DECEMBER 1975
VOL. 54 NO. 10

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, J. D. deButts, Chairman and Chief Executive Officer, R. D. Lilley, President, C. L. Brown, Executive Vice President and Chief Financial Officer, F. A. Hutson, Jr., Secretary. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 1038, 195 Broadway, New York, N. Y. 10007. Subscriptions $15.00 per year; single copies $1.75 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 54 December 1975 Number 10

Quasi-Optical Polarization Diplexing of Microwaves

By T. S. CHU, M. J. GANS, and W. E. LEGG

(Manuscript received July 17, 1975)

The feasibility of a microwave quasi-optical polarization diplexer has been demonstrated using photo-etched copper strips with thin mylar backing as a practical example. The insertion losses of the principal polarization and the cross polarization have been calculated and measured. The conducting strips must be aligned in a preferred direction, namely, perpendicular to the plane of incidence, to minimize cross-polarized radiation, whereas orientation of the plane of the grid with respect to the beam direction is not restricted. The measured cross-polarized radiation agrees with predictions from simple theoretical models of a magnetic current sheet for the transmission mode and an electric current sheet for the reflection mode. This type of diplexer has been successfully employed in studying the polarization properties of the 20-GHz signal from the ATS-6 satellite.

I. INTRODUCTION

To achieve frequency reuse by employing orthogonal polarizations in a radio communication system, it is essential to avoid cross polarization in the feed patterns that illuminate the antennas; this relies upon the diplexing of two orthogonal polarizations with high isolation. Waveguide-type polarization couplers perform diplexing well where a carrier with an associated bandwidth of about 10 percent is involved. However, it is difficult to provide a low-loss ($\lesssim 0.1$ dB) waveguide diplexer to separate effectively and simultaneously the two polarizations in each of two widely separated common carrier bands, such as 18 and 30 GHz. The difficulty stems from the vulnerability of an oversized waveguide to higher-order modes in the higher-frequency band. Contamination by only 1 percent of the power in higher-order modes may cause unacceptable cross-polarized radiation in the feed.

The necessity for overcoming this problem has led to the suggestion of using a closely spaced wire grid as a quasi-optical polarization diplexer. The purpose of this paper is to describe feasibility studies of this quasi-optical approach. Not only should the insertion loss be small and the feed pattern distortion slight in the principal polarization, but minimization of the cross-polarized radiation should also be achieved. Our measurements have shown that the wires must be oriented in a *preferred direction* to minimize the cross-polarized radiation. This property is explained theoretically by utilizing a magnetic current sheet for transmission through the grid and an electric current sheet for reflection from the grid. This preferred direction requires that the wires be perpendicular to the plane of incidence determined by the beam axis and the grid normal, as shown in Fig. 1b. One notes that the classical application of a polarizer, which consists of a wire grid parallel to an aperture plane, always satisfies this condition.

In Section II we calculate the insertion loss and cross-polarized radiation using simple theoretical models. Section III describes the measurements and the comparison between the calculated and measured cross-polarized radiation of wire grids in various configurations. Section IV discusses applications and includes concluding remarks.

To avoid confusion about the definition of cross polarization,[1] the following two explicit expressions

$$\hat{p}_1 = \hat{\theta} \cos \phi - \hat{\phi} \sin \phi \tag{1}$$

$$\hat{p}_2 = \hat{\theta} \sin \phi + \hat{\phi} \cos \phi \tag{2}$$

are defined as the two orthogonal polarization vectors which are the cross polarization of each other. The carat "^" indicates unit vector, and (θ, ϕ) are spherical coordinates as shown in Fig. 1a; $\hat{p}_1$ and $\hat{p}_2$ within the immediate vicinity of the Z axis are in the nominal X and Y directions, respectively. The usual antenna pattern measurements yield directly the patterns for the two polarization components under this definition. If a feed pattern with the polarization vector (1) or (2) (often called a balanced-feed radiation[2]) illuminates a paraboloid with its axis oriented in the Z direction, the reflected field in the aperture of the paraboloid is free of cross polarization.

II. THEORETICAL CALCULATIONS

2.1 *Insertion losses*

Implementation of quasi-optical polarization diplexers requires prediction of the insertion losses for both the principal and the unwanted

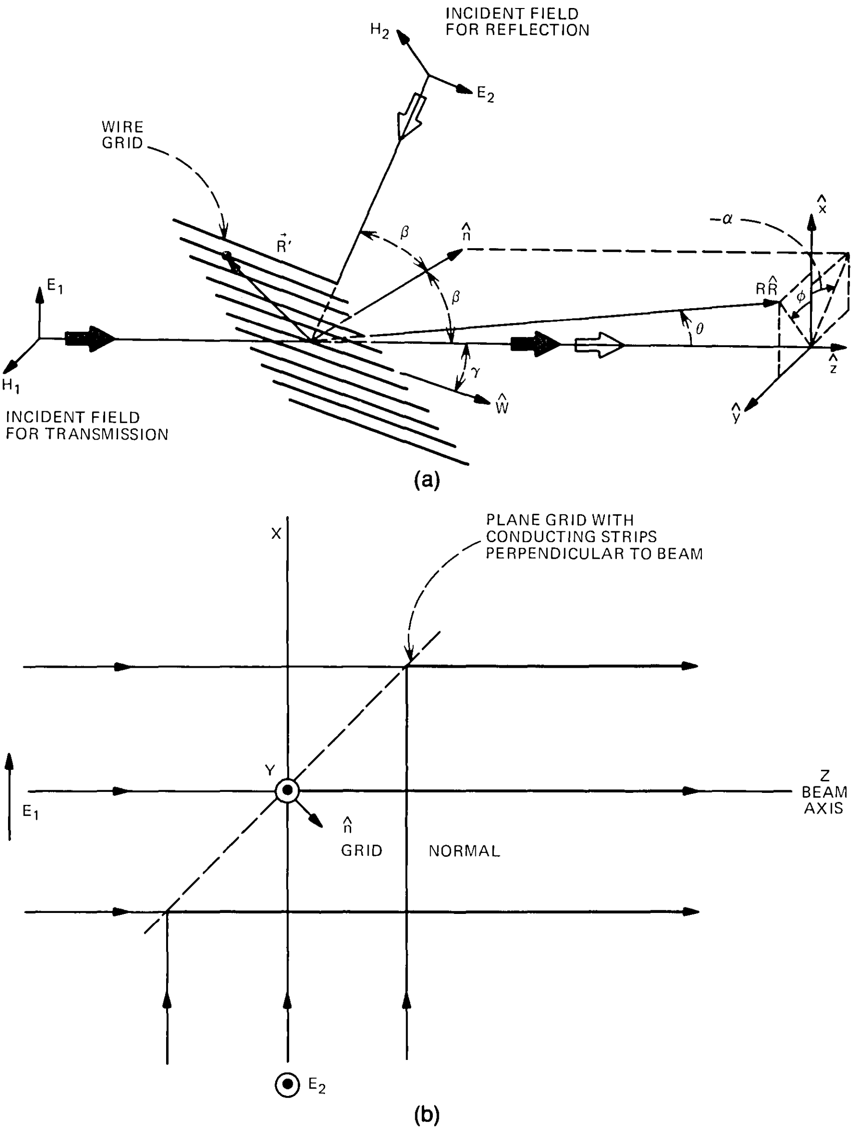


Fig. 1—(a) Geometry of a quasi-optical polarization diplexer. (b) Preferred configuration of quasi-optical polarization diplexing.

cross polarization. If the polarization grid is made of uniformly spaced copper strips with a thin mylar backing, the calculation is facilitated by the equivalence between the periodic grating and the capacitive diaphragm in a parallel-plate waveguide.[3] Neglecting the effect of the thin mylar layer, the power reflection coefficient for an incident wave polarized perpendicular to the strips is approximately given by[3,4]

$$R_{\perp} = \frac{B^2 \cos^2 \theta}{4 + B^2 \cos^2 \theta}, \tag{3}$$

where

$$B = \frac{4b}{\lambda} \ln \sec \left(\frac{\pi d}{2b} \right)$$

is the shunt susceptance, b the grating period, d the width of the conducting strip, λ the wavelength, and θ the angle of incidence between the direction of propagation and the normal to the grating plane. Equation (3)[*] is based upon the low-frequency approximation ($b \ll \lambda$) for a grating of infinitely thin perfectly conducting strips. Using Babinet's principle, the power transmission coefficient "$T_{\parallel}$" for an incident wave, polarized parallel to the plane determined by the strip and the propagation direction, is also approximately given by eq. (3) provided the strip width d in the expression for B is replaced by the spacing $(b - d)$.

If we employ the numerical value, $b = 0.5$ mm, $d = 0.2$ mm, $\lambda = 1.05$ cm, and $\theta = 45°$, as used in the experiment[†] later, substitution into eq. (3) yields the insertion loss of the cross polarization

$$\begin{aligned} -10 \log_{10} R_{\perp} &= 36.8 \text{ dB} \\ -10 \log_{10} T_{\parallel} &= 28.8 \text{ dB}, \end{aligned}$$

which is equivalent to an insertion loss of only 0.001 dB and 0.006 dB for the transmitting and reflecting principal polarizations, respectively. Typically, the cross polarization in a Cassegrain feed aperture that illuminates a polarization diplexer is of the order of -20 dB or less; thus, the improvement provided by use of a quasi-optical polarization diplexer reduces the residual cross-polarized components in the grid aperture to negligible values.

However, the residual cross polarization discussed above is only part of the possible cross-polarized radiation; the following calculations show that the co-polarized field in the grid aperture may also give rise to off-axis cross-polarized radiation if the strips are not oriented in a preferred direction.

[*] Equation (3) is known to be accurate when the strips are parallel to the plane of incidence for all strip widths. However, to the authors' knowledge, the rigorous demonstration, in the literature, of its accuracy for other strip directions is restricted to the cases $(b - d)/b \ll 1$, or $d/b \ll 1$.

[†] Equation (3), for the idealized grid, shows that $d = b/2$ provides the minimum value of the quantity, max $[R_{\perp}, T_{\parallel}]$. Therefore, the grid was designed to have both copper strip width and gap spacing equal to 0.25 mm. However, this specification was close to the resolution limit of the fabrication process in use at the time, which resulted in a grid with the above measured dimensions.

2.2 Magnetic current sheet

For the case of transmission through an arbitrarily oriented wire grid, a magnetic-current-sheet model is used to calculate the radiation.

If a wire grid is placed in front of a radiating aperture that produces a wave collimated in the Z direction, the on-axis radiation is linearly polarized perpendicular to the conducting direction of the wire grid. We shall find the polarization properties of the off-axis radiation.

Let the plane of the wire grid be oriented in an arbitrary direction, as shown in Fig. 1a, with the following unit normal

$$\hat{n} = \hat{x} \sin \beta \cos \alpha + \hat{y} \sin \beta \sin \alpha + \hat{z} \cos \beta. \tag{4}$$

In order that an incident electric field, $\mathbf{E}_1 = E_1 \hat{x}$, may pass freely through the wire grid, the direction of the conducting wires

$$\hat{w} = \sin \gamma \, \hat{y} + \cos \gamma \, \hat{z} \tag{5}$$

must be the same as that of the equivalent magnetic current density $2\hat{n} \times \mathbf{E}_1$ (see the appendix):

$$\hat{n} \times \mathbf{E}_1 = E_1 \sqrt{\cos^2 \beta + \sin^2 \beta \sin^2 \alpha} \left[\frac{\cos \beta \hat{y}}{\sqrt{\cos^2 \beta + \sin^2 \beta \sin^2 \alpha}} - \frac{\sin \beta \sin \alpha \hat{z}}{\sqrt{\cos^2 \beta + \sin^2 \beta \sin^2 \alpha}} \right], \tag{6}$$

i.e.,

$$\sin \gamma = \frac{\cos \beta}{\sqrt{\cos^2 \beta + \sin^2 \beta \sin^2 \alpha}}$$

and

$$\cos \gamma = \frac{-\sin \beta \sin \alpha}{\sqrt{\cos^2 \beta + \sin^2 \beta \sin^2 \alpha}}.$$

The far-zone electric-field radiation of a magnetic current sheet can be written

$$\mathbf{E}_1 = -\frac{jk}{2\pi R} e^{-jkR} \int [\hat{R} \times (\hat{n} \times \mathbf{E}_1)] e^{jk\mathbf{R}' \cdot \hat{R}} dA, \tag{7}$$

where k is the free-space phase constant, and R and $\hat{R}$ are the distance to the far-field point and the corresponding unit direction vector. The points on the magnetic current sheet are defined by $\mathbf{R}'$. The polarization is determined by the bracketed vector product in eq. (7),

$$\mathbf{P}_1 = \hat{R} \times (\hat{n} \times \mathbf{E}_1) = E_1 \sqrt{\cos^2 \beta + \sin^2 \beta \sin^2 \alpha} \, [\hat{\theta}(-\cos \phi \sin \gamma) + \hat{\phi}(-\sin \theta \cos \gamma + \cos \theta \sin \phi \sin \gamma)]. \tag{8}$$

The dot product of eqs. (1) and (8) gives the principal polarization component

$$\mathbf{P}_1 \cdot \hat{p}_1 = -E_1 \cos \beta [1 - \sin^2 \phi (1 - \cos \theta) - \sin \theta \sin \phi \cot \gamma]. \tag{9}$$

If we substitute the above product for the bracket in eq. (7), it is seen that the on-axis radiation is the same as that without the wire grid, while the off-axis radiation is only slightly perturbed. Here the factor $\cos \beta$ accounts for the larger slanted area of the grid.

Now the dot product of eqs. (2) and (8) gives the cross-polarization component

$$\mathbf{P}_1 \cdot \hat{p}_2 = -E_1 \cos \beta [\sin \phi \cos \phi (1 - \cos \theta) + \sin \theta \cos \phi \cot \gamma]. \tag{10}$$

The cross polarization on axis vanishes, as expected. When the direction of the conducting wire is perpendicular to the beam axis, i.e., $\gamma = 90°$, only second-order cross polarization, as represented by the first term in eq. (10), is present with maxima in the $\phi = 45°$ planes. This second-order cross polarization is negligibly small for narrow feed patterns. However, this residue can become a considerable item for broad feed patterns, as will be demonstrated by an experiment described later.

When the direction of the conducting wire is not perpendicular to the beam axis, i.e., $\gamma \neq 90°$, the second term in eq. (10) represents first-order cross-polarization lobes with maxima in the $\phi = 0$ plane. This term is kept small if both θ and $(90° - \gamma)$ are small. Therefore, fine adjustment to reduce the residual cross polarization can be accomplished by rotation of the grid in its own plane in the case of a narrow feed pattern.

2.3 Electric current sheet

Next, consider the case of reflection from the wire grid. The reflected field is entirely due to the fields radiated by electric currents flowing in the wires. If the grid is fine enough, these currents will flow only in the direction of the wires. Thus, we may use an electric current sheet model to compute the field reflected from the wire grid. To obtain perfect reflection from the grid, the direction of the conducting wires must be the same as the induced electric current direction $\hat{n} \times \mathbf{H}_2$, where $\mathbf{H}_2$ is the incident magnetic field for this case. The far-zone electric field of an electric current sheet can be written

$$\mathbf{E}_2 = \frac{jkZ_0}{2\pi R} e^{-jkR} \int \{\hat{R} \times [\hat{R} \times (\hat{n} \times \mathbf{H}_2)]\} e^{jk\mathbf{R}' \cdot \hat{R}} dA, \tag{11}$$

where Z_0 is the free-space impedance. The polarization is determined by the vector product inside the bracket in eq. (11):

$$\mathbf{P}_2 = \hat{R} \times \{\hat{R} \times (\hat{n} \times \mathbf{H}_2)\} = |\hat{n} \times \mathbf{H}_2| [\hat{\theta}(\sin\theta \cos\gamma - \cos\theta \sin\phi \sin\gamma) + \hat{\phi}(-\cos\phi \sin\gamma)]. \quad (12)$$

The above polarization is orthogonal to that of eq. (8). The principal and cross-polarization components are obtained by the dot products of eq. (12) with eqs. (2) and (1), respectively.

$$\mathbf{P}_2 \cdot \hat{p}_2 = -H_2 \cos\beta [1 - \sin^2\phi(1 - \cos\theta) - \sin\theta \sin\phi \cot\gamma] \quad (13)$$

$$\mathbf{P}_2 \cdot \hat{p}_1 = H_2 \cos\beta [\sin\phi \cos\phi (1 - \cos\theta) + \sin\theta \cos\phi \cot\gamma]. \quad (14)$$

Here the relation $|\hat{n} \times \mathbf{H}_2| \sin\gamma = H_2 \cos\beta$ follows the symmetry with respect to $\hat{n}$ between the incident wave and reflected wave of which the magnetic field is equal to $H_2\hat{x}$.

