January 1985 Vol. 64 No. 1 Part 1

AT&T TECHNICAL JOURNAL

A JOURNAL OF THE AT&T COMPANIES

Resource Allocation

Failure Rate Model

Queueing

Jitter in Repeaters

Echo Cancellation

Input-Output Expansions

AT&T TECHNICAL JOURNAL (ISSN 8756-2324) is published ten times each year by AT&T, 550 Madison Avenue, New York, NY 10022; C. L. Brown, Chairman of the Board; T. O. Davis, Secretary. The Computing Science and Systems section and the special issues are included as they become available. Subscriptions: United States—1 year $35; foreign—1 year $45.

Back issues of the special, single-subject supplements may be obtained by writing to the AT&T Customer Information Center, P.O. Box 19901, Indianapolis, Indiana 46219, or by calling (800) 432-6600. Back issues of the general, multisubject issues may be obtained from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Payment for foreign subscriptions or single copies must be made in United States funds, or by check drawn on a United States bank and made payable to the AT&T Technical Journal and sent to AT&T Bell Laboratories, Circulation Dept., Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

Printed in U.S.A. Second-class postage paid at Short Hills, NJ 07078 and additional mailing offices. Postmaster: Send address changes to the AT&T Technical Journal, Room 1E335, 101 J. F. Kennedy Pky, Short Hills, NJ 07078.

AT&T
TECHNICAL
JOURNAL

VOL. 64 JANUARY 1985 NO. 1, PART 1

AT&T Technical Journal
Vol. 64, No. 1, January 1985
Printed in U.S.A.

Optimal Resource Allocation for Two Processes*

By C. COURCOUBETIS† and P. VARAIYA‡

(Manuscript received May 9, 1984)

Two processes compete for access to n resources. A scheduling policy allocates the resource when the processes request it simultaneously. The objective is to minimize the average value of a state-dependent cost. The optimal policy can be calculated explicitly for the case of one resource. In the general case $n > 1$, an adaptive scheduling algorithm is proposed. The algorithm measures average transition times of the system and converges to the optimal policy.

I. INTRODUCTION

Two processes share n resources. A process operates in one of these states: thinking, requesting, or holding a given resource. The thinking and resource holding times are geometric, with means depending on the process and the resource.

Time for the system is discrete. A decision has to be made when the two processes simultaneously request the same resource k, in which case a scheduling policy assigns the resource to process 1 with probability $u_k \in [0, 1]$. A resource cannot stay idle when there is a process requesting it. The problem is to choose $u^* \in [0, 1]^n$, which minimizes the average value of a cost depending on the state of the system.

A simple example of such a system is the case of two processes sharing the same broadcast facility. Such a process goes through the following phases. It thinks for an arbitrary amount of time (any

* Research supported by the Office of Naval Research Contract N00014-80-C-0507.
† AT&T Bell Laboratories. ‡ University of California at Berkeley.

activity other than broadcasting), requests the resource (broadcast channel), uses the resource for an arbitrary time (broadcasts), and then resumes thinking. If the resource was busy upon request, the process has to wait until the resource is released. If two processes request the resource simultaneously (during the same slot), the scheduler of the resource will decide, according to the scheduling policy of the system, which process will proceed first.

In Section III the problem is solved explicitly for the case $n = 1$ and when the cost is taken to be resource idle time. It is shown that for large ratios of average thinking times, u^* does not depend on the holding times and contrasts with the familiar "$c\mu$" results, where the average waiting cost is minimized when priority is given to processes with large waiting cost and small holding times (e.g., Refs. 1 through 6). It turns out that in that case the optimal policy consists in serving the process with the least thinking time first.

This is a new result for nonpreemptive systems. Similar results appeared in the literature concerning n processes sharing a single resource in a preemptive-resume way. For such systems, the "least thinking time first" scheduling rule has been proved optimal (see Refs. 7 and 8). For example, in Ref. 8 the author considers a "mirror image" problem, where the goal is to minimize the utilization (repair time) of the scarce resource (repairman). In this case the policy of serving the process with longest thinking time first is optimal. Because preemptive policies were considered and a single resource was shared, results are not applicable to the problem considered here.

In Section IV we consider the case of n resources and an arbitrary state-dependent cost. Since Section III suggests that an analytic expression for the optimal policy u^* as a function of the parameters of the processes would be extremely hard or even impossible to get, we proceed with an adaptive algorithm to compute u^*. This algorithm converges in a finite number of steps to u^* by using measurements of the average transition times of the system. The analysis is based on the results contained in Refs. 9 through 15, which are reviewed in Section 4.1. In Section 4.4 the structure of the problems for which the results apply is generalized. We permit a process to consist of an arbitrary set of thinking states, an arbitrary set of resource holding states for each resource in the system (one set of states per resource), and a number of resource request states (one state per resource). Finally, in Section V we present some open problems.

II. MODEL DESCRIPTION

In this section we give a formal description of the model for $n = 1$. The state of process i ($i = 1, 2$) at time t ($t \geq 0$) is X_t^i. If $X_t^i = 0$,

then process i is thinking. If $X_t^i = 1$ [resp. 2], then process i is requesting [resp. holding] the resource. See Fig. 1.

The resource can accommodate only one process at any time. When the two processes are simultaneously requesting the resource, then the resource is assigned with probability $u \in [0, 1]$ to process 1.

Denote by p_i the probability that process $i (i = 1, 2)$ will complete its thinking time in a given time unit. Similarly define q_i for completion of service (holding). From this definition, one finds that the process $X_t = (X_t^1, X_t^2)$, $t = 0, 1, \cdots$, is a Markov chain with transition probability matrix $P = \{P(x, y) \mid x, y \in X = \{0, 1, 2\}^2\}$ defined as follows:

$$P(00, 10) = p_1(1 - p_2),\ P(00, 01) = (1 - p_1)p_2,\ P(00, 11) = p_1p_2;$$

$$P(01, 12) = p_1,\ P(10, 21) = p_2;$$

$$P(01, 02) = (1 - p_1),\ P(10, 20) = (1 - p_2);$$

$$P(11, 21) = u,\ P(11, 12) = 1 - u;$$

$$P(02, 12) = p_1(1 - q_2),\ P(02, 00) = (1 - p_1)q_2;$$

$$P(20, 21) = (1 - q_1)p_2,\ P(20, 00) = q_1(1 - p_2);$$

$$P(12, 20) = q_2,\ P(21, 02) = q_1. \tag{1}$$

The diagonal elements of P are defined so that the rows sum to one. For any choice of u, X_t is ergodic.

Define as π_u the invariant measure of X_t, and let k: $X \to R$ be the state-dependent cost vector. Then the expected cost-per-unit time is

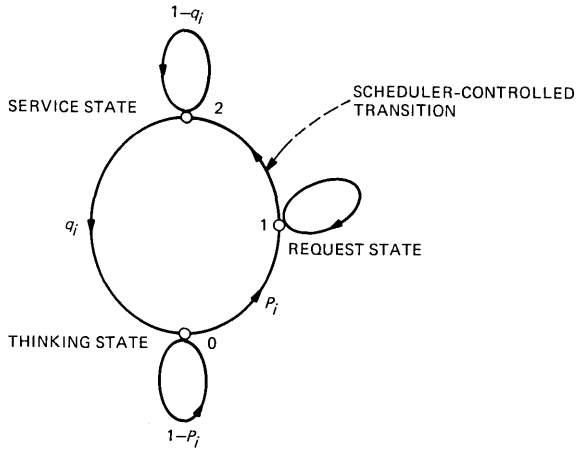


Fig. 1—Process i ($i = 1, 2$).

$$J(u) := \lim_{T\to\infty} \frac{1}{T+1} E \sum_{t=0}^{T} k(X_t) = \sum_{x\in X} \pi_u(x)k(x). \tag{2}$$

Our goal is to find u^* that minimizes (2).

III. MAXIMIZING RESOURCE UTILIZATION

Throughout this section we will assume $k(00) = 1$, $k(ij) = 0$, $ij \neq 00$.

Let π be the invariant probability measure associated with P. Then the expected idle time of the resource is

$$J(u) = \sum_{i,j\in\{0,1\}} \pi_u(ij). \tag{3}$$

Direct computation of π_u shows that (3) takes the following form:

$$J(u) = \frac{Au + B}{Cu + D}, \tag{4}$$

where A, B, C, D are functions of $\{p_1, p_2, q_1, q_2\}$. From (4) it follows that

$$\operatorname{sign}\left(\frac{\partial J(u)}{\partial u}\right) = \operatorname{sign}(AD - BC). \tag{5}$$

In what follows, we will choose $u^* \in [0, 1]$, which minimizes (4). From (5) it follows that u^* is an extreme point of $[0, 1]$. We state now the key results, which are proved in the Appendix.

Fact:

$$AD - BC = p_1p_2q_1^2q_2^2F_1F_2,$$

where $F_i = F_i(p_1, p_2, q_1, q_2)$, $i = 1, 2$.

Theorem 1 implies that the sign of $AD - BC$ is determined by the sign of F_1.

Theorem 1:

$$F_2 \geqslant 0 \quad \textit{for all} \quad p_1, p_2, q_1, q_2 \in [0, 1].$$

We state now the main theorem.

Theorem 2: 1. The curve $\{(p_1, p_2) \in [0, 1]^2 \mid F_1(p_1, p_2, q_1, q_2) = 0\}$ *lies between the lines* $p_1 = p_2$ *and* $p_2 = p_1/2$ *for* $q_1 \leqslant q_2$, *and between* $p_1 = p_2$ *and* $p_1 = p_2/2$ *for* $q_2 \leqslant q_1$.

2. $p_1 \geqslant 2p_2$ *implies* $F_1 \geqslant 0$, *and* $p_2 \geqslant 2p_1$ *implies* $F_1 \leqslant 0$, *for all* $q_1q_2 \in [0, 1]$.

The following corollary follows now from Theorem 2 and (5):

Corollary 2.1: Let u^* *be the value of the control minimizing (4). Then the following hold:*

1. $p_1 \geqslant 2p_2$ *implies* $u^* = 1$, *and* $p_2 \geqslant 2p_1$ *implies* $u^* = 0$;

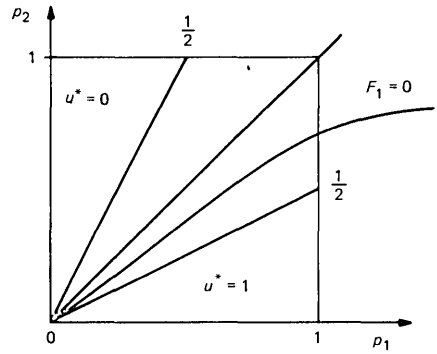


Fig. 2—Optimal policy u^* as a function of the parameter p_1, p_2.

2. If $p_1 = p_2$, *then* $q_2 \leqslant q_1$ *implies* $u^* = 0$, *else* $u^* = 1$;
3. If $q_1 = q_2$, *then* $p_1 \geqslant p_2$ *implies* $u^* = 1$, *else* $u^* = 0$.

Theorem 2 and Corollary 2.1 are illustrated in Fig. 2, where $F_1 = 0$ is considered for some fixed values of q_1, q_2, $q_2 < q_1$. The surprising result here is that there exist "safe" regions of the parameters, where u^* does not depend on the service times of the processes.

IV. ESTIMATION OF u^*

In this section we give a method for calculating u^* for an arbitrary cost vector k. The algorithm we propose requires no a priori knowledge of the parameters of the processes and is performed adaptively to the system. It starts by applying some arbitrary policy u^0, and then by monitoring the evolution of the system and estimating certain transition times, it updates the policy used until u^* is reached; this is always achieved in a finite number of steps, depending on the number of the shared resources. Since the algorithm is adaptive, it could be effectively used in systems with slowly varying parameters. This, together with the simplicity of the computation involved, makes this approach an interesting alternative to a direct calculation of u^* by using standard policy iteration algorithms (see Ref. 15). In Section 4.1 some general results are stated. We discuss in Section 4.2 the case of two processes and one resource, and in Section 4.3 the case of two processes and n resources.

4.1 Some results in Markov decision theory

Propositions 1, 2, and 3 consist of results already known and are stated without proof in order to make the present work self-contained. The key propositions used to calculate u^* are Propositions 4, 5, and 6, which consist of new results; their proofs are included in the Appendix.

Let X_t be a Markov chain on $\{1, \cdots, s\}$. If $X_t = i$, then any control $u \in U(i)$ may be used, where $U(i)$ is a compact set. A stationary strategy is any element $u = (u(1), \cdots, u(s)) \in U = U(1) \times \cdots \times U(s)$. Let $P(u)$ be the $s \times s$ transition probability matrix describing X_t, and assume that X_t consists of a single ergodic class and that $P(u)$ is continuous on U. Let $Q(u) := P(u) - I$, and let $k := (k(1), \cdots, k(s))'$ be the cost vector, not depending on u. The cost to be minimized is $J(u) = \lim_{T\to\infty} 1/(T+1) E \sum_{t=0}^{T} k(X_t)$. The following two propositions give the optimality conditions for u.

Proposition 1: (see, e.g., Ref. 14, Lemma 3.1) For $u \in U$ consider the s linear equations in the $s + 1$ variables $\gamma \in R$, $c \in R^s$,

$$\gamma \perp = Q(u)c + k, \tag{6}$$

where $\perp = (1, \cdots, 1)'$. Then:

1. If (γ, c) is a solution, then $\gamma = J(u)$.
2. If (γ, c) is a solution, then so is $(\gamma, c + \delta\perp)$ for all δ.
3. A solution always exists and is almost unique in the sense of (2).

Let $H(c, u) = Q(u)c + k$, $H = (H_1, \cdots, H_s)'$. Note that $H_i(c, u) = H_i(c, u(i))$ depends only on $u(i)$. Let $h(c) = (h_1, \cdots, h_s)$, $h_i(c) = \min \{H_i(c, v) \mid v \in U(i)\}$.

Proposition 2: (see, e.g., Theorems 3.1 and 3.2) The control u is optimal (minimizing $J(u)$) iff there exist (γ, c) such that $\gamma\perp = h(c) = H(c, u)$.

The following propositions will be used in the part dealing with the estimation of u^*.

Consider the equation

$$V_a = (I - aP(u))^{-1}k. \tag{7}$$

Then $V_a(i)$ is the expected discounted cost with the discount factor a, $a \in [0, 1]$, starting from state i, i.e., $V_a(i) = E_i\{\sum_{t=0}^{\infty} a^t k(X_t)\}$. Let $Z^a := (I - aP(u))^{-1}$.

Proposition 3: (see, e.g., Ref. 9, Chapter 3) Z^a is the fundamental matrix of the absorbing chain X_t^a defined on $\{0, 1, \cdots, s\}$ with the $(s + 1) \times (s + 1)$ transition probability matrix P_a such that

$$P_a(i, j) = aP(i, j), \qquad i, j \in \{1, \cdots, s\},$$

$$P_a(i, 0) = 1 - a \quad \text{for} \quad i \neq 0, \qquad P_a(0, 0) = 1.$$

Let $\bar{N}_i^a[j]$ be the expected number of visits to state j starting from i of X_t in a geometrically distributed interval of time with mean $(1 - a)^{-1}$. Then $z_{ij}^a = \bar{N}_i^a[j]$.

Proposition 4:

$$\lim_{a \to 1}[V_a(i) - V_a(k)] = c(i) - c(k),$$

where c satisfies (6).

Proposition 5: Let $\bar{T}_{ij}$ be the expected time before the first visit to state j by X_t starting from i. Then

$$\lim_{a \to 1}[\bar{N}_i^a[j] - \bar{N}_k^a[j]] = \begin{cases} \dfrac{\bar{T}_{kj} - \bar{T}_{ij}}{\bar{T}_{jj}} & \text{if} \quad k \neq j \neq i \\ -\dfrac{\bar{T}_{ij}}{\bar{T}_{jj}} & \text{if} \quad k = j \neq i. \end{cases}$$

The next proposition is key, since it relates the dual variable c to transition times of the chain X_t, which can be estimated quite readily. The proposition can be readily obtained from Propositions 3, 4, 5, and eq. (7).

Proposition 6:

$$c(i) - c(k) = \sum_{j=1}^{s} \left[\frac{\bar{T}_{kj} - \bar{T}_{ij}}{\bar{T}_{jj}} \right] k(j) + k(i) - k(k).$$

4.2 Optimal resolution of conflict for two processes and one resource

In this section we provide the following results. In Theorem 3 we characterize the value of u^*. In Theorem 4 we relate the value of u^* with the sign of a quantity that can be estimated using Proposition 6 from the behavior of the system. We finally propose Algorithm 1, which uses these results to adaptively calculate u^*.

Theorem 3: u^* can always be restricted to the set $\{0, 1\}$.

Proof: Let (γ^*, c^*) be the variables in (6) corresponding to the optimal policy u^*. Then by Proposition 2 and the fact that u enters only in the row of P corresponding to the state (11),

$$\begin{aligned} \gamma^* &= \min\{-c^*(11) - (1 - u)c^*(12) \\ &\qquad + uc^*(21) + k(11) \,|\, u \in U\} \\ &= \min\{[c^*(12) - c^*(11) + k(11)] \\ &\qquad + u[c^*(21) - c^*(12)] \,|\, u \in U\} \end{aligned} \tag{8}$$

and the minimum is achieved at u^*. Let $A = c^*(21) - c^*(12)$. Then $A > 0$ implies $u^* = 0$, $A < 0$ implies $u^* = 1$, and $A = 0$ implies that any $u \in [0, 1]$ will do. □

Theorem 4: Let (γ, c) be a solution to (6). Then $\text{sign}[c(21) - c(12)] =$ const. *for all $u \in U$.*

Proof: Suppose there exists a u_0 such that $c^{u_0}(21) - c^{u_0}(12) = 0$. Then γ^{u_0}, c^{u_0} are optimal dual variables since the optimality condition (8) is trivially satisfied. Then every $u \in U$ will satisfy the optimality conditions with $c^*(21) - c^*(21) = c^{u_0}(21) - c^{u_0}(12) = 0$; hence u is also optimal. From this and Proposition 4.1 it follows that $c^{u_1}(21) - c^{u_1}(12) = c^{u_0}(21) - c^{u_0}(12) = 0$, for all $u_1 \in U$. This, with the next Fact, proves Theorem 4.

Fact: The vector c of (6) can be chosen to be a continuous function of u.

This can be proved as follows. Let π_u be the invariant probability measure of $P(u)$. Since $\gamma = \pi_u k$, by using (6) we get

$$\pi_u k \perp - k = Q(u)c. \tag{9}$$

Since π_u is continuous in u (by the ergodicity of $P(u)$), and $Q(u)$ is of rank $s - 1$ for all $u \in [0, 1]$, then there is always a solution $\bar{c} = (\bar{c}(1), \cdots, \bar{c}(s-1), 0)$ of (9) continuous in u, and by the "almost" uniqueness of c the Fact follows. □

Corollary 4.1: Let (γ, c) *be a solution to* (6). *Then* $c(21) - c(12) > 0$ *implies* $u^* = 0$; $c(21) - c(12) < 0$ *implies* $u^* = 1$; *and if* $c(21) - c(12) = 0$, *any* u^* *will do.*

Corollary 4.1 suggests the following algorithm to estimate u^* adaptively.

Algorithm 1:

1. Start the system with an arbitrary u.
2. Use Proposition 6 to estimate the sign of $c(21) - c(12)$.
3. Use Corollary 4.1 to choose u^*.

We will conclude this section with an example. Consider the case of minimizing the probability of conflict, i.e., the probability that both processes request simultaneously. In this case, $k(i) = 0$ for $i \neq 11$. It follows that

$$c(21) - c(12) = k(11) \frac{\bar{T}_{12,11} - \bar{T}_{21,11}}{\bar{T}_{11,11}},$$

and

$$\text{sign}[c(21) - c(12)] = \text{sign}[\bar{T}_{12,11} - \bar{T}_{21,11}].$$

An intuitive justification for the above equation is the following. Minimizing $P(11)$ is equivalent to maximizing $\bar{T}_{11,11}$, which is equivalent to giving priority to the process so that the busy period of the system corresponding to that process being serviced first is maximized. It is easy to see by using renewal arguments that this is equivalent to choosing the largest among the $\bar{T}_{12,11}$, $\bar{T}_{21,11}$.

4.3 The case of n resources

In this section we will generalize some of the previous results. Each process has a thinking state 0 as before, and pair $(1_k, 2_k)$ of request and resource holding states for every resource k. Hence there are n conflict states, and the control u is the n-tuple $(u_1, \cdots, u_n)$, where u_k is the probability of assigning resource k to process 1 in the conflict state $(1_k, 1_k)$. See Fig. 3. Again, we are concerned with the estimation of u^*. To simplify notation, we will prove the theorems for $n = 2$; the same proofs hold for any n larger than 2. We follow the sequence of the previous section.

Theorem 5: u^* can always be selected from $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$.

Proof: Assume that each process consists of the states $(0, 1_1, 2_1, 1_2, 2_2)$, where states 1_1, 2_1 are associated with resource 1, and 1_2, 2_2 with resource 2. The proof is similar to the proof of Theorem 3, since the optimality conditions are

$$\gamma^* = \min\{[c^*(1_12_1) - c^*(1_11_1) + k(1_11_1)] + u_1[c^*(2_11_1) - c^*(1_12_2)] \mid u_1 \in U_1\}, \tag{10a}$$

$$= \min\{[c^*(1_22_2) - c^*(1_21_2) + k(1_21_2)] + u_2[c^*(2_21_2) - c^*(1_22_2)] \mid u_2 \in U_2\}. \quad \square \tag{10b}$$

The following algorithm is a policy iteration algorithm that starts with an arbitrary initial choice of u and in a finite number of repetitions converges to u^*. Step 2 corresponds to the value determination operation, and Step 3 to the policy improvement operation. The novel feature of this algorithm is the simple and adaptive execution of the value determination operation by using Proposition 6. The ergodicity of the system ensures that every improvement of the policy corresponds to a strict decrease in cost (see Refs. 15 and 16). Define $A_1 :=$

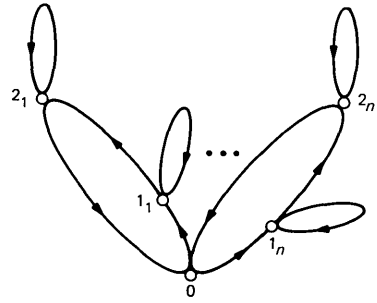


Fig. 3—Process i $(i = 1, 2)$, n resources.

$c(21) - c(12)$ and $A_2 := c(2'1') - c(1'2')$. (Observe that A_1, A_2 are the coefficients of u_1, u_2 in eq. [10].)

Algorithm 2:

1. Start with some arbitrary $u^k = (u_1^k, u_2^k)$.
2. Using Proposition 6, estimate the sign of A_1^k and A_2^k, where $A_i^k = A_i(u^k)$, $i = 1, 2$.
3. Choose in an extreme way the u_i^{k+1}, $i = 1, 2$, in order to decrease each $A_i^k u_i^{k+1}$ separately (i.e., if $A_i(u^k) > 0$, choose $u_i^{k+1} = 0, \cdots$). If $u^{k+1} \neq u^k$, go to Step 2 by using $u = u^{k+1}$. If $u^{k+1} = u^k$, then $u^k = u^*$.

The convergence follows, since every choice of a different u^k corresponds to a strict decrease of the cost (by ergodicity), and there is a finite number of choices for u^k (finite extreme points in U). The essential difference between the cases $n = 1$ and $n > 1$ is that Theorem 4 cannot be generalized for $n > 1$. This prevents us from inventing a "one step" algorithm for the estimation of u^*.

Note that although not addressed in this paper, the problem of estimating the $\bar{T}_{ij}$'s is of great importance for the algorithm. Substantial errors in the estimation procedure could lead to a wrong choice of u^* by the algorithm.

4.4 A generalization

One can notice that throughout Sections 4.2 and 4.3, the only parts of the structure of P used were the rows corresponding to conflict states. This leads to a generalization of the form of the processes. A process can consist of an arbitrary set S_0 of thinking states and a pair $(1_k, S_2^k)$ for each resource k, where S_2^k is an arbitrary set of service states. The only constraint on the transition diagram of the process is that there is a unique state in S_2^k to which a process k can transit from state 1_k.

V. CONCLUSIONS

One open problem is the relation between the distributions of service and thinking times, and the safe regions of Section III. In other words, is there a general rule suggesting that for large ratios of thinking time of the processes, the choice of u^* is independent of service times? This would be nice, since no further calculations are needed to obtain the optimal u^*.

Another open problem is the generalization to m processes sharing n resources. A realistic model for this situation would suggest a decentralized information structure for each resource scheduler. By this we mean that the scheduler of resource r should base its decision on "local" information only, i.e., the state of resource r and the identity of processes requesting resource r. If one uses Markov Decision Theory

as was done here, the optimal decision of scheduler r concerning conflict between processes i, j would depend on "global" information about the system, i.e., the state of all other processes even if they are not involved in this particular conflict. This is unsatisfactory. Instead one would like to obtain a rule for making local decisions that are optimal "in the average", by "smoothing out" what happens in the rest of the system.

As a final open question, we state the generalization of Section III. One should be able to prove the existence of the safe regions of Theorem 2 without explicitly calculating the invariant probability measure.

VI. ACKNOWLEDGMENTS

The authors would like to thank Professor Jean Walrand for many helpful discussions and for his shortened version of the proof of Proposition 5.

REFERENCES

1. J. Harrison, "Dynamic Scheduling of a Multiclass Queue: Discount Optimality," Oper. Res., *23* (March–April 1975), pp. 270–82.
2. J. Harrison, "A Priority Queue With Discounted Linear Costs," Oper. Res., *23* (March–April 1975), pp. 260–9.
3. G. Klimov, "Time-Sharing Service Systems, I," Theory Probab. Appl., *19* (September 1974), pp. 532–51.
4. I. Meilijson and G. Weiss, "Multiple Feedback at a Single-Server Station," Stochastic Processes and Applications, *5* (1977), pp. 195–205.
5. L. Schrage, "A Proof of Optimality of the Shortest Remaining Processing Time Discipline," Oper. Res., *16* (1968), pp. 687–90.
6. I. Meilijson and U. Yechiali, "On Optimal Right-of-Way Policies at a Single Server Station When Insertion of Idle Times is Permitted," Stochastic Processes and Applications, *6* (November 1977), pp. 25–32.
7. P. Nash and R. Weber, "Dominant Strategies in Stochastic Allocation and Scheduling Problems," in *Deterministic and Stochastic Scheduling*, M. A. H. Dempster, ed., Dordrecht, The Netherlands: Reidel, 1982, pp. 343–53.
8. D. Smith, "Optimal Repair of a Series System," Oper. Res., *26*, No. 4 (July–August 1978), pp. 653–62.
9. J. Kemeny and J. Snell, *Finite Markov Chains*, New York: Springer Verlag, 1976.
10. N. Jaiswal, *Priority Queues*, New York: Academic Press, 1968.
11. P. Billingsley, *Statistical Inference for Markov Processes*, Chicago: The University of Chicago Press, 1961.
12. J. Kemperman, *The Passage Problem for a Stationary Markov Chain*, Chicago: The University of Chicago Press, 1961.
13. V. Borkar and P. Varaiya, "Adaptive Control of Markov Chains, I: Finite Parameter Set," IEEE Trans. Automat. Contr., *24*, No. 6 (December 1979), pp. 953–7.
14. P. Varaiya, "Optimal and Suboptimal Control for Markov Chains," IEEE Trans. Automat. Contr., *23*, No. 3 (June 1978), pp. 388–94.
15. S. Ross, *Applied Probability Models with Optimization Applications*, San Francisco: Holden Day, 1970.
16. H. Mine and S. Osaki, *Markovian Decision Processes*, New York: Elsevier, 1970.

APPENDIX

Proof of Theorem 1: After calculating, we have

$$AD - BC = p_1 p_2 q_1^2 q_2^2 F_1 F_2,$$

where

$$F_1 = p_1^3p_2q_2 - p_1^2p_2q_2 - p_1^2q_2 - p_1p_2^3q_1 - p_1p_2^2q_1 + p_2^2q_1 + p_1p_2^2 - p_2^2 - p_1^2p_2 - p_2 + p_1 + p_1^2 \quad (11)$$

$$F_2 = q_1q_2[p_1p_2^2 + p_1^2p_2 - p_1p_2 - p_1 - p_2 + 1] + (p_2 - p_1p_2)q_2 + (p_1 - p_1p_2)q_1. \quad (12)$$

Let $L(p_1, p_2) := p_1^2p_2 + p_1p_2^2 - p_1p_2 - p_1 - p_2 + 1$. Then by (12)

$$F_2 = q_1q_2L + (p_2 - p_1p_2)q_2 + (p_1 - p_1p_2)q_1.$$

Consider $G_1(q_1) := F_2(p_1, p_2, q_1, q_2)$. Then

$$G_1(q_1) = q_1[q_2L + (p_1 - p_1p_2)] + (p_2 - p_1p_2)q_2.$$

Since $G_1(q_1)$ is linear in q_1, to prove $G_1 \geq 0$ for all $p_1, p_2, q_1, q_2 \in [0, 1]$, it is enough to show $G_1(0) \geq 0$ and $G_1(1) \geq 0$. But $G_1(0) = (p_2 - p_1p_2)q_2 \geq 0$; hence we only have to prove $G_1(1) \geq 0$. Let $G_2(q_2) := G_1(1)$. Using a similar argument, since G_2 is linear in q_2 and $G_2(0) = p_1 - p_2p_2 \geq 0$, we only have to show $G_2(1) \geq 0$. But this holds since $G_2(1) = p_1p_2(p_1 + p_2 - 2) - p_1p_2 + 1 \geq 0$, as one can readily check. □

Proof of Theorem 2: Consider the function $G_1(p_2) = F_1(p_1, p_2, q_1, q_2)$ and $q_1 \leq q_2$. Then proving the theorem is equivalent to proving the following:

1. $G_1(p_2)$ has a unique root p_2^0 in the interval $[p_1/2, p_1]$.
2. $G_1(p_2)$ has no root in the intervals $[0, p_1/2]$, $[p_1, 1]$.
3. $p_2 \leq p_2^0$ implies $G_1(p_2) \geq 0$, and $p_2 \geq p_2^0$ implies $G_1(p_2) \leq 0$.

One can now prove 1 through 3 since

$$G_1(1) \leq 0,\ G_1(p_1) \leq 0,\ G_1(p_1/2) \geq 0,\ G_1(0) \geq 0,\ p_2^0p_2^1p_2^2 > 1,$$

where p_2^0, p_2^1, p_2^2 are the roots of $G_1(p_2) = 0$, and $F_1(a, b, c, d) = -F_1(b, a, d, c)$, as one can easily check.

Proof of Proposition 4: Define $\bar{c} := (c(1), \cdots, c(s-1), 0)$, and $Q := (P - I)$. Then

$$\gamma\perp = Q\bar{c} + k \quad \text{has a unique solution} \quad (\gamma, \bar{c}). \quad (13)$$

Let $y_a := V_a - V_a(s)\perp$. Then by using (7) and subtracting, we get

$$(1 - a)V_a(s)\perp = (aP - I)y_a + k. \quad (14)$$

By multiplying (7) by the invariant measure π, we get $\pi[(1 - a)V_a - k] = 0$, which, together with the ergodicity assumption (i.e., all components of π are >0) and the fact that $V_a \geq 0$, implies that no component of V_a tends to ∞ as $a \to 1$. Let $\gamma_a := (1 - a)V_a(s)$. By the previous argument it follows that there is a converging subsequence

of γ_a as $a \to 1$, and for this subsequence let $\gamma^* := \lim_{a\to\infty}\gamma_a$. By using (14) we find

$$\gamma_a \perp = (aP - I)y_a + k. \tag{15}$$

Consider (15) as $a \to 1$ along the above subsequence. Since it has a unique solution y_a for every a, it follows that there is a unique $y^* := \lim_{a\to 1} y_a$, and (13) implies that $\gamma = \gamma^*$ and $\bar{c} = y^*$. The proof now follows by observing that $V_a(i) - V_a(k) = y_a(i) - y_a(k)$, and $c(i) - c(k) = \bar{c}(i) - \bar{c}(k)$ by Proposition 1. □

Proof of Proposition 5: Let X_t be the chain under consideration, $X_t \in \sum_1 := \{1, \cdots, s\}$, and define Y_t, $t = 0, 1, \cdots$, $Y_t \in \sum_2 := \{0, 1\}$ such that $P[Y_{t+1} = 1 \mid Y_t = 0] = 1 - a$, $P[Y_{t+1} = 1 \mid Y_t = 1] = 1$, i.e., is an absorbing chain with one being the absorbing state. We also define the following:

$N_i^a[j]$:= Number of visits by X_t to state j starting from i before absorption,

$T_1 := \text{Min}\{n \geq 0 \mid X_n = j\}$,

$T_2 := \text{Min}\{n \geq 0 \mid Y_n = 1\}$,

z := Number of visits to state j before absorption. □

Lemma: $\bar{N}_i^a[j] = E_{(j,0)}[z]P_i[T_2 > T_1]$.

Proof: Start the system (X_t, Y_t) from state $(i, 0)$ and count visits to state j before absorption. One can always start counting from time T on, since no visit to state j occurs before T. The count z will not be identically zero only when $(X_t, Y_t) = (j, 0)$, and this occurs with probability $P_i[T_2 > T_1]$. By the strong Markov property, one can always restart the system from state $(j, 0)$, and the result follows since the new expected count will be $E_{(j,0)}[z]$.

Note that $\bar{N}_j^a[j] = E_{(j,0)}[z]$ since $P_j[T_2 > T_1] = 1$.

Let $\beta_{ij}^a := P_i\{X_t^a = j \text{ for some } t > 0\}$. Then by using the previous lemma, it follows that

$$\bar{N}_i^a[j] = \bar{N}_j^a[j]\beta_{ij}^a, \qquad \text{for} \quad i \neq j,$$

and by using similar renewal arguments,

$$\bar{N}_j^a[j] = 1 + \beta_{jj}^a \bar{N}_j^a[j];$$

hence

$$\bar{N}_j^a[j] = \frac{1}{1 - \beta_{jj}^a}.$$

Since $\beta_{ij}^a = E[a^{T_{ij}}]$, by using dominated convergence we obtain

$$\frac{\partial}{\partial a}\beta_{ij}^a = E[T_{ij}a^{T_{ij}-1}]$$

and

$$\lim_{a\to 1}\frac{\partial}{\partial a}\beta_{ij}^a = \bar{T}_{ij}.$$

Therefore if $i \neq j \neq k$, then

$$\begin{aligned}\lim_{a\to 1}[\bar{N}_i^a[j] - \bar{N}_k^a[j]] &= \lim_{a\to 1}[(\beta_{ij}^a - \beta_{kj}^a)\bar{N}_j^a[j]] \\ &= \lim_{a\to 1}\frac{\beta_{ij}^a - \beta_{kj}^a}{1-\beta_{jj}^a} \\ &= \frac{\bar{T}_{kj} - \bar{T}_{ij}}{\bar{T}_{jj}},\end{aligned}$$

by de l'Hospital's rule. The proof for $k = j$ is similar. □

AUTHOR

Costas Courcoubetis, Diploma (Electrical and Mechanical Engineering), 1977, National Technical University of Athens, Greece; M.S., and Ph.D. (Electrical Engineering and Computer Science), 1980 and 1982, respectively, University of California at Berkeley; AT&T Bell Laboratories, 1982—. He was a Research Assistant in the Electronic Research Lab at U.C. Berkeley, and an Associate Member of Technical Staff at Bell Laboratories during the summer of 1981. He is currently a Member of Technical Staff in the Mathematics and Statistics Research Center. His interests are distributed algorithms and computation, queueing theory, performance analysis of computer systems, and design and verification of computer communication protocols.

AT&T Technical Journal
Vol. 64, No. 1, January 1985
Printed in U.S.A.

An Infant Mortality and Long-Term Failure Rate Model for Electronic Equipment

By D. P. HOLCOMB* and J. C. NORTH†

(Manuscript received June 29, 1984)

This paper describes the reliability model used by system designers at AT&T Bell Laboratories to predict component and equipment reliability. A decreasing-failure-rate Weibull model describes the high incidence of early-life failures, or infant mortality. This is combined with the constant-failure-rate (exponential) model traditionally and widely used for the long term. Formal modeling of both early-life and long-term reliability is needed to manage the development and manufacture of reliable products. The effects of temperature and electrical stress on failure rate are taken into account. A model for the effect of integrated circuit dynamic burn-in on reliability is also described.

I. INTRODUCTION

Reliability describes the ability of a system to continue to perform its required function to the satisfaction of the user. Predicting the reliability of a new electronic system is an important part of the system design process. If the design will not meet reliability objectives, it must be improved by using more reliable components, adding system redundancy (tolerance to component failures), or performing burn-in or other screening.

The ability to satisfy customers is the most critical factor for a viable product. But reliability has other economic impacts as well. Repair costs, both during a warranty period and beyond, depend on

* AT&T Information Systems. † AT&T Bell Laboratories.

the reliability. Reliability also affects the numbers of spares needed in the field to accommodate expected repair needs.

System reliability depends on the reliability of its components in their application environment. Unfortunately, it is impossible to predict the operating life for any individual electronic component. It is, however, possible to treat large populations of such components statistically with acceptable results. For example, the number of failures among a large population of components and each component's probability of survival or failure can be estimated. The statistical behavior of the components then determines the statistical behavior of the entire system. In this way the reliability of an electronic system can be estimated.

This paper describes the techniques and models currently used by system designers at AT&T Bell Laboratories to predict the failure probabilities of electronic components. There are other widely used reliability prediction models. Most are based on a simple constant-failure-rate model. One such widely used method is described in Ref. 1, MIL-HDBK-217D. For many purposes, these models are inadequate. Early-life reliability predictions provide information needed to balance requirements, design, screening, and field support activities for commercial products. By not reflecting the relatively high failure rates associated with early equipment life, the constant-failure-rate models do not provide the information needed to manage these early-life reliability issues.

The model described here is more realistic. It has elements in common with the conceptual "bathtub curve" reliability model. Under this model, a relatively large number of defects can be expected early in equipment life. This is referred to as "infant mortality". The likelihood of failure then falls dramatically to a low, constant level called long-term reliability. This low failure rate is the behavior expected of mature products. Eventually, components can degrade and the incidence of failure increases during "wear-out". The model presented here reflects both infant mortality and long-term behavior. Integrated circuit reliability dominates the reliability of modern electronic equipment. Because properly designed and manufactured silicon integrated circuits do not experience "wear-out" behavior, it is not modeled.

To go along with the model, we also have tabulated, elsewhere, reliability estimates for a wide variety of components. These estimates are based on AT&T Bell Laboratories data whenever possible. Otherwise, estimates are obtained from MIL-HDBK-217D, the *de facto* industry standard.

As we already mentioned, predictions of reliability are useful in estimating its impact on both customer satisfaction and economic

viability. Both the early-life and long-term aspects of reliability are important and need to be addressed. The need to address long-term reliability is well known. In addition, formal modeling of early-life reliability provides information essential to managing the design, manufacture, reliability testing, and screening programs needed to assure that initial product reliability will satisfy customers.

II. INFANT MORTALITY (SHORT-TERM RELIABILITY)

Infant mortality is characterized by an initially high, but rapidly decreasing, failure rate. The early failures come from a small fraction of the components considered to be weak. These weak units contain defects (usually manufacturing defects) that are not immediately fatal but that will cause failure in a relatively short time. Examples of these defects are poor internal electrical connections, the presence of contaminants, and insulating layers that are too thin.

Failures due to infant mortality can appear in two different ways. In one, failures occur during operation after some time. These are called "device operating failures" or DOFs. The failures are time dependent. The infant mortality part of the reliability model describes their occurrence.

In addition to the DOFs, initial failures are found at various first tests, including first circuit pack tests, first system test, or when the system is first tested after shipment to the field. These failures are called "dead on arrivals" or DOAs. They cannot be related to operating time. A component can test as satisfactory, be assembled into equipment, and then fail to work. No operation has occurred. Instead, these failures may be thought of as event dependent rather than time dependent. Somehow, handling during equipment manufacture has induced failure of the weak component. DOAs are not reflected in the reliability model. Although their existence is well recognized, we do not know how to quantitatively predict their occurrence.

During infant mortality, components exhibit a "high incidence of failure", high relative to later life (the long term). This should be kept in perspective. Only a very small fraction of components actually fail (typically much less than one percent). See the discussion in Section VII on calculating numbers of failures for more details.

III. LONG-TERM RELIABILITY

Infant mortality failures are mostly caused by defects. Even in the long term, some failures continue to appear due to manufacturing defects; however, other failures occur due to more fundamental component properties. The important aspect, though, is that the failure rate is low and relatively constant in the long term. This is the behavior

observed in large populations of mature components. Failures occur at a fairly constant rate within the entire population; therefore, it can be treated as a homogeneous population of components having constant failure rates.

IV. RELIABILITY DEFINITIONS

Before looking at the specific reliability model, we should review some basic reliability definitions. Reliability models are based on the probabilities of survival or failure of a component or system. These probabilities can be described by one of several common functions.[2] Assume that a component starts to operate at time $t = 0$. Then $F(t)$ represents the probability that the component fails at or before time t. This is called a *cumulative distribution function* and it has the properties

$$\begin{aligned} F(t) &= 0 && \text{for} \quad t < 0, \\ 0 \le F(t) &\le F(t') && \text{for} \quad 0 \le t \le t', \\ F(t) &\to 1 && \text{for} \quad t \to \infty. \end{aligned}$$

The *reliability function*, $R(t)$, gives the probability of surviving past time t. It is related to $F(t)$:

$$R(t) = 1 - F(t).$$

This function is the source of the usual definition of reliability as "the probability of surviving". The derivative of $F(t)$ is a *probability density function* represented by $f(t)$:

$$f(t) = \frac{d}{dt} F(t)$$

or

$$F(t) = \int_0^t f(x)dx.$$

In practice, the instantaneous failure rate or hazard rate $\lambda(t)$ is often more useful than the functions just mentioned. From this point on, *failure rate* will mean instantaneous failure rate:

$$\lambda(t) = \frac{f(t)}{R(t)} = -\frac{d}{dt} \ln[R(t)],$$

which implies that

$$R(t) = e^{-\int_0^t \lambda(x)dx}.$$

The failure rate $\lambda(t)$ of a unit has the following interpretation: If the

unit has survived until t, the probability of failing in a small time interval Δt at t is $\Delta t\, \lambda(t)$.

V. COMPONENT FAILURE RATE MODEL

The reliability model described here applies to individual components. We use the failure rate function to describe the reliability model, since it is the most convenient.

The basic reliability model consists of two parts, shown by the heavy lines in Fig. 1. During the infant mortality period, the failure rate is described by a two-parameter Weibull model. The Weibull failure rate[3] can be expressed as

$$\lambda(t) = \lambda_1 t^{-\alpha}.$$

This distribution appears as a straight line when plotted on logarithmic scales, as in Fig. 1. The slope of the line is $-\alpha$ and the intercept at $t = 1$ hours is λ_1. The failure rate is initially high but decreases rapidly.

Beyond 10,000 hours the model assumes that the failure rate is constant. The exponential model, which implies a constant failure rate, is used. The long-term failure rate is simply

$$\lambda(t) = \lambda_L.$$

Defining the Weibull-to-exponential switchover point to be at ex-

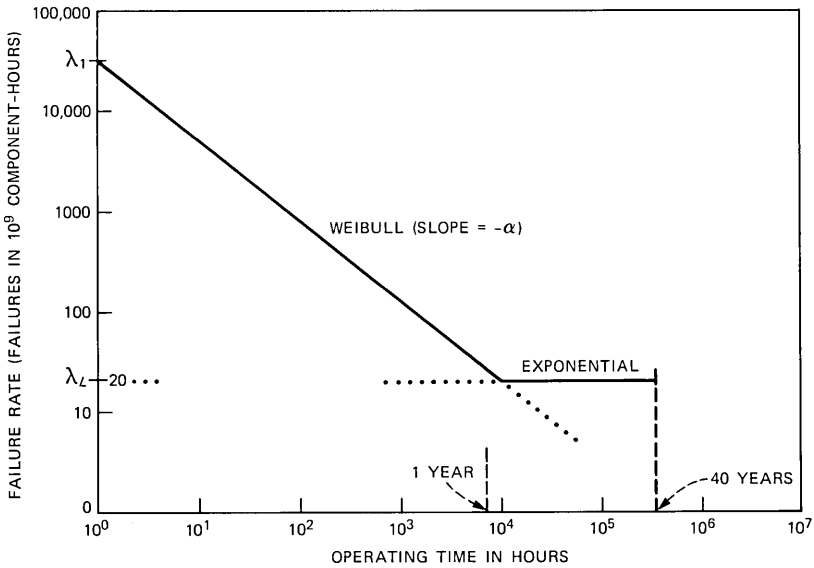


Fig. 1—A component failure rate model (shown by solid lines) combining a Weibull model and an exponential distribution. The failure rates shown here are typical, but are not intended to correspond to any particular component.

actly 10,000 hours is arbitrary but reasonable. Weibull behavior has been observed to persist for at least a year (8760 hours). At that point, the failure rate is changing very slowly. Therefore, any time somewhat greater than one year could have been chosen as the switchover point with little impact on the modeled failure rates; 10,000 hours was a conveniently round number.

