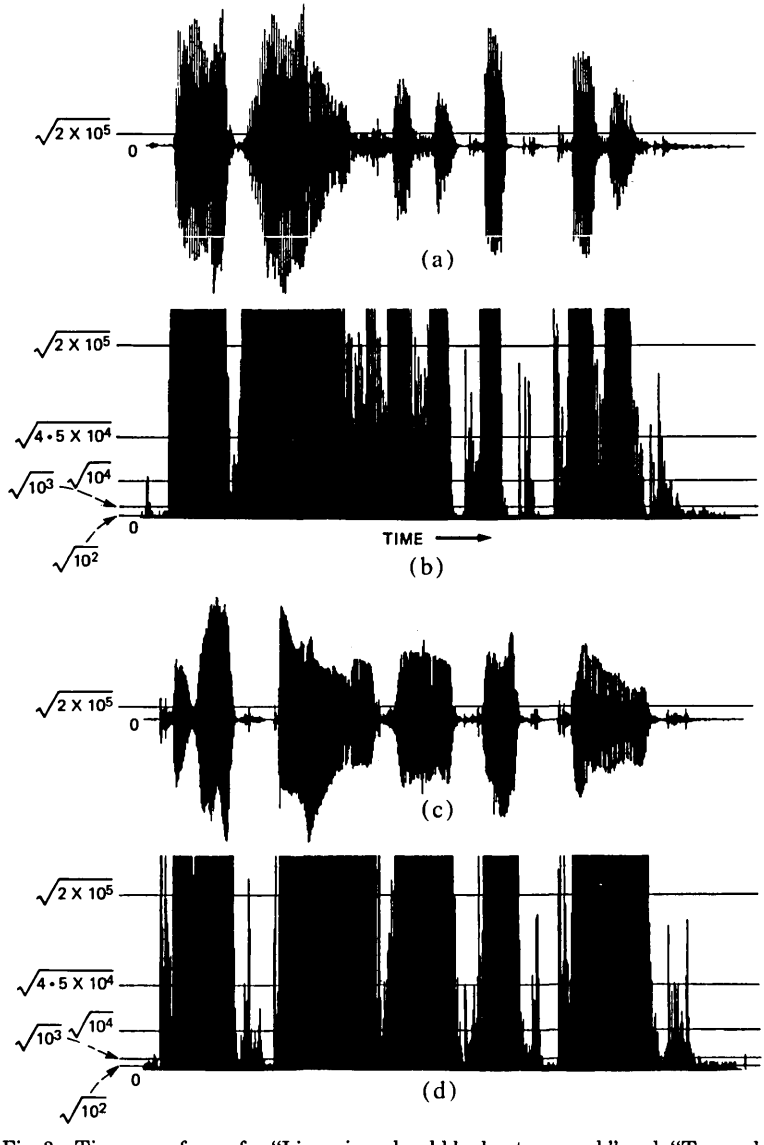


Fig. 3—Time waveforms for "Live wires should be kept covered," and, "To reach the end he needs much courage," are shown in (a) and (c). The corresponding positive amplitudes of the waveforms for the low-level sounds (high amplitudes truncated), together with various values of $\sqrt{T}$, are displayed in (b) and (d), respectively.

deteriorated in the presence of additive noise on the input speech and on the transmitted SSDT/FIS signal. We assumed that block synchronization between transmitter and receiver was correctly maintained at all times.

4.1 Selection of K

The two sentences were processed sequentially. The speech samples were divided into blocks of N samples, where N could be either 8, 16, 32, 64, 128, or 256, resulting in 4864, 2432, 1216, 608, 304, and 152 blocks of samples, respectively. For each value of N the probability density function (PDF) was computed for the correlation coefficient ρ, and plotted in Fig. 4. The PDFs were found to have similar shapes for $N =$ 16 to 256, although the shape marginally altered for $N = 8$. For smaller values of N, there is a translation in the position of the PDF to lower values of ρ. This arises because of the definition of ρ given by eq. (3). The maximum possible value of ρ for $N = 4$, 3, and 2 is 0.809, 0.707, and 0.5, respectively. We will subsequently show that $N = 4$ is the smallest block size of interest in this transmission system; therefore, we do not display PDFs in Fig. 4 for $N < 4$.

In the SSDT/FIS system, with the threshold T set to zero, the signal used to transmit data is the original speech signal, for which the curves in Fig. 4 apply. However, if $T > 0$, more blocks of speech are rejected for the conveyance of data. Therefore, TBR decreases, and the resulting blocks available for data transmission have PDFs for the correlation coefficient that are different from those in Fig. 4. At this stage, we will confine the discussion to the case of $T = 0$, i.e., where TBR has its highest values, and the curves in Fig. 4 are relevant.

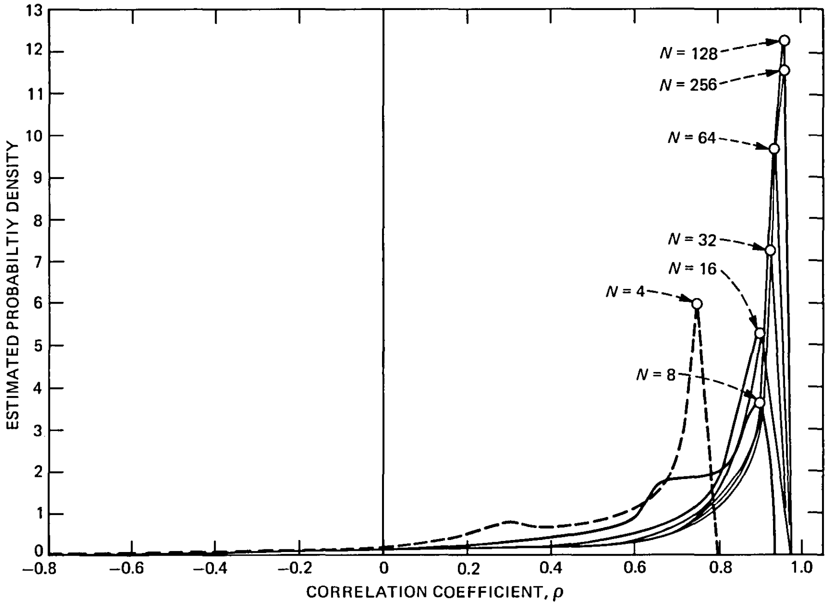


Fig. 4—Probability density function for the correlation coefficient ρ for different values of N.

Returning to these curves, we draw attention to their most negative correlation coefficient, $\rho_{\min}$, values, as they can have a significant effect on the number of bit errors. The variation of $\rho_{\min}$, and the maximum correlation coefficient $\rho_{\max}$, as a function of N is displayed in Fig. 5. We recall from our discussion in Section III, that if $\rho < K$, $\sigma_x^2 > T$, no data is transmitted. Assuming an ideal channel, and given that $\hat{\rho} = \rho < -K$, the system is fooled into believing a logical 1 was transmitted and a bit error occurs. Clearly, if K is selected such that $\rho < -K$ does not exist, then no bit errors are possible over an ideal channel. To avoid bit errors we arrange for

$$K > |\rho_{\min}|, \tag{12}$$

and the choice of K to avoid bit errors as a function of N must, therefore, be below the curve $|\rho_{\min}|$, e.g., for $N = 64$, $K > 0.43$. For $N < 16$, $\rho_{\min}$, and $\rho_{\max}$ both decrease with decreasing N, and for $N = 4$ we have the interesting situation that $|\rho_{\min}| = \rho_{\max}$, which means that if Inequality eq. (12) is satisfied no data will be transmitted as $\rho > K$ cannot exist. The value, $N = 4$, therefore, marks the lower limit of the block size for combined speech and data transmission over an ideal channel without the occurrence of bit errors.

Reducing K from $\rho_{\max}$ increases the number of speech blocks that can be considered for the conveyance of binary data, but if $K \leq |\rho_{\min}|$, bit errors ensue. Thus, in order to transmit the greatest amount of

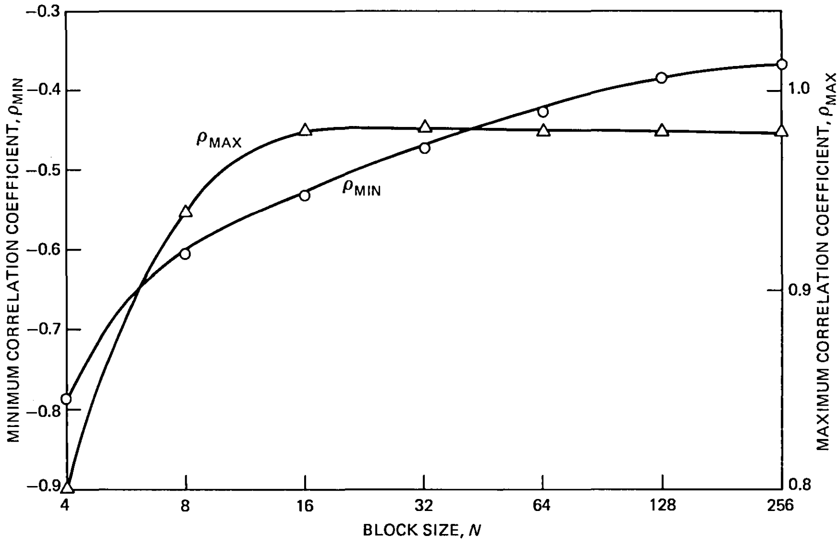


Fig. 5—Variation of maximum ($\rho_{\max}$) and minimum ($\rho_{\min}$) correlation coefficient values for different block sizes (N).

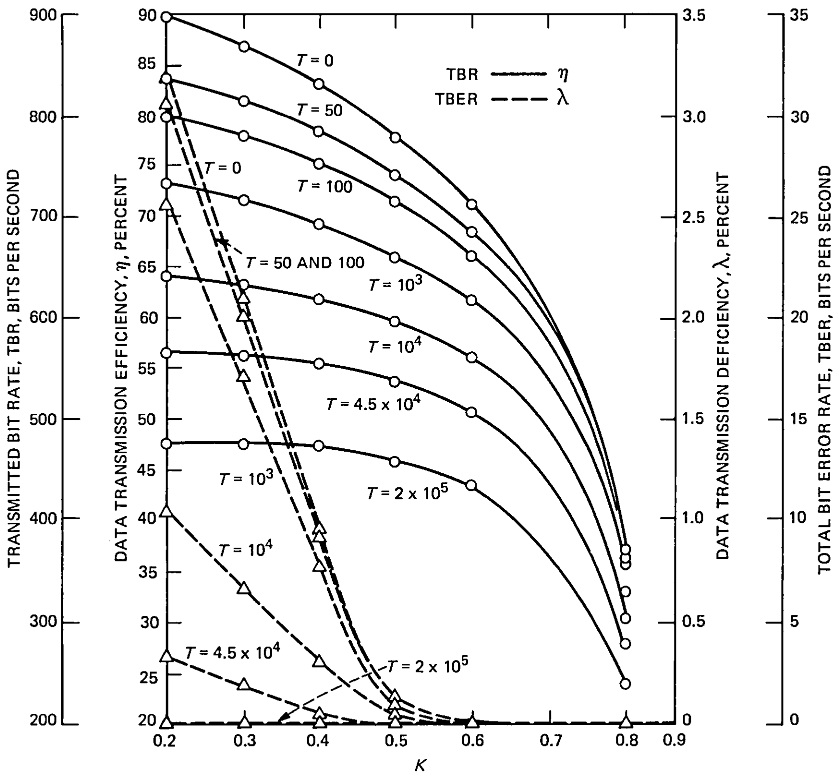


Fig. 6—Variation of data transmission efficiency (η) and data transmission deficiency (λ), as a function of K and T for $N = 8$.

data without errors over an ideal channel, K is bounded by

$$|\rho_{\min}| < K < \rho_{\max}. \tag{13}$$

However, the negative tails of the PDFs are long and of low amplitude and, hence, $K < |\rho_{\min}|$ can be used provided the penalty of a low TBER can be tolerated.

4.2 Ideal channel

The effects of parameters K and T on the data transmission efficiency η, the TBR, the data transmission deficiency λ, and the total TBER, for block sizes of 8 and 32, is shown in Figs. 6 and 7, respectively. These two block sizes were selected because $N = 8$ provides the highest data transmission rate in the absence of false bits, and $N = 32$ has fewer false bits than does $N = 8$ at low values of K, while having a relatively high TBR. Because the shape of the curves in Figs. 6 and 7 are similar, we refrain from showing curves for other values of N.

The curve for $T = 0$ is of interest as it provides the highest values of

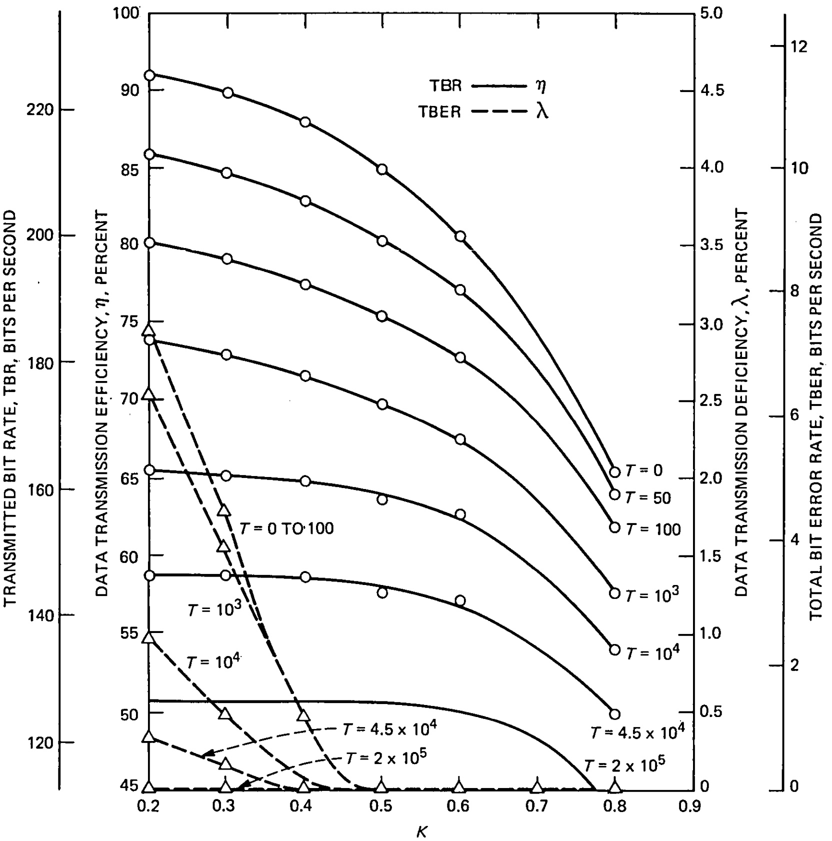


Fig. 7—Variation of data transmission efficiency (η) and data transmission deficiency (λ), as a function of K and T for $N = 32$.

η, and relates to the discussion in Section 4.1. If k is increased beyond 0.8 the curve falls rapidly, becoming zero for $K \geq \rho_{\max}$. As K is reduced below 0.2, η climbs towards 100 percent, but λ becomes excessive. Consider the operating condition: $T = 0$ and $K = |\rho_{\min}|$. For $N = 8$, $K = 0.6$, $\eta = 71$ percent, $\lambda = 0$, yielding a TBR of 710 b/s and a TBER of zero. By reducing K to 0.2 while maintaining $T = 0$, η is increased to 90 percent, or to a TBR of 898 b/s. However, λ is now 3.35 percent, giving a TBER = FBR of 33.5 b/s. These FBRs arise because K is below $|\rho_{\min}|$. The system can operate with low values of K, 0.2 say, provided T is increased. By raising the value of T, blocks which occur during silence and unvoiced periods are not considered for data transmission. For still higher values of T, blocks existing during silence, unvoiced, and low amplitude voiced sounds, are rejected for the conveyance of data. The values of T used in our experiments (other than $T = 0$), namely 50, 100, 10^3, 10^4, 4.5×10^4, 2×10^5; correspond to power levels that are 43.4, 40.4, 30.4, 20.3, 13.8, and 7.36 dB, respectively, below the mean square value MS_x of the combined speech signals. The $\sqrt{T}$ thresholds, except $\sqrt{50}$, are shown in Fig. 3(b) and (d), and for reference $(2 \times 10^5)^{1/2}$ is shown in Fig. 3(a) and (c).

The effect of using non-zero values of T is a modification of the shape of the PDF of the correlation coefficient ρ—specifically, the truncation of its long negative tail. Consequently, $\rho_{\min}$, a negative value in Fig. 3, is made significantly more positive. For example, when $N = 8$, $\rho_{\min} = -0.6$, -0.56 and -0.07 for $T = 0$, 10^3 and 2×10^5, respectively. When $T = 2 \times 10^5$, no false bits occur, irrespective of K, as shown in Fig. 6, but η decreases significantly to 48 percent, giving a TBR of 480 b/s. Clearly, for the ideal channel, there is no advantage in making T anything other than zero and $K = 0.6$. We will find that in the presence of channel noise T must have a high value if λ is to be contained.

Although increasing N generally produces higher values of η, as can be seen in Fig. 7, the larger block size results in a significant reduction in TBR. The effect of N on λ is seen to be small; therefore, we recommend the use of $N = 8$.

The effect of block size N on TBR and TBER for different values of T is shown in Fig. 8, where $K = 0.5$. The data transmission efficiency is approximately independent of N for a given T, and consequently TBR is inversely proportional to N. [See eq. (8)]. False-bit errors occur for $N = 16$ and 8 as $\rho_{\min} < -0.5$, unless T is increased to $\simeq 4.5 \times 10^4$. By using this high value of T, TBR is seen to fall from 530 b/s to 19.5 b/s as N is increased from $N = 8$ to 256, the TBER being maintained at zero. Clearly, from a data transmission point of view, high values of N are undesirable, although they do increase the listening fatigue of an eavesdropper, as described in Section 5.2.

The small block size of $N = 4$ that spans a duration of 0.5 ms can be used to transmit data without error, provided a large value of T is used to remove the long tail in the correlation coefficient PDF of Fig. 4. We found that if $T = 2 \times 10^5$, $\lambda = 0$, and $\eta = 31.6$ percent. This represents a TBR of 632 b/s and a TBER of zero.

4.2.1 Effect of background noise

To simulate a noisy environment, we added a random noise sequence having a power σ_{ni}^2 to the speech sequence. The effect of this background noise power on the data transmission efficiency η and TBR for different values of threshold T is shown in Figs. 9 and 10 for $N = 8$, $K = 0.6$, and $N = 32$, $K = 0.5$, respectively. An additive power level of $\sigma_{ni}^2 = 10^k$, $k = 0, 1, 2, \cdots$, corresponds to a power level of 60.4-10 k, dB, below the mean square value MS_x of the speech signal. As expected, the highest value of η occurs when $T = 0$, as the only criterion applied

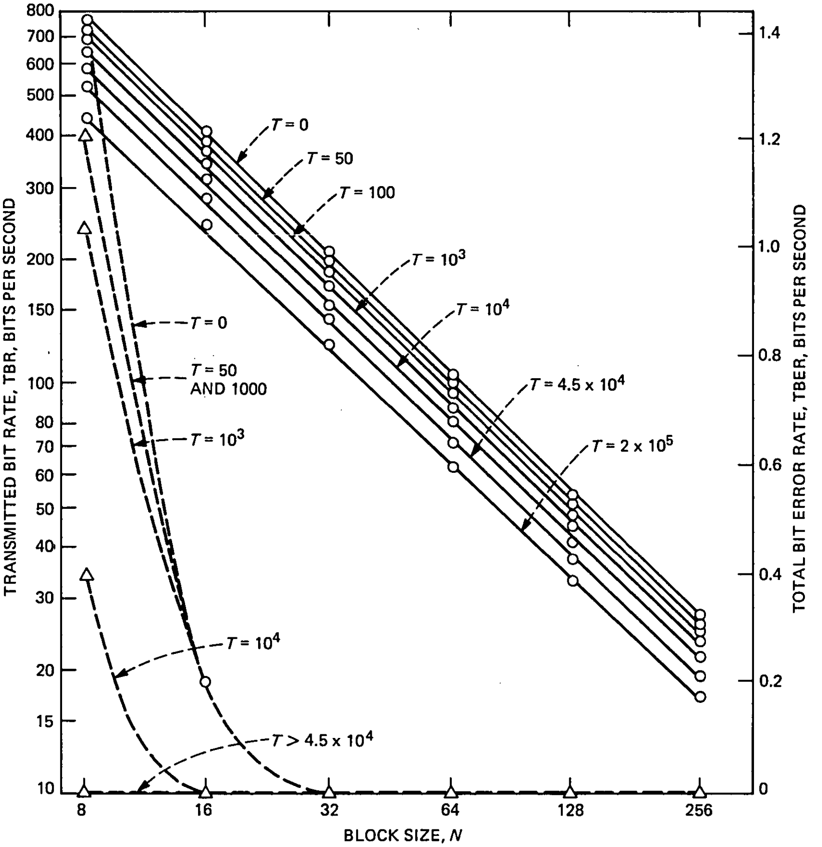


Fig. 8—Variation of data transmission efficiency (η) and data transmission deficiency (λ), as a function of block size (N) for different values of T, $K = 0.5$.

in the selection of blocks to convey data is based on whether the correlation coefficient ρ is above K. Increasing σ_{ni}^2 causes the speech to be decorrelated; therefore, less blocks have $\rho > K$, and consequently η decreases. Increasing T means that fewer blocks fulfill the condition that y of eq. (5) is a logical 1. The blocks discarded are generally those containing low-level speech, and it is these blocks that experience greatest decorrelation. Thus, as σ_{ni}^2 increases, η remains constant as the decorrelative effect is masked by the value of T. When σ_{ni}^2 approaches T, blocks not rejected because their mean value is $>T$ are now abandoned because of ρ being too small due to the decorrelation. Consequently, η versus σ_{ni}^2 is no longer a constant, and η coalesces with the $T = 0$ curve as σ_{ni}^2 is further increased. This occurs because the controlling factor in block rejection is now the correlation criterion. For $\sigma_{ni}^2 > T$, η decreases at approximately 6.8 percent per decade

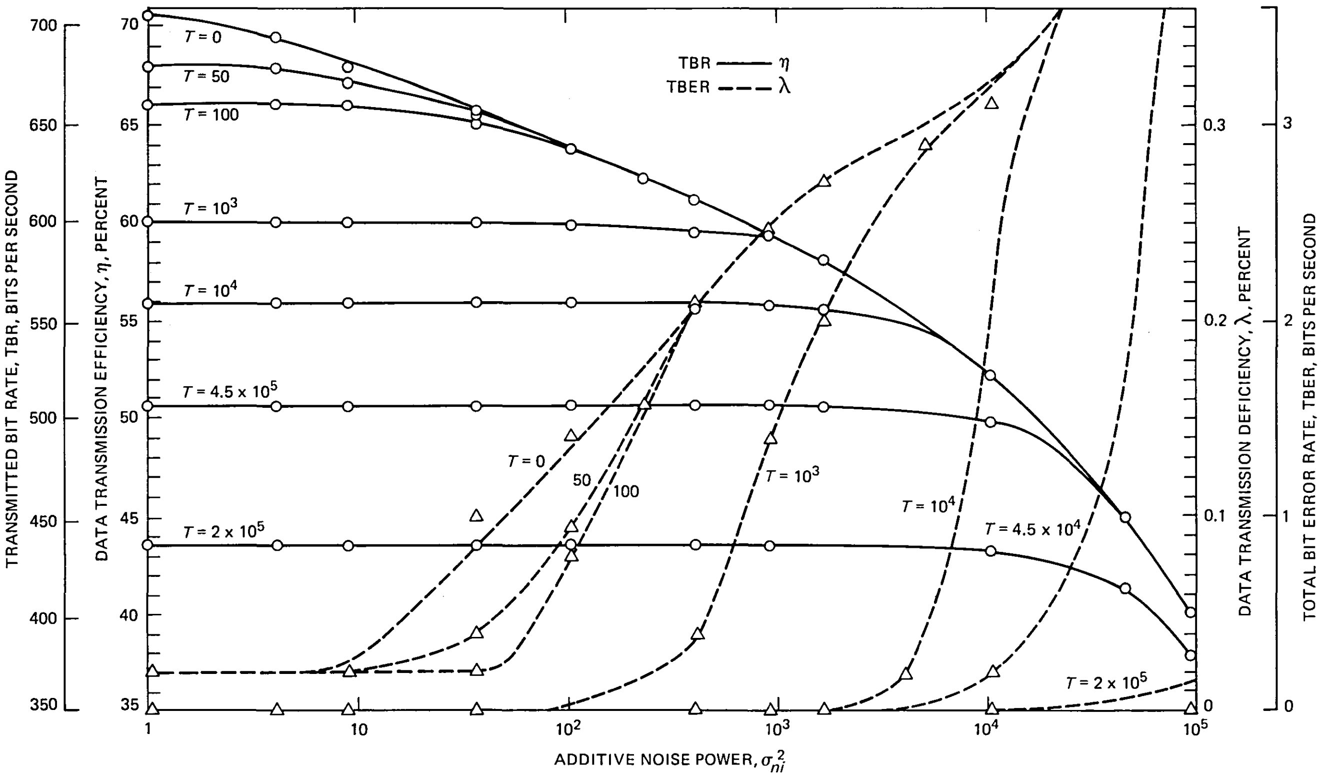


Fig. 9—Effect of additive background noise power (σ_{ni}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function of threshold T. $K = 0.6$, $N = 8$.

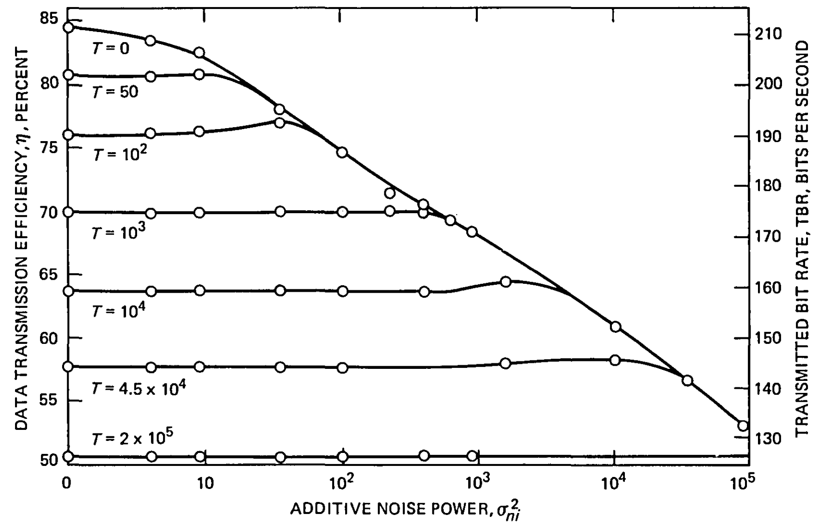


Fig. 10—Effect of additive background noise power (σ_{ni}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function of threshold T. $K = 0.5$, $N = 32$, $\lambda = 0$.

increase in σ_{ni}^2 when N is 32, and at a rate approaching this value when N is 8.

The data transmission deficiency λ and the TBER was found to be zero for the case of $N = 32$, over the range of power levels shown in Fig. 10. However, when the block size was 8, data errors occurred. The variation of λ and TBER with σ_{ni}^2 for different values of T is shown in Fig. 9. The figure demonstrates that data errors can be avoided for $\sigma_{ni}^2 < 10^4$ by setting T to 2×10^5, although the data transmission efficiency falls to 43.5 percent, i.e., TBR = 435 b/s.

4.3 Noisy channel

When no noise was added to the speech signal, but the channel was noisy with additive channel noise power σ_{nc}^2, η was uneffected. However, λ increased due to blocks that did not contain data ($\sigma_x^2 < T$) but had their power increased to $\sigma_x^2 + \sigma_{nc}^2 \geq T$, and if $\hat{\rho}$ or $\hat{\rho}_s$ exceeded K, data errors ensued. The variation of λ and TBER with σ_{nc}^2 for various values of T is displayed in Figs. 11 and 12, for $N = 8$ and 32, respectively. As λ and TBER had zero values for large values of T when $N = 32$, we present a zero line in Fig. 12, and the lines from this base to the other points on the curve are dotted. Notice in Fig. 12 that no data errors were recorded over the entire range of σ_{nc}^2 when $T = 2 \times 10^5$, and from Figure 10 this value of T corresponded to $\eta = 50.5$ percent. Thus, by using $N = 32$, and a background noise power and additive channel noise power up to 10^5, i.e., up to 10.4 dB below MS_x, we found that 126 b/s can be transmitted without error. When $N = 8$, $T = 2 \times 10^5$, and both types of noise are present up to 10^3, TBR = 435 b/s, but TBER is no longer zero, having a value between 0.2 to 3.5 b/s. The various combinations of TBR and TBER can be deduced from Figs. 9 through 12.

V. SPEECH TRANSMISSION PERFORMANCE

Emphasis has been given to data transmission because we wanted to investigate if it could be reliably achieved using speech as a carrier signal. In the previous section, we presented results showing that it was possible to transmit data without transmission errors, and consequently the recovered speech signal was unimpaired by conveying the data. However, we have also observed that the data rate can be increased if bit errors can be tolerated. Thus, we now address the problem of how the bit errors affect the recovered speech signal, and specifically ask, How serious is the degradation of speech quality and intelligibility when the total TBER approaches the maximum values found in our experiments?

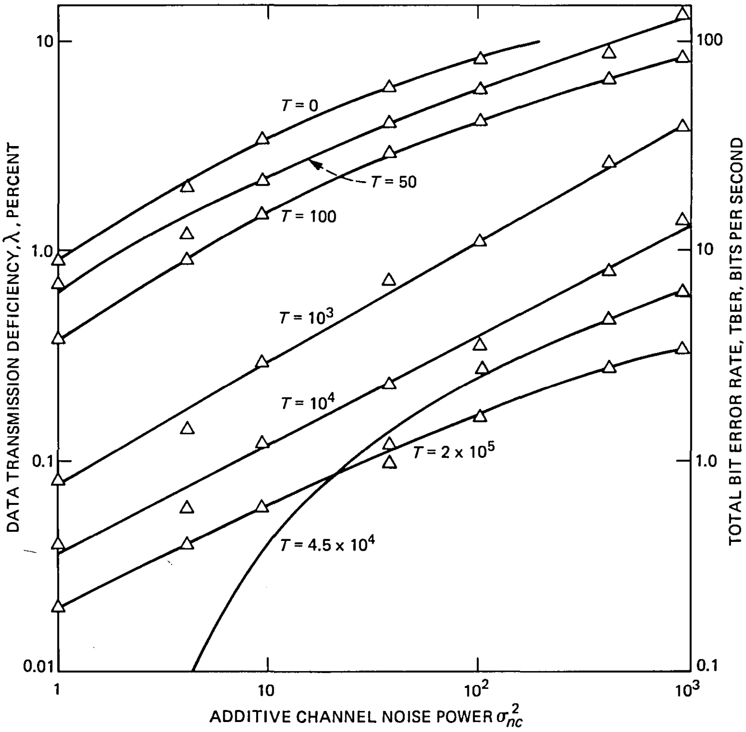


Fig. 11—Effect of additive channel noise power (σ_{nc}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function threshold T. $K = 0.6$, $N = 8$.

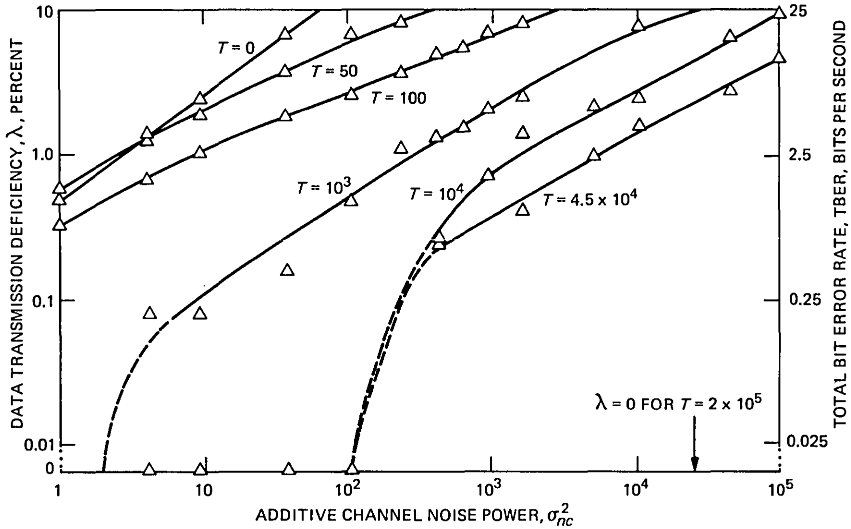


Fig. 12—Effect of additive channel noise power (σ_{nc}^2) on data transmission efficiency (η) and data transmission deficiency (λ), as a function of threshold T. $K = 0.5$, $N = 32$, $\eta = 63.7$.

5.1 Objective measure

To provide an objective measure of the degradation of the recovered speech signal that accrues solely from the effect of data errors and not from the presence of additive background or channel noise, we elected to use segmental s/n (SEG-S/N). This ratio is used[5,6] as an objective measure because its value corresponds more closely to perceived quality than those of conventional s/n measurements, i.e., those using the ratio of mean square signal power to mean square noise power, determined over the duration of the entire signal. The reason for this resides in the computation of SEG-S/N, which is performed as follows. The input speech sequence $\{x_k\}$ is divided into contiguous blocks of 128 samples, i.e., into periods of 16 ms. Only those blocks, for example, M, whose rms value exceeds Γ dB (here −60 dB) relative to the peak value, have their s/n calculated. The s/n for the jth block is

$$\text{s/n}_j = 10 \log_{10} \left\{ \frac{\sum_{i=1}^{128} x_{128j+i}^2}{\sum_{i=1}^{128} (x_{128j+i} - \hat{x}_{128j+i})^2} \right\}$$

$$j = 1, 2, \cdots, M, \tag{14}$$

where $\{\hat{x}_k\}$ is the recovered speech sequence. There is no point in considering $s/n_j < -10$ dB, or $> +80$ dB, because the speech quality is not perceived worse or better than -10 dB or $+80$ dB, respectively. Hence,

$$s/n_j = -10 \text{ dB for } s/n_j \leq -10 \text{ dB}$$

and

$$s/n_j = +80 \text{ dB for } s/n_j \geq +80 \text{ dB} \tag{15}$$

for $j = 1, 2, \cdots, M$. The number of data blocks of length N contained in the s/n_j calculation block size of 128, decreases from 16 to 0.5 as N is increased from 8 to 256, respectively. For a given TBER, the effect of increasing N is to cause the number of blocks not having the maximum s/n_j of 80 dB to decrease, but the decrease in s/n_j is more substantial in those blocks of 128 samples where erroneous scrambling occurred with all the samples, compared to those blocks where only, 8 samples were erroneously scrambled.

The segmental s/n is computed as the average of s/n_j; $j = 1, 2, \cdots, M$, namely,

$$\text{SEG-S/N} = \frac{1}{M} \sum_{j=1}^{M} s/n_j . \tag{16}$$

The Γ threshold enables us to ignore blocks of low-level speech in the calculation of SEG-S/N. Even if these blocks are erroneously scrambled at the receiver output, their removal from the calculation is justified on the basis that the effect on such a low-level sound is imperceptible.

Let us now consider the case of speech conveying data through an ideal channel without the introduction of data errors. Here SEG-S/N is 80 dB, as s/n_j, for all j, is forced to 80 dB. When data errors occur, the SEG-S/N does not fall greatly below 80 dB, implying that the degradation in speech quality because of the presence of data errors is small. Perceptual observations substantiate this implication, confirming our decision to use SEG-S/N as an objective measurement of performance.

When the input speech is contaminated by statistically independent background noise, the recovered speech at the receiver is the sum of the original speech and noise signals, provided there are no data errors. Whenever a data error occurs, both the speech and the noise in the block are scrambled. In the case of additive channel noise, and for no data errors, the recovered speech is, again, the sum of the original speech and the noise signals, except when the data is a logical 1 when the output noise signal is scrambled. The effect of data errors, excluding states B and F, is to cause the output signal to be the sum of the scrambled original speech signal and the scrambled noise signal. Perceptually, the scrambled noise signal is the same as the original random noise signal; therefore, its effect must be removed in evaluating the loss of SEG-S/N because of data errors. This is achieved by noting those blocks where data errors occurred in a noisy environment, and then scrambling the original speech in those same blocks to give the sequence $\{\hat{x}_k\}$ used in eq. (14). By this method, we are able to separate the distortion in the output signal, caused by the additive noise, from the effect of the data errors that were precipitated by this noise.

5.1.1 Objective results

The occurrence of a bit error results in a block of speech samples at the output of the receiver being erroneously scrambled. Clearly, if this happens to a block containing high amplitude samples, significant distortion ensues. From Section IV we have seen that the reduction or elimination of data errors can be achieved by increasing T. However, this is not our purpose here. We wish to generate data errors and observe their effect on the recovered speech. Consequently, in selecting conditions to illustrate the reduction in SEG-S/N caused by data errors, we have opted for $T = 0$. Table II shows the SEG-S/Ns for some of the worst data-error conditions shown in Figs. 5 to 12, plus a high data-error case when $N = 256$. These conditions were selected to show that in spite of the TBER values being unacceptably high for most data communications systems, the effect of the data errors on speech quality is small. Indeed, SEG-S/N remains above 66 dB for all the conditions depicted in Table II. Therefore, we refrain from presenting detailed measurements of speech distortion that is barely perceptible. However, we do discuss the error conditions, but cannot make general deductions from the few entries in the table, particularly as there is not a consistent theme. For example, for the noisy channel there are three values of N, but they each have a different K, so comparisons must be tempered with caution.

In Table I we have included the recovered speech status at the receiver for each of the data-error states. Two states, B and F, do not cause the recovered speech signal to have the samples in the erroneous blocks scrambled. When additive channel noise is present, error states D and F occur more frequently than the other states. Thus, the distortion in the output speech results mainly from state D, i.e., when a transmitted logical 1 is ignored. As states D and F are likely to occur with approximately the same probability (see Table II), the error rate in terms of distorting the recovered speech, can be considered to be reduced by a factor of two. However, state D is associated with a loss of a data signal caused by the channel noise increasing the correlation of a block of speech samples. Because data was transmitted, the speech can be voiced (although $T = 0$ for Table II) in which case the speech distortion may be substantial.

Table II—SEG-SN for some high data-error conditions, T = 0

Condition	Block Size N	Corr. Threshold K	Additive Noise Power σ_n^2	Occurrence of Data-Error States						Data Transmission Deficiency λ, %	TBER b/s	Data Transmission Efficiency η, %	TBR b/s	SEG-S/N dB
				A	B	C	D	E	F					
Ideal channel, no additive noise	8	0.2	ZERO	154	0	0	0	0	0	3.17	31.7	89.7	897	72.2
Noisy channel	8	0.6	918	6	32	1	302	1	286	12.8	128	80.0	800	66.7
Ideal channel, additive background noise	8	0.6	9.18×10^4	34	0	0	0	0	0	0.47	4.7	40.0	400	76.1
Ideal channel, no additive noise	32	0.2	ZERO	36	0	0	0	0	0	2.96	7.4	91.3	228	74.5
Noisy channel	32	0.5	918	0	0	0	99	0	85	15.5	38.8	84.5	211	75.6
Noisy channel	256	0.2	4.3×10^3	1	0	0	10	0	10	14.6	4.56	93.4	29	77.4

The data errors resulting from the ideal channel, with or without additive background noise, cause state A to apply, as shown by the examples in Table II. Although the output speech blocks are erroneously scrambled every time a data error occurs, the distortion is confined to blocks that often contain unvoiced sounds as the data error is the result of no data being transmitted, but a logical 1 being falsely generated.

Waveforms for the worst condition shown in Table II, namely, additive channel noise, $N = 8$, $T = 0$, $K = 0.6, \sigma_{nc}^2 = 918$, are displayed in Fig. 13. The high amplitude signal levels are seen to be substantially unaffected by the data errors whose effects are often immersed in the channel noise, and are therefore not perceptibly annoying.

5.2 Informal listening experiences

Informal listening tests were performed for the conditions listed in Table II. The recovered two sentences of speech, stripped of noise, with blocks of speech erroneously scrambled where data errors occurred, suffered only minor distortions. For the ideal channel, minor distortions resembling a "sshing" sound, occurred on three occasions for the cases of $N = 8$ and 32. A quiet noise, like additive white noise, was perceived for $N = 8$ when either background, or channel noise (the worst condition) were present. For the noisy channel condition, $N = 32$ produced the effect of barely perceptible scratches, while $N = 256$ yielded the least distortion, where the degradations were reminiscent of barely audible metallic clicks.

When the noisy output signal containing the effects of data errors was compared to the original speech plus additive noise, the effect of the data errors was imperceptible in the case of the substantial additive input noise power $\sigma_{ni}^2 = 9.18 \times 10^4$, N being 8. The effect of unwanted scrambling when the channel was noisy ranged from barely perceptible, $N = 256$ and 32, to nonannoying crackles when $N = 8$ andσ_{nc}^2 was only 918.

The conclusion is that for the data-error rates of practical significance, the degradation in speech quality is insignificant.

The transmitted signal sounded like distorted speech, plus white noise for the case of $N = 8$, and an ideal channel. The effect of additive background or channel noise was to reduce the fatiguing effects, as if the distortion had been removed from the speech and the background noise increased. When $N = 32$, the channel ideal, the distortion was increased as this block size corresponds to 4 ms, approximately half a pitch period. The speech sounded as if speaking and gargling were being performed simultaneously. The act of adding noise marginally reduced listening fatigue. The scrambled signal was found to be just intelligible when $N = 256$.

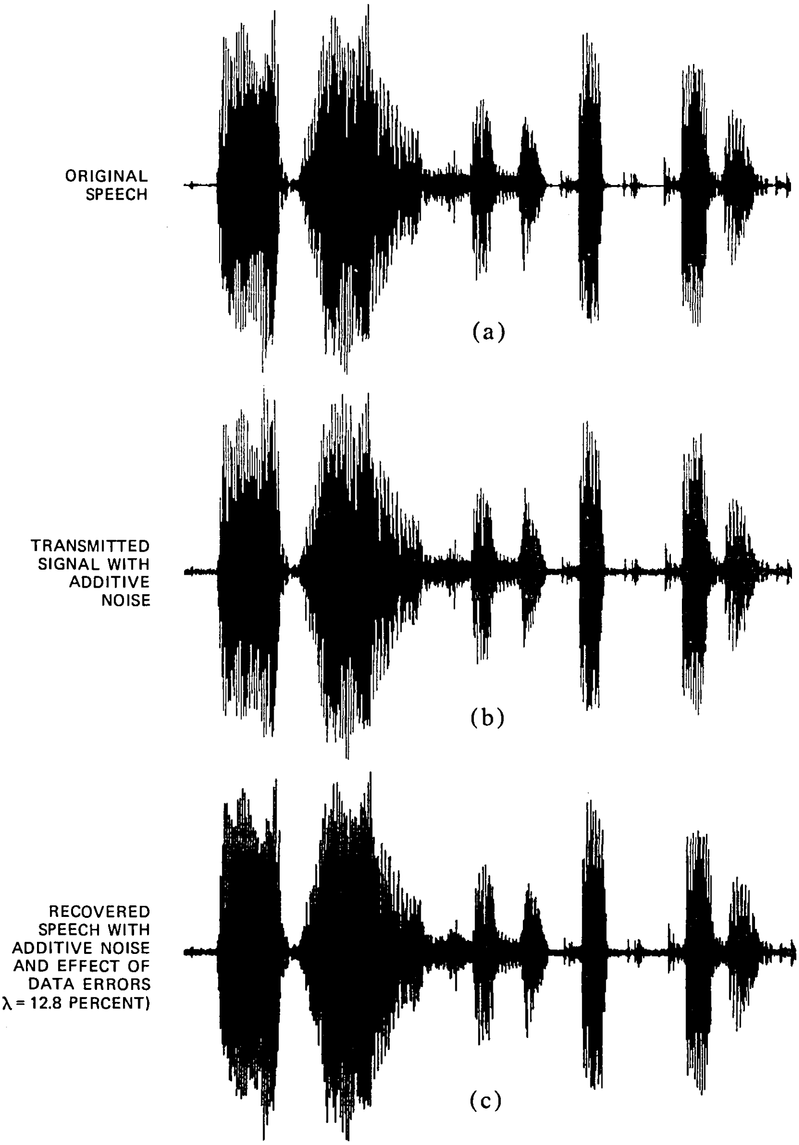


Fig. 13—Effect of additive channel noise. (a) Original speech. (b) Transmitted signal with additive channel noise having $\sigma_{nc}^2 = 920$. (c) Recovered speech having data errors ($\lambda = 12.8$ percent) and additive noise. $N = 8$, $T = 0$, $K = 0.6$.

VI. DISCUSSION

We started by enunciating a principle: that data could be transmitted by making it the scrambling key, and casting the receiver in the role of code breaker. Every time the receiver guesses the key, it obtains the correct data and the correct speech. The speech is made an unwitting data carrier, while the data gets a free ride. The implications of this concept are considerable. Continuous users, or providers, of telephone traffic can, at the expense of additional terminal equipment, surreptitiously transmit teleprinter data, with the proviso that the bandwidth of the speech signal and the block size are appropriate for the channel bandwidth.

To demonstrate the principle, we wanted a scrambling technique that was easy to implement, and where the bandwith of the scrambled speech was not larger than that of the original speech. Frequency inversion scrambling aptly fulfilled these requisites, where the scrambling is achieved by merely altering the polarity of every other speech sample. We have shown that by using this form of scrambling, it is possible to transmit speech and data simultaneously, and to receive the data without errors and the speech without distortion, even in the presence of additive-channel and background noise. Provided some data errors can be tolerated, the data rate can be substantially increased as shown in Figs. 6–12. Even at high data-bit rates the distortion in the speech was found to be minimal, as the results in Table II indicate.

We have not presented results for dispersive channels, although we did do some experiments. The effect of such channels was to significantly alter the PDF of the correlation coefficient of the received signal compared to that of the transmitted signal. The power in the blocks of speech was also changed by the dispersive properties of the channel. As the data detection procedure is based on a measurement of power and correlation in a block of N samples, where the correlation is usually the most important factor, the dispersive channel results in an unacceptably high TBER. Thus, in the presence of dispersive channels, equalization of the channel must be performed.

The speech used in our experiments were two sentences whose waveforms are displayed in Fig. 3. Therefore, the results will differ when other sentences are used, but not significantly, as the sentences used consisted of over thirty-eight thousand samples. The system proposed here is for conveying data on speech or short silences. When prolonged silences occur, we envisage data being transmitted by conventional modern techniques.

The basic principle established, the way forward is to find scrambling methods that will be easier to break with certainty, and will operate via dispersive channels without the necessity of channel equalization.

VII. ACKNOWLEDGMENT

The authors thank D. J. Goodman for his constructive criticism of this work.

REFERENCES

1. D. Kahn, *The Code-Breakers*, New York: Macmillan, 1967.
2. G. Guanella, "Automatic Speech Scrambling," Brown, Boveri & Company Pub. CH-E7.30038.2E.
3. S. C. Kak and N. S. Jayant, "On Speech Encryption Using Waveform Scrambling," B.S.T.J., *56,* No. 5 (May–June 1977), pp. 781–808.
4. N. S. Jayant et al., "A Comparison of Four methods for Analog Speech Privacy," IEEE Trans. Commun., *COM-29,* No. 1 (January 1981), pp. 18–23.
5. B. McDermott, C. Scagliola, and D. J. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM," B.S.T.J., *57,* No. 5 (May–June 1978), pp. 1597–618.
6. J. M. Tribolet et al., "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.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

Frequency Scaling of Speech Signals by Transform Techniques

By D. MALAH* and J. L. FLANAGAN

(Manuscript received May 15, 1981)

The general framework of short-time Fourier analysis, modification, and synthesis is used to describe in a unified way several known techniques for frequency scaling of speech signals. Subsequently, a frequency domain harmonic scaling technique is studied in detail with emphasis on improving its performance and its implementation efficiency. This technique is particularly attractive for 2:1 scaling by use of a sign tracking algorithm which avoids the need for explicit phase computation and unwrapping. The implementation efficiency is achieved by using the fast Fourier transform algorithm, embedded decimation and interpolation, and an extended version of a recently developed weighted overlap-add synthesis scheme. The improvement in quality is achieved by improved sign tracking and elaborate design and selection of the analysis and synthesis prototype filters (data windows). Results of computer simulations, for a variety of adverse acoustical environment conditions, indicate that the system is highly robust but its quality for clean speech is lower than with a time domain harmonic scaling technique which uses pitch information. In applications which do not permit pitch transmission, a hybrid scheme which combines the two techniques is found to yield a better quality than either system alone.

I. INTRODUCTION

Frequency scaling of speech signals is a useful method for reducing the bandwidth requirements in analog and digital speech transmission systems.[1-5] In analog systems the frequency compressed signal is transmitted at reduced bandwidth. In digital systems the frequency

* On leave from the Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa, Israel.

compressed signal is waveform coded to provide reduced bit-rate transmission.[4-7] A general block diagram of such a digital system is shown in Fig. 1. In this figure, the schematic spectral representation of the input speech signal is shown to consist of a spectral envelope with pronounced resonances (formant peaks) and of a fine structure because of pitch harmonics in voiced speech. The spectral envelope of the compressed signal is a scaled version of the input spectral envelope. However, different frequency scaling techniques may result in different fine structures.

Since we do not refer at this point to any specific technique, Fig. 1 does not show the fine structure of the compressed signal. This suggests that frequency scaling techniques can be classified according to the way the fine spectral structure is scaled. In particular, one can distinguish between narrow-band techniques, such as the phase vocoder[2] and time-domain harmonic scaling (TDHS)[4] techniques, which aim at separating and scaling the individual pitch harmonics, and wide-band techniques, such as the analytic signal rooting (ASR) technique[3] and the more recent constant Q transform (CQT) method,[8,9] which aim at directly scaling the spectral envelope. The much earlier Vobanc and CODIMEX systems also fall into the latter category.[10,11]

Another way of classification is to distinguish between time- and frequency-domain techniques. To provide useful quality, time-domain techniques require pitch tracking, as done in the TDHS technique.[4]

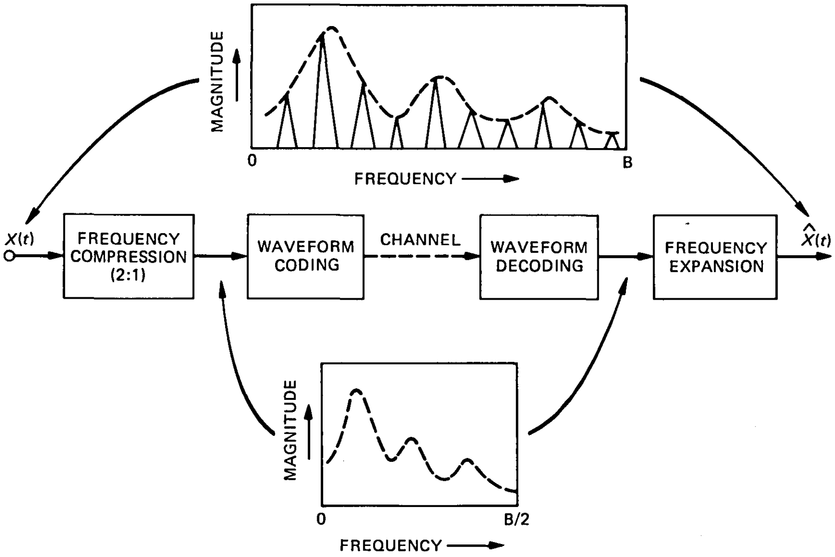


Fig. 1—General block diagram of a digital coding system which applies frequency scaling for bit rate reduction.

Once the pitch is known, the time-domain operations are simple and result in good quality scaled and reconstructed speech.[5-7] Frequency-domain techniques are typically much more complex but do not require explicit pitch tracking. However, they usually have a lower quality because of errors made in resolving phase ambiguity and from the need, in general, to scale both the phase and amplitude signals, as will be elaborated later on. For applications in which pitch tracking is not desired or not possible, because of adverse acoustical environment conditions, the use of an efficient frequency domain technique is of much interest.

In this work we present an efficient implementation of a frequency-domain harmonic scaling (FDHS) technique which is based on an improved version of the technique presented in Ref. 12. Frequency-domain harmonic scaling is a narrow-band technique which aims at scaling the individual pitch harmonics. It is particularly attractive for 2:1 scaling, since in this case a sign tracking algorithm avoids the need for explicit phase computation and unwrapping (that is, eliminating 2π phase ambiguities), which in general is a difficult and error-prone task.[13] The efficient implementation is based on the recently developed weighted overlap-add method for short-time Fourier analysis/synthesis, which allows block processing using the fast Fourier transform (FFT) algorithm.[14] It is extended here to include analysis and synthesis windows which are both longer than the FFT block, or the transform size.

The general framework of the short-time Fourier transform (STFT) as developed in several recent works[14-16] also provides a unified description of other known frequency scaling techniques, and helps to relate them to the FDHS technique. This unified description is given in the following section, and is followed by a detailed description of the FDHS technique. Section IV gives the details of the implementation scheme and Section V discusses design considerations and simulation results. Section VI presents a hybrid technique which combines TDHS and FDHS. This combination is designed for applications in which it is feasible to extract the pitch at the transmitter but for which the transmission of pitch data is either impossible or is to be avoided.

II. A UNIFIED DESCRIPTION OF FREQUENCY SCALING TECHNIQUES

A general scheme for frequency scaling is presented in this section. It is based on viewing the frequency scaling operation as a modification of the short-time spectrum of the speech signal. This scheme is then used to describe in a unified way several known frequency scaling techniques. Our attention in describing the different techniques will be mainly focused on the nature of the spectral modifications used by each technique, and not necessarily on the way they are implemented.

The relation between the short-time Fourier transform (STFT) and a filter-bank analysis is well established.[2,15] For the convenience of this presentation, the filter bank which is used to divide the two-sided speech spectrum into sub-bands is assumed to consist of N complex bandpass filters. The center frequency of the kth filter is denoted by ω_k and its complex (or analytic) output signal by $z_k(t)$. It is also assumed that each bandpass filter has a real low-pass filter prototype, which means that the complex impulse response $h_k(t)$, of the kth filter, is given by

$$h_k(t) = w_k(t)\exp(j\omega_k t), \tag{1}$$

where $w_k(t)$ is the impulse response of the low-pass prototype of $h_k(t)$. Note that in general the prototype filters need not be identical, but we assume that the bandpass filters are contiguous and are arranged symmetrically about $\omega = 0$, so that the filter centered at $\omega = -\omega_k$ has the same prototype filter as the one centered at $\omega = \omega_k$. This way the summation of the outputs from a pair of corresponding (conjugate) complex filters results in a real signal.

The output signal from the kth complex filter has the general form

$$z_k(t) = A_k(t)\exp[j\theta_k(t)], \tag{2}$$

where $A_k(t)$ is the amplitude, or envelope, function and $\theta_k(t)$ is the phase function. The phase function can be written as a sum of two components

$$\theta_k(t) = \omega_k t + \phi_k(t), \tag{3}$$

where the meaning of $\phi_k(t)$ is elaborated below. By substituting eq. (3) into eq. (2), we see that $z_k(t)$ can be interpreted as being the result of the simultaneous modulation of the amplitude and phase of the complex carrier signal $\exp(j\omega_k t)$ by the amplitude and phase signals $A_k(t)$ and $\phi_k(t)$, respectively. The instantaneous frequency of $z_k(t)$ is given by the phase derivative $\dot{\theta}_k(t) = \omega_k + \dot{\phi}_k(t)$, so that $\dot{\phi}_k(t)$ is seen to be the deviation of the instantaneous frequency from the center frequency ω_k. Frequency scaling of $z_k(t)$ by a factor q ($q < 1$ for compression and $q > 1$ for expansion), is achieved if the center frequency ω_k is scaled or shifted to $q\omega_k$ and the bandwidth of $z_k(t)$, about ω_k, is also scaled by q. It is well known that the bandwidth of a signal which is characterized by simultaneous amplitude and phase modulation of a carrier signal is a function of both modulating signals.[17] Hence, just scaling the instantaneous frequency deviation $\dot{\phi}_k(t)$ by a factor q does not result, in general, in the exact scaling of the bandwidth of $z_k(t)$ by q. The lack of adequate analytical models which describe the time variations of the amplitude and phase modulating signals—for an input speech signal—has resulted in a variety of frequency scaling techniques which use different analysis filter banks and different modifications of the modulating signals. In the following, the modifications applied to the modulation signals by different frequency scaling techniques are used to analyze and compare the different techniques. We first, however, show that the modification of $A_k(t)$ and $\phi_k(t)$ corresponds to the modification of the short-time spectrum of the speech signal.

Since $z_k(t)$ is the output signal from a bandpass filter having an impulse response $h_k(t)$ it can be expressed as the convolution between the input speech signal $x(t)$ and $h_k(t)$. From eq. (1) this results in

$$z_k(t) = X(\omega_k, t)\exp(j\omega_k t), \tag{4}$$

where

$$X(\omega_k, t) = \int_{-\infty}^{\infty} x(\tau)w_k(t-\tau)\exp(-j\omega_k\tau)d\tau. \tag{5}$$

Comparing eq. (4) with eq. (2) and using eq. (3), we have

$$X(\omega_k, t) = A_k(t)\exp[j\phi_k(t)]. \tag{6}$$

This shows that the amplitude and phase modulations of the carrier $\exp(j\omega_k t)$ are fully described by the composite modulation function $X(\omega_k, t)$. Additional understanding of these modulation functions can be gained from the studies in Refs. 18 and 19. The expression for $X(\omega_k, t)$ in eq. (5) shows that $X(\omega_k, t)$ is equal to the value of the STFT of $x(t)$ at the frequency $\omega = \omega_k$, if $w_k(t)$ is the window function used to weight the input signal.[2,15] It should be emphasized again that in the present discussion the different bandpass filters covering the speech band have, in general, different prototype filters. Hence, for each bandpass filter one can define an STFT which, if evaluated at the center frequency of that filter, gives the corresponding composite modulation function. If all bandpass filters have identical prototype low-pass filters, only a single STFT is needed to find the value of $X(\omega_k, t)$ for each k, by evaluating the STFT at each center frequency. With this understanding, we will refer to $X(\omega_k, t)$ as the STFT of $x(t)$ at $\omega = \omega_k$, even for the general case of nonidentical prototype filters.

Denoting the frequency scaled version of $z_k(t)$ by $z_{qk}(t)$ and the corresponding modified STFT by $X_q(\omega_k, t)$, we have

$$z_{qk}(t) = X_q(\omega_k, t)\exp(jq\omega_k t). \tag{7}$$

The magnitude and phase components of $X_q(\omega_k, t)$ are accordingly denoted by $A_{qk}(t)$ and $\phi_{qk}(t)$, respectively. As noted above, the modification of $X(\omega_k, t)$ needed for exact frequency scaling of speech signals is not known, and the bandwidth of the individual sub-band signals is usually only partially scaled by any given technique. Hence, to avoid excessive interband aliasing when the partially scaled sub-band signals are combined, additional filtering of $z_{qk}(t)$ may be needed. The filtering of $z_{qk}(t)$ can be performed either by bandpass filters having a bandwidth which is q-times the bandwidth of the analysis filters, or equivalently by low-pass filtering the modified STFT by the corresponding low-pass prototype filters. Figure 2 shows a general block diagram for frequency scaling which is based on modifying the STFT of the input signal as discussed above. The impulse response of the synthesis low-pass filters which are used to band-limit the output signals in each channel are denoted by $w_{qk}(t)$. These scaled-bandwidth synthesis filters can generally be obtained from $w_k(t)$ by the relation $w_{qk}(t) = w_k(qt)$. In the diagram of Fig. 2, only the details of the kth channel are given since all the other channels are similar (see solid line). The filtered modified STFT is denoted in Fig. 2 by $\hat{X}_q(\omega_k, t)$. The output scaled speech signal $\hat{y}_q(t)$ is given by

$$\hat{y}_q(t) = \sum_k \hat{z}_{qk}(t) = \sum_k \hat{X}_q(\omega_k, t)\exp(jq\omega_k t), \tag{8}$$

where, as seen in Fig. 2, $\hat{z}_{qk}(t)$ is the kth-channel scaled and filtered bandpass signal. The summation in eq. (8) is over the N sub-bands.