Since eqs. (13) and (14) are identical to eqs. (9) and (10) except for a proportionality constant, the properties of the off-axis cross-polarized radiation described in the previous case are also valid for this orthogonal case in reflection.

III. EXPERIMENT

3.1 Insertion loss

The insertion loss of the combinations of a dual-mode horn and a wire grid were measured at 28.5 GHz and 19 GHz in both transmission and reflection. Figure 2 shows the sketch of the 28.5-GHz experimental model. The dual-mode horn has been described elsewhere.[5] The wire grid was made by photo-etching a copper-covered mylar sheet; copper

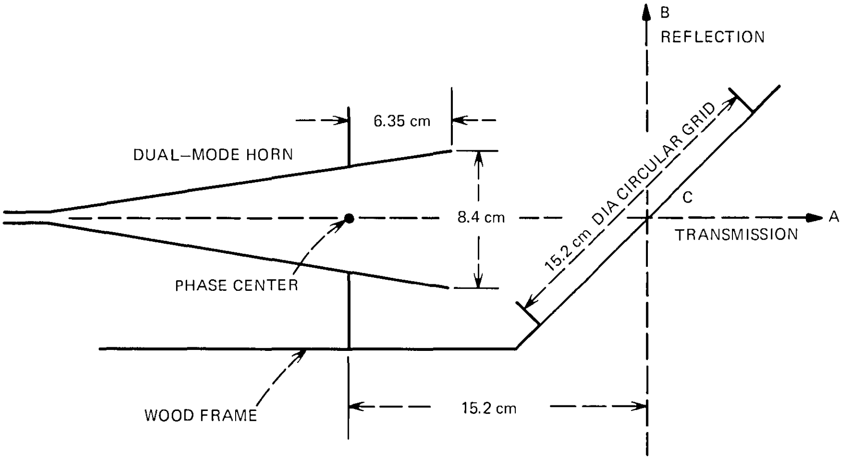


Fig. 2—Schematic of experimental assembly for 28.5 GHz.

strips 0.2 mm wide and 0.018 mm thick are spaced 0.3 mm apart on a mylar sheet 0.013 mm thick. The grid, mounted in a circular wooden rim, can be rotated in its own plane, which is oriented at 45° with respect to the horn aperture. The 19-GHz experiment uses a scaled dual-mode horn and 24.7-cm-diameter circular grid with the same copper strips and the same mylar sheet described earlier.

The measured minimum insertion loss on axis, for both transmission and reflection, was found to be only about 0.1 dB for the principal polarization; the discrepancy with the calculated 0.001 and 0.006 dB can be explained by measuring error and slight pattern distortion due to diffraction around the grid.

The maximum insertion loss of the cross-polarized field, on axis, at 19 and 28.5 GHz for both transmission and reflection is shown in Table I. The symbols ∥ and ⊥ indicate that the grid wires are parallel and perpendicular, respectively, to the plane of incidence.

The measured data are only in qualitative agreement with the approximate prediction from eq. (3). However, the effect of the mylar sheet (0.013 mm thick with a dielectric constant of 3), imperfect polarization of the horn radiation, and diffraction around the grid have been neglected in the approximate calculation. It was observed that the measured insertion loss of the cross polarization depends somewhat upon the spacing between the horn and the grid.

3.2 Radiation patterns at 28.5 GHz

The measured cross polarization in the radiation patterns is found to be negligible if the conducting strips are aligned in the preferred direction normal to the beam. But for the conducting strips in non-preferred directions, such as those parallel to the plane of incidence, maximum cross-polarized radiation is obtained in the transverse planes—AC for transmission and BC for reflection—both perpendicular to the plane of Fig. 2. To illustrate the predictions of the theoretical models in the preceding section, we present the measured

Table I — Measured insertion loss of cross-polarized fields

Grid-Wire Position	28.5 GHz		19 GHz	
	Transmission	Reflection	Transmission	Reflection
∥[*]	24 dB	25	32.5	30
⊥[†]	28 dB	30	38	34
Eq. (3)	28.8	36.8	32.3	40.3

[*] Conducting strips are parallel to the plane of incidence.
[†] Conducting strips are perpendicular to the plane of incidence.

transverse plane patterns at 28.5 GHz for four combinations of horn polarization and conducting strip directions.

The transverse plane patterns in Figs. 3 and 4 were measured with the radiation transmitted through the grid. In Fig. 3 the horn polarization is perpendicular to, and the conducting strips parallel to, the plane of Fig. 2. The average of the cross-polarization lobe maxima is about 20 dB below that of the principal polarization in the same direction ($\theta = 6°$), and agrees well with the prediction of eq. (10) [relative to eq. (9) with $\gamma = 45°$] as shown by the dotted curves. In Fig. 4,

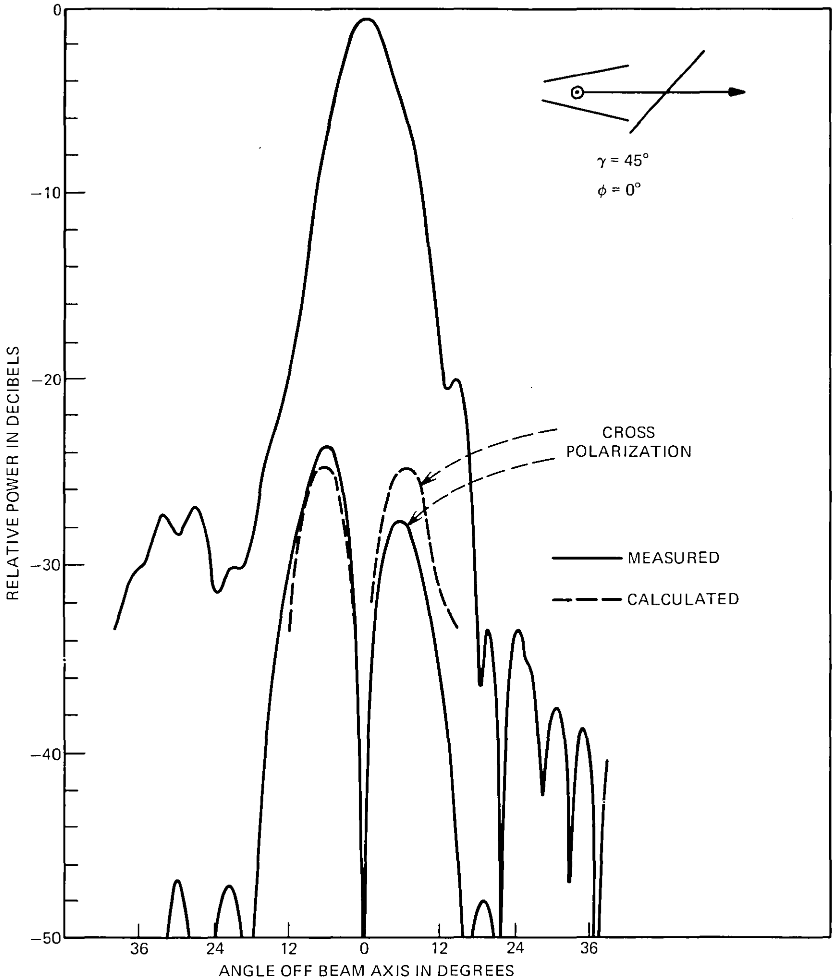


Fig. 3—Radiation patterns of a transmitting grid at 28.5 GHz with conducting strips parallel to the plane of incidence.

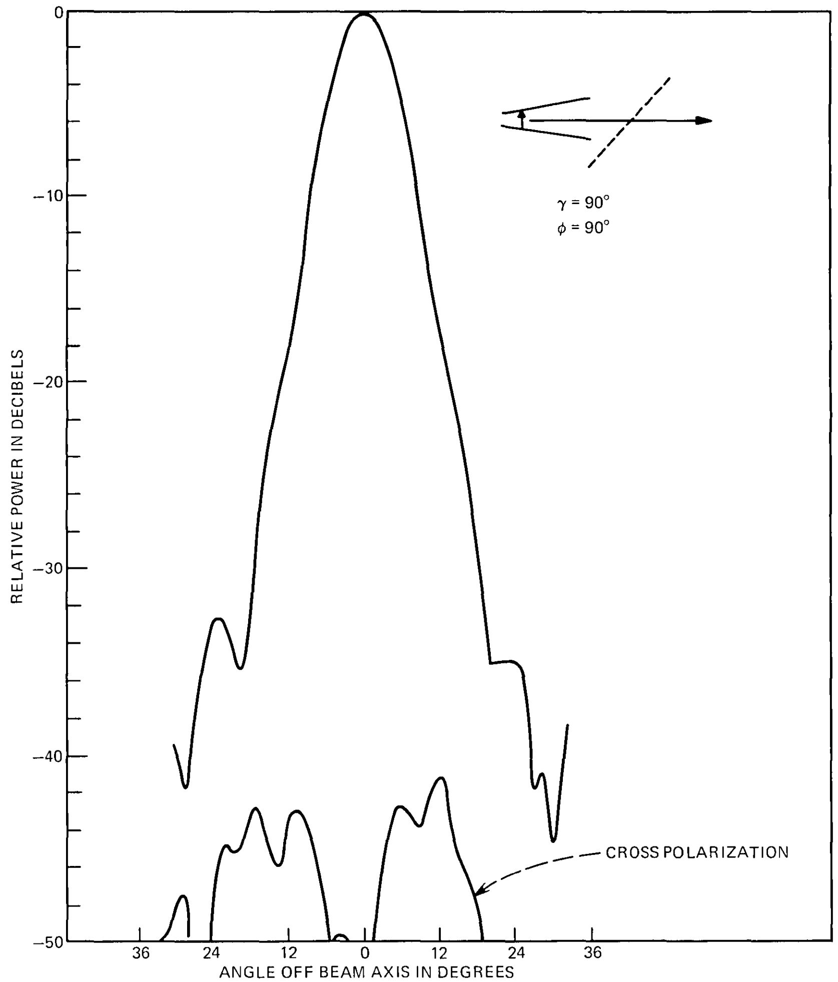


Fig. 4—Measured radiation patterns of a transmitting grid at 28.5 GHz with conducting strips perpendicular to the plane of incidence.

the horn polarization is parallel to, and the conducting strips perpendicular to, the plane of Fig. 2. The measured cross polarization of less than −40 dB essentially confirms the theoretical prediction of negligible cross polarization from eq. (10) ($\gamma = 90°$), since the measuring accuracy of the cross-polarization level is reliable down to about −40 dB.

The transverse plane patterns in Figs. 5 and 6 were measured with the radiation reflected from the grid. In Fig. 5 both the horn polarization and the conducting strips are parallel to the plane of Fig. 2,

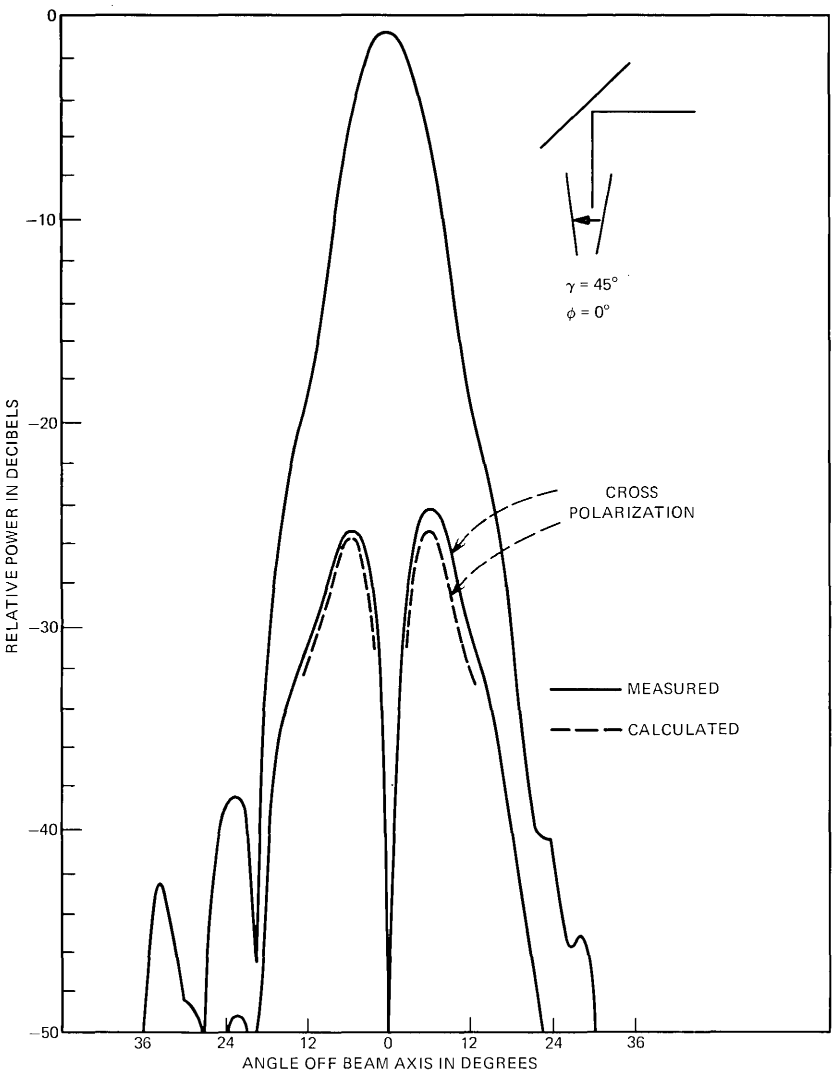


Fig. 5—Radiation patterns of a reflecting grid at 28.5 GHz with conducting strips parallel to the plane of incidence.

and the measured cross polarization is essentially the same as that of the transmitting case in Fig. 3. In Fig. 6, both the horn polarization and the conducting strips are perpendicular to the plane of Fig. 2, and the measured cross polarization of less than −40 dB is similar to that of Fig. 4. Thus, the results show that in employing quasi-optical polarization diplexers the off-axis cross-polarized radiation

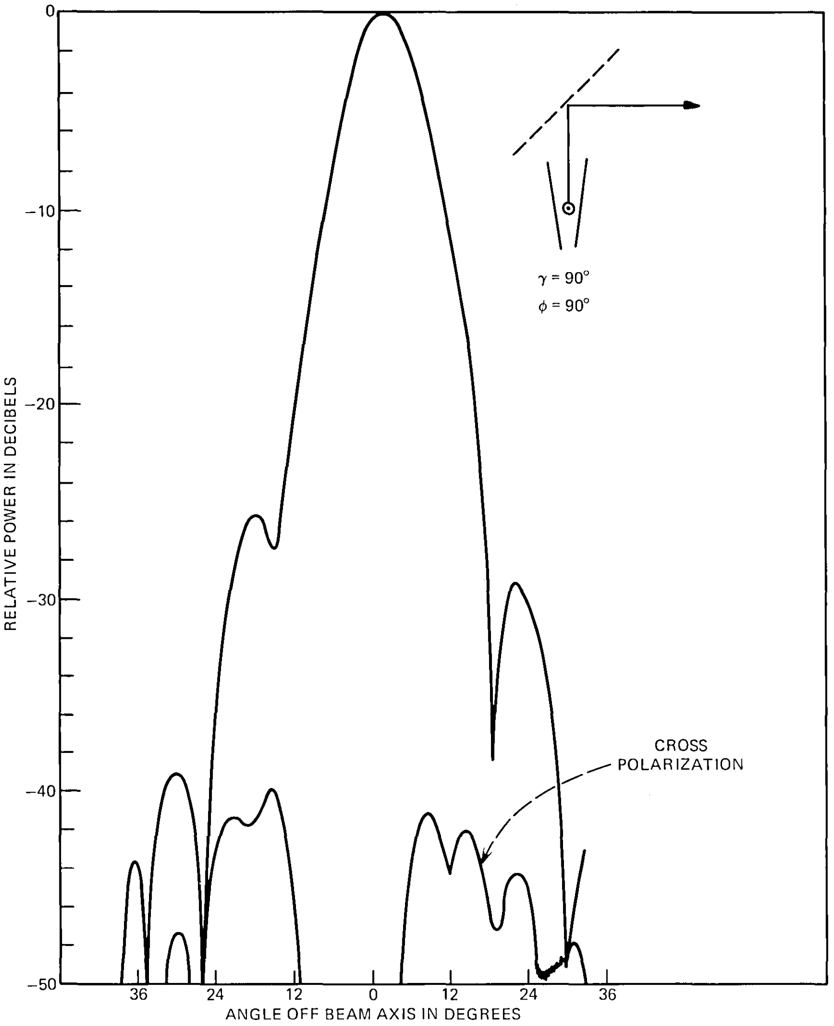


Fig. 6—Measured radiation patterns of a reflecting grid at 28.5 GHz with conducting strips perpendicular to the plane of incidence.

can be suppressed only if the conducting wires are perpendicular to the beam (i.e., perpendicular to the plane of incidence).