There are two distinct sources of information on component reliability: direct monitoring of performance in the factory or field, and accelerated life tests. Factory or field data give a real measure of component reliability in the short term (a year or so), and some data are also available for the long term. Accelerated life tests provide information about reliability expected in the very long term (tens to hundreds of years).

Reliability studies seldom continue for more than two or three years. Therefore, primarily infant mortality is observed. Plotting the logarithm of the observed failure rate versus the logarithm of operating time usually gives a straight line. The straight line means that a Weibull distribution describes the failure rate behavior well, as we see in Fig. 2. We use the Weibull model to describe infant mortality because of such data.

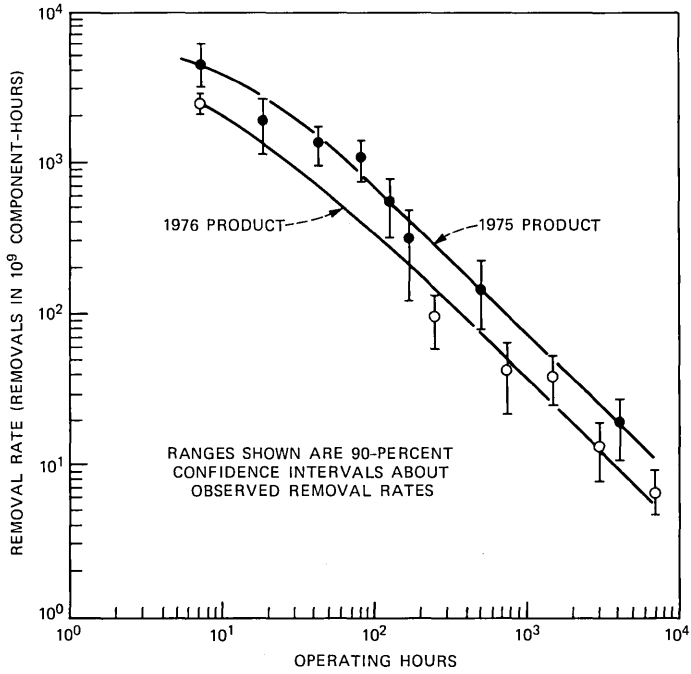


Fig. 2—Infant mortality removal rate of beam-lead sealed-junction T^2L integrated circuits (small-scale integration/medium-scale integration) (see Ref. 4).

As just mentioned, we can directly measure the time dependence of infant mortality failures. However, measuring the lifetime distribution of the main population in the long term is impractical using normal operating conditions, since so few components fail. Therefore, we use accelerated test conditions to estimate the lifetime distribution and, hence, the failure rate, at normal use conditions. The test results imply that only a very small fraction of semiconductor components will fail at normal use conditions within a forty-year service life. This is a very important result. It means that wear-out *will not occur* during the service life. (In fact, wearout of semiconductor components should never occur.) Figure 3 illustrates this point. For semiconductors, a lognormal model describes the main lifetime distribution in accelerated tests.[5] Two such lognormal failure rate curves, extrapolated to normal use, are shown in Fig. 3. These examples represent values in the range usually observed.

Accelerated testing does not define the long-term failure rate very precisely during the service life. Accelerated life conditions are far removed from normal use conditions; therefore, a long extrapolation is required to estimate real field performance. This is inherently a tricky business. Furthermore, in accelerated life tests, the sample size is generally small. In such cases the tests cannot accurately show the distribution of the first few percent of the failures. However, only the lowest few percent of the population failure times will occur within

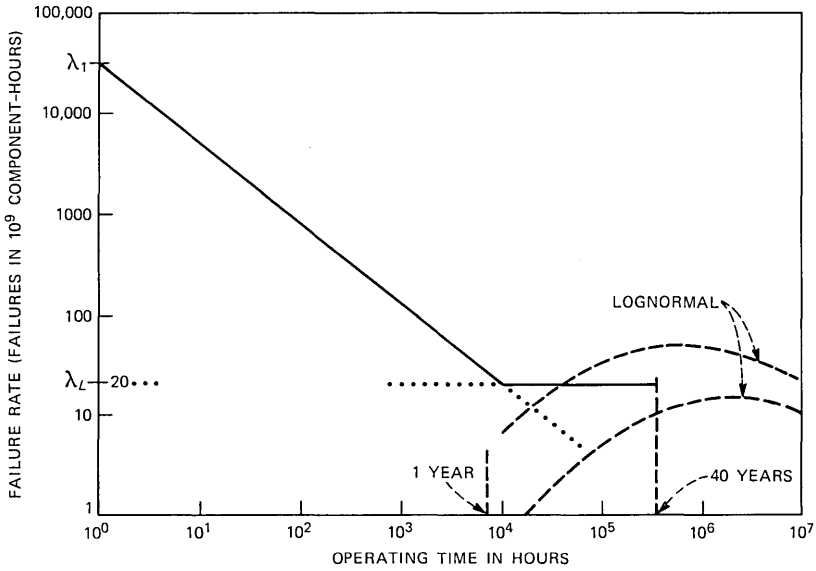


Fig. 3—The two lognormal curves (dashed lines) show possible relationships between accelerated-stress results and the basic failure rate model.

the service life. Therefore, accelerated testing does not accurately predict the failure rate distribution during the service life. For the lognormal examples shown in Fig. 3, only two percent of the components will have failed before the maximum failure rate is reached (in Fig. 1, at forty to four hundred years).

Since neither direct observations nor accelerated life testing defines the time dependence of the long-term failure rate, how then do we justify using the constant-failure-rate exponential model? We have used the exponential model largely by default, since it is almost universally used to model long-term failure rates as, for example, in MIL-HDBK-217D. The model is a reasonable compromise between the decreasing failure rate of the Weibull model and the increasing failure rate of the lognormal curve.

Field tracking does not define the failure rate during the very long term, but it can give failure rates at the end of infant mortality and the beginning of the long term. From those we can estimate the long-term failure rates. Accelerated life testing results are used in addition to tracking data. From these results we predict the total fraction of the population estimated to fail within the service life, which must be consistent with the failure rate estimates. If it is not, the estimates must be reevaluated.

5.1 Sources of component failure rates

If field tracking or accelerated testing data are available, the procedure just described is used to estimate the long-term failure rate. Where we do not have any relevant data, we use MIL-HDBK-217D numbers. With few exceptions, our estimates for semiconductor components come from AT&T Bell Laboratories data and those for nonsemiconductors come from MIL-HDBK-217D.

The infant mortality data that exist are solely from semiconductor components. These data form the basis for our estimates of the infant mortality parameters for all semiconductor components.

We do not have good infant mortality data on nonsemiconductor components. We do, however, believe that infant mortality exists for these components. Because we do not have good data, we chose a value of $\alpha = 0.6$ at the lower end of the range observed in equipment (between 0.6 and 0.9). Furthermore, we assume the infant mortality failure rate at 10,000 hours equals the long-term failure rate. These two assumptions, taken together, describe the existence of infant mortality, to a modest extent, in nonsemiconductor components.

5.2 Effect of temperature

Up to this point, we have only described the failure rate for a component as a function of time. In reality, a component's operating

environment will also affect the failure rate. The component's operating temperature is one such environmental factor. It can have a strong effect on the failure rate.

We base the component failure rate estimates on an assumed 40°C "typical" ambient temperature, since this is the temperature at which most of the available field tracking data are taken. Ambient temperature refers to the temperature in the immediate vicinity of the component. In most cases the failure rate estimates can be used directly. In cases where the temperature departs significantly from the typical value, its effect must be taken into account.

The effect of temperature is usually modeled through its effect on the rate of aging of a component. At a higher temperature, failures will generally occur sooner and, therefore, at a greater rate. The difference in rates of component aging at two different temperatures is described by an acceleration factor. If T_2 is a higher temperature than T_1, then the Arrhenius relationship gives the following acceleration factor:

$$A(T_2, T_1) = e^{\frac{E_a}{k_B}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)},$$

where k_B is the Boltzmann constant and E_a is the "activation energy". T_1 and T_2 are in units of degrees Kelvin. This Arrhenius relationship is well understood for chemical reactions. However, its use in the current context is based purely on empirical evidence. Therefore, the constant E_a does not really have physical meaning as an activation energy. Rather, it should be considered as an empirical curve-fitting constant.

With the exponential model, a factor of $A(T_2, T_1)$ increase in the rate of aging leads to a factor of $A(T_2, T_1)$ increase in the constant failure rate, that is:

$$\lambda_{T_2} = A(T_2, T_1)\lambda_{T_1}.$$

Under the Weibull model for infant mortality, the effect of temperature on the failure rate is not as simple. A factor of $A(T_2, T_1)$ increase in the rate of aging leads (after some careful algebra) to a factor of $A(T_2, T_1)^{1-\alpha}$ increase in the failure rate:

$$\lambda_{T_2}(t) = A(T_2, T_1)^{1-\alpha}\lambda_{T_1}(t).$$

It should be noted that the activation energy, and hence the acceleration factor, in the long term is not necessarily the same as during infant mortality. It depends on whether the expected cause of failures (failure mechanism) is the same. Table I lists the activation energies used.

Table I—Activation energies for selected components

Component	E_a (eV)	Reference
Infant Mortality		
All components	0.4	4
Long Term		
Discrete semiconductor	0.4	Unpublished work
Bipolar integrated circuits	0.4	Unpublished work
Metal-Oxide Semiconductor (MOS) integrated circuits	0.5	Unpublished work
Ceramic capacitors	1.0	6
Plastic capacitors—metallized and foil	0.12	6
Film resistors—metal or carbon	0.08	6
Carbon resistors	0.34	6

5.3 Effect of electrical stress

The level of electrical stress at which a component operates can also affect the failure rate. The higher the level of electrical stress, the more quickly we expect a component to fail. The effect of electrical stress, as for temperature, is modeled with an acceleration factor. This gives the difference in rate of aging at different values of applied electrical stress. The relationship we use to describe S, the acceleration factor, is

$$S(p_2, p_1) = e^{m(p_2 - p_1)},$$

where p_2 and p_1 are stress levels. These stress levels are given as a percentage of the maximum specified level. The electrical parameters that constitute electrical stress are different for different types of components (see Table II).

The stress constants (m) are based on information in MIL-HDBK-217D for long-term operation. There is no effect of electrical stress on integrated circuits, since the applied voltage is specified and assumed to be constant. Due to lack of information, no effects of electrical stress on failure rates during infant mortality are modeled.

The level of applied electrical stress assumed for typical operation is 25 percent. If actual levels of applied electrical stress significantly differ from 25 percent, then the acceleration must be taken into account. As with temperature acceleration, the electrical stress acceleration factor simply multiplies the constant long-term failure rate, as follows:

$$\lambda_{p_2} = S(p_2, p_1)\lambda_{p_1}.$$

Figure 4 shows how the combined effects of elevated operating temperature and high electrical stress can affect a component's failure rate. Both temperature acceleration (A_{LT}) and acceleration due to electrical stress (S) can affect the long-term failure rate. Only temperature acceleration (A_{IM}) affects the failure rate during infant mortality.

Table II—Electrical stress dependence (S) of selected components

Component	Electrical Stress Parameter	Range of m Values
Integrated circuits	Not applicable	—
Resistors	Power	0.006–0.024
Capacitors	Voltage	0.024–0.150
Switches	Current	0.013

Note that these acceleration factors change the time at which the long-term failure rate is reached in the model.

5.4 Effect of dynamic burn-in screening

One widely used method to reduce the impact of infant mortality on equipment is the "screening" method. This refers to some activity performed on components or equipment to screen or "weed out" infant mortality failures *before customer use.* Components or equipment are stressed in some way and then tested. Any failures are removed or repaired. By inducing these failures to occur prior to use, customers should experience fewer equipment failures. Some commonly used screens are thermal cycling, high-voltage stress, or simply electrical operation (burn-in).

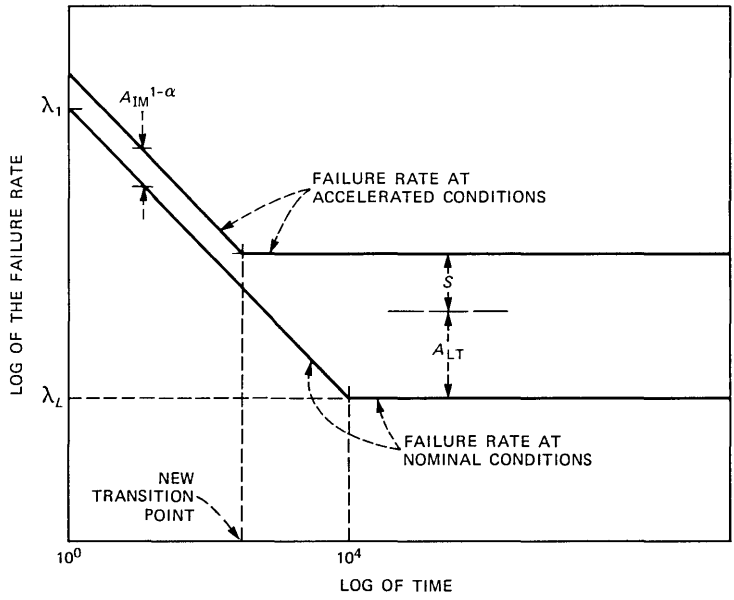


Fig. 4—Effect of temperature (A_{LT} and A_{IM}) and electrical stress (S) on failure rate model.

A variety of screens are believed to be effective in some cases. However, we have formally modeled the effects of only dynamic burn-in. This refers to the electrical operation of components or systems in a manner simulating eventual use. It consists of powering up and dynamically exercising the component or system for a period of time. This is distinguished from static burn-in in which power is applied but no dynamic exercise takes place. The burn-in may or may not occur at an elevated temperature.

We model only dynamic burn-in because we do not have a sufficiently good understanding of the quantitative effects of other screens. Even the understanding of the effects of dynamic burn-in is poor. Mathematically, our burn-in model follows naturally from the basic failure rate model. However, there is little direct evidence to substantiate the model.

One assumption provides the basis for the dynamic burn-in model. We assume that the failure rate of a component depends on the length of time of previous operation, *wherever that operation occurred.* As is clear from our model in Fig. 1, the more operating time a component has accumulated within the infant mortality period, the lower will be its failure rate. Therefore, a manufacturer can reduce the failure rate a customer will experience by operating components or equipment for a period of time before shipment.

Calculating the effect of dynamic burn-in is a matter of calculating the effective operating time to which the burn-in is equivalent. The effective operating time, t_{eff}, corresponds to operation at the nominal 40°C. Then, if t represents the amount of operating time after the burn-in, the failure rate at 40°C is

$$\lambda(t) = \lambda_1(t + t_{\text{eff}})^{-\alpha}.$$

Operation after burn-in might occur at a temperature higher than 40°C. In that case, the above failure rate is modified by the temperature acceleration factor already discussed.

Figure 5 illustrates the modeled effect of dynamic burn-in. The straight dashed line shows the basic infant mortality failure rate. The solid curve gives the failure rate after a burn-in equivalent in time to t_{eff}. Note that this curve is a simple replotting of the dashed line but starting at age t_{eff} rather than at zero age.

To calculate t_{eff}, we again make use of the temperature acceleration factor, A_{IM}, introduced earlier. If burn-in occurs at a temperature above 40°C for x hours, then the effective burn-in time is $t_{\text{eff}} = A_{\text{IM}}x$. ($A_{\text{IM}}$ is the acceleration factor, for the burn-in temperature, relative to 40°C.)

Burn-in can be performed at any of several stages. Components can be burned in. Burn-in can also occur at the circuit pack or at the

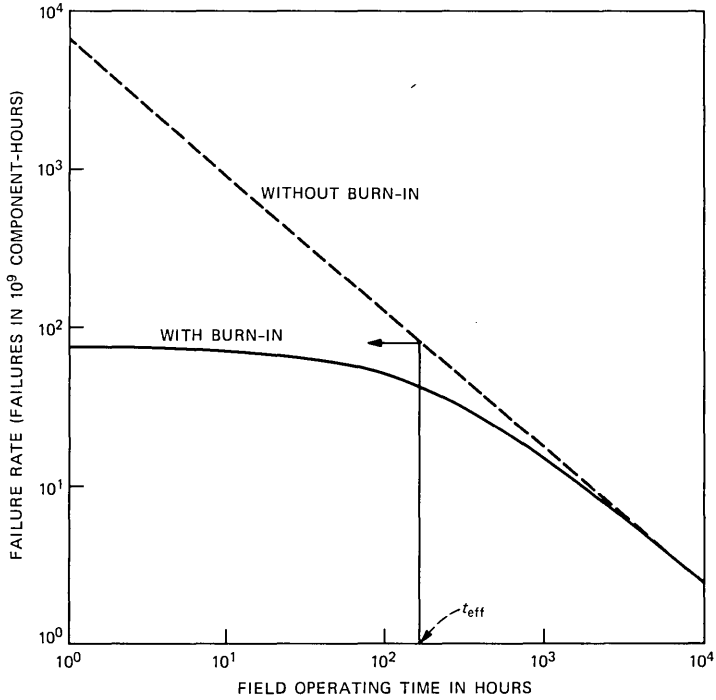


Fig. 5—Effect of burn-in on early system failure rate.

system level. Moreover, equipment can be burned in at more than one of these stages. Under the model, the effect of burn-in is cumulative. With more than one burn-in, the effective burn-in times at each stage are added to give the total effective burn-in time which is

$$(t_{\text{eff}})_{\text{total}} = \sum_{i}^{\text{all stages}} (t_{\text{eff}})_i.$$

Caution should be exercised in using this additive assumption of burn-in treatments. In some instances the initial failure rate of a population of devices during a second burn-in was larger than the final failure rate during the first burn-in. This "setback" may have been due to the testing of the devices and insertion of the devices into circuit boards, which was done between the two burn-in treatments. This additional handling may damage some devices (e.g., by electrostatic discharge) causing them to fail sooner than they would have otherwise.

VI. FAILURE RATES FOR EQUIPMENT

Up to this point, we have described the failure rate models applied to individual components. It is a simple matter to combine the com-

Table III—Environmental application factors

Environment	E
Permanent structures, environmentally controlled	1.0
Ground shelters, not temperature controlled	1.1 (Ref. 7)
Manholes, poles	1.5 (assumed)
Vehicular-mounted equipment	8.0 (Ref. 7)

ponent failure rates to estimate the failure rates of equipment. Basically, component failure rates are added together to give equipment failure rates. There is, however, one final modification we make to the summed failure rates. This modification accounts for a failure rate effect, which we do not understand well enough to apply at the component level. Rather, we apply it at the equipment level.

The modification involves an environmental application factor, E. It reflects environmental factors other than temperature that affect the equipment failure rates. Values of E are listed in Table III. These cover the usual environments for AT&T telecommunications equipment. They are based on information in Ref. 7.

With the inclusion of the equipment-level application factor, equations giving equipment failure rates can be written. For infant mortality the equipment failure rate is

$$\lambda_{\text{total}}(t) = E \sum_{i}^{\text{all components}} (A_{\text{IM}})_i^{1-\alpha}(\lambda_1)_i[t + (t_{\text{eff}})_i]^{-\alpha_i}.$$

The long-term failure rate is

$$\lambda_{\text{total}} = E \sum_{i}^{\text{all components}} (A_{\text{LT}})_i S_i(\lambda_L)_i.$$

VII. PREDICTING THE NUMBER OF INFANT MORTALITY FAILURES IN A TIME INTERVAL

Being able to predict the numbers of failures during the infant mortality period is important for a number of reasons. The estimates are useful for anticipating customer reaction. They can be used to understand warranty repair costs and to plan repair strategies. If any of these factors appear undesirable or unacceptable, the design, screening, or requirements of the products can be reevaluated. If this is done at an early enough stage, changes can be made when the impact on cost is low. Predictions also provide standards against which component or equipment performance can be measured. Once production begins, the results of ongoing reliability testing can be compared to the standards to show where to concentrate ongoing efforts to make improvements.

One calculation that is particularly useful is predicting the number

of failures out of some population during a stated time interval. The fraction of units failing between t_1 and t_2 is

$$\text{fraction fail} = \int_{t_1}^{t_2} \lambda(t)dt,$$

where $\lambda(t)$ is the unit's time-dependent failure rate. This approach is strictly correct when all failures in a population are repaired. In addition, it is correct if they are replaced with units of the same age. However, it is also a good approximation even if failing units are replaced by units of different age or not at all, if the fraction failing is small (less than a few percent).

The following example illustrates the strong effect of infant mortality. Typical failure rate parameters for integrated circuits are λ_L = 10 FITs, $\alpha = 0.8$, and λ_1 = 16,000 FITs. (One FIT equals one failure per 10^9 component-hours.) In the long term, we would expect about 10^{-8} failures per component-hour, or 7.2×10^{-6} failures per component-month. In a system made up of 10,000 such components, we would expect 0.072 failures per system during one month. In the infant mortality period, the situation is drastically different. In the *first* month, we would expect

$$\text{fraction fail} = (1.6 \times 10^{-5}) \int_0^{720} t^{-0.8}dt,$$

or roughly 0.0003 failures per component. This gives three failures per system in the first month. The number of failures expected during the first month is roughly forty times higher than expected in one month in the long term. Clearly, such an effect, if not anticipated, could lead to nasty surprises.

VIII. COMPARISON WITH MIL-HDBK-217D

Many of the concepts embodied in our failure rate model are also used in MIL-HDBK-217D, the *de facto* industry standard for failure rate prediction. There are, however, some important differences between the two models as well. As we already mentioned, the largest difference is in our formal inclusion of a model for infant mortality. This enhancement overcomes the major shortcoming of the MIL-HDBK-217D methodology, especially as applied to commercial products.

A second difference lies in the magnitudes of long-term failure rate estimates for integrated circuits. Our numbers are based on predivestiture Bell System experience. They are generally lower than those in MIL-HDBK-217D. Our data show that the MIL-HDBK-217D failure rates for integrated circuits are unrealistically high, especially those

for Large-Scale Integration (LSI) components. A final difference is in our dynamic burn-in model, which predicts an effect of burn-in on the infant mortality but not on long-term failure rates. MIL-HDBK-217D assumes that burn-in impacts long-term failure rates.

IX. SUMMARY

We have described the basic component reliability model used at AT&T Bell Laboratories to predict component and equipment reliability. It is a two-part model. A Weibull model describes infant mortality. An exponential model describes long-term behavior, beyond roughly the first year. The infant mortality part of the model is very important. It quantitatively describes the initially high, but rapidly decreasing, early-life failure rates. Recognizing such behavior is becoming critical as increasingly complex electronic systems are being sold to a variety of customers. Widely recognized methods of predicting reliability, such as MIL-HDBK-217D, do not model infant mortality effects.

Given the basic component reliability models and failure rate estimates, failure rates of equipment can be easily estimated. These equipment estimates, including infant mortality, are essential when planning for new, competitive product offerings and manufacturing them to have well-controlled reliability.

REFERENCES

1. MIL-HDBK-217D, *Military Handbook—Reliability Prediction of Electronic Equipment*, November 1, 1980 (draft), Department of Defense, Wash. DC 20301.
2. P. D. T. O'Connor, *Practical Reliability Engineering*, New York: Wiley, 1981, pp. 24–7.
3. N. L. Johnson and S. Kotz, *Continuous Univariate Distributions—1*, New York: Wiley, 1970, pp. 250–71.
4. D. S. Peck, "New Concerns About Integrated Circuit Reliability," Reliability Physics, 16th Annual Proceedings, 1978, pp. 1–6.
5. D. S. Peck, "Semiconductor Reliability Predictions From Life Distribution Data," in *Semiconductor Reliability*, Schwop & Sullivan, Eds., New York: Reinhold, 1961, pp. 51–67.
6. J. G. Gribbons, private communication based on approximations to Ref. 1.
7. R. C. Winans, *Physical Design of Electronic Systems*, Volume IV, Englewood Cliffs, NJ: Prentice-Hall, 1972, by Bell Telephone Laboratories, Inc., p. 316.

AUTHORS

Douglas P. Holcomb, B.S. (Engineering Physics), 1975, M. Eng. (Engineering Physics), 1976, Cornell University; AT&T Bell Laboratories, 1977–1982; AT&T Information Systems, 1983—. Mr. Holcomb joined the technical staff of AT&T Bell Laboratories in 1977, where, through 1981, he worked in the field of component reliability. Since then he has been Supervisor of a group having quality assurance responsibilities for PBX telecommunications products and, more recently, of a group developing a quality management system for data communications products.

James C. North, B.S. (Physics, Math), 1959, Capital University; Ph.D. (Solid State Physics), 1965, Purdue University; Research Assistant Professor, University of Illinois Physics Department, 1965; AT&T Bell Laboratories, 1968—. Mr. North joined AT&T Bell Laboratories, where he worked on ion implantation, channeling, backscattering and device development. In 1979, he became Supervisor of the Device Quality and Reliability Analysis group in the Quality Assurance Center, and since 1982 has been Head of the Device Reliability and Purchase Specifications Department.

AT&T Technical Journal
Vol. 64, No. 1, January 1985
Printed in U.S.A.

Waiting Time Convexity in the M/G/1 Queue

By D. L. JAGERMAN*

(Manuscript received June 8, 1984)

Strong bounds are obtained on the complementary waiting time distribution for the M/G/1 queue using the α-convexity structural characteristic of the distribution. This notion is discussed and a sufficient condition is obtained.

I. INTRODUCTION

This paper investigates the α-convexity[1] of the complementary waiting time distribution in the M/G/1 queue and shows how a sufficient condition for α-convexity is obtained, in terms of the service time distribution. This structure characteristic will permit strong bounds to be obtained on the complementary waiting time distribution. For convenience, the definition of α-convexity and some of its properties are given below.

II. α-CONVEXITY

A function $f(x)$ is said to be α-convex on an interval I if $e^{\alpha x}f(x)$ is convex on I. Of course, ordinary convexity corresponds to $\alpha = 0$. A sufficient condition for α-convexity is

$$e^{-\alpha x}\frac{d^2}{dx^2}(e^{\alpha x}f(x)) \geq 0, \qquad x \in I. \tag{1}$$

This is the same as

$$f''(x) + 2\alpha f'(x) + \alpha^2 f(x) \geq 0, \qquad x \in I. \tag{2}$$

*AT&T Bell Laboratories.

A function may be α-convex without being convex; for example, consider $f(x) = x^3$, which is α-convex for $x \geq 0$ and $\alpha \geq 0$. For $\alpha = 1$, however, x^3 is α-convex for $-3-\sqrt{3} \leq x \leq -3+\sqrt{3}$ as seen by use of (2).

The α-convexity of a function may permit stronger bounds than convexity to be obtained on the function or on integrals of the function. For example, let $p(x) \geq 0$ and let $f(x)$ be convex on I; then Jensen's inequality[2] states

$$\int_I f(x)p(x)dx \geq f(\mu) \int_I p(x)dx,$$

$$\mu = \int_I xp(x)dx \Big/ \int_I p(x)dx. \tag{3}$$

If $f(x)$ is α-convex on I, then, since

$$\int_I f(x)p(x)dx = \int_I e^{\alpha x}f(x)e^{-\alpha x}p(x)dx, \tag{4}$$

one has

$$\int_I f(x)p(x)dx \geq e^{\alpha\mu}f(\mu) \int_I e^{-\alpha x}p(x)dx,$$

$$\mu = \int_I xe^{-\alpha x}p(x)dx \Big/ \int_I e^{-\alpha x}p(x)dx. \tag{5}$$

This result can be stronger than (3).

An example is provided by

$$K = \int_0^\infty \frac{e^{-x}}{1+x}dx, \tag{6}$$

whose value is $K = 0.5963$. From (3), one has $K \geq 0.5$. Since, for $x \geq 0$, $1/(1+x)$ is α-convex for all α, one may apply (5) to obtain

$$K \geq \frac{e^{\frac{\alpha}{\alpha+1}}}{\alpha+2}, \tag{7}$$

which, for $\alpha = (\sqrt{5}-1)/2$, yields $K \geq 0.5596$.

Let $\tilde{f}(s)$ be the Laplace transform of a function $f(x)$, that is,

$$\tilde{f}(s) = \int_0^\infty e^{-sx}f(x)dx, \tag{8}$$

and let the transform be absolutely convergent for $s > 0$. Then the

approximation sequence, $f_n(x)$ $(n = 0, 1, 2, \cdots)$, introduced in the Laplace inversion theory, is given by[1]

$$f_n(x) = \frac{(-1)^n}{n!} s^{n+1} \tilde{f}^{(n)}(s) \Big|_{s=\frac{n+1}{x}}, \tag{9}$$

from which, in particular,

$$f_0(x) = \frac{1}{x} \tilde{f}\left(\frac{1}{x}\right),$$

$$f_1(x) = -\frac{4}{x^2} \tilde{f}'\left(\frac{2}{x}\right). \tag{10}$$

The approximation sequence may be used to obtain bounds on $f(x)$; thus, if $f(x)$ is convex for $x \geq 0$, then

$$f(x) \leq f_{n+1}(x) \leq f_n(x), \qquad x \geq 0, \qquad n \geq 0. \tag{11}$$

If $f(x)$ is α-convex on $x \geq 0$, then the bound of (11) may be strengthened. Let $\tilde{f}(s - \alpha)$ be absolutely convergent for $s > 0$; then it is the transform of a function $g(x)$ for which

$$f(x) = e^{-\alpha x} g(x). \tag{12}$$

Application of (11) now provides the inequality

$$f(x) \leq e^{-\alpha x} g_{n+1}(x) \leq e^{-\alpha x} g_n(x), \qquad x \geq 0, \qquad n \geq 0. \tag{13}$$

This is a much tighter inequality than (11), especially for the tail of $f(x)$, and constitutes the main tool for bounding the M/G/1 waiting time distribution.

The Bernstein theorem,[3] which states that $f(x) \geq 0$ if $\tilde{f}(s)$ is completely monotone and, conversely, may be used to translate condition (1) in terms of $\tilde{f}(s)$. Thus let $f''(x)$ be continuous on $(0, \infty)$; then $f(x)$ is α-convex on $(0, \infty)$ if and only if

$$(s + \alpha)^2 \tilde{f}(s) - (s + 2\alpha) f(0+) - f'(0+) \tag{14}$$

is completely monotone in s on $(0, \infty)$ and is absolutely convergent for $s > 0$.

A function, $f(x) > 0$, is said to be log-convex if $\ln f(x)$ is convex on some interval I. The condition for log-convexity is

$$f''(x) f(x) - f'(x)^2 \geq 0, \qquad x \in I. \tag{15}$$

In particular, log-convexity implies convexity; hence $e^{\alpha x} f(x)$ is convex, and a log-convex function is α-convex for all α. The converse is also

true. This follows from (2) on observing that the discriminant of the quadratic in α is $f'(x)^2 - f''(x)f(x)$; hence α-convexity for all α implies (15) and the log-convexity of $f(x)$. An interesting corollary of this is that the sum of log-convex functions is log-convex since, clearly, the sum of functions convex for the same α is again convex for this α. This theorem and eq. (14) permit ascertaining the log-convexity of $f(x)$ from its Laplace transform.

III. α-CONVEXITY IN M/G/1

The starting point for this investigation of α-convexity in M/G/1 is the Pollaczek-Khintchine formula.[4] Let $B(x)$ be the service time distribution and $F(x)$ the complementary waiting time distribution; also let $\hat{B}(s)$, $\hat{F}(s)$ be the corresponding Laplace-Stieltjes transforms. Then

$$\hat{F}(s) = \frac{\rho s - \lambda[1 - \hat{B}(s)]}{s - \lambda[1 - \hat{B}(s)]}, \qquad \rho < 1, \tag{16}$$

in which λ is the arrival rate, μ is the service rate, and $\rho = \lambda/\mu$ is the offered load. It is convenient to use the forward recurrence time distribution, $\theta(x)$, corresponding to $B(x)$. Since the Laplace-Stieltjes transform, $\hat{\theta}(s)$, of $\theta(x)$ is

$$\hat{\theta}(s) = \mu \frac{1 - \hat{B}(s)}{s}, \tag{17}$$

one has

$$\hat{F}(s) = \rho \frac{1 - \hat{\theta}(s)}{1 - \rho\hat{\theta}(s)},$$

$$\tilde{F}(s) = \frac{\rho}{s} \frac{1 - \hat{\theta}(s)}{1 - \rho\hat{\theta}(s)}, \tag{18}$$

in which $\tilde{F}(s)$ is the corresponding Laplace transform. Clearly,

$$\theta(s) \sim \frac{\mu}{s}, \qquad s \to \infty; \tag{19}$$

hence

$$\hat{F}(s) \sim \rho - \frac{\mu\rho(1 - \rho)}{s}, \qquad s \to \infty. \tag{20}$$

Thus,

$$F(0+) = \rho, \qquad F'(0+) = -\mu\rho(1 - \rho). \tag{21}$$

The information is now available to apply condition (14). That expres-

sion now takes the form

$$\frac{\tilde{N}(s)}{1 - \rho\hat{\theta}(s)}, \tag{22}$$

$$\tilde{N}(s) = \mu\rho(1 - \rho) + \frac{\alpha^2\rho}{s} - \left\{\rho(1 - \rho)(s + \lambda + 2\alpha) + \frac{\alpha^2\rho}{s}\right\}\hat{\theta}(s).$$

The function $1/[1 - \rho\hat{\theta}(s)]$ is the Laplace-Stieltjes transform of a monotone increasing function on $(0, \infty)$. If we write $\tilde{N}(s)$ in the form

$$\tilde{N}(s) = \frac{\alpha^2\rho}{s} + \lambda(1 - \rho)\hat{B}(s) - \frac{\lambda(1 - \rho)(\lambda + 2\alpha)}{s}[1 - \hat{B}(s)] - \frac{\lambda\alpha^2}{s^2}[1 - \hat{B}(s)]. \tag{23}$$

we see that $\tilde{N}(s)$ is a Laplace transform. For the function of (22) to be completely monotone, it is therefore sufficient that $\tilde{N}(s)$ be the transform of a nonnegative function. If we let $b(x)$ be the service-time density function, and $r(x)$ the corresponding rate function, that is,

$$r(x) = \frac{b(x)}{1 - B(x)}, \tag{24}$$

this condition may be written in the following two forms:

$$\frac{\alpha^2\rho}{\lambda(1 - \rho)} + b(x) \geqslant (\lambda + 2\alpha)[1 - B(x)] + \frac{\alpha^2}{1 - \rho}\int_0^x [1 - B(u)]du,$$

$$r(x) + \frac{\alpha^2}{\mu(1 - \rho)}\frac{1 - \theta(x)}{1 - B(x)} \geqslant \lambda + 2\alpha \tag{25}$$

to assure the α-convexity of $F(x)$.

The case $\alpha = 0$ of (25) yields the interesting result that convexity of $F(x)$ is guaranteed by

$$r(x) \geqslant \lambda, \qquad x \geqslant 0, \tag{26}$$

which also implies

$$1 - B(x) \leqslant e^{-\lambda x}, \qquad x \geqslant 0. \tag{27}$$

Application of (25) will now be made to the following class of distribution functions $B(x)$:

$$B(x) = \int_0^\infty (1 - e^{-xu}) dG(u) \tag{28}$$

in which $G(x)$ is a distribution function on $(0, \infty)$. Condition (25) now becomes

$$\int_0^\infty e^{-xu} \left(u - \lambda - 2\alpha + \frac{\alpha^2}{1 - \rho} \frac{1}{u} \right) dG(u) \geqslant 0. \tag{29}$$

Now the integrand is nonnegative if

$$2u \geqslant \lambda + 2\alpha + \sqrt{(\lambda + 2\alpha)^2 - \frac{4\alpha^2}{1 - \rho}}, \tag{30}$$

which suggests the introduction of the quantity c defined by

$$c = \inf_x [x;\ G(x) > 0]. \tag{31}$$

Thus condition (30) need be satisfied only for $u \geqslant c$, and hence

$$2c \geqslant \lambda + 2\alpha + \sqrt{(\lambda + 2\alpha)^2 - \frac{4\alpha^2}{1 - \rho}} \tag{32}$$

assures (29) and the α-convexity of $F(x)$. One implication of (32) is the following constraint on α:

$$\alpha \leqslant c(1 - \rho) - \sqrt{c(1 - \rho)(\lambda - c\rho)}. \tag{33}$$

Application of (33) to the exponential case $B(x) = e^{-\mu x}$ yields

$$\alpha \leqslant \mu(1 - \rho), \tag{34}$$

which, of course, is consistent with the known result

$$F(x) = \rho e^{-\mu(1-\rho)x}. \tag{35}$$

As another illustration, consider

$$B(x) = 1 - \frac{1}{2} e^{-x} - \frac{1}{2} e^{-2x} \tag{36}$$

for which $c = 1$, $\mu = 4/3$. One has

$$\alpha \leqslant 1 - \frac{3}{4}\lambda - \frac{1}{2}\sqrt{\lambda\left(1 - \frac{3}{4}\lambda\right)}. \tag{37}$$

Since

$$\hat{B}(s) = \frac{1}{2}\frac{1}{s+1} + \frac{1}{s+2}, \tag{38}$$

use of (18) yields

$$\tilde{F}(s) = \frac{\lambda}{4}\frac{3s+5}{s^2 + (3-\lambda)s + 2 - \frac{3}{2}\lambda}. \tag{39}$$

If we let

$$\gamma = \frac{-3 + \lambda + \sqrt{1+\lambda^2}}{2},$$

$$\delta = \frac{-3 + \lambda - \sqrt{1+\lambda^2}}{2},$$

$$A = \frac{3\sqrt{1+\lambda^2} + 1 + 3\lambda}{2},$$

$$B = \frac{3\sqrt{1+\lambda^2} - 1 - 3\lambda}{2}, \tag{40}$$

then we have

$$F(x) = \frac{\lambda}{4\sqrt{1+\lambda^2}}(Ae^{\gamma x} + Be^{\delta x}). \tag{41}$$

This distribution is, in fact, log-convex; thus (37) is overly restrictive. This might have been expected since (25) is only a sufficient condition for α-convexity.

IV. BOUNDS

When values of α have been determined by use of the complete monotonicity of $\tilde{N}(s)$ in (23), or by use of (25), then (13) may be applied to $\tilde{F}(s)$ in (18). It is, of course, advantageous to use as large a value of α as possible consistent with the requirement that $\tilde{F}(s - \alpha)$ be absolutely convergent for $s > 0$. Applying (10) and (13) to (18)

Table I—Numerical inversion for different α

	$F(x)$			
X	Exact Results	$\alpha = 0$	$\alpha =$ 0.34549	$\alpha =$ 0.69098
1	0.1718	0.1815	0.1753	0.1727
2	0.0834	0.0997	0.0883	0.0841
3	0.0414	0.0590	0.0460	0.0416
4	0.0207	0.0368	0.0245	0.0208
5	0.0103	0.0239	0.0133	0.0104
6	0.0052	0.0161	0.0073	0.0052
7	0.0026	0.0112	0.0041	0.0026
8	0.0013	0.0080	0.0023	0.0013
9	0.0007	0.0058	0.0013	0.0007

provides the following explicit bounds:

$$F(x) \leq \frac{e^{-\alpha x}}{1-\alpha x}\left[1 - \frac{1-\rho}{1-\rho\hat{\theta}\left(\frac{1}{x}-\alpha\right)}\right],$$

$$F(x) \leq \frac{4e^{-\alpha x}}{(2-\alpha x)^2}\left[1 - \frac{1-\rho}{1-\rho\hat{\theta}\left(\frac{2}{x}-\alpha\right)}\right] + \frac{4e^{-\alpha x}}{2x-\alpha x^2}\frac{\rho(1-\rho)\hat{\theta}'\left(\frac{2}{x}-\alpha\right)}{\left[1-\rho\hat{\theta}\left(\frac{2}{x}-\alpha\right)\right]^2}. \quad (42)$$

As a numerical example, the approximation $F_4(x)$ was calculated for $\tilde{F}(s)$ of (39) with $\lambda = 0.5$. Table I shows the exact results obtained from (41). It also shows the inversion with $\alpha = 0$ (that is, no α-enhancement used), the results with the value of α obtained from (37) (0.34549), and, finally, the calculation using the optimum choice, $\gamma = -\alpha$, i.e., $\alpha = 0.69098$.

As expected, the table shows improved accuracy as α is increased and, in particular, in the tracking of the tail behavior. This, of course, is the primary goal of α-enhancement. Observe that the approximate values are larger than the exact values, as implied by the α-convexity of $F(x)$ and (13).

REFERENCES

1. D. L. Jagerman, "An Inversion Technique for the Laplace Transform With Application to Approximation," B.S.T.J., *57*, No. 3 (March 1978), pp. 669–710.
2. G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, Cambridge: Cambridge University Press, 1959.

3. J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, Mathematical Surveys No. 1, Providence, RI: American Mathematical Society, 1943.
4. R. B. Cooper, *Introduction to Queueing Theory*, Second Edition, Amsterdam: North Holland, 1981.

AUTHOR

David L. Jagerman, B.E.E., 1949, Cooper Union; M.S., and Ph.D. (Mathematics), 1954 and 1962, respectively, New York University; AT&T Bell Laboratories, 1963—. Mr. Jagerman has been engaged in mathematical research on quadrature, interpolation, and approximation theory, especially related to the theory of widths and metrical entropy, with application to the storage and transmission of information. For the past several years, he has worked on the theory of difference equations and queueing, especially with reference to traffic theory and computers. He is currently preparing a text on difference equations with application to stochastic models.

AT&T Technical Journal
Vol. 64, No. 1, January 1985
Printed in U.S.A.

Accumulation of Jitter: A Stochastic Model

By C. CHAMZAS*

(Manuscript received March 29, 1984)

A problem that has been considered extensively in the past is the accumulation of jitter in a chain of regenerative repeaters. For simplicity it is usually assumed that all repeaters in the chain are identical. However, this is not the case in real systems, where considerable differences among the repeaters of a chain have been observed. These random variations are due mainly to manufacturing tolerances, aging effects, temperature changes, etc. In this work, we examine the accumulation of systematic and random jitter along the chain, when the repeater transfer functions are subjected to random variations. We derive expressions for the expected value and variance of the output jitter spectrum in terms of average values of the repeater's jitter transfer function. In addition we find the expected value and the variance of the RMS jitter. Finally we examine two special cases where the timing circuit employs either a phased-locked loop or a surface acoustic wave filter, and derive some asymptotic relations for the power spectral density and root-mean-square value of the accumulated systematic jitter.

I. INTRODUCTION

In a chain of self-timed regenerative repeaters for data transmission with Pulse Amplitude Modulation (PAM), each regenerator extracts the timing information (clock) directly from the received pulse train. Ideally the output of the timing circuit should be a sine wave with frequency equal to the baud rate of the data. In practice, however, imperfections of the circuits and noise in the transmission channel disturb the timing recovery operation so that the phase of the recovered

* AT&T Bell Laboratories.

clock is randomly deviated from the desired input clock phase. Such deviations, called *timing jitter*, produce a position modulation of the regenerated signal; the timing jitter tends to accumulate along the chain and degrade the system's performance. If the jitter introduced by the repeater's timing recovery circuitry is the *same* for each repeater (for example, if it depends on the data pattern), it is called *systematic jitter*. Otherwise it is called *random jitter*.