It is clear from the above discussion, and from the block diagram in Fig. 2, that the choice of the filter bank and the STFT modification are the key issues for any given technique. While the block diagram in Fig. 2 provides a basis for comparing different techniques, the actual implementations can differ, either because of historical reasons or the availability of more efficient or convenient ways for implementation.

We turn now to the description of several known frequency scaling techniques in terms of the STFT modification used by each technique. This will exemplify the above discussion and will provide a proper perspective for discussing the FDHS technique and its properties and implementation.

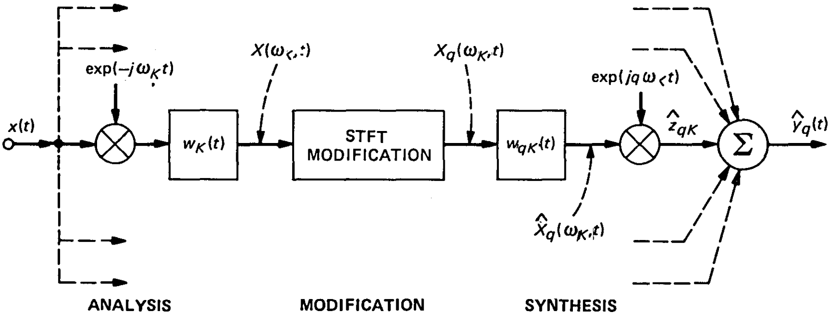


Fig. 2—General block diagram of a frequency-scaling system based on short-time spectral modification (kth channel shown in solid line).

2.1 *Analytic signal rooting*

In the analytic signal rooting technique[3] the number of bandpass filters is chosen to match the formant structure of speech signals so that, preferably, no more than one formant is present in each sub-band. The approach taken in Ref. 3, as well as in the earlier CODIMEX system,[11] is to obtain $z_{qk}(t)$ by raising the analytic signal $z_k(t)$ to the power of q. If $q < 1$, this corresponds to taking the $1/q$ root of $z_k(t)$ which is the origin for the name of this technique. Using the relations in eqs. (4) and (7), we find that the STFT modification performed by this technique is

$$X_q(\omega_k, t) = [X(\omega_k, t)]^q. \tag{9}$$

In terms of the modulating amplitude and phase signals, this modification corresponds to

$$A_{qk}(t) = [A_k(t)]^q, \tag{10a}$$

and

$$\phi_{qk}(t) = q\phi_k(t). \tag{10b}$$

To understand the effect of this modification, we note that since each sub-band is to contain no more than one formant, it can be expected that most often one-pitch harmonic, the one closest to the peak of the formant, is dominant to the other harmonics in that sub-band. From the analysis in Refs. 20 and 21 one can conclude that the phase-scaling operation in eq. (10b) scales the instantaneous frequency of the dominant harmonic in each band by q, but the other lower amplitude harmonics are shifted in such a way that their spacing from the dominant harmonic remains unchanged. The result of this translation is that the fine structure spectral components are not necessarily harmonic, although their spacing is equal to the pitch frequency. The scaling of the amplitude signals in the way given by eq. (10a) can be shown to scale the magnitude of the nondominant components (harmonics) relative to the amplitude of the dominant component. It also affects the intermodulation terms generated by the phase scaling. For $q < 1$, the effect is to reduce the magnitude of the nondominant components relative to the dominant one and hence, effectively, to reduce the bandwidth of the frequency-scaled formants. To avoid excessive interband aliasing, it is particularly important in this technique to use the band-limiting low-pass filters $w_{qk}(t)$ following the modification.

For more effective scaling of the amplitude signals, and with respect to the CQT which uses constant-Q bandpass filters, consider the approach by Ravindra.[9] He suggests that the $A_k(t)$ be spectrally analyzed for each k by an additional bank of filters and that the bandwidth be scaled by scaling the phase in each sub-band—this can be repeated in a tree-like structure.[9] The implementation complexity of this approach, however, appears to be exorbitant.

2.2 Phase vocoder

The phase vocoder, as its name indicates, can be used directly as a vocoder system in which the phase derivative and magnitude of the input signal STFT are coded and transmitted.[1,2,22,23] The phase vocoder technique can also be applied for frequency scaling and this aspect is considered here.[2]

In the phase vocoder, the number of bandpass filters is chosen to match the harmonic structure of voiced speech.[2] This means that a relatively large number of filters is used so that, preferably, no more than one-pitch harmonic is present in each sub-band. The fact that individual harmonics are separately scaled, allows us to infer the characteristics of the modulation signals in each band from known speech properties. In particular, since pitch and vocal tract variations are relatively slow, the bandwidth of each pitch harmonic is quite narrow, as shown by the "pitch teeth" in the input spectrum shown in Fig. 1. In view of this fact, one would expect that even if the pitch harmonics are only shifted to the proper frequencies, without scaling the bandwidth of each pitch tooth, acceptable compression can be achieved (i.e. only a small interharmonic aliasing is expected), provided that the compression ratio is limited to 2 or at most 3. Indeed, this finds support in the results obtained with the TDHS technique which we discuss later.[4,5] However, to shift the pitch harmonics to the proper locations requires knowledge of the pitch frequency or, equivalently, the deviation of each pitch harmonic from the center frequency of the sub-band in which it is located.

Let Ω_k be the pitch-harmonic frequency in the kth sub-band, with center frequency ω_k, and $\Delta\Omega_k$ the deviation of the pitch harmonic from the center frequency; i.e., $\Delta\Omega_k = \Omega_k - \omega_k$. Then, the phase derivative $\dot{\phi}_k(t)$ can be expressed as

$$\dot{\phi}_k(t) = \Delta\Omega_k + \dot{\Psi}_k(t), \tag{11}$$

where $\dot{\Psi}_k(t)$ describes the contribution of the phase variations to the bandwidth of the pitch harmonic in the kth sub-band. In the phase vocoder technique, the phase derivative $\dot{\phi}_k(t)$ is scaled by q, so that in addition to shifting each pitch tooth to its proper location, a partial scaling of its bandwidth is obtained since $\dot{\Psi}_k(t)$ is scaled as well. The amplitude modulation signals are not modified in this technique. However, since individual harmonics are analyzed, the amplitude signal in each band varies slowly [see (a) of Fig. 5 in Ref. 19], and its contribution to the pitch-tooth bandwidth is expected to be small.

Accordingly, the modified amplitude and phase signals are given by

$$A_{qk}(t) = A_k(t), \tag{12a}$$

and

$$\phi_{qk}(t) = \int_{t_0}^{t} q\dot{\phi}_k(\tau)d\tau. \tag{12b}$$

It is observed from eq. (12b) that the constant phase term $\phi_k(t_0)$ is discarded [note: $\phi_k(t) = \int_{t_0}^{t} \dot{\phi}_k(\tau)d\tau + \phi_k(t_0)$]. This can have an effect on the shape of the scaled signal waveform, but because of the relative insensitivity of the ear to a fixed phase distortion, it was not judged to be perceptually significant.[2]

Since in this technique individual pitch harmonics are scaled and the interband aliasing is expected to be small, use of the output synthesis filters, denoted by $w_{qk}(t)$ in Fig. 2, is less compelling than for the ASR technique, but it can still be useful.

It should be noted that the phase vocoder technique can perform time-scale variations of speech signals simply by playing back the signal which has been frequency-scaled by a factor q at $(1/q)$-times the original speed. This restores the original frequency range but scales the signal's time duration by q. This useful property is not shared by the ASR technique because of the way the pitch harmonics are shifted and because of the nonlinear scaling of the amplitude signals. On the other hand, the ASR technique can be useful for restoring speech distorted by a helium atmosphere, where scaling of the formants without changing the perceived pitch of the signal is desired.[1,3]

We turn now to the more recently developed time-domain harmonic scaling (TDHS) technique.[4] Although this technique is most efficiently implemented in the time domain, it was formulated and derived within the STFT framework.

2.3 Time-domain harmonic scaling

As noted in the discussion on the phase vocoder technique, compression factors of up to 3 can possibly be obtained even if the bandwidth of each pitch harmonic is not scaled, provided that the pitch harmonics are shifted to the correct frequency locations. In the phase vocoder this necessitates scaling the phase derivative of the STFT so that $\Delta\Omega_k$, the frequency deviation of the pitch harmonic in the kth sub-band from the center frequency ω_k, is scaled by q. The approach taken by the TDHS technique is to incorporate pitch information which is obtained by a separate pitch detector, into the scaling process.[4] If the pitch frequency is known, the bandwidth of each bandpass filter can be made equal to the pitch frequency and the center frequency of each bandpass filter can be aligned with the corresponding pitch harmonic, so that $\Delta\Omega_k = 0$ for all the bandpass filters which cover the speech band. Here, in principle, the number of bandpass filters also varies with the pitch frequency and is equal to the number of pitch harmonics in the given speech band. Hence, if only shifting of the pitch harmonics is desired, as schematically shown in Fig. 3 (for $q = 1/2$ and $q = 2$), without scaling the pitch-teeth bandwidth, there is no need to modify the modulating amplitude and phase signals (i.e. the STFT), but only to scale the carrier, or center, frequencies. Thus,

$$X_q(\omega_k, t) = X(\omega_k, t), \tag{13}$$

with the understanding that ω_k is chosen to coincide with the pitch harmonic Ω_k in the kth sub-band. Using eq. (13) in eq. (8), and assuming that no synthesis filters are used, the output-scaled signal $y_q(t)$ is given by

$$y_q(t) = \sum_k X(\omega_k, t)\exp(jq\omega_k t). \tag{14}$$

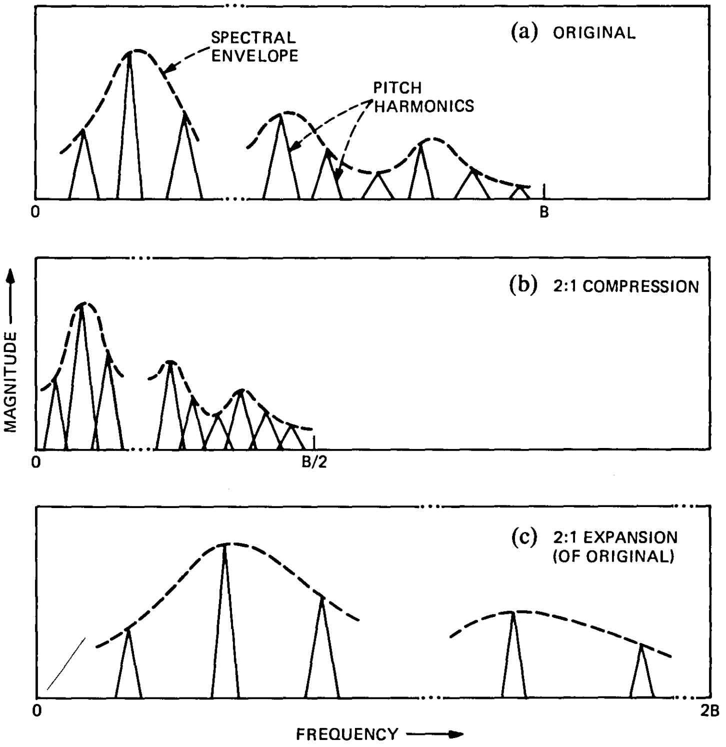


Fig. 3—Schematic spectral representation of frequency scaling by shifting of the pitch harmonics.[5]

Equation (14) can be used to derive an explicit linear relation between $y_q(t)$ and $x(t)$ by substituting the right-hand side of eq. (5) for $X(\omega_k, t)$ in eq. (14). Discretization of this relation yields the TDHS algorithms,[4-6] which were successfully applied in conjunction with several waveform coding systems for an effective reduction in the required transmission bit-rate of speech signals.[5-7] Like the phase vocoder, the TDHS technique can also be used for time scaling of speech signals by playing back the signal at a different speed or, equivalently, at a different sampling rate.

Before we conclude this section, we would also like to mention two additional time- or frequency-scaling systems. One is the scaling system devised in Ref. 24 to which we will refer in a later section; the other is the recent CQT method to which we already referred in Section 2.1.[8,9] In the CQT method a constant Q analysis filter bank is used (i.e., the $w_k(t)$'s in Fig. 2 are not identical) which resembles the type of spectral analysis performed by the ear. Originally the modification consisted of scaling only the unwrapped phase in each band.[8] However, since some bands may contain more than one-pitch harmonic, the amplitude signals may vary significantly and need to be scaled as well. [See (a) of Fig. 7 in Ref. 19.] As mentioned earlier, the approach proposed in Ref. 9 is to analyze the amplitude signal in each band with another bank of filters and to scale the phase signals.

III. FREQUENCY-DOMAIN HARMONIC SCALING TECHNIQUE

The basic FDHS technique is given in Ref. 12. In this section, we relate this technique to the techniques described in the previous section and give its details, including modifications which we introduce in the original sign tracking algorithm.[12]

As in the phase vocoder, the FDHS technique aims at scaling the individual pitch harmonics of voiced speech signals. However, in FDHS, the total phase is scaled (including the constant phase term) which is discarded in the phase vocoder technique. Also, as in other narrowband techniques the amplitude signals remain unmodified.

Therefore, the modified STFT amplitude and phase components are given by

$$A_{qk}(t) = A_k(t), \tag{15a}$$

and

$$\phi_{qk}(t) = q\phi_k(t). \tag{15b}$$

This type of modification has been the underlying modification of several early techniques which are described and analyzed in Ref. 21. The more recent techniques reported in Refs. 24 and 25 are also based on this modification, if only frequency compression or expansion is considered. However, the way the FDHS technique performs the phase modification is specific to this technique as is elaborated in the following.

For noninteger values of q, the phase modification in eq. (15b) must be performed on the unwrapped phase. Otherwise, phase ambiguity of a multiple of 2π, which results from computing the principal value of the phase, may give rise to an incorrect modified phase value. In the phase vocoder technique, the difficult task of explicit phase unwrapping is avoided by directly computing the phase derivative from the real and imaginary components of the STFT and their time derivatives.[2] In the ASR technique, the scaling factors are restricted to $q = 1/2$ for compression and $q = 2$ for expansion.[3] The explicit phase division by 2 for $q = 1/2$ is then avoided by expressing the scaled signal in each band in terms of the input signal and its envelope (amplitude function), with a sign which is determined by means of a simple sign tracking algorithm.[3]

Since compression and expansion by a factor of 2 is of most practical interest (speech quality at higher scaling factors degrades rapidly), the approach of using 2:1 scaling and avoiding explicit phase computation and unwrapping, by means of a proper sign tracking algorithm, is adopted also by the FDHS technique presented in Ref. 12. We now present the details of the FDHS technique and the modifications introduced in the original sign tracking algorithm.

Let $a_k(t)$ and $b_k(t)$ be the real and imaginary parts, respectively, of the input signal STFT, $X(\omega_k, t)$, and $a_{qk}(t)$, $b_{qk}(t)$ the corresponding real and imaginary parts of the modified STFT $X_q(\omega_k, t)$. From eqs. (6) and (15) we find,

$$a_k(t) = A_k(t)\cos\phi_k(t); \quad a_{qk}(t) = A_k(t)\cos[q\phi_k(t)]. \tag{16a}$$

$$b_k(t) = A_k(t)\sin\phi_k(t); \quad b_{qk}(t) = A_k(t)\sin[q\phi_k(t)]. \tag{16b}$$

By restricting q to be 1/2 or 2 and using basic trigonometric relations, the following algorithms were derived in Ref. 12.

3.1 Compression ($q = 1/2$)

Using half-angle trigonometric relations, one finds from eq. (16)

$$a_{k/2}(t) = \mathrm{SGN}[a_{k/2}(t)]\{[A_k(t)/2][A_k(t) + a_k(t)]\}^{1/2}, \tag{17a}$$

$$b_{k/2}(t) = \mathrm{SGN}[b_{k/2}(t)]\{[A_k(t)/2][A_k(t) - a_k(t)]\}^{1/2}, \tag{17b}$$

where the signs [SGN$(\cdot)$] are determined according to the quadrant in which the complex vector (or phasor) $[a_{k/2}(t) + jb_{k/2}(t)]$ is present at any given time instant t. Once initialized, a consistent sign, or quadrant, tracking algorithm is given by the following rules or logic.[12]

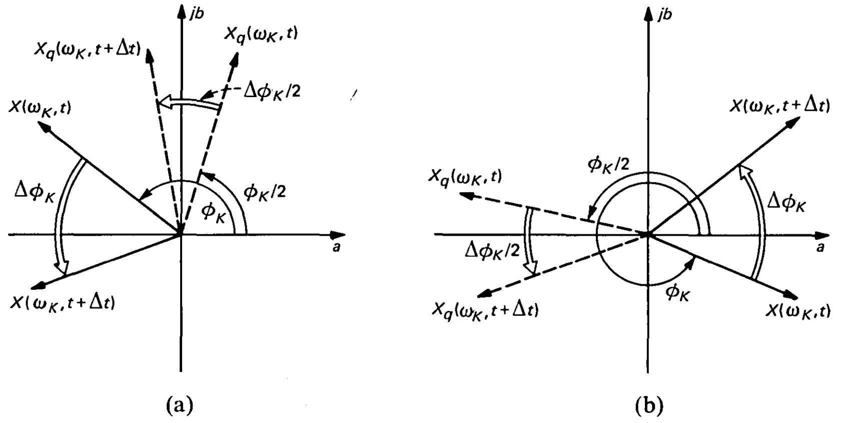


Fig. 4—Demonstration of the validity of the change of sign rules in eq. (18): (a) $\text{SGN}[a_{k/2}(t + \Delta t)] = -\text{SGN}[a_{k/2}(t)]$; (b) $\text{SGN}[b_{k/2}(t + \Delta t)] = -\text{SGN}[b_{k/2}(t)]$.

$$\text{SGN}[a_{k/2}(t + \Delta t)] = \begin{cases} -\text{SGN}[a_{k/2}(t)] & \text{if } A \cap B \text{ is true} \\ \text{SGN}[a_{k/2}(t)], & \text{otherwise} \end{cases} \tag{18a}$$

$$\text{SGN}[b_{k/2}(t + \Delta t)] = \begin{cases} -\text{SGN}[b_{k/2}(t)] & \text{if } \bar{A} \cap B \text{ is true}, \\ \text{SGN}[b_{k/2}(t)], & \text{otherwise} \end{cases} \tag{18b}$$

where A and B are logical variables such that

$$A \text{ is true if } a_k(t + \Delta t) < 0,$$

$$B \text{ is true if } \text{SGN}[b_k(t + \Delta t)] = -\text{SGN}[b_k(t)].$$

The validity of these rules is demonstrated in the phasor diagram shown in Fig. 4. In this figure, the solid-line vectors represent the input STFT at times t and $t + \Delta t$, with a corresponding phase change of $\Delta\phi_k = \Delta\Omega_k \Delta t$, where $\Delta\Omega_k$ is the deviation of the harmonic from the center frequency ω_k. The dashed-line vectors represent the modified STFT and rotate at half the angular velocity. In (a) of Fig. 4 we illustrate a situation in which the input vector crosses the real axis; that is, condition B is "true." Since the real part of the input vector is negative (i.e. condition A is also "true"), the rule in eq. (18a) states that the real part of the modified STFT changes its sign during the time interval Δt, as is indeed the case—the dashed-line vector crosses the imaginary axis. In the same way, (b) of Fig. 4 illustrates a situation in which the imaginary part of the modified STFT changes its sign in agreement with the rule in eq. (18b). Similar diagrams can show that the rules are valid even if the direction of rotation of the phasors shown in Fig. 4 is reversed.*

* The direction of rotation depends on the location of the pitch harmonic with respect

Originally, it was proposed to initialize the sign tracking algorithm described above by assigning positive values to the signs of $a_{k/2}$ and $b_{k/2}$ in all bands,[12] or for that matter, any arbitrary assignment will do as well. However, further analysis shows that such an initialization can result in shifting some pitch harmonics to incorrect frequencies and thereby breaking up the harmonic structure. This situation results when the signs of $a_{k/2}$ and $b_{k/2}$ in a given band happen to be initialized such that the sign of one is correct and the sign of the other is reversed. The sign tracking algorithm in eq. (18) will then cause the scaled signal vector (phasor) to reverse its direction of rotation. This can be verified with the help of a phasor diagram such as in Fig. 4. For simplicity, let the input signal be of the form $A\exp(j\Omega_0 t)$, such that $\Omega_0 = \omega_k + \Delta\Omega$, where ω_k is the center frequency of kth band, and $\Delta\Omega \leq \Delta\omega$ (the bandwidth of that band). Then, the inversion of rotation direction discussed above will result in a scaled signal of the form $A\exp[j(\omega_k/2 - \Delta\Omega/2)t] = A\exp[j(\Omega_0/2 - \Delta\Omega)t]$ instead of the desired signal $A\exp[j(\omega_k/2 + \Delta\Omega/2)t] = A\exp[j(\Omega_0/2)t]$. The situation can be even worse when two adjacent filters share a single harmonic;* that is, the harmonic lies in the transition bands of the two filters as schematically shown in (a) of Fig. 5. In this case, the following conditions may arise: (*i*) The sign initialization in both bands is correct so that the components from the two bands sum up to

$$A_1\exp[j(\omega_1/2)t]\exp[j(\Delta\Omega_1/2)t] + A_2\exp[j(\omega_2/2)t]\exp[-j(\Delta\Omega_2/2)t] = (A_1 + A_2)\exp[j\Omega_0/2)t],$$

where ω_1 and ω_2 are assumed to be the center frequencies of the two bandpass filters under consideration, $\Delta\Omega_1$ and $\Delta\Omega_2$ are the deviations of the given harmonic, of frequency Ω_0, from ω_1 and ω_2, respectively, and A_1 and A_2 are the corresponding signal amplitudes in each band. For a uniform filter-bank design, $A_1 + A_2 = A$, where A is the input signal amplitude. This is the desired result as depicted in (b) of Fig. 5. (*ii*) The sign initialization in one band, say the second one, is incorrect but such that only one of the signs of $a_{k/2}$ or $b_{k/2}$ is reversed. As discussed earlier, if we use phasor description, this will reverse the direction of rotation of the scaled component in that band and will result in

$$A_1\exp[j(\omega_1/2)t]\exp[j(\Delta\Omega_1/2)t] + A_2\exp[j(\omega_2/2)t]\exp[j(\Delta\Omega_2/2)t] = A_1\exp[j(\Omega_0/2)t] + A_2\exp[j(\Omega_0/2 + \Delta\Omega_2)t].$$

to the center frequency ω_k of the sub-band in which it is located (i.e. on the sign of $\Delta\Omega_k$). If $\Delta\Omega_k > 0$, the phasor rotates in a counterclockwise direction, as in Fig. 4, whereas if $\Delta\Omega_k < 0$, its direction of rotation is reversed.

* It is assumed that the side lobes of the filters in the filter bank are sufficiently small so that only two adjacent filters which share the same harmonic are considered.

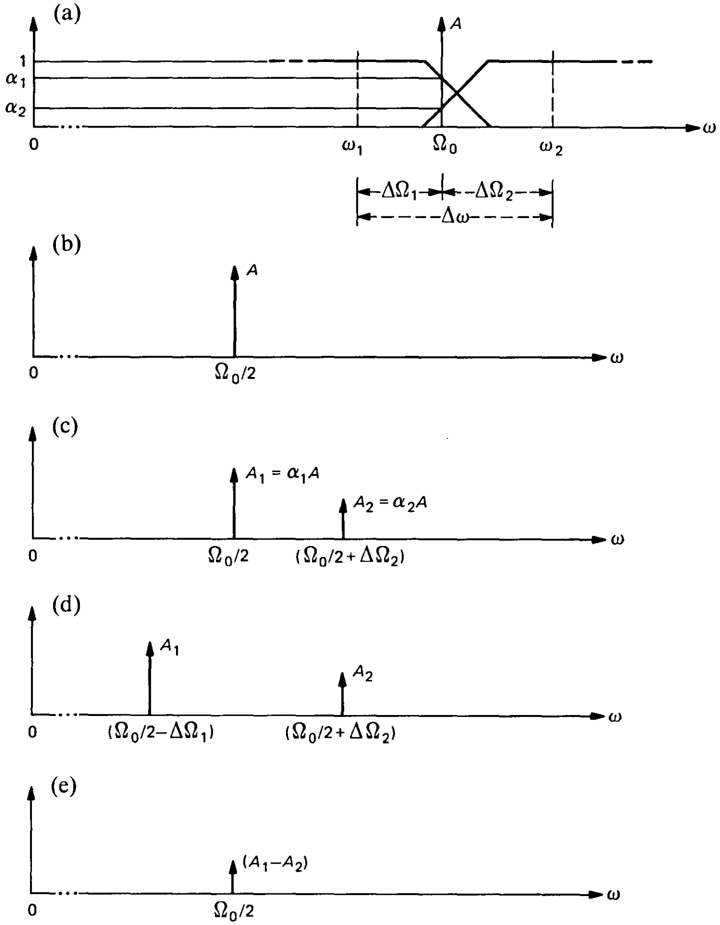


Fig. 5—Schematic representation of the effect of sign initialization on frequency compression ($q = 1/2$) of a harmonic shared by two adjacent filters. (a) Original harmonic. (b) Condition (*i*): correct sign initialization. (c) Condition (*ii*): sign inversion in one band of one of $(a_{k/2}, b_{k/2})$. (d) Condition (*iii*): sign inversion in both bands of one of $(a_{k/2}, b_{k/2})$. (e) Condition (*iv*): sign inversion in one band of both $(a_{k/2}, b_{k/2})$.

This result is shown in (c) of the figure. (*iii*) The sign initialization in both bands is incorrect, and in both bands it is such that only one of the signs of $a_{k/2}$ or $b_{k/2}$ is reversed. In this case, as shown in (d) both components are shifted to incorrect locations. (*iv*) The signs in one band, say the second one, are initialized incorrectly, such that both signs of $a_{k/2}$ and $b_{k/2}$ are reversed. In this case the rotation direction can be shown to remain unchanged, but there is a π phase shift in the phase of the scaled component in this band. Hence, when the two components from the two adjacent bands are recombined in the synthesis stage, the following scaled signal results

$$(A_1 - A_2)\exp[j(\Omega_0/2)t].$$

This is shown in (e). As will be discussed later, designing the filters for minimal overlap, or narrow transition bands, mitigates the amplitude interactions. (*v*) The signs in both bands are initialized incorrectly, but in both bands both signs of $a_{k/2}$ and $b_{k/2}$ are reversed. In this case there is a π shift in both bands and this results in a constant π phase shift in the recombined scaled signal as well, resulting in

$$-(A_1 + A_2)\exp[j(\Omega_0/2)t] = -A\exp[j(\Omega_0/2)t].$$

This situation is of no concern. It may affect the compressed signal waveform but is not found to have any perceptual effect (if we listen to the time compressed version of the signal). Furthermore, in the expansion process the phase is multiplied by 2 and this constant phase shift is cancelled. Similarly, if the harmonic is located within a single band, a sign inversion is produced in the scaled harmonic, which is of no concern.

We have seen that certain errors in sign initialization can have an adverse effect on the resulting scaled signal, and particularly so when a harmonic is shared by two filters. The best approach for initialization would appear to be the use of phase unwrapping along the frequency axis at the time of initialization, so that correct initial signs are assigned to each band. However, this is generally a difficult and error prone task, which typically requires a high-frequency resolution.[13] Also, since speech is nonstationary, reinitialization is needed quite often and a complex phase unwrapping technique will defeat the whole purpose of using a simple sign tracking algorithm. Therefore, we have examined the implications of using the principal value of the phase in each band for assigning the initial signs of $a_{k/2}$ and $b_{k/2}$. To do this, there is no need to actually compute the phase but only to examine the sign of $b_k(t)$ at the instant of initialization. Since the principal value of the phase is assumed to be in the range 0 to 2π, the initial position of the vector $[a_{k/2}(t) + jb_{k/2}(t)]$ must be assumed to be in the first or second quadrant. Hence, a sign initialization according to the principal value of the phase is given by the following simple rules:

$$\mathrm{SGN}[a_{k/2}(t_0)] = \begin{cases} 1 & \text{if } b_k(t_0) > 0 \\ -1 & \text{if } b_k(t_0) < 0 \end{cases}. \tag{19a}$$

$$\mathrm{SGN}[b_{k/2}(t_0)] = 1. \tag{19b}$$

An analysis of this sign initialization shows that either both signs of $a_{k/2}$ and $b_{k/2}$ are correct or both are reversed. If the harmonic is located within a single filter, the scaled harmonic will be shifted to the correct frequency with a possible phase shift of π – which is of no concern.

However, if the harmonic is shared by two adjacent filters, the sign initialization according to eq. (19) can give rise to conditions (*iv*) and (*v*) discussed above. In summary, we face a problem only when a harmonic is shared by two adjacent filters and condition (*iv*) exists. As shown in (e) of Fig. 5, under this condition the harmonic is shifted to the correct frequency location but its amplitude is generally attenuated and it can even be canceled, if $A_1 = A_2$. This possibility of attenuation, or even cancellation, is still of concern. After further study of the problem, we found it possible to devise a procedure for matching the signs in the two bands so that no attenuation of the scaled signal will occur. (A π phase shift of the recombined harmonic is still possible, but as explained earlier this is of no concern). We now explain this sign matching procedure.

For the purpose of simplifying the explanation, consider again a single input complex tone at frequency Ω_0, of the general form $A\exp[j(\Omega_0 t + \phi_0)]$, with Ω_0 satisfying $\omega_1 < \Omega_0 < \omega_2$, where ω_1 and ω_2 are again the center frequencies of two adjacent bands. Following phase scaling and sign initialization according to eq. (19), the modified STFT signals $X_1(t) = X_q(\omega_1, t)$ and $X_2(t) = X_q(\omega_2, t)$ (with $q = 1/2$) are given by

$$X_1(t) = S_{x_1}A_1\exp[j(\Delta\Omega_1 t + \phi_0)/2],$$

$$X_2(t) = S_{x_2}A_2\exp[-j(\Delta\Omega_2 t - \phi_0)/2],$$

where A_1 and A_2 are the magnitudes of the components in the two bands. Further, $A_1 + A_2 = A$, $\Delta\Omega_1 = \Omega_0 - \omega_1$, $\Delta\Omega_2 = \omega_2 - \Omega_0$, and S_{x_1}, S_{x_2} can only take the values of (+1) or (−1) and are used to denote a possible sign inversion because of an incorrect sign initialization. Therefore, the problem is to make S_{x_1} and S_{x_2} equal to each other so that the two components will recombine without attenuation, a sign inversion being inconsequential. To perform this matching of S_{x_1} and S_{x_2} we examine the function $R(t)$ given by

$$R(t) \triangleq X_1(t)/X_2(t) = S_R(A_1/A_2)\exp[j(\Delta\omega/2)t], \tag{20}$$

where $\Delta\omega = \omega_2 - \omega_1 = \Delta\Omega_1 + \Delta\Omega_2$ is the fixed frequency separation between the centers of the two given bands, and

$$S_R = \begin{cases} 1 & \text{if } S_{x_1} = S_{x_2} \\ -1 & \text{if } S_{x_1} \neq S_{x_2} \end{cases}.$$

Our goal then is to have a sign initialization for which $S_R = 1$. Since $R(t)$ is independent of the unknown value of Ω_0, we can find the value of S_R by examining $R(t)$ at any of the specific time instants $t = t_n =$

$nT_0 = n(2\pi/\Delta\omega)$, (where n is an integer) which results in

$$R(t_n) = S_R(A_1/A_2)(-1)^n. \tag{21}$$

Hence, S_R can be determined from*

$$S_R = \text{SGN}[R(t_n)](-1)^n. \tag{22}$$

In practical situations where the filters are not ideal, such as when leakage from other bands is present, and the signal is not purely periodic, $R(t)$ at $t = t_n$ is not necessarily real as is implied in eq. (21). Therefore, we use the sign of the real part of $R(t_n)$; that is,

$$S_R = \text{SGN}[\text{Re}\{R(t_n)\}](-1)^n. \tag{23}$$

To evaluate the right-hand side of eq. (23), there is no need to fully compute $R(t_n)$, as shown below. Let $X_1(t_n) = \alpha_1 + j\beta_1$ and $X_2(t_n) = \alpha_2 + j\beta_2$—for convenience, the explicit time dependence of α_1, β_1, α_2, and β_2 is suppressed. Then, by the definition in eq. (20),

$$\text{SGN}[\text{Re}\{R(t_n)\}] = \text{SGN}(\alpha_1\alpha_2 + \beta_1\beta_2). \tag{24}$$

Hence, assuming that we know two bands which share a single harmonic, the complete initialization procedure (at an appropriate time instant $t_n = nT_0$) consists of first initializing the two signs of $a_{k/2}$ and $b_{k/2}$ according to eq. (19) and then evaluating S_R from eqs. (23) and (24). If $S_R = 1$, the two components are matched. If $S_R = -1$, the signs in one band should be inverted. If the signal is stationary, the above initialization procedure needs to be done only once as the correct sign will continue to be tracked by the sign tracking algorithm in eq. (18). However, even for the stationary case one has to first determine which are the two bands which share a single harmonic because, in general, both bands on the two sides of the band to be initialized may contain signal components. One way to find out which is the correct pair of bands is to compute the phase of $R(t_n)$ for each of the two bands and pick the band which has the smaller phase angle. Again, there is no need to explicitly compute the value of the phase of $R(t_n)$ for the two bands in question, but it is sufficient to compute the ratio $(\alpha_2\beta_1 - \alpha_1\beta_2)/(\alpha_1\alpha_2 + \beta_1\beta_2)$ for each of the two possible pairs, and pick the pair for which this ratio has the smaller magnitude. However, in simulations we have found that, given sufficient frequency resolution (relative to the separation between harmonics), it is also adequate to simply choose the band in which the signal magnitude is larger.

In addition to the above pairing issue, since speech is nonstationary

* By choosing t_n to be a multiple of $2T_0$, the term $(-1)^n$ can be avoided. However, for better tracking of pitch frequency variations we prefer to reduce the time interval between possible sign initializations to T_0.

one is also faced with the problem of detecting the following conditions: onset of speech, transitions from unvoiced to voiced, and the appearance of a pitch harmonic in a band because of pitch variation. In all of these cases, the signs should be reinitialized according to the procedure discussed above. We have found in simulations that satisfactory automatic tracking of the above conditions is obtained by reinitializing the signs in any band k for which

$$|X(\omega_k, t_n)|/|X(\omega_k, t_{n-1})| \geq E_T, \tag{25}$$

where E_T is a preset threshold (typically, as elaborated in Section V, it is set to correspond to an energy increase of 10 dB in the time interval $T_0 = 16$ ms). The mechanization of this initialization process is further detailed in Section 5.1.

3.2 Expansion ($q = 2$)

From eq. (16) we have

$$a_{2k}(t) = A_k(t)\cos[2\phi_k(t)], \tag{26a}$$

and

$$b_{2k}(t) = A_k(t)\sin[2\phi_k(t)]. \tag{26b}$$

Using double-angle trigonometric relations and eq. (26) one finds,

$$a_{2k}(t) = \{2[a_k(t)]^2 - [A_k(t)]^2\}/A_k(t), \tag{27a}$$

and

$$b_{2k}(t) = 2\, a_k(t)b_k(t)/A_k(t), \tag{27b}$$

where it is assumed that $A_k(t) \neq 0$. If $A_k(t) = 0$, then, directly from eq. (26), $a_{2k}(t) = b_{2k}(t) = 0$.

It is seen that because the phase is multiplied by an integer ($q = 2$), a phase ambiguity which is a multiple of 2π is of no concern in frequency expansion. Note also that if the compressed signal is expanded (for signal reconstruction), A_k, a_k, and b_k, should be replaced in eq. (27) by $A_{k/2}$, $a_{k/2}$, and $b_{k/2}$, respectively.

In comparing the STFT modifications, as expressed by eq. (15) for this technique and by eq. (10) for the ASR technique, one may question if it is useful to use eq. (10a) in place of eq. (15a). The answer is no. The reason is that, again, with practical filters a pitch harmonic may be shared by two adjacent filters. Hence, if a nonlinear modification, such as that in eq. (10a), is applied to the amplitude signal in each band, the two signals (which are components of the same harmonic) will in general not recombine to the correct magnitude. Again, since the analysis is narrow-band the scaling of the amplitude signal is generally of secondary importance.

IV. EFFICIENT DISCRETE-TIME IMPLEMENTATION

The basis for the discrete time implementation of the FDHS technique is the general block diagram for frequency scaling in Fig. 2. However, in this form (following discretization) it is highly inefficient because a large amount of computation is needed to perform the filtering analysis and synthesis operations. It has been demonstrated in several works[14-16,26] that a discrete Fourier transform (DFT) formulation, using the FFT algorithm, can be used for STFT analysis and synthesis with its accompanying large saving in computation. We particularly found the weighted overlap-add (WOLA) method given in Ref. 14, to be most suitable for our application. However, we had to extend the particular scheme given in Fig. 2 of Ref. 14 to accommodate situations in which both the analysis and synthesis windows (the prototype low-pass filter impulse responses) have durations longer than the transform size N. This arises from the need to use analysis filters with narrow transition bands, and, hence long duration, when performing frequency compression. This design minimizes filter overlap and lowers the probability of obtaining harmonics shared by adjacent filters, and, therefore, lessens the effect of possible incorrect sign initialization in such bands.

We begin by showing that a discrete-time version of the block diagram in Fig. 2 of this paper has the basic form considered in Ref. 14, so that the WOLA scheme developed is indeed suitable for our application.[14]

Figure 6 shows a discrete-time form of the block diagram in Fig. 2 which is made to match Fig. 3 in Ref. 14 (with somewhat different notation). This requires some clarification, which is given next.

The discrete-time input signal $x(nT)$ represents samples of $x(t)$ which is assumed to be sampled at or above its Nyquist rate, with T being the sampling interval. The signal band is divided into N equally spaced contiguous sub-bands with center frequencies $\omega_k = 2\pi k/(NT)$, $k = 0, 1, \cdots, N-1$. The discrete-time STFT, $X(\omega_k, nT)$, is obtained for each value of k, by modulating $x(nT)$ by $\exp(-j\omega_k nT) = W_N^{-nk}$, where

$$W_N \triangleq \exp(j2\pi/N), \tag{28}$$

and filtering the modulated signal by the low-pass filter $h(nT)$, which is the sampled version of $w_k(t)$. Here all N prototype filters $w_k(t)$ are identical. Since $h(nT)$ is approximately band-limited to $\Delta\omega/2 = \pi/NT$, its output signal can be decimated. To reduce frequency-domain aliasing, and because of considerations related to the sign tracking algorithm used by the FDHS technique, the decimation factor R, an integer, typically satisfies $R < N$. The decimated STFT, $X(\omega_k, nRT)$, is modified according to the modification algorithms of the FDHS technique, using discretized forms of eqs. (17) and (27) for compression

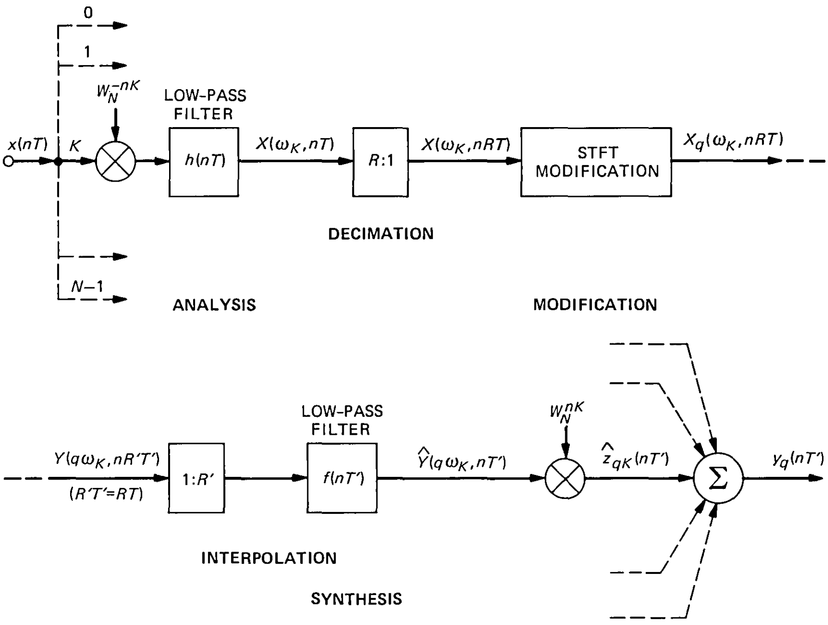


Fig. 6—Discrete-time form of the block diagram in Fig. 2, including decimation and interpolation of the sub-band signals.

and expansion, respectively. The modified STFT, $X_q(\omega_k, nRT)$, can be considered to be the STFT of the desired scaled signal at the scaled center frequency $q\omega_k$, and as being sampled at a rate which corresponds to the scaled signal bandwidth, i.e. with a sampling interval $T' = T/q$. It is suitable, therefore, to rename $X_q(\omega_k, nRT)$ as $Y(q\omega_k, nR'T')$, where R' is related to R through the requirement $R'T' = RT$ (i.e., $R' = qR$), so that the original time scale is maintained. For synthesizing the output scaled signal from the decimated and modified STFT signals, $Y(q\omega_k, nR'T')$, $k = 0, 1, \cdots, N-1$, these signals must first be interpolated by a factor R'. The interpolation is done in the scheme shown in Fig. 6 by inserting $(R'-1)$ zeroes between adjacent samples. This is represented by the box labeled 1:R' in Fig. 6. The result is processed with a low-pass filter $f(nT')$ having a nominal bandwidth of $\pi/(NT')$.

Since the interpolation is not ideal, we denote the interpolated STFT by $\hat{Y}(q\omega_k, nT')$. Modulation of the base-band signals with the complex sequence $\exp(jq\omega_k nT')$ results in the complex bandpass signals $\hat{z}_{qk}(nT')$, $k = 0, 1, \cdots, N-1$. Using $\omega_k = 2\pi k/NT$ and $T' = T/q$, we note that $\exp(jq\omega_k nT') = W_N^{nk}$, where W_N is defined in eq. (28). Thus, the input and output modulating sequences are complex conjugates of each other and are identical to the discrete transform kernels used in the DFT. Finally, by summing the N complex bandpass signals the frequency scaled output signal $\hat{y}_q(nT')$ is obtained.

Thus, we have seen that the general block diagram given in Fig. 2 can be implemented in the discrete-time form shown in Fig. 6. From the identity between this figure and Fig. 3 in Ref. 14, the more efficient WOLA scheme presented in Ref. 14 can be directly applied for the implementation of the block diagram in Fig. 6. However, as mentioned earlier, it is import t that frequency compression with the FDHS technique be performed with long duration window functions (longer than N). Hence, the scheme shown in Fig. 2 of Ref. 14 needs to be modified to accommodate this paticular situation. For clarity of presentation, and since the scheme shown in Fig. 2 of Ref. 14 can be used for expansion without any change, we briefly explain this scheme before we show its modification.

For convenience, we denote the analysis and synthesis window sequences by $h(n)$ and $f(n)$, respectively, dropping the explicit sampling intervals used in the previous notation. We now duplicate in Fig. 7 the scheme shown in Fig. 2 of Ref. 14, with some changes in notation to match the notation used in this paper.

According to this scheme, the input data samples are shifted into an input data buffer of length N, R samples at a time, corresponding to the R:1 decimation shown in Fig. 6. The input data block is weighted by $h(-n)$, which has its origin at the center of the block (see Fig. 7). The weighted data block is transformed using the FFT algorithm. The resulting STFT has its time reference at the beginning of the block (i.e., a sliding time reference); hence, a linear phase shift, corresponding to the time interval between the fixed time origin, say, $t_0 = 0$, and the beginning of the transformed data block, must be introduced. In addition, since the time origin of the analysis window is at the center of the block, a circular rotation of the weighted data by $N/2$ points is also needed.[14] Performing this rotation by phase modification in the frequency domain results in an overall phase modification by $(-1)^k W_N^{-nRk}$, $k = 0, 1, \cdots, N = 1$, as shown in Fig. 7. The result is the desired discrete-time STFT, $X(\omega_k, nT)$, in a fixed time reference, which is to be modified for frequency scaling. The modified STFT $Y(q\omega_k, nR'T')$ is translated back to the sliding time reference by applying the complementary phase modification $(-1)^k W_N^{nR'k}$. The output scaled signal is obtained by inverse transforming the modified STFT (in the sliding time reference), weighting the resulting data block by the synthesis window $f(n)$, and overlap-adding the weighted block to the output buffer. The output buffer shifts out R' samples for every R samples that are shifted into the input buffer. Also, to facilitate the overlap-add operation, R' zeroes are shifted into the output buffer as the processed signal samples are shifted out. Since the input and

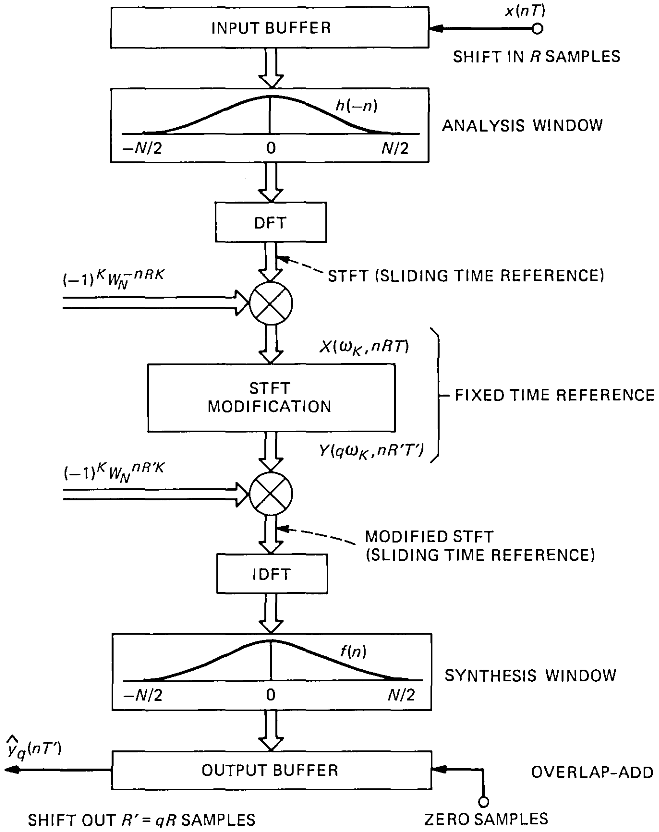


Fig. 7—A WOLA block implementation scheme for short-time Fourier analysis, modification, and synthesis.[14]

output sampling intervals are T and T', respectively, the time scale is not altered and the flow of data is uninterrupted. However, if time scaling is desired, the output data can be stored first and then replayed at the original sampling rate, resulting in a time-scaled signal (by the factor $q = R'/R$) which occupies the original frequency band.

We turn now to the more general situation in which $h(n)$ and $f(n)$ have longer durations than N samples. Let L_h and L_f denote the durations of $h(n)$ and $f(n)$, respectively, and let $L_h = m_h N$, $L_f = m_f N$, where m_h and m_f are positive integers. (If necessary, zeroes can be appended to the impulse responses to satisfy these conditions.) The appropriate scheme for this case is shown in Fig. 8 and is explained below.

It has been established in earlier works,[15,22,26] as well as repeated in Ref. 14, that if $L_h > N$, the N data points to be transformed are

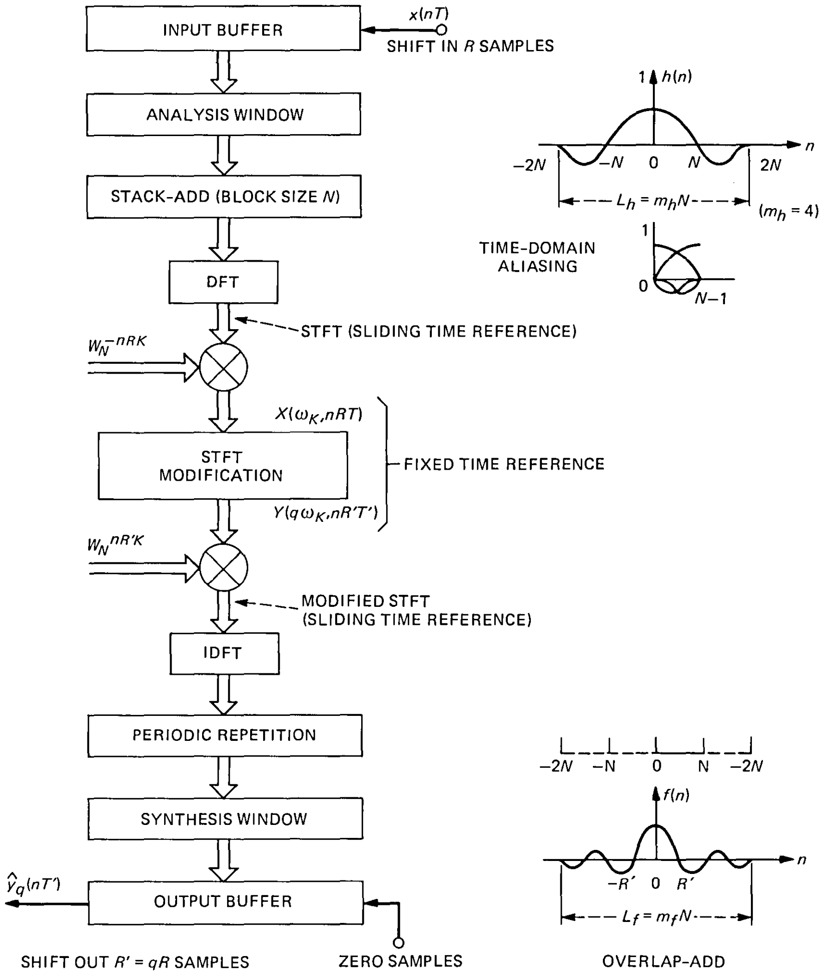


Fig. 8—Modified implementation scheme to accommodate analysis and synthesis windows which are longer than the transform block size N.

obtained by time aliasing the weighted L_h data points into N points. This can be seen as a stack-add operation of the m_h data segments, each of duration N, as shown in Fig. 8. Because of the stacking operation, the time origin of the transform is seen to be aligned with the time origin of the data window, and, hence, no circular rotation by N is needed; that is, there is no need to multiply the transformed data by $(-1)^k$. As before, the phase modification by W_N^{-nRk} is needed to obtain $X(\omega_k, nRT)$ in the fixed time reference. Following the STFT modification, the conversion of the sliding time reference, and the inverse transformation, N data points are obtained to which the synthesis window of duration $L_f = m_f N$ is to be applied. To do this, we periodically repeat the given block of N data points m_f times.* To be consistent with the analysis window weighting, the center of the synthesis window is aligned with the beginning of the given data block, as shown in Fig. 8. The weighted data is then overlapped-added to the output data buffer as before.

Before we conclude this section, several comments are in order. First, it should be noted that the scheme in Fig. 8 can be used also for the case considered in Fig. 7; that is, when the windows have a length which is less or equal to N. This is done by appending zeroes on each side of each window so as to make the duration of each $2N$ (i.e., $m_h = m_f = 2$). The stack-add operation will provide the circular shift by $N/2$ points, as needed in Fig. 7, and hence eliminate the need for multiplying by $(-1)^k$, as is the case in Fig. 8.

Second, it is observed that the implementation schemes presented above can, in principle, be used for scaling an input signal by any rational factor R'/R. However, for frequency compression with $q = 1/2$, the FDHS technique provides a particularly efficient realization of the STFT modification because the sign tracking algorithm avoids the need for explicit phase unwrapping. Expansion by integer factors is the simplest operation because the principal value of the phase can be used and the filter design considerations are simple. For other rational scaling factors it appears that one has to resort to phase unwrapping with its attendant complexity.

Third, in noncoding applications it may be desired to obtain a frequency compressed signal at the original sampling rate, i.e., oversampled. The needed interpolation can be embedded in the above implementation schemes. This is done simply by enlarging the modified STFT transform size ($1/q$ times) by padding with zeroes (at the center of the transform block). Following the inverse transform (IDFT) and the weighting by a suitable $f(n)$ of the longer data block, it is overlapped-added to the output data buffer. The data in both the input and output data buffers is accordingly shifted by R samples at a time.

Finally, it is of interest to point out that the scheme shown in Fig. 8 offers a generalization of the TDHS technique[4] as explained below.

Let us apply the principles of TDHS, as elaborated in Section III, to the scheme in Fig. 8. In principle, this means using a pitch-adaptive analysis filter bank—N is made equal to the pitch period and $h(n)$ is varied accordingly—and applying no modification to the STFT, except for scaling the center frequencies, as expressed by eq. (13) in Section III. Since the STFT is not modified, there is actually no need to use a

* This takes into account the underlying periodicity in the IDFT result because of the discretization in frequency. It can also be concluded from eq. (14) in Ref. 14.

transform—the phase shifts can be done by circular rotations in the time domain.[14] The part of the scheme in Fig. 8 between the stacking and the weighting by $f(n)$ collapses to a counter clock wise circular rotation in the time domain by $(R' - R)n$ Mod N_p samples, where N_p is the pitch period. The implementation scheme of the generalized TDHS technique is shown in Fig. 9.

The TDHS algorithms in Ref. 4 correspond to specific choices of R, R', and $f(n)$. For example, for 2:1 compression the values are $R = 2$, $R' = 1$ and $f(n)$ is a unit impulse; that is, $f(n) = \delta(n)$. The new generalized TDHS scheme above offers more flexibility through the possible use of more general synthesis windows, $f(n)$. It also could provide means for additional interband filtering such as $w_{qk}(t)$ in Fig. 2, which could improve the quality of the scaled speech signal, although at the expense of additional computation. This generalization of the TDHS technique is now under further investigation.

V. DESIGN CONSIDERATIONS AND SIMULATIONS

The implementation scheme in Fig. 8 applies both to compression and expansion. However, the type of frequency scaling performed affects the design requirements for the analysis/synthesis system. For this reason, we discuss below the design of the compression and expansion systems separately.

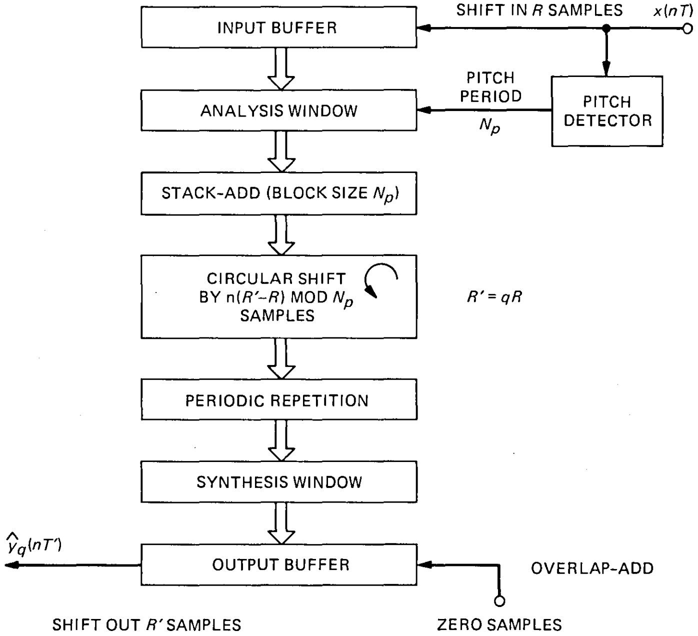


Fig. 9—Implementation scheme of the *generalized* TDHS technique.

5.1 Compression system design

A basic assumption in the development of the FDHS technique is that each filter in the analysis filter bank contains no more than one pitch harmonic. Hence, the number of filters in the filter bank should equal or exceed the largest number of harmonics expected in the given speech band. To accommodate low-pitch male speakers, we have chosen the number of filters to be $N = 128$. With a sampling rate of 8 kHz, which is typical for telephone bandwidth speech, this corresponds to a nominal frequency resolution of 62.5 Hz. This may not be sufficient resolution for very low-pitch speakers but was chosen as a compromise because of another conflicting requirement elaborated below. The prototype low-pass filter $h(n)$ has, therefore, a nominal bandwidth of 31.25 Hz. Besides the usual requirement that the transform of $h(n)$ have low side lobes to minimize interband leakage, it is extremely important that the transition band of $h(n)$ be as narrow as possible to minimize the overlap between the bandpass filters. The importance of minimizing the overlap stems from the difficulties in identifying bands which share a harmonic with an adjacent band and in matching the signs in those bands to avoid attenuation of the corresponding harmonic, as we discussed in detail in Section 3.1. By reducing the overlap between the filters, the probability of occurrence of this condition is lowered. In addition, it is also lowered by using the minimum number of filters necessary to separate the pitch harmonics. For this reason, we did not increase the transform size to 256 and have chosen N to be 128. For female voices, the best results are obtained with $N = 64$. However, since N is fixed the value of $N = 128$ was found to match both male and female voices. Other considerations in designing $h(n)$ are as follows.

In principle, one can reduce the transition-band width by increasing the filter length L_h. However, since speech is not a stationary signal and has a typical quasi-stationarity interval of several tens of ms, the length of $h(n)$ should be limited accordingly. Therefore, we have limited its length to 512 samples, i.e., to a duration of 64 ms for 8-kHz sampling rate, with the main lobe duration being 32 ms. With $L_h = 512$ we have $m_h = L_h/N = 4$ and, for this reason ($m_h > 1$), it was necessary to develop the scheme in Fig. 8, as an extension of the scheme in Fig. 7 from Ref. 14. To meet the requirements for low side lobes and a narrow transition band, we have chosen $h(n)$ to be an optimal equiripple filter[27] and designed it with the filter design package[28] available on our computer system.*

* The maximum filter length that can be designed on our system is 511. A 512-point filter is formally defined by appending a zero.

An additional requirement on the analysis filter bank is that its overall response be uniform. This way the two components of a harmonic which is shared by two filters will sum up to the correct magnitude. In terms of $h(n)$, a necessary and sufficient condition for the filter bank to be uniform is[22]

$$h(nN) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}. \tag{29}$$

While this condition is easily satisfied by windowing for example a $\sin(x)/x$ type response, the optimal equiripple filter does not necessarily satisfy this condition. Therefore, we had to use a trial and error approach and repeat the design several times until our design goals were met. Figure 10 presents the results of the final design which we denote by $h_0(n)$. The analysis window $h_0(n)$ is shown in (a) of Fig. 10, its frequency response in (b), and the composite filter bank frequency response in (c). The peak-to-peak ripple of 0.2 dB in the composite response is quite acceptable. The composite frequency response of a filter bank having a prototype low-pass filter $h(n)$ is simply found by transforming the sequence $d(n)$ defined by

$$d(n) = \begin{cases} h(n) & \text{if } n \text{ Mod } N = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{30}$$

Before we consider issues related to the modification of the STFT signals, we would like first to consider the issues involved in designing a synthesis window $f(n)$ that provides an adequate reconstruction of the input signal when no spectral modifications are performed. Since we deal with the case where $L_h > N$ (i.e., $m_h > 1$), the synthesis filter has the difficult task of undoing the time aliasing, because of the stacking operation, as well as the frequency aliasing, because of the decimation of the signal in each band by the factor R.