Owing to limitation of the measuring accuracy, it is difficult to measure the second-order cross polarization, the $(1 - \cos \theta)$ term in eqs. (10) and (14), of the wire grid for narrow feed patterns. Therefore, we conducted an experiment with a broad feed pattern to check this term, which grows rapidly when θ increases. The radiation patterns

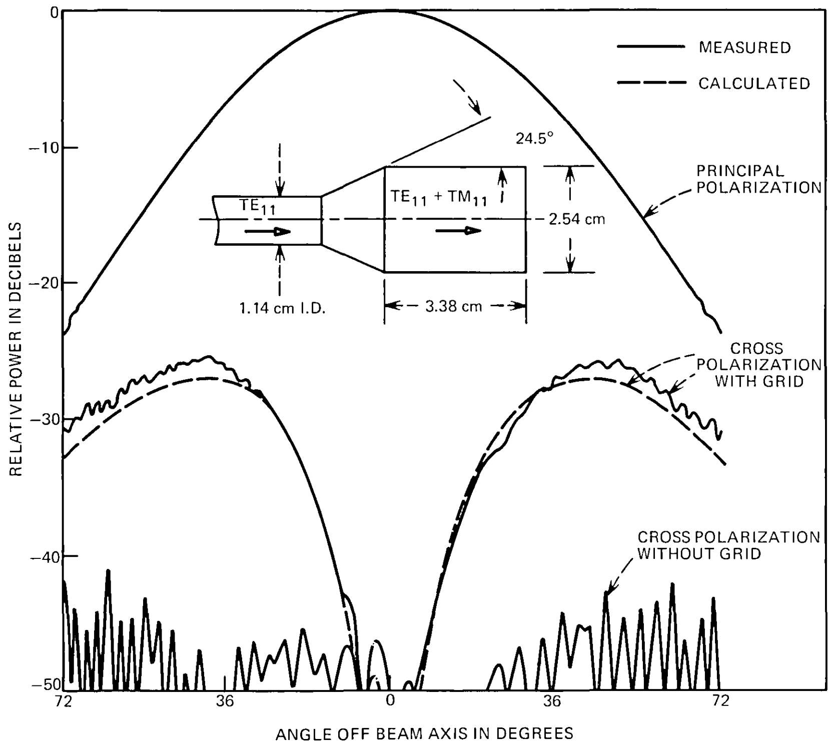


Fig. 7—Radiation patterns of a small dual-mode circular aperture feed ($D/\lambda = 1.39$ at 16.5 GHz).

of a small dual-mode horn[6] (see inset in Fig. 7) were measured with and without the wire grid.

In the absence of the grid, the measured 45° plane patterns of principal and cross polarization are as shown in Fig. 7. The measured patterns in other planes (not shown) exhibited circular symmetry in the co-polarized radiation pattern, and less than −40 dB in cross polarization everywhere. When the small dual-mode aperture was covered by a wire grid, the measured pattern in co-polarization remains essentially the same as without the grid; however, the cross polarization in the $\phi = 45°$ plane rises to −26 dB as shown in Fig. 7. The calculated cross polarization, which is plotted as a dotted curve, shows good agreement with the measured pattern. The cross-polarized radiation of a small grid-covered aperture is similar to that of a dipole.

Thus, the above results demonstrate that the wire grid is a good polarizer for large apertures, whereas improper use of the wire grid can even enhance the cross-polarized radiation of a small aperture.

IV. DISCUSSION

It has been observed[2] that there will be no cross polarization in the main reflector aperture of an offset near-field Cassegrainian antenna provided no cross polarization illuminates the subreflector. This ideal condition can be approximately realized by application of a quasi-optical polarization diplexer to the feed of an offset Cassegrainian antenna with large effective F/D ratio. The quasi-optical polarization diplexer can also be used on a symmetrical Cassegrainian antenna as demonstrated by its application in an earth-station receiver[7,8] for the 20-GHz ATS-6 signal.

The basic philosophy of the quasi-optical polarization diplexer can be simply stated as a cleaning up of the two orthogonal polarizations simultaneously just before illuminating the subreflector. This cleaning process is especially desirable if the feed is an offset reflector with relatively small F/D ratio. But the conducting wires must be oriented in a preferred direction, perpendicular to the plane of incidence, to avoid the off-axis cross-polarized radiation. For the broad feed pattern of a small dual-mode aperture, the second-order cross-polarized radiation from a classical polarizer may *exceed* that of the dual-mode aperture *without* the polarizer. The accuracy of the theoretical predictions demonstrates the utility of equivalent current sources for such analyses.

To avoid excessive spill-over loss and pattern distortion, the quasi-optical diplexer should be made conservatively large, typically with edge illumination less than -20 dB. The main disadvantages of quasi-optical feed systems appear to be bulkier volume and heavier weight compared with conventional waveguide diplexing feed systems, especially when both polarization and frequency diplexing are performed by quasi-optical components. But these components have an advantage in handling high power without difficulty. Equation (10) indicates that fine tuning ($\gamma \approx 90°$) of the residual polarization response may be accomplished by rotation of the grid in its own plane.

APPENDIX

To calculate the field transmitted through the wire grid, a magnetic current equivalent source is chosen, the choice being governed by the following considerations.

Given the tangential electric and/or magnetic field on the bounding surface of a source-free region, we may place equivalent sources on the bounding surface to correctly reproduce the original field in the source-free region:[9]

(*i*) Magnetic current: $\mathbf{K} = \mathbf{E}_t \times \hat{n}$, backed by a perfect electric conductor on the bounding surface.

(*ii*) Electric current: $\mathbf{J} = \hat{n} \times \mathbf{H}_t$, backed by a perfect magnetic conductor on the bounding surface.

(*iii*) Combination:

$$\left\{\begin{array}{l} \mathbf{K} = \mathbf{E}_t \times \hat{n}, \text{ magnetic current} \\ \\ \mathbf{J} = \hat{n} \times \mathbf{H}_t, \text{ electric current} \end{array}\right\} \text{operating in free space.}$$

All three equivalent sources give identical results if they are based on the true fields, $\mathbf{E}_t$ and $\mathbf{H}_t$. However, in the case of transmission through a grid, we do not know the true magnetic field $\mathbf{H}_t$ on the source-free side of the grid. If the polarizer is fine enough, one can be sure, though, that the tangential electric field is perpendicular to the wires. Thus, the field transmitted through the grid can be predicted most accurately by the magnetic current equivalent source backed by an electric conducting plane on the grid. Since a tangential magnetic current imaged in an electric conducting plane is equal to itself, we may include the effect of the electric conducting plane by using twice the magnetic current, 2**K**, operating in free space.

The far field radiated by the magnetic current density, 2**K**, in free space is

$$\mathbf{E}_1 = -\frac{jk}{4\pi R} e^{-jkR} \int_{\substack{\text{grid} \\ \text{gaps}}} [(2\mathbf{K}) \times \hat{R}] e^{jk\mathbf{R}' \cdot \hat{R}} dA. \tag{15}$$

If the grid spacing is very small compared with wavelength, then, as the magnetic current $2\mathbf{K} = 2\mathbf{E}_t \times \hat{n}$ varies from zero on the grid wires to maximum in the space between, the other terms in the integrand of eq. (15) are essentially constant. Thus, we may replace the fluctuating **K** with its average value, $\mathbf{K}_{\text{avg}}$,

$$\mathbf{E}_1 = -\frac{jk}{4\pi R} e^{-jkR} \int [(2\mathbf{K}_{\text{avg}}) \times \hat{R}] e^{jk\mathbf{R}' \cdot \hat{R}} dA. \tag{16}$$

Although we know the direction of **K**, we do not know its magnitude unless the reflection coefficient of the grid is known. In the usual case, the grid is designed to introduce negligible insertion loss for the desired polarization, whence $\mathbf{E}_1$ on axis should equal that present when no grid is used. In this case, the magnitude of $2\mathbf{K}_{\text{avg}}$ would have to be such that

$$2\mathbf{K}_{\text{avg}} = 2\mathbf{E}_1 \times \hat{n} \text{ (negligible insertion loss)} \tag{17}$$

in order that eq. (16) will result in the correct on-axis value for $\mathbf{E}_1$. By substituting eq. (17) into eq. (16), we arrive at eq. (7), the desired equation for computing the field transmitted through the wire grid.

REFERENCES

1. A. C. Ludwig, "The Definition of Cross Polarization," IEEE Trans., *AP-21* (January 1973), pp. 116–119.
2. T. S. Chu and R. H. Turrin, "Depolarization Properties of Off-Set Reflector Antennas," IEEE Transactions, *AP-21* (May 1973), pp. 339–345.
3. R. E. Collin, *Field Theory of Guided Waves*, New York: McGraw-Hill, 1960, p. 366.
4. J. R. Wait, "Reflection at Arbitrary Incidence from a Parallel Wire Grid," Appl. Sci. Res., Sec. B, *4*, No. 6, 1954–1955, pp. 393–400.
5. M. J. Gans and R. A. Semplak, "Some Far-Field Studies of an Offset Launcher," B.S.T.J., *54*, No. 9 (September 1975), pp. 1319–1340.
6. R. H. Turrin, "Dual Mode Small-Aperture Antennas," IEEE Trans. Antennas and Propagation, *AP-15* (March 1967), pp. 307–308.
7. D. A. Gray, "Depolarization of ATS-6 Satellite 20 GHz Beacon Transmitted Through Rain," presented at the 1975 USNC-URSI Spring Meeting, June 3–5, Urbana, Illinois.
8. R. H. Turrin, personal communication.
9. V. H. Rumsey, "Some New Forms of Huygens' Principle," IRE Transactions, *AP-7* (Suppl.) (December 1959), p. S103.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 10, December 1975
Printed in U.S.A.

Material Structure of Germanium-Doped Optical Fibers and Preforms

By H. M. PRESBY, R. D. STANDLEY, J. B. MacCHESNEY, and P. B. O'CONNOR

(Manuscript received July 8, 1975)

The structural characteristics of preforms and optical fibers fabricated by modified chemical vapor deposition were studied by optical, interference, and scanning electron microscopy. It was observed that the structural features resulting from the deposition process are preserved through subsequent processing and appear in the fiber with the exception of a region at the center of the fiber. Here, selective evaporation of dopant material from the inner surface of the deposit results in a refractive index depression on the axis of the optical waveguide.

I. INTRODUCTION

The chemical vapor deposition process, in which oxides are deposited and simultaneously fused on the inner surface of a fused silica tube, has become a valuable technique for fabricating low-loss[1,2] and graded-index optical fibers.[3] In modifications and refinements of this technique, higher depositional rates and very low-loss single-mode fibers[4] have also been achieved.

An important question that arises in utilizing this process concerns the correlation of the deposited material structure in the preform to that in the resulting optical fiber. Can one be confident, for example, that the same distribution of refractive index that is introduced into the preform by changing the material composition of the deposited layers exists in the fiber pulled from this preform? This determination is necessary if one is to reliably fabricate those graded-index profiles required to achieve a minimum of pulse dispersion.[5] This is due to the fact that the shaping of the index profile is quite critical because the reduction-in-pulse-dispersion-vs-profile curves exhibit a singularity-like behavior in the region of the optimum index distribution.[6]

Evidence for the preservation of the deposited profile has recently been reported, based on the observation of a linear increase in refractive index in a fiber which was pulled from a preform in which the dopant concentration was increased in the same manner.[7] In this paper, we present results of optical, interference, and scanning electron microscope studies[8] of a graded, near-parabolic, index fiber and preform as further aid in understanding the transition of material from the preform to the fiber state. A main conclusion of this study is that the structural features resulting from the preform deposition process are preserved and, after suitable scale transformation, appear in the fiber. Due consideration should therefore be given to depositional characteristics that may ultimately affect transmission behavior.

II. OBSERVATIONS

The preform originates from an approximately $\frac{1}{2}$-m-long 12×14-mm fused-quartz tube which is collapsed into a rod after the deposition process. In the structure studied here, the deposition started with an initial layer of borosilicate to prevent impurity diffusion into the core. This layer was deposited with 41 traversals of a oxyhydrogen burner which provides the heat to react the BCl_3, $SiCl_4$, and O_2 starting materials. The core deposition process consists of systematically increasing the flow of $GeCl_4$ while holding the flows of $SiCl_4$ and BCl_3 constant, thus producing an increasing GeO_2 content and associated increased refractive index with increasing deposit thickness. The $GeCl_4$ flow was increased 11 times, in such a manner as to produce a near-parabolic index variation from the cladding interface to the center of the core. The number of torch traversals during each of the 11 steps was controlled to make the thickness of each step approximately equal. After collapse, a length of preform was pulled into a fiber with an overall diameter of $\sim 100\ \mu m$ by the use of an electric furnace.

A slice transverse to the axis of the remaining length of preform was made and then polished to a thickness of approximately 10 mills for interference and optical microscopic observations. Transverse samples of the fiber were also prepared for interference-microscope and scanning-electron-microscope studies.[8] In the latter case, after a short length of fiber is scored and broken to ensure a flat end, it is etched in a 25-percent solution of hydrofluoric acid for several minutes and then flash-coated. The last step is performed to prevent charge build-up on the sample during scanning-electron microscope observations.

