Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, AT&T Bell Laboratories Technical Journal, Room 1H321, 101 J. F. Kennedy Pky, Short Hills, NJ 07078. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of AT&T. Comments selected for publication may be edited for brevity, subject to author approval.

Heavy-Traffic Approximations for Service Systems With Blocking

By W. WHITT*

(Manuscript received October 27,1983)

This paper develops approximations for the blocking probability and related congestion measures in service systems with s servers, r extra waiting spaces, blocked customers lost, and independent and identically distributed service times that are independent of a general stationary arrival process (the G/GI/s/r model). The approximations are expressed in terms of the normal distribution and the peakedness of the arrival process. They are obtained by applying previous heavy-traffic limit theorems and a conditioning heuristic. There are interesting connections to Hayward's approximation, generalized peakedness, asymptotic expansions for the Erlang loss function, the normal-distribution method, and bounds for the blocking probability. For the case of no extra waiting space, a renewal arrival process and an exponential service-time distribution (the GI/M/s/0 model), a heavy-traffic local limit theorem by A. A. Borovkov implies that the blocking depends on the arrival process only through the first two moments of the renewal interval as the offered load increases. Moreover, in this situation Hayward's approximation is asymptotically correct.

I. INTRODUCTION AND SUMMARY

In this paper we introduce and investigate approximations for congestion measures in G/GI/s/r service systems, which have s servers, r extra waiting spaces, the first-come first-served discipline, and independent and identically distributed (i.i.d.) service times with a general distribution that are independent of a general stationary

---

arrival process. Customers arriving when all s servers are busy and all r waiting spaces are full are blocked; they leave without receiving service and without affecting future arrivals (no retrials). We primarily focus on the case $r = 0$ (except for Section VII). We present approximate expressions for the proportion of arriving customers that are blocked (call congestion) and the proportion of time that the system is full (time congestion). We also approximate the distributions of the number of customers in the system at arrival epochs and at arbitrary times.

We obtain our approximations by applying previous heavy-traffic limit theorems[1-4] and a conditioning heuristic (Section III). As with much of the earlier work on this problem, we are not able to present a completely rigorous development, but we believe that we have a novel perspective that provides additional insight. There are interesting connections to earlier work, including Hayward's approximation,[5] generalized peakedness,[6] asymptotic expansions for the Erlang loss function,[7-9] the normal distribution method,[10-14] and bounds for the blocking probability.[15-17]

Perhaps our most important contribution is to point out the significance of a heavy-traffic local limit theorem by Borovkov [Theorem 15(2) of Ref. 2], which was first published in Russian in 1972. ("Local" means that the limit is for the probability mass function instead of the cumulative distribution function.) For GI/M/s/0 models (no extra waiting space, renewal arrival process, and exponential service-time distribution), this theorem provides a rigorous basis for the approximations under heavy loads. For example, this theorem implies that Hayward's approximation [(22) in Section 6.2] is asymptotically correct as the offered load increases. Of course, this property is consistent with extensive numerical evidence, but apparently no mathematical proof has been given before.

Here is how this paper is organized. In Section II we review a heavy-traffic limit theorem for G/GI/∞ models that we will apply, which is also due to Borovkov.[1] In Section III we introduce a conditioning heuristic and apply it with the limit theorem in Section II to obtain an approximation for the distribution of the number of busy servers at an arbitrary time in the associated G/GI/s/0 system. In Section IV we use a conservation law plus the approximation in Section III to generate an approximation for the blocking probability in G/GI/s/0 systems. We also discuss an approximation for the distribution of the number of busy servers at arrival epochs. In Section V we state Borovkov's local heavy-traffic limit theorem for GI/M/s/0 models that supports the approximations. We also make several conjectures about related theorems.

In Section VI we discuss connections to other work. We indicate

that the normal approximation for the blocking probability in $M/M/s/0$ systems has a long history, going all the way back to Erlang.[18] In the Appendix we also give a simple proof of the heavy-traffic local limit theorem for $M/M/s/0$ systems, using the elementary central limit theorem and Stirling's formula.[19] In Section VI we also discuss connections to Hayward's approximation,[5,6] bounds for the blocking probability,[15-17] and previous normal approximations.[10-14]

In Section VII we indicate how the approach can be extended to systems with finite waiting rooms, drawing on Halfin and Whitt.[4] In doing so, we obtain a modification of Hayward's approximation for the case of a finite waiting room [see (42)]. Finally, in Section VIII we give the results of experiments testing the approximations for $G/M/s/0$ systems. As observed by Rahko,[10-12] Hertzberg,[13] and Delbrouck,[14] the normal approximation tends to work quite well except in low loads.

We close this introduction by noting that the general blocking problem discussed here continues to generate considerable attention; several related papers were presented at the Tenth International Teletraffic Congress at Montreal.[6,12,20-24] Another recent related contribution is Newell.[25]

II. THE INFINITE-SERVER MODEL IN HEAVY TRAFFIC

Consider the $G/GI/\infty$ service system, which has infinitely many servers and independent and identically distributed service times that are independent of a general stationary arrival process. Let $A(t)$ count the number of arrivals in the interval $[0, t]$ for $t \geq 0$. We assume that $A(t)$ satisfies a central limit theorem; i.e.,

$$[A(t) - \lambda t]/(\lambda c^2 t)^{1/2} \rightarrow N(0, 1) \qquad (1)$$

as $t \rightarrow \infty$, where $\rightarrow$ denotes convergence in distribution, $N(a, b)$ denotes a random variable with the normal distribution having mean a and variance b, and λ is the arrival rate. When $A(t)$ is a renewal process, c^2 in (1) is the squared coefficient of variation (variance divided by the square of the mean) of the renewal interval. More generally, the denominator in (1) typically is asymptotically equivalent to the standard deviation of $A(t)$, so that

$$c^2 = \lim_{t \to \infty} \text{var}[A(t)]/\lambda t = \lim_{t \to \infty} \text{var}[A(t)]/EA(t). \qquad (2)$$

The parameter c^2 in (1) and (2) is the basis for approximating the arrival process by a renewal process via the asymptotic method in Ref. 26.

Let μ^{-1} be the mean and $G(t)$ the cumulative distribution function (cdf) of the service-time distribution. Let $\alpha = \lambda/\mu$ be the offered load.

Our approximations are developed by considering limits as the offered load α increases. We fix the service-time cdf $G(t)$ and change α by changing λ.

Let X_α be the equilibrium number of busy servers in the G/GI/∞ system at an arbitrary time, as a function of α. (We assume that a unique equilibrium distribution exists—see Section 2.3 of Franken et al.[27]) Borovkov[1] proved the following heavy-traffic limit theorem for X_α. He showed that if (1) holds [actually a slightly stronger functional limit theorem for $A(t)$], then

$$(X_\alpha - \alpha)/\sqrt{\alpha z} \to N(0, 1) \tag{3}$$

as $\alpha \to \infty$, where

$$z = 1 + (c^2 - 1)\eta \tag{4}$$

and

$$\eta = \mu \int_0^\infty [1 - G(t)]^2 dt; \tag{5}$$

also see Refs. 3 and 6. Actually, Borovkov did not directly treat the equilibrium variable X_α, so that there is a further interchange of limits to get (3). For practical purposes, we can regard (3) as a consequence of (1).

To interpret (3) through (5), recall that $EX_\alpha = \alpha$ for all α, even with a general stationary arrival process. The parameter z in (4) is the asymptotic peakedness of the arrival process $A(t)$ with respect to the service-time cdf $G(t)$ because

$$z = \lim_{\alpha \to \infty} \text{var}(X_\alpha)/EX_\alpha = \lim_{\alpha \to \infty} \text{var}(X_\alpha)/\alpha; \tag{6}$$

see Ref. 6 and references there. In particular, z in (4) is $z_G(0+)$ in Section 3.3.1 of Ref. 6. Note that η in (5) has the maximum value of 1 attained by a deterministic service-time distribution (unit mass on μ^{-1}) and can assume any value in the interval (0, 1]. For example, $\eta = 2/3$, $1/2$, and $1/n$, respectively, when the distribution is uniform, exponential and concentrated on two points with mass n^{-1} on 0.

We have defined three parameters characterizing variability: c^2, η, and z. The parameter c^2 in (1) and (2) is a measure of the variability of the arrival process (over large time intervals). The parameter η in (5) is a measure of the variability of the service-time distribution and the parameter z in (4) is a second measure of variability of the arrival process (as measured by the G/GI/∞ model with service-time cdf G having variability parameter η in heavy traffic). The heavy-traffic peakedness z in (4) is increasing in c^2, but whether it is increasing or decreasing in η depends on the sign of $c^2 - 1$. For a Poisson process,

$c^2 = 1$. In general, increased service-time variability as expressed by decreasing η in (5) tends to make z in (4) closer to one. Hence, if the arrival process is more variable or bursty than a Poisson process in the sense that $c^2 > 1$, then z in (4) is increasing in η, which means that the peakedness increases as the variability of the service-time distribution decreases. When $c^2 > 1$, the deterministic service-time cdf gives the largest heavy-traffic peakedness among all service-time cdf's with the same mean. This phenomenon seems to have been first discussed by Wolff[28] (see also Section 3.3.1 of Ref. 6). Related results about GI/G/1 queues are contained in Whitt.[29]

III. THE CONDITIONING HEURISTIC

We now use the heavy-traffic limit theorem (3) for the G/GI/∞ system to produce approximations for the associated G/GI/s/0 loss system, which has s servers, no extra waiting space, and the same arrival process and service-time distribution. Our starting point is a basic property of the Markovian M/M/s/0 models in which the number of customers in the system evolves as a birth-and-death process. The equilibrium probability $p_k(s_1)$ that there are k customers in an M/M/s_1/0 model is just the conditional probability that there are k customers in an M/M/s_2/0 model given that there are no more than s_1 customers, for any s_2 with $s_1 \leq s_2 \leq \infty$; in other words, $p_k(s_1)$ is obtained by truncating and renormalizing $p_k(s_2)$:

$$p_k(s_1) = p_k(s_2) \bigg/ \sum_{j=0}^{s_1} p_j(s_2) \tag{7}$$

for $0 \leq k \leq s_1 \leq s_2$. Truncation formula (7) is an immediate consequence of the known formula for $p_k(s)$, but also is easily derived for more general birth-and-death processes (see p. 68 of Heyman and Sobel).[30]

Let Y_α be the equilibrium number of busy servers at an arbitrary time in the G/GI/s/0 model. (We also assume existence and uniqueness—see Section 2.3.2 of Ref. 27.) As an approximation, we assume that Y_α is related to X_α of Section II by the same conditioning formula (7). In particular, we suggest the heuristic approximation

$$P(Y_\alpha = k) \approx P(X_\alpha = k)/P(X_\alpha \leq s) \tag{8}$$

for $0 \leq k \leq s$. This conditioning approximation is no doubt an old idea, which would be hard to trace; it was used previously by Jagerman to develop approximate blocking formulas in the case of nonstationary Poisson traffic.[31]

Next we obtain a further approximation by invoking the heavy-traffic limit theorem for X_α in (3). For this purpose, let $\Phi(t)$ be the standard normal cdf, i.e., of $N(0, 1)$, and $\phi(t)$ the associated density. We combine (3) and (8) to obtain

$$P(Y_\alpha = k) \approx \frac{P[k - 0.5 \leq N(\alpha, \alpha z) \leq k + 0.5]}{P[N(\alpha, \alpha z) \leq s + 0.5]}$$

$$\approx \frac{\Phi\left(\dfrac{k - \alpha + 0.5}{\sqrt{\alpha z}}\right) - \Phi\left(\dfrac{k - \alpha - 0.5}{\sqrt{\alpha z}}\right)}{\Phi\left(\dfrac{s - \alpha + 0.5}{\sqrt{\alpha z}}\right)}$$

$$\approx (1/\sqrt{\alpha z})\phi\left(\frac{k - \alpha}{\sqrt{\alpha z}}\right) \bigg/ \phi\left(\frac{s - \alpha + 0.5}{\sqrt{\alpha z}}\right). \qquad (9)$$

The time congestion, say B_T, is defined as $B_T = P(Y_\alpha = s)$, so that an approximation for it is obtained from (9) simply by setting $k = s$.

In further support of the conditioning heuristic, we note that it is also valid for the diffusion process limits that arise in several cases of heavy traffic, e.g., for the stochastic-process version of (3) in the case of exponential service-time distributions (see Ref. 1 and 3).

We have applied the conditioning heuristic to the distributions at an arbitrary time instead of at arrival epochs. Since Poisson arrivals see time averages,[27,30,32] these two distributions are the same for M/M/s/0 systems. Also, the heavy-traffic limits for these distributions are the same for G/GI/∞ system. (See Ref. 3 for the renewal arrival process case.) However, these distributions are definitely not the same for the G/GI/s/r system or even the GI/M/s/0 special case. The conditioning heuristic seems to perform much better when applied to the distribution at an arbitrary time, as will be clear from the next two sections. This might not be surprising, but a good explanation is still needed.

IV. A CONSERVATION LAW AND THE BLOCKING APPROXIMATION

Let B_C be the probability that an arriving customer in the G/GI/s/0 system is blocked (call congestion). A basic conservation law enables us to express B_C in terms of EY_α. In particular, since the average rate of accepted arrivals equals the average departure rate, not counting lost calls (see Heyman[17]),

$$\lambda(1 - B_C) = \mu EY_\alpha. \qquad (10)$$

Hence, we can combine (9) and (10) to obtain an approximation for B_C.

It is well known and easy to show that

$$E[N(0, 1) | N(0, 1) \leq \theta] = -\phi(\theta)/\Phi(\theta), \qquad (11)$$

so that

$$EY_\alpha \approx \alpha - \sqrt{\alpha z}\phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right)\bigg/\Phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right) \qquad (12)$$

and

$$B_C = 1 - \alpha^{-1}EY_\alpha \approx \sqrt{z/\alpha}\phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right)\bigg/\Phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right) \approx zB_T \quad (13)$$

(e.g., see Appendix A of Delbrouck[14]).

We thus suggest (9) as an approximation for $P(Y_\alpha = k)$, (9) with $k = s$ for an approximation for B_T, and (13) as an approximation for B_C.

Let Z_α be the equilibrium number of busy servers in the G/GI/s/0 system at arrival epochs, again as a function of the offered load α. For G/M/s/0 systems (exponential service-time distributions, but still general stationary arrival process), we have the exact relationship

$$P(Z_\alpha = k - 1) = kP(Y_\alpha = k)/\alpha \qquad (14)$$

for $1 \leq k \leq s$ (p. 113 of Franken et al.[27]), which is a refinement of (10), because (10) is obtained from (14) by simply summing over k. We thus propose (14) as an approximation for G/M/s/0 systems and also G/GI/s/0 systems. Improved approximations for the nonexponential service-time distribution case can perhaps be obtained from the relationships in Section 4.3.2 of Ref. 27, but this does not appear easy.

These approximations are expressed in terms of the asymptotic peakedness z in (4). However, if this parameter is not available, other expressions for the peakedness can be used instead.[6] For example, the general formula for the peakedness of a renewal arrival process with respect to an exponential service-time distribution is given here in (27).

Formulas (4) and (5) can also be used to calculate a revised peakedness if there is a change in the service-time distribution, as suggested in Section 6 of Ref. 33 and as has been done in practice. Suppose that z_1 has been previously determined based on the service-time cdf G_1, but now we are going to consider the same G/GI/s/0 system with new service-time cdf G_2. For this purpose let $\eta_i = \eta(G_i)$ in (5). Using (4), we obtain an approximation for c^2, namely,

$$c^2 = 1 + (z_1 - 1)/\eta_1. \qquad (15)$$

We can thus approximate z_2 based on (4), (5) and

$$z_2 = 1 + (c^2 - 1)\eta_2 = 1 + (z_1 - 1)\eta_2/\eta_1. \qquad (16)$$

More general transformations based on Mellin transforms have also been developed by Jagerman for the entire peakedness functionals

$z[F]$ considered in Ref. 6 to describe the effect of changing the service-time distribution.[34]

Approximation formulas (9), (13), and (14) can also be used to generate approximate multiplicative correction factors to be used with the exact $M/M/s/0$ formulas. For example, we can obtain our approximation by multiplying the exact blocking probability for the $M/M/s/0$ system by the ratio $B_C(z)/B_C(1)$, where $B_C(z)$ is B_C in (13) as a function of z in (4). This procedure is slightly more complicated, but it is exact for $M/M/s/0$ systems.

V. BOROVKOV'S HEAVY-TRAFFIC LOCAL LIMIT THEOREM

A rigorous justification of (13) is provided by Theorem 15 (2), p. 226, of Borovkov.[2] For the $GI/M/s/0$ model, Borovkov established that

$$\lim_{\alpha \to \infty} \sqrt{\alpha} B_C = \sqrt{z}\phi(\beta/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{17}$$

if $(s - \alpha)/\sqrt{\alpha} \to \beta$ as $\alpha \to \infty$, where $z = (c^2 + 1)/2$ as specified in (4) in the case of an exponential service-time distribution.

Borovkov identifies z in (17) as $(c^2 + 1)/2$ rather than the heavy-traffic peakedness in (4), but we believe (4) is appropriate for generalizations to nonexponential service-time distributions and nonrenewal arrival processes.

Borovkov's arguments can also be applied to yield a related local limit theorem for $P(Z_\alpha = k)$ in $GI/M/s/0$ systems, namely,

$$\lim_{\alpha \to \infty} \sqrt{\alpha} P(Z_\alpha = k) = \sqrt{z}\phi(\beta'/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{18}$$

if $(s - \alpha)/\sqrt{\alpha} \to \beta$ and $(k - \alpha)/\sqrt{\alpha} \to \beta' < \beta$ as $\alpha \to \infty$. We can then apply (14) to deduce that

$$\lim_{\alpha \to \infty} \sqrt{\alpha} P(Y_\alpha = k) = (1/\sqrt{\alpha z})\phi(\beta'/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{19}$$

and

$$\lim_{\alpha \to \infty} \sqrt{\alpha} B_T = (1/\sqrt{\alpha z})\phi(\beta/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{20}$$

under the same asymptotic conditions. The limit in (19) coincides with Theorem 16, p. 232, of Borovkov,[2] which is stated there without proof.

Hence, we have theoretical justification for all our approximations in the case of $GI/M/s/0$ systems. We conjecture that (17) through (20) are still valid for general stationary arrival processes and nonexponential service-time distributions, i.e., under the conditions for Borovkov's $G/GI/\infty$ theorem (3). Since (14) is valid for general nonrenewal arrival

processes but not for general service-time distributions, the conjecture seems much more likely to be true for the generalization of the arrival process than for the generalization of the service-time distribution.

VI. CONNECTIONS TO OTHER WORK

6.1 Beginning with Erlang

For the Markovian M/M/s/0 model as well as the M/G/s/0 model, the exact blocking probability is given by the classic Erlang loss (B) formula

$$B(s,\ \alpha) = (\alpha^s/s!) \Big/ \sum_{k=0}^{s} (\alpha^k/k!). \tag{21}$$

Since Poisson arrivals see time averages,[27,30,32] $B_C = B_T$ and (18) and (19) coincide in this case. In this case, approximation (9) reduces to a rather standard normal approximation which was evidently known in 1924 by Erlang.[18] The M/M/s/0 case of the limit theorem (17) plus various asymptotic expansion improvements were given by Brockmeyer[7] and Vaulot.[8] Asymptotic expansion improvements also follow from Theorem 14 of Jagerman[9] and are discussed on p. 88 of Delbrouck.[14] A simple proof of the M/M/s/0 version of (17) using the central limit theorem and Stirling's formula[19] is given here in the Appendix.

6.2 The Hayward approximation

A relatively simple approximation for the blocking probability B_C in G/M/s/0 and even G/G/s/0 systems as a function of s, α and z proposed by Hayward is

$$B_C \equiv B_C(s,\ \alpha,\ z) \approx B_C(s/z,\ \alpha/z,\ 1) = B(s/z,\ \alpha/z) \tag{22}$$

using (21) (see Fredericks[5] and Eckberg[6]). It is significant that the approximation (9) and the limit theorem (17) are consistent with (22); for those expressions, $B(s,\ \alpha,\ z) \approx B(s/z,\ \alpha/z,\ 1)$.

There are several possible interpretations and applications. We can interpret (17) through (20) as additional evidence in support of the Hayward approximation. The limit theorems and approximations provide additional evidence that the Hayward approximation should perform well under heavy loads. The Hayward approximation is particularly appealing, given that there are convenient computer programs to calculate $B(s,\ \alpha)$ in (21) extended to nonintegral s, as have been developed by Jagerman.[35]

On the other hand, we can use the Hayward approximation as an additional justification for (9). We can derive (9) by applying the limit

theorem for M/M/s/0 systems described in Section 6.1 together with the Hayward approximation.

The connection also suggests that improved approximations can be obtained for G/M/s/0 systems and possibly G/G/s/0 systems by applying the asymptotic expansions of Brockmeyer,[7] Vaulot,[8] or Jagerman[9] for $B(s, \alpha)$ after applying the Hayward approximation to treat the peakedness z. In particular, formula (5) in Jagerman,[9] which is based on Theorem 14 there, seems particularly promising in combination with Hayward's approximation. However, testing remains to be done. Of course, improved approximations would also be obtained from asymptotic expansions related to (17). This seems to be a promising direction of research.

6.3 Bounds for the blocking probability

Sobel[16] and Heyman[17] recently established a lower bound for the blocking probability in a G/G/s/r system when $\rho > 1$, namely,

$$B_C \geq 1 - \rho^{-1}, \tag{23}$$

and observed that the lower bound often is a good approximation when $\rho > 1$.

We partly explain why the lower bound is a good approximation by showing that it appears in limiting lower and upper bounds as $\alpha \to \infty$ in our heavy-traffic approximation (9) for G/G/s/0 systems. (This paper is a revised version of Ref. 11 in Sobel,[16] where our result is mentioned.)

We use the familiar bounds for the tail of a normal distribution

$$(x^{-1} - x^{-3})\phi(x) < 1 - \Phi(x) = \Phi(-x) < x^{-1}\phi(x), \qquad x > 0; \tag{24}$$

see p. 175 of Feller.[19] To make the connection to (9) and (17), let $x = (\alpha - s)/\sqrt{\alpha z}$. Since $x\sqrt{z/\alpha} = (1 - \rho^{-1})$, and

$$1 \leq \frac{x\phi(x)}{\Phi(-x)} \leq (1 - x^{-2})^{-1} \tag{25}$$

by (24), from (9) we obtain the approximate bounds as $x \to \infty$

$$(1 - \rho^{-1}) \leq B_C \leq (1 - \rho^{-1})(1 - x^{-2})^{-1}, \tag{26}$$

which are useful for $\rho \geq 1$. The distance between the bounds goes to zero as $x \to \infty$.

Holtzman also establishes bounds for B_C in the GI/M/s/0 system.[15] In fact, he described the range of all possible values of B_C given only the offered load α and the peakedness z' of a renewal process, which is

$$z' = [1 - \phi(\mu)]^{-1} - \alpha, \tag{27}$$

where

$$\phi(\mu) = \int_0^\infty e^{-\mu t}dF(t) \tag{28}$$

with $F(t)$ the cdf of an interarrival time. By (17), we know that the range approaches 0 as $\alpha \to \infty$ with $(s - \alpha)/\sqrt{\alpha} \to \beta$. The z' in (27) approaches z in (4) and, by (9) and (17), B_C depends only on α, s and z for large α.

6.4 The normal-distribution method

The approximations here for $P(Y_\alpha = k)$, B_T, and B_C in (9), (13), and (14) coincide with the normal-distribution method (NDM) of Rahko[10-12] and Hertzberg[13] and the normal approximation for the Bernoulli-Poisson-Pascal (BPP) approximation of Delbrouck,[14] but the analysis here is different.

6.5 Light-traffic approximations and interpolations

The approximations here have been developed by considering the service systems in heavy traffic, i.e., as the offered load increases. Improved approximations for lighter loads may be possible by considering the service systems in light traffic, i.e., as the offered load decreases. Better approximations might be obtained by making interpolations between light and heavy traffic. This seems to be another promising direction for future research. Previous work on light-traffic approximations for queues is contained in Bloomfield and Cox,[36] Newell,[37] and Burman and Smith.[38,39] Interpolations between light and heavy traffic have been considered by Burman and Smith[39] and Reiman and Simon.[40] The hybrid approximations for queues with superposition arrival processes developed by Albin[41] and used in Ref. 42 to approximate networks of queues are also in this spirit.

VII. FINITE WAITING ROOMS

Corresponding approximations can be developed for $G/GI/s/r$ systems with r extra waiting spaces. We can apply heavy-traffic limit theorems for $G/GI/s/\infty$ systems (see Ref. 4), together with the conditioning heuristic of Section III. The conditioning relationship (7) is also valid for $M/M/s/r$ systems with different values of r. As Ref. 4 describes, there are several possible heavy-traffic limit theorems to apply. For $GI/M/s/\infty$ systems, we suggest the heavy-traffic limit theorems in Ref. 4 with $(s - \alpha)/\sqrt{\alpha} \to \beta$ as $\alpha \to \infty$ or, equivalently, $(1 - \rho)\sqrt{s} \to \beta$, where $\alpha = \lambda/\mu$ and $\rho = \alpha/s$. This leads to a promising analog of Hayward's approximation (22) for the case of a finite waiting room, namely, (42) below.

From Section 4 of Ref. 4, it is evident that the extension to nonexponential service times is more difficult when $r > 0$, but we conjecture that the GI/M/s/∞ results in Ref. 4 extend to G/M/s/∞ systems (nonrenewal arrival processes) and that the conditioning heuristic is valid in heavy traffic for G/M/s/r systems. In support of this, the conservation relationships (10) and (14) extend to G/M/s/r systems. (The factor k on the right side of (14) is replaced by min$\{k, s\}$.) The corresponding heavy-traffic local limit theorem for M/M/s/r systems is easy to prove using the methods of Ref. 4 or the Appendix. We conjecture that heavy-traffic local limit theorems corresponding to (17) through (20) are also valid for GI/M/s/r systems as well as for the more general G/M/s/r systems. Moreover, we conjecture that the form of the limits will coincide with what we get by applying the conditioning heuristic to the GI/M/s/∞ limits in Ref. 4.

In this section, let X_α and Y_α be the equilibrium number of customers in G/M/s/∞ and G/M/s/r systems, respectively, at an arbitrary time. For α large with $(1 - \rho)\sqrt{s} = \beta$, the approximations derived from Ref. 4, where the limit theorem is proved only for renewal arrival processes (see Propositions 1 and 2 and Theorems 1 and 4), are

$$P(X_\alpha \geq s) \equiv \gamma \approx [1 + \beta'\Phi(\beta')/\phi(\beta')]^{-1}, \tag{29}$$

$$P(X_\alpha > s + r \,|\, X_\alpha \geq s) \approx \eta \equiv e^{-\beta'r/\sqrt{s}}, \tag{30}$$

$$P(Y_\alpha \geq s) \equiv \xi \approx \gamma(1 - e^{-\beta'r/\sqrt{s}})/(1 - e^{-\beta'r/\sqrt{s}}), \tag{31}$$

$$\sqrt{s}P(Y_\alpha = k \,|\, Y_\alpha \leq s) \approx \phi(\beta' + \delta)/\Phi(\beta'), \tag{32}$$

$$\sqrt{s}P(Y_\alpha = k \,|\, Y_\alpha \geq s) \approx (\gamma\beta'/\xi)e^{-\beta'\delta}, \tag{33}$$

and

$$\sqrt{s}B_T \approx \beta'e^{-\beta'r/\sqrt{s}} \tag{34}$$

for $(k - s)/\sqrt{s} = \delta$ and $\beta' = \beta/z$ with z being the peakedness in (4).

Since

$$E(\min\{X_\alpha, s\}) = \alpha = s\rho \tag{35}$$

for all G/G/s/∞ systems, e.g., by (4.2.3) of Ref. 27,

$$E(\min\{Y_\alpha, s\}) = \frac{sP(s \leq X_\alpha \leq s + r) + s\rho - sP(X_\alpha \geq s)}{P(X_\alpha \leq s + r)}$$

$$\approx [s(\gamma - \gamma\eta) + s(\rho - \gamma)]/(1 - \gamma\eta)$$

$$\approx s(\rho - \gamma\eta)/(1 - \gamma\eta), \tag{36}$$

for γ and η in (29) and (30), so that by (10)

$$B_C = 1 - \alpha^{-1}E(\min\{Y_\alpha, s\}) \approx 1 - \rho^{-1}(\rho - \gamma\eta)/(1 - \gamma\eta)$$

$$\approx (1 - \rho)\gamma\eta/\rho(1 - \gamma\eta) \tag{37}$$

and

$$\sqrt{s}B_C \approx \beta\gamma\eta/\rho(1 / \gamma\eta) \approx (z/\rho)\sqrt{s}B_T \tag{38}$$

so that $B_C \approx zB_T$ as in (13), although the correction with ρ in (38) may be useful.

Also note that

$$\eta \equiv \eta(\rho, s, r, z) \equiv \eta(\rho, r, z) \approx \eta(\rho, r/z, 1) \tag{39}$$

and

$$\gamma \equiv \gamma(\rho, s, r, z) \equiv \gamma(\rho, s, z) \approx \gamma(\rho, s/z^2, 1), \tag{40}$$

so that

$$B_C \equiv B_C(\rho, s, r, z) \approx B_C(\rho, s/z^2, r/z, 1) \tag{41}$$

or, equivalently,

$$B_C \equiv B_C(\alpha, s, r, z) \approx B_C(\alpha/z^2, s/z^2, r/z, 1) \tag{42}$$

in the manner of Hayward's approximation in Section 6.2. If we express B_C in terms of $1 - \rho$, then we have the alternate expression

$$B_C \equiv B_C(1 - \rho, s, r, z) \approx B_C[(1 - \rho)/z, s, r, 1]. \tag{43}$$

We can achieve (42) by fixing s, r and μ and then changing λ so that $(1 - \rho)/z$ is unchanged while z is replaced by 1. As with Hayward's approximation, we can use M/M/s/r formulas when $z = 1$.

Note that the normalizations in (41) through (43) are not the same as in Hayward's approximation in Section 6.2. Comparing (42) with (22), we see that now α and s are divided by z^2 in (42) instead of z, so that evidently the peakedness has a much greater impact when there is a finite waiting room. This might be expected because now the boundary where losses occur is further away from the center of mass, with ρ required to be less than one. The different normalization of r and s might be expected because this approximation is based on r being of order $\sqrt{s}$.

Of course, the approximations developed in this section need to be tested, which has not yet been done. From the theory, we know that the modification of Hayward's approximation for finite waiting rooms in (42) should work well when ρ is high but less than one, s is large, and r is of order $\sqrt{s}$, but it remains to determine the actual range over which the approximation is good.

VIII. TESTING THE BLOCKING APPROXIMATION

In this section we present numerical comparisons to test the approximation for the approximate blocking probability B_C in (13). Our testing here is confined to $G/M/s/0$ systems. Since formula (13) coincides with the normal approximations of Rahko[10-13] and Delbrouck,[14] their comparisons are relevant too.

Tables I through IV compare the heavy-traffic approximation (13) with exact $M/M/s/0$ results and the equivalent random method. We selected seven different blocking probabilities: 0.001, 0.01, 0.05, 0.10, 0.20, 0.40, and 0.60. The higher numbers were selected so that we could test the lower bound in (23), which is only applicable when $\alpha > s$. We also selected four different (z, s)-pairs: (1, 5), (1, 50), (2, 50), and (10, 400). In each case, we used the charts on pages 23–32 of Wilkinson[43] to determine the corresponding load in Erlangs dictated by the equivalent random method for the given blocking probabilities and parameters s and z. When the peakedness was not 1 (Tables III and IV), we also calculated Hayward's approximation (22). For Hayward's approximation we often used the more detailed graphs in Appendix A of Cooper.[44] (The results were also checked using Jagerman's computer programs.[35]) In parentheses next to the lower bound is the upper bound obtained from (26). It should be noted that the upper bound in (26) is an upper bound for our normal approximation, not necessarily on the true blocking probability.

In interpreting the tables, remember that the equivalent random method is only an approximation too when the arrival process is not Poisson. Moreover, from Holtzman[15] we know that the range of possible blocking probabilities consistent with the partial information provided by the peakedness can be quite wide. (As we indicated in Section 6.3, this is not true in heavy traffic.) Hence, there often is little reason to prefer the numerical accuracy of exact calculations according to the Erlang loss formula (21) over the normal approximation (13).

Table I—The blocking probability $B_C(s, \alpha, z)$ for $z = 1$ and $s = 5$

Load in Erlangs α	Blocking Probability	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho > 1$
0.77	0.001	0.000	
1.37	0.01	0.003	
2.22	0.05	0.05	
2.86	0.10	0.11	
4.0	0.20	0.23	
6.6	0.40	0.39	0.24 (0.33)
11.5	0.60	0.49	0.57 (0.77)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

Table II—The blocking probability $B_C(s, \alpha, z)$ for $z = 1$ and $s = 50$

Load in Erlangs α	Blocking Probability	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho > 1$
32.5	0.001	0.001	
38.0	0.01	0.01	
44.5	0.05	0.05	
49.7	0.10	0.10	
59.0	0.20	0.21	0.15 (0.16)
82.0	0.40	0.42	0.39 (0.40)
122.0	0.60	0.59	0.59 (0.61)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

As we expected, the quality of the approximation, as measured against the exact formula when $z = 1$ or the equivalent random method when $z \neq 1$, improves as α or s increases and z decreases. The parameter α/z gives a good indication of the quality to be expected; i.e., the quality depends approximately on α/z and improves as α/z increases. The heavy-traffic approximation tends to degrade as the number of servers gets beyond two or three standard deviations ($\sqrt{\alpha z}$) away from the G/GI/∞ mean (the load α).

Table V presents some of Kuczura's[45] results (as displayed in his Figures 1–3) for GI+M/M/s/0 systems (having an arrival process that is a superposition of a renewal process and a Poisson process) together with our heavy-traffic approximation. A significant feature of these systems is that the blocking experienced by the customers in the different streams is not the same. Since the arrival rates in the two streams are identical in each case, the blocking experienced by an arbitrary customer, which is what our approximations are for, is the average of the blocking probabilities associated with the separate streams.

Table III—The blocking probability $B_C(s, \alpha, z)$ for $z = 2$ and $s = 50$

Load in Erlangs α	Approximate Blocking Probability (Equiv. Rand.)	Hayward's Approximation $B(25, \alpha/2)$	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho > 1$
26.5	0.001	0.001	0.001	
32.6	0.01	0.01	0.01	
40.2	0.05	0.05	0.06	
45.8	0.10	0.10	0.11	
55.5	0.20	0.20	0.21	0.10 (0.10)
78.9	0.40	0.40	0.38	0.37 (0.38)
120.0	0.60	0.59	0.58	0.58 (0.61)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

Table IV—The blocking probability $B_C(s, \alpha, z)$ for $z = 10$ and $s = 400$

Load in Erlangs α	Approximate Blocking Probability (Equiv. Rand.)	Hayward's Approximation $B(40, \alpha/10)$	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho > 1$
256	0.001	0.002	0.001	
290	0.01	0.01	0.01	
340	0.05	0.05	0.05	
385	0.10	0.10	0.10	
460	0.20	0.20	0.22	0.13 (0.14)
640	0.40	0.40	0.41	0.38 (0.39)
900	0.60	0.56	0.56	0.56 (0.57)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

To obtain the heavy-traffic approximation, it is necessary to specify the peakedness of the arrival process. When the arrival process is the superposition of independent renewal processes, at least one of which is not Poisson, the superposition process is not a renewal process (see Ref. 26 and its references). However, the peakedness of the superposition process is clearly the convex combination of the individual peakedness values. Suppose there are n independent streams with λ_i the arrival rate and z_i the peakedness of stream i. Then clearly

Table V—The blocking probability for a GI+M/M/s/0 model with $s = 25$: comparison with Kuczura[45]

System	Load in Erlangs, α		
	20	25	30
M/M/s/0			
Arbitrary arrival	0.050	0.144	0.245
Heavy traffic	0.054	0.148	0.229
D + M/M/s/0			
Poisson arrival	0.045	0.15	0.28
Renewal arrival	0.027	0.11	0.20
Arbitrary arrival	0.036	0.13	0.24
Heavy traffic ($z = 0.75$)	0.037	0.124	0.214
Hayward ($z = 0.75$)	0.035	0.126	0.232
GI + M/M/s/0 [$z(G) = 2$]			
Poisson arrival	0.058	0.14	0.21
Renewal arrival	0.094	0.20	0.31
Arbitrary arrival	0.076	0.17	0.26
Heavy traffic ($z = 1.5$)	0.084	0.186	0.274
Hayward ($z = 1.5$)	0.077	0.172	0.268
GI + M/M/s/0 [$z(G) = 3$]			
Poisson arrival	0.06	0.14	0.20
Renewal arrival	0.15	0.26	0.38
Arbitrary arrival	0.115	0.20	0.29
Heavy traffic ($z = 2$)	0.12	0.21	0.30
Hayward ($z = 2$)	0.101	0.195	0.286
Equivalent random method ($z = 2$)	0.105	0.200	0.285

Note: All entries except the heavy-traffic, equivalent-random method, and Hayward values are from Kuczura.[45]

$$\lambda = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad \sum_{i=1}^{n} \lambda_i z_i / \lambda. \qquad (44)$$

We use (44) to obtain the peakedness values for the heavy-traffic approximation given in Table V. For the case in which the peakedness of the superposition process is $z = 2$, we also compared the blocking probabilities with those obtained using the equivalent random method, Chart 2 on page 24 of Wilkinson,[43] and Hayward's approximation using Jagerman's program.[35] The results seem to be good, about the same as those in Tables I through IV.

IX. ACKNOWLEDGMENTS

REFERENCES

1. A. A. Borovkov, "On Limit Laws for Service Processes in Multi-Channel Systems," Siberian Math. J., 8 (1967), pp. 746–763 (English translation).
2. A. A. Borovkov, Stochastic Processes in Queueing Theory, New York: Springer-Verlag, 1976.
3. W. Whitt, "On the Heavy-Traffic Limit Theorem for GI/G/∞ Queues," Adv. Appl. Prob., 14, No. 1 (March 1982), pp. 171–190.
4. S. Halfin and W. Whitt, "Heavy-Traffic Limits for Queues With Many Exponential Servers," Oper. Res., 29, No. 3 (May–June 1981), pp. 567–588.
5. A. A. Fredericks, "Congestion in Blocking Systems—A Simple Approximation Technique," B.S.T.J., 59, No. 6 (July–August 1980), pp. 805–827.
6. A. E. Eckberg, "Generalized Peakedness of Teletraffic Processes," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 4.4b.3.
7. E. Brockmeyer, "Approximative Formulae for Loss and Improvement," Appendix 2 in The Life and Work of A. K. Erlang, E. Brockmeyer, H. L. Halstrom, and A. Jensen, eds., Copenhagen: Danish Academy of Sciences, 1948, pp. 120–126.
8. E. Vaulot, "Les Formules d' Erlang et leur Calcul Pratique," Ann. Telecom., 6 (1951), pp. 279–286.
9. D. L. Jagerman, "Some Properties of the Erlang Loss Functions," B.S.T.J., 53, No. 3 (March 1974), pp. 525–551.
10. K. Rahko, "The Dimensioning of Local Telephone Traffic Routes Based on the Distribution of the Traffic Carried," Acta Polytechnica Scandinavica, 1967, El 14, 95.
11. K. Rahko, "Dimensioning of Traffic Routes According to the EERT-Method and Corresponding Methods," Proc. Eighth Inter. Teletraffic Congress, Melbourne, November 1976, p. 143.
12. K. Rahko, "Some Methods for Approximation of Congestion," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.6.
13. S. Hertzberg, "Approximative Dimensioning Formulas with the Normal Distribution Method," Reports 3/79 and 5/82, Telecommunication Switching Laboratory, Helsinki University of Technology, 1979 and 1982 (in Swedish).
14. L. E. N. Delbrouck, "A Unified Approximate Evaluation of Congestion Functions for Smooth and Peaky Traffics," IEEE Trans. Comm., COM-29, No. 2 (February 1981), pp. 85–91.
15. J. M. Holtzman, "The Accuracy of the Equivalent Random Method With Renewal Inputs," B.S.T.J., 52, No. 2 (May 1973), pp. 1673–1679.
16. M. Sobel, "Simple Inequalities for Multiserver Queues," Management Sci., 26, No. 9 (September 1980), pp. 951–956.

17. D. P. Heyman, "Comments on a Queueing Inequality," Management Sci., *26*, No. 9 (September 1980), pp. 956–959.
18. A. K. Erlang, "On the Rational Determination of the Number of Circuits," *The Life and Works of A. K. Erlang*, E. Brockmeyer, H. L. Halstrom, and A. Jensen, eds., Copenhagen: Danish Academy of Technical Sciences, 1948, pp. 216–221.
19. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, 2nd Ed., New York: John Wiley and Sons, 1968.
20. J. DeBoer, "Limits and Asymptotes of Overflow Curves," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.4.
21. A. A. Fredericks, "Approximating Parcel Blocking Via State Dependent Birth Rates," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.2.
22. K. Lindberger, "Simple Approximations of Overflow System Quantities for Additional Demands in the Optimization," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.3.
23. B. Sanders, W. H. Haemers, and R. Wilcke, "Simple Approximation Techniques for Congestion Functions for Smooth and Peaked Traffic," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 4.4b.1.
24. E. A. van Doorn, "Some Analytical Aspects of the Peakedness Concept," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 4.4b.5.
25. G. F. Newell, "The $M/M/\infty$ Service System With Ranked Servers in Heavy Traffic," Institute of Transportation Studies, University of California at Berkeley, 1983.
26. W. Whitt, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," Oper. Res., *30*, No. 1 (January–February 1982), pp. 125–147.
27. P. Franken, D. König, U. Arndt and V. Schmidt, *Queues and Point Processes*, Berlin: Akademic-Verlag, 1981.
28. R. W. Wolff, "The Effect of Service Time Regularity on System Performance," in *Computer Performance*, K. M. Chandy and M. Reiser, eds., Amsterdam: North-Holland, 1977, pp. 297–304.
29. W. Whitt, "Minimizing Delays in the GI/G/1 Queue," Oper. Res., *32*, No. 2 (March–April 1984), to be published.
30. D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research*, Vol. 1, New York: McGraw-Hill, 1982.
31. D. L. Jagerman, "Nonstationary Blocking in Telephone Traffic," B.S.T.J., *54*, No. 3 (March 1975), pp. 625–661.
32. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30*, No. 2 (March–April 1982), pp. 223–231.
33. W. Whitt, "Approximating a Point Process by a Renewal Process: The View Through a Queue, An Indirect Approach," Management Sci., *27*, No. 6 (June 1981), pp. 619–636.
34. D. L. Jagerman, unpublished work.
35. D. L. Jagerman, unpublished work.
36. P. Bloomfield and D. R. Cox, "A Low Traffic Approximation for Queues," J. Appl. Prob., *9*, No. 4 (December 1972), pp. 832–840.
37. G. F. Newell, *Applications of Queueing Theory*, Second Edition, London: Chapman Hall, 1982.
38. D. Y. Burman and D. R. Smith, "A Light Traffic Theorem for Multiserver Queues," Math. Oper. Res., *8*, No. 1 (February 1983), pp. 15–25.
39. D. Y. Burman and D. R. Smith, "Asymptotic Analysis of a Queueing Model With Bursty Traffic," B.S.T.J., *62*, No. 6 (July–August 1983), pp. 1433–1453.
40. M. I. Reiman and B. Simon, unpublished work.
41. S. L. Albin, "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes of Queues," Oper. Res., *32* (1984).
42. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *63*, No. 9 (November 1983), pp. 2779–2815.
43. R. I. Wilkinson, *Nonrandom Traffic Curves and Tables*, Traffic Studies Center, Bell Laboratories, 1970.
44. R. B. Cooper, *Introduction to Queueing Theory*, New York: The Macmillan Company, 1972.
45. A. Kuczura, "Loss Systems with Mixed Renewal and Poisson Inputs," Oper. Res., *21*, No. 3 (May–June 1973), pp. 787–795.

APPENDIX

A Heavy-Traffic Local Limit Theorem for the Erlang Loss Formula

For the elementary M/G/s/0 system, where the blocking probability is given by the Erlang loss formula (21), approximations (9) and (13) follow easily from well-known limit theorems for the Poisson distribution. We present one possible argument here.

Let $p(k, \lambda)$ denote the probability mass function of the Poisson distribution with mean λ, i.e.,

$$p(k, \lambda) = e^{-\lambda}\lambda^k/k!. \tag{45}$$

The Erlang loss formula then can be expressed as

$$B(s, \alpha) = p(s, \alpha) \Big/ \sum_{k=0}^{s} p(k, \alpha). \tag{46}$$

As before, let $\phi(x)$ and $\Phi(x)$ be the density and cdf, respectively, of a standard normal random variable, $N(0, 1)$.

Theorem: In an M/G/s/0 system,

$$\lim_{\alpha \to \infty} \sqrt{\alpha}B([\alpha + s\sqrt{\alpha}], \alpha) = \phi(s)/\Phi(s),$$

where $[x]$ is the greatest integer less than or equal to x.

Proof: Since a Poisson random variable with mean λ has the same distribution as the sum of n i.i.d Poisson random variables with mean λ/n, the central limit theorem can be applied to obtain

$$\lim_{\alpha \to \infty} \sum_{k=0}^{[\alpha+s\sqrt{\alpha}]} p(k, \alpha) = \Phi(s)$$

(see Problem 9, p. 194, and Example X(c), p. 245, of Feller[19]). Hence, it remains to show that

$$\lim_{\alpha \to \infty} \sqrt{\alpha}p[(\alpha + s\sqrt{\alpha}), \alpha] = \phi(s) = (2\pi)^{-1/2}e^{-s^2/2}.$$

We can establish this result using Stirling's formula (p. 52 of Feller[19]). As in Stirling's formula, let the symbol $\sim$ below mean that the ratio of the two sides tends to 1 as $\alpha \to \infty$. We have

$$\sqrt{\alpha}p[(\alpha + s\sqrt{\alpha}), \alpha] = \frac{\sqrt{\alpha}e^{-\lambda}(\alpha^{(\alpha+s\sqrt{\alpha})})}{(\alpha + s\sqrt{\alpha})!}$$

$$\sim \frac{\sqrt{\alpha}e^{-\alpha}(\alpha^{(\alpha+s\sqrt{\alpha})})}{\sqrt{2\pi}(\alpha + s\sqrt{\alpha})^{(\alpha+s\sqrt{\alpha})}(\alpha + s\sqrt{\alpha})^{1/2}e^{-(\alpha+s\sqrt{\alpha})}}$$

$$\sim \frac{e^{s\sqrt{\alpha}}}{\sqrt{2\pi}(1 + s/\sqrt{\alpha})^{\alpha}(1 + s/\sqrt{\alpha})^{s\sqrt{\alpha}}}$$

$$\sim \frac{e^{-s^2}}{\sqrt{2\pi}} \frac{e^{s\sqrt{\alpha}}}{(1 + s/\sqrt{\alpha})^{\alpha}},$$

where

$$\log(e^{s\sqrt{\alpha}}(1 + s/\sqrt{\alpha})^{-\alpha}) = s\sqrt{\alpha} - \alpha \log(1 + s/\sqrt{\alpha})$$

$$= s\sqrt{\alpha} - \alpha \left(\frac{s}{\sqrt{\alpha}} - \frac{1}{2} \left(\frac{s}{\sqrt{\alpha}} \right)^2 + o(\alpha) \right) = \frac{s^2}{2} + 0(1).$$

Hence,

$$e^{s\sqrt{\alpha}}(1 + s/\sqrt{\alpha})^{\alpha} \sim e^{s^2/2}$$

and

$$\sqrt{\alpha}p[(\alpha + s\sqrt{\alpha}), \alpha] \sim (2\pi)^{-1/2}e^{-s^2/2}$$

as claimed.

AUTHOR

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; AT&T Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At AT&T Bell Laboratories he is in the Operations Research Department in the Systems Analysis Center.

Blocking Probability in a Switching Center With Arbitrary Routing Policy

By B. GOPINATH,* J.-M. GARCIA,† and P. VARAIYA‡

(Manuscript received March 15, 1983)

We model the arrival of calls at a switch where they are assigned to any one of the available idle outgoing links. A call is blocked if all links are busy. It may be lost after assignment to an idle link with a probability that depends on the link. For the case of Poisson arrivals and exponential holding times, all show that the distribution of passage time to the blocked state is independent of the assignment policy. As a consequence, the blocking probability and the law of the overflow traffic are also independent of the assignment policy.

I. INTRODUCTION

Telephone calls arrive at a switching center in a Poisson stream of rate λ. When a call arrives, the switch, in accordance with a pre-specified policy, assigns the call to one of the idle outgoing links. If the call is assigned to the idle link i, there is a probability $1 - \epsilon_i$ that the call is unsuccessful, and probability $\epsilon_i > 0$ that it is successful. An unsuccessful call is immediately lost and link i remains idle. If the call is successful, link i immediately becomes busy and remains in that state for a holding or conversation time, which is an exponentially distributed random time with mean $1/\mu$. At the end of this holding time link i returns to the idle state again.

The preceding paragraph describes a model that approximates the behavior of a single node in a telephone network. In an actual network,

---

* AT&T Bell Laboratories; present affiliation Bell Communications Research, Inc. † Toulouse, France. ‡ University of California, Berkeley, CA.

for a call to be placed successfully a circuit or path of idle links must be established between the call's originating node and the destination node. The node where the call arrives assigns it to one of its idle links, say, link i. As the call proceeds from node to node towards its destination, it may encounter a node all of whose outgoing links are busy, in which case the call is lost. The probability that this may happen is $1 - \epsilon_i$. This probability depends on the entire path that the call follows. It is assumed here, however, that the probability depends only on the initial link i. Such a link-by-link analysis is a common approximation in traffic theory.[1]

Suppose the model has n outgoing links indexed $i \in \Omega := \{1, 2, \cdots, n\}$. The state of the links or of the switching node is the subset $I \subset \Omega$ of idle links. So the state space is 2^Ω and there are 2^n states. The state where there is no idle link, namely state ϕ, is the blocked state. In a manner specified precisely in the next section, each switching policy u defines an irreducible Markov chain on this state space. Let $\tau_u(I)$ be the first time that, in equilibrium, the chain starting in state I reaches the blocked state ϕ, and let ρ_u be the time to return to ϕ after leaving ϕ.

In this paper we show the surprising fact that the distributions of $\tau_u(I)$ and ρ_u are independent of the policy u. Several consequences follow from this: First, the blocking probability $p_u(\phi)$, an important performance measure, is independent of u. Second, the process of blocked calls[4] has a law that does not depend on the policy u either. Hence, in designing a switching policy one need not worry about blocked calls since they cannot be affected by the policy anyway. The overflow process, which consists of both lost and blocked calls, does of course depend on the policy.

This paper is organized as follows. Section II presents the Markov chain description. Section III gives the calculation of the distribution of the passage time $\tau_u(I)$. An algorithm is given to evaluate these distributions. Section IV contains two examples. For the special case where there are only two possible values for the loss probabilities ϵ_i, the algorithm simplifies. For this case a formula for the blocking probability is conjectured. The formula extends the well-known Erlang formula and it has been verified for many examples. However, a proof of its correctness is not yet available.

II. THE MARKOV CHAIN DESCRIPTION

A policy u prescribes, for each state I, the probability with which an arriving call is assigned to an idle link $i \in I$. Thus u is specified by an array $u = \{u(I, i), i \in I \neq \phi\}$ such that $u(I, i) \geq 0$, $\sum_i u(I, i) = 1$.

Each u defines a transition rate matrix $\mathbf{R}_u = \{R(I, J), I \subset \Omega, J \subset \Omega\}$

whose nonzero elements are

$$\mathbf{R}_u(I, I - i) = \lambda_i u(I, i), \quad i \in I$$

$$\mathbf{R}_u(I, I + j) = \mu, \quad j \notin I$$

$$\mathbf{R}_u(I, I) = - \sum_{i \in I} \lambda_i u(I, i) - (n - |I|)\mu, \tag{1}$$

where $\lambda_i := \epsilon_i \lambda$, $I - i := I \setminus i$, $I + j := I \cup \{j\}$, and $|I|$ is the cardinality of I. The first expression in (1) gives the rate with which an idle link becomes busy, the second gives the rate with which a busy link becomes idle, the third gives the diagonal term $\mathbf{R}_u(I, I) = \sum_{J \neq I} \mathbf{R}_u(I, J)$. For later reference observe that the column of $\mathbf{R}_u$ corresponding to the blocked state has elements

$$\mathbf{R}_u(\phi, \phi) = -n\mu, \qquad \mathbf{R}_u(\{i\}, \phi) = \lambda_i,$$

$$\mathbf{R}_u(I, \phi) = 0 \quad \text{for} \quad |I| > 1. \tag{2}$$

Since each $\epsilon_i > 0$ by assumption, it follows that the chain defined by $\mathbf{R}_u$ is irreducible and possesses a unique invariant positive probability measure $\{p_u(I), I \subset \Omega\}$. In particular, $p_u(\phi)$ is the equilibrium blocking probability. It will be shown to be independent of u. This is surprising in view of the fact that, if the ϵ_i are all different, then for every $I \neq \phi$, $p_u(I)$ depends on u.