It was shown in Refs. 16 and 24 that for exact reconstruction the following relation between $h(n)$ and $f(n)$ should hold

$$\sum_{s=-\infty}^{\infty} f(n - sR)h(pN - n + sR) = \delta(p), \quad \text{for all } n, \tag{31}$$

where $\delta(p) = 1$ for $p \doteq 0$ and zero otherwise. If we assume that N is divisible by R, then eq. (31) is periodic in n with period R and constitutes a set of R conditions (for $n = 0, 1, \cdots, R - 1$). In particular, for $R = N$, it is seen from eq. (31) that the design of $f(n)$ is equivalent to the problem of designing N inverse sub-filters.[16,24] Since $h(n)$ and $f(n)$ are both of finite duration (FIR), exact reconstruction cannot be obtained. However, if $R < N$ there is less aliasing in the frequency domain and the problem is relaxed. In view of the equivalent scheme in Fig. 6, we expect that for $R < N$ the use of a reasonably good

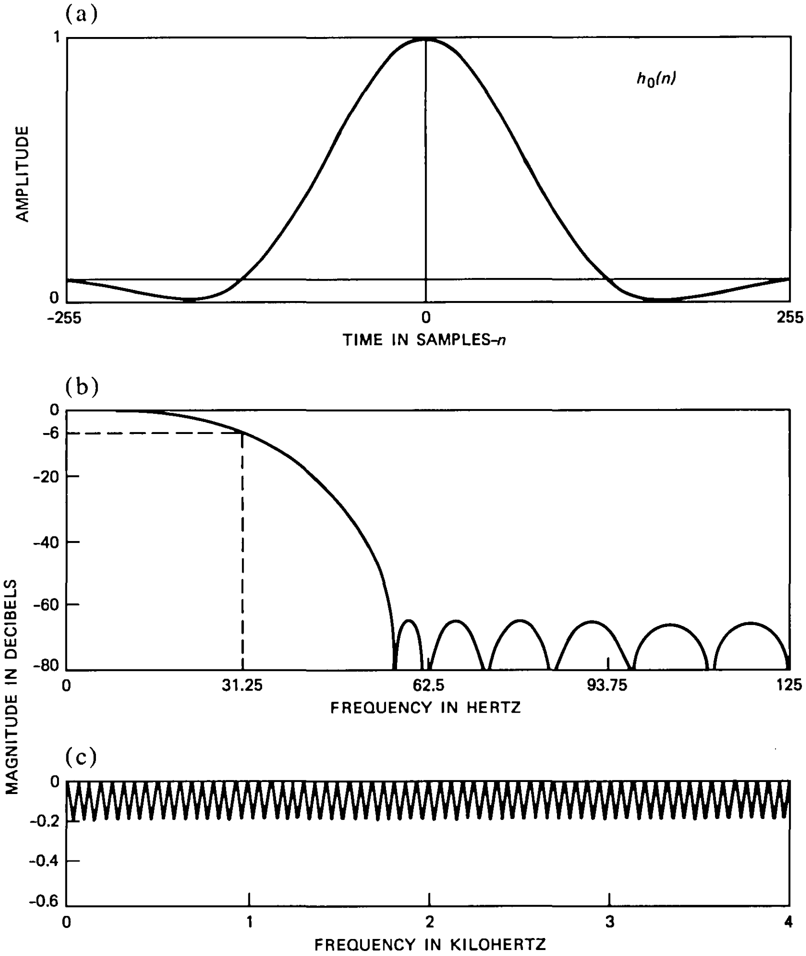


Fig. 10—Designed analysis prototype filter $h_0(n)$. (a) Impulse response. (b) Frequency response. (c) Composite filter-bank frequency response.

interpolation filter will also result in an acceptable reconstruction error. Therefore, we have initially selected four interpolation filters which have simple analytical respresentations. This facilitates the variation of their bandwidth by a change of a parameter. These windows are symmetrical and can also be made to have the usually desired property of interpolation filters, namely,

$$f(nR) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}. \qquad (32)$$

The four synthesis window functions considered are a rectangular window $[w_R(n)]$, a Hanning window $[w_H(n)]$, a $\sin(x)/x$ function multiplied by a Hanning window $[w_{HS}(n)]$, and a special window function $[w_M(n)]$ which was derived in Ref. 4 and has some interesting properties.* The analytical representations of these windows are

$$w_R(n) = \begin{cases} 1 & |n| < L, \\ 0 & |n| \geq L \end{cases};$$

$$w_H(n) = \begin{cases} \frac{1}{2}[1 + \cos(\pi n/L)] & |n| \leq L, \\ 0 & |n| > L \end{cases} \tag{33}$$

$$w_{HS}(n) = \begin{cases} \frac{1}{2}[1 + \cos(2\pi n/L_f)] \cdot [\sin(\pi n/L)/(\pi n/L)] & |n| \leq L_f/2 \\ 0 & |n| > L_f/2 \end{cases}, \tag{34}$$

and

$$w_M(n) = \begin{cases} \frac{L}{L_f} \sin(\pi n/L) \cot(\pi n/L_f) & |n| \leq L_f/2, \\ 0 & |n| > L_f/2 \end{cases} \tag{35}$$

where L is an integer which determines the bandwidth of the synthesis window and L_f is its length, which is assumed to be an even multiple of L. In particular, the rectangular and Hanning windows in eq. (33) have a duration of $L_f = 2L$. Note that if L is chosen to be equal to R, then eq. (32) is satisfied. The above four window functions are shown in Fig. 11 for several values of L_f/L. The reconstruction error expected from using these synthesis windows, with the analysis window $h(n) = h_0(n)$ shown in Fig. 10, was found by evaluating the left-hand side of eq. (31) and computing its deviation from the desired value on the right $[\delta(p)]$. An average mean-square error (mse) was defined as follows. Let $V_n(p)$, n = 0, 1, $\cdots$, $R - 1$ be defined by the left-hand side of eq. (31) under the assumption that N is an integer multiple of R and let

$$\epsilon_n^2 \triangleq \sum_p [V_n(p) - G\,\delta(p)]^2 \quad n = 0, 1, \cdots, R - 1, \tag{36}$$

* This window is described by eq. (53) in Ref. 4 (even case) and has the interesting property of satisfying an equation like (32) in both the time and frequency domains. Thus, as a prototype filter it results in a uniform filter-bank response, whereas in a WOLA-type operation it results in a uniform time response [see eq. (52) in Ref. 4]. For a particular choice of parameters, it becomes identical to the well-known Hanning window.[4]

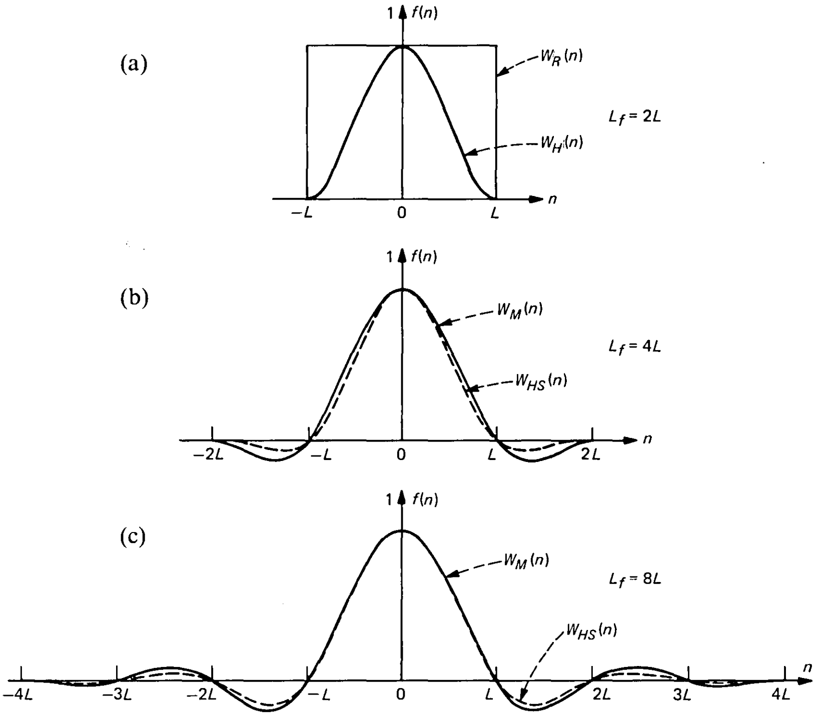


Fig. 11—Synthesis window functions. (a) Rectangular [$w_R(n)$] and Hanning [$w_H(n)$] windows. (b) $w_M(n)$ and $w_{HS}(n)$ windows for $L_f = 4L$. (c) w_M and w_{HS} windows for $L_f = 8L$.

where G is a constant given by

$$G = \frac{1}{R} \sum_{n=0}^{R-1} V_n(0), \tag{37}$$

and, in general, is not equal to 1 as assumed in eq. (31). We have introduced G since we consider a reconstruction which differs from the original signal by a constant gain term as being errorless. An average mse is now defined by

$$\epsilon^2 = \frac{1}{R} \sum_{n=0}^{R-1} \epsilon_n^2. \tag{38}$$

It can be shown that the choice of G according to eq. (37) minimizes ϵ^2; that is, $\partial\epsilon^2/\partial G = 0$. The values of ϵ^2, in dB, which were obtained for the above synthesis windows, using the analysis window $h_0(n)$ are shown in Fig. 12 as a function of R, which is given in terms of the transform size N. For each value of R, ϵ^2 was computed for different values of L. The minimum value of ϵ^2 so obtained is shown in Fig. 12.

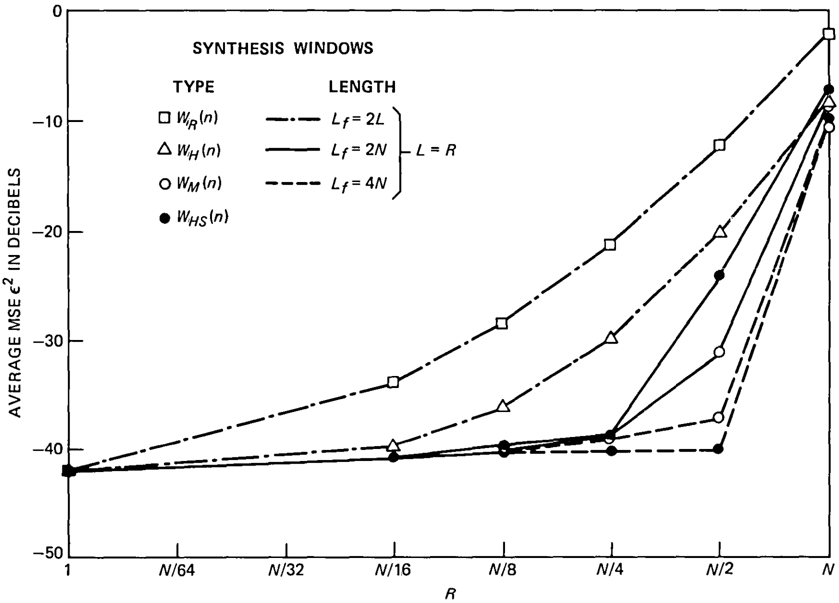


Fig. 12—Minimum values of the average mse ϵ^2 for different synthesis windows, paired with the analysis window $h_0(n)$ in (a) of Fig. 10, as a function of the decimation factor R, or input data shift. N is the transform block size.

In fact, for the above windows, the minimum is attained when $L = R$; that is, eq. (32) is satisfied.

Note also that in Fig. 12 the length of the synthesis windows $w_R(n)$ and $w_H(n)$ varies with R, since $L = R$ and $L_f = 2L$. However, the length of $w_{HS}(n)$ and $w_M(n)$ is assumed to be fixed—either $L_f = 2N$ or $L_f = 4N$—but the bandwidth still varies with R since $L = R$. It is seen from Fig. 12 that with all four windows R must be less than N. To save computation we wish, of course, to use the largest possible value of R. For $R = N/2$ and $L_f = 4N = 512$ the preferred window is $w_{HS}(n)$, whereas for $R = N/2$ and $L_f = 2N$ it is $w_M(n)$ (in both cases $L = R = 64$). To reduce the output buffer size and to save computation, we preferred using $L_f = 2N$. The resulting reconstruction error (below −30 dB) is quite acceptable. The other two windows [$w_R(n)$ and $w_H(n)$] result in reconstruction errors too high to be useful with this value of R.

The above results are for an analysis-synthesis system without the modification of the STFT signals. Since the frequency scaling affects the output sampling rate (the interpolation factor becomes $R' = qR$, where q is the scaling factor) and the sign tracking algorithm limits the maximum value of R (as elaborated later), the choice of R and the synthesis window function must be carefully reconsidered to account for the spectral modification.

For proper operation of the sign tracking algorithm, the phase in each frequency cell or band should not change by more than $\pi/2$ between observations in order that the crossing of the complex signal vector from one quadrant to another will be accounted for and the signs of its real and imaginary parts be tracked correctly. The maximum phase change occurs when a harmonic has the largest deviation $\Delta\Omega_{\max}$ from the center frequency. Taking in account the overlap between filters, we let $\Delta\Omega_{\max} = \Delta\omega$, which is the difference between adjacent center frequencies. During R samples, the phase can change by up to $\Delta\omega RT$ and, hence, with $\Delta\omega = 2\pi/(NT)$ we get the condition that $2\pi R/N \leq \pi/2$; that is, $R \leq N/4$. Therefore, with $N = 128$ we have $R \leq 32$, and to save computation we choose $R = 32$.

It is noted that this choice of R was dictated by the FDHS modification technique used and not by the analysis-synthesis system. However, while a transform of the input data must be taken every 32 samples for proper sign tracking, the modified STFT can actually be computed at a lower rate corresponding to a decimation factor $R_M > R$. This obtains because the analysis-synthesis system is capable of acceptable reconstruction for larger values of the decimation factor. Furthermore, since $q = 1/2$ (2:1 compression), the interpolation factor, or the output data shift, is $R' = R_M/2$ so that the modified STFT needs to be computed only every $N = 128$ samples. This corresponds to $R' = N/2 = 64$ which, as shown earlier, can be accommodated by the analysis-synthesis system with an acceptable reconstruction error.

In summary, the following windows and parameters are used. The transform size is $N = 128$; the analysis window is $h(n) = h_0(n)$ of Fig. 10 ($L_h = 512$); the input data shift is $R = N/4 = 32$; the STFT modification is computed every $R_M = N = 128$ samples; the output data shift is $R' = R_M/2 = N/2 = 64$; and the synthesis window chosen is $f(n) = w_M(n)$ of eq. (35), with $L_f = 2N = 256$, and $L = R' = 64$. This selection of parameters and window functions is supported by simulation results with both synthetic and natural speech signals as will be illustrated subsequently.

Before we present the design considerations for the expansion system, we wish to explain an additional issue related to the selection of the synthesis window and give details of the STFT modification. In Fig. 2, the synthesis filters were assumed to have a bandwidth which is q times the bandwidth of the analysis filters. In the implementation described above, this corresponds to using synthesis filters with $L = N$ and not $L = R'$ as we have chosen, which widens the synthesis filter bandwidth since $R' < N$. However, with practical synthesis filters the use of $L = N$ was found to cause inband attenuation and an increase in the reconstruction error. For $R' = N/2 = 64$, the use of the above-selected window functions, $w_M(n)$ for $L_f = 256$ or $w_{HS}(n)$ for $L_f = 512$

with $L = R'$ gave better or as good results as several other window functions having bandwidth in the normalized frequency range of π/R' to π/N. However, for smaller values of R' an advantage was found in using $L = N/2 > R'$ because of the additional filtering provided by the synthesis window.

We discuss now the details of the STFT modification for frequency compression. The modification is performed in each band according to the expressions in eq. (17) at time instants $t = nR_MT$, where R_M is the modification decimation factor, chosen to be $R_M = N = 128$. The sign tracking is done according to eq. (18) at a higher rate, namely at time instants $t = nRT$, with $R = N/4 = 32$. The condition for sign initialization in each band is determined by eq. (25). The ratio on the left-hand side of eq. (25) is computed at time instants $t = t_n = nT_0$, where $T_0 = NT$, which is also the STFT modification update interval since $R_M = N$. If the ratio exceeds the threshold value E_T in a given band, the signs in that band are initialized. On the basis of simulations, E_T was chosen to correspond to an energy increase of 10 dB during the time interval $T_0 = 16$ ms for 8-kHz sampling rate. This choice of E_T was found to indicate quite reliably the onsets of speech, the transitions from unvoiced to voiced, and the crossing of a pitch harmonic into a given band from a neighboring band. At the same time, this value was found to be sufficiently high to prevent reinitialization because of the normal amplitude fluctuations in each band during sustained voiced intervals.

If the initialization condition is met, the initialization process consists, on the basis of the discussion in Section 3.1, of the following operations. First, the signs in all the cells, or bands, which need to be initialized are set according to eq. (19); that is, according to the principal value of the phase in each cell. Then, starting from the lowest frequency cell to be initialized, a "pairing" process is performed. That is, for each cell to be initialized one of its adjacent cells is picked for sign matching to provide for the situation in which a pitch harmonic is shared by two adjacent bands. This is done either by picking the band which gives an $R(t_n)$ [see eq. (20)] with a smaller phase angle (see Section 3.1) or, as preferred in our simulations and to save computation, by picking the band in which the signal magnitude is larger. Note that if cell k_0 is to be initialized and cell $k_0 + 1$ is picked for sign matching, then, even if cell $k_0 + 1$ is also to be initialized, this cell is skipped in the pairing and matching process since it was already chosen to be matched to cell k_0. Finally, the sign matching of the chosen pairs is performed by computing S_R for each pair, using eqs.(23) and (24), and inverting the signs in one of the bands of those pairs for which S_R is found to have a negative value.

5.2 Expansion system design

Expansion by an integer factor avoids the phase ambiguity problems and is, therefore, a relatively easy task. The main requirement is sufficient frequency resolution in the spectrum analysis. Since there is no concern if the filters overlap, the resolution requirement is easily fulfilled by using even the simple Hanning window $h(n) = [1 + \cos(\pi n/N)]/2$, $n \in [-N/2, N/2]$, with $L_h = N = 256$. The Hanning window also satisfies eq. (29) and, therefore, the resulting filter bank is uniform. Without the STFT modification, the analysis-synthesis system can now be made a unity system—no reconstruction error—since eq. (31) can be exactly satisfied. In particular, if $f(n)$ is a rectangular window of length N, eq. (31) becomes

$$\sum_{s=-L_R}^{L_R-1} h(sR+n) = 1 \qquad n = 0, 1, \cdots, R-1, \tag{39}$$

where $L_R = N/R$, with R being assumed again to divide N. The condition in eq. (39) is satisfied within a constant gain factor by the Hanning window for any $R \leq N/2$ which divides N. The gain factor is $L_R/2$. The analysis-synthesis implementation then becomes the well-known overlap-add (OLA) implementation.[14,15,29]

Let us now consider the effect of STFT modification for frequency expansion on the analysis-synthesis system design. Since in this case $q = 2$, so that $R' = 2R$, limiting R' to $N/2$ requires limiting R to $N/4$. However, since the rectangular synthesis window has high sidelobes in its frequency response, it did not provide sufficiently good speech quality when frequency expansion was performed. To improve the filtering provided by $f(n)$, we have considered again using the window functions in eqs. (34) and (35). Very good quality expanded speech was obtained with $R = N/8 = 32$ and $w_M(n)$ of eq. (35), using $L_f = N = 256$ and $L = N/2 = 128$. For this choice of parameters, $w_M(n)$ becomes identical to the Hanning window $w_H(n)$. Note that the output data shift or interpolation factor is $R' = 2R = N/4$, but $L = N/2$, i.e., $L > R'$, which is a condition that provides for additional filtering as discussed in Section 5.1.

To summarize, the expansion system is implemented with a transform size of $N = 256$ and analysis and synthesis windows that are both Hanning windows of length N. The input and output data shifts are given by $R = N/8 = 32$ and $R' = 2R = 64$, respectively, and the STFT modification is done according to eq. (27) at time instants $t = nRT$, i.e., every $R = 32$ samples.

5.3 Simulations

The FDHS system was simulated on a laboratory computer which is equipped with an integral array processor (Data General-Eclipse AP/130). The array processor facilitated fast computation of the needed array operations, such as transforms, windowing, stack-adding, and overlap-adding. The input signal was generally telephone bandwidth speech (200-3200 Hz) sampled at 8 kHz. In addition to natural speech input, a synthetic vowel with fixed pitch was used. It was synthesized by periodically repeating a single pitch period (51 samples) from the speech of a male speaker. This synthetic signal was valuable in the development of the system, in checking assumptions, and in selecting parameters and window functions. By way of illustration, and to support our selection of parameters, Fig. 13 shows in part (a) the spectrum of the synthetic vowel, and in parts (b) to (l) the spectra obtained for different windows and parameter values, as detailed in the figure caption and explained below. For reference, an ideally

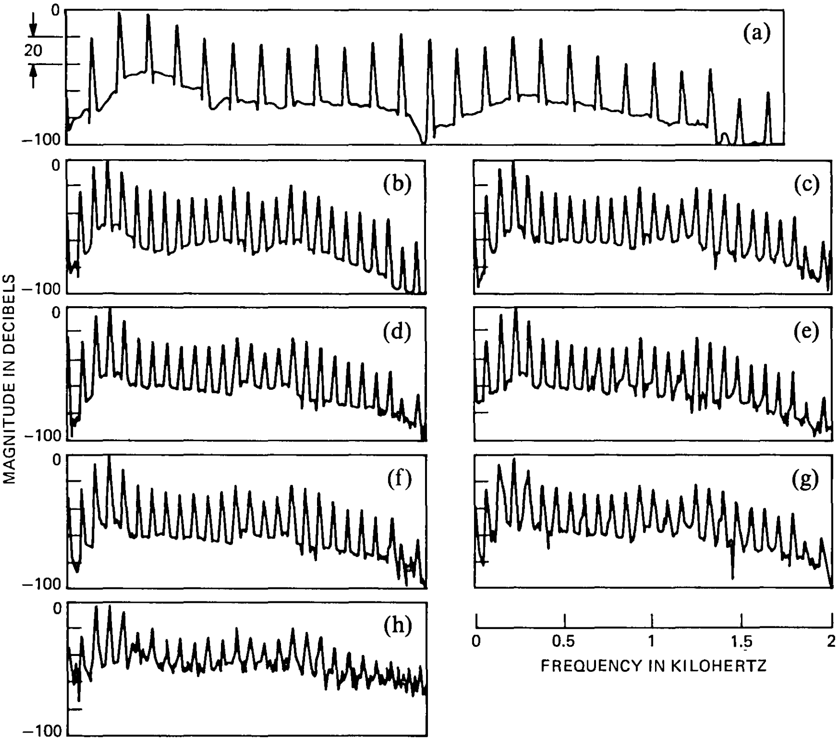


Fig. 13—Spectral representation of original and processed synthetic vowel for different system parameters and window functions. (a) Original. (b) Ideal 2:1 compression. (c) Compression with the selected system parameters: $N = 128$, $h(n) = h_0(n)$, $R = 32$, $R_M = 128$, $R' = 64$, $f(n) = w_M(n)$ with $L_f = 256$, $L = 64$. (d) Same as in (c), except $R_M = R = 32$. (e) Same as in (c), except $R_M = R = 64$. (f) Same as in (c) but $f(n) = w_{HS}(n)$ with $L_f = 512$, $L = 64$. (g) Same as in (c) but $f(n) = w_H(n)$ with $L = 64$ ($L_f = 128$). (h) Same as in (c) but $f(n) = w_R(n)$ with $L = 64$ ($L_f = 128$).

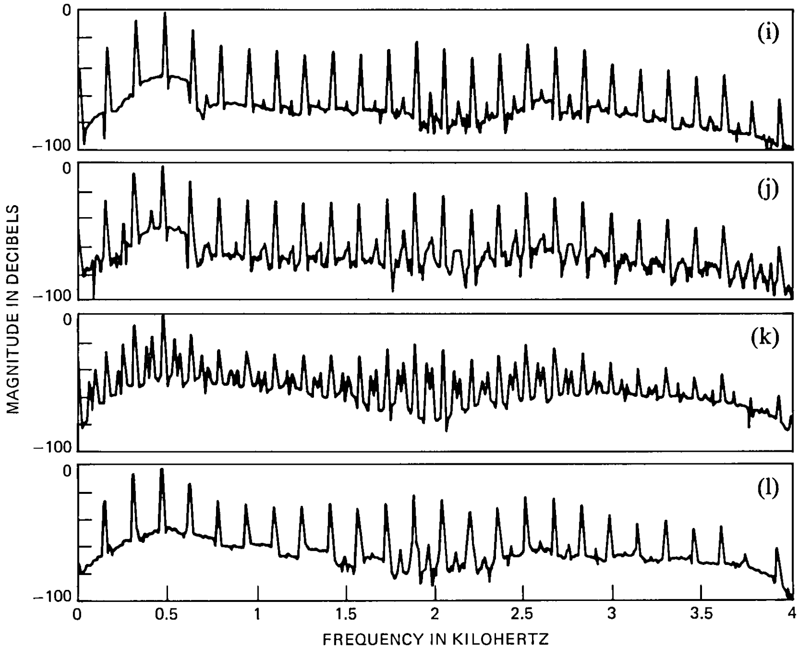


Fig. 13—(Contd.) (i) Expansion of the ideally compressed signal in (b) using the selected expansion system parameters: $h(n) = w_H(n)$ with $L = 128$, ($L_f = N = 256$), $R = 32$, $R' = 64$, $f(n) = w_H(n)$ with $L = 128$, ($L_f = 256$). (j) Same as in (i) but with $f(n) = w_R(n)$ ($L = 128$) and $R = 64$. (k) Same as (i) except $R = 64$. (l) Expansion of the compressed signal in (c) using the selected system parameters as in (i).

compressed spectrum is shown in part (b). Since the synthetic signal is periodic with a known period, ideal 2:1 compression can be obtained by discarding every other period and reducing the sampling rate by a factor of 2. Practically, (b) is obtained by windowing (we used a Hamming window for the spectral analysis) a segment of speech with half the number of samples than in the segment analyzed for producing the spectra in (a). Since the input signal is exactly periodic, the shape of the spectral teeth is determined only by the frequency response of the window. In natural speech, which is only quasiperiodic, the amplitude and phase modulations of the pitch harmonics, as discussed in Section II, widen further the spectral teeth. This can be observed in later illustrations.

In (c) of Fig. 13 the spectrum of the compressed signal, using the selected system parameters detailed in Section 5.1, is shown. Because the signal is purely periodic, the results are particularly good. A frequency domain signal-to-distortion ratio (SDR) computation* results

* The frequency-domain distortion measure used is based on the mse between the spectral envelopes of the input and processed signals, as measured in sub-bands with a

here in SDR = 42 dB. To illustrate that no harm is done by using $R_M > R$ (see Section 5.1), part (d) shows, for comparison with (c), the result of using $R_M = R = 32$. The difference is minute—only 0.1 dB higher SDR between the processed and ideal spectrum. Part (e) shows the spectrum obtained when $R = R_M = 64$ is used. The problem in this case is that for some harmonics the sign tracking is inadequate because R is too large. The effect on the spectra is clearly seen. Part (f) shows the result obtained when a longer, $L_f = 512$, synthesis window is used. On the basis of Fig. 12, $w_{HS}(n)$ is now preferred and used. The results are somewhat better than in (c), an improvement of 2 dB in SDR, but this does not seem to justify the doubling of the output buffer size and the additional computation. Parts (g) and (h) reaffirm the unsuitability of the simple Hanning and rectangular windows, respectively, as synthesis windows in the given compression system. The loss in SDR amounts to 18 and 28 dB, respectively, as compared to (c).

The remaining parts of Fig. 13 illustrate some of the results obtained with the expansion system. Part (i) shows the spectrum that results from expanding the ideally compressed signal in (b), using the selected system parameters as detailed in Section 5.2. As expected, the results are better than for compression and the frequency-domain SDR is above 60 dB. Parts (j) and (k) show the results obtained with rectangular and Hanning synthesis windows, respectively, using $R = 64$. It is clearly seen that this value of R is too large and its use results in significant SDR reductions for both windows—up to 35-dB reduction for the Hanning window, as compared to (i). For $R = 32$, the selected Hanning window performs best, as seen in (i), offering an 8-dB advantage in SDR over the rectangular window.

Finally, part (l) shows the resulting spectrum when the compressed signal in (c) is expanded back, using the selected system parameters as in (i). The reduction in SDR because of the expansion process was found to be only about 0.5 dB; that is, a 41.5-dB SDR is obtained when the spectrum in (l) is compared to the spectrum in (a).

The importance of correct sign initialization is demonstrated in Fig. 14. In addition to showing the ideally compressed spectrum of the synthetic vowel in (a)—identical to part (b) of Fig. 13—it also shows the spectrum of the frequency compressed signal obtained with three different sign initialization approaches. Part (b) of Fig. 14 shows the result obtained when all the signs are initialized to be positive; part (c), when the signs are initialized according to eq. (19) (i.e., using only

bandwidth that increases with frequency, similar to articulation index measurements. The actual band allocation used was taken from Ref. 31. This measure was found useful for the synthetic signal, as it correlated well with our observations. For natural speech, however, this was not so and we had to rely on subjective listening and visual examination of the spectral representations.

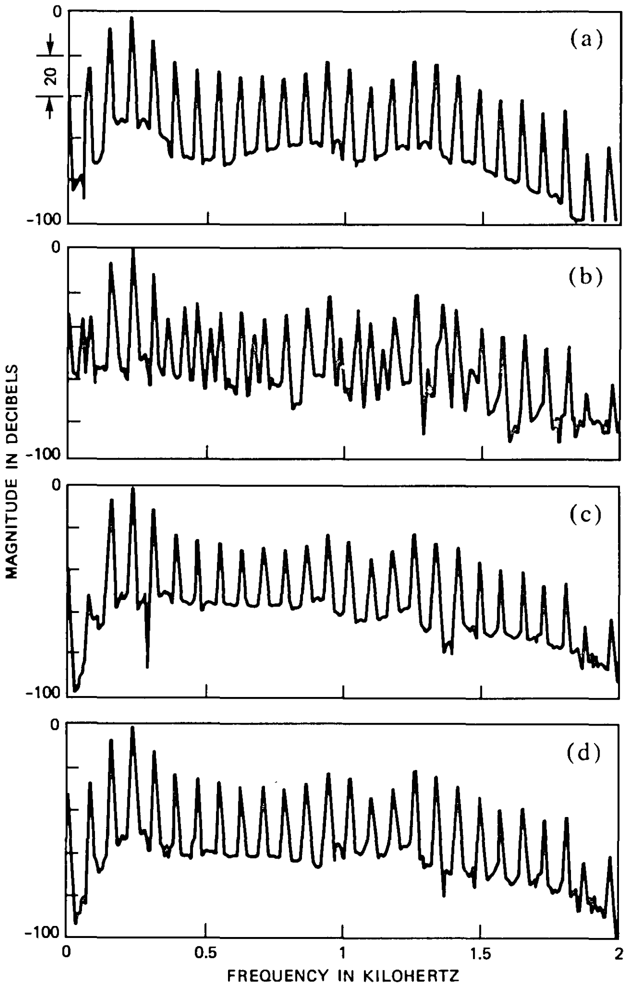


Fig. 14—Demonstration of the effect of different sign initialization approaches. (a) Spectrum of the ideally compressed synthetic vowel. (b) All-positive sign initialization. (c) Sign initialization according to principal value of the phase in each band. (d) Sign initialization according to the algorithm developed in Section 3.1.

the principal value of the phase); and part (d), according to the final initialization algorithm which includes the pairing and matching operations described earlier and also shown in (c) of Fig. 13. The results indeed validate the assumptions and analysis performed in Section III—namely, the possible generation of undesired spectral components, in addition to the attenuation of the desired components. Part (b) shows the result when a random or all-positive sign initialization is used; part (c), the possible cancellation of harmonics because of incorrect matching of the signs in two adjacent bands which share the same harmonic; and part (d), the improved results obtained by using the developed sign initialization algorithm.

Similar results were obtained with natural speech, although errors in the initialization do occur due to speech nonstationarity and deviations from the harmonic model. Yet, in all the simulations performed, the developed sign initialization algorithm always resulted in better speech quality. As an illustration, Fig. 15 shows the spectra obtained for a segment of voiced speech. Parts (a) to (c) show the spectra of the original, compressed, and reconstructed signal using the developed FDHS system. For comparison, parts (d) and (e) show the corresponding

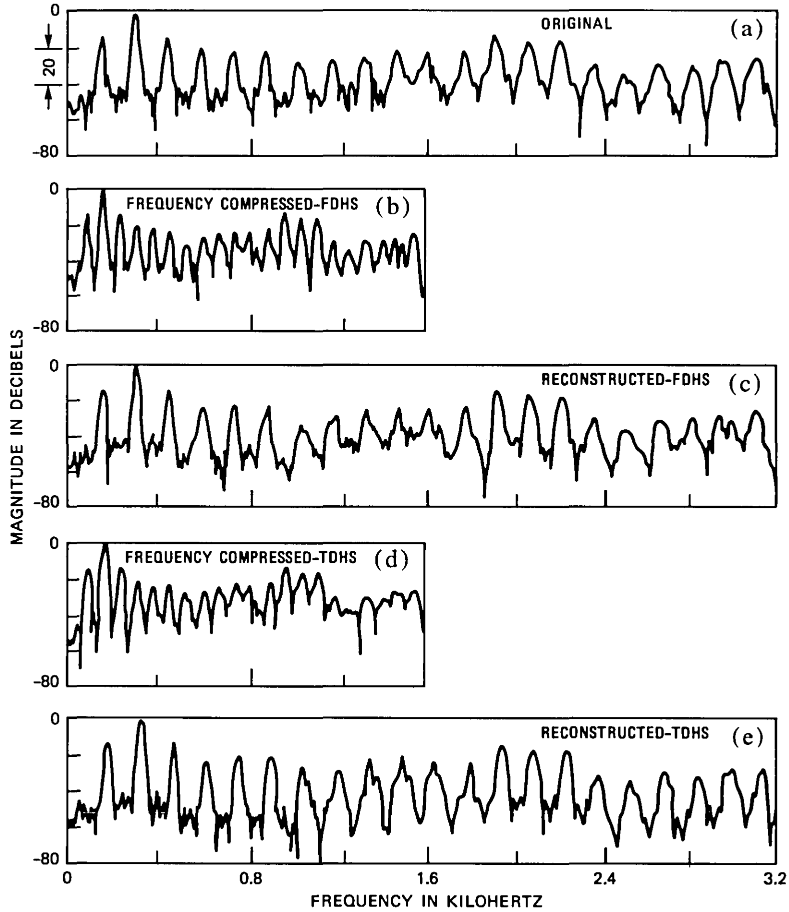


Fig. 15—Spectral representation of original and processed voiced speech segment for male speaker. (a) Original. (b) Compressed 2:1 using FDHS. (c) Expansion of compressed signal (reconstructed) using FDHS. (d), (e) Same as (b) and (c), respectively, but using TDHS.

results obtained with the time-domain harmonic scaling (TDHS) technique[4] using a cepstral pitch detector[30] implemented on the same array processor. The TDHS system was judged to give higher quality speech. Further discussion on the comparison between the two systems is given later. For an additional demonstration of the results obtained by the two systems, Fig. 16 shows the corresponding time waveforms, and Fig. 17 presents spectrograms of the complete processed sentence. To

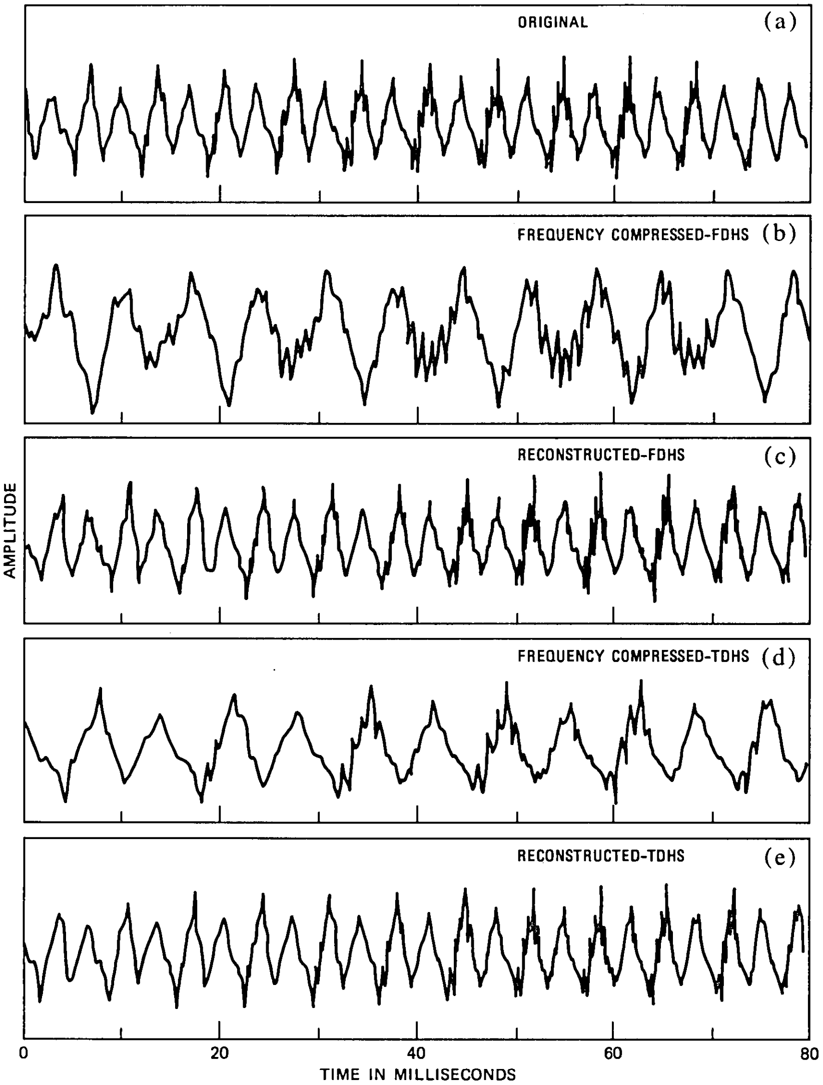


Fig. 16—Time-domain waveforms corresponding to the spectral representations given in Fig. 15.

illustrate the capability of the two systems to change the time scale, without changing the frequency scale, parts (f) and (g) of Fig. 17 show spectrograms of the time-compressed signals.

5.3.1 Performance with degraded inputs

To examine the robustness of the FDHS technique to adverse acoustical environment conditions, we ran simulations with noisy speech signals (down to 0-dB s/n), speech with severe room reverberation, and speech from three speakers speaking simultaneously. In simulations with noisy speech, the FDHS system appeared to be quite robust

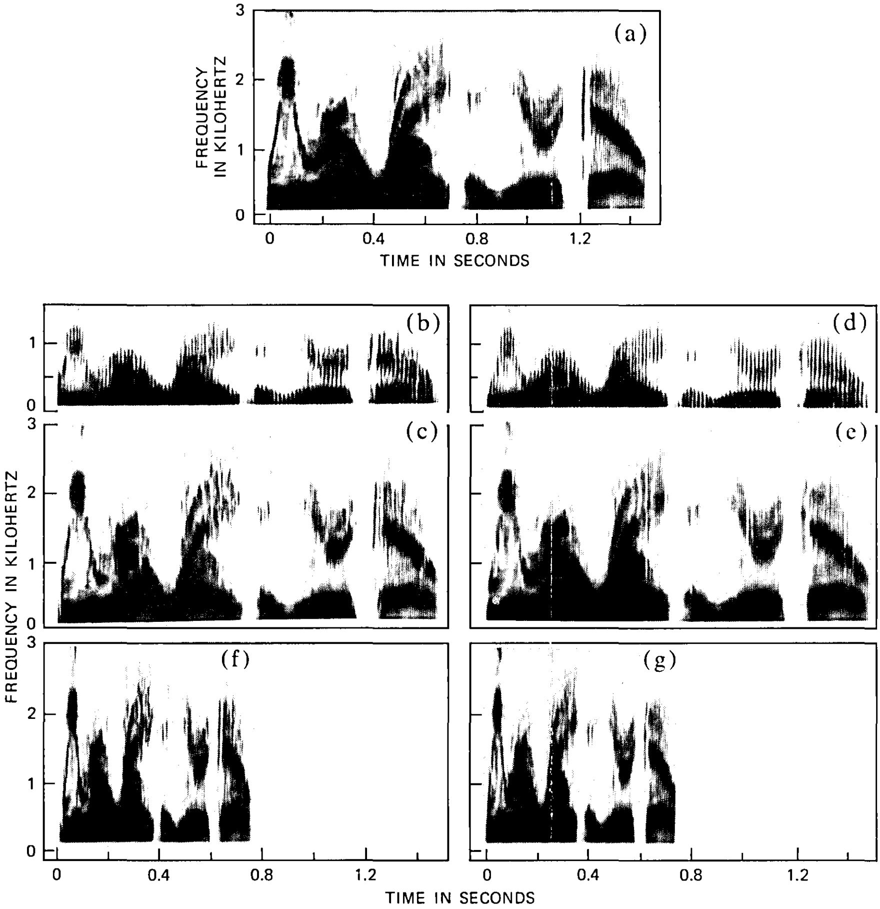


Fig. 17—Spectrograms of original and processed speech by FDHS and TDHS systems ("We were away a year ago," male speaker.) (a) Original. (b) Frequency compressed—FDHS. (c) Frequency-expanded (reconstructed)—FDHS. (d) Frequency-compressed—TDHS. (e) Frequency-expanded (reconstructed)—TDHS. (f) Time-compressed—FDHS. (g) Time-compressed—TDHS.

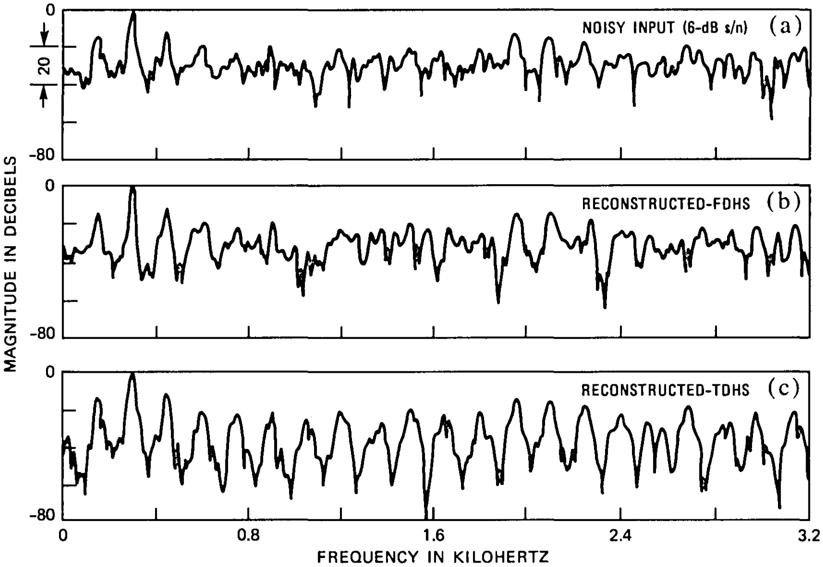


Fig. 18—Spectral representation of original and processed noisy voiced speech segment [white noise at 6-dB s/n added to signal in (a) of Fig. 16]. (a) Original. (b) Reconstructed—FDHS. (c) Reconstructed—TDHS.

with only some reduction in signal crispness as compared to processing clean speech. No change in the general nature of the noise was observed. This is in contrast to results obtained with the TDHS technique which tends to structure the noise caused by the pitch-dependent, time-domain weighting process. This is illustrated in Fig. 18 which shows the spectra of the original noisy signal (white noise at 6-dB s/n) and the reconstructed signals by the FDHS and TDHS systems. The structuring of the spectra, according to the speech pitch, caused by the TDHS technique is evident. Although this structuring provides a filtering effect similar to comb filtering, the structured noise is usually more annoying to listen to. For further illustration, Fig. 19 presents spectrograms of the complete processed noisy sentence by the two systems.

The results of processing speech with severe room reverberation, and multiple speakers speech, indicate that the FDHS system is highly robust to these conditions. In fact, these conditions tend to mask processing artifacts and the reconstructed signal sounds very similar to the original. In the TDHS system, the structuring of the uncorrelated reverberation components is perceivable. On the other hand, the TDHS system was found to be quite robust to multiple speakers speech.[5] Spectrograms of a processed sentence, recorded with severe room reverberation, are shown in Fig. 20, for the two systems.

To summarize, by the above simulation results we have demonstrated the validity of our assumptions in deriving the sign initialization algorithm and the selection of the analysis-synthesis window functions and parameters. The FDHS system was found to be robust to environment conditions, although its quality for clean speech signals is judged to be lower than with the TDHS. On the other hand, the robustness of the TDHS system is largely dependent on the type of pitch detector used (particularly with noisy signals), and it suffers from artifacts

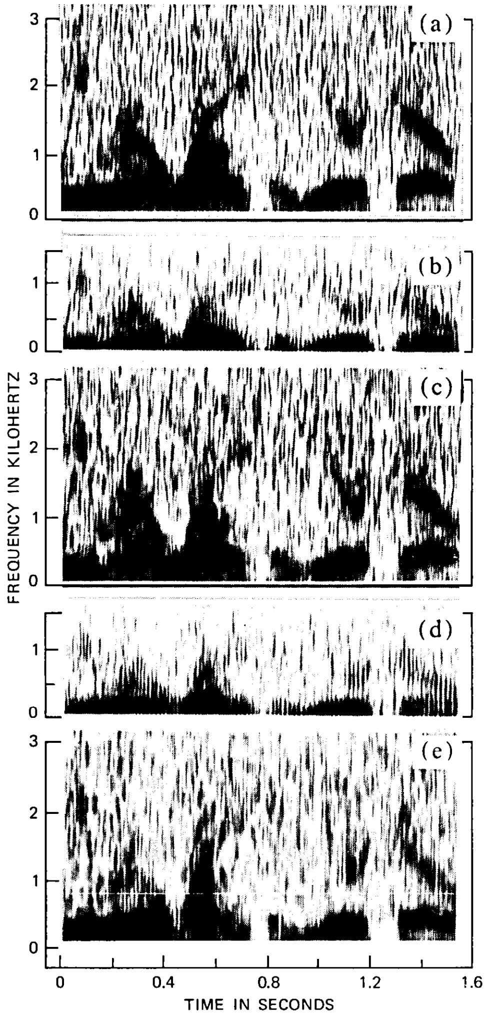


Fig. 19—Spectrograms of original and processed noisy speech by FDHS and TDHS systems. Parts (a) through (e) as in Fig. 17, but with white noise at 6-dB s/n added to original speech signal.

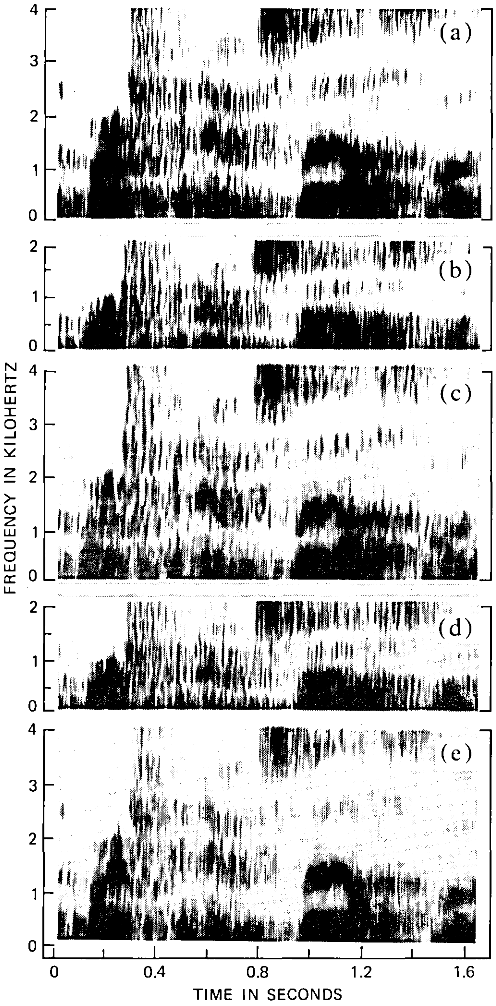


Fig. 20—Spectrogram of processed speech with severe room reverberation by FDHS and TDHS systems ["The birch canoe slid on the smooth (plank)," male speaker]. (a) Input signal. (b) Compressed—FDHS. (c) Reconstructed—FDHS. (d) Compressed—TDHS. (e) Reconstructed—TDHS.

introduced by structuring wide-band noise or uncorrelated reverberation.

It should be mentioned, however, that while the FDHS system, in general, provided good communications speech quality, an increase in reconstructed signal degradation was found for some low-pitch male speakers, typically below 80 Hz. One possible reason is insufficient frequency resolution of the analysis filter bank. At the expense of a slight input bandwidth reduction, this condition can be corrected somewhat by reducing the sampling rate from 8 kHz to 6.4 kHz since increasing N from 128 to 256 is not desirable as explained earlier. Another source of degradation appears to be the fact that if the pitch teeth are wide, and the pitch frequency is low, the original speech is characterized by a small separation of the pitch teeth, and the model assumed in the derivation of the FDHS technique is not as applicable. Further study of this problem is needed.

VI. A HYBRID TDHS-FDHS SYSTEM

The simulation results presented in Section V and our earlier experience with TDHS[4-6] indicate the advantage of TDHS over FDHS, provided that the acoustical environment conditions allow adequate pitch extraction, the noise structuring is acceptable, and pitch data can be transmitted. If pitch data cannot be transmitted, as would be the case with analog channels or digital channels with tandeming of waveform coders, or if it is not desirable to transmit the data, use of TDHS requires reextraction of the pitch at the receiver. Since pitch extraction from the compressed signal is more difficult because fewer pitch periods per unit time are available, the quality of the reconstructed signal is degraded. Since expansion alone with the FDHS system provides almost transparent speech quality without the need for explicit pitch extraction, we examined the possibility of using a hybrid system, such as shown in Fig. 21. In this system the compression is done by TDHS and the expansion by FDHS. Simulations using this system supported this approach, and the overall speech quality obtained was judged to be better than by TDHS without pitch transmission or by FDHS alone. For illustration, Fig. 22 shows spectrograms of a processed sentence by the different systems. It should also be noted that the proposed hybrid system has an advantage with noisy signals as well, as long as pitch extraction at the transmitter is feasible. The advantage is that the structuring of the noise in the reconstructed signal is avoided, since this structuring occurs mainly in the expansion stage of the TDHS, which is replaced by FDHS in the hybrid system. Furthermore, the hybrid system should also be more tolerant to channel errors than TDHS alone since these errors appear as noise in the compressed signal which does not affect the FDHS expansion system much, but could obviously affect the TDHS expansion process.

VII. CONCLUSION

A unified description of several frequency scaling techniques has been given in terms of the short-time spectral modifications which they produce. This description helps in understanding the properties and limitations of these techniques and their relation to FDHS.

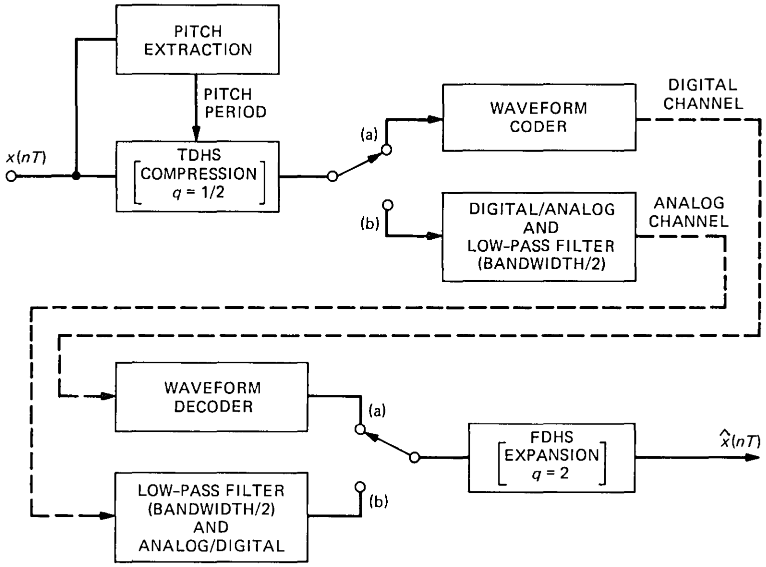


Fig. 21—Block diagram of hybrid frequency scaling system TDHS-FDHS. (a) Digital channel. (b) Analog channel.

The results of this work, with respect to the FDHS technique,* provide a substantial improvement of the earlier version of this technique as reported in Ref. 12. The improvement is manifested in both the quality achieved, and in implementation efficiency. The improvement in quality is the result of using a filter bank with higher frequency resolution and less overlap between filters, as well as the use of the dynamic sign initialization and matching algorithm developed in the present work. The improvement in implementation efficiency is achieved by the use of a block implementation of the short-time Fourier analysis-synthesis system. In this system, the FFT, the embedded decimation and interpolation, and the WOLA synthesis scheme—described in Ref. 14 and extended here to include the case of analysis and synthesis window having longer duration than the transform size—provide a large saving in computation.

It was seen that the introduction of STFT modifications can greatly affect the characteristics of the analysis-synthesis system and its design. The detailed design considerations of the window functions and the selection of the decimation and interpolation factors given here can be useful also in other applications involving spectral modi-

* An early presentation of the results was given in a talk that included an audio tape demonstration.[32]

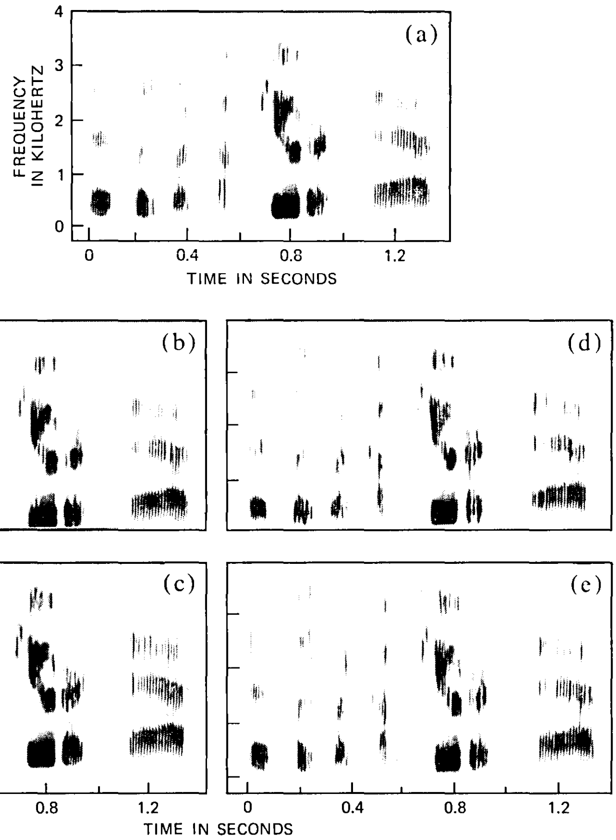


Fig. 22—Spectrograms of original and processed speech by FDHS, TDHS, and hybrid TDHS-FDHS systems ["This is a computer test (of a digital speech coder)," male speaker.] (a) Original. (b) Reconstructed—FDHS. (c) Reconstructed—TDHS. (d) Reconstructed—TDHS with pitch reextraction from compressed signal. (e) Reconstructed—hybrid TDHS-FDHS system.

fications of signals, such as in some of the techniques described in Section II.

The developed FDHS system is particularly amenable to an array-processor implementation. From simulations of the system for a variety of adverse acoustical environment conditions, the FDHS technique appears to be robust and provides reconstructed speech of good communications quality. The main degradation, as mentioned earlier, is introduced in the compression stage, since the expansion operation provides almost transparent quality. Unlike the simpler TDHS technique, the FDHS is not explicitly dependent on pitch extraction and is, hence, more robust. However, for clean speech, compression with TDHS results in better speech quality than with FDHS. On the other hand, for noisy speech signals, in addition to possible failure of the pitch detector at high-noise levels, the TDHS expansion process tends to structure the noise which can be perceptually annoying.

In applications where pitch extraction at the transmitter is feasible but where pitch data transmission is to be avoided, the hybrid TDHS-FDHS system, in which compression is performed by TDHS and expansion by FDHS, provides better overall speech quality than the TDHS or FDHS systems alone. The additional advantages of the hybrid system, such as reduction of noise structuring and higher immunity to channel errors, as compared to TDHS alone, and the lower complexity, as well as higher quality, as compared to FDHS alone, singles out the hybrid system as the best solution for a variety of applications.

An interesting outcome of the general implementation scheme, shown in Fig. 8, is the generalization of the TDHS technique to include both analysis and synthesis windows. This generalization has potential for further improving the TDHS performance.

The FDHS technique was developed on the basis of the quasiharmonic nature of voiced speech. Deviations from this model cause a reduction in processed speech quality. To achieve higher than communications quality with the FDHS technique, further study and ununderstanding are needed of the simultaneous amplitude and phase modulation processes of the individual pitch harmonics, and of the nonstationary characteristics of speech signals.

REFERENCES

1. J. L. Flanagan, *Speech Analysis, Synthesis and Perception.* New York: Springer Verlag, 1972.
2. J. L. Flanagan and R. M. Golden, "Phase Vocoder," B.S.T.J., *45* No. 9 (November 1966), pp. 1493-509.
3. M. R. Schroeder, J. L. Flanagan, and E. A. Lundry, "Bandwidth Compression of Speech by Analytic-Signal Rooting," Proc. IEEE, *55* (March 1967), pp. 396-401.
4. D. Malah, "Time Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27,* No. 2 (April 1979), pp. 121-33.
5. D. Malah, R. E. Crochiere, and R. V. Cox, "Performance of Transform and Subband Coding Systems Combined with Harmonic Scaling of Speech," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-29* (April 1981), pp. 273-83.
6. D. Malah, "Combined Time Domain Harmonic Compression and CVSD for 7.2 kb/s Transmission of Speech Signals," Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing (April 1980), pp. 504-7.
7. J. L. Melsa and A. K. Pande, "Medium-Band Speech Encoding Using Time Domain Harmonic Scaling and Adaptive Residual Coding," Proc. 1981 IEEE Int. Conf. Acoust., Speech, Signal Processing (April 1981), pp. 603-6.
8. J. E. Youngberg, "Rate/Pitch Modification Using the Constant-Q Transform," Proc. 1979 IEEE Int. Conf. Acoust., Speech, Signal Processing (1979), pp. 748-51.
9. H. Ravindra, "Speech Articulation Rate Change Using Recursive Bandwidth Scaling," Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing (April 1980), pp. 352-5.
10. B. P. Bogert, "The Vobanc – A Two-to-One Speech Bandwidth Reduction System," J. Acoust. Soc. Am. *28,* No. 3 (May 1956), pp. 399-404.
11. J. L. Daguet, "Speech Compression CODIMEX System," IEEE Trans. Audio, *AU-11,* No. 2 (March-April 1963), pp. 63-71.
12. J. L. Flanagan and S. W. Christensen, "Technique for Frequency Division/Multi-

plication of Speech Signals," J. Acoust. Soc. Am. *60,* No. 4 (October 1980), pp. 1061–8.