The problem of jitter accumulation in a chain of regenerative repeaters has been considered extensively in the past.[1,2] For simplicity it is usually assumed that all repeaters in the chain are identical. However, this is not the case in real systems, where considerable differences have been observed among repeaters.[3] These random variations are mainly because of manufacturing tolerances, aging effects, temperature changes, etc. In this work we examine the accumulation of systematic and random jitter along the chain, when the repeater's jitter transfer functions are subjected to random variations. Previous papers[1,4] have treated the relationship between the jitter transfer function and the retiming circuit parameters.

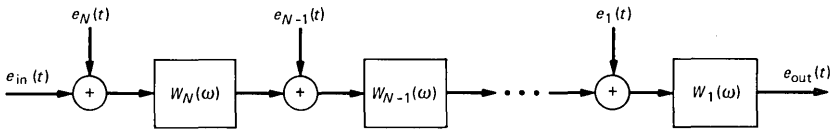
In this paper, we derive expressions for the expected value and the variance of the output jitter spectrum in terms of averages of jitter transfer functions of the ensemble of regenerators. In addition we find the expected value and the variance of the Root-Mean-Square (RMS) jitter.

We also examine two special cases where the timing circuit employs either a Phased-Locked Loop (PLL) or a Surface Acoustic Wave (SAW) filter. Finally we consider the case of a retiming circuit exhibiting random variations in its phase only and derive some asymptotic relations for the power spectral density and the RMS value of the accumulated systematic jitter. Of particular interest perhaps is relation (40), where we show that the RMS value of the accumulated systematic jitter for a nonpeaking case is approximately equal to $\sqrt{0.36*N/\alpha}$, where N is the number of regenerators and α the phase slope of the average jitter transfer function.

II. THE THEORY

2.1 The basic model

A brief review of jitter accumulation is presented here for completeness. The model we use to study the jitter accumulation is similar to the linear model attributed to Chapman, as described by Byrne et al.[1] Questions regarding the validity of this linear model will not be considered here. This problem will be addressed in a forthcoming paper. An oversimplified treatment is also given in Appendix B. According to this model, the timing jitter produced in a regenerative



$W_m(\omega)$: JITTER TRANSFER FUNCTION FOR THE mth REPEATER.

Fig. 1—Chapman's model for timing jitter accumulation in a chain of repeaters.

section is the filtered sum of the jitter coming from the previous section plus an additional equivalent jitter $e_i(t)$ inserted at the local input. This is illustrated in Fig. 1, where $W_i(\omega)$ is the jitter transfer function of the ith repeater. The major differences between our model and the ones used previously are that we do not assume $W_i(\omega)$ to be the same for every repeater.

The input jitter, $e_i(t)$, can be separated into its two components, the random part $e_{ir}(t)$ and the systematic part $e_{is}(t)$, i.e.,

$$e_i(t) = e_{ir}(t) + e_{is}(t). \tag{1}$$

The random component $e_{ir}(t)$ is different for each repeater and is usually caused by random sources, e.g., line noise, thermal noise, crosstalk, etc. The systematic component $e_{is}(t)$ is the *same* for each repeater and is usually pattern dependent.

Let $\Phi(\omega)$, $\Phi_r(\omega)$, $\Phi_s(\omega)$ be the two-sided power spectra of $e_i(t)$, $e_{ir}(t)$, and $e_{is}(t)$, respectively. Assuming that $e_{ir}(t)$ and $e_{is}(t)$ are statistically independent, then

$$\Phi(\omega) = \Phi_r(\omega) + \Phi_s(\omega). \tag{2}$$

If $S(\omega)$ is the spectrum of the output jitter then we can write

$$S(\omega) = \Phi_r(\omega)T_r(\omega) + \Phi_s(\omega)T_s(\omega), \tag{3}$$

where[1]

$$T_r(\omega) = |W_1(\omega)|^2 + |W_1(\omega)W_2(\omega)|^2 + \cdots + |W_1(\omega) \cdots W_N(\omega)|^2 \tag{4}$$

is the total transfer function for random jitter, and

$$T_s(\omega) = |W_1(\omega) + W_1(\omega)W_2(\omega) + \cdots + W_1(\omega)W_2(\omega) \cdots W_N(\omega)|^2 \tag{5}$$

is the total transfer function for systematic jitter.

If we assume that

$$W_j(\omega) = W(\omega) \quad j = 1, 2, \cdots, N, \tag{6}$$

then we obtain the known relations[1]

$$T_r(\omega) = |W(\omega)|^2 \frac{1 - |W(\omega)|^{2N}}{1 - |W(\omega)|^2} \tag{7}$$

and

$$T_s(\omega) = |W(\omega)|^2 \left| \frac{1 - W^N(\omega)}{1 - W(\omega)} \right|^2. \tag{8}$$

As we will show, relations (7) and (8) provide us with a satisfactory approximation when the fine structure of $S(\omega)$ is not important, as in the evaluation of the RMS jitter,

$$\sigma_o^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) d\omega. \tag{9}$$

However, in some applications the detailed structure of $S(\omega)$ is essential, and the random variations of $W_i(\omega)$ must be taken into consideration. Hence it is important to describe $W_i(\omega)$ with a probabilistic model instead of the deterministic approach used in the past.

Another advantage of using this approach is that we can directly examine the impact of tolerances in manufacturing the repeater components on the accumulation of jitter.

2.2 The stochastic model

In this section we shall consider $W_i(\omega)$ to be a random variable, independent and identically distributed (i.i.d.). Thus $S(\omega)$, the spectrum of the jitter at the output of N regenerative repeaters, is also a random variable and it should be described with its expected value and variance.

With the assumption that $e_i(t)$ is a stationary stochastic process we obtain from (2) and (3)

$$E\{S(\omega)\} = \Phi_s(\omega) E\{T_s(\omega)\} + \Phi_r(\omega) E\{T_r(\omega)\} \tag{10a}$$

$$\begin{aligned} \mathrm{V\hat{a}r}^2\{S(\omega)\} &= \Phi_s^2(\omega) \mathrm{V\hat{a}r}^2\{T_s(\omega)\} + \Phi_r^2(\omega) \mathrm{V\hat{a}r}^2\{T_r(\omega)\} \\ &\quad + 2\Phi_s(\omega)\Phi_r(\omega)[E\{T_s(\omega)T_r(\omega)\} - E\{T_s(\omega)\}E\{T_r(\omega)\}], \end{aligned} \tag{10b}$$

where $E\{X\}$ and $\mathrm{V\hat{a}r}^2\{X\}$ are, respectively, the expected value and the variance of the random variable X. The spectra $\Phi_s(\omega)$ and $\Phi_r(\omega)$ of the input systematic and random jitter will be considered known, because we can measure them experimentally. Usually $\Phi_s(\omega)$ and $\Phi_r(\omega)$ are constant for the low frequencies where $T_s(\omega)$ and $T_r(\omega)$ are significant (white noise). Typical values for Φ_s and Φ_r are between 1 and 100 deg^2/MHz.

Thus for the evaluation of the $E\{S(\omega)\}$ and $\mathrm{Var}\{S(\omega)\}$ in (10) we need to find $E\{T_s(\omega)\}$, $E\{T_r(\omega)\}$, $\mathrm{Var}\{T_s(\omega)\}$, $\mathrm{Var}\{T_r(\omega)\}$ and $E\{T_s(\omega)T_r(\omega)\}$. The expressions for their analytical evaluation are

given below for a chain of N repeaters. Their derivations are given in Appendix A. For simplicity hereafter we shall omit the dependence of the various functions on the parameter ω, except if it is not obviously implied. Since the derivation of $\mathrm{Var}\{T_s\}$ and $E\{T_sT_r\}$ is complicated, we simplified our analysis by making use of the Central Limit Theorem and assuming that the two random variables X and Y defined as $X + jY = \sum_{k=1}^{N} W_1W_2 \cdots W_k$ are jointly normal. Hence relations (15) and (16) are not exact, because of the above approximation.

For a chain of N repeaters

$$\bar{T}_r = E\{T_r\} = B\,\frac{1 - B^N}{1 - B} \tag{11}$$

$$\bar{T}_s = E\{T_s\} = B\,\frac{1 - B^N}{1 - B} + 2\,\mathrm{Re}\,\frac{B}{1 - B/W}\left\{\frac{1 - W^N}{1 - W} - \frac{1 - B^N}{1 - B}\right\} \tag{12}$$

$$E\{T_r^2\} = C\,\frac{1 - C^N}{1 - C} + 2\,\frac{C}{1 - C/B}\left\{\frac{1 - B^N}{1 - B} - \frac{1 - C^N}{1 - C}\right\} \tag{13}$$

$$\hat{\mathrm{Var}}\{T_r\} = E\{T_r^2\} - E^2\{T_r\} \tag{14}$$

$$\hat{\mathrm{Var}}\{T_s\} \approx 2\left[E^2\{X^2\} + E^2\{Y^2\} + 2E^2\{XY\} - (E^2\{X\} + E^2\{Y\})^2\right] \tag{15}$$

$$E\{T_sT_r\} - E\{T_s\}E\{T_r\} \approx 2\left[E\{X\}E\{XT_r\} + E\{Y\}E\{YT_r\} - (E^2\{X\} + E^2\{Y\})E\{T_r\}\right], \tag{16}$$

where

$$W = E\{W_i\} \tag{17a}$$

$$Z = E\{W_i^2\} \tag{17b}$$

$$B = E\{|W_i|^2\} \tag{17c}$$

$$C = E\{|W_i|^4\} \tag{17d}$$

and

$$D = E\{|W_i|^2W_i\}. \tag{17e}$$

With X and Y defined as

$$X + jY = Q = \sum_{k=1}^{N} W_1W_2 \cdots W_K \tag{17f}$$

we have

$$E\{X\} = \mathrm{Re}\,E\{Q\},\ E\{Y\} = \mathrm{Im}\,E\{Q\} \tag{17g}$$

$$E\{X^2\} = [\bar{T}_s + \mathrm{Re}\,E\{Q^2\}]/2 \tag{17h}$$

$$E\{Y^2\} = [\bar{T} - \mathrm{Re}\,E\{Q^2\}]/2 \tag{17i}$$

$$E\{XY\} = \text{Im } E\{Q^2\}/2 \tag{17j}$$

$$E\{XT_r\} = \text{Re } E\{QT_r\} \tag{17k}$$

$$E\{YT_r\} = \text{Im } E\{QT_r\}, \tag{17l}$$

where

$$E\{Q\} = W\frac{1 - W^N}{1 - W} \tag{17m}$$

$$E\{Q^2\} = Z\frac{1 - Z^N}{1 - Z} + 2\frac{Z}{1 - Z/W}\left[\frac{1 - W^N}{1 - W} - \frac{1 - Z^N}{1 - Z}\right] \tag{17n}$$

$$E\{QT_r\} = D\frac{1 - D^N}{1 - D} + 2\frac{D}{1 - D/W}\left[\frac{1 - W^N}{1 - W} - \frac{1 - D^N}{1 - D}\right]. \tag{17o}$$

With the above formulas we can evaluate the $E\{S(\omega)\}$ and $\text{Var}\{S(\omega)\}$ in (10) in terms of the averages $W(\omega)$, $Z(\omega)$, $B(\omega)$, $C(\omega)$, and $D(\omega)$ of the individual jitter transfer functions $W_i(\omega)$. These quantities can be estimated either experimentally from a sufficient number of samples of $W_i(\omega)$, or numerically, using a Monte Carlo technique, from an appropriate model of $W_i(\omega)$. The second approach will be used in our examples. If $W_i(\omega) = W(\omega)$ $i = 1, 2, \cdots, N$ then $B = |W(\omega)|^2$, $W = W(\omega)$, and after some algebra we can show that relation (12) is equivalent to relation (8).

Another quantity, beyond $S(\omega)$, which is important in jitter accumulation, is σ_o, the RMS value of the output jitter. We shall also evaluate its expected value and its variance.

The expected value is obtained easily from (9) and (10a) as

$$E\{\sigma_o^2\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} E\{S(\omega)\}d\omega. \tag{18}$$

The evaluation of the variance of σ_o^2 is more difficult. This is done in Appendix A. However, the result is complicated because it involves the evaluation of $R(u, v) = E\{W_i(u)W_i(v)\}$, the autocorrelation of $W_i(\omega)$. To avoid the additional computations needed for the calculation of $R(u, v)$, we obtain an upper and lower bound for $\text{Var}\{\sigma_o^2\}$ by assuming that $W_i(\omega_1)$ and $W_i(\omega_2)$ are, respectively, highly correlated or uncorrelated. Then we can derive (see Appendix A) that

$$\frac{1}{2\pi}\left|\int_{-\infty}^{\infty} \text{Var}\{S(\omega)\}d\omega\right|^{1/2} < \text{Var}\{\sigma_o^2\} < \frac{1}{2\pi}\int_{-\infty}^{\infty} \text{Var}\{S(\omega)\}d\omega, \tag{19}$$

where $\text{Var}\{S(\omega)\}$ is obtained from (10b). The average value of the two bounds appears to be a good estimator of the true variance of the RMS. With the relations given above we can now evaluate $E\{S(\omega)\}$, $\text{Var}\{S(\omega)\}$, $E\{\sigma_o^2\}$, $\text{Var}\{\sigma_o^2\}$.

In deriving $E\{S(\omega)\}$ in (10a) and $\mathrm{Var}\{S(\omega)\}$ in (10b), we have made the additional assumption that $\Phi(\omega)$, the spectrum of the additive jitter $e_i(t)$, is the same for all the regenerators. This appears to contradict our assumption of different $W_i(\omega)$, since it is well known that $\Phi(\omega)$ is highly correlated with the transfer function of the regenerator. This problem can be resolved if we define $e_{is}(t)$, the systematic component of $e_i(t)$, as the part that is identical, within a constant, for all the repeaters, i.e., $e_{is}(t) = \alpha_i e_s(t)$, where α_i depends on $W_i(\omega)$. The random component $e_{ir}(t)$, the remaining part of $e_i(t)$, has a spectrum $\Phi_i(\omega)$ and is statistically independent for each repeater. Then in relation (5) we have to replace $W_i(\omega)$ by $\alpha_i W_i(\omega)$, while in relation (4) we must replace $|W_i(\omega)|^2$ by $\Phi_i(\omega)|W_i(\omega)|^2$. Using these substitutions, all the remaining expressions are still valid. The evaluation of α_i and $\Phi_i(\omega)$ in terms of the transfer function $W_i(\omega)$ is a difficult problem directly related to the problem of expressing the statistical properties of $e_i(t)$ in terms of the regenerator's transfer function. This question has been considered in Refs. 4 through 6.

III. APPLICATIONS

In the following three sections we shall present some numerical applications of the above theory. In Section 3.1 we shall examine chains employing phase-locked loops in their timing circuits.[4,6] In Section 3.2 the timing circuit employs a SAW filter.[7,8] Finally, in Section 3.3 we shall examine a special case, where $W_i(\omega)$ is assumed to exhibit a random variation only in its phase while its amplitude is identical for all repeaters. This case is of interest when we have a maximally flat filter and we want to examine the spectrum of the accumulated jitter for low frequencies.

All the numerical results are normalized by assuming $\Phi_s(\omega) = \Phi_r(\omega) = 1\ \mathrm{deg}^2/\mathrm{MHz}$.

3.1 Timing circuits with PLL

Phase-locked loops have been used extensively in the timing circuits of regenerative repeaters and their jitter performance has been examined thoroughly.[3,4,6] If the timing extractor employs a PLL, its jitter transfer function $W(\omega)$ is equal to the phase transfer function of the PLL, under the assumption that the output of the phase detector is small. This is equivalent to the assumption of having small alignment jitter. (*Alignment jitter* refers to the deviations in alignment between the clock embedded in the incoming data stream of the regenerator and the timing clock derived from the data stream by the timing circuit of the regenerator.) The above requirement is usually satisfied, validating the linear model of the PLL. The phase transfer function of a PLL used in the literature is typically described with a second-order

model. If we want to consider parasitic elements in the PLL, then, as M. W. Hall suggested,[3] a fifth-order model is more realistic, especially for high frequencies. However, the spectrum of the accumulated jitter is significant only in the low frequencies. Hence the second-order model provides an adequate description of the PLL in this frequency range, because it coincides with the fifth-order model in this range of frequencies.

Thus the jitter transfer function $W_i(\omega)$ of the ith repeater can be modeled as

$$W_i(s/j) = \frac{2\zeta_i\omega_{ni}s + \omega_{ni}^2}{s^2 + 2\zeta_i\omega_{ni}s + \omega_{ni}^2} \quad s = j\omega, \tag{20}$$

where ζ_i is the damping coefficient of the PLL and ω_{ni} its natural frequency.

We shall assume that the average values of ζ_i and ω_{ni} are $\bar{\zeta}_i = 6$ and $\bar{\omega}_{ni} = 4.5$ kHz. Also, the bandwidth of the PLL is assumed to vary from 20 kHz up to 80 kHz. To simulate the above conditions we assumed that ζ_i is uniformly distributed from 4 to 8 and ω_{ni} is also uniformly distributed from 2 kHz to 7 kHz. Using the above numbers, we calculated $W(\omega)$, $Z(\omega)$, $B(\omega)$, $C(\omega)$, $D(\omega)$ (see 17a, b, c, d, e) by numerically averaging 1000 samples of $W_i(\omega)$. In Fig. 2 we show the amplitude and phase of $W(\omega)$ as well as two samples of $W_i(\omega)$ with $\zeta_i = 6$ and $\omega_{ni} = 2$ kHz and 7 kHz. The shaded area shows the permissible range of $W_i(\omega)$. Bandwidth varies between 16 kHz and 110 kHz, while phase slope varies between −0.6 deg/kHz and −3.2 deg/kHz. In Fig. 3a we plot $E\{T_s(\omega)\}$, and $E\{T_s(\omega)\} \pm \text{Var}\{T_s(\omega)\}$ for a chain of 50 repeaters. Also in the same graph we show the $T_s(\omega)$ evaluated by using relation (8), i.e., assuming that all repeaters have identical PLLs with $W(\omega) = E\{W_i(\omega)\}$. In Fig. 3b we show the same results for 200 repeaters. Figure 4 shows the same curves for the random jitter component.

To check our theoretical results, we did a complete numerical Monte-Carlo simulation of 200 chains and computed the various parameters, including $E\{\sigma_o^2\}$ and $\text{Var}\{\sigma_o^2\}$. The resulting graphs were indistinguishable from those shown in Figs. 3 and 4; the numerical values obtained for $E\{\sigma_o^2\}$ and $\text{Var}\{\sigma_o^2\}$ are shown in Table I. All the numbers agree with the results of our theoretical analysis. Also, the expected value of the RMS of the accumulated jitter, systematic and random, is shown in Fig. 5 with the $\pm 3\ \text{Var}\{\sigma_o^2\}$ curves (99.7 percent confidence). Also in the same graph we plot the RMS values for the accumulated jitter for a chain having identical repeaters with jitter transfer function equal to the average value $E\{W_i(\omega)\}$.

We can draw the following conclusions from the previous example: (1) The model of identical repeaters having as jitter transfer function

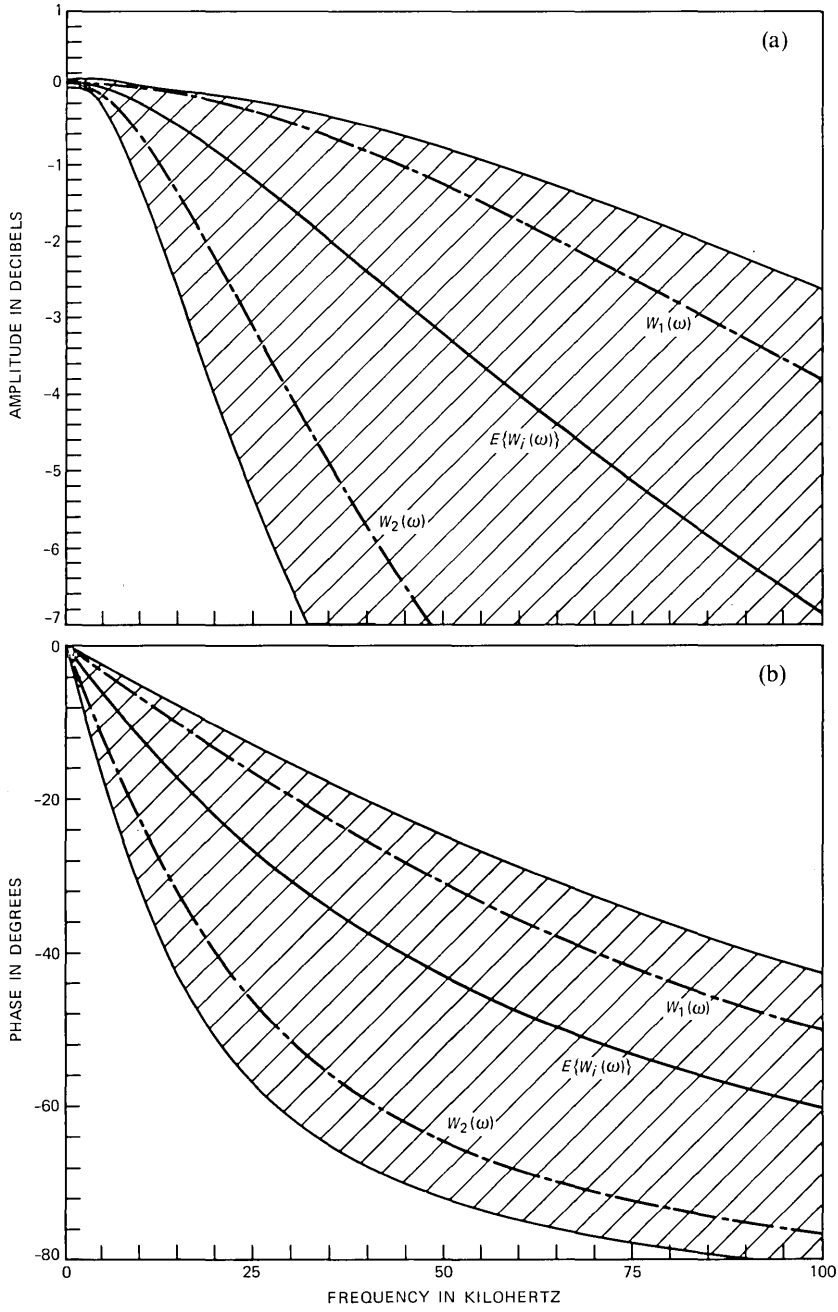


Fig. 2—(a) Amplitude and (b) phase of $W_1(\omega)$, $W_2(\omega)$, $E\{W_i(\omega)\}$ for the PLL model. $W_1(\omega)$ is obtained with $\zeta = 6$ and $\omega_{ni} = 7$ kHz. $W_2(\omega)$ is obtained with $\zeta = 6$ and $\omega_{ni} = 2$ kHz. Shaded area indicates the permissible range of $W_i(\omega)$.

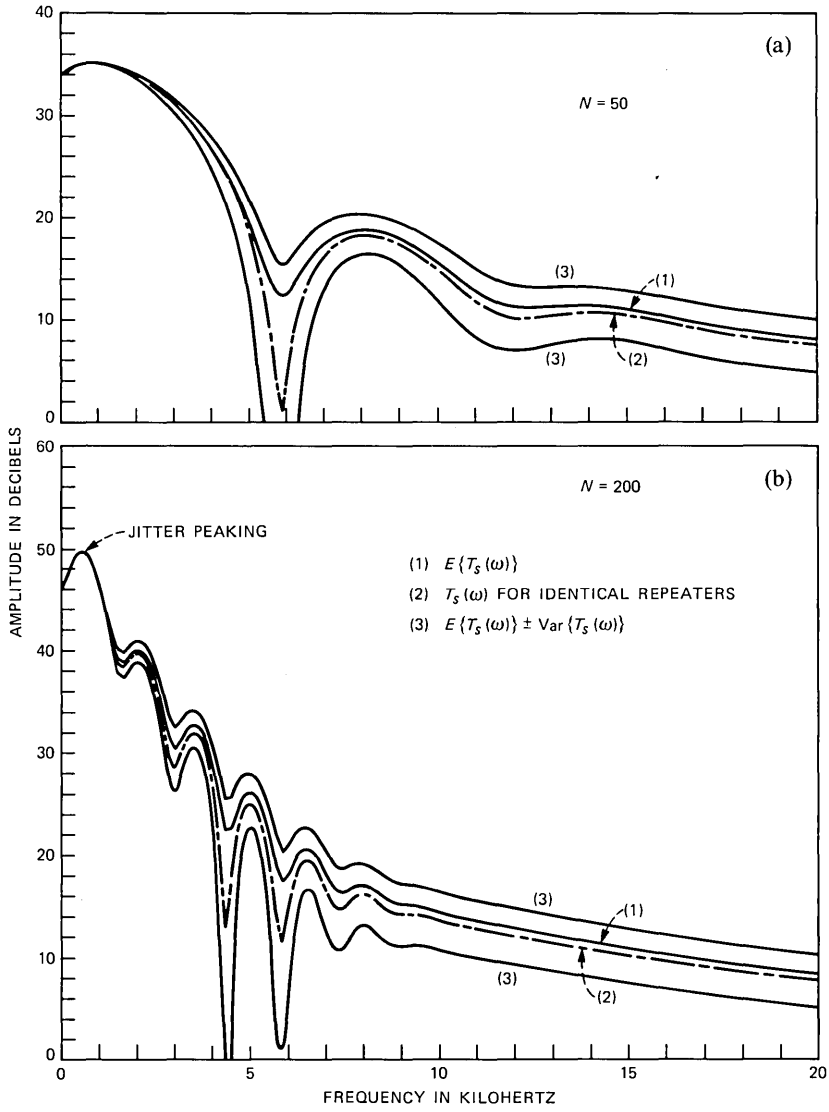


Fig. 3—Total transfer function $T_s(\omega)$ for systematic jitter for (a) 50 repeaters and (b) 200 repeaters with PLL using the stochastic model and the model with identical repeaters.

the average jitter transfer function underestimates slightly the average RMS value and the power spectrum density. (2) The variance of the RMS jitter was only 5 percent (see Table I and Fig. 5) even if the bandwidth of the PLL jitter transfer function varied from 16 kHz up to 110 kHz. This implies that the model of identical repeaters provides

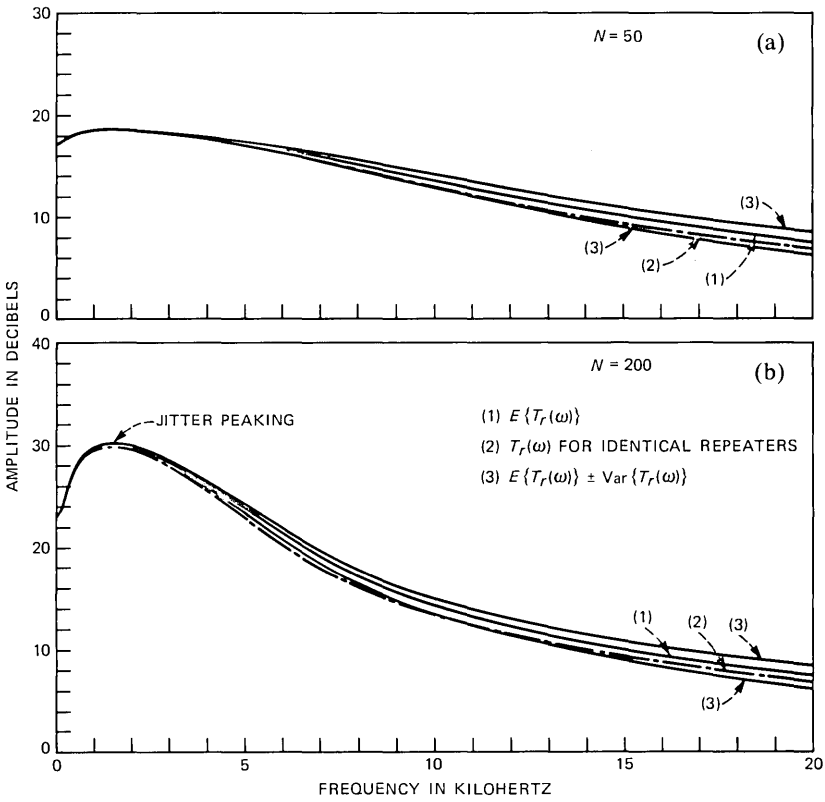


Fig. 4—Total transfer function $T_r(\omega)$ for random jitter for (a) 50 repeaters and (b) 200 repeaters with PLL using the stochastic model and the model with identical repeaters.

us with a reliable estimation of the RMS jitter. (3) The total transfer function $T_r(\omega)$ has a smaller variance than $T_s(\omega)$.

3.2 Timing circuits with SAW filters

The introduction of optical-fiber transmission systems has made possible data transmission rates of several hundred megabits per second. This introduced significant changes in the construction of the timing extracting circuits in the regenerative repeaters. The popular PLLs had to be replaced, because their implementation above 100 Mb/s has been difficult, especially in integrated circuit form. Currently, SAW filters[7] have emerged as their replacements. This actually represents a return to passive filtering after many years of using the PLL.

In this section we analyze a system where the tuned filter in the timing circuit is a passive SAW filter. In the present analysis we will

Table I—Theoretical and numerical values for $E\{\sigma_o^2\}$ and $Var\{\sigma_o^2\}$ for chains with 50, 100, 200, and 300 repeaters

	$E\{\sigma_o^2\}$ deg^2 *						$Var\{\sigma_o^2\}$					
	Using Random Model		Using Identical Repeaters		Numerical Result (200 chains)		Lower Bound		Upper Bound		Numerical Result (200 chains)	
N	Syst.	Ran.	Syst.	Ran.	Syst.	Ran.	Syst.	Ran.	Syst.	Ran.	Syst.	Ran.
50 PLL	18.57	1.44	18.20	1.36	18.61	1.45	0.18	0.01	1.85	0.18	1.06	0.15
SAW	28.94	8.45	28.82	8.00			0.08	0.05	0.93	0.90		
100 PLL	51.49	2.70	50.57	2.50	51.57	2.72	0.56	0.01	4.50	0.26	2.27	0.18
SAW	58.17	13.94	57.98	13.01			0.23	0.08	1.53	1.47		
200 PLL	188.85	7.73	184.55	7.05	189.16	7.77	2.68	0.05	17.82	0.60	8.61	0.36
SAW	116.67	22.85	116.38	20.97			1.02	0.11	3.27	2.18		
300 PLL	540.64	22.41	522.45	20.12	550.53	22.68	10.12	0.21	62.89	1.80	32.93	1.25
SAW	175.19	30.43	174.83	27.58			2.02	0.15	4.87	2.80		

* The input jitter is $\Phi_s(\omega) = \Phi_r(\omega) = 1$ deg^2/MHz.

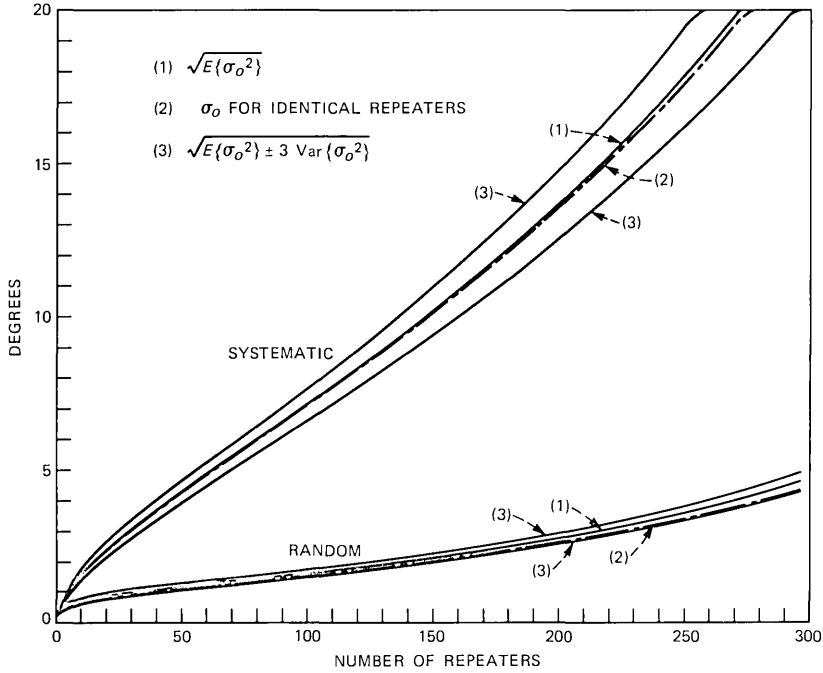


Fig. 5—RMS values of systematic and random jitter for repeaters with PLLs. Curves show: (1) stochastic model; (2) conventional model with identical repeaters; and (3) 99.7-percent confidence interval.

not consider the effects introduced by the prefilter and nonlinear device, which often precede the tuned circuit. It can be shown that this simplification is not restrictive.[2] A simple proof is also presented in Appendix B.

Let $H(\omega) = A(\omega)e^{j\phi(\omega)}$ be the transfer function of the SAW filter. Then, as has been shown,[2] the jitter transfer function of the regenerator is approximately given by

$$W(\omega) = \frac{H(\omega - \omega_0)e^{j\phi(\omega_0)} + H(\omega + \omega_0)e^{-j\phi(\omega_0)}}{2A(\omega_0)}, \quad \text{for} \quad |\omega| < \omega_0, \qquad (21)$$

where ω_0 is the baud rate of the data. Relation (21) is valid under the assumption that the accumulated jitter does not have large components at high frequencies. The exact conditions for the validity of (21) will be studied in a forthcoming work. A simple derivation of (21) is given also in Appendix B.

Let us now define the normalized low-pass equivalent of $H(\omega)$ as

$$H_L(\omega) = \frac{H(\omega + \omega_0)e^{-j\phi(\omega_0)}}{A(\omega_0)} \quad \text{for} \quad \omega > -\omega_0. \qquad (22)$$

Thus $W(\omega)$ is now given by

$$W(\omega) = \frac{[H_L(\omega) + H_L^*(-\omega)]}{2}, \tag{23}$$

where $*$ denotes complex conjugate.

A model for transversal SAW filters related to its design electrical characteristics, i.e., number of fingers, distance of transducers, terminating impedances, etc, is given in Ref. 8. Since in the present simulation we are interested only in its transfer function, we will use a simpler representation, and model $H_L(\omega)$, the low-pass equivalent of the SAW filter, as a two-pole filter. Its transfer function is

$$H_L(s) = \frac{s_1 s_2}{(s - s_1)(s - s_2)} e^{-cs}, \tag{24}$$

where

$$s_i = -a_i + jb_i = -(1 \pm \epsilon)a + jb_i \qquad i = 1, 2$$

are its two poles and c is a linear phase slope used to correct the phase of the model, since SAW filters are not minimum phase filters. For jitter studies our model provides an adequate description for SAW filters.

Denoting by BW the bandwidth of the SAW filter, ω_c its center frequency, and ω_0 the baud rate of the data, the various SAW filter parameters can be defined in terms of α, b_i, and the asymmetry factor ϵ as follows:

$$\begin{aligned}
\omega_c - \omega_0 &= (b_1 + b_2)/2 && \text{mistuning} \qquad (25)\\
b &= (b_1 - b_2)/2 && a = (a_1 + a_2)/2\\
\alpha &= 2(\omega_c - \omega_0)/BW && \text{detuning factor}\\
\omega_n^2 &= a^2 + b^2 && \text{natural frequency}\\
\zeta &= a/\omega_n && \text{damping factor}\\
Q &= \omega_c/BW && \text{filter's quality factor.} \qquad (26)
\end{aligned}$$

The natural frequency ω_n can be determined approximately from the BW (for $\epsilon \ll 1$) by

$$\omega_n^2 = \left(\frac{BW}{2}\right)^2 \frac{1}{-(1 - 2\zeta^2) + ((1 - 2\zeta^2)^2 + 1)^{1/2}}. \tag{27}$$

For the simulation we used the following numerical values for the range of bandwidth BW, static offset, damping factor ζ, and asymmetry factor ϵ:

$$160 \text{ kHz} < \frac{BW}{2} < 240 \text{ kHz}$$

$$|f_c - f_0| < 50 \text{ kHz}$$

$$0.60 < \zeta < 0.80$$

$$-0.1 < \epsilon < 0.1$$

$$c = -0.2 \text{ deg/kHz}. \tag{28}$$

We assumed the above parameters to be uniformly distributed between their upper and lower limits. If the baud rate of the data is 300 MHz, then the above values correspond to filters having Q's between 625 and 940. To illustrate the relation of the above parameters to $H_L(\omega)$ and $W(\omega)$, we plot in Fig. 6 the passband of the SAW filter and the corresponding jitter transfer functions for $\zeta = 0.65$ and various combinations of ϵ and α. Notice that the frequency scale has been normalized to $BW/2$, the bandwidth of $H_L(\omega)$.

In Fig. 6a we plot the passband of the SAW filters for $\zeta = 0.65$ and $\epsilon = 0$, 0.1. In Fig. 6b we plot the corresponding jitter transfer function

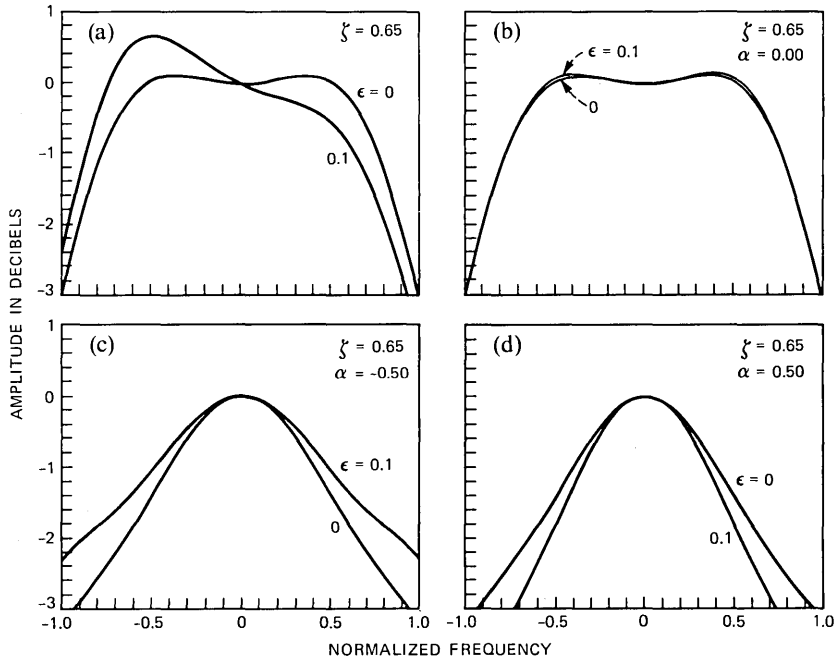


Fig. 6—(a) Underdamped SAW filter with $\zeta = 0.65$ and asymmetry factor $\epsilon = 0$ and 0.1. Jitter transfer functions obtained with various detuning factors $\alpha = 2(\omega_c - \omega_o)/BW$ for (b) $\alpha = 0$ (0.105 dB jitter peaking), (c) $\alpha = -0.50$, and (d) $\alpha = 0.50$.

$W(\omega)$ when $f_c = f_0(\alpha = 0)$. For this case the asymmetry is almost canceled. This is expected because from relation (23) we can see that asymmetries in $H_L(\omega)$ that are odd with respect to f_c will be canceled when we form $W(\omega)$. In Fig. 6c we plot the corresponding jitter transfer function when the detuning factor is $\alpha = -0.5$, i.e., the baud rate is 100 kHz less than the center frequency of the filter. Figure 6d shows the jitter transfer function for $\alpha = 0.5$. In Fig. 7 we plot $E\{W_i(\omega)\}$ (1000 samples) as well as two extreme samples of $W_i(\omega)$. Most of the $W_i(\omega)$ will be between $W_1(\omega)$ and $W_2(\omega)$. The shaded area shows the permissible range of $W_i(\omega)$. In Fig. 8 we plot the average total transfer function for systematic jitter, $E\{T_s(\omega)\}$, and $E\{T_s(\omega)\} \pm \mathrm{Var}\{T_s(\omega)\}$ for 50 and 200 repeaters. The variance of $T_s(\omega)$ in this example is much less than the variance of the example with PLL. This is due to the narrow distribution of the SAW filter phase slopes (−0.5 deg/kHz to −0.8 deg/kHz). Finally, in Fig. 9 we plot the expected value of the RMS of the accumulated jitter, systematic and random, with the $\pm 3\ \mathrm{Var}\{\sigma_0^2\}$ curves (99.7 percent confidence). Some numerical values for the accumulated RMS jitter and its variance are shown in Table I for $N = 50$, 100, 200, and 300. To find the true RMS value we have to multiply the numbers shown in Table I with the $\Phi_s(\omega)$ and $\Phi_r(\omega)$ measured in deg^2/MHz. For example, if $N = 200$ and a PLL is used, then assuming $\Phi_s = 20$ deg^2/MHz and $\Phi_r = 5$ deg^2/MHz, we obtain

$$\sigma_o = [20 \cdot 189 + 5 \cdot 7.73]^{1/2} = 61.8 \text{ degrees.} \tag{29}$$

Using the same numbers for the SAW filters we obtain

$$\sigma_o = [20 \cdot 117 + 5 \cdot 22.9]^{1/2} = 49.5 \text{ degrees.} \tag{30}$$

Thus the simulated SAW filters accumulate less jitter, even with a bandwidth larger than the bandwidth of PLLs. This is due to the fact that PLLs have an inherent jitter peaking and because their phase is smaller than the SAW filters [see eq. (40)]. To obtain the corresponding peak-to-peak values we usually multiply the RMS value with a peak-to-peak/RMS factor. Typical values for this factor are between 8 and 15.

From the above example it becomes clear that, using our random model, we can tolerate larger manufacturing variances, because we can accept retiming circuits exhibiting substantial jitter peaking. For example, $W_1(\omega)$ (0.4 dB jitter peaking) in Fig. 7 can be accepted if the expected average jitter transfer function $W(\omega)$ of the manufactured SAW filters possess a moderate jitter peaking (i.e., less than 0.1 dB).

3.3 Timing circuits with random phase

In this section we will consider the dependence of the jitter power spectrum on phase variations. This case is of interest when we want

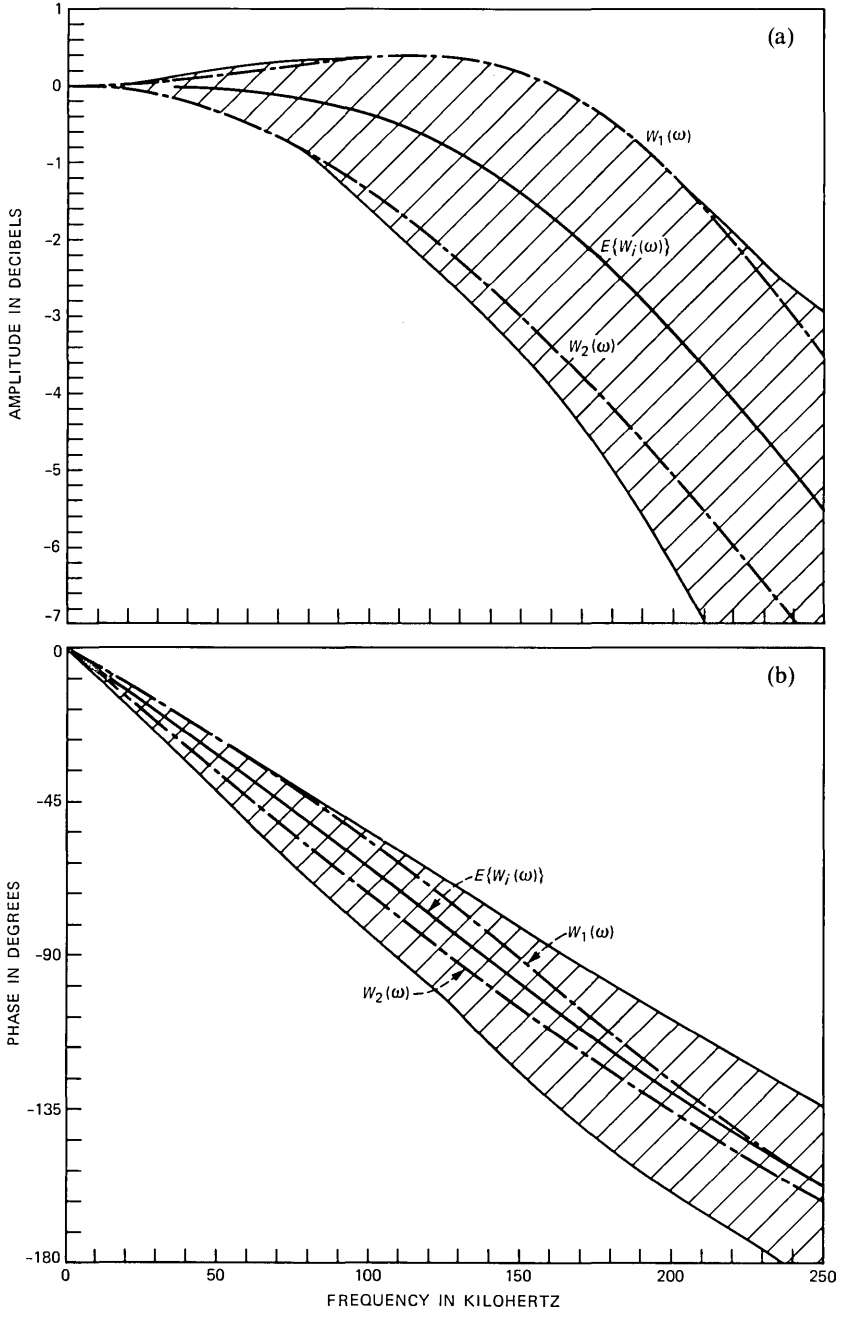


Fig. 7—(a) Amplitude and (b) phase of $E\{W_i(\omega)\}$, $W_1(\omega)$, $W_2(\omega)$ for the SAW model. $W_1(\omega)$ is obtained with BW/2 = 240 kHz, $\alpha = 0$, $\zeta = 0.60$, and $\epsilon = 0.1$ (0.4 dB jitter peaking). $W_2(\omega)$ is obtained with BW/2 = 160 kHz, $\alpha = 0.25$, $\zeta = 0.60$, and $\epsilon = 0.1$. Area with lines indicates the permissible range of $W_i(\omega)$.