An overall view of the preform sample observed with conventional optical microscopy is shown in Fig. 1a. The sample is $\sim$7.2 mm in diameter with a core diameter of 4 mm. The irregular shape is due to pieces of the cladding which broke off during the cutting and polishing procedure. It should be noted that the preform is under considerable stress because of the difference in the expansion coefficients of the

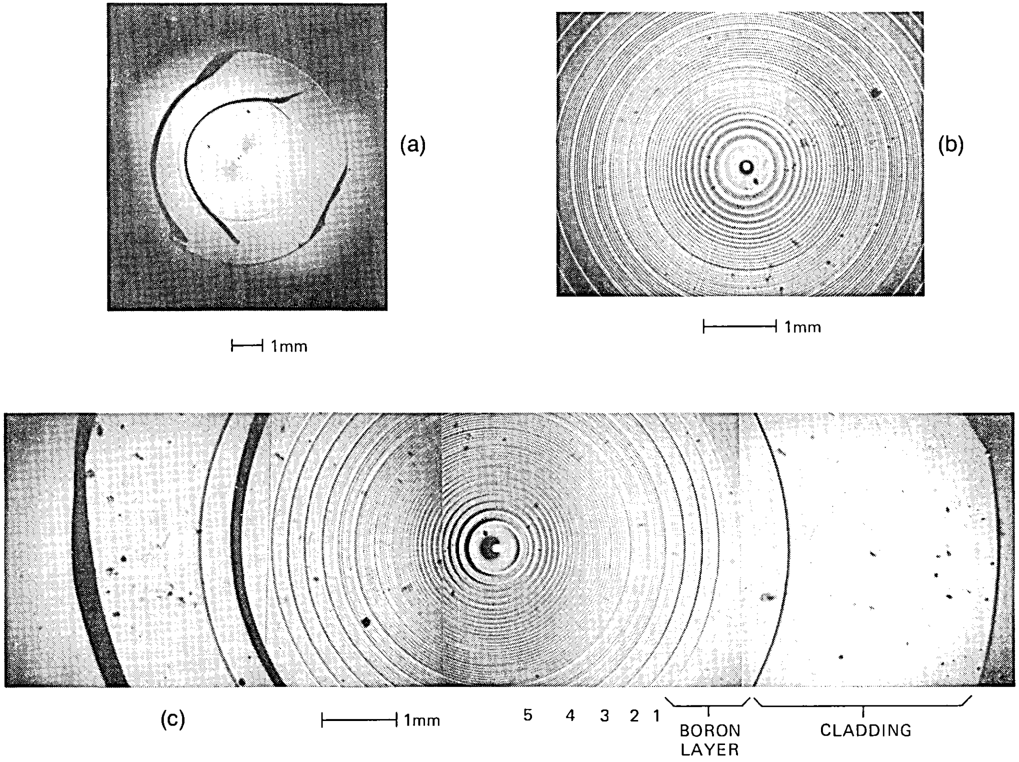


Fig. 1—Transmitted light photomicrographs of transverse section of preform. (a) Entire section. (b) Central core region. (c) Entire section at higher magnification.

core and the cladding. If this stress is not to be relieved by shattering upon cutting, extreme care must be used. In this case, the preform was cut with a diamond wire saw with controlled lubrication. Even so, the preform did crack, as seen by the dark curved line in the left-hand section of the sample.

Figures 1b and 1c obtained by optical microscopy show expanded views of the central region and of a portion of the entire cross section. The cladding, the borosilicate layer, and the first five germania-borosilicate steps are labeled. The regions between the steps are quite distinct, as are the individual layers within each step. Each of these layers, as noted, corresponds to a traversal of the oxyhydrogen burner along the tube. During step 3, for example, the $GeCl_4$ flow was maintained constant for nine traversals of the burner, producing the nine layers observed in Fig. 1c.

An expanded view of one-half the preform sample as observed by interference microscopy is shown in Fig. 2. The refractive-index difference between the cladding and a point in the core of the preform is given by the fringe displacement at that point times the wavelength of observation and divided by the thickness of the sample. One observes the straight parallel fringes in the fused silica cladding on the right indicating the uniform composition of this region, as expected. The drop in the level of the fringes indicates the termination of the cladding and the start of the borosilicate step that has a lower index of refraction than pure fused silica. Again, in this region, which extends for about 375 μm, the composition is relatively uniform and no evidence is seen of the 41 layers which comprise this step. This tends to indicate that some boron diffusion occurs, smoothing out the individual layers. It does not appear, however, that boron diffuses into the cladding, as evidenced by the relatively sharp transition occurring $\sim$10 μm between the cladding and this step.

At the termination of the pure borosilicate layer, germania deposition commences. The first three of these steps are labeled. Note, in particular, that step 3 exhibits nine sinusoid-like variations which, as discussed previously, correspond to nine torch traversals used in depositing this step. The reason for this index variation within each layer may be due either to a difference in composition of the particles reacting homogeneously (i.e., in the gas phase) and heterogeneously (i.e., on the surface of the tube),[7] or to temperature variation effects, depositing different concentrations as the torch passes. These compositional variations become quite sizable as the number of layers in each step decreases towards the center of the core. Note that in all steps the integrity of these layers is maintained and that relatively sharp boundaries exist between the steps. The transitions are most

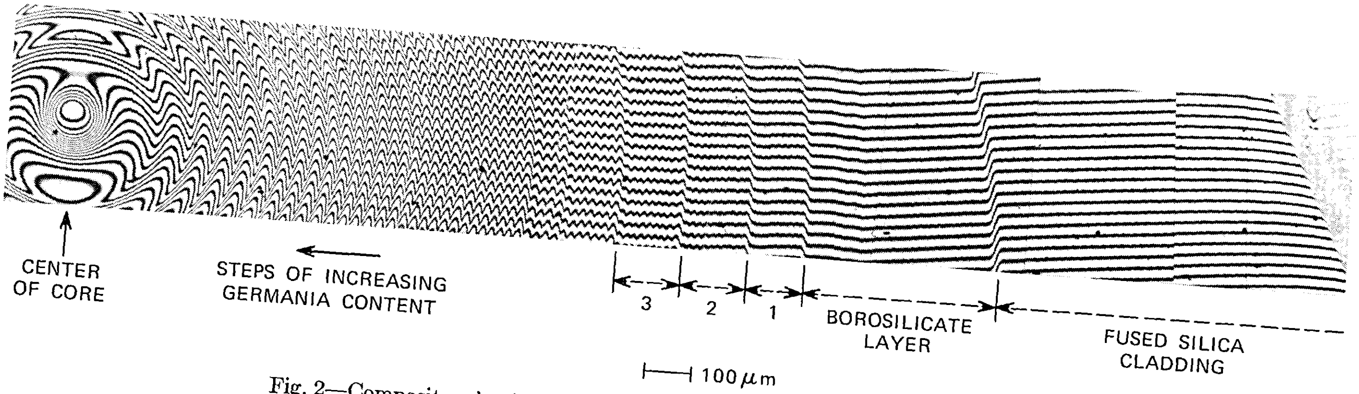


Fig. 2—Composite microinterferrogram of segment of one-half the preform sample.

clearly seen between the first few steps in which the change in germanium content is largest. These observations indicate that little, if any, germanium diffusion occurs between adjacent steps or within a given step itself.

The situation is quite different at the center of the core. Despite the fact that the germanium concentration was varied in a smoothly increasing manner in the last several steps, a large disturbance in the center is observed, with a corresponding dip in the refractive index. We believe this is due to the evaporation of germania at the inner surface of the deposit during the elevated temperatures experienced in the collapse process. The fact that the last two or three layers appear to broaden towards the center also indicates the existence of some germanium flow extending beyond the immediate central segment. It may be possible to compensate for this effect by a germanium overdoping in this region.

Quantitative index measurements indicate a maximum index difference between core and cladding of approximately $\Delta n_m = 0.016$ and an index difference between the cladding and the borosilicate layer of $\Delta n = 0.004$. Both values are in very good agreement with the corresponding measurements made from the microinterferrogram of the fiber, which is shown in Fig. 3a, and indicate a preservation, for the most part of the material composition through the pulling process.

Note in Fig. 3a the uniform cladding containing straight parallel fringes, the subsequent drop owing to the borosilicate layer, and the gradual grading of the index profile to a maximum near the axis. The grading was near-parabolic and has been related to a reduction in pulse dispersion for this fiber.[9] These regions are again shown in the transmitted-light photomicrograph of Fig. 3b. The core which appears as the bright central area is surrounded by the relatively dark borosilicate layer. Beyond that is the grayish cladding with an overall diameter of 96 μm.

The resolution of the interference microscope is not sufficient to resolve any layer or step structure in the fiber. To obtain greater resolution, we made use of scanning electron microscopy. Preparation of the fiber samples was described previously, and results of observations are shown in Figs. 4 and 5.

Figure 4 presents three micrographs taken at increasing magnifications centered on the axis, and Fig. 5 is a composite photograph of a section of one-half the fiber at somewhat greater magnification. The main points to be noted are the preservation of the step and layer structure in the fiber and the appearance of the elevated region near the axis. This feature is a region that did not etch as rapidly as the surrounding area, because of a lack of contained dopant and agrees

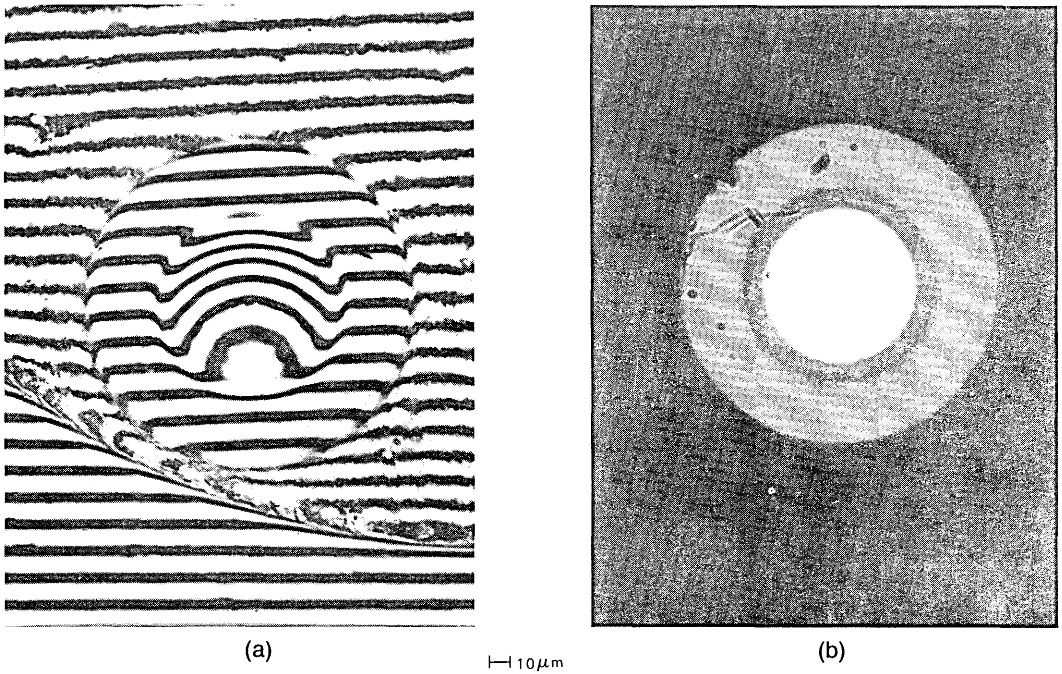


Fig. 3—(a) Microinterferrogram. (b) Photomicrograph of fiber.

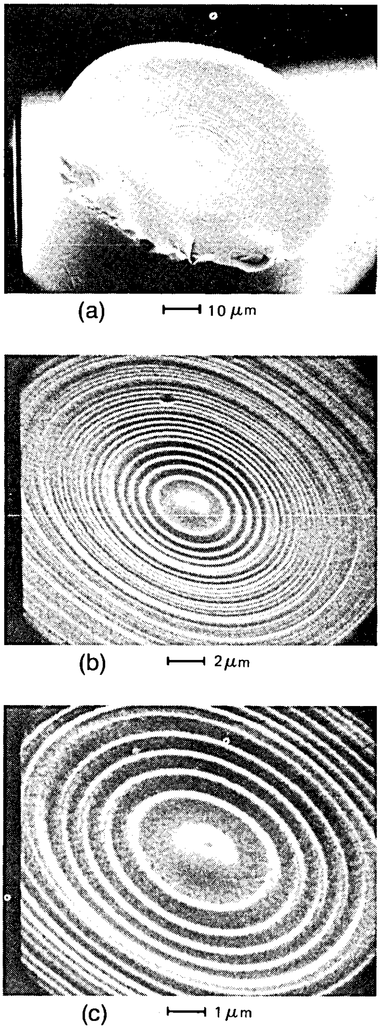


Fig. 4—Scanning-electron-beam-microscope photographs of fiber at increasing magnification.

with the preform observation of germanium departure during collapse. The appearance of the distinct step and layer structure displays the further lack of germanium diffusion even during the pulling process and indicates the maintenance of compositional and structural integrity from the deposition through the pulling process, with the exception of the central region.

We further investigated this latter region in another fiber prepared in the same manner by chemical vapor deposition. In this fiber, the

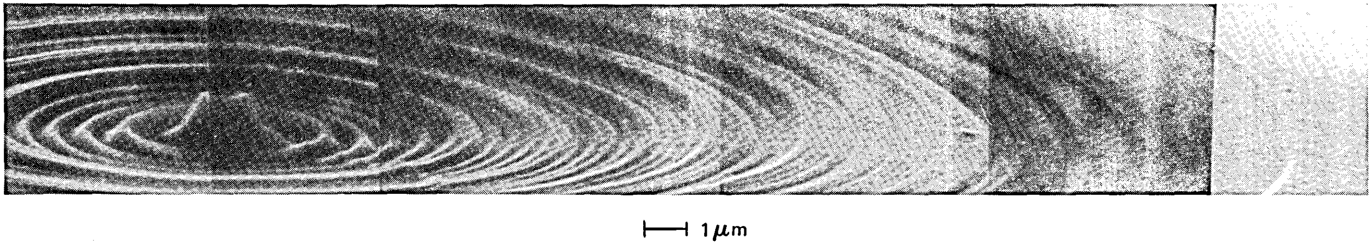


Fig. 5—Composite scanning-electron-beam-microscope photograph of segment of one-half fiber.

$GeCl_4$ flow was increased at each of ten steps to produce a linear variation in concentration. The number of layers in each step varied somewhat from the fiber considered above.[7] Scanning-electron-microscope photographs of etched fiber samples are shown at increasing magnification in Fig. 6.

The depression of the index on the axis because of the loss of dopant is quite pronounced in this fiber and appears as the micron-or-so in

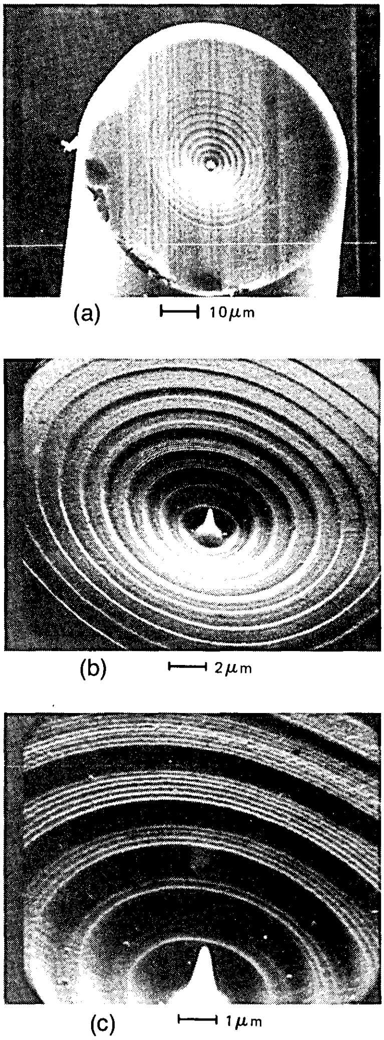


Fig. 6—Scanning-electron-beam-microscope photograph of fiber having linear refractive-index profile.

diameter, raised, tapered pip in the center. Note also the preservation of the distinct step and layer structure. The refractive index depression is also observed in this case in the microinterferrogram of the fiber shown in Fig. 7. It is seen as a dip in the fringe normally passing through the center of the fiber. Slight modulations of the refractive index can also be observed at each of the ten steps in this case owing to the relatively large change in germanium concentration between the steps.