13. J. M. Tribolet, "A New Phase Unwrapping-Algorithm," IEEE Trans Acoust., Speech, Signal Processing, *ASSP-25,* No. 2 (April 1977), pp. 170–7.
14. R. E. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/ Synthesis," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-28,* No. 1 (February 1980), pp. 99–102.
15. J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proc. IEEE, *65* (November 1977), pp. 1558–64.
16. M. Portnoff, "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-28,* No. 1 (February 1980), pp. 55–69.
17. R. E. Kahn and J. B. Thomas, "Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation," IEEE Trans. Inf. Theory, *IT-11,* No. 4 (October 1965), pp. 516–20.
18. J. L. Flanagan, "Parametric Coding of Speech Spectra," J. Acoust. Soc. Am., *68,* No. 2 (August 1980), pp. 412–9.
19. J. L. Flanagan and S. W. Christensen, "Computer Studies on Parametric Coding of Speech Spectra," J. Acoust. Soc. Am. *68,* No. 2 (August 1980), pp. 420–30.
20. R. E. Bogner and J. L. Flanagan, "Frequency Multiplication of Speech Signals," IEEE Trans. Audio Electroacoust., *AU-17* (September 1969), pp. 202–8.
21. R. E. Bogner, "Frequency Division in Speech Bandwidth Reduction," IEEE Trans. Comm. Tech., *COM-13,* No. 4 (December 1965), pp. 438–51.
22. R. W. Schafer and L. R. Rabiner, "Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis," IEEE Trans. Audio Electroacoust., *AU-21* (June 1973), pp. 165–74.
23. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Englewood Cliffs, New Jersey: Prentice Hall, 1978, Chapter 6.
24. M. R. Portnoff, Time Scale Modification of Speech Based on Short-Time Fourier Analysis, Ph.D Dissertation, Massachusetts Inst. Tech., Cambridge, April 1978.
25. S. Seneff, "High Quality System for Speech Transformations," (Abstract) 98th Meeting, Acoust. Soc. Am., J. Acoust. Soc. Am., Suppl. 1, *66* (Fall 1979), p. S22.
26. M. R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-24* (June 1976), pp. 243–8.
27. L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing,* Englewood Cliffs, New Jersey: Prentice Hall, 1975, Chapter 3.
28. J. H. McClellan, J. W. Parks, and L. R. Rabiner, "FIR Linear Phase Filter Design Program," in *Programs for Digital Signal Processing,* New York: IEEE Press, 1979, Chapter 5.
29. J. B. Allen, "Short-Term Spectral Analysis Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-25* (June 1977), pp. 235–8.
30. A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am., *41,* No. 2 (February 1967), pp. 293–309.
31. D. H. Klatt, "A Digital Filter Bank for Spectral Matching," Proc. IEEE Int. Conf. Acoust., Speech Signal Processing (April 1976), pp. 573–6.
32. D. Malah and J. L. Flanagan, "Efficient Implementation of a Frequency Domain Technique for Frequency Scaling of Speech Signals," (Abstract) 100th Meeting, Acoust. Soc. Am., November 1980, J. Acoust. Soc. Am., Suppl. 1, *68* (Fall 1980), p. S87.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

McDonald's Problem—An Example of Using Dijkstra's Programming Method

By D. K. SHARMA

(Manuscript received April 20, 1981)

We use Dijkstra's programming method to solve the so-called McDonald's problem and show how to rigorously introduce file input/output operations in the program. The steps involved are quite simple and the paradigm suggested is applicable to the wider class of problems that involve sequentially processing file records. For these problems, the programs developed using the data structure design methodology are generally considered to be the most desirable. We show that Dijkstra's method can yield the same program. Unlike other methodologies, it also yields a correctness proof, which is extremely valuable in understanding the program and in modifying it.

I. INTRODUCTION

In this paper, we use Dijkstra's method of simultaneously developing a program and a proof of its correctness to solve the so-called McDonald's warehouse problem. The problem briefly is to read a card file and print an inventory report. Our interest in it stems from the fact that it requires sequentially processing the records of a file—a task common to a large class of problems. As we solve the problem, we illustrate how to rigorously introduce file input/output operations within the framework of Dijkstra's method. This aspect of the solution is intended to be a paradigm that is applicable to the above-mentioned class of problems.

Our solution serves one other purpose. The McDonald's warehouse problem is often used to compare the effectiveness of different programming methodologies in developing programs that sequentially process files. As discussed in Ref. 1, the programs developed using the data structure design methodology are generally considered to be the most desirable. This is primarily because the structure of the resulting program closely reflects the structure of the input data. The program developed in this paper is identical to that developed by the data structure design methodology, except that the latter has no correctness proof associated with it. The proof-related assertions in a program are not only helpful in understanding it, but also in systematically modifying it.

The solution discussed below is quite simple and regular—as a paradigm it can be systematically applied to other similar problems. It is obtained in three steps. In Step 1, we develop the program and its proof assuming that the records of the file are available in an array. This version of the program also uses a few symbols that are related to its proof, assuming that their values are readily available. These assumptions are made purely for the sake of convenience in developing the program and are removed in the next two steps. In Step 2, we (*i*) introduce additional program variables, (*ii*) modify the assertions to reflect the introduction of the new program variables, and (*iii*) add appropriate statements that make the assertions hold. Thus, we are guaranteed that the program remains correct through all the modifications. This is done to eliminate from the program text the symbols whose values are not readily available. In Step 3, we introduce the file operations. This is done by replacing suitably chosen initialization statements by *openfile* and by replacing certain other groups of assignment statements, by *readfile* or *writefile.*

Section II presents the problem, first informally and then formally. This is followed by the three steps of the solution in Section III and by a summary in Section IV.

II. PROBLEM SPECIFICATION

McDonald's food warehouse receives and distributes food items. Each shipment received or distributed is recorded on a punched card that contains the name of the item and the change in the quantity of the item due to that shipment. The change is recorded as a positive integer when items are received and negative otherwise. These cards are alphabetically sorted according to item names by another program.

The problem is to write a program to read the sorted card file and print an inventory report. The report should show the net change in the inventory of each item transacted and the number of distinct food items transacted. At this point, the reader may wish to devise his or her own solution and later compare it with the solution derived below.

The following is a more formal statement of the problem, which is used in developing the solution in Section III.

Assume that the transaction file contains $M - 1$ cards, and the ith card contains two fields, $f(i)$ and $q(i)$, where $1 \leq i \leq M - 1$, $f(i)$ is a positive integer representing the name of the food item on the ith card, and $q(i)$ is an integer representing the change in the quantity of $f(i)$. Note that, without any loss of generality, we have assumed $f(i)$'s to be integers instead of identifiers.

The file has been sorted in the nondescending order of $f(i)$'s. For the sake of discussion, we augment the file with an extra card signifying "end of file" (EOF) condition for which $f(M)$ = EOF and $q(M) = 0$. The value EOF is assumed to be greater than all the other $f(i)$'s to maintain the sorted nature of the file.

Clearly, all the cards of a particular food item are grouped together in the file. Let N be the number of distinct food items, that is, the number of groups in the augmented file. Let m_n be the index of the first card of the nth group, where $1 \leq n \leq N$. That is,

$$m_1 = 1$$

and $m_n = \min\{i : m_{n-1} < i \leq M \quad \text{and } f(i-1) < f(i)\} \quad$ for $1 < n \leq N$.

The m_i's have the following property

$$1 = m_1 < m_2 \cdots < m_{N-1} < m_N = M.$$

Define p_n to be the net change in the quantity of the nth food item, where $1 \leq n < N$. We do not define p_N, which corresponds to the fictitious card used to augment the file. We can express p_n as

$$p_n = \sum_{\substack{i \\ m_n \leq i < m_{n+1}}} q(i) \quad \text{for } 1 \leq n < N.$$

Note that p_n's are the values to be printed in the report.

We can now define the goal of the program as: Print $N - 1$ lines such that the nth line contains the name of the nth food item [(i.e., $f(m_n)$] and the net change in the quantity of the food item, i.e., p_n. Then print a line containing $N - 1$, the number of food items transacted. Note that N is the number of groups in the augmented file.

III. SOLUTION

In the following, we first develop a program to print the first $N - 1$ lines of the report and add the last line later.

We assume that the reader has at least a cursory knowledge of the methodology described in Refs. 1 and 2. See Ref. 3 for a brief tutorial.

We develop the iterative solution in three steps. In Section 3.1, we assume that N is known, and the following are available as arrays:

$$f(i) \quad \text{and} \quad q(i) \quad \text{for} \quad 1 \leq i \leq M, \text{ and}$$

$$m_i \quad \text{for} \quad 1 \leq i \leq N.$$

In Section 3.2, we remove the above assumptions one by one, as described in Section I. This is done by introducing new program variables, etc., while still maintaining program correctness. This culminates in a program in which N need not be known, and only one element each from the arrays $f(i)$ and $q(i)$ appears in the loop.

In Section 3.3, we identify statements that can be replaced by file operations. This substitution is quite mechanical and yields a program that sequentially reads the card file and prints the first $N - 1$ lines of the report. The post-assertion of this program is then used to print the last line of the report.

3.1 Step 1

In the notation of Ref. 1, the result assertion is given by

$$\text{R1: } (\mathbf{A}\, i : 1 \le i < N : p_i \text{ has been printed}).$$

This should be read as follows: for all i such that $1 \le i < N, p_i$ has been printed.

The iterative statement will be the main part of the program. Its loop invariant $P1$ is obtained by weakening the result assertion $R1$, that is, replacing the constant N by a variable n. Thus, we have

$$P1\text{: } (\mathbf{A}\, i : 1 \le i < n \le N : p_i \text{ has been printed})$$

$$\text{and} \quad (P1 \wedge n = N) \equiv R1.$$

The first version of the program is as below.

Solution 1.1

```
$n := 1$; $\{P1\}$
do $n \ne N \rightarrow$
   Increase $n$ under the invariance of $P1$.
od $\{P1 \wedge n = N\}$.
```

This program begins with an initialization step that trivially establishes $P1$. The loop increases n and keeps $P1$ invariant; therefore, at the end we can assert $P1 \wedge n = N$, which is equivalent to $R1$.

To show termination, choose the termination function

$$t = N - n.$$

The value of t is initially $N - 1$, each iteration of the loop reduces it, and $t \ge 0$. The loop, therefore, must terminate.

We now add a statement to increment n:

```
$n := 1$; $\{P1\}$
do $n \ne N \rightarrow$
   $S1$; $\{Q\}$
   $n := n + 1$ $\{P1\}$
od $\{P1 \wedge n = N\}$.
```

Here, Q is the "weakest precondition such that the execution of '$n := n + 1$' will establish $P1$." It is obtained by replacing all occurrances of n in $P1$ by $n + 1$, and the resulting expression is denoted by $P1|_n^{n+1}$. Thus,

$$Q = wp(\text{“}n := n + 1\text{”}, P1) = P1|_n^{n+1},$$

where $wp(S,P)$ is the weakest precondition in which execution of S will establish P.

The statement $S1$, starting execution in state $P1$, must establish $P1|_n^{n+1}$. This can be simply done by computing the value of p_n and printing it. Thus, $S1$ can be refined as:

```
S1: {$P1$}
    $S2$;        {$sum = p_n \wedge P1$}
    print (sum) {$P1|_n^{n+1}$},
```

where the program variable *sum* has been introduced.

We now refine $S2$. It has the property

$$\{P1\}\ S2\ \{R2 \text{ and } P1\},$$

where

$$R2\text{: } sum = p_n = \sum_{m_n \le i < m_{n+1}} q(i).$$

Statement $S2$ will be an iterative program, and to get its loop invariant $P2$, replace the constant m_{n+1} by a variable m. Thus,

$$P2\text{: } sum = \sum_{m_n \le i < m \le m_{n+1}} q(i)$$

$$\text{and} \quad (P2 \wedge m = m_{n+1}) \equiv R2.$$

Using the same technique as before, $S2$ is refined as

```
S2: $m := m_n$; $sum := 0$;{$P2$}
    do $m \ne m_{n+1} \rightarrow$
      S3; {$P2|_m^{m+1}$}
      $m := m + 1${$P2$}
    od {$P2 \wedge m = m_{n+1}$}.
```

Notice that the initializations establish $P2$ in the beginning and the post-assertion is equivalent to $R2$. Statement $S3$ must have the property

$$\{P2\}\ S3\ \{P2|_m^{m+1}\}$$

and can be easily shown to be $sum := sum + q(m)$.

This gives us Solution 1.2, after assembling all the pieces.

Solution 1.2

```
$n :=$ 1;$\{P1\}$
do $n \neq N \rightarrow$
  $m :=$ $m_n$; $sum :=$ 0; $\{P2 \wedge \mathbf{m = m_n}\}$
  do $m \neq m_{n+1} \rightarrow$
    $sum :=$ $sum + q(m)$;
    $m :=$ $m + 1\{P2\}$
  od; $\{P2 \wedge \mathbf{m = m_{n+1}}\}$
  print (sum);
  $n :=$ $n + 1\{P1 \wedge \mathbf{m = m_n}\}$
od $\{P1 \wedge \mathbf{m = m_n} \wedge n = N\}$.
```

We have added "$m = m_n$" to the assertions where it happens to hold. This is indicated in bold and is used in the next section.

3.2 Step 2

Solution 1.2 is unsatisfactory since it explicitly uses N, m_n, and m_{n+1}, which are not available *a priori.* We modify it in the following to eliminate this deficiency.

These modifications are done by (*i*) introducing additional program variables, (*ii*) slightly modifying the assertions to reflect the introduction of new variables, and (*iii*) adding appropriate statements to make the assertions hold. Thus, the modified program is guaranteed to be correct. We believe that this technique is applicable not just to this problem but also to the wider class of problems wherein files are processed sequentially or where the initial versions of the programs refer to quantities not readily available.

We make the above-mentioned three changes as follows. In the first change, we eliminate the assignment statement $m := m_n$. Note that the rest of the outer loop maintains $m = m_n$ invariant; so we could add this term to the loop invariant $P1$ and eliminate the assignment statement. An extra initialization statement, $m := 1$, would then become necessary to establish the new loop-invariant in the beginning. The program is still correct. See Solution 2.1 below.

In the second change, we modify the outer loop guard $n \neq N$. Since

$$(n \neq N) \equiv [f(m_n) \neq f(m_N)]\,,$$

$$m_N = M, \text{ and}$$

$$m = m_n$$

hold before the outerloop, its guard can be replaced by $f(m) \neq f(M)$. This change does not affect any of the assertions.

In the third change, we modify the inner loop guard, $m \neq m_{n+1}$. For $m_n \leq m < m_{n+1}$,

$$(m \neq m_{n+1}) \equiv [f(m) = f(m_n)]\,.$$

Therefore, we can use $f(m) = f(m_n)$ as the guard in place of $m \neq m_{n+1}$, without disturbing anything else. We now replace $f(m_n)$ by a program variable F; the guard becomes $f(m) = F$. The meaning of the inner loop is kept unchanged by requiring $F = f(m_n)$ to be true before the inner loop. This requirement is trivially met by adding the assignment $F := f(m)$ at that point; notice that $m = m_n$ is true before the loop, as discussed above.

The above additions appear in bold print in the following program. *Solution 2.1*

```
$n := 1; \mathbf{m} := 1; \{P1 \wedge \mathbf{m} = \mathbf{m_n}\}$
$\mathbf{do\ f(m) \neq f(M) \rightarrow}$
   $sum := 0; \mathbf{F := f(m)}; \{P2 \wedge m = m_n \wedge \mathbf{F = f(m_n)}\}$
   $\mathbf{do\ f(m) = F \rightarrow}$
      $sum := sum + q(m); m := m + 1\{P2\}$
   $\mathbf{od};\ \{P2 \wedge m = m_{n+1}\}$
   $print\ (sum);$
   $n := n + 1\{P1 \wedge \mathbf{m = m_n}\}$
$\mathbf{od}\ \{P1 \wedge \mathbf{m = m_n} \wedge n = N\}.$
```

This program, still correct, does not explicitely use N or m_n; notice that $f(M)$ is the special value EOF. They are, however, an integral part of the proof and the assertions. The two assignment statements involving n could be removed from the program without affecting it. But we retain them, as the final value of n is of interest in printing the Nth line of the report.

3.3 Step 3

Solution 2.1 uses $f(m)$ and $q(m)$ as if they are available as arrays, but they are not. We eliminate their use in two steps and introduce file operations instead. This technique is useful not only in solving McDonald's problem, but in a wider class of problems that involve sequential file processing.

In Step 1, we replace $f(m)$ and $q(m)$ by the variables f and q, respectively. This necessitates asserting $f = f(m)$ and $q = q(m)$ before the statements where the substitution is made. Just as in the previous section, the assertions are made to hold by introducing appropriate assignment statements.

The two above assertions can become false only after a statement that modifies m. Therefore, we introduce the assignment statements $f := f(m)$ and $q := q(m)$ after each statement that modifies m. Solution 2.1 has only two such instances. The program after these modifications is as follows.

Solution 3.1

```
$n := 1; m := 0;$
$m := m + 1; f := f(m); q := q(m);$ $\{P2 \wedge m = m_n \wedge A\}$
do $f \neq f(M)\rightarrow$
   $sum := 0; F := f;$ $\{P2 \wedge m = m_n \wedge F = f(m_n) \wedge A\}$
   do $f = F\rightarrow$
      $sum := sum + q;$
      $m := m + 1;$
      $f := f(m);$
      $q := q(m)\{P2 \wedge A\}$
   od; $\{P2 \wedge m = m_n \wedge A\}$
   print $(sum)$;
   $n := n + 1\{P1 \wedge m = m_n \wedge A\}$
od$\{P1 \wedge m = m_n \wedge n = N\}$,
```

where A is $f = f(m) \wedge q = q(m)$, and the initialization of m has been broken up into two statements in the beginning of the program.

We now introduce the file operations in this program: replace $m := 0$ by *openfile*; $m := n + 1$, $f := f(m)$; $q := q(m)$ by *read*(f, q); and $f(M)$ by EOF. Also, n equals N at the end of the program, so we can add the statement to print $n - 1$, the number of items transacted. The resulting program, without the assertions, is as follows.

Solution 3.2

```
$n := 1;$ openfile; read$(f, q)$;
do $f \neq$ EOF$\rightarrow$
   $sum := 0; F := f;$
do $f = F\rightarrow$
      $sum := sum + q;$
      read$(f, q)$
   od;
   print $(sum)$;
   $n := n + 1$
od;
print $(n - 1)$.
```

This solution is identical to that obtained by the data structure design technique. It is considered to be a desired solution: its structure reflects the problem closely, it does not treat any card specially, it neatly handles all the groups one after the other, and it can be modified to add special processing at the beginning or at the end of a group.[3]

IV. SUMMARY

In this paper, we solved the McDonald's warehouse problem to show

how to effectively use Dijkstra's method to develop programs that process records in a file sequentially.

The solution was developed in three steps. We first assumed that the records of the file were available in an array, and developed the program disregarding the file operations. This program also used certain symbols whose values are not readily available. These symbols were removed in the next step by introducing additional program variables and modifying the program under correctness assertions so that the next step could be carried out. Finally, we replaced one initialization statement by *openfile*, and selected groups of statements by *readfile* or *writefile*. The example discussed in this paper involved only reading a file, but the same techniques apply when a file is written, too.

With a proper choice of invariants, the programs thus developed are comparable to those obtained by the data structure design, a technique that is considered to yield good programs for such problems.

V. ACKNOWLEDGMENTS

Thanks are due W. L. Bain, Jr. and N. Gehani, for many useful comments on an earlier draft of this paper, and G. D. Bergland, for bringing this problem to the author's attention.

REFERENCES

1. Dijkstra, E. W., *A Discipline of Programming*, Englewood Cliffs: Prentice-Hall, 1976.
2. Gries, D., "An Illustration of Current Ideas on the Derivation of Correct Proofs and Correct Programs," IEEE Trans. Software Engineering, *SE-2*, No. 4 (December 1976), pp. 238–44.
3. Bergland, G. D., "Structural Design Methodologies," 15th Annual Design Automation Conf., June 21, 1978, Las Vegas, Nevada, Design Automation Conf. Proc., June 1978.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

Spectral Properties and Band-Limiting Effects of Time-Compressed TV Signals in a Time-Compression Multiplexing System

By K. Y. ENG and O.-C. YUE

(Manuscript received March 30, 1981)

*Time-compression multiplexing (*TCM*) has recently been proposed for application in multiple* TV *transmissions through satellites. It is advantageous over frequency-division multiplexing because of its relative immunity to nonlinear transponder effects. Here we study two important and fundamental aspects of* TCM*—the spectral properties and band-limiting effects on the time-compressed signal. We derive the output spectrum of a time-compressed signal and show that if the original input signal is assumed to be: (i) band-limited to B Hz, (ii) segmented into T-second intervals before time compression by a factor of* $\alpha(\alpha \geq 1)$*, and (iii)* $1/T \ll B$*, then essentially all the spectral power in the output time-compressed signal is contained in the bandwidth* $|f| \leq \alpha B$ *Hz. This result is applicable to the* TV *case. Numerical examples on various types of spectra are also presented. Using the* TV *example, we further demonstrate that the ripples created by low-pass filtering the time-compressed signal up to* αB *Hz are small, and interburst interference due to these ripples can be kept negligible with a small guard time (about 2 percent of the burst duration) between different signal bursts. We also provide a brief discussion on some interesting spectral properties of time-compressed signals in spectrum-expansion applications.*

I. INTRODUCTION

Time-compression multiplexing (TCM) is a technique whereby multiple signals can be multiplexed together in a common communication channel for transmission.[1,2] A simple illustration of this method is shown in Fig. 1 where $x(t)$ is a continuous waveform intended for transmission. It is first divided into segments of T seconds each; and each segment is time compressed by a factor $\alpha(\alpha \geq 1)$, resulting in a

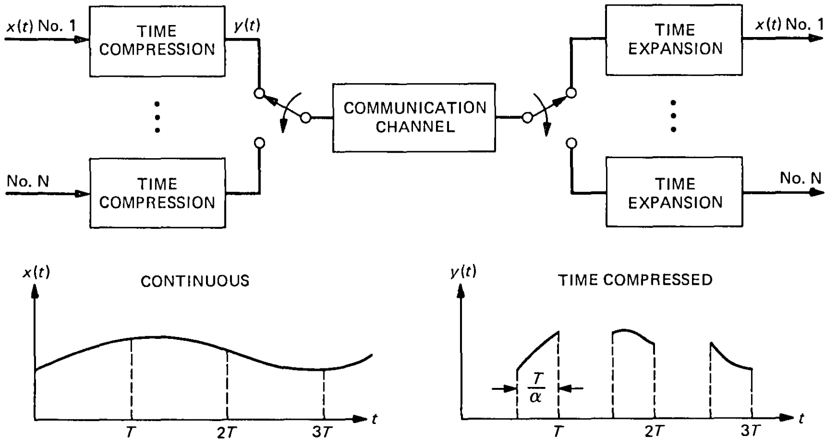


Fig. 1—A simple illustration for TCM.

bursty signal $y(t)$ with a burst duration of T/α seconds. A total of α such time-compressed signals can then be time multiplexed together for transmission. In the particular case of TV transmissions through satellites, TCM is advantageous over frequency-division multiplexing (FDM) because various degrading effects (e.g., intermodulation and intelligible crosstalk) due to transponder nonlinearities can be avoided by employing TCM. In a more general context, TCM is more efficient than FDM whenever time division can be accomplished more efficiently than frequency division. In this paper, we study two important and fundamental aspects of TCM—the spectral properties and the band-limiting effects of the time-compressed signal.

If we assume that the original signal $x(t)$ in Fig. 1 is band-limited to B Hz, then time compressing it by a factor of α in the infinite time duration, i.e., transforming $x(t)$ to $x(t/\alpha)$, would mean a frequency spectrum expansion by the same factor (α). However, as shown in the diagram, $x(t)$ is segmented into T-second intervals before time compression on each segment. Doing so, it is no longer obvious what the spectrum should look like or what the bandwidth expansion factor would be. It is clear though that the spectral power in $y(t)$ is nonzero beyond αB Hz due to the segmentation; and it is desirable that this power beyond αB Hz be small to maintain spectral efficiency in TCM. We derive and discuss an explicit expression for the output spectrum of the time-compressed signal $y(t)$ (see Section II) and show by numerical examples (Section III) that all the significant power is contained in the frequency bandwidth below αB Hz, thus, confirming the long-speculated result that the bandwidth expansion factor in TCM is the same as the time-compression factor.

To ensure compliance with the out-of-band emission requirements, signals are often filtered before transmission. Such a band-limiting operation on the time-compressed signal truncates its small but nonzero power beyond its passband (αB Hz) and creates ripples in its time waveform. The ripples following the trailing edge of each burst are important because they lead to interburst interference in the system. We demonstrate using a computer simulation (Section IV) that in the specific case of TV transmission, (*i*) such a band-limiting effect is minimal as long as all the spectral components below αB Hz are transmitted without distortion and (*ii*) the interburst interference can be kept negligible by introducing a small guard time in the order of two percent of the burst duration between different time-compressed TV signals. These encouraging results on both the bandwidth expansion and band-limiting effects assure us of the basic attractiveness of using TCM to transmit TV signals in nonlinear satellite channels.

It is well-known that time compression can also be used as a means to obtain spectrum expansion, e.g., Henry's spectrum expander used in radio astronomy.[3] In such a case, the key concern is that of spectrum distortion as analyzed thoroughly by Rowe.[4] We extend our results to examine this problem in an appendix, and some simple and interesting spectral properties pertinent to the spectrum expansion application are discussed.

II. SPECTRUM OF THE TIME-COMPRESSED SIGNAL

2.1 Derivation

Referring to Fig. 2, let $x(t)$ be an input signal to an ideal time compressor which performs the time compression on each T-second segment of the input waveform as discussed before. This is mathematically equivalent to first time compressing $x(t)$ in the infinite time duration, resulting in $x_c(t)$, by the required time-compression factor $\alpha(\alpha \geq 1)$, and then time-shifting segments of T/α-second duration in $x_c(t)$ to various proper time instants to arrive at $y(t)$, the desired time-compressed output as shown in the diagram. We are interested in the spectrum (or Fourier transform) of $y(t)$, denoted by $Y(f)$.

By the above definition, $y(t)$ is related to $x_c(t)$ by

$$y(t) = \sum_{k=-\infty}^{\infty} x_c[t - k(T - \tau)]\text{rect}_\tau(t - kT), \tag{1}$$

where

$$\tau \triangleq \frac{T}{\alpha}, \quad (\alpha \geq 1) \tag{2}$$

and

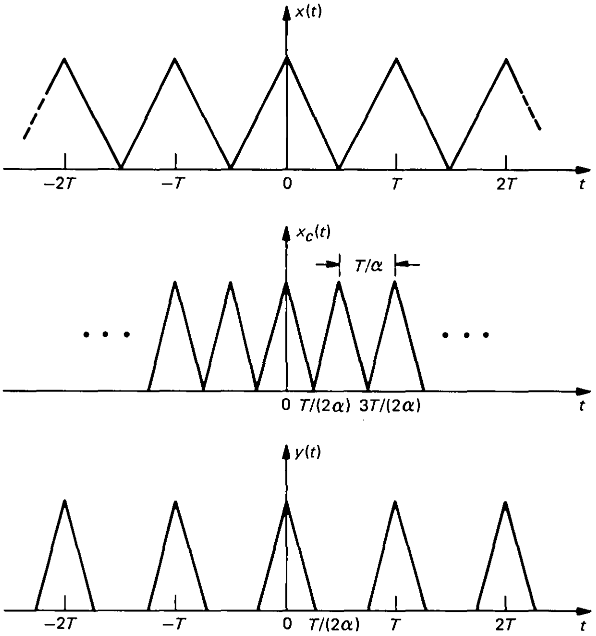


Fig. 2—An illustrative time-compression sequence.

$$\text{rect}_\tau(t) \triangleq \begin{cases} 1, & |t| \le \dfrac{\tau}{2}, \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

Using the following,

$$x_c(t) \leftrightarrow X_c(f), \tag{4}$$

$$\text{rect}_\tau(t) \leftrightarrow \tau \text{ sinc } \pi f \tau, \tag{5}$$

where ↔ denotes Fourier transform pair, and

$$\text{sinc } x \triangleq \frac{\sin x}{x}, \tag{6}$$

the Fourier transform of $y(t)$ can be written as

$$Y(f) = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} X_c(g) \exp[-j2\pi g k(T-\tau)]$$

$$\times \tau \text{ sinc}[\pi(f-g)\tau] \exp[-j2\pi(f-g)kT]dg. \tag{7}$$

Assuming that the summation and integration can be interchanged, the above becomes

$$Y(f) = \int_{-\infty}^{\infty} X_c(g)\tau \operatorname{sinc}[\pi(f-g)\tau] \sum_{k=-\infty}^{\infty}$$

$$\times \exp[-j2\pi k(-g\tau + fT)]dg. \quad (8)$$

Applying the well-known identity of

$$\sum_{k=-\infty}^{\infty} \exp(-j2\pi kfT) = \sum_{k=-\infty}^{\infty} \delta(fT - k) = \frac{1}{T}\sum_{k=-\infty}^{\infty} \delta\left(f - \frac{k}{T}\right), \quad (9)$$

$Y(f)$ is simplified as

$$\begin{aligned} Y(f) &= \int_{-\infty}^{\infty} X_c(g)\tau \operatorname{sinc}[\pi(f-g)\tau] \sum_{k=-\infty}^{\infty} \delta(fT - g\tau - k)\,dg \\ &= \sum_{k=-\infty}^{\infty} X_c\left(\frac{fT-k}{\tau}\right) \operatorname{sinc}\left[\pi\left(f - \frac{fT-k}{\tau}\right)\tau\right] \\ &= \sum_{k=-\infty}^{\infty} X_c\left[\frac{T}{\tau}\left(f - \frac{k}{T}\right)\right] \operatorname{sinc} \pi[f(\tau - T) + k]. \end{aligned} \quad (10)$$

Using (2) and

$$x_c(t) = x(\alpha t) \leftrightarrow X_c(f) = \frac{1}{\alpha} X\left(\frac{f}{\alpha}\right), \quad (11)$$

where $X(f)$ is the Fourier transform of $x(t)$, we get the final result of

$$Y(f) = \frac{1}{\alpha}\sum_{k=-\infty}^{\infty} X\left(f - \frac{k}{T}\right) \operatorname{sinc}\left[\pi\left(f - \frac{f}{\alpha} - \frac{k}{T}\right)T\right]. \quad (12)$$

The above expression relates the output spectrum $Y(f)$ to the input spectrum $X(f)$ and is the basic relationship we work with in the rest of the paper. The preceding derivation is the simplest that we are aware of (see Appendix A and Ref. 5 for comparison). We now proceed to discuss some simple properties of $Y(f)$.

2.2 Simple properties

We are interested in the properties of $Y(f)$ that convey information about the spectral occupancy problem (or the bandwidth expansion factor) in TCM systems. In this regard, we must note that $Y(f)$ derived above is just the Fourier transform of one single time-compressed signal in the TCM system. If there are N users for the channel (i.e., N time-compressed signals for transmission), then the total signal in the transmission channel is (without post-time-compression filtering):

$$z(t) = \sum_{i=1}^{N} y_i(t), \quad (13)$$

where each $y_i(t)$ is a time-compressed signal resembling $y(t)$ considered previously. The power spectrum of the total signal in the channel is

$$P(f) = |Z(f)|^2 = \left| \sum_{i=1}^{N} Y_i(f) \right|^2, \tag{14}$$

where $Z(f)$ and $Y_i(f)$ are the Fourier transforms of $z(t)$ and $y_i(t)$, respectively. It is well-known that

$$P(f) = \sum_{i=1}^{N} |Y_i(f)|^2 \tag{15}$$

only when the $y_i(t)$ are all uncorrelated. In the particular case of TV transmission, the various time-compressed TV signals are not totally uncorrelated because of the presence of sync pulses, color subcarrier bursts, and so on. However, from the point of view of spectral occupancy (i.e., the total power contained in some passband), consideration of $Y(f)$ alone is sufficient. In any event, $P(f)$ is calculable from the above if it is needed.

Without loss of generality, we may normalize $T = 1$, and $Y(f)$ becomes

$$Y(f) = \frac{1}{\alpha} \sum_{k=-\infty}^{\infty} X(f-k) \operatorname{sinc}\left[\pi\left(f - \frac{f}{\alpha} - k\right)\right]. \tag{16}$$

The simplest property observable from the above is the output dc component in $Y(f)$, i.e.,

$$\begin{aligned} Y(0) &= \frac{1}{\alpha} \sum_{k=-\infty}^{\infty} X(-k) \operatorname{sinc}(\pi k) \\ &= \frac{1}{\alpha} X(0), \end{aligned} \tag{17}$$

which holds for any general $X(f)$ (see Ref. 5 for comparison).

Let us now examine the bandwidth property of $y(t)$. We assume that the input signal is band-limited to B Hz, i.e., $X(f)$ is zero for $|f| > B$. With the normalization of $T = 1$, $X(f)$ is band-limited to $M = B/T$. Dropping the multiplying constant of $1/\alpha$ for convenience, and at a particular frequency $f = \alpha f_x$ (recall that α is the time-compression factor),

$$Y(\alpha f_x) = \sum_{k=-\infty}^{\infty} X(\alpha f_x - k) \operatorname{sinc} \pi[f_x(\alpha - 1) - k], \tag{18}$$

where the sinc function provides weightings for various points in $X(f)$. Since the sinc function is maximum when its argument is zero, it is sensible to perform the summation starting with the value of k that maximizes the sinc function. Denoting such a value of k by k_0, it is given by solving

$$f_x(\alpha - 1) - k_0 = 0. \tag{19}$$

The solution is

$$k_0 = f_x(\alpha - 1) + \epsilon, \quad |\epsilon| \leq 0.5, \tag{20}$$

where ϵ is necessary because k_0 is restricted to be integer, and k_0 is unique, except for the case $\epsilon = \pm 0.5$. It should be emphasized that ϵ depends on both f_x and α. Using this expression for k_0, $Y(\alpha f_x)$ can be written as

$$\begin{aligned} Y(\alpha f_x) &= \{X(\alpha f_x - k) \operatorname{sinc} \pi[f_x(\alpha - 1) - k]\}_{k=k_0} \\ &\quad + \sum_{k \neq k_0} X(\alpha f_x - k) \operatorname{sinc} \pi[f_x(\alpha - 1) - k] \\ &= X(f_x - \epsilon) \operatorname{sinc} \pi(-\epsilon) \\ &\quad + \sum_{K \neq 0} X(f_x - \epsilon + K) \operatorname{sinc} \pi(-\epsilon + K), \end{aligned} \tag{21}$$

where K in the summation is taken as $K = \pm 1$, ± 2, and so on. A graphical representation of the above is depicted in Fig. 3.

To get immediate insight into the bandwidth property, we consider the following two cases:

Case 1: α = integer, f_x = integer. Under this assumption, $\epsilon = 0$ and we have a simple relationship of

$$Y(\alpha f_x) = X(f_x). \tag{22}$$

This means that every integer value of f in $X(f)$ is mapped exactly onto αf in $Y(f)$. Without the normalization on T, this is equivalent to saying that every integer multiple of $1/T$ in $X(f)$ is mapped exactly onto α/T in $Y(f)$ as shown in Fig. 4. If the condition that $1/T \ll B$ holds, it is almost certain that the spectrum $Y(f)$ is simply the

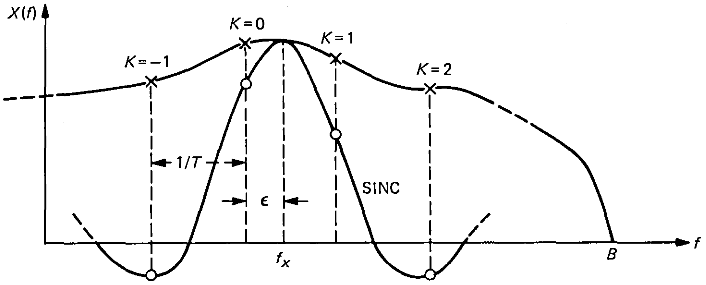


Fig. 3—Graphical illustration for the summation in the expression for the output spectrum of a time-compressed signal.

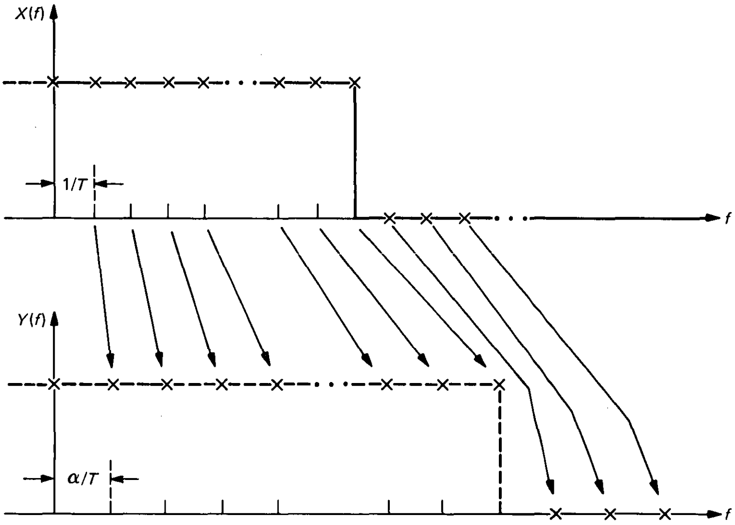


Fig. 4—Frequency-spectrum expansion property through time compression.

frequency-expanded version of $X(f)$ by the factor α with some insignificant sidebands for $|f| > \alpha B$ Hz. In the particular case of TCM/TV, T is taken as a scan line duration, yielding $1/T \approx 15.73$ kHz. With $B = 4.2$ MHz, $BT \approx 267$, using α as the bandwidth expansion factor is, therefore, a good rule of thumb for the TV case. We note that the condition of α and f_x being integers is merely an artifact due to normalization. Therefore, we emphasize that this case is indeed more general than it appears to be. For instance, for any given set of α, we can always find a set of f_x such that this condition holds.

Case 2: $\alpha \neq$ integer, $f_x \neq$ integer. ϵ is generally nonzero here, and $Y(\alpha f_x)$ is given by a weighted sum of $X(f_x - \epsilon + k)$ [see (21) and Fig. 3] with $X(f_x - \epsilon)$ as the main contributor. An alternative view is that $Y(\alpha f_x)$ is the weighted average of $X(f_x - \epsilon)$ and its neighboring points. The output spectrum $Y(f)$ is again a frequency-expanded version of $X(f)$, except for ripples created by the averaging process. It is noted that ϵ can be zero here resulting in $Y(\alpha f_x) = X(f_x)$. This occurs whenever $f_x(\alpha - 1)$ is an integer in (20), e.g., $\alpha = 3.5$, $f_x = 106.8$.

The foregoing discussion by and large answers the basic question on bandwidth expansion in TCM systems. The result of using the time-compression factor α as the bandwidth-expansion factor makes sense for most cases where the input spectrum $X(f)$ can be modeled as continuous and band-limited. The inclusion of peculiar nulls and delta functions in $X(f)$ would complicate the matter, but it can be examined in detail using the equations provided above. As to the precise shape of $Y(f)$ in comparison to $X(\alpha f)$, which is relevant in the spectrum expansion application, some interesting discussions are given in Appendix B.

III. NUMERICAL EXAMPLES OF OUTPUT SPECTRUM

We present in this section specific numerical examples of output spectra calculated from (12). There are four different types of $X(f)$ in the examples:

(*i*) Rectangular

$$X(f) = \begin{cases} 1, & |f| \leq B, \\ 0, & \text{otherwise}. \end{cases} \tag{23}$$

B is normalized to be 1 Hz in the calculation, and T is taken as $267/B$ seconds (the TV case). Results for five different values of the time-compression factor α are plotted in Fig. 5. The peak-to-peak ripples in the output passband ($|f| \leq B$) are less than 0.5 dB and the sidebands are more than 25 dB down in the vicinity beyond the edge of the passband and drop off very rapidly. There is no doubt that most of the spectral power is contained in the bandwidth $|f| \leq B$.

(*ii*) Triangular

$$X(f) = \begin{cases} 1 - \dfrac{|f|}{B}, & |f| \leq B, \\ 0 \quad , & \text{otherwise}. \end{cases} \tag{24}$$

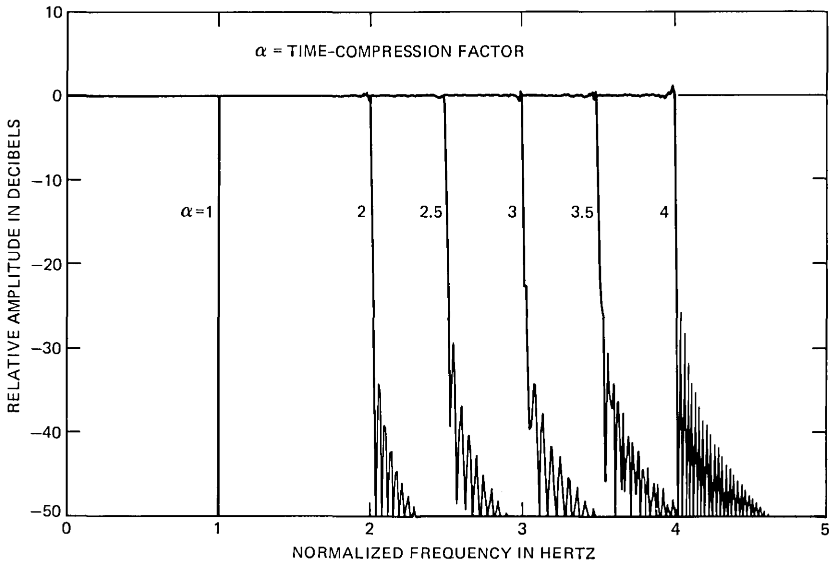


Fig. 5—Output spectrum of time-compressed signal with a rectangular input spectrum.

B is normalized to be 1 Hz, and $T = 267/B$ seconds. The results are plotted in Fig. 6. Here the sidebands are so low that they are not observable in the diagram. Of course, this is due to the taper-off characteristic in $X(f)$. Some small ripples are again present inside the bandwidth $|f| \leq \alpha B$.

(*iii*) Half-Cosine

$$X(f) = \begin{cases} \cos\left(\dfrac{\pi}{2}\dfrac{|f|}{B}\right), & |f| \leq B, \\ 0 \quad , & \text{otherwise.} \end{cases} \tag{25}$$

B is again 1 Hz and $T = 267/B$ seconds. The results are plotted in Fig. 7, and the same observations as in (*ii*) apply here.

(*iv*) Truncated Half-Cosine

The equation for $X(f)$ is the same as in (*iii*) above, except that B is 0.9362 Hz, which corresponds to a 20-dB taper at $X(B)$ as compared to $X(0)$. T is again taken as 267 seconds. The results are shown in Fig. 8 where the glitches at the edge of the output passband (i.e., $f \approx \alpha B$) are evident. Note that the sidebands outside the passband are much lower compared to those in (*i*) above.

IV. BAND-LIMITING EFFECTS

When the original input signal $x(t)$ is band-limited to B Hz, we have shown that most of the power in the output time-compressed signal

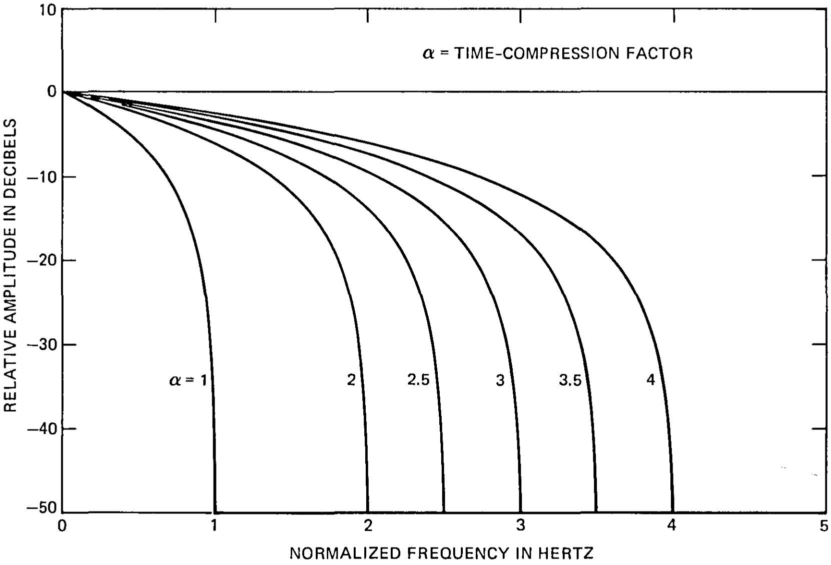


Fig. 6—Output spectrum of time-compressed signal with a triangular input spectrum.

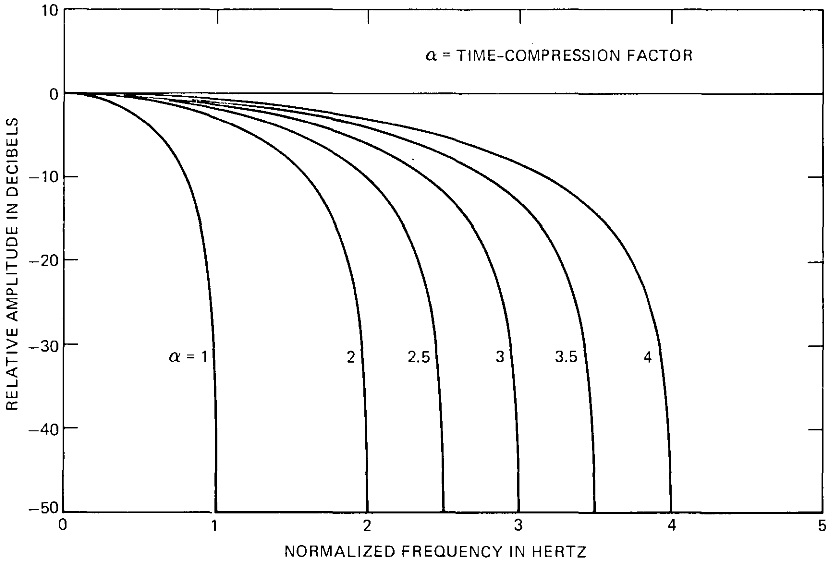


Fig. 7—Output spectrum of time-compressed signal with a half-cosine input spectrum.

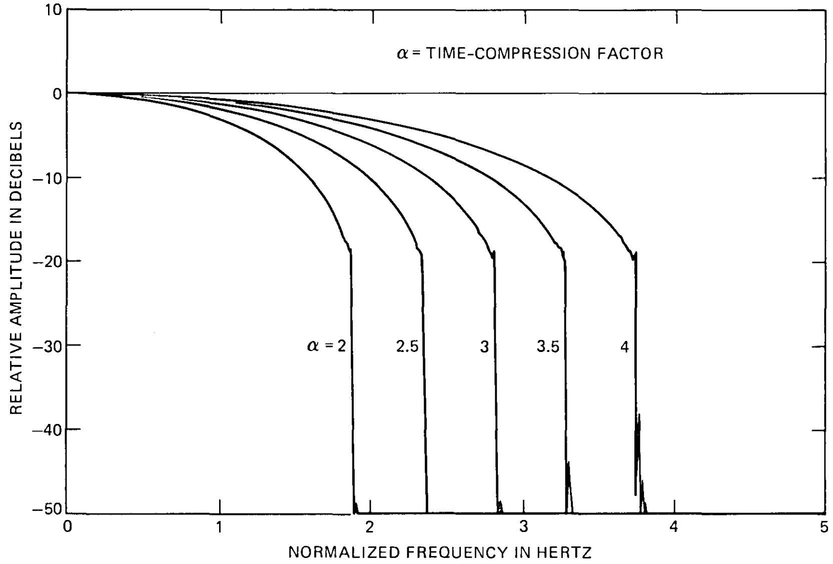


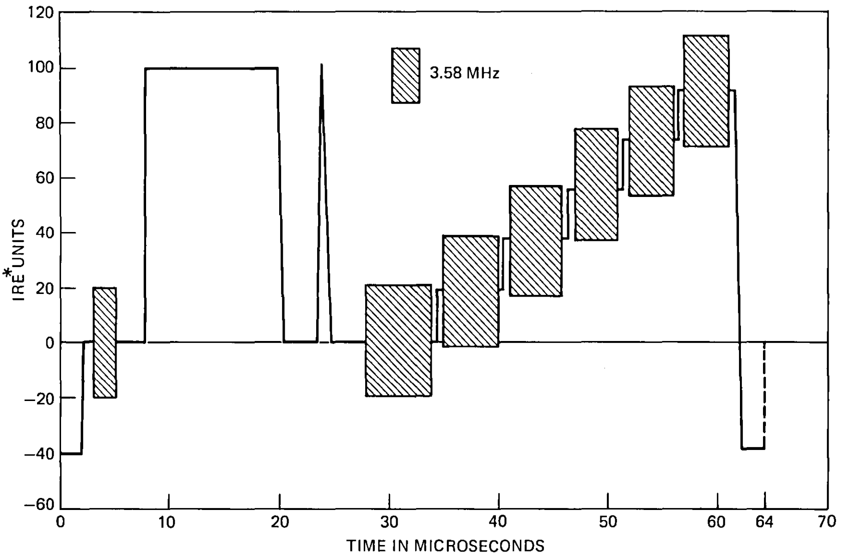
Fig. 8—Output spectrum of time-compressed signal with a 20-dB tapered half-cosine spectrum.

$y(t)$ is confined to $|f| \leq \alpha B$ Hz. Band-limiting $y(t)$ up to αB Hz therefore should hardly affect the fidelity of the signal itself. However, such filtering does create ripples following each time-compressed signal burst, where a sharp edge occurs in the time waveform. These ripples constitute interburst interference and could be a potential source of degradation in a TCM system. We demonstrate via a simulation on a TV signal in this section that the problem can be alleviated by introducing a small guard time (about 2 percent of the burst duration) between time-compressed signal bursts from different users.

In our computer simulation, we generate a TV test signal which is similar to the composite test signal in Ref. 6. A scan line (64 μs) of this is shown in Fig. 9. We first band-limit the test signal by a low-pass (LP) filter with zero delay and a raised-cosine amplitude roll-off of

$$H(f) = \begin{cases} 1 & , |f| \leq F_1 + \dfrac{1-r}{2t_0}, \\ \dfrac{1}{2}\left\{1 - \sin\left[\dfrac{t_0\pi(|f| - F_1)}{r} - \dfrac{\pi}{2r}\right]\right\} & \\ & , \dfrac{1-r}{2t_0} \leq |f| - F_1 \leq \dfrac{1+r}{2t_0}, \\ 0 & , \text{otherwise}. \end{cases} \tag{26}$$

where the parameters F_1, t_0 and r ($F_1 \geq 0$, $t_0 \geq 0$; $0 \leq r \leq 1$) control the



* INSTITUTE OF RADIO ENGINEERS

Fig. 9—A modified composite test signal.

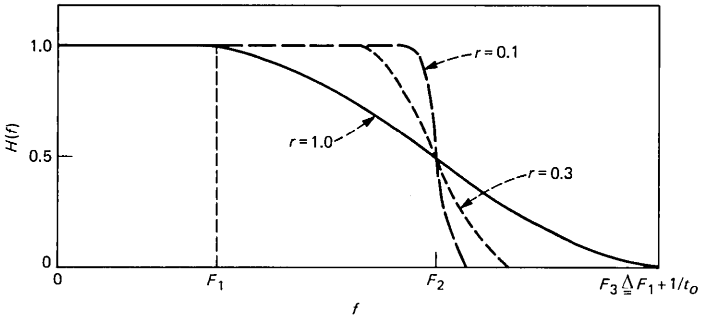


Fig. 10—A low-pass filter with a raised-cosine roll-off.

filter shape. Instead of using t_0 and r to define the filter shape, we use the two critical frequencies defined by (Fig. 10):

$$F_2 \triangleq F_1 + \frac{1}{2t_0}; \tag{27}$$

$$F_3 \triangleq F_1 + \frac{1+r}{2t_0}. \tag{28}$$

The specific values for F_1, F_2, and F_3 are 4.2 MHz, 4.8 MHz, and 5 MHz, respectively.

After the initial band-limiting, we perform an ideal time compression over each scan line, and the signal is then compressed (with $\alpha = 2$) into time bursts, each of which is 32 μs long. We then filter the time-compressed signal by the LP filter referred to above with F_1, F_2, and F_3 at 8.4 MHz, 9.6 MHz, and 10 MHz, respectively. The ripples following each 32-μs signal burst are observed, and the data are plotted in Fig. 11 where the ripple magnitude is defined as the ratio of the peak-to-peak ripple voltage to the peak-to-peak picture voltage. As seen from the figure, a guard time of 0.5 μs is sufficient for controlling the interburst interference. This amounts to about 2 percent of the burst duration.

V. CONCLUSION

We have studied two important and fundamental aspects of TCM—the spectral properties and band-limiting effects of the time-compressed signal. We find under the assumptions that (*i*) the original input signal $x(t)$ is band-limited to B Hz, (*ii*) $x(t)$ is segmented into T-second intervals before time compression by a factor of $\alpha(\alpha \geq 1)$, and (*iii*) $1/T \ll B$, then essentially all the spectral power in the output time-compressed waveform is confined to frequencies below αB Hz. This result is immediately applicable to the TV case. Numerical ex-

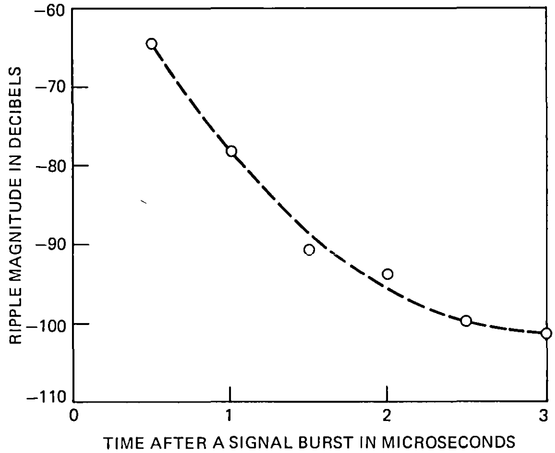


Fig. 11—Ripple magnitude following a time-compressed signal burst in the TV simulation.

amples on various types of spectra verify this and show that the spectral sidebands beyond αB Hz are very low. We also find that filtering of a time-compressed TV signal creates small ripples following each signal burst, but the interburst interference due to these ripples can be kept negligible with a small guard time (about 2 percent of the burst duration) between different signal bursts.

REFERENCES

1. J. E. Flood and D. I. Urquhart-Pullen, "Time-Compression-Multiplex Transmission," Proc. IEE, *111,* No. 4 (April 1964), pp. 647–68.
2. D. H. Morgen and E. N. Protonotarios, "Time Compression Multiplexing for Loop Transmission of Speech Signals," IEEE Trans. Commun., *COM-22,* No. 12 (December 1974), pp. 1932–9.
3. P. S. Henry, "Variable Resolution Capability for Multichannel Filter Spectrometers," Rev. Sci. Instrum., *50,* No. 2 (February 1979), pp. 185–92.
4. Harrison E. Rowe, "Signal and Noise Response of a Spectrum Expander," IEEE Trans. Circ., Syst., *CAS-27,* No. 9 (September 1980), pp. 804–15.
5. E. N. Aniebona, "The Spectrum of a Time-Compression-Multiplex (TCM) Signal and its Transmission Bandwidth Requirements," Proc. ICC (1976) pp. 43-1–6.
6. NTC Report No. 7, The Public Broadcasting Service, Washington, D. C., June 1975, Revised January 1976.

APPENDIX A

An Alternate Derivation for the Spectrum of the Time-Compressed Signal

Referring to Fig. 2, the input signal can be written as

$$x(t) = \sum_{i=-\infty}^{\infty} x(t)\,\mathrm{rect}_T(t - iT), \tag{29}$$

where $\mathrm{rect}_T(t)$ is defined in (3). The output time-compressed signal is

$$y(t) = \sum_{i=-\infty}^{\infty} x\left\{\alpha\left[t - iT\left(1 - \frac{1}{\alpha}\right)\right]\right\} \mathrm{rect}_\tau (t - iT), \tag{30}$$

where α is the time-compression factor, and τ is as defined in (2). Note that

$$\mathrm{rect}_\tau (t - iT) \leftrightarrow \tau \,\mathrm{sinc}(\pi f \tau) \exp(-j2\pi fiT), \tag{31}$$

$$x\left\{\alpha\left[t - iT\left(1 - \frac{1}{\alpha}\right)\right]\right\} \leftrightarrow \frac{X\left(\dfrac{f}{\alpha}\right)}{\alpha} \exp\left[-j2\pi fiT\left(1 - \frac{1}{\alpha}\right)\right]. \tag{32}$$

The spectrum of $y(t)$ involves the convolution of the two right sides in (31) and (32) and is

$$\begin{aligned} Y(f) = \tau \sum_i \int_{-\infty}^{\infty} & X(\beta) \exp[-j\beta iT(\alpha - 1)] \\ & \times \mathrm{sinc}[\pi\tau(f - \beta\alpha)] \exp[-j(2\pi f - \beta\alpha) iT] d\beta. \end{aligned} \tag{33}$$

Using the identity of (9), the above can be rewritten as

$$Y(f) = \tau \sum_i \int_{-\infty}^{\infty} X(\beta) \,\mathrm{sinc}[\pi\tau(f - \beta\alpha)]\delta\left[T\left(f - \frac{\beta}{2\pi}\right) - i\right] d\beta. \tag{34}$$

With a change of variable of

$$\sigma = T\left(f - \frac{\beta}{2\pi}\right), \tag{35}$$

$Y(f)$ becomes

$$\begin{aligned} Y(f) &= \frac{1}{\alpha} \sum_i \int_{-\infty}^{\infty} X\left(\frac{Tf - \sigma}{T}\right) \mathrm{sinc}\left[\pi\tau\left(f - \frac{Tf - \sigma}{T}\alpha\right)\right] \delta(\sigma - i) d\sigma \\ &= \frac{1}{\alpha} \sum_i X\left(f - \frac{i}{T}\right) \mathrm{sinc}\left[\pi\tau\left(f - f\alpha + \frac{i\alpha}{T}\right)\right] \\ &= \frac{1}{\alpha} \sum_i X\left(f - \frac{i}{T}\right) \mathrm{sinc}\left[\pi\left(f - \frac{f}{\alpha} - \frac{i}{T}\right)T\right], \end{aligned} \tag{36}$$

which is the same as (12).

A simple way to check $Y(f)$ is of course to let $\alpha = 1$ in (36), which yields

$$\begin{aligned} Y(f) &= \sum_i X\left(f - \frac{i}{T}\right) \mathrm{sinc}\ \pi(-i) \\ &= X(f), \end{aligned} \tag{37}$$

as expected. Another way to check $Y(f)$ involves letting $\alpha \to \infty$. In doing so, we have to assume that the energy in $x_c(t)$ is preserved through the time compression, i.e.,

$$\int_{-\infty}^{\infty} x_c^2(t)\,dt = \int_{-\infty}^{\infty} x^2(t)\,dt. \tag{38}$$

This means that the multiplying factor $1/\alpha$ in (36) can be dropped, and with $\alpha \to \infty$,

$$Y(f) = \sum_i X\left(f - \frac{i}{T}\right) \operatorname{sinc}\left[\pi\left(f - \frac{i}{T}\right)T\right]. \tag{39}$$

We now try to verify (39) by a different means. Using $\alpha \to \infty$ as described above, the output time waveform is

$$\begin{aligned} y(t) &= \sum_i \delta(t - iT) \int_{iT - \frac{T}{2}}^{iT + \frac{T}{2}} x(\tau)\,d\tau \\ &= \sum_i \delta(t - iT)\bar{y}(iT). \end{aligned} \tag{40}$$

Note that if $\bar{Y}(f) \leftrightarrow \bar{y}(t)$, then

$$\frac{1}{T}\sum_k \bar{Y}\left(f - \frac{k}{T}\right) \leftrightarrow \sum_i \delta(t - iT)\bar{y}(iT). \tag{41}$$

Consider now a general definition of

$$\begin{aligned} \bar{y}(t) &\triangleq \int_{t - \frac{T}{2}}^{t + \frac{T}{2}} x(\tau)\,d\tau \\ &= \int_{-\infty}^{\infty} x(\tau)\,\mathrm{rect}_T(t - \tau)\,d\tau. \end{aligned} \tag{42}$$

The Fourier transform of $\bar{y}(t)$ is then

$$\bar{Y}(f) = X(f)[T \operatorname{sinc}(\pi f T)]. \tag{43}$$

Using the above, (40) and (41), we get the desired result of (39).