Suppose the chain starts in state $I \neq \phi$ at time 0 and let $\tau_u(I)$ denote the first time the chain reaches ϕ. The distribution of $\tau_u(I)$ can be obtained as follows. Let $\hat{\mathbf{R}}_u$ be the rate matrix obtained from $\mathbf{R}_u$ by making ϕ an absorbing state:

$$\hat{\mathbf{R}}_u(\phi, j) \equiv 0, \qquad \hat{\mathbf{R}}_u(I, J) \equiv \mathbf{R}_u(I, J), \qquad I \neq \phi. \tag{3}$$

Let $\mathbf{e}_J = \{\mathbf{e}_J(I), I \subset \Omega\}$ denote the (column) vector whose only nonzero component is $\mathbf{e}_J(J) = 1$. Let $p_t = \{p_t(J)\}$ be the solution of

$$p_t(J) = \sum_K p_t(K)\hat{\mathbf{R}}_u(K, J), \qquad p_0 = \mathbf{e}_I. \tag{4}$$

Then one has

$$\text{Prob}\{\tau_u(I) \leq t\} = p_t(\phi),$$

i.e., $p_t(\phi)$ is the cumulative distribution function of the first passage time $\tau_u(I)$. From (4) it follows that the Laplace transform $F_u(I, s)$ of $p_t(\phi)$ is given by

$$F_u(I, s) = \mathbf{e}_I'(s\mathscr{I} - \hat{\mathbf{R}}_u)^{-1}\mathbf{e}_\phi, \tag{5}$$

where $\mathscr{I}$ is the identity matrix.

The next section is devoted to the main result:

Theorem 1: $(s\mathscr{I} - \hat{\mathbf{R}}_u)^{-1}\mathbf{e}_\phi$ does not depend on u.

The theorem has several consequences. First, it shows that the distribution of the passage time $\tau_u(I)$ does not depend on u. Second, suppose the chain corresponding to $\mathbf{R}_u$ leaves state ϕ at time 0-. Then at time 0 it enters one of the states $\{1\}, \cdots, \{n\}$, each with probability $1/n$. Let ρ_u be the first time that the chain returns to ϕ. The Laplace transform $G_u(s)$ of $\text{Prob}\{\rho_u \le t\}$ is simply

$$G_u(s) = \frac{1}{n} \sum_{i=1}^{n} F_u(\{i\}, s),$$

which is independent of u. In particular, the expected value $\bar{\rho}$ of ρ_u is independent of u. The expected holding time in state ϕ is $1/n\mu$, and so the blocking probability

$$p_u(\phi) = \frac{1}{1 + n\mu\bar{\rho}}$$

is also independent of u. Finally, the process of blocked calls can be described by a sequence of independent intervals $S_1, T_1, S_2, T_2, \cdots$. Each S_i has the same distribution as $\rho_u(\phi)$, and during this interval there is no blocked call. Each T_i has the same distribution as the holding time in state ϕ, and during this interval all arriving calls are blocked. Thus the statistics of the blocked call process are not affected by u.

III. CALCULATION OF $(s\mathscr{I} - R_u)^{-1}\mathbf{e}_\phi$

Throughout this section a fixed policy is considered and so the suffix u is dropped.

Using (1) through (3) write

$$\hat{\mathbf{R}} = A_\lambda + A_\mu + B,$$

with the nonzero elements of these matrices being

$$A_\lambda(I, I - i) = \lambda_i u(I, i), \quad i \in I \quad \text{and} \quad I - i \ne \phi$$

$$A_\lambda(I, I) = -\sum_{i \in I} \lambda_i u(I, i), \quad I \ne \phi$$

$$A_\mu(I, I + j) = \mu, \quad j \notin I \quad \text{and} \quad I \ne \phi$$

$$A_\mu(I, I) = -(n - |I|)\mu, \quad I \ne \phi$$

$$B(\{i\}, \phi) = R(\{i\}, \phi) = \lambda_i.$$

Then

$$A_\lambda(\phi, \, \cdot\,) \equiv A_\lambda(\cdot\,, \, \phi) \equiv A_\mu(\phi, \, \cdot\,) \equiv A_\mu(\cdot\,, \, \phi) \equiv 0, \qquad (6)$$

and for any vector $x = \{x(I)\}$,

$$(A_\lambda x)(I) := \sum_j A_\lambda(I, J) x(J) = \sum_{\substack{i \in I \\ I - i \neq \phi}} \lambda_i u(I, i) x(I - i)$$

$$- \sum_{i \in I} \lambda_i u(I, i) x(I), \qquad (7)$$

$$(A_\mu x)(I) = \mu \sum_{j \notin I} x(I + j) - (n - |I|) \mu x(I), \quad I \neq \phi$$

$$= 0, \qquad I = \phi. \qquad (8)$$

First one calculates $\hat{\mathbf{R}}^d \mathbf{e}_\phi$, $d \geq 0$.

From (6)

$$\mathbf{v}_1 := \hat{\mathbf{R}} \mathbf{e}_\phi = B \mathbf{e}_\phi \qquad (9)$$

so its only nonzero components are $\mathbf{v}_1(\{i\}) = \lambda_i$. In particular, since $\mathbf{v}_1(\phi) = 0$, it follows that

$$\hat{\mathbf{R}}^d \mathbf{e}_\phi = (A_\lambda + A_\mu)^{d-1} \mathbf{v}_1, \qquad d \geq 1. \qquad (10)$$

The appendix introduces vectors $\mathbf{w}_d$ and $\mathbf{z}_d$, $d \geq 1$. [See eqs. (16) through (19) and (24) through (26).] Define vectors $\mathbf{v}_d$ by

$$\mathbf{v}_d(I) = (-1)^{|I| \cdot d - 1} \mathbf{w}_d(I), \qquad 1 \leq d \leq n + 1. \qquad (11)$$

It can be checked that $\mathbf{v}_1$ given by (9) and (11) are the same.

Next one evaluates $A_\lambda \mathbf{v}_d$.

Lemma 1: $\qquad\qquad \mathbf{v}_{d+1} = A_\lambda \mathbf{v}_d, \qquad 1 \leq d \leq n \qquad (12)$

$$\mathbf{v}_{n+1} = - \sum_{r=1}^{n} \mathbf{z}_r(\Omega) \mathbf{v}_{n+1-r}. \qquad (13)$$

Proof: Since $\mathbf{w}_d(\phi) = \mathbf{v}_d(\phi) = 0$ and since $A_\lambda(\cdot\,, \, \phi) \equiv 0$, therefore, trivially,

$$\mathbf{v}_{d+1}(\phi) = (A_\lambda \mathbf{v}_d)(\phi) = 0.$$

Now suppose $I = \{i\}$. From (7) we get

$$(A_\lambda \mathbf{v}_d)(\{i\}) = -\lambda_i \mathbf{v}_d(\{i\})$$

$$= -\lambda_i (-1)^d \mathbf{w}_d(\{i\})$$

$$= (-1)^{d+1} \mathbf{w}_{d+1}(\{i\}), \quad \text{by Lemma 3,}$$

$$= \mathbf{v}_{d+1}(\{i\}).$$

Next suppose $|I| \geq 2$. From (7)

$$(A_\lambda \mathbf{v}_d)(I) = \sum_{i \in I} \lambda_i u(I, i) \mathbf{v}_d(I - i) - \sum_{i \in I} \lambda_i u(I, i) \mathbf{v}_d(I)$$

$$= \sum_{i \in I} u(I, i)(-1)^{|I|+d} [\lambda_i \mathbf{w}_d(I) + \lambda_i \mathbf{w}_d(I - i)]$$

$$= (-1)^{|I|+d} \left[\sum_{i \in I} u(I, i) \right] \mathbf{w}_{d+1}(I), \quad \text{by Lemma 3,}$$

$$= (-1)^{|I|+d} \mathbf{w}_{d+1}(I) = \mathbf{v}_{d+1}(I).$$

Finally,

$$\mathbf{v}_{n+1}(I) = (-1)^{|I|+n} \mathbf{w}_{n+1}(I)$$

$$= -(-1)^{|I|+n} \sum_{r=1}^{n} (-1)^r \mathbf{z}_r(\Omega) \mathbf{w}_{n+1-r}(I), \quad \text{by Lemma 7,}$$

$$= - \sum_{r=1}^{n} \mathbf{z}_r(\Omega) \mathbf{v}_{n+1-r}(I). \qquad \square$$

Next one evaluates $A_\mu \mathbf{v}_d$. Define

$$q_r := \sum_{i \in \Omega} \lambda_i^r.$$

Lemma 2: $A_\mu \mathbf{v}_d = -\mu \sum_{r=1}^{d-1} (-1)^r q_r \mathbf{v}_{d-r} - \mu(n - d)\mathbf{v}_d, \qquad d \ge 1.$ (14)

Proof: From (7), and for $I \ne \phi$,

$$(A_\mu \mathbf{v}_d)(I) = \mu \sum_{j \notin I} \mathbf{v}_d(I + j) - (n - |I|)\mu \mathbf{v}_d(i)$$

$$= \mu(-1)^{|I|+d} \left[\sum_{j \notin I} \mathbf{w}_d(I + j) + (n - |I|)\mathbf{w}_d(I) \right]$$

$$= \mu(-1)^{|I|+d} \left[\sum_{j \notin I} \sum_{r=1}^{d-1} \lambda_j^r \mathbf{w}_{d-r}(I) + \sum_{r=1}^{d-1} \sum_{j \in I} \lambda_j^r \mathbf{w}_{d-r}(I) \right.$$

$$\left. + (n - d)\mathbf{w}_d(I) \right], \quad \text{by Lemmas 4 and 6,}$$

$$= \mu(-1)^{|I|+d} \left[\sum_{r=1}^{d-1} q_r \mathbf{w}_{d-r}(I) + (n - d)\mathbf{w}_d(I) \right]$$

$$= - \left[\sum_{r=1}^{d-1} (-1)^r q_r \mathbf{v}_{d-r}(I) + (n - d)\mathbf{v}_d(I) \right]. \qquad \square$$

These two lemmas yield the next corollary.

Corollary 1: For each $d \geq 1$, there exist coefficients $\alpha_{d1}, \cdots, \alpha_{dd\wedge n}$ not depending on u such that

$$\hat{\mathbf{R}}^d \mathbf{e}_\phi = \sum_{e=1}^{d\wedge n} \alpha_{de} \mathbf{v}_e,$$

where $d \wedge n := \min(d, n)$.

Proof: This is apparent from eqs. (10) through (14) and the fact that the vectors $\mathbf{w}_d$, $\mathbf{z}_d$, and q_r do not depend on u.

Corollary 1 implies the existence of coefficients $\alpha_0(s), \cdots, \alpha_n(s)$ such that

$$(s\mathscr{I} - \hat{\mathbf{R}})^{-1} \mathbf{e}_\phi = \alpha_0 \mathbf{e}_\phi + \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n.$$

These coefficients can be readily calculated. Multiplying both sides by $(s\mathscr{I} - \hat{\mathbf{R}})$,

$$\mathbf{e}_\phi = \alpha_0(s\mathscr{I} - \hat{\mathbf{R}})\mathbf{e}_\phi + \sum_1^n \alpha_d(s\mathscr{I} - \hat{\mathbf{R}})\mathbf{v}_d$$

$$= \alpha_0 s\mathbf{e}_\phi - \alpha_0 \mathbf{v}_1 + \sum_1^n \alpha_d s\mathbf{v}_d - \sum_1^n \alpha_d A_\lambda \mathbf{v}_d - \sum_1^n \alpha_d A_\mu \mathbf{v}_d$$

$$= \alpha_0 s\mathbf{e}_\phi + \sum_1^n (\alpha_d s - \alpha_{d-1})\mathbf{v}_d - \alpha_n \sum_1^n \mathbf{z}_{n+1-d}(\Omega)\mathbf{v}_d$$

$$+ \mu \sum_1^n \alpha_d \sum_{r=1}^{d-1} (-1)^r q_r \mathbf{v}_{d-r} + \mu \sum_1^n \alpha_d(n - d)\mathbf{v}_d.$$

Rearranging terms and equating the coefficients of $\mathbf{e}_\phi, \mathbf{v}_1, \cdots, \mathbf{v}_n$ gives

$$\alpha_0 = s^{-1}$$

$$\alpha_{d-1} = [s + (n - d)\mu]\alpha_d + \mu \sum_{r=1}^{n-d} (-1)^r q_r \alpha_{d+r}$$

$$- \mathbf{z}_{n+1-d}(\Omega)\alpha_n, \qquad 1 \leq d \leq n.$$

The last n equations are in "triangular" form and can be solved recursively to yield

$$\alpha_{n-r} = Q_r(s)\alpha_n,$$

where $Q_r(s)$ is a monic polynomial in s of degree r. This gives

$$\alpha_0 = \frac{1}{s} = Q_n(s)\alpha_n.$$

Hence

$$(s\mathcal{I} - \hat{\mathbf{R}})^{-1}\mathbf{e}_\phi = \frac{1}{s}\,\mathbf{e}_\phi + \frac{1}{sQ_n(s)}\,[Q_{n-1}(s)\mathbf{v}_1 + \cdots + Q_1(s)\mathbf{v}_{n-1} + \mathbf{v}_n].$$

Since the right-hand side does not depend on u, this proves Theorem 1, announced in the previous section.

IV. EXAMPLES

Suppose the n links are grouped into two "trunks," trunk 1 with n_1 and trunk 2 with n_2 links, $n_1 + n_2 = n$. The loss probability of every link in trunk i is the same, namely, $1 - \epsilon_i$. In this special case one may construct a Markov chain with state (i_1, i_2), where $0 \le i_1 \le n_1$ and $0 \le i_2 \le n_2$ are the number of idle links in trunks 1 and 2, respectively. The number of states reduces to $(n_1 + 1)(n_2 + 1)$. (Of course, if links within the same trunk are distinguished, then 2^n states are needed as before.) This reduction carries over to the argument in the appendix, and so one can define the "generating" vectors $\mathbf{w}_d$, $\mathbf{v}_d$ with components indexed by the reduced state (i_1, i_2). The recursive definition of $\mathbf{w}_d$ can be explicitly solved to obtain

$$\mathbf{w}_d(i_1, i_2) = \sum_{k=i_1}^{d-i_2} B(k, i_1)B(d - k, i_2)\lambda_1^k\lambda_2^{d-k},$$

where

$$B(k, i) = \sum_{j=0}^{k-i} \binom{k - 2 - j}{i - j}.$$

Also, as before,

$$\mathbf{v}_d(i_1, i_2) = (-1)^{i_1+i_2+d-1}\mathbf{w}_d(i_1, i_2), \qquad 1 \le d \le n$$

$$\mathbf{v}_{n+1}(i_1, i_2) = -\sum_{r=1}^{n} \mathbf{z}_r(n_1, n_2)\mathbf{v}_{n+1-r}(i_1, i_2),$$

where

$$\mathbf{z}_d(i_1, i_2) = \sum_{k=0}^{d} C(k, i_1)C(d - k, i_2)\lambda_1^k\lambda_2^{d-k},$$

$$C(k, i) = \binom{i}{k}.$$

Let $p_u(i_1, i_2)$ denote the probability that in equilibrium $i_1(i_2)$ links in trunk 1(2) are idle. This probability generally depends on u. However, $p_b := p_u(0, 0)$, the blocking probability, is independent of the assignment policy. Calculating this explicitly in terms of the parameters λ_1, λ_2, μ, and for small values of n_1, n_2 suggests the general formula:

Table I—Blocking probabilities

	Policy	p_1	p_2	p_b
$n_1 = 9, n_2 = 9$	u_1	0.178	0.82×10^{-3}	0.502
$y = 10$	u_2	0.84×10^{-3}	0.122	0.502
$n_1 = 9, n_2 = 19$	u_1	0.498	0.807×10^{-3}	0.63×10^{-3}
$y = 20$	u_2	0.14×10^{-3}	0.44×10^{-1}	0.63×10^{-3}

Notation: $y = \lambda\mu^{-1}, p_1 = \sum_{i_2} p(0, i_2), p_2 = \sum_{i_1} p(i_1, 0)$

$$p_b^{-1} = \sum_{d=0}^{n_1+n_2} \sum_{i=0}^{d} \gamma(i, d) y_1^{-i} y_2^{-d+i}, \tag{15}$$

where $y_i := \lambda_i \mu^{-1}$, $i = 1, 2$. The coefficients γ are given by the recursion

$$\gamma(i, 0) = 1, \quad \text{all } i$$

$$\gamma(i, d) = 0, \quad i > d \quad \text{or} \quad i < 0$$

$$\gamma(i, d + 1) = (n_1 - i + 1)\gamma(i - 1, d) + (n_2 - d + i)\gamma(i, d).$$

The formula (15) reduces to the Erlang loss formula when $y_1 = y_2$. It is true for all y_1, y_2 and $n_1 + n_2 \leq 3$, and it has been verified for many particular cases. However, a proof of correctness is not available.

Consider now two numerical examples. In both $\epsilon_1 = 0.81$, $\epsilon_2 = 0.7$. Let u_1, respectively u_2, denote the "overflow" policy that assigns every call to trunk 1, respectively trunk 2, if it has an idle link. The table below shows that the probability of blocking a single trunk does vary considerably with policy, but the probability of blocking both trunks does not vary.

REFERENCES

1. J. M. Holtzman, "Analysis of Dependence Effects in Telephone Trunking Network," B.S.T.J., 50, No. 8 (October 1971), pp. 2647–62.
2. R. I. Wilkinson, "Theories of Toll Traffic Engineering in the U.S.A.," B.S.T.J., 35, No. 2 (March 1956), pp. 421–514.
3. R. Nesenbergs, "A Hybrid of Erlang B and C Formulas and Its Application," IEEE Trans. Commun., COM-27, No. 1 (1979), pp. 59–68.
4. A. Kuczura, "The Interrupted Poisson Process as an Overflow Process," B.S.T.J., 52, No. 3 (March 1973), pp. 437–48.

APPENDIX

For $d \geq 1$ define vectors $\mathbf{w}_d = \{\mathbf{w}_d(I), I \subset \Omega\}$ as follows:

$$\mathbf{w}_d(I) = 0, \quad |I| > d \tag{16}$$

$$\mathbf{w}_d(I) = \prod_{i \in I} \lambda_i, \quad |I| = d \tag{17}$$

$$\mathbf{w}_d(I) = \lambda_i \mathbf{w}_{d-1}(I) + \lambda_i \mathbf{w}_{d-1}(I - i), \tag{18}$$

where $\quad i := \min\{j \mid j \in I\}, \quad 0 < |I| < d,$

$$\mathbf{w}_d(\phi) = 0. \tag{19}$$

Lemma 3: $\quad \mathbf{w}_d(I) = \lambda_j \mathbf{w}_{d-1}(I) + \lambda_j \mathbf{w}_{d-1}(I - j),$

$$j \in I, \quad I \subset \Omega, \quad d \geq 2. \tag{20}$$

Proof: By direct verification we see that (20) is true for $d = 2$. Suppose it is true for $d \leq e$, and consider $d = e + 1$. If $|I| > e + 1$, (20) is trivial.

Suppose $|I| = e + 1$. Then, from (17),

$$\mathbf{w}_{e+1}(I) = \prod_{i \in I} \lambda_i.$$

On the other hand, if $j \in I$, then

$$\lambda_j \mathbf{w}_e(I) + \lambda_j \mathbf{w}_e(I - j) = 0 + \lambda_j \prod_{i \in I-j} \lambda_i = \prod_{i \in I} \lambda_i$$

and so (20) again holds.

Finally, suppose $0 < |I| < e + 1$, and let $i = \min\{j \mid j \in I\}$. Then by (18),

$$\mathbf{w}_{e+1}(I) = \lambda_i \mathbf{w}_e(I) + \lambda_i \mathbf{w}_e(I - i). \tag{21}$$

If $j = i$ then (20) again holds. Suppose $j \neq i$. By induction hypothesis,

$$\mathbf{w}_e(I) = \lambda_j \mathbf{w}_{e-1}(I) + \lambda_j \mathbf{w}_{e-1}(I - j)$$

$$\mathbf{w}_e(I - i) = \lambda_j \mathbf{w}_{e-1}(I - i) + \lambda_j \mathbf{w}_{e-1}(I - i - j).$$

Substitution in (21) gives

$$\mathbf{w}_{e+1}(I) = \lambda_i \lambda_j [\mathbf{w}_{e-1}(I) + \mathbf{w}_{e-1}(I - j) + \mathbf{w}_{e-1}(I - i)$$

$$+ \mathbf{w}_{e-1}(I - i - j)]$$

$$= \lambda_j [\lambda_i \mathbf{w}_{e-1}(I) + \lambda_i \mathbf{w}_{e-1}(I - i)] + \lambda_j [\lambda_i \mathbf{w}_{e-1}(I - j)$$

$$+ \lambda_i \mathbf{w}_{e-1}(I - j - i)]$$

$$= \lambda_j \mathbf{w}_e(I) + \lambda_j \mathbf{w}_e(I - j), \quad \text{by induction hypothesis.} \quad \square$$

Lemma 4: $\quad \mathbf{w}_d(I + j) = \sum_{r=1}^{d-1} \lambda_j^r \mathbf{w}_{d-r}(I), \quad j \notin I.$

Proof: From (20)

$$\mathbf{w}_d(I + j) = \lambda_j \mathbf{w}_{d-1}(I + j) + \lambda_j \mathbf{w}_{d-1}(I).$$

The assertion follows by iterating on the first term on the right. $\quad \square$

Let $N(I, d)$ denote the set of all $|I|$-tuples $n = \{n_i \mid i \in I\}$ such that $n_i \geq 1$ and $\sum n_i = d$.

Lemma 5: $\quad \mathbf{w}_d(I) = \sum_{n \in N(I,d)} \prod_{i \in I} \lambda_i^{n_i}.$

Proof: If $d = 1$, the assertion is immediate. Suppose it is true for some $d \geq 1$. From (20) and $j \in I$

$$\mathbf{w}_{d+1}(I) = \lambda_j[\mathbf{w}_d(I) + \mathbf{w}_d(I - j)]$$

$$= \sum_{n \in N(I,d)} \lambda_j \prod_{i \in I} \lambda_i^{n_i} + \sum_{n \in N(I-j,d)} \lambda_j \prod_{i \in I-j} \lambda_i^{n_i}.$$

The first term is the sum over all $n \in N(I, d + 1)$ with $n_j > 1$, while the second term is the sum over all $n \in N(I, d + 1)$ with $n_j = 1$. Hence the assertion is true for $d + 1$. □

Lemma 6:
$$(d - |I|)\mathbf{w}_d(I) = \sum_{r=1}^{d-1} \sum_{j \in I} \lambda_j^{d-r} \mathbf{w}_r(I).$$

Proof: By Lemma 5, the right-hand side equals

$$\sum_{r=1}^{d-1} \sum_{j \in I n \in N(I,r)} \lambda_j^{d-r} \prod_{i \in I} \lambda_i^{n_i}. \tag{22}$$

Each term in (22) is of the form

$$\prod_{i \in I} \lambda_i^{m_i} \tag{23}$$

for some $m \in N(I, d)$. Hence the assertion will be proved if it is shown that each of the terms (23) appears exactly $(d - |I|)$ times in (22). Fix m. Then (23) appears in the sum (22)

$$\sum_{n \in N(I,r)} \lambda_j^{d-r} \prod_{i \in I} \lambda_i^{n_i}$$

if and only if $m_j > d - r$, or $r > d - m_j$. Therefore, the term (23) appears in (22) exactly $\sum_{j \in I} (m_j - 1) = d - |I|$ times, as required. □

Next, for $1 \leq d \leq n$ define vectors $\{\mathbf{z}_d(I), I \in \Omega\}$ as follows:

$$\mathbf{z}_1(I) = \sum_{i \in I} \lambda_i \tag{24}$$

$$\mathbf{z}_d(I) = 0, \qquad 0 \leq |I| < d \tag{25}$$

$$\mathbf{z}_d(I + j) = \lambda_j \mathbf{z}_{d-1}(I) + \mathbf{z}_d(I). \tag{26}$$

Lemma 7: Let $d = |J|$ and $I \subset J$. Then for $e \geq 1$

$$\mathbf{w}_{d+e}(I) + \sum_{r=1}^{d} (-1)^r \mathbf{z}_r(J)\mathbf{w}_{d+e-r}(I) = 0.$$

Proof: If $I = \phi$ the assertion is trivial by (19). We first prove the assertion for $I = J \neq \phi$ using induction on d. If $d = 1$ and $I = \{i\}$, then

$$\mathbf{w}_{1+e}(I) = \lambda_i \mathbf{w}_e(I), \qquad \text{by (20)}$$

$$= \mathbf{z}_1(I)\mathbf{w}_e(I), \qquad \text{by (24).}$$

Suppose the assertion is true for $1, \cdots, d$. Let $|I| = d$ and $j \notin I$. Then by (20),

$$\mathbf{w}_{d+1+e-r}(I + j) - \lambda_j \mathbf{w}_{d+e-r}(I + j) = \lambda_j \mathbf{w}_{d+e-r}(I).$$

Multiplying both sides by $\mathbf{z}_r(J)$ for $r \geq 1$ and summing for $r = 0, \cdots, d + 1$ gives

$$\mathbf{w}_{d+1+e}(I + j) - \lambda_j \mathbf{w}_{d+e}(I + j)$$

$$+ \sum_{i=1}^{d+1} (-1)^r \mathbf{z}_r(J)[\mathbf{w}_{d+1+e-r}(I + j) - \lambda_j \mathbf{w}_{d+e-r}(I + j)]$$

$$= \lambda_j[\mathbf{w}_{d+e}(I) + \sum_{r=1}^{d+1} (-1)^r \mathbf{z}_r(J)\mathbf{w}_{d+e-r}(I)]$$

$$\lambda_j[\mathbf{w}_{d+e}(I) + \sum_{r=1}^{d} (-1)^r \mathbf{z}_r(J)\mathbf{w}_{d+e-r}(I)],$$

since $\mathbf{z}_{d+1}(J) = 0$ by (25)

$$= 0 \text{ by induction hypothesis.}$$

Hence,

$$0 = \mathbf{w}_{d+1+e}(I + j) - \lambda_j \mathbf{w}_{d+e}(I + j) + \sum_{r=2}^{d+1} (-1)^r [\mathbf{z}_r(J)$$

$$+ \lambda_j \mathbf{z}_{r-1}(J)]\mathbf{w}_{d+1+e-r}(I + j) - \mathbf{z}_1(J)\mathbf{w}_{d+e}(I + j)$$

$$- (-1)^{d+1}\mathbf{z}_{d+1}(J)\lambda_j \mathbf{w}_{e-1}(I + j)$$

$$= \mathbf{w}_{d+1+e}(I + j) + \sum_{r=1}^{d+1} (-1)^r \mathbf{z}_r(J + j)\mathbf{w}_{d+1+e-r}(I + j),$$

using (26) and the fact that $\mathbf{z}_{d+1}(J) = 0$ by (25).

This completes the proof for the case when $|I| = |J|$. For the case $|I| < |J|$ the proof proceeds as in the last step using induction on $|J|$ for fixed I. $\square$

AUTHOR

Bhaskarpillai Gopinath, M.Sc. (Mathematics), 1964, University of Bombay; Ph.D. (E.E.), 1968, Stanford University; Research Associate, Stanford University, 1967–1968; Alexander von Humboldt Research Fellow, University of Göttingen, 1971–1972; Gordon McKay Professor, University of California, Berkeley, 1980–1981; AT&T Bell Laboratories, 1968–1983. Present affiliation Bell Communications Research, Inc. Mr. Gopinath is engaged in research in communications and computer science.

A Vector Quantizer Combining Energy and LPC Parameters and Its Application to Isolated Word Recognition

By L. R. RABINER,* M. M. SONDHI,* and S. E. LEVINSON*

(Manuscript received July 25, 1983)

The theory of vector quantization (VQ) of linear predictive coding (LPC) coefficients has established a wide variety of techniques for quantizing LPC spectral shape to minimize overall spectral distortion. Such vector quantizers have been widely used in the areas of speech coding and speech recognition. The conventional vector quantizer utilizes only spectral shape information and essentially disregards the energy or gain term associated with the optimal LPC fit to the signal being modeled. In this paper we present a method of incorporating LPC spectral shape and energy into the code-book entries of the vector quantizer. To do this, we postulate a distortion measure for comparing two LPC vectors that uses the weighted sum of an LPC shape distortion and a log energy distortion. Based on this combined distortion measure, we have designed and studied vector quantizers of several sizes for use in isolated word speech recognition experiments. We found that a fairly significant correlation exists between LPC shape and signal energy. Hence, an LPC shape combined with energy vector quantizer with a given distortion requires far fewer code-book entries than one in which LPC shape and energy are quantized separately. Based on isolated word recognition tests on both a 10-digit and a 129-word airlines vocabulary, we found improvements in recognition accuracy by using the VQ with both LPC shape and energy over that obtained using a VQ with LPC shape alone.

* AT&T Bell Laboratories.

I. INTRODUCTION

The idea of quantizing linear predictive coding (LPC) coefficient sets using a vector quantizer (VQ), rather than a scalar quantizer, has been studied for several years.[1-5] Previously developed VQ algorithms have been widely used with good sucess in the areas of speech coding and isolated word recognition.[2-9]

The "standard" VQ algorithm quantizes the spectral shape of the LPC vector to one of M^* code-book entries, where M^* represents the number of LPC prototype vectors needed to span the space of LPC vectors with a given distortion criterion. This type of vector quantizer disregards the gain or energy associated with the LPC vector; instead it codes only the spectral shape. For LPC vocoder applications the gain of the signal is generally coded independently of the LPC spectral shape. This effectively assumes that spectral shape and signal gain are independent of each other. For standard implementations of isolated word recognition systems, signal gain information generally has not been used.[6-11]

Recently, Brown and Rabiner[12] improved the performance of an LPC dynamic time-warping (DTW) word recognition system by incorporating gain information into the conventional distortion measure. Their results indicated a substantial reduction in word error rates on a moderate-size vocabulary of words (129 words) used in an airlines reservation and information system.

In this paper we extend the work of Brown and Rabiner and show how gain information can be incorporated into the vector quantizer design algorithm to yield a set of code-book entries with both spectral shape and gain information. We show that since a nonzero correlation exists between spectral shape and gain, the number of code-book entries required to obtain prescribed levels of distortion for both spectral shape and gain is significantly fewer than would be necessary to determine the same distortion levels using separate code books for spectral shape and gain. We use the newly designed code books in an isolated word speech recognition system based on the theory of hidden Markov models (HMM) and demonstrate that this system's performance surpasses one employing LPC shape code books of the same size.

The organization of this paper is as follows. In Section II we briefly review the standard LPC VQ design algorithm. In Section III we demonstrate how we implemented the VQ design procedure using both LPC shape and energy. In Section IV, we describe results of several isolated word recognition tests using the combined LPC shape and energy VQ. In Section V we summarize the results.

II. REVIEW OF THE LPC SHAPE VQ DESIGN ALGORITHM

Assume we are given a section of a speech signal, $s(n)$ $n = 0, 1, \cdots,$

$N - 1$, with z-transform $S(z)$. From this we derive the pth-order LPC model, $\hat{S}(z)$, of the form

$$\hat{S}(z) = \frac{G}{1 - \sum\limits_{i=1}^{p} a_i z^{-i}}, \tag{1}$$

where $\mathbf{a}' = \{1, a_1, a_2, \cdots, a_p\}$ is the optimal pth-order LPC model and G is the model gain. It is readily shown that G can be written in the form

$$G = \sqrt{\sum_{n} e^2(n)} = \sqrt{\mathbf{a}'\mathbf{V}\mathbf{a}}, \tag{2}$$

where $e(n)$ is the error between the true speech samples $s(n)$ and the predicted speech samples $\hat{s}(n)$, (i.e., those obtained from the model) and $\mathbf{V}$ is the Toeplitz autocorrelation matrix of the actual speech signal, with the first row given by

$$V(m) = \sum_{n} s(n)\, s(n + m), \qquad m = 0, 1, \cdots, p. \tag{3}$$

The zeroth autocorrelation coefficient, $V(0)$, is conventionally called the signal energy.

If we want to compare two LPC models, e.g., $A_T(z)$ and $A_R(z)$, of the forms

$$A_T(z) = \frac{G_T}{1 - \sum\limits_{k=1}^{p} a_k^T z^{-k}} \tag{4a}$$

and

$$A_R(z) = \frac{G_R}{1 - \sum\limits_{k=1}^{p} a_k^R z^{-k}}, \tag{4b}$$

several related LPC distortion measures (distance metrics) have been proposed including:

1. The Itakura-Saito measure of the form

$$d_{\mathrm{IS}}(A_T, A_R) = \left(\frac{\mathbf{a}_R' \mathbf{V}_T \mathbf{a}_R}{\mathbf{a}_T' \mathbf{V}_T \mathbf{a}_T} - 1 \right) + \ln \left(\frac{G_T^2}{G_R^2} \right). \tag{5}$$

2. The log likelihood measure of the form

$$d_{\mathrm{LLR}}(A_T, A_R) = \ln \left(\frac{\mathbf{a}_R' \mathbf{V}_T \mathbf{a}_R}{\mathbf{a}_T' \mathbf{V}_T \mathbf{a}_T} \right). \tag{6}$$

3. The gain normalized measure of the form

$$d_{\mathrm{GN}}(A_T, A_R) = \left(\frac{\mathbf{a}_R' \mathbf{V}_T \mathbf{a}_R}{\mathbf{a}_T' \mathbf{V}_T \mathbf{a}_T} - 1\right). \tag{7}$$

It is readily seen that d_{LLR} and d_{GN} are essentially identical when values of d are close to zero,[13] and differ primarily for large values of d. Both d_{LLR} and d_{GN} are independent of signal energy, since the only term in the expressions of eqs. (6) and (7) that depends on signal energy is $\mathbf{V}_T$, which is cancelled out because it appears in both the numerator and the denominator. Although the d_{IS} measure of eq. (5) has a signal energy dependent term (G_T^2/G_R^2) and contains some energy information, experimentation by several researchers[2,6,7] indicates that d_{LLR} and d_{GN} are much better for designing a VQ than d_{IS}. In our own work,[5,9] we have used d_{GN} exclusively.

Using the distortion measure of eq. (7), one can define a distortion (distance) between a training LPC vector ($\mathbf{a}_T$) and a VQ code-book vector ($\tilde{\mathbf{a}}_R$). An algorithm for choosing a set of M^* code-book vectors, $\tilde{\mathbf{a}}_R$, that minimize the distortion of a set of training vectors from the code-book entries can then be devised, for example,

$$\bar{D}_I(M^*) = \min_{\{\tilde{\mathbf{a}}_R\}} \left[\frac{1}{I} \sum_{L=1}^{I} \min_{1 \leqslant m \leqslant M^*} [d_{\mathrm{GN}}(\mathbf{a}_T, \tilde{\mathbf{a}}_R)]\right], \tag{8}$$

where we have simplified the distance notation of eq. (7) to represent the distance between a test and a reference LPC vector (rather than a test and a reference LPC model). Various iterative algorithms for implementing the minimization of eq. (8) have been proposed and work well over a wide range of conditions.[1-6] The optimum code books (which we will call spectral shape code books) are generated by a method similar to the K-means algorithm. Starting with an initial guess of M^* entries, each LPC vector of the training set is assigned to the closest entry. The centroids of the M^* subsets (clusters) obtained in this manner are used as new trial entries in the code book, and the iteration is continued until some stopping criterion is satisfied. Generally, initial solutions for any desired value of M^* are obtained by first finding solutions for smaller values of M^* and then splitting some or all of the code-book entries. This way one can start from a value of $M^* = 1$, where the solution is simply the centroid of the training set vectors, and generate code books for higher values of M^*, by either splitting every cluster[1] or one cluster at a time.[5]

III. MODIFICATIONS OF THE DISTORTION MEASURE TO INCLUDE SIGNAL ENERGY

If we denote any of the LPC shape distortions of eqs. (6) and (7) as

$d_{\text{LPC}}(T, R)$, then a straightforward way of including signal energy in the overall distortion is to form the sum

$$d(T, R) = d_{\text{LPC}}(T, R) + \alpha f(d_E(T, R)), \tag{9}$$

where $d_E(T, R)$ is an energy distortion, $f(x)$ is a nonlinearity applied to the energy distortion, and α is a multiplicative factor on the energy distortions. If we define the (unnormalized) test energy as E_T, and the (unnormalized) reference energy as E_R, then

$$E_T = 10 \, \log_{10}(V_T(0)) \tag{10a}$$

$$E_R = 10 \, \log_{10}(V_R(0)). \tag{10b}$$

A normalized energy $(\hat{E}_T, \hat{E}_R)$ can be obtained making all energy values relative to a local peak energy (e.g., for isolated word recognition we make it relative to the peak energy within a word). Thus,

$$\hat{E}_T = E_T - (E_T)_{\text{MAX}}, \tag{11a}$$

$$\hat{E}_R = E_R - (E_R)_{\text{MAX}}, \tag{11b}$$

and an energy distortion, d_E, can then be defined as

$$d_E(T, R) = |\hat{E}_T - \hat{E}_R|. \tag{12}$$

The nonlinearity, $f(x)$, is used to give a smaller weight to small energy distortions. The form we have used is

$$f(x) = \begin{cases} 0, & |x| \leqslant \text{CLIP} \\ x, & |x| > \text{CLIP}, \end{cases} \tag{13}$$

where CLIP is a threshold chosen by appropriate experimentation. The use of a clipping threshold on energy distortion was proposed previously by Silverman and Dixon[14] for use in speech spectra classification studies.

The combined distortion measure of eq. (9) has the property that, as α is made small, it approaches the LPC shape distance and, as α is made large, it becomes proportional to the energy distortion alone.

Given the combined distortion measure of eq. (9), the parameters α and CLIP must be chosen in order to implement the computation. Based on results in Ref. 12, the optimum value of α is expected to be in the range 0.1 to 0.3. Similarly, based on previous experimentation,[14,15] a reasonable value of CLIP is in the range 0 to 6 dB.

3.1 Application of the combined distortion measure to VQ

It is straightforward to use the combined distortion measure of eq. (9) in the VQ design algorithm of Section II. The resulting VQ code-

book vectors are then characterized by an LPC vector, along with a normalized log energy value. To understand some of the properties of these code-book vectors, a simple set of experiments was carried out on a set of 10,000 frames of speech derived from spoken isolated words of a 129-word vocabulary of airline terms. The single words were spoken by 100 different talkers (50 male and 50 female) over a standard dialed-up telephone line.

Figure 1 shows an energy histogram of the 10,000 frames of speech. The peak level of the normalized energy of any frame is, by definition, 0 dB, and a dynamic range of about 60 dB for energy can be seen in this figure. The first experiment used the training set to design a conventional LPC shape VQ [i.e., α was set to 0 in eq. (9)]. A VQ with $M^* = 16$ was designed and each of the 10,000 training vectors was assigned to one of the $M^* = 16$ code-book entries. After convergence to the best set of code-book vectors, energy histograms of each of the 16 subsets of training vectors were made, and the results are shown in
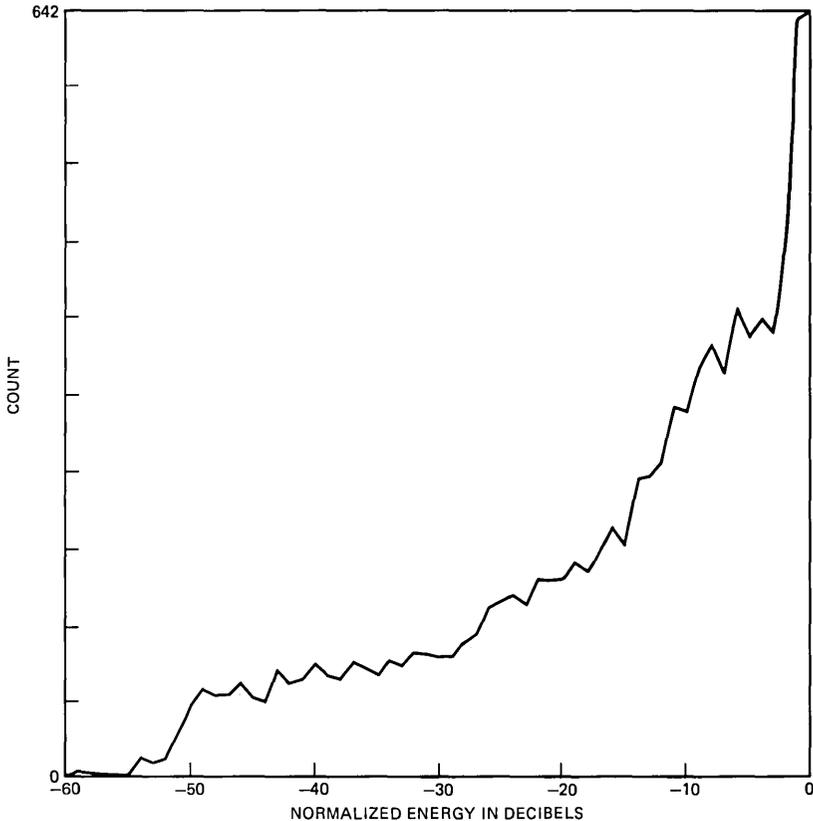

Fig. 1—Energy histogram of 10,000 frames of speech from isolated words.

Fig. 2. If signal energy and LPC shape were totally independent, we would expect each of the 16 energy histograms of Fig. 2 to be essentially identical. This is clearly not the case, as some of the energy histograms are peaked near 0 dB (i.e., strong vowels), other histograms are peaked near -50 dB (i.e., silence or weak fricatives), and still other histograms peak somewhere between these upper and lower limits.

The energy histograms of Fig. 2 indicate a fairly high degree of correlation between LPC spectral shape and normalized signal energy. Hence, one would expect that using a VQ designed from the combined distortion measure would be more efficient than using a separate VQ for LPC shape and a separate quantizer for energy.

A second set of experiments was run on the 10,000 vector training set in which the average LPC distortion, $\bar{d}_{\mathrm{LPC}}$, was determined as a function of M^* (the VQ size) for the case of $\alpha = 0$ (no energy in the distortion measure). Similarly, the average distortion, $\bar{d}_E$, was determined as a function of M^* for the case of $\alpha = \infty$ (no LPC in the distortion measure) and CLIP $= 0$. The results obtained are given in Table I. The last column of Table I gives the value of α^* where

$$\alpha^* \bar{d}_E = \bar{d}_{\mathrm{LPC}}, \tag{14}$$
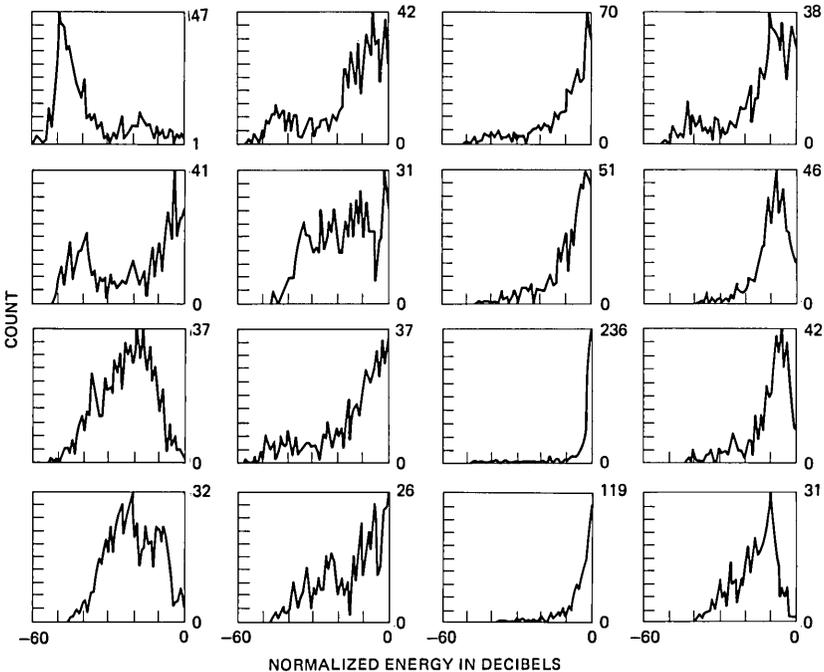
Fig. 2—Energy histograms of the 16 code words of an $M^* = 16$ shape LPC vector quantizer.

that is, the value of α (as a function of M^*) such that the average distortions due to energy and LPC would be equal. The results in Table I show that the average LPC distortion decreases slowly as M^* increases, whereas the average energy distortion almost halves with each doubling of M^*. The halving of the average distortion with each doubling of the size of the quantizer for a scalar variable is a well-known effect for scalar quantizers.[16] The last column in Table I shows that α^* increases dramatically as M^* increases. Hence, for small values of M^*, α values have to be very small or the VQ essentially becomes an energy quantizer. For larger values of M^*, the value of α is not overly important, since the LPC distortion dominates.

Based on the above discussion, LPC shape combined with energy VQs were designed for three sets of conditions, namely:

1. $\alpha = 0.1$, CLIP = 0
2. $\alpha = 0.3$, CLIP = 0
3. $\alpha = 0.3$, CLIP = 6 (dB).

The results $(\bar{d}_{LPC}, \bar{d}_E)$, as functions of M^*, are shown in Table II. For the first set of conditions, the resulting VQ achieves compromise values of the shape and energy distortions. For example, when $M^* = 64$, the LPC shape average distortion is comparable to an $M^* = 16$ VQ, based on LPC shape alone (see Table I), and the energy average distortion is comparable to an $M^* \cong 10$ VQ, based on energy alone. When α is raised to 0.3 (the second set of conditions), the energy distortion is lowered at the expense of increased LPC shape distortion for a given M^*. Thus, for $M^* = 64$ the LPC shape average distortion is now comparable to an $M^* \cong 7$ LPC shape VQ, and the energy average distortion to an $M^* \cong 18$ energy VQ. When a reasonable clipping threshold is used (CLIP = 6), the influence of the energy distortion term is significantly reduced, since all vectors within CLIP dB of each other (with similar LPC shapes) contribute a zero energy distance. Hence, for $M^* = 64$, the LPC average distortion is compa-

Table I—Average distortions as a function of M^* for VQs designed with $\alpha = 0$ (column 2), and $\alpha = \infty$, CLIP = 0 (column 3), and the α^*, which gives equal distortion

M^*	$\bar{d}_{LPC}$	$\bar{d}_E$	α^*
2	0.784	5.88	0.13
4	0.579	2.97	0.19
8	0.428	1.47	0.29
16	0.317	0.75	0.42
32	0.218	0.37	0.59
64	0.196	0.19	1.0
128	0.149	0.09	1.65

Table II—Average distortions, as a function of M^*, for several combined LPC shape plus energy VQs (10,000 training vectors)

M^*	$\alpha = 0.1$ CLIP = 0		$\alpha = 0.3$ CLIP = 0		$\alpha = 0.3$ CLIP = 6	
	$\bar{d}_{\text{LPC}}$	$\bar{d}_E$	$\bar{d}_{\text{LPC}}$	$\bar{d}_E$	$\bar{d}_{\text{LPC}}$	$\bar{d}_E$
2	0.93	8.11	1.28	5.96	1.28	4.40
4	0.73	5.22	1.18	3.22	1.20	0.99
8	0.62	3.29	0.79	2.37	0.78	0.14
16	0.49	2.34	0.70	1.29	0.50	0.04
32	0.38	1.80	0.55	0.97	0.36	0.01
64	0.31	1.30	0.45	0.69	0.26	0.01
128	0.26	0.98	0.36	0.49	0.21	0.008

rable to an $M^* \cong 29$ LPC shape VQ, and the energy average distortion is comparable to an $M^* > 128$ energy VQ.

The trend of the results of Table II reveals that for small values of M^*, the combined VQ tries to reduce energy distortion at the expense of LPC distortion, while at larger values of M^*, the VQ primarily reduces the LPC distortion. Since energy is correlated with LPC shape and vice versa, a reduction in one distortion will always bring about a reduction in the other distortion. Figure 3 illustrates a series of energy histograms of the condition 2 VQ ($\alpha = 0.3$, CLIP = 0) for the training subsets of the $M^* = 16$ case. The effects of using energy in the combined distortion measure are seen in that each histogram is tight around some average energy for the code word. Some energies are high (vowel-like sounds), some are mid-range, and some are low level (weak fricatives or silence). When CLIP is set to some nonzero value (e.g., 6 dB), most of the clusters will have a very low energy distortion. Since the spread of the energy histograms is ±6 dB or less from the average energy for the cluster in most cases, the energy distortion term is due to vectors outside the cluster.

IV. APPLICATION OF THE COMBINED VQ TO WORD RECOGNITION

To further evaluate the effectiveness of combining energy with LPC shape in the VQ, a series of isolated word recognition tests was carried out using the hidden Markov model (HMM) recognition algorithm described in Ref. 9. In this system an LPC analysis of each speech frame is carried out, and each LPC vector is vector quantized. For each word in the vocabulary, an HMM is designed using a training set of VQ outputs for the word. In normal usage each word HMM is scored by means of a Viterbi algorithm that computes the probability of the sequence of VQ outputs from the specified word HMM. The word model with the highest probability score is declared to be the spoken word.
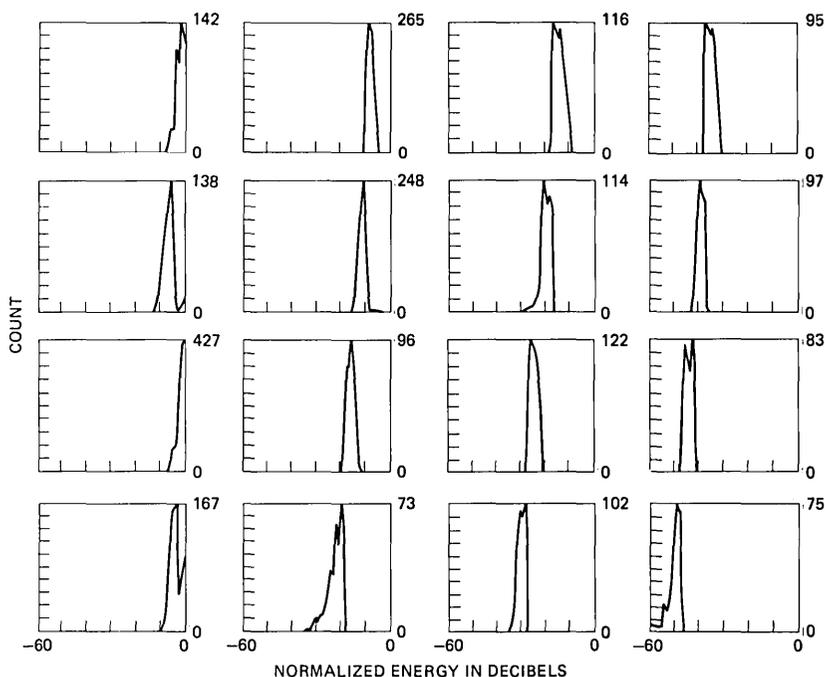
Fig. 3—Energy histograms of the 16 code words of an $M^* = 16$ LPC shape combined with energy vector quantizer designed with $\alpha = 0.3$ and CLIP = 0 dB.

The standard HMM word recognizer can be trivially modified to work with the combined LPC and energy VQ. The only change is in the quantization of the training set and of the unknown test. New HMM models are computed for the combined VQ and the standard scoring algorithm is still used in the recognizer.

The HMM recognizer using the combined LPC plus energy VQ has been tested on two vocabularies, namely the set of 10 digits, and a 129-word airlines vocabulary. The results of these tests are presented in the next sections.

4.1 Recognition results on digits vocabulary

For the 10-digit vocabulary, the training set consisted of each digit spoken once by each of 100 different talkers (50 male, 50 female). The test set consisted of a separate set of 100 tokens produced in the same way by the same 100 talkers. The test recordings were made about one month after the training recordings. All words were recorded over dialed-up telephone lines.

Three sets of VQ parameters were used in the recognition system, namely:

1. $\alpha = 0$ (no energy)
2. $\alpha = 0.1$, CLIP $= 0$
3. $\alpha = 0.1$, CLIP $= 6$ dB.

For each set of VQ parameters a set of HMM parameters were computed for each word model for the following conditions:

$$M^* = 32, 64, 128, \text{ and } 256$$

$$N = \text{Number of states in Markov model} = 5, 8, \text{ and } 10.$$

The results of the recognition tests (in terms of digit error rates) are given in Table III. The baseline LPC shape VQ (Table IIIa) has error rates of from 3 to 6 percent, depending on N and M^*. Generally, the larger the value of M^*, the lower the error rate for the recognizer. No strong dependence on N is seen in these results. The results of Tables IIIb and c show about a 1-percent reduction in error rate for the recognizers using the LPC shape combined with the energy VQ, for values of M^* of 128 and 256. For the smaller values of M^* (32 and 64), there is no consistent improvement in accuracy with the combined VQ. This result is probably due to the high distortion of the LPC shape quantization for these small size VQs.