It is important to note that, despite the structural features that exist on such a small scale, the losses of these fibers were less than 5 dB/km in the region of 1.0 μm. This is presumably due to the fact that these features are very uniform in the direction parallel to the axis of the fiber and hence do not contribute in a large way to scattering losses. Small-scale variations of such features if existing, however, could form a lower limit on losses achievable with fibers fabricated by this technique.

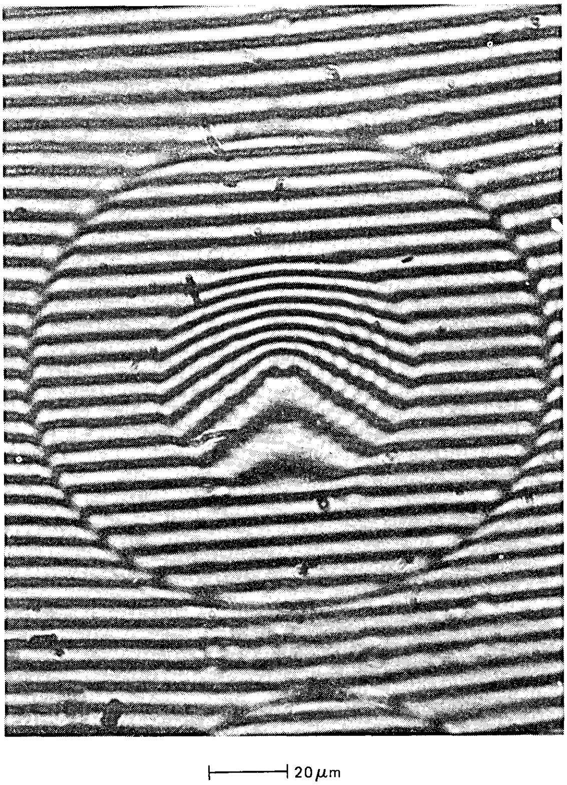


Fig. 7—Microinterferrogram of linear refractive-index profile fiber showing index depression in center.

In summary, structural features resulting from the preparation of both the preforms and graded-index optical fiber by the chemical vapor deposition process have been observed by optical, interference, and scanning electron microscopy. These features can be directly related to steps in the fiber fabrication. It was observed that structures present in the preform were preserved through the drawing and were present in the fiber. For instance, optical and interference observations indicated that germania concentrations varied within each deposited layer, and this variation was not substantially altered by subsequent processing steps except in the centermost layers. Here, flow of the deposit during collapse and vaporization of germania, probably as GeO, led to a depleted region of lower refractive index at the fiber center.

REFERENCES

1. J. B. MacChesney, P. B. O'Connor, F. V. DiMarcello, J. R. Simpson, and P. D. Lazay, Proceedings of Xth International Congress on Glass, Kyoto, Japan, July 6, 1974, pp. 40–45.
2. W. G. French, J. B. MacChesney, P. B. O'Connor, and G. W. Tasker, "Optical Waveguides with Very Low Losses," B.S.T.J., *53*, No. 5 (May–June 1974), pp. 951–954.
3. J. B. MacChesney, P. B. O'Connor, and H. M. Presby, "A New Technique for the Preparation of Low-Loss and Graded-Index Optical Fibers," Proc. IEEE, *62*, No. 9 (September 1974), pp. 1280–1281.
4. W. G. French and G. W. Tasker, "Fabrication of Graded Index and Single Mode Fibers with Silica Cores," Digest of Papers Presented at Topical Meeting on Optical Fiber Transmission, Williamsburg, Virginia, January, 1975, pp. TuA2-1—TuA2-3.
5. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52*, No. 9 (November 1973), pp. 1563–1578.
6. E. A. J. Marcatili, "Theory and Design of Fibers for Transmission," Digest of Papers Presented at Topical Meeting on Optical Fiber Transmission, Williamsburg, Virginia, January, 1975, pp. TuC4-1—TuC4-4.
7. P. B. O'Connor, H. M. Presby, J. B. MacChesney, and L. G. Cohen, to be published by J. Am. Ceram. Soc.
8. C. A. Burrus and R. D. Standley, "Viewing Refractive-Index Profiles and Small-Scale Inhomogeneities in Glass Optical Fibers: Some Techniques," Applied Optics, *13*, No. 10 (October 1974), pp. 2365–2369.
9. L. G. Cohen, P. Kaiser, J. B. MacChesney, P. B. O'Connor, and H. M. Presby, "Transmission Properties of a Low-Loss Near-Parabolic-Index Fiber," Appl. Phys. Lett., *26*, No. 8 (April 1975), pp. 472–474.

Copyright © 1975 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 10, December 1975
Printed in U.S.A.

An Efficient Linear-Prediction Vocoder

By M. R. SAMBUR

(Manuscript received April 29, 1975)

A primary interest in any method for producing synthetic speech is to minimize the number of bits per second required to generate acceptable quality speech. An efficient method for transmitting the linear-prediction parameters has been found by using the techniques of differential PCM. *Using this technique, speech transmission is achieved employing fewer than 1500 bits/s. Further reductions in the linear-prediction storage requirements can be realized at a cost of higher system complexity by transmission of the most significant eigenvectors of the parameters. This technique in combination with differential* PCM *can lower the storage to 1000 bits/s.*

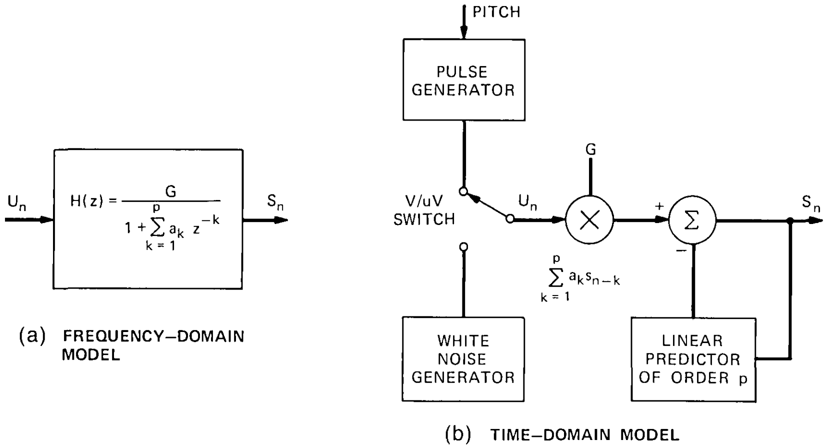
I. INTRODUCTION

The method of linear prediction has proved quite popular and successful for use in speech compression systems.[1–4] In this method, speech is modeled as the output of an all-pole filter $H(z)$ that is excited by a sequence of pulses separated by the pitch period for voiced sounds, or pseudo-random noise for unvoiced sounds. These assumptions imply that within a frame of speech the output speech sequence is given by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + u_n,$$

where p is the number of modeled poles, u_n is the appropriate input excitation, and the a_k's are the coefficients characterizing the filter (linear prediction coefficients). Figure 1 illustrates the frequency-domain, as well as the equivalent time-domain, model of linear-prediction speech production. To account for the nonstationary character of the speech waveform, the parameters a_k of the modeled filter are periodically updated during successive speech frames.* Generation of speech in this method requires a knowledge of the pitch, the filter

* A frame is a segment of speech thought adequate to assume stationarity of the speech process. Typical frame lengths employed range from 10 to 30 ms.


PITCH
PULSE GENERATOR
G
V/uV SWITCH
U_n
+
−
Σ
S_n
$\sum_{k=1}^{p} a_k s_{n-k}$
WHITE NOISE GENERATOR
LINEAR PREDICTOR OF ORDER p
U_n
$H(z) = \frac{G}{1+\sum_{k=1}^{p} a_k z^{-k}}$
S_n

(a) FREQUENCY–DOMAIN MODEL

(b) TIME–DOMAIN MODEL

Fig. 1—Discrete model of speech production as employed in linear prediction.

parameters, and the gain of the filter (amplitude of input excitation) in each speech frame.

A primary interest in any method for producing synthetic speech is to minimize the number of bits per second needed to generate acceptable quality speech. The smaller the information storage requirements (bits per second), the more attractive the system becomes for the important applications of voice answer-back and speech transmission.[5] To achieve the minimum storage requirement for a given system, an efficient means of quantizing the generating parameters must be determined. Using conventional pulse code modulation (PCM) techniques in which the amplitude of each parameter is uniformly quantized into 2^B levels, it has been found necessary to allot at least five bits ($B = 5$) of information for both pitch and gain and at least 11 bits for each a_k.[1] The corresponding storage requirements for this method of quantization of the linear-prediction (LPC) parameters is unacceptable for many applications, and an improved scheme for quantizing the parameters is needed.

For the usual 12-pole linear-prediction representation, the dominant portion of storage is allotted to the filter coefficients (132 bits per frame in the PCM method of information transmission). The extremely fine quantization of the a_k's is necessary because small perturbations in the coefficients can sometimes cause radical changes in the important pole frequencies of the modeled filter $H(z)$ and may even cause the filter to become unstable (poles outside the unit circle). To overcome the limitations of quantizing the predictor (filter) coefficients, it has been found quite profitable to transmit different but informationally equivalent sets of parameters.[4,6] The most suitable parameters have been experimentally determined to be the log-area ratio coefficients g_i.[4] These coefficients are nonlinearly related to the filter coefficients by

$$g_i = \log \frac{1 + k_i}{1 - k_i}, \tag{1}$$

where the k_i's are termed the parcor coefficients.[1,2,4,6] If we denote $a_i^{(j)}$ as the ith linear prediction coefficient for a jth-pole linear-prediction model, then

$$k_i = a_i^{(i)}. \tag{2}$$

The parcor coefficients have the very important property that if

$$|k_i| < 1, \qquad i = 1, \cdots, p, \tag{3}$$

then it is guaranteed that the linear-prediction filter is stable.[4] Thus, small perturbations in the parcor coefficients or log-area coefficients will not affect the stability of the modeled filter.

Since the log-area coefficients are nonlinearly related to the filter coefficients, transmission of the log-area parameters is equivalent to a nonuniform quantization of the linear-prediction coefficients. By transmitting the log-area parameters, the number of bits allotted to the filter parameters can be effectively reduced by nearly $\frac{1}{2}$.[3,4,6] In this paper, we offer two additional methods of quantization of the necessary synthesis parameters (pitch, gain, and filter coefficients) that can even further reduce the storage requirements of a linear-prediction vocoder. One proposed method of quantization uses the technique of differential PCM (DPCM) to transmit the linear-prediction parameters. This scheme exploits the fact that the LPC parameters are themselves predictable from previous samples. Using this method, speech transmission that is practically equivalent to the unquantized synthesis can be achieved using fewer than 2000 bits/s.

The second method of transmission exploits the redundancy between the linear-prediction parameters. The LPC parameters can be predicted not only from the given parameter's past values, but also in some sense from values of the other parameters. The suggested scheme involves the transmission (using DPCM techniques) of the most significant eigenvectors of the log-area parameters. For the typical speech utterance, the space of the 12 log-area coefficients can be effectively represented by only the first five or six eigenvectors. The transmission requirement for this method is fewer than 1200 bits/s.

The organization of this paper is as follows. In Section II, we briefly discuss the concept of DPCM coding. In Section III, we show that DPCM coding offers a significant improvement over PCM coding for transmission of the linear-prediction parameters. In Section IV, the results are presented of a synthetic speech experiment using the proposed DPCM scheme. In Section V, we discuss several methods of DPCM coding that are more suitable for real-time implementation. Included in this section is a discussion of adaptive quantization (ADPCM) and adaptive DPCM prediction. In Section VI, we discuss the method of orthogonal linear prediction. The results of synthetic speech experiments are included in this section. Finally, we conclude with a summary and discussion of the results presented in the paper.

II. DIFFERENTIAL PULSE CODE MODULATION

The idea of differential PCM is similar in philosophy to the concept employed in linear-prediction speech analysis. In DPCM, we assume that the transmitted parameter in a given frame of interest can be estimated by a linear combination of the parameter in previous frames.[7] If we let x_r denote the value of the transmission parameter x in the rth frame (where x can represent pitch, gain, log-area coefficients, or whatever), then this assumption implies

$$x_r \approx \hat{x}_r = \sum_{i=1}^{n} b_i x_{r-i}, \tag{4}$$

where n is the order of the DPCM prediction analysis. The DPCM technique calls for the transmission of the difference between the predicted value $\hat{x}_r$ and the true value x_r.

Figure 2 illustrates the structure of the DPCM coding system. In the implementation of a DPCM scheme, a feedback path around the quantizer is used to ensure that the error in the reconstructed (quantized) signal $\tilde{x}_r$ is precisely the quantization error for the difference signal $e_r = x_r - \hat{\tilde{x}}_r$, where $\hat{\tilde{x}}$ is the predicted value based upon the quantized signal $\tilde{x}_r$. The predictor coefficients b_i are chosen to minimize the power of the difference signal e_r. The mathematical analysis required for the solution of the optimum set of b_i's is exactly the same as the analysis for the calculation of the linear-prediction coefficients, a_i, $i = 1, \cdots, p$. The determination of the b_i's is made by solving the familiar correlation equations:

$$\sum_{i=1}^{n} b_i \sum_{r=n}^{N} x_{r-i} x_{r-k} = -\sum_{r=n}^{N} x_r x_{r-k}, \qquad 1 \leqq k \leqq n, \tag{5}$$

where N is the number of frames in the utterance.

The advantage of DPCM coding is obvious when one realizes that, if x_r can be accurately estimated from previous samples, the information necessary for transmission (as expressed by the difference signal $x_r - \hat{\tilde{x}}$) is necessarily less than the information required for coding the signal without exploiting its predictability. The advantage of

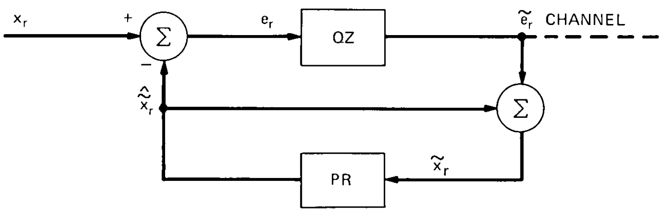


Fig. 2—Differential PCM (QZ = quantizer; PR = predictor; = $\hat{\hat{x}}_r = \sum_{j=1}^{n} b_j x_{r-j}$).

DPCM coding can be precisely specified by noting that, for a given fineness of quantization, the quantization error is proportional to the variance of the signal present at the quantizer.[7] Thus, the improvement in performance (as measured by the frequently used standard of signal-to-quantization-error ratio) using DPCM strategy over straight PCM coding is given by the ratio of the variance (power) of x_r to that of the difference signal

$$G = \frac{\langle x_r^2 \rangle}{\langle (x_r - \hat{x})^2 \rangle}. \tag{6}$$

Using the optimum predictors b_i, the resulting gain over PCM is approximately* given by

$$G_{\text{opt}} = \left(1 - \sum_{k=1}^{n} \frac{b_i C_i}{C_0}\right)^{-1} = \frac{\langle x_r^2 \rangle}{\langle (x_r - \hat{x}_r)^2 \rangle}, \tag{7}$$

where

$$C_i = \sum_{r=n}^{N} x_r x_{r-i}. \tag{8}$$

For equal standards of synthetic speech quality, the gain obtained by using a DPCM strategy over that of PCM coding can be traded off to reduce the rate of information transmission. Of course, for $G < 1$, it is disadvantageous to use DPCM coding. However, for the transmission of parameters that are reasonably smooth in their variation from one transmission frame to the next, it is guaranteed that DPCM coding is superior to PCM coding. In the next section, we demonstrate the efficiency of DPCM techniques for the coding of the linear-prediction speech parameters.

III. DPCM IMPROVEMENT IN CODING LPC PARAMETER

To illustrate the efficiency of DPCM techniques in the coding of the synthesis parameters, Fig. 3 shows the improvement factor G_{opt} in decibels as a function of the number of DPCM predictors. The figure

* The gain is approximate because the effects of the quantizer in Fig. 2 are ignored.