APPENDIX B

Additional Spectral Properties of the Time-Compressed Signal

With $T = 1$, we show in (21) that the spectral output of $f = \alpha f_x$ in $Y(f)$ is

$$\begin{aligned} Y(\alpha f_x) = {} & X(f_x - \epsilon) \operatorname{sinc} \pi(-\epsilon) \\ & + \sum_{K \neq 0} X(f_x - \epsilon + K) \operatorname{sinc} \pi(-\epsilon + K), \end{aligned} \tag{44}$$

where ϵ depends on both α and f_x. Whenever $\epsilon = 0$, we get

$$Y(\alpha f_x) = X(f_x), \tag{45}$$

which is a desirable result for spectrum-expansion applications. Consider now a simple example, where $\alpha = 2.1$ and $X(f)$ consists of two spectral lines at $f = \pm 5$ Hz (a constant sine wave). We expect two spectral lines in $Y(f)$ at $f = \pm 2.1 \times 5 = \pm 10.5$ Hz. But using $f_x = 5$ in (44), we have

$$\begin{aligned} Y(10.5) &= X(5 - 0.5) \operatorname{sinc} \pi(0.5) \\ &\quad + \sum_{K \neq 0} X(5 - 0.5 + K) \operatorname{sinc} \pi(-0.5 + K) = 0, \end{aligned} \tag{46}$$

which is so because each term containing $X(f)$ in the summation is identically zero as $X(f)$ admits nonzero values only at $f = \pm 5$ Hz. This absence of output at $f = 10.5$ is due to the fact that $Y(f)$ admits nonzero outputs only at $f =$ integer, or equivalently in multiple spacings of $1/T$ in the unnormalized case. The segmentation of the input signal $x(t)$ into T-second intervals can therefore be viewed as making the frequency resolution in $Y(f)$ $1/T$.

As mentioned previously, when $\epsilon \neq 0$ in (44), the output $Y(\alpha f_x)$ can be viewed as some weighted average of $X(f_x - \epsilon)$ and its neighboring points. How close is this average to the value $X(f_x)$? We do not have a satisfactory answer, but the following discussion is interesting. Let us change the notation in (44) to

$$\begin{aligned} \hat{X}(f_x) &= X(f_x - \epsilon) \operatorname{sinc} \pi(-\epsilon) \\ &\quad + \sum_{K \neq 0} X(f_x - \epsilon + K) \operatorname{sinc} \pi(-\epsilon + K), \end{aligned} \tag{47}$$

where $\hat{X}(f_x)$ denotes an interpolated or estimated value for $X(f)$. A physical interpretation on the above was presented in Fig. 3. An alternate view is to rewrite the second term in the above as

$$\sum_{K \neq 0} X(f_x - \epsilon + K) \operatorname{sinc} \pi[(f_x - \epsilon + K) - f_x], \tag{48}$$

which is graphically interpreted in Fig. 12. We now do the following manipulations:

$$\begin{aligned} \hat{X}(f_x) &= \sum_K X(f_x - \epsilon + K) \operatorname{sinc} \pi[(f_x - \epsilon + K) - f_x] \\ &= \sum_K X(f_x - \epsilon + K) \int_{-1/2}^{1/2} \exp\{j2\pi t[(f_x - \epsilon + K) - f_x]\}\, dt \\ &= \int_{-1/2}^{1/2} \sum_K X(f_x - \epsilon + K) \exp[j2\pi t(f_x - \epsilon + K)] \\ &\quad \times \exp(-j2\pi f_x t)\, dt. \end{aligned} \tag{49}$$

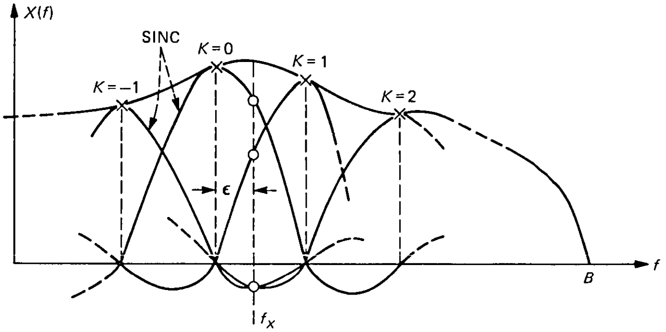


Fig. 12—Alternative representation for the summation in the output spectrum of a time-compressed signal.

We note that the term

$$\sum_K X(f_x - \epsilon + K) \exp[j2\pi t(f_x - \epsilon + K)] \tag{50}$$

is a Fourier series with a period of unity. Denoting this Fourier series by $\hat{x}(t)$, we see that

$$\begin{aligned}\hat{x}(t) &= \exp[j2\pi t(f_x - \epsilon)] \sum_K X(f_x - \epsilon + K) \exp(j2\pi Kt) \\ &= \sum_k x(t-k) \exp[j2\pi k(f_x - \epsilon)],\end{aligned} \tag{51}$$

which is a complicated summation of various time- and phase-shifted versions of $x(t)$.

We now extend our previous results to a limited case applicable to Refs. 5, 6. Consider the special case where α is an integer, denoted by N. We take the time-compressed waveform $y(t)$ in Fig. 2, time shift it by iT/N ($i = 0$ to $N-1$) and add all N waveforms together. The resultant waveform has no blank time interval and is basically a staggering of N time compressed $y(t)$ which we have considered so far. This can be written as

$$z(t) = \sum_{i=0}^{N-1} y\left(t - \frac{iT}{N}\right), \quad N \geq 2. \tag{52}$$

Its Fourier transform is

$$\begin{aligned}Z(f) &= Y(f) \sum_{i=0}^{N-1} \exp\left(-j2\pi f \frac{iT}{N}\right) \\ &= Y(f) \frac{1 - \exp(-j2\pi fT)}{1 - \exp\left(-j2\pi f \frac{T}{N}\right)},\end{aligned} \tag{53}$$

where $Y(f)$ is the Fourier transform of $y(t)$. The magnitude of the second complex factor is

$$\left|\frac{1 - \exp(-j2\pi fT)}{1 - \exp\left(-j2\pi f\dfrac{T}{N}\right)}\right| = \sqrt{\frac{1 - \cos 2\pi fT}{1 - \cos\left(2\pi f\dfrac{T}{N}\right)}}. \tag{54}$$

The above term vanishes whenever only the numerator vanishes, and is nonzero when both numerator and denominator vanish simultaneously. This means that the above term is zero when

$$f = \frac{k}{T}, \quad (k = \text{integer}), \tag{55}$$

except at those points where

$$f = k\frac{N}{T} \tag{56}$$

holds. Therefore, one may infer that the frequency resolution in this case is N/T as compared to $1/T$ in the single $y(t)$ case.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

Transmission Cathodoluminescence

By A. K. CHIN, H. TEMKIN, and S. MAHAJAN

(Manuscript received May 11, 1981)

The application of transmission cathodoluminescence (TCL) in evaluating the quality of luminescing materials is reviewed. This scanning electron microscope technique is particularly useful in imaging localized nonuniformities, such as dislocations and inclusions, in semiconductors. Understanding and analyzing such material defects are of practical importance since they greatly affect device performance and lifetime. After a detailed description of TCL, the advantages of this technique in comparison to defect etching, cathodoluminescence imaging, and electron beam-induced current (EBIC) is presented. Since TCL is simple to perform, this technique can be used to evaluate and monitor material growth and device processing procedures. Although TCL may be applied to any luminescing material, this paper demonstrates its usefulness for the GaAs/GaAlAs and InP/InGaAsP materials systems which provide most sources and detectors for optical communication.

I. INTRODUCTION

Diode lasers, light-emitting diodes (LEDs), and photodiodes are integral components of the current lightwave transmission systems. These devices are mainly based on the III-V materials systems, e.g., GaAs/GaAlAs and InP/InGaAsP. Both material and device development are recent, relative to the highly developed Si-based technology. Material defect analysis is required to evaluate and monitor material growth and device processing procedures, especially during the early stages of development.

The performance and reliability of optoelectronic devices are determined by the quality of materials and fabrication processes. Threading dislocations have been shown to increase the leakage current of both InP and GaAs photodiodes;[1,2] these dislocations are also sources of microplasmas in avalanche photodiodes.[3-5] The quantum efficiency of GaP and GaAlAs:Si LEDs is strongly dependent on the dislocation density.[6,7] Threading dislocations and inclusions are known sources of dark line and dark spot defects in degraded LEDs and solid-state lasers fabricated from the GaAs/GaAlAs and InP/InGaAsP materials systems.[8-12] Finally, stresses from dielectric coatings are known to induce defects and accelerate device degradation.[12,13]

Characterization techniques are required to assess the effects of growth and processing procedures on material or device quality. Because of the variety of defects, materials, and devices, a number of characterization techniques have been developed. These techniques are listed and compared in Table I.

The purpose of this paper is to review the most recently developed technique of transmission cathodoluminescence (TCL) and to compare it with other defect analysis techniques currently in use.[10,14-23] Transmission cathodoluminescence is performed on a scanning electron microscope (SEM); the magnification and depth of field available in an SEM are exploited.

1.1 Defect analysis techniques

The technique with the highest magnification and resolution, transmission electron microscopy (TEM), reveals the nature and origin of

Table I—Comparison of various characterization techniques

Technique	Disadvantages
IR microscopy	Low resolution Dependent on detector sensitivity
Scanning photoluminescence	Requires removal of contact layer Dependent on detector sensitivity Difficult to identify features because of lack of an optical image
Scanning photocurrent	Requires p-n junction or Schottky barrier Requires electrical contacts Difficult to identify features because of lack of an optical image
CL	Requires removal of contact layer Dependent on detector sensitivity
EBIC	Requires p-n junction or Schottky barrier Requires electrical contacts
TCL	Dependent on detector sensitivity
Defect etching	Material, orientation, dopant, and etch dependent Destructive
X-ray topography	Low resolution Slow Restricted to relatively small samples
TEM	Difficult sample preparation Destructive Small volume of material probed

material defects. However, it cannot be used routinely for screening defective materials because it is destructive, the sample preparation is very difficult, and the volume of material probed is small relative to device dimensions. X-ray topography is relatively slow, has limited resolution, and is limited to small samples because of internal stresses generally associated with device wafers.[24,25] The evaluation of wafers by electron beam-induced current (EBIC) scan, cathodoluminescence (CL) scan, photoluminescence (PL) scan, photocurrent scan, and infrared microscopy have also been reported.[9,26–32] The electron excitation techniques—EBIC and CL scan—are more useful in comparison to the conceptually similar photoexcitation techniques, i.e. photocurrent and PL scan, because of the availability of an exactly corresponding secondary-electron (SE) image to isolate surface details. Even when this disadvantage is overcome by careful experimental design, dislocations which are responsible for poor device performance and reliability cannot be readily imaged using these techniques, although some specific examples may be found. For example, dark spots in CL images have been identified to be dislocations only for the case of GaP, GaAs, and CdTe.[6,33–35] Additionally, it is difficult to observe dislocations in GaAs:Se.[27] Moreover, dislocations have not been imaged in GaAs:Si.

Etching is the simplest and, thus, the most widely used method for revealing defects. Once the identity of etch features has been established, evaluation of material quality by etching is relatively fast. Additionally, etching has sufficient resolution to reveal a defect density as high as $\sim 10^7$ cm^{-2}. However, aside from being destructive, etching is material, material orientation, and dopant dependent. For example, the KOH etch readily reveals dislocations on the {100} and {111} surfaces of GaAs.[36,37] However, it cannot be used for the {110} surface which is the cleavage plane and, thus, forms the mirror faces of solid-state lasers. Huber etch works well for the {100} surface of InP but not for the {110} surface.[38] The A-B etch reveals dislocations on the {111} surface of InGaAsP, but no etch has been developed for the technologically important {100} surface.[39,40] To estimate dislocation densities of InGaAsP layers in a device structure, the InP confining layers had to be etched.[40] Finally, the R-C and A-B etchants are successful for the {111}B face of GaP but cannot be used for other orientations.[39,41,42]

The interpretation of etch patterns is often difficult even with known defect etchants.[21] Figure 1 shows the response of an InP crystal to four different etchants reported to reveal dislocations: (a) Huber etch, (b) 6:6:1, (c) RRE, and (d) modified A-B etch.[38,43–45] The sample is the {100} surface of a crystal doped with Sn to 2×10^{18} cm^{-3} and is a typical substrate for device wafers. A brief comparison shows large differences in these etch patterns. Dislocation and saucer pits are well

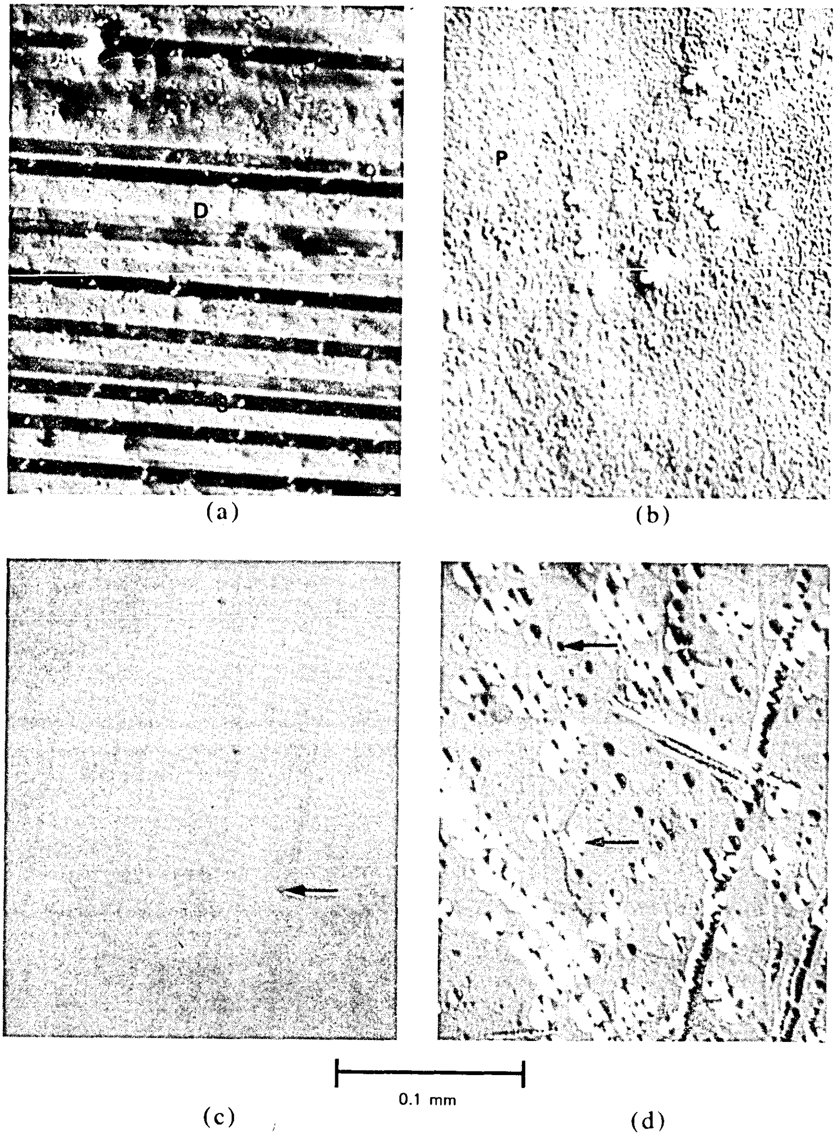


Fig. 1—Optical micrographs showing the response of InP:Sn to different etchants: (a) Huber etch, (b) 6:6:1, (c) RRE, and (d) the modified A-B etch. D, S, and P refer to dislocation pit, saucer pit, and protrusions, respectively.

delineated with the Huber etch but are not discernible with the modified A-B etch. On the other hand, growth striations, which are periodic variations in impurity concentration, are particularly well defined after A-B etching. The etch patterns of the other two etches are in between that of the Huber and modified A-B etches. Therefore, the use of an etchant to reveal defects is often unreliable.

Transmission cathodoluminescence reveals defects as dark spots, similar to the CL and EBIC techniques.[14] The characteristic images of dislocations have been identified by a one-to-one correspondence with a known dislocation etchant.[14,15] Dislocations are roughly the same size and have a comma shape. The comma shape is due to the dislocation intersecting the surface of the sample at an oblique angle. Inclusions whose contrast is due to nongeneration of CL radiation have random shapes and sizes. The improved collection efficiency and measurement geometry of TCL in comparison to CL allows defects to be readily imaged in many more materials. The materials to which TCL has been applied is listed in Table II.

II. EXPERIMENTAL

2.1 Transmission cathodoluminescence technique

A schematic of the TCL technique is shown in Fig. 2. Transmission cathodoluminescence is accomplished by placing a light detector underneath a sample. The CL radiation generated on the top surface provides illumination which is transmitted through the sample and detected on the opposite side. Both sides of the sample must be optically smooth to reduce surface artifacts in TCL images. The TCL method has an efficient light collection geometry and no modification of the SEM is required. On the other hand, additional optics are required in the CL technique to guide the CL to an external detector. In Fig. 2, surface defects with their lower CL efficiency are detected as a decrease in CL radiation transmitted through the sample. Volume defects, i.e. defects which lie beneath the electron-hole (e-h) excitation

Table II—Different substrates and device wafers examined by TCL

	GaAs/GaAlAs	InP/InGaAsP
Impurity striations	GaAs:Te GaAs:Si	InP:S InP:Se InP:Te InP:Sn InP:Zn
Dislocations; defects	GaAs:Te GaAs:Si GaAlAs:Si LED wafer GaAs/GaAlAs LED wafer GaAs/GaAlAs laser wafer	InP:S InP:Se InP:Te InP:Sn InP:Zn InGaAsP InP/InGaAsP LED wafer InP/InGaAsP laser wafer
Degradation	GaAlAs:Si optoisolator LED GaAlAs data link LED	InP/InGaAsP transmission LED

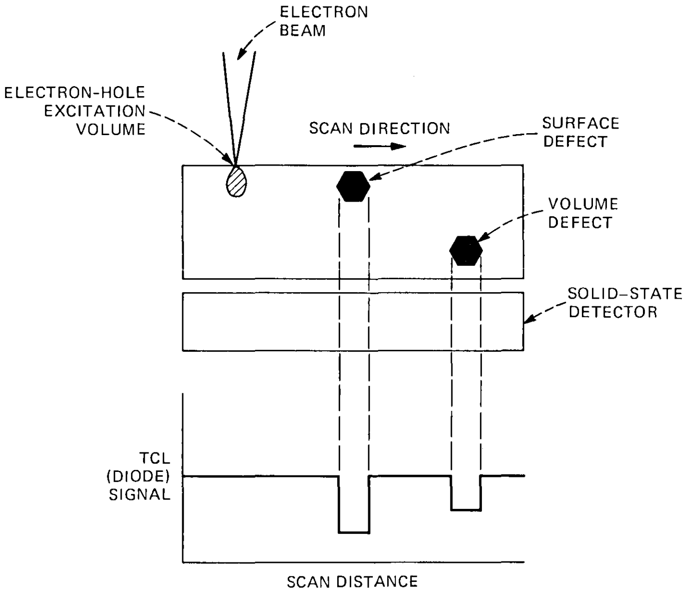


Fig. 2—Schematic of TCL measurement. Both surface and volume defects are detected by the solid-state detector as a decrease in luminescent radiation. The surface defect has a lower luminescing efficiency and the volume defect shadows the dector from the luminescing surface.

volume of the electron beam, are detected by their interaction with the surface-generated CL radiation passing through the sample to the detector below. A void in the sample volume will increase light transmission because of reduced absorption. A volume defect will decrease light transmission because of increased absorption or scattering. Using this method, defects deep within a thick sample can be observed. Although the transparency of a sample to its own CL radiation is material dependent, in the following we give TCL data on samples whose thickness is comparable to that of substrates and device wafers.

2.2 Experimental consideration

The data given was obtained with an ETEC Corporation autoscan SEM, but any SEM with similar capabilities may be used.[46] Two solid-state detectors were used for the TCL measurements, depending on the spectral region of interest. For wavelengths less than ~1.0 μm, a silicon photodiode was used; for wavelengths shorter than ~1.6 μm, a germanium photodiode was selected. These detectors have low noise equivalent power (NEP), fast response at zero bias, and output linearity with light input. The performance specifications of the photodiodes are listed in Table III.

For typical SEM parameters of 20 keV and 10^{-7}A for the electron

Table III—Characteristics of photodiodes used in TCL experiments

	Silicon Photodiode	Germanium Photodiode
Active area	0.20 cm^2	0.20 cm^2
NEP	8×10^{-13} W/$\sqrt{\text{Hz}}$ @ 850 nm	6×10^{-11} W/$\sqrt{\text{Hz}}$ @ 1300 nm
Responsivity	0.35 A/W @ 850 nm	0.22 A/W @ 1300 nm

beam, the luminescent power radiated by the sample is approximately 10^{-5}W, assuming a CL conversion efficiency of 10^{-2}. This conversion efficiency is typical of GaAs but is lower for other materials.[47] Most of this radiation is collected by the detector in the TCL configuration since it is placed 2–3 mm from the luminescing source. The solid angle subtended by the photodiode in the TCL geometry is $\sim\pi$, whereas the collection angle in the usual CL mode is $\sim 10^{-2}\pi$.

Since the silicon and germanium diodes have an NEP of 10^{-12} W/$\sqrt{\text{Hz}}$ and 10^{-10} W/$\sqrt{\text{Hz}}$, respectively, the detectors are capable of sensing 10^{-10}W and 10^{-8}W, respectively, of TCL power at a bandwidth of 10^5 Hz required for SEM imaging. Approximately two orders of magnitude greater power is required to easily image material defects without the use of difficult noise reduction techniques such as lock-in amplification coupled with beam blanking. This power is readily available from the materials listed in Table II.

2.3 Resolution

The resolution of the TCL technique is limited by the type of defect observed. The contrast from dislocations is due to nonradiative recombination; therefore, it is limited by the minority carrier diffusion length. The contrast at inclusions is due mainly to nongeneration of CL, and the resolution of these defects is limited by the electron beam conditions as in SE imaging. Finally, volume defects can be resolved to roughly the CL emission wavelength since the defect is effectively observed by optical means.

Figure 3 shows an example of two different defects appearing in the same InP:S sample. The large dark spots are due to dislocations intersecting the surface, whereas the much smaller dark spots are inclusions located within an e-h excitation volume of the surface. The identification of defects using TCL will be discussed in Section 2.4. The dislocation image is large because of the large minority carrier diffusion length. By imaging these dislocations, the minority carrier diffusion length is readily estimated in this sample to be ~5 μm. The images of inclusions show that defects as small as 1–2 μm can be imaged under the beam conditions of 20 keV and 10^{-7}A.

Figure 4 is an example of a volume defect imaged by TCL. Figure 4a is the SE image of an unusual feature appearing in an InP/InGaAsP

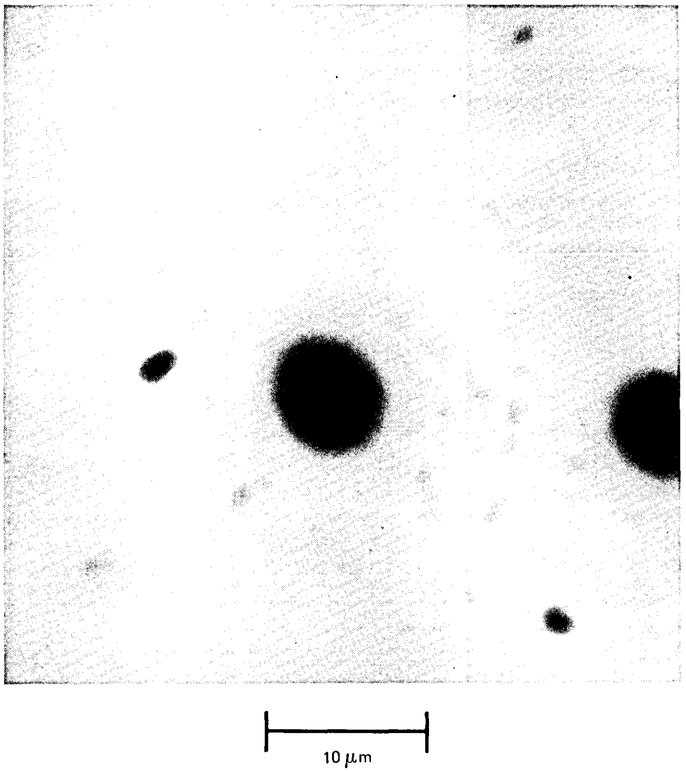


Fig. 3—Transmission cathodoluminescence image of heavily S-doped InP substrate. The large and small dark spots are dislocations and inclusions, respectively.

wafer. A CL picture of the same region, Fig. 4b, shows the area to be nonluminescent. It is only the TCL image, Fig. 4c, that explains the origin of this dark area. One can clearly see a bright spot in the center of the dark region associated with a void in the epitaxial layers. Since the void does not appear in the CL image, it lies below the e-h excitation volume.

2.4 Interpretation of TCL images

In general, defects appear as dark spots in a TCL image. To interpret images, dark spots must be identified. For the CL technique, dark spots in images of GaAs:Te, GaP, and CdTe were identified to be dislocations by demonstrating an exact one-to-one correspondence between these dark spots and dislocations revealed as etch pits produced by a known defect etchant.[6,23,34,35] We have followed this procedure for the TCL technique and applied it to GaAlAs:Si, GaAs:Te, InP:Te, and InP: S.[14,15]

Two methods were used to show that dark spots observed in a TCL

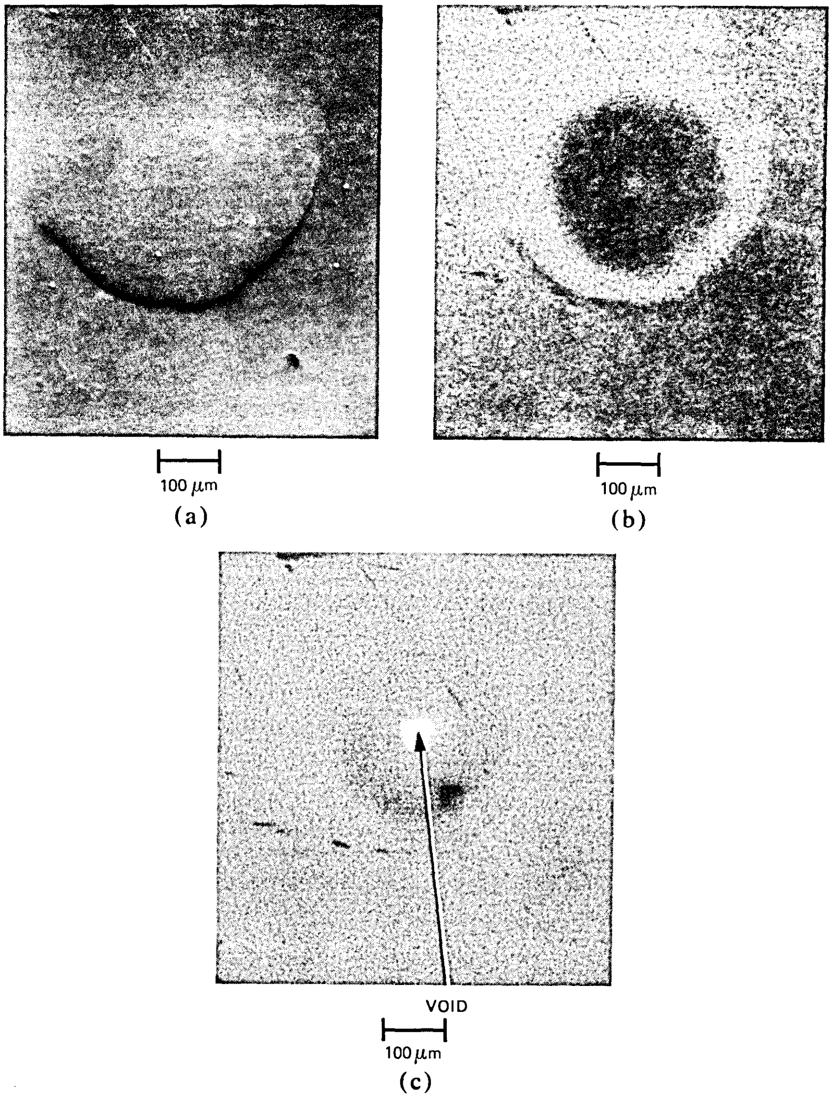


Fig. 4—(a) Secondary electron image of a surface feature on an InP/InGaAsP wafer. (b) Cathodoluminescence image of the region shown in (a). The dark circular region shows that the surface feature is nonluminescent. (c) Transmission cathodoluminescence image of the region shown in (a). The bright dot in the center of the dark area is a void.

image represent grown-in dislocations. In the first method, a TCL image of a sample was taken prior to etching to reveal dislocations. By comparison of the etch pattern with the TCL image of the same region, an exact one-to-one correspondence was established. In the second method, a {100} wafer was etched to reveal dislocations as etch pits.

Subsequently, the samples were cleaved along the {110} cleavage planes. By examining with TCL the {110} surface where it intersects an etch pit, a dark region immediately below each pit was found. Thus, each dislocation intersecting the {100} surface corresponds to a dark spot in a TCL image. The technique of first etching the surface and then showing dark spots at dislocation pits is not useful since surface features always appear in either a CL or TCL image.

In Fig. 5a, a TCL image of a Si-doped $Ga_{1-x}Al_xAs$ layer grown by liquid phase epitaxy (LPE) on a {100} oriented GaAs substrate is shown.[48] The aluminum concentration varies from $x = 0.00$ on the top surface to $x = 0.30$ at the substrate-epitaxial layer interface. This gradient occurs over a thickness of ~250 μm. The sample has been bromine-methanol polished to produce smooth, damage-free surfaces. Only dust particles are observed on the top GaAlAs surface when examined by SE imaging and no features could be seen in normal CL using an S1 photomultiplier. This observation is in agreement with previous measurements on similar material.[7] However, when examined by TCL, a number of randomly distributed dark spots appear. The size of these dark spots varies from 5 to 10 μm. At high magnification, their shape closely resembles commas; that is, the central dark region is connected with a fainter tail. The average density of these dark spots is approximately 2×10^4 cm^{-2}.

Figure 5b shows the SE image of the region shown in Fig. 5a. The

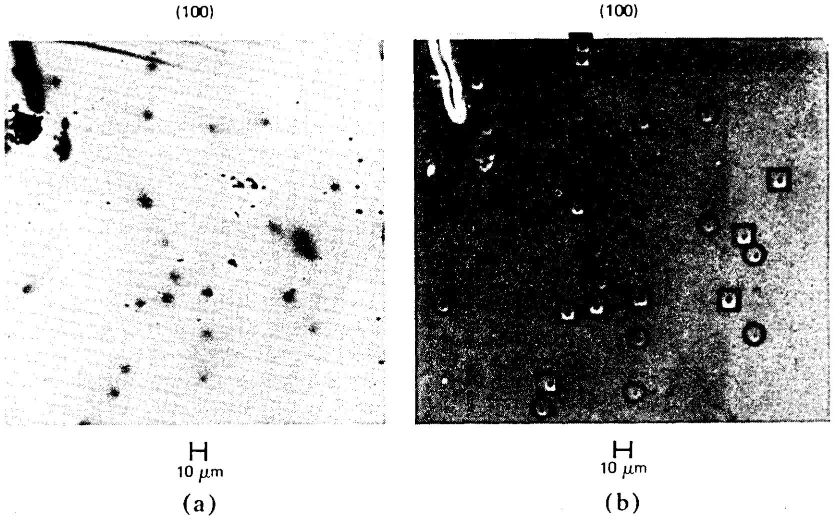


Fig. 5—Transmission cathodoluminescence image of a GaAlAs:Si epitaxial layer grown on {100} GaAs substrate. The dislocations appear as dark spots approximately 10 μm in diameter. (b) Secondary electron image of the area shown in (a) after 30 min in 300°C KOH etchant. The circles surround the small (1- to 2-μm diameter) dislocation pits. The squares surround the large (5 to 10 μm) dislocation pits.

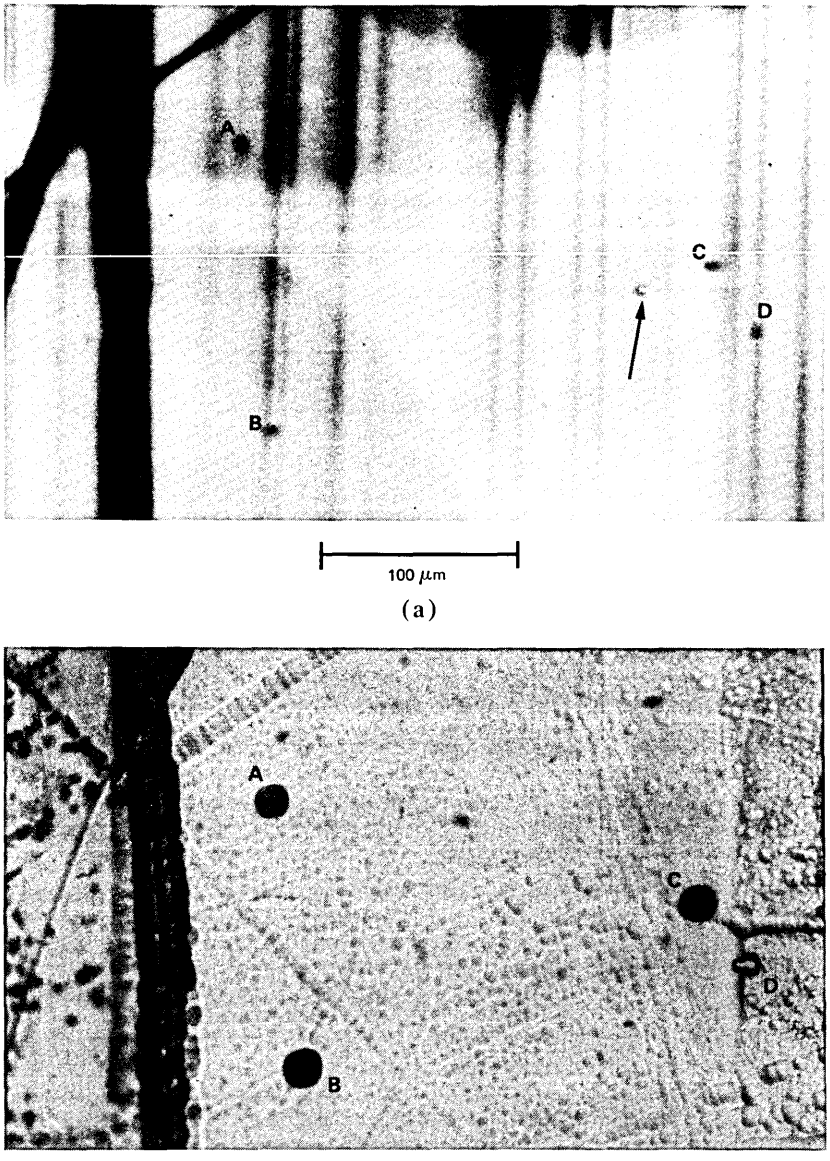
sample has been etched in molten KOH at 300°C for 30 min to reveal dislocations. The dislocation pits in {100} GaAlAs, as revealed by this etchant, are hexagonal in shape and terminate in a central point.[36] Two distinct dislocation pit sizes are observed independent of etching time. The pit size has been related to the dislocation inclination.[24] A careful comparison of Figs. 5a and 5b shows a one-to-one correspondence between the large (5- to 10-μm diameter) dislocation pits surrounded by a square in the figure and the TCL dark spots. In addition, apparently similar TCL spots also correspond to the small (1- to 2-μm diameter) dislocation pits surrounded by a circle.

Figure 6a is a TCL image of a Te-doped InP crystal. The vertical unevenly spaced lines are the impurity-induced growth striations. Their contrast is due to the dependence of the CL efficiency on doping density and not due to unresolved defects.[49] The dark band on the left of the figure is a scratch intentionally introduced to specify the location. Neither the striations nor the dark spots indicated by letters A through D are apparent on either a CL or an SE image. At higher magnifications, the dark spots have comma shapes as was found for the GaAlAs:Si sample in Fig. 5. The feature indicated by an arrow is a dust particle which appears also in the SE image.

Figure 6b is a Nomarski optical photograph of the region shown in Fig. 6a after Huber etching to reveal dislocations. The magnification of Figs. 6a and 6b are slightly different. The four dislocation pits are labeled A through D for their corresponding TCL dark spot; an exact correspondence between dark spots and dislocation pits is found by comparing Figs. 6a and 6b. As compared to the freshly etched surface of Fig. 1a, Huber etching followed by TCL scanning produces a fairly poor surface. In addition, the growth striations are not clearly revealed. It is believed that carbon deposition during the TCL examination interferes with the etching process. The carbon is due to electron-beam decomposition of residual organics within the sample chamber. Etch pit D is somewhat obscure because the carbon film deposited over that area was twice as thick, a result of overlapping scan regions.

Figure 7a shows an SE image of the cleaved {110} surface of an InP: S sample after Huber etching and cleaving. Several dislocation pits produced on the {100} surface are intersected. A TCL scan of the corresponding area is shown in Fig. 7b. A dark region extending into the sample beneath each pit is observed, indicating that the pit corresponds to a dark spot on a TCL image of the {100} surface. Since the dark regions do not extend further into the crystal, the observed images could be due to portions of dislocation loops.

The above examples of correspondence between etch features and TCL images brings out several advantages of TCL over etching. First, TCL can be applied to any crystal face, whereas defect etchants are



(a)

(b)

Fig. 6—(a) Transmission cathodoluminescence image of {100} surface of InP:Te. (b) Secondary electron image of the area shown in (a) after Huber etching to reveal dislocations. Corresponding features in (a) and (b) are labeled A through D.

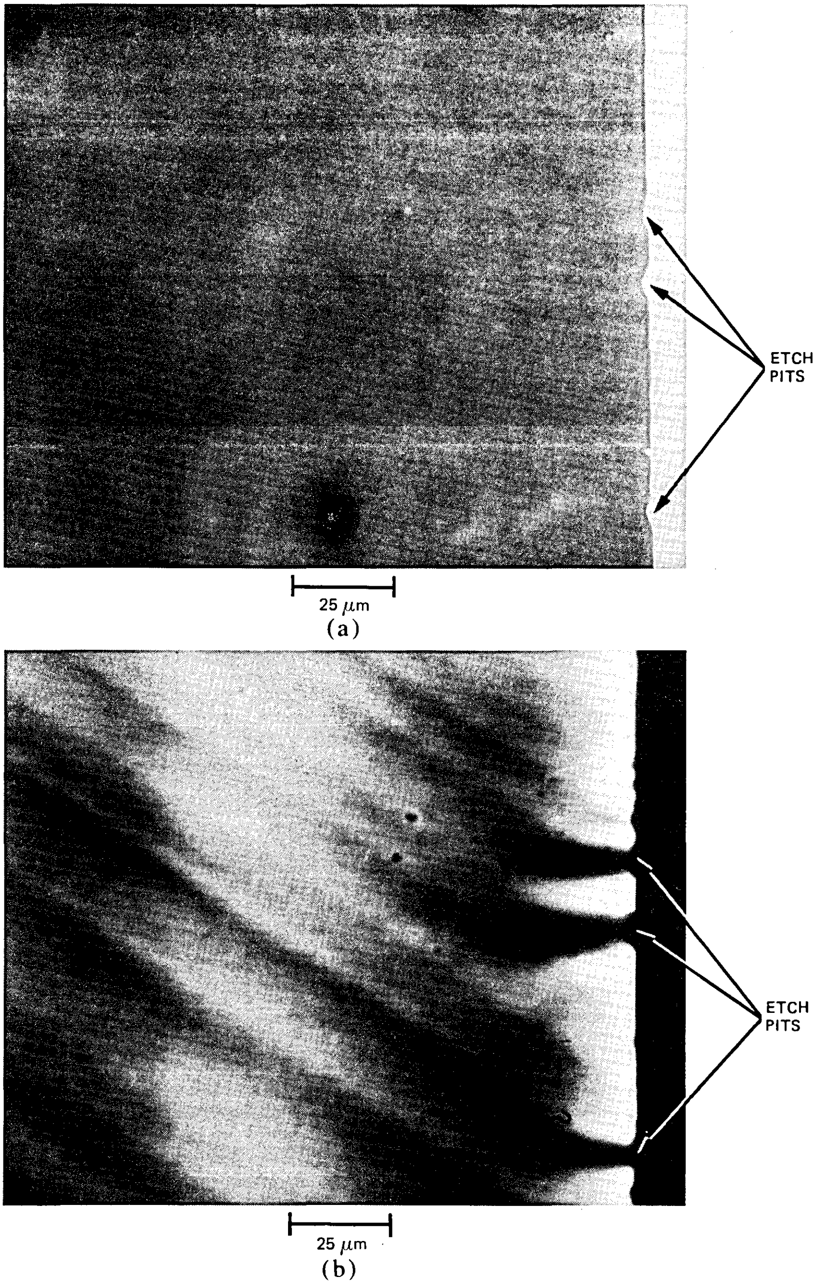


Fig. 7—(a) Secondary electron image of {110} surface of InP:S. Several etch pits on the {100} surface produced by Huber etching are intersected on the right-hand edge of the sample. (b) Transmission cathodoluminescence image of area shown in (a) displaying a dark region below each etch pit.

extremely orientation sensitive. Neither the KOH nor the Huber etch reveals defects on the {110} surfaces used as light-emission facets for lasers. Secondly, a defect etchant must be developed for each semiconductor, whereas TCL requires, at most, a change in detectors to accommodate the spectral range. Additionally, the development of a defect etchant is difficult and time consuming. Finally, etching is extremely sensitive to the surface preparation. The etchant may not work properly if the surface has been contaminated.

III. SUBSTRATES/EPILAYERS

The common substrates used for optoelectronic devices are GaP, GaAs, InP, and GaSb. Since good photodetectors at 1.6 μm are unavailable, dislocations in GaSb have not been imaged by any luminescence technique. In the case of GaP, dislocations appear readily under examination by CL or TCL. Using CL, however, dislocations in GaAs are readily imaged only for the case of Te doping; dislocations in GaAs:Se are difficult to image and have not been imaged in GaAs:Si.[27] Finally, dislocations in InP have not been observed for any dopant with CL scanning. However, using TCL, dislocations are easily observed for the commonly available GaAs and InP substrates.[10,14,15] These materials are listed in Table II. We will demonstrate the usefulness of the TCL technique on selected GaAs and InP substrates.

3.1 GaAs substrates

Figures 8a and 8b are, respectively, the SE and TCL images of a 375-μm-thick GaAs:Si substrate commonly used for the growth of LED and laser wafers. The white spot in the upper right of Fig. 8a is a dirt particle that appears dark in the TCL image; the dirt particle effectively blocks the electron beam from the semiconductor and prevents CL generation. The remaining comma-shaped dark spots with no corresponding features in the SE image are due to dislocations intersecting the surface at an oblique angle; the direction of the tail indicates the sense of inclination. This TCL image is typical of GaAs substrates grown by the gradient freeze technique; growth striations show weak contrast and are rarely observed. The dislocation image is unchanged when the dopant is changed from silicon to tellurium, another standard dopant for GaAs substrates.

3.2 InP substrates/epilayers

The recent work of Seki et al. has demonstrated that the macroscopic perfection of InP crystals grown by the liquid encapsulated Czochralski (LEC) technique can be improved by heavy doping.[50,51] The order of effectiveness in reducing the dislocation density is Zn, S, and Te. It is of considerable technological interest to assess whether or not

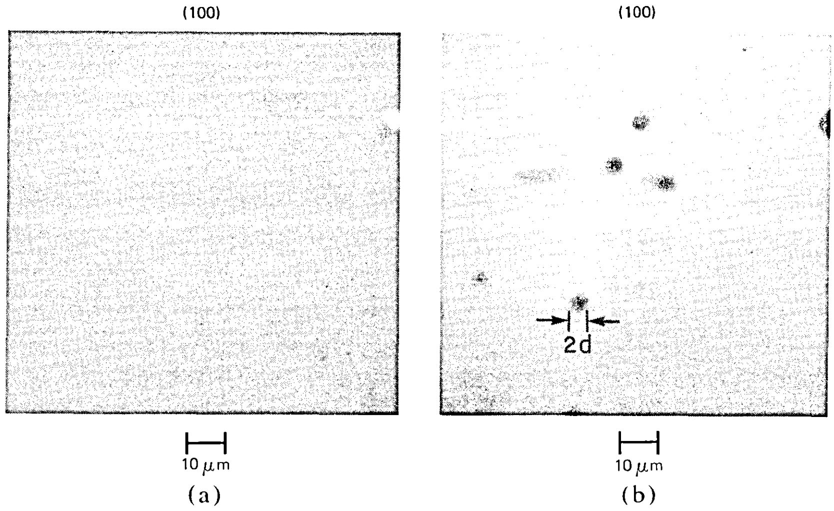


Fig. 8—(a) Secondary electron image of the {100} surface of a GaAs:Si substrate. The white spot on the upper right is a dirt particle. (b) Transmission cathodoluminescence image of the area shown in (a) displaying four dislocations. The diameter of the dislocation image is approximately twice the minority carrier diffusion length.

the macroscopic perfection of such highly doped substrates can be replicated into isoepitaxial layers grown by LPE. This is especially important since it was shown by Mahajan et al. that although heavy zinc doping reduces the dislocation density, a high density of Zn related precipitates results.[20] Similar effects have also been found for heavy Ge doping of InP.[52] Although individual precipitates, which are ~675 Å in dimension, cannot be resolved using TCL, precipitate clusters can be imaged.[20] Transmission cathodoluminescence imaging in this example is important since neither the precipitates nor their clusters show enough strain contrast to be observed either in X-ray topography or in Huber etching.[20] TEM is required for a detailed study, but TCL is useful for a simple, macroscopic investigation.

To facilitate the location of corresponding areas of substrate and epilayers in the TCL study,[23] a pattern was produced on one side of the wafers by evaporating gold through a shadow mask. This pattern may be observed irrespective of which side is excited with the electron beam. A similar CL or EBIC study would not be easy since corresponding areas of a wafer on opposite sides are difficult to locate in these techniques. In the TCL study, each side of the wafers is examined with the electron beam since maximum contrast and resolution of defects are obtained when the defects lie within the e-h excitation volume.

As ascertained by etch pitting, the quality of epilayers grown on the Zn-doped substrates was satisfactory and compared well with that of the substrate.[23] Occasionally, etch pits in the form of clusters were observed. The TCL study, however, revealed additional features within the epilayers. These results are presented in Fig. 9. Figure 9a is the SE image of the 10-μm-thick InP:Sn epitaxial layer. This epilayer is a typical buffer layer used in device wafers. The small white features are indium-rich beads caused by localized surface decomposition. The beads appear as dark spots in the TCL image of Fig. 9b. In addition, however, a high density of light grey circular areas, whose origin at

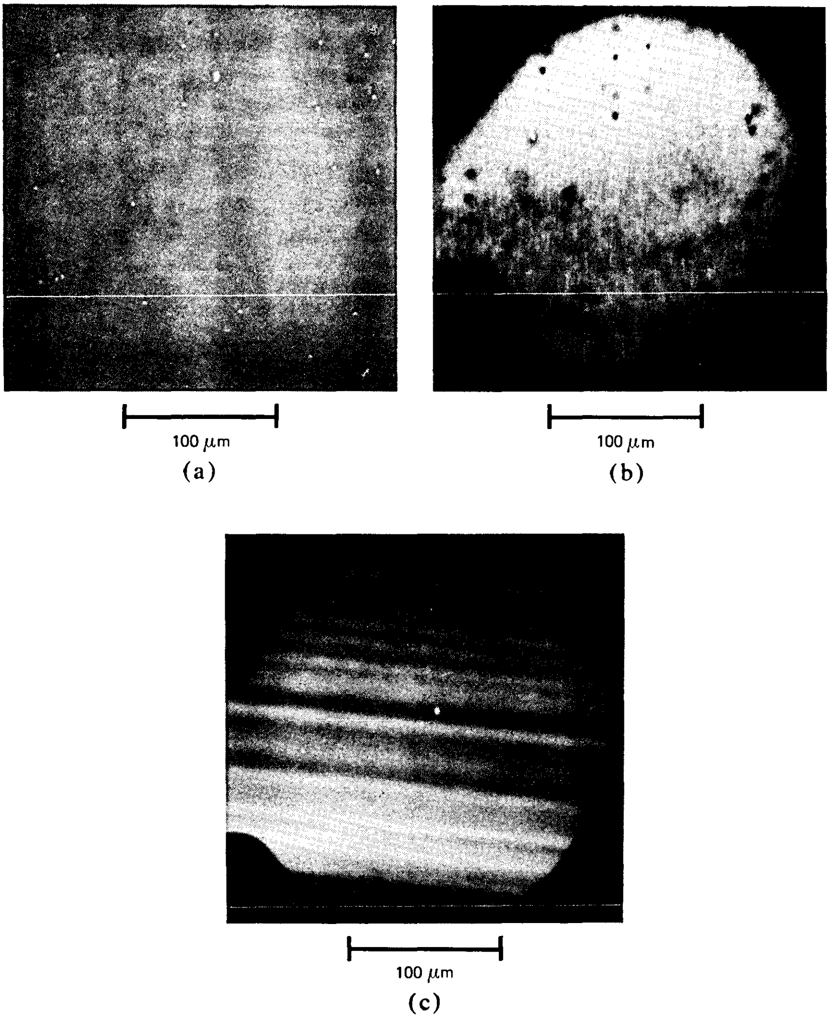


Fig. 9—(a) Secondary electron image of a 10-μm-thick InP:Sn epitaxial layer. (b) Transmission cathodoluminescence image of region shown in (a). (c) Transmission cathodoluminescence image of the underlying InP:Zn substrate.

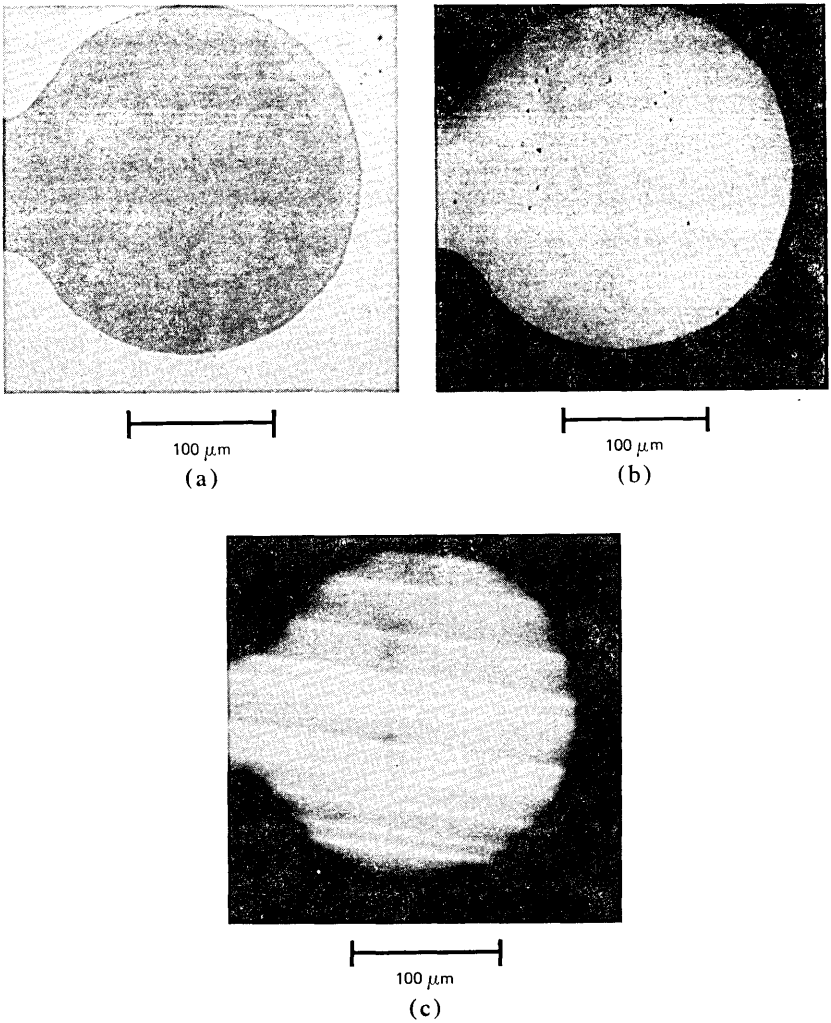


Fig. 10—(a) Secondary electron image of a 10-μm-thick InP:Sn epitaxial layer. (b) Transmission cathodoluminescence image of region shown in (a). (c) Transmission cathodoluminescence image of the underlying InP:S substrate.

present is uncertain, is apparent. These defects are likely related to the precipitates within the substrate, although corresponding features are not present in the substrate of Fig. 9c. Only growth striations and two defects are observed in the substrate. The dark border in Figs. 9b and 9c is due to the gold mask evaporated on the substrate side. A mirror inversion about a horizontal axis is required to exactly compare the image of the substrate with that of the epilayer.

Figure 10 shows images of a 10-μm-thick InP:Sn layer grown on a heavily S-doped substrate. Figure 10a shows the SE image of the epilayer surface. The indium-rich beads on the surface are again evident. Figure 10b demonstrates that this portion of the epilayer is dislocation free and correlates very well with the quality of the substrate as illustrated in Fig. 10c. Nearly horizontal light and dark bands are growth striations, and darker regions on some of these bands may represent local dopant variations.

As a result of these studies, it was confirmed that macroscopically dislocation free InP may be obtained by heavy Zn or S doping. However, defect-free isoepitaxial layers may be grown only on the S-doped substrates. The defects in the epilayers on InP:Zn are probably a result of precipitates within the substrate. These precipitates have been observed by TEM in low-dislocation InP:Zn but not InP:S. Therefore, it is recommended that InP:S crystals having low-dislocation density be used as substrates for the growth of device wafers.

3.3 InGaAsP epitaxial layers

For InGaAsP layers whose bandgap corresponds to 1.3 μm, the Si photodetector normally used to detect transmitted CL was replaced with the Ge photodiode. Although CL radiation transmitted through the sample was detected by the Si diode in all cases, defects appeared for most of the InGaAsP samples studied only in the TCL image obtained with the Ge detector. Apparently, for these samples, the defects affected the CL emission and optical properties in the spectral region beyond the Si bandgap, i.e. wavelengths longer than 0.9 μm.

Figure 11 shows an SE and TCL image of a 2-μm-thick n-type, nominally undoped, InGaAsP epilayer grown on a 15-mil-thick, {111} oriented InP substrate. This TCL image was obtained using the Ge photodiode. The InP substrate does not contribute to the TCL image since it is transparent to the InGaAsP CL centered at 1.3 μm. Dark lines correspond to surface scratches or cracks seen in an SE image. The lower region of Fig. 11b is darker than the upper region because of a slight nonuniformity in the CL or optical properties of the quaternary layer. The dark spots, approximately 2 μm in diameter and similar in shape, have no corresponding surface features. By analogy with the TCL study of dislocations in GaAs and InP, it is suggested that these spots are caused by dislocations in the epitaxial layer. Additionally, it has been shown by etching studies that dislocations in the substrate are sources of dislocations in quarternary layers. The average density of dark spots imaged by TCL, 2–3 $\times$ 10^4 cm^{-2}, is approximately the same as that of the InP substrate. Finally, the minority carrier diffusion length of InGaAsP has been measured to be 1 to 2 μm in agreement with the diameter of the dark spots.[53]

It should be noted that the Ge detector is unsuitable for TCL imaging

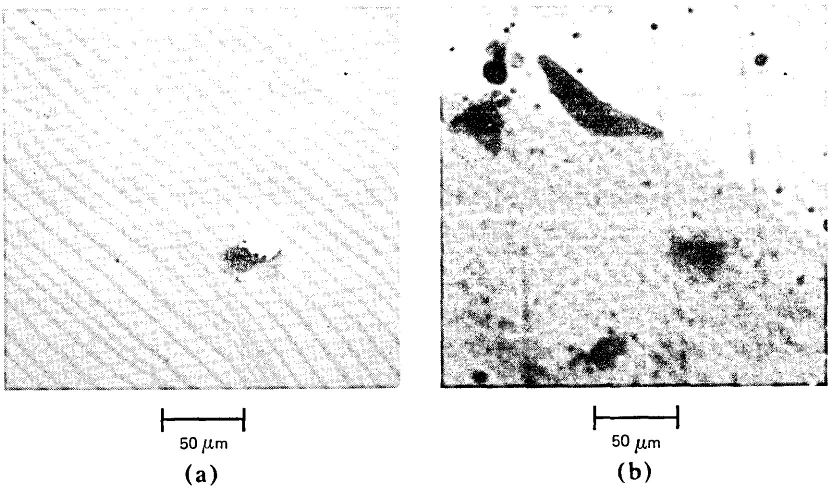


Fig. 11—(a) Secondary electron image of a 2-μm-thick n-type InGaAsP layer grown on a 15-mil-thick {111} oriented InP substrate. (b) Transmission cathodoluminescence image of region shown in (a). The dark spots not apparent in the SE image are dislocations.

of p-type InGaAsP, which has at least an order of magnitude lower efficiency than an n-type material. Imaging of p-type InGaAsP may be possible using lock-in amplification coupled with beam blanking of the electron beam.

IV. DEVICE WAFERS

The TCL imaging of device wafers differs from that of substrates because of the presence of a p-n junction and heterojunctions within the wafer. The built-in electric field at the p-n junction (depletion region) separates the generated electrons and holes and prevents recombination.[54] Thus, in the CL, EBIC, and TCL modes, the depletion region is probed indirectly. In either the CL or TCL mode, the image is complicated by this depletion width if the p-n junction intersects the e-h excitation volume. For the plane view configuration where the electron beam is orthogonal to the plane of the p-n junction, CL radiation is generated in the layers confining the p-n junction, i.e. the layers above and below the p-n junction. Thus, the image is composed of three components: the CL image of the two confining layers and variations in the depletion width. Electron beam-induced current scanning, however, is primarily composed of variations in the depletion width and comparison of EBIC images with CL and TCL images may be used to locate the position of defects within the wafer.

In this section, we will demonstrate the usefulness of TCL in evaluating the quality of device wafers used in fabricating LEDs and lasers based on both the GaAs/GaAlAs and InP/InGaAsP material systems. Presently, TCL is used to screen out GaAs/GaAlAs wafers that produce unreliable LEDs; highly reliable LEDs are incorporated as light sources in the Western Electric (WE) 1250A optical transmitter. Additionally, TCL is used to screen out GaAs/GaAlAs laser wafers with a high density of rake lines without the removal of the contacting layer. Finally, TCL is used to screen out InP/InGaAsP LED and laser wafers with misfit dislocations.

4.1 GaAlAs LEDs

High radiance GaAs/GaAlAs LEDs and lasers operating at a current density of $J \gtrsim 10^3$ A/cm^2 can degrade rapidly by the development of dark line defects (DLDs).[8,31,55] Since these optoelectronic devices are intended for optical communication systems, a high yield of reliable devices is necessary. By detecting defects which act as sources for DLDs during processing, wafers with an unacceptable defect density may be rejected early and, thus, save processing time and cost.

A preliminary study involving single heterostructure LEDs shows that approximately 12 percent of the devices fail a stress test (burn-in) because of the growth of DLDs, whereas the remaining LEDs have a lifetime in excess of 10^6 hours.[56] Since DLDs in GaAs/GaAlAs-based devices are observed to originate at material defects, a high yield of long-lived devices may not be consistently achieved due to inevitable variations in substrate quality, epitaxial growth conditions, and processing procedures. A nondestructive technique such as TCL selects device quality material, evaluates growth and processing procedures, and reveals the degradation mechanism of fabricated devices.

The schematic of the GaAlAs LED used in the WE 1250A transmitter and the epitaxial structure is shown in Fig. 12. The growth of the GaAlAs layers by LPE and the device processing have been described elsewhere.[57] The 50-μm-thick n-type GaAlAs carrier confinement layer also serves as the window and mechanical support layer. This planar

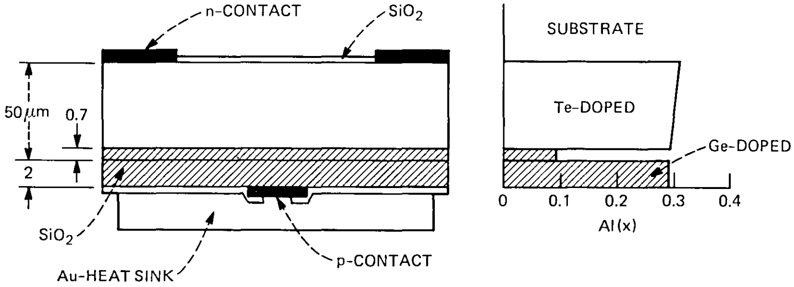


Fig. 12—Schematic of planar DH LED and device wafer. The shaded region of the device is Ge-doped.

structure facilitates device processing but the epitaxial growth is difficult. Because of the thickness of the first-to-grow window layer, the aluminum concentration gradient is large and nonequilibrium growth conditions must be controlled in forming the active and final confinement layers.