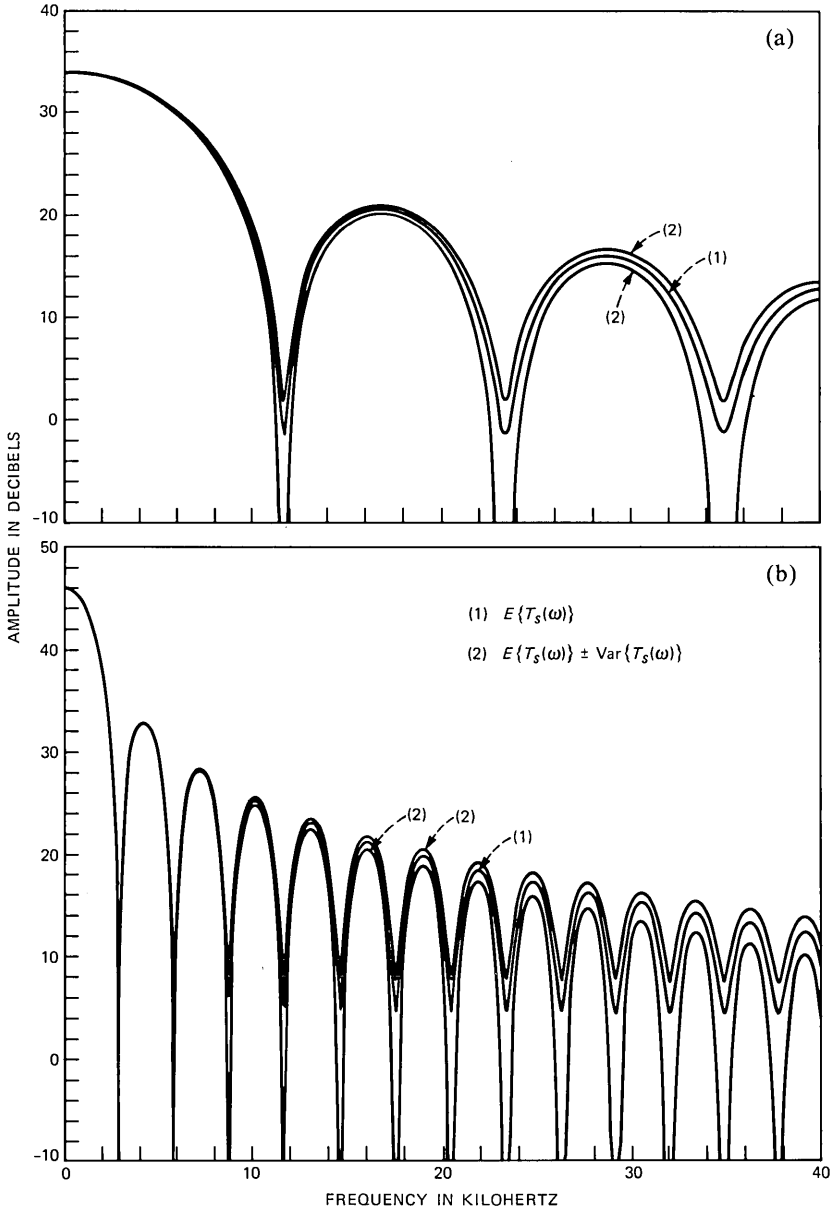


Fig. 8—Total systematic jitter transfer function $T_s(\omega)$ and its variance for 50 and 200 repeaters.

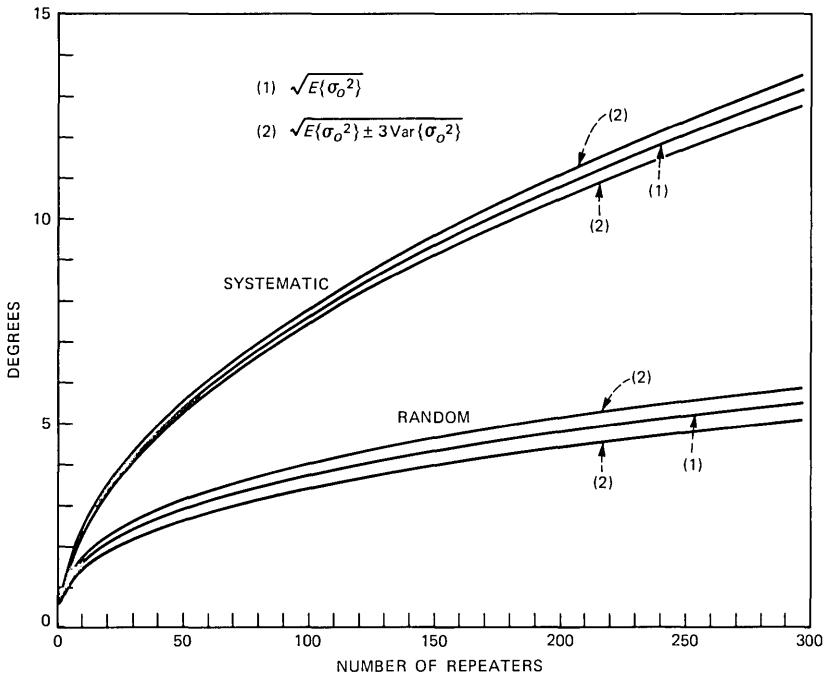


Fig. 9—RMS values of systematic and random jitter for repeaters with SAW filters.

to estimate the jitter spectral density for low frequencies.[9] We shall assume that

$$W_i(\omega) = A(\omega)e^{j\{\phi(\omega)+\eta_i(\omega)\}}, \tag{31}$$

where $\eta_i(\omega)$ are random variables with zero mean and independent and identically distributed (i.i.d.). Let $\Phi_\eta(X)$ be the characteristic density function of η_i, i.e.,

$$\Phi_\eta(X, \omega) = E\{e^{j\eta_i(\omega)X}\} = \int_{-\infty}^{\infty} f_\eta(X, \omega)e^{j\eta X}d\eta. \tag{32}$$

Then [see relations (17)] all the needed statistics for $W_i(\omega)$ can be obtained analytically if $\Phi_\eta(X, \omega)$ is known, i.e.,

$$\begin{aligned} W(\omega) &= E\{W_i(\omega)\} = A(\omega)e^{j\phi(\omega)}\Phi_\eta(1, \omega) \\ Z(\omega) &= E\{W_i^2(\omega)\} = A^2(\omega)e^{j2\phi(\omega)}\Phi_\eta(2, \omega) \\ D(\omega) &= E\{\,|\,W_i(\omega)\,|^2\,W_i(\omega)\} = A^3(\omega)e^{j\phi(\omega)}\Phi_\eta(1, \omega). \end{aligned} \tag{33}$$

If $\eta_i(\omega)$ is assumed to have zero mean and to be uniformly distributed between $-\Psi(\omega)$ and $\Psi(\omega)$, then

$$\Phi_\eta(X, \omega) = \frac{\sin \Psi(\omega)X}{\Psi(\omega)X}. \tag{34}$$

If $\eta_i(\omega)$ is assumed to be zero mean and Gaussian distributed with variance $\sigma(\omega)$, then

$$\Phi_\eta(X, \omega) = e^{-X^2\sigma^2(\omega)/2}. \tag{35}$$

Using (34), (35), (33), and (10) we can obtain $E\{T_s(\omega)\}$ and $\mathrm{Var}\{T_s(\omega)\}$. In Fig. 10 we plot $E\{T_s(\omega)\}$ for a transversal filter with maximally flat

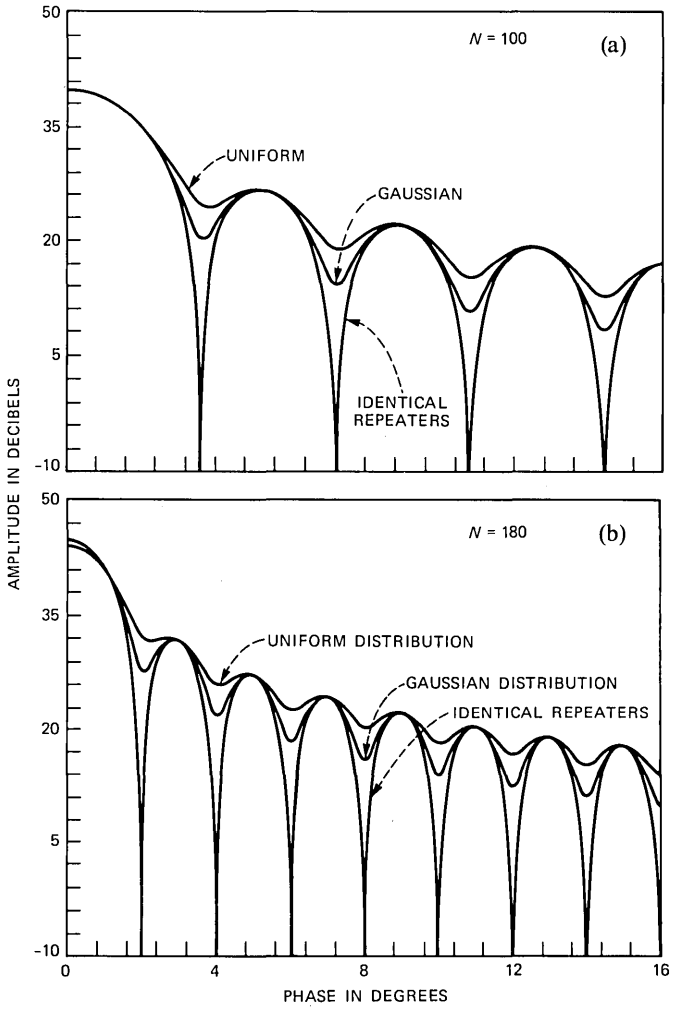


Fig. 10—Systematic jitter power density versus phase for identical repeaters, repeaters with uniform phase distribution, and repeaters with Gaussian phase distribution for (a) $N = 100$ repeaters and (b) $N = 180$ repeaters.

transfer function ($\zeta = \sqrt{2}/2$) as a function of its phase $\phi(\omega)$ for 100 and 180 regenerators for the cases $\eta_i = 0$, η_i uniformly distributed with $\Psi_i(\omega) = 8.1°$ and η_i gaussian distributed with $\sigma_i(\omega) = 8.1°/3$.

In Fig. 11 we plot $E\{T_s\}$ and its variance for the case of uniform distribution.

Finally, we would like to note the following approximate relations.

1. Large variance approximation—If the variance of the phase is large for $\omega = \omega_0$, then $W(\omega_0) \ll 1$ because $\Phi_\eta(1, \omega_0)$ is small [see relations (33) and (34)], and from relation (12) we obtain

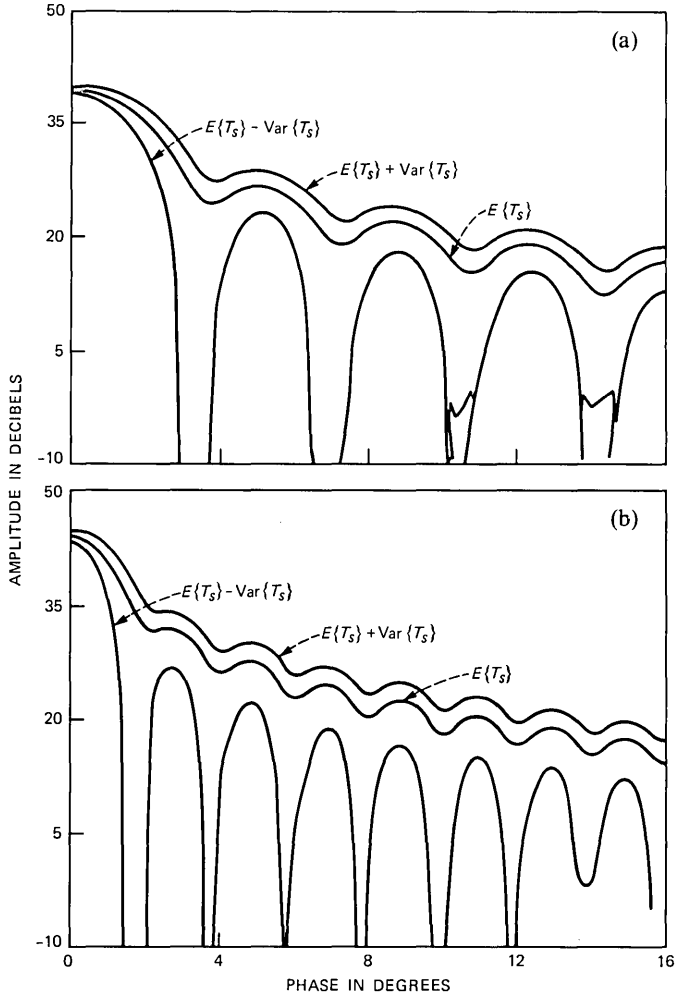


Fig. 11—Mean and variance of systematic jitter power density versus phase for repeaters with uniform phase distribution for (a) $N = 100$ repeaters and (b) $N = 180$ repeaters.

$$\lim_{\sigma(\omega_0)\to\infty} E\{T_s(\omega_0)\} = E\{T_r(\omega_0)\}. \tag{36}$$

This is expected because the randomness of the phase removes the coherent accumulation of the systematic jitter for $\omega = \omega_0$.

2. Low frequencies and long chain approximation—For large N the jitter energy is concentrated near $\omega = 0$. For this region, we can assume that $A(\omega) \simeq 1$ and approximate relation (12) with

$$E\{T_s\} \approx \text{Re}\left\{\frac{2[1 - W^N]W - N(1 - W^2)}{(1 - W)^2}\right\} \quad \omega \approx 0. \tag{37}$$

3. High frequencies and long chain—For high frequencies where $A(\omega) \ll 1$ and large N) we obtain that $E\{T_s\} \approx E\{T_r\} = E\{|W_i|^2\}$.

4. Small-phase variance approximation—If we assume that the variance $\sigma(\omega)$ of the phase $\phi(\omega)$ is small, after some algebra we can obtain the following relation:

$$E\{T_s\} \approx A_N + \sigma^2(\omega) \sum_{n=1}^{N} A_n, \tag{38}$$

where

$$A_n(\omega) = |A(\omega)|^2 \left|\frac{1 - A^n(\omega)e^{j\phi(\omega)}}{1 - A(\omega)e^{j\phi(\omega)}}\right|^2$$

and

$$\sigma^2(\omega) \ll \frac{1}{N}.$$

For low frequencies, $\omega \approx 0$, we can assume $A(\omega) \approx 1$ and the term $A_n(\omega)$ in relation (38) can be approximated with

$$A_n(\omega) = \left|\frac{\sin\dfrac{n\phi(\omega)}{2}}{\sin\dfrac{\phi(\omega)}{2}}\right|^2. \tag{39}$$

5. Linear phase, low frequencies, small variance approximation—Let us assume $\phi(\omega) \approx -\alpha\omega$ for $\omega \approx 0$. Then $\sigma^2(\omega) = S^2\omega$, where S is the variance of α, and relation (39) becomes

$$A_n(\omega) = \left|\frac{\sin\dfrac{n\bar{\alpha}\omega}{2}}{\sin\dfrac{\bar{\alpha}\omega}{2}}\right|^2.$$

For large N and assuming no jitter peaking and small-phase variance

we can approximate the variance σ_o^2 of the systematic jitter at the output of N regenerators with

$$E\{\sigma_o^2\} = \Phi_s(0)\left[\frac{0.36}{\bar{\alpha}}N + \frac{S^2}{3}\left(\frac{0.36}{\bar{\alpha}}\right)^3 N^2\right] \text{deg}^2, \qquad (40)$$

where

$\Phi_s(0)$ is the spectral density in deg^2/MHz
$\bar{\alpha}$ is the average phase slope in deg/kHz
S^2 is the variance of α
N is the number of regenerators.

Relation (40) shows that for long chains of regenerators exhibiting no jitter peaking the RMS value of the systematic accumulated jitter is determined mainly by the dc phase slope of the jitter transfer function and not by its shape. This is illustrated in Fig. 12, where we evaluate the RMS value of the systematic accumulated jitter for chain, consisting of identical regenerators having the jitter transfer function the $W(\omega)$ shown in Fig. 7. A curve using numerical integration is compared versus the curve predicted by the simple formula $\sigma_0 = 0.6\sqrt{N/\alpha}$ (α = 0.63 deg/kHz). For a first-order filter with band-

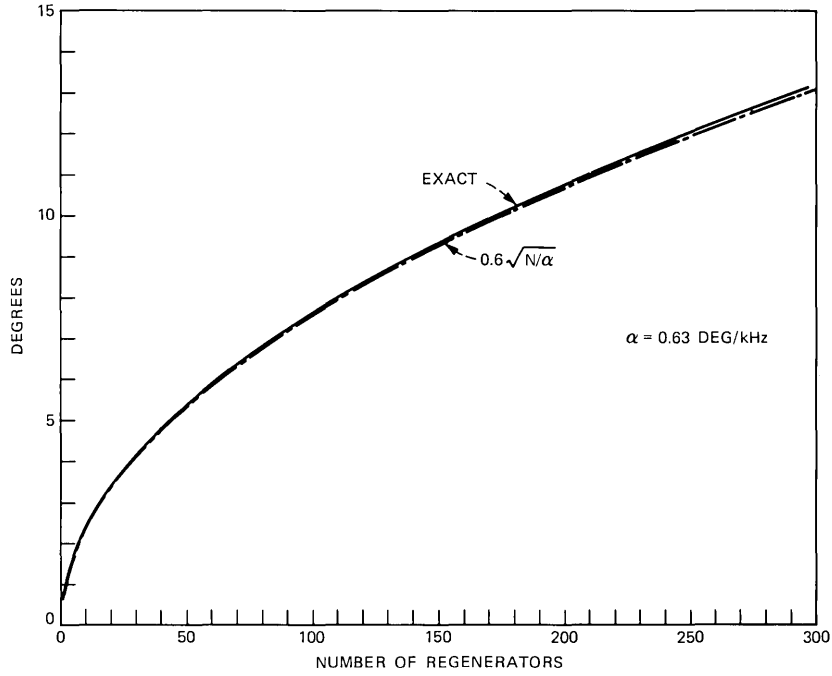


Fig. 12—RMS systematic jitter accumulation using numerical integration and relation (40).

width B, we have $B = 0.36/\alpha$ and the above relation becomes $\sigma_o = \sqrt{BN}$. This is the widely used relation (69) given in Ref. 1, which is valid only for first-order filters.

IV. CONCLUSION

We have presented a generalized model for the accumulation of jitter. This model differs from those used previously in that we do not assume all the repeaters have the same transfer functions. We have derived analytical expressions for the variance and the mean of the accumulated jitter in terms of the repeater's jitter transfer function $W_i(\omega)$.

We have also presented some numerical results for PLL and SAW filters. From our numerical simulations we found that for long chains the variance of the RMS jitter is about 5 percent. This implies that we can still reliably estimate RMS jitter by assuming that the jitter transfer function of all the repeaters is equal to the average jitter transfer function. Another result obtained from our modeling is that retiming circuits exhibiting large jitter peaking are acceptable if their average jitter transfer function does not have jitter peaking.

V. ACKNOWLEDGMENTS

I would like to thank R. L. Rosenberg for his help on modeling the SAW filters and for his comments on various aspects of this work. Also I would like to express my appreciation to P. K. Runge for his continuous encouragement.

REFERENCES

1. C. J. Byrne, B. J. Karafin, and D. B. Robinson, Jr. "Systematic Jitter in a Chain of Digital Regenerators," B.S.T.J., *42* (November 1963), pp. 2679–714.
2. U. Mengali and G. Pirani, "Jitter Accumulation in PAM Systems," IEEE Trans. Commun., *COM-28*, No. 8 (August 1980), pp. 1172–83.
3. M. W. Hall, private communication.
4. E. Rosa, "Analysis of Phase-Locked Timing Extraction Circuits for Pulse Code Transmission," IEEE Trans. Commun., *COM-22* (September 1974), pp. 1236–49.
5. L. E. Franks and J. P. Bubrouski, "Statistical Properties of Timing Jitter in a PAM Timing Recovery Scheme," IEEE Trans. Commun., *COM-22* (July 1974), pp. 913–20.
6. D. L. Duttweiler, "The Jitter Performance of Phase-Locked Loops Extracting Timing From Baseband Data Waveforms," B.S.T.J., *55* (January 1976), pp. 37–58.
7. R. L. Rosenberg, C. Chamzas, and D. A. Fishman, "Timing Recovery With SAW Transversal Filters in the Regeneration of Undersea Long-Haul Fiber Transmission Systems," to be published in publication IEEE J. Selected Areas Commun. (Joint Issue with IEEE/OSA Journal of Lightwave Technology), special issue on Undersea Lightwave Communications, December 1984.
8. D. A. Fishman, R. L. Rosenberg, and C. Chamzas, unpublished work.
9. C. D. Anderson and D. L. Keller, "SL Supervisory Analysis of Jitter-Peaking Effects in Digital Long-Haul Transmission Systems Using SAW Filter Retiming," to be published in IEEE J. Selected Areas Commun. (Joint Issue with IEEE/OSA Journal of Lightwave Technology), special Issue on Undersea Lightwave Communications, December 1984.
10. A. Papoulis, *Signal Analysis*, New York: McGraw-Hill, 1977.

APPENDIX A

Stochastic Evaluation of Jitter Parameters

For simplicity the dependence on ω is omitted in most of the following formulas.

A.1 Evaluation of $E\{T_r(\omega)\}$

The transfer function $T_r(\omega)$ for the random jitter is given by [see (4)]

$$T_r = |W_1|^2 + |W_1 W_2|^2 + \cdots + |W_1 W_2 \cdots W_N|^2. \tag{41}$$

With the assumption that the W_i are i.i.d., we obtain the expected value of T_r as

$$E\{T_r\} = B + B^2 + B^3 + \cdots + B^N = B\,\frac{1 - B^N}{1 - B}, \tag{42}$$

where

$$B = E\{|W_i|^2\}$$

and this is the desired relation (11).

A.2 Evaluation of $E\{T_s(\omega)\}$

The transfer function $T_s(\omega)$ for the systematic jitter is given [see (5)] by

$$T_s = \left| \sum_{k=1}^{N} W_1 W_2 \cdots W_k \right|^2. \tag{43}$$

The expected value of T_s may be written

$$\begin{aligned} E\{T_s\} &= E\left\{ \sum_{k=1}^{N} W_1 W_2 \cdots W_k \sum_{m=1}^{N} W_1^* W_2^* \cdots W_m^* \right\} \\ &= E\left\{ \sum_{k=1}^{N} |W_1 W_2 \cdots W_k|^2 \right. \\ &\qquad \left. + 2\,\mathrm{Re} \sum_{k=1}^{N} \sum_{m=1}^{k-1} |W_1 W_2 \cdots W_m|^2 W_{m+1} W_{m+2} \cdots W_k \right\}. \end{aligned}$$

Defining

$$W = E\{W_i\}, \qquad B = E\{|W_i|^2\} \tag{44}$$

and using that W_i are i.i.d. we obtain

$$\begin{aligned} E\{T_s\} &= \sum_{k=1}^{N} B^k + 2\,\mathrm{Re} \sum_{k=1}^{N} \sum_{m=1}^{k-1} B^m W^{k-m} \\ &= B\,\frac{1 - B^N}{1 - B} + 2\,\mathrm{Re} \sum_{k=1}^{N} B W^{k-1}\,\frac{1 - (B/W)^{k-1}}{1 - (B/W)} \end{aligned}$$

or

$$E\{T_s\} = B\frac{1-B^N}{1-B} + 2\,\mathrm{Re}\,\frac{B}{1-B/W}\left\{\frac{1-W^N}{1-W} - \frac{1-B^N}{1-B}\right\}, \quad (45)$$

which is the desired relation (12).

For large ω where $|W_i(\omega)| \ll 1$ we can obtain from (42) and (45) that

$$E\{T_s\} \approx E\{T_r\} \approx B(\omega). \quad (46)$$

A.3 Evaluation of Var{T_r}

To evaluate the variance of $T_s(\omega)$, the transfer function of the random jitter, we need only to find $E\{T_r^2\}$, since

$$\mathrm{Var}\{T_r\} = (E\{T_r^2\} - E^2\{T_r\})$$

and $E\{T_r\}$ has been evaluated in (42):

$$E\{T_r^2\} = E\left\{\sum_{k=1}^{N} |W_1W_2\cdots W_k|^2 \sum_{m=1}^{N} |W_1W_2\cdots W_m|^2\right\}. \quad (47)$$

Defining $B(\omega)$ as in (44) and

$$C(\omega) = E\{|W_i(\omega)|^4\}, \quad (48)$$

we obtain

$$\begin{aligned} E\{T_r^2\} &= E\left\{\sum_{k=1}^{N} |W_1W_2\cdots W_k|^4 \right. \\ &\qquad \left. + 2\sum_{k=1}^{N}\sum_{m=1}^{N-1} |W_1W_2W_m|^4 |W_{m+1}W_{m+2}\cdots W_k|^2\right\} \\ &= \sum_{k=1}^{N} C^k + 2\sum_{k=1}^{N}\sum_{m=1}^{N-1} C^m B^{k-m} \end{aligned}$$

or

$$E\{T_r^2\} = C\frac{1-C^N}{1-C} + 2\frac{C}{1-C/B}\left\{\frac{1-B^N}{1-B} - \frac{1-C^N}{1-C}\right\}, \quad (49)$$

and this is relation (13).

A.4 Evaluation of Var{$T_s(\omega)$}

A direct evaluation of the Var$\{T_s\}$ is possible but the resulting formula is lengthy. To simplify our analysis we will make an additional assumption.

Let us define [see (17f)]

$$Q = \sum_{k=1}^{N} W_1W_2\cdots W_k = X + jY \quad (50)$$

and

$$\bar{X} = E\{X\}, \qquad \bar{Y} = E\{Y\}.$$

Then, if N is large, since W_i are independent we can assume that X and Y are jointly normal. This assumption will permit us to avoid calculations of fourth-order statistics.

Since X, Y are jointly normal, the following relations are valid (see Ref. 10, p. 374):

$$\begin{aligned} E\{X^4\} &= 3E^2\{X^2\} - 2\bar{X}^4 \\ E\{Y^4\} &= 3E^2\{Y^2\} - 2\bar{Y}^4 \\ E\{X^2Y^2\} &= 2E^2\{XY\} - 2\bar{X}^2\bar{Y}^2 + E\{X^2\}E\{Y^2\}. \end{aligned} \tag{51}$$

Then

$$\begin{aligned} \hat{\text{Var}}\{T_s\} &= E\{T_s^2\} - E^2\{T_s\} \\ &= E\{X^2 + Y^2)^2\} - E^2\{T_s\} \\ &= E\{X^4\} + E\{Y^4\} + 2E\{X^2Y^2\} - E^2\{T_s\} \\ &= 3E^2\{X^2\} + 3E^2\{Y^2\} - 2\bar{X}^4 - 2\bar{Y}^4 \\ &\quad + 4E^2\{XY\} - 4\bar{X}^2\bar{Y}^2 + 2E\{X^2\}E\{Y^2\} \\ &\quad - E^2\{X^2\} - E^2\{Y^2\} - 2E\{X^2\}E\{Y^2\} \end{aligned}$$

or

$$\hat{\text{Var}}\{T_s\} = 2\{E^2\{X^2\} + E^2\{Y^2\} + 2E\{XY\}\} - 2(\bar{X}^2 + \bar{Y}^2)^2, \tag{52}$$

which is relation (15).

To evaluate (52) we need $\bar{X}$, $\bar{Y}$, $E\{X^2\}$, $E\{Y^2\}$, $E\{XY\}$. We evaluate these terms below.

Since $Q = X + jY$ we have

$$\begin{aligned} E\{Q^2\} &= E\{X^2\} - E\{Y^2\} + 2jE\{XY\} \\ E\{T_s\} &= E\{|Q|^2\} = E\{X^2\} + E\{Y^2\}. \end{aligned}$$

Thus

$$\begin{aligned} \bar{X} &= \text{Re } E\{Q\} \\ \bar{Y} &= \text{Im } E\{Q\} \\ E\{X^2\} &= \{E\{T_s\} + \text{Re } E\{Q^2\}\}/2 \\ E\{Y^2\} &= \{E\{T_s\} - \text{Re } E\{Q^2\}\}/2 \\ E\{XY\} &= \text{Im } E\{Q^2\}/2. \end{aligned} \tag{53}$$

Hence, we need evaluate only $E\{Q\}$, $E\{Q^2\}$, since $E\{T_s\}$ has already been calculated in (45). From (50) and (44) we obtain

$$E\{Q\} = W\frac{1 - W^N}{1 - W}. \tag{54}$$

Also, following a similar method with the evaluation of $E\{T_s\}$, we can find that

$$E\{Q^2\} = Z\frac{1 - Z^N}{1 - Z} + 2\frac{Z}{1 - Z/W}\left\{\frac{1 - W^N}{1 - W} - \frac{1 - Z^N}{1 - Z}\right\}, \tag{55}$$

where

$$Z(\omega) = E\{W_i^2(\omega)\}$$

and this relation completes the evaluation of $\hat{\text{Var}}\{T_s\}$.

A.5 Evaluation of $E\{T_sT_r\}$

A direct evaluation of $E\{T_sT_r\}$ is possible but the derivation and the resulted formula are lengthy. To facilitate our analysis, let us assume that $Q = X + jY$ and T_r are jointly normal. Then

$$E\{T_sT_r\} = E\{X^2T_r + Y^2T_r\}. \tag{56}$$

But since X, Y and T_r are assumed jointly normal we have

$$\begin{aligned} E\{X^2T_r\} &= 2\bar{X}E\{XT_r\} - 2\bar{X}^2E\{T_r\} + E\{X^2\}E\{T_r\} \\ E\{Y^2T_r\} &= 2\bar{Y}E\{YT_r\} - 2\bar{Y}^2E\{T_r\} + E\{Y^2\}E\{T_r\} \end{aligned} \tag{57}$$

and

$$\begin{aligned} &E\{T_sT_r\} - E\{T_s\}E\{T_r\} \\ &\qquad = 2\{\bar{X}E\{XT_r\} + \bar{Y}E\{YT_r\} - E\{T_r\}(\bar{X}^2 + \bar{Y}^2)\}. \end{aligned} \tag{58}$$

All the above terms have been evaluated in (53) except for $E\{XT_r\}$ and $E\{YT_r\}$. Since $Q = X + jY$ and T_r is real, it is enough to find $E\{QT_r\}$. From relations (50) and (4) we obtain

$$\begin{aligned} E\{QT_r\} &= E\left\{\sum_{k,m=1}^{N} W_1 \cdots W_k / W_1 \cdots W_m |^2\right\} \\ &= \sum_{k=1}^{N}\sum_{m=1}^{K} D^m W^{k-m} + \sum_{m=2}^{N}\sum_{k=1}^{m-1} D^k W^{m-k} \end{aligned}$$

or

$$E\{QT_r\} = D\frac{1 - D^N}{1 - D} + 2\frac{D}{1 - DW}\left[\frac{1 - W^N}{1 - W} - \frac{1 - D^N}{1 - D}\right], \tag{59}$$

and then

$$\begin{aligned} E\{XT_r\} &= \text{Re } E\{QT_r\} \\ E\{YT_r\} &= \text{Im } E\{QT_r\}. \end{aligned} \tag{60}$$

A.6 Evaluation of $\mathrm{Var}\{\sigma_o^2\}$

We have defined

$$\sigma_o^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)d\omega. \tag{61}$$

Hence

$$\begin{aligned}\mathrm{Var}^2\{\sigma_o^2\} &= E\{\sigma_o^4\} - E^2\{\sigma_o^2\} \\ &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int [E\{S(u)S(v)\} - E\{S(u)\}E\{S(v)\}]dudv \\ &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int I(u, v)dudv. \end{aligned}\tag{62}$$

To facilitate our calculations we will assume that the random term is small compared to the systematic term and we shall approximate the spectrum $S(\omega)$ as

$$S(\omega) \approx \Phi_s(\omega)T_s(\omega). \tag{63}$$

A direct evaluation of $\mathrm{Var}\{\sigma_o^2\}$ is possible but requires knowledge of

$$\begin{aligned} R_1(u, v) &= E\{W_i(u)W_j(v)\} \\ R_2(u, v) &= E\{W_i(u)W_j^*(v)\}. \end{aligned}\tag{64}$$

A simpler approach is to evaluate an upper and lower bound for $\mathrm{Var}\{\sigma_o^2\}$. We can obtain an upper bound if we assume that $S(u)$ and $S(v)$ are highly correlated, i.e.,

$$E\{(S(u) - \bar{S}(u))(S(v) - \bar{S}(v))\} = \mathrm{Var}\{S(u)\}\mathrm{Var}\{S(v)\}, \tag{65}$$

and a lower bound if we assume that $S(u)$ and $S(v)$ are uncorrelated, i.e.,

$$E\{(S(u) - \bar{S}(u))(S(v) - \bar{S}(v))\} = \hat{\mathrm{Var}}\{S(u)\}\delta(u - v). \tag{66}$$

Then

$$\frac{1}{2\pi}\left|\int_{-\infty}^{\infty}\hat{\mathrm{Var}}\{S(\omega)\}d\omega\right|^{1/2} < \mathrm{Var}\{\sigma_o^2\} < \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{Var}\{S(\omega)\}d\omega. \tag{67}$$

$\mathrm{Var}\{S(\omega)\}$ has already been evaluated in (10b). It is expected that for long chains (large N) the true value will be closer to the lower bound, while for short chains (small N), it will be closer to the upper bound. From our numerical simulations we found the average of the two bounds to be a good estimator for $\mathrm{Var}\{\sigma_o^2\}$.

APPENDIX B

On the Jitter Transfer Function of a Tuned Circuit

In most applications the jitter transfer function of a timing recovery circuit is approximated with the phase transfer function of the timing passive bandpass filter. In this appendix we derive in a simple way the phase transfer function of an arbitrary filter. The limits of the applicability of the derived formula are discussed. Finally, the results are extended for the case when a prefilter followed by a squarer is used.

B.1 Phase transfer function

In this part the phase transfer function of a narrow passive bandpass filter is shown to be

$$W(\omega) = \frac{H(\omega - \omega_0)e^{j\phi(\omega_0)} + H(\omega + \omega_0)e^{-j\phi(\omega_0)}}{2A(\omega_0)}, \quad \text{for } |\omega| < \omega_0, \qquad (68)$$

where $H(\omega) = A(\omega)e^{j\phi(\omega)}$ is the transfer function of the bandpass filter and ω_0 is the baud rate of the received data. In Fig. 13 we illustrate the above relation.

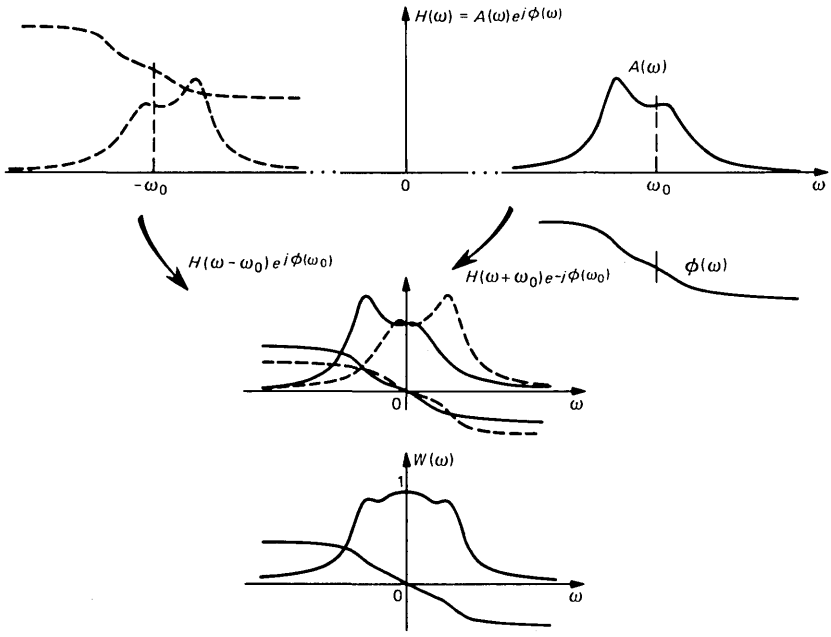


Fig. 13—Construction of the jitter transfer function $W(\omega)$ from $H(\omega)$, the transfer function of the retiming circuit filter.

Proof:

Let

$$f(t) = e^{j[\omega_0 t + e(t)]} \tag{69}$$

be the input of the bandpass filter and $e(t)$ the input jitter. We decompose $e(t)$ into its low-frequency and high-frequency parts, $e_L(t)$ and $e_H(t)$, respectively. We define $e_L(t)$ as the component of $e(t)$ with frequencies much less than the bandwidth of the bandpass filter. Thus

$$e(t) = e_L(t) + e_H(t). \tag{70}$$

Then if $|e_H(t)| \ll 1$ we can write $f(t)$ as

$$f(t) = e^{j[\omega_0 t + e_L(t)]}[1 + je_H(t) - \cdots], \tag{71}$$

which is a narrowband process centered at ω_0. Let $H_L(\omega)$ be the normalized low-pass equivalent of $H(\omega)$, where

$$H_L(\omega) = \frac{H(\omega + \omega_0)}{A(\omega_0)} e^{-j\phi(\omega_0)} \quad \omega > -\omega_0. \tag{72}$$

Then the output of the filter is[4]

$$g(t) = f(t)*h(t) = e^{j\omega_0 t}A(\omega_0)e^{j\phi(\omega_0)}\{h_L(t)*e^{je(t)}\}, \tag{73}$$

where $*$ indicates convolution and $h(t)$, $h_L(t)$ are the impulse responses corresponding to $H(\omega)$ and $H_L(\omega)$. The output phase is the phase of the term $h_L(t)*\exp(je(t))$, plus the static phase shift of $e^{j\phi(\omega_0)}$.

Using that [see (72)]

$$\int_{-\infty}^{\infty} h_L(t)dt = H_L(0) = 1$$

and because of the definition of $e_L(t)$, we can assume $e_L(t)$ to be constant compared with $h_L(t)$, i.e.,

$$h_L(t)*e^{je_L(t)} \approx H_L(0)e^{je_L(t)} = e^{je_L(t)}.$$

We therefore obtain

$$\begin{aligned} h_L(t)*e^{je(t)} &\approx e^{je_L(t)}\{1 + je_H(t)*h_L(t)\} \\ &\approx e^{j[e_L(t)+e_H(t)]*h_L(t)} \\ &= e^{j[e(t)*h_L(t)]}. \end{aligned} \tag{74}$$

Thus, since $e(t)$ is real,

$$\begin{aligned} g(t) &= A(\omega_0)e^{j[\omega_0 t + \phi(\omega_0)]}e^{j[e(t)*h_L(t)]} \\ &= A(\omega_0)e^{-[e(t)*\mathrm{Im}h_L(t)]}e^{j[\omega_0 t + \phi(\omega_0) + e(t)*\mathrm{Re}h_L(t)]}, \end{aligned} \tag{75}$$

where

$A(\omega_0)e^{-e(t)*\mathrm{Im}h_L(t)}$ is an amplitude modulation term. Thus the phase transfer function is the Fourier transform of $w(t) = \mathrm{Re}\{h_L(t)\}$, that is,

$$W(\omega) = \frac{H_L(\omega) + H_L^*(-\omega)}{2}, \tag{76}$$

and using (72) we obtain, finally, (68).

To derive (9) we have made two assumptions:

$$\text{(a)} \quad |e_H(t)| \ll 1 \quad \text{and} \quad |e_H(t)*h_L(t)| \ll 1 \tag{77}$$

$$\text{(b)} \quad |e_L(t)| \approx \text{constant}. \tag{78}$$

When $W(\omega)$ is going to be used to examine the accumulation of jitter in a chain of repeaters, the validity of (10) and (11) must be questioned for every repeater in the chain. For the Nth repeater, $e_L(t)$ represents the accumulated jitter appearing in the clock of the $(N-1)$-th repeater, while $e_H(t)$ represents the additional jitter generated by the Nth repeater section, and it is always very small. Thus, assumptions (a) are valid in general, while assumption (b) is true only if jitter peaking does not occur. As is known, jitter peaking occurs if $\max|W(\omega)| > 1$, and in such a case the accumulated jitter grows exponentially. Thus relation (68) can be used to evaluate jitter accumulation when we have no jitter peaking. If jitter peaking is present, then

$$e_L(t) \approx A_0 \cos \omega_p t,$$

where ω_p is the peaking frequency, i.e.,

$$|W(\omega_p)| = \max|W(\omega)|$$

and (68) can be used only if $|e(t)| \ll 1$. This limits the applicability of the formula to short chains of regenerators. It is our feeling that in the case of jitter peaking, the linear model in (68) will overestimate jitter accumulation for long chains, because the nonlinear model will shift energy from the peaking frequency band to other bands. Preliminary simulations appear to agree with the above statement.

Relation (68) also suggests that

1. Filters with symmetric ripples in their passband are undesirable because they will always create jitter peaking. This is due to the normalization factor $2A(\omega_0)$ in (68).

2. Even with a monotonic filter, jitter peaking can occur if the data frequency, f_0, is placed away from the filter's center frequency f_c.[7] Define the detuning parameter as follows:

$$\alpha = \frac{\omega_0 - \omega_c}{B},$$

where B is the bandwidth of the low-pass filter $H_L(\omega)$. Then jitter peaking usually occurs if $\alpha > 1$.

These results are illustrated in Fig. 14, where the phase transfer function of a second-order Butterworth is plotted for various α.

B.2 Timing circuit with prefilter and squarer

In case a prefilter and a squarer are used we can modify the above analysis and also take the above circuits into consideration.

With ω_0 denoting the baud rate of the received data, the component that is going to generate the clock is located at $\omega_0/2$. Therefore, if $e(t)$ is the jitter present in the input data, the component of the input located at $\omega_0/2$ can be represented as

$$f(t) = e^{j[\omega_0 t + e(t)]/2}. \tag{79}$$

Notice that $e(t)$ also contains the jitter generated by the repeater.