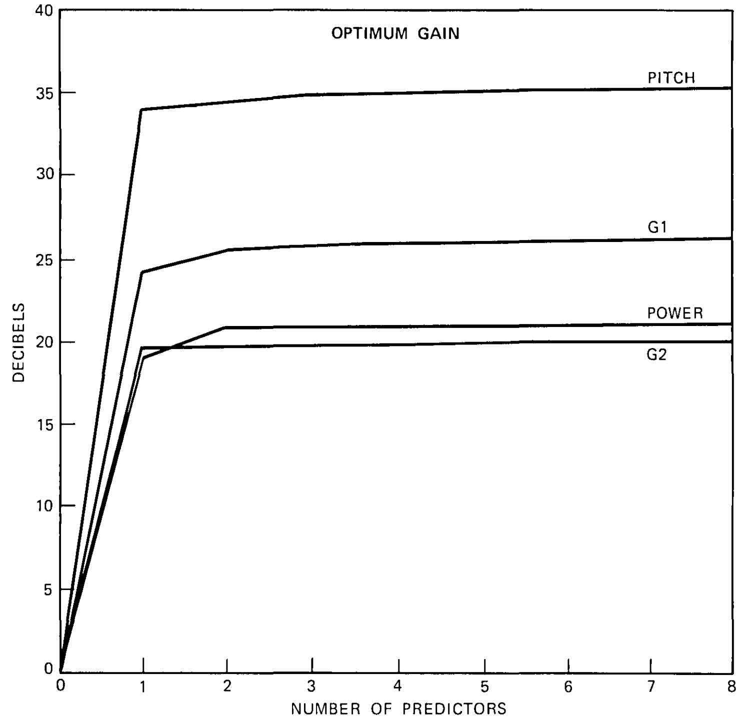


Fig. 3—G_{opt} for the sentence, "May we all learn a yellow lion roar."

shows G_{opt} for the first two log-area coefficients (g_1 and g_2),* pitch period and power† for the all-voiced utterance, "May we all learn a yellow lion roar." The improvement factor was calculated by considering each particular parameter across the entire sentence and then calculating the optimum predictors using eq. (5) and G_{opt} using eq. (7). The results depicted in Fig. 3 are for a male speaker, but the results are typical of those obtained for other male and female speakers. For the complete ensemble of parameters necessary to produce synthetic speech (12 log-area coefficients, pitch, and power),‡ the set of improvement factors were all significantly greater than 1.

Figure 4 shows a typical plot of the improvement factor calculated for a sentence containing unvoiced sounds, "Few thieves are never

* The parameters were calculated at the rate of 50 samples per second. The filter parameter was calculated by the covariance method (Ref. 1), and pitch was measured by a method developed by B. S. Atal (Ref. 8).

† Power is defined as the energy in the speech frame. For the synthetic system employed, it is more convenient to transmit power instead of the amplitude of the input excitation.

‡ Log-area coefficients were transmitted because of their optimum quantization properties.

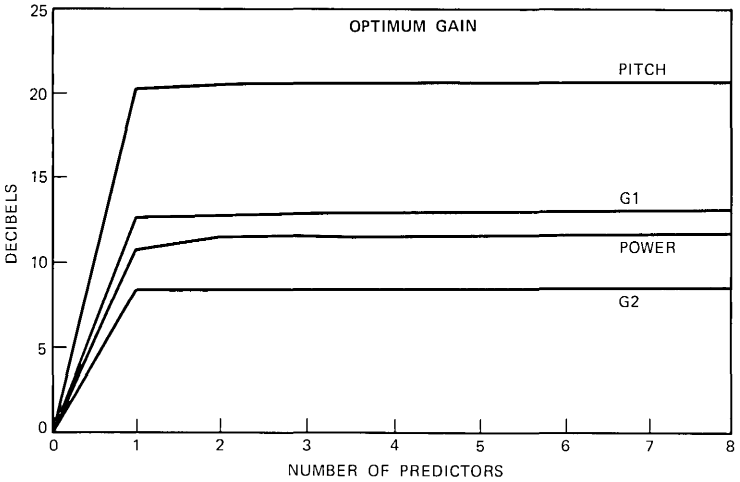


Fig. 4—G_{opt} for the sentence, "Few thieves are never sent to the jug."

sent to the jug." In this sentence, the DPCM improvement over PCM coding is not as dramatic as for the all-voiced sentence. The reason for the decreased values of G_{opt} is that the synthesis parameters tend to change sharply during the unvoiced-voiced transition. Thus, during the transition region there is an abrupt reduction in the correlation between successive samples, and very little information can be gained about the signal from past values. Another reason for the reduced values of G_{opt} is that the variation of the LPC parameters during un-

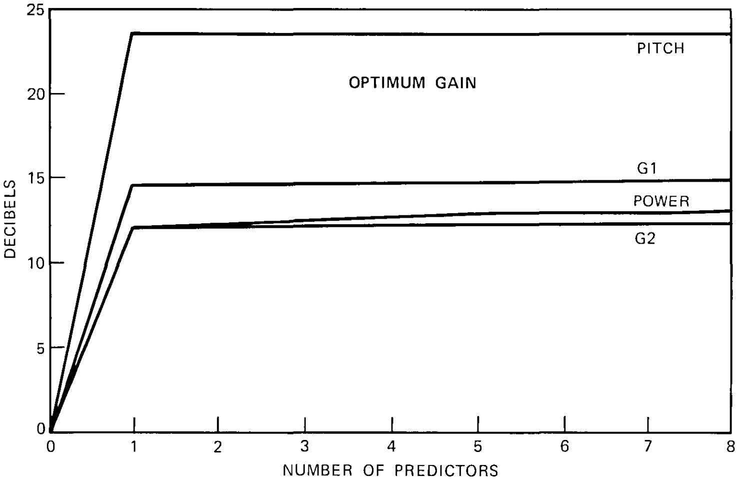


Fig. 5—G_{opt} for the sentence, "Few thieves are never sent to the jug." A separate DPCM analysis is used in each unvoiced and voiced segment.

voiced sounds is inherently more random than during voiced sounds and is thus less predictable. Fortunately, during unvoiced regions the quality of the synthesized speech is more tolerant to quantization noise than during voiced regions.[4] Thus, the diminished values of the G_{opt}'s is not as significant as might at first appear.

One method of increasing the improvement factor for utterances containing unvoiced sounds is to update the DPCM predictors whenever the spectral properties of the speech signal change from unvoiced to voiced sounds. Figure 5 shows G_{opt} for the same sentences as were used to obtain the results of Fig. 4, but in this figure the optimum DPCM predictors were separately calculated for each different section of unvoiced and voiced speech. The improvement factor for this form of DPCM coding is about 5 dB better than a single calculation of the predictors. In a later section of the paper, we discuss another method for updating or adapting the DPCM predictors to the changing spectral properties of the speech signal.

IV. SPEECH SYNTHESIS

4.1 Synthesizer

The improvement factors for the LPC parameters demonstrate that DPCM coding is superior to PCM coding. To confirm the results of the G_{opt} experiments, a synthetic speech system was constructed in the manner illustrated in Fig. 2. To take advantage of the fact that the improvement factor saturates near $n = 1$ (Figs. 3 and 5), only a simple first-order DPCM system was used. The optimum predictor was recomputed for each separate unvoiced and voiced region and the LPC parameters were calculated at a rate of 50 samples per second. The speech was synthesized using the formulation discussed by Atal and Hanauer.[1] After quantization, the parameters were geometrically interpolated (linear interpolation on a logarithmic scale) to allow pitch synchronous resetting of the synthesizer.

The quantizer used in the DPCM coding of the synthesis parameter was a nonuniform quantizer that was designed to exploit the properties of each parameter's error signal. An experimental investigation has indicated that the difference signal for pitch, power, and g_1 are most suitably modeled by a zero mean gamma density,

$$P_{e_r}(e_r) = \frac{\sqrt{k}}{2\sqrt{\pi |e_r|}} \exp\left(-k|e_r|\right), \tag{9}$$

where

$$\sigma = \frac{\sqrt{0.75}}{k}.$$

The higher-order log-area coefficients are more Laplacian in character:

$$P_{e_r}(e_r) = \frac{1}{2\beta} \exp\left(-\frac{|e_r|}{\beta}\right),$$

where

$$\sigma = \sqrt{2}\beta.$$

A signal with a gamma distribution is highly concentrated near its mean, but can also readily achieve values more than three standard deviations from its mean. A Laplacian signal is less concentrated than a gamma signal near its mean value. Figure 6 illustrates the statistical characteristics of a zero mean, unit standard, deviation signal with a gamma density, a Laplacian density, and a gaussian density. Figure 7 shows a comparison between the calculated distributions for the difference signal of several typical synthesis parameters and their approximated distributions.

For a gamma-behaved signal, the properties of the optimum quantizer are summarized in Table I.[9] The x_i values in the table define the ends of quantizer input ranges, and the y_i values are the corresponding outputs. Thus, for a two-bit quantizer, an input between 0 and 1.205 is quantized as 0.302. Similarly, an input between 0.229 and 0.588 for a four-bit scheme is quantized as 0.386. The properties of the optimum quantizer for Laplacian signals are summarized in Table II.[9] Included in these tables is the expected mean square between the difference

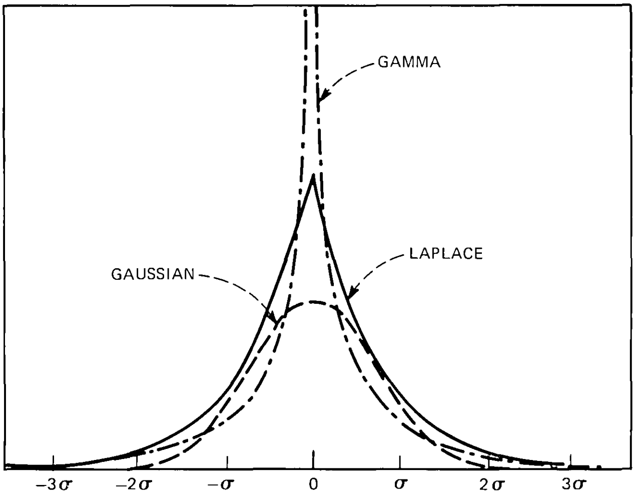


Fig. 6—Comparison of a gaussian, gamma, and Laplacian density with zero mean and unit standard deviation.

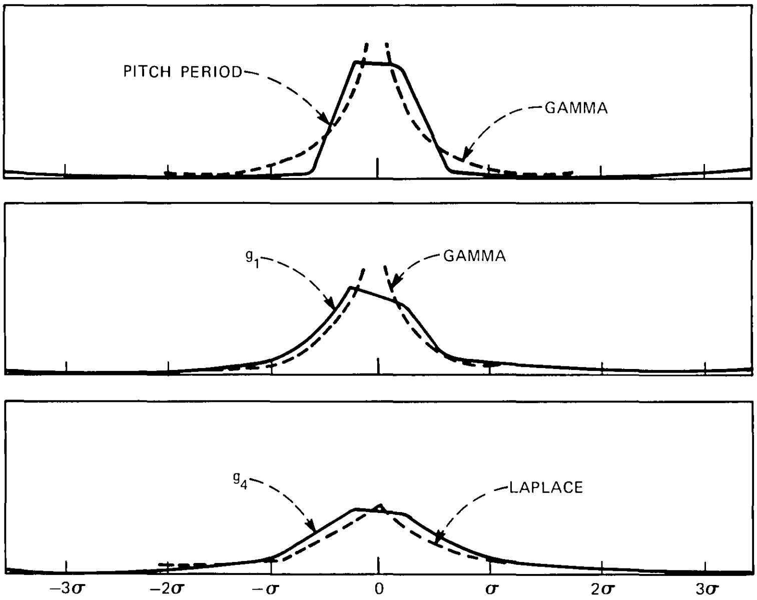


Fig. 7—Comparison between calculated density and approximated density for difference signals.

signal and the quantized difference. Thus, for a four-bit quantization of a gamma signal, the mean square error is 0.0196.

Tables I and II are constructed for signals with unit standard deviation. To obtain the levels y_i and boundaries x_i for signals with standard

Table I — Optimum quantizers for signals with gamma density ($\mu = 0$, $\sigma^2 = 1$)

B	1		2		3		4		5	
i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
1	∞	0.577	1.205	0.302	0.504	0.149	0.229	0.072	0.101	0.033
2			∞	2.108	1.401	0.859	0.588	0.386	0.252	0.169
3					2.872	1.944	1.045	0.791	0.429	0.334
4					∞	3.779	1.623	1.300	0.630	0.523
5							2.372	1.945	0.857	0.737
6							3.407	2.798	1.111	0.976
7							5.050	4.015	1.397	1.245
8							∞	6.085	1.720	1.548
9									2.089	1.892
10									2.517	2.287
11									3.022	2.747
12									3.633	3.296
13									4.404	3.970
14									5.444	4.838
15									7.046	6.050
16									∞	8.043
MSE	0.6680		0.2326		0.0712		0.0196		0.0052	

Table II — Optimum quantizers for signals with Laplace density ($\mu = 0$, $\sigma^2 = 1$)

B	1		2		3		4		5	
i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
1	∞	0.707	1.102	0.395	0.504	0.222	0.266	0.126	0.147	0.072
2			∞	1.810	1.181	0.785	0.566	0.407	0.302	0.222
3					2.285	1.576	0.910	0.726	0.467	0.382
4					∞	2.994	1.317	1.095	0.642	0.551
5							1.821	1.540	0.829	0.732
6							2.499	2.103	1.031	0.926
7							3.605	2.895	1.250	1.136
8							∞	4.316	1.490	1.365
9									1.756	1.616
10									2.055	1.896
11									2.398	2.214
12									2.804	2.583
13									3.305	3.025
14									3.978	3.586
15									5.069	4.371
16									∞	5.768
MSE	0.5		0.1765		0.0548		0.0154		0.00414	

deviation different from unity, simply multiply the given values by the actual standard deviation.[*] The standard deviation for each parameter can be approximated as the rms power of the unquantized error signal. The rms value of the unquantized error signal is obtained directly from the calculation of the optimum DPCM predictors and is given by

$$\sigma^2 = C_0 - \sum_{i=1}^{n} b_iC_i.$$[†]

4.2 Experimental results

Four sentences were synthesized in the experimentation:

A. Few thieves are never sent to the jug.
B. May we all learn a yellow lion roar.
C. It's time we rounded up that herd of Asian cattle.
D. Should we chase those young outlaw cowboys?

High-quality recordings of these sentences were made by two male and two female speakers, and these utterances were used to obtain the analysis data for the DPCM coding method.

[*] To obtain the mean square error, multiply the values by the signal variance.

[†] Since the properties of the unquantized error signal are explicitly known, it is sometimes advantageous to use a more complex nonuniform quantizer to truly optimize the transmission system.

Various schemes were tested for assigning bit rates for each individual error signal. From informal listening experiments, it was determined that synthetic speech that was negligibly different from the unquantized synthesis could be generated according to the following bit assignment:

Pitch:	3 bits/frame
Power:	3 bits/frame
Unvoiced-voiced:	1 bit/frame
g_1:	4 bits/frame
g_2:	4 bits/frame
g_3:	4 bits/frame
g_4:	4 bits/frame
g_5:	3 bits/frame
g_6:	3 bits/frame
g_7:	2 bits/frame
g_8:	2 bits/frame
g_9:	2 bits/frame
g_{10}:	1 bit/frame
g_{11}:	1 bit/frame
g_{12}:	1 bit/frame

The total number of bits dedicated to the complete set of LPC parameter is only 38 bits/frame or 1900 bits/s. On the average, an additional 100 bits/s are required to transmit the necessary DPCM information (DPCM predictors, standard deviations, and initial values of the LPC parameters). As can be observed from Figs. 8, 9, and 10, the spectrogram of the DPCM synthetic speech closely resembles that of the unquantized synthetic speech but requires only a fraction of the storage.