An examination of the individual errors of the recognizer yielded no consistent pattern of errors that were corrected by the combined VQ. Since the total number of digit errors was so small (on the order of 20 out of 1000 trials), it is difficult to identify how energy information is helping in this system.

4.2 Recognition results on the 129-word airlines vocabulary

The second recognition test of the combined VQ was conducted on

Table III—Digit error rate scores for different values of N and M^*

N	M^*			
	32	64	128	256
(a) Digit error rates in percent for $\alpha = 0$ (no energy)				
5	6.0	3.9	3.8	3.8
8	6.1	3.9	4.1	4.0
10	4.8	3.1	3.0	3.0
(b) Digit error rates in percent for $\alpha = 0.1$, CLIP $= 0$				
5	5.0	4.3	2.7	2.3
8	5.3	4.0	2.1	2.4
10	5.6	2.4	2.0	2.4
(c) Digit error rates in percent for $\alpha = 0.1$, CLIP $= 6$ dB				
5	5.7	4.5	2.8	2.9
8	4.1	3.0	2.7	3.2
10	5.4	4.8	2.2	2.2

a 129-word airlines system vocabulary. The training set again consisted of each word spoken once by each of 100 talkers (50 male, 50 female) over dialed-up telephone lines. The test set consisted of each word spoken once by each of 20 new talkers (i.e., not included in the training set), again over dialed-up telephone lines.

Four recognition tests were performed under the following conditions:

1. A standard dynamic time-warping (DTW) LPC-based recognizer without VQ.

2. A DTW recognizer with an LPC-shape VQ using $M^* = 128$.

3. An HMM recognizer with an LPC-shape VQ using $M^* = 256$, with $N = 10$ states in each Markov model.

4. An HMM recognizer with the combined LPC-shape and energy VQ using $M^* = 128$, with $N = 10$ states in each Markov model.

For each test the average word error rate γ of the recognizer was measured as a function of the best β candidates. An error rate of γ for the best β candidates means the correct word was *not* in the β top recognition choices of the system γ percent of the time. Results for values of β from one (conventional word error rate) to 6 are shown in Fig. 4. The results show that the DTW recognizer without a VQ performed the best. However, the HMM recognizer with a combined VQ of the size $M^* = 128$ performed almost identically to the DTW recognizer with an LPC shape VQ of the size $M^* = 128$, and significantly better than the HMM recognizer with an LPC shape VQ of the size $M^* = 256$. The results strongly suggest that energy is a powerful discriminator for polysyllabic vocabularies, and when used in conjunction with an LPC shape VQ of moderate size (equivalent M^*, for the shape of about 32) performance is better than that achieved with even a significantly larger VQ based on LPC shape alone.

Even more convincing evidence for the advantages of the combined VQ was obtained by implementing a connected word HMM recognizer for Levinson's[17] airlines vocabulary and syntax for an airlines reservation task. When the combined LPC shape and energy VQ was used on fluent-rate sentences by six talkers in a series of informal tests of the connected word HMM recognizer, sentence accuracies of about 60 percent and word accuracies greater than 90 percent were achieved. Applying the same size VQ with LPC shape alone, sentence accuracy fell to about 5 percent, and word accuracy to about 15 percent! The energy contour constraints in the VQ appear to provide an elementary form of parsing of the sentence into syllables and words. Such time-synchronization markers are clearly necessary for ensuring any reasonable degree of accuracy in a connected word recognition task based on simple isolated word models.
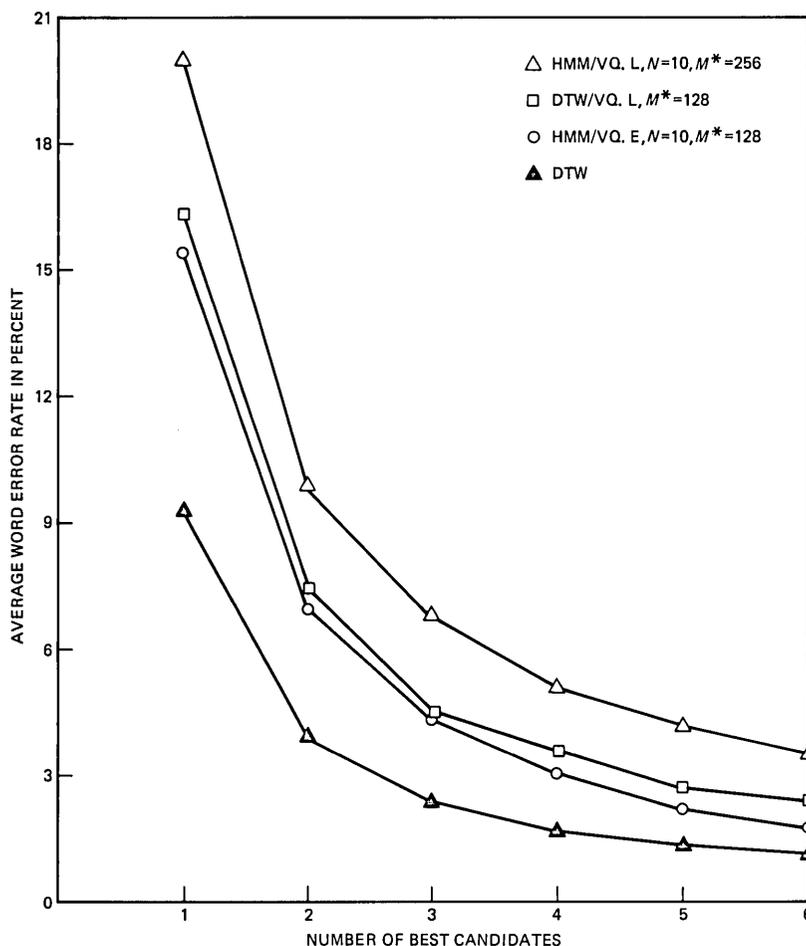
Fig. 4—Average word error rate in percent as a function of the number of best candidates for the airlines vocabulary for each of four recognition systems.

V. SUMMARY

Our results with the combined LPC shape and energy VQ indicate that the addition of energy directly into the VQ design algorithm provides an efficient method of incorporating energy constraints into an isolated word recognition system. Tests with isolated word recognizers indicate improved performance using the combined VQ over that for the LPC shape alone, provided that a sufficiently large VQ size, M^*, is used (i.e., $M^* \geq 128$).

The incorporation of energy into the VQ design algorithm is straightforward. All that is needed is a modification of the distortion

measure to include both LPC distortion and energy distortion. The form we have chosen for the combined distortion measure is based on our knowledge of distance metrics and our experience with energy contours for recognition purposes. Alternative definitions of combined distortion measures may well provide further improvements in performance.

Our results on the combined VQ algorithm indicate an improved efficiency of quantization by combining LPC shape and energy into a single VQ, over that obtained for separate LPC shape and energy quantizers. Our results could also be applied to LPC voice coding with reductions in the bit rate required to achieve desirable levels of quantization of the LPC vector and gain terms.

REFERENCES

1. Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantization," IEEE Trans. Commun., COM-28, No. 1 (January 1980), pp. 84–95.
2. B. Juang, D. Wong, and A. H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-30, No. 2 (April 1982), pp. 294–303.
3. A. Buzo et al., "Speech Coding Based Upon Vector Quantization," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-28, No. 5 (October 1980), pp. 562–74.
4. D. Wong, B. Juang, and A. H. Gray, Jr., "An 800 Bit/S Vector Quantization LPC Vocoder," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-30, No. 5 (October 1982), pp. 770–80.
5. L. R. Rabiner, M. M. Sondhi, and S. E. Levinson, "Note on the Properties of a Vector Quantizer for LPC Coefficients," B.S.T.J., 62, No. 8 (October 1983), pp. 2603–16.
6. A. Buzo, H. Martinez, and C. Rivera, "Discrete Utterance Recognition Based Upon Source Coding Techniques," Proc. ICASSP-82 (May 1982), pp. 539–42.
7. J. E. Shore and D. Burton, "Discrete Utterance Speech Recognition Without Time Normalization," Proc. ICASSP-82 (May 1982), pp. 907–10.
8. R. Billi, "Vector Quantization and Markov Source Models Applied to Speech Recognition," Proc. ICASSP-82 (May 1982), pp. 574–77.
9. L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition," B.S.T.J., 62, No. 4 (April 1983), pp. 1075–105.
10. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-23, No. 1 (February 1975), pp. 67–72.
11. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition—Theory and Selected Applications," IEEE Trans. Commun., COM-29, No. 5 (May 1981), pp. 621–59.
12. M. K. Brown and L. R. Rabiner, "On the Use of Energy in LPC-Based Recognition of Isolated Words," B.S.T.J., 61, No. 10 (December 1982), pp. 2971–87.
13. J. M. Tribolet, L. R. Rabiner, and M. M. Sondhi, "Statistical Properties of an LPC Distance Measure," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-27, No. 5 (October 1979), pp. 550–8.
14. H. F. Silverman and N. R. Dixon, "A Comparison of Several Speech-Spectra Classification Methods," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-24, No. 4 (August 1976), pp. 289–95.
15. B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-31, No. 4 (August 1983), pp. 793–807.
16. N. S. Jayant and P. Noll, Digital Coding of Waveforms, Englewood Cliffs, NJ: Prentice-Hall, 1984.
17. S. E. Levinson, "The Effects of Syntactic Analysis on Word Recognition Accuracy," B.S.T.J., 57, No. 5 (May–June 1978), pp. 1627–44.

AUTHORS

Stephen E. Levinson, B.A. (Engineering Sciences), 1966, Harvard; M.S. and Ph.D. (Electrical Engineering), University of Rhode Island, Kingston, Rhode Island, 1972 and 1974, respectively; General Dynamics, 1966–1969; Yale University, 1974–1976; AT&T Bell Laboratories, 1976—. From 1966 to 1969, Mr. Levinson was a design engineer at Electric Boat Division of General Dynamics in Groton, Connecticut. From 1974 to 1976, he held a J. Willard Gibbs Instructorship in Computer Science at Yale University. In 1976, he joined the technical staff at AT&T Bell Laboratories, where he is pursuing research in the areas of speech recognition and cybernetics. Member, Association for Computing Machinery, Acoustical Society of America, editorial board of Speech Technology; associate editor, IEEE Transactions on Acoustics, Speech and Signal Processing.

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

Man Mohan Sondhi, B.Sc. (Physics), Honours degree, 1950, Delhi University, Delhi, India, D.I.I.Sc. (Communications Engineering), 1953, Indian Institute of Science, Bangalore, India; M.S., 1955 and Ph.D. (Electrical Engineering), 1957, University of Wisconsin, Madison, Wisconsin; AT&T Bell Laboratories, 1962—. Before joining AT&T Bell Laboratories, Mr. Sondhi worked for a year at the Central Electronics Engineering Research Institute, Pilani, India, and taught for a year at the University of Toronto. At AT&T Bell Laboratories his research has included work on speech signal processing, echo cancellation, adaptive filtering, modeling of auditory and visual processes, and acoustical inverse problems. From 1971 to 1972, Mr. Sondhi was a guest scientist at the Royal Institute of Technology, Stockholm, Sweden.

Generic Approaches to the Design of Network Services Circuits

By M. MALEK-ZAVAREI,* M. C. CHOW,* and J. D. WILLIAMS*

(Manuscript received August 8, 1983)

This paper describes recent developments in modeling and analysis tools that are basic to a generic approach to the design of network services circuits for voiceband applications. Providing network services is a major part of the effort to support the tactical needs of telecommunications users. The design of network services circuits is the process of choosing configurations of transmission facilities and equipment for realizing circuits within appropriate performance specifications. Techniques for design have been developed under the general heading of "standard designs" and, until recently, have largely taken their structure from the capabilities of specific equipment.

I. INTRODUCTION

Providing public switched services constitutes a major part of the telecommunications industry. These services, collectively referred to as plain old telephone service (POTS), are provided over the public switched telecommunications network to residential and business customers. Another segment of this industry involves the offering of special services. A customer service is termed a special service if it requires special procedures and/or equipment for design, installation, or maintenance as compared to POTS. Special services provide arrangements and features that are not ordinarily available in the public network. They are generally provided to business customers by a

---

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

variety of facilities such as dedicated private communication lines, switched-service arrangements and adjuncts to POTS. Figure 1 illustrates some typical special services circuits.

We refer to POTS and special services collectively as network services. The design of network services circuits is one of the major processes in the provisioning of these services. It involves the selection of proper transmission facilities as well as the selection and placement of necessary network equipment so that the resulting circuit meets all transmission and signaling requirements. In the case of special services circuits, the design is more complicated because, generally, such circuits have more stringent transmission and signaling requirements than POTS circuits. However, they must be built from communication facilities often intended for POTS.

Most network services are extended to customers' premises over a network of transmission facilities referred to as the loop plant. This network is constructed to serve large local areas. Thus, especially in analog voice transmission, a large variation exists in the transmission characteristics of the loops.[1] This further complicates the design of network services circuits, particularly special services circuits where

EO — END OFFICE (LOCAL CENTRAL OFFICE)
FX — FOREIGN EXCHANGE
OPS — OFF-PREMISES STATION
PBX — PRIVATE BRANCH EXCHANGE
PBX-CO — PBX-CENTRAL OFFICE
WATS — WIDE AREA TELECOMMUNICATIONS SERVICE
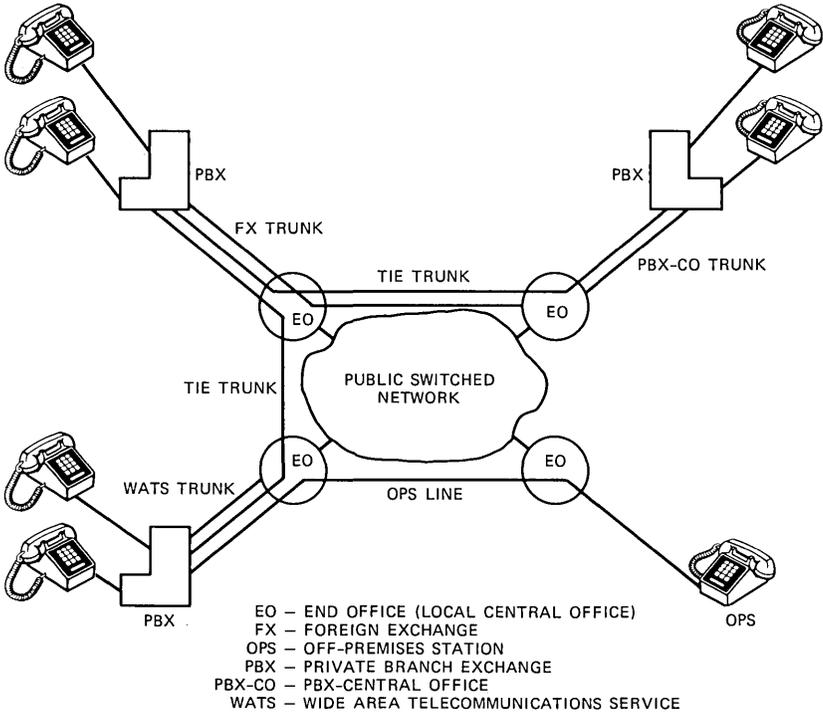
Fig. 1—Typical special services circuits.

the transmission characteristics of each link in the circuit must be tightly controlled so that the transmission requirements (such as loss, attenuation distortion, echo, and noise) in the overall circuit are kept within acceptable bounds.[2]

The general steps in the design of a network services circuit are as follows:

1. Selection of loop facilities (if any)
2. Selection of interoffice transmission facilities (i.e., routing)
3. Selection of design architecture
4. Selection of network equipment
5. Setting network equipment (i.e., adjusting equipment parameters to proper values)
6. Validation of the circuit (i.e., making sure that the circuit meets all transmission and signaling requirements).

Standard methods have been developed for the design of network services circuits.[3-5] These methods, referred to as standard design procedures, are sets of guidelines and algorithms that enable the designer to perform the above steps in the design systematically. To date, however, these procedures have been heavily dependent on the capabilities of specific network equipment rather than on their functions. That is, the procedures assume that only certain specific equipment items, whose characterizations have been stored, can be used in the design. The drawback of this is that whenever a new piece of equipment is developed, design procedures must be revised in order to incorporate it into the circuit design process. This often causes a considerable time lag between the availability of a new piece of equipment and its incorporation into the design.

To alleviate these shortcomings, a generic design concept must be developed. Such a design concept would be based on the functional features of equipment items rather than their specific technology. We discuss the generic design concept in Section II, where we also identify the theoretical basis required for the realization of the generic design process. The remainder of the paper is concerned with developing this theoretical basis. Equipment models are discussed in Section III. An analysis tool which can be used to standardize various equipment requirements in network services circuits is presented in Section IV. An illustrative example is taken up in Section III and is followed through in Section IV. A glossary of terms used in the text with their definitions is given in Appendix A. The emphasis of the discussion is on the transmission aspects of voice-grade network services circuits.

II. THE GENERIC DESIGN CONCEPT

Generic design methods are based on the functional features of network equipment. Thus they remove the dependence of design on

specific equipment technology. Steps in the generic design process include identification of appropriate facility/equipment configurations, development of standards for equipment functions, and determination of design objectives. The required subprocesses in the generic design process are listed below:

1. Estimation of equipment/facility performance capabilities
2. Definition and priority ordering of various design options (or strategies)
3. Determination of equipment/facility compatibility rules
4. Identification of various transmission/signaling options
5. Development of methods for performance validation (i.e., simulation)
6. Packaging and deployment of methods (e.g., software development).

Figure 2 shows the process flow in the provisioning of network services circuits. The top part of this figure (Databases) shows various databases that contain information about transmission facilities and equipment as well as design architectures and design objectives. The boxes in the middle row of this figure (Design Steps) show the tasks that have to be performed by the circuit designer. They start with the Service Order and culminate in the Circuit Work Order (used by the installers). Cumulative outputs of the design processes are shown above these boxes. The bottom row of Fig. 2 (Design Algorithms) shows various algorithms needed to perform the design. The designer uses these algorithms in conjunction with the databases to perform the design steps. For example, the designer accesses the loop-inventory database and uses the loop-selection algorithm to implement the loop-selection step of the design, i.e., to select a loop (or loops) for the circuit being designed.

The generic design process has a hierarchical information structure. The top level of such a hierarchy contains robust constructs for network services circuits. In such constructs one deals with the features required of various equipment rather than their identities. Each equipment is represented by a feature code which is a collection of feature elements (such as gain and equalization) that provide the appropriate transmission and/or signaling treatment for the equipment's intended use in a circuit. To fully develop these constructs, models are needed for the equipment that can provide adequate and representative characterization of equipment behavior. In such models, referred to as generic equipment models, the primary characterization of equipment would be functional rather than physical. Specific equipment could be coupled into the generic design process by linking the equipment to feature codes through a validation step.

The lower levels of the information hierarchy further specify the
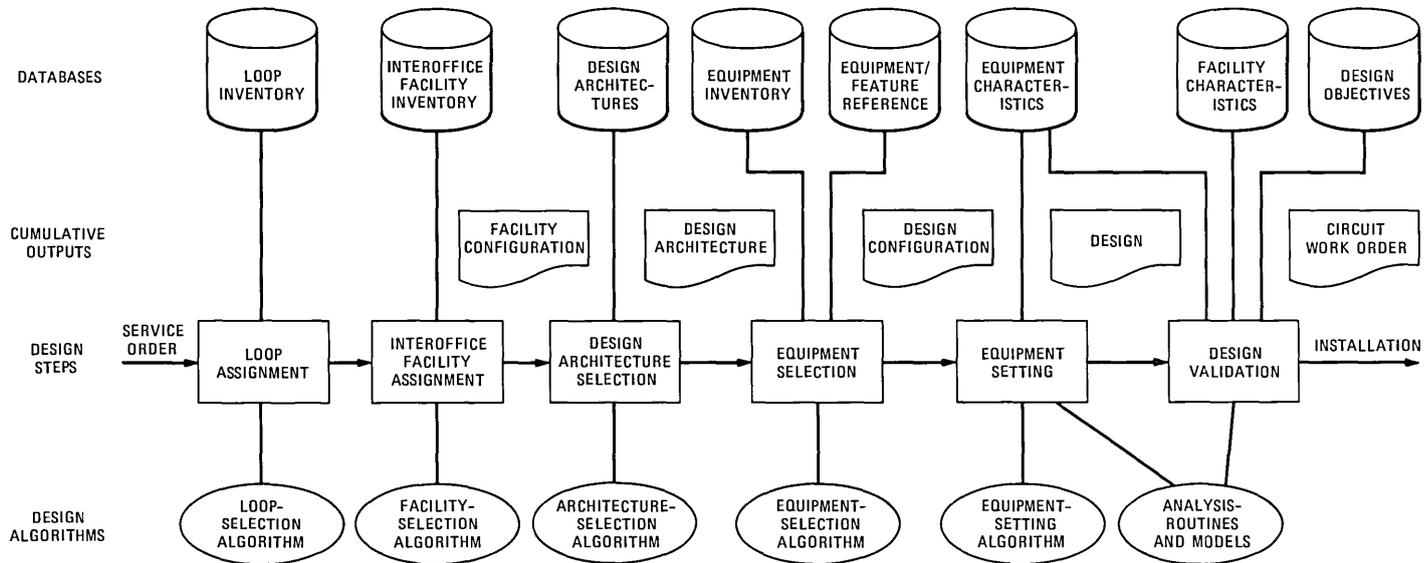
Fig. 2—Process flow in the provisioning of network services circuits.

requirements of the equipment in the circuit based on circuit configuration and service requirements. Thus the information hierarchy will be stable at the top level and only the lower levels may change due to equipment technology and type of service.

Another requirement of the generic design process is an analysis tool for evaluating various design strategies, determining equipment functional specifications, and predicting circuit performance. Such an analysis tool should be capable of simulating circuit performance based on nominal facility and eqiupment specifications. It should be able to evaluate the effects of variations in the transmission facility and equipment specifications due to manufacturing tolerances, quantization errors, temperature changes, aging, etc. The analysis tool would be used to determine correct bounds on the elements of equipment features in various design configurations. This, in turn, would allow the equipment to be categorized (or standardized) according to the bounds on the elements of their features.

The benefits of the generic design process can be summarized as follows:

1. It alleviates shortcomings in the application of the present design methods which are limited to specific network equipment. That is, it facilitates the incorporation of suitable new equipment into the circuit design process.

2. It promotes the standardization of equipment functions. This, in turn, can serve as guidelines for equipment developers.

3. It reduces the lead time required to incorporate new equipment into the circuit design process. Indeed, it enables parallel (rather than serial) interaction with the equipment developers.

4. It facilitates coordination of planning and provisioning.

We discuss equipment models in Section III. An analysis tool for the generic design process is presented in Section IV. Various information management tools would be needed to develop and deploy the generic design process. In this connection, some software tools have already been developed.[6] However, we are not concerned with this aspect of the generic design process here.

III. MODELS

A generic segment of a network services circuit is a portion of the circuit that can be functionally characterized apart from any specific equipment technology. For example, consider the network services circuit with the three 2-wire cable facilities and two repeaters shown in Fig. 3. Any facility alone, any repeater alone, or any repeater with one or two adjacent facilities can be considered as a generic segment. The models to be described in this section are for 2-wire to 2-wire, 2-wire to 4-wire (or 4-wire to 2-wire) and 4-wire to 4-wire generic
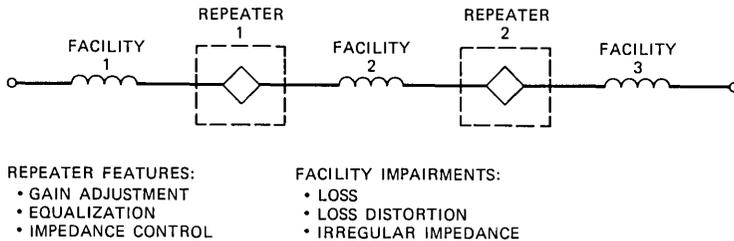
Fig. 3—A 2-wire voice-grade circuit.

segments. A generic equipment model is an equipment model in which primary characterization of the equipment is functional (e.g., transducer and return losses) rather than physical (e.g., resistors, inductors, and capacitors).

3.1 Equipment and facility considerations

As Fig. 3 indicates, the functional features of voice-grade network equipment are gain adjustment, equalization, and impedance control. The facility impairments to be considered in generic modeling are loss, loss (or attenuation) distortion, and irregular impedance. In developing generic equipment models, we consider the following generic segment measures: 1-kHz gain (or loss); gain-frequency response; 1-kHz phase delay; envelope delay distortion (EDD); and impedance control (return loss). Table I summarizes the generic segment measures.[7] Column 1 lists the generic segment measures; column 2 lists the parameter from which each measure is derived; column 3 lists the definition and the dimension of each measure; and column 4 lists the typical form of specification of each generic segment measure. For example, the 1-kHz gain is defined as $10 \log[g_{vf}(1 \text{ kHz})]$, where g_{vf} is the forward voltage gain of the equipment. This measure is typically specified by a nominal value with symmetric tolerances as 3.5 ± 1.0 dB. The balance return loss and the structural return loss in Table I are typically specified by minimum and median weighted values.[8]

3.2 Parameterization

To represent a linear network, one can use any one of the following sets of parameters: the open-circuit impedance (or the z) parameters; the short-circuit admittance (or the y) parameters, the transmission (or the ABCD) parameters; the scattering (or the s) parameters, the hybrid (or the h) parameters. In this study we will use unnormalized voltage scattering parameters[9] to characterize a generic segment for the following two reasons. First, not all networks possess the z or the y parameters; however, any linear network can be represented by a scattering matrix. Second, most of the commonly used network func-

Table I—Generic segment measures, their derivatives, definitions, and typical specifications

Generic Segment Measure	Parameter Derived From	Definition and Dimension	Typical Form of Specification
1-kHz gain	Forward gain g_{vf} Reverse gain g_{vr}	$10 \log[g_v(1 \text{ kHz})]$ Decibels (see note 1)	Nominal value with symmetric tolerances
Gain-frequency response	Forward gain g_{vf} Reverse gain g_{vr}	$10 \log[g_v(f)]$ Decibels (see note 1)	Template over band 400–2800 Hz and referenced to 1 kHz
Phase delay	Forward gain g_{vf} Reverse gain g_{vr}	$T_p(1 \text{ kHz})$ Microseconds (see note 2)	Maximum value at 1 kHz
Envelope delay distortion	Forward gain g_{vf} Reverse gain g_{vr}	$T_e(f) - T_e(f_{ref})$ Microseconds (see note 3)	Maximum difference of envelope delay between any two frequencies over 500–3000 Hz band
Balance return loss	Input: Balance reflection coefficient δ_1 Output: Balance reflection coefficient δ_2	$-20 \log[\,|\delta_1|\,]$ Decibels	Minimum weighted values; e.g., echo and singing (high and low) return loss
Structural return loss	Input: Structural reflection coefficient ρ_{11} Output: Structural reflection coefficient ρ_{22}	$-20 \log[\,|\rho_{11}|\,]$ Decibels	Minimum weighted values

1. g_v (equipment voltage gain) can be g_{vf} or g_{vr}
2. $T_p(f) = -\dfrac{1}{\omega} \arg(s_{21})$
3. $T_e(f) = \dfrac{d}{d\omega} [-\omega \cdot \arg(s_{21})]$

tions, e.g. the input and the output impedances, the transducer loss, the return loss, and so on, can be easily related to the scattering parameters (i.e., the elements of the scattering matrix).

As an example consider a 2-wire generic segment, a 2-port network, which is represented by a 2×2 scattering matrix $\{s_{ij}\}$, $i, j = 1, 2$. It can be easily shown that the input and output impedances are given by[7]

$$Z_{\text{in}} = \frac{1 + s_{11}}{1 - s_{11}} Z_{01} \qquad (1)$$

$$Z_{\text{out}} = \frac{1 + s_{22}}{1 - s_{22}} Z_{02}, \qquad (2)$$

where Z_{0i} $(i = 1, 2)$ is the impedance seen looking from the repeater into facility i with a resistive termination (nominally 600 or 900 ohms). Also the forward transducer gain is given by[7]

$$tg_f = \frac{\text{Re}(Z_{01})\text{Re}(Z_{02})}{|Z_{02}|^2} \, |s_{21}|^2, \tag{3}$$

and the reverse transducer gain is given by

$$tg_r = \frac{\text{Re}(Z_{01})\text{Re}(Z_{02})}{|Z_{01}|^2} \, |s_{12}|^2. \tag{4}$$

The return loss RL_i $(i = 1, 2)$ at port i can be shown to be[7]

$$RL_i = -20 \log |s_{ii}|. \tag{5}$$

One can see that the s parameters (i.e., s_{11}, s_{12}, s_{21}, and s_{22}) of a 2-wire generic segment can be obtained from these network functions. It should be mentioned that an input reference impedance and an output reference impedance are required in order to represent a linear 2-port network by a scattering matrix.[9] In this study, Z_{01} and Z_{02} represent the input and the output reference impedances (see Fig. 4).

3.3 Reference environments

There are two reference environments that are commonly used to specify transmission performance measures: (1) the driving-point (or impedance control) environment, and (2) the gain-loss related environment. The development of generic equipment models is based on these two environments. However, the final model is characterized in the gain-loss related environment.

The impedance control measures, such as the structural return loss,[8] balance return loss, and equal-level echo path loss (ELEPL),[7] are specified in the driving-point environment. The 1-kHz loss, the gain-frequency response, the phase delay and the envelope delay distortion (Table I) are specified in the gain-loss related environment.

The reference impedances in the gain-loss related environment are called the primary input and output reference impedance (Fig. 4). Those in the driving-point environment are called the secondary reference impedances.

3.4 Gain definition

The following approximation has been widely used by network services circuit design engineers: *the end-to-end circuit transducer loss is equal to the sum of the facility transducer losses minus repeater gains.* For a 2-wire network services circuit with only one repeater, the statement becomes valid if the following definition for forward repeater voltage gain is adopted for a repeater:[7]

$$g_{vf} = \frac{s_{21}}{(1 + \alpha_1)(1 - \alpha_2)} \sqrt{R_1/R_2}, \tag{6}$$

where

| Z_{01} OR Z_1 | W_{01} OR R_1 | W_{02} OR R_2 | Z_{02} OR Z_2 |
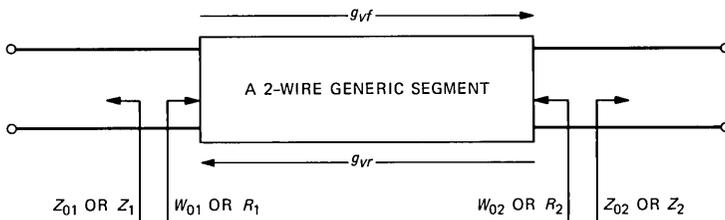
BASIC PARAMETERS:

Z_{01}/Z_{02} = primary input/output references impedances
Z_1/Z_2 = secondary input/output reference impedances
W_{01}/W_{02} = nominal input/output impedances
R_1/R_2 = nominal input/output resistances
ρ_{11}/ρ_{22} = input/output structural reflection coefficients
δ_1/δ_2 = input/output balance reflection coefficients
g_{vf}/g_{vf} = forward/reverse voltage gain

DEFINED AND DERIVED PARAMETERS:

$$\eta_i = \frac{W_{oi} - Z_{oi}}{W_{oi} + Z_{oi}}; \ \gamma_i = \frac{Z_i - Z_{oi}}{Z_i + Z_{oi}}; \ \alpha_i = \frac{R_i - Z_{oi}}{R_i + Z_{oi}}$$

$$\hat{s}_{11} = \frac{\eta_i + \rho_{ii}}{1 + \eta_i\rho_{ii}}; \ \gamma_{ib} = \frac{\gamma_i + \delta_i}{1 + \gamma_i\delta_i}; \ (i = 1, 2)$$

Fig. 4—A 2-wire generic segment and its associated parameters.

$$\alpha_i = \frac{R_i - Z_{0i}}{R_i + Z_{0i}}, \qquad i = 1, 2, \tag{7}$$

$Z_{01}(Z_{02})$ is the primary input (output) reference impedance, and R_i is the nominal input ($i = 1$) or the nominal output ($i = 2$) resistance.[7] R_1 and R_2 are often associated with the generic segment in the gain-loss environment. They are used for purposes of gain administration. They do not represent the true input and output resistances. For a circuit with two repeaters (Fig. 3), even with the gain definition given in eq. (6) the statement given earlier in this section is not precise any more. Section 3.7 will discuss this in detail.

A generic equipment model, for a 2-wire generic segment, reflecting common performance measures, has been developed based on the gain definition given in eq. (6). This model is represented by the following 2×2 scattering matrix:

$$\{s_{ij}\} = \begin{bmatrix} \hat{s}_{11} + \dfrac{1 - \gamma_{1b}\hat{s}_{11}}{2\gamma_{1b}}[1 - \sqrt{1 + x}] & g_{vr}(1 + a_2) \\ \cdot(1 - a_1)\sqrt{R_1/R_2} & \\ g_{vf}(1 + a_1)(1 - a_2)\sqrt{R_2/R_1} & \hat{s}_{22} + \dfrac{1 - \gamma_{2b}\hat{s}_{22}}{2\gamma_{2b}} \\ \cdot[1 - \sqrt{1 + x}] & \end{bmatrix}, \tag{8a}$$
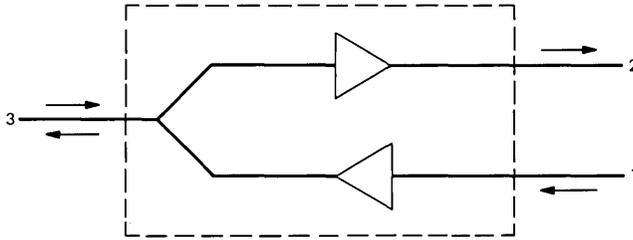
Fig. 5—A 4-wire to 2-wire (or 2-wire to 4-wire) generic segment.

where $\hat{s}_{ii}$, γ_{ib}, $i = 1, 2$, g_{vf}, and g_{vr} are defined in Fig. 4 and

$$x = \frac{4\gamma_{1b}\gamma_{2b}s_{12}s_{21}}{(1 - \gamma_{1b}\hat{s}_{11})(1 - \gamma_{2b}\hat{s}_{22})}. \tag{8b}$$

Detailed information on this 2-wire model can be found in Ref. 7.

3.5 Two-wire to four-wire and four-wire to four-wire segments

In this section we first describe the model of a 4-wire to 2-wire (or 2-wire to 4-wire) generic segment (Fig. 5). The "receive" path is designated as port 1, the "transmit" path as port 2, and the 2-wire port as port 3. The segment can be represented by a 3×3 scattering matrix.

In most applications of interest, we can assume that both the transmit and the receive paths are unidirectional. This assumption results in three zero elements: $s_{12} = s_{13} = 0$. Thus, the resulting scattering matrix becomes

$$\{s_{ij}\} = \begin{bmatrix} s_{11} & 0 & 0 \\ s_{21} & s_{22} & s_{23} \\ s_{31} & 0 & s_{33} \end{bmatrix}. \tag{9}$$

Next, we will relate the matrix elements with the transmission performance measures. The diagonal elements are related to the return losses as follows:

$$s_{ii} = \rho_{ii} \quad (i = 1, 2) \tag{10}$$

and

$$s_{33} = \frac{\eta_3 + \rho_{33}}{1 + \eta_3\rho_{33}}, \tag{11}$$

where

$$\text{return loss at port } i = -20 \log|\rho_{ii}|, \tag{12}$$

$$\eta_3 = \frac{W_{03} - Z_{03}}{W_{03} + Z_{03}}, \tag{13}$$

and Z_{03} is the primary reference impedance at port 3 while W_{03} is the nominal input impedance at port 3.

The transmit path voltage gain, g_{23}, is related to s_{23} as follows:

$$s_{23} = g_{23}(1 + \alpha_3)\sqrt{R_1/R_3}, \tag{14}$$

where R_1 is the nominal resistance at port 1 or 2, R_3 is the nominal resistance at port 3,

$$\alpha_3 = \frac{W_{03} - Z_{03}}{W_{03} + Z_{03}}, \tag{15}$$

and g_{23} is the voltage gain from port 3 to port 2. The receive path voltage gain, g_{32}, is related to s_{31} as follows:

$$s_{31} = g_{32}(1 - \alpha_3)\sqrt{R_3/R_1}. \tag{16}$$

The element s_{21} can be obtained as follows. When port 3 is terminated by a balance impedance Z_{3b}, the Equal-Level Echo Path Loss from port 1 to port 2 is given by

$$ELEPL12 = -20 \log|s_{21B}| - (T - R), \tag{17}$$

where $-20 \log|s_{21B}|$ represents the voltage gain, in dB, from port 1 to port 2, T is the transmitted signal level at 1 kHz at port 1, and R is the received signal level at 1 kHz at port 2. We define

$$\delta_3 = \frac{Z_{3b} - Z_3}{Z_{3b} + Z_3} \tag{18}$$

$$\gamma_3 = \frac{Z_3 - Z_{03}}{Z_3 + Z_{03}} \tag{19}$$

$$\gamma_{3b} = \frac{Z_{3b} - Z_{03}}{Z_{3b} + Z_{03}}, \tag{20}$$

where Z_{03} is the primary reference impedance and Z_3 is the secondary reference impedance at port 3. It can be shown that

$$\gamma_{3b} = \frac{\gamma_3 + \delta_3}{1 + \gamma_3\delta_3}. \tag{21}$$

The reference impedances for the final model are $\{Z_{01}, Z_{02}, Z_{03}\}$, while the reference impedances for s_{21B} are $\{Z_{01}, Z_{02}, Z_{3b}\}$. The conversion technique for converting the scattering parameters under a change of reference impedances must be used.[6] The conversion factors are

$$\Gamma_1 = \Gamma_2 = 0 \quad \text{and} \quad \Gamma_3 = \gamma_{3b}, \tag{22}$$

and it can be shown that

$$s_{21B} = s_{21} + \frac{\gamma_{3b}s_{23}s_{31}}{1 - \gamma_{3b}s_{33}}$$

or

$$s_{21} = s_{21B} - \frac{\gamma_{3b}s_{23}s_{31}}{1 - \gamma_{3b}s_{33}}. \tag{23}$$

Thus, a 2-wire to 4-wire (or 4-wire to 2-wire) segment is represented by a 3×3 matrix, eq. (9), with its elements described in (10), (11), (14), (16), and (23).

A 4-wire to 4-wire segment (Fig. 6), in reality, can be considered as two unidirectional 2-port networks. A 2×2 matrix can be used to represent either the transmit or the receive path:

$$\{s_{ij}\} = \begin{bmatrix} s_{11} & 0 \\ s_{21} & s_{22} \end{bmatrix}, \tag{24}$$

where s_{ii} ($i = 1, 2$) is a function of return loss at port i,

$$s_{21} = g_{21}\sqrt{R_1/R_2}, \tag{25}$$

and g_{21} is the gain from port 1 to port 2.

3.6 Example

We will use the network services circuit shown in Fig. 3 as an example for the following studies:

1. Circuit (end-to-end) transducer loss of a 2-repeater circuit
2. Applications of generic equipment models to 2-wire network services circuit design
3. Sensitivity of transmission performance measures to return loss deviations.

It can be shown that for a 2-repeater/3-facility 2-wire network services circuit the transducer loss, L (in dB), is given as[7]

$$L = \sum_{i=1}^{3} L_i - \sum_{j=1}^{2} G_j + L_e + 10 \log \frac{R_{21}R_{22}}{R_{11}R_{12}}, \tag{26}$$
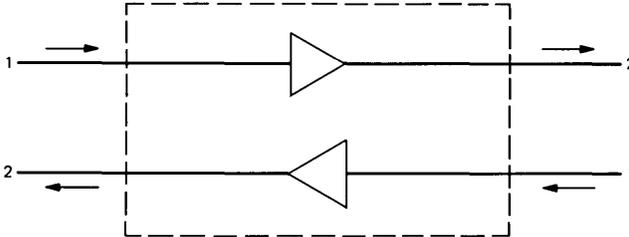

Fig. 6—A 4-wire to 4-wire generic segment.

where L_i is the transducer loss, in dB, of facility i; G_i is the voltage gain, in dB, of repeater j. Also

$$L_e = 20 \log |t_e| \qquad (27a)$$

$$t_e \simeq 1 - \rho_{221}\rho_{112}s_{212}s_{122}, \qquad (27b)$$

where ρ_{221} is the output reflection coefficient of repeater 1, ρ_{112} is the input reflection coefficient of repeater 2, s_{212} is the s_{21} of repeater 2, s_{122} is the s_{12} of repeater 2, and R_{1i} (R_{2i}) is the nominal input (output) resistance of repeater i.

In practical applications we have $R_{11} = R_{21}$ and $R_{12} = R_{22}$. Therefore eq. (26) becomes

$$L = \sum_{i=1}^{3} L_i - \sum_{j=1}^{2} G_j + L_e. \qquad (28)$$

Thus, for a 2-repeater/3-facility 2-wire network services circuit, *the circuit transducer loss is equal to the sum of facility transducer losses minus the sum of repeater gains plus an error term, L_e.* In most applications, L_e has a value between 0 and 0.3 dB.

The models developed in this section have been applied to several cases. The 2-wire network services circuit shown in Fig. 3 has been analyzed. A chi-square random number generator was used to generate return losses meeting specified minimum and median requirements. (The rationalization for using chi-square distribution can be found in Ref. 10.) A Monte Carlo method was applied for the circuit simulation.[11] The Universal Cable Circuit Analysis Program (UNICCAP) was used to analyze the specific repeaters, used in the circuit.[12] The results were in complete agreement with the results obtained from generic equipment models.

The transducer loss or the attenuation distortion requirements of a network services circuit are specified by upper and lower templates for gain-frequency response of the circuit. Let $L(f)$ be the transducer loss at any frequency f. Then the attenuation distortion $D(f)$ is defined as

$$D(f) = L(f) - L(1 \text{ kHz}). \qquad (29)$$

Figure 7a shows the upper template, D_u, the lower template, D_ℓ, and the median curve, D_m, for a typical attenuation distortion objective. For analytic purposes, these templates are smoothed as shown in Fig. 7b. Note that with a proper scale for the vertical axis in Fig. 7b, the templates specify the transducer loss instead of the attenuation distortion requirement of the circuit.

The results for the circuit shown in Fig. 3 are given in Fig. 8. It is clear from Fig. 8 that this 2-repeater/3-facility network services circuit provides a satisfactory performance.
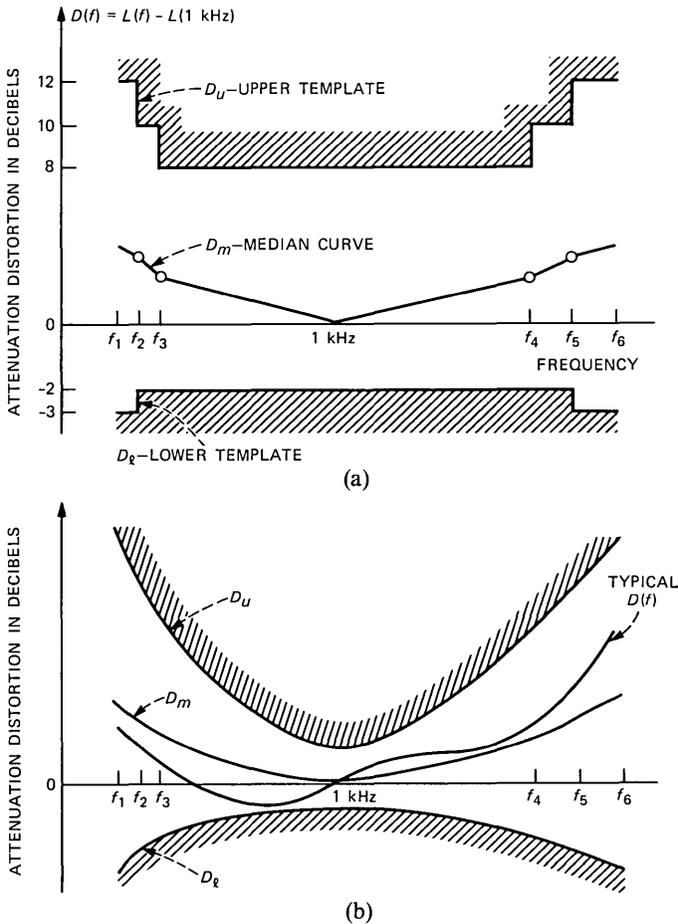
Fig. 7—Template and median curves for attenuation distortion objective: (a) Attenuation distortion objective; (b) smoothed attenuation distortion objective.

A sensitivity study of transmission performance was also made for this example. It was found that deviating all the return losses 6 dB from their nominal values would result in a circuit that would no longer provide satisfactory performance. Figure 9 shows the gain-frequency response in this case. This confirms with practical experience that return loss (or impedance mismatch) is a critical transmission performance measure.

The results obtained from this study indicate that the generic equipment models developed here are valid and useful. It is easy to apply them to assess a candidate design process in the generic design of network services circuits.
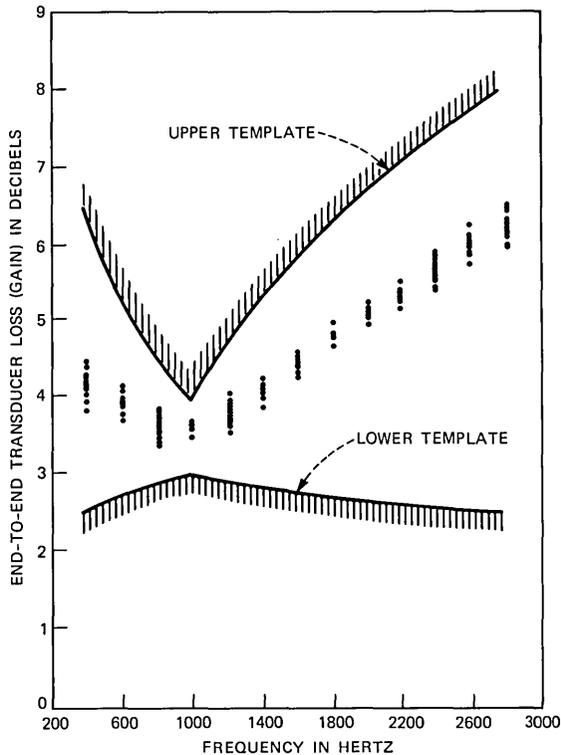
Fig. 8—Gain-frequency response with typical return-loss performance.

IV. THE ANALYSIS TOOL

Network services circuits are designed based on the *nominal* information about the transmission facilities and equipment. However, the parameters of the transmission facilities and equipment are subject to inevitable variations around their nominal values. These variations may be due to temperature changes, quantization, aging, etc. Statistical characterizations of these variations are often available from previous data. The design objective must be satisfied despite these variations.

The problems encountered here involve the determination of constraints on some parameters of the circuit while constraints on the remaining parameters of the circuit are specified. More specifically, let us assume that the design configuration, the design strategy, and the design objective are known. The associated problems can then be divided into the following three types:

1. Given the specifications of the transmission facilities, determine the type of equipment to be used. (Note that if generic equipment models are available for each design configuration, then the type of
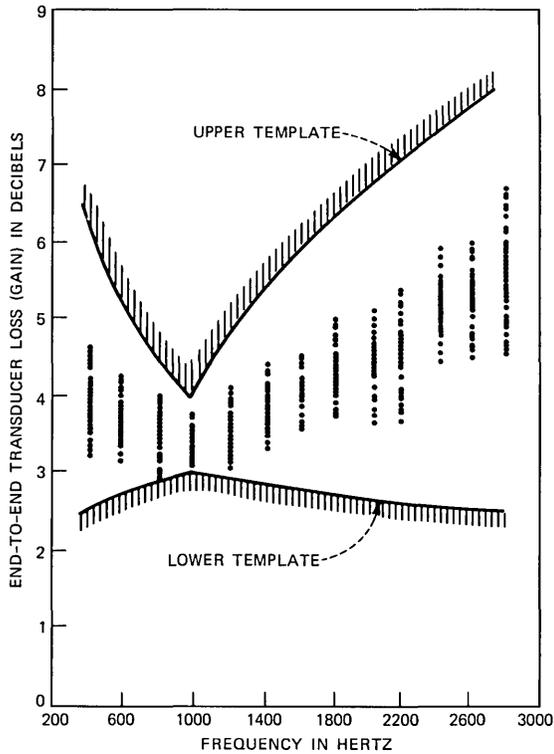
Fig. 9—Gain-frequency response with degraded return-loss performance.

equipment will refer to the constraints on the parameters of the models.)

2. Given the specifications of the equipment, determine the maximum application range of the facility sections.

3. Given partial specifications of the transmission facilities and equipment, determine the constraints on the remaining equipment parameters and the maximum range of the remaining facility sections.

Regardless of the type of problem, the analysis tool should be capable of determining the constraints on the unknown parameters of the circuit based on the available information about the specified parameters, such that the design objective is satisfied. In the following subsections we first discuss the nature of the objective functions in such problems. We then describe the analysis tool used to determine the optimal constraints. Finally, we illustrate the results by an example.

4.1 Nature of objective functions

Network services circuits must be designed such that circuit requirements on each overall performance measure are satisfied. Due to

variations in the transmission facility and equipment parameters, circuit requirements cannot be fixed at exact deterministic values. (Such exact values are referred to as *target* or *design-center* values.) The circuit requirements are, of necessity, statistical in nature; that is, variations around design target values are specified statistically.

An example of a circuit requirement is that related to circuit transducer loss at 1 kHz, $L(1000)$, as described blow. If the target circuit transducer loss at 1 kHz is $\bar{L}(1000)$, then it may be specified that in PC_i percent of circuit samples, the transducer loss is not to deviate by more than Δ_i dB from the design target value ($i = 1$, $2, \cdots$); that is, $|L(1000) - \bar{L}(1000)| \leq \Delta_i$. We define

$$NP_1 = |L(1000) - \bar{L}(1000)| \qquad (30)$$

as a normalized circuit parameter and specify statistical bounds on NP_1. The collection of such statistical circuit requirements on the normalized circuit parameters is collectively referred to as the *design objective*. Figure 10 shows an element of the design objective for a typical special services circuit, i.e., a foreign exchange (FX) line. In this figure, which refers to circuit transducer loss, we have $NP_1 = |L(1000) - 3.5|$ and

$$NP_1 < 0.5 \text{ dB}, \qquad PC_1 = 95; \qquad NP_1 \leq 1.0 \text{ dB}, \qquad PC_2 = 5. \quad (31)$$

Note that the design target value for NP_1 is zero. The cumulative
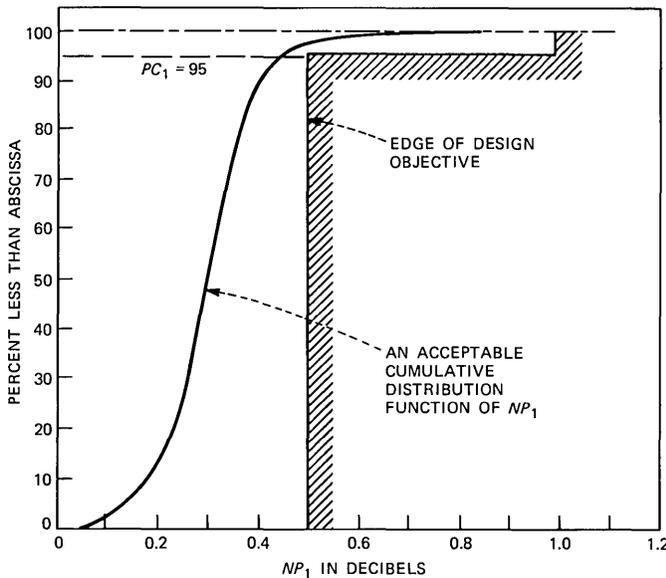
Fig. 10—Representative transducer-loss design objective for an FX line.

distribution of NP_1 for a large number of circuits designed by the application of a certain design strategy is referred to as the design performance. A typical design performance curve is also shown in Fig. 10. Note that any design performance curve that falls to the left of the "edge of the design objective" in Fig. 10 is acceptable.