Let us also define as $p(t)$ the output of the prefilter when the symbol 1 is transmitted, and let $P(\omega) = B(\omega)e^{j\Psi(\omega)}$ be the Fourier Transform of $p(t)$, i.e.,

$$p(t) \overset{FT}{\longleftrightarrow} P(\omega) = B(\omega)e^{j\Psi(\omega)}. \tag{80}$$

Then, using the same assumptions we used in deriving relations (74) and (75) and defining

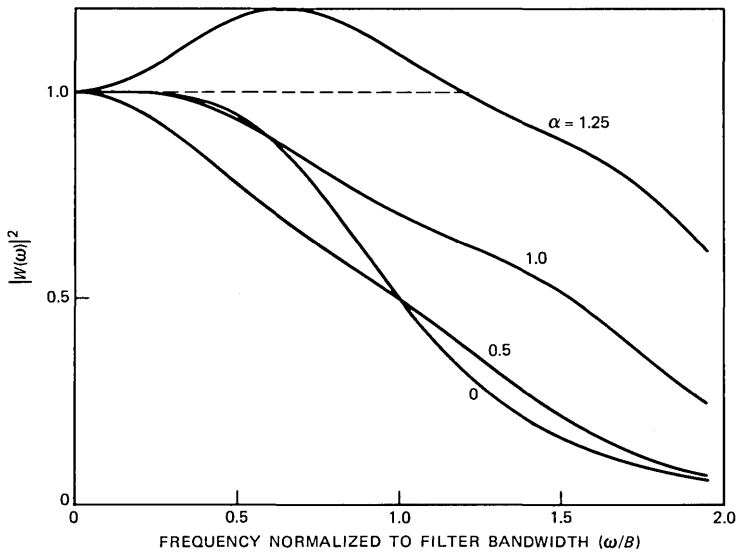


Fig. 14—Jitter transfer function of a second-order Butterworth filter for various values of the detuning parameter α.

$$P_L(\omega) = \frac{P(\omega + \omega_0/2)}{B(\omega_0/2)} e^{-j\Psi(\omega_0/2)} \tag{81}$$

as the low-pass equivalent of $P(\omega)$, we obtain

$$\begin{aligned} g(t) &= [f(t)^* p(t)]^{2*} h(t) \\ &\approx B(\omega_0/2) e^{j[\omega_0 t + 2\Psi(\omega_0/2)]} e^{j[e(t)^* p_L(t)]} * h(t) \\ &\approx A(\omega_0) B(\omega_0/2) e^{j[\omega_0 t + 2\Psi(\omega_0/2) + \phi(\omega_0)]} e^{j[e(t)^* p_L(t)^* h_L(t)]}. \end{aligned} \tag{82}$$

Reasoning as in (76) we obtain

$$W(\omega) = \frac{P_L(\omega) H_L(\omega) + P_L^*(-\omega) H_L^*(-\omega)}{W(0)}. \tag{83}$$

Since $P(\omega)$ is much wider than $H_L(\omega)$, we can assume that

$$P_L(\omega) H_L(\omega) \approx H_L(\omega), \tag{84}$$

which implies that the presence of a prefilter will not change significantly the jitter transfer function of the repeater. However, the jitter generated within the repeater may significantly depend on the prefilter and squarer.

AUTHOR

Christodoulos Chamzas, Diploma degree (Electrical and Mechanical Engineering), the National Technical University of Greece, Athens, Greece, 1974; M.S., 1975, Ph.D., 1979 (Electrical Engineering), Polytechnic Institute of New York, Farmingdale, N.Y.; AT&T Bell Laboratories, 1982—. From 1979 to 1982 Mr. Chamzas was an Assistant Professor with the Department of Electrical Engineering at Polytechnic Institute of New York, where he is currently a part-time Visiting Professor for the Imaging Institute. At AT&T Bell Laboratories he has been working on problems in high-speed pseudorandom noise generators, the Submarine Lightwave System and adaptive echo cancelers. His primary interests are in signal processing and communication systems. Member, Technical Chamber of Greece, Sigma Xi.

AT&T Technical Journal
Vol. 64, No. 1, January 1985
Printed in U.S.A.

Nonlocal Input-Output Expansions

By I. W. SANDBERG*

(Manuscript received July 16, 1984)

In a recent series of papers by this writer on the existence, determination, and properties of power-series-like expansions for expressing a nonlinear system's outputs in terms of its inputs, the emphasis is primarily on *locally* convergent expansions. Here we report on related general results concerning nonlocal expansions, including in particular material concerning the *size* of the region of convergence. One of the results given provides useful *necessary and sufficient* conditions under which f^{-1} has a generalized power-series expansion, where f is a certain important general type of invertible map (that, for example, might take one set of complex-valued signals defined on $[0, \infty)$ into another).

I. INTRODUCTION

In a recent series of papers including Refs. 1 through 3 on the existence, determination, and properties of power-series-like expansions for expressing a nonlinear system's outputs in terms of its inputs, the emphasis is primarily on *locally* convergent expansions. Here, in Section II, we report on related general results concerning nonlocal expansions, including in particular material concerning the *size* of the region of convergence.

More specifically, Theorem 1 in Section II gives necessary and sufficient conditions under which f^{-1} has a generalized power series expansion (see Section II for the details) when f is an invertible locally Lipschitz map between certain general subsets of two complex Banach

* AT&T Bell Laboratories.

spaces. Theorem 2 provides an algorithm for obtaining the expansion whenever it exists.

In Section 2.4 we use Theorems 1 and 2 to prove results concerning a system model considered in Ref. 3 and in earlier papers (e.g., Ref. 2). This model is characterized by five operators: a nonlinear operator N and linear operators, A, B, C, and D. The system input v and corresponding output w are related by the equations

$$y = Nx \tag{1}$$

$$x = Av + Cy \tag{2}$$

$$w = Dv + By, \tag{3}$$

where x and y, respectively, can be interpreted as the input and output of the nonlinear portion of the system.* In Ref. 3 it is assumed that y, x, v, and w are n-vector valued and defined for $t \geq 0$, where as usual n is an arbitrary positive integer; here these quantities are allowed to belong to a general complex Banach space $\mathscr{B}$, but we shall be interested mainly in the case where $\mathscr{B}$ is the space L_∞ of bounded functions considered in Ref. 3. Theorem 4 in Section 2.4 shows that, in an interesting and important setting, w can be expressed as a certain power series in v that converges for $Av \in V$, where V is *any* open ball in $\mathscr{B}$ centered at the origin such that there is an open subset V_0 of $\mathscr{B}$ for which $(I - CN)$ (I is the identity operator on $\mathscr{B}$) is a homeomorphism of V_0 onto V, with $(I - CN)^{-1}$ locally Lipschitz on V in the sense of Section 2.1. Theorem 4 is actually somewhat more general than is indicated above, and it together with its relation to earlier work is discussed in Section 2.4.2. A pertinent example is given in Appendix B.

II. HOMOGENEOUS POLYNOMIALS AND EXPANSIONS FOR MAPS BETWEEN SUBSETS OF COMPLEX BANACH SPACES

2.1 Preliminaries

Throughout the paper, $\mathscr{B}_0$ and $\mathscr{B}$ are Banach spaces such that there is a linear homeomorphism q of $\mathscr{B}$ onto $\mathscr{B}_0$. (Thus, $\mathscr{B}_0$ and $\mathscr{B}$ are taken to be isomorphic. We shall be interested mainly in the case in which in fact $\mathscr{B}_0 = \mathscr{B}$.) The same symbols $\|\cdot\|$ and θ, respectively, are used to denote the norm and zero element in $\mathscr{B}_0$ as well as in $\mathscr{B}$. Unless stated otherwise, $\mathscr{B}_0$ and $\mathscr{B}$ are assumed to be over the field of *complex* scalars.

---

* The equations of a very large class of systems can be written in the form (1) through (3), with N memoryless. It is not necessary that the equations of the system to be studied be given *at the outset* in the form (1) through (3) (see Ref. 3, Appendices I and II). In Ref. 3, y, x, and w are taken to belong to an *extended* space $L_{\infty e}$. For the purposes of this paper extended space concepts are not needed.

A map g from a nonempty open subset U of $\mathscr{B}$ or of $\mathscr{B}_0$ into $\mathscr{B}_0$ or $\mathscr{B}$, respectively, is *locally Lipschitz* on U if for each $a \in U$ there are a positive number c_a and an open ball $\beta_a \subset U$ centered at a such that $\|g(u_1) - g(u_2)\| \leq c_a \|u_1 - u_2\|$ for u_1 and u_2 in β_a. We use $d^m g(a)$ to denote the mth-order Fréchet derivative (Ref. 4, pp. 149, 181) of g at a point $a \in U$, assuming it exists. A sufficient condition for g to be locally Lipschitz on U is that it be continuously Fréchet differentiable in U.

For V_0 and V any nonempty open subsets of $\mathscr{B}_0$ and $\mathscr{B}$, respectively, $H(V_0, V)$ denotes the set of all homeomorphisms h of V_0 onto V such that h is locally Lipschitz on V_0. Our results are concerned with this class of maps with V assumed to be a *c-star* about some $p \in \mathscr{B}$, by which is meant that $V = \{x \in \mathscr{B}: x = p + W\}$, where W is a subset of $\mathscr{B}$ with the property that $zw \in W$ for $w \in W$ and any complex scalar z with $|z| \leq 1$. The concept of a c-star is of importance in studies of the region of convergence of abstract power series expansions (see Ref. 5, Theorems 26.5.9 and 26.6.1). In this paper we could have restricted attention to the case where V is an open ball centered at p, which clearly is a c-star about p, but this was not done because no significant simplification results.

Given any positive integer m, by an *m-linear* map Q from $\mathscr{B}^m$ into $\mathscr{B}_0$, we mean that $Q(h_1, \cdots, h_m)$ is linear (i.e., additive and homogeneous) separately in each h_j. Such a map is *symmetric* if $Q(h_1, \cdots, h_m)$ is symmetric in the variables $h_1, \cdots, h_m$. A map $M(\cdot)$ from $\mathscr{B}$ into $\mathscr{B}_0$ is called a *homogeneous polynomial* of degree m if there exists an m-linear map Q from $\mathscr{B}^m$ into $\mathscr{B}_0$ such that $M(h) = Q(h, \cdots, h)$ for all h*. We now come to a definition of central importance in our results.

For $p \in \mathscr{B}$ and $V \subset \mathscr{B}$ an open c-star about p, let $\mathscr{P}(p, V)$ denote the set of all maps g from V into $\mathscr{B}_0$ such that there are homogeneous polynomials $g_m(p, \cdot)$ of degree m ($m = 1, 2, \cdots$), from $\mathscr{B}$ into $\mathscr{B}_0$, with the properties that

$$\sum_{m=1}^{\infty} g_m(p, w)$$

converges in $\mathscr{B}_0$ for each $(p + w) \in V$, and

$$g(p + w) = g(p) + \sum_{m=1}^{\infty} g_m(p, w), \qquad (p + w) \in V. \tag{4}$$

---

* This definition of a homogeneous polynomial $M(\cdot)$ is not the same as, but is equivalent to, the one given in Ref. 5, that $M(zh) = z^m M(h)$ and $M(p + zh) = \sum_{j=1}^{m} M_j(p, h) z^j$ for p and h in $\mathscr{B}$ and any complex scalar z, where the M_j do not depend on z. Also, the first definition given above describes the same class of maps $M(\cdot)$ if "m-linear" is replaced with "symmetric m-linear".

The set $\mathscr{P}(p, V)$ is of course a set of maps g that admit a generalized power series expansion in the sense indicated. The expansion (4) for any $g \in \mathscr{P}(p, V)$ is *unique* in the sense that if

$$g(p + w) = g(p) + \sum_{m=1}^{\infty} h_m(p, w), \qquad (p + w) \in V \tag{5}$$

[by which is meant, in particular that the sum in (5) converges] with each $h_m(p, \cdot)$ a homogeneous polynomial of degree m, then $g_m(p, \cdot) = h_m(p, \cdot)$ for all m. This follows from a simple argument due to Graves (Ref. 6, p. 174). (See also Ref. 1, Section 2.7.)

Finally, we say that g belongs to $\mathscr{P}_F(p, V)$ if g belongs to $\mathscr{P}(p, V)$ and for each m there is a *continuous* symmetric m-linear Q_m from $\mathscr{B}^m$ into $\mathscr{B}_0$ that depends on p such that $g_m(p, h) = Q_m(h, \cdots, h)$ for all h. In particular, then each $g_m(p, \cdot)$ is bounded in the sense that there is a positive constant ρ_m such that $\|g_m(p, h)\| \leq \rho_m \|h\|^m$ for all m and h, with p fixed, and every $g_m(p, \cdot)$ is Fréchet differentiable on $\mathscr{B}$.

2.2 Inverses of maps in $H(V_0, V)$ and generalized power series expansions

Our first result, Theorem 1 below, provides a complete characterization of those f's in $H(V_0, V)$, with V a c-star, for which f^{-1} has a generalized power series expansion of the type described in the preceding section.

Theorem 1: Let V_0 and V be nonempty open subsets of $\mathscr{B}_0$ and $\mathscr{B}$, respectively, with V a c-star about some point p in $\mathscr{B}$. Let $f \in H(V_0, V)$. Then, $f^{-1} \in \mathscr{P}(p, V)$ if and only if f is Fréchet differentiable on V_0 and f^{-1} is locally Lipschitz on V. In addition, we have $f^{-1} \in \mathscr{P}_F(p, V)$ whenever $f^{-1} \in \mathscr{P}(p, V)$.

2.2.1 Proof of Theorem 1

We first prove the following lemma.

Lemma 1: Let V_0 and V be nonempty open subsets of $\mathscr{B}_0$ and $\mathscr{B}$, respectively, with $\mathscr{B}_0$ and $\mathscr{B}$ over the same field, either real or complex. Let h be an invertible Fréchet continuously differentiable map of V_0 onto V, with h^{-1} locally Lipschitz on V. Then $d(h^{-1})(\cdot)$ (the Fréchet derivative of h^{-1}) exists and is continuous throughout V.

Proof of Lemma 1: With q the homeomorphism mentioned in Section 2.1, define s on V_0 by $s(x) = qh(x)$ for $x \in V_0$. It is not difficult to verify that s is a continuously Fréchet differentiable invertible map of $V_0 \subset \mathscr{B}_0$ onto the open subset* $q(V)$ of $\mathscr{B}_0$, and that s^{-1} is locally Lipschitz on $q(V)$. Since the inverses h^{-1} and s^{-1} of h and s, respec-

* The set $q(V)$ is *open* by the open mapping theorem, or because it is simply the inverse image of the open set V under the continuous map q^{-1}.

tively, are related by $h^{-1} = s^{-1}q$, it follows that the lemma holds in general if it holds for $\mathscr{B} = \mathscr{B}_0$. We therefore assume throughout the remainder of the proof that $\mathscr{B}$ and $\mathscr{B}_0$ are the same spaces.

Let $a \in V_0$ be arbitrary, and let $c_{h(a)}$ and $\beta_{h(a)}$, respectively, be a positive constant and an open ball in V centered at $h(a)$ such that $\|h^{-1}(u_1) - h^{-1}(u_2)\| \leqslant c_{h(a)}\|u_1 - u_2\|$ for u_1 and u_2 in $\beta_{h(a)}$. Let $S = \{x \in \mathscr{B}: (x + a) \in V_0\}$. Let β denote the open ball in $\mathscr{B}$ centered at the origin, with the same radius as $\beta_{h(a)}$.

Define $H_a: S \to \mathscr{B}$ by $H_a(x) = h(x + a) - h(a)$, $x \in S$. The set S is open in $\mathscr{B}$. Thus, by the continuity of H_a, the inverse image $H_a^{-1}(\beta)$ is open in S and hence in $\mathscr{B}$. [The fact that $H_a^{-1}(\beta)$ is an open subset of $\mathscr{B}$ is used in connection with (6) and (7) below, and at the end of the proof.]

Let $h_a: H_a^{-1}(\beta) \to \beta$ be given by $h_a(x) = H_a(x)$, $x \in H_a^{-1}(\beta)$. By the invertibility of h, h_a is an invertible map of $H_a^{-1}(\beta)$ onto β. Let h_a^{-1} denote its inverse, and define $g: \beta \to \mathscr{B}$ by

$$g(u) = u - dh(a)h_a^{-1}u + w, \qquad u \in \beta$$

for $w \in \beta$, where $dh(a)$ is the Fréchet derivative of h at a.

By the mean value theorem in Ref. 4, p. 160, and the continuity of $dh(\cdot)$,

$$\frac{\|h_a(z_1) - h_a(z_2) - dh(a)(z_1 - z_2)\|}{\|z_1 - z_2\|} \to 0 \tag{6}$$

as $\max(\|z_1\|, \|z_2\|) \to 0$ with $z_1 \neq z_2$.* Therefore, with $\sigma > 0$ such that $\sigma c_{h(a)} < 1$, there is a $\delta > 0$ for which $\{z \in \mathscr{B}: \|z\| \leqslant \delta\} \subset H_a^{-1}(\beta)$ and

$$\frac{\|h_a(z_1) - h_a(z_2) - dh(a)(z_1 - z_2)\|}{\|z_1 - z_2\|} \leqslant \sigma \tag{7}$$

for $z_1 \neq z_2$, $\|z_1\| \leqslant \sigma$, and $\|z_2\| \leqslant \delta$. Since $\|z\| \leqslant c_{h(a)}\|u\|$ for $u \in \beta$ and $z = h_a^{-1}u$ [because then $h(a + z) = u + h(a)$, which gives $z = h^{-1}(u + h(a)) - h^{-1}(h(a))$], we have

$$\frac{\|u_1 - u_2 - dh(a)[h_a^{-1}u_1 - h_a^{-1}u_2]\|}{\|h_a^{-1}u_1 - h_a^{-1}u_2\|} \leqslant \sigma \tag{8}$$

for $u_1 \neq u_2$, $\|u_1\| \leqslant \rho$, and $\|u_2\| \leqslant \rho$, where $\rho = \min(\rho_\beta/\alpha, \delta(c_{h(a)})^{-1})$, ρ_β is the radius of β, and α is any number in $(1, \infty)$.

Now let u_1 and u_2 belong to β, and define x_1 and x_2 by $x_1 = h_a^{-1}u_1$ and $x_2 = h_a^{-1}u_2$. Clearly, $h(x_1 + a) = h(a) + u_1$ and $h(x_2 + a) = h(a) + u_2$, which gives $\|x_1 - x_2\| \leqslant c_{h(a)}\|u_1 - u_2\|$. This, together with (8) and

---

* With regard to Ref. 4, p. 160, $h_a(z_1) - h_a(z_2) - dh(a)(z_1 - z_2) = [h(a + z_1) - dh(a)z_1] - [h(a + z_2) - dh(a)z_2]$.

$\sigma c_{h(a)} < 1$, shows that the map g defined above is a contraction on the set $S_0 = \{u \in \mathscr{B}: \|u\| \leq \rho\}$. Since $h_a^{-1}(\theta) = \theta$, it is easy to see that for all $w \in \mathscr{B}$ with $\|w\| \leq \rho_0$ and $\rho_0 > 0$ sufficiently small, g maps S_0 into itself. By the contraction-mapping fixed point theorem, g has a unique fixed point* in S_0, and thus there is a unique solution $u \in S_0$ of $dh(a)h_a^{-1}u = w$, for each such w.

By the linearity of $dh(a)$, we see that for each $w \in \mathscr{B}$ there is a unique $x \in \mathscr{B}$ such that $dh(a)x = w$, the uniqueness following from the fact that h_a maps an open ball in $\mathscr{B}$ centered at θ into S_0 [recall that $h_a(\theta) = \theta$, that h_a is continuous, and that $H_a^{-1}(\beta)$ is open in $\mathscr{B}$].† This shows that $dh(a)$ is an invertible map of $\mathscr{B}$ onto $\mathscr{B}$. By the boundedness of $dh(a)$, $dh(a)^{-1}$ is bounded (Ref. 9, p. 119). Since $a \in V_0$ is arbitrary, and $dh(\cdot)$ is assumed to be continuous, it follows (Ref. 4, p. 273) that h^{-1} is continuously Fréchet differentiable on V, which proves the lemma.‡

Returning now to the proof of the theorem, suppose initially that $f^{-1} \in \mathscr{P}(p, V)$. Since the series associated with f^{-1} is a (G)-power series in the sense of Ref. 5, p. 773, this series is (G)-differentiable (i.e., Gâteaux differentiable) in V (Ref. 5, p. 773). By the (G)-differentiability and continuity of f^{-1} (which implies that f^{-1} is analytic), it follows Ref. 5, Theorem 3.17.1 that f^{-1} is Fréchet differentiable throughout V. Using the following lemma,§ which is proved in Appendix A and which is used also in Section 2.3, $d(f^{-1})(\cdot)$ is continuous.

Lemma 2: If g is a Fréchet differentiable map of a nonempty open subset V of $\mathscr{B}$ into a complex Banach space $\mathscr{B}_1$, then g is twice Fréchet differentiable on V.

---

* This part of the proof is along the lines of proofs of the classical implicit function theorem (see, for example, Ref. 7, pp. 194–5). Lemma 1 is related to a theorem stated in Ref. 8, p. 165, but the argument given there does not prove the theorem. (In the terminology of Ref. 8, it is not shown that the definition of a linearization leads to the limit given. However, the proof above shows that the theorem in Ref. 8, p. 165 is true under the additional hypotheses that L_x exists and is continuous in a neighborhood of $x = x_0$.)

† A modification of an argument given in Ref. 8, p. 165 could also have been used to obtain the uniqueness.

‡ With regard to Lemma 1 and the condition in Section 2.1 that there is a linear homeomorphism q that maps $\mathscr{B}$ onto $\mathscr{B}_0$, we note that, in the absence of that condition, the existence of $dh(\cdot)$ and $dh^{-1}(\cdot)$ as indicated in the lemma *implies* (see Proposition 1 in Section 2.3.1) that the condition holds. This shows that there is no loss of generality in assuming that the condition is met.

§ This lemma, which is an analog of a standard proposition in the classical theory of functions of a complex variable, is probably a known result, but we have not encountered it in the literature. However, for related material concerning *Gâteaux* variations, see Ref. 5. The lemma provides additional understanding concerning some results proved in Ref. 1 and other recent papers. While it does not strengthen these results, it shows that some hypotheses actually follow from others that were introduced (see, e.g., Ref. 1, Theorem 2, which is an early result concerning nonlocal expansions).

Thus f^{-1} is locally Lipschitz on V, and, by Lemma 1, $df(\cdot)$ exists on V_0. Using the analyticity of f^{-1}, by Ref. 5, Theorems 3.16.2, 26.3.4 and 26.3.6, one has $f^{-1} \in \mathscr{P}_F(p, V)$.

Now assume that f is (F)-differentiable, and thus continuously (F)-differentiable, on V_0 and that f^{-1} is locally Lipschitz on V. Using Lemma 1, $d(f^{-1})(\cdot)$ exists in V. In particular, f^{-1} is (G)-differentiable in V, and by Ref. 5, Theorems 3.16.2 and 26.3.5, we have $f^{-1} \in \mathscr{P}(p, V)$, which completes the proof of the theorem.

2.2.2 Comments

Proposition 1 in Section 2.3.1 and the proof of Theorem 1 show that under the hypotheses of the theorem, $f^{-1} \in \mathscr{P}(p, V)$ implies also that $df(a)$ is a homeomorphism of $\mathscr{B}_0$ onto $\mathscr{B}$ for every $a \in V_0$.*

For the extreme case in which both $\mathscr{B}_0$ and $\mathscr{B}$ are just the space of complex numbers with the absolute value norm, the condition that V is an open c-star about p reduces to the requirement that V is an open disk in the complex plane centered at p. In that case, by standard results in the classical theory of functions of a complex variable, $f^{-1} \in \mathscr{P}(p, V)$ implies that the ordinary derivative $(f^{-1})'(z)$ exists for $z \in V$. In addition, since f^{-1} is one-to-one on V, it follows from a known result (Ref. 10, Theorem 16-23) that $(f^{-1})'(z) \neq 0$ for $z \in V$, which shows that $f'(\cdot)$ exists on V_0. Similarly, using Ref. 10, Theorem 16-23, it follows from the hypothesis that $f'(z)$ exists for $z \in V_0$ that $f^{-1} \in \mathscr{P}(p, V)$.† Theorem 1 can be viewed as a Banach space relative of these propositions.

2.3 Construction of the expansion of f^{-1}

Here we give an algorithm, along the lines of Theorem 2 of Ref. 1, for expanding f^{-1}.

Theorem 2: Let the hypotheses of Theorem 1 be met, and assume that $f^{-1} \in \mathscr{P}(p, V)$. Then for each $l = 1, 2, \cdots$, the lth-order Fréchet derivative $d^l f(x)$ exists for $x \in V_0$, and

$$f^{-1}(p + h) = f^{-1}(p) + \sum_{m=1}^{\infty} g_m(p, h), \qquad (p + h) \in V \tag{9}$$

where

$$g_1(p, h) = (df[f^{-1}(p)])^{-1}h$$

---

* And this adds to material in Ref. 1, Section 2.6 concerning the necessity of a certain invertibility condition.

† In these observations, the hypothesis that f is locally Lipschitz is not needed. On the other hand, the Lipschitz condition is obviously a consequence of $f^{-1} \in \mathscr{P}(p, V)$.

(the inverse exists by Proposition 1, below), and

$$g_m(p, h) = -(df[f^{-1}(p)])^{-1} \sum_{l=2}^{m} (l!)^{-1} \cdot \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l f[f^{-1}(p)] g_{k_1}(p, h) g_{k_2}(p, h) \cdots g_{k_l}(p, h),$$

$$m \geq 2.^* \quad (10)$$

2.3.1 Proof of Theorem 2

Proof: Since by Theorem 1 and Proposition 1 (which appears below), $f^{-1} \in \mathscr{P}(p, V)$ implies that $df(a)$ and $[df(a)]^{-1}$ exist for each $a \in V_0$, Lemma 2, the proof of Theorem 2 of Ref. 1, and Theorem 3.16.2 of Ref. 5, show that Theorem 2 holds.

Proposition 1: If V_0 and V are nonempty open subsets of $\mathscr{B}_0$ and $\mathscr{B}$, respectively, and f is a homeomorphism of V_0 onto V such that f and f^{-1} are Fréchet differentiable on their respective domains, then $df(a)$ is a homeomorphism of $\mathscr{B}_0$ onto $\mathscr{B}$ for any $a \in V_0$.

The proposition is a well-known result (see, for example, Ref. 11, p. 175, Problem 6) provable using the relations $f^{-1}[f(x)] = x$ and $f[f^{-1}(y)] = y$ for x and y in V_0 and V, respectively, and the chain rule (see Ref. 11, pp. 171–2) for differentiating a composite function.

2.4 Theorems concerning the system model

In this section attention is focused on the system model described in Section I. We use I to denote the identity map on $\mathscr{B}$.

Let A, B, C, and D be linear maps of $\mathscr{B}$ into itself, with B and C bounded. Let N be a map from a subset S of $\mathscr{B}$ into $\mathscr{B}$, for which there are nonempty open subsets V_0 and V of $\mathscr{B}$ such that $V_0 \subset S$, V is a c-star about some point $p \in \mathscr{B}$, and $(I - CN)$ is a homeomorphism of V_0 onto V, with $(I - CN)^{-1}$ locally Lipschitz on V. (It is not difficult to give important examples in which these hypotheses on N are met. This is illustrated in Appendix B.)

Let $\mathscr{I}$ (for "set of inputs") denote the collection of all $v \in \mathscr{B}$ such that $Av \in V$, and assume that there is a $v_0 \in \mathscr{B}$ for which $Av_0 = p$. We see that for each $v \in \mathscr{I}$, there are unique x, y, and w in V_0, $\mathscr{B}$, and $\mathscr{B}$, respectively, such that

$$x = Av + Cy$$

$$w = Dv + By$$

$$y = Nx.$$

---

* In (10), $\sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}}$ denotes a sum over all positive integers $k_1, \cdots, kl$ that add to m.

Let W be the map from $\mathscr{I}$ into $\mathscr{B}$ defined by the condition that $W(v) = w$ when $v \in \mathscr{I}$. In other words, let W be given by

$$W(v) = Dv + B(I - CN)^{-1}Av, \qquad v \in \mathscr{I}. \tag{11}$$

Since N differentiable on V_0 implies (by Lemma 2) that it is continuously differentiable on V_0 and thus locally Lipschitz there, by Theorem 1 and the boundedness of C we have the following:

Theorem 3: Let N be Fréchet differentiable on V_0. Then $(I - CN)^{-1} \in \mathscr{P}_F(p, V)$.

Theorem 3 shows that, under merely the condition indicated, W has an expansion about v_0 in terms of homogeneous polynomials that is valid* for all $v \in \mathscr{I}$. With regard to actually determining the expansion for W, we have the following.

Theorem 4: Let N be Fréchet differentiable on V_0. Then for each $l = 2, 3, \cdots$ the Fréchet derivative $d^l N(\cdot)$ exists in some open neighborhood $\mathscr{N}_l$ of the point x_0 in V_0 that satisfies $x_0 - CN(x_0) = Av_0$, and for $v \in \mathscr{I}$, we have

$$W(v) = BN(x_0) + Dv + \sum_{m=1}^{\infty} BY_{(m)}(v - v_0) \tag{12}$$

where $Y_{(1)}, Y_{(2)}, \cdots$ are the homogeneous polynomials defined by the relations†

$$X_{(1)}(v - v_0) = [I - CdN(x_0)]^{-1}A(v - v_0),$$

$$X_{(m)}(v - v_0) = [I - CdN(x_0)]^{-1}C \sum_{l=2}^{m} (l!)^{-1}$$

$$\sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l N(x_0)X_{(k_1)}(v - v_0) \cdots X_{(k_l)}(v - v_0) \tag{13}$$

for $m \geq 2$, and

$$Y_{(m)}(v - v_0) = \sum_{l=1}^{m} (l!)^{-1} \sum_{\substack{k_1+k_2+\cdots+k_l=m \\ k_j>0}} d^l N(x_0)X_{(k_1)}(v - v_0)$$

$$\cdots X_{(k_l)}(v - v_0) \tag{14}$$

for $m \geq 1$.

2.4.1 Proof of Theorem 4

Proof: Let $\mathscr{B}^2$ denote the Banach space $\mathscr{B} \times \mathscr{B}$ with norm max $(\|\cdot\|, \|\cdot\|)$. Define G: $V_0 \times \mathscr{B} \to \mathscr{B}^2$ by

* Here the boundedness of B is used to ensure that $\sum_{m=0}^{\infty} Bg_m[p, A(v - v_0)]$ converges to $B\sum_{m=0}^{\infty} g_m[p, A(v - v_0)$, where $\sum_{m=0}^{\infty} g_m[p, A(v - v_0)]$ is the series for $(I - CN)^{-1}Av$.

† The inverse of $I - CdN(x_0)$ exists; see Section 2.2.2.

$$G_1(p_1, p_2) = p_1 - CNp_1$$

$$G_2(p_1, p_2) = Np_1 - p_2$$

for $(p_1, p_2) \in V_0 \times \mathscr{B}$. With β a nonempty open ball in $\mathscr{B}$ centered at θ, let

$$\mathscr{D} = \{(p_1, p_2) \in V_0 \times \mathscr{B} : G(p_1, p_2) = (q_1, q_2), (q_1, q_2) \in V \times \beta\}$$

[i.e., let $\mathscr{D} = G^{-1}(V \times \beta)$]. We see that $\mathscr{D}$ is *open* in $\mathscr{B}^2$ (because V and β are open, and G is continuous and defined on an open subset of $\mathscr{B}^2$), and that $V \times \beta$ is a *c-star* in $\mathscr{B}^2$ centered at (p, θ). The restriction F of G to $\mathscr{D}$ is a differentiable homeomorphism of $\mathscr{D}$ onto $V \times \beta$. Using this fact, it is not difficult to show that a proof of Theorem 4 can be obtained by proceeding as in Ref. 2, proof of part (ii) of Theorem 1, but with our Theorem 2 with $\mathscr{B} = \mathscr{B}_0$ employed instead of Lemma 1 in Ref. 2.* (It is easy to verify that for F as described above, F^{-1} is locally Lipschitz.)

2.4.2 Discussion

Under the conditions on N of Theorem 4, N has the representation

$$N(x) = N(x_0) + \sum_{l=1}^{\infty} (l!)^{-1} d^l N(x_0)(x - x_0)^l \tag{15}$$

in which the series converges (for example) uniformly for x in some sufficiently small ball in S centered at x_0,† and of course (15) provides an important interpretation of the maps $d^l N(x_0)$ that appear in (13) and (14).

It is often useful to observe that the operator $(I - CN)^{-1}$ in (11) can naturally be identified with the "feedback part" of the system represented by (1), (2), and (3) (see Ref. 3, Figs. 1 and 2). Aside from considerations concerning the differentiability hypothesis on N, and the existence of a v_0 as described, Theorem 4 shows that the series representation given by (12) holds for $Av \in V$, whenever V is an open c-star about some p such that the equation $x - CNx = u$ of the feedback portion is, so to speak, uniquely locally-Lipschitz-solvable in some open subset of S for every u in V. Implicit in this is the assumption that $\mathscr{B}$ is a *complex* Banach space. Since the inputs and outputs of most nonlinear systems of direct interest are *real* valued functions, the main point of Theorem 4 with regard to applications is that an

* See also Ref. 3, Theorem 1, with regard to the expression for the Y_m.

† See Ref. 5, Theorem 3.17.1, and Ref. 11, p. 198. Also, notice that the convergence to $N(x)$ of the right side of (15) for $x_0 = p$ and $x \in V$ is implicit in Theorem 4, since C and D can be taken to be the zero map and A and B can be assumed to be the identity map.

expansion exists, and can be constructed as specified, when the original system equations can be "complexified" so that the hypotheses concerning V_0, V, and $(I - CN)$, and of the theorem, are met. It is not difficult to give specific examples. One example is provided by the material in Appendix B.

Theorem 4 with $N(\theta) = p = v_0 = x_0 = \theta$ bears directly on the main result in Ref. 3, which concerns expansions involving iterated integrals, for the case in which $\mathscr{B}$ is the set L_∞ of bounded complex n-vector valued functions on the interval $[0, \infty)$.[3*] Our theorem shows that the locally convergent expansions described there converge in fact for $Av \in V$ for any V and V_0 such that our solvability hypotheses with $p = \theta$ are met, assuming merely that $S = \Gamma$, where Γ is the domain of definition of N in Hypothesis B.2 of Ref. 3.

In this connection, if $\mathscr{B} = L_\infty$ and N is memoryless in the sense that $(Ns)(t) = \eta[s(t), t]$, $t \geq 0$ for $s \in S$, with $S = \Gamma$ and η and Γ, respectively, a function and a domain of the kind described in Ref. 3, Hypothesis B.2, and if, for simplicity, $n = 1$, then $[d^l N(x_0) X_{(k_1)}(v - v_0) \cdots X_{(k_l)}(v - v_0)](t)$ is just the product $Ml(t)[X_{(k_1)}(v - v_0)](t) \cdots [X_{(k_l)}(v - v_0)(t)]$ for $t \geq 0$, where $Ml(t) = \partial^l \eta(z, t)/\partial z^l \big|_{z=x_0(t)}$ (see Ref. 2, Lemma 3).

REFERENCES

1. I. W. Sandberg, "Expansions for Nonlinear Systems," B.S.T.J., *61*, No. 2 (February 1982), pp. 159–99.
2. ——, "Volterra Expansions for Time-Varying Nonlinear Systems," B.S.T.J., *61*, No. 2 (February 1982), pp. 201–25.
3. ——, "On Volterra Expansions for Time-Varying Nonlinear Systems," IEEE Trans. Circuits Syst., *CAS-30* (February 1983), pp. 61–7.
4. J. Dieudonné, *Foundations of Modern Analysis*, New York: Academic Press, 1969.
5. E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, Providence: Amer. Math. Soc. Coll. Publ. XXXI, 1957.
6. L. M. Graves, "Riemann Integration and Taylor's Theorem in General Analysis," Trans. Amer. Mathematical Soc., *29* (January 1927), pp. 163–77.
7. L. A. Liusternik and V. J. Sobolev, *Elements of Functional Analysis*, New York: Frederick Unger Publishing Co., 1961.
8. J. C. Willems, *The Analysis of Feedback Systems*, Cambridge: M.I.T. Press, 1971.
9. I. J. Maddox, *Elements of Functional Analysis*, London: Cambridge University Press, 1970.
10. T. M. Apostol, *Mathematical Analysis*, Reading: Addison-Wesley, 1957.
11. T. M. Flett, *Differential Analysis*, London: Cambridge University Press, 1980.
12. A. A. M. Saleh, "Matrix Analysis of Mildly Nonlinear Multiple-Input, Multiple-Output Systems with Memory," B.S.T.J. *61*, No. 9 (November 1982), pp. 2221–43.
13. R. J. P. de Figueiredo, "A Generalized Fock Space Framework for Nonlinear System and Signal Analysis," IEEE Trans. Circuits and Syst., *CAS-30* (September 1983), pp. 637–47.
14. M. Fliess, M. Lamnabhi, and F. Lamnabhi-Lagarrique, "An Algebraic Approach to Nonlinear Functional Expansions," IEEE Trans. Circuits and Syst., *CAS-30* (August 1983), pp. 554–70.

* For recent work related in a general sense, see, for example, Refs. 12 to 15. The complexification needed to use the results in Ref. 3 is ordinarily trivial, because of the nature of the hypotheses involved.

15. S. Boyd, L. O. Chua, and C. A. Desoer, unpublished work.
16. R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Reading, MA: Addison-Wesley, 1983.

APPENDIX A

Proof of Lemma 2

Proof: Choose $a \in V$ and, using the continuity of g, let r and M be positive numbers such that $\mathscr{D} \subset V$ and $\|g(x)\| \leq M$ for $x \in \mathscr{D}$, where $\mathscr{D} = \{x \in \mathscr{B}: \|x - a\| < r\}$. Since g is analytic on $\mathscr{D}$ in the sense of Ref. 5, Definition 3.17.2, there are continuous symmetric m-linear maps Q_m from $\mathscr{B}^m$ into $\mathscr{B}_1$, which depend on a, such that

$$g(a) + \sum_{m=1}^{\infty} (m!)^{-1} Q_m(k, \cdots, k)$$

converges absolutely to $g(a + k)$ for $\|k\| < r$ (see Ref. 5, Theorems 3.16.2, 26.3.4, 26.3.6, and 26.6.6).

With $h \in \mathscr{B}$, consider the formal series

$$\sum_{m=1}^{\infty} (m!)^{-1} m Q_m(h, \cdots, h, \cdot) \tag{16}$$

whose terms are bounded linear maps from $\mathscr{B}$ to $\mathscr{B}_1$. Choose any $\beta_1 > 1$ and let $\beta = (\beta_1/r)$. Using an analog (Ref. 5, Theorem 3.16.3) of the Cauchy bounds and the m-linearity of Q_m,

$$\|Q_m(k, \cdots, k)\| \leq M(m!)(\beta\|k\|)^m$$

for $k \in \mathscr{B}$. Since (see Ref. 16, Proposition 2.2.11)

$$\sup\{\|Q_m(k_1, \cdots, k_m)\| : \|k_1\| \leq 1, \cdots, \|k_m\| \leq 1\}$$
$$\leq m^m(m!)^{-1}\sup\{\|Q_m(k, \cdots, k)\| : \|k\| \leq 1\} \tag{17}$$

we have

$$\|Q_m(k_1, \cdots, k_m)\| \leq m^m M\beta^m \|k_1\| \cdots \|k_m\| \tag{18}$$

for any $k_1, \cdots, k_m$ in $\mathscr{B}$. By Sterling's formula for $m!$,

$$m! > (2\pi)^{1/2} m^{1/2} m^m e^{-m}. \tag{19}$$

Thus, by (18) and (19) one has for the mth term in (16),

$$\|(m!)^{-1} m Q_m(h, \cdots, h, \cdot)\|$$
$$= \sup\{\|(m!)^{-1} m Q_m(h, \cdots, h, k)\| : \|k\| \leq 1\}$$
$$\leq (2\pi)^{-1/2} m^{1/2} M(e\beta) \|e\beta h\|^{(m-1)}.$$

It easily follows that there is a positive $r_1 < r$ such that (16) converges absolutely to a bounded linear map $L(a, h)$ for $\|h\| < r_1$. Since $L(a, h)$

is an (F)-power series in h in the sense of Ref. 5, Theorem 26.64, the map $L(a, \cdot)$ is analytic and hence Fréchet differentiable (Ref. 5, Theorem 3.17.1) on $\{x \in \mathscr{B}: \|x\| < r_1\}$. In particular, $L(a, \cdot)$ is Fréchet differentiable at θ.

In addition, for $\|h\| < \alpha r_1$ and $\|\delta\| < \alpha r_1$ where $\alpha \in (0, 1/2)$,

$$\begin{aligned}
&\|g(a + h + \delta) - g(a + h) - L(a, h)\delta\| \\
&\quad \leqslant \sum_{m=2}^{\infty} \|(m!)^{-1}[Q_m(h + \delta, \cdots, h + \delta) - Q_m(h, \cdots, h) \\
&\qquad - mQ_m(h, \cdots, h, \delta)]\| \\
&\quad \leqslant \sum_{m=2}^{\infty} (m!)^{-1} 2^m \|Q_m\| (\alpha r_1)^{(m-2)} \|\delta\|^2, \qquad (20)
\end{aligned}$$

where $\|Q_m\|$ is the left side of (17). Using (18) and (19), it is a simple matter to verify that for α sufficiently small the extreme right side of (20) is $o(\|\delta\|)$ as $\|\delta\| \to 0$. This shows that $dg(a + h) = L(a, h)$ for all h in some neighborhood of θ, and thus that $dg(\cdot)$ is Fréchet differentiable at the arbitrary point $a \in V$.

APPENDIX B

An Example

Here we give a simple example of a map $(I - CN)$ and associated sets V_0 and V that meet the conditions at the beginning of Section 2.4. The construction of the example involves a contraction-mapping technique that is well known and often useful. We shall use L_∞ to denote the set of bounded complex-valued Lebesgue measurable functions defined on $[0, \infty)$ with the usual sup norm.*

Let $S = \mathscr{B} = L_\infty$, and let us take C to be an operator, such as a convolution operator, with induced norm ρ. We assume that $\rho > 0$. Let N be defined by $(Nx)(t) = x(t)^3$ for $t \geqslant 0$ and $x \in \mathscr{B}$, and take F: $\mathscr{B} \to \mathscr{B}$ to be given by $Fx = u + CNx$ for every $x \in \mathscr{B}$, where u is an element of $\mathscr{B}$.

We note that for x_0 in the closed ball $\beta(r)$ of radius r in $\mathscr{B}$ centered at θ, $dN(x_0)h(t) = 3x_0(t)^2h(t)$, $t \geqslant 0$, for arbitrary $h \in \mathscr{B}$, and therefore $\|CdN(x_0)\| \leqslant 3\rho r^2$. Thus, F is a contraction mapping on $\beta(r)$ for $0 < r < r_0$, where $r_0 = (3\rho)^{-1/2}$. Also, F takes $\beta(r)$ into itself if

$$\|u\| + \sup\{\|CNx\| : x \in \beta(r)\} \leqslant r,$$

* In other words, here L_∞ is the set L_∞ in Ref. 3 with $n = 1$. (The definition of L_∞ ordinarily found in the system-theoretic literature involves instead the essential sup norm. That our L_∞ in this paper, or in Ref. 3, is *complete* follows from the Cauchy criterion for uniform convergence and the standard proposition that pointwise limits of measurable functions are measurable.)

which is met when $\|u\| \leq q(r)$, where $q(r) = r - \rho r^3$. These observations and the monotonicity of q on $(0, r_0)$ motivate the following.

Take V to be the open ball in $\mathscr{B}$ of radius $q(r_0)$ centered at θ. Given any $u \in V$ there is an $r_1 \in (0, r_0)$ such that $\|u\| \leq q(r)$ for all $r \in [r_1, r_0)$. Hence, by the contraction-mapping theorem, for any such u and all associated r there is a unique solution $x \in \beta(r)$ of the equation $x - CNx = u$. Also, for any open ball β_a in V centered at some point a, there is an $r \in (0, r_0)$ such that $\|u\| \leq q(r)$ for $u \in \beta_a$, and it easily follows that the solutions x_1 and x_2 corresponding to any u_1 and u_2 in β_a satisfy $\|x_1 - x_2\| \leq (1 - 3\rho r^2)^{-1} \|u_1 - u_2\|$.