As the bit rate for the DPCM linear prediction vocoder is lowered below the value of 2000 bits/s, the quality of the synthesis slowly begins to deviate from that of the unquantized synthesis. Since the log-area parameters are approximately ordered in terms of their sensitivity, the most expandable bits are those allotted to the lower-ordered g_i's.[4] Depending on the speaker and the utterance, the bit rate can be lowered to between 900 and 1500 bits/s and still allow acceptable quality synthesis.* Figures 11, 12, and 13 illustrate the above examples for a bit-rate of 1400 bits/s (3; 3; 1; 4, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1). The synthetic speech in these examples is slightly degraded from the unquantized synthesis, but the speech is still readily understood and the vocal attributes of the speaker are still apparent. It should be appre-

* Acceptable quality speech synthesis is defined as speech containing all the information content of the original without containing any annoying degradation in speech quality.

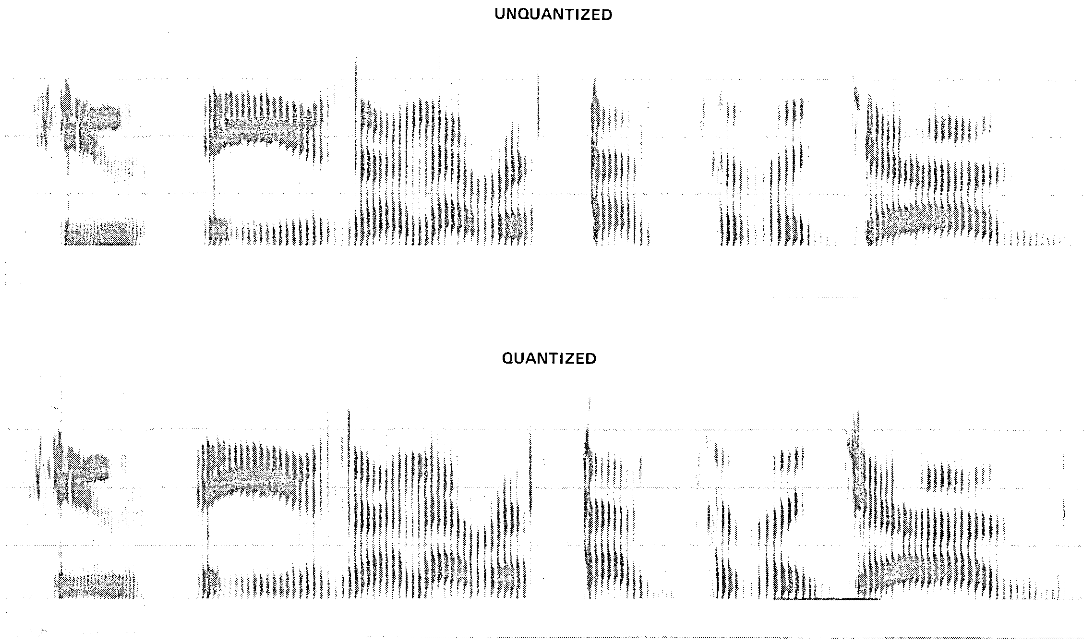


Fig. 8—2000-bits/s quantization of "Few thieves are never sent to the jug." (Male speaker, LG.)

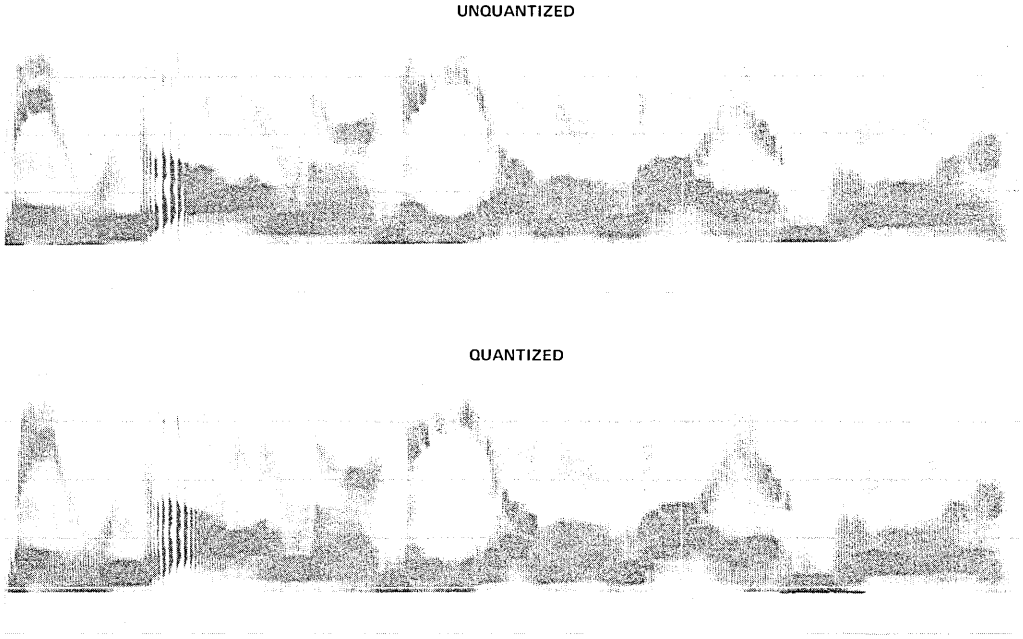


Fig. 9—2000-bits/s quantization of "May we all learn a yellow lion roar." (Female speaker, BG.)

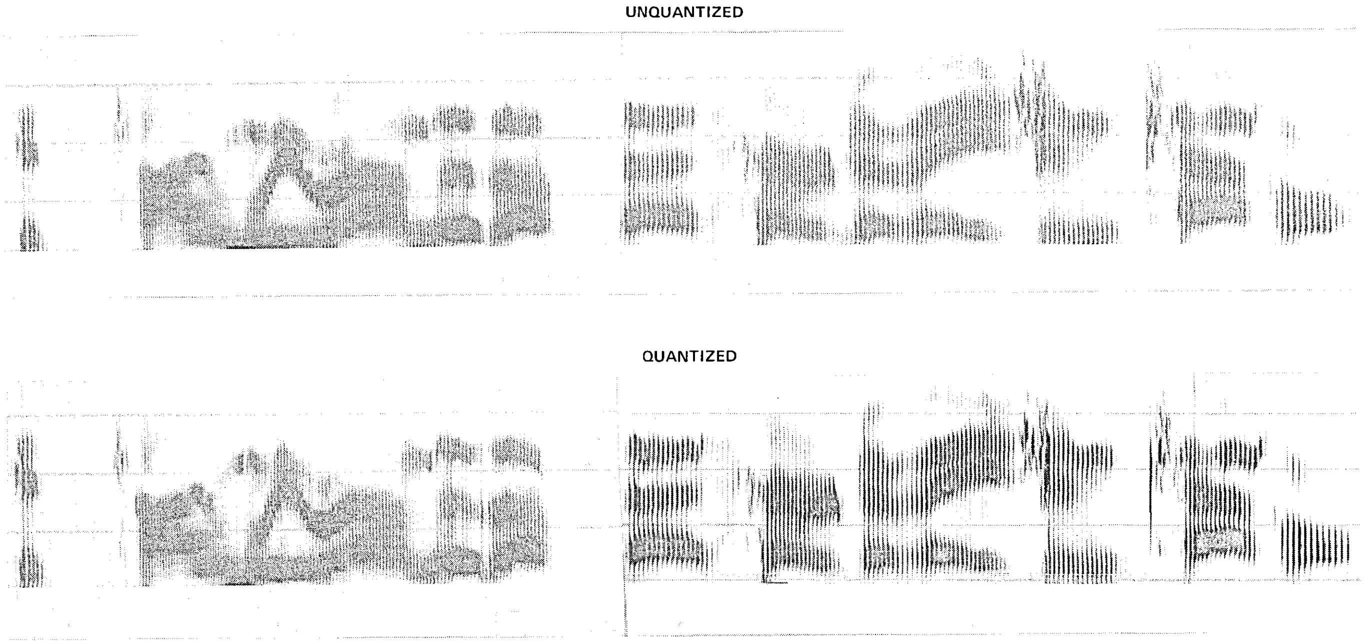


Fig. 10—2000-bits/s quantization of "It's time we rounded up that herd of Asian cattle." (Male speaker, PB.)

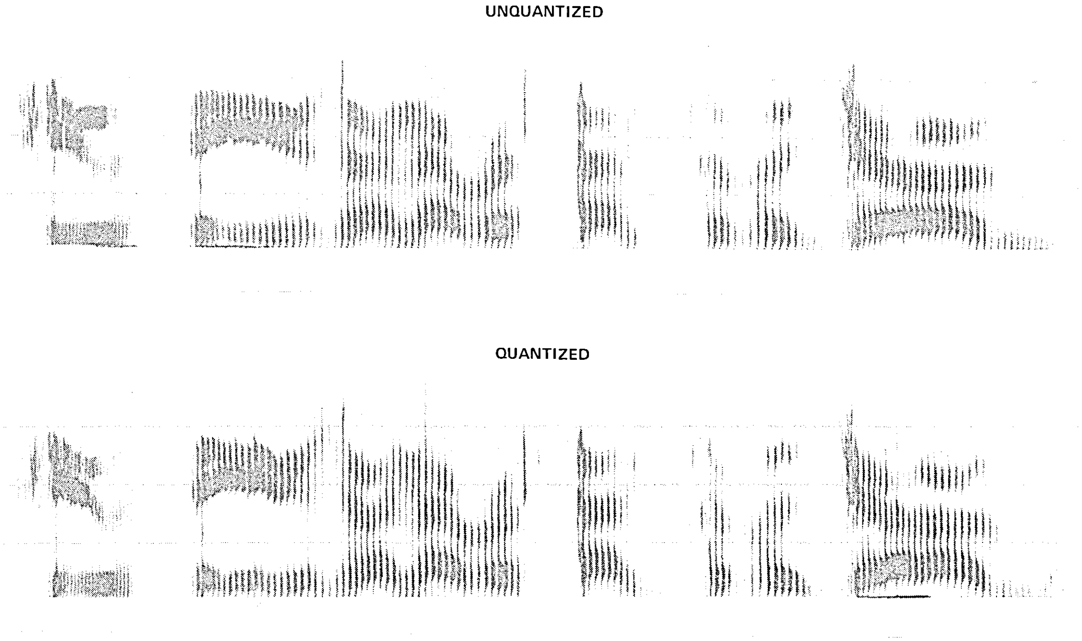


Fig. 11—1400-bits/s quantization of "Few thieves are never sent to the jug." (Male speaker, LG.)

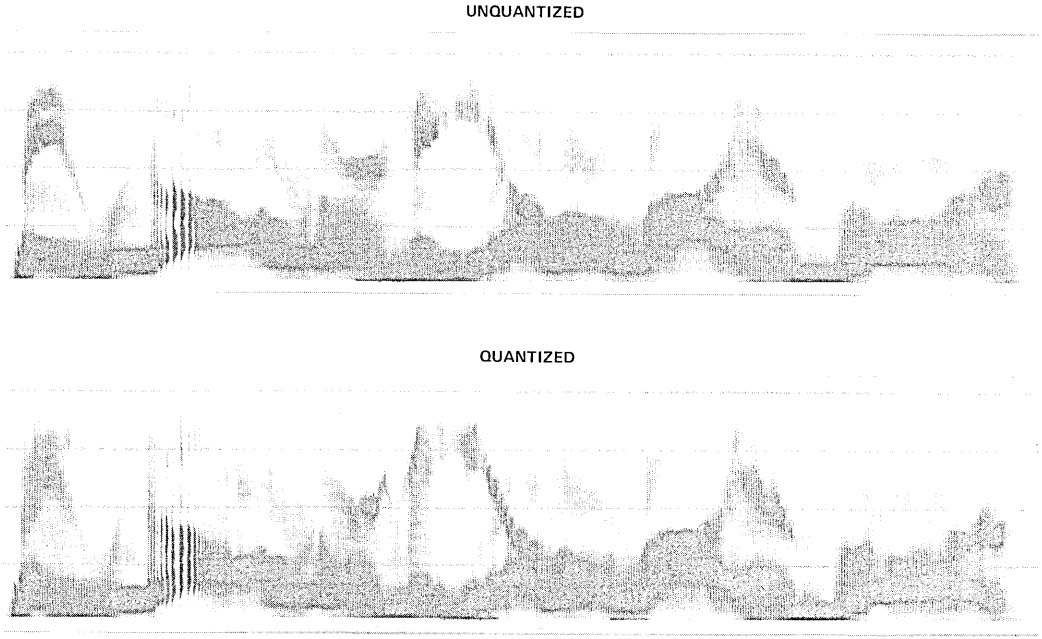


Fig. 12—1400-bits/s quantization of "May we all learn a yellow lion roar." (Female speaker, BM.)

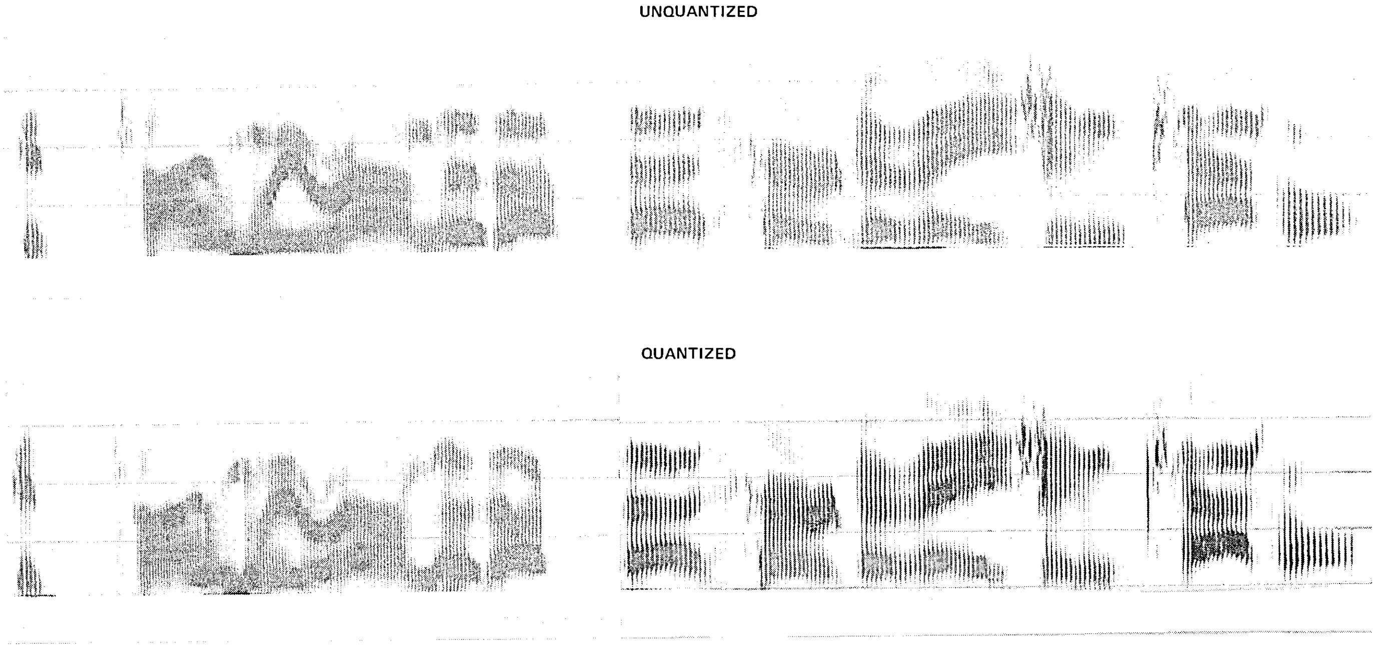


Fig. 13—1400-bits/s quantization of "It's time we rounded up that herd of Asian cattle." (Male speaker, PB.)

ciated that the necessary storage requirements to produce acceptable quality synthetic speech in this method are nearly $\frac{1}{3}$ the requirement for the PCM transmission of the LPC parameters (see Section I).