Another example of a circuit requirement is that related to attenuation distortion. We define the normalized circuit parameter in this case as

$$NP_2 = \min_{400 \le f \le 2800} \psi(f), \tag{32}$$

where $\psi(f)$ is the *latent loss distortion* of the circuit, defined as follows:

$$\psi(f) = \begin{cases} \dfrac{D(f) - D_\ell}{D_m - D_\ell} & \text{for} \quad D_\ell \le D(f) \le D_m \\[2ex] \dfrac{D_u - D(f)}{D_u - D_m} & \text{for} \quad D_m \le D(f) \le D_u \\[2ex] D(f) - D_\ell & \text{for} \quad D(f) < D_\ell \\[2ex] D_u - D(f) & \text{for} \quad D(f) > D_u. \end{cases} \tag{33}$$

In the above equation, $D(f)$ is the difference between circuit transducer loss at frequency f and at 1000 Hz given in eq. (29), D_ℓ and D_u are the lower and upper bounds on the loss-versus-frequency characteristic of the circuit, and D_m is the median curve between D_ℓ and D_u. Figure 11 shows the upper and lower templates for gain-frequency response of a foreign exchange (FX) line where, for convenience, the curves D_ℓ and D_u are made piecewise linear. Figure 12 shows the statistical bounds on NP_2. In this figure we have
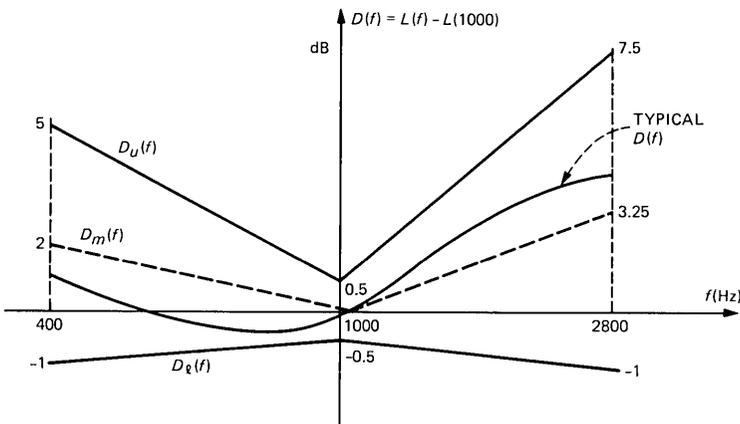

Fig. 11—Piecewise linear template and median curve for attentuation distortion design objective of an FX line.
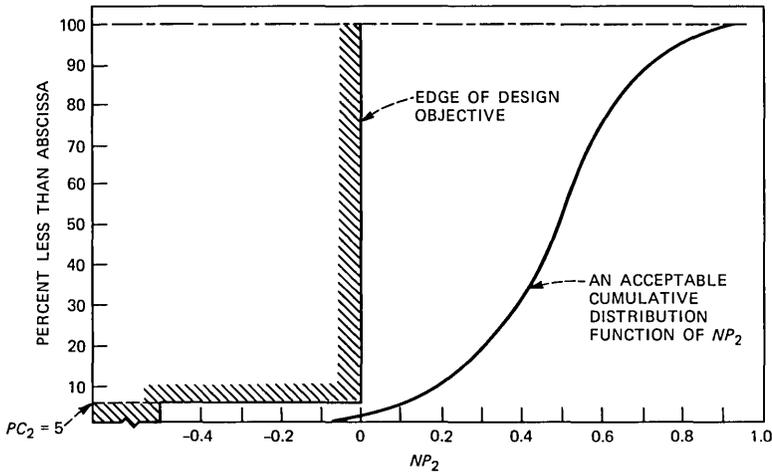
Fig. 12—Representative attenuation distortion design objective for an FX line.

$$NP_2 > -0.5, \quad PC_1 = 5; \quad NP_2 \geq 0, \quad PC_2 = 95. \quad (34)$$

Note that the target value for NP_2 (as well as the maximum value it can assume) is 1.

Equation (34) indicates that a 5-percent deviation from the design requirement related to attenuation distortion will be tolerated. Another way of interpreting this is that the design process will produce circuits that meet requirements in 95 percent of the time (assuming ergodicity).

4.2 Methods of optimization

The optimization problem under consideration can be stated as follows. Given the nominal values of and the statistical information about some parameters of the circuit, determine the constraints on the remaining parameters of the circuit such that the design objective is satisfied.

We use a Monte Carlo procedure[11] to simulate the circuit, i.e., to include the deviations from nominal parameter values, and to evaluate the design performance. To illustrate the procedure, consider a circuit including a set of unspecified parameters p_i, $i = 1, 2, \cdots, n$ (as well as a set of specified parameters) that are subject to variations. The circuit is assumed to meet the design target when parameters are at their nominal values. The value of each parameter can vary according to a certain statistical distribution within a range containing its nominal value. Parameter values may have upper bounds, lower bounds, or both upper and lower bounds. The problem is to determine

bounds on the values of parameters p_i which result from meeting the design objective. The basic method to be used consists of the following three steps:

1. Determine the nominal parameter values $\bar{p}_i$, i = 1, 2, $\cdots$, n.

2. Choose suitable initial bounds for parameter values around their nominal values.

3. Analyze and evaluate the circuits resulting from the random choice of parameter values within their corresponding (initialized or specified) bounds.

4. Determine how these bounds should be modified to improve the design performance.

Steps 3 and 4 are repeated until the design objective is met. These steps involve statistical analysis of the results and their interpretation.

After determining the nominal values for the unspecified circuit parameters, one of the following three basic approaches can be employed:

1. Choose arbitrary initial bounds on parameter values and use iterations to tighten or loosen the bounds.

2. Choose sufficiently large initial bounds on parameter values and use iterations to tighten the bounds.

3. Choose sufficiently small initial bounds on parameter values and use iterations to loosen the bounds.

We use a combination of the first and second approaches. More specifically, we use the first approach to determine suitable initial bounds on parameter values. Note that the nominal values of parameters must fall within these initial bounds. The initial bounds on parameter values must be sufficiently loose so that no unnecessary constraint is placed on any parameter. At the same time, they should not be so loose as to cause the procedure not to converge in a reasonable number of iterations. We then gradually tighten these bounds, through iterations, until the design objective is met.

The flowchart of the above procedure is shown in Fig. 13. In this figure, the INITIALIZER is a software module which provides initial bounds on parameters p_i based on the properties of these parameters and the circuit. The Monte Carlo circuit analysis program in each pass uses samples of parameters p_i, generated by proper random number generators, and analyzes the resulting circuit to determine whether or not it meets the requirements. The CIRCUIT ANALYSIS MODULE in the flowchart of Fig. 13 is a software module which, for given design configuration, facility specifications, and equipment parameters, analyzes the circuit (e.g., by using the methods of Section III) and determines the normalized circuit parameters. The results of each pass through the algorithm are stored. The INTERPRETER in Fig. 13 is a software module that uses the stored results to determine
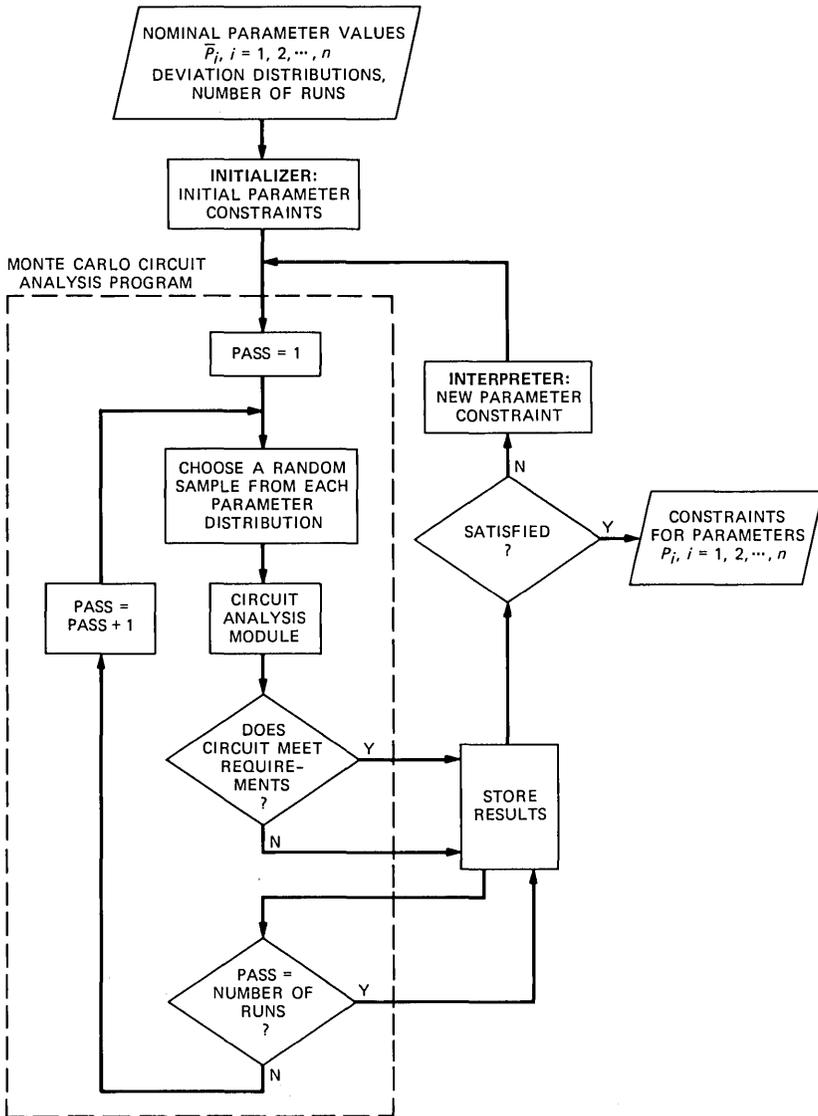
Fig. 13—Statistical analysis procedure.

the parameter whose bounds must be modified, as well as the new bounds for it such that a larger proportion of circuits will meet the requirements in the next Monte Carlo run. The algorithms used by the INTERPRETER are explained in Appendix B. The Monte Carlo runs are repeated until a set of bounds are produced for parameter values that meet the design objective.

4.3 The equipment requirements optimization process

One of the objectives of the generic design process is to facilitate the incorporation of general network equipment products into the circuit design process. To accomplish this, one has to determine the admissible ranges of values for equipment parameters used in various design configurations. In this section we will explain a computer-aided approach developed to determine the widest possible ranges for the values of equipment parameters in a given design configuration. More specifically, we assume that the design configuration and the facility specifications are known. We want to find maximum tolerances for the equipment parameters based on the given design strategy and design objective. Let us indicate the equipment parameters by p_{ij}; $i = 1, 2, \cdots, n_j$; $j = 1, 2, \cdots, N$, where n_j denotes the number of parameters related to equipment E_j, and N denotes the total number of equipment items in the circuit. The basic flowchart for the equipment requirements optimization process is shown in Fig. 14. The process consists of four steps:

1. Determine the nominal circuit solution.
2. Check the design strategy.
3. Initialize the process.
4. Analyze and interpret.

These procedures will be explained below.

4.3.1 Determine the nominal circuit solution

In this procedure the facility specifications are fixed at their nominal values. The design strategy is then applied to determine the nominal values of the equipment parameters, i.e., $\bar{p}_{ij}$, $i = 1, 2, \cdots, n_j$; $j = 1, 2, \cdots, N$. This corresponds to the initial setting of the equipment in the circuit.

4.3.2 Check the design strategy

This procedure determines whether the design strategy will lead to a satisfactory circuit when all the equipment parameters are fixed at their nominal values. Through a Monte Carlo procedure the circuit is subjected to variations in the lengths of facility sections, and the design performance (i.e., the cumulative distributions of various transmission performance measures) is determined. The block indicated as PERFORMANCE COMPARATOR in the flowchart of Fig. 14 compares the design performance P with the design objective Φ. If the design objective is met (indicated by $P \geq \Phi$), the process will continue to the initialization procedure. Otherwise, it is concluded that the design strategy is not suitable and the process will be aborted.
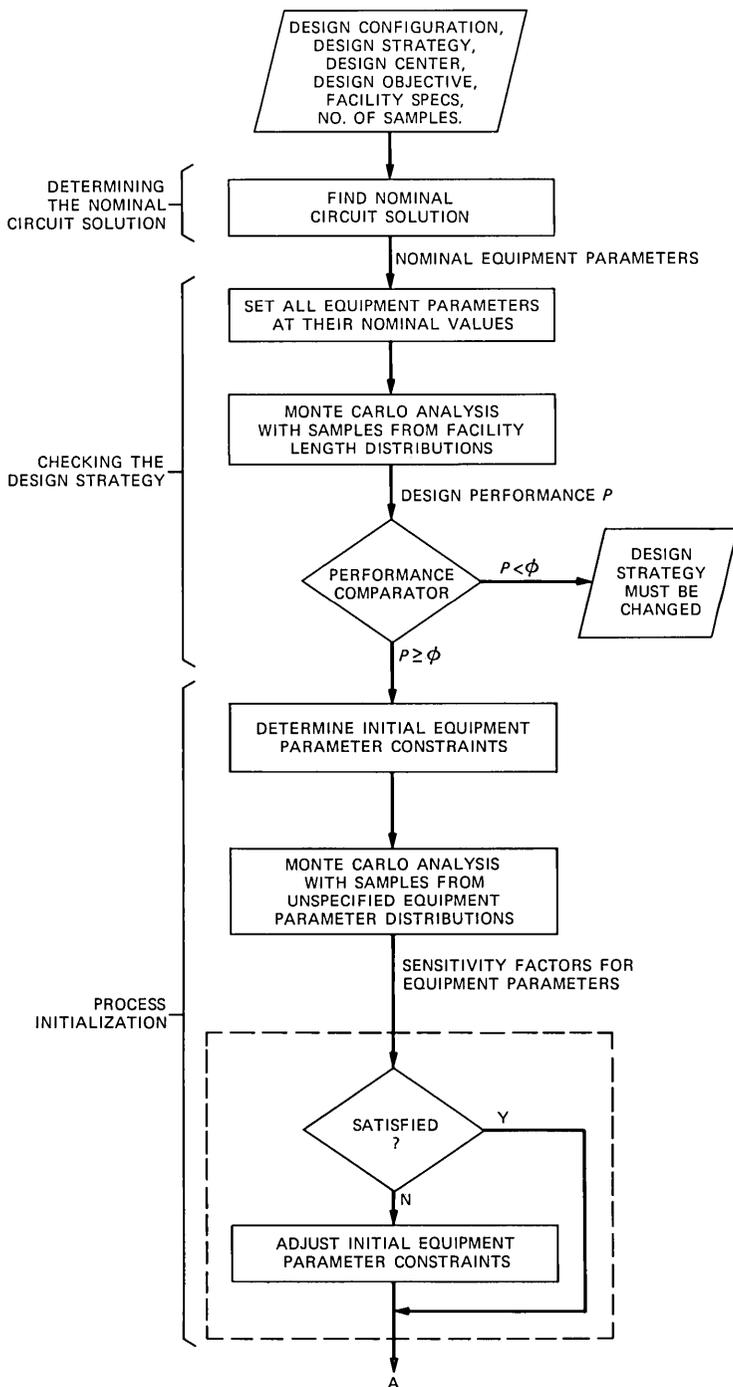
Fig. 14—Basic flowchart of the equipment requirement optimization process.

A

MONTE CARLO ANALYSIS WITH
SAMPLES FROM BOTH FACILITY
LENGTH AND EQUIPMENT
PARAMETER DISTRIBUTIONS

ϕ: DESIGN OBJECTIVE

DESIGN PERFORMANCE P

ANALYSIS AND
INTERPRETATION

PERFORMANCE
COMPARATOR

$P \geq \phi$

DETERMINE NEW CONSTRAINTS
FOR ALL UNSPECIFIED
EQUIPMENT PARAMETERS FROM
THEIR HISTOGRAMS "Y" AND "N"

$P < \phi$

INTERPRETER:
MODIFY CONSTRAINT FOR
AT LEAST ONE UNSPECIFIED
EQUIPMENT PARAMETER

OPTIMAL
CONSTRAINTS
FOR
UNSPECIFIED
EQUIPMENT
PARAMETERS

4.3.3 Initialize the process

The initialization procedure determines suitable initial constraints for equipment parameters. The method of choosing these initial constraints is as follows:

1. For one-sided equipment parameters with upper bounds, we let

$$p_{ij} = \bar{p}_{ij} + \delta_{ij}, \tag{35}$$

where random variable δ_{ij} represents variations around the nominal parameter value $\bar{p}_{ij}$. It is chosen from a uniform distribution whose mean and standard deviation are both equal to $x_{ij}|\bar{p}_{ij}|$, where x_{ij} is an input constant.

2. For one-sided equipment parameters with lower bounds, we let

$$p_{ij} = \bar{p}_{ij} - \delta_{ij}, \tag{36}$$

where the random variable δ_{ij} is chosen from a uniform distribution whose mean and standard deviation are as given above.

3. For two-sided equipment parameters, we use p_{ij} given by eq. (35), where the random variable δ_{ij} is chosen from a uniform distribution with zero mean, and the standard deviation $x_{ij}|\bar{p}_{ij}|$, where x_{ij} is an input constant. Initially all factors x_{ij} are chosen equal to x, which is an input to the process (e.g., $x = 0.1$).

To check the suitability of the initial constraints on equipment parameters based on the input value of x, a Monte Carlo analysis is performed in which samples are taken from the parameters of unspec-

ified equipment just defined. Note that at this point facility lengths will be fixed at their nominal values. A bad choice for x may overconstrain some equipment parameters, or may cause the problem not to converge to a suitable solution in a reasonable number of runs. Decision on the suitability of the value of x is based on the values of the sensitivity factors M_{ij} of parameters p_{ij} as explained below. (See Appendix B for the definition of sensitivity factor.) We compare the sensitivity factors M_{ij} with two input values M_{min} and M_{max}. (Typical values: $M_{min} = 0.4$, $M_{max} = 1.6$.)

1. If $M_{ij} < M_{min}$, it is concluded that equipment parameter p_{ij} is undersensitive to variations; thus the initial range for this parameter must be loosened.

2. If $M_{ij} > M_{max}$, it is concluded that p_{ij} is oversensitive to variations and the corresponding initial range must be tightened.

3. If $M_{min} \leq M_{ij} \leq M_{max}$, the initial range for p_{ij} will not be changed. Note that the initial range for equipment parameter p_{ij} is related to the value of x_{ij}. To adjust the initial range, new values for x_{ij} are determined according to the following algorithm, which has proved to work satisfactorily. For $i = 1, 2, \cdots, n_j$; $j = 1, 2, \cdots, N$; if $M_{ij} < M_{min}$,

$$x_{ij}^{new} = (1 + M_{min} - M_{ij})x_{ij}^{old}, \qquad (37)$$

if $M_{ij} > M_{max}$,

$$x_{ij}^{new} = (1 + M_{max} - M_{ij})x_{ij}^{old}; \qquad (38)$$

and if $M_{min} \leq M_{ij} \leq M_{max}$,

$$x_{ij}^{new} = x_{ij}^{old}. \qquad (39)$$

4.3.4 Analyze and interpret

This part of the process involves iterations during which the initial equipment parameter ranges are tightened just enough to achieve satisfactory design performance. The procedure is based on a Monte Carlo analysis where samples are chosen from both facility length and equipment parameter distributions. After each Monte Carlo run, the block indicated as PERFORMANCE COMPARATOR in the flowchart of Fig. 14 compares the design performance P with the design objective Φ. If the design objective is not met, the constraint for at least one parameter will be modified (as will be explained) and the Monte Carlo analysis procedure is repeated. The Monte Carlo runs continue until the design objective is met, at which point the final constraints for equipment parameters are determined.

The procedure for the modification of constraints for equipment parameters is based on the algorithms described in Appendix B. This procedure is as follows. Let

$$\max_{i,j} M_{ij} = M_{IJ}, \tag{40}$$

where M_{IJ} is the sensitivity factor corresponding to equipment parameter p_{IJ}. Consider the following cases:

1. p_{ij} is one-sided with an upper limit. In this case the intersection of "Y" and "N" histograms for p_{ij} will be the new upper limit (see Fig. 15a).

2. p_{ij} is one-sided with a lower limit. In this case the intersection of "Y" and "N" histograms for p_{ij} will be the new lower limit (see Fig. 15b).

3. p_{ij} is two-sided. In this case the intersection of "Y" and "N" histograms for p_{ij} defines the new range for this parameter as shown in Figs. 15c and 15d. (If "Y" histogram is on the right side of "N" histogram, the new range will be determined from the "Y" histogram, as shown in Fig. 15c.)

4.3.5 Incremental analysis

It is desirable to let the nominal characteristics (e.g., loss) of each facility section vary within a range. In such a case we would like to determine the corresponding ranges for the nominal values of equipment parameters. Also, it will still be desirable to determine the loosest possible constraints on equipment parameters for each set of nominal facility specifications. An example (Section 4.4) will demonstrate that the *nominal* values of equipment parameters depend only on the *nominal* facility specifications, and that the optimal *variations* around the nominal values of equipment parameters depend only on the *variations* around nominal facility specifications.

4.4 Example

This section demonstrates the equipment requirements optimization process on a foreign exchange (FX) line. The circuit under consideration consists of three 2-wire metallic facility sections and two repeaters, as shown in Fig. 3. In general, metallic facility sections can be specified by their physical parameters such as length, gauge, and loading characteristics. Using the Universal Cable Circuit Analysis Program (UNICCAP),[12] the loss for each facility section can then be calculated for standard terminations at any frequency. We assume, for convenience, that the facility sections in Fig. 3 are specified via their respective nominal losses at frequencies of 400, 1000, and 2800 Hz. Further, we assume that the loss at frequency f of each facility section F_i is subject to variation $\Delta_i(f)$. Thus, the losses for the facility sections are given by

$$L_i(f) = \bar{L}_i(f) + \Delta_i(f), \qquad i = 1, 2, 3$$
$$f = 400, 1000, 2800 \text{ Hz}, \tag{41}$$
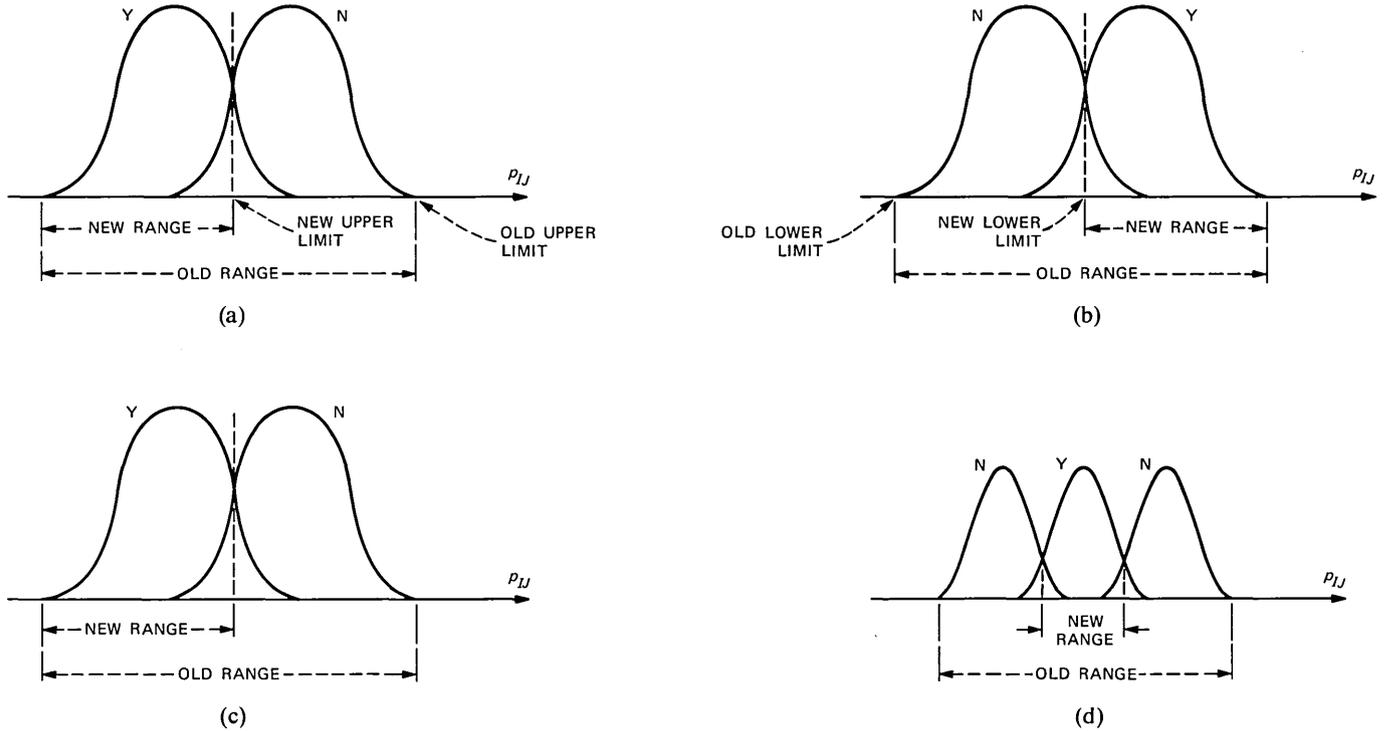
Fig. 15—Typical "Y" and "N" histograms for equipment parameter p_{IJ}. (a) One-sided parameter with upper limit. (b) One-sided parameter with lower limit. (c), (d) Two-sided parameter.

where $\bar{L}_i(f)$ indicates the nominal loss of facility section F_i at frequency f and $\Delta_i(f)$ is a random variable indicating variations around $\bar{L}_i(f)$. The type of distribution for $\Delta_i(f)$ can be specified as well. Here we assume that $\Delta_i(f)$, $i = 1, 2, 3$ are normally distributed with zero means and with specified standard deviations. We indicate the lower and the upper limits of nominal loss of facility F_i by $\bar{L}_{li}$ and $\bar{L}_{ui}$, respectively. These values as well as the values of $\Delta_i(f)$ for the example under consideration are given in Table II.

The equipment parameters used in this example are the gains of the repeaters at frequencies 400, 1000, and 2800 Hz, i.e.,

$$p_{11} = G_1(400), \quad p_{21} = G_1(1000), \quad p_{31} = G_1(2800)$$

$$p_{12} = G_2(400), \quad p_{22} = G_2(1000), \quad p_{32} = G_2(2800), \qquad (42)$$

where $G_j(f)$ indicates the gain of repeater E_j at frequency f. The above parameters are all two-sided, i.e., an upper bound and a lower bound must be determined for each. Similar to eq. (41), we can write

$$G_j(f) = \bar{G}_j(f) + \delta_j(f), \qquad j = 1, 2$$

$$f = 400, 1000, 2800 \text{ Hz}. \qquad (43)$$

We are looking for the values $\bar{G}_j(f)$ as well as the characteristics of random variables $\delta_j(f)$.

The normalized circuit parameters NP_1 and NP_2 and the design objective defined in eqs. (31) to (33) and depicted in Figs. 10 and 12 will be used in this example. Using eqs. (28), (41), and (43), we can easily show that

$$NP_1 = \left| \sum_{i=1}^{3} \Delta_i(1000) - \sum_{j=1}^{2} \delta_j(1000) \right|. \qquad (44)$$

Table II—Facility specifications

Facility Section	Nominal Loss and Std. Dev. (in dB)	At 400 Hz	At 1000 Hz	At 2800 Hz
1	$\bar{L}_{l1}(f)$	3	5	9
	$\bar{L}_{u1}(f)$	5	7	11
	$\Delta_1(f)$	0.2	0.1	0.3
2	$\bar{L}_{l2}(f)$	7	9	11
	$\bar{L}_{u2}(f)$	9	11	13
	$\Delta_2(f)$	0.1	0.05	0.2
3	$\bar{L}_{l3}(f)$	4	7	11
	$\bar{L}_{u3}(f)$	6	9	13
	$\Delta_3(f)$	0.2	0.1	0.3

Notes: $L_i(f) = \bar{L}_i(f) + \Delta_i(f)$ and $\bar{L}_{li}(f) \leq \bar{L}_i(f) \leq \bar{L}_{ui}(f)$, $i = 1, 2, 3; f = 400, 1000, 2800 \text{ Hz}.$

Also,

$$D(400) = D_m(400) + \sum_{i=1}^{3} \{\Delta_i(400) - \Delta_i(1000)\}$$

$$- \sum_{j=1}^{2} \{\delta_j(400) - \delta_j(1000)\}, \quad (45)$$

$$D(1000) = 0, \quad (46)$$

and

$$D(2800) = D_m(2800) + \sum_{i=1}^{3} [\Delta_i(2800) - \Delta_i(1000)]$$

$$- \sum_{j=1}^{2} [\delta_j(2800) - \delta_j(1000)], \quad (47)$$

where D_m is defined in Fig. 11, and $\Delta_i(f)$ and $\delta_j(f)$ are defined in eqs. (41) and (43), respectively. Thus the latent loss distortion $\psi(f)$ in eq. (32) can be calculated for $f = 400$, 1000, and 2800 Hz, and the normalized circuit parameter NP_2 in eq. (32) can be expressed in terms of $\Delta_i(f)$ and $\delta_j(f)$.

Design strategy refers to the transmission plan which, for specified facility sections, determines the nominal values of equipment parameters (i.e., the values of equipment parameters when all circuit parameters are at design-target values and facility sections are not subjected to deviations). The transducer loss of the circuit of Fig. 3 at frequency f can be calculated as follows:

$$L(f) \approx L_1(f) + L_2(f) + L_3(f) - G_1(f) - G_2(f), \quad (48)$$

where $L_i(f)$ indicates the loss of facility section F_i at frequency f Hz. Equation (48) is the same as eq. (28) where the error term L_e has been neglected for convenience. The above formula holds, of course, also for nominal values:

$$\bar{L}(f) \approx \bar{L}_1(f) + \bar{L}_2(f) + \bar{L}_3(f) - \bar{G}_1(f) - \bar{G}_2(f). \quad (49)$$

A typical design strategy for the example under consideration is as follows. For the gain adjustment aspect of the strategy, we use the following formulas:[13]

$$\bar{G}_1(1000) = \bar{L}_1(1000) + \frac{\bar{L}_2(1000)}{2} - \frac{3.5}{2} \quad (50)$$

$$\bar{G}_2(1000) = \bar{L}_3(1000) + \frac{\bar{L}_2(1000)}{2} - \frac{3.5}{2}, \quad (51)$$

where 3.5 dB is the design-target value for the 1-kHz circuit transducer loss. Solution of eqs. (50) and (51) results in the nominal values of

repeater gains at 1 kHz. For the equalization aspect of the design strategy we use the following formulas [which are similar to eqs. (50) and (51)]:

$$\bar{G}_1(f) - \bar{G}_1(1000) = \bar{D}_1(f) + \frac{\bar{D}_2(f)}{2} - \frac{D_m(f)}{2} \qquad (52)$$

$$\bar{G}_2(f) - \bar{G}_2(1000) = \bar{D}_3(f) + \frac{\bar{D}_2(f)}{2} - \frac{D_m(f)}{2} \qquad (53)$$

for $f = 400$ and 2800 Hz, where

$$\bar{D}_i(f) = \bar{L}_i(f) - \bar{L}_i(1 \text{ kHz}) \qquad (54)$$

represents the attenuation distortion corresponding to facility section F_i [cf. eq. (29)] and $D_m(f)$ is defined in Fig. 11. Solution of eqs. (52) and (53) results in the nominal values of repeater gains at frequencies 400 and 2800 Hz. Thus the implementation of the design strategy described above results in the nominal values of the six equipment parameters, i.e., $\bar{G}_j(f)$, $j = 1, 2$, $f = 400, 1000, 2800$ Hz based on the specified nominal values $\bar{L}_i(f)$ for facility sections. If the nominal values $\bar{L}_i(f)$ are allowed to vary in the range $[\bar{L}_{li}(f), \bar{L}_{ui}(f)]$, then the corresponding ranges $[\bar{G}_{lj}(f), \bar{G}_{uj}(f)]$ for each repeater nominal gain, $\bar{G}_j(f)$, must be found. To do this, we allow the nominal values $\bar{L}_i(f)$ to vary randomly within their specified ranges $[\bar{L}_{li}(f), \bar{L}_{ui}(f)]$ according to their specified (or assumed distributions). For each set of nominal facility losses, we determine the corresponding nominal repeater gains. Then the lower bounds and the upper bounds of nominal repeater gains will be determined.

A FORTRAN program, named EROP, was written for this example. Table III shows the corresponding output for the input specified in Table II. The number of iterations to achieve the final solution was 15.

Table III—EROP output for example in Section 4.4

Re-peater	Nominal Gain (in dB)	At 400 Hz	At 1000 Hz	At 2800 Hz
1	$\bar{G}_{l1}(f)$	3.750	7.881	11.469
	$\bar{G}_{u1}(f)$	6.692	10.701	13.939
2	$\bar{G}_{l2}(f)$	4.752	10.033	13.226
	$\bar{G}_{u2}(f)$	7.616	12.574	15.939

Deviations of repeater gains (in dB):

$-0.2 \leq \delta_1(400) \leq 2.0$, $-0.2 \leq \delta_1(1000) \leq 0.4$, $-2.8 \leq \delta_1(2800) \leq 1.6$
$-1.6 \leq \delta_2(400) \leq 0.2$, $-0.2 \leq \delta_2(1000) \leq 0.2$, $-1.2 \leq \delta_2(2800) \leq 2.8$

Note: $G_j(f) = \bar{G}_j(f) + \delta_j(f)$ and $\bar{G}_{lj}(f) \leq \bar{G}_j(f) \leq \bar{G}_{uj}(f)$, $j = 1, 2$; $f = 400, 1000, 2800$.

V. CONCLUSION

A perspective on the design of network services circuits has been discussed and recent developments in modeling and analysis tools have been reported. Models that have been addressed cover voiceband network equipment and are parameterized on a functional basis in terms of commonly used transmission measures. Tests show the models are consistent with the behavior of network equipment products. The optimization problem has been studied under general scenarios of fixed facility specifications, fixed equipment specifications, and partial specifications of facilities and equipment. Corresponding optimization methods have been developed and implemented in mechanized analysis tools to provide a viable procedure for determining needed standards on network equipment. These developments are providing much of the foundation needed for a generic approach to the design of network services circuits. Although the discussion has centered around voiceband cables and repeaters, the optimization techniques reported here can also be used in digital applications.

REFERENCES

1. L. M. Manhire, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., 57, No. 1 (January 1978), pp. 35–59.
2. F. C. Link and M. Malek-Zavarei, "Impact of Loop Engineering Data on the Design of Special Services," Proc. Int. Symp. Subscriber Loops and Services, Toronto, Canada, September 10–24, 1982, pp. 164–8.
3. W. S. Butler, "Mechanized Design of Special Services," Proc. IEEE Int. Conf. Comm., June 1973, pp. 50-9–11.
4. W. B. Humphries, F. F. Klapproth, and W. J. Kopp, "Simplifying the Design of Special Service Circuits," Bell Lab. Rec., 53, No. 4 (April 1975), pp. 188–96.
5. M. Malek-Zavarei and C. J. Powell, "Standard Design of Special Services Circuits—Methodologies and Processes," Proc. IEEE Int. Conf. Comm. (June 1982), pp. 1D.3.1–4.
6. T. C. Addison, S. Y. Lee, and J. J. O'Rourke, "The MEDITS Software Tools to Support Special Services Systems Engineering," Proc. IEEE Info. Com. (April 18–21, 1983), pp. 600–7.
7. M. C. Chow, "A Transmission Model Reflecting Common Performance Measures for a 2-wire Generic Segment," Proc. 13th Annual Pittsburgh Conf. on Modeling and Simulation, April 22, 1982, pp. 611–16.
8. Telecommunications Transmission Engineering, Vol. 3, Second Edition, AT&T, 1977.
9. H. J. Carlin and A. B. Giordano, Network Theory: An Introduction to Reciprocal and Nonreciprocal Circuits, Englewood Cliffs, N.J.: Prentice-Hall, 1964.
10. S. Basu and M. Malek-Zavarei, "Enhancement of Transmission Models for Network Services Circuit Design," Proc. Phoenix Conf. on Computers and Communications, March 19–21, 1984.
11. C. L. Semmelman, E. D. Walsh, and G. T. Daryanani, "Linear Circuits and Statistical Design," B.S.T.J. 50, No. 4 (April 1971), pp. 1149–72.
12. L. S. DiBiaso and R. A. McDonald, "UNICCAP: Calculated Help for Transmission Engineers," Bell Lab. Rec., 51, No. 6 (June 1973), pp. 170–3.
13. E. A. Nelson and J. Wu, "Generic Transmission Loss Constraints for a 2-wire Metallic Transmission Model," Proc. 13th Annual Pittsburgh Conf. on Modeling and Simulation, April 1982, pp. 605–10.
14. A. Papoulis, Probability, Random Variables and Stochastic Processes, New York: McGraw-Hill, 1965.
15. N. J. Elias, "The Application of Statistical Simulation to Automated Analog Test

Development," IEEE Trans. Circuit Theory, *CAS-26,* No. 7 (July 1979), pp. 513–17.

APPENDIX A

Glossary

This appendix describes the terminology used in the main text of the paper. The terms are arranged in alphabetical order.

Circuit—The telecommunications transmission path providing the network services extending from one end (customer location or central office) to another.

Circuit parameters—Parameters that are important measures of circuit transmission characteristics (e.g., circuit transducer loss).

Design architecture—A circuit topology consisting of generic transmission facilities and equipment.

Design center (or design target)—Deterministic values that are specified targets for circuit parameters.

Design configuration—A circuit topology consisting of physical equipment and transmission facilities.

Design objective—Specified bounds on the statistical distributions of normalized circuit parameters.

Design performance—Statistical distributions of normalized circuit parameters.

Design strategy—A transmission plan that determines the nominal values of equipment parameters.

Equipment functional specifications—Constraints for values of equipment parameters with their corresponding distributions within the constraints.

Equipment parameters—A set of quantities whose values determine the characteristics of the equipment.

Facility configuration—A circuit configuration that represents a physical combination of transmission facilities without specifying the junction points between facilities.

Facility section—A portion of a circuit consisting of a transmission facility between two pieces of equipment.

Facility specifications—The type and the physical specifications of each facility section in the circuit, or equivalently, the transmission parameters of each facility section that are relevant to the measurement of circuit transmission characteristics.

Generic equipment model—A model whose primary parametric characterization is functional (e.g., transducer and return losses) rather than physical (e.g., resistance, inductance, and capacitance).

Generic segment—A circuit segment that is functionally characterized apart from any specific equipment technology.

Generic segment measures—The causes whose effects are the transmission performance measures.

Incremental analysis—Circuit analysis where only variations of parameters around their nominal values are considered.

Nominal circuit solution—The collection of nominal equipment parameters.

Nominal equipment parameters—The values of equipment parameters determined on the basis of the given facility specifications (with no deviations) and transmission plan such that all the circuit parameters are at design-center values.

Normalized circuit parameters—Parameters used to measure circuit performance. These parameters are derived from circuit parameters.

One-sided equipment parameters—Equipment parameters that are constrained either from above or from below.

Transmission performance measures—Measures that can characterize the transmission performance of a circuit.

Two-sided equipment parameters—Equipment parameters that are constrained both from above and from below.

APPENDIX B

The 'Interpreter' in the Statistical Optimization Program

B.1 Algorithm for modification of parameter ranges

For each random sample of parameters p_i, $i = 1, 2, \cdots, n$, we determine whether or not the resulting circuit meets the worst-case requirements defined by the design objective. Therefore, at the end of each Monte Carlo run we have two histograms (or probability density functions) of the values of each unspecified parameter corresponding to the circuits that meet or do not meet such requirements. We indicate these histograms by "Y" and "N" as shown in Fig. 16. From these histograms we will determine the ranges of values for parameters p_i over which the proportion of circuits meeting such requirements predominate.

Figure 16 illustrates the case where parameter p_i was assigned values in the original range, i.e., $a < p_i < c$, and the circuits with low values of p_i generally met the requirements. A new range for p_i is determined from the intersections of "Y" and "N" histograms as shown in Fig. 16. By restricting the values of parameter p_i to the new range (i.e., $a < p_i < b$) while the ranges of other parameters are unchanged, an increase will be expected in the proportion of circuits meeting the requirements in the next Monte Carlo run. This can be proved by the application of Bayes' Theorem of Statistics,[14] which relates a posteriori probability of events A_k subject to the occurrences of event B, $P_B(A_k)$, to a priori probabilities $P(A_j)$:

$$P_b(A_k) = \frac{P_{A_k}(B)P(A_k)}{P(B)} = \frac{P_{A_k}(B)P(A_k)}{\sum_{j=1}^{n} P_{A_j}(B)P(A_j)}. \tag{55}$$

In eq. (55), events A_j, $j = 1, 2, \cdots, n$ are mutually exclusive and their union equals the universal event, and event B is arbitrary. Now consider the events

$$A_1 = Y: \text{circuit meets requirements}$$

$$A_2 = N: \text{circuit does not meet requirements}$$

$$B: a < p_i < b. \tag{56}$$

Note that

$$P(N) = 1 - P(Y). \tag{57}$$
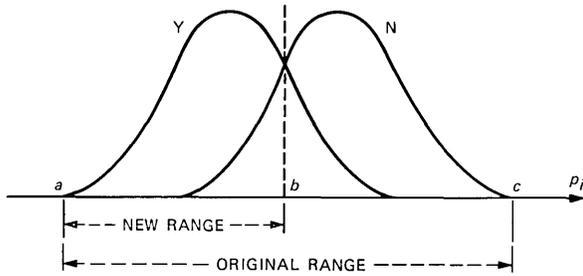
Fig. 16—Histograms of parameter p_i corresponding to circuits that meet the requirements "Y" or do not meet the requirements "N".

Thus, for $k = 1$ and $n = 2$, eq. (55) becomes

$$\underset{a < pi < b}{P\,(Y)} = \frac{P_Y(p_i < b)P(Y)}{P_Y(a < p_i < b)P(Y) + P_N(a < p_i < b)P(N)}. \quad (58)$$

We want to show that the above algorithm for range modification at the end of each Monte Carlo run leads to an improvement on the probability of circuits meeting the requirements. That is, we want to prove that

$$\underset{a<p_i<b}{P(Y)} > P(Y). \quad (59)$$

Histograms "Y" and "N" in Fig. 16 represent conditional probability densities $P_Y(p_i)$ and $P_N(p_i)$ for parameter (random variable) p_i in circuits meeting or not meeting the requirements, respectively. We have, from Fig. 16,

$$P_Y(p_i) > P_N(p_i) \quad \text{for} \quad a < p_i < b. \quad (60)$$

Thus,

$$P_Y(a < p_i < b) = \int_a^b P_Y(p_i)dp_i > \int_a^b P_N(p_i)dp_i$$
$$P_N(a < p_j < b). \quad (61)$$

Using eqs. (57) and (61) in eq. (58) and noting that $0 < P(Y) < 1$ yields eq. (59).

B.2 Algorithm for choosing the parameter for treatment

At the end of each Monte Carlo run we must determine the parameter whose range of values most influences the proportion of circuits meeting the requirements. This can be accomplished by a sensitivity factor M_i for each parameter p_i as follows:[15]
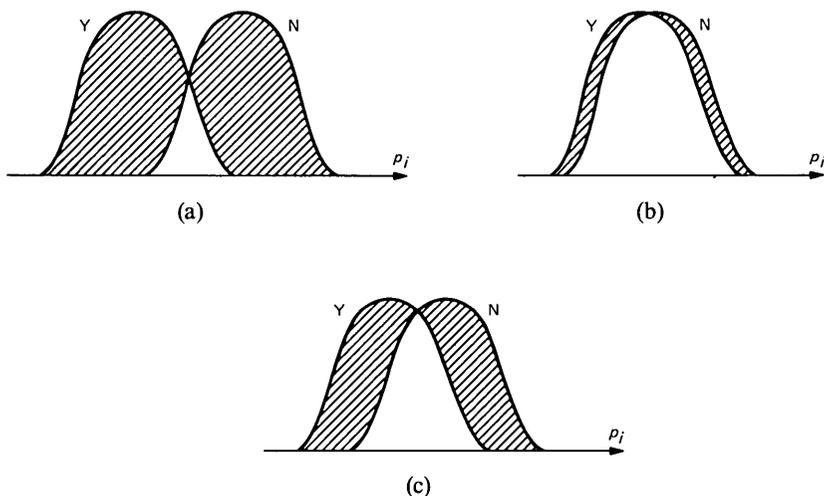
Fig. 17—Sensitivity factor M_i for parameter p_i. (a) $M_i \approx 2$. (b) $M_i \approx 0$. (c) $0 < M_i < 2$.

$$M_i = \int_{-\infty}^{+\infty} [P_Y(p_i) - P_N(p_i)]dp_i. \tag{62}$$

Sensitivity factor M_i is a measure of overlap of the probability density functions $P_Y(p_i)$ and $P_N(p_i)$. In fact, it is the area of the nonoverlapping parts of these density functions. It varies between zero for the case of complete overlap and two for the case of no overlap.

Figure 17 shows three different cases of overlap. Figure 17a corresponds to the case where circuits with sufficiently low values of p_i meet the requirements independent of variations in any of the other parameters. That is, in this case, parameter p_i is the determining factor in whether or not the circuit meets the requirements.

Figure 17b corresponds to the situation where the value of parameter p_i has no significant effect on whether or not the circuit meets the requirements. That is, the determining factor must be sought among other parameters. The case shown in Figure 17c is between the above two extreme situations.

At the end of each Monte Carlo run, the parameter with the largest sensitivity factor will be chosen for range modification, while keeping the ranges of values for other parameters unchanged.

AUTHORS

Ming-Chwan Chow, B.S.E.E., 1960, National Taiwan University; M.S.E.E., 1966, and Ph.D. (Electrical Engineering), 1968, University of Missouri-Rolla; AT&T Bell Laboratories, 1968–1983. Present affiliation Bell Communications

Research, Inc. Mr. Chow has worked on video signal transmission and studies of efficient encoding algorithm for video signals, high-speed multiplexing, echo control, time-slot interchanging, modulation techniques for optical fiber systems, and switched networks data communication performance. He has also studied the generic approach to standard design methods for special services. Mr. Chow teaches graduate courses in the Electrical Engineering department of Fairleigh Dickinson University. He is the author of technical papers in the areas of variable-length encoding for video signals and in the areas of generic models for special services circuits. He holds two U.S. patents. Member, IEEE.

Manu Malek-Zavarei, B.S.E.E., 1965, University of Tehran; M.S.E.E., 1968, and Ph.D., 1970, University of California, Berkeley; Bell Laboratories, 1970–1971; Professor, Dept. of Electrical Engineering, Shiraz University, Iran, 1971–1980; Hughes Aircraft Co., 1976–1977; Visiting Professor of Electrical and Computer Engineering, University of New Mexico, 1980–1981; AT&T Bell Laboratories, 1981–1983. Present affiliation Bell Communications Research, Inc., where he is Manager of the Standard Designs Systems Engineering District. He is also Adjunct Professor of Electrical and Computer Engineering at Stevens Institute of Technology. Mr. Malek-Zavarei is the author and coauthor of 3 books and over 40 technical papers. Senior Member, IEEE; member, Sigma Xi, Eta Kappa Nu.

John D. Williams, B.S.E.E., 1966, Syracuse University; M.S.E.E., 1968, Polytechnic Institute of Brooklyn; and Ph.D. (Electrical Engineering), 1978, Rutgers University; AT&T Bell Laboratories, 1966–1983. Present affiliation Bell Communications Research, Inc. Mr. Williams has worked on voice-frequency transmission, and modeling approaches and analysis tools for the study of transmission systems. He became Supervisor of the Special Services Standard Designs Group in 1981. Mr. Williams is presently Manager of the Transmission Studies and Analysis District in the Operations Systems—Interoffice and Special Services Center. Member, Tau Beta Pi, Eta Kappa Nu.

1980 Bell System Noise Survey of the Loop Plant

By D. V. BATORSKY* and M. E. BURKE†

This paper presents the principal findings of measurements taken in 1980 of voltages and current induced in Bell System loops by commercial power lines. These findings give designers of new, solid state, electronic switching and terminal equipment a detailed characterization of power induced signals. The last Bell System noise survey was in 1964, and since then power system load and telephone loop length have increased. Use of shielded cable has grown, and the number of poorly balanced party lines has decreased. The net result is that the loop plant contribution to main-station levels of induced longitudinal and metallic voltages indicate only a slight increase at the 90th percentile of 1 to 2 dB since 1964—an increase too small to be considered statistically significant. When the central office and loop plant contributions are combined, main station metallic noise shows a decrease at the 90th percentile of 6 dB since 1964. Because of diurnal variations, distributions of peak noise levels on a loop over a 24-hour period are 4 to 6 dB higher than corresponding distributions based on one-time measurements made during business hours. This survey provides the first characterization of induced longitudinal current based on measured data. Finally, as a result of the stratified, two-stage sampling scheme used for the survey, differences among urban, suburban, and rural environments have been identified, and the unique behavior of long rural loops has been highlighted.

I. INTRODUCTION

1.1 Survey motivation and objectives

Since the earliest days of telephony, commercial power systems have imposed significant constraints on the design of the Bell System

---

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

network. In fact, the evolution from open wire to shielded cable with twisted pairs was in part a response to the need for reduced susceptibility to induced noise. Despite the tremendous progress in reducing plant susceptibility through use of shielded cables, improved mitigation techniques,[1] and better-balanced plant, induction in the loop plant remains a major concern because of increasingly stringent design specifications and changing service requirements. For example, modern solid state transmission equipment is often more vulnerable to damage from longitudinal voltages than were earlier generations of equipment. In addition to existing Bell System objectives, several public utility commissions have recently established noise standards for subscriber loops. This latter development, along with the restructuring of the Bell System, implies a need to develop programs to assess the performance of both power and telephone systems with respect to noise parameters. A detailed knowledge of noise in the loop plant is necessary to address these and similar issues.

Since the last Bell System noise survey, which was conducted in 1964,[2] significant changes have occurred in the power and telephone plants. The open-wire plant, which still served a significant percentage of loops in 1964, is virtually nonexistent today, and poorly balanced party lines, once a major cause of noisy loops, are rarely encountered. In contrast to these positive developments, power system loads have doubled since 1964, and homes, businesses, and utilities increasingly employ nonlinear devices, which can be potent sources of telephone interference. In addition, subscriber loop lengths have increased approximately 14 percent since 1964.[3] The net effect of such changes cannot easily be predicted, but it is clear that the plant conditions present in 1964 no longer exist. Also, the 1964 noise data provide no information on longitudinal currents, noise spectra, central office voltages, or the effects of diurnal variations, all of which are now considered important to design and performance specifications.[4,5] Until now these parameters have been considered only by more localized measurement programs.[6]

Thus, the pressing need for up-to-date noise characterizations and the limitations of previous systemwide noise data led the Electrical Protection and Interference Department of Bell Laboratories to conduct the 1980 Bell System noise survey. The result of that survey is a characterization of induced noise at subscriber and central office ends of telephone pairs in the Bell System, with special consideration of the noisier long pairs.

1.2 Scope of paper

This paper discusses the design and presents the results of the 1980 Bell System noise survey. The presentation emphasizes the informa-

tion most relevant to establishing induced noise characterizations. Section II briefly introduces the physical mechanisms and models associated with induction in the loop plant. Section III describes the survey design, defines measured and derived noise parameters, and concludes with a summary of the principal results. Section IV presents noise data measured at the central office interface, and Section V presents data measured at the main station interface. Finally, Section VI provides concluding comments. The appendix expands on the survey design, details the method of sample selection, and illustrates the mathematical technique used to derive means and cumulative distributions.

II. OVERVIEW OF POWER INDUCTION

Power distribution lines are the principal source of induction in telephone lines, and are a natural consequence of both systems serving a common public and frequently sharing rights of way. Power line induction is a function of the magnitude of the net three-phase power currents and the distance over which telephone lines are exposed to these currents.[7,8] Urban, suburban, and rural areas represent significantly different environments with respect to these parameters.[6] At one extreme are the densely populated urban areas, which tend to have relatively well-balanced, three-phase power systems and relatively short telephone loops. At the other extreme are rural areas, which are typically characterized by single-phase power distribution and long loops. Suburban areas generally encompass a wide range of intermediate conditions.

The induced longitudinal (or common mode) disturbances,[9,10] i.e., noise-to-ground and longitudinal current (as defined in Section 3.2), that result from power line exposures are essentially equal for both wires of a pair.[10] Metallic noise (defined in Section 3.2) results primarily from electrical imbalances or asymmetries in the circuits formed by each wire of the pair. Since the metallic noise level depends on the longitudinal excitation, a measure of imbalance is the ratio of metallic noise to longitudinal noise (see Section 3.2 for a specific definition).

Imbalances result from differences between series resistances and shunt capacitances of the two wires of a pair. (Components of this capacitance are discussed by Miller.[11]) These imbalances determine the level of noise appearing on a subscriber circuit. Therefore, wire pair cables are manufactured to minimize asymmetries, and circuit terminations are designed to achieve good balance.[10,12]

Power induction influences telephone system design in two distinct ways. The 60- and 180-Hz spectral components, which usually dominate the induced signals, affect dissipation ratings, power arrange-

ments, and signaling characteristics of telecommunication equipment interfacing with the loop plant.[6,13,14] The voiceband harmonics of 60 Hz have an impact on transmission quality. The harmonics usually arise from nonlinearities of the power distribution transformers, which, because of symmetry conditions, predominantly generate odd multiples of 60 Hz.[8,15-17] Among these harmonics, the "triple-odd" harmonics of 60 Hz, i.e., odd multiples of 180 Hz, are of particular interest because they tend to add in phase on all three phases of a multigrounded-neutral (MGN) power distribution system.