Thus, with V_0 the intersection of the open ball $\beta_0(r_0)$ of radius r_0 centered at θ with the inverse image of V under $(I - CN)$, we see that V_0 is open and that $(I - CN)$ is a homeomorphism of V_0 onto the open c-star V about the point θ, with $(I - CN)^{-1}$ locally Lipschitz on V.*

AUTHOR

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; AT&T Bell Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of radar systems for military defense, synthesis and analysis of active and time-varying networks, with several fundamental studies of properties of nonlinear systems, and with some problems in communication theory and numerical analysis. His more recent interests have included compartmental models, the theory of digital filtering, global implicit-function theorems, and functional expansions for nonlinear systems. IEEE Centennial Medalist, Former Vice Chairman IEEE Group on Circuit Theory, and Former Guest Editor IEEE Transactions on Circuit Theory Special Issue on Active and Digital Networks. Fellow and member, IEEE; member, American Association for the Advancement of Science, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, National Academy of Engineering.

---

* A similar example can be given that addresses the more general case in which $(Nx)(t) = \eta[x(t)]$, $t \geq 0$ ($x \in \mathscr{B}$) with η an entire function such that $\eta(0) = 0$.

AT&T Technical Journal
Vol. 64, No. 1, January 1985
Printed in U.S.A.

Effects of Channel Impairments on the Performance of an In-band Data-Driven Echo Canceler

By J. J. WERNER*

(Manuscript received July 11, 1984)

High-speed ($\geq$ 4.8 kb/s) echo-cancellation-based full-duplex direct distance dialing modems usually have to deal with two echos: the near echo, which is generated at the modem location, and the far, or talker, echo, which has been looped back to the modem after passing through a carrier system. The near echo propagates through a channel that is essentially linear, and thus it can, at least in theory, be perfectly canceled by an echo canceler. On the other hand, the far-echo channel is generally plagued by impairments that can seriously degrade the performance of an echo canceler. In this paper we study the effects of these channel impairments on the performance of an in-band data-driven echo canceler. This echo canceler has been found to be particularly well suited for full-duplex voice-grade data transmission applications. Both analytical and real-time experimental results are presented. It is shown that frequency offset, even in small amounts, is by far the most damaging of the channel impairments that are commonly encountered in carrier systems. The degradation of performance due to phase jitter can be significant. However, this can only happen under simultaneous worst-case conditions of phase jitter and signal power levels, and these cases might not be statistically significant. Worst-case nonlinearities in the echo channel do not degrade substantially the performance of the echo canceler.

I. INTRODUCTION AND SUMMARY

In this paper we study the effects of channel impairments on the performance of a data-driven in-band echo canceler. This echo can-

* AT&T Information Systems, Inc.

celer is intended to be used in two-wire, high-speed (≥ 4.8 kb/s), full-duplex, voice-grade data transmission. Details about this application for 4800 b/s Direct Distance Dialing (DDD) operation are given in Ref. 1. Performance degradations due to finite precision effects in the digital implementation of the echo canceler are studied in a companion paper.[2]

The use of echo cancellation in two-wire, full-duplex data transmission will be described with reference to Fig. 1.[3-6] Figure 1a shows a simplified version of a typical connection over the switched network. Echoes arise because of impedance mismatches in the hybrid couplers that make the connections between two-wire and four-wire transmission facilities. Consequently, some energy leaks directly through the first hybrid encountered by the transmitted signal. This signal is called the *near echo*. Similarly, some energy leaks through the hybrid at the other end of the four-wire circuit and is looped back to the modem through the carrier system. This signal is called the *talker* or *far echo*. These two echoes are added to the signal transmitted by the far-end modem, and they appear as interference at the receiver's input.

The echo canceler in Fig. 1b synthesizes a replica of the channel traversed by the echoes. It can take its input from various points of the transmitter. For example, a voice-type canceler takes its input from the output of the transmitter, whereas a data-driven canceler gets its input directly from the data symbols at the input of the transmitter. Under ideal conditions, if the canceler and the echo channel have the same inputs, they should also have the same outputs,

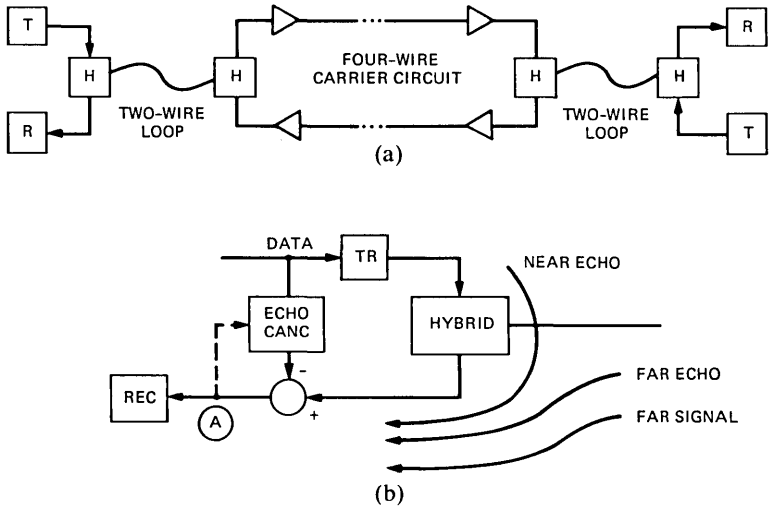


Fig. 1—(a) Typical dialed connection. (b) Use of data-driven echo canceler at station location.

and the signal, after subtraction at point A, should consist only of the wanted far-end data signal. These ideal conditions exist, to some extent, for the near-end echo because the hybrid introduces mainly linear distortion (for which the canceler can compensate perfectly if its memory span and the digital precision are large enough). However, this is not true for the far echo that has propagated through a channel that is generally time varying and nonlinear. The updating algorithms and structure used in the basic canceler can only compensate partially, if at all, for these impairments, and some residual echo will appear at point A in Fig. 1b. Some of the effects of these channel impairments for voice-type cancelers are studied in Refs. 7 through 11.

Because the near echo and the far echo have different characteristics, we will find it convenient to break the canceler into two parts, a near canceler and a far canceler. The requirements for these two cancelers are quite different. Due to the delay characteristics of the far-echo channel, the far canceler requires a larger memory span than the near canceler. However, under worst-case conditions of hybrid leakage and signal levels, the near canceler has to provide a much larger echo attenuation than the far canceler. In both cases these attenuations generally have to be achieved under "double-talking"* (full-duplex transmission) conditions and, in the case of the far canceler, in the presence of time-varying and nonlinear echo-channel impairments.

Our purpose in this paper is to study the performance degradation of the far canceler in the presence of the impairments that are most commonly encountered in carrier systems. The echo canceler used in the study is the so-called data-driven, Nyquist, in-band canceler. Its structure is derived in the next section. The experimental and analytical work presented here consisted of studying the signals at point A, in Fig. 1b, when different types of impairments were inserted in the far-echo channel. In the experimental setup, the echo channel was simulated by using the appropriate laboratory equipment, and real-time performance measurements were achieved by using an echo canceler implemented on an in-house developed digital signal processor. We now summarize our findings. Frequency offset, even in small amounts (a fraction of a hertz), is found to be the most damaging of all the impairments. Possible corrective actions are studied in a forthcoming paper. The performance of the canceler can also be significantly degraded by phase jitter. However, in practice, this can

* The signal at point A in Fig. 1b is used to adapt the echo canceler's tap coefficients. In steady-state operation this signal will consist mostly of the far signal, which will thus act as a strong noise component in the adaptation algorithm.[10] An arrangement proposed recently allows, in principle, the elimination of the double-talker from the adaptation algorithm.[12]

only happen under certain simultaneous worst-case (round-trip) conditions of jitter and signal power levels, and these cases might not be statistically significant. Worst-case nonlinearities did not degrade substantially the performance of the canceler.

The paper is organized as follows. The structure of the echo canceler used in this study is derived in the next section. Its convergence properties in the presence of a double-talker are analyzed in Section III. The effects of phase jitter and frequency offset are studied in Sections IV and V, respectively. Finally, in Section VI we present real-time performance results obtained with an echo canceler implemented on a bit-slice processor.

II. IN-BAND ECHO CANCELER STRUCTURE

The echo canceler structure derived in this section is of the Nyquist or interpolating type. That is, it cancels the echo at all frequencies. It is also called an in-band canceler because it synthesizes passband filters rather than equivalent baseband filters. This canceler is particularly attractive because it is less complex to implement than either the voice-type canceler or the Nyquist data-driven canceler described in Ref. 6.

A two-dimensional (in-phase and quadrature) modulated signal is generally represented by the expression

$$s(t) = \text{Re}\left\{\sum_n A_n g(t - nT)e^{j\omega_c t}\right\}, \tag{1}$$

where $A_n = a_n + jb_n$ is the discrete-valued multilevel complex symbol to be transmitted, $g(t)$ is a Nyquist pulse, $1/T$ is the symbol rate, and $\omega_c/2\pi$ is the carrier frequency. In the usual case where the highest frequency component in $g(t)$ is smaller than the carrier frequency, the complex signal in brackets in (1) is an analytic signal $Z(t)$, where

$$Z(t) = s(t) + j\tilde{s}(t) = \sum_n A_n g(t - nT)e^{j\omega_c t}, \tag{2}$$

and where $\tilde{s}(t)$ is the Hilbert transform of $s(t)$. Equation (2) can be rewritten as

$$Z(t) = \sum_n A_n e^{j\omega_c nT} g(t - nT)e^{j\omega_c (t-nT)} \tag{3}$$

$$Z(t) = \sum_n A'_n R(t - nT), \tag{4}$$

where

$$A'_n = A_n e^{j\omega_c nT} \tag{5}$$

$$R(t) = g(t)e^{j\omega_c t}. \tag{6}$$

When the signal $s(t)$ is transmitted through a channel with impulse response $h(t)$, the analytic signal corresponding to the output signal is

$$Z_1(t) = Z(t)^* h(t) = \sum_n A'_n R_1(t - nT), \tag{7}$$

where * denotes convolution and

$$R_1(t) = R(t)^* h(t), \tag{8}$$

and where $Z_1(t)$ and $R_1(t)$ are analytic signals. The signal at the output of the channel is the real part of $Z_1(t)$, i.e.,

$$s_1(t) = \sum_n [a'_n r_1(t - nT) - b'_n \tilde{r}_1(t - nT)], \tag{9}$$

where $R_1(t) = r_1(t) + j\tilde{r}_1(t)$.

Thus, the echo $s_1(t)$ is obtained by feeding the symbols a'_n and b'_n to in-phase and quadrature bandpass filters with impulse responses $r_1(t)$ and $\tilde{r}_1(t)$, respectively. This suggests the following structure for a digitally implemented data-driven echo canceler. We can feed the rotated symbols a'_n and b'_n to two transversal filters with variable tap coefficients and tap delay spacings of T'. The signal $s_1(t)$ in (9) is sampled at a rate $1/T'$, which will be assumed to be at least twice the highest frequency of $s_1(t)$, so that the wanted far signal, which is added to $s_1(t)$ in full-duplex operation, can be reconstructed after cancellation. A standard Mean-Squared Error (MSE) criterion can then be used to adapt the tap coefficients. Ideally, after adaptation, the tap coefficients of the two transversal filters converge to the sampled values of $r_1(t)$ and $\tilde{r}_1(t)$ in (9). The echo canceler structure is shown in Fig. 2. Notice from (4), (5), and (6) that the same structure can be used to implement the transmitter, in which case the tap coefficients are fixed.[13]

Referring again to Fig. 2, the transmitter generates the symbols a'_n and b'_n at the symbol rate $1/T$, which is always smaller than the sampling rate $1/T'$. Therefore the canceler's delay lines in Fig. 2 will only be sparsely filled. In fact, they will contain $L - 1$ zeros for each

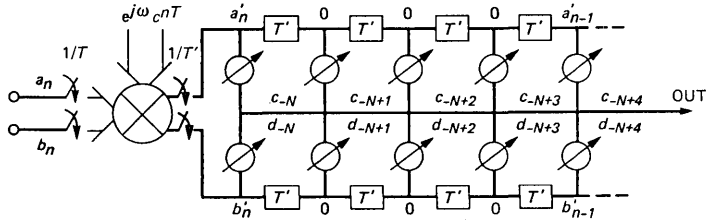
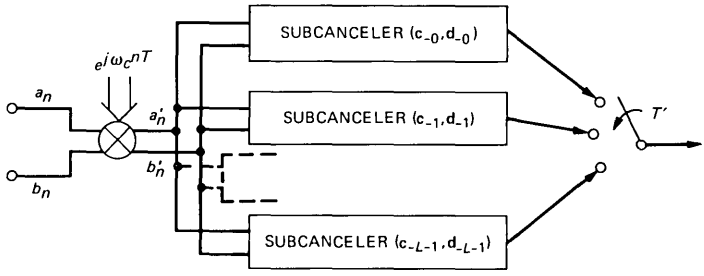
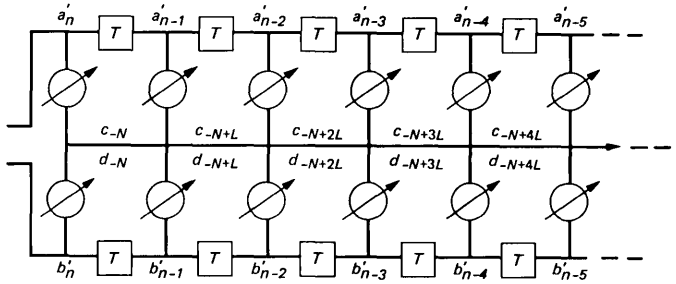


Fig. 2—In-band data-driven echo canceler structure.



(a)



(b)

Fig. 3—(a) Modified canceler structure. (b) Subcanceler structure.

nonzero symbol, where $L = T/T'$ is the number of sampling periods per symbol interval.* A direct implementation of this canceler structure would waste computation power since multiplication by zero need not be performed. A more efficient, but equivalent, structure can be derived by observing that a different subset of coefficients is used for the computation of the output when the symbols move down the delay line in Fig. 2. There are L such subsets that are each used once per symbol interval. All the subsets of coefficients are correlated with the same input vectors $(a'_n, a'_{n-1}, \cdots)$ and $(b'_n, b'_{n-1}, \cdots)$ in a given symbol period. Therefore the canceler can be considered as a parallel combination of L "subcancelers" having all the same delay line but different sets of coefficients, as shown in Fig. 3a. One of the subcancelers is shown in Fig. 3b. The taps of the transversal filters are now spaced at T rather than T', and none of the entries in the delay lines are zero. The ith subcanceler consists of the tap vectors $\mathbf{c}_i$ and $\mathbf{d}_i$ whose elements

* Theoretically many other interpolation schemes can be envisioned. However, the scheme described here leads to the simplest possible practical implementation of the echo canceler.

are $c_{-N+i+kL}$ and $d_{-N+i+kL}$, respectively, where $i = 0 \cdots L - 1$. The outputs of the canceler are generated at the sampling rate $1/T'$ by computing, in a cyclic fashion, the outputs of the L subcancelers shown in Fig. 3a. This model applies also to the echo channel and we can consider that it consists of a parallel combination of L subchannels. Convergence of the whole canceler can then be achieved by having each subcanceler converge to the corresponding subchannel. The algorithms used during adaptation are described in the next section.

III. CONVERGENCE PROPERTIES IN THE PRESENCE OF A DOUBLE-TALKER

In order to adapt the tap coefficients of the canceler, we will separately minimize the MSE between the outputs of each subcanceler and the corresponding echo subchannel. That is, each subcanceler's output is computed once per symbol interval, an error is derived, and a stochastic-gradient algorithm is used to update the tap coefficients. Nyquist cancellation is obtained by cyclicly repeating these operations at the sampling rate $1/T'$ for all the subcancelers. We analyze this adaptation scheme by assuming that the subcancelers adapt independently. This assumption was found to be in excellent agreement with our experimental results. For simplicity of notation, we will also assume that the inputs of the echo canceler are the original symbols a_n and b_n rather than the rotated symbols a'_n and b'_n. We now define the following quantities, at the nth symbol instant:

$\mathbf{a}_n^T = [a_n, a_{n-1}, a_{n-2}, \cdots]$ = in-phase data vector.

$\mathbf{b}_n^T = [b_n, b_{n-1}, b_{n-2}, \cdots]$ = quadrature data vector.

$\mathbf{c}_i^T = [c_{-N+i}, c_{-N+i+L}, c_{-N+i+2L}, \cdots]$

= vector of in-phase tap coefficients of ith subcanceler.

$\mathbf{d}_i^T = [d_{-N+i}, d_{-N+i+L}, d_{-N+i+2L}, \cdots]$

= vector of quadrature tap coefficients of ith subcanceler.

$\mathbf{r}_{1i}^T = [r_1(iT'), r_1(T + iT'), r_1(2T + iT'), \cdots]$

= vector of in-phase samples of ith echo subchannel.

$\mathbf{r}_{2i}^T = [r_2(iT'), r_2(T + iT'), r_2(2T + iT'), \cdots]$

= vector of quadrature samples of ith echo subchannel.

The superscript T designates transposed vectors where all these vectors are assumed to be infinite. For finite-length cancelers shorter than the echo-channel impulse response, we simply insert zeros on the right in the definitions of $\mathbf{c}_i^T$ and $\mathbf{d}_i^T$. The outputs, $s_e(nT + iT')$ and

$s_c(nT + iT')$, of the ith subchannel and the ith subcanceler at time $nT + iT'$ are

$$s_e(nT + iT') = \mathbf{a}_n^T \mathbf{r}_{1i} + \mathbf{b}_n^T \mathbf{r}_{2i} + \xi_{n,i} \tag{10}$$

$$s_c(nT + iT') = \mathbf{a}_n^T \mathbf{c}_i + \mathbf{b}_n^T \mathbf{d}_i, \tag{11}$$

where $\xi_{n,i}$ is an additive interference signal that is uncorrelated with the signal to be canceled. This interference will generally consist of the desired (far-end) data signal and some additive noise. The error $e_{n,i}$ between the outputs of the ith subchannel and the ith subcanceler is

$$e_{n,i} = s_e(nT + iT') - s_c(nT + iT'), \qquad i = 0, 1, \cdots, L - 1, \tag{12}$$

and we want to minimize the MSE

$$\begin{aligned} E_i &= \langle (e_{n,i})^2 \rangle = \langle [s_e(nT + iT') - s_c(nT + iT')]^2 \rangle \\ &= \langle [\mathbf{a}_n^T(\mathbf{r}_{1i} - \mathbf{c}_i) + \mathbf{b}_n^T(\mathbf{r}_{2i} - \mathbf{d}_i) + \xi_{n,i}]^2 \rangle, \end{aligned} \tag{13}$$

where $\langle \cdot \rangle$ denotes the expectation of the quantity inside the brackets. Note that E_i cannot be smaller than the irreducible noise $\langle \xi_{n,i}^2 \rangle$.

The Minimum MSE (MMSE) is achieved when the N_i complex tap coefficients of the ith subcanceler are equal to the corresponding subchannel complex sampled values. The MMSE is given by

$$\min E_i = A \sum_{k>N_i} [r_1^2(kT + iT') + r_2^2(kT + iT')] + \langle \xi_i^2 \rangle, \tag{14}$$

where the data-symbol power is given by $A = \langle a^2 \rangle = \langle b^2 \rangle$.

This MMSE is obviously not the same for all the subcancelers. The adjustment algorithms for the updating of the subcanceler tap coefficients are obtained by taking the gradient of the MSE in (13) with respect to the tap vectors $\mathbf{c}_i$ and $\mathbf{d}_i$

$$\frac{\partial E_i}{\partial \mathbf{c}_i} = -2\langle \mathbf{a}_n e_{n,i} \rangle = \mathbf{r}_{1i} - \mathbf{c}_i \tag{15}$$

$$\frac{\partial E_i}{\partial \mathbf{d}_i} = -2\langle \mathbf{b}_n e_{n,i} \rangle = \mathbf{r}_{2i} - \mathbf{d}_i. \tag{16}$$

As is usual in practice, the gradients, with respect to the squared error rather than the MSE, are used for the adjustment of the tap coefficients. The corresponding stochastic tap adjustment algorithms are then

$$\mathbf{c}_{n+1,i} = \mathbf{c}_{n,i} + \alpha \mathbf{a}_n e_{n,i} \tag{17}$$

$$\mathbf{d}_{n+1,i} = \mathbf{d}_{n,i} + \alpha \mathbf{b}_n e_{n,i}, \tag{18}$$

where α is the step size of the adjustments and $i = 0, 1, \cdots, L - 1$.

Inspection of (11), (17), and (18) shows that each subcanceler requires a total of about $4N$ multiplications and additions for the filtering and updating operations. Therefore the implementation of the whole canceler requires $4LN$ multiplications and additions per symbol period. An analysis similar to those given in Refs. 4, 6, and 14 can be used to study the convergence properties of the MSE as a function of time. As is usually the case, the analysis assumes that the data vectors $\mathbf{a}_n$ and $\mathbf{b}_n$ in the subcancelers are uncorrelated between successive tap adjustments. Under these conditions it is shown in Appendix A that the MSE decreases as

$$\langle e_{n,i}^2 \rangle = (1 - 2\alpha A + 2\alpha^2 N_i A^2)^n \langle e_{0,i}^2 \rangle + \frac{1 - (1 - 2\alpha A + 2\alpha^2 N_i A^2)^n}{1 - (1 - 2\alpha A + 2\alpha^2 N_i A^2)} \cdot 2\alpha A \langle \xi_{n,i}^2 \rangle, \tag{19}$$

where N_i is the number of complex taps in the ith subcanceler.

In the derivation of (19) it was assumed that the first term on the right in (14) was zero. That is, it was assumed that the canceler was long enough to cover the memory span of the echo channel, and that no degradation was introduced by the finite precision in the digital implementation. Effects of finite precision are studied in detail in Ref. 2.

For the expression in (19) to converge, we require

$$|1 - 2\alpha A + 2\alpha^2 N_i A^2| < 1 \tag{20}$$

so that the step size α has to satisfy

$$0 < \alpha < \frac{1}{N_i A}. \tag{21}$$

The step size that provides the fastest speed of convergence to the corresponding steady-state MSE is obtained by setting the derivative of the expression in (20) to zero, i.e.,

$$\alpha_{\text{opt}} = \frac{1}{2N_i A}, \tag{22}$$

or one-half the maximum step size. From (19) it is clear that the ith subcanceler's steady-state MSE for a given step size is given by

$$\langle e_{\infty,i}^2 \rangle = \frac{\langle \xi_i^2 \rangle}{1 - \alpha N_i A}, \tag{23}$$

and the overall steady-state MSE averaged over a symbol period is

$$\langle e_\infty^2 \rangle = \frac{1}{L} \sum_{i=0}^{L-1} \frac{\langle \xi_i^2 \rangle}{1 - \alpha N_i A}. \tag{24}$$

If we assume that all the subcancelers have the same number N of complex taps and if we define

$$\langle \xi^2 \rangle = \frac{1}{L} \sum_{i=0}^{L-1} \langle \xi_i^2 \rangle \tag{25}$$

as the average power of the interfering signal in a symbol period, we can rewrite (24) in the form

$$E \equiv \langle e_\infty^2 \rangle = \frac{\langle \xi^2 \rangle}{1 - \alpha NA}. \tag{26}$$

The signal-to-noise ratio (s/n) achievable in front of the receiver can be derived from (26). The uncorrelated term $\langle \xi^2 \rangle$ consists of the noise in the channel with power P_u and the far signal with power P_S. The residual MSE E is the sum of P_S and some interfering "noise", I, defined by

$$I = E - P_S = \frac{\alpha NAP_S + P_u}{1 - \alpha NA}. \tag{27}$$

The steady-state received signal-to-noise ratio is then

$$s/n \equiv \frac{P_S}{I} = \frac{1 - \alpha NA}{\alpha NA + P_u/P_s}. \tag{28}$$

Notice from (22) and (28) that the maximum achievable s/n is 0 dB when the optimum step size for speed of convergence is used, and the channel noise is assumed to be zero ($P_u = 0$). Although the noise in (22) through (28) is not Gaussian, this expression has proved to be very useful for predicting the echo canceler's performance for various design parameters. Notice that, in the absence of channel impairments other than noise, the s/n is *not* a function of the relative powers of the echo and the far signal before cancellation. This result, of course, is not valid in a finite precision environment.

IV. EFFECT OF PHASE JITTER AND DOUBLE-TALKING

The distant echo, which has propagated through carrier systems, is likely to exhibit phase jitter. A Quadrature Amplitude Modulation (QAM) signal with phase jitter is usually defined as the real part of

$$Z_0(t) = \sum_n A_n G_1(t - nT) e^{j(\omega_c t + \Phi(t))}, \tag{29}$$

where $\Phi(t)$ is the phase jitter.

This complex signal is generally not an analytic signal; however, for small enough $\Phi(t)$, it is a very good approximation to an analytic signal. If we assume $|\Phi(t)| \ll 1$, we can write

$$Z_0(t) = Z_1(t) e^{j\Phi(t)} \cong Z_1(t) \cdot [1 + j\Phi(t)], \tag{30}$$

where $Z_1(t)$ is the analytic signal of the jitter-free, but possibly otherwise distorted, QAM signal. Taking the real part of (33), we get

$$s_0(t) = s_1(t) - \tilde{s}_1(t) \cdot \Phi(t). \tag{31}$$

It can be easily shown that $s_1(t)$ and $\tilde{s}_1(t)$ are uncorrelated. Therefore the signal $\tilde{s}_1(t) \cdot \Phi(t)$ can be considered as uncorrelated noise that is added to the jitter-free QAM signal $s_1(t)$. The phase jitter $\Phi(t)$ is usually slowly time varying compared to the symbol rate, so that we can assume that it remains constant in a given symbol period, i.e.,

$$\Phi(nT + iT') \cong \Phi_n = \text{constant for given } n \text{ and } i = 0, i, (L-1).$$

Sampling the output signal (34) at time $nT + iT'$, squaring, and taking the average, we get

$$\langle s_0^2(nT + iT') \rangle = \langle s_1^2(nT + iT') \rangle + \langle \tilde{s}_1^2(nT + iT') \rangle \cdot \langle \Phi_n^2 \rangle. \tag{32}$$

If we define the average powers in a symbol period

$$P_0 \triangleq \frac{1}{L} \sum_{i=0}^{L-1} \langle s_0^2(nT + iT') \rangle \tag{33}$$

$$P_1 \triangleq \frac{1}{L} \sum_{i=0}^{L-1} \langle s_1^2(nT + iT') \rangle = \frac{1}{L} \sum_{i=0}^{L-1} \langle \tilde{s}_1^2(nT + iT') \rangle, \tag{34}$$

we can rewrite (36) as

$$P_0 = P_1(1 + \langle \Phi_n^2 \rangle). \tag{35}$$

Thus the power, P_0, of the far echo is generally time varying due to the term $\langle \Phi_n^2 \rangle$. However, it can be shown that the power, P_1, of the jitter-free echo is a constant provided that $L \geq 2$. The expression, (26), for the residual MSE after convergence of the canceler is repeated here as

$$E = \langle e_\infty^2 \rangle = \frac{\langle \xi^2 \rangle}{1 - \alpha NA}, \tag{36}$$

where $\langle \xi^2 \rangle$ is the average power in the uncorrelated interference signal. One of the components of $\langle \xi^2 \rangle$ is the quantity $P_1 \cdot \langle \Phi_n^2 \rangle$, which is due to the phase jitter, and another component is the desired information-bearing signal received from the other modem. The average power of this later signal, defined as in (33), will be denoted by P_s. There is, of course, the ever-present additive Gaussian noise, and there are some other uncorrelated components that will generally depend on the specific architecture of the system. The average power of all these signals will be denoted by P_u. Equation (36) can now be written as

$$E = \frac{P_s + P_1 \cdot \langle \Phi_n^2 \rangle + P_u}{1 - \alpha NA}. \tag{37}$$

This is the input power seen by the receiver after echo cancellation has taken place. The useful power for the receiver is P_s and the rest of the power is interfering "noise", I, defined by

$$I = E - P_s = \frac{\alpha NAP_s + P_1 \cdot \langle \Phi_n^2 \rangle + P_u}{1 - \alpha NA}. \tag{38}$$

The steady-state received s/n in a given symbol period is then

$$s/n = \frac{P_s}{I} = \frac{1 - \alpha NA}{\alpha NA + (P_1/P_s) \cdot \langle \Phi_n^2 \rangle + P_u/P_s}. \tag{39}$$

Notice that the s/n is now a function of the echo-to-signal ratio before cancellation. The phase jitter, $\Phi(t)$, is usually modeled as a slowly varying sine wave. As a consequence, the expression for the s/n is also time varying, and in order to compute the worst-case s/n, we have to use the largest value of $\langle \Phi_n^2 \rangle$ in (39). Assume that $\Phi(t)$ is a simple sine wave

$$\Phi(t) = C \cos 2\pi f_0 t. \tag{40}$$

If B is the peak-to-peak phase jitter in degrees, then the maximum value of $\langle \Phi_n^2 \rangle$ is

$$\max \langle \Phi_n^2 \rangle = C^2 = \left(\frac{\pi}{360} B \right)^2. \tag{41}$$

This expression does not depend on frequency and can be used even if the jitter cannot be modeled as a sine wave. However, if we want to check experimental results against analytical results, we have to use the time average of $\langle \Phi_n^2 \rangle$, since laboratory equipment will generally average power measurements over long periods of time. In this case the quantity

$$\langle \Phi_n^2 \rangle = \frac{C^2}{2} = \frac{1}{2} \left(\frac{\pi}{360} B \right)^2 \tag{42}$$

should be used in (39).

V. EFFECT OF FREQUENCY OFFSET

The analytic signal of a QAM signal affected by frequency offset can be represented by

$$Z_0(t) = \sum_n A_n G_1(t - nT) e^{j(\omega_c t + \omega_1 t)}, \tag{43}$$

where ω_1 is the radian frequency offset. Frequency offset can be considered as a special case of phase-jitter if we replace $\Phi(t)$ in (29) by $\omega_1 t$. However, the expansion (30) does not hold anymore, even for a very small ω_1. In order to study the effect of frequency offset on the

in-band canceler we can use the analysis given in Section III and rewrite (43) in the following way:

$$Z_0(t) = \sum_n A'_n R_1(t - nT), \tag{44}$$

where, if we assume $\omega_c T = k2\pi$,

$$A'_n = A_n e^{j\omega_1 nT} \tag{45}$$

and $R_1(t)$ consists of in-phase and quadrature bandpass filters. The sampled output of the ith subchannel is, from (12),

$$s_e(nT + iT') = \mathbf{a}'^T_n \mathbf{r}_{1i} + \mathbf{b}'^T_n \mathbf{r}_{2i}, \tag{46}$$

where we assume, for the time being, that $\xi_{n,i} = 0$.

If we define

$$\mathbf{A}_n = \mathbf{a}_n + \mathbf{j}\mathbf{b}_n \quad \text{and} \quad \mathbf{R}_{1i} = \mathbf{r}_{1i} + \mathbf{j}\mathbf{r}_{2i}, \tag{47}$$

we can rewrite (46) in the two following ways:

$$\begin{aligned} 2s_e(nT + iT') &= (e^{jn\Delta}\mathbf{A}^T_n)\cdot\mathbf{R}^*_{1i} + \mathbf{R}^T_{1i}\cdot(e^{jn\Delta}\mathbf{A}_n)^* \\ &= \mathbf{A}^T_n\cdot(e^{-jn\Delta}\mathbf{R}_{1i})^* + (e^{-jn\Delta}\mathbf{R}^T_{1i})\cdot\mathbf{A}^*_n, \end{aligned} \tag{48}$$

where we have defined $\omega_1 T = \Delta$, and * denotes the complex conjugate.

The two expressions in (48) lead to two different interpretations of the effect of frequency offset on a QAM signal. First we can consider that the symbol vectors $\mathbf{A}_n$ are rotated by small increments Δ at the input of a time-invariant channel. Alternatively we can assume that the vectors $\mathbf{A}_n$ are the channel's inputs and that the sampled channel's impulse responses $\mathbf{R}_{1i}$ are rotated by increments—Δ. The canceler's performance in the presence of frequency offset is studied in Appendix B, where we have assumed that the nonrotated symbols $\mathbf{A}_n$ are the inputs of both the canceler and the channel, and that the algorithms given in Section III are used for updating the tap coefficients. Under these conditions it is shown that the mean steady-state tap coefficients satisfy

$$\langle \mathbf{C}_{n,i} \rangle = \mathbf{R}_{1i} e^{-jn\Delta} \frac{\alpha A e^{j\Delta}}{1 - (1 - \alpha A)e^{j\Delta}}. \tag{49}$$

The quantity on the right is equal to the ith subchannel's complex impulse response at time $(nT + iT')$ multiplied by a constant complex number. Therefore, the canceler's complex taps rotate and track the channel's complex impulse response with a fixed phase lag and a different amplitude. It is interesting to note that the mean-tap values go to zero when the step size α goes to zero. This most peculiar behavior was also observed experimentally. The ith subcanceler's steady-state MSE after cancellation is derived in Appendix B. It is given by

$$\langle e_{n,i}^2 \rangle = \langle e_{n,i}^2(\Delta = 0) \rangle + \frac{1 - \alpha A}{1 - \alpha NA} \cdot \frac{A\Delta^2}{\alpha^2 A^2 + (1 - \alpha A)\Delta^2} \cdot \mathbf{R}_{1i}^T \mathbf{R}_{1i}^*, \quad (50)$$

where $\langle e_{n,i}^2(\Delta = 0) \rangle$ is the MSE in the absence of frequency offset.

The MSE for the whole canceler is obtained by averaging (50) over all the subcancelers. The resulting quantity is not very useful, because it is dependent on the channel characteristics. This is not the case for the Echo-Return-Loss Enhancement (ERLE), which is also studied in Appendix B. The ERLE is defined as the ratio between the power of the uncanceled echo and the power of the residual echo. It is given by

$$\text{ERLE} = -10 \log \left(\frac{1 - \alpha A}{1 - \alpha NA} \cdot \frac{\Delta^2}{\alpha^2 A^2 + (1 - \alpha A)\Delta^2} \right). \quad (51)$$

This expression is the same for all the subcancelers, and therefore it gives also the ERLE for the whole canceler. In the derivation of (51) it was assumed that there was no interfering, uncorrelated signal at the output of the channel.

VI. EXPERIMENTAL RESULTS

Experimental results were obtained by using an in-band data-driven echo canceler implemented on a bit-slice processor. Figure 4 shows

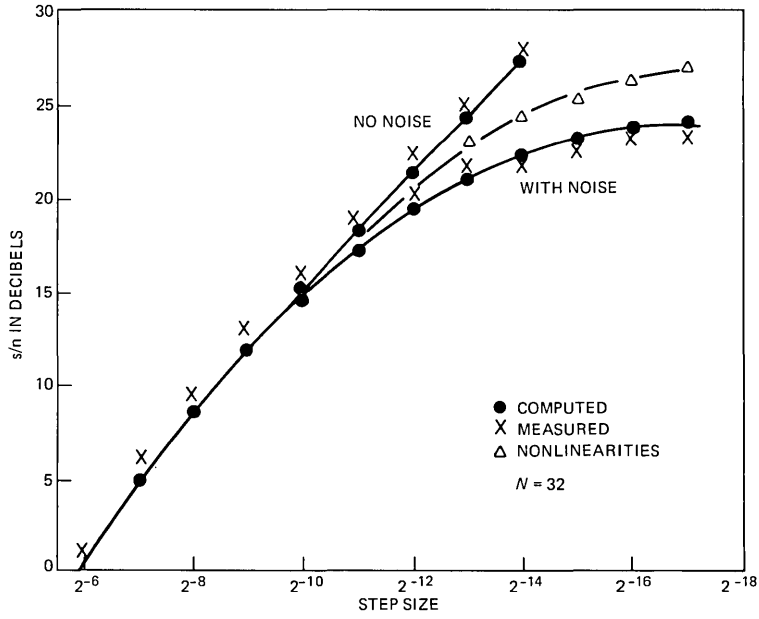


Fig. 4—The s/n in the presence of noise or nonlinearities.

performance curves giving the achievable s/n as a function of the step size for various types of channel impairments. A simulated far signal was added to the echo before cancellation, and the s/n was measured after cancellation (point A in Fig. 1b). The echo and the far signal had the same power before cancellation. The straight curve corresponds to the case where no impairments, other than linear distortion, were present in the echo channel. Notice that the achievable s/n, under these ideal conditions, increases by 3 dB when the step size decreases by a factor of two. One of the two curves that flatten out was obtained by adding noise to the echo and the far signal. In this case the noise power was 24 dB below the echo power. As one might expect, this curve goes asymptotically to an s/n of 24 dB when the step size becomes small. In both the preceding cases the theoretical curves were computed by using (28). The third curve was obtained by introducing nonlinearities in the echo channel. The amount of nonlinearities used in this experiment corresponded to the worst case reported in the 1969–70 DDD connection survey.[15]

Figure 5 shows similar results when phase jitter was introduced in the echo channel. The theoretical curves were obtained by using (39) and (42). The frequency of the phase jitter used in the experiments was 120 Hz, and the far signal and far echo had again the same power. If the far echo's power were X dB below the power of the far signal, then the flat portion of the curves would move up by X dB, as seen

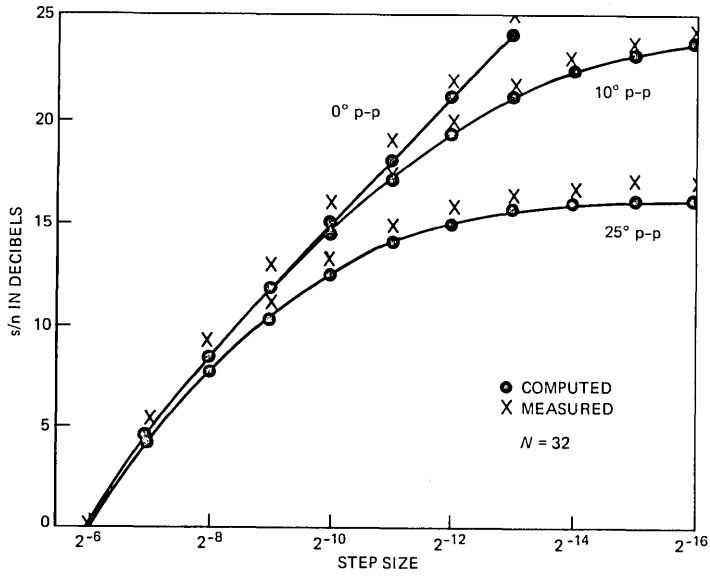
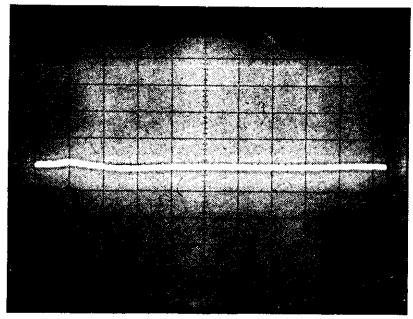


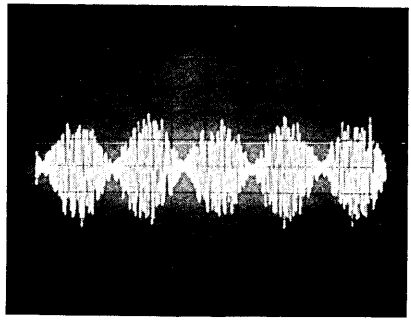
Fig. 5—The s/n degradation due to phase jitter.

from (39). A word of caution is in order here about the usage of all the results discussed so far for predicting receiver performance. The noise in the s/n is not Gaussian, and, thus, standard probability of error formulas assuming Gaussian noise should not be used in a naive way. However, we found that the preceding results were very good indicators for predicting relative receiver performance for various design parameters and channel impairments. One has also to consider that the residual echo is amplitude modulated by the phase jitter when this impairment is present in the echo channel, as seen from (31) and Fig. 6. The probability of error, in this case will be mostly influenced by the maxima of the residual echo, rather than by its average power.

Figures 7 and 8 show experimental results obtained when frequency offset was present in the echo channel. The Echo-Return-Loss Enhancement (ERLE) shown in Fig. 7 is defined as the ratio between the echo power before cancellation and the power of the residual echo after cancellation. No double-talker was present in these experiments.



(a)



(b)

Fig. 6—Residual mean-squared error (a) without phase jitter and (b) in the presence of phase jitter.

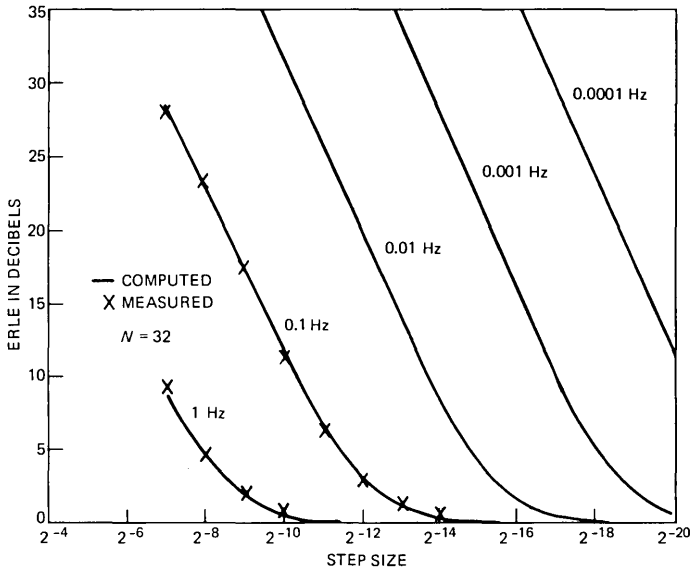


Fig. 7—ERLE degradation due to frequency offset.

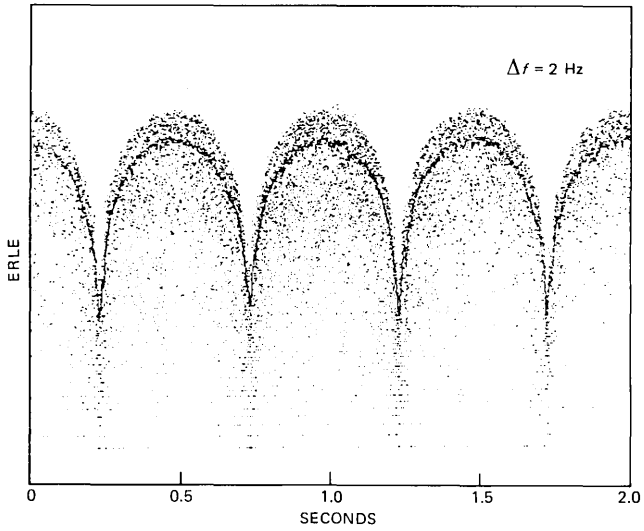


Fig. 8—Evolution of the ERLE in the presence of frequency offset and with frozen taps.

The theoretical curves were computed by using (51). The agreement between the experimental and analytical results is seen to be excellent. Notice the intuitively satisfying result that the tracking capability of the echo canceler improves with large step sizes. Figure 7 was obtained

by conducting the following experiment. The echo canceler was first converged when no frequency offset was present in the echo channel. The canceler's taps were then frozen and a frequency offset of 2 Hz was introduced in the channel. Notice the periodic evolution of the ERLE. As was mentioned in the preceding section, the effect of frequency offset can be modeled by a periodic rotation of the sampled values of the in-phase and quadrature impulse responses of the echo channel. Thus, the ERLE will pass through a minimum when these sampled values are in phase with the complex tap coefficients of the echo canceler. It will pass through a maximum when the sampled impulse responses and the complex tap coefficients are out of phase by 180 degrees.