The device wafers are examined with the epilayers in the as-grown conditions. The substrate side of the wafers, normally rough polished, is bromine-methanol polished after epigrowth to reduce scattering of the TCL radiation. The electron beam of the SEM is scanned over the episide of the wafers since the quality of the epimaterial is of main interest. The typical 375-μm-thick GaAs substrate does not have sufficient absorption to reduce the TCL intensity below the limit required for good imaging using the silicon photodiode. This is in contrast to the CL or PL scan geometry where a thin (1 to 2 μm) GaAs contact layer is sufficient to reduce the signal below the detection limit.[30] No processing problems because of TCL examination of device wafers have been found; the wafers must be soaked in concentrated sulfuric acid to remove the electron beam deposited carbon prior to standard processing.

Figure 13 is a TCL image of a wafer that produced a high yield of reliable devices. Almost no dark spots corresponding to defects are observed; the faint light and dark bands are the growth lamellae normally observed in wafers grown by LPE. Figure 14a is a TCL image of an LED wafer grown under nonequilibrium conditions. Defects in the substrate are not imaged since the CL efficiency of GaAlAs is much higher than that of GaAs. The epilayers were grown on low-dislocation ($<10^3$ cm^{-2}) Si-doped substrates. However, as shown in Fig. 14b, an order of magnitude higher defect density ($\gtrsim 10^4$ cm^{-2}) is found in the TCL image of the epilayers. These defects cannot be observed using Nomarski contrast microscopy and are, thus, “hidden.” An etching study using KOH was attempted to find the nature and origin of the defects. The etchant study was performed carefully so as not to remove the 2-μm-thick, last-to-grow layer. The etch pattern obtained did not distinctly reveal dislocation pits but resembled the TCL image. A similar etch study of a grooved sample showed these defects to be localized to the active and last-to-grow layers.

A TEM evaluation of the sample shown in Fig. 14a ascertained that the defects are dislocation tangles located within the epilayer. Figure 14b displays one of the dislocation tangles. The bar indicates 1 μm. A comparison of Figs. 14a and 14b shows that the TCL defect image is roughly 10 to 20 times their actual size, a result of the large minority carrier diffusion length. It is apparent from Fig. 14b that the dislocations are very closely spaced contiguous to the center, whereas interdislocation spacing increases when going toward the edges. Further-

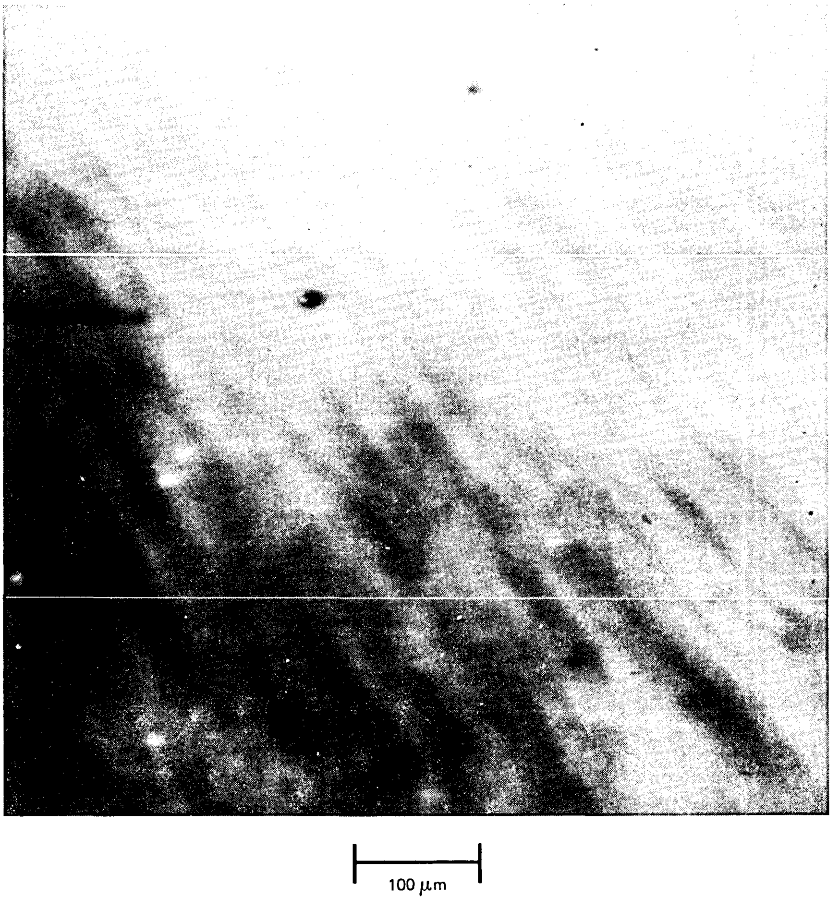


Fig. 13—Transmission cathodoluminescence image of a planar DH LED wafer-producing reliable LEDs.

more, stereomicroscopy reveals that the dense dislocation networks are essentially planar in character and are located within the active and last-to-grow layers. Since the networks appear to intersect along the [110] direction, their habit planes may be $(\bar{1}11)$ and $(1\bar{1}1)$ planes. The topology observed by stereomicroscopy is consistent with this suggestion.

A plausible explanation for the origin of dislocation networks is the following. It is suggested that at the temperature at which the growth of the thick window layer was completed, the active layer melt was supercooled in excess of 5°C. Faceted growth on {111} planes is initiated locally by this supercooling. Since the compositions of layers growing on the (001), $(\bar{1}11)$, and $(1\bar{1}1)$ planes are likely to be different, accommodation dislocation networks will be generated at the conflu-

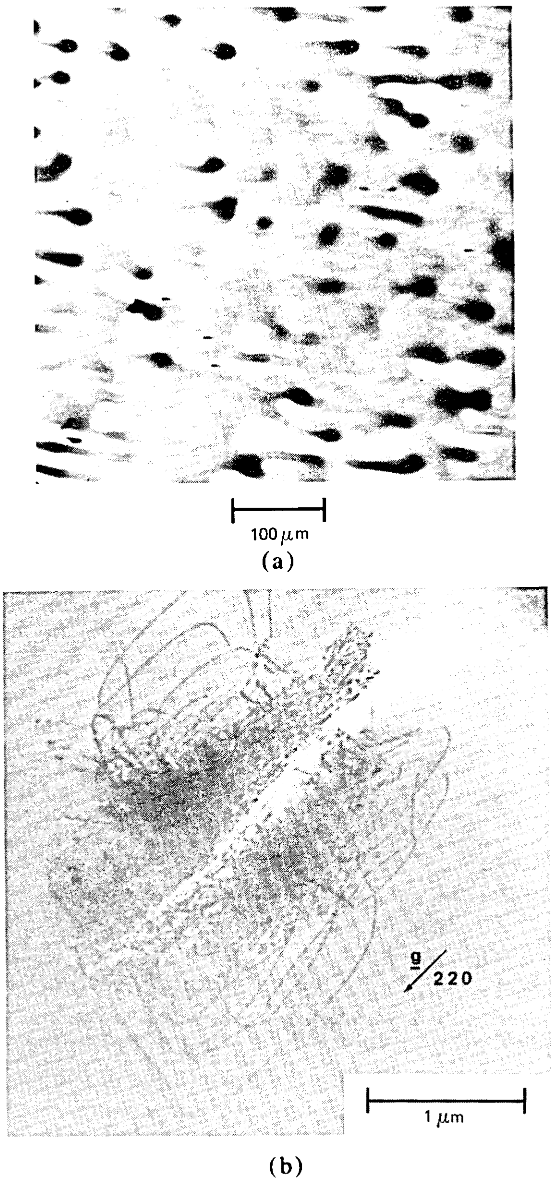


Fig. 14—(a) Transmission cathodoluminescence image of a planar DH LED wafer grown under nonequilibrium conditions. The dark areas are dislocation tangles. (b) Transmission electron microscopy image of an individual dislocation cluster in the samples shown in (a).

ence of these different growth fronts. This growth problem encountered in the early stages of the planar double heterostructure (DH) LED development, results in a low yield of reliable devices. After adjusting the composition of the melts so that growth of the active and last-to-grow layers is under near-equilibrium conditions, dislocation tangles in the device wafers were no longer observed.

4.2 GaAs/GaAlAs lasers

The use of TCL as a nondestructive screening technique for defects in GaAlAs DH laser material has been recently reported by Gaw and Reynolds.[18] Every wafer is examined with TCL prior to processing. Wafers with a high density of rake lines are rejected since devices with this defect either do not lase or have an excessive threshold current. A nondestructive screening technique for this growth defect is especially important since ~45 percent of all material grown exhibits rake lines.[18]

The LPE laser material was prepared for the 0.83-μm lasers used in the FT-3 transmitter subsystem.[18] It has the usual four-layer DH with a Te-doped n-ternary confinement layer. Starting from the n-GaAs substrate, the layers have the compositions (i) N-GaAlAs ($x = 0.34$ to 0.44), (ii) P-GaAlAs ($x = 0.00$ to 0.10), (iii) P-GaAlAs ($x = 0.34$ to 0.44), and p-GaAs. The thickness of the p-active layer is 0.10 to 0.20 μm and the total thickness of the p-layers is ~1.8 μm. The p-GaAs is used for a contact layer. This contact layer prevents wafer scanning by either CL or PL because of absorption within this layer. Photocurrent scan may be used if contacts are applied, but TCL has a higher resolution in addition to the advantages provided by the SEM.

The TCL imaging of laser wafers is performed as previously for the LED wafers. In these wafers the back surface is polished prior to growth and, thus, TCL involves no additional processing.

Figure 15 compares the TCL image of a slice having severe rake lines with the almost featureless SE image of the top surface. Wafers with such a TCL image are rejected from further processing. Figure 16a and 16b shows the SE and TCL image of a wafer free of rake lines. In the absence of gross defects, the dislocations and dark spot defects which have lower contrast may be observed. By correlating the TCL evaluation with broad area and stripe geometry laser yields, TCL was shown to be an effective nondestructive screening technique for rake lines.

4.3 InP/InGaAsP LED and lasers

In InP/InGaAsP heteroepitaxy, the possibility of lattice mismatch exists. Misfit dislocations at the substrate-epilayer interface form when lattice mismatch is in excess of 0.06 percent.[58] These misfit dislocations have a deleterious effect on device performance and reliability.[40,59] In

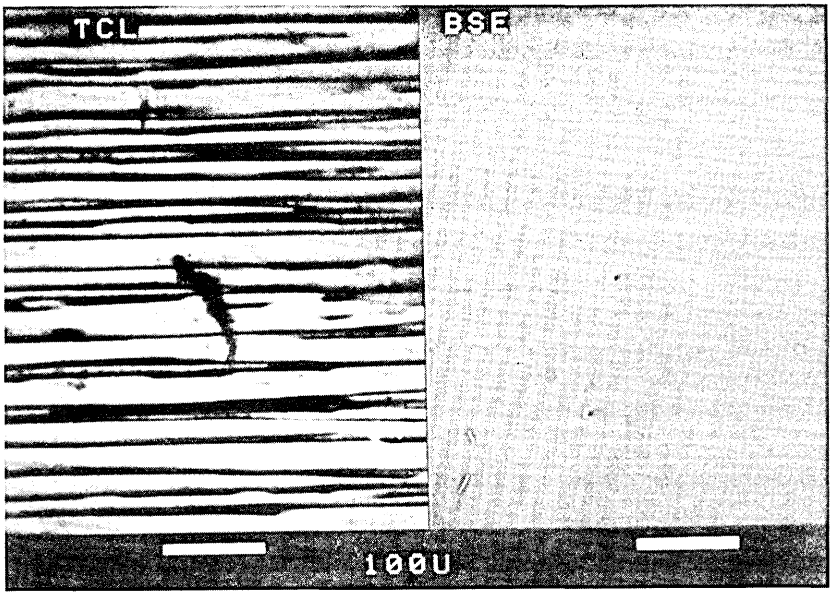


Fig. 15—Comparison of the TCL image of DH laser material having severe rake lines with the almost featureless SE image of the top surface of the slice.

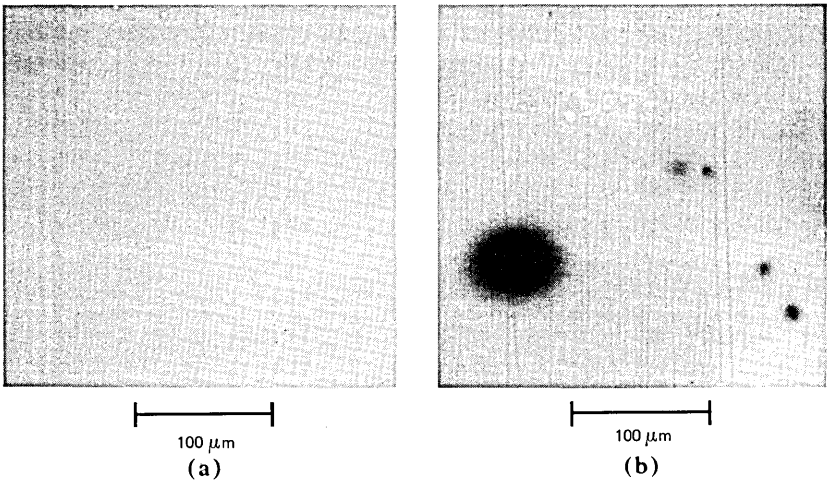


Fig. 16—(a) Secondary electron image of a DH laser wafer. (b) Transmission cathodoluminescence image of dislocations and dark spot defect in area shown in (a).

this section, we show that TCL can be used to screen out device wafers with misfit dislocations present.

Figure 17 presents the results obtained on a DH InP/InGaAsP/InP wafer intended for 1.3-μm devices. The wafer was lattice mismatched with $\Delta a/a = 0.12$ percent. The active layer is ~0.5 μm thick and may be used for either LEDs or lasers. In Fig. 17a, the SE image of the top epitaxial layer shows growth lamellae, hillocks, and surface scratches. The epilayers were grown on a {111} oriented substrate. Most of the

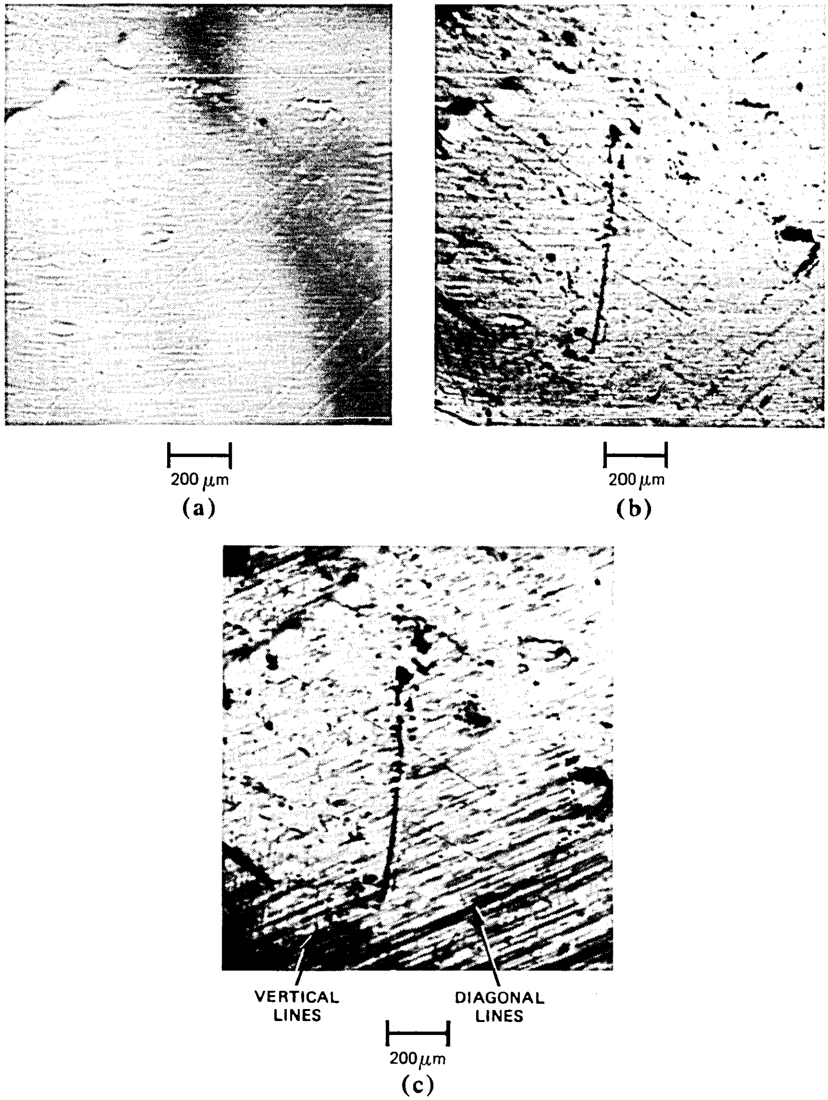


Fig. 17—(a) Secondary electron image of a DH InP/InGaAsP wafer. Growth lamellae, hillocks, and surface scratches are observed. (b) Cathodoluminescence image of the region shown in (a). The dark areas are surface defects. (c) Transmission cathodoluminescence image of the region shown in (a). The vertical and diagonal lines not seen in the CL image are misfit dislocations.

features of Fig. 17a are reproduced in Fig. 17b which shows a CL picture of the same region. The CL image was taken with an S1 photomultiplier which has a spectral response similar to that of an Si photodiode. Dark nonluminescing regions are clearly visible. Finally, in Fig. 17c, we show the TCL picture obtained with an Si detector. Additional features not observed in either SE or CL modes of the SEM are the large number of sharp diagonal and vertical dark lines which form an angle of ~60° between them. These lines are not within the electron beam range on the surface since they do not appear in the CL image. It is suggested that they are misfit dislocation at the InP buffer layer-InGaAsP active layer interface. Similar TCL imaging of lattice mismatched epilayers grown on {100} oriented substrates reveals two sets of orthogonal dark lines, providing additional support for this interpretation. Additionally, X-ray topographs of lattice mismatched {111} oriented wafers show images of misfit dislocations similar to those shown in Fig. 17c.[40,59]

V. DEGRADATION

Rapid degradation of optoelectronic devices has been due to defect growth. The overall device and especially the light-emitting region are often of extremely small dimension. Etching to reveal defects followed by visual examination is extremely difficult because of problems in sample handling and resolution. It is also difficult to perform defect analysis using the techniques of PL scan, infrared microscopy, and photocurrent scan for similar reasons. Although TEM analysis may be the only technique that can resolve defects, preparation of samples for TEM becomes an art. On the other hand, SEM techniques are both convenient and have adequate resolution and magnification in most cases. However, EBIC and CL analysis may be prohibited by the geometry of the device. The region of interest is often inaccessible to the electron beam and the sample must be demounted from the headers.

Figure 18 shows three device structures where EBIC and CL evaluation is difficult. Figure 18a shows either a planar LED or laser structure; the p-contact represents a dot for the LED and a stripe for the laser. Figure 18b shows an LED used for optoisolators. In the planar LED, the window layer is ~50–100 μm thick, thus, preventing the electron-beam excited carriers from reaching the active layer. In the laser structure, the device is usually mounted episide down for better heat-sinking. Even when the top *n*-metallization is partially removed to access the semiconductor, the thick substrate prevents the carriers from reaching the p-n junction. For the LED shown in Fig. 18b, the p-n junction is again located far from the excitation source. Thus, without demounting the devices, EBIC and CL provide little information about the p-n junction or active layer where the light is generated in electrolumi-

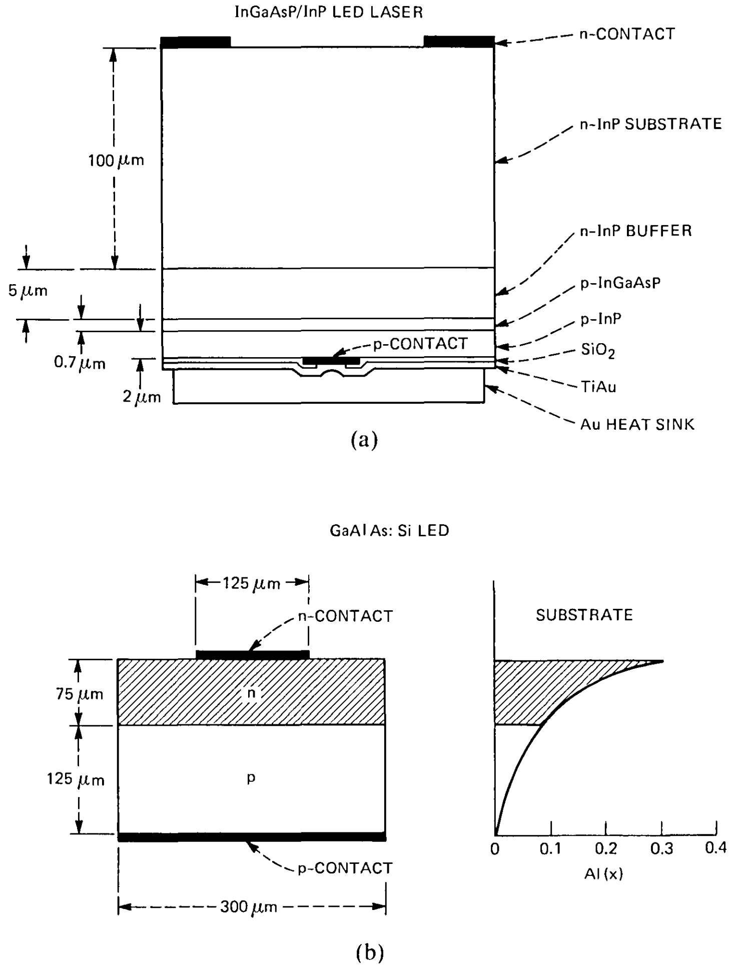


Fig. 18—(a) Schematic of InP/InGaAsP LED or laser structure. (b) Schematic of a GaAlAs:Si LED structure.

nescence. Once a device is demounted for analysis, it is impractical to rebond it for EBIC analysis. For the demounted devices, TCL is advantageous over CL because of the greater sensitivity of the technique in imaging defects. Examples will be shown where TCL is used to evaluate device degradation.

Formation of dark spots and dark lines oriented in the ⟨100⟩ and ⟨110⟩ directions is a degradation mode for LEDs and lasers fabricated from many semiconductors, e.g., GaP,[60] GaAs/GaAlAs,[8,31,54,55] GaAsP,[61,62] and InP/InGaAsP.[19,40,63,64] Dark line defects oriented along the ⟨110⟩ direction grow by glide of dislocations on the {111} slip planes,[65-68] whereas the growth of ⟨100⟩ DLDs may involve both glide and climb of dislocations.[8,69,70] Studies have shown that formation of ⟨100⟩ DLDs require minority carrier injection, whereas ⟨110⟩ DLDs can be induced with either minority carrier injection or stress.[66]

Studies of stress-induced DLDs in GaAs/GaAlAs DH crystals provide a useful background. These studies showed that a threshold stress is required to form ⟨110⟩ DLDs.[65–68] This threshold level decreases for increasing minority carrier injection.[65,66,68] For no injection, the threshold stress is ~8 × 10^9 dynes/cm^2.[66] This value rapidly decreases to ~2 × 10^9 dynes/cm^2 under the lower excitation level of 3W/cm^2 from a 6471 Å Kr-ion laser beam, equivalent to a current density of $J = 1$ A/cm^2.[66] A lower stress level of ~6 × 10^8 dynes/cm^2 is found for $J =$ 58 A/cm^2.[68] These stress values are comparable in magnitude to other typical sources of stress. Stresses from dielectric coatings can be as high as 10^{10} dynes/cm^2.[62] Stresses because of dicing damage are 2–7 × 10^7 dynes/cm^2,[61] and the internal stress on a mounted LED, assuming a 100°C temperature difference between header and LED, can be ~10^9 dynes/cm^2.[71] Care in processing and handling is required if stress-induced DLD formation is to be avoided.

5.1 InP/InGaAsP LED

The degradation of InP/InGaAsP LEDs by the formation of ⟨110⟩ DLDs has recently been reported by Temkin et al.[19] It was demonstrated that this mode of degradation occurs for temperatures in excess of 190°C even without current bias. Although DLDs may be observed in electroluminescence and EBIC scanning, TCL was used to show that the DLDs originated at defects within the InP buffer layer. Additionally, the DLDs were confined to the InP buffer layer and did not extend into the quarternary active layer during aging.

The wafer consists of an Sn-doped InP buffer layer (n ~ 2 × 10^{18} cm^{-3}, 2 to 4 μm thick) followed by a InGaAsP active layer (not intentionally doped, 0.7 μm thick) and an Zn-doped InP confining layer (p ~ 3 × 10^{18} cm^{-3}, 2 μm thick). The p-n junction, formed by Zn outdiffusion from the confining layer, is placed at the interface between the buffer and quarternary layer. The lattice mismatch, Δa/a, between the InGaAsP active layer and InP was kept below 2 × 10^{-4}, and the wafer, thus, has no misfit dislocations to begin with. After thinning the wafer to ~100 μm, metallizations were electron-beam evaporated using shadow masks. The Be/Au p-contact, 50 μm in diameter, was made directly to the p-confining layer; the Au-Sn-Au n-contact was deposited during the same evaporation run.[72] Finally, a 20-μm-thick gold heat-sink, separated from the semiconductor by a 2000-Å-thick layer of SiO_2, was plated on the p-side of the wafer to assure low thermal impedance of the diode. The finished devices were mounted with epoxy on heat-sink headers and wire bonded. A schematic of the device structure is shown in Fig. 18a.

The electroluminescence (EL) images of an LED before and after the 200°C aging (300 h and no bias) are shown in Fig. 19. Figure 19a shows a uniform and well-defined light spot imaged by an infrared-sensitive vidicon tube. No inclusions or processing-induced defects could be seen. Figure 19b shows the EL pattern of the LED degraded to 40 percent of its initial light output. A large number of ⟨110⟩ DLDs could be seen throughout the entire contact area. Similar EL patterns have been seen in all of the 30 devices aged by 200°C storage (no bias) degradation.

To find the position of DLD sources within the device structure, diodes which were carefully removed from their heat-sink headers were examined by TCL. After demounting the devices, the *p*-contact and the confining *p*-InP layer were removed using appropriate etchants, and devices were evaluated by TCL. Two series of TCL scans were performed. In the first, the exposed opening in the *n*-metallization was scanned by the electron beam of the SEM and the signal detected by a Ge-PIN placed under the device. The resulting image, in which CL was generated in the InP substrate is shown in Fig. 20a. Inclusion-like defects with associated DLDs can be seen. This image is similar to an EBIC image of the device. In the second TCL scan, the chip was turned upside down with the quaternary (*Q*) active layer facing the electron beam and the detector placed directly below the *n*-InP. In this configuration, in which the CL signal originated in the *Q*-layer, inclusions and DLDs were not observed. After removal of the active layer with a selective stop-etch, the exposed surface of the *n*-InP buffer layer was

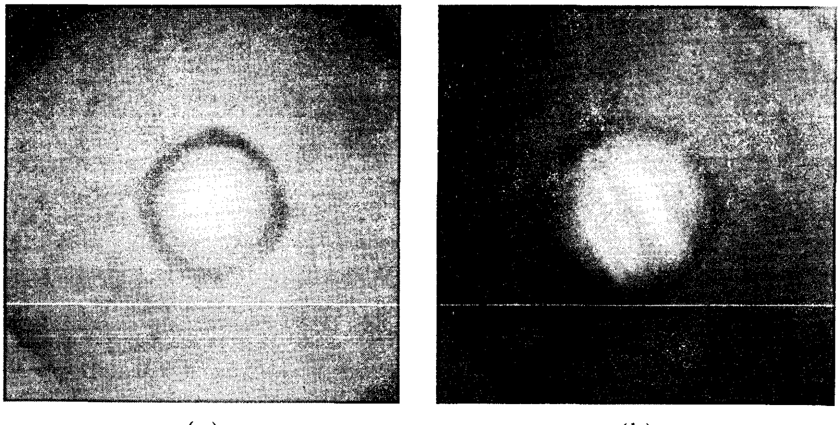


Fig. 19—(a) The EL pattern of an LED before aging. (b) The EL pattern after storage aging at 200°C for 300 h with no bias.

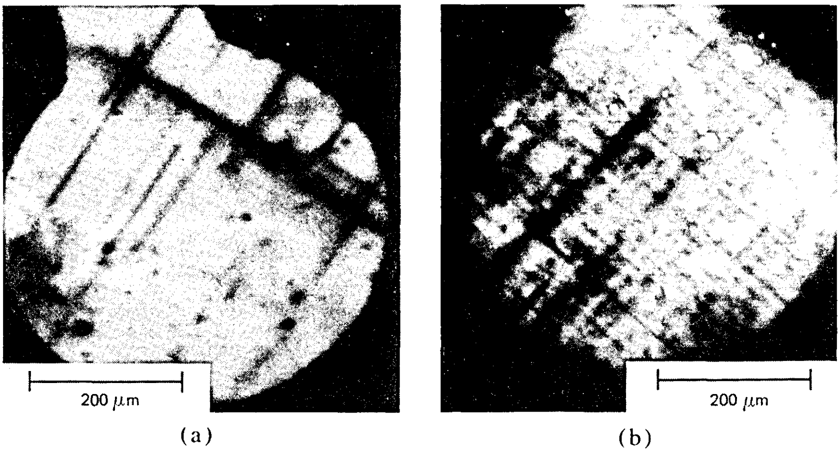


Fig. 20—(a) Transmission cathodoluminescence image obtained on a degraded device. The n-InP surface is imaged in the opening in the metallization. (b) Transmission cathodoluminescence image of the n-InP buffer layer with the quarternary active layer removed by a selective etchant.

imaged and the result is shown in Fig. 20b. Again, only the area above the opening in the n-metallization was imaged. The prominent DLDs observed previously are still visible (a mirror inversion is needed for comparison between Fig. 20a and 20b). However, a number of additional features can be seen. These consist of a much greater density of inclusion-like defects and a large density of shorter DLDs originating at those defects. This DLD pattern is similar to the images obtained by EL. Thus, the DLDs and their sources appear to be confined to the InP buffer layer and its interfaces with the InP substrate and the Q-active layer. Since these DLDs grow without current injection, they are believed to be stress induced.

5.2 Planar GaAlAs LED

For planar GaAlAs LEDs, the TCL technique is especially important since dislocations and DLDs cannot be observed in either the EBIC or CL mode of the SEM. Although good quality CL images of the LEDs were obtained on our apparatus, dislocations are still not observed probably because of a lack of instrument sensitivity. Balk et al. showed that a highly efficient CL apparatus is required to detect dislocations in GaAs: Se;[73] a similar apparatus may be required to detect dislocations in our device wafers. Also, because of the thickness of the window layer, the EBIC signal is not from primary e-h pairs generated by the electron beam, but from secondary e-h pairs generated within a minority carrier diffusion length of the p-n junction by recombination radiation from the primary electron-holes. Thus, the effect of dislocations on the EBIC signal is likely to be below the noise level.

Figure 21 is a TCL image of the back surface of an LED that degraded during the 100 h burn-in. The 50-μm-diameter area defines the contact or light-emitting region. A ⟨100⟩ DLD is observed in the contact region along with two ⟨110⟩ DLDs. These DLDS were observed in the EL image. However, the source of the two ⟨110⟩ DLDs, not seen in EL, is a dislocation lying at the intersection of the DLDs. This dislocation, appearing as the dark spot indicated by the arrow in Fig. 21, intersects the back surface outside the contact area but probably threads through the light-emission region below the surface. The source of the ⟨100⟩ DLD could not be distinguished in this image; this DLD was already well developed as shown by the larger width in comparison to the ⟨110⟩ DLDs. The high density of dark spots representing the intersection of

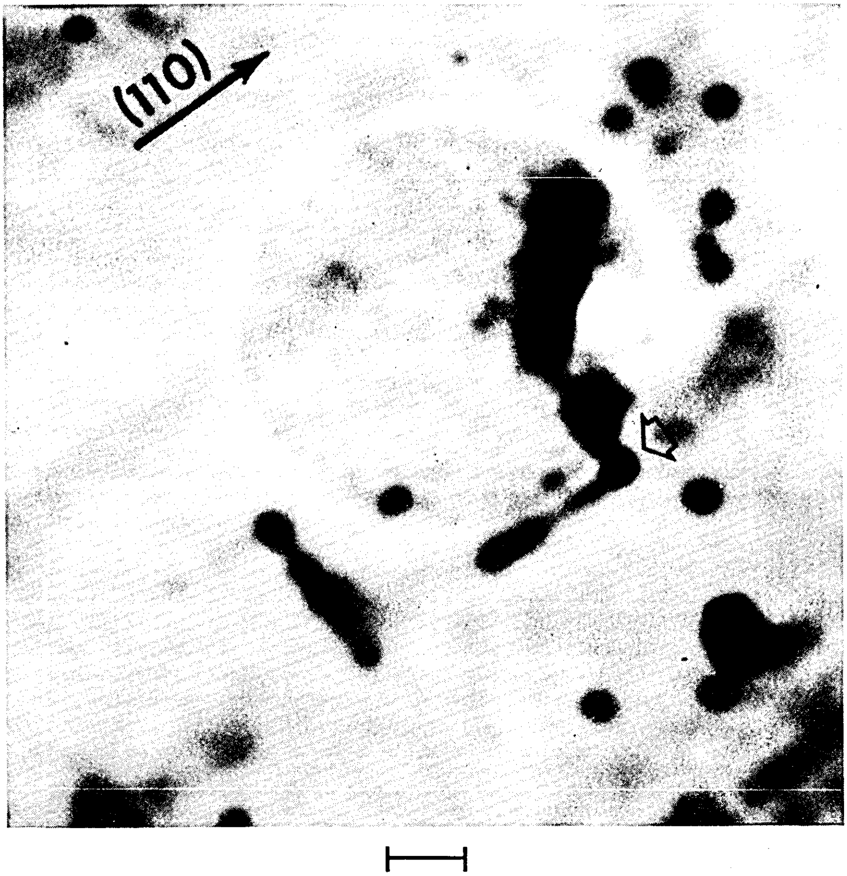


Fig. 21—Dark line defects oriented along ⟨100⟩ and ⟨110⟩ directions in a planar DH LED induced by current injection during burn-in. The circular region defines the 50-μm-diameter back surface contact.

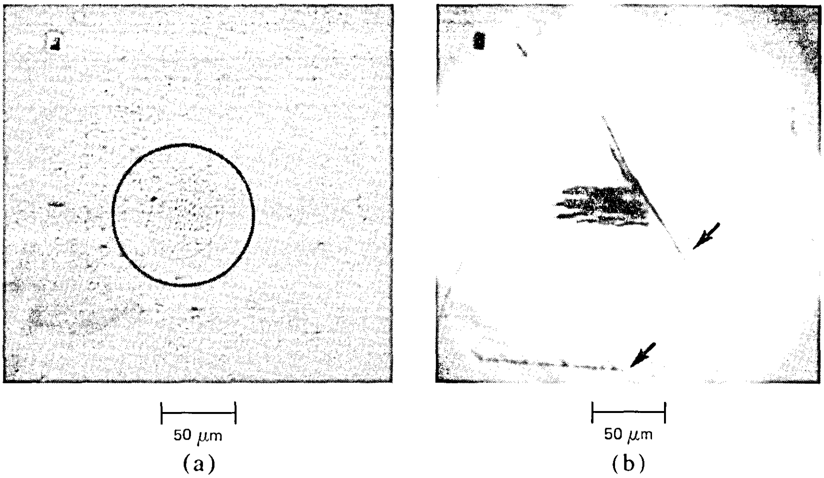


Fig. 22—(a) Secondary electron image of the p-surface of a planar DH LED. The circle encloses the contact. (b) Transmission cathodoluminesence image showing ⟨110⟩ DLDs induced by current injection during long-term aging. The source of DLDs are the two scratches indicated by the arrows.

dislocations with the surface in the vicinity of the contact and the identification of a dislocation as the source of the ⟨110⟩ DLDs in this sample suggest that a dislocation may also be the source of the ⟨100⟩ DLD.

The majority of the DLDs observed in samples degraded during burn-in are similar to those shown in Fig. 21. Also typical of degraded LEDs is the high density of dislocations. It is emphasized that no dislocations were observed in the contact area of undegraded LEDs subjected to the burn-in.

Figure 22 shows the p-surface of a device that catastrophically degraded after ~5000 h of operation at room temperature under 6 kA/cm^2 bias. From the TCL image, Fig. 22b, the method of degradation can be reconstructed. Two scratches were initially present, i.e. the two lines away from the p-contact indicated by the arrows. During operation, a ⟨110⟩ DLD started from one of the scratches and propagated towards the contact. Upon reaching the contact area, enclosed by a circle in the SE image of Fig. 22a, the DLDs multiplied rapidly because of the high-current density. The catastrophic degradation at $\sim 5 \times 10^3$ h agrees with the slow growth of ⟨110⟩ DLDs under the low injection conditions away from the p-contact. Thus, TCL examination identifies a previously unreported method by which LEDs degrade; defects away from the contact area may initiate DLDs that grow into the contact area and rapidly degrade the device.

GaAlAs:Si LED

Graded-bandgap, homojunction GaAlAs:Si LEDs are presently used as high-efficiency light sources for optoisolators.[17,48] Bias tests of these devices, accelerated by applying a thermal stress, demonstrated high reliability; with 30 mA forward current bias, a mean-time-to-failure of $\sim 10^5$ h at 250°C and $\sim 10^8$ h at 25°C was determined. The EL pattern of degraded diodes appears to differ from that of unaged devices only in relative brightness. Both degraded and undegraded LEDs exhibit uniform light emission; no dark lines or spots are observed. Recently, however, aging studies have shown that some LEDs degrade rapidly when aged at 200°C even without current bias; degradation to 10 percent of initial efficiency occurred within 400 h. Examination of the EL image of many of the most heavily degraded devices showed the presence of ⟨110⟩ oriented DLDs. Thus, the drop in LED efficiency is attributed to nonradiative recombination at the DLDs. Since dark line formation has not been previously observed in graded bandgap GaAlAs:Si LEDs, TCL was used to investigate the origin and growth behavior of the DLDs.

To determine the source of the DLDs, the degraded LEDs were first visually inspected on the headers using the SEM and Nomarski interference microscopy. Some damage was found, but it was considered neither unusual nor excessive. After demounting the LEDs from the headers, small pyramids of various sizes were found on the *p*-surface of some of the LEDs.

Figure 23a is an SE image of an unaged LED with several pyramids.

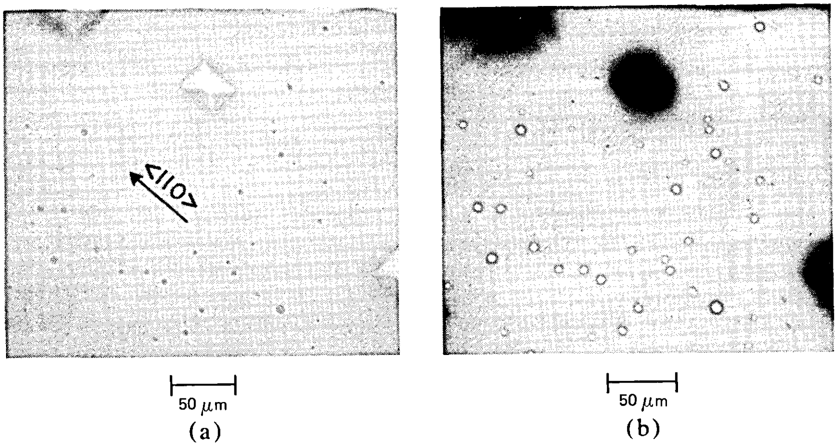


Fig. 23—(a) Secondary electron image of the *p*-surface of an unaged LED with pyramids. The sides of the pyramids are parallel to the ⟨110⟩ directions. (b) Transmission cathodoluminesence image of the region shown in (a). The dark area surrounding each pyramid is a dislocation network. The short dark lines are dislocations extending along the ⟨110⟩ directions.

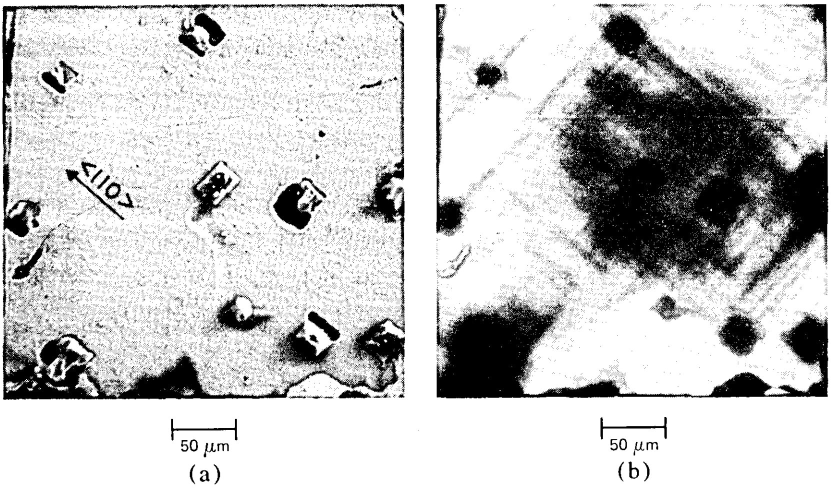


Fig. 24—(a) Secondary electron image of the *p*-surface of a degraded LED with pyramids. (b) image of region shown in (a). Dark line defects oriented along ⟨110⟩ direction initiate at the pyramids.

X-ray microanalysis on the SEM showed that these pyramids are composed mainly of silicon. Figure 23b is the TCL image of the region shown in Fig. 23a. The pyramids appear dark because of the low luminescence efficiency of these regions. In addition to the dark regions, dark lines extending from the pyramids along the ⟨110⟩ directions are observed. These dark regions and lines are part of a dislocation network and have no corresponding features in the SE image, Fig. 23a. These dislocation networks are probably a result of lattice mismatch between the GaAlAs epilayer and the Si pyramids.

Figures 24a and 24b are, respectively, the SE and TCL images of the *p*-surface of typical degraded, bonded LEDs. Although Figs. 23 and 24 correspond to two different LEDs, comparison of these figures show clearly that with aging, the dislocation networks originating at the pyramids enlarge along the ⟨110⟩ directions and form ⟨110⟩ DLDs.

The above study was carried out exclusively on the *p*-surface because of the presence of pyramids. To determine whether the DLDs originate only on the *p*-surface, the top *n*-surface and {100} sides of the LEDs need to be examined. In unaged devices, the usual dislocations, stacking faults, and processing damage were found on these surfaces. In degraded devices, irrespective of the degree of degradation, TCL images of the *n*-surface were similar to those of unaged LEDs, but ⟨110⟩ DLDs were found on the {100} side faces. By examining the {100} planes of partially and highly degraded LEDs, the DLDs were found to initiate at the *p*-surface and propagate towards the p-n junction; the DLDs did not extend beyond the p-n junction into the n-region even in the most heavily degraded LEDs ($\eta/\eta_o < 0.01$).

Figures 25a and 25b show the TCL image of the {100} side of a partially ($\eta/\eta_o \sim 0.6$) and heavily degraded ($\eta/\eta_o \sim 0.1$) LED, respectively. The LED in Fig. 25a did not show DLDs in the EL image, whereas several DLDs were apparent in the EL image of the LED in Fig. 25b. The explanation of the EL images is evident by comparing Figs. 25a and 25b. In both figures, the p-n junction appears as a dark line since the built-in electric field separates the e-h pairs generated by the electron beam and prevents recombination. The n-layer is brighter than the p-layer because of higher intrinsic luminescence efficiency rather than a result of degradation. In Fig. 25a, the DLDs initiate at the p-surface and extend up the sides. The DLDs intersect the {100} p-surface at an angle of 45 degrees, suggesting that the DLDs lie on the {111} slip planes. Dark line defects do not appear in the EL image since they have not propagated to the p-n junction where the EL image is generated. In Fig. 25b, the DLDs propagate up to the p-n junction but do not penetrate into the n-layer. Thus, DLDs appear in the EL image, but not in a TCL or EBIC scan of the n-layer. Figure 25 also demonstrates the ineffectiveness of EBIC in obtaining similar images. Electron beam-

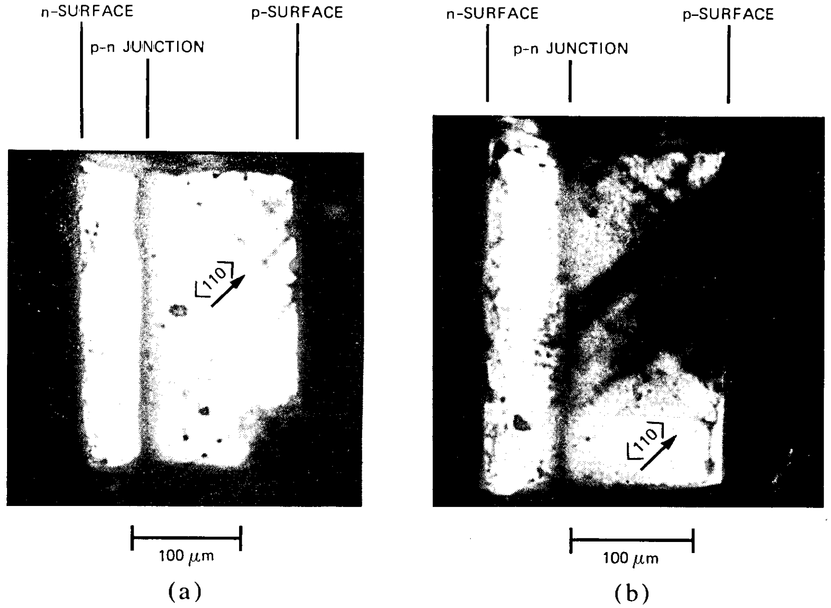


Fig. 25—(a) Transmission cathodoluminesence image of ⟨100⟩ side of a degraded LED. Dark line defects oriented along ⟨110⟩ direction initiate at the p-surface and propagate along the ⟨111⟩ planes toward the p-n junction. (b) Transmission cathodoluminesence image of ⟨100⟩ side of a degraded LED. ⟨110⟩ DLDs terminate at the p-n junction.

induced current images the region approximately twice the minority carrier diffusion length on either the *p* or *n* region. However, as shown in Fig. 25, the DLDs form far from the p-n junction. Additionally, the EBIC signal varies dramatically at the p-n junction. The small perturbations on the EBIC signal produced by the DLDs are hard to image on such a strongly varying background signal.

The DLDs stop their growth upon reaching the p-n junction probably because of lower stress in the *n*-layer. Since the etch used to remove the saw damage during die separation removes more of the *n*-layer than the *p*-layer, the *n*-layer may have a lower surface damage-induced stress as found for GaP LEDs. A high internal electric field at the p-n junction is unlikely to hinder the propagation of the DLDs. Similar stress induced ⟨110⟩ DLDs in GaAs/GaAlAs wafers have been shown to propagate through the p-n junction.[74] Further study, e.g., an examination of degraded LEDs using TEM or a reliability study of LEDs that have been etched more uniformly, is required to determine the exact cause of this effect.

As a result of this TCL study, wafers with Si pyramids evident in a visual inspection are rejected prior to processing since the wafers produce unreliable LEDs.

VI. SUMMARY

We have demonstrated a new SEM imaging technique which we have called transmission cathodoluminescence (TCL). Defects within the e-h excitation volume are imaged as in the familiar CL mode but with a much higher collection efficiency and, thus, with a much higher detection sensitivity. In addition, this technique can be used to image defects within the sample volume that changes the optical properties of the materials; these defects cannot be imaged with the CL mode. We have demonstrated the TCL technique on a variety of defects in a variety of semiconducting materials used in optical communications. Transmission cathodoluminesence can be used to screen out low-quality substrates and device wafers to save processing time and cost, provide rapid evaluation of growth and processing procedures, and evaluate degraded devices where other defect revealing techniques are either unsuitable or very difficult.

VII. ACKNOWLEDGMENTS

This work was done in collaboration with various colleagues in Bell Laboratories. We would like to thank W. A. Bonner, A. A. Ballman, V. G. Keramidas, D. R. Ketchow, and A. G. Dentai for growing many of the crystals. We thank R. H. Saul, O. G. Lorimor, G. Y. Chin, J. H. Wernick for helpful discussions. We are grateful to C. L. Zipfel for the reliability studies, R. J. Roedel for the etching studies, and D. D. Roccasecca, E. Lassiter, and K. B. Bauers for processing.

REFERENCES

1. T. P. Lee and C. A. Burrus, "Dark Current and Breakdown Characteristics of Dislocation-Free InP Photodiodes," Appl. Phys. Lett. *36,* No. 7 (April 1980), pp. 587–9.
2. K. Takahashi, "Comparison of Etch Pits and V-I Characteristics in n-GaAs {100} Crystals," Jpn. J. Appl. Phys. *19,* No. 4 (April 1980), pp. 773–4.
3. T. P. Lee et al., "High Avalanche Gain in Small-Area InP Photodiodes," Appl. Phys. Lett. *35,* No. 7 (October 1979), pp. 511–3.
4. J. E. Lawrence, "Electrical Properties of Copper Segregates in Silicon p-n Junctions," J. Electrochem. Soc. *112,* No. 8 (August 1965), pp. 796–800.
5. F. Capasso et al., "Investigation of Microplasmas in InP Avalanche Photodiodes," Proc. Int. Elec. Device Meeting, Paper No. 27.4 (December 1979), pp. 647–9.
6. C. Workhoven, C. van Opdorp, and A. T. Vink, "Non-Radiative Recombination in *n*-type LPE GaP," Inst. of Phys. Proc. GaAs and Related Compounds, *33a,* (September 1976) pp. 317–25.
7. R. J. Roedel et al., "The Effects of Dislocations in $Ga_{1-x}Al_xAs$:Si Light-Emitting Diodes," J. Electrochem. Soc. *126,* No. 4 (April 1979), pp. 637–41.
8. P. Petroff and R. L. Hartman, "Defect Structure Introduced During Operation of Heterojunction GaAs Lasers," Appl. Phys. Lett. *23,* No. 8 (October 1973), pp. 469–71.
9. W. D. Johnston, Jr. et al., "Spatially Resolved Photoluminescence Characterization and Optically Induced Degradation of $In_{1-x}Ga_xAs_yP_{1-y}$ DH Laser Material," Appl. Phys. Lett. *33,* No. 12 (December 1978), pp. 992–4.
10. A. K. Chin et al., "Evaluation of Defects and Degradation in GaAs-GaAlAs Wafers Using Transmission Cathodoluminescence," J. Appl. Phys. *51,* No. 2 (February 1980), pp. 978–83.
11. K. Ishida and T. Kamejima, "TEM Study of Dark Line Defect Growth From Dislocation Clusters in (GaAl)As-GaAs Double Heterostructure Lasers," J. Elec. Mat. *8,* No. 1 (January 1979), pp. 57–73.
12. A. R. Goodwin et al., "The Effects of Processing Stresses on Residual Degradation in Long-Lived $Ga_{1-x}Al_xAs$ Lasers," Appl. Phys. Lett. *34,* No. 10 (May 1979), pp. 647–9.
13. B. Wakefield, "Strain-Enhanced Luminescence Degradation in GaAs/GaAlAs Double-Heterostructure Lasers Revealed by Photoluminescence," J. Appl. Phys. *50,* No. 12 (December 1979), pp. 7914–6.
14. A. K. Chin, H. Temkin, and R. J. Roedel, "Transmission Cathodoluminescence: A New SEM Technique to Study Defects in Bulk Semiconductor Samples," Appl. Phys. Lett. *34,* No. 7 (April 1979), pp. 476–8.
15. A. K. Chin, S. Mahajan, and A. A. Ballman, "Imaging of Dislocations in InP Using Transmission Cathodoluminescence," Appl. Phys. Lett. *35,* No. 10 (November 1979), pp. 784–6.
16. A. K. Chin et al., "Evaluation of Defects in InP and InGaAsP by Transmission Cathodoluminescence," J. Appl. Phys. *50,* No. 9 (September 1979), pp. 5707–9.
17. A. K. Chin et al., "Stress-Induced Dark Line Defect Formation in GaAlAs:Si LEDs," J. Electrochem. Soc. *128,* No. 3 (March 1981), pp. 661–9.
18. C. A. Gaw and C. L. Reynolds, Jr., "Transmission Cathodoluminescence as a Screening Technique for Rake Lines in (Al,Ga)As DH Laser Material," Elec. Lett. *17,* No. 8 (April 1981), pp. 285–6.
19. H. Temkin, C. L. Zipfel, and V. G. Keramidas, "High Temperature Degradation of InGaAsP/InP LEDs," J. Appl. Phys. *52,* No. 8 (August 1981), pp. 5377–80.
20. S. Mahajan et al., "The Characterization of Highly-Zinc-Doped InP Crystals," Appl. Phys. Lett. *35,* No. 2 (July 1979), pp. 165–8.
21. S. Mahajan and A. K. Chin, "The Status of the Current Understanding of InP and InGaAsP Materials," J. Crystal Growth *54* (July 1981), pp. 138–49.
22. A. K. Chin, H. Temkin, and G. Y. Chin, "Method of Analyzing Localized Nonuniformities in Luminescing Materials," U.S. Patent 4-238-686, issued December 1980.
23. S. Mahajan et al., "Perfection of Homoepitaxial Layers Grown on (001) InP Substrates," Appl. Phys. Lett. *38,* No. 4 (February 1981), pp. 255–8.
24. S. Komiya and T. Kotani, "Direct Observation of Dislocations in $Ga_{1-x}Al_xAs$-GaAs

Grown by the LPE Method," J. Electrochem. Soc. *125,* No. 12 (December 1978), pp. 2019–24.

25. S. Kishino et al., "X-ray Topographic Study of Dark-Spot Defects in GaAs-$Ga_{1-x}Al_xAs$ Double-Heterostructure Wafers," Appl. Phys. Lett. *27,* No. 4 (August 1975), pp. 207–9.
26. P. G. McMullin, "Quality Evaluation of GaAs-AlGaAs Heterostructure Wafers Using the Electron Beam Induced Current Technique," Proc. SEM Symp., Ill. Inst. Tech. Res. Inst. (April 1976), pp. 543–50.
27. D. A. Shaw and P. R. Thornton, "Cathodoluminescent Studies of Laser Quality GaAs," J. Mat. Science *3,* No. 5 (September 1968), pp. 507–18.
28. F. R. Nash et al., "Laser-Excited Photoluminescence of Three-Layer GaAs Double-Heterostructure Laser Material," Appl. Phys. Lett. *27,* No. 4 (August 1975), pp. 234–7.
29. F. R. Nash, W. R . Wagner, and R. L. Brown, "Threshold Current Variations and Optical Scattering Losses in (Al,Ga)As Double-Heterostructure Lasers," J. Appl. Phys. *47,* No. 9 (September 1976), pp. 3992–4005.
30. R. A. Logan et al., "Doping Effects on Rake-Line Formation in LPE Growth of $Al_xGa_{1-x}As$ DH Lasers," J. Appl. Phys. *50,* No. 9 (September 1979), pp. 5970–7.
31. S. D. Hersee and D. J. Stirland, "Degradation Mechanisms in High-Radiance LEDs," Inst. of Physics Proc. GaAs and Related Compounds, *33a* (September 1976), pp. 370–8.
32. M. E. Drougard, "Optical Inhomogeneities in Gallium Arsenide," J. Appl. Phys. *37,* No. 4 (March 1966), pp. 1858–66.
33. J. M. Titchmarsh et al., "Carrier Recombination at Dislocations in Epitaxial Gallium Phosphide Layers," J. Mat. Science *12,* No. 2 (February 1977), pp. 341–6.
34. H. C. Casey, "Investigation of Inhomogeneities in GaAs by Electron-Beam Excitation," J. Electrochem. Soc. *114,* No. 2 (February 1967), pp. 153–8.
35. K. Nakagawa, K. Maeda, and S. Takeuchi, "Observation of Dislocations in Cadmium Telluride by Cathodoluminescence Microscopy," Appl. Phys. Lett. *34,* No. 9 (May 1979), pp. 574–5.
36. J. G. Grabmaier and C. B. Watson, "Dislocation Etch Pits in Single Crystal GaAs," Phys. Status Solidi *32,* No. 1 (January 1969), pp. K13–5.
37. M. Ishii et al., "Etch Pit Observation of Very Thin {001}-GaAs Layer by Molten KOH," Jpn. J. Appl. Phys. *15,* No. 4 (April 1976), pp. 645–50.
38. A. Huber and N. T. Linn, "Révélation Métallographique des Defauts Cristallins dans InP," J. Crystal Growth *29,* No. 1 (May 1975), pp. 80–4.
39. M. S. Abrahams and C. J. Buiocchi, "Etching of Dislocations on the Low-Index Faces of GaAs," J. Appl. Phys. *36,* No. 9 (September 1965), pp. 2855–63.
40. S. Yamakoshi et al., "Degradation of High Radiance InGaAsP/InP LEDs at 1.2 – 1.3 μm Wavelength," Int. Elec. Device Meeting, Paper No. 5.6 (December 1980), pp. 122–5.
41. J. L. Richards and A. J. Crocker, "Etch Pits in Gallium Arsenide," J. Appl. Phys. *31,* No. 3 (March 1960), pp. 611–2.
42. D. B. Darby and G. R. Booker, "Scanning Electron Microscope EBIC and CL Micrographs of Dislocations in GaP," J. Mat. Science *12,* No. 9 (September 1977), pp. 1827–33.
43. G. B. Mullin et al., "Crystal Growth and Properties of Group IV Doped Indium Phosphide," J. Crystal Growth *13/14* (May 1972), pp. 640–6.
44. R. C. Clarke, D. S. Robertson, and A. W. Vere, "A Preliminary Study of Dislocations in Indium and Gallium Phosphides," J. Mat. Science *8,* No. 9 (September 1973), pp. 1349–54.
45. T. Iizuka, "Etching Studies of Impurity Precipitates in Pulled GaP Crystals," J. Electrochem. Soc. *118,* No. 7 (July 1971), pp. 1190–4.
46. ETEC Corp., Haywood California.
47. D. B. Holt, *Quantitative Scanning Electron Microscopy,* D. B. Holt, M. D. Muir, P. R. Grant, and J. M. Boswarva, Eds. New York: Academic Press, 1974, pp. 335–86.
48. L. R. Dawson, "High-Efficiency Graded-Band-Gap $Ga_{1-x}Al_xAs$ Light-Emitting Diodes," J. Appl. Phys. *48,* No. 6 (June 1977), pp. 2485–92.
49. S. Mahajan et al., "Characterization of Growth Striations in InP," Amer. Inst. Metallurgical Engrs. Meeting, February 1979.
50. Y. Seki, J. Matsui, and H. Watanabe, "Impurity Effect on the Growth of Dislocation-Free InP Single Crystals," J. Appl. Phys. *47,* No. 7 (July 1976), pp. 3374–6.
51. Y. Seki, H. Watanabe, and J. Matsui, "Impurity Effect on Grown-In Dislocation Density of InP and GaAs Crystals." J. Appl. Phys. *49,* No. 2 (February 1978), pp. 822–8.