Diurnal variations in power system loads not only produce significant changes in induction levels but may also cause changes in spectral content.[6] For instance, an increase in the power system load increases the 60-Hz component of the current and produces greater 60-Hz induction in the telephone plant. However, harmonic levels may be reduced, because the increased power system current is accompanied by an increased voltage drop in the power lines. As a result of the increased drop, the voltage excitation of distribution transformer cores is slightly reduced, which leads to reduced harmonic generation.[13,16]

Although the above paragraphs highlight those aspects of induction phenomena most relevant to a basic analysis of the survey data, an appropriate measure of exposure length requires further discussion. From physical considerations it can be argued that total loop length* is the most direct measure of potential exposure, particularly in the case of measurements from the central office to an on-hook station. On the other hand, if measurements are made at a main station location, the choice between working length and total length is less obvious. However, examination of survey data indicated that main station noise distributions that used either working or total loop length as a parameter provided required information. Hence, we chose total loop length as the length parameter to be used in this paper. Figure 1 shows the distributions of loop lengths of assigned working pairs.†

Once the length parameter had been chosen, it was desirable to define long loops since they were most likely to have the highest induction.[2] For the 1980 survey, long loops were arbitrarily defined as those having total lengths greater than 20 kft; all remaining loops were considered short. Two main considerations suggested this breakpoint. First, based on the 1973 loop plant survey,[3] this definition

---

* Total length is the sum of working loop length plus all bridged tap length. Working length is the length of wire pair in which dc current flows when a station goes off hook; bridged tap length is the additional length of wire pair bridged onto the working length.[3]

† Throughout this paper, distributions are characterized with respect to the population of working assigned pairs, unless stated otherwise. In 1980 there were approximately 89,900,000 working assigned pairs in the Bell System, which excludes Southern New England Telephone and Cincinnati Bell.
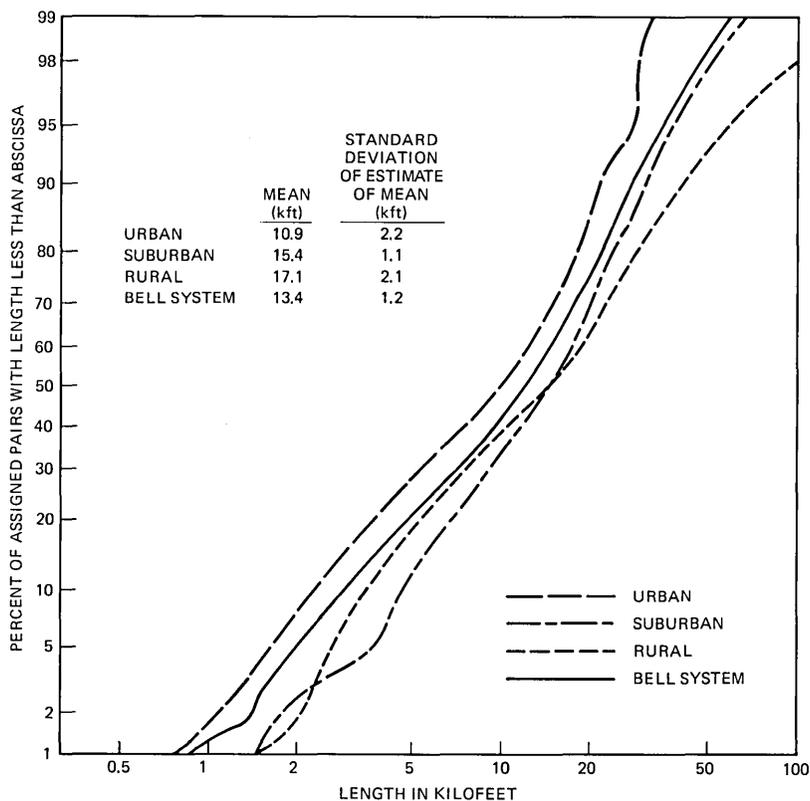
Fig. 1—Total loop length for urban, suburban, rural, and Bell System loop populations.

assured the selection of a reasonable number of long loops without unduly increasing the complexity or size of the sample. Second, the authors believed that a number of applications of the survey results would require separating the loop population at approximately 20 kft, since that is approximately the dividing line between loaded and nonloaded loops. For example, loop electronics that operate on nonloaded loops would serve the short loop population. Also, the increased use of subscriber carrier systems should continue to reduce the proportion of physical loops exceeding 20 kft in total length.

Since induced signal levels are a function of both exposure length and details of the power system configuration, the measured levels for a given length have a large variance.[7] In addition, for wire pairs significantly under 10 kft in total length, currents in local grounds will contribute a component to the noise[14] that is not length dependent. As a result, induced noise on wire pairs in areas with heavy power usage, such as industrial areas near power plants, may not show a

strong length dependence until longer exposure lengths are reached. For very long exposures there is a decrease in the rate of increase of induced noise. Wire pairs can be modeled as lossy transmission lines.* As exposure lengths increase much beyond 30 kft, the mean level of induced signals increases much more with length than for exposures of intermediate length. This is a direct result of the attenuation of signals induced remotely from the observation point.

III. OVERVIEW OF SURVEY

3.1 Survey design

A major objective of the 1980 noise survey was to characterize statistically the induced voltages and currents present at central office and subscriber loop interfaces. An additional goal was to identify the small percentage of Bell System loops experiencing high induction levels, without using a large sample size. To accomplish this, we stratified the Bell System into urban, suburban, and rural wire centers. Urban wire centers were defined as those serving more than 1000 assigned pairs per square mile (ap/sq.mi.), suburban centers, between 45 and 1000 ap/sq.mi., and rural centers, less than 45 ap/sq.mi. This stratification assured representation of noisy loops, which were assumed to be the long loops (over 20 kft in total length) in the suburban and rural environments.

The survey sample was selected using a stratified, two-stage sampling scheme. The 36 sampled wire centers were geographically dispersed as shown in the appendix: 6 were urban wire centers, 18 suburban, and 12 rural. The loops to be tested were randomly selected from physical loops in a wire center. To limit the scope of the loop survey, we excluded loops providing special services or served by carrier systems. We tested 30 to 40 assigned pairs in each wire center. Table I summarizes the number of tested loops in the survey and compares the proportions of loops in the survey and Bell System populations. A relatively high proportion of rural loops was used to identify high induction loops. The appendix gives details on the survey design.

Once the measurements were completed, cumulative distributions and means for the various noise parameters were obtained for urban, suburban, and rural strata and for the Bell System. In addition, distributions were calculated for short loops in each strata and long loops in the nonurban strata. The statistics were calculated using a general ratio technique as described in the appendix.

---

* These models can be developed from the results in a paper by Parker,[9] Appendix A of a paper by Wilson,[10] or Appendix E of a book by Smith.[18]

Table I—Composition of tested loop and Bell System loop populations

	Urban		Suburban		Rural		Subtotal	
	Number of Tested Loops	Percent of Total Surveyed Loops	Number of Tested Loops	Percent of Total Surveyed Loops	Number of Tested Loops	Percent of Total Surveyed Loops	Number of Tested Loops	Percent of Total Surveyed Loops
(a) Loops in Survey								
Long loops	20	1.6	194	15.5	155	12.3	369	29.4
Short loops	161	12.8	450	35.8	276	22.0	887	70.6
Total	181	14.4	644	51.3	431	34.3	1256	100
(b) Percent of Bell System Loops in 1980								
Long loops	6		12		3.5		21.5	
Short loops	43		29		6.5		78.5	
Total	49		41		10		100	

3.2 Noise parameters and their measurements

The measured and derived noise quantities, which are illustrated for a particular configuration in Fig. 2, are defined as follows:

1. *Noise to ground*, N_g, is the average of the tip and ring conductor voltages to ground. Conventionally, it is measured using a 3-kHz flat or a C-message weighted filter. The resulting rms values are expressed in units of dBrn and dBrnC, respectively. Figure 1 shows the conversion from rms voltage to dBrn; the same formula holds for the conversion from C-message weighted rms voltage to dBrnC.

2. *Metallic noise*, N_m, is the voltage across a nominal 600-ohm resistor connected between the tip and ring conductors of a pair. The magnitude of the metallic noise is expressed in units of dBrn or dBrnC. C-message weighted metallic noise is a direct electrical measure of the noise perceived by the customer.[19]

3. *Longitudinal current*, I_L, is the sum of the tip and ring short-circuit currents to ground. During the survey this current was determined from the voltage across a 10-ohm resistor connecting tip and ring to ground.

4. *Longitudinal impedance* is calculated as the ratio of longitudinal voltage to longitudinal current. Only the magnitude of the impedance is available from the survey data.

5. *Balance, B,* is 20 times the logarithm of the ratio of the C-message
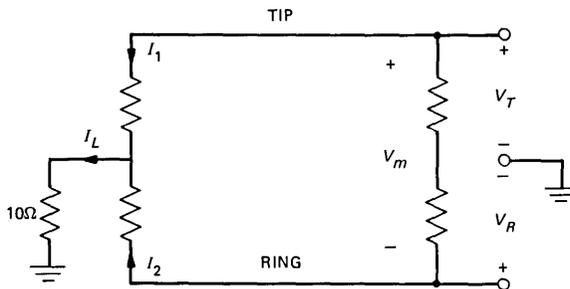
LOOP CONDUCTORS

N_m — METALLIC NOISE
= 20 log $(|V_m/24.5\ \mu V|)$ dBrn

N_g — NOISE TO GROUND
= 20 log $(|V_g/24.5\ \mu V|)$ dBrn,

WHERE:

V_g = $(V_T + V_R)/2$ V(rms)

I_L — LONGITUDINAL CURRENT
= $(I_1 + I_2)$ A(rms)

B — BALANCE
= 20 log $(|V_g/V_m|)$ dB
= $N_g - N_m$

Fig. 2—Measured noise parameters.

weighted noise-to-ground voltage to the C-message weighted metallic noise voltage, or, equivalently, N_g-N_m (dBrnC).

A one-time set of measurements of the above parameters was made at main station and central office interfaces on 1256 loops at random times between 8:30 AM and 4:30 PM local time. These business hour measurements were made essentially simultaneously at both ends of the loop. The one-time business hour measurements at the main station provided the data most directly comparable to that of the 1964 survey.

To characterize the diurnal variation and spectral content of the noise parameters, automated measurements were made at central office and main station interfaces for at least 24 hours. Automated measurements at the central office were made for all noise parameters for about 950 of the tested loops. At the main station, automated noise-to-ground measurements were made for 195 loops.

Corresponding to one-time and automated measurements, two types of cumulative distributions are used to describe the electrical parameters. Consistent with previous surveys, business hour distributions are calculated from data obtained by the one-time measurements made on each loop. Daily maximum and daily median distributions are calculated from the maximum and median levels reached on each loop in a 24-hour period as determined by the automated measurements. In the following, business hour distributions are discussed unless stated otherwise.

3.3 Principal findings

Tables II and III summarize induction levels at main station and central office interfaces in terms of 3-kHz flat and C-message weighted noise to ground, longitudinal current, and metallic noise. (Distributions for the more important parameters are given in Sections IV and V. However, a normal distribution can be assumed for the other parameters as long as it is understood that the actual tails of the distributions are not modeled well. The distributions presented in Sections IV and V provided the basis for understanding the limitations of such models.) The estimates in Tables II and III generally have 90-percent confidence intervals of ±2 dB. The appendix discusses the derivation of these confidence intervals. Maximum levels in Tables II and III represent measured data obtained by considering both the 1980 survey and its pilot, conducted in 1979.

In addition to these statistics, other findings are as follows:

1. Noise to ground, metallic noise, and longitudinal current at the central office have been characterized for the first time (see Table II). The 90th percentile of the longitudinal current is −12 dBmA (0.25 mA).

Table II—Central office noise to ground,
longitudinal current, and metallic noise as
measured to on-hook stations

	Mean	Median	90th Percentile	Maximum*
(a) Central Office Noise to Ground (dBrn)				
Urban	65	64	79	100
Suburban	73	72	88	114
Rural	79	81	92	110
Bell System	**70**	**68**	**87**	**114**
Bell System (DM)†		**73**	**91**	
(b) Central Office C-Message Weighted Noise to Ground (dBrnC)				
Urban		<40	47	77
Suburban		46	66	85
Rural		54	69	83
Bell System		**41**	**61**	**85**
Bell System (DM)†		**45**	**62**	
(c) Central Office Longitudinal Current (dBmA)				
Urban		−40	−26	−2
Suburban		−30	−8	25
Rural		−23	0	21
Bell System		**−34**	**−12**	**25**
Bell System (DM)†		**−32**	**−7**	
(d) Central Office C-Message Weighted Metallic Noise (dBrnC)				
Urban			10	38
Suburban			17	46
Rural			19	53
Bell System			**15**	**53**
Bell System (DM)†			**21**	

* Largest values measured at any time as part of the survey.
† Results are from distributions of daily maximums on measured loops. All other statistics except maximums are for one-time measurements made during business hours.

2. The median of the *main station noise to ground* as measured during business hours was 61 dBrnC. Although this represents a measured increase of 2 dB since 1964, the increase cannot be considered statistically significant. The median of the daily maximum noise to ground, which had not been previously characterized, was 67 dBrnC. This difference in medians indicates a substantial diurnal variation in noise magnitudes. The 540-Hz component (ninth harmonic of 60 Hz) typically dominates C-message weighted noise to ground.

3. The 90th percentile of *main station metallic noise* on the loop plant, as measured during business hours, is 11 dBrnC, and 4 percent of the loops exceed the traditional reference levels of 20 dBrnC. When contrasted to the comparable 1964 distribution calculated from Gresh,[2] there is an increase of 1 dB in the 90th percentile value. However, this increase cannot be considered statistically significant.

Table III—Main station noise to ground, longitudinal current, and metallic noise as measured to main distributing frames

	Mean	Median	90th Percen- tile	Maximum*
(a) Main Station Noise to Ground (dBrn)				
Urban	86	87	103	111
Suburban	96	95	108	134
Rural	96	95	108	117
Bell System	**91**	**92**	**106**	**134**
Bell System (DM)[†]		**96**	**110**	
(b) Main Station C-Message Weighted Noise to Ground (dBrnC)				
Urban	—	57	70	90
Suburban	68	67	83	94
Rural	70	69	85	101
Bell System	**—**	**61**	**80**	**101**
Bell System (DM)[†]		**67**	**82**	
(c) Main Station Longitudinal Current (dBmA)				
Urban	9	9	20	30
Suburban	16	16	28	47
Rural	**15**	**17**	**28**	**47**
Bell System	**13**	**13**	**25**	**47**
(d) Main Station C-Message Weighted Metallic Noise (dBrnC)				
Urban			5	31
Suburban			12	37
Rural			**16**	**51**
Bell System			**11**	**51**

* Largest values measured at any time as part of the survey.
† Results are from distributions of daily maximums on measured loops. All other statistics except maximums are for one-time measurements made during business hours.

4. The 90th percentile of main station metallic noise measured to a central office quiet termination during business hours is 13 dBrnC. Comparable results for 1964[2] indicate a 90th percentile of 19 dBrnC. For reasonable assumptions on the variance of the 1964 estimates, it is concluded the 90th percentile level has decreased approximately 6 dB. This measurement of main station metallic noise includes both loop plant and central office contributions. Since the loop component has remained approximately constant, the improvement is attributed to a reduction in the central office contribution to the metallic noise.

5. *Central office metallic noise* measurements are usually of interest as an indication of metallic noise at the main station. Simultaneous measurements at the two interfaces indicate that central office metallic noise measured to an on-hook main station is generally higher than main station metallic noise measured to the quiet termination of the central office. Based on the measured central office diurnal variations,

daily maximum distributions of main station metallic noise are between 3- and 6-dB higher than the business hour distributions for corresponding percentiles.

6. Power lines induce significantly different levels of noise to ground and longitudinal currents in urban and nonurban loops. This difference exists because nonurban environments have longer loops and typically higher net power currents than urban environments. Urban short loop distributions of longitudinal noise parameters are typically 6- to 10-dB lower than the nonurban short loop distributions. The importance of loop length is illustrated by the fact that for a nonurban stratum the median noise to ground for short loops (with total lengths less than or equal to 20 kft) is about 15-dB lower than for long loops (with lengths longer than 20 kft).

IV. CENTRAL OFFICE RESULTS

The need for information on noise parameters at the interface between the central office and the loop has grown as new generations of solid state loop and trunk terminating equipment have appeared.[4,5] The 1980 Bell System noise survey provides the first relatively complete characterization of induction at this interface. Since central office loop testing systems, such as the Mechanized Loop Testing system (MLT),[20] usually make noise measurements to on-hook main stations, data obtained under similar conditions (see Fig. 3) are of primary concern. However, since central office metallic noise depends on the metallic termination at the main station, it can be argued that central office metallic noise with an off-hook station is a more relevant measure of transmission quality. For this reason, metallic noise data obtained with an ungrounded 632Ω resistor replacing the on-hook station set (see Fig. 3a) are also considered.

4.1 Central office noise to ground

The central office measurements confirm the postulates underlying the induction model used during the survey design (see Section II). Those postulates are: urban environments generally have shorter loops than nonurban; induction increases with loop length; urban loops, as a result of well-balanced power systems, have lower induction-per-unit length than nonurban loops; and nonurban long loops have the highest induction levels. Business hour distributions of central office C-message noise to ground given in Fig. 4 support the conclusion that urban loops have less induction that nonurban loops. The distributions of total loop lengths given in Fig. 1* confirm that urban environments have a higher proportion of short loops than nonurban environments.

---

* In every plot of cumulative distributions, a straight line would imply a normal distribution.

CENTRAL OFFICE

LOOP CONDUCTORS

316Ω

V_m

316Ω

ON-HOOK*
STATION

(a)

V_T 316Ω

V_R 316Ω

(b)

I_L

10Ω

(c)
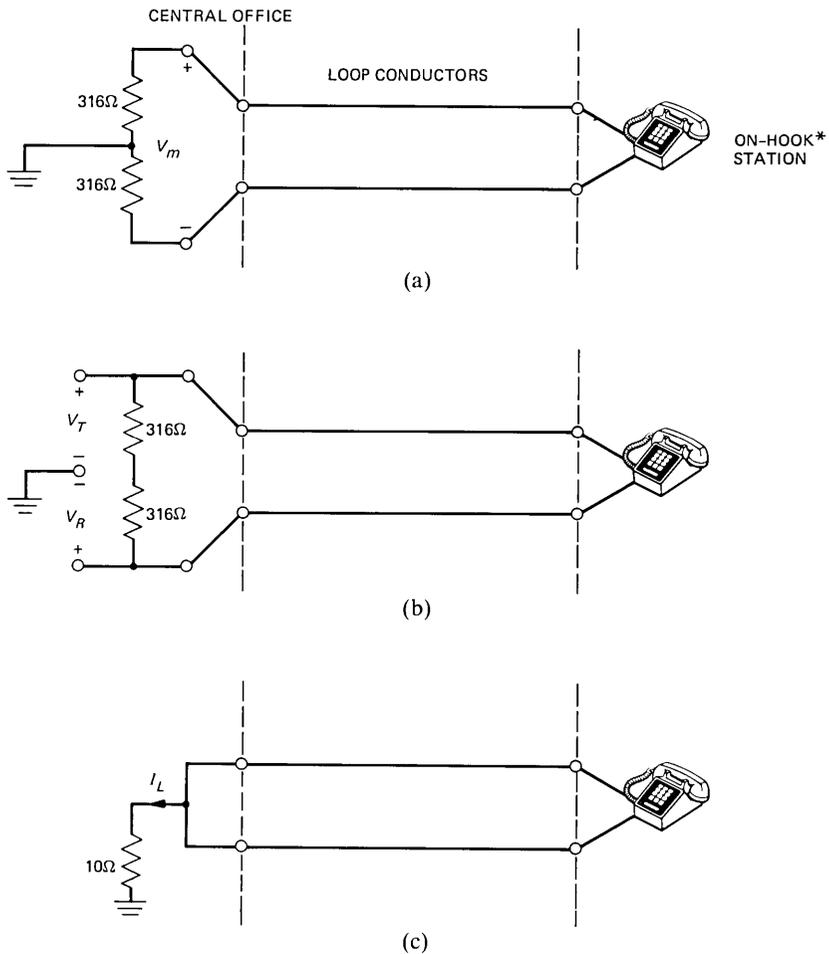
*IN ON-HOOK CONDITION, ONLY THE
RINGER BRIDGES THE LINE.

Fig. 3—Central office measurement configurations used in survey.

As a basis for further consideration of the postulates, Fig. 5 presents the noise-to-ground distributions for nonurban long and short loops, where long loops are greater than 20 kft in total length and short loops are less than or equal to 20 kft in total length. The urban long loop distribution, which is not shown, had a median of 42 dBrnC and a 95th percentile of 63 dBrnC. Because the variance associated with the estimator of this distribution is large, as a result of the limited number of urban long loops in the sample (Table Ia), the distribution is not plotted. (Because of the small sample size, urban long-loop distribu-

The chart contains the following data table:

	MEDIAN (dBrnC)	STANDARD DEVIATION (dB)
URBAN	–	–
SUBURBAN	46	14
RURAL	54	12
BELL SYSTEM	41	15

Legend:
- – – – URBAN
- – · – SUBURBAN
- – – – RURAL
- ——— BELL SYSTEM

Axis labels: PERCENT OF ASSIGNED PAIRS WITH NOISE LESS THAN ABSCISSA (vertical), NOISE TO GROUND IN dBrnC (horizontal)
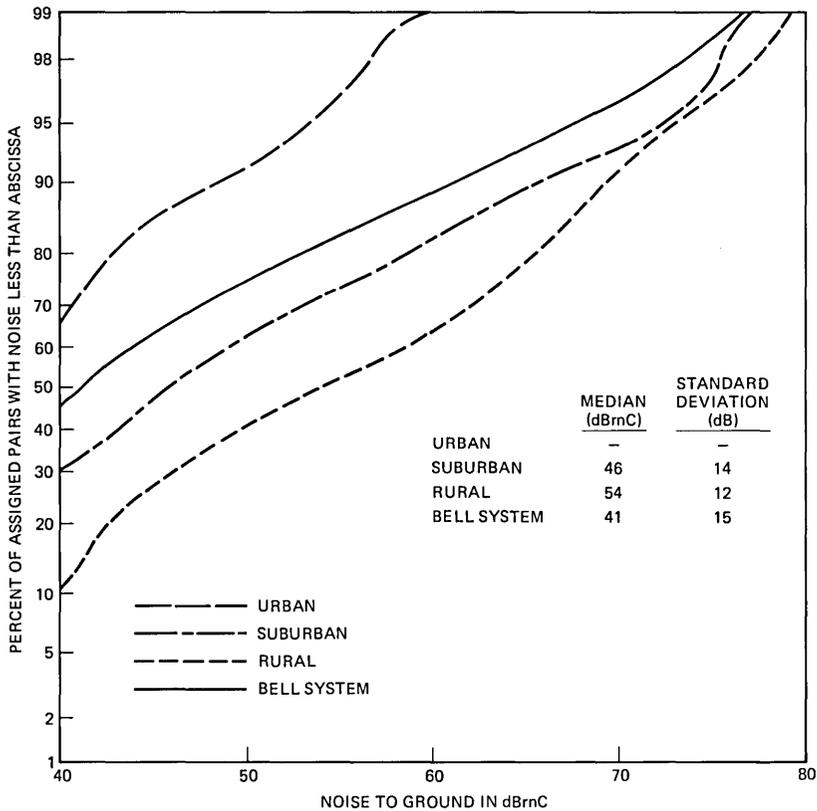
Fig. 4—Central office C-message weighted noise to ground for urban, suburban, rural, and Bell System loop populations.

tions will not be plotted at all in this paper.) However, a reasonable*
normal fit to the distribution confirms that urban long loops have
lower induction levels than nonurban long loops. A comparison of the
nonurban long and short loops on Fig. 5 shows that induction generally
increases with loop length. Based on the short loop distributions, the
urban loops are seen to have lower induction levels than the nonurban
environments. This is consistent with the assumption that urban
power systems are typically well balanced.

Figure 6 illustrates the effects of diurnal variation by presenting the
distributions of the 24-hour median and maximum levels of C-message
and 3-kHz flat weighted noise to ground. The daily median and

---

* A reasonable fit implies that based on a Kolmogorov-Smirnov test, the fit could
not be rejected at the 5-percent significant level.[21]
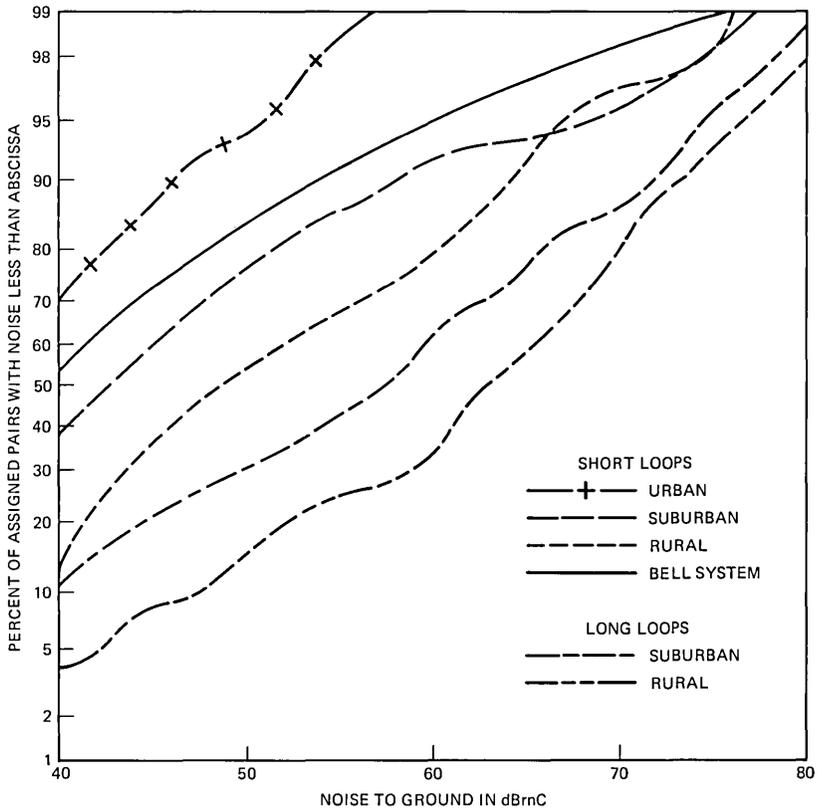
Fig. 5—Central office C-message weighted noise to ground for long (>20 kft) and short (≤20 kft) loops.

maximum distributions are derived from the median and largest reading, respectively, of the 24 hourly measurements made on each loop. The 3-kHz flat and C-message weighted distributions of maximums are displaced positively from the distributions of the daily medians by approximately 6 and 4 dB, respectively. The Bell System distribution of the 24-hour median levels is nearly identical to the business hour distribution of noise to ground calculated independently from one-time measurements, as can be seen by a comparison of C-message weighted distributions on Figs. 4 and 6.

Figure 7 presents the distributions of the noise-to-ground harmonic levels at the odd multiples of 60 Hz. (Even harmonics of significant amplitudes are sufficiently rare that their presence constitutes a powerful noise diagnostic tool.) The Bell System distribution of each component has been characterized by its median and quartiles. For each component the distribution was derived from the daily median
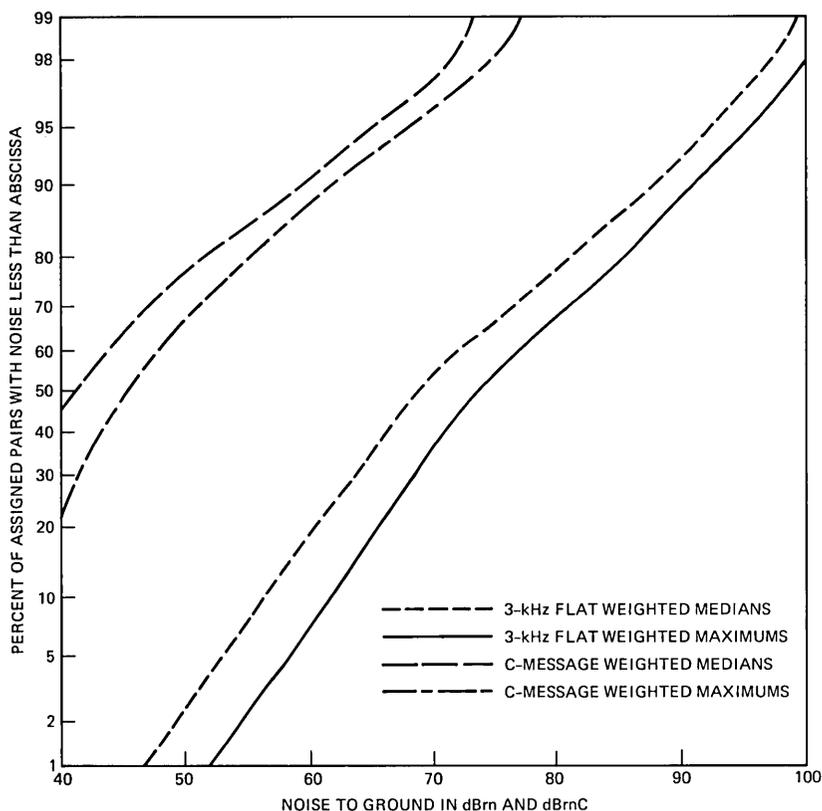
Fig. 6—Central office median and maximum noise to ground in a 24-hour period.

levels occurring on each loop. The spectral components decrease monotonically as frequency increases, except for the 9th and possibly the 15th harmonics, which are triple odd harmonic components, as defined in Section II. The ninth harmonic (540 Hz) generally proves to be the primary interfering frequency when C-message weighting is applied.

4.2 Central office longitudinal current

Because the majority of central office terminations have a low longitudinal impedance under operating conditions, longitudinal current provides a more direct characterization of longitudinal induction at the central office interface than does longitudinal voltage. Despite the importance of longitudinal current, only estimates based on the 1964 survey were available previously. The 1980 current data are summarized in Table II, and distributions are given in Fig. 8.

Longitudinal current and noise to ground are not independent
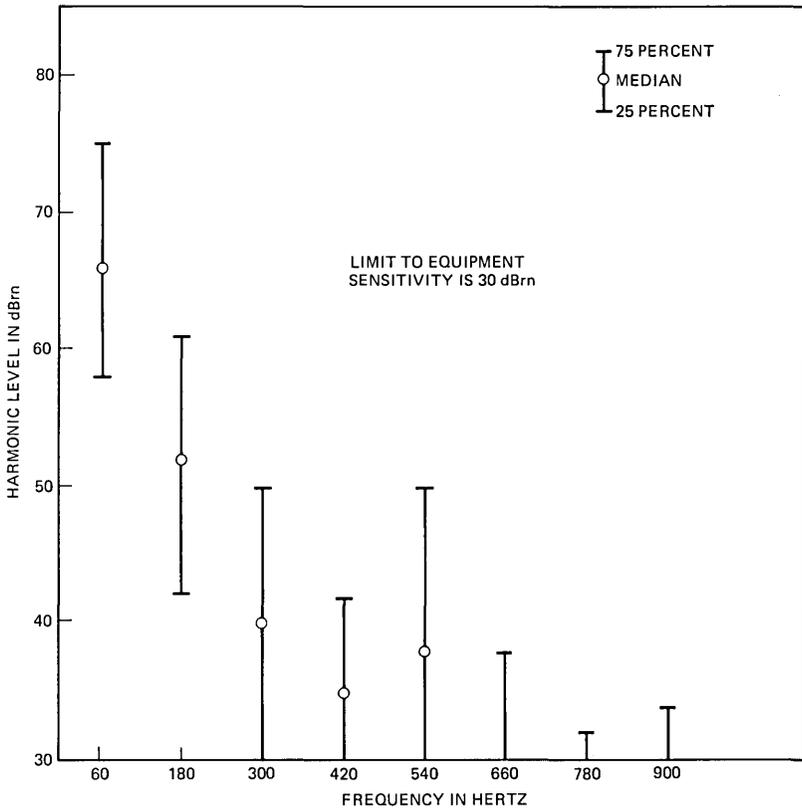
Fig. 7—Distributions of spectral levels of central office noise to ground.

variables, since they are related by the longitudinal input impedance of the loop. The magnitude of the longitudinal impedance is defined as the ratio of the magnitudes of central office noise to ground to longitudinal current. A scatter plot of longitudinal impedance at 540 Hz versus total loop length is presented in Fig. 9 for a subset of loops included in the survey, and is indicative of the behavior of the total loop population. Figure 9 shows that the impedance decreases essentially linearly with increasing loop length to beyond 30 kft, where transmission line loss effects become important. (A transmission line model with reasonable assumptions on loop resistance and capacitance to ground can be used to derive comparable results.)[18]

The longitudinal impedance at the central office is determined by the capacitance to ground of the cable pair, and the slope in Fig. 9 can be used to estimate the cable capacitance-per-unit length. The capacitance corresponding to the slope of Fig. 9 is 0.037 μF/kft (0.2 μF/mi).

Since longitudinal impedance decreases with loop length whereas
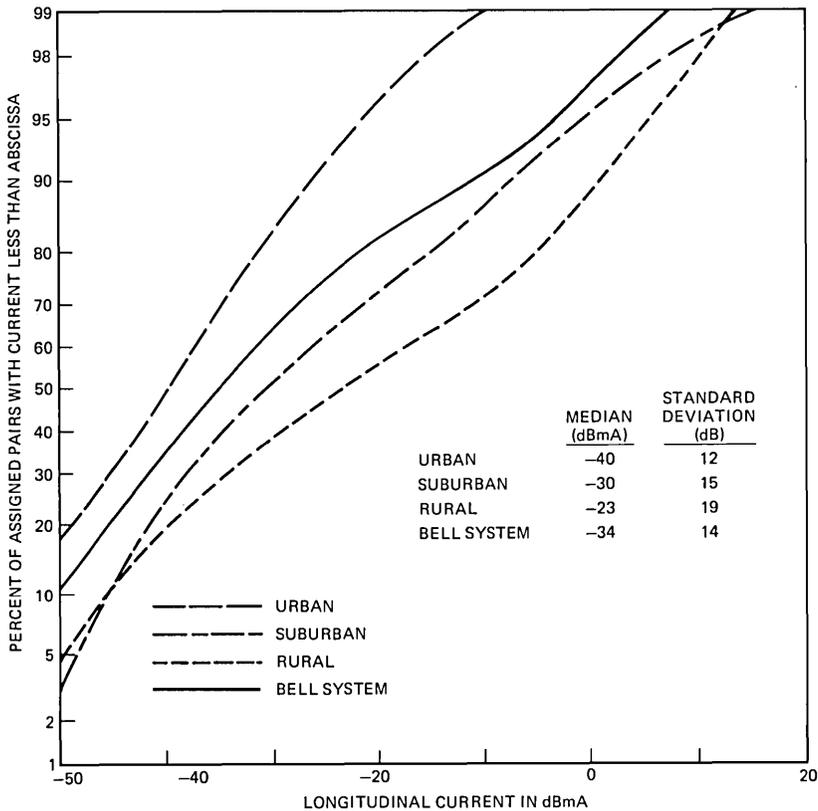
Fig. 8—Central office longitudinal current for urban, suburban, rural, and Bell System loop populations.

noise to ground generally increases with loop length (Section 3.1), longitudinal current should exhibit a stronger dependence on loop length than does noise to ground. To illustrate this point, Fig. 10 presents current distributions for nonurban long and short loops. The separation between current distributions for long and short loops in either nonurban environment is about 10 dB greater than the separation of corresponding noise-to-ground distributions (see Fig. 5).

The capacitive nature of the longitudinal impedance for a large percentage of loops, i.e., 95 percent are under 30 kft, causes the spectrum of the longitudinal current to differ from the spectrum of its driving source, the noise to ground. As a direct consequence, the separation between 3-kHz flat and C-message noise-to-ground measurements on each loop differs from the separation of the corresponding longitudinal current measurements. Figure 11 shows that a separation of 10 to 15 dB exists between 3-kHz flat and C-message weighted
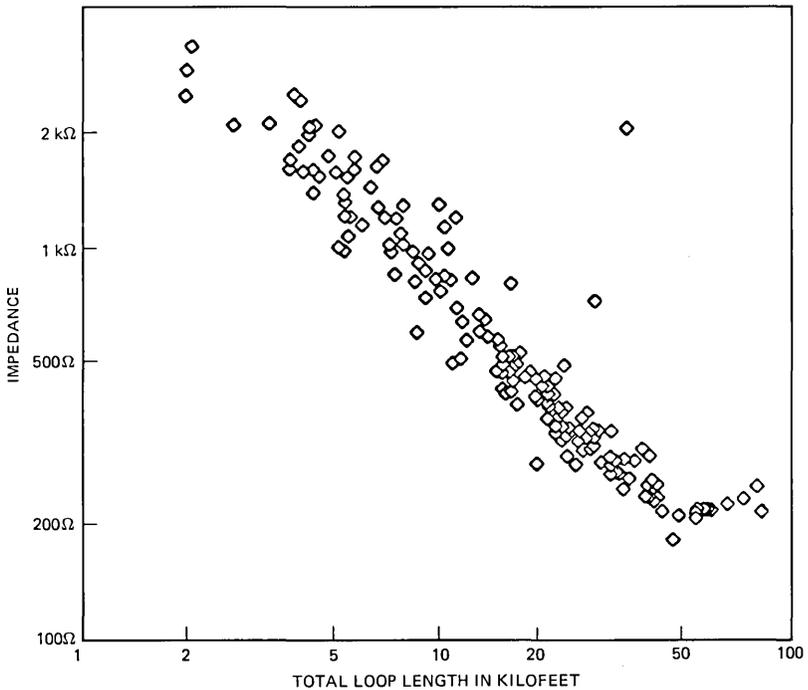
Fig. 9—Central office longitudinal impedance at 540 Hz versus total loop length.

distributions of longitudinal current. This contrasts sharply with the approximately 30-dB difference between 3-kHz flat and C-message weighted distributions of noise to ground (see Fig. 6).

4.3 Central office metallic noise

Metallic noise is produced by induced longitudinal voltages and currents acting on imbalances that are located at loop terminations and distributed throughout the loop. In addition, coupling from other loops in the cable can contribute to measured metallic noise. Figure 12 provides the central office distributions of metallic noise measured with on-hook station sets (see Fig. 3a). Figure 13 provides the central office distributions of metallic noise measured with a 632Ω termination replacing the on-hook station and simulating an off-hook station set. Included on the latter plot is the Bell System distribution of metallic noise for the on-hook condition. In general, it is found that the on-hook distributions for the various strata show significantly higher levels than the corresponding off-hook distributions.

Since induction increases with loop length and the imbalances are generally cumulative, metallic noise levels are generally greater on longer loops. Figure 14 presents the on-hook distributions of C-
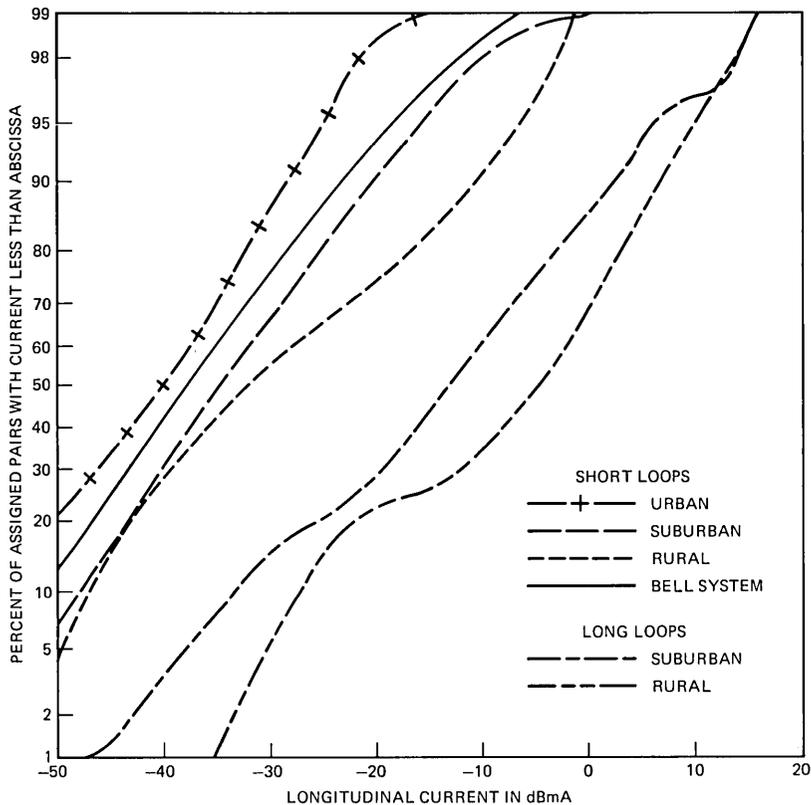
Fig. 10—Central office longitudinal current for long (>20 kft) and short (≤20 kft) loops.

message weighted metallic noise for nonurban long loops and for short loops. The distributions for the long and the short loops are widely separated at low noise levels, but with the exception of long rural loops, approach one another in the region of the upper tails. The long rural loops with greater than 30-dBrnC metallic noise represent about 0.15 percent of the total loops in the Bell System.

Bell System metallic noise distributions of daily median and maximum levels for on-hook conditions at the main station show the same general relationship observed for the noise to ground. Harmonic levels of central office metallic noise show a behavior similar to the noise-to-ground harmonic curve (e.g., a local peak occurs at 540 Hz).

V. MAIN STATION RESULTS

The 1980 noise survey measurements made at main station interfaces update and expand data on noise voltages, and provide new data
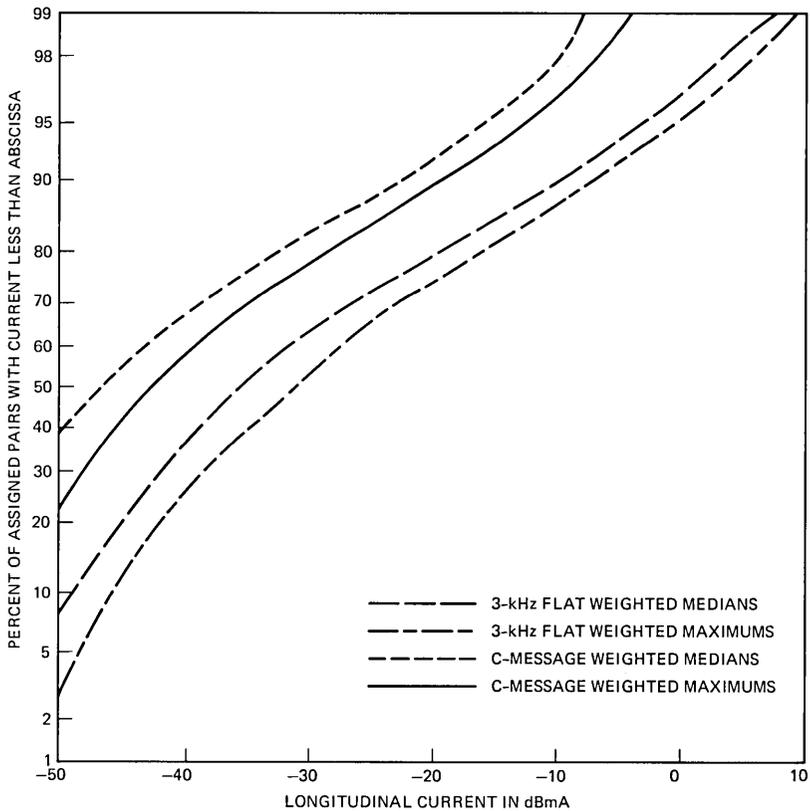
Fig. 11—Central office median and maximum longitudinal current in a 24-hour period.

on longitudinal currents. Since under normal operating conditions the loop has a balanced, grounded termination at the central office, main station longitudinal and metallic voltage measurements were made to balanced terminations at the central office. Two balanced office terminations were used. To determine the loop plant contribution to metallic noise, a well-balanced, grounded resistive termination was installed at the main distributing frame with the central office disconnected from the loop (see Fig. 15a). The second termination was established by calling into the quiet termination. The quiet termination is a 900Ω balanced test termination located at the central office. Metallic noise measurements made to the latter termination are influenced by noise present on intraoffice cables and any imbalance associated with these cables or the office quiet termination. To determine the maximum longitudinal currents that may exist at the main station,

NOISE SURVEY   795
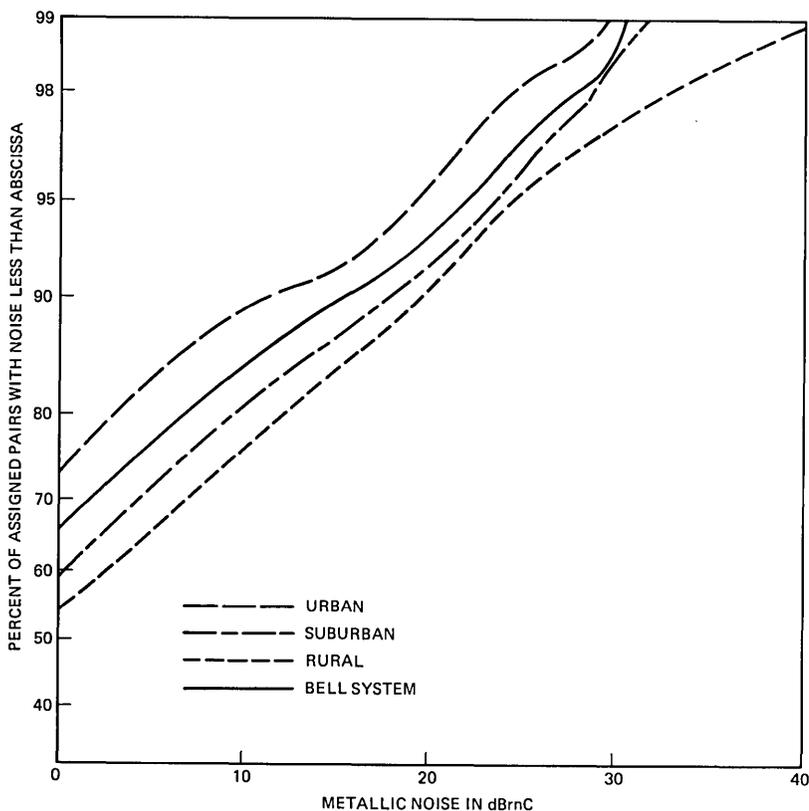
Fig. 12—Central office C-message weighted metallic noise to on-hook main station for urban, suburban, rural, and Bell System loop populations.

the short-circuit longitudinal currents were measured with tip and ring grounded through 10 ohms at the central office (see Fig. 15b).

Table II summarizes noise to ground, longitudinal current, and metallic noise data as measured at main station interfaces with a balanced, grounded termination at the main distributing frame. (Measurements to the quiet termination are considered in Sections 5.1 and 5.3.) The Bell System noise to ground medians of 92 dBrn and 61 dBrnC given in Tables IIa and b are found to be 2 dB higher than computed 1964 median levels. The 90th percentile of the metallic noise in Table IId is 1 dB higher than the 90th-percentile level of 10 dBrnC computed from 1964 distributions.[2] Thus a consistent pattern of higher 1980 levels is indicated; however, the confidence intervals on these estimates do not permit a statistically meaningful determination of the increase. Computations were necessary in order to compare the 1980 and 1964 data because Gresh presented the 1964 distributions
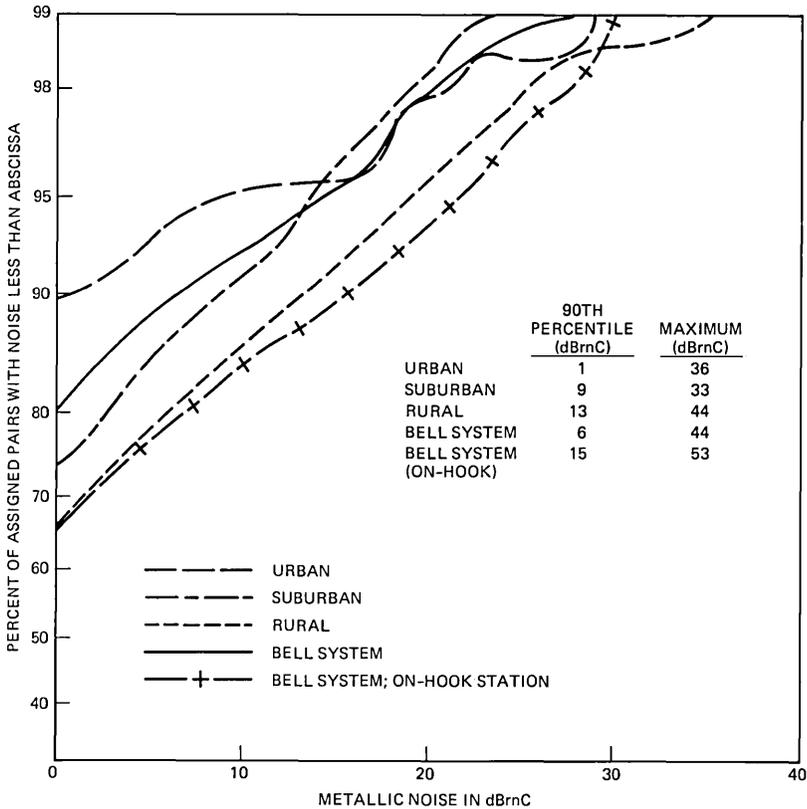
Fig. 13—Central office C-message weighted metallic noise to off-hook main station for urban, suburban, rural, and Bell System loop populations.

with respect to the population of main stations, whereas this paper presents distributions with respect to the population of assigned pairs. Transformation of 1964 distributions to an assigned pairs basis required using the distributions for the individual and party lines as published by Gresh and calculating weighted means of the percentiles for each noise level. The weights are the proportion of assigned pairs serving individual and party lines in 1964, 88 percent and 12 percent, respectively. The transformed distributions differed only slightly from those obtained directly from Gresh's distributions for individual lines.

5.1 Main station to ground

Figure 16 presents distributions for C-message weighted, main station, noise to ground as taken in 1980 during business hours. Four statistics are shown for each curve: the median, the mean, the standard deviation of the estimate of the mean (sdem), and the standard
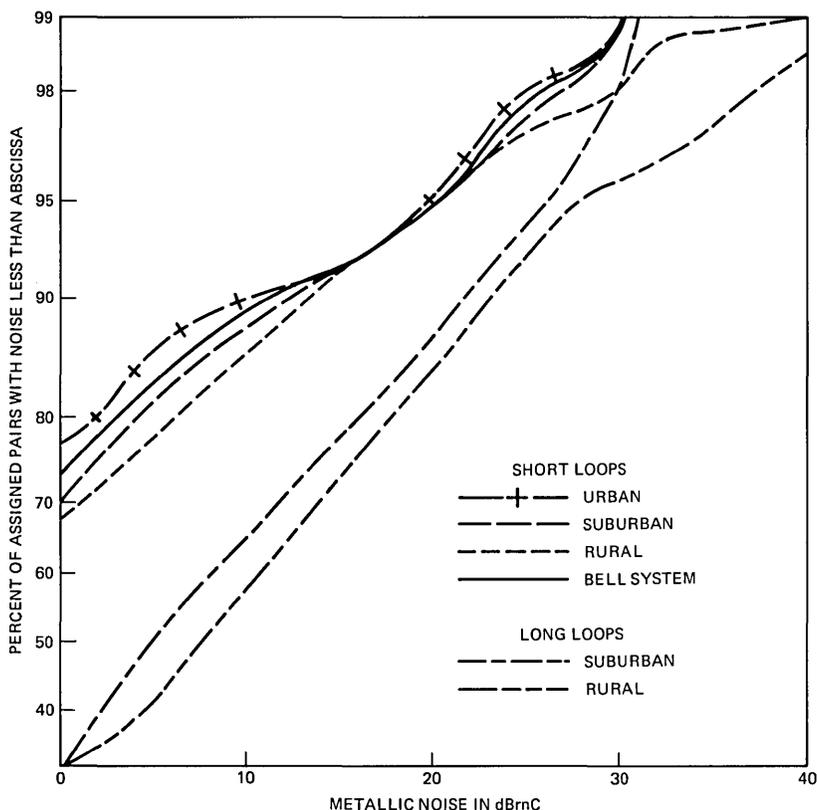
Fig. 14—Central office C-message weighted metallic noise to on-hook station for long (>20 kft) and short (≤20 kft) loops.

deviation of the distribution(s). The 90-percent confidence interval for the mean is 1.67 times the sdem. As an example of the significance of these distributions, Fig. 16 shows that the Bell System lines that exceed 90 dBrnC, a commonly recognized maintenance objective, are largely in nonurban areas. The percentage of Bell System loops exceeding 90 dBrnC is estimated to be 1.4 percent; the upper bound to the 95-percent confidence interval for this estimate is 4 percent.

Noise-to-ground measurements presented in this section were made to balanced resistive terminations at the main distributing frame (see Fig. 15a). The derived distributions are identical to those obtained using quiet terminations in the office. This implies that no significant longitudinal voltage source is present in intraoffice wiring.