We conclude this section with some brief comments on the implications of the preceding results on the practical implementation of high-speed, full-duplex DDD modems. The effects of noise, nonlinearities, and phase jitter are very similar, as seen in Figs. 4, 5, and 7. In each case a comfortable s/n can be achieved by using a small enough step size in the tap adjustment algorithm. This is certainly true for the 4800 b/s modem described in Ref. 1 that requires an s/n of about 14 dB in order to achieve a bit-error rate of 10^{-5}. It is usually assumed that the far echo is always at least 10 dB below the far signal. Thus, under these worst-case conditions, the flat portions of the curves in Fig. 5 would all move up by 10 dB. In this case, even a 25-degree peak-to-peak phase jitter in the far echo would not deteriorate significantly the canceler's performance, provided that the step size is chosen small enough. On the other hand, a small step size will limit the echo canceler's capability of tracking frequency offset in the far echo, as shown in Fig. 7. A frequency offset as small as 0.01 Hz can seriously degrade the echo canceler's performance. It is generally accepted that the United States' domestic network does not introduce frequency offset in the far echo. However, this is not necessarily true in other countries. Frequency offset compensation techniques are known, but they are quite expensive to implement.[4,11]

REFERENCES

1. J. J. Werner, "An Echo-Cancellation-Based 4800 bps Full-Duplex DDD Modem," IEEE J. Selected Areas Commun., *SAC-2*, No. 5 (September 1984), pp. 722–30.
2. J. M. Cioffi and J. J. Werner, "The Effect of Biases on Digitally Implemented Data-Driven Echo Cancelers," AT&T Tech. J., this issue.
3. V. G. Koll and S. B. Weinstein, "Simultaneous Two-Way Data Transmission Over a Two-Wire Circuit," IEEE Trans. Commun., *COM-21*, No. 2 (February 1973), pp. 143–7.
4. K. H. Mueller, "A New Digital Echo Canceller for Two-Wire Full-Duplex Data Transmission," IEEE Trans. Commun., *COM-24*, No. 9 (September 1976), pp. 956–67.
5. D. D. Falconer, K. H. Mueller, and S. B. Weinstein, "Echo Cancellation Techniques for Full-Duplex Data Transmission on Two-Wire Lines," Proc. NTC, Dallas, December 1976.

6. S. B. Weinstein, "A Passband Data-Driven Echo Canceller for Full-Duplex Transmission on Two-Wire Circuits," IEEE Trans. Commun. (July 1977), pp. 654–66.
7. J. R. Rosenberger and E. J. Thomas, "Performance of an Adaptive Echo Canceller in a Noisy, Linear, Time Invariant Environment," B.S.T.J., *50*, No. 3 (March 1971), pp. 785–813.
8. E. J. Thomas, "An Adaptive Echo Canceller in a Non-Ideal Environment (Nonlinear or Time Variant)," B.S.T.J., *50*, No. 8 (October 1971), pp. 2779–95.
9. H. G. Suyderhoud and M. Onufry, "Performance of a Digital Adaptive Echo Canceller in a Simulated Satellite Circuit Environment," AIAA 4th Commun. Satellite Syst. Conf., Washington, April 24–26, 1972.
10. R. D. Gitlin and S. B. Weinstein, "The Effects of Large Interference on the Tracking Capability of Digitally Implemented Echo Cancellers," IEEE Trans. Commun., *COM-26*, No. 6 (June 1978), pp. 833–9.
11. R. D. Gitlin and J. S. Thompson, "A Phase Adaptive Structure for Echo Cancellation," IEEE Trans. Commun., *COM-26*, No. 8 (August 1978), pp. 1211–20.
12. D. D. Falconer, "Adaptive Reference Echo Cancellation," IEEE Trans. Commun. (September 1982), pp. 2083–94.
13. J. J. Werner, "Modulated Passband Signal Generator," U.S. Patent No. 4,015,222, March 29, 1977.
14. N. A. M. Verhoecks et al., "Digital Echo Cancellation for Baseband Data Transmission," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27*, No. 6 (December 1979), pp. 768–81.
15. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1311–48.

APPENDIX A

Convergence of the MSE

For notational convenience we will delete the index i in the equations but it will be understood that the analysis applies to a subcanceler and not the whole canceler. The updating algorithms are, from (19) and (20),

$$\mathbf{c}_{n+1} = \mathbf{c}_n + \alpha \mathbf{a}_n e_n \tag{52}$$

$$\mathbf{d}_{n+1} = \mathbf{d}_n + \alpha \mathbf{b}_n e_n. \tag{53}$$

If we define the complex tap weights and data symbols

$$\mathbf{C}_n \triangleq \mathbf{c}_n + j\mathbf{d}_n \tag{54}$$

$$\mathbf{A}_n \triangleq \mathbf{a}_n + j\mathbf{b}_n, \tag{55}$$

we can rewrite (52) and (53) in the compact form

$$\mathbf{C}_{n+1} = \mathbf{C}_n + \alpha \mathbf{A}_n e_n. \tag{56}$$

Subtracting both sides of (56) from the complex vector

$$\mathbf{R} = \mathbf{r}_1 + j\mathbf{r}_2, \tag{57}$$

we get the impulse-response error signal

$$\epsilon_{n+1} = \epsilon_n - \alpha \mathbf{A}_n e_n + \mathbf{F}_n, \tag{58}$$

where we have defined the canceler tap errors as

$$\epsilon_n \triangleq \mathbf{R} - \mathbf{C}_n = [\mathbf{r}_1 - \mathbf{c}_n + j(\mathbf{r}_2 - \mathbf{d}_n)]. \tag{59}$$

In (58) we have also included a complex vector $\mathbf{F}_n$, which will be needed in Appendix B, to account for frequency offset. In this Appendix this quantity is taken to be zero.

The canceler error at the nth iteration is, from (13),

$$e_n = (\mathbf{r}_1^T - \mathbf{c}_n^T)\mathbf{a}_n + (\mathbf{r}_2^T - \mathbf{d}_n^T)\mathbf{b}_n + \xi_n \tag{60}$$

$$= \frac{1}{2}(\boldsymbol{\epsilon}_n^T \mathbf{A}_n^* + \mathbf{A}_n^T \boldsymbol{\epsilon}_n^*) + \xi_n, \tag{61}$$

where the asterisk * denotes the complex conjugate. The MSE at the nth iteration becomes

$$\langle e_n^2 \rangle = \frac{1}{4}\langle (\boldsymbol{\epsilon}_n^T \mathbf{A}_n^* + \mathbf{A}_n^T \boldsymbol{\epsilon}_n^*)^2 \rangle + \langle \xi_n^2 \rangle. \tag{62}$$

The tap error vectors $\boldsymbol{\epsilon}_n$ depend only on the vectors $\mathbf{A}_{n-1}$ as shown in (58). If we assume that consecutive data vectors $\mathbf{A}_{n-1}$ and $\mathbf{A}_n$ are uncorrelated, and that the $a_n's$ are uncorrelated with the $b_n's$, we can rewrite (62) as

$$\langle e_n^2 \rangle = A\langle \boldsymbol{\epsilon}_n^T \boldsymbol{\epsilon}_n^* \rangle + \langle \xi_n^2 \rangle, \tag{63}$$

where $A = \langle a_n^2 \rangle = \langle b_n^2 \rangle$ is the variance in the symbols. The MSE at the $(n + 1)$th iteration is

$$\langle e_{n+1}^2 \rangle = A\langle \boldsymbol{\epsilon}_{n+1}^T \boldsymbol{\epsilon}_{n+1}^* \rangle + \langle \xi^2 \rangle. \tag{64}$$

Using (58), we get

$$\langle e_{n+1}^2 \rangle = A\langle (\boldsymbol{\epsilon}_n^T + \mathbf{F}_n^T - \alpha e_n \mathbf{A}_n^T) \cdot (\boldsymbol{\epsilon}_n^* + \mathbf{F}_n^* - \alpha e_n \mathbf{A}_n^*) \rangle + \langle \xi^2 \rangle \tag{65}$$

$$= A\langle \boldsymbol{\epsilon}_n^T \boldsymbol{\epsilon}_n^* \rangle - \alpha A\langle e_n(\mathbf{A}_n^T \boldsymbol{\epsilon}_n^* + \boldsymbol{\epsilon}_n^T \mathbf{A}_n^*) \rangle + 2\alpha^2 A^2 N\langle \boldsymbol{\epsilon}_n^2 \rangle + Af(\mathbf{F}_n), \tag{66}$$

where we have used

$$\mathbf{A}_n^T \mathbf{A}_n^* = 2AN, \tag{67}$$

and have defined

$$f(\mathbf{F}_n) = \langle \mathbf{F}_n^T(\boldsymbol{\epsilon}_n^* - \alpha e_n \mathbf{A}_n^*) + (\boldsymbol{\epsilon}_n^T - \alpha e_n \mathbf{A}_n^T)\mathbf{F}_n^* \rangle. \tag{68}$$

Setting $\mathbf{F}_n$ to zero and using (61) and (63) in (66), we get the equation

$$\langle e_{n+1}^2 \rangle = (1 - 2\alpha A + 2\alpha^2 N A^2) \cdot \langle e_n^2 \rangle + 2\alpha A\langle \xi^2 \rangle. \tag{69}$$

The solution to this recurrence equation is

$$\langle e_n^2 \rangle = (1 - 2\alpha A + 2\alpha^2 N A^2)^n \langle e_0^2 \rangle + \frac{1 - (1 - 2\alpha A + 2\alpha^2 N A^2)^n}{1 - (1 - 2\alpha A + 2\alpha^2 N A^2)} \cdot 2\alpha A\langle \xi^2 \rangle. \tag{70}$$

APPENDIX B

Effect of Frequency Offset

It was shown in Section VI that the effect of frequency offset can be accounted for by rotating the channel's sampled impulse responses at the symbol rate by angles $\Delta = \omega_1 T$, where ω_1 is the radian frequency offset. We can then rewrite (56) in the following way:

$$\mathbf{R}_{n+1} - \mathbf{C}_{n+1} = \mathbf{R}_n - \mathbf{C}_n - \alpha \mathbf{A}_n e_n + \mathbf{R}_{n+1} - \mathbf{R}_n \tag{71}$$

$$\boldsymbol{\epsilon}_{n+1} = \boldsymbol{\epsilon}_n - \alpha \mathbf{A}_n e_n + \mathbf{F}_n, \tag{72}$$

where we have defined

$$\boldsymbol{\epsilon}_n = \mathbf{R}_n - \mathbf{C}_n = \mathbf{R} e^{-jn\Delta} - \mathbf{C}_n \tag{73}$$

$$\mathbf{F}_n = \mathbf{R}_{n+1} - \mathbf{R}_n = \mathbf{R} e^{-jn\Delta}(e^{-j\Delta} - 1). \tag{74}$$

The analysis given in Appendix A can now be carried through up to (69) by keeping $\mathbf{F}_n$ different from zero. The MSE at the $(n+1)$th iteration is now

$$\langle e_{n+1}^2 \rangle = (1 - 2\alpha A + 2\alpha^2 N A^2)\langle e_n^2 \rangle + 2\alpha A \langle \xi^2 \rangle + Af(\mathbf{F}_n), \tag{75}$$

where from (68)

$$f(\mathbf{F}_n) = \langle \mathbf{F}_n^T(\boldsymbol{\epsilon}_n^* - \alpha e_n \mathbf{A}_n^*) + (\boldsymbol{\epsilon}_n^T - \alpha e_n \mathbf{A}_n^T)\mathbf{F}_n^* \rangle. \tag{76}$$

Using (72) and (73), we can rewrite this expression in the following way:

$$f(\mathbf{F}_n) = \mathbf{F}_n^T(\mathbf{R}_n^* - \langle \mathbf{C}_{n+1}^* \rangle) + (\mathbf{R}_n^T - \langle \mathbf{C}_{n+1}^T \rangle)\mathbf{F}_n^*. \tag{77}$$

The evaluation of this quantity requires the knowledge of the mean-tap fluctuations $\langle \mathbf{C}_n \rangle$. Replacing e_n in (56) by its value in (61) and taking the average gives

$$\begin{aligned} \langle \mathbf{C}_{n+1} \rangle &= \langle \mathbf{C}_n \rangle + \frac{\alpha}{2} \langle \mathbf{A}_n(\boldsymbol{\epsilon}_n^T \mathbf{A}_n^* + \mathbf{A}_n^T \boldsymbol{\epsilon}_n^*) \rangle \\ &= \langle \mathbf{C}_n \rangle + \alpha A \langle \boldsymbol{\epsilon}_n \rangle. \end{aligned} \tag{78}$$

From (73) we get

$$\langle \mathbf{C}_{n+1} \rangle = \langle \mathbf{C}_n \rangle (1 - \alpha A) + \alpha A \mathbf{R} e^{-jn\Delta}. \tag{79}$$

After some algebra, the solution of this recurrence equation is found to be

$$\langle \mathbf{C}_{n+1} \rangle = \langle \mathbf{C}_0 \rangle (1 - \alpha A)^{n+1} + \alpha A \mathbf{R} e^{-jn\Delta} \cdot \frac{1 - (1 - \alpha A)^{n+1} e^{j(n+1)\Delta}}{1 - (1 - \alpha A) e^{j\Delta}}. \tag{80}$$

In steady-state operation, as n goes to infinity, we have

$$\langle \mathbf{C}_{n+1} \rangle = \alpha A \mathbf{R} e^{-jn\Delta} \cdot \frac{1}{1 - (1 - \alpha A)e^{j\Delta}}. \tag{81}$$

We can now get an expression for the steady-state value of $f(\mathbf{F}_n)$ in (77). After some algebra

$$f(\mathbf{F}_n) = -2R^2(1 - \alpha A) \cdot \frac{3 - 4 \cos \Delta + \cos 2\Delta - 2\alpha A(1 - \cos \Delta)}{1 + (1 - \alpha A)^2 - 2(1 - \alpha A) \cos \Delta}, \tag{82}$$

where we have defined

$$R^2 = \mathbf{R}^T \mathbf{R}^*. \tag{83}$$

In the usual case where $\Delta \ll 1$, we can use the approximation

$$\cos \Delta \cong 1 - \frac{\Delta^2}{2} \tag{84}$$

in which case (82) simplifies to

$$f(\mathbf{F}_n) \cong \frac{2\alpha A \Delta^2 (1 - \alpha A) \cdot R^2}{\alpha^2 A^2 + (1 - \alpha A)\Delta^2}. \tag{85}$$

This quantity is a constant, which does not depend on time. The solution to the recurrence (75) is then, in steady-state operation,

$$\langle e_n^2 \rangle = \langle e_n^2(\Delta = 0) \rangle + \frac{1 - \alpha A}{1 - \alpha NA} \cdot \frac{A\Delta^2 R^2}{\alpha^2 A^2 + (1 - \alpha A)\Delta^2}, \tag{86}$$

where $\langle e_n^2(\Delta = 0) \rangle$ designates the steady-state value of the MSE given in (70) in the absence of frequency offset.

We can use this expression to find the value of the ERLE in the presence of frequency offset. We will assume that there is no other impairment in the channel. The output of the channel is then

$$s_{e,n} = \frac{1}{2} (\mathbf{A}_n^T \mathbf{R}_n^* + \mathbf{R}_n^T \mathbf{A}_n^*), \tag{87}$$

and the mean-squared output of the channel is

$$\langle s_{e,n}^2 \rangle = A \mathbf{R}_n^T \mathbf{R}_n^* = AR^2. \tag{88}$$

The MSE after cancellation is given in (86), where we put the first term on the right equal to zero. The ERLE is then

$$\begin{aligned} \text{ERLE} &= 10 \log \langle s_{e,n}^2 \rangle - 10 \log \langle e_n^2 \rangle \\ &= -10 \log \left(\frac{1 - \alpha A}{1 - \alpha NA} \cdot \frac{\Delta^2}{\alpha^2 A^2 + (1 - \alpha A)\Delta^2} \right). \end{aligned}$$

Notice that the ERLE goes to zero when the step size α goes to zero.

AUTHOR

Jean-Jacques Werner, Ing. Deg., 1965, INSA, Lyon, France; M.S. (Electrical Engineering), 1967, Laval University, Canada; Eng. Sc.D. (Electrical Engineering), 1973, Columbia University; Bell Laboratories, 1973–1982; AT&T Information Systems, 1983—. At Bell Laboratories and AT&T Information Systems, Mr. Werner has worked on problems in data transmission and digital signal processing. Member, Sigma Xi, IEEE.

AT&T Technical Journal
Vol. 64, No. 1, January 1985
Printed in U.S.A.

Effects of Biases on Digitally Implemented Data-Driven Echo Cancelers

By J. M. CIOFFI* and J. J. WERNER†

(Manuscript received July 10, 1984)

In this paper the effects of biases, of a hardware origin, on the performance of a digitally implemented data-driven echo canceler are studied both analytically and experimentally. It is shown that, as a consequence of any such bias, the canceler tap weights can randomly drift; however, in contrast to voice-type cancelers and fractionally spaced equalizers, the data-driven canceler will not drift into instability. Nevertheless, the canceler's performance can be severely degraded, even for very small amounts of bias. The main result of this paper is a quantitative study of the canceler's performance as a function of the biases and the canceler's various design parameters, such as the number of tap coefficients and the step size used in the tap adjustment algorithm. Although the study concentrates on the biases introduced by two's-complement arithmetic, the results are general enough to be used with any type of arithmetic, provided that the biases introduced by these different types of arithmetic are known. Some of the analytical results have been verified experimentally, in real time, on a digital signal processor constructed at AT&T Bell Laboratories and AT&T Information Systems. Specifically, it is shown how the bias introduced by rounding the product of commercially available two's-complement multipliers can be eliminated by a proper choice of the values of the canceler's input symbols.

I. INTRODUCTION AND SUMMARY

The "tap-drifting" problem in fractionally spaced equalizers and voice-type echo cancelers is a manifestation of the presence of small

* AT&T Bell Laboratories; now at IBM Research. † AT&T Information Systems, Inc.

biases in the digital implementation of adaptive algorithms. These biases allow some of the equalizer and canceler tap coefficients to slowly grow in magnitude. As a result, the intermediate accumulated sums in the filtering computations also grow. Ultimately, either the tap coefficients or the intermediate-accumulated sum exceeds the boundaries of the digital representation ("overflow"), and the algorithms become unstable. The sensitivity of these algorithms to biases has been investigated, and several corrective actions against tap drifting have been proposed and successfully implemented.[1,2] Tap drifting can, in principle, be eliminated in its hardware origin by removing all the biases from the digital implementation. However, the elimination of biases is not always possible, especially if off-the-shelf devices are used in the design. For example, most commercially available arithmetic units use two's-complement arithmetic, which introduces biases in the computations when the digital words are reduced in length.

In this paper, we present a study of the tap-drifting problem in a data-driven canceler. This canceler is very attractive for two-wire full-duplex data transmission applications.* In its simplest configuration, for 4800-b/s full-duplex operation, the data-driven canceler requires only additions and subtractions for the computation of the filtering and updating algorithms. In this case no multiplier is required in the implementation of the canceler. However, a multiplier is required for higher speeds of transmission ($\geq$9600 b/s).

For most implementations, the effect of bias can be mitigated by using enough precision in the digital computations. However, such an approach will generally not be cost-effective, and in most practical digital implementations there will be some small biases due to the finite precision used in the computations. It is shown in this paper that, unlike voice-type cancelers,[4] data-driven cancelers cannot be driven to instability by these biases. Nevertheless, tap drifting always occurs to some extent and, after a sufficient period of time, introduces a degradation in the canceler's performance. This phenomenon has been studied both analytically and experimentally. We give a general formula that permits the degradation in the canceler's performance due to digital biases to be predicted. Real-time experimental results were obtained on a digital signal processor. The frequency components in the distortion introduced by the biases are shown to include spectral lines located at the origin and at multiples of the symbol rate. It is also shown that the performance of the biased canceler is degraded when the step size used in the updating of the tap coefficients de-

* The motivations for using echo cancellation in these applications are explained in Ref. 3. Section II of Ref. 3 also presents more details about the echo canceler structure used in the following study.

creases. This is in direct contrast to the behavior of the unbiased, infinite-precision canceler for which performance improves with decreasing step sizes. This phenomenon has also been observed in other applications.[5-7]

Special attention has been given to the degradation introduced by commercially available two's-complement multipliers. The naive use of such multipliers is shown to degrade the canceler's performance to unacceptable levels, even when large precision is used in the digital implementation. However, it is proved in this paper that proper choice of the values of the data symbols can completely eliminate the bias associated with the rounding of a two's-complement product. Two sets of binary symbols having this desired property are described.

The paper is organized as follows. In the next section the data-driven canceler used in the analysis and experiments is briefly described. In Section III we discuss the mathematical modeling of the biases, when two's-complement arithmetic is utilized in the canceler's updating algorithm. Quantitative results for the degradation in the canceler's performance are obtained in Section IV. The frequency-domain characteristics of the distortion introduced by the biases are studied in Section V. Finally, in Section VI we present some experimental results obtained on a digital signal processor, and we compare these results to the analytical results.

II. CANCELER DESCRIPTION

The data-driven passband echo canceler described in Refs. 8 and 3 synthesizes a signal of the form

$$s(t) = \text{Re}\left\{\sum_n A_n g(t - nT)e^{jw_c t}\right\}, \tag{1}$$

where $A_n = a_n + jb_n$ is the complex symbol to be transmitted, $g(t)$ is a (possibly complex) baseband signal, $\omega_c/2\pi$ is the carrier frequency, and Re denotes the real part of the quantity in brackets. It is shown in Ref. 8 that this signal can be generated by using the structure given in Fig. 1. The canceler consists of two transversal filters whose taps are spaced at intervals T', where $1/T'$ has to be at least twice the highest frequency in the signal $s(t)$ in (1). This condition makes the canceler Nyquist, that is, it can generate an exact replica of $s(t)$ at all frequencies.

After convergence, the tap coefficients are equal to the sampled values of the impulse responses of the in-phase and quadrature passband filters. For this reason the canceler is called an "in-band" canceler, distinguishing it from other possible structures that synthesize baseband equivalents of the channel. The symbol rotation at the input of the canceler ensures phase continuity of the carrier. For most

cases of practical interest, the relationship between ω_c and T is such that the rotated symbols a'_n and b'_n are similar to the symbols a_n and b_n. That is, if a_n and b_n are binary levels $\{\pm a\}$, then the rotated symbol levels will also be binary. Thus, the primes have been dropped in the ensuing analysis. The Digital-to-Analog (D/A) converter and low-pass filter at the output of the canceler perform the usual interpolation functions needed for further analog processing.

Since inputs are accepted by the canceler at a rate of $1/T'$, while the data symbols are only presented at a rate $1/T$, L − 1 zero symbols, where $L = T/T'$, are inserted between successive nonzero inputs to the canceler. Thus, only one of every L complex taps is active for each filter iteration, as shown in Fig. 1. The unnecessary computations associated with the zero symbols can be eliminated by grouping taps that act simultaneously into L parallel subcancelers, as seen in Fig. 2. Similarly, the echo channel can be considered as a parallel combination of L subchannels. Convergence of the canceler is achieved by minimizing the Mean-Squared Error (MSE) between the outputs of each subchannel and the corresponding subcanceler. The subcancelers are assumed to adapt independently, and this assumption was found to be in excellent agreement with experimental results. The MSE for the whole canceler is obtained by averaging MSEs of all the subcancelers. A more detailed analysis of the subcanceler structure is given in Ref. 8. The echo canceler's performance in the presence of channel impairments is studied in a companion paper.[3] Some of the definitions used in Ref. 3 that are needed in the sequel are briefly repeated. The error for the ith subcanceler at time $nT + iT'$ is

$$e_{n,i} = (\underline{r}_{1i} - \underline{c}_{n,i})^T \underline{a}_n + (\underline{r}_{2i} - \underline{d}_{n,i})^T \underline{b}_n + \xi_i, \qquad (2)$$

where $i = 1, 2, \cdots, L$ and the superscript T denotes a transposed vector. In (2), $\underline{r}_{1i}$ and $\underline{r}_{2i}$ are the sampled in-phase and quadrature impulse response vectors of the ith subchannel; $\underline{c}_{n,i}$ and $\underline{d}_{n,i}$ are the in-phase and quadrature tap coefficient vectors of the ith subcanceler; and ξ_i is a signal that is uncorrelated with the data symbols (the far-end signal and noise). The following definitions are also needed:

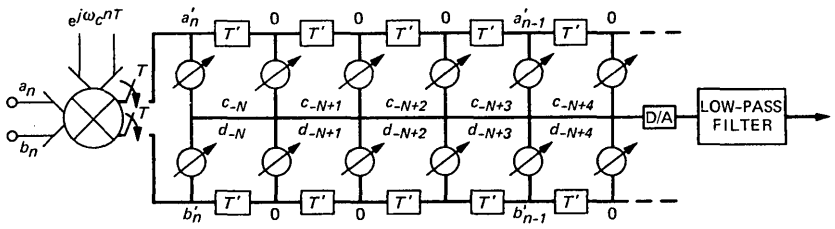
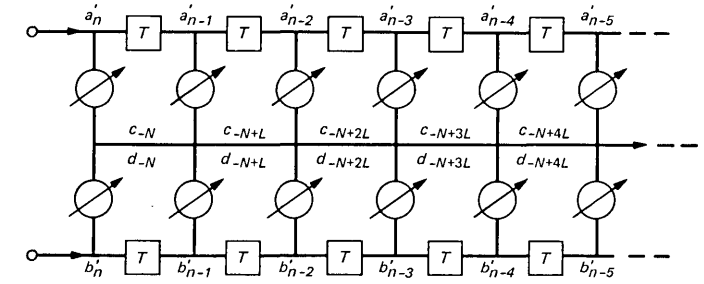
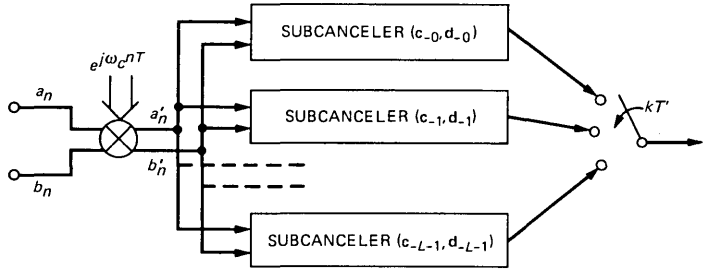


Fig. 1—In-band echo canceler structure.



(a)



(b)

Fig. 2—(a) Subcanceler structure. (b) Modified canceler structure.

$$\underline{A}_n = \underline{a}_n + j\underline{b}_n = \text{vector of complex input symbols,}$$

$$\underline{C}_{n,i} = \underline{c}_{n,i} + j\underline{d}_{n,i} = \text{vector of complex tap coefficients of the } i\text{th subcanceler.}$$

The above summary of the subcanceler structure is sufficient for the purpose of describing the effects of biases in this paper.

III. MATHEMATICAL MODEL OF BIAS

The stochastic-gradient algorithm for the adjustment of the ith subcanceler's complex tap coefficients is given by

$$\underline{C}_{n+1,i} = \underline{C}_{n,i} + \gamma e_{n,i}\underline{A}_n, \qquad i = 1, 2, \cdots, L, \tag{3}$$

where γ is the step size, and the error $e_{n,i}$ is given in (2). All of the entries (real and imaginary) in $\underline{A}_n$ are assumed to be discrete-valued symbols with zero mean. Biases are usually introduced in the evaluation of the correction term,

$$\underline{G}_{n,i} = \gamma e_{n,i}\underline{A}_n. \tag{4}$$

In the following, it is assumed that this expression is computed by using fractional two's-complement arithmetic, since this is the type of arithmetic most commonly used in digital signal processing. (Some of the properties of two's-complement arithmetic are discussed in Appendix B.) In a cost-conscious implementation, the quantity $\gamma e_{n,i}$ should be evaluated first, because it is the same for all the tap coefficients. First consideration is given to the case where (4) is computed without utilizing a multiplier, i.e., the scaled error, $\gamma e_{n,i}$, is obtained by using arithmetic right shifts since γ is always less than unity to ensure stability. Due to the finite precision of the digital representation, some of the lower bits of $e_{n,i}$ may be lost during the shifting operation, and consequently, a positive number will always decrease in magnitude and a negative number will always increase in magnitude. Therefore, negative bias is introduced, on the average, during the computation of $\gamma e_{n,i}$ and during "multiplication" by data symbols having value less than unity (since this operation also corresponds to right shifts). If all of the symbols are binary and chosen equal to ± 1, the updating algorithm in (3) is simply implemented by either adding or subtracting the scaled error to each tap coefficient. An explicit bias occurs during these operations if the quantity $\gamma e_{n,i}$ has "fallen out" of its register length. In this case a positive number becomes a true zero, but a negative number remains equal in magnitude to the Least-Significant Bit (LSB) of the digital representation, thus again introducing a negative mean bias. The correction factor in (4) can now be rewritten as

$$\underline{G}_{n,i} = (\gamma e_{n,i} + \Delta_{1n,i})\underline{A}_n + \underline{\Delta}_{2n,i} = \gamma e_{n,i}\underline{A}_n + \underline{B}_{n,i}, \qquad (5)$$

where $\Delta_{1n,i}$ is a random variable representing biases introduced in the evaluation of $\gamma e_{n,i}$; $\underline{\Delta}_{2n,i}$ is a complex random vector representing biases introduced in the computation of $(\gamma e_{n,i})\ \underline{A}_n$; and $\underline{B}_{n,i}$ is the total bias. The other quantities are assumed to be represented with infinite precision.

An expression similar to (5) is obtained when a two's-complement multiplier is used to compute the correction factor in the updating of the tap coefficients. When a negative product is truncated, the resulting number is always increased in magnitude. Conversely, when the product is positive, it is always decreased in magnitude. Therefore, truncation introduces a negative mean bias. Two's-complement rounding, on the other hand, always selects the number that is closest in magnitude to the exact product, independent of the product's sign. One exception, however, occurs when the double-precision product is equally close to two single-precision numbers. In this case the most positive number is always selected, independently of the product's

sign, and a positive bias is introduced in the computations.* This situation can occur quite frequently in data-driven echo cancelers using symbol values of $\pm 1/2$ or $\pm 1/4$. A remarkable property of data-driven cancelers is that the mean bias associated with this rounding can be eliminated by using the proper levels for the symbols. These levels must be chosen in a way such that the bias situation can never occur. Two sets of symbols which have the desired property are described in Appendix B.

The model represented by (5) also holds for other types of arithmetic. Furthermore, other effects such as biases in the Analog-to-Digital (A/D) conversion can also be accounted for by properly defining the random variables in $\Delta_{1n,i}$ and $\underline{\Delta}_{2n,i}$. In general, the statistics of these random variables will depend on γ, the statistics of $e_{n,i}$ and $\underline{A}_n$, and the type of digital implementation utilized. The characterization of the statistics of $\Delta_{1n,i}$ and $\underline{\Delta}_{2n,i}$ is a formidable problem that will not be addressed here. However, reasonable approximations permit the study of the canceler's performance degradation in the presence of digital biases. It will be shown later that both the mean and the variance of the random variables in (5) influence the canceler's performance. However, the effect of any nonzero-mean bias will, in general, be predominant. This quantity is discussed next.

From (5) the mean gradient estimate is given by

$$\langle \underline{G}_{n,i} \rangle = \langle \gamma e_{n,i} \underline{A}_n \rangle + \underline{B}_i, \tag{6}$$

where

$$\underline{B}_i = \langle \underline{B}_{n,i} \rangle = \langle \Delta_{1n,i} \underline{A}_n + \underline{\Delta}_{2n,i} \rangle. \tag{7}$$

The mean values of $\Delta_{1n,i}$ and $\underline{\Delta}_{2n,i}$ are generally not zero, as explained in the preceding discussion. In obtaining quantitative results, some assumptions concerning the mean values are made. First, all of the components of the vector $\langle \underline{\Delta}_{2n,i} \rangle$ corresponding to the ith subcanceler are assumed equal. This is a reasonable assumption since the same scaled error is used for the updating of all of the taps of a given subcanceler. The weighting of this term by $\underline{A}_n$ should yield the same average bias in steady-state operation. It is not assumed that the vector $\langle \underline{\Delta}_{2n,i} \rangle$ is the same for all the subcancelers, since the error, $e_{n,i}$, is a sample of a cyclostationary process whose statistics will depend upon the index i. As a consequence, the statistics of the product, $\gamma e_{n,i} A_n$, generally vary for different subcancelers.

* This type of bias is present in most of the commercially available two's-complement multipliers. However, the bias can be removed in a new design at the cost of a slight increase in the chip's complexity. This is achieved by first detecting the bias condition and then changing the rounding rules according to the sign of the product.

The influence of the term $\langle \Delta_{1n,i} \underline{A}_n \rangle$ in (7) is somewhat more difficult to assess. If the assumption is made that $\Delta_{1n,i}$ is uncorrelated with $\underline{A}_n$ and that $\langle \underline{A}_n \rangle$ is zero, then this quantity is also zero. However, there can be reasonably long sequences of symbols during which the mean value of $\underline{A}_n$ is not zero. Biases can accumulate during these sequences and introduce a degradation in the canceler's performance. We will not pursue this problem any further.

IV. EXCESS ERROR DUE TO BIAS

We will now show how biases can produce an increase in the MSE. It is assumed that no other impairments are present except, perhaps, for some uncorrelated noise added to the input signal. The mean tap coefficient fluctuations are investigated first. In the beginning of this section, the subscript i is dropped in the equations, but it will be understood that the analysis applied to a subcanceler. Combining (3) and (6), the mean tap vector evolves according to

$$\langle \underline{C}_{n+1} \rangle = \langle \underline{C}_n \rangle + \langle \gamma e_n \underline{A}_n \rangle + \underline{B}. \tag{8}$$

Using (2), we can write this expression as

$$\langle \underline{C}_{n+1} \rangle = (1 - \gamma A) \langle \underline{C}_n \rangle + (\underline{r}_1 + j\underline{r}_2) \gamma A + \underline{B}, \tag{9}$$

where $A = \langle a_n^2 \rangle = \langle b_n^2 \rangle$, and the a_n's and b_n's are assumed to be uncorrelated. The steady-state tap values are given by

$$\langle \underline{C}_\infty \rangle = \underline{r}_1 + j\underline{r}_2 + \frac{\underline{B}}{\gamma A}, \tag{10}$$

where the term $\underline{B}/\gamma A$ represents the tap deviation due to the mean bias $\underline{B}$. With $\Delta \underline{C}$ denoting the tap deviation from the optimum setting, one obtains

$$\Delta \underline{C} = \langle \underline{C}_\infty \rangle - \underline{C}_{\text{opt}} = \frac{\underline{B}}{\gamma A}. \tag{11}$$

Note that the bias is the same for each tap weight and that decreasing the step size γ will result in an increased mean tap deviation. This contrasts with the well-known results that decreasing the step size, for an infinite-precision canceler (without bias), will, in general, improve steady-state performance. The deviation is also proportional to the bias, which agrees with intuition. Recall that in a finite-precision environment the step size cannot be arbitrarily close to zero, and therefore the tap deviation in (11) cannot approach infinity. The only way to effectively have a zero step size is to stop updating, in which case the bias no longer affects the algorithms. When quantitative results are discussed in the following sections, the step sizes used in practice will usually be found to be several orders of magnitude larger

than the bias $\underline{B}$. Thus the tap offset in (11) is small. As a consequence, the tap coefficients will never overflow, and a data-driven canceler cannot become unstable. However, as will be shown later, a very small offset can severely degrade the MSE.

It is interesting to contrast the preceding results with those obtained for a voice-type canceler, which behaves similarly to fractionally spaced equalizers.[2] In this case it can be shown that the mean tap offset becomes

$$\Delta\underline{C} = \mathscr{A}^{-1}\frac{\underline{B}}{\gamma}, \tag{12}$$

where $\mathscr{A}$ is the input autocorrelation matrix of the data signal. Performing a spectral expansion of the mean tap deviation yields

$$\Delta\underline{C} = \frac{1}{\gamma}\sum_i \frac{1}{\gamma_i}\underline{f_i^T B f_i}, \tag{13}$$

where γ_i and $\underline{f_i}$ are the ith eigenvalue and eigenvector of $\mathscr{A}$, respectively. Small eigenvalues, corresponding to input frequency ranges of little or no energy, can make this term large, especially if $\underline{f_i^T B}$ is not small. Hence, for the voice-driven canceler, distortion due to biases can be expected to be concentrated in frequency ranges corresponding to little input energy. Furthermore, the tap offset in (13) can be much larger than the offset given in (11), so that some tap coefficients can overflow and make the canceler unstable.

The mean-squared error for the data-driven subcanceler is derived in Appendix A.* The expression for the ith subcanceler's steady-state MSE is given by

$$\langle e_{\infty,i}^2\rangle = \frac{\langle \xi_i^2\rangle + \dfrac{N}{\gamma}\langle \beta_i^2\rangle + \dfrac{2N}{\gamma^2 A}\langle \beta_i\rangle^2}{1-\gamma NA}, \tag{14}$$

where $\langle \xi_i^2\rangle$ is the minimum mean-squared error, N is the number of taps, and β_i is a random variable corresponding to one component (real or imaginary) of the vector $\underline{B}_{n,i}$ in (5). In the derivation of (14), it was assumed that $\langle \beta_i^2\rangle$ and $\langle \beta_i\rangle$ were constants and the same for all the entries of the vector $\underline{B}_{n,i}$. The MSE averaged over all L subcancelers is

* Another expression for the MSE (also derived in Appendix A) is obtained in the resonance case, when the step size has its optimum value for speed of convergence. This step size can only be used at start-up, when there is no double-talker, in which case convergence is so fast that the effects of biases are negligible. Therefore, we will not discuss this second case any further, and the MSE discussed in the sequel corresponds to the nonresonant case.

$$\langle e_\infty^2 \rangle = \frac{\langle \xi^2 \rangle + \dfrac{LN}{\gamma}\langle \beta^2 \rangle + \dfrac{2LN}{\gamma^2 A}\langle \beta^2 \rangle}{1 - \gamma NA}, \tag{15}$$

where we have defined

$$\langle \xi^2 \rangle = \frac{1}{L}\sum_{i=0}^{L-1} \langle \xi_i^2 \rangle, \qquad \langle \beta^2 \rangle = \frac{1}{L}\sum_{i=0}^{L-1} \langle \beta_i^2 \rangle, \tag{16}$$

and

$$\langle \beta \rangle^2 = \frac{1}{L}\sum_{i=1}^{L-1} \langle \beta_i \rangle^2.$$

Two quantities of interest can be derived from this expression. The first one is the *Bias-Performance Ratio* (BPR), which is defined as the ratio of the uncanceled echo's power to the MSE after cancellation, in the absence of a double-talker. This quantity is similar to the Echo-Return-Loss Enhancement (ERLE) studied in Ref. 3 and is particularly useful in comparing analytical and experimental results. The other quantity of interest is the signal-to-noise ratio (s/n), which is the ratio of the wanted signal power to the power of any additive noise after cancellation and which determines the receiver error rate. (From the receiver's viewpoint, the wanted signal is the double-talker, which acts as noise in the tap-adjustment algorithm.) From (15) the BPR is given by

$$\text{BPR} = \frac{P_e(1 - \gamma NA)}{\dfrac{LN}{\gamma}\langle \beta^2 \rangle + \dfrac{2LN}{\gamma^2 A}\langle \beta \rangle^2}, \tag{17}$$

where P_e is the power of the uncancelled echo. In deriving an expression for the s/n, it is assumed that the only component of the interfering signal is the wanted signal, i.e., there is no additive noise. The total signal power after cancellation is the MSE, (15), the power in the wanted signal is $P_s = \langle \xi^2 \rangle$, and the remaining power is noise, so that the s/n becomes

$$\text{s/n} = \frac{P_s}{\langle e_\infty^2 \rangle - P_s} = \frac{1 - \gamma NA}{\gamma NA + \dfrac{LN}{\gamma P_s}\langle \beta^2 \rangle + \dfrac{2LN}{\gamma^2 A P_s}\langle \beta \rangle^2}. \tag{18}$$

Under normal conditions of operation, the step size γ is very small, relative to $1/NA$, and the expressions in (17) and (18) can be approximated by

$$\text{BPR} \approx \frac{P_e A \gamma^2}{2LN\langle \beta \rangle^2} \tag{19}$$

and

$$\text{s/n} \approx \frac{P_s A \gamma^2}{2LN\langle\beta\rangle^2}, \tag{20}$$

where it has been assumed that $\langle\beta\rangle \neq 0$. The BPR and the s/n have the same expressions, except that the echo's power, P_e, in (19) is replaced by the wanted signal's power, P_s, in (20). Both quantities decrease with decreasing step size, which is in agreement with the preceding observations made about the mean tap behavior for small step size.

Equations (14) through (20) reflect degradation in the canceler's performance only under steady-state conditions. These steady-state quantities are not influenced by initial conditions. The complete equations for the MSE evolution with time are given by (55) and (64) in Appendix A. These equations are strongly influenced, in their transient terms, by the canceler's initial state. An initial condition of particular interest is that which the condition that exists after the canceler has converged with no bias, and then a bias is introduced. As was discussed in Section III, this situation can arise as a result of certain nonrandom short-term statistics of the scrambled data sequence, A_n. Under these circumstances, the iterative MSE, given in (55), reduces to

$$\begin{aligned}\langle e_n^2\rangle = {} & \frac{\langle\xi^2\rangle}{1-\gamma NA} \\ & + (1-[1-2\gamma A+2N\gamma^2A^2]^n)\,\frac{\dfrac{NL}{\gamma}\langle\beta^2\rangle + \dfrac{2NL}{\gamma^2 A}\langle\beta\rangle^2}{(1-\gamma NA)} \\ & + ([1-\gamma A]^n - [1-2\gamma A+2N\gamma^2A^2]^n)\,\frac{4N\langle\beta^2\rangle}{\gamma^2 A(2N\gamma A-1)},\end{aligned} \tag{21}$$

and the evolution of the s/n is approximated by

$$s/n_n = \frac{P_s}{\langle e_n^2\rangle - P_s} \cong \frac{P_s\gamma^2 A}{P_s(\gamma NA) + [1-(1-2\gamma A)^n](2NL)\langle\beta\rangle^2}, \tag{22}$$

where it is assumed that the uncorrelated signal $\langle\xi^2\rangle$ consists only of the wanted signal with power, P_s. The approximation in (22) is obtained for a small step size. We will not pursue further the study of transient effects.

V. FREQUENCY ANALYSIS OF BIAS DISTORTION

For the data-driven canceler, distortion due to biases introduces spectral components concentrated in narrow frequency bands centered

around integer multiples of the symbol rate. To facilitate investigation of this phenomenon, the mean tap vector is again written in terms of its deviation from optimum

$$\langle \underline{C}_\infty \rangle = \underline{C}_{\text{opt},i} + \Delta \underline{C}_i, \tag{23}$$

where the subscript i is reintroduced to designate a subcanceler. The corresponding frequency response is the Fourier transform of the sequence obtained from the vector $\underline{C}_\infty$, and it is defined as

$$\langle C_{\infty,i}(\omega) \rangle = C_{\text{opt},i}(\omega) + \Delta C_i(\omega). \tag{24}$$

Since all components of the mean-bias vector $\underline{B}_i$ are assumed equal and

$$\Delta \underline{C}_i = \frac{\underline{B}_i}{\gamma A}, \tag{25}$$

then

$$\begin{aligned} \Delta C_i(\omega) &= \frac{B_i}{\gamma A} \sum_{n=0}^{N-1} e^{-j\omega nT}, \\ &= \frac{B_i}{\gamma A} e^{-j\omega(N-1)T/2} \frac{\sin[\omega NT/2]}{\sin[\omega T/2]}, \end{aligned} \tag{26}$$

where B_i is a component of $\underline{B}_i$ and $1/T$ is the symbol rate.

The amplitude of the distortion is the familiar periodic sinc function. From (26) it is seen that the distortion is concentrated at integer multiples of the symbol rate. As mentioned earlier, the bias vectors for each subcanceler need not be the same even if the bias components within any particular subcanceler are assumed constant. The corresponding tap-deviation spectrum can be expressed in terms of the tap-deviation spectra, $\Delta C_i(\omega)$, of each of the L subcancelers as

$$\Delta C(\omega) = \sum_{i=0}^{L-1} \Delta C_i(\omega) e^{-j\omega iT'}, \tag{27}$$

where $T' = T/L$.