52. G. T. Brown, B. Cockayne, and W. R. MacEwan, "The Growth of Dislocation-Free Ge-doped InP," J. Crystal Growth *51,* No. 2 (February 1891), pp. 369–72.
53. S. Sakai, M. Umeno, and Y. Amemiya, "Measurement of Diffusion Coefficient and Surface Recombination Velocity for p-InGaAsP Grown on InP," Jpn. J. Appl. Phys. *19,* No. 1 (January 1980), pp. 109–13.
54. D. B. Wittry and D. F. Kyser, "Cathodoluminescence at p-n Junctions in GaAs," J. Appl. Phys. *36,* No. 4 (April 1965), pp. 1387–9.
55. S. O. Hara et al., "Defect-Induced Degradation in High-Radiance Lamps," Inst. of Phys. Proc. GaAs and Related Compounds, *33a* (September 1976), pp. 379–87.
56. S. Yamakoshi et al., "Degradation of High Radiance $Ga_{1-x}AlAs$ LED's," Appl. Phys. Lett. *31,* No. 9 (November 1977), pp. 627–9.
57. V. G. Keramidas et al., unpublished work.
58. K. Oe, Y. Shinoda, and K. Sugiyama, "Lattice Deformations and Misfit Dislocations in GaInAsP/InP Double-Heterostructure Layers," Appl. Phys. Lett. *33,* No. 11 (December 1978), pp. 962–4.
59. S. Yamakoshi et al., "Reliability of High Radiance InGaAsP/InP LED's Operating in the 1.2 – 1.3 μm Wavelength," IEEE J. Quantum Electron. *AE-17,* No. 2 (February 1981), pp. 167–73.
60. M. Iwamoto and A. Kasami, "Observation of Dark Line Defects in GaP Green LED's Under External Uniaxial Stress," Appl. Phys. Lett. *28,* No. 10 (May 1976), pp. 591–2.
61. N. Shimano, "Degradation of $GaAs_{0.9}P_{0.1}$ LED's Operating at High Current Densities," Jpn. J. Appl. Phys. *17,* No. 8 (August 1978), pp. 1323–30.
62. N. Shimano, Y. Kawai, and M. Sakuta, "Degradation of $GaAs_{0.9}P_{0.1}$ Light-Emitting Diodes for Optical Fiber Communication with Internal Stress," J. Appl. Phys. *51,* No. 2 (February 1980), pp. 1227–32.
63. S. Mahajan et al., "The Mechanism of Optically Induced Degradation in InP/ $In_{1-x}Ga_xAs_yP_{1-y}$ Heterostructures," Appl. Phys. Lett. *34,* No. 10 (May 1979), pp. 717–9.
64. O. Ueda et al., "Transmission Electron Microscope Observation of Dark-Spot Defects in InGaAsP/InP Double-Heterostructure Light-Emitting Diodes Aged at High Temperatures," Appl. Phys. Lett. *36,* No. 4 (February 1980), pp. 300–1.
65. S. Kishino et al., "Dark-Line Defects Induced by Mechanical Bending in GaAs-$Ga_{1-x}Al_xAs$ Double-Heterostructure Wafers," Appl. Phys. Lett. *29,* No. 8 (October 1976), pp. 488–90.
66. H. Nakashima et al., "Growth and Propagation Mechanism of ⟨110⟩-Oriented Dark Line Defects in GaAs-$Ga_{1-x}Al_xAs$ Double Heterostructure Crystals," J. Appl. Phys. *48,* No. 7 (July 1977), pp. 2771–5.
67. K. Ishida, T. Kamejima, and J. Matsui, "Nature of ⟨110⟩ Dark-Line Defects in Degraded (GaAl)As-GaAs Double-Heterostructure Lasers," Appl. Phys. Lett. *31,* No. 6 (September 1977), pp. 397–9.
68. T. Kamejima, K. Ishida, and J. Matsui, "Injection-Enhanced Dislocation Glide Under Uniaxial Stress in GaAs-(GaAs)As Double Heterostructure Laser," Jpn. J. Appl. Phys. *16,* No. 2 (February 1977), pp. 233–40.
69. P. W. Hutchinson and P. S. Dobson, "Defect Structure of Degraded GaAlAs-GaAs Double Heterostructure Lasers," Phil. Mag. *32,* No. 4 (October 1975), pp. 745–54.
70. P. Petroff and R. L. Hartman, "Rapid Degradation Phenomenon in Heterojunction GaAlAs-GaAs Lasers," J. Appl. Phys. *45,* No. 9 (September 1974), pp. 3899–903.
71. G. Zaeschmar and R. S. Speer, "Mechanical-Stress-Induced Degradation in Homojunction GaAs LED's," J. Appl. Phys. *50,* No. 9 (September 1979), pp. 5686–93.
72. H. Temkin et al., "Ohmic Contacts to *p*-type InP Using Be-Au Metallization," Appl. Phys. Lett. *36,* No. 6 (March 1980), pp. 444–6.
73. L. J. Balk, E. Kubalek, and E. Menzel, "Investigation of As-Grown Dislocations in GaAs Single Crystals in the SEM," Proc. SEM Symp., Ill. Inst. Tech. Res. Inst. (April 1976), pp. 257–64.
74. K. Ikeda et al., "Degradation of GaAs-(Al,Ga)As Double Heterostructure Light Emitting Diodes," Inst. of Phys. Proc. GaAs and Related Compounds *24* (September 1974), pp. 174–80.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

Detecting the Occurrence of an Event by FM Through Noise

By V. E. BENEŠ

(Manuscript received March 26, 1981)

The occurrence of an event at a random time τ is signaled through white noise by an FM signal whose modulation $h(t - \tau)$ is a causal pulse triggered at τ. Nonlinear filtering is used to find exact expressions for the chance that $\tau > t$, and the expectation of τ, each conditioned on the observed noisy FM signal over $(0, t)$. The former quantity can be used to minimize the probability of error in guessing—from the observations over $(0, t)$—whether τ has occurred by t.

I. INTRODUCTION

The theory of frequency modulation has always been beset by analytical difficulties, and nowhere have these been more in evidence than in the area of optimal demodulation of noisy FM signals. Recent advances in nonlinear filtering, however, make it possible to solve certain problems of detection and estimation quite explicitly. We report on such a class of problems here.

The basic problem setup is this: an event of interest occurs at a random time τ. Its occurrence is signaled by sending a pulse of shape $h(\cdot)$, starting at τ; that is, we send

$$s(t) = \begin{cases} 0 & t < \tau \\ h(t - \tau) & t \geq \tau \end{cases},$$

where $h(\cdot)$ is some causal, integrable pulse. The signal $s(t)$ is transmitted by FM; the waveform is

$$\cos\left[\theta + \omega t + \int_0^t s(u)du\right]$$

for a carrier frequency ω and initial phase θ. In transmission this wave suffers the degradation of having white noise added to it; thus, we observe a signal y_t defined by

$$dy_t = \cos\left[\theta + \omega t + \int_0^t s(u)du\right]dt + db_t,$$

with b_t a Brownian motion independent of τ. We would like to construct a nonlinear filter acting causally on y_t to estimate optimally at each time t whether $\tau < t$ or not, and if so, by how much. This filter will be obtained by solving the nonlinear filtering problem of determining the conditional probability

$$p_0(t) = P\{\tau > t \mid y_s, 0 \le s \le t\}$$

and the conditional density (u = distance back from t to τ)

$$p_1(t, u) = P\{\tau \epsilon d(t-u) \mid y_s, 0 \le s \le t\}, \quad 0 \le u \le t.$$

Such a filter (p_0, p_1) represents a summary, without loss, of all the information in the "past" $\sigma\{y_s, 0 \le s \le t\}$ that is relevant to whether τ occurred by time t, and if so, how far back. In particular, the filter (p_0, p_1) yields least-squares estimates of τ, by integration over u, according to the formula

$$E\{\tau \mid y_s, 0 \le s \le t\} = p_0(t)\frac{\int_t^\infty uf(u)du}{1 - F(t)} + \int_0^t (t-u)p_1(t, u)du,$$

where F is the a priori distribution of τ, and $f = F'$ its density. The first term predicts where τ will be, on the average, when it has not yet occurred by t; the second "postdicts" τ when it has already happened by time t. Indeed, the first term is $E\{\tau 1_{\tau>t} \mid y_s, 0 \le s \le t\}$ and the second is $E\{\tau 1_{\tau \le t} \mid y_s, 0 \le s \le t\}$.

II. NOTATIONS

Let x_t be the process $1_{\tau \le t}$ so that

$$x_t = \begin{cases} 0 \text{ if the event has not occurred by time } t. \\ 1 \text{ if the event occurred by time } t. \end{cases}$$

Then with $X_t = \int_0^t x_s ds$, the signal $s(t)$ can be written as

$$s(t) = \int_0^t h(t-s)dx_s = \begin{cases} H(X_t) & t \ge \tau \\ 0 & t < \tau \end{cases}$$

and the FM signal as

$$\cos[\theta + \omega t + H(X_t)],$$

where $H = \int_0 h(s)ds$.

III. FILTERING EQUATIONS

Our approach is Bayesian: foreknowledge of distr $\{\tau\}$ is used to calculate the conditional probabilities p_0 and p_1. We assume for simplicity, and with only slight loss of generality, that τ has a known a priori distribution F with a differentiable density f. The "rate" at which τ is occurring during $(t, t + h)$, given that it has not yet happened, is just

$$\lambda(t) = \frac{f(t)}{1 - F(t)}, > 0.$$

We assume at first that the phase θ is known at the receiver. Then, the filtering or Zakai equations for unnormalized versions ρ_0 and ρ_1 of p_0 and p_1, respectively, are just

$$d\rho_0 = \lambda(t)\rho_0 dt + \cos(\theta + \omega t)\rho_0 dy_t,$$

$$d\rho_1 = \frac{\partial \rho_1}{\partial u} + \cos[\theta + \omega t + H(u)]\rho_1 dy_t,$$

with initial conditions $\rho_0(0) = 1$,

$$\lim_{t \downarrow 0} \int_0^t \rho_1(t, u)du = 0$$

and boundary condition $\rho_1(t, 0) = \rho_0(t)\lambda(t)$.

It can be seen that since the process x_t is transient in character these equations are coupled in one direction only: ρ_1 depends on ρ_0 via the boundary condition, but ρ_0 in no way depends on ρ_1. Thus, it will be possible to solve for ρ_0 first, and then for ρ_1. We first transform the problem into one without stochastic differentials. This is done by the now familiar device[1] of looking for a solution of the form

$$\rho_0(t) = \exp[y_t \cos(\theta + \omega t)]q_0(t)$$

$$\rho_1(t, u) = \exp\{y_t \cos[\theta + \omega t + H(u)]\}q_1(t, u), \quad 0 \leq t \leq u,$$

where q_0 and q_1 are differentiable functions, though not necessarily C^1. This form for ρ_0 and ρ_1 indicates that the rough or martingale dependence of these functions on $y(\cdot)$ is confined to the exponent as shown, while their dependence on $y(\cdot)$ via q_0 and q_1 is of a much smoother integrated form, as will be seen.

Applying Ito's formula to the postulated form, with quadratic variation $d\langle y \rangle_t = dt$ since the observation process is a translation of the Wiener process, we find these nonstochastic PDEs for q_0 and q_1:

$$\dot{q}_0 = q_0 \left(-\lambda(t) + \omega y_t \sin(\theta + \omega t) - \frac{1}{2}\cos^2(\theta + \omega t) \right), \; q_0(0) = 1$$

$$\frac{\partial q_1}{\partial t} = -\frac{\partial q_1}{\partial u} + q_1 \left(y_t[\omega + h(u)]\sin[\theta + \omega t + H(u)] \right.$$
$$\left. - \frac{1}{2}\cos^2[\theta + \omega t + H(u)] \right), \; 0 \leq u \leq t.$$

The boundary condition is

$$q_1(t, 0) = q_0(t)\lambda(t)\,.$$

The first equation is an ODE solvable as

$$q_0(t) = \exp\left(-\int_0^t \lambda(s)ds + \int_0^t [\omega y_s \sin(\theta + \omega s) - \frac{1}{2}\cos^2(\theta + \omega s)]ds\right)$$

$$= [1 - F(t)]\exp\left(\int_0^t [\omega y_s \sin(\theta + \omega s) - \frac{1}{2}\cos^2(\theta + \omega s)]ds\right).$$

The second is a first-order PDE solvable by characteristics as

$$q_1(t, u) = A(t - u)\exp\left(\int_0^t \{\omega y_s \sin[\theta + \omega s + H(s - t + u)] - \frac{1}{2}\cos^2[\theta + \omega s + H(s - t + u)] + y_s h(s - t + u)\sin[\theta + \omega s + H(s - t + u)]\}ds\right),$$

where $A(\cdot)$ is an arbitrary function. To obtain A we let $u{\downarrow}0$, and we use $h(s - t + u) = 0$ and $H(s - t + u) = 0$ for $s \le t - u$ to find

$$q_1(t, 0) = A(t)\exp\left(\int_0^t [y_s\omega \sin(\theta + \omega s) - \frac{1}{2}\cos^2(\theta + \omega s)]ds\right),$$

$$= q_0(t)\lambda(t),$$

by the boundary condition. Thus, $A(t) = f(t)$, and we obtain

$$q_1(t, u) = f(t - u)\exp\left(\int_0^{t-u} [\omega y_s \sin(\theta + \omega s) - \frac{1}{2}\cos^2(\theta + \omega s)]ds + \int_{t-u}^t \{[\omega + h(s - t + u)]y_s \sin[\theta + \omega s + H(s - t + u)] - \frac{1}{2}\cos^2([\theta + \omega s + H(s - t + u)]\}ds\right).$$

We remark that this is the unconditional density $f(t - u)$ that τ occur at $t - u$, multiplied by a positive factor depending on the pulse shape h and the observation $\{y_s, 0 \le s \le t\}$. The $\cos^2$ integral can be evaluated explicitly, leading to some simplification, and to approximate formulas for large carrier frequencies.[2]

The normalization

$$p_0(t) + \int_0^t p_1(t, u)du = 1$$

is achieved by dividing each of ρ_0 and ρ_1 by

$$\rho_0(t) + \int_0^t \rho_1(t, u)du,$$

where

$$\rho_0(t) = [1 - F(t)]\exp\biggl(y_t\cos(\theta + \omega t) + \int_0^t [\omega y_s\sin(\theta + \omega s) - \frac{1}{2}\cos^2(\theta + \omega s)]ds\biggr)$$

$$\rho_1(t, u) = f(t - u)\exp\biggl[y_t\cos[\theta + \omega t + H(u)] + \int_0^{t-u} [\omega y_s\sin(\theta + \omega s) - \frac{1}{2}\cos^2(\theta + \omega s)]ds + \int_{t-u}^{t} \biggl[[\omega + h(s - t + u)]y_s\sin[\theta + \omega s + H(s - t + u)] - \frac{1}{2}\cos^2[\theta + \omega s + H(s - t + u)]\biggr)ds\biggr].$$

If, as is likely, the phase θ is not known at the receiver, then it must be integrated out in both ρ_0 and ρ_1 prior to normalization, a process that mars the relatively neat formulas obtained for ρ_0 and ρ_1 for θ known. With θ uniform over $(-\pi, \pi)$ and independent of τ, familiar Bessel function approximations again arise.[2]

IV. HAS τ OCCURRED YET? THE OPTIMAL GUESS

In the kind of system under study here, a task of primary interest is to guess at t whether τ has happened yet. Such a guess is represented mathematically by a random process v_t, taking the value 1 for a decision that τ has not occurred, and a value 0 for a decision that it has, and adapted to the past observations $\rho\{y_s,\ 0 \le s \le t\}$. The probability of error is just

$$P\{\tau \le t \;\&\; v_t = 1\} + P\{\tau > t \;\&\; v_t = 0\},$$

which can be written as

$$\begin{aligned} &E1_{\tau>t}(1 - v_t) + Ev_t(1 - 1_{\tau>t}) \\ &= E1_{\tau>t} - 2E1_{\tau>t}v_t + Ev_t \\ &= E(1_{\tau>t} - v_t)^2, \end{aligned}$$

the mean square error in approximating $1_{\tau>t}$ by v_t. Thus, the chance of error is the least if v_t is chosen to minimize this mean square error. Noting that $p_0(t) = E\{1_{\tau>t} | y_s, 0 \le s \le t\}$, we can write this mean square error as

$$E\{p_0(t) - 2p_0(t)v_t + v_t\}$$

and conclude that a minimizing v_t is

$$v(t) = \begin{cases} 1 & \text{if } p_0(t) > \dfrac{1}{2} \\ 0 & \text{if } p_0(t) \le \dfrac{1}{2}. \end{cases}$$

It follows that by watching $p_0(\cdot)$ we can make a best guess as to whether τ has occurred yet or not, best in the sense of minimizing the chance of being wrong.

V. THE CONDITIONAL EXPECTATION OF τ

As we observe the signal y_t, we may be interested in predicting τ on the basis of the information seen so far. More precisely, since it is possible that at $t > 0$ τ has already occurred, we want to simultaneously predict and "postdict" τ by calculating the two terms in

$$\begin{aligned} \hat{\tau} &= E\{\tau | y_s, 0 \le s \le t\} \\ &= E\{\tau 1_{\tau>t} | y_s, 0 \le s \le t\} + E\{\tau 1_{\tau \le t} | y_s, 0 \le s \le t\}. \end{aligned}$$

The second term is clearly given in terms of p_1 by

$$\int_0^t (t-u)p_1(t, u)du, \quad (u = \text{distance back to } \tau \text{ from } t)\,.$$

We claim that the first is just

$$p_0(t)\frac{\displaystyle\int_t^\infty u dF(u)}{1 - F(t)}.$$

For with $\sigma\{y_s, 0 \le s \le t\} = Y_0^t$ for short, we have

$$\begin{aligned} E\{\tau 1_{\tau>t} | y_0^t\} &= E\{\tau 1_{\tau>t} | y_0^t v\tau > t\} | y_0^t\} \\ &= E\{E\{\tau | y_0^t v\tau > t\} 1_{\tau>t} | y_0^t\} \\ &= E\{E\{\tau | \tau > t\} 1_{\tau>t} | y_0^t\} \\ &= p_0(t)E\{\tau | \tau > t\} \end{aligned}$$

since the additional y_s information in $Y_0^t \cup \tau > t$ is irrelevant to τ when it is known that $\tau > t$. That is, since $x_t = 1_{\tau>t}$ is a Markov process, all the information $\sigma\{x_s, y_s, 0 \leq s \leq t\}$ is irrelevant to $\{x_u, u > t\}$ when it is known that $x_t = 1$, i.e., $\tau > t$.

REFERENCES

1. B. L. Rozovsky, "Stochastic Partial Differential Equations Arising in Nonlinear Filtering Problems," Uspekhi Matem. Nauk, *27* (1972), pp. 213–4.
2. V. E. Beneš, "Least Squares Estimator for Frequency-Shift Position Modulation in White Noise," B.S.T.J., *59*, No. 7 (September 1970), pp. 1289-96.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

Fault-Simulation Methods—Extensions and Comparison

By Y. H. LEVENDEL and P. R. MENON

(Manuscript received December 16, 1980)

In this paper, we compare four different methods of fault simulation in terms of their handling of arbitrary numbers of logic values, modeling levels, and detailed timing. The methods considered are parallel, deductive, multilist, and concurrent simulation methods. Since some of the methods, in their current forms, are unable to handle all the problems under consideration, we have proposed extensions to the methods wherever necessary before making the comparisons. While all the methods considered are capable of solving the problems with the same degree of accuracy, the concurrent simulation method appears to be the simplest and most flexible.

I. INTRODUCTION

Different techniques for the efficient simulation of faults in digital circuits have been published. Among these, the best known are parallel simulation,[1-3] deductive simulation,[4] and concurrent simulation.[5] A few papers analyzing some aspects of these methods have also been published.[6-9]

This paper and two others[10,11] comprise a series attempting a comprehensive analysis of fault simulation methods. It is hoped that they will provide a basis for the selection of fault-simulation methods to satisfy specific requirements.

In this paper, we consider three aspects of circuit modeling and their effects on the fault-simulation method used. First, we consider the number of logic values needed to accurately model logic devices and its impact on the simulation method. Next, the effectiveness of the different methods for simulating at different levels (e.g., gate level, functional level, subsystem level, etc.) is considered. Finally, we discuss the modeling of timing effects, such as rise and fall times and high-frequency rejection.

Our study covers four methods of fault simulation: parallel, deductive, multilist, and concurrent. In their current forms, some of the methods are not capable of handling all the problems we consider.

Therefore, we have attempted to extend the existing methods, wherever necessary, before making the comparisons between methods. Before proceeding to the analysis of the methods, we present a brief description of each method.

Historically, parallel simulation was the first method that simulated a number of faults simultaneously.[1] This method, which is perhaps the most widely used, takes advantage of word-oriented operations in the host computer and packs together several faulty circuit values into one or more computer words. Although this method is quite efficient, multiple passes are required for simulating large numbers of faults.

Deductive simulation attempts to eliminate the need for multiple passes by computing normal signal values in the circuit and deducing the faulty values by manipulating lists of faults.[4] Associated with each signal is a fault list, which is a set of faults, any one of which will cause the signal value to be different from the normal value. The effects of faults are propagated through the circuit by an algebra of sets.

The multilist method associates two or more lists of faults with each signal.[10,12] Conceptually, the number of lists associated with a signal is equal to the number of logic values simulated. Thus, for two logic values, there will be a 0-list and a 1-list associated with each signal, the former being the set of faults in whose presence (individually) the signal will have the value 0, and the latter those that result in a value of 1. Set algebra is necessary for manipulating these lists also. However, unlike the deductive method, the equations for computing the output lists of a device from its input lists are dependent only on the function performed by the device and not on the signal values.

In concurrent simulation, any fault that causes the inputs, outputs, or internal state of a device to be different from their normal values is represented conceptually by a copy of the device. During simulation, if the inputs, outputs, and state of a faulty copy become identical to those of the fault-free copy, the faulty copy is deleted. Thus, faulty copies are created and deleted during simulation. The evaluation of faulty copies is essentially the same as fault-free copies, and no set algebra is involved. Concurrent simulation can also handle a large number of faults simultaneously.

It is interesting to note that all the above methods, except parallel simulation, use some form of data compression for storing faulty signal values. On the other hand, parallel simulation attempts to compute simultaneously the fault-free signal value and a number of faulty signal values associated with each lead in the circuit.

II. NUMBER OF LOGIC VALUES

Three-valued logic systems have been widely used for analyzing essentially binary systems.[13,14] Three logic values are also used in logic simulation, where 0 and 1 represent the two discrete values being modeled and a third value, u, denotes that a particular value is unknown.

Recently, tri-state busing has become a widespread technique used in many LSI designs. Difficulties in modeling effects associated with CMOS technology have been reported.[15] One effect is the memory associated with a disabled bus. That is, the disabled bus remembers the previous logic value on the bus. A solution consists of adding special circuitry to regular gates, making possible the use of a simulator with only three logic values.[15] An alternate solution is the addition of three more logic values, namely z_0, z_1, and z, for representing the states of disabled buses, with previous value equal to 0, 1, and unknown, respectively.[16] Transistor-transistor logic (TTL) tri-state technology requires the addition of only one logic value, z.[16]

Bus contention, another typical, potentially destructive tri-state effect, cannot be modeled by added circuitry. A solution consists of adding one more logic value representing a conflict state, a, as shown in the following example.

Consider a driver inverter and a bus configuration in TTL tri-state technology (Fig. 1). When line e is enabled, the gate operates as an inverter, when e is disabled the output of the gate is in a high-impedance state. When used in a bus configuration, two enabled inverters create a conflict (bus contention), if they are in opposite states. The set of logic values $\{0, 1, u, a, z\}$ is sufficient to model these effects, since tri-state devices in TTL technology do not have the memory property mentioned above.

Table I shows how the bus configuration of Fig. 1 can be simulated using the above set of five logic values. Since the bus will be connected to the output of drivers, which can produce four out of the five logic values, only four logic values are used for modeling the bus.

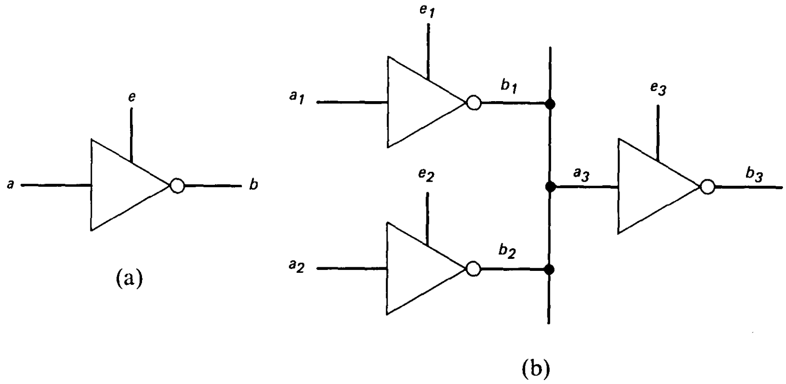


Fig. 1—(a) TTL Driver-inverter. (b) Bus configuration.

Table I—(a) Tristate inverter output
(b) State of tristate bus

				e_i		
		0	1	*u*	*a*	*z*
	0	*z*	1	*u*	*u*	*u*
	1	*z*	0	*u*	*u*	*u*
a_i	*u*	*z*	*u*	*u*	*u*	*u*
	a	*z*	*u*	*u*	*u*	*u*
	z	*z*	*u*	*u*	*u*	*u*

(a)

			b_1		
		0	1	*u*	*z*
	0	0	*a*	*u*	0
b_2	1	*a*	1	*u*	1
	u	*u*	*u*	*u*	*u*
	z	0	1	*u*	*z*

(b)

If an ordinary gate could be connected directly to a bus, the model should allow five logic values for the gate inputs, but requires only three logic values for its output. Table II shows the behavior of such an AND gate with inputs x and y, and output t.

The use of larger sets of logic values, though necessary to correctly model modern technology, has a serious impact on the method of simulation used. The following sections deal with this problem.

2.1 Parallel simulation

When using a switching algebra (i.e., two logic values) parallel simulation can be implemented by associating one computer word with each line in the circuit. One bit of this word represents the signal value on a line in the fault-free circuit and the remaining bits represent values on the same line in the presence of different single faults.

Table II—AND gate with five input logic values

				x		
		0	1	*u*	*a*	*z*
	0	0	0	0	0	0
	1	0	1	*u*	*u*	*u*
y	*u*	0	*u*	*u*	*u*	*u*
	a	0	*u*	*u*	*u*	*u*
	z	0	*u*	*u*	*u*	*u*

Table III—Coding for three logic values

α_i^0	α_i^1	a_i
0	0	unknown
0	1	1
1	0	0
1	1	unused

When a three-valued system is used, each of the circuits simulated in parallel must be coded using two binary digits. A commonly used method consists of associating two words with each line a, namely the 0-word, α^0, and the 1-word, α^1.[17] The coding used is shown in Table III, where the subscript i refers to the ith bit of each word.

Examples of its use are shown in Fig. 2. Here, and elsewhere in this paper, lower-case roman letters are used to denote leads and Greek letters represent words. For the gates of Fig. 2, we have

$$\begin{cases} \gamma^0 = \alpha^0 + \beta^0 \\ \gamma^1 = \alpha^1 \cdot \beta^1 \end{cases}$$

$$\begin{cases} \delta^0 = \alpha^0 \cdot \beta^0 \\ \delta^1 = \alpha^1 + \beta^1 \end{cases}$$

$$\begin{cases} \xi^0 = \alpha^1 \\ \xi^1 = \alpha^0, \end{cases}$$

where · and + represent the bitwise AND and OR operations on complete words.

This method can be extended for any number of logic values. For instance, consider the AND gate of Fig. 3, using the set of logic values $\{0, 1, u, a, z\}$. The binary coding scheme requires three computer words for each line, and three codes (out of eight) are not used. For

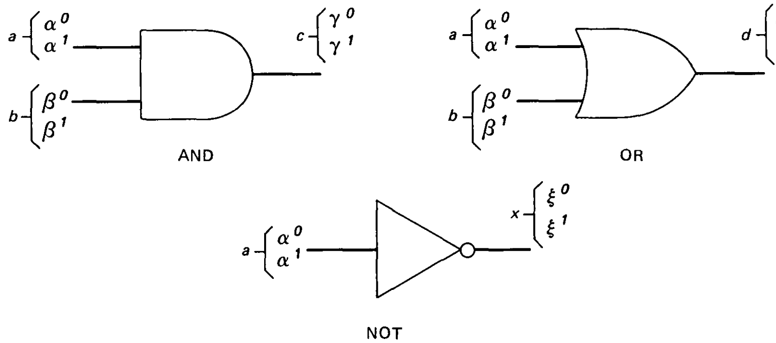


Fig. 2—Use of coding to represent signal values on gates.

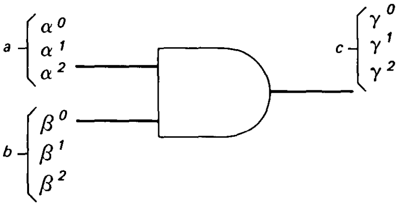


Fig. 3—AND gate representation for five logic values.

any choice of code, it is possible to calculate the gate output from switching expressions of the following form:

$$\gamma^0 = f(\alpha^0, \alpha^1, \alpha^2, \beta^0, \beta^1, \beta^2)$$
$$\gamma^1 = g(\alpha^0, \alpha^1, \alpha^2, \beta^0, \beta^1, \beta^2)$$
$$\gamma^2 = h(\alpha^0, \alpha^1, \alpha^2, \beta^0, \beta^1, \beta^2).$$

The original set of logic values and operations do not constitute a Boolean algebra. The coding scheme establishes a mapping of non-Boolean functions into switching operations that can be applied on full computer words, thus, allowing parallel processing.

By using a coding of $n - 1$ variables to represent n logic values, it is possible to obtain simpler equations for computing the outputs of gates. For example, consider the gate of Fig. 3 and a coding using four words α^0, α^1, α^a, and α^z to represent five logic values. The code is such that $\alpha_i^j = 1$, $j = 0$, 1, a, or z, iff $a_i = j$. All the variables will be zero, if and only if $a_i = u$. With this coding, the following equations are obtained for the gate of Fig. 3:

$$\gamma^0 = \alpha^0 + \beta^0$$
$$\gamma^1 = \alpha^1 \cdot \beta^1$$
$$\gamma^a = 0$$
$$\gamma^z = 0.$$

This type of coding can be used for any number of logic values.

2.2 Multilist simulation

It has been shown that for a three-valued logic system, three lists X^0, X^1, and X^u can be associated with each line x.[10,12] Each list, X^i, ($i = 0, 1, u$) represents the faults which cause line x to have the value i. For each line x, all the lists X^i are disjoint and any list is the complement of the union of the other two (i.e., the union of the three lists is the set of all faults being simulated).

For the gates of Fig. 2, we have

$$\begin{cases} C^1 = A^1 \cap B^1 \\ C^0 = A^0 \cup B^0 \end{cases} \qquad C^u = \overline{C^0 \cup C^1}$$

$$\begin{cases} D^1 = A^1 \cup B^1 \\ D^0 = A^0 \cap B^0 \end{cases} \qquad D^u = \overline{D^0 \cup D^1}$$

$$\begin{cases} X^1 = A^0 \\ X^0 = A^1 \end{cases} \qquad X^u = \overline{X^0 \cup X^1},$$

where $^{-}$, $\cup$, and $\cap$, are set complement, union, and intersection, respectively.

When five logic values are used, we need five lists; for instance, A^0, A^1, A^u, A^a, and A^z are associated with line a.

For the AND gate of Fig. 3, we have

$$\begin{aligned} C^0 &= A^0 \cup B^0 \\ C^1 &= A^1 \cap B^1 \\ C^a &= \{\ \ \} \\ C^z &= \{\ \ \} \\ C^u &= \overline{(C^0 \cup C^1 \cup C^z \cup C^a)} = \overline{A^0 \cup B^0 \cup A^1 \cap B^1}. \end{aligned}$$

For the inverter of Fig. 1, we have

$$\begin{aligned} B^0 &= E^1 \cap A^1 \\ B^1 &= E^1 \cap A^0 \\ B^z &= E^0 \\ B^u &= E^u \cup E^a \cup E^z \cup (E^1 \cap (\overline{A^0 \cup A^1})) \\ B^a &= \{\ \ \}, \end{aligned}$$

and for the bus configuration of Fig. 1

$$\begin{aligned} A_3^0 &= (B_1^0 \cap B_2^0) \cup (B_1^0 \cap B_2^z) \cup (B_2^0 \cap B_1^z) \\ A_3^1 &= (B_1^1 \cap B_2^1) \cup (B_1^1 \cap B_2^z) \cup (B_2^1 \cap B_1^z) \\ A_3^u &= B_1^u \cup B_2^u \\ A_3^z &= B_1^z \cap B_2^z \\ A_3^a &= (B_2^0 \cap B_1^1) \cup (B_2^1 \cap B_1^0). \end{aligned}$$

This method can be generalized to any gate type and any number of logic values as follows: Let us assume that we wish to simulate a function $f(x_1, x_2, \cdots, x_n)$, where each input and the output may assume any one of k values, denoted by $1, 2, \cdots, k$, and that the function is defined by a table which specifies the values of f for all combinations of values of x_i.

(i) We associate a variable x_i^j with each variable x_i, such that $x_i^j = 1$ if and only if $x_i = j$, $1 \le j \le k$. Similarly, we associate k variables f^j with f.

(ii) For each i, $1 \le i \le k$, we obtain an expression

$$f^i = \sum P_j,$$

where P_j are products of literals x_l^m, representing all combinations of values for which $f = i$. For example, if the table has an entry

$$x_1 = 1,\, x_2 = 0,\, x_3 = z,\; f = 1,$$

the expression for f^1 will contain the term

$$x_1^1 x_2^0 x_3^z.$$

(iii) Replace all lower-case letters in the equation for f^i by the corresponding upper-case letters, representing lists, and retain the superscripts and subscripts. Replace products by intersection and sums by union.

2.3 Deductive simulation

Deductive simulation is well defined for two logic values, and is also applicable to three logic values with some loss of information.[4] Specifically, if the signal value in the normal circuit is known, (i.e., 0 or 1), but the value in the presence of a fault α is unknown (denoted by u), the fault α is included in the fault list as a star fault;[4,18] that is, it is unknown whether the particular signal value in the presence of the fault α will be different from the fault-free value. It was shown in Refs. 10 and 12 that there are cases where the circuit value in the normal circuit may be unknown, but the value in the presence of a fault may be known. Since the deductive method cannot represent this case, the results obtained may be less accurate than with other methods.[10,12]

A modification of the deductive method that leads to accurate three-valued fault simulation was presented in Ref. 10. It uses the coding of Table III for representing each signal value by a pair of binary variables. A pair of equations can then be derived, as in Section 2.1, for computing the coded outputs for each gate type. These equations can be viewed as defining a transformation of the original circuit with three signal values into two circuits that will have only binary signals. These two circuits can be simulated using the two-valued deductive method. The fault-free and faulty signal values on any lead in the original circuit can be determined from the signal values and fault lists associated with the corresponding pair of leads in the transformed circuits.

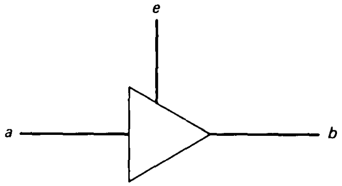


Fig. 4—Tristate bus driver.

The same approach can be used for performing deductive fault simulation with any number of logic values. If k logic values are to be simulated, $\lceil \log_2 k \rceil$ binary variables will be used to represent them, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. The equations for the coded outputs of different gate types can be derived from their truth tables, and used in deductive simulation.

As as example, consider the bus driver of Fig. 4 to be simulated with four logic values, namely, 0, 1, z (high impedance) and u (unknown). The behavior of the device is specified in Table IV.

Using the coding of Table V, we shall represent the signals a, e, and b of the bus driver by a_0 and a_1, e_0 and e_1, and b_0 and b_1. The output equations b_0 and b_1 can be derived from Tables IV and V.

$$b_0 = e_0 \cdot \bar{e}_1 + a_0 \cdot \bar{e}_0 \cdot e_1$$

$$b_1 = e_0 \cdot \bar{e}_1 + a_1 \cdot \bar{e}_0 \cdot e_1$$

For any combination of input values and fault lists, the output values and fault lists can be computed as in Ref. 19.

Denoting the fault list associated with each variable by the corresponding upper-case letter, let the input values and fault lists for the circuit of Fig. 4 be as follows:

$$a_0 = 0; \quad A_0 = \{1, 3\}$$

$$a_1 = 1; \quad A_1 = \{3\}$$

$$e_0 = 0; \quad E_0 = \{2, 4\}$$

$$e_1 = 1; \quad E_1 = \{4, 5\}$$

Table IV—Behavior of bus driver

		e			
		0	1	z	u
a	0	z	0	u	u
	1	z	1	u	u
	z	z	z	u	u
	u	z	u	u	u

Table V—Coding for tristate devices

x_0	x_1	x
0	0	u
0	1	1
1	0	0
1	1	z

Let us assume that all the faults being considered are external to the device, and we wish to propagate the effects of the faults through the device. The input conditions are: $a = 1$, $e = 1$. Since the fault 1 is contained only in the fault list A_0, it will cause a_0 to become 1, and therefore a to become z. On the other hand, fault 3 is contained both in A_0 and A_1, and will cause both a_0 and a_1 to be inverted; the value of a in the presence of fault 3 will be 0. Similarly, fault 2 will result in $e = z$, fault 4 in $e = 0$, and fault 5 in $e = u$.

For the above set of values, the output values and fault lists can be computed using the equations for b_0 and b_1 and the method presented in Ref. 19 as follows:

$$b_0 = 0$$

$$b_1 = 1$$

$$B_0 = (A_0 \cap \overline{E_0} \cap \overline{E_1}) \cup (E_0 \cap E_1) = \{1, 3, 4\}$$

$$B_1 = (A_1 \cup E_0 \cup E_1) \cap \overline{(E_0 \cap E_1)} = \{2, 3, 5\}.$$

Denoting the value of b in the presence of fault α by $b(\alpha)$, we can obtain the following faulty values from the values b_0 and b_1 and fault lists B_0 and B_1.

$$b(1) = z; \quad b(2) = u; \quad b(3) = 0; \quad b(4) = z; \quad b(5) = u.$$

These can be verified by computing the output for each faulty combination of inputs using Table IV.

The modified deductive method discussed above does not lose any information about the normal and faulty circuits and is as accurate as any of the other methods. It requires only $\lceil \log_2 n \rceil$ lists compared to the n lists needed for multilist simulation. However, fault list computations depend on signal values and may be more complex than in the multilist method.

2.4 Concurrent simulation

There is no limitation on the number of logic values in this simulation method since faulty and fault-free circuits are treated independently. As long as the primitive elements of the circuit are well defined, the evaluation of faulty circuits presents no difficulty.

2.5 Summary of results

The results of Section II are represented in Table VI.

Deductive simulation with three logic values (indicated by * in Table VI) requires the introduction of the concept of star faults. Deductive fault simulation for more than three logic values (indicated by † in Table VI) could be defined by using a transformed circuit as proposed in Section 2.3. However, the complexity of such a procedure does not seem to justify its use.

From the point of view of simulating more than three logic values, concurrent simulation represents the simplest, most flexible simulation method.

III. MODELING LEVELS

Three levels of modeling and their effects on the simulation method used will be considered: gates, higher-level primitives, and user-defined functions.

3.1 Gate-level simulation

All the fault-simulation methods presented here were initially developed for simulating circuits modeled at the gate level. Therefore, none of the methods presents any problem, provided only two (or three) logic values are to be simulated. The differences due to the number of logic values needed have already been discussed in Section II, and the effects of detailed timing analysis are discussed in Section IV.

3.2 High-level primitives

It is often convenient to model devices such as flip-flops, multiplexors, counters, and shift registers as high-level primitives rather than as interconnections of gates. For purposes of simulation, such devices may be described by tables, Boolean equations, or algorithms. The

Table VI—Summary of Results: logic values

	Parallel	Multilist	Deductive	Concurrent
Switching algebra	One word per line	Two lists per line	Well defined	Well defined
Three logic values	Two words per line	Three lists per line	Well defined but pessimistic*	Well defined
Five logic values	Three words per line	Five lists per line	Undefined†	Well defined
n logic values	$\lceil \log_2 n \rceil$ words per line	n lists per line	Undefined†	Well defined

* Deductive simulation with three logic values.

† Deductive fault simulation for more than three logic values.

type of representation that is most convenient to use will usually depend on the simulation method.

3.2.1 Parallel simulation

Several solutions are possible. The input values for individual faults can be determined from the input word(s), and the outputs of the high-level primitive can be evaluated for each case. The output values must then be packed so that the parallel simulation method may be used elsewhere. While this approach may be satisfactory for predominantly gate-level circuits which also contain a few high-level primitives, the overhead associated with converting to single-fault simulation and back to parallel simulation may not be acceptable.

When only two logic values are involved, the primitive can be represented by Boolean (switching) expressions. The operators in these expressions can be treated exactly like gates in parallel simulation of gate-level circuits. When more than two logic values are simulated, the description of the primitive may be in the form of tables. Using a coding of the type described in Section II, switching algebraic expressions for the coded output words (i.e., the 0-word, 1-word, etc.) can be obtained in terms of the coded words associated with the inputs and state variables. These equations can then be used to compute the coded output words.

3.2.2 Multilist simulation

The function realized by a high-level primitive can be represented by tables. From these tables, equations for the output lists in terms of lists associated with inputs and state variables can be obtained as discussed in Section 2.2 and used for simulation.

3.2.3 Deductive simulation

As in the case of parallel simulation, one approach is to simulate each high-level primitive for one fault at a time and use the results to construct output fault lists for use outside the primitive. Alternatively, outputs and the next state values of state variables may be represented by Boolean equations, which are used for fault-list computations in the same manner as gates. When more than two logic values are to be simulated, a binary coding can be used as discussed in Section 2.3, and equations for each coded bit can be used for fault-list computations.

Another possibility is to use tables that specify the fault-list computations for every combination of input values.[4] These tables can be constructed from the tables specifying the primitive, as shown by the example of Fig. 5.

The behavior of the function is represented in Table VII, where p_1, q_1 are the initial states of the flip-flops and p_2, q_2 are the next states.

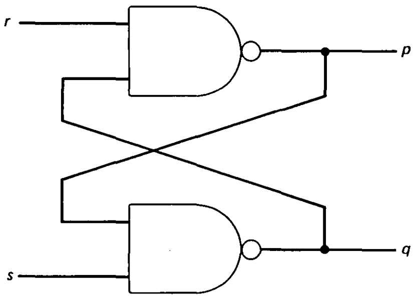


Fig. 5—NAND SR latch.

Table VII—Behavior of SR latch

	p_1	q_1	r	s	p_2	q_2
1	0	1	0	0	1	1
2	0	1	0	1	1	0
3	0	1	1	0	0	1
4	0	1	1	1	0	1
5	1	0	0	0	1	1
6	1	0	0	1	1	0
7	1	0	1	0	0	1
8	1	0	1	1	1	0
9	1	1	0	0	1	1
10	1	1	0	1	1	0
11	1	1	1	0	0	1
12	1	1	1	1	u	u

The table does not include 00 as an initial state because it cannot be produced directly.

The fault-list propagation is summarized in Table VIII and does not include local faults. P_1, Q_1, R, S, P_2, and Q_2 are the fault lists associated with p_1, q_1, r, s, p_2, and q_2. The star faults in the table are to be added to both the fault lists, P_2 and Q_2. We shall demonstrate the procedure used for deriving Table VIII, by showing how line 1 of the table was obtained.

Consider line 1 of Table VII. To get a change in p_2, we need a change into lines 3, 4, 7, or 11 and for a change in q_2, we need a change into lines 2, 6, 8, or 10, which produces:

$$\begin{aligned} P_2 &= \bar{P}_1\bar{Q}_1R\bar{S} \cup \bar{P}_1\bar{Q}_1RS \cup P_1Q_1R\bar{S} \cup P_1\bar{Q}_1R\bar{S} \cup \bar{P}_1Q_1(\Phi) \\ &= R\bar{S} \cup \bar{P}_1R \\ Q_2 &= \bar{P}_1\bar{Q}_1\bar{R}S \cup P_1Q_1RS \cup P_1Q_1\bar{R}S \cup P_1\bar{Q}_1\bar{R}S \cup \bar{P}_1Q_1(\Phi) \\ &= \bar{R}S \cup Q_1S, \end{aligned}$$

where juxtaposition represents intersection. Since a change to line 12 is needed to cause p_2 and q_2 to become unknown, the star faults are

Table VIII—Fault-list equations for SR latch

	P_2	Q_2	Star Faults
1	$R\bar{S}\cup\bar{P}_1R$	$\bar{R}S\cup Q_1S$	$P_1\bar{Q}_1RS$
2	$RS\cup\bar{P}_1R$	$S\cup\bar{P}_1R$	$P_1\bar{Q}_1R\bar{S}$
3	$R\cup Q_1S$	$RS\cup Q_1S$	$P_1\bar{Q}_1\bar{R}S$
4	$R\cup Q_1\bar{S}$	$R\bar{S}\cup Q_1\bar{S}$	$P_1\bar{Q}_1\bar{R}\bar{S}$
5	$R\bar{S}\cup RP_1$	$\bar{R}S\cup\bar{Q}_1S$	$\bar{P}_1Q_1RS$
6	$RS\cup P_1R$	$S\cup P_1R$	$\bar{P}_1Q_1R\bar{S}$
7	$R\cup\bar{Q}_1S$	$RS\cup R\bar{Q}_1$	$\bar{P}_1Q_1\bar{R}S$
8	$\bar{R}S\cup P_1\bar{R}$	$S\cup P\bar{R}$	$\bar{P}_1Q_1\bar{R}\bar{S}$
9	$R\bar{S}\cup Q_1R$	$\bar{R}S\cup SP_1$	$\bar{P}_1\bar{Q}_1RS$
10	$RS\cup Q_1R$	$S\cup Q_1R$	$\bar{P}_1\bar{Q}_1R\bar{S}$
11	$R\cup Q_1S$	$RS\cup Q_1S$	$\bar{P}_1\bar{Q}_1\bar{R}_1S$
12	{}	{}	$\bar{P}_1\bar{Q}_1\bar{R}\bar{S}$

given by

$$P_1\bar{Q}_1RS.$$

We have used $\bar{P}_1Q_1$, which corresponds to the initial state 00 in the faulty circuit, as a don't-care state (Φ) to simplify the expressions.

3.2.4 Concurrent simulation

In concurrent simulation, the same method is used to evaluate fault-free and faulty circuit signal values. Therefore, no transformations of representation are necessary, and any representation that leads to efficient simulation may be chosen.

3.3 User-defined functions

Our discussion of Section 3.2 also applies to user-defined functions. The main difference is that the tables or equations used for representing the functions must be generated from descriptions in a high-level language such as the function definition language in LAMP.[20]

A typical construct in such a language is the cause-effect statement. Such statements can be nested to many levels. The techniques discussed in Section 3.2 can be used for simulating user-defined functions using the parallel, multilist, or deductive method by first replacing cause-effect statements by equivalent equations. For example, the statement

$$\text{if } x \quad \text{then} \quad z = a \quad \text{else} \quad z = b$$

can be replaced by

$$z = a\cdot x + b\cdot\bar{x} + a\cdot b.$$

The redundant term $a\cdot b$ has been introduced to produce the correct result $z = 1$ for the case $a = b = 1$ and $x = u$.[21] Otherwise, the pessimistic result $z = u$ will be produced for this case.

Concurrent simulation does not require the transformation of cause-effect statements into equations. For each fault and each combination of inputs and state, only those computations enabled by the conditions need be performed. The operations in a function definition need not be restricted to logical operations. Therefore, it is not necessary to generate Boolean equations corresponding to arithmetic operations, as would be necessary in the other methods considered. Thus, it appears that the concurrent method would allow simulation of functions defined at a higher level than is possible with the other methods.

3.4 Summary of results

The results of this section are summarized in Table IX. The concurrent method is clearly superior in its ability to simulate different levels of models.

IV. TIMING

In this section, we study different effects related to timing, and their impact on the four simulation methods under consideration. We shall consider the effects of different rise and fall times associated with signal changes, suppression of short pulses to model inertial delays, and the simulation of faults which affect the magnitude of delays associated with devices. We shall restrict our discussion to logic simulation with two and three logic values.

4.1 Rise and fall times

The delays associated with 0 to 1 and 1 to 0 transitions of a signal, called here the rise and fall times, are not necessarily equal.[22,23] All the methods of fault simulation under discussion simulate a number of signals simultaneously, some of which may be rising and some falling. To simulate this effect accurately, a mechanism is necessary for allowing rising and falling signals to change at different times.

Let t_0 be a time before any change occurs on the line under consideration. Due to differences in the rise and fall times, signal changes may occur on the line at times t_1 and t_2, where $t_0 < t_1 \leq t_2$. Thus, at time t_2, all signal changes associated with the particular event would have occurred. The effect of different rise and fall times can be

Table IX—Summary of results: modeling levels

	Parallel	Multilist	Deductive	Concurrent
Gate level	1	1	1	1
Higher level primitives	2	2	2	1
User defined functions	2	2	2	1

Note: 1 = No transformations required.
2 = Transformation into equations required.

simulated accurately by computing the values of the signals at time t_1 based on the values at t_0 and t_2, namely, the initial and final values for the particular set of transitions.

In the following sections, we shall consider four simulation methods and examine the results produced by their different models at three points in time, namely, before $t_1(t_0)$, between t_1 and $t_2(t_1)$, and after $t_2(t_2)$.

4.1.1 Parallel simulation

Let ξ_0, ξ_1, and ξ_2 be the words associated with a line x at times t_0, t_1, and t_2, in two-valued parallel simulation. In three-valued simulation, two words denoted by superscripts 0 and 1 will be associated with the line for each of the above times, and the coding of Table III will be used.

Case 1: Rise time < fall time. We have

$$\xi_1 = \xi_0 + \xi_2$$

for two-valued simulation, and

$$\xi_1^1 = \xi_0^1 + \xi_2^1$$
$$\xi_1^0 = \xi_0^0 \cdot \xi_2^0$$

for three-valued simulation where + and · represent bitwise OR and AND performed on full words.

Case 2: Fall time < rise time. We have

$$\xi_1 = \xi_0 \cdot \xi_2$$

for two-valued simulation, and

$$\xi_1^1 = \xi_0^1 \cdot \xi_2^1$$
$$\xi_1^0 = \xi_0^0 + \xi_2^0$$

for three-valued simulation.

The preceding formulas can be verified by checking all nine possible transitions between the set {0, 1, u} and itself.

4.1.2 Three-list method

Let X_0^i, X_1^i, and X_2^i be the i-lists at times t_0, t_1, and t_2, defined above, for $i = 0, 1, u$. Using the same arguments as in Section 4.1.1, we obtain the lists for time t_1 as given below.

Case 1: Rise time < fall time. We have

$$X_1^1 = X_0^1 \cup X_2^1$$
$$X_1^0 = X_1^0 \cap X_2^0.$$

Case 2: Fall time < rise time. We have

$$X_1^1 = X_0^1 \cap X_2^1$$

$$X_1^0 = X_0^0 \cup X_2^0 .$$

In both cases, $X_1^u = (\overline{X_1^0 \cup X_1^1})$.

4.1.3 Deductive simulation

Deductive simulation with different rise and fall times has been discussed by Kjelkerud and Thessen.[24] Here we present an alternate method.

Let the times t_0, t_1, and t_2 be as defined earlier and let x_i and X_i represent the signal values and fault lists at those times, $i = 0, 1, 2$. If the rise time is less than the fall time, all 0 to 1 transitions will occur at t_1. Therefore, we have

$$x_1 = x_0 + x_2 .$$

Similarly, if fall time < rise time, 1 to 0 transitions will occur at t_1, and $x_1 = 1$ if and only if it remains at 1 throughout the transitions. Therefore, for this case

$$x_1 = x_0 \cdot x_2 .$$

The fault lists X_1 at time t_1 for different signal changes in the fault-free circuit and different relative values of rise and fall times can be determined from these equations. They are summarized in Table X.

4.1.4 Concurrent simulation

In fact, concurrent simulation is a trivial case, because fault-free and and faulty circuits are simulated independently. Rising and falling edges will still occur in distinct event waves, but the treatment of these events is individual.

4.2 High-frequency rejection

High-frequency rejection consists of eliminating short pulses for modeling the effect of inertial delays. We consider events occurring at

Table X—Fault-list equations for handling different rise and fall times

Fault-Free Circuit	Rise Time < Fall Time	Fall Time < Rise Time
Rising edge $x_0 = 0; x_2 = 1$	$x_1 = 1$ $X_1 = \overline{X}_0 \cap X_2$	$x_1 = 0$ $X_1 = X_0 \cap \overline{X}_2$
Falling edge $x_0 = 1; x_2 = 0$	$x_1 = 1$ $X_1 = X_0 \cap \overline{X}_2$	$x_1 = 0$ $X_1 = \overline{X}_0 \cap X_2$
Constant one $x_0 = 1; x_2 = 1$	$x_1 = 1$ $X_1 = X_0 \cap X_2$	$x_1 = 1$ $X_1 = X_0 \cup X_2$
Constant zero $x_0 = 0; x_2 = 0$	$x_1 = 0$ $X_1 = X_0 \cup X_2$	$x_1 = 0$ $X_1 = X_0 \cap X_2$

times t_0, t_1, and t_2, where $t_0 < t_1 \leq t_2$ and the logic values at these times. If $t_2 - t_1$ is less than the magnitude of the inertial delay, then the change at t_1 must be rejected to suppress short pulses and the value between t_1 and t_2, x_1, will be replaced by a corrected logic value, x_{1n}. If x_0, x_1, and x_2 are the computed signal values at t_0, t_1, and t_2 respectively, and if $t_2 - t_1$ is less than the inertial delay, the corrected signal value at time t_1 is given by:

$$x_{1n} = x_0 x_1 + x_1 x_2 + x_0 x_2.$$

The method for performing high-frequency rejection can be derived from this equation.

In case there are more than two events within the range of the inertial delay, the treatment elaborated above must be repeated for each pair of events within that range. For example, consider three events occurring at times t_1, t_2, and t_3. The following triples will be considered: (t_0, t_1, t_2), (t_0, t_1, t_3), (t_1, t_2, t_3), which represent three pairs of events.

4.2.1 *Parallel simulation*

For two-valued parallel simulation, the word ξ_1 has to be replaced by

$$\xi_{1n} = \xi_0 \xi_1 + \xi_1 \xi_2 + \xi_0 \xi_2$$

and for three-valued parallel simulation, ξ_1^0 and ξ_1^1 are replaced by

$$\xi_{1n}^1 = \xi_0^1 \xi_1^1 + \xi_1^1 \xi_2^1 + \xi_0^1 \xi_2^1$$
$$\xi_{1n}^0 = \xi_0^0 \xi_1^0 + \xi_1^0 \xi_2^0 + \xi_0^0 \xi_2^0 .$$

The coding defined in Table III was used to obtain the above equations for three-valued parallel simulation.

4.2.2 *Three-list methods*

Using the equation for x_{1n} given above, we obtain the following fault-list equations:

$$X_{1n}^0 = (X_0^0 \cap X_1^0) \cup (X_1^0 \cap X_2^0) \cup (X_0^0 \cap X_2^0)$$
$$X_{1n}^1 = (X_0^1 \cap X_1^1) \cup (X_1^1 \cap X_2^1) \cup (X_0^1 \cap X_2^1)$$
$$X_{1n}^u = (\overline{X_1^0 \cup X_1^1}).$$

4.2.3 *Deductive simulation*

The deductive fault list X_{1n} can be obtained from the equation for the new signal value x_{1n}, the values of x_0, x_1, and x_2, and the associated fault lists. The fault list X_{1n} can be computed in the same manner as fault propagation through functional blocks.[19] In fact, high-frequency rejection may be thought of as being performed by a filter whose equation is given above.

The fault-list computations for the eight possible patterns of x_0, x_1, and x_2 are summarized in Table XI.

As an example, consider the case $x_0 = 0$, $x_1 = 1$, $x_2 = 1$(line 4 in Table XI). We have

$$x_{1n} = x_0x_1 + x_1x_2 + x_0x_2 = a + b + c.$$

The fault lists associated with the terms $a = x_0x_1$, $b = x_1x_2$, and $c = x_0x_2$ for the specified values are:

$$A = X_0 \cap \bar{X}_1; \quad B = X_1 \cup X_2; \quad C = X_0 \cap \bar{X}_2 .$$

Therefore,

$$X_{1n} = \bar{A} \cap B \cap \bar{C} = (\bar{X}_0 \cup X_1) \cap (X_1 \cup X_2) \cap (\bar{X}_0 \cup X_2)$$
$$= (\bar{X}_0 \cap X_1) \cup (\bar{X}_0 \cap X_2) \cup (X_1 \cap X_2) .$$

4.2.4 Concurrent simulation

In this case, each faulty signal value is computed separately. Therefore, high-frequency rejection can be performed on each signal individually, using the equation for x_{1n} given in the preceding section.

4.2.5 Suppression of short-duration detections

We have considered the suppression of short pulses produced independently by each fault-free or faulty signal value. However, we did not consider the case of a short pulse of detection, when neither the faulty nor the fault-free signal incurs a pulse. This is illustrated by the case where $x_0 = 1$, $x_1 = 1$, $x_2 = 0$, for the fault-free signal and $x_0 = 1$, $x_1 = 0$, $x_2 = 0$, for the faulty signal. This causes a short detection between t_1 and t_2. For deductive simulation, this short detection may be eliminated by using the formula

$$X_{1n} = (X_0 \cap X_1) \cup (X_0 \cap X_2) \cup (X_1 \cap X_2)$$

independently of the fault-free signal pattern and after the high-

Table XI—Fault-list equations for high-frequency rejection

x_0	x_1	x_2	x_{1n}	X_{1n}
0	0	0	0	$(X_0 \cap X_1) \cup (X_1 \cap \underline{X}_2) \cup (X_0 \cap \underline{X}_2)$
0	0	1	0	$(X_0 \cap \underline{X}_1) \cup (\underline{X}_1 \cap \bar{X}_2) \cup (X_0 \cap \bar{X}_2)$
0	1	0	0	$(\underline{X}_0 \cap \bar{X}_1) \cup (\bar{X}_1 \cap X_2) \cup (\underline{X}_0 \cap X_2)$
0	1	1	1	$(\underline{\bar{X}}_0 \cap X_1) \cup (X_1 \cap X_2) \cup (\underline{\bar{X}}_0 \cap X_2)$
1	0	0	0	$(\bar{X}_0 \cap \underline{X}_1) \cup (\underline{X}_1 \cap X_2) \cup (\bar{X}_0 \cap X_2)$
1	0	1	1	$(X_0 \cap \bar{X}_1) \cup (\bar{X}_1 \cap \underline{X}_2) \cup (X_0 \cap \underline{X}_2)$
1	1	0	1	$(X_0 \cap X_1) \cup (X_1 \cap \bar{X}_2) \cup (X_0 \cap \bar{X}_2)$
1	1	1	1	$(X_0 \cap X_1) \cup (X_1 \cap X_2) \cup (X_0 \cap X_2)$

frequency rejection has been performed. The term X_i is the set of faults detected at time t_i.

The same method may be used for all the other simulation algorithms described earlier.

4.3 Delay faults

A fault that affects the transport delay associated with a signal is called a delay fault. Consider a fault that causes a delay to change from d to d'. When the signal at the site of such a fault changes, the signal value corresponding to the particular faulty circuit must be delayed by d' instead of d.