Figure 17 presents the distribution of noise to ground for nonurban long and for short loops, where short loops are less than or equal to 20 kft in total length and long loops are greater than 20 kft in total

CENTRAL OFFICE | LOOP CONDUCTORS | MAIN STATION

(a)

(b)
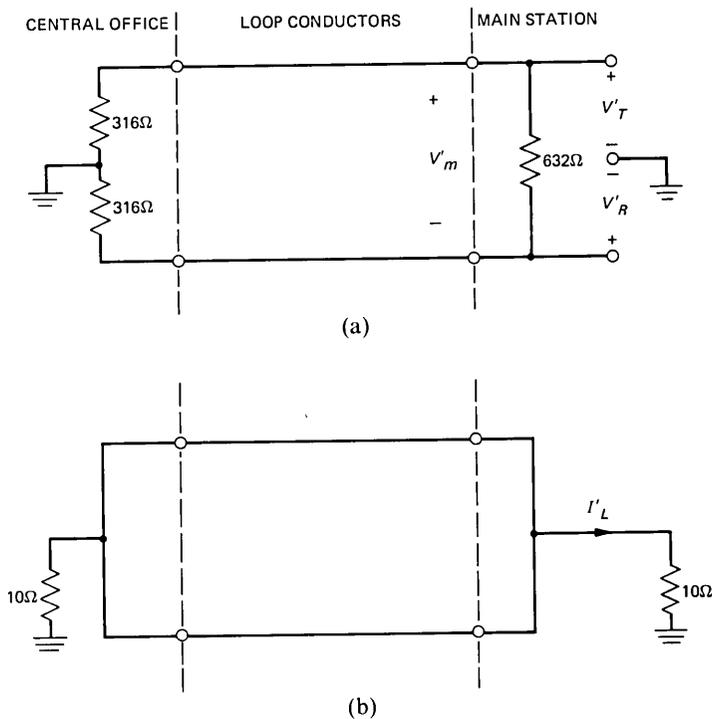
Fig. 15—Main station measurement configurations. (a) Metallic noise and noise to ground measurements. (b) Longitudinal current measurement.

length. The medians of long and short loops for a given nonurban population differ by 13 to 15 dB. As was the case for central office measurements (see Section 3.1), this dependence on loop length accounts for much of the difference between urban and nonurban distributions of noise to ground.

For typical subscriber loop terminations, which have a ground at the central office and no ground at the main station, the 3-kHz flat weighted noise to ground is the most direct measure of power exposure. Figure 18 presents the distributions of this important parameter. The 3-kHz flat weighted distributions are displaced by approximately 30 dB from the corresponding C-message weighted distributions of Fig. 16.

The changes in power load responsible for the central office diurnal noise variations also produce diurnal variations at the main station. The magnitude of the effect on 3-kHz flat and C-message weighted noise to ground is illustrated by Fig. 19, as well as by Table II. This information and the curves of Figs. 16 and 17 show that the business hour distributions are significantly lower than the daily maximum
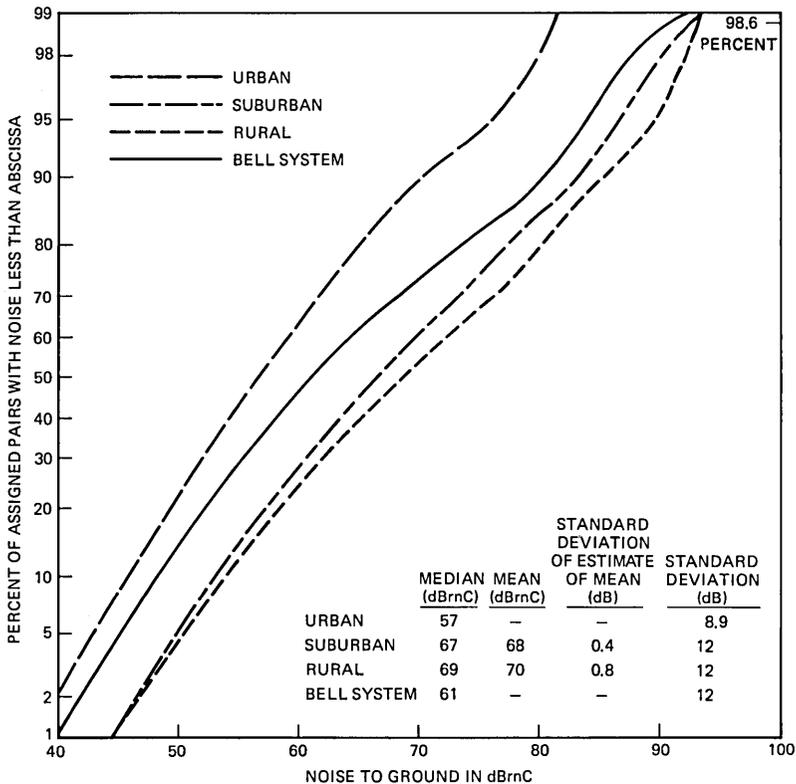
Fig. 16—Main station C-message weighted noise to ground for urban, suburban, rural, and Bell System loop populations.

distributions. At the 90th-percentile points, the difference is estimated to be 4 dB for the 3-kHz flat weighted data and 2 dB for the C-message weighted data. The differences remain essentially constant for percentiles greater than the 90th. These findings are consistent with the results of Heirman.[6]

Figure 20 describes the relative spectral content of main station noise to ground. For Fig. 20 all hourly measurements on all loops for which diurnal measurements were made were equally weighted. The harmonic levels for each waveform were divided by the 3-kHz flat weighted magnitude to obtain a normalized amplitude. An analysis of these normalized harmonic levels indicates that the spectral components at 60 and 180 Hz dominate the 3-kHz flat weighted voltage spectrum. When C-message weighting is applied to the measured data, the 540-Hz component is generally dominant. The remaining odd harmonics between 300 and 900 Hz (the 5th through the 15th har-
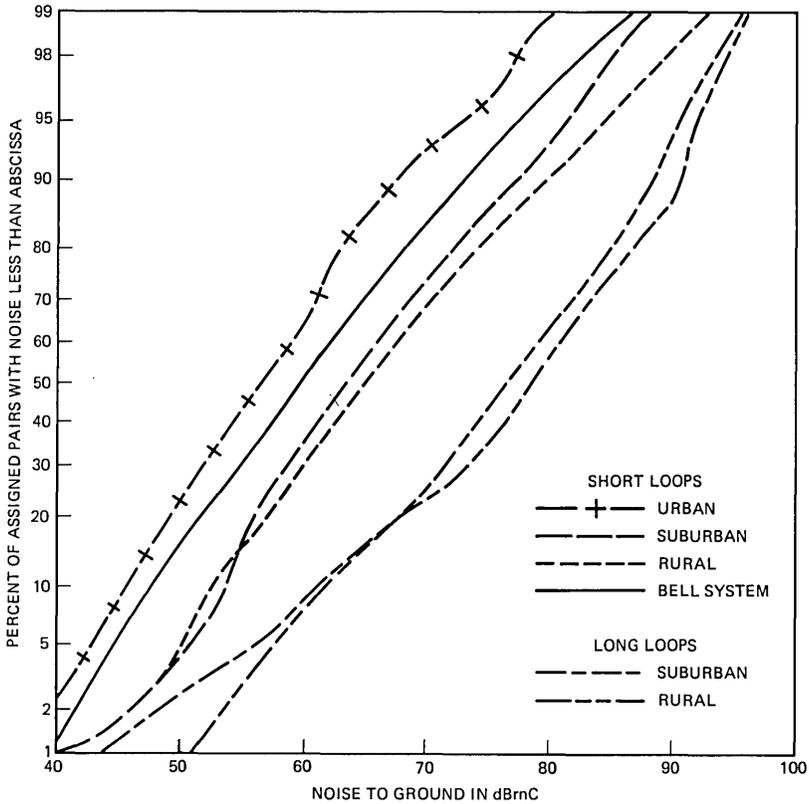
Fig. 17—Main station C-message weighted noise to ground for long (>20 kft) and short (≤20 kft) loops.

monic) are of lesser but approximately equal importance in determining C-message weighted noise levels.

5.2 Main station longitudinal current

Station equipment employing terminations with a low impedance to ground can be subjected to substantial induced longitudinal currents. The first measured characterization of these main station currents is summarized in Fig. 21. As a point of reference, the current on 1 percent of Bell System assigned pairs exceeded 80 mA (38 dBmA). Currents measured at the main station are significantly higher than at the central office because of the change in termination conditions (compare Figs. 3c and 15b).

5.3 Main station metallic noise

Figure 22 presents the main station C-message weighted metallic noise distributions obtained with a termination at the main distrib-
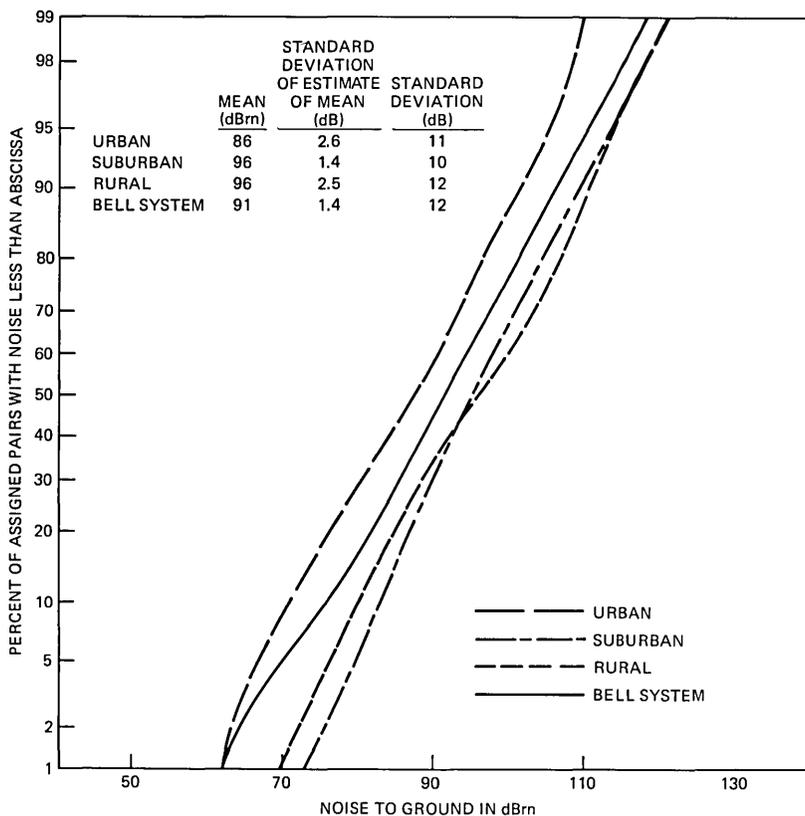
Fig. 18—Main station noise to ground for urban, suburban, rural, and Bell System loop populations.

uting frame. Since 20 and 30 dBrnC are standard levels of reference,[22] the corresponding percentiles are indicated on Fig. 22. Figure 23 presents distributions for the nonurban long loops and short loops. Although nonurban long loops generally have higher longitudinal induction levels than short loops, Fig. 23 shows that except for the long rural loop distributions, the difference between the various metallic noise distributions becomes smaller in the upper tail of the distributions. A reasonable conjecture explaining the observed convergence of distributions in the region of 30 dBrnC is that operating company engineers make special efforts to limit metallic noise on high induction loops.

Up to this point the discussion has centered on noise measured with a balanced resistive termination at the main distributing frame. However, Bell System recommended objectives for metallic noise refer to
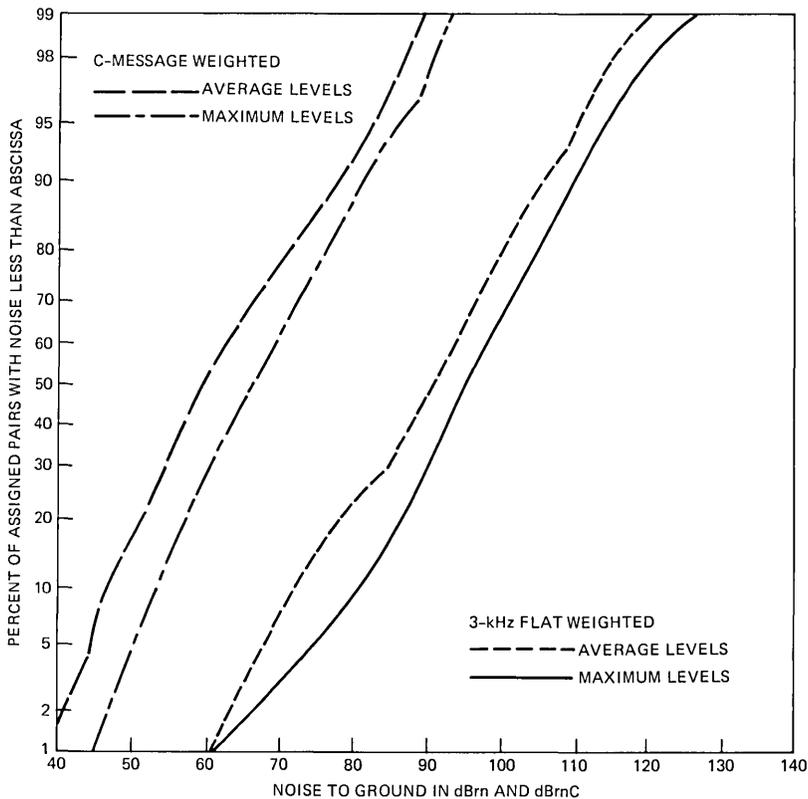
Fig. 19—Maximum and average main station noise to ground in a 24-hour period.

noise as measured to the central office quiet termination. Figure 24 presents main station distributions of metallic noise measured under this condition. Since the quiet termination is accessed by a call to the central office, the metallic noise is influenced by the imbalances and noise introduced by battery plant and intraoffice cables, and noise within the central office switch. The effect of these sources can be seen in Fig. 25, which presents the Bell System distributions for metallic noise measured with the main distributing frame and quiet terminations. For a given percentile, metallic noise measured to the quiet termination is generally greater than the metallic noise measured to the main distributing frame. The two distributions show a maximum separation of 6 to 7 dB but approach one another in the upper tail of the distributions. This apparent convergence, which was not evident in results obtained in 1964,[2] could be attributed to statistical uncertainties in the tails of the data.

In recent years the relationship between measurements of metallic
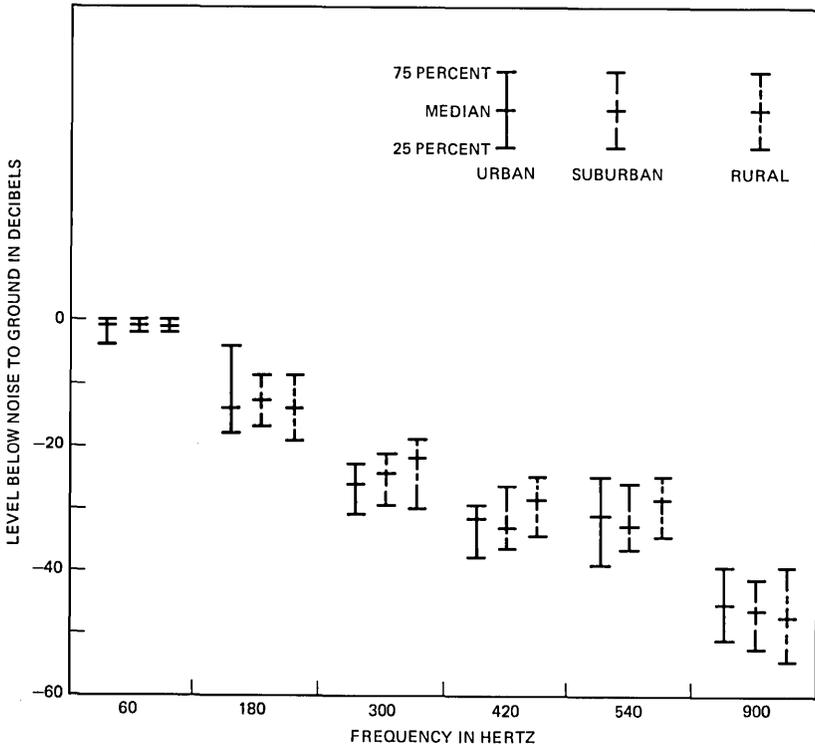
Fig. 20—Distributions of normalized spectral components of main station noise to ground.

noise at the central office and at the main station has been of considerable interest. Although much work remains to be done, Fig. 25 provides the basis for a comparison of four metallic noise distributions. Two curves present distributions of central office metallic noise measured with on-hook and simulated off-hook main station terminations (see Section 4.3). The second set of two curves present distributions of main station metallic noise obtained with quiet and main distributing frame terminations. Figure 25 shows that the distribution of central office metallic noise measured to an on-hook station is generally higher than the other distributions and that all the distributions approach one another in the upper tails. As can be seen, the two central office distributions bracket the main station measurements in the range of greatest interest. This suggests that the application of special main station terminations may permit estimates of the upper tail of the subscriber noise distributions from central office measurements.
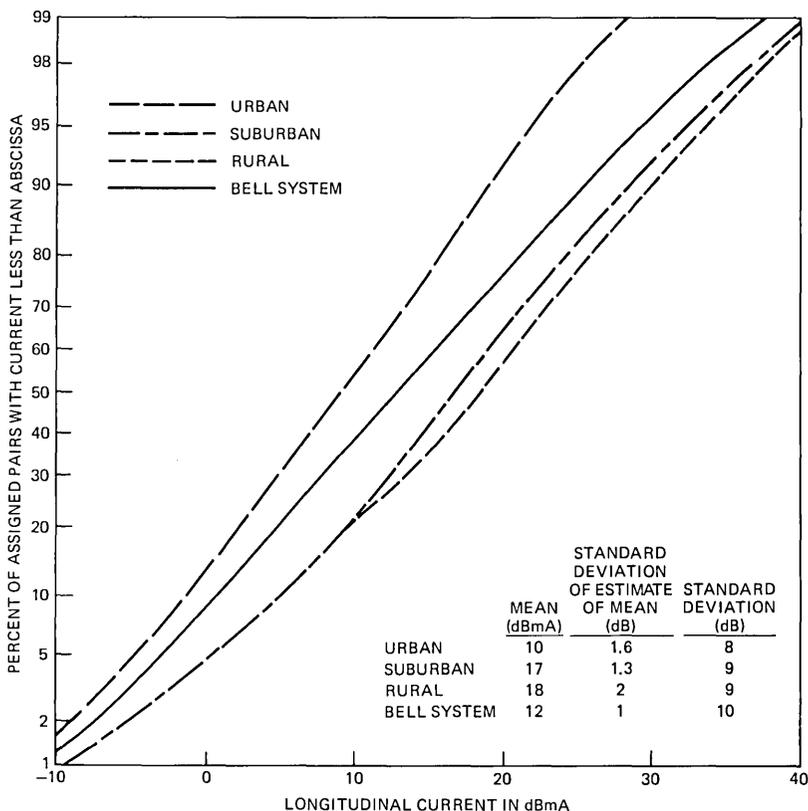
Fig. 21—Main station longitudinal current for urban, suburban, rural, and Bell System loop populations.

5.4 Balance

Balance is a measure of loop plant susceptibility to power system induction that complements the direct measures of power induction such as noise to ground and metallic noise. Balance as used in grade of service studies[22] and used by operating company engineers is defined as the difference in decibels between the main station C-message weighted noise to ground and metallic noise. This definition is implicitly based on the assumption that power induction is the main source of metallic noise. If this assumption is not satisfied—a condition that occurs when a significant component of metallic noise comes from sources such as crosstalk or office noise—then the result of a balance calculation may be misleading. Thus, to avoid this difficulty while still characterizing the susceptibility of the 1980 loop plant, balance was calculated only for those loops with a main station metallic noise of
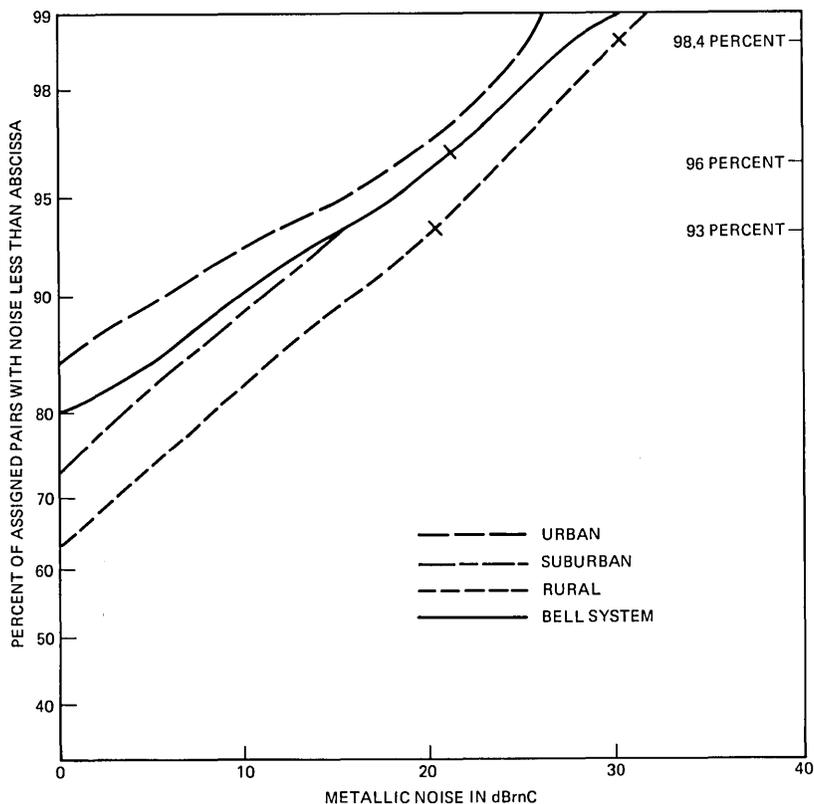
Fig. 22—Main station C-message weighted metallic noise as measured to main distributing frame for urban, suburban, rural, and Bell System loop populations.

greater than 0 dBrnC, a procedure which eliminated about 50 percent of the tested loops.

A scatter plot of balance versus noise to ground is given in Fig. 26 and suggests an interesting phenomenon. As we discussed in Section 4.3, metallic noise is generally below 30 dBrnC even on loops with high induction levels. Figure 26 indicates that the minimum balance improves as the induction increases and that the improvement in most instances is just sufficient to limit the worst-case metallic noise to 30 dBrnC. Since the increasing minimum balance cannot be attributed to a physical mechanism, this behavior suggests that operating company personnel have taken special measures to limit metallic noise to levels recommended in Bell Operating Company requirements. The cluster of plotted points on Fig. 26 with high induction and high balance indicates that a balance of 70 to 80 dB is achievable on noisy loops.
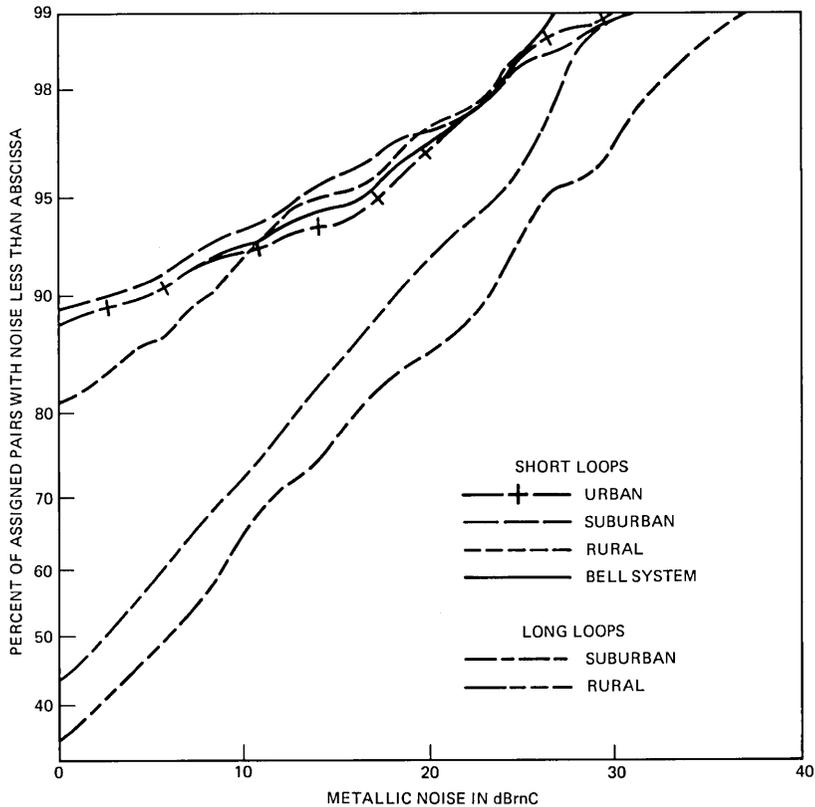
Fig. 23—Main station C-message weighted metallic noise as measured to main distributing frame for long (>20 kft) and short (≤20 kft) loops.

VI. SUMMARY

This report of the findings of the 1980 noise survey describes induced noise voltages and currents observed at subscriber and central office loop interfaces. The noise-to-ground, metallic noise, and longitudinal current measurements made as part of the survey have resulted in statistical descriptions of amplitude distributions, diurnal variations, and spectra of these quantities. As a result of the stratification used, differences between urban and nonurban environments have been identified, and the unique behavior of rural long loops has been highlighted. Moreover, it is evident that the operating company engineer plays an important role in controlling the metallic noise at subscriber locations.

This comprehensive characterization of the response of the loop plant to power line induction satisfies a pressing need for the data

|	90TH PERCENTILE (dBrnC)	MAXIMUM (dBrnC)
URBAN	11	26
SUBURBAN	16	69
RURAL	18	56
BELL SYSTEM	13	69
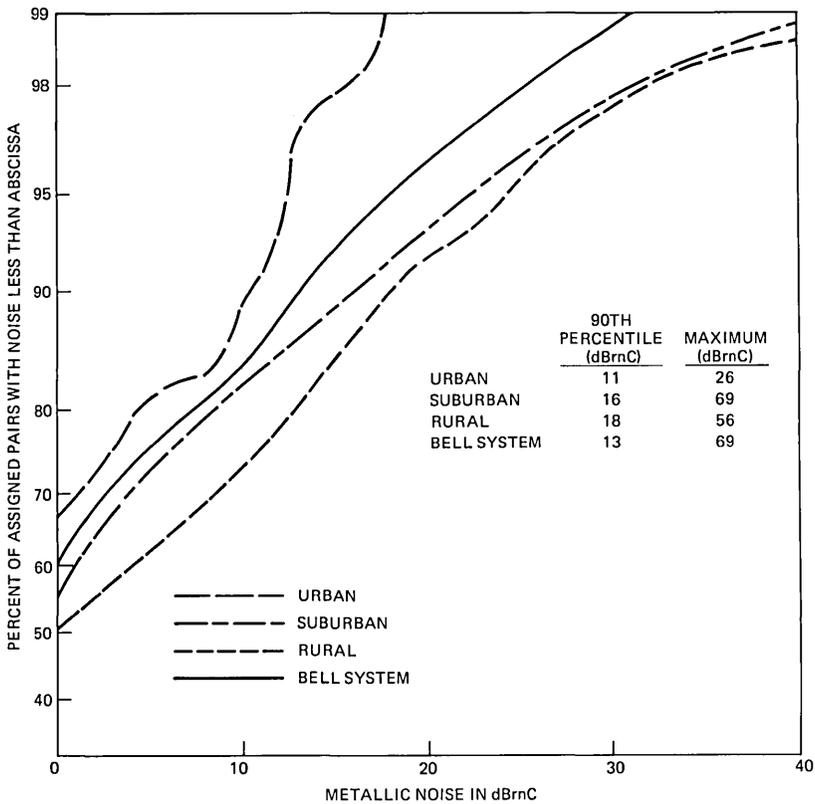
Fig. 24—Main station C-message weighted metallic noise as measured to quiet termination for urban, suburban, rural, and Bell System loop populations.

necessary to establish loop noise performance criteria and equipment design standards. In the long term, the insights into the inductive interference process provided by the 1980 survey will also lead to improved techniques for diagnosis and mitigation of induced noise problems.

VII. ACKNOWLEDGMENTS
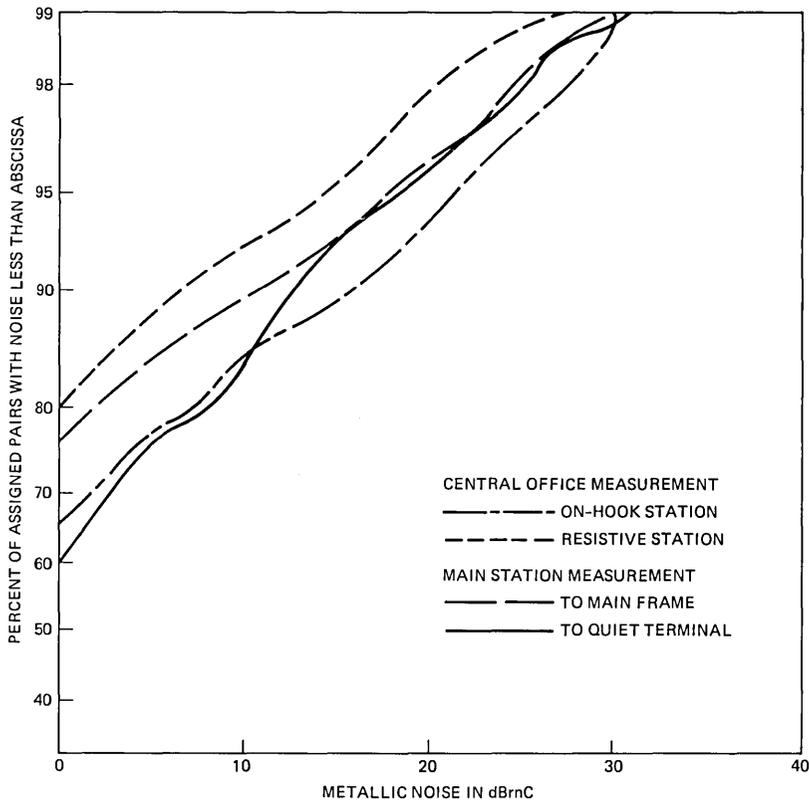
Fig. 25—Comparison of central office and main station metallic noise distributions.

Maher in formulating the statistical design and analysis techniques used. We also thank A. R. Eckler for his review of the appendix. Finally, the authors dedicate this paper to the memory of D. W. McLellan, who initially proposed the noise survey.

APPENDIX

Sample Design

This appendix describes the sample design of the 1980 noise survey. Section I details the method of selecting the sample and concludes with a description of the final sample. Section II describes the statistical techniques that were applied to the measured data and illustrates them with an example. The basic concepts of the sample design described here have been investigated in more detail by I. Nasell[23] and J. A. Maher.[24]
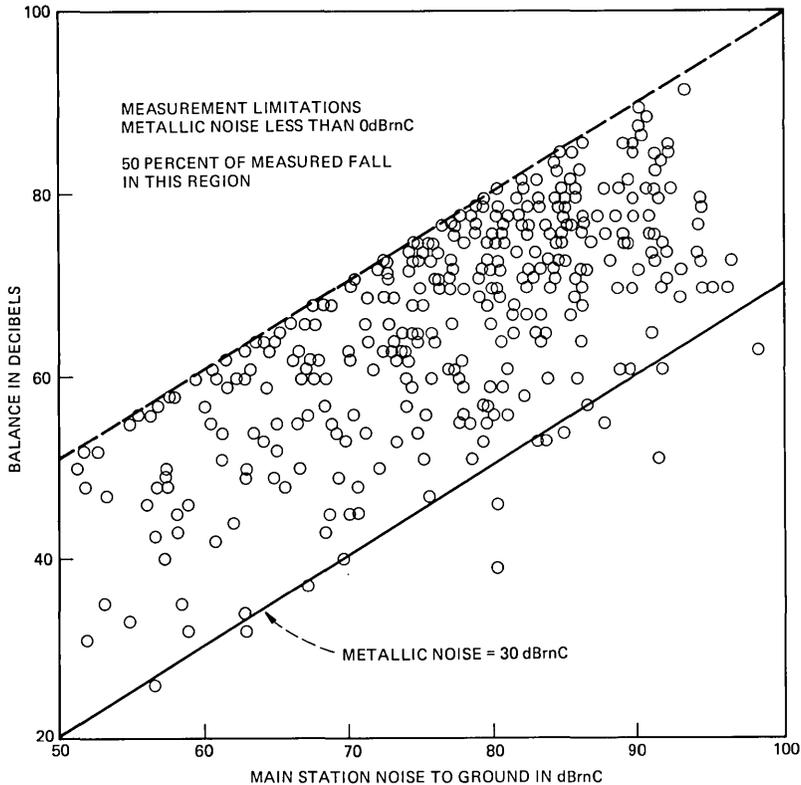
Fig. 26—Main station balance versus C-message weighted noise to ground.

A.1 Sample selection

A stratified, two-stage sampling plan was used to select sample units from the universe of working assigned pairs in the Bell System. Wire centers were stratified according to their assigned pairs per square mile (ap/sq.mi.). Urban centers served more than 100 ap/sq.mi., rural centers less than 45 ap/sq.mi., and suburban centers the intermediate densities. In the first stage of sampling, wire centers in each stratum were chosen with a probability proportional to size; in the second stage, assigned pairs were chosen randomly in each office. If a multi-party line was selected, then one of the main station locations was randomly chosen as the test point.

The next several paragraphs describe the method used to implement this sampling plan. The first step was to obtain a description of all the wire centers in the Bell System through the use of 1978 plant utilization data and a 1976 survey of the Bell System Outside Plant

Engineering records. Using the CLLI (Common Language Location Identifier) code as an index, comparisons were made between these two databases to obtain the number of assigned pairs and the area served by each wire center. (Neither database had sufficient information to be used alone.) Since CLLI codes can be reassigned as wire centers are created and restructured, it was necessary to verify that a unique wire center had been identified. If the 1976 and 1978 assigned pairs differed by less than 12 percent, i.e., the expected growth rate, it was assumed that a unique identification had been made.

The next step in implementation of the sampling procedure required defining stratum boundaries. This process was based in part on a characterization of central offices from which the sample of long loops had been selected for the 1964 loop survey.[2] The results of that study interpreted in light of the expected differences among the urban, suburban, and rural induction environments suggested a boundary
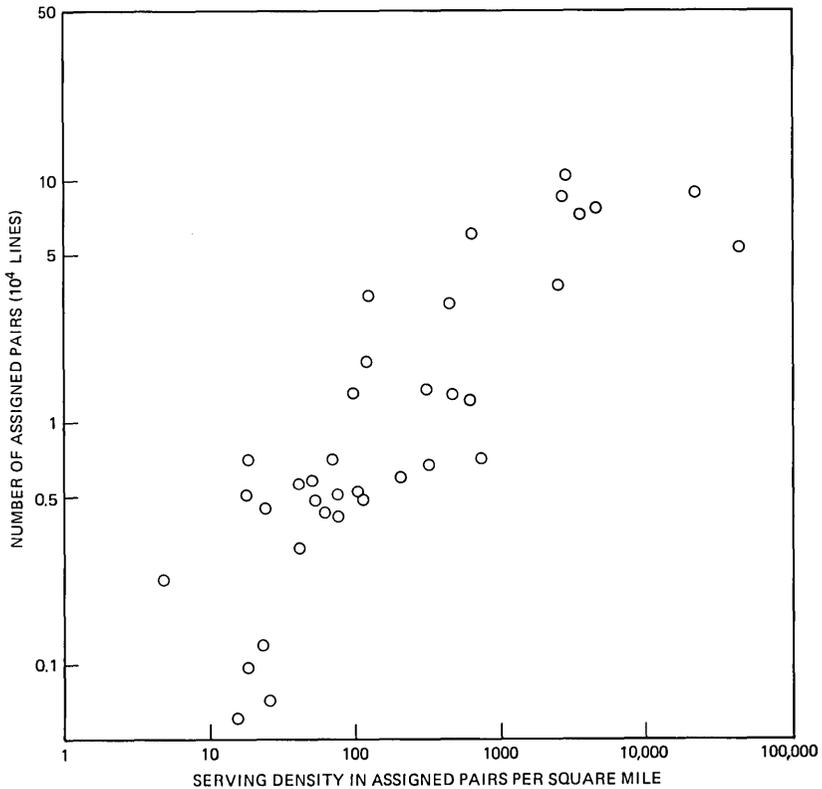
Fig. 27—Sampled wire center size versus serving density.

between rural and suburban offices of 45 ap/sq.mi. The choice of the urban boundary of 1000 ap/sq.mi. was based on the desire to balance the number of assigned loops in the urban and suburban categories.

Finally, a pilot survey in 1979, which included wire centers from the three strata, provided the basis for determining the final sample size. The criterion for the 1980 survey was that the mean noise to ground at the main station for each stratum should be determined with a 90-percent confidence interval of 2 dB. Although it is not immediately evident, post-survey analysis showed that the sample size arrived at based on this criterion was sufficiently large to include the noisier, longer loops of the rural and suburban populations. The final sample consisted of approximately 1256 loops, and the testing took 10 months.

Tables IV and V and Figs. 27 and 28 summarize the characteristics of the sample and locate the test sites. Table IV presents the proportion of loops and wire centers in each stratum and in the final sample. As we discussed in Section A.2, knowledge of the proportions of loops in each stratum is necessary to estimate statistical noise parameters. The relatively high porportion of suburban and rural loops in the sample results from the survey goal of characterizing the longer, noisier Bell System loops in a statistically meaningful way.

Surveys of the Bell System frequently use office size rather than office density as a criterion for stratification. Figure 27 illustrates the

U – URBAN SITE (6 SITES)
S – SUBURBAN SITE (18 SITES)
R – RURAL SITE (12 SITES)
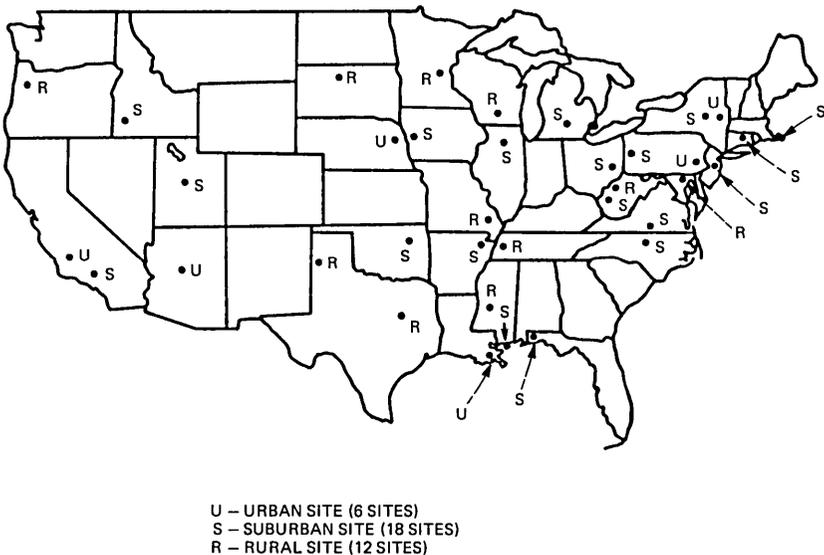
Fig. 28—Location of survey test sites.

Table IV—Percent of loops and central offices in urban, suburban, and rural environments

Sample	Urban (%)	Suburban (%)	Rural (%)
Assigned pairs in Bell System	49	41	10
Wire centers in Bell System	13	34	53
Loops in survey	14	51	34
Wire centers in survey	17	50	33

range of office sizes and densities present in the 1980 survey. Larger offices generally have higher serving densities, but as might be expected, a high correlation does not exist.

A.2 Estimates of statistics

This section describes the method used to estimate the means and cumulative distributions of the various noise parameters and the method used to estimate the variances of these estimates. The data used to form these estimates are assumed to have no significant experimental bias.

Ratio estimators were used to estimate means and cumulative distributions.[23] The form of the estimator given below is appropriate to the case of wire centers selected with a probability proportional to size and a random sampling of pairs within the centers.[23] The size assumed during the sample selection was the number of assigned pairs in each wire center in 1978. During the survey more recent information on office size was obtained. The formulas presented here and used in this report assume that the size of each wire center did not change from 1978 to the time of the survey. Nasell allows for a correction if wire center size did change; several distributions recalculated using this correction did not change significantly.

The forms of the estimators are summarized below. The statistic to be calculated for the Bell System population is assumed to be f. This statistic could be the mean of a parameter, or it could be the proportion of loops that have a parameter with a value less than a given level, x. Note that if f is this proportion of loops, then $f(x)$ is the cumulative distribution function. The ratio estimate of f can be written

$$f = \frac{\sum_{i=1}^{3} \left((w_i/n_i) \sum_{j} \bar{n}_{ij} \right)}{\sum_{i=1}^{3} \left((w_i/n_i) \sum_{j} \bar{d}_{ij} \right)}, \tag{1}$$

with

Table V—Summary of 1980 survey sites

Operating Company	Wire Center	CLLI	Environment	Test Dates
NET	Harwich, Massachusetts	HRWCMAMA	Suburban	4/28–5/2
SNET	Coventry, Connecticut	CNTYCTOO	Suburban	5/5–5/9
NJ Bell	Burlington, New Jersey	BURLNJBU	Suburban	2/18–2/22
Bell of PA	Sharpsville, Pennsylvania	SRVLPASH	Suburban	4/21–4/25
	Philadelphia, Pennsylvania	PHLAPAPE	Urban	12/1–12/5
C & P	Keedysville, Maryland	KDVLMDKV	Rural	3/17–3/21
	Montgomery, West Virginia	MTGMMVMG	Rural	4/7–4/11
	Oakhill, West Virginia	OKHLWVCH	Suburban	3/24–3/28
	Danville, Virginia	DAVLVADA	Suburban	3/31–4/4
Southern Bell	Pleasant Garden, North Carolina	GNBONCPL	Suburban	11/17–11/21
	Gulf Breeze, Florida	GLBRFLMA	Suburban	11/10–11/14
South Central Bell	Yazoo City, Mississippi	YZCYMSMA	Rural	10/20–10/24
	Ripley, Tennessee	RPLYTNMA	Rural	10/13–10/17
	Biloxi, Mississippi	BILXMSED	Suburban	10/27–10/31
	New Orleans, Louisiana	NWORLAMA	Urban	11/3–11/7
Southwestern Bell	Hale Center, Texas	HLCTTXHC	Rural	9/8–9/11
	Tyler, Texas	TYLRTXCH	Rural	9/15–9/19
	Campbell, Missouri	CMPBMOCH	Rural	9/29–10/3
	West Memphis, Arkansas	WMMPARMA	Suburban	10/6–10/10
	Tulsa, Oklahoma	TULSOKFI	Suburban	9/22–9/26
Pacific Bell	El Toro, California	ELTRCA11	Suburban	8/18–8/22
	Gardena, California	GRDNCA01	Urban	8/11–8/15
Pacific Northwest Bell	Dallas, Oregon	DLLSOR58	Rural	8/4–8/8
Mountain Bell	Springville, Utah	SPVLUTMA	Suburban	7/21–7/25
	Phoenix, Arizona	PHNXAZMA	Urban	8/25–8/29
	Twin Falls, Idaho	TWFLIDMA	Suburban	7/28–8/1
Illinois Bell	Bolingbrook, Illinois	BGBKILBK	Suburban	5/19–5/23
Ohio Bell	Uhrichsville, Ohio	UHVLOH92	Suburban	4/14–4/18
Michigan Bell	Lansing, Michigan	LNNGMIMN	Suburban	5/12–5/16
Northwestern Bell	Council Bluffs, Iowa	CNBLIADT	Suburban	6/16–6/20
	East Soderville, Minnesota	SDVLMNSO	Rural	5/26–5/30
	Mobridge, South Dakota	MBRGSDCO	Rural	6/9–6/13
	Omaha, Nebraska	OMAHNEIZ	Urban	6/23–6/27
New York Tel	Albany, New York	ALBYNYSS	Urban	12/8–12/10
	Clarksville, New York	CLVLYNCK	Rural	12/11–12/16
Wisconsin Bell	Richmond, Wisconsin	RCMDWI11	Rural	

$$\bar{n}_{ij} = \left(\sum_k u_{ijk}x_{jk}\right)/m_{ij} \tag{2}$$

and

$$\bar{d}_{ij} = \left(\sum_k u_{ijk}\right)/m_{ij}, \tag{3}$$

where the sum over index i is a sum over the strata, the sum over index j is a sum over offices in a stratum, and the sum over k is a sum over loops in an office. The proportion of assigned pairs in stratum i is w_i. The number of loops measured in office j of stratum i is m_{ij}. The number of offices in stratum i is n_i.

The parameters u_{ijk} allow for the determination of f for any pre-specified subpopulation of tested loops, e.g., long or short loops. If the tested loop is a member of the subpopulation, $u_{ijk} = 1$; otherwise $u_{ijk} = 0$. The special case of a ratio to size estimator assumes all tested loops are to be considered and $u_{ijk} = 1$ for all loops.

If the mean of a parameter is required, then x_{jk} is equal to the measured value of the parameter. If the proportion of loops with a parameter less than x is required, then $x_{jk} = 1$ if the measured value is less than x, and $x_{jk} = 0$ if the measured level is greater than or equal to x.

For a given stratum the statistic corresponding to f is

$$f_i = \frac{\sum_j \bar{n}_{ij}}{\sum_j \bar{d}_{ij}}, \qquad i = 1, 2, 3. \tag{4}$$

At the office level the statistic corresponding to f is

$$f_{ij} = \bar{n}_{ij}/\bar{d}_{ij}. \tag{5}$$

This report considers only the variance of the ratio-to-size estimator. For this case the variance is[23]

$$v^2(f) = (w_1)^2 v_1^2(f_1) + (w_2)^2 v_2^2(f_2) + (w_2)^2 v_3^2(f_3), \tag{6}$$

where $v_i^2(f_i)$ is the variance of f_i, $i = 1, 2, 3$.
The variance of f_i is

$$v_i^2(f_i) = \frac{1}{n_i(n_i - 1)} \left[\sum_{j=1}^{n_i} (f_{ij} - f_i)^2\right], \tag{7}$$

where n_i is the number of offices in stratum i.

	MEDIAN (dBrnC)	MEAN (dBrnC)	STANDARD DEVIATION OF ESTIMATE OF MEAN (dB)	STANDARD DEVIATION (dB)
URBAN	57	–	–	8.9
SUBURBAN	67	68	0.4	12
RURAL	69	70	0.8	12
BELL SYSTEM	61	–	–	12
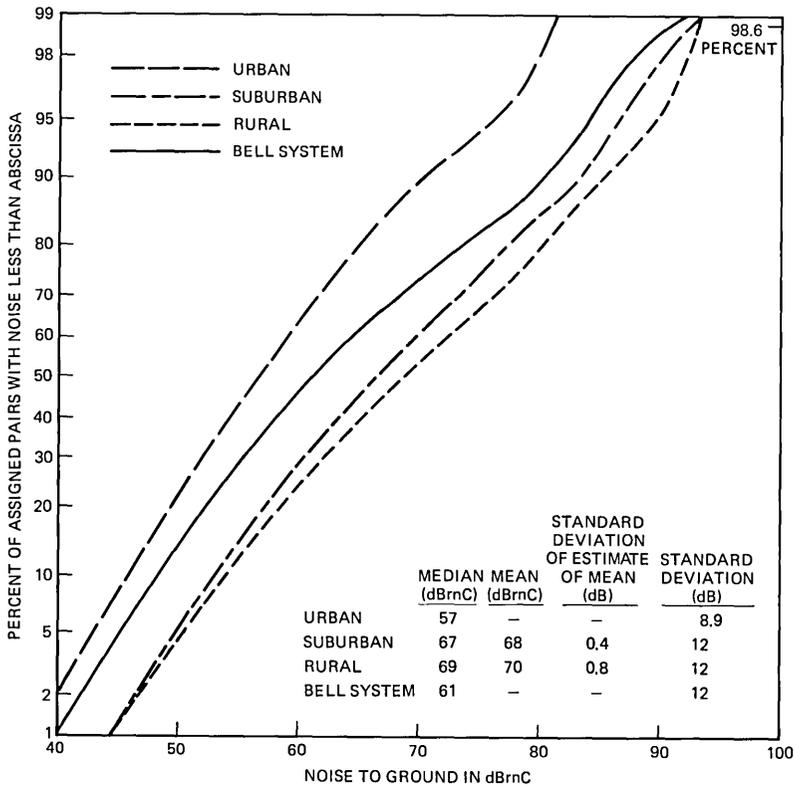
Fig. 29—Main station C-message weighted noise to ground for urban, suburban, rural, and Bell System loop populations.

As an illustration, Fig. 29 shows the distribution of the C-message noise to ground at the main station for the urban, suburban, rural, and Bell System environments. As is the case in Fig. 29, Bell System distributions for noise parameters generally lie between the distributions for urban and suburban environments with a larger variance than either.

Table VI presents statistics relevant to the median and 90th percentile of the rural and Bell System C-message weighted noise to ground. Two statistically equivalent confidence intervals are considered.[25] One confidence interval assumes that the quantile, x, is given and that a 90-percent confidence interval for the estimate of the percentile, $f(x)$, is to be calculated. This confidence interval is symmetrically located with respect to the estimate of the percentile and has a length of twice 1.67 times the sdem, where the sdem is the square root of the variance, v^2. The second confidence interval assumes that the percentile is given and that a 90-percent confidence interval for the quantile is to be determined. Since this confidence interval is not

Table VI—C-message weighted noise-to-ground statistics

(a) Median C-Message Noise to Ground

	Rural Loops		Bell System Loops	
Quantile	69	dBrnC	61	dBrnC
LLCI*	67	dBrnC	60	dBrnc
ULCI†	70.5	dBrnC	62	dBrnC
CIP‡	±5%		±3%	

(b) 90th Percentile C-Message Noise to Ground

	Rural Loops		Bell System Loops	
Quantile	85	dBrnC	80	dBrnC
LLCI*	84	dBrnC	78.5	dBrnC
ULCI†	86.5	dBrnC	80.5	dBrnC
CIP‡	±2%		±1.5%	

* Lower limit of 90% confidence interval for quantile.
† Upper limit of 90% confidence interval for quantile.
‡ 90% confidence intervals for percentile.

necessarily symmetric, upper and lower limits of the interval must be given.

Table VI indicates that for assigned pairs in rural environments, the 90th percentile of the C-message weighted noise to ground at the main station is 85 dBrnC. The confidence interval on the percentile means that for the given noise level of 85 dBrnC, the estimate of the percentile would have fallen between 88.7 and 91.3 percent for 90 percent of the possible samples that could be chosen by the sampling scheme described here. Equivalently, for the given percentile of 90 percent, the estimate of the noise level would have fallen between 78.5 and 80.5 dBrnC for 90 percent of possible samples.

REFERENCES

1. Bell Laboratories Staff, March 1974, unpublished work.
2. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., 48, No. 10 (December 1969), pp. 3337–83.
3. L. M. Manhire, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., 57, No. 1 (January 1978), pp. 35–59.
4. H. J. Beuscher and L. W. Richards, "No. 2 ESS Design for Operation in High 60 Hz Induction Environments," Conference Record, 1976 Nat. Telecommun. Conf., pp. 28.1.1–28.1.5.
5. P. R. Gray and D. G. Messerschmitt, "Integrated Circuits for Local Digital Switching Line Interfaces," IEEE Commun. Soc. Mag. (May 1980), pp. 12–23.
6. D. N. Heirman, "Time Variations and Harmonic Content of Inductive Interference in Urban/Suburban and Residential/Rural Telephone Plants," IEEE Trans. Commun., COM-23, No. 12 (December 1975), pp. 1484–95.
7. J. C. Parker, Jr., "A Probabilistic Characterization of 60 Hz Induction from Power Distribution Lines," IEEE 1976 Nat. Telecommun. Conf., November 19–December 1, 1976, pp. 12.1–1 to 12.1–6.

8. B. Szabados and E. J. Burgess, "Optimizing Shunt Capacitor Installation Using Inductive Co-ordination Principles," IEEE Trans. Power App. Syst., *PAS-96* No. 1 (January–February 1977), pp. 222–6.
9. J. C. Parker, Jr., "Analytical Foundation for Low-Frequency Power-Telephone Interference," B.S.T.J., *57*, No. 5 (May–June 1978), pp. 663–97.
10. D. S. Wilson, "A Statistical Method for Determining Customer Noise Due to Wire Pair Unbalances," Proc. 26th Int. Wire and Cable Symposium, November 15–17, 1977, pp. 440–53.
11. G. Miller, "The Effect of Longitudinal Imbalance on Crosstalk," B.S.T.J., *54*, No. 7 (September 1975), pp. 1227–51.
12. "IEEE Standard Test Procedure for Measuring Longitudinal Balance of Telephone Equipment Operating in the Voice Band," IEEE Std. 455-1976, (September 30, 1976).
13. R. E. Witter, "Harmonic Analysis of a Rural Electric System Serving Overexcited Distribution Transformers," Conf. Paper, IEEE 1978 Rural Elec. Power Conf., Minneapolis, MN, April 30–May 2, 1978.
14. F. G. Doell and J. St. Arnand, "60 Hertz Power Harmonic Interference from a Multigrounded Neutral System Into Telephone Circuits," IEEE Electromagnet. Compat. Symposium Rec., San Francisco, CA, July 16–18, 1974, p. 162–7.
15. B. Szabados, E. J. Burgess, and W. A. Noble, "Harmonic Interference Corrected by Shunt Capacitors on Distribution Feeders," IEEE Power Eng. Soc. Summer Meeting, Portland, OR, July 1976, p. 234ff.
16. Edison Electric Institute and Bell Telephone System, "The Telephone Influence Factor of Supply Systems Voltages and Currents," Joint Subcommittee on Development and Research, Supply Engineering Report 33, EEI Publication 60-68, New York, NY September 1960.
17. D. Crevier and A. Mercier, "Estimation of Higher Frequency Network Equivalent Impedances by Harmonic Analysis of Natural Waveforms," IEEE Power and Eng. Soc., Mexico City, July 17–22, 1977.
18. A. A. Smith, Jr., *Coupling of External Electromagnetic Fields to Transmission Lines*, Wiley, 1977.
19. A. J. Aikens and D. A. Lewinski, "Evaluation of Message Circuit Noise," B.S.T.J., *34*, No. 4 (July 1960), pp. 879–909.
20. F. J. Uhrhane, "Mechanized Loop Testing Strategies and Techniques," B.S.T.J., *61*, No. 6 (July–August 1982), pp. 1209–34.
21. H. W. Lilliefors, "On the Kolmogrov-Smirnov Test for the Exponential Distribution with Mean Unknown," J. Amer. Statist. Ass. (March 1969).
22. D. A. Lewinski, "A New Objective for Message Circuit Noise," B.S.T.J., *43*, No. 2 (March 1964), pp. 719–40.
23. I. Nasell, private communication.
24. J. A. Maher, private communication.
25. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. II, 3rd Edition, Hafner, 1973.