If each $\Delta C_i(\omega)$ has the same shaping, but with a different magnitude, i.e.,

$$\Delta C_i(\omega) = k_i \Delta C_0(\omega), \qquad i = 0, \cdots, L-1, \tag{28}$$

where $k_0 = 1$ and the k_m are proportionality constants, then (27) becomes

$$\Delta C(\omega) = \Delta C_0(\omega) \sum_{i=0}^{L-1} k_i e^{-j\omega iT'}. \tag{29}$$

For the special case $k_m = 1$ for all m, the distortion becomes

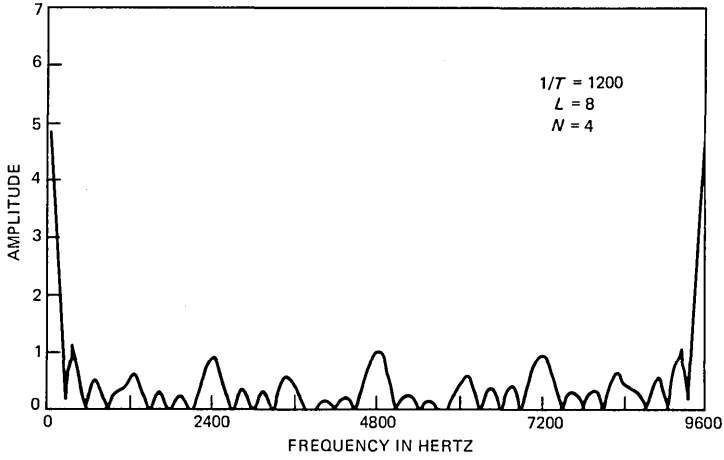


Fig. 3—Frequency distortion of mean tap coefficients.

concentrated at integer multiples of $1/T'$. Since all subcanceler biases are equal in this case, the structure becomes equivalent to one canceler at rate $1/T'$. As the value k_m varies, bias distortion is concentrated at integer multiples of the symbol rate, $1/T$.* One such example is shown in Fig. 3, where we have arbitrarily chosen values of k_m equal to 1, 0.5, 0.25, 0.5, 1, 0.5, 0.75, and 0.5. The study of many other examples shows that as a general rule of thumb, the spectral lines introduced by biases in a data-driven echo canceler composed of L subcancelers will be concentrated at integer multiples of the symbol rate. Experimental evidence verifying these findings will be presented in the next section.

VI. EXPERIMENTAL RESULTS

Several experiments were conducted on a digital signal processor, using a 12-bit two's-complement multiplier, to verify some of the qualitative and quantitative results obtained in the preceding sections. The echo canceler used in the experiments was operating at a sampling rate of 9600 samples per second, and a symbol rate of 1200 bauds. Thus, it could be implemented by using eight subcancelers. The validity of the expression for the BPR in (17) was verified by artificially inserting biases in the updating algorithm of an in-band echo canceler. A positive quantity equal to 2^{-11} was added periodically to the correction factor before updating the tap coefficients. Therefore, if this quantity was added every mT seconds, the equivalent mean bias per

* In certain rare cases, the biases can insert nulls at some particular multiples of the symbol rate.

symbol update was $(2^{-11})/m$. This bias was chosen large enough so that we could distinguish its effect from other digital effects such as round-off noise. The influence of the bias of the two's-complement multiplier was eliminated by using one of the sets of symbols described in Appendix B.

The measured values for the BPR are given in Fig. 4 for different values of m and as a function of the step size. The BPR has also been computed by using (17) and the corresponding curves are shown in Fig. 4. In Fig. 5, curves are also shown for the case in which the bias of the two's-complement multiplier was not eliminated by a proper choice of the data symbols. No artificial bias was added, and the computed curves were obtained by assuming a mean bias of 2^{-13}. (A

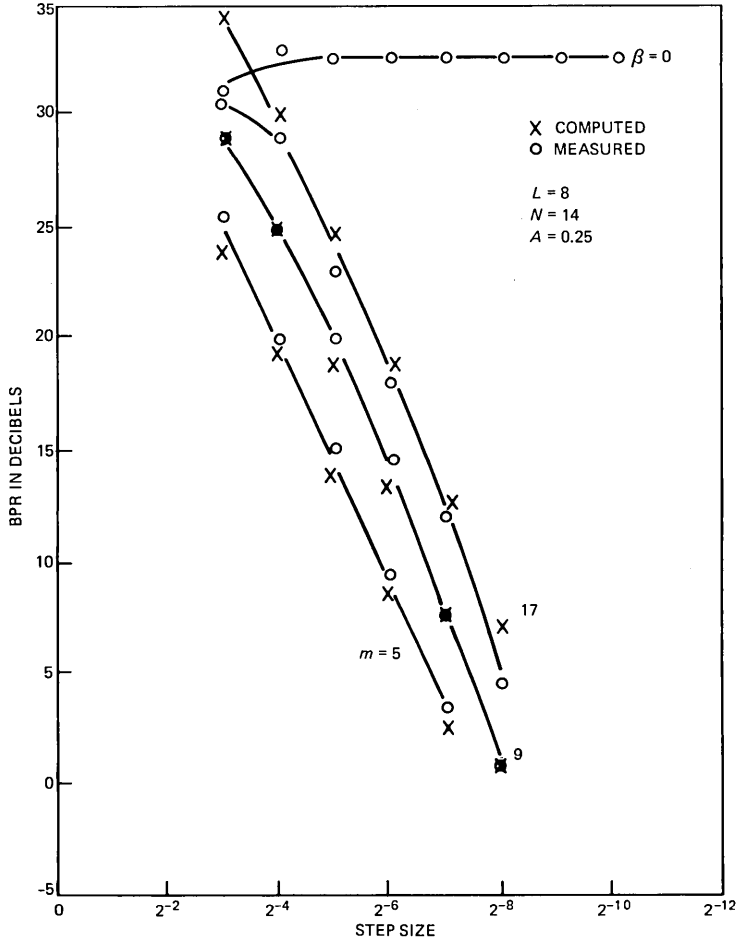


Fig. 4—Performance degradation due to simulated biases.

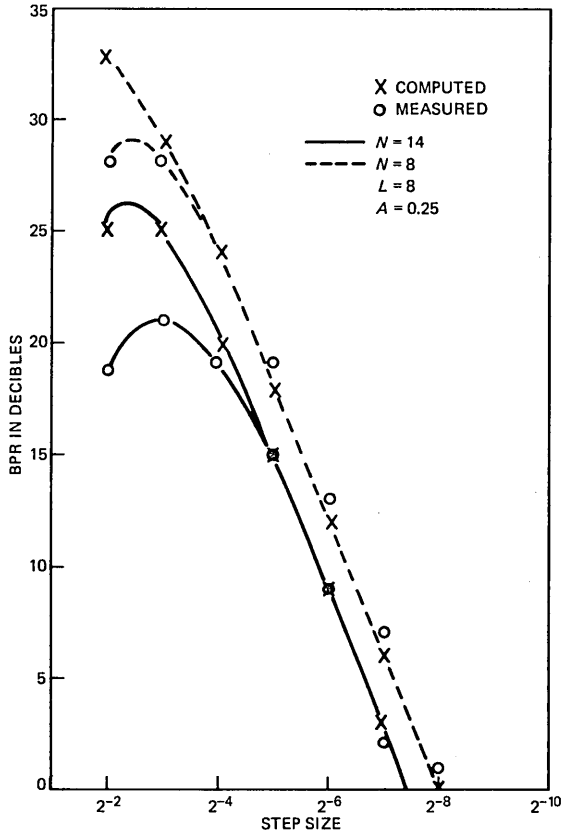


Fig. 5—Performance degradation due to rounding of a two's-complement multiplier.

bias of 2^{-12} was introduced whenever the bias situation occurred. With symbol values $\pm 1/2$, this situation was likely to occur half the time so that a mean bias of 2^{-13} was introduced in the algorithm.) For reasonably small step sizes, both the experimental and the theoretical curves decrease 6 dB for each factor-of-two decrease in step size. This is consistent with the expression for the BPR given in (19), and similar behavior can be expected for the s/n as shown in (20). Both these quantities go to zero when the step size goes to zero. This is in direct contrast to the behavior of a bias-free, infinite precision canceler for which performance improves with decreasing step sizes.

Although the two's-complement multiplier rounding bias was eliminated in the processor by using the sets of symbols described in Appendix B, there remained some other very small, unexplained biases. The effect of these biases on the BPR was negligible, as shown in Fig. 4. Nevertheless, they could be observed by studying the spec-

trum of the residual echo when no double-talker was present. This spectrum is the flat trace in Fig. 6, and the bell-shaped curve depicts the spectrum of the uncanceled echo. (Due to the small magnitude of the residual biases, it took several minutes of the canceler's operation to obtain the spectrum in Fig. 6. Immediately after convergence the peaks were very weak.) Notice that the peaks around the origin and at multiples of the symbol rate (1200 bauds in this case) produce exactly the kind of spectrum that was predicted by the analysis in Section V. These peaks would be much larger if the optimum sets of symbols described in Appendix B were not utilized. In Fig. 6, we also show the spectrum of the residual echo of a voice-type canceler that was implemented on the processor. Notice that for the voice-canceler spectrum, the residual echo's energy accumulates in the frequency regions where the uncanceled echo has little energy. This energy increases with time and ultimately, without some sort of compensation, the canceler diverges after several minutes of operation. Finally, for

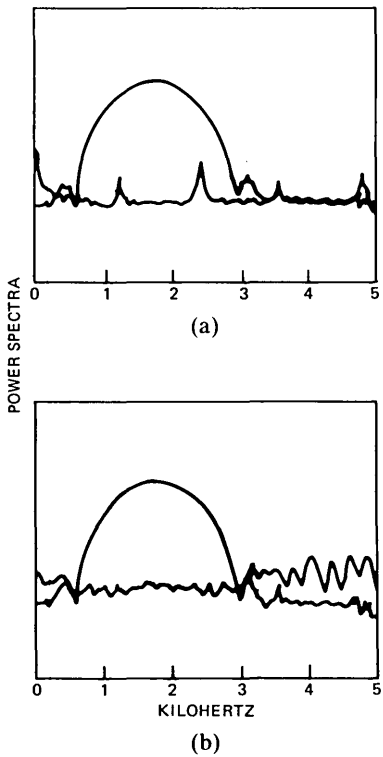


Fig. 6—Power spectrum of the residual error. (a) Data-driven echo canceler. (b) Voice-type echo canceler.

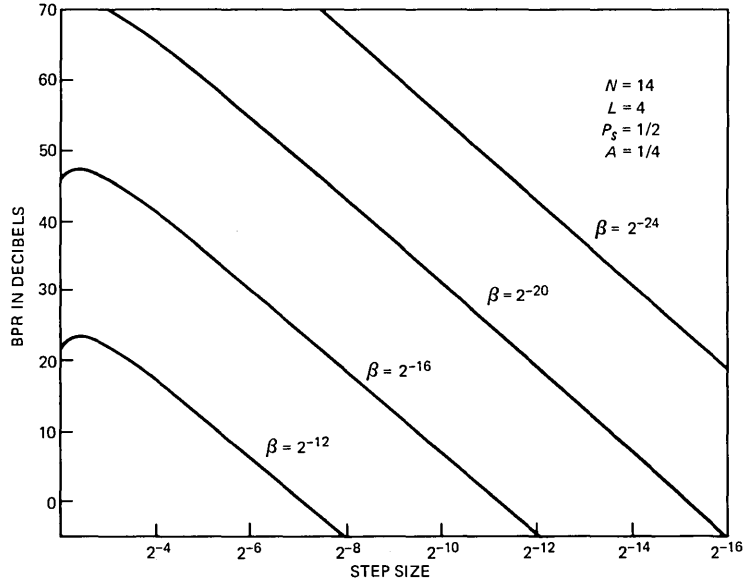


Fig. 7—Bias-performance ratio for various amounts of biases.

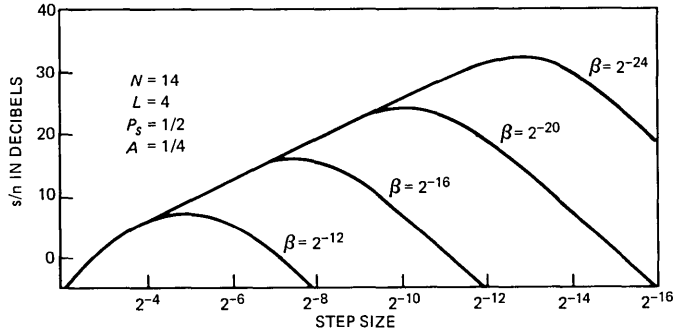


Fig. 8—Signal-to-noise degradation due to biases.

completeness, computed curves giving the BPR and the s/n for various values of biases are given in Figs. 7 and 8.

REFERENCES

1. J. J. Werner, "Control of Coefficient Drift for Fractionally Spaced Equalizers," US Patent 4,384,355, May 17, 1983.
2. R. D. Gitlin, H. C. Meadors, and S. B. Weinstein, "The Tap Leakage Algorithm: An Algorithm for the Stable Operation of a Digitally Implemented Fractionally Spaced Adaptive Equalizer," B.S.T.J., *61*, No. 8 (October 1982), pp. 1817–39.
3. J. J. Werner, "Effects of Channel Impairments on the Performance of an In-band Data-Driven Echo Canceler," AT&T Tech. J., this issue.
4. A. Weiss and D. Mitra, "Digital Adaptive Filters: Conditions for Convergence, Rates

of Convergence, Effects of Noise and Errors Arising from the Implementation," IEEE Trans. Inform. Theory, *IT-25*, No. 6 (November 1979), pp. 637–52.
5. R. D. Gitlin, J. E. Mazo, and M. G. Taylor, "On the Design of Gradient Algorithms for Digitally Implemented Adaptive Filters," IEEE Trans. Circuit Theory, *CT-20*, No. 2 (March 1973).
6. R. D. Gitlin and S. B. Weinstein, "On the Required Tap-Weight Precision for Digitally Implemented, Adaptive, Mean-Squared Equalizers," B.S.T.J., *58*, No. 2 (February 1979), pp. 301–21.
7. C. Caraiscos and B. Liu, "A Roundoff Error Analysis of the LMS Adaptive Algorithm," IEEE Trans. ASSP, *ASSP-32*, No. 2 (February 1984), pp. 34–41.
8. J. J. Werner, "An Echo-Cancellation-Based 4800 bps Full-Duplex DDD Modem," IEEE J. Selected Areas Commun., *SAC-2*, No. 5 (September 1984).

APPENDIX A

MSE Evolution

The following is an analysis of the evolution of the MSE for a data-driven echo canceler in the presence of digital biases. It is assumed that the input sequences $\underline{a}_n$ and $\underline{b}_n$ are white and uncorrelated with the bias, $\underline{\beta}_n$, and each other. It is also assumed that the canceler spans the entire length of the echo impulse response. The analysis applies to a subcanceler, but for simplicity of notation, the index i will be deleted.

The following definitions are used in the derivations:

$$\underline{B}_n = \underline{\beta}_n + j\underline{\beta}_n = \text{complex bias vector}^*$$

$$\underline{C}_n = \underline{c}_n + j\underline{d}_n = \text{complex tap vector}$$

$$\underline{A}_n = \underline{a}_n + j\underline{b}_n = \text{complex data vector}$$

$$\gamma = \text{step size in the adjustment algorithm}$$

$$\underline{r}_1 = \text{sampled in-phase channel impulse response vector} \tag{30}$$

$$\underline{r}_2 = \text{sampled quadrature channel impulse response vector} \tag{31}$$

$$\xi = \text{sampled uncorrelated interfering signal.} \tag{32}$$

The error at the nth iteration is given by

$$e_n = [\underline{r}_1 - \underline{c}_n]^T \underline{a}_n + [\underline{r}_2 - \underline{d}_n]^T b_n + \xi. \tag{33}$$

The in-phase and quadrature tap error vectors are defined by

$$\underline{\in}_{1,n} \triangleq \underline{r}_1 - \underline{c}_n \tag{34}$$

$$\underline{\in}_{2,n} \triangleq \underline{r}_2 - \underline{d}_n \tag{35}$$

---

* For ease of notation the real and imaginary parts of $\underline{B}_n$ are taken to be equal. Although this assumption is not strictly true, it will not modify the end result of the analysis.

so that (33) can be rewritten as

$$e_n = \underline{\in}_{1,n}^T \underline{a}_n + \underline{\in}_{2,n}^T \underline{b}_n + \xi, \tag{36}$$

and the MSE becomes

$$\langle e_{n+1}^2 \rangle = \langle \underline{\in}_{1,n+1}^T \underline{a}_{n+1} \underline{a}_{n+1}^T \underline{\in}_{1,n+1} \rangle + \langle \underline{\in}_{2,n+1}^T \underline{b}_{n+1} \underline{b}_{n+1}^T \underline{\in}_{2,n+1} \rangle + \langle \xi^2 \rangle. \tag{37}$$

Letting

$$A \triangleq \langle a_n^2 \rangle = \langle b_n^2 \rangle \tag{38}$$

and making the usual assumption that successive data vectors are uncorrelated, one obtains

$$\langle e_{n+1}^2 \rangle = A \langle \in_{1,n+1}^T \underline{\in}_{1,n+1} \rangle + A \langle \in_{2,n+1}^T \underline{\in}_{2,n+1} \rangle + \langle \xi^2 \rangle. \tag{39}$$

The biased stochastic-gradient algorithm can be expressed as

$$\underline{c}_{n+1} = \underline{c}_n + \gamma e_n \underline{a}_n + \underline{\beta}_n \tag{40}$$

$$\underline{d}_{n+1} = \underline{d}_n + \gamma e_n \underline{b}_n + \underline{\beta}_n . \tag{41}$$

Subtracting both sides of (40) and (41) from $\underline{r}_1$ and $\underline{r}_2$ yields

$$\underline{\in}_{1,n+1} = \underline{\in}_{1,n} - \gamma e_n \underline{a}_n - \underline{\beta}_n \tag{42}$$

$$\underline{\in}_{2,n+1} = \underline{\in}_{2,n} - \gamma e_n \underline{b}_n - \underline{\beta}_n . \tag{43}$$

Using (42) and (43) in (39) gives

$$\langle e_{n+1}^2 \rangle = A \langle (\underline{\in}_{1,n} - \gamma e_n \underline{a}_n - \underline{\beta}_n)^T (\underline{\in}_{1,n} - \gamma e_n \underline{a}_n - \underline{\beta}_n) \rangle + A \langle (\underline{\in}_{2,n} - \gamma e_n \underline{b}_n - \underline{\beta}_n)^T (\underline{\in}_{2,n} - \gamma e_n \underline{b}_n - \underline{\beta}_n) \rangle + \langle \xi^2 \rangle . \tag{44}$$

After some algebra, and using (36), (38), and $\underline{a}_n^T \underline{a}_n = \underline{b}_n^T \underline{b}_n = NA$, we have

$$\langle e_{n+1}^2 \rangle = \langle e_n^2 \rangle [1 - 2\gamma A + 2\gamma^2 N A^2] + 2\gamma A \langle \xi^2 \rangle + 2AN \langle \beta_n^2 \rangle - 2A \langle \underline{\beta}_n^T \underline{\in}_{1,n} \rangle - 2A \langle \underline{\beta}_n^T \underline{\in}_{2,n} \rangle. \tag{45}$$

To proceed further, $\langle \underline{\beta}_n^T \underline{\in}_{1,n} \rangle$ and $\langle \underline{\beta}_n^T \underline{\in}_{2,n} \rangle$ must be evaluated. We will assume that $\underline{\beta}_n$ is uncorrelated with $\underline{\in}_{1,n}$ and $\underline{\in}_{2,n}$.

Using (42) and (36), and assuming again that the current tap error vector and the current data vector are independent, we have

$$\langle \underline{\beta} \rangle^T \langle \underline{\in}_{1,n+1} \rangle = (1 - \gamma A) \langle \underline{\beta} \rangle^T \langle \underline{\in}_{1,n} \rangle - N \langle \beta \rangle^2, \tag{46}$$

where it is assumed that the mean vector $\langle \underline{\beta}_n \rangle$ is a constant vector whose entries are all equal to $\langle \beta \rangle$. Solving the iteration (46) yields

$$\langle \underline{\beta} \rangle^T \langle \underline{\in}_{1,n} \rangle = (1 - \gamma A)^n \langle \underline{\beta} \rangle^T \underline{r}_1 - \frac{1 - (1 - \gamma A)^n}{\gamma A} N \langle \beta \rangle^2. \tag{47}$$

In (50) it is assumed that $\langle \underline{\in}_{1,0} \rangle = \underline{r}_1$, that is, the canceler is originally started with all of its tap coefficients equal to zero. A similar expression holds for the quadrature component.

The difference equation for the MSE in (45) now becomes

$$\langle e_{n+1}^2 \rangle = \langle e_n^2 \rangle [1 - 2\gamma A + 2N\gamma^2 A^2] + 2\gamma A \langle \xi^2 \rangle + 2AN \langle \beta_n^2 \rangle - 2A \left[(1 - \gamma A)^n \langle \underline{\beta}^T \rangle \underline{r}_1 - \frac{1 - (1 - \gamma A)^n}{\gamma A} N \langle \beta \rangle^2 \right] - 2A \left[(1 - \gamma A)^n \langle \underline{\beta}^T \rangle \underline{r}_2 - \frac{1 - (1 - \gamma A)^n}{\gamma A} N \langle \beta \rangle^2 \right]. \quad (48)$$

The above equation, although notationally complex, is only a first-order difference equation of the form

$$\langle e_{n+1}^2 \rangle = \langle e_n^2 \rangle \lambda_1 + k_1 - k_2 \lambda_2^n, \quad (49)$$

where

$$\lambda_1 = 1 - 2\gamma A + 2N\gamma^2 A^2 \quad (50)$$

$$\lambda_2 = 1 - \gamma A \quad (51)$$

$$k_1 = 2\gamma A \langle \xi^2 \rangle + 2AN \langle \beta^2 \rangle + \frac{4}{\gamma} N \langle \beta \rangle^2 \quad (52)$$

$$k_2 = 2A \langle \underline{\beta}^T \rangle \underline{r}_1 + 2A \langle \underline{\beta}^T \rangle \underline{r}_2 + \frac{4}{\gamma} N \langle \beta \rangle^2, \quad (53)$$

and where it is assumed that $\langle \beta_n^2 \rangle$ is a constant and equal to $\langle \beta^2 \rangle$. Assuming stability,

$$|\lambda_1| < 1 \quad \text{and} \quad |\lambda_2| < 1, \quad (54)$$

there are two cases to be considered in solving (53).

Case 1: $\lambda_1 \neq \lambda_2$ (nonresonant).

Assume

$$\langle e_n^2 \rangle = C_1 + C_2 \lambda_1^n + C_3 \lambda_2^n. \quad (55)$$

Then solving (49) by substitution, and assuming the initial condition $\langle e_0^2 \rangle$ gives

$$C_1 = \frac{k_1}{1 - \lambda_1} \quad (56)$$

$$C_3 = \frac{-k_2}{\lambda_2 - \lambda_1} = \frac{k_2}{\lambda_1 - \lambda_2} \quad (57)$$

$$C_2 = \langle e_0^2 \rangle - C_1 - C_3. \quad (58)$$

In terms of the original parameters in (48),

$$C_1 = \frac{\langle \xi^2 \rangle + \frac{N}{\gamma}\langle \beta^2 \rangle + \frac{2N}{\gamma^2 A}\langle \beta \rangle^2}{(1 - \gamma NA)} \tag{59}$$

$$C_3 = \frac{\frac{2}{\gamma}[\langle \underline{\beta}_T \rangle \underline{r}_1 + \langle \underline{\beta}^T \rangle \underline{r}_2] + \frac{4N}{\gamma^2 A} N \langle \beta \rangle^2}{(2N\gamma A - 1)} \tag{60}$$

$$C_2 = \langle e_0^2 \rangle - C_1 - C_3. \tag{61}$$

All of the parameters appearing in the expression for the MSE in (55) are now known. In the steady state, as n goes to infinity the MSE becomes

$$\langle e_\infty^2 \rangle = \frac{\langle \xi^2 \rangle + \frac{N}{\gamma}\langle \beta^2 \rangle + \frac{2N}{\gamma^2 A}\langle \beta \rangle^2}{1 - \gamma NA}. \tag{62}$$

Case 2: Resonance ($\gamma_1 = \gamma_2$).
In this case, from (50) and (51)

$$\gamma = \frac{1}{2NA}. \tag{63}$$

Assume

$$\langle e_n^2 \rangle = C_1 + C_2 \gamma_1^n + n C_e \gamma_1^n. \tag{64}$$

Solving (49) again by substitution, we get, after some algebra,

$$\begin{aligned}\langle e_n^2 \rangle &= 2\langle \xi^2 \rangle + 4NA\langle \beta^2 \rangle + 16N^2 A \langle \beta \rangle^2 \\ &\quad + [\langle e_0^2 \rangle - 2\langle \xi^2 \rangle - 4NA\langle \beta^2 \rangle - 16N^2 A \langle \beta \rangle^2] \left(1 - \frac{1}{2N}\right)^n \\ &\quad - n[2A[\langle \underline{\beta}^T \rangle \underline{r}_1 + \langle \underline{\beta}^T \rangle \underline{r}_2] + 8N^2 A \langle \beta \rangle^2] \left(1 - \frac{1}{2N}\right)^{n-1}.\end{aligned} \tag{65}$$

In this case the steady-state MSE is

$$\langle e_\infty^2 \rangle = 2\langle \xi^2 \rangle + 4NA\langle \beta^2 \rangle + 16N^2 A \langle \beta \rangle^2. \tag{66}$$

APPENDIX B

Bias-Free Two's-Complement Multiplication for a Data-Driven Canceler

In this Appendix, it is shown that the bias that occurs when rounding the product of a two's-complement multiplier can be eliminated by properly choosing the values of the binary input symbols. Let α be an $(l+1)$-bit fractional two's-complement number represented as

$$\alpha = a_0 a_1 a_2 \cdots a_l, \tag{67}$$

where $a_i = 0$ or 1, and a_l is the LSB of the number.
The numerical value of α is given by*

$$\alpha = -a_0 + \sum_{i=1}^{l} a_i 2^{-i}, \tag{68}$$

and the product of two such numbers becomes

$$\begin{aligned}\alpha \cdot x &= (-a_0 + \sum_{i=1}^{l} a_i 2^{-i})(-x_0 + \sum_{k=1}^{l} x_k 2^{-k}) \\ &= a_0 x_0 - a_0 \sum_{k=1}^{l} x_k 2^{-k} - x_0 \sum_{i=1}^{l} a_i 2^{-i} + \sum_{i=1}^{l} \sum_{k=1}^{l} a_i x_k 2^{-(i+k)}. \end{aligned} \tag{69}$$

This can be rewritten as a two's-complement number

$$\alpha \cdot x = -c_0 + \sum_{i=1}^{2l} c_i 2^{-i}, \tag{70}$$

where the summation on the right now goes to $2l$, making the length of the product $(2l + 1)$. A two's-complement multiplier usually has one of two means of reducing this product to a $(l + 1)$-bit number. In truncation, all the bits corresponding to $i \geq l + 1$ are discarded. From (70), it is seen that such an operation always decreases the magnitude of a positive number, and increases the magnitude of a negative number, thus introducing a negative bias. In rounding a number, $2^{-(l+1)}$ is added to the product in (70) and the result is truncated to $(l + 1)$ bits. This always selects the $(l + 1)$-bit number that is closest in magnitude to the true product. An ambiguity arises, however, when this product is equidistant from two $(l + 1)$-bit numbers. It is seen from (69) that rounding, in this case, always increases the magnitude of a positive number and decreases the magnitude of a negative number, thus introducing a positive bias in the arithmetic.

It is assumed that the numbers $\pm\alpha$ and x in (69) represent the symbols and the scaled error, respectively, in the updating algorithm. The bias situation in rounding arises when the following conditions occur in (70):

$$c_{l+1} = 1 \quad \text{and} \quad c_i = 0 \quad \text{for} \quad i > l + 1. \tag{71}$$

Inspection of (69) shows that these conditions are equivalent to

$$\sum_{i=1}^{l} \sum_{k=l+1-i}^{l} a_i x_k 2^{-(i+k)} = 2^{-(l+1)}. \tag{72}$$

---

* It is readily verified that, with this definition, the largest negative number that can be represented is -1, and the largest positive number is $+1 - \text{LSB}$.

The rounding bias can be eliminated by choosing the a_i's in such a manner as to never satisfy (72), regardless of the choice of the x_k's. The search for all of the possible numbers, α, having this property is tedious and will not be pursued here. Rather, two sets of symbols are proposed and shown to have the desired property. The proof requires that the scaled error be reasonably small. More specifically, the magnitude of x must be strictly less than one-half. This assumption is always satisfied in practice, even during start-up.

The first two binary symbols to be considered are defined by

$$\alpha = 0100 \cdots 001 = 2^{-1} + 2^{-l} \tag{73}$$

and

$$-\alpha = 1011 \cdots 111 = -1 + \sum_{i=2}^{l} 2^{-i}. \tag{74}$$

These numbers are equal to $\{\pm((1/2) + \text{LSB})\}$. Considering the positive symbol and replacing the a_i's by their value in (72) yields

$$x_l 2^{-(l+1)} + \sum_{k=1}^{l} x_k 2^{-(l+k)} = 2^{-(l+1)}. \tag{75}$$

This equation can only be satisfied under the condition $x_1 = 1$ and $x_i = 0$ for $i > 1$. From (68), it is seen that the only two numbers satisfying this condition are $\pm 1/2$. These numbers were discarded earlier as valid solutions. For a negative symbol (72) becomes

$$\sum_{i=2}^{l} \sum_{k=l+1-i}^{l} x_k 2^{-(i+k)} = 2^{-(l+1)}. \tag{76}$$

A sequence of x_k's cannot be synthesized to satisfy this equation. The lowest-order bit on the left corresponds to the power $2l$ of 1/2. There is only one term on the left in (76) contributing to this bit's value: $i = k = l$. Since there is no such term on the right, this bit must be zero ($x_l = 0$). The value of second-lowest bit corresponding to a power $(2l-1)$ depends upon two terms: $i = l$, $k = l - 1$ and $i = l - 1$. Therefore, its value is $(x_l + x_{l-1})$, which must be zero modulo-2. Since $x_l = 0$, this implies $x_{l-1} = 0$. Reasoning by induction shows that all of the x_i's must be zero for $i \geq 2$. The highest-order bit corresponds to the $(l + 1)$th power of 1/2. Combining this with the preceding result, we see that only one term contributes to its value; $i = l$, $k = 1$. Therefore, to satisfy (76), x_1 must equal one. As previously noted, the only two numbers that produce the bias situation are $\pm 1/2$, which are not valid solutions.

The second set of symbols that eliminates the multiplier's bias is given by

$$\alpha = 0111 \cdots 111 = 1 - 2^{-l} \tag{77}$$

and

$$-\alpha = 1000 \cdots 001 = -1 + 2^{l}. \tag{78}$$

These numbers are equal to $\{\pm(1\text{-LSB})\}$. Proof by contradiction, similar to the one utilized previously, can be used to show that these symbols eliminate the multiplier's bias. The proof is left as an exercise for the reader.

AUTHORS

John M. Cioffi, B.S. (Electrical Engineering), 1978, University of Illinois; M.S. and Ph.D. (Electrical Engineering), Stanford University, in 1979 and 1984, respectively; Bell Laboratories, 1979–81; IBM Research, 1984—. In 1978 Mr. Cioffi received the Jordan and Davidson awards from the University of Illinois as the outstanding senior in Engineering and the Valedictorian of Engineering, respectively. He is the author of numerous publications in the areas of adaptive filtering and data transmission and storage. Member, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi, Phi Eta Sigma, Sigma Xi, IEEE.

Jean-Jacques Werner, Ing. Deg., 1965, INSA, Lyon, France; M.S. (Electrical Engineering), 1967, Laval University, Canada; Eng. Sc.D. (Electrical Engineering), 1973, Columbia University; Bell Laboratories, 1973–1982; AT&T Information Systems, 1983—. At Bell Laboratories and AT&T Information Systems, Mr. Werner has worked on problems in data transmission and digital signal processing. Member, Sigma Xi, IEEE.

PAPERS BY AT&T BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Chang J. M., Maxemchuk N. F., **Reliable Broadcast Protocols.** ACM T Comp 2(3):251–273, Aug 1984.

Cody W. J. et al., **A Proposed Radix-Independent and Word-Length-Independent Standard for Floating-Point Arithmetic.** IEEE Micro 4(4):86–100, Aug 1984.

Fishburn P. C., **Elements of Risk Analysis in Non-Linear Utility Theory.** Infor 22(2):81–97, May 1984.

Kapur R. N., Browne J. C., **Techniques for Solving Block Tridiagonal Systems on Reconfigurable Array Computers.** SIAM J Sci 5(3):701–719, Sep 1984.

Pierce J. R., **More on the Fifth Generation (Letter).** Abacus-NY 1(4):8, Sum 1984.

Prell E. M., Sheng A. P., **Building Quality and Productivity Into a Large Software System.** IEEE Softw 1(3):47–54, Jul 1984.

Whitt W., **The Amount of Overtaking in a Network of Queues.** Networks 14(3):411–426, Fal 1984.

ENGINEERING

Acampora A. S., Hluchyj M. G., **A New Local Area Network Architecture Using a Centralized Bus.** IEEE Comm M 22(8):12–21, Aug 1984.

Alferness R. C., Buhl L. L., **Low-Crosstalk Waveguide Polarization Multiplexer/Demultiplexer for $\lambda = 1.32$ μm.** Optics Lett 9(4):140–142, 1984.

Arnold H. W., Bodtmann W. F., **Switched-Diversity FSK in Frequency-Selective Rayleigh Fading.** IEEE Veh T 33(3):156–163, Aug 1984.

Baumert R. J., Cameron L. E., Wilson R. A., **A Mixed EFL I^2L Digital Telecommunication Integrated Circuit.** IEEE Device 31(2):160–165, 1984.

Chin A. K., Caruso R., Young M. S. S., Vonneida A. R., **Uniformity Characterization of Semi-Insulating GaAs by Cathodoluminescence Imaging.** Appl Phys L 45(5):552–554, Sep 1, 1984 .

Chraplyvy A. R., Marcuse D., Henry P. S., **Carrier-Induced Phase Noise in Angle-Modulated Optical-Fiber Systems.** J Lightw T 2(1):6–10, Feb 1984.

Fisher R. E., **UHF Television Interference Associated With Cellular Mobile Telephone Systems.** IEEE Veh T 33(3):244–249, Aug 1984.

Forrest S. R., **Gain-Bandwidth-Limited Response in Long-Wavelength Avalanche Photodiodes.** J Lightw T 2(1):34–39, Feb 1984.

Haight R., Bokor J., Storz R. H., Stark J. B., Bucksbaum P. H., Freeman R. R., **Photoemission Apparatus Using XUV Harmonics of a Picosecond KRF Laser.** P Soc Photo 476:61–64, 1984.

Hartman D. H., **Multiple Reflection Noise Production on Transmission-Systems With Periodic or Quasi-Periodic Impedance Disturbances—A Time-Domain Approach.** IEEE Circ S 31(10):866–875, Oct 1984.

Hayes J. R., Leheny R. F., Temkin H., Gossard A. C., Wiegmann W., **Electroluminescence From a Heterojunction Bipolar Transistor.** Appl Phys L 45(5):537–539, Sep 1, 1984 .

Levy U., Logan R. A., Niv Y., **Laser Cathode-Ray Tube Operation at Room Temperature.** Appl Phys L 45(5):497–499, Sep 1, 1984 .

Linke R. A., **Direct Gigabit Modulation of Injection Lasers—Structure-Dependent Speed Limitations.** J Lightw T 2(1):40–43, Feb 1984.

Liu P. L., Ogawa K., **Statistical Measurements as a Way to Study Mode Partition in Injection Lasers.** J Lightw T 2(1):44–48, Feb 1984.

McCaughan L., **Low-Loss Polarization-Independent Electrooptical Switches at Lambda = 1.3 μm.** J Lightw T 2(1):51–55, Feb 1984.

Seth S. C., Agrawal V. D., **Characterizing the LSI Yield Equation From Wafer Test Data.** IEEE Comp A 3(2):123–126, Apr 1984.

Stone J., Marcuse D., **Direct Measurement of Second-Order Dispersion in Short Optical Fibers Using White-Light Interferometry.** Electr Lett 20(18):751–752, Aug 30, 1984 .

Winters J. H., **Optimum Combining in Digital Mobile Radio With Cochannel Interference.** IEEE Veh T 33(3):144–155, Aug 1984.

Yeh Y. S., Wilson J. C., Schwartz S. C., **Outage Probability in Mobile Telephony With Directive Antennas and Macrodiversity.** IEEE Veh T 33(3):123–127, Aug 1984.

MANAGEMENT/ECONOMICS

Doshi B. T., Lipper E. H., **The Throughput Performance of a Prioritized LIFO Service Discipline.** Oper Res L 3(2):75–80, Jun 1984.

PHYSICAL SCIENCES

Batlogg B., Remeika J. P., Cooper A. S., Fisk Z., **Magnetism and Superconductivity in $CeCu_2Si_2$ Single Crystals.** J Appl Phys 55(6):2001–2003, 1984.

Belfiore L. A., Schilling F. C., Tonelli A. E., Lovinger A. J., Bovey F. A., **Magic Angle Spinning Carbon-13 NMR Spectroscopy of Three Crystalline Forms of Polybutene-1.** Polym Prepr 25(1):351–353, 1984.

Beni G., Hackwood S., **Intermittent Turbulence and Period Doubling at the Corrosion Passivity Transition in Iron.** J Appl Elec 14(5):623–626, Sep 1984.

Benton J. L., Levinson M., Macrander A. T., Temkin H., Kimerling L. C., **Recombination Enhanced Defect Annealing in N-InP.** Appl Phys L 45(5):566–568, Sep 1, 1984 .

Bergmann E. E., McCaughan L., Watson J. E., **Coupling of Intersecting Ti-$LiNbO_3$ Diffused Waveguides.** Appl Optics 23(17):3000–3003, Sep 1, 1984 .

Caton J. A., Heywood J. B., Mendillo J. V., **Hydrocarbon Oxidation in a Spark-Ignition Engine Exhaust Port.** Comb Sci T 37(3–4):153–169, 1984.

Davis G. P., Moore C. A., Gottscho R. A., **Dynamics of Laser Stimulated Etching of Germanium by Bromine.** J Appl Phys 56(6):1808–1811, Sep 15, 1984.

Donnelly V. M., Karlicek R. F., **Excimer Laser Enhancement and Probing of III-V Compound Semiconductor Chemical Vapor Deposition.** P Soc Photo 476:102–109, 1984.

Dutta N. K., Napholtz S. G., Yen R., Brown R. L., Shen T. M., Olsson N. A., Craft D. C., **1.3 μm InGaAsP DCPBH Multiquantum-Well Lasers.** Electr Lett 20(18):727–728, Aug 30, 1984.

Freund R. S., Donohue D. E., Fisanick G. J., **Polarization Line Shape of Balmer-β From Electron Impact Dissociation of H^2.** J Chem Phys 80(5):1754–1759, 1984.

Gershenfeld N., **The Measurement of Optically Thick Atomic Vapor Densities by the Nonlinear Least-Squares Fitting of Absorption or Fluorescence Spectra (Letter).** Nucl Inst A 224(3):570–572, Jul 15, 1984 .

Harrus A., Mihalisin T., Batlogg B., **Saturation of Resistivity in the $La_{1-x}Ce_xRh_2$ Mixed Valent Kondo System.** J Appl Phys 55(6):1993–1995, 1984.

Harrus A., Timlin J., Mihalisin T., Batlogg B., **Resistivity, Susceptibility, and Superconductivity in Mixed Valent $Ce(Rh_{1-x}Ru_x)_2$[a].** J Appl Phys 55(6):1990–1992, 1984.

Hodeau J. L., Marezio M., Remeika J. P., **The Structure of $[Er(1)_{1-x}, Sn(1)_x]$ $Er(2)_4Rh_6Sn(2)_4Sn(3)_{12}Sn(4)_2$, a Ternary Reentrant Superconductor.** Act Cryst B 40(Feb):26–38, 1984.

Johnson A. M., Glass A. M., Olson D. H., Simpson W. M., Harbison J. P., **High Quantum Efficiency A-Si-H Picosecond Transit Time Limited Schottky-Barrier Photodetectors.** J Non-Cryst 66(1–2):381–386, Jul 1984.

Kahane C., Frerking M. A., Langer W. D., Encrenaz P., Lucas R., **Measurement of the Formaldehyde Ortho to Para Ratio in Three Molecular Clouds.** Astron Astr 137(2):211–222, Aug 1984.

Kash K., Shah J., **Carrier Energy Relaxation in $In_{0.53}Ga_{0.47}$ as Determined From Picosecond Luminescence Studies.** Appl Phys L 45(4):401–403, Aug 15, 1984 .

Lang D. V., Cohen J. D., Harbison J. P., Chen M. C., Sergent A. M., **Resolution of the A-Si-H DLTS Energy Scale Controversy.** J Non-Cryst 66(1–2):217–222, Jul 1984.

Levy R. A., Green M. L., Gallagher P. K., **Characterization of LPCVD Aluminum for VLSI Processing.** J Elchem So 131(9):2175–2182, Sep 1984.

Mazumdar P., Bhagat S. M., Manheimer M. A., Chen H. S., **Low Field Magnetization Studies on $(Fe_xNi_{1-x})_{75}P_{16}B_6Al_3$ Ribbons.** J Appl Phys 55(6):1685–1687, 1984.

McNevin S. C., Becker G. E., **CF_4/Silicon Surface Reactions: Evidence for Parallel Etching Mechanisms From Modulated Ion Beam Studies.** J Vac Sci B 2(1):27–33, 1984.

Mims W. B., **Elimination of the Dead-Time Artifact in Electron Spin-Echo Envelope Spectra.** J Magn Res 59(2):291–306, Sep 1984.

Murarka S. P., **Effect of Oxygen Contamination on the Properties of Cosputtered Tantalum Silicide.** Appl Phys L 45(4):392–394, Aug 15, 1984 .

Smith J. L., Fisk Z., Willis J. O., Batlogg B., Ott H. R., **Impurities in the Heavy-Fermion Superconductor UBe_{13}.** J Appl Phys 55(6):1996–2000, 1984.

Sodini C. G., Ko P. K., Moll J. L., **The Effect of High Fields on MOS Device and Circuit Performance.** IEEE Device 31(10):1386–1393, Oct 1984.

Stern M. B., Craighead H. G., Liad P. F., Mankiewich P. M., **Fabrication of 20-nm Structures in GaAs.** Appl Phys L 45(4):410–412, Aug 15, 1984 .

Tonelli A. E., Schilling F. C., **^{13}C-NMR Chemical Shifts and the Microstructure of Propylene-Vinyl Chloride Copolymer With Low Propylene Content.** Polym Prepr 25(1):334–335, 1984.

Tyson J. A., **Comparison of Space Telescope and 4-Meter Ground-Based Telescope—Faint Galaxy Detection and Photometry.** Pub Ast S P 96(581):566–573, Jul 1984.

Weschler C. J., **Indoor Outdoor Relationships for Nonpolar Organic Constituents of Aerosol Particles.** Env Sci Tec 18(9):648–652, Sep 1984.

Weschler C. J., **Sulfur-Dioxide Content of Mount St. Helens Ash.** J Geo Res-A 89(ND3):4891–4894, Jun 20, 1984 .

Wicksted J. P., Shapiro S. M., Chen H. S., **Investigation of the Ferromagnetic-Spin Glass Transition in α-$(Fe_{77}Cr_{23})_{75}P_{16}B_6Al_3$.** J Appl Phys 55(6):1697–1699, 1984.

Yamada Y., Fujii Y., Akahama Y., Endo S., Narita S., Axe J. D., McWhan D. B., **Lattice-Dynamical Properties of Black Phosphorus Under Pressure Studied by Inelastic Neutron Scattering.** Phys Rev B 30(5):2410–2413, Sep 1, 1984 .

Zucker J. E., Pinczuk A., Chemla D. S., Gossard A., Wiegmann W., **Optical Vibrational Modes and Electron-Phonon Interaction in GaAs Quantum Wells.** Phys Rev L 53(13):1280–1283, Sep 24, 1984 .

SOCIAL AND LIFE SCIENCES

Shute S. J., Starr S. J., **Effects of Adjustable Furniture on VDT Users.** Human Fact 26(2):157–170, Apr 1984.

CONTENTS, FEBRUARY 1985

Part 1

Part 2

COMPUTING SCIENCE AND SYSTEMS

Numerical Computation of Delays in Clocked Schedules
M. H. Ackroyd

Analysis of Clocked Schedules—High-Priority Tasks
B. T. Doshi

AT&T TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


AT&T