4.3.1 Parallel simulation

Two aspects of delay faults must be considered: injection of delay faults and the propagation of the effects of delay faults. Let us assume that a gate which is the site of a delay fault has been evaluated at time t, and the jth bit of the word represents the circuit with the delay fault. Let the normal and faulty delays be d and d', respectively, and let $d < d'$. Let $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $\mathbf{x}' = (x'_1, x'_2, \cdots, x'_n)$ be the vectors representing the old and new values, respectively. If for any $i \neq j$, $x'_i \neq x_i$, a vector $(x'_1, x'_2, \cdots, x'_{j-1}, x_j, x'_{j+1}, \cdots, x'_n)$ will be scheduled to be applied to the gate output at time $t + d$. If $x'_j \neq x_j$, the vector $\mathbf{x}'$ will be scheduled to be applied to the gate output at time $t + d'$. Similarly, if $d' < d$ and $x'_j \neq x_j$, the vector $(x_1, x_2, \cdots, x_{j-1}, x'_j, x_{j+1}, \cdots, x_n)$ will be scheduled for time $t + d'$. If $x'_i \neq x_i$ for any $i \neq j$, then the vector $\mathbf{x}'$ will be scheduled to be applied at time $t + d$. The vectors for updating at the different times can be obtained from the old and new vectors by appropriate masks and logical operations.

From the above discussion it should be clear that the effect of delay faults is to cause the signal values on the same lead in the presence of delay faults to change at different times. When one or more values in a vector change, the gates to which the signal fans out in the fault-free and all faulty circuits are scheduled for evaluation in parallel. Therefore, no special treatment is necessary for propagating delay faults. Since any signal change at the inputs of a device, faulty or fault-free, will cause an evaluation of the device (faulty and fault-free), delay faults will tend to increase the number of evaluations required.

4.3.2 Three-list method

The equations required for simulating delay faults using the three-list and deductive methods can be derived by treating the delay fault as an internal fault in a functional block. These equations can then be used for simulating delay faults without explicitly modeling them as faults in functions.

Let α be a delay fault which causes the delay associated with a signal to change from d to d'. Let f_α be a fault variable,[19] which has the fault-free value of 0, but the fault α causes it to become 1. We shall represent the input and the output of the function used for modeling the delay fault by x and z, respectively. We assume that the evaluation is being done at time $t = 0$, and the value of the input x, t_1 units of time earlier is represented by $x(-t_1)$. Two different cases must be considered:

Case 1: $d < d'$. The value of z at time d is given by

$$\begin{array}{c} \text{if } f_\alpha \text{ then } z = x(d - d') \\ \text{else } z = x, \end{array}$$

which can be transformed into the equation

$$z(d) = f_\alpha \cdot x(d - d') + \bar{f}_\alpha \cdot x.$$

Case 2: $d' < d$. The value of z at time d' can be represented by a function as in Case 1, and transformed into the following equation:

$$z(d') = f_\alpha \cdot x + \bar{f}_\alpha \cdot x(d' - d).$$

The equations for the three-list method can be obtained from these equations using the method discussed in Section 2.2.

Case 1: $d < d'$

$$Z^1(d) = [\{\alpha\} \cap X^1(d - d')] \cup [X^1 \cap \overline{\{\alpha\}}]$$

$$Z^0(d) = [\overline{\{\alpha\}} \cap X^0] \cup [X^0 \cap X^0(d - d')] \cup [X^0(d - d') \cap \{\alpha\}]$$

$$Z^u(d) = [\overline{Z^1(d) \cup Z^0(d)}].$$

Case 2: $d' < d$

$$Z^1(d') = [\{\alpha\} \cap X^1] \cup [X^1(d' - d) \cap \overline{\{\alpha\}}]$$

$$Z^0(d') = [\overline{\{\alpha\}} \cap X^0] \cup [X^0 \cap X^0(d' - d)] \cup [X^0(d' - d) \cap \{\alpha\}]$$

$$Z^u(d') = [\overline{Z^1(d') \cup Z^0(d')}].$$

4.3.3 Deductive simulation

The functional equations derived in Section 4.3.2 can be used for deriving deductive fault lists for delay faults. The fault-list equations will depend on signal values as shown in Table XII.

4.3.4 Concurrent simulation

Since each fault is handled separately, the simulation of delay faults does not need any special processing.

Table XII—Fault-list equations for delay faults

x	$x(d-d')$	$z(d)$	$Z(d)$
0	0	0	$[X(d-d') \cap \{\alpha\}] \cup \left[X \cap \overline{\{\alpha\}}\right]$
0	1	0	$[\overline{X(d-d')} \cap \{\alpha\}] \cup \left[X \cap \overline{\{\alpha\}}\right]$
1	0	1	$\left[X \cap \overline{\{\alpha\}}\right] \cup [X(d-d') \cap \{\alpha\}] \cup [X \cap \overline{X(d-d')}]$
1	1	1	$[X(d-d') \cap \{\alpha\}] \cup [X \cap X(d'-d)] \cup \left[X \cap \overline{\{\alpha\}}\right]$

Case 1: $d < d'$

x	$x(d'-d)$	$z(d')$	$Z(d')$
0	0	0	$[X \cap \{\alpha\}] \cup \left[X(d'-d) \cap \overline{\{\alpha\}}\right]$
0	1	1	$[\bar{X} \cap X(d'-d)] \cup [\bar{X} \cap \{\alpha\}] \cup \left[X(d'-d) \cap \overline{\{\alpha\}}\right]$
1	0	0	$[\bar{X} \cap \{\alpha\}] \cup \left[X(d'-d) \cap \overline{\{\alpha\}}\right]$
1	1	1	$\left[X(d'-d) \cap \overline{\{\alpha\}}\right] \cup [X \cap \{\alpha\}] \cup [X \cap X(d'-d)]$

Case 2: $d' < d$

4.4 Summary of results

Changes in the fault-free and faulty values may occur at different times for the same line due to different rise and fall times and to delay faults. In the case of parallel simulation, a change in a single faulty or fault-free value on a line leads to computations involving the whole word (or pair of words). In the three-list and deductive methods, the addition or deletion of a single fault will require recomputation of complete lists. On the other hand, concurrent simulation treats each event, faulty or fault-free, independent of all other events and, therefore, should require less computation. High-frequency rejection is also simpler in concurrent simulation than in the other methods.

V. CONCLUSION

We have compared parallel, multilist, deductive, and concurrent simulation methods with regard to their ability to simulate more than two logic values, different levels of simulation, and accurate timing analysis. All the methods, except deductive, can handle any number of logic values without significant changes in the method. An extension of the deductive method to an arbitrary number of logic values is presented. Concurrent simulation appears to be the most convenient method of simulating an arbitrary number of logic values.

All the methods, except concurrent, require the transformation of functional descriptions of high-level devices into Boolean equations. No such transformation is required for concurrent simulation. In fact, it is not even necessary to restrict operations in functional descriptions to Boolean operations if concurrent simulation is used.

All the methods are capable of handling different rise and fall times, performing high-frequency rejection and simulating delay faults. Since concurrent simulation handles each event separately, these functions can be performed more easily and efficiently than the other methods.

In addition to the aspects discussed here, two factors that must be considered in selecting a simulation method are storage requirements and speed. A detailed analysis of the speed and the storage requirements of these methods is made in Ref. 11 based upon statistical data gathered from deductive simulation.

REFERENCES

1. S. Seshu, "The Logic Analyzer and Diagnosis Programs," Coordinated Science Laboratory, Rept. R-226, 1964.
2. S. Seshu, "On an Improved Diagnosis Program," IEEE Trans. Electronic Computers, *EC-14*, No. 1 (February 1965), pp. 76–9.
3. S. A. Szygenda, "TEGAS-2- Anatomy of a General Purpose Test Generation and Simulation System for Digital Logic," Proc. 9th ACM-IEEE Design Automation Workshop (June 1972), pp. 116–27.
4. D. B. Armstrong, "A Deductive Method of Simulating Faults in Logic Circuits," IEEE Trans. Computers, *C-21*, No. 5 (May 1972), pp. 464–71.
5. E. G. Ulrich and T. G. Baker, "Concurrent Simulation of Nearly Identical Digital Networks," Computer, *7*, No. 4 (April 1974), pp. 39–44.
6. H. Y. Chang et al., "Comparison of Parallel and Deductive Simulation Methods," IEEE Trans. Computers, *C-23*, No. 11 (November 1974), pp. 1132–8.
7. Y. H. Levendel and W. C. Schwartz, "Impact of LSI on Logic Simulation," Proc. of COMPCON, San Francisco, February 1978.
8. M. Abramovici, M. A. Breuer, and K. Kumar, "Concurrent Fault Simulation and Functional Level Modeling," Proc. 14th Design Automation Conference (June 1977), pp. 128–37.
9. F. Ozguner, W. E. Donath, and C. W. Cha, "On Fault Simulation Techniques," J. Design Automation and Fault Tolerant Computing," *3,* No. 2 (April 1979), pp. 83–92.
10. Y. H. Levendel and P. R. Menon, "Comparison of Fault Simulation Methods — Treatment of Unknown Signal Values," J. of Digital Systems, *4,* No. 4 (Winter 1980), pp. 443–59.
11. Y. H. Levendel and P. R. Menon, unpublished work.
12. Y. H. Levendel and P. R. Menon, "Unknown Signal Values in Fault Simulation," Proc. 9th International Symposium on Fault Tolerant Computing (June 1979), pp. 125–8.
13. M. Yoeli and S. Rinon, "Application of Ternary Algebra to the Study of Static Hazards," *J. ACM, 11,* No. 1 (January 1964), pp. 84–97.
14. E. B. Eichelberger, "Hazard Detection in Combinational and Sequential Switching Circuits," Proc. 5th Annual Symp. on Switching Circuit Theory and Logical Design (1964), pp. 111–20.

15. R. L. Wadsack, "Fault Modeling and Logic Simulation of CMOS and MOS Integrated Circuits," B.S.T.J. *57,* No. 5 (May–June 1978), pp. 1449–74.
16. Y. H. Levendel, P. R. Menon, and C. E. Miller, "Accurate Simulation Models for TTL Totempole and MOS Gates and Tristate Devices," B.S.T.J., *60,* No. 7 (September 1981), pp. 1271–87.
17. Y. H. Levendel and M. A. Breuer, "Vector Representation of Switching and Three-Valued Functions," Proc. Eighth Internat. Symp. on Multi-valued Logic (May 1978), pp. 163–70.
18. S. G. Chappell, C. H. Elmendorf, and L. D. Schmidt, "LAMP: Logic-Circuit Simulators," B.S.T.J., *53,* No. 8 (October 1974), pp. 1451–76.
19. P. R. Menon and S. G. Chappell, "Deductive Fault Simulation with Functional Blocks," IEEE Trans. Computers, *C-27,* No. 8 (August 1978), pp. 689–95.
20. S. G. Chappell et al., "Functional Simulation in the LAMP System," J. Design Automation and Fault Tolerant Computing, *1,* No. 3 (May 1977), pp. 203–16.
21. K. Wu, Ph.D. Dissertation, "Synthesis of Accurate and Efficient Functional Modeling Techniques for Performing Design Verification of VLSI Digital Circuits," Univ. of Texas, Austin, December, 1979.
22. S. G. Chappell and S. S. Yau, "Simulation of Large Asynchronous Logic Circuits Using an Ambiguous Gate Model," Proc. Fall Joint Computer Conf. (1971), pp. 651–61.
23. S. A. Syzgenda, D. M. Rouse, and E. W. Thompson, "A Model and Implementation of a Universal Time Delay Simulator for Large Digital Nets," Proc. Spring Joint Computer Conf. (1970), pp. 207–16.
24. E. Kjelkerud and O. Thessen, "Techniques for Generalized Deductive Fault Simulation," J. Design Automation and Fault Tolerant Computing, *1,* No. 10 (October 1974), pp. 377–90.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 9, November 1981
Printed in U.S.A.

A Fault-Collapsing Analysis in Sequential Logic Networks

By S.-J. CHANG and M. A. BREUER*

(Manuscript received September 11, 1980)

Although a sequential circuit M reduces to a combinational network C_M after all feedback paths have been cut, an application of Bossen and Hong's checkpoint labeling procedure to C_M does not necessarily yield a minimal solution. The set of checkpoints so obtained will include all feedback lines. In this paper, it is shown that these feedback lines are not necessary checkpoints under a "delay equivalence" relation. In addition to this, we also show that not every fanout branch is a necessary checkpoint. Any "singular fanout branches" can be removed from consideration. The results of our analysis lead to a minimal checkpoint labeling procedure for sequential logic networks.

I. INTRODUCTION

Because there are $3^w - 1$ possible multiple stuck faults in a logic network containing w distinct locations where signals may fail (each location may assume one of the three possible states: normal, stuck-at-0, or stuck-at-1), test generation and simulation procedures often resort to fault collapsing techniques to reduce the number of faults which need to be considered. To date reported results in the literature deal only with combinational logic. This prompted our interests in the research to be discussed in this paper.

Bossen and Hong introduced a fault collapsing technique for combinational logic networks called checkpoint labeling procedure.[1] Checkpoints, as defined by them, are a number of specified points in the network such that any multiple fault in the network is equivalent to some multiple fault among these specified checkpoints.[1] The checkpoints defined in Ref. 1 are a minimal set of points having the property

* M. A. Breuer is with the Departments of Electrical Engineering and Computer Science, University of Southern California, Los Angeles, California.

just stated. A natural question is, What are the checkpoints in a sequential circuit? It is well known that by cutting all the feedback lines, a sequential network can be mapped into a combinational network. A reasonable approach would be to apply Bossen and Hong's labeling procedure to the resulting combinational network. The checkpoints obtained would then include all the feedback lines. In this paper, we show that in general all these feedback lines need not be checkpoints under a relation called "delay-equivalence" to be defined later. We will see that the number of checkpoints can be greatly reduced for a highly sequential logic network if our results are utilized. This, in turn, greatly reduces computational complexities for multiple fault analysis in sequential networks. A possible application of our results can be found in a paper by Chang, Su, and Breuer.[2]

II. FAULT COLLAPSING IN SEQUENTIAL NETWORKS

The checkpoints in a circuit are specially designated signal lines. Checkpoints are defined so that for an arbitrary stuck-at-fault α there exists at least one equivalent fault defined on the checkpoints. Hence, if there are $u \ll w$ checkpoints, we need only consider $3^u - 1$ multiple faults. Bossen and Hong have developed a labeling procedure for specifying the minimal set of checkpoints in a combinational circuit. Their results seem to be an extension of the work of Shertz and Poage.[3,4]

For convenience, Bossen and Hong's procedure is as follows:

(*i*) All the primary inputs that do not fan out are checkpoints.

(*ii*) All the fanout branches are checkpoints.

(*iii*) NOT gates are considered as lines.

Although a sequential circuit M reduces to a combinational network C_M after all feedback paths have been cut, Bossen and Hong's labeling procedure does not necessarily yield a minimum number of checkpoints for such circuits.

Consider a synchronous sequential circuit M represented by Huffman's model as shown on Fig. 1, where C_M is the combinational portion of M, and D is the set of delay elements. We assume that faults in C_M are restricted to stuck-at type and faults in delay elements (D *f/f*'s) result in stuck outputs. Gates and flip-flops of M are connected by edges. An edge of M is either a line or a branch. Namely, it is either a primary input, a primary output, a fanout stem, or a fanout branch. Also we shall assume the following:

Assumption 1: The output value of a gate is a function of each of its inputs.

Assumption 2: The number of edges in M is finite.

Assumption 3: There are no "inaccessible" edges in M. An edge of

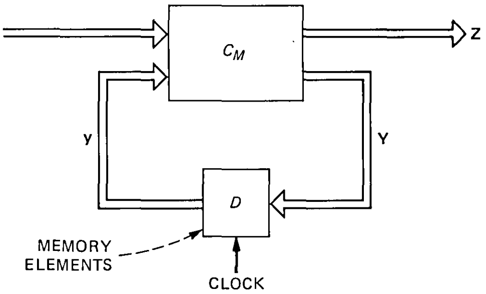


Fig. 1—Huffman's model of a sequential circuit M.

M is said to be inaccessible if it is not a primary input and, moreover, if it is not driven by any gates or memory elements of M.

Almost all practical sequential circuits satisfy the above assumptions. Figure 2 shows a fictitious integrated circuit chip. Edge 5 is inaccessible.

Restricting our attention to stuck-at-faults, we define a single fault as exactly one edge stuck-at-1 (s-a-1), or stuck-at-0 (s-a-0), and a multiple fault as a collection (one or more) of single faults, each associated with a different edge. Also, we shall not consider intermittent faults.

By applying Bossen and Hong's procedure to the combinational network C_M, the set of checkpoints so obtained will include all feedback lines denoted by the vector **Y**. However, we shall show that this subset of checkpoints is not necessary under a delay-equivalence relation. Later we shall consider asynchronous sequential circuits.

III. SEQUENTIAL NETWORK WITH D FLIP-FLOPS

Definition 1

Two faults u and v in M are said to be *delay-equivalent* (*d-equivalent*) *of order* k, if M with u is equivalent to M with v after an application of an input sequence of length at least k.

If the specific value of k is not of interest to us, we shall simply say

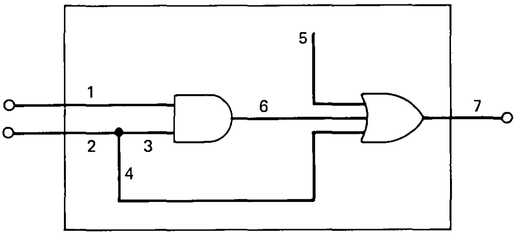


Fig. 2—A fictitious integrated circuit chip.

that u and v are d-equivalent. To demonstrate this concept, let us refer to the sequential circuit shown in Fig. 3. The fault a s-a-0 is d-equivalent of order 1 to the multiple fault b s-a-0 and c s-a-0.

Lemma 1: *Any multiple fault in a delay flip-flop is d-equivalent of order* 1 *to a stuck input.*

Proof: Let y and Y denote, respectively, the output and input of a delay *f/f*. Then, $y(q + 1) = Y(q)$ for all q. Thus, a stuck output is d-equivalent of order 1 to a stuck input. Also, a stuck input and output is d-equivalent of order 1 to a stuck input only, and is d-equivalent of order 0 to a stuck output only.

Because of this lemma, we shall consider a delay *f/f* as a 1-input gate under the d-equivalence relation.

Definition 2

(i) A sequence of edges of M, denoted by $[s_1, s_2, \cdots, s_i, \cdots, s_n]$, where $s_i \neq s_j$ for all $i \neq j$, is said to be a *forward path*, if for each $i < n$ either (a) s_i is a fanout stem and s_{i+1} is a fanout branch of s_i, or (b) there exists a gate of M, say g, such that s_i and s_{i+1} are, respectively, an input and the output of g.

If $s_1 = \alpha$ and $s_n = \beta$, the path $[s_1, s_2, \cdots, s_n]$ is said to be a *forward path from* α *to* β. If β is a primary output, the path is said to be a *terminal forward path of* α.

(ii) A sequence of edges of M, denoted by $[s_1, s_2, \cdots, s_i, \cdots, s_n]$, where $s_i \neq s_j$ for all $i \neq j$, is said to be a *backward path*, if for each $i < n$ either (a) s_i is a fanout branch and its stem is s_{i+1}, or (b) there exists a gate of M, say h, such that s_i and s_{i+1} are, respectively, the output and an input of h.

If $s_1 = \alpha$ and $s_n = \beta$, the path is said to be a *backward path from* α *to* β. If β is a primary input, then the sequence $[s_1, s_2, \cdots, s_i, \cdots, s_n]$ is said to be a *terminal backward path of* α.

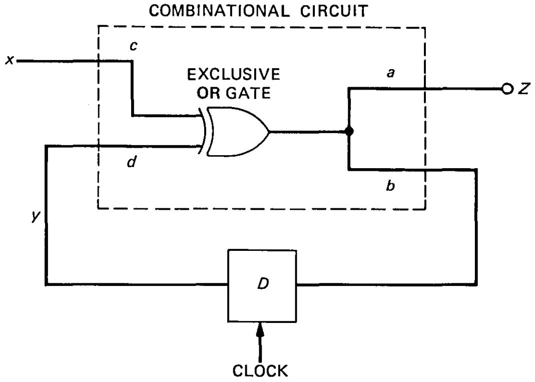


Fig. 3—A simple sequential circuit.

A backward path from α to β, denoted by p, is said to be a backward path of α, if there exists no edge Γ in M such that a backward path from α to Γ contains every element of p. The path p is said to *terminate* at β.

Clearly, if β is a primary input of M, then a backward path from α to β is also a terminal backward path of α.

The path $[s_1, s_2, \cdots, s_n]$ is said to *pass through* each s_i. Conversely, each s_i is said to be *contained* in the path. Branch s_i is also said to be an *element* of the path $[s_1, s_2, \cdots, s_n]$.

As an example, consider M_1 on Fig. 4. The sequences [12, 18, 6, 8, 9, 11, 20] and [12, 17, 19, 5, 8, 9, 11, 20] are forward paths of edge 12. However, the sequence [12, 18, 6, 8, 10, 12, 15, 16, 19, 5, 8, 9, 11, 20] is not a forward path of 12 because edges 12 and 8 appear twice in the sequence. All backward paths of edge 20 are shown in the graph of Fig. 5. Note that this graph is a tree with edge 20 as the root, and the leaves are either primary inputs or fanout branches.

Definition 3

If every edge of M possesses at least one terminal forward path, then M is said to be a *regular* sequential circuit. Otherwise, M is said to be *irregular.*

We will focus our attention on regular sequential circuits.

Definition 4

A *sensitized path* from an edge $\alpha = x_i$ to an output β in M is a terminal forward path from α to β, along with constant signal values assigned to some of the other edges such that changing the logic value of x_i will change the logic value of β.

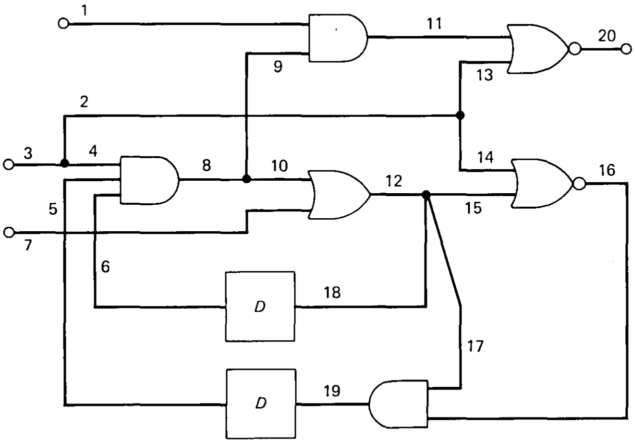


Fig. 4—Sequential circuit M_1.

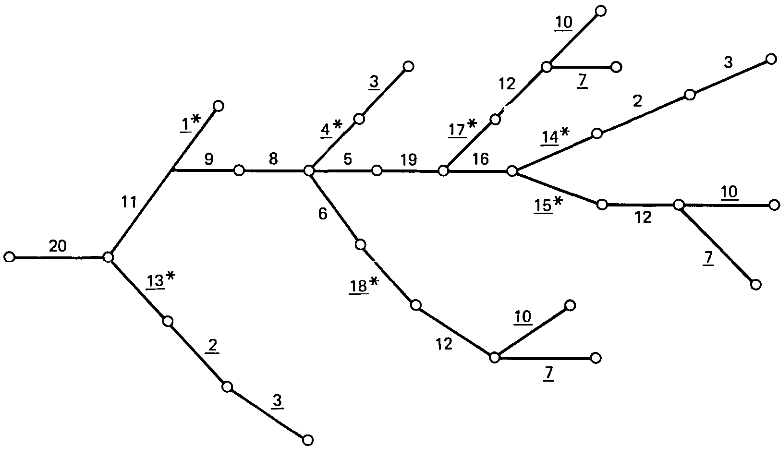


Fig. 5—Backward paths of edge 20 of M_1. Underline indicates either a primary input or nonsingular fanout branch. Starred edges are members of $K(20)$.

In the circuit shown in Fig. 4, we can sensitize the edge 1 to the output 20 by setting edge 9 to s-a-1 and edge 13 to s-a-0.

Lemma 2: Let g be a gate in M, a regular sequential circuit, and x_1, $x_2, \cdots, x_n$ be the inputs to g. For each i, there always exists a multiple fault in M which will sensitize x_i to an output.

Proof: Since M is regular there exists at least one terminal forward path from x_i to an output β. Using the concept of Boolean difference,[5] the condition for sensitizing the output of g to the input x_i is given by the equation

$$\frac{dg}{dx_i} = g_0 \oplus g_1 = g_0\bar{g}_1 + \bar{g}_0 g_1 = 1,$$

where

$$g_0 = g(x_1, \cdots, x_{i-1}, 0, x_{i+1}, \cdots, x_n),$$

and

$$g_1 = g(x_1, \cdots, x_{i-1}, 1, x_{i+1}, \cdots, x_n).$$

Now select any minterm of $g_0\bar{g}_1$ or $\bar{g}_0 g_1$. If this minterm specifies $x_j = \sigma_j$, where σ_j is either 0 or 1, then let the x_j input to g be stuck at σ_j. Now assume that for the terminal forward path from x_i to β, the output of g is an input to gate h. Repeating the process just described, h can be sensitized to g. This chaining process is continued until β is reached.

Definition 5

Let α and β be two edges of M. Edge α is said to *dominate* edge β, if both α s-a-0 and α s-a-1 will cause M to become independent of the signal on β regardless of whether or not other edges are normal.

To illustrate this, let us consider Fig. 2. Edge 6 dominates edges 1 and 3. Edge 7 dominates all other edges. However, edge 2 does not dominate edge 1, because when edge 3 s-a-1 and edge 4 s-a-0, edge 7 is a function of edge 1 independent of the state of edge 2.

Theorem 1: *Edge α dominates edge β, if and only if all terminal forward paths of β contain α.*

Proof: Suppose all forward paths of β pass through α. A stuck α will block any signal on β. Thus, M becomes independent of β if α is stuck. Therefore, α dominates β. Suppose there exists one terminal forward path of β that does not contain α. Then, by Lemma 2, one can always find a multiple fault on M, which will sensitize the signal on β to at least one of the primary outputs of M. Thus, M is dependent on β and α does not dominate β. This proves the only if part.

Theorem 2: *Dominance relation induces a partial ordering.*

Proof: Every forward path of α passes through α. Thus, α dominates α. Let α_1 dominate α_2 and α_2 dominate α_3. By Theorem 1, all forward paths of α_3 and α_2 pass through α_2 and α_1, respectively. Thus, all forward paths of α_3 also pass through α_1. This implies that α_1 also dominates α_3. Suppose α dominates β and β dominates α. By Theorem 1, all forward paths of β pass through α and all forward paths of α pass through β. This means that if $\alpha \neq \beta$, none of α and β has a forward path. Therefore, α dominates β and β dominates α implies $\alpha = \beta$. Namely, dominance relation is reflexive, transitive, and antisymmetric. We conclude that it induces a partial ordering.

In a combinational network, a fanout branch never dominates other fanout branches of its stem. However, in a sequential circuit a fanout branch may dominate other fanout branches of its stem. Consider, for example, the sequential circuit M_1, which is shown on Fig. 4. Branches 9 and 10 are fanout branches of stem 8. Since every forward path of 10 passes through 9, branch 9 dominates branch 10.

Definition 6

A fanout branch that dominates other fanout branches of its stem is said to be *singular*; otherwise, it is said to be *nonsingular.*

Theorem 3: (*i*) *A stem possesses at most one singular branch.* (*ii*) *Every singular branch dominates all other fanout branches of its stem.*

Proof: Let a stem which possesses singular branches be denoted by β and its fanout branches be denoted by $\alpha_1, \alpha_2, \cdots, \alpha_n$, where α_1 is singular. By definition, α_1 must dominate at least one α_i ($\neq\alpha_1$). If α_j, where $j \neq 1$ or i, is also singular, then α_i will have at least one forward path that does not pass through α_1 because every forward path of α_i contains the fanout stem β. Thus, α_i will no longer be dominated by α_1. This contradiction proves (*i*). Suppose some α_k is not dominated by

α_1. Then, α_k possesses at least one forward path that does not pass through α_1. It follows that α_i, for all $i \neq 1$, also possesses at least one forward path that does not pass through α_1. Part (*ii*) of the theorem follows as a result of Theorem 1.

Definition 7

The *kernel set* of edge α of M, denoted by $K(\alpha)$, is a set of edges of M such that

(*i*) every backward path of α contains exactly one member of $K(\alpha)$,

(*ii*) members of $K(\alpha)$ are either primary inputs that do not fanout or nonsingular fanout branches, and

(*iii*) every forward path from each member of $K(\alpha)$ to α does not contain any other nonsingular fanout branches.

Let us consider M_1 again (see Fig. 4). All backward paths of edge 20 are shown on Fig. 5. Each underlined edge is either a primary input or nonsingular fanout branch. Every starred edge is an element of the kernel set of edge 20 [i.e., $K(20)$]. From Fig. 5 we have $K(20) = \{1, 4, 13, 14, 15, 17, 18\}$. The following kernel sets can be easily verified: $K(1) = \{1\}$; $K(12) = \{7, 10\}$; and $K(8) = \{4, 14, 15, 17, 18\}$.

Lemma 3: *Let α and β be any pair of edges of M.*

(*i*) *A stuck α is d-equivalent to some multiple fault among its kernel set $K(\alpha)$.*

(*ii*) *If $K(\alpha)$ and $K(\beta)$ are disjoint, then any multiple fault among α and β is d-equivalent to some multiple fault among $K(\alpha) \cup K(\beta)$.*

Proof: Part (*i*) follows jointly from the fact that any multiple faults of a logic gate are equivalent to multiple faults among the input lines only,[1] and the definition of $K(\alpha)$ which requires that every backward path of α contains exactly one member of $K(\alpha)$. Part (*ii*) follows part (*i*) of the lemma.

Lemma 4: *If α does not dominate β, then either* (*i*) *there exists no forward path from β to α, or* (*ii*) *every forward path from β to α contains at least one nonsingular branch.*

Proof: If statement (*i*) is true, then α does not dominate β.

Suppose there exists a forward path from β to α, denoted by $p = [s_1, \cdots, s_n]$, that does not contain any nonsingular branch. Then s_i is either a fanout free stem or a singular branch. Let s_i be singular. Then, s_{i-1} is a fanout stem. By Theorem 3, all other fanout branches of s_{i-1} are dominated by s_i. Therefore, all forward paths of s_{i-1} pass through s_i. Thus, all forward paths of β also pass through α. This implies that α dominates β. Suppose every s_i is a stem without fanout, then p is a fanout free path. But by Theorem 1, this also implies that α dominate β. The contradiction proves the lemma.

Theorem 4: *If α does not dominate β and β does not dominate α, then their kernel sets are disjoint.*

Proof: Let $\sigma \epsilon K(\beta)$. Then, σ falls into exactly one of the following categories.

(*i*) σ lies in a backward path of β which does not contain any edge in common with any backward path of α.

(*ii*) σ is contained in a forward path from α to β.

(*iii*) σ is contained in a backward path of β, say p, that does not contain α but p, and a backward path of α, say q, contains common branches.

Clearly, if σ belongs to category (*i*), $\sigma \notin K(\alpha)$. If σ belongs to category (*ii*), by Lemma 4 and category (*iii*) of the definition of $K(\alpha)$, $\sigma \notin K(\alpha)$. So, consider only category (*iii*). Let $p = [s_1, s_2, \cdots, s_n]$, and $q = [t_1, t_2, \cdots, t_m]$. Category (*iii*) implies that there exists u and v such that $s_u = t_v$ and $s_i \neq t_j$ for all $i < u, j < v$. Namely, s_u (or t_v) is a fanout stem. If s_{u-1} and t_{v-1} are nonsingular, by definition σ lies in a forward path from s_{u-1} to β; that is, $\sigma \notin K(\alpha)$. If s_{u-1} is singular, by Theorem 3 t_{v-1} is not. Thus, branch t_{v-1} is dominated by s_{u-1} and so is α. If s_{u-1} is dominated by β then so is α. This contradicts the assumption. Therefore, s_{u-1} is not dominated by β. By Lemma 4, there exists $k \leq u - 1$ such that s_k is a nonsingular branch. This implies that $\sigma \notin K(\alpha)$. Therefore, $K(\beta)$ and $K(\alpha)$ are disjoint.

Theorem 5: *In a regular sequential circuit, every singular branch is not a necessary checkpoint, but all nonsingular fanout branches are necessary checkpoints.*

Proof: Let β be a fanout stem with fanout branches $\alpha_1, \cdots, \alpha_n$, where α_1 is singular. By Theorem 3, every $\alpha_i \neq \alpha_1$, is dominated by α_1. Thus, a stuck α_1 edge is *d*-equivalent to a stuck β edge. As to a stuck α_i edge, for any $i \neq 1$, it is not *d*-equivalent to either a stuck β edge, or stuck α_1 edge, or stuck α_j edge for $j \neq i$. Therefore, all nonsingular fanout branches of β are necessary checkpoints.

Theorem 6: *For any irregular sequential circuit M there exists a regular sequential circuit M^* such that M and M^* are equivalent, and furthermore, M^* preserves fault behavior of M under stuck-type fault assumption.*

Proof: Being irregular, M possesses two sets of edges, denoted by Ω and Ω^*, where each member Ω does not have any terminal forward paths, and each member of Ω^* possesses at least one terminal forward path. It follows that no primary output of M is a function of the signals on any member of Ω under fault-free or any stuck-type fault conditions. Therefore, removing all members of Ω will not alter the normal or abnormal functional behavior of M. The resulting circuit consists of only edges in Ω^* and is regular.

To illustrate this theorem, consider sequential circuit M_2 on Fig. 6a. One can easily verify that $\Omega = \{4, 5, 8, 11, 13, 14, 16\}$ and $\Omega^* = \{1, 2, 3, 6, 7, 8, 9, 10, 12, 15, 17\}$. After removing Ω from M_2 and

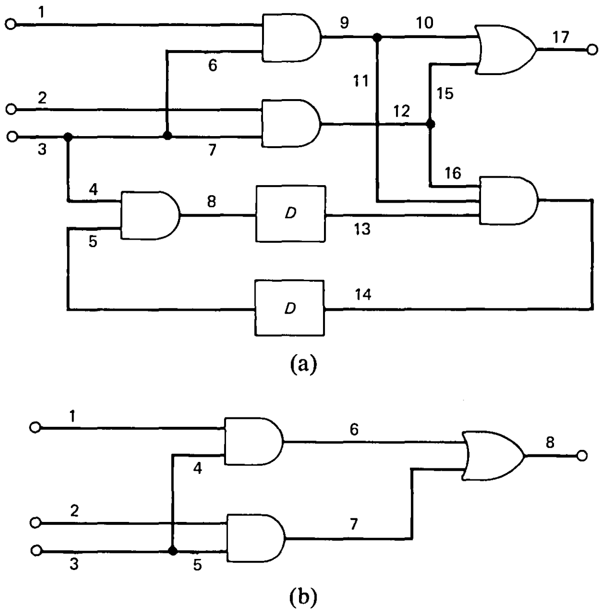


Fig. 6—Use of Theorem 6. (a) Sequential circuit. (b) Circuit M_3.

relabeling, one obtains the sequential circuit M_3, shown on Fig. 6b, which is actually a combinational circuit.

Theorem 7: *In a sequential circuit M, either* (*i*) *the kernel set for any edge in M exists, or* (*ii*) *M is irregular.*

Proof: Suppose α is an edge of M. Let a backward path of α be denoted by $p = [s_1, s_2, \cdots, s_n]$, where $s_1 = \alpha$. Then, p belongs to one of the following categories.

(*i*) p is a terminal backward path; namely, s_n is a primary input of M.

(*ii*) There exists $\tau < n$ such that s_n and $s_{\tau-1}$ are fanout branches of τ. (See Fig. 7a.)

(*iii*) There exists a gate g such that α is an input of g and s_n is the output of g. (See Fig. 7b.)

(*iv*) s_n is inaccessible.

Category (*iv*) never occurs because of Assumption 3.

Suppose p falls into either category (*i*) or (*ii*). Then every backward path of α contains at least one nonsingular fanout branch or a primary input. This implies that $K(\alpha)$ does exist if case (*iii*) does not apply.

Now consider case (*iii*). Suppose there exists no s_i, $1 \leq i \leq n$, that is, a fanout branch. This means that every element of p does not have a terminal forward path. Thus, M is irregular.

Suppose there exists some s_i, $1 \leq i \leq n$, that is a fanout branch. If s_i

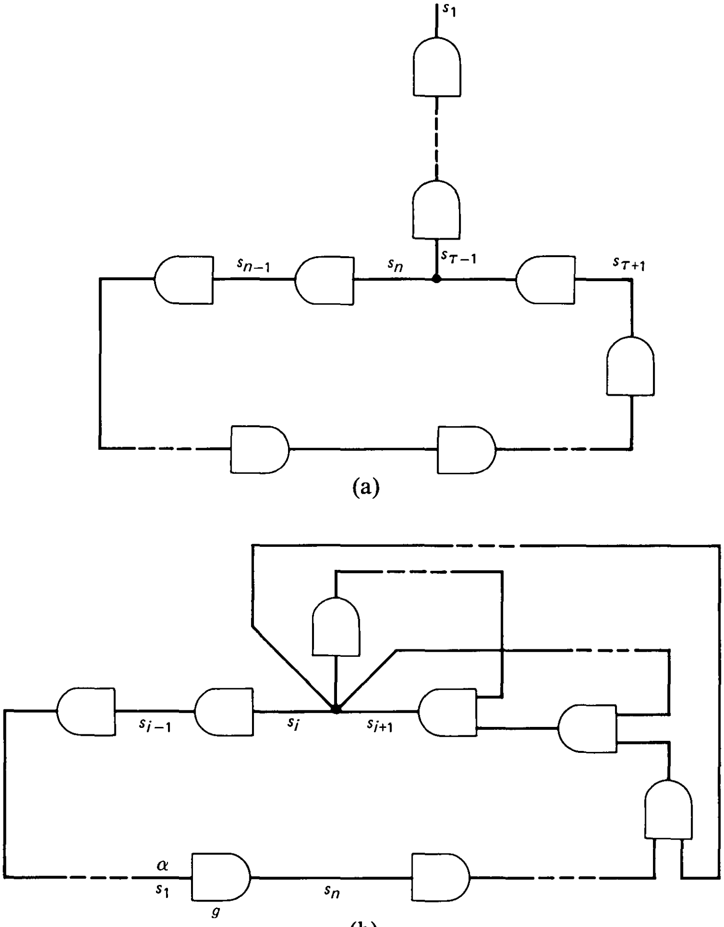
is singular, then s_{i+1} is the stem of s_i. Moreover, s_{i+1} is dominated by s_i. If all other fanout branches of p are also singular, then by the transitivity of dominance, s_i would be dominated by s_{i+1}. This would imply that neither s_i nor s_{i+1} has a forward path. Thus, we conclude that some s_j must be a nonsingular fanout branch and, hence, the theorem follows.

From the foregoing analysis, we now derive a checkpoint labeling procedure for a sequential circuit M.

Step 1. If M is irregular, convert M into M^* as in the proof of Theorem 6.

Step 2. All primary inputs that do not fanout are checkpoints.

Step 3. All nonsingular fanout branches are checkpoints.



(a)

(b)

Fig. 7—Proof of Theorem 7. (a) Category (*ii*); (b) Category (*iii*).

Step 4. NOT gates are considered as lines and delay *f/f*'s are considered as 1-input gates.

Theorem 8: *The above procedure yields the necessary and sufficient checkpoints to represent all the multiple faults in M which are detectable.*

Proof: Clearly, faults among the edges of Ω of Theorem 6 are undetectable and can be ignored. Let ϵ be a multiple fault defined over a set of edges of M^*, denoted by $A = \{\alpha_i\}$. Let $B = \{\beta_j\}$ be constructed from A by removing all elements of A which are dominated by some other elements of A. Then B is the largest subset of A such that no member of B is dominated by any member of B. Because any signal on a dominated edge will be blocked by the edges which dominate it, ϵ is equivalent to a multiple fault among B. Furthermore, since $\beta_i \in B$ does not dominate $\beta_j \in B$ for all $i \neq j$, $K(\beta_i)$ and $K(\beta_j)$ are disjoint. By Lemma 3, a multiple fault among B is equivalent to a multiple fault among the following collection of edges of M^*:$\Gamma = \cup_{\beta_i \in B} K(\beta_i)$. Since Γ is a subset of the collection of checkpoints W given by the above procedure, ϵ is equivalent to a multiple fault among W. This proves the sufficiency. The necessity follows from Theorems 5 and 6.

In comparison with the set of checkpoints obtained by applying Bossen and Hong's procedure to the combinational network C_M of Fig. 1, we see that we have removed from consideration the following edges: (i) all feedback lines represented by the vector Y; and (ii) all singular fanout branches.

Therefore, our work greatly simplifies fault analysis of cyclic logic networks.

It is important to point out that Definition 6 and Theorem 1 can be used to identify the singular edges.

IV. ASYNCHRONOUS SEQUENTIAL CIRCUITS

Having found a minimal checkpoint labeling procedure for synchronous sequential circuits with delay flip-flops as memory elements, we now consider asynchronous sequential circuit. The structure of an asynchronous sequential circuit can be represented as shown in Fig. 8.

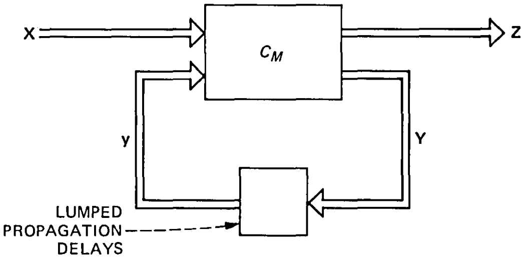


Fig. 8—Asynchronous sequential circuit.

It can be seen that the model is almost identical to the one shown in Fig. 1. The only difference is that the clocked delay flip-flops of Fig. 1 are now lumped propagation delays of the feedback lines. Because of this, one might conjecture that the checkpoint labeling procedure is applicable without modification to asynchronous sequential circuits. This conjecture is indeed true.

Consider Step 3 of the labeling procedure. It states that a delay flip-flop is considered as a 1-input gate. Under stuck-type fault assumption, this is equivalent to saying that a delay flip-flop is considered as an edge in the procedure. Therefore, we ascertain that the foregoing analysis carries over to an asynchronous case under the permanent stuck-at fault assumption.

V. SUMMARY AND DISCUSSION

In our analysis, we have developed a minimal checkpoint labeling procedure for squential circuits. This procedure is applicable to both synchronous sequential circuits with delay flip-flops as memory elements and asynchronous sequential circuits. Since most of the clocked flip-flops in use today, such as *JK f/f*, latch *f/f*, etc., are actually made of asynchronous sequential circuits, the procedure is applicable to any synchronous sequential circuits whose memory elements are either delay flip-flops or the aforementioned clocked flip-flops.

REFERENCES

1. D. C. Bossen and S. J. Hong, "Cause-Effect Analysis for Multiple Fault Detection in Combination Networks," IEEE Trans. Computers (November 1971), pp. 1252–7.
2. S. J. Chang, S. Y. H. Su, and M. A. Breuer, "Detection and Location of Multiple Stuck-Type Failures in Synchronous Sequential Circuits," IEEE Computer Soc. Repository, Report No. R-72-233.
3. D. R. Shertz, "On the Representation of Digital Faults," Report R-418, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois.
4. J. F. Poage, "Derivation of Optimum Tests to Detect Faults in Combinatorial Circuits," *Mathematical Theory of Automata,* Polytechnic Institute of Brooklyn, pp. 483–528, 1963.
5. M. A. Breuer and A. D. Friedman, "Diagnosis and Reliable Design of Digital Systems," Woodland Hills, California: Computer Science Press, 1976.

CONTRIBUTORS TO THIS ISSUE

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Laboratories, 1953—. Mr. Beneš has pursued mathematical research in traffic theory, stochastic processes, frequency modulation, combinatorics, servomechanisms, stochastic control, and filtering. In 1959-60 he was visiting lecturer in mathematics at Dartmouth College. In 1971 he taught stochastic processes at SUNY Buffalo, and in 1971-72 he was Visiting MacKay Lecturer in electrical engineering at the University of California in Berkeley. He is the author of *General Stochastic Processes in the Theory of Queues* (Addison-Wesley, 1963), and of *Mathematical Theory of Connecting Networks and Telephone Traffic* (Academic Press, 1965). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mathematical Association of America, IEEE, Phi Beta Kappa.

Melvin A. Breuer, B.S. (Engineering), 1959, M.S., 1961, University of California, Los Angeles; Ph.D. (Electrical Engineering), 1965, University of California, Berkeley; University of Southern California, Los Angeles, 1965—. At the University of Southern California, Los Angeles, Mr. Breuer is Professor of Electrical Engineering. His main interests are in the area of switching theory, computer-aided design of computers, fault diagnosis, and simulation. In addition to his research activities, Mr. Breuer has been at the forefront of developing courses on design automation and fault tolerant computing at universities and at a number of research institutes. He is the editor and co-author of *Design Automation of Digital Systems: Theory and Techniques*, Prentice-Hall; editor of *Digital System Design Automation: Languages, Simulation and Data Base*, Computer Science Press; co-author of *Diagnosis and Reliable Design of Digital Systems*, Computer Science Press; and editor-in-chief of the *Journal of Design Automation and Fault Tolerant Computing.* Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

Shih-Jeh Chang, Diploma E.E., 1959, Taipei Institute of Technology, M.S.E.E., 1966, Ph.D., 1972, University of Southern California; Bell Laboratories, 1972—. Early in his career at Bell Laboratories, Mr. Chang worked on the No. 1 Automatic Message Accounting and Recording Center (AMARC) in the area of system architecture, fault recognition and recovery. He was instrumental in the early phase design of 3B cache and memory management system. He also worked in the area of a *UNIX** base real-time computer system performance analysis and improvement and as a *UNIX* system consultant. Currently he is engaged in No. 5 ESS maintenance software design and development and multimodule office software development coordination and integration. Member, IEEE, Eta Kappa Nu.

Aland K. Chin, B.A., 1972, Brandeis University; M.S., 1975, Ph.D., 1977, Cornell University; Senior Research Engineer, Honeywell Electro-Optics Center, 1977–1978; Bell Laboratories, 1978—. Mr. Chin is involved in the design, processing, and characterization of light-emitting diodes for optical communication systems. Member, American Physical Society, American Association for the Advancement of Science, Phi Beta Kappa, New York Academy of Science, Electrochemical Society.

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

James L. Flanagan, B.S. (Electrical Engineering) 1948, Mississippi State University; M.S., 1950, Sc.D., 1955, Massachusetts Institute of Technology. Bell Laboratories 1957—. Mr. Flanagan is Head, Acoustics Research Department. He has project responsibilities for digital voice encoding, speech recognition and synthesis, electroacoustic systems and transducers. Fellow, IEEE, Acoustical Society of America. Member, National Academy of Engineering.

Ralph E. Frazee, Jr., (Electronic Engineering); Western Electric, 1969—. At the Western Electric Engineering Research Center, Princeton, N.J., Mr. Frazee has been primarily engaged in research on manufacturing process control and inspection systems. His current interests include measurement, characterization, and control of optical fiber manufacturing processes.

* Registered trademark of Bell Laboratories.

Allen Gersho, B.S.(E.E.), 1960, Massachusetts Institute of Technology; M.S., 1961, and Ph.D., 1963, Cornell University; Bell Laboratories, 1963–1980; University of California, Santa Barbara, 1980—. At Bell Laboratories, Mr. Gersho was a member of the technical staff at the Mathematics and Statistics Research Center where he was engaged in basic and applied research in digital communications, data transmission, quantization, and signal processing. At the University of California, he holds a professorship in Electrical and Computer Engineering. He has served as Editor of the IEEE Communications Magazine and is currently an Associate Editor of Communication Theory, IEEE Transactions on Communications. Member, Board of Governors of the IEEE Communications Society.

Ytzhak Levendel, B.S.E.E., 1971, Technion-Israel; M.S.C.S., 1974, The Weitzman Institute of Science; Ph.D., 1976, University of Southern California; Bell Laboratories 1976—. Mr. Levendel has done research in fault diagnosis and is currently involved in the development of a logic and test design aid system. Member, IEEE, Eta Kappa Nu.

Tong L. Lim, A.B., B.E., 1973, Dartmouth College; M.S.E.E., 1974, Ph.D. (E.E.), 1976, University of California, Berkeley; University of Pennsylvania, 1976–1978; Bell Laboratories, 1978—. Mr. Lim has worked on problems related to radiophysics and radar and sonar signal processing. His current interests are in data transmission and graphics communication. Member, IEEE, Sigma Xi; Associate Editor, IEEE Communications Magazine.

Subhash Mahajan, B.S. (Physics, Chemistry, and Mathematics), 1959, Panjab University, B.E. (Metallurgy), 1961, Indian Institute of Science, Bangalore, and Ph.D. (Materials Science), 1965, University of California, Berkeley; University of Denver, 1965–1968; United Kingdom Atomic Energy Authority Research Fellow, Atomic Energy Research Establishment, Harwell, England, 1968–1971; Bell Laboratories, 1971—. Mr. Mahajan has done research in the nucleation and growth of deformation twins in metallic crystals, mechanisms of phase transformations and plastic deformation, and in the role of deformation twins in the nucleation of fraction. The central theme of his research has been to establish the correlations between structure and properties in materials. Most recently he has been involved in understanding the interrelationships between substructure and growth parameters in III-V compound semiconductors. He is presently a group supervisor in the Materials Research Laboratory. Member, AIME, Electrochemical Society, Sigma Xi, New York Academy of Sciences.

David Malah, B.S. (E.E.), 1964, M.S. (E.E.), 1967, Technion-Israel Institute of Technology, Ph.D. (E.E.), 1971, University of Minnesota; University of New Brunswick, Canada N.B., 1970–72; Technion-Israel Institute of Technology, 1972—; sabbatical leave, Bell Laboratories, 1979–81. At the University of Minnesota, Mr. Malah was Assistant Professor in the Electrical Engineering Department. At Technion-Israel, he is Associate Professor of Electrical Engineering. During 1975–1979, Mr. Malah was in charge of the Signal Processing Laboratory at Technion-Israel and was involved in research on speech and image communications, and real-time digital hardware implementations. Additionally, he has done research on discrete time systems, digital signal processing techniques, speech communications, and adaptive filtering. While on leave he is conducting research on narrow-band communications and digital signal processing techniques in the Acoustics Research Department at Bell Laboratories.

Premachandran R. Menon, B.S. (Electrical Engineering), 1954, Banaras Hindu University; Ph.D. (Electrical Engineering), 1962, University of Washington; Bell Laboratories, 1963—. Mr. Menon has done research in switching theory and fault diagnosis and is currently involved in the development of a logic simulation system. Member, IEEE.

Markus S. Mueller, E.E. Diploma, 1970, Ph.D. (Electronic Engineering), 1976, Swiss Federal Institute of Technology, Zurich, Switzerland; Swiss Federal Institute of Technology, Zurich, 1971–1976; GenRad, Zurich, 1976–1978; Bell Laboratories, 1978–1981. Mr. Mueller was teaching and research assistant at the Swiss Federal Institute of Technology, where he worked in various fields, including filter theory, data transmission, and adaptive signal processing. He was a product specialist with GenRad for computer controlled automatic test systems. At Bell Laboratories, he was a member of the Data Theory Group in the Data Systems and Technology Department and his interest was in data communication, digital signal processing, and adaptive systems.

Jack Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; Bell Laboratories, 1961—. Mr. Salz first worked on remote line concentrators for the electronic switching system. Since 1968, he has supervised a group engaged in theoretical studies in data communications, and he is currently a member of the Communications Methods Research Department. During the academic year 1967–68, he was on leave as Professor of Electrical Engineering at the University of Florida. In Spring, 1981, he was a visiting lecturer at Stanford University. Member, Sigma Xi.

Dhiraj K. Sharma, B. Tech. (Electrical Engineering), 1971, I.I.T. Kanpur, India; M.S. and Ph.D. (Electrical Engineering), 1972 and 1975, California Institute of Technology; Bell Laboratories 1975—. Mr. Sharma has worked on efficient encoding of video signals and identification of structural parameters of tall buildings. His current interests are in screen-based human interfaces, communication and synchronization in distributed systems, and programming languages. Member, ACM, Sigma Xi.

David H. Smithgall, B.S. (Electrical Engineering), 1967, M.S., 1968, Ph.D., 1970, Cornell University; Western Electric, 1970—. At the Western Electric Engineering Research Center, Princeton, N.J., Mr. Smithgall has been engaged primarily in research in process characterization and control. Since 1973 he has been involved in lightguide fiber measurement and characterization of the fiber and preform fabrication processes. Member, IEEE.

Raymond Steele, B.S., 1959, Ph.D. (Electrical Engineering) 1975, Durham University, Durham, England; E. K. Cole, Ltd., 1959–1961; Cossor Radar, Electronics, Ltd., 1961–1962; The Marconi Company, 1962–1965; Royal Naval College, 1965–1968; Loughborough University of Technology, 1968–1979; Bell Laboratories, 1979—. At E. K. Cole, Ltd., Cossor Radar, Electronics, Ltd., and The Marconi Company, all located in Essex, England, Mr. Steele was engaged in research and development. As a member of the Lecturing Staff at the Royal Naval College in London he lectured on telecommunications for the NATO and external London University degree courses. At the Loughborough University of Technology in Loughborough, England, he directed a research group in digital encoding of speech and television signals, in addition to serving as Senior Lecturer. Before joining Bell Laboratories on a full time basis, Mr. Steele served as a part-time consultant to the Laboratories' Acoustics Research Department. Presently, he is a member of the Communications Methods Research Department. He is the author of *Delta Modulation Systems* published in 1975.

Henryk Temkin, M.S. 1971, Yeshiva University, Ph.D. (Physics), 1975, Stevens Institute of Technology; Cornell University, 1975–1977; Bell Laboratories, 1977—; Mr. Temkin is working on developing light-emitting diodes for optical communications. Member, American Physical Society, Electrochemical Society.

Diane Vitello, A.A. (Mathematics) Brookdale Community College; Bell Laboratories, 1962—. Ms. Vitello is with the Communications Methods Research Department. She has written computer programs for radio research projects in antenna design, satellite tracking, and fiber optic transmission. She has also written several user-oriented programs to aid in the design and analysis of terrestrial and satellite digital radio systems. Ms. Vitello is currently continuing undergraduate work in Mathematics at Monmouth College.

On-Ching Yue, B.E.E., 1968, Cooper Union; M.S.E.E., 1971, Rochester Institute of Technology; Ph.D. (Information Sciences), 1977, University of California at San Diego; General Dynamics/Electronics Division, 1968–1977; Bell Laboratories, 1977—. Mr. Yue has worked in the areas of underwater acoustics and microwave imaging. Since joining Bell Laboratories, he has been a member of the Radio Research Laboratory, studying the effect of interference on digital communications systems, including intersymbol, adjacent satellite, and multiuser interferences. Member, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi; Senior Member, IEEE.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

On Ternary Self-Dual Codes of Length 24. J. Leon, V. Pless, and N. J. A. Sloane, IEEE Trans Information Theory, *IT-27* (March 1981), pp 176-80.

ENGINEERING

Broadband Astigmatic Compensation. T. S. Chu, IEEE/AP-S Int Symp Digest (June 15, 1981), pp 131-4.

Digital Satellite Communications-Problems and Possibilities. V. I. Johannes, Conf Proc, Commun Techniques Seminar (March 24, 1981), pp 6-1-6.4.

Plots and Tests for Goodness of Fit With Randomly Censored Data. V. N. Nair, Biometrika, *68,* No. 1 (April 1981), pp 99-103.

Process Simulation in Two Dimensions. B. R. Penumalli, 1981 IEEE Int Solid-State Circuits Conf Digest of Technical Papers, *XXIV* (February 1981), pp 212-3.

Per-Channel, Memory Oriented, Transmultiplexer With Logarithmic Processing—Architecture and Simulation. C. F. Kurth, K. J. Bures, and P. R. Gagnon, Int Commun Conf, 1981, Conf Record, *1,* No. 7.3 (June 15, 1981), pp 7.3.1-5.

Per-Channel, Memory Oriented, Transmultiplexer With Logarithmic Processing—Lab Model Implementation and Testing. C. F. Kurth, P. R. Gagnon, and K. J. Bures, Int Commun Conf, 1981 Conf Record, *1,* No. 7.4 (June 15, 1981), pp 7.4.1-4.

Ultraviolet Radiation Curable Conformal Coatings for Printed Circuit Board Assemblies. G. B. Fefferman and T. S. Hsu, Proc Printed Circuit World Convention II, *1* (June 1981), pp 303-15.

PHYSICAL SCIENCES

Boundary-Layer Model of Field Effects in a Bistable Liquid-Crystal Geometry. J. Cheng, R. N. Thurston, and D. W. Berreman, J Appl Phys, *52* (April 1981), pp 2756-65.

Electronic Charge Density of V_3Si. L. F. Mattheiss and D. R. Hamann, Solid State Commun, *38,* No. 8 (May 1981), pp 689-94.

Electronic Stacking-Fault States in Silicon. L. F. Mattheiss and J. R. Patel, Phys Rev, *23,* No. 10 (May 15, 1981), pp 5384-96.

Equilibrium and Stability of Liquid Crystal Configurations in an Electric Field. R. N. Thurston and D. W. Berreman, J Appl Phys, *52* (January 1981), pp 508-9.

Erythrocyte Protoporphyrin/Heme Ratio by Hematofluorometry. A. A. Lamola, J. Eisinger, and W. E. Blumberg, Clinical Chem, *26* (1980), p 677.

Generation of High-Brightness Coherent Radiation in the Vacuum Ultraviolet by Four-Wave Parametric Oscillation in Mercury Vapor. J. Bokor, R. R. Freeman, R. L. Panick, and J. C. White, Opt Lett, *6,* No. 4 (April 1981), pp 182-4.

Interplay of Structure and Magnetic Properties in MnSi: A Concentrated Spin Glass. T. M. Hayes, J. W. Allen, J. B. Joyce, and J. J. Hauser, Phys Rev, *23,* No. 9 (May 1, 1981), pp 4691-7.

Light Scattering by Photoexcited Two-Dimensional Electron Plasma in GaAs-(A1Ga) as Heterostructures. A. Pinczuk, J. Shah, A. C. Gossard, and W. Wiegmann, Phys Rev Lett, *46,* No. 20 (May 18, 1981), pp 1341-4.

Liquid Crystals and Geodesics. R. N. Thurston and F. J. Almgren, J De Phys (Paris), *42* (March 1981), pp 413-7.

Orientation and Implantation Effects on Stacking Faults During Silicon Buried Layer P Processing. M. Robinson, G. A. Rozgonyi, T. E. Seidel, and M. H. Read, J Electrochem Soc, *128* (April 1981), pp 926-9.

The Propagation of Disclinations in Bistable Switching. J. Cheng and R. N. Thurston, J Appl Phys, *52* (April 1981), pp 2766-75.

Sensitivity Analysis of Oscillating Reactions. 1. The Period of the Oregonator. D. Edelson and V. M. Thomas, J Phys Chem, *85* (May 1981), pp 1555–8.

Stability of Nematic Liquid Crystal Configurations. R. N. Thurston, J De Phys (Paris), *42* (March 1981), pp 419–25.

Unit Cell of the γ-Phase of Poly(Vinylidene Fluoride). A. J. Lovinger, Macromolecules, *14*, No. 2 (March 1981), pp 322–5.

Unit Sphere Description of Liquid Crystal Configurations. R. N. Thurston, J Appl Phys, *52* (April 1981), pp 3040–52.

CONTENTS, DECEMBER 1981

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering*, *Applied Mechanics Review*, *Applied Science & Technology Index*, *Chemical Abstracts*, *Computer Abstracts*, *Current Contents/ Engineering, Technology & Applied Sciences*, *Current Index to Statistics*, *Current Papers in Electrical & Electronic Engineering*, *Current Papers on Computers & Control*, *Electronics & Communications Abstracts Journal*, *The Engineering Index*, *International Aerospace Abstracts*, *Journal of Current Laser Abstracts*, *Language and Language Behavior Abstracts*, *Mathematical Reviews*, *Science Abstracts* (*Series A, Physics Abstracts*; *Series B, Electrical and Electronic Abstracts*; *and Series C, Computer & Control Abstracts*), *Science Citation Index*, *Sociological Abstracts*, *Social Welfare, Social Planning and Social Development*, and *Solid State Abstracts Journal*. Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Bell System