AUTHORS

Donald V. Batorsky, BSEE and MSEE, 1965, Ph.D. (Electrophysics), 1972, Polytechnic Institute of Brooklyn; AT&T Bell Laboratories, 1965–1983. Present affiliation Bell Communications Research, Inc. At AT&T Bell Laboratories Mr. Batorsky worked on the analysis of Electromagnetic Pulse coupling to large communication systems, low-frequency inductive interference into distribution telephone plants, and wideband access planning. Currently, he is a District Manager in the Distribution Capabilities and Architecture Division of Bell Communications Research, Inc., where he is responsible for loop cost modeling and architecture.

Michael E. Burke, B.S. (Electrical Engineering), 1977, Case Western Reserve University; S.M. (Electrical Engineering), 1979, Massachusetts Institute of Technology; AT&T Bell Laboratories, 1977—. Mr. Burke has worked on the interaction between power and telephone lines. He is currently working in the Exploratory Loop Systems Department. Member, IEEE, Eta Kappa Nu, Sigma Xi.

On the Accuracy of Forecasting Telephone Usage Demand

By M. N. YOUSSEF*

This paper examines the accuracy of telephone-usage-demand forecasts produced by three state-of-the-art forecasting techniques: Box-Jenkins, Akaike state space, and an autoregressive spectrum estimation. The study considers 35 actual monthly demand time series and 300 simulated realizations. Principal results are that: (1) correct identification of the nonstationary behavior of telephone demand is crucial to forecast performance, (2) overparameterization or underparameterization of the stationary aspects of a process has little or no impact on the accuracy of the forecast, and (3) forecasts based on the naive random-walk method compare favorably to those produced by sophisticated techniques. Also, strengths and weaknesses of the investigated techniques are revealed through data analysis. It is further argued that the traditional method for assessing the accuracy of the usage forecast based on average busy season quantities is biased towards underforecasting.

I. INTRODUCTION

Errors in forecasting per-telephone usage demand at the switching-machine level can be quite costly for the Bell operating companies. It has been estimated that a net 3-percent overforecast would yield an increase in capital requirements of about $800 million.[1] The result of underforecasting is degraded service.

The usage variable to be projected is known as the hundred call seconds per main station (CCS/M).[†] This variable is required to

---

* AT&T Bell Laboratories.

[†] One CCS is one hundred call seconds of usage per hour and is equal to 1/36 Erlang.

determine central office relief timing and sizing strategies. In this regard forecasts are needed up to five years ahead.

Historically, the most commonly used CCS/M forecasting technique in the Bell operating companies has been linear regression trending. The result has often been a large forecast error. In this paper, we examine three state-of-the-art time-series techniques. These are the Autoregressive Integrated Moving Average (ARIMA) class of models and the Box-Jenkins philosophy, a state space method developed by Akaike, and a robust autoregressive spectrum estimation and factorization approach.

The Box-Jenkins technique provides a procedure for model building.[2] But, because the procedure requires user judgment, it can lead different users to identify different models for the same set of data. The question naturally arises as to what situations lead to different identified models and what impact this has on forecasting accuracy.

On the other hand, attempts have been made to eliminate judgment by developing procedures that automatically determine the order of the most representative model. The state space approach presented in this paper and its information criterion (known as the AIC) have been advocated to do just that, assuming that the given process is stationary.[3]

Furthermore, efforts were spent in seeking procedures that may circumvent the identification stage entirely. The nonparametric method for estimating the spectrum, and hence the coefficients of autoregressive models, is a result of these efforts. It is examined here since it may provide a viable alternative.[4]

As a first step in assessing the three techniques, we selected 35 time-series data on CCS/M from more than 100 series that were available. In this selection we avoided time series that were contaminated with many outliers or with large jumps. The selection was strictly based on visual inspection of the time-series graphs.

Limiting the investigation to "clean" data does not hinder the significance of its findings. Our primary objective is to answer the question: Is there a method better than a naive one that can improve the CCS/M forecast even for the most regular offices? Moreover, concerning outliers, steps have been taken to guarantee better data-collection procedures via reliable mechanization. As for jumps, we were able to associate a causative factor with each jump observed in the data. Such information ought to be incorporated in the forecasting mechanism.

Results obtained from applying these techniques to the selected time series show that none of the techniques was able to provide a marked improvement in forecast accuracy—even for 1 year ahead—when compared to a simple-minded forecast that uses data from 1

year past. Moreover, the models identified via the state space approach were different from those identified by the Box-Jenkins technique. Existing literature concerning the performance of the AIC criterion in identifying the order of a pure autoregressive model indicates that the AIC has a tendency to overparameterize.[5] Our results show that the criterion mostly underparameterizes and is inconsistent in general.

As a first impression, one may attribute these findings to many factors. Some possibilities include the short history data, outliers, failure to reduce the data to stationary time series, nonconstant variance of the noise process and/or the applicability of the criterion used in comparing forecasting accuracy of the various methods.

The Bell operating companies measure of forecast accuracy for the CCS/M is the percent difference between a forecast and an actual for a quantity known as Average Busy Season (ABS). The ABS is defined as the average of the highest three observations in a year. Forecast accuracy is only relevant for ABS values. It is the ABS forecast that is used for planning and engineering.

A simulation experiment was conducted to verify and explain results obtained from analyzing the CCS/M data. Three typical processes identified from the CCS/M time series were chosen. In the simulation, one hundred realizations were generated from each of the three processes. Using the simulated data to assess the various approaches, findings similar to those obtained from analyzing actual data were disclosed. These findings, however, provided additional insight into a number of aspects of the problem.

This paper is organized as follows. Section II outlines the three investigated techniques. Section III discusses our experience in applying these techniques to the CCS/M data, and illustrates, where possible, the effect of overfitting, underfitting, and misspecification of model parameters on the forecast accuracy. Section IV describes the simulation experiment and assesses the performance of the three approaches in an artificially structured ideal environment for the CCS/M processes. Section V summarizes the results and outlines a number of observations that were noted during the investigation.

II. METHODOLOGY

The three investigated approaches for the analysis and forecasting of the CCS/M variable are all claimed to be powerful. They are also complicated and expensive (in varying degrees). They depend heavily on elaborate software packages that are not normally available in standard computer systems.

Our reason, then, for investigating these approaches was to understand the properties of the data in order to suggest a simple method. Hypothetically, using state-of-the-art methods would allow for quan-

tifying the maximum forecast accuracy that can be achieved, or the maximum loss that may be incurred due to simplification.

In the next three subsections, we outline these methods. The following presentation assumes that the reader is familiar with autoregressive and moving average processes.

2.1 The Box-Jenkins approach for seasonal data

For seasonal time series of period S Box and Jenkins suggest representation by members of the following class

$$\phi_p(B)\Phi_P(B^S)(1 - B)^d(1 - B^S)^D(z_t - \mu) = \theta_q(B)\Theta_Q(B^S)a_t, \quad (1)$$

where

$$z_t = \text{the time series data at time t}$$
$$\mu = \text{the location parameter}$$
$$B = \text{the backshift operator such that } B^k z_t = z_{t-k}$$
$$a_t = \text{independent random variable with mean zero and variance } \sigma_a^2, \text{ known as white noise}$$
$$d \text{ and } D = \text{the degrees of differencing required for achieving process stationarity}$$
$$\phi_p(B), \Phi_P(B), \theta_q(B) \text{ and } \Theta_Z(B) = \text{polynomials in } B \text{ of order } p, P, q, \text{ and } Q, \text{ respectively.}$$

The class of models in (1) is known as multiplicative seasonal Autoregressive Integrated Moving Average (ARIMA) of order $(p, d, q) \times (P, D, Q)_S$.

The Box-Jenkins approach for fitting models of the form (1) involves a three-part cycle of identification, estimation, and diagnostic checking. The cycle is ended once an adequate model is derived. The basic tools for model identification are two descriptive functions known as the sample Autocorrelation Function (ACF) and the sample Partial Autocorrelation Function (PACF). Nonstationarity of a series is identified from the behavior of its ACF. For nonstationary series, the serial correlation in the ACF remains large at large lags. In this situation, the integrated part of the ARIMA representation allows for differencing to induce stationarity. Once stationarity is achieved, the problem is to select reasonable values for p, q, P, and Q. This is also done by examining the shape of the ACF (and PACF) of the stationary series. In the estimation stage, the parameters ϕ's, Φ's, θ's, and Θ's of the identified model are calculated on the basis of the minimization of the sum of squared errors $\{a_t\}$.

After the model is fitted, the residuals $\{\hat{a}_t\}$ are checked for whiteness. If no pattern is detected in the autocorrelations of residuals, the assumption that the a_t's are white noise is accepted. If, on the other

hand, a pattern is observed, such a pattern would provide information on how to modify the model. In the latter case, the identification, estimation, and diagnostic checking cycle is repeated.

We note from (1) that the Box-Jenkins approach provides a capability for representing a wide class of nonstationary time series. Through differencing an appropriate number of times, the nonstationary behavior can be removed. The use of differencing, however, requires judgment. Nevertheless, we adopted this technique to transform all 35 CCS/M time-series data before employing any of the other approaches. These other approaches deal only with the representation of stationary processes. But, we tested the impact of adopting this technique on the results via the simulation experiment.

2.2 State space approach

While the state space approach can accommodate multivariate systems, for the purpose of this study, we address the univariate case only. From control theory, a linear, discrete-time, time-invariant system can be represented by

$$v_{t+1} = Fv_t + gu_{t+1}$$

$$x_t = hv_t, \tag{2}$$

where v_t is an $r \times 1$ state vector, and $\mathbf{F}$, $\mathbf{g}$, $\mathbf{h}$ are $r \times r$, $r \times 1$, and $1 \times r$ matrices, respectively. This representation assumes x_t (scalar) to be the output and u_t to be the input of the system.

Akaike showed that, when x_t and u_t are stochastic Gaussian processes an analogous representation to (2) exists.[6,7] The new representation is called Markovian and can be obtained from the analysis of canonical correlations between the set of the present and future output and the set of the present and past input. We consider in this paper the case where there is a feedback from the output to the input, that is, when $u_t = x_t$. This Markovian representation takes the form

$$v_{t+1} = Fv_t + ga_{t+1}$$

$$x_t = hv_t, \tag{3}$$

where a_{t+1} is the innovation of x_t at time $t + 1$. It is defined by

$$a_{t+1} = x_{t+1} - x_{t+1|t}. \tag{4}$$

Here $\{a_t\}$ is white noise with autocovariance $c_r = \mathbf{E}(a_t \cdot a_{t-r})$. Also, $x_{t+1|t}$ is the projection of x_{t+1} onto the linear space $[R(t-)]$ spanned by the components of the present and the past $(x_t, x_{t-1}, \cdots)$. It is the one-step-ahead predictor of x_t. The space that is spanned by the components of the predictors $x_{t+i|t}$ is called the predictor space $[R(t+|t-)]$. The components of the state vector, v_t, are elements of this space,

which is assumed to have a finite dimension. In this regard v_t provides a specification of the predictor space.

Since any basis of the predictor space can play a role of the state vector, $\mathbf{F}$, $\mathbf{g}$, and $\mathbf{h}$ in (3) can have different structures. However, a state vector whose elements are the first maximum set of independent components of the predictor space has the smallest possible dimension. This "minimal" representation defines a canonical representation of the system. Its dimension is equal to the number of nonzero canonical correlation coefficients. We are interested in the canonical representation because it corresponds to the concept of parsimony in the ARMA representation, and hence to the problem of identifiability of the process.

To show the relation between the Markovian representation and the ARMA representation, suppose an ARMA model for x_t is given in the form

$$x_t + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} = a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} \qquad (5)$$

or

$$\phi(B)x_t = \theta(B)a_t.$$

The zeros of $\phi(B)$ and $\theta(B)$ are constrained to lie outside the unit circle to ensure stationarity and invertibility of the process. Hence x_t can be expressed as

$$x_t = \sum_{i=0}^{\infty} \psi_i a_{t-i}, \qquad (6)$$

where

$$\psi_0 = 1 \qquad \text{and} \qquad \sum_{i=0}^{\infty} \psi_i^2 < \infty.$$

Since $x_{t+i|t} = x_{t+i}$ for $i = 0, -1, -2, \cdots$ and $a_{t+i} = 0$ for $i = 1, 2, \cdots$, eq. (5) can be expressed by

$$x_{t+i|t} + \phi_1 x_{t+i-1|t} + \cdots + \phi_p x_{t+i-p|t}$$

$$= a_{t+i|t} + \theta_1 a_{t+i-1|t} + \cdots + \theta_q a_{t+i-q|t}. \qquad (7)$$

For $i \geq q + 1$, the right-hand side of (7) vanishes. Thus for $r \geq \max (p, q + 1)$, (7) becomes

$$x_{t+r|t} = -\phi_1 x_{t+r-1|t} - \phi_2 x_{t+r-2|t} - \cdots - \phi_r x_{t|t}, \qquad (8)$$

where

$$\phi_i = 0 \qquad \text{for} \qquad r > p.$$

Also, from (6) one can get

$$x_{t+i+1|t+1} - x_{t+i+1|t} = \Psi_i a_{t+1}. \tag{9}$$

From (8) and (9) one can see that the state vector $v_t = (x_{t|t} x_{t+1|t} \cdots x_{t+r-1|t})'$ provides a Markovian representation in the form given by (3), where

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & & 1 \\ -\phi_r & -\phi_{r-1} & \cdots & -\phi_1 \end{bmatrix}$$

$$\mathbf{g} = [1 \quad \psi_1 \quad \cdots \quad \psi_{r-1}]'$$

$$\mathbf{h} = [1 \quad 0 \quad \cdots \quad 0]'.$$

Here, the ψ's are obtained from the relation

$$\theta_i = \psi_i + \phi_1 \psi_{i-1} + \cdots + \phi_p \psi_{i-p}, \qquad i = 1, 2, \cdots, q,$$

where

$$\psi_j = 0 \qquad \text{for} \qquad j < 0.$$

Thus, given an ARMA representation of a stationary process, one can derive a Markovian representation. Similarly, it can be shown that any stationary process that has a Markovian representation has an ARMA representation. Also, the minimal Markovian representation—known as canonical representation—corresponds to a parsimonious ARMA model.

In practice, the canonical representation needs to be identified from a finite length of a time series. The theoretical canonical correlation coefficients are replaced by the sample canonical correlation coefficients. A criterion is required to determine how small the sample canonical correlation coefficients should be to be judged zero. An information criterion (AIC) was suggested by Akaike in this regard. The AIC criterion is defined by

$$\text{AIC} = -2(\text{maximum log likelihood}) + 2(p + q).$$

An approximation of the likelihood function is calculated from the sample autocovariances. The log likelihood measures the goodness of model fit. The term $2(p + q)$ gives preference to a structure with the least number of free parameters. It was argued that the canonical structure has Minimum AIC (MAICE). There is no theoretical proof of the optimality of the MAICE in this context however.

A procedure for automatically identifying the canonical structure from time-series data is described in Ref. 8. In this procedure, the

amount of past to be used in the canonical-correlation analysis is determined from fitting a sequence of autoregressive models. It is equal to the order of the model that has MAICE. The canonical correlations between the past data and the future with increasing number of steps are then calculated. These calculations determine a structure of the state vector by which variables that yield large correlations are included. Usually, the provided structure of the state vector is not necessarily adequate.[7] Besides, estimates provided by this type of analysis are poor. However, this structure and the estimated parameters may be used as initial guesses. Computation of the likelihood function for this structure and all possible choices of the state vector are then performed. The MAICE structure is hence selected.

For maximum likelihood estimation, an approximation was developed since direct maximization of the likelihood function for each possible structure is a formidable task. The developed algorithm is based on Davidon's variance algorithm,[7] which requires the evaluation of the gradient and the inversion of the approximate Hessian of the log likelihood. Our experience with the application of this approach to telephone usage data will be further discussed in Section 3.3.

2.3 Autoregressive spectrum estimation approach

This approach involves the fitting of an autoregressive process of order p—where p is unknown—with the coefficients being estimated in the frequency domain. The power spectral density function and its factorization are used to estimate model parameters. The order, p, may be determined using a stopping rule criterion, such as Parzen's[9] or Akaike's.[10] The following is a brief account of an autoregressive spectral estimation approach. The reader is also referred to the work in Refs. 9, 11, 12, and 13.

Let x_t be a zero mean, discrete time, stationary stochastic process with autocovariance C_τ at lag τ. Then

$$C_\tau = \text{cov}(x_t, x_{t-\tau}).$$ (10)

The power spectral density function $f(\omega)$ at frequency ω is defined as

$$f(\omega) = \frac{1}{2\pi} \sum_\tau C_\tau e^{-i\omega\tau}, \qquad |\omega| \leq \pi.$$ (11)

Since $\{x_t\}$ can be represented by an autoregressive process in the form

$$\sum_{j=0}^{\infty} \phi_j x_{t-j} = a_t,$$ (12)

where

$$\phi_0 = 1 \quad \text{and} \quad a_t \sim N(0, \sigma_a^2),$$

the spectral density of the process is

$$f(\omega) = \frac{\sigma_a^2}{2\pi} \left| \sum_{j=0}^{\infty} \phi_j e^{ij\omega} \right|^{-2} = \frac{\sigma_a^2}{2\pi[A(\omega)]^2}, \tag{13}$$

where

$$A(\omega) = \sum_{j=0}^{\infty} \phi_j e^{ij\omega}$$

is the transfer function of ϕ_j's.

For a known $f(\omega)$ and an infinite series x_t, factorization of the spectrum provides values for the parameters ϕ_j's. A procedure for factorization of $f(\omega)$ that can determine model coefficients is described in Ref. 4.

In practice, however, the spectral density function is unknown and there is only a finite realization, say of length T, of a time series. The concept of a "windowed" estimate for $f(\omega)$ is used for carrying out the factorization and subsequently the estimation of ϕ_j's. The following discussion outlines the method. Briefly,

$$C_\tau^T = \frac{1}{T} \sum_{t=0}^{T-r-1} x_t x_{t+r} \qquad 0 \leqslant \tau < T \tag{14}$$

gives the sample covariance function. An estimation of the spectral function can be expressed by

$$f^T(\omega) = \sum_\tau \frac{1}{2\pi} C_\tau^T \, k\left(\frac{\tau}{p}\right) e^{-i\tau\omega}, \tag{15}$$

where p is the truncation point that gives the order of the model, and $k(v)$ is called the lag window. Usually, $f^T(\omega)$ and subsequently ϕ_j's are estimated at N points equally spaced in $(0, \pi)$. The equation $f^T(\omega_i)$, $\omega_i = 2\pi i/2N$ for $(i = 0, \cdots, N - 1)$ is called the "windowed" estimate of $f(\omega)$.

The spectral analysis and estimation of model parameters are done by three applications of the Fast Fourier Transform (FFT) algorithm.[13] The software for carrying out this analysis is currently available in the Statistical Computing Library (STATLIB).[14]

Our use of the analysis in the frequency domain is limited to the estimation of the model parameters $(\phi_j, j = 1, \cdots, p)$. The advantage of this approach is that the problem of model identification can be avoided for the stationary part of the process.

III. APPLICATION OF METHODOLOGY TO THE CCS/M DATA

Central office CCS/M monthly data were sampled for 35 switching entities. Preliminary analysis of these data was performed to check if

the requirements on the methodology used were met. Data transformations were employed where needed. Forecasting models from each of the three techniques were fitted to the transformed data for the period between May 1969 and April 1975. Forecasts were then generated for the next 24 months and converted to ABS values. The percent forecast error—which is defined as the percent difference between actual and forecast for ABS values—for 1 and 2 years ahead were hence calculated. Percent forecast errors were also calculated for the random-walk method, which considers the future ABS of a given series to be its most recent ABS value. This naive method is selected to serve as a benchmark in assessing forecast performance for the other compared methods.

3.1 Preliminary data analysis

As a first step in the analysis, we plotted the raw data for each of the time series. Figures 1 through 3 show time-series graphs for three of the series (A, B, C). We consider the behavior exhibited in Figs. 1 and 2 to be typical of the CCS/M data and that of Fig. 3 to be less typical. The sample autocorrelation function for Series A is also shown in Fig. 4. Preliminary assessments of the time-series graphs indicate (1) there are obvious periodicities at lag 12, (2) apparent upward or downward trends can be located in few series (Figs. 2 and 3), (3) for series that are not contaminated with outliers and/or jumps, stationarity can be attained by differencing the data at lag 12, and (4) transformation of differenced data is unnecessary. Further tests indicate that the assumption of process linearity can be accepted in general.

As mentioned above, the method adopted to induce stationarity is the first step in the identification process of the Box-Jenkins procedure. By applying the appropriate differencing, a time zeries $\{z_t\}$ can be converted to a stationary series $\{x_t\}$. Our analysis showed that homogeneous nonstationarity of the data can be removed by removing the seasonal component (Figs. 4 and 5). Thus, each of the CCS/M series was differenced at lag 12 prior to the application of the forecasting techniques.

3.2 Application of Box-Jenkins technique

Forecasting models for the selected time series, with varying characteristics of the CCS/M variable, were individually identified using the Box-Jenkins approach. In the following, we briefly discuss the application of the method to Series A (Fig. 1). From the autocorrelation function (Fig. 5) of the differenced series $\{x_t\}$

$$x_t = (1 - B^{12})z_t, \tag{16}$$
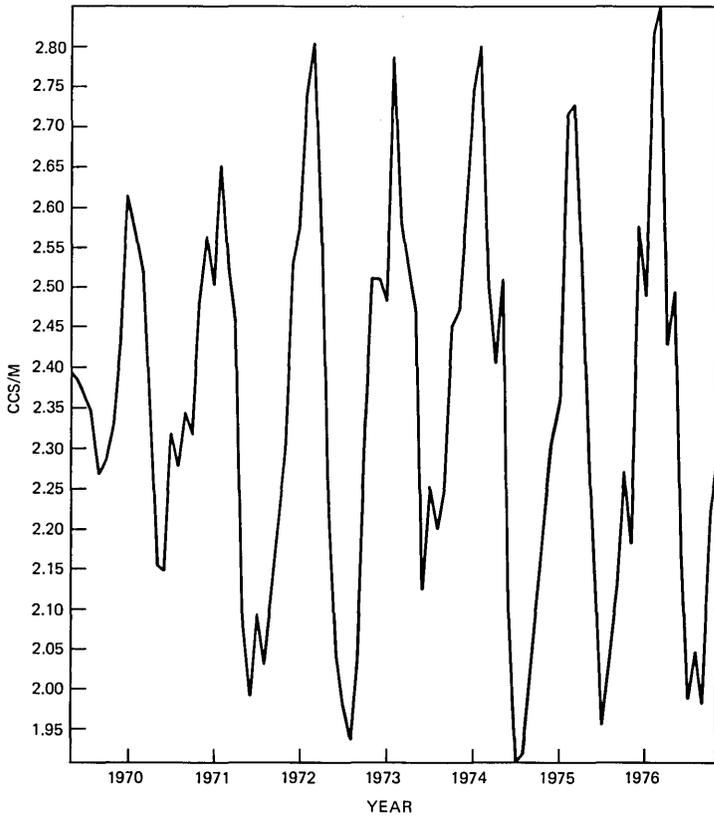
Fig. 1—Monthly data on CCS/M, Series A. Visible seasonal variation—typical behavior.

we note that $\{x_t\}$ is not a white noise process since there is still useful predictive information that can be further explained. As a formal test of hypothesis on the randomness of x_t, we calculated the statistic Q as

$$Q = n \sum_{k=1}^{36} r_k^2(x_t) = 78,$$

where r_k is the kth autocorrelation coefficient of x_t and n (= 60) is the number of points in the differenced series.

For x_t to be white noise, at the 5-percent level

$$Q_{\text{conf}} < \chi_{0.95}^2(36) = 50.$$

Since $Q > Q_{\text{conf}}$, we considered modifications of the model in (16). We note, from the sample autocorrelation function of $\{x_t\}$, that the correlation coefficient at lag 12 is large. This suggests a moving average term of order 12. The autocorrelation pattern suggests a low-order
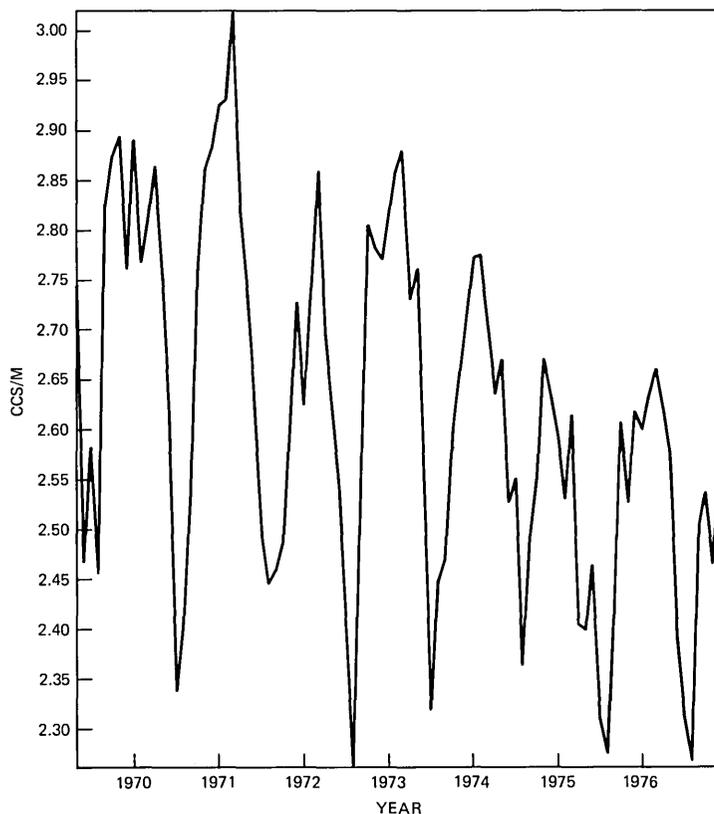
Fig. 2—Monthly data on CCS/M, Series B. Apparent trend variation for peak months—typical behavior.

autoregressive term. Thus we considered the model

$$(1 - \phi B)x_t = (1 - \Theta B^{12})a_t. \tag{17}$$

After estimation, (17) becomes

$$(1 - 0.63B)(1 - B^{12})z_t = (1 - 0.47B^{12})a_t, \tag{18}$$

where $\hat{\sigma}_a = 0.12$. Further inspection of the autocorrelation function of the residuals (Fig. 6) suggested no further modification [$Q = x^2(34) = 18$].

3.2.1 Alternative ARMA models

Models other than those individually identified by the Box-Jenkins procedure for the examined series were investigated. Alternative models were considered in an attempt to understand the effect of overfitting, underfitting, differencing, and modeling a spurious linear trend on the forecast of CCS/M. For the sake of space, these effects
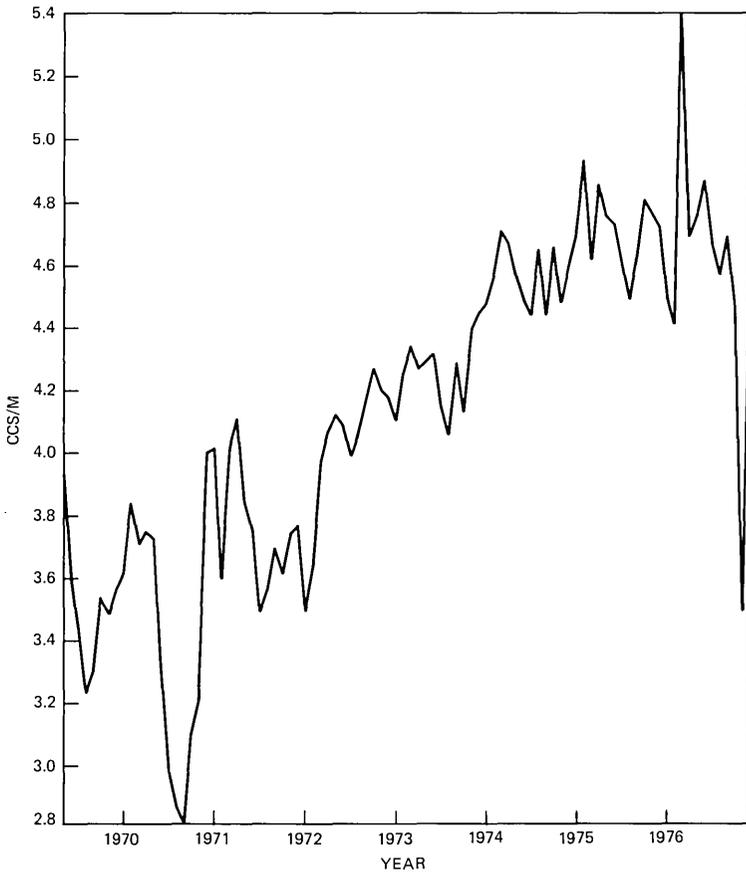
Fig. 3—Monthly data on CCS/M, Series C. Atypical behavior.

will be illustrated graphically by showing the data, and a model fit and its extrapolation, over two to four years for some selected alternatives.

To test the effect of overfitting an MA(1) term to the previously identified form for Series A, consider

$$(1 - \phi B)(1 - B^{12})x_t = (1 - \theta B)(1 - \Theta B^{12})a_t. \qquad (19)$$

The above model specification was examined, since it was identified for several other CCS/M series, using the Box-Jenkins procedure.

After estimation of model parameters, (19) becomes

$$(1 - 0.61B)(1 - B^{12})z_t = (1 + 0.05B)(1 - 0.47B^{12})a_t, \qquad (20)$$

where

$$\hat{\sigma}_a = 0.12, \qquad Q = 19.0.$$
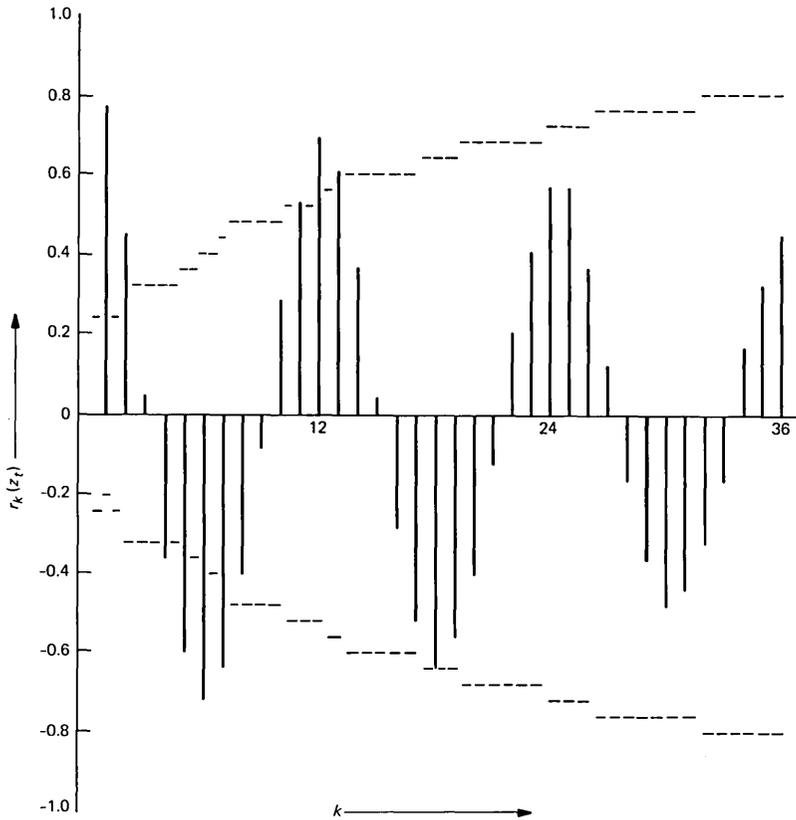
Fig. 4—Estimated autocorrelation, Series A.

The estimated value for the overfitting parameter, θ, is not significantly different from zero, and the goodness-of-fit measure does not indicate inadequacy. And, as expected, this model generates forecasts that are identical to those generated by the original model (18).

Now, consider Series B (Fig. 2). The originally identified Box-Jenkins model for this series is

$$(1 - 0.79B)(1 - B^{12})z_t$$

$$= (1 - 0.33B)(1 - 0.50B^{12})a_t, \qquad \hat{\sigma}_a = 0.12. \quad (21)$$

Three alternatives are discussed. First

$$(1 - \phi B)(1 - \Phi B^{12})z_t = a_t \qquad (22a)$$

with estimated parameters

$$\phi = 0.99, \qquad \Phi = 0.33, \qquad \sigma_a = 0.12.$$
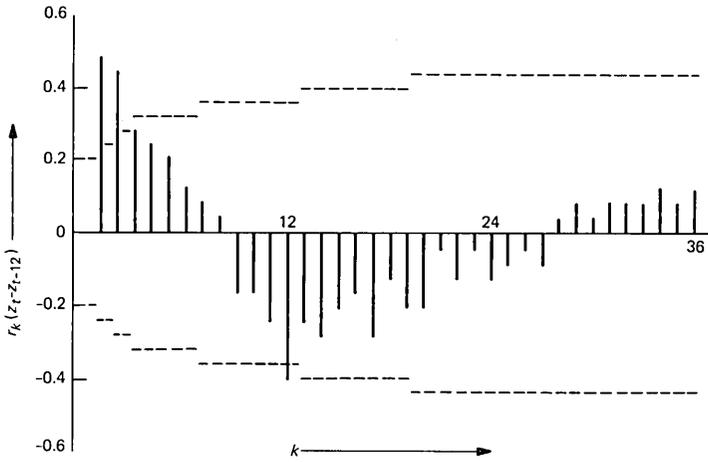
Fig. 5—Estimated autocorrelation, Series A, differenced at Lag 12.
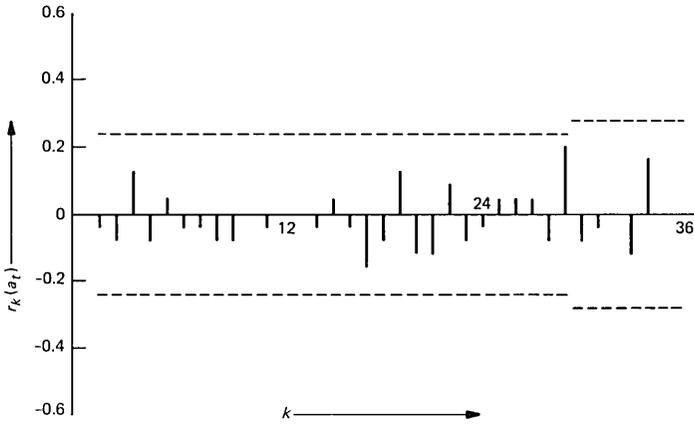
Fig. 6—Estimated autocorrelation of residuals, Series A.

This alternative is considered to determine if the original model, (21), is overdifferencing the time series and, if so, the impact on its forecast.

While both models, (21) and (22a), do not indicate inadequacy from the fitting point of view, their forecasts are markedly different. Figures 7 and 8 give an impression of the forecast performance for these models. It is clear that for model (22a) the seasonality in the forecast damps out quickly. This implies considerable underforecasting of peak months—and consequently of ABS values—for the CCS/M data.

The two other alternatives for Series B were suggested from its graph (Fig. 2). The identified Box-Jenkins model for this series, after differencing for seasonality, represents a stationary process, despite
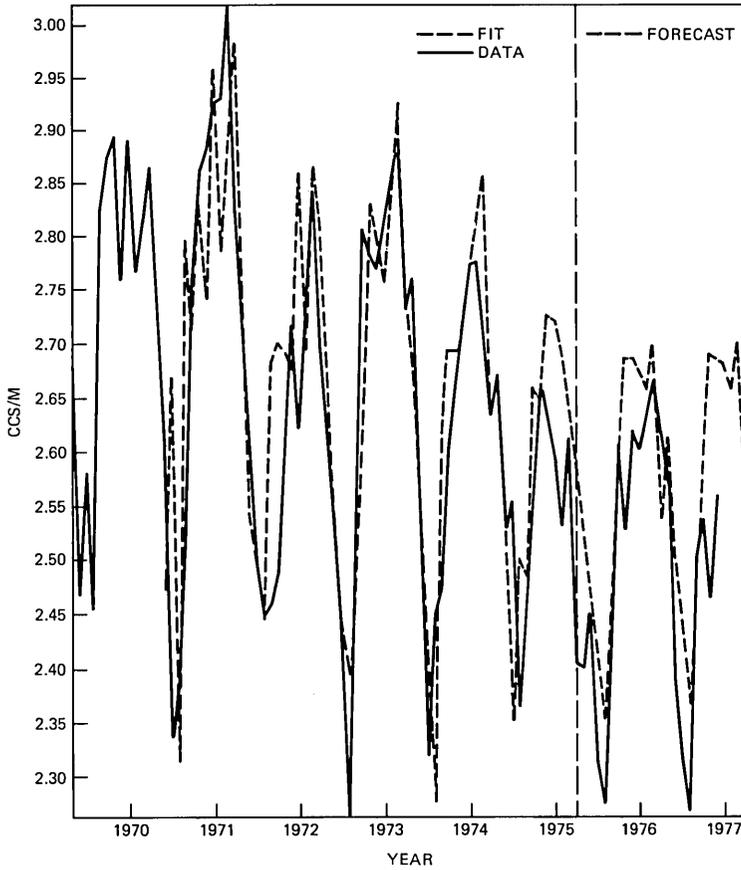
Fig. 7—Data, fitting, and forecast, Series B, model (21).

an apparent linear downward trend. This apparent trend is modeled as short-term fluctuation and is represented by ARMA(1, 1) parameters. If this trend is to be extended into the future, one must consider other representations. For example,

$$(1 - B)(1 - B^{12})z_t = (1 - \theta B)(1 - \Theta B^{12})a_t, \qquad (22b)$$

where

$$\hat{\theta} = 0.5, \qquad \hat{\Theta} = 0.53, \qquad \sigma_a = 0.12,$$

and

$$(1 - B^{12})z_t = \theta_0 + (1 - \theta B)(1 - \Theta B^{12})a_t, \qquad (22c)$$

where

$$\hat{\theta}_0 = -0.04, \qquad \hat{\theta} = -0.32, \qquad \hat{\Theta} = 0.56, \qquad \sigma_a = 0.12.$$
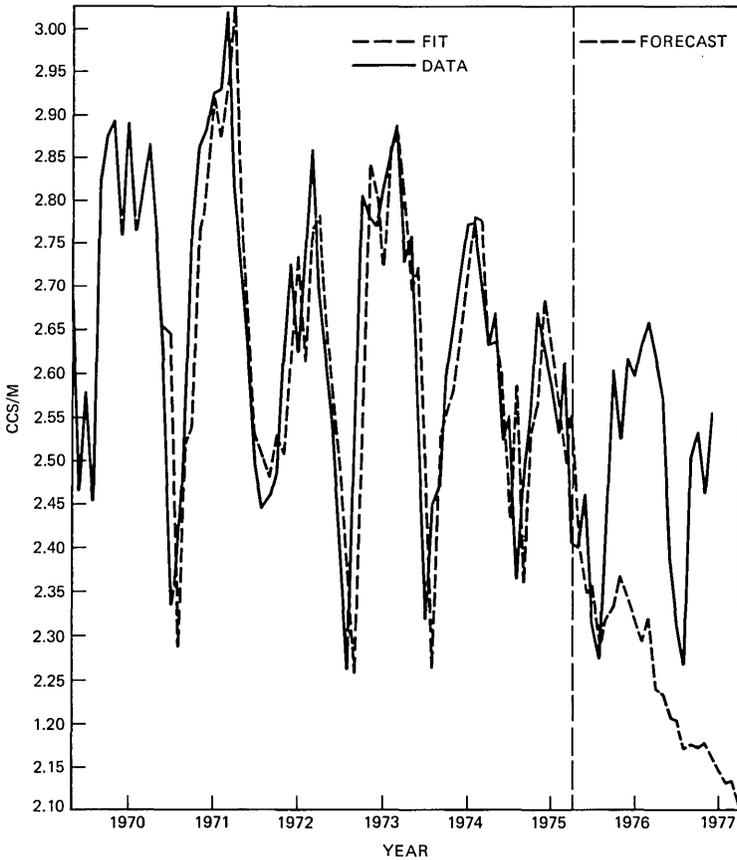
Fig. 8—Data, fitting, and forecast, Series B, model (22a).

Model (22b) simply suggests an exponential smoothing for the year-to-year variation followed by another exponential smoothing for the month-to-month variation. Model (22c) includes a deterministic constant to account for a slope, considered a useful alternative to differencing. It is more restrictive, however, in the sense that it assumes a common slope for the 12 parallel lines (one for each month) connecting different years.

Based on the autocorrelation function of the residuals and on the goodness of fit tests, models (22b) and (22c) may represent Series B. From the forecast point of view, however, the trend effect was much exaggerated by model (22b) and somewhat exaggerated by model (22c) (Figs. 9 and 10). The original model did react to the apparent trend [through the AR(1) term], but was more conservative in its extrapolation (Fig. 7). Because of our experience with the CCS/M data, we favor the original model. We have learned from the behavior of many
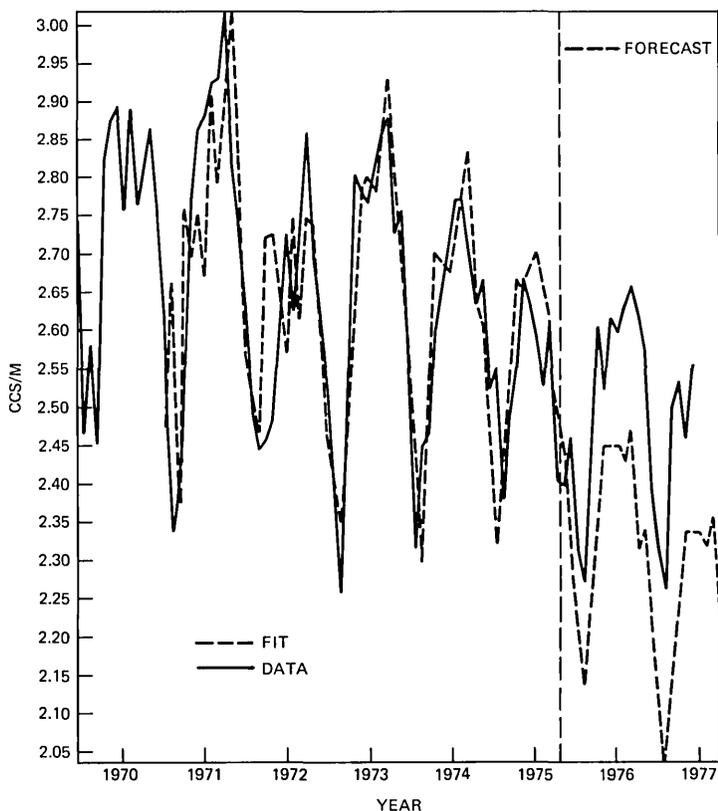
Fig. 9—Data, fitting, and forecast, Series B, model (22b).

time series for CCS/M that a turning point will eventually occur. We emphasize here that one must be careful in allowing for a linear trend representation in a model unless one is certain that this behavior has a physical interpretation. An apparent linear trend may arise simply from the repeated summation of independent disturbances.

3.3 Application of state space approach

Akaike et al. implemented the state space method into a computer package.[8] The original package appeared in 1974 (known as TIMSAC-74) and has been revised several times since. The version we employed here was revised in March 1977.* More recently, the method was implemented by SAS in a procedure called STATESPACE.[15] It is

---

*A copy of TIMSAC programs was obtained from the Mathematical Sciences Division of the University of Tulsa, Oklahoma. The university distributes these programs for the authors.
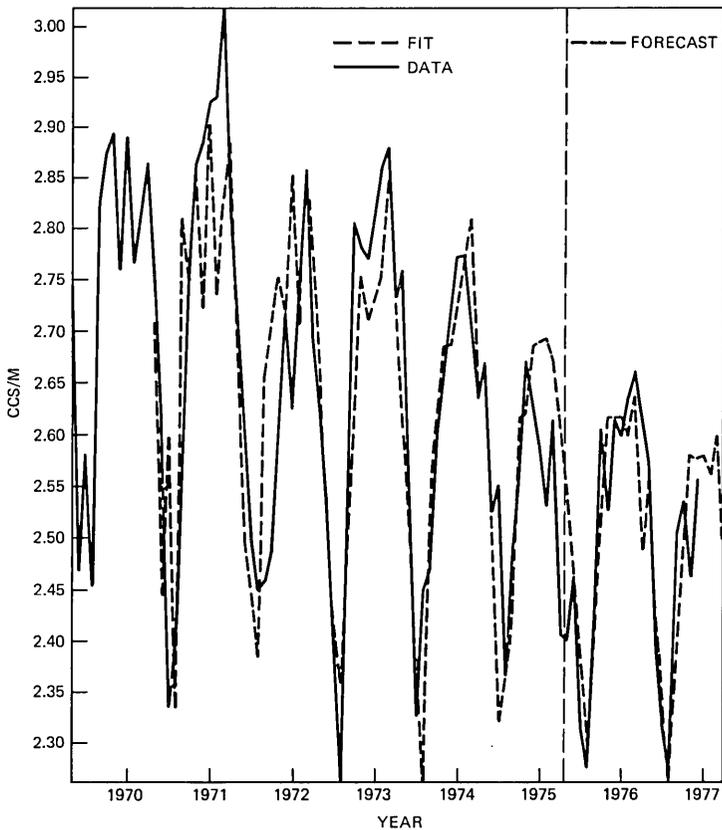
Fig. 10—Data, fitting, and forecast, Series B, model (22c).

assumed that models are limited to the block identifiability condition.[7] In this regard, an identified form using this procedure is equivalent to an ARMA form of order $(p, p - 1)$.

Using TIMSAC programs, models were identified for 15 of the 35 series that were previously identified by the Box-Jenkins approach. While preliminary analysis of these data (Section 3.1) indicated that for the differenced series $\{x_t\}$ there is a dependency between x_t and x_{t-12} (-0.47 for Series A), which suggests a moving average term of order 12, none of the automatically identified models included this term.

We also noted that the numerical procedure for the maximum likelihood computation did not converge easily. It required a large number of iterations. The problem was due to the improper initial values—estimated by the canonical correlation analysis—of model parameters. In many cases, these values specified noninvertible

models. For six of the series, numerical breakdowns were encountered as a result of the singularity of the Hessian of the mean log-likelihood function. At the breakdown points, the MAICE model for all series was noise, ARMA(0, 0). As for the other nine series, six were identified as noise [ARMA(0, 0)], two as MA(1) and one as an AR(1). One of the series that was identified as noise is the differenced data for Series A. Neither the sample autocorrelation function of this data (Fig. 4) nor the goodness-of-fit measure is compatible with such representation. We note here that the identified model from the canonical correlation analysis for this series is an AR(1) with $\phi = 0.55$ and AIC $= -314$. The competing model ARMA(0, 0) was preferred because it has AIC $= -316$.

3.4 Application of autoregressive approach

Until now, we have been discussing model identification and forecast performance while disregarding the data problems of gross outliers and large jumps. Since these problems can be held accountable for forecast inaccuracies, the idea of eliminating them from the data prior to modeling seemed plausible.

Two procedures for removing jumps were implemented. The first is based on a moving median technique to locate discontinuities in the data; and the other uses a robust regression method to remove the discontinuities found.[16] The latter procedure replaces the time series points before a jump by new values at the most recent level of the process. It also substitutes suspected single outlying points by more probable values.

After modifying the 35 investigated time series using these algorithms, we applied the autoregressive spectrum estimation approach to the cleaned data. A detailed description of this approach and its implementation is given in Ref. 15. Since the considered series are fairly short (less than 250 points),[12] their forecasts were generated from parametric autoregressive models, with their parameters estimated in the frequency domain. The procedure applied here uses the Fast Fourier Transform (FFT) algorithm to estimate the spectrum. The bias that may occur in these estimates due to the limited length of the series was handled by using Tukey's "twicing."[17]

The ABS forecast accuracies obtained using this approach were not significantly different from those found by the random-walk method even for the 1 year ahead, in spite of the screening of the data, the removal of jumps, and the robust estimation of models parameters.

IV. SIMULATIONS

Our objectives in carrying out the simulation study are first to determine whether any of the examined methodology can automati-

cally identify the actual process in an ideal environment. A second objective is to understand the relationship between over- or underparameterization of a stationary process and model adequacy. A third objective is to assess the forecast performance of the various methods and determine if a use of the "best" method is worthwhile. The three simulated processes are

$$\text{Process I} \quad (1 - 0.6B)(1 - B^{12})z_t$$
$$= (1 - 0.5B^{12})a_t, \quad \sigma_a = 0.12$$
$$\text{Process II} \quad (1 - 0.8B)(1 - B^{12})z_t$$
$$= (1 - 0.3B)(1 - 0.5B^{12})a_t, \quad \sigma_a = 0.12$$
$$\text{Process III} \quad (1 - B^{12})z_t = a_t, \quad \sigma_a = 0.25.$$

Processes I and II correspond to the two previously identified models for Series A and Series B, respectively. Process III is a seasonal random-walk process (where B is replaced by B^{12}), which was also observed in many of the time-series data.

For each of the three processes, 100 realizations were generated with 252 observations in each. The white noise $\{a_t\}$ was generated from a normal random variate with mean zero and variance equal to those estimated from the actual data. The starting values of these realizations $(z_1, z_2, \cdots, z_{13})$ were taken from the actual CCS/M time series. In all realizations, the first 96 points were discarded, the next 120 points were used for fitting, and the last 36 observations were saved for ex-post analysis. In the following we discuss the simulation results in terms of our three objectives.

4.1 Identification

In this discussion, we exclude the Box-Jenkins approach since its identification stage is based on judgment, while the processes to be identified are known a priori. The state space procedure and the autoregression spectrum estimation procedure were applied to the simulated realizations after differencing at lag 12. Table I gives summary statistics on model identification of time series generated from Processes I and II using the state space procedure. The tabulated statistics are the number of realizations an ARMA(p, q) model is identified, where p and q are given at certain values and/or specified ranges. From these results we note that the procedure did not recognize the distant terms (coefficients of a_{t-12} and a_{t-13}) except for one series. For this series, however, the identified model was ARMA (13, 12), with nonzero estimates for all 25 coefficients. Further inspection of Table I also indicates that the state space procedures did not identify correct parsimonious ARMA models.

We must also indicate that, as with actual data, we encountered the same computational difficulties. These difficulties were, in major part,

Table I—Frequency distribution of the order of the fitted process using Akaike state space procedures

Fitted Order	Frequency	
	Process I	Process II
$p = q = 0$	19	14
$p = 1; q = 0$	16	9
$p = 1; q = 1$	10	18
$p = 0; q = 1$	0	0
$p = 2; q = 1$	15	8
$p = 2; q = 2$	2	8
$p = 0,1; q = 2$	2	2
$3 \leq \max(p, q) \leq 5$	26	29
$6 \leq \max(p, q) \leq 11$	6	8
$\max(p, q) = 12$	3	4
$\max(p, q) = 13$	1	0
Total	100	100

connected with the numerical approximations for the maximization of the likelihood function. The singularity of the Hessian was due to improper assumed values for model parameters. A large number of iterations was required for convergence in most cases.

On the autoregressive spectrum estimation procedure, Table II presents the goodness of fit of its models to the simulated realizations. Since the procedure assumes pure autoregressive models, we will not discuss whether the assumed order over- or underparameterizes a given series. Instead, we compare the estimated value for the standard error of the residual ($\hat{\sigma}_{AS}$ in Table II) for each of the simulated series with the true value of its process. An average estimate of 0.13 (compared to $\sigma_e = 0.12$) for the standard error for Processes I and II realizations indicates that the assumed models may account for most of the predictive information in the simulated data.

4.2 Effect of overparameterizations

To demonstrate the effect of overparameterization, we used (19), denoted hereafter by S, to model each of the 300 simulated series. This model specification was developed as a result of Box-Jenkins analysis of actual CCS/M time series. Table II tabulates the estimated values for the associated parameters ($\hat{\sigma}_S$, $\hat{\phi}_1$, $\hat{\theta}_1$, $\hat{\Theta}_{12}$). Inspection of the estimated standard error ($\hat{\sigma}_S$) indicates that these models can account for the predictive variabilities of the time series (average $\hat{\sigma}_e = 0.12$ for Processes I and II). To compare the estimated values of model coefficients to their true values, however, one must understand the redundancy between autoregressive and moving average terms in a mixed model. For simplicity, consider the form

$$(1 - \phi B)y_t = (1 - \theta B)a_t. \tag{23}$$

Table IIa—Standard errors for AS, RW, and S models and estimates of the S model parameters, with $\sigma_e = 0.12$—Process I realizations

Series	$\hat{\sigma}_{AS}$	$\hat{\sigma}_{RW}$	$\hat{\sigma}_S$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\Theta}$	Series	$\hat{\sigma}_{AS}$	$\hat{\sigma}_{RW}$	$\hat{\sigma}_S$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\Theta}$
A001	0.132	0.158	0.121	0.59	0.00	0.52	A051	0.124	0.168	0.126	0.65	0.07	0.41
02	0.120	0.163	0.114	0.51	−0.29	0.40	052	0.128	0.156	0.122	0.60	0.05	0.41
03	0.126	0.154	0.124	0.39	−0.20	0.32	053	0.117	0.149	0.104	0.53	−0.06	0.59
04	0.123	0.153	0.107	0.53	−0.13	0.55	054	0.117	0.159	0.113	0.68	0.15	0.47
05	0.123	0.189	0.122	0.74	−0.03	0.34	055	0.123	0.167	0.121	0.57	−0.09	0.48
06	0.138	0.171	0.135	0.54	−0.06	0.49	056	0.143	0.196	0.131	0.65	−0.15	0.51
07	0.124	0.143	0.120	0.58	0.05	0.43	057	0.145	0.171	0.124	0.59	0.04	0.79
08	0.137	0.180	0.126	0.84	0.31	0.64	058	0.130	0.161	0.131	0.52	0.01	0.41
09	0.137	0.179	0.131	0.63	0.00	0.52	059	0.143	0.188	0.136	0.65	−0.04	0.54
010	0.133	0.169	0.129	0.71	0.14	0.39	060	0.126	0.159	0.116	0.69	0.15	0.54
011	0.144	0.173	0.128	0.63	0.07	0.59	061	0.145	0.159	0.129	0.63	0.12	0.64
012	0.120	0.165	0.116	0.68	0.06	0.42	062	0.135	0.183	0.127	0.69	0.14	0.48
013	0.142	0.193	0.130	0.68	0.08	0.60	063	0.111	0.155	0.115	0.57	−0.14	0.18
014	0.128	0.175	0.132	0.54	−0.04	0.47	064	0.139	0.173	0.130	0.73	0.25	0.53
015	0.118	0.142	0.113	0.33	−0.25	0.43	065	0.148	0.177	0.128	0.48	−0.13	0.55
016	0.131	0.159	0.117	0.55	−0.08	0.57	066	0.138	0.175	0.125	0.62	0.12	0.57
017	0.120	0.171	0.119	0.70	−0.03	0.42	067	0.142	0.179	0.127	0.60	−0.06	0.59
018	0.136	0.165	0.134	0.47	−0.06	0.54	068	0.138	0.178	0.117	0.48	−0.31	0.64
019	0.123	0.173	0.119	0.74	0.16	0.47	069	0.140	0.204	0.137	0.63	−0.04	0.50
020	0.126	0.166	0.124	0.63	0.09	0.48	070	0.123	0.145	0.121	0.55	0.04	0.38
021	0.131	0.171	0.126	0.57	−0.11	0.39	071	0.136	0.177	0.130	0.72	0.23	0.43
022	0.132	0.167	0.125	0.63	0.04	0.45	072	0.130	0.156	0.129	0.63	0.20	0.45
023	0.125	0.171	0.122	0.72	0.35	0.44	073	0.126	0.172	0.124	0.77	0.14	0.38
024	0.138	0.174	0.124	0.79	0.29	0.56	074	0.126	0.143	0.120	0.68	0.20	0.44
025	0.130	0.163	0.125	0.43	−0.23	0.39	075	0.116	0.165	0.114	0.48	−0.36	0.45
026	0.103	0.154	0.113	0.69	0.13	0.29	076	0.146	0.193	0.128	0.83	0.22	0.67
027	0.127	0.182	0.120	0.68	−0.02	0.53	077	0.137	0.179	0.133	0.61	−0.04	0.55
028	0.121	0.151	0.120	0.76	0.31	0.37	078	0.122	0.154	0.121	0.38	−0.20	0.51
029	0.120	0.134	0.113	0.38	−0.15	0.36	079	0.126	0.150	0.114	0.57	0.08	0.51
030	0.127	0.158	0.121	0.43	−0.06	0.52	080	0.132	0.169	0.121	0.63	−0.08	0.49
031	0.115	0.151	0.115	0.66	−0.09	0.32	081	0.132	0.171	0.129	0.70	0.11	0.47
032	0.124	0.163	0.118	0.70	0.01	0.52	082	0.127	0.170	0.116	0.66	0.03	0.43
033	0.133	0.182	0.133	0.81	0.32	0.31	083	0.140	0.170	0.128	0.39	−0.17	0.51
034	0.139	0.179	0.134	0.58	0.01	0.57	084	0.116	0.139	0.116	0.48	−0.11	0.40
035	0.147	0.191	0.130	0.65	−0.01	0.54	085	0.127	0.174	0.117	0.61	−0.02	0.63
036	0.122	0.148	0.110	0.58	−0.01	0.45	086	0.115	0.167	0.119	0.50	−0.30	0.35
037	0.125	0.153	0.112	0.44	−0.15	0.47	087	0.125	0.154	0.115	0.25	−0.48	0.54
038	0.123	0.150	0.118	0.32	−0.20	0.46	088	0.140	0.167	0.123	0.76	0.31	0.51
039	0.114	0.140	0.120	0.50	0.16	0.41	089	0.125	0.168	0.110	0.75	0.04	0.51
040	0.136	0.188	0.123	0.70	0.02	0.68	090	0.111	0.133	0.106	0.69	0.26	0.38
041	0.132	0.196	0.116	0.66	−0.09	0.51	091	0.121	0.174	0.123	0.73	0.13	0.24
042	0.113	0.140	0.115	0.46	−0.14	0.29	092	0.128	0.160	0.114	0.74	0.19	0.63
043	0.131	0.185	0.129	0.77	0.20	0.38	093	0.133	0.175	0.129	0.62	−0.08	0.37
044	0.146	0.185	0.124	0.57	−0.11	0.67	094	0.122	0.164	0.115	0.61	−0.02	0.55
045	0.136	0.171	0.129	0.50	−0.02	0.60	095	0.130	0.174	0.117	0.72	0.11	0.51
046	0.132	0.166	0.117	0.62	−0.01	0.61	096	0.137	0.187	0.125	0.64	−0.01	0.46
047	0.130	0.167	0.126	0.70	0.17	0.48	097	0.121	0.162	0.113	0.61	−0.12	0.48
048	0.138	0.172	0.130	0.43	−0.15	0.46	098	0.140	0.187	0.128	0.60	−0.07	0.55
049	0.135	0.188	0.129	0.74	0.11	0.46	099	0.137	0.174	0.134	0.50	−0.01	0.46
050	0.132	0.173	0.119	0.60	0.05	0.57	100	0.119	0.157	0.116	0.51	−0.10	0.44

This can be written as

$$y_t = [1 + (\phi - \theta)B + (\phi^2 - \theta\phi)B^2 + (\phi^3 - \theta\phi^2)B^3 + \cdots]a_t. \quad (24)$$

For $\phi = \theta = 0$, (23) represents a noise process. But when $\phi = \theta \neq 0$ it still represents a noise process, since all terms of the polynomial (24) cancel out.

For the model specification (19), similar redundancy exists between the AR(1) coefficient, ϕ, and the MA(1) coefficient, θ, and to a lesser degree it exists between ϕ and the moving average coefficients at lags 12 and 13. The effect of these redundancies is particularly clear from the estimated coefficients of Process III realizations ($\hat{\phi} = \hat{\theta} \neq 0$).

Series	$\hat{\sigma}_{AS}$	$\hat{\sigma}_{RW}$	$\hat{\sigma}_S$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\Theta}$	Series	$\hat{\sigma}_{AS}$	$\hat{\sigma}_{RW}$	$\hat{\sigma}_S$	$\hat{\phi}$	$\hat{\theta}$	Θ
B001	0.132	0.156	0.122	0.77	0.28	0.52	B051	0.124	0.175	0.126	0.80	0.28	0.42
02	0.120	0.165	0.114	0.73	0.06	0.43	052	0.128	0.162	0.122	0.83	0.38	0.41
03	0.126	0.159	0.125	0.76	0.32	0.33	053	0.116	0.149	0.104	0.65	0.15	0.59
04	0.123	0.155	0.107	0.81	0.29	0.55	054	0.118	0.165	0.113	0.85	0.42	0.48
05	0.124	0.197	0.122	0.85	0.18	0.35	055	0.124	0.171	0.120	0.87	0.40	0.52
06	0.138	0.175	0.134	0.70	0.19	0.51	056	0.143	0.207	0.131	0.84	0.16	0.53
07	0.124	0.145	0.120	0.85	0.43	0.43	057	0.145	0.175	0.124	0.86	0.44	0.78
08	0.137	0.190	0.127	0.92	0.46	0.65	058	0.130	0.161	0.131	0.76	0.35	0.42
09	0.137	0.178	0.131	0.81	0.28	0.52	059	0.143	0.197	0.136	0.87	0.30	0.56
10	0.134	0.177	0.129	0.90	0.46	0.42	060	0.126	0.169	0.115	0.85	0.40	0.54
011	0.144	0.177	0.127	0.91	0.52	0.62	061	0.146	0.162	0.128	0.90	0.54	0.47
012	0.120	0.172	0.116	0.79	0.24	0.44	062	0.135	0.191	0.127	0.81	0.33	0.49
013	0.142	0.201	0.130	0.83	0.31	0.61	063	0.110	0.154	0.115	0.71	0.11	0.19
014	0.129	0.178	0.132	0.71	0.22	0.47	064	0.139	0.176	0.131	0.86	0.43	0.52
015	0.117	0.135	0.111	0.44	−0.01	0.44	065	0.148	0.172	0.128	0.73	0.22	0.56
016	0.130	0.157	0.117	0.81	0.30	0.57	066	0.138	0.179	0.126	0.75	0.33	0.60
017	0.120	0.157	0.119	0.85	0.22	0.43	067	0.142	0.117	0.127	0.73	0.16	0.57
018	0.136	0.160	0.133	0.60	0.16	0.55	068	0.138	0.176	0.117	0.65	−0.07	0.64
019	0.123	0.184	0.120	0.84	0.33	0.48	069	0.139	0.207	0.137	0.76	0.17	0.50
020	0.126	0.170	0.124	0.77	0.32	0.48	070	0.123	0.145	0.121	0.73	0.30	0.38
021	0.131	0.168	0.125	0.71	0.15	0.39	071	0.135	0.182	0.131	0.85	0.41	0.43
022	0.132	0.169	0.125	0.81	0.31	0.45	072	0.130	0.159	0.129	0.84	0.48	0.47
023	0.125	0.149	0.120	0.93	0.66	0.46	073	0.124	0.184	0.124	0.88	0.34	0.38
024	0.138	0.189	0.125	0.90	0.45	0.53	074	0.127	0.150	0.120	0.87	0.47	0.44
025	0.130	0.159	0.125	0.62	0.07	0.39	075	0.116	0.167	0.115	0.94	−0.09	0.44
026	0.104	0.162	0.115	0.83	0.31	0.29	076	0.147	0.208	0.127	0.94	0.42	0.71
027	0.127	0.190	0.119	0.88	0.34	0.55	077	0.137	0.189	0.133	0.84	0.31	0.55
028	0.121	0.160	0.120	0.90	0.51	0.39	078	0.122	0.152	0.120	0.80	0.41	0.51
029	0.120	0.135	0.113	0.77	0.37	0.35	079	0.127	0.155	0.113	0.80	0.40	0.52
030	0.127	0.161	0.123	0.62	0.20	0.51	080	0.132	0.170	0.121	0.78	0.16	0.49
031	0.115	0.163	0.115	0.85	0.40	0.33	081	0.132	0.178	0.129	0.80	0.30	0.49
032	0.124	0.166	0.118	0.86	0.28	0.52	082	0.127	0.177	0.117	0.81	0.26	0.43
033	0.134	0.204	0.132	0.92	0.49	0.35	083	0.139	0.168	0.128	0.71	0.29	0.50
034	0.138	0.179	0.133	0.88	0.46	0.61	084	0.118	0.138	0.116	0.70	0.22	0.40
035	0.147	0.201	0.130	0.79	0.23	0.55	085	0.126	0.173	0.117	0.75	0.21	0.63
036	0.122	0.143	0.109	0.67	0.18	0.45	086	0.114	0.171	0.119	0.72	0.08	0.39
037	0.124	0.148	0.112	0.61	0.14	0.48	087	0.124	0.145	0.116	0.42	−0.17	0.51
038	0.122	0.141	0.118	0.53	0.11	0.43	088	0.140	0.171	0.125	0.85	0.45	0.50
039	0.114	0.139	0.120	0.77	0.53	0.40	089	0.125	0.177	0.111	0.87	0.25	0.51
040	0.136	0.198	0.123	0.85	0.26	0.68	090	0.111	0.143	0.106	0.83	0.45	0.37
041	0.132	0.204	0.116	0.81	0.15	0.51	091	0.121	0.177	0.122	0.81	0.31	0.26
042	0.113	0.138	0.115	0.66	0.14	0.28	092	0.128	0.166	0.114	0.80	0.38	0.63
043	0.131	0.199	0.123	0.87	0.38	0.39	093	0.133	0.175	0.129	0.74	0.12	0.36
044	0.146	0.188	0.125	0.81	0.25	0.67	094	0.121	0.171	0.115	0.75	0.21	0.56
045	0.136	0.168	0.128	0.63	0.19	0.60	095	0.130	0.179	0.116	0.88	0.40	0.53
046	0.132	0.168	0.117	0.76	0.23	0.60	096	0.137	0.187	0.124	0.74	0.19	0.46
047	0.131	0.176	0.127	0.83	0.36	0.50	097	0.121	0.159	0.113	0.71	0.08	0.46
048	0.138	0.174	0.131	0.83	0.44	0.48	098	0.140	0.189	0.128	0.76	0.19	0.55
049	0.134	0.200	0.129	0.89	0.35	0.47	099	0.135	0.173	0.133	0.65	0.26	0.46
050	0.132	0.181	0.119	0.78	0.32	0.57	100	0.120	0.162	0.116	0.82	0.36	0.43

Based on these results, we conclude that (19) can represent the
generated processes adequately. In other words, overparameterization
by a moving average term at lag 1, as in Process I, and furthermore
by an autoregressive term at lag 1 and moving average terms at lags
12 and 13, as in Process III, did not affect the standard error of the
residuals significantly.

4.3 Forecast accuracy

For this evaluation, we consider forecasts generated by: (a) the
model specification in (19) (S), (b) the autoregressive spectral esti-

Table IIc—Standard errors for AS, RW, and S models and estimates of the S model parameters, with $\sigma_e = 0.25$—Process III realizations

Series	$\hat{\sigma}_{AS}$	$\hat{\sigma}_{RW}$	$\hat{\sigma}_S$	$\hat{\phi}$	$\hat{\theta}$	$\bar{\Theta}$	Series	$\hat{\sigma}_{AS}$	$\hat{\sigma}_{RW}$	$\hat{\sigma}_S$	$\hat{\phi}$	$\hat{\theta}$	$\bar{\Theta}$
C001	0.239	0.251	0.242	−0.63	−0.75	0.08	C051	0.236	0.251	0.254	0.11	0.11	−0.04
02	0.226	0.229	0.226	−0.25	−0.48	−0.07	052	0.244	0.253	0.252	−0.24	−0.22	−0.07
03	0.250	0.263	0.258	0.87	0.94	−0.17	053	0.207	0.209	0.208	−0.80	−0.79	0.17
04	0.227	0.223					054	0.215	0.226	0.227	0.02	0.09	0.06
05	0.242	0.256	0.252	0.66	0.51	−0.17	055	0.229	0.246	0.249	0.08	0.00	−0.02
06	0.263	0.275	0.275	0.89	0.98	−0.08	056	0.260	0.270	0.270	−0.05	−0.25	0.00
07	0.241	0.245	0.248	−0.63	−0.59	−0.06	057	0.261	0.264	0.266	−0.30	−0.30	0.16
08	0.254	0.261	0.264	−0.75	−0.67	−0.01	058	0.242	0.261	0.263	0.00	0.04	−0.00
09	0.255	0.260	0.262	0.90	0.87	0.10	059	0.269	0.273	0.276	0.00	−0.12	−0.01
010	0.260	0.264	0.266	−0.68	−0.65	−0.16	060	0.225	0.231	0.234	0.45	0.44	0.09
011	0.255	0.257	0.259	−0.56	−0.52	0.13	061	0.267	0.266	0.269	0.23	0.30	0.01
012	0.239	0.236	0.239	0.49	0.45	−0.10	062	0.248	0.254	0.258	0.32	0.35	0.03
013	0.253	0.265	0.267	−0.46	−0.40	0.14	063	0.232	0.238	0.233	−0.07	−0.20	−0.28
014	0.239	0.255	0.257	−0.65	−0.57	−0.02	064	0.253	0.266	0.268	0.78	0.79	0.01
015	0.223	0.234	0.230	0.09	0.06	0.05	065	0.258	0.262	0.257	0.57	0.65	0.18
016	0.239	0.236	0.237	−0.56	−0.63	0.13	066	0.246	0.256	0.252	−0.76	−0.61	0.13
017	0.234	0.240	0.240	0.60	0.45	−0.05	067	0.254	0.255	0.258	0.43	0.36	0.13
018	0.257	0.277	0.270	−0.88	−0.92	−0.04	068	0.240	0.248	0.244	−0.07	−0.29	0.19
019	0.226	0.245	0.243	0.77	0.77	−0.00	069	0.267	0.274	0.277	0.66	0.58	−0.09
020	0.242	0.248	0.251	−0.86	−0.81	−0.07	070	0.234	0.249				
021	0.256	0.257					071	0.257	0.265	0.266	−0.67	−0.53	−0.04
022	0.247	0.254	0.257	0.40	0.36	−0.04	072	0.241	0.257	0.254	−0.69	−0.50	0.00
023	0.236	0.252	0.252	−0.64	−0.49	0.01	073	0.244	0.253	0.254	0.82	0.75	−0.13
024	0.261	0.257	0.261	−0.24	−0.17	−0.03	074	0.241	0.244	0.244	−0.45	−0.31	−0.01
025	0.256	0.263	0.265	0.19	0.12	−0.08	075	0.220	0.246	0.239	−0.08	−0.37	−0.06
026	0.209	0.236	0.227	−0.46	−0.38	−0.16	076	0.267	0.267	0.264	0.91	0.86	0.25
027	0.228	0.239	0.240	−0.17	−0.29	0.09	077	0.256	0.267	0.268	0.84	0.82	0.01
028	0.232	0.243	0.242	−0.60	−0.43	−0.05	078	0.225	0.245	0.245	0.47	0.57	0.01
029	0.234	0.237	0.239	−0.45	−0.41	−0.07	079	0.233	0.234	0.233	0.28	0.38	−0.00
030	0.239	0.247	0.243	−0.24	−0.17	0.02	080	0.240	0.242	0.239	−0.22	−0.40	0.08
031	0.239	0.241	0.238	0.04	0.01	−0.18	081	0.251	0.263	0.267	0.82	0.80	−0.09
032	0.235	0.243	0.246	0.31	0.23	−0.01	082	0.240	0.244	0.249	0.09	0.08	−0.01
033	0.253	0.263	0.264	0.92	0.86	−0.04	083	0.266	0.266	0.270	−0.73	−0.78	−0.05
034	0.246	0.267	0.267	0.49	0.52	0.09	084	0.231	0.232	0.235	0.76	0.83	−0.04
035	0.249	0.259	0.259	0.44	0.37	0.18	085	0.222	0.239	0.241	−0.09	−0.08	0.16
036	0.227	0.238	0.234	0.09	0.07	−0.02	086	0.239	0.242	0.239	0.24	0.08	−0.06
037	0.227	0.229	0.232	−0.68	−0.68	0.05	087	0.230	0.242	0.239	−0.62	−0.82	0.02
038	0.231	0.242	0.241	0.72	0.80	0.02	088	0.257	0.260	0.258	−0.29	−0.10	0.06
039	0.216	0.244	0.237	0.32	0.53	−0.09	089	0.234	0.232	0.232	0.70	0.56	0.03
040	0.247	0.253	0.254	0.81	0.73	0.09	090	0.220	0.220	0.219	0.03	0.17	−0.12
041	0.234	0.240	0.240	0.40	0.25	0.07	091	0.243	0.245	0.238	0.76	0.79	−0.14
042	0.228	0.231	0.231	0.20	0.18	−0.20	092	0.228	0.231	0.229	−0.85	−0.72	0.13
043	0.267	0.275	0.276	0.42	0.42	−0.13	093	0.251	0.260	0.262	0.38	0.28	−0.09
044	0.254	0.251	0.245	−0.91	−0.99	0.21	094	0.227	0.238	0.238	0.80	0.82	0.01
045	0.239	0.262	0.262	0.32	0.39	0.16	095	0.231	0.235	0.226	1.00	1.05	0.11
046	0.241	0.234	0.237	−0.75	−0.81	0.10	096	0.260	0.260	0.264	0.14	0.10	−0.03
047	0.246	0.258	0.261	−0.17	−0.07	0.01	097	0.227	0.230	0.232	0.18	0.03	−0.03
048	0.254	0.259	0.255	−0.96	−1.02	−0.03	098	0.258	0.260	0.264	−0.09	−0.14	0.05
049	0.247	0.266	0.265	0.92	0.88	−0.02	099	0.255	0.266	0.270	−0.56	−0.53	−0.01
050	0.246	0.245	0.249	0.08	0.09	0.02	100	0.231	0.233	0.236	−0.09	−0.10	−0.06

mation method (AS) and, (c) the random-walk model (RW) for the simulated series.

While minimization of the mean square errors during the model-fitting phase guarantees minimum mean square errors for the future of the simulated data (post-sample phase), we consider ex-post analysis to assess the effect of conversion of the monthly forecast to ABS values on forecast accuracy.

The ABS quantity is made up of the average of the three highest months in a year. The busy season months can vary from one year to the next. A nonbusy season month of one year may become a busy season month of the next year as a result of some added white noise.
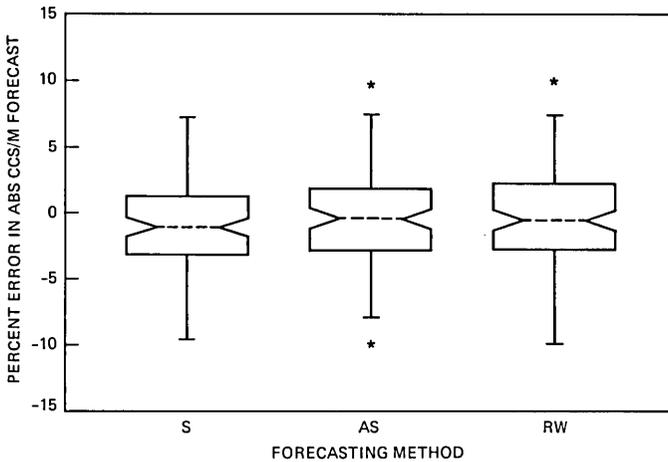
Fig. 11—Notched box plot for one-year-ahead forecast error—Process I realizations.
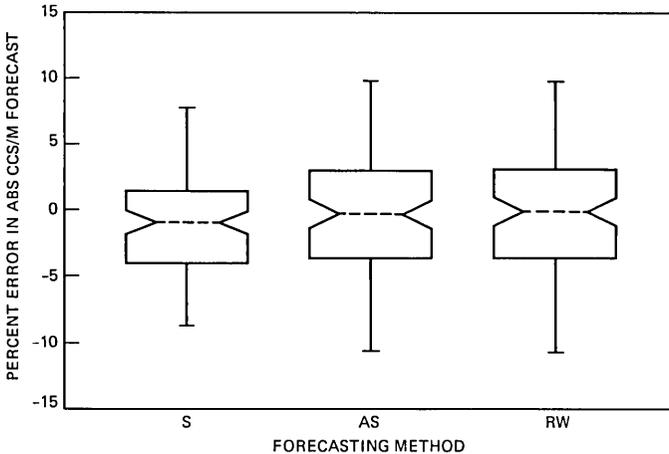

Fig. 12—Notched box plot for two-year-ahead forecast error—Process I realizations.

Since the effect of this aggregation is not a simple analytical problem, we resort to ex-post analysis of the simulated realizations for explanation.

Figures 11 through 16 present these comparisons in the form of box plots* of the percent forecast accuracy[18] for series that are grouped by

---

*In a box plot the middle line is the median and the upper and lower lines of the box are the upper and lower quartiles, respectively. The box whiskers are drawn out to the nearest values within 1.5 quartiles. Points outside the whiskers are considered outliers and are indicated by asterisks. The notches in the side of a box represents a rough 95-percent confidence interval for the median.
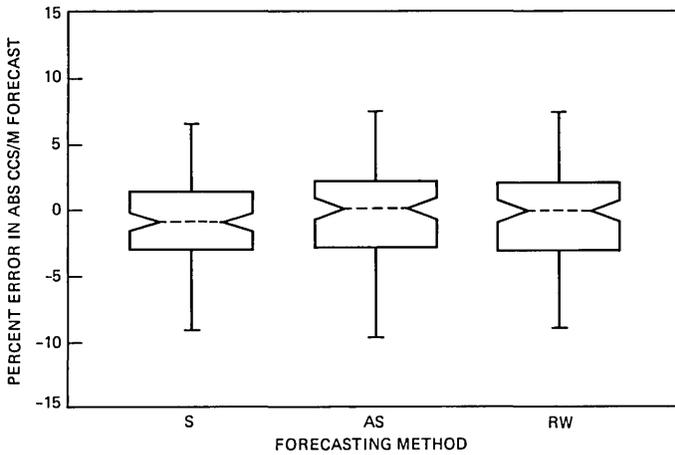
Fig. 13—Notched box plot for one-year-ahead forecast error—Process II realizations.
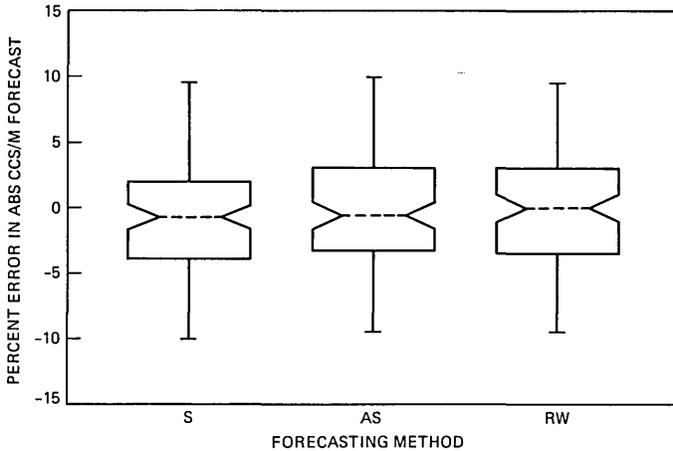
Fig. 14—Notched box plot for two-year-ahead forecast error—Process II realizations.

their process type for one and two years ahead. Based on these graphs, we draw a general conclusion that there are no significant differences in the forecasting accuracies among the three methods.

We did notice, however, that an optimal model that represents the actual process, and is expected to provide best forecasts, results in low prediction of the ABS values. More specifically, the S form is the model specification for Process II. Thus, it is expected to generate "best" forecasts. Figures 13 and 14, however, indicate a bias towards underforecast of ABS values for optimal models. The estimated medians of the percent errors for one- and two-years-ahead forecasts are
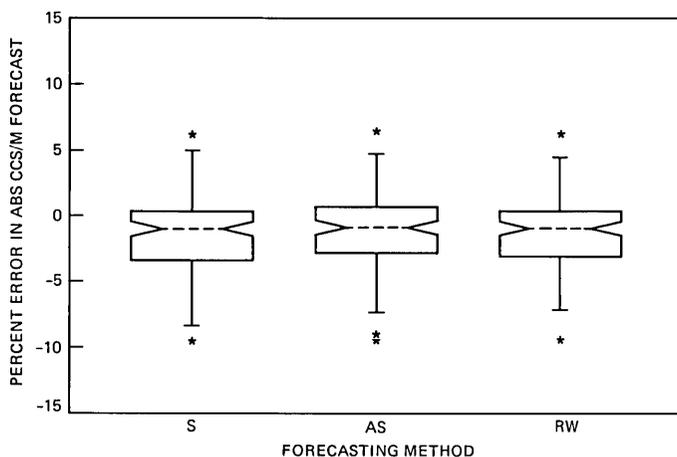
Fig. 15—Notched box plot for one-year-ahead forecast error—Process III realizations.
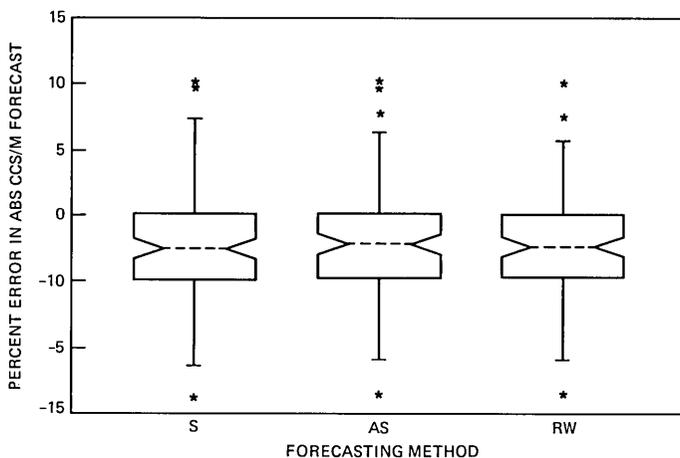
Fig. 16—Notched box plot for two-year-ahead forecast error—Process III realizations.

−1.0 and −0.7, respectively. This observation is also repeated for Process III. The random-walk model underforecasts ABS values for realizations generated by the random-walk process. Figures 15 and 16 demonstrate this observation. The estimated medians for one and two years ahead were −1.0 and −2.4, respectively (notice that here $\sigma_a = 0.25$). In this regard, the ABS forecast is not optimal.

In the following we give a heuristic argument on the above observation and explain why the ABS forecast is biased. When the ABS forecast is made at time t, $(A\hat{B}S_t)$, it is conventionally derived from the monthly forecast by

Table III—Simulation results. Frequency method A outperforms Method B. There are 100 realizations generated from each process

Method A	Process No.	Method B								
		1 Year Ahead			2 Years Ahead			3 Years Ahead		
		S	AS	RW	S	AS	RW	S	AS	RW
S		—	57	52	—	62*	58	—	50	47
AS	I	43	—	45	37	—	40	49	—	39
RW		45	40	—	39	45	—	51	43	—
S		—	51	51	—	58	52	—	50	43
AS	II	46	—	45	39	—	42	50	—	43
RW		45	38	—	45	41	—	54	41	—
S		—	41	32	—	52	36	—	39	42
AS	III	50	—	41	42	—	41	45	—	45
RW		46	48	—	44	43	—	44	41	—

* Indicates significance at 5-percent level.

$$A\hat{B}S_t = ABS(\hat{z}_t, \hat{z}_{t+1}, \cdots, \hat{z}_{t+12}).$$

Suppose z_t is white noise and the maximum forecast for the next k months is required. The best forecast is the expected value of the maximum of k-order statistics from the normal distribution, which is a positive number. On the other hand, the maximum of the predictions of $z_{t+1}, z_{t+2}, \cdots, z_{t+k}$ is zero. This observation indicates the bias in $A\hat{B}S$. However, one can adjust for this bias since it can be estimated directly from the simulation results.

For further evaluation of forecast accuracy, Table III gives the frequency that Method A outperforms Method B. This table was generated using a sign test.[19] This test does not take into account the magnitude of the error.

V. SUMMARY AND CONCLUSIONS

Through extensive data analysis of actual and simulated time series we answered, in this paper, numerous crucial questions: specific questions concerning the selection of a best univariate method for forecasting telephone usage demand, and general questions on the weaknesses and strengths of the three state-of-the-art techniques.

On the specific questions, our results indicate that none of the investigated techniques outperformed a naive random walk in forecasting the ABS CCS/M variable. By means of simulation results we show that there is no gain in using optimal models (models that represent the exact process) for generating point forecasts over a random-walk model for lead times of more than one year. Moreover, we pointed out a bias in the ABS measure which can be adjusted for, using the simulation result.

Concerning the general issues, we demonstrated that identification

of the nonstationary behavior is critical to the forecast accuracy. It is this behavior that has a long-lasting effect on a process. While accuracy is sensitive to this identification, proper techniques are lacking. Of the three investigated methods, only the Box-Jenkins approach gives some guidelines, but it requires user's judgment.

The simulation results also indicate that the AIC criterion did not identify parsimonious models for the investigated processes. They point out that the nonconvergence of the numerical procedures for the state space approach can add considerable complexities and can increase computer costs—all of which must be considered when assessing the feasibility of the method as an automatic forecasting technique.

REFERENCES

1. N. D. Blair and W. S. Hayward, Jr., "New Methods for Administrative Area and Local Entity Forecasting," Ninth ITC Conf, Spain, October 1979.
2. G. E. P. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, San Francisco: Holden-Day, 1976, p. 575.
3. H. Akaike, "A New Look at the Statistical Model Identification," IEEE Trans. Automat. Contr., *AC-19* (December 1974), pp. 716–23.
4. R. H. Jones, "Identification and Autoregressive Spectrum Estimation," IEEE Trans. Automat. Contr., *AC-19* (December 1974), pp. 894–8.
5. R. W. Katz, "On Some Criteria for Estimating the Order of a Markov Chain," ASA Annual Meeting, Washington, D.C., August 1979.
6. H. Akaike, "Markovian Representation of Stochastic Processes and its Application to the Analysis of Autoregressive Moving Average Processes," Ann. Inst. Statist. Math., *26*, No. 3 (1974), pp. 363–87.
7. H. Akaike, "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," in R. K. Mehra and D. G. Lainiotis, eds., *System Identification: Advances and Case Studies*, New York: Academic Press, 1976.
8. H. Akaike, E. Arahata, and T. Ozaki, "TIMSAC-74-A Time Series Analysis and Control Program Package −(1)," Computer Science Monographs, Tokyo, March 1975.
9. E. Parzen, "An Approach to Time Series Analysis," Annals of Mathematical Statistics, *32*, No. 4 (December 1961), pp. 951–89.
10. H. Akaike, "Fitting Autoregressive Models for Prediction," Ann. Inst. Statist. Math., *21*, No. 2 (1969), pp. 243–7.
11. P. Whittle, *Prediction and Regulation by Linear Least Square Methods*, Princeton: D. Van Nostrand, 1963.
12. B. J. Bhansali, "Asymptotic Properties of the Wiener-Kolmogorov Predictor," JRSS, Series B, *36*, No. 1 (1974), pp. 61–73.
13. J. W. Tukey, "An Introduction to the Calculations of Numerical Spectrum Analysis," in B. Harris, ed., *Spectral Analysis of Time Series*, New York: Wiley, 1967.
14. W. M. Brelsford, D. A. Relles, *STATLIB: A Statistical Computing Library*, Englewood Cliffs, NJ: Prentice-Hall, 1981, p. 427.
15. D. DeLong, "A State Space Approach to Modeling Vector Time Series," ASA Annual Meeting, Houston, Texas, August 1981.
16. B. Kleiner, R. D. Martin, and D. J. Thomson, "Robust Estimation of Power Spectra," JRSS, Series B, *41*, No. 3 (1979), pp. 313–51.
17. J. W. Tukey, *Exploratory Data Analysis*, Princeton: Addison Wesley, 1977.
18. R. McGill, J. W. Tukey, and W. A. Larson, "Variations of Box Plots," American Statistician, *32* (February 1978), pp. 12–16.
19. A. M. Mood, "On the Asymptotic Efficiency of Certain Nonparametric Two Sample Tests," Annals of Mathematical Statistics, *25* (September 1954), pp. 514–22.

AUTHOR

Mary N. Youssef, B.Sc. (Mathematics and Astronomy), 1958, Cairo University, Egypt; M.A. (Mathematical Statistics), 1964, Columbia University; M.Sc.

(Computer Science), 1967, Stanford University; Ph.D. (Statistics), 1970, Oregon State University; AT&T Bell Laboratories, 1970–1983; IBM, 1983—. From 1958 to 1965 Mrs. Youssef was an instructor at Cairo University and the Institute of National Planning, Egypt. At AT&T Bell Laboratories she was engaged in various studies on traffic analysis forecasting and optimal design of switching systems. From 1981 to 1983, she was on leave as Associate Professor of Computer Information Systems at the City University of New York. Currently she is an Advisory Engineer with IBM. Senior Member, IEEE; Member, Sigma Xi, ACM, ORSA.

PAPERS BY AT&T BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Bank R. E., Rose D. J., Fichtner W., **Numerical Methods for Semiconductor-Device Simulation.** SIAM J Sci 4(3):416–435, 1983.

Buhler J., Graham R., **Fountains, Showers, and Cascades—Jugglings Quintessential Combinations of Algebra and Acrobatics.** The Science 24(1):44–51, 1984.

Chung F. R. K., Erdos P., **On Unavoidable Graphs.** Combinatori 3(2):167–176, 1983.

Chung F. R. K. et al., **Perfect Storage Representations for Families of Data Structures.** SIAM J Alg 4(4):548–565, 1983.

Cohen A., Mallows C. L., **Assessing Goodness of Fit of Ranking Models to Data.** Statisticia 32(4):361–373, 1983.

Coloncas W. A., Kirkman D., **Interfaces Between Protocol Layers on a Multiprocessor System.** J Telecomm 2(2):211–218, 1983.

Conway J. H., Sloane N. J. A., **Complex and Integral Laminated Lattices.** T Am Math S 280(2):463–490, 1983.

Evangelist W. M., **Software Complexity Metric Sensitivity to Program Structuring Rules.** J Syst Soft 3(3):231–243, 1983.

Fichtner W., Rose D. J., Bank R. E., **Semiconductor-Device Simulation.** SIAM J Sci 4(3):391–415, 1983.

Fishburn P. C., **Ellsberg Revisited—A New Look at Comparative Probability.** Ann Statist 11(4):1047–1059, 1983.

Gilstein C. Z., Leamer E. E., **The Set of Weighted Regression Estimates.** J Am Stat A 78(384):942–948, 1983.

Kaufman L., **Matrix Methods for Queuing Problems.** SIAM J Sci 4(3):525–552, 1983.

Sloane N. J. A., **The Packing of Spheres.** Sci Am 250(1):116–125, 1984.

Tarjan R. E., **Space-Efficient Implementations of Graph Search Methods.** ACM T Math 9(3):326–339, 1983.

ENGINEERING

Baker R. G., Palumbo T. A., **Contact Materials—A Science in Transition.** Plat Surf F 70(12):63–66, 1983.

Chen H. S., Krause J. T., Inoue A., Masumoto T., **The Effect of Quench Rate on the Young Modulus of Fe-Based, Co-Based, Ni-Based, and Pd-Based Amorphous Alloys.** Scrip Metal 17(12):1413–1414, 1983.

Coon D. D. et al., **Readout Electronics for a New Mode of Infrared Detection.** P Soc Photo 395:198–204, 1983.

Craighead H. G., **Highly Absorbing Coatings Using Graded Refractive Indexes and Textured Surfaces.** P Soc Photo 401:356–362, 1983.

Current M. I., Wagner A., **Status of the Production Use of Ion Implantation for IC Manufacture and Requirements for Competitive Application of Focused Ion Beams.** P Soc Photo 393:177–189, 1983.

Dragone C., **Unique Reflector Arrangement With Very Wide Field of View for Multibeam Antennas.** Electr Lett 19(25-2):1061–1062, 1983.

Dutta N. K., Hartman R. L., Tsang W. T., **Gain and Carrier Lifetime Measurements in AlGaAs Multiquantum Well Lasers.** IEEE J Q El 19(11):1613–1616, 1983.

Dziedzic J. M., Ashkin A., Stolen R. H., Simpson J. R., **Tension-Adjusted In-Line Optical-Fiber Attenuator.** Optics Lett 8(12):647–649, 1983.

Gordon E. I., Nash F. R., Hartman R. L., **Purging—A Reliability Assurance Technique for New Technology Semiconductor Devices.** IEEE Elec D 4(12):465–466, 1983.

Harrod W. L., Lubowe A. G., **The Bellpac Packaging System.** West Elec E 27(3): 5–15, 1983.

Hasegawa A., **Amplification and Reshaping of Optical Solitons in a Glass Fiber. 4. Use of the Stimulated Raman Process.** Optics Lett 8(12):650–652, 1983.

Joyce W. B., **Expression for the Fermi Energy in Narrow-Bandgap Semiconductors.** IEEE J Q El 19(11):1625–1627, 1983.

Kaplan M. L., Forrest S. R., Schmidt P. H., Venkatesan T., Lovinger A. J., **Unique Optical and Electronic Properties of Thin-Film, Nonpolymeric Compounds for Device Applications.** J Elec Mat 12(6):989–1001, 1983.

Korner J., **OPEC or a Basic Problem in Source Networks.** IEEE Info T 30(1):68–77, 1984.

Mollenauer L. F., Stolen R. H., **The Soliton Laser.** Optics Lett 9(1):13–15, 1984.

Ong E., Tai K. L., Vadimsky R. G., Kemmerer C. T., **Germanium-Selenium (Ge-Se) Based Resist Systems for Submicron VLSI Application.** P Soc Photo 394:39–48, 1983.

Rainal A. J., **Power Spectrum of FM Clicks (Letter).** IEEE Info T 30(1):122–124, 1984.

Saul R. H., Chen F. S., **Reliability Assurance for Devices With a Sudden-Failure Characteristic.** IEEE Elec D 4(12):467–468, 1983.

Schwarz S. A., Russek S. E., **Semiempirical Equations for Electron Velocity in Silicon. 1. Bulk.** IEEE Device 30(12):1629–1633, 1983.

Schwarz S. A., Russek S. E., **Semiempirical Equations for Electron Velocity in Silicon. 2. MOS Inversion Layer.** IEEE Device 30(12):1634–1639, 1983.

Tenne R., Chin A. K., **Application of Cathodoluminescence Imaging for the Investigation of CdTe—The Effect of Photoetching.** Mater Lett 2(2):143–146, 1983.

Tomita A., Duff D. G., Cheung N. K., **Mode Partition Noise Caused by Wavelength-Dependent Attenuation in Lightwave Systems.** Electr Lett 19(25-2):1079–1080, 1983.

Tsang W. T., Olsson N. A., Logan R. A., **Optoelectronic Logic Operations by Cleaved-Coupled-Cavity Semiconductor Lasers.** IEEE J Q El 19(11):1621–1625, 1983.

Venkatesan T., Brown W. L., Murray C. A., Marcantonio K. J., Wilkens B. J., **Dynamics of Ion-Beam Modification of Polymer Films.** Polym Eng S 23(17):931–934, 1983.

Wagner A., **Applications of Focused Ion Beams.** P Soc Photo 393:167–176, 1983.

Williams J. H., Kahn E. B., Lee S. S., **Effects of Specimen Resonances on Acoustic-Ultrasonic Testing.** Mater Eval 41(13):1502–1510, 1983.

PHYSICAL SCIENCES

Arwin H. et al., **Dielectric Function of Thin Polypyrrole and Prussian Blue Films by Spectroscopic Ellipsometry.** Synth Metal 6(4):309–316, 1983.

Arwin H., Aspnes D. E., Rhiger D. R., **Properties of $Hg_{0.71}Cd_{0.29}Te$ and Some Native Oxides by Spectroscopic Ellipsometry.** J Appl Phys 54(12):7132–7138, 1983.

Aspnes D. E., **The Analysis of Optical Spectra by Fourier Methods.** Surf Sci 135(1–3):284–306, 1983.

Celler G. K., **Laser Crystallization of Thin Si Films on Amorphous Insulating Substrates.** J Cryst Gr 63(3):429–444, 1983.

Chang R. P. H., Darack S., Lane E., Chang C. C., Allara D., Ong E., **Plasma Enhanced Beam Deposition of Thin Films at Low Temperatures.** J Vac Sci B 1(4):935–942, 1983.

Cleveland W. S., Freeny A. E., Graedel T. E., **The Seasonal Component of Atmospheric CO_2—Information From New Approaches to the Decomposition of Seasonal Time Series.** J Geo R-O A 88(NC15):934–946, 1983.

Comin F., Rowe J. E., Citrin P. H., **Structure and Nucleation Mechanism of Nickel Silicide on Si(111) Derived from Surface Extended-X-Ray-Absorption Fine-Structure.** Phys Rev L 51(26):2402–2405, 1983.

Craighead H. G., White J. C., Howard R. E., Jackel L. D., Behringer R. E., Sweeney J. E., Epworth R. W., **Contact Lithography at 157 nm With an F_2 Excimer Laser.** J Vac Sci B 1(4):1186–1189, 1983.

Cross M. C., Hohenberg P. C., Lucke M., **Forcing of Convection Due to Time-Dependent Heating Near Threshold.** J Fluid Mec 136(Nov): 269–276, 1983.

Degani J., Besomi P., Wilt D. P., Nelson R. J., Wilson R. B., **Optical Evaluation of Indium Gallium-Arsenide Phosphide Double-Heterostructure Material for Injection Lasers.** J Appl Phys 54(12):7114–7118, 1983.

Dewey R. J. et al., **Search for C-120 Emission at High Galactic Latitude.** Astronom J 88(12):1832–1834, 1983.

Dutta N. K., **Effect of Uniaxial Stress on Optical Gain in Semiconductors.** J Appl Phys 55(2):285–288, 1984.

Dynes R. C., Lee P. A., **Localization, Interactions, and the Metal-Insulator Transition.** Science 223(4634):355–360, 1984.

Feibelma P. J., Hamann D. R., **Electronic Structure of a Poisoned Transition-Metal Surface.** Phys Rev L 52(1):61–64, 1984.

Ferreira I. B., King A. R., Jaccarino V., Cardy J. L., Guggenhem H. J., **Random-Field-Induced Destruction of the Phase Transition of a Diluted Two-Dimensional Ising Antiferromagnet: $Rb_2CO_{0.85}Mg_{0.15}F_4$.** Phys Rev B 28(9):5192–5198, 1983.

Fleming R. M., Schneemeyer L. F., **Transient Electrical Response of $K_{0.30}$ MoO_3.** Phys Rev B 28(12):6996–6999, 1983.

Furukawa T. et al., **Dielectric Hysteresis and Nonlinearity in a 52/48 Mol Percent Copolymer of Vinylidene Fluoride and Trifluoroethylene.** Macromolec 16(12):1885–1890, 1983.

Gane P. A. C., Leadbetter A. J., Wrighton P. G., Goodby J. W., Gray G. W., Tajbakhsh A. R., **The Phase Behavior of Bis-(4′-N-Heptyloxybenzylidene)-1,4-Phenylenediamine (HEPTOBPD), Crystal J and K Phases.** Molec Cryst 100(1–2):67–74, 1983.

Goodby J. W., **Synthesis, Properties, and Applications of Ferroelectric Smectic Liquid Crystals.** Ferroelectr 49(1–4):275–284, 1983.

Greene B. I., Fleury P. A., Carter H. L., Farrow R. C., **Microscopic Dynamics in Simple Liquids by Subpicosecond Birefringences.** Phys Rev A 29(1):271–274, 1984.

Hong M., Hull G. W., Fuchs E. D., Holthuis J. T., **Multifilamentary Superconducting (NbTa)-Sn Composite Wire by Solid Liquid Reaction for Possible Application Above 20 Tesla.** Mater Lett 2(2):165–168, 1983.

Howard R. E., Craighead H. G., Jackel L. D., Mankiewich P. M., Feldman M., **Electron-Beam Lithography From 20 to 120 keV With a High-Quality Beam.** J Vac Sci B 1(4):1101–1104, 1983.

Huse D. A., Fisher M. E., **Commensurate Melting, Domain-Walls, and Dislocations.** Phys Rev B 29(1):239–270, 1984.

Inoue A., Masumoto T., Chen H. S., **Correlation Between Superconducting Characteristics and Normal Electrical Resistivity for Zr-Si and Zr-Ge Alloys in Metastable BCC and Amorphous Phases.** J Phys F 13(12):2603–2614, 1983.

Johnson P. D., Farrell H. H., Smith N. V., **Carbon KVV Auger Spectroscopy Using a Plane-Grating Monochromator at the National Synchrotron Light Source.** Vacuum 33(10-1):775–778, 1983.

Larson R. G., **Convection and Diffusion of Polymer Network Strands.** J Non-Newt 13(3):279–308, 1983.

Lovinger A. J., Forrest S. R., Kaplan M. L., Schmidt P. H., Venkatesan T., **Structural and Morphological Investigation of the Development of Electrical Conductivity in Ion-Irradiated Thin Films of an Organic Material.** J Appl Phys 55(2):476–482, 1984.

Marcus M. A., **Structure of the Nematic-Isotropic Interface in Landau Theory.** Molec Cryst 100(3–4):253–262, 1983.

Mattheiss L. F., **Plasma Energies for $BaPb_{1-x}Bi_xO_3$.** Phys Rev B 28(12):6629–6633, 1983.

Moerner W. E., Chraplyvy A. R., Sievers A. J., Silsbee R. H., **Persistent Nonphotochemical Spectral Hole Dynamics for an Infrared Vibrational Mode in Alkali-Halide Crystals.** Phys Rev B 28(12):7244–7259, 1983.

Murarka S. P., Kinsbron E., Fraser D. B., Andrews J. M., Lloyd E. J., **High-Temperature Stability of PtSi Formed by Reaction of Metal With Silicon or by Cosputtering.** J Appl Phys 54(12):6943-6951, 1983.

Nalewajek D., Wudl F., Kaplan M. L., Bereman R. D., Dorfman J., Bordner J., **The Chemistry of the 1,2-Dicyanocyclopentadienide Anion—Metal-Complexes and Nucleophilic-Substitution Reactions.** Inorg Chem 22(26):4112-4116, 1983.

Nassau K. et al., **The Germanium-Rich Region of the GeO$_2$-La$_2$O$_3$ Phase Diagram.** Phys C Glas 24(6):150-154, 1983.

Ong E., Baker R. M., Hale L. P., **A Two-Layer Photoresist Process for Patterning High-Reflectivity Substrates.** J Vac Sci B 1(4):1247-1250, 1983.

Raghavan P., Ding Z. Z., Raghavan R. S., **Hyperfine Field at Strontium Nuclei in Nickel.** Hyper Inter 15(1-4):317-320, 1983.

Robinson M., Lischner D. J., Celler G. K., **Large Area Recrystallization of Poly-silicon With Tungsten-Halogen Lamps.** J Cryst Gr 63(3):484-492, 1983.

Roth H. D., Schillinger M. L., Mukai T., Miyashi T., **The Electron-Transfer Induced Interconversion of Cyclobutanes and Diolefins.** Tetrahedr L 24(52):5815-5818, 1983.

Ryan J. T., Matsuoka S., **Viscoelastic Properties of Filled Polycarbonate.** J Elastom P 15(4):209-216, 1983.

Schweizer K. S., Stillinger F. H., **Phase Transitions Induced by Proton Tunneling in Hydrogen-Bonded Crystals—Ground-State Theory.** Phys Rev B 29(1):350-360, 1984.

Stern M. B., Liao P. F., **Reactive Ion Etching of GaAs and InP Using SiCl$_4$.** J Vac Sci B 1(4):1053-1055, 1983.

Tarascon J. M., DiSalvo F. J., Eibschutz M., Murphy D. W., Waszczak J. V., **Preparation and Chemical and Physical Properties of the New Layered Phases Li$_x$Ti$_{1-y}$M$_y$S$_2$ With M = V, Cr, or Fe.** Phys Rev B 28(11):6397-6406, 1983.

Tarascon J. M., Disalvo F. J., Murphy D. W., Hull G., Waszczak J. V., **New Superconducting Ternary Molybdenum Chalcogenides In Mo$_6$Se$_8$, TlMo$_6$S$_8$, and TlMo$_6$Se$_8$.** Phys Rev B 29(1):172-180, 1984.

Vansaarloos W., Llebot J. E., Rubi J. M., **Comment on a Generalized Langevin Equation for 1/f-Noise (Letter).** Physica B&C 122(2):246-248, 1983.

Wilkens B., Venkatesan T., **High-Resolution Mass-Spectrometry of Liquid-Metal Ion Sources.** J Vac Sci B 1(4):1132-1136, 1983.

Wu I. W. et al., **The Microstructure and Critical Current Characteristic of a Bronze-Processed Multifilamentary Nb$_3$Sn Superconducting Wire.** J Appl Phys 54(12):7139-7151, 1983.

SOCIAL AND LIFE SCIENCES

Chang J. J., Julesz B., **Displacement Limits for Spatial-Frequency Filtered Random-Dot Cinematograms in Apparent Motion.** Vision Res 23(12):1379+, 1983.

CONTENTS, JULY–AUGUST 1984