V. REAL-TIME DPCM TRANSMISSION

The DPCM scheme developed in Section III suffers from the drawback that the calculation of the DPCM predictors and the quantizer step size are delayed until all the LPC parameters are available. For real-time speech synthesis, it is desirable that the process of parameter transmission be done concurrently with the measurement of the LPC parameters. In this section, we discuss several schemes for achieving real-time transmission while still retaining almost the performance of the optimum DPCM strategy.

5.1 Average statistical system

The first means of obtaining a real-time system is based upon the observation that the optimum DPCM first-order predictor for many of the LPC parameters is nearly equal to one [$b_1 = 1.0$ in eq. (4)]. Thus, the optimum linear prediction of the parameter x_r is approximately x_{r-1}. Table III is a comparison of the improvement factors G_{opt} obtained for $b_1 = 1.0$ and b_1 set equal to the optimum value. The overall improvement factors for $b_1 = 1.0$ are not significantly different from the optimum values, and the delay in calculating the optimum b_1 can be avoided by simply letting $b_1 = 1.0$.

To design the optimum quantizer, it is necessary to know the standard deviation of the signal to be quantized. However, our statistical studies have indicated that the standard deviation of the various difference signals are quite stable across different utterances and different speakers. Table IV shows the measured standard deviations for each difference signals computed with $b_1 = 1.0$. Table IV also

Table III — Comparison of G_{opt} in decibels with b_1 set equal to optimum value and $b_1 = 1.0$. Sentence A is "Few thieves are never sent to the jug" and sentence B is "May we all learn a yellow lion roar."

	Pitch	Power	g_1	g_2
Sentence A				
b_1 = Optimum	23.7	12.2	14.7	12.2
b_1 = 1.0	20.2	10.1	14.1	11.0
Sentence B				
b_1 = Optimum	33.8	19.0	24.0	19.6
b_1 = 1.0	33.1	18.8	23.9	19.2

Table IV — Measured standard deviations for the synthesis parameters

	Updated	No Updating
Pitch Period	13.01	16.5
Power	27×10^5	27×10^5
g_1	0.697	0.959
g_2	0.729	0.830
g_3	0.509	0.559
g_4	0.510	0.554
g_5	0.413	0.446
g_6	0.417	0.430
g_7	0.386	0.406
g_8	0.385	0.406
g_9	0.377	0.399
g_{10}	0.346	0.364
g_{11}	0.332	0.342
g_{12}	0.322	0.328

contains the standard deviation for a system in which the prediction scheme is not updated for each unvoiced and voiced region.

Using the standard deviations listed in Table IV and the quantizer discussed in Section IV, a robust transmission scheme is achieved. For example, the difference signal for the pitch period can be accurately quantized for differences as small as two samples or as large as 50 for three-bit quantization.* The synthetic speech quality for the average statistical system compares quite favorably to the optimum scheme, and has the added advantage of real-time implementation.

5.2 Adaptive system

5.2.1 Adaptive DPCM prediction

The DPCM predictors can also be calculated without knowing the entire sequence of parameters by an adaptive method that is based upon the technique of "steepest descent."[11] In this scheme, an initial estimate of the DPCM predictors is determined and then a new set of predictors is calculated to reduce the prediction error. The perturbation in the predictors is in a direction opposite the gradient of the prediction error taken with respect to the DPCM predictor vector. The resulting perturbation is given by

$$\delta^r(b_j) = B \cdot \operatorname{sgn}(e_r) \cdot \tilde{x}_{r-j} / \sum_{k=1}^{n} |\tilde{x}_{r-k}|, \tag{10}$$

where B is the adaptation rate (typically, $B = 0.09$), and $\tilde{x}_r$ is the

* If a nonlinear smoothing algorithm (Ref. 10) is applied to the raw pitch measurements, the variance of the corresponding difference signal is reduced by more than $\frac{1}{2}$. A two-bit quantization can then be used for pitch without diminishing the quality of the synthesis.

quantizer value of the parameter. For the prediction of the $(r+1)$th sample of the parameter, the DPCM predictors are given by

$$b_j^{r+1} = b_j^r + \delta^r(b_j). \tag{11}$$

For a quantizer with $B \geqq 2$, it can be shown that the adaptation scheme will match the changing spectral properties of the speech signal and result in near-optimum performance.[12] For the two methods given above, it should be noted that, in addition to the on-line calculation of the DPCM predictors, it is unnecessary to transmit the predictors.

5.2.2 Adaptive quantization

In the previous section, the quantizer was constructed to take advantage of the known properties or average statistical properties of each parameter's difference signal. In this part of the paper, we introduce an alternate technique for estimating the signal variance. This method is based upon an adaptive approach developed by Cummiskey, Jayant, and Flanagan.[13] In their scheme, a simple uniform quantization of the difference signal is used, but the step size for every new input is varied by a factor depending on which quantizer slot was occupied by the previous sample. Numbering the quantizer slots in the manner shown in Fig. 14, the updated step size Δ_{r+1} is calculated from the previous step size Δ_r by

$$\Delta_{r+1} = \Delta_r \cdot M(|H_r|), \tag{12}$$

where $H_r = 1, 2, \cdots, B$ and the multiplier function $M(\;)$ is a time-invariant function of the quantizer slot number.

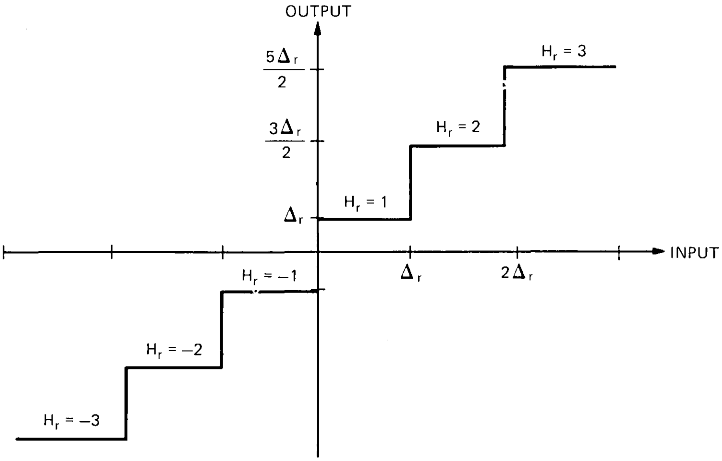


Fig. 14—Numbering of quantizer slots for adaptive quantization.

To adequately match the step size to the signal variance, the multiplier function must be properly chosen. Table V shows the multiplier functions found to be experimentally optimum for quantizing speech waveforms. Using this adaptive scheme (ADPCM) and these multipliers, the quantization of the difference signals can also be efficiently achieved even when the initial step size is a poor estimate of signal variance. Table VI is a comparison of the signal-to-noise ratio for the adaptive scheme with a crude initial estimate of step size and the optimum quantizer discussed in Section IV. The results in Table VI are an encouraging demonstration that it is not necessary to know the statistical structure of the difference signal to efficiently quantize the signal. In fact, it can be shown that, if the properties of the signal are nonstationary, the adaptive method is more suitable than the scheme used in Section IV.

It should be noted that the above scheme does not apply for one-bit quantization ($B = 1$). A simple strategy for one-bit quantization has been developed by Jayant.[14] Let c_r and c_{r-1} denote the values of successive bits in a one-bit scheme, then

$$\Delta_r = \Delta_{r-1} P^{c_r c_{r-1}}, \tag{13}$$

where P has the typical value $P = 1.5$. Although this method was developed for quantizing speech waveforms, it performs quite well in quantizing the parameter difference signals. A comparison of this method and the optimum technique is shown in Table VII. Again, the adaptive scheme works well even with a poor initial estimate of signal variance.

5.3 Synthesis

To subjectively evaluate the performance of the adaptive methods suggested in this section, several speech utterances were synthesized. The synthesis scheme was again the one described by Atal and Hanauer,[1] but an adaptive quantizer and a second-order adaptive predic-

Table V — Step size multipliers for B = 2, 3, and 4 (Ref. 7)

	2	3	4
$M(1)$	0.80	0.90	0.90
$M(2)$	1.60	0.90	0.90
$M(3)$		1.25	0.90
$M(4)$		1.75	0.90
$M(5)$			1.20
$M(6)$			1.60
$M(7)$			2.00
$M(8)$			2.40

Table VI — Comparison of the signal-to-noise ratio for the adaptive quantizer with crude initial estimate of step size and the optimum gaussian signal uniform quantizer. The analysis is for the sentence, "May we all learn a yellow lion roar."

Bits	g_1		g_2		g_3	
	Adaptive	Optimum	Adaptive	Optimum	Adaptive	Optimum
2	12.6	13.6	18.3	18.4	15.6	16.5
3	18.0	20.4	21.8	21.8	19.2	20.0
4	22.8	21.9	24.9	23.9	24.0	23.1

tion DPCM technique was used to transmit the LPC parameters. The initial estimates of the predictors were $b_1 = 1.0$ and $b_2 = 0.0$. A second-order analysis was performed because adaptive prediction makes the G_{opt} function saturate at a larger value than a nonadaptive predictor.[7] The initial estimate of the quantizer step size was set equal to the standard deviations of the parameters listed in Table IV. For parameters in which the quantizer uses only one bit, a first-order system with $b_1 = 1.0$ was used.

Employing the bit assignment cited in Section IV, the quality of the synthetic speech was only slightly worse than the optimum scheme. Figure 15 shows a comparison of one example of the optimum scheme and the adaptive method. To achieve the performance of the optimum scheme, it has been found necessary to allot approximately one bit more per frame to the most sensitive parameters (usually pitch and power).

VI. ORTHOGONAL LINEAR PREDICTION

In the DPCM method of transmission, the value of the parameter x_r is predicted from previous values of the given parameter. However,

Table VII — Comparison of the signal-to-noise ratio for a one-bit adaptive quantizer and optimum one-bit gaussian signal uniform quantizer

g_1		g_2		g_3	
Adaptive	Optimum	Adaptive	Optimum	Adaptive	Optimum
7.3	8.5	8.8	9.9	8.2	9.7

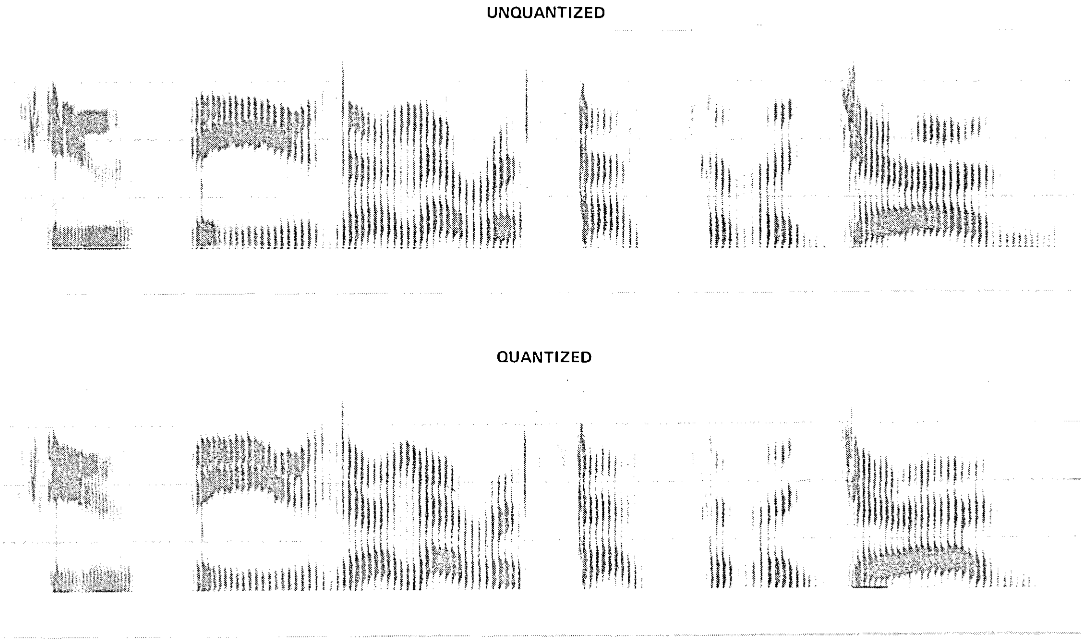


Fig. 15—Adaptive quantization (2000 bits/s) of "Few thieves are never sent to the jug." (Male speaker, LG.)

the LPC parameters have been experimentally determined to be quite redundant.[15] Thus, the parameter x_r can be predicted not only from its own past values but also in some sense from the values of the *other* LPC parameter. A more efficient method of transmission can then be obtained by exploiting all the available information about a given parameter.

One means of exploiting the redundancy among the LPC parameters is to generate a set of orthogonal parameters that are linear combinations of the original set. The new parameters are uniquely (one to one) related to the LPC parameters and are calculated to be independent of each other and therefore do not contain any mutual information. If the original parameters are redundant, only a small subset of the orthogonal parameter will demonstrate any significant frame-to-frame variation. The process of obtaining the appropriate orthogonal parameters is referred to as an eigenvector analysis.[16] The orthogonal parameters are termed eigenvectors, and each vector's statistical variance is termed the eigenvalue of the eigenvector.

To determine the eigenvectors, we first calculate the covariance matrix of the log-area parameters **R** across the utterance. If we denote g_{ij} as the ith log-area parameter in the jth frame, then the elements of R are

$$r_{ik} = \frac{1}{N-1} \sum_{j=1}^{N} (g_{ij} - m_i)(g_{kj} - m_k),$$

where

$$m_i = \frac{1}{N} \sum_{j=1}^{N} g_{ij}$$

and N is the number of frames in the utterance. Given the covariance matrix, the set of eigenvalues λ_i are found by solving the set of simultaneous equations

$$|\mathbf{R} - \lambda\mathbf{I}| = 0,$$

where **I** is the identity matrix and $|\mathbf{A}|$ denotes the determinant of the matrix **A**. The eigenvectors Φ_i are then found as solutions of the equation

$$\lambda_i\Phi_i = \mathbf{R}\Phi_i.$$

To illustrate the behavior of the LPC parameters and the corresponding orthogonal parameters, Table VIII contains a listing of the typical variance (eigenvalues) of each calculated eigenvector parameter across the four utterances examined. The redundancy in the original log-area coefficients is reflected in the fact that more than 90 percent of the total statistical variance is contained in the first five or six eigenvectors.

Table VIII — Measured eigenvalues for the four sentences analyzed:

A. Few thieves are never sent to the jug.
B. May we all learn a yellow lion roar.
C. It's time we rounded up that herd of Asian cattle.
D. Should we chase those young outlaw cowboys?

	A	B	C	D
1	2.62	2.23	1.75	2.75
2	1.44	0.80	1.29	1.58
3	0.67	0.54	0.85	0.6
4	0.44	0.38	0.52	0.36
5	0.25	0.32	0.31	0.28
6	0.21	0.24	0.17	0.22
7	0.10	0.12	0.13	0.16
8	0.09	0.10	0.10	0.15
9	0.08	0.08	0.08	0.09
10	0.06	0.05	0.06	0.06
11	0.03	0.04	0.04	0.06
12	0.02	0.01	0.02	0.03

The higher numbered orthogonal parameters have a relatively small variance and can therefore be considered essentially constant throughout the utterance. Thus, the total information in the 12 log-area parameters can be effectively represented in the space of only the first six eigenvectors.

The redundancy in the LPC parameters is not surprising in view of the fact that the speech signal can be synthesized with only three formant parameters (F_1, F_2, F_3). Thus, the information contained in the 12 log-area coefficients are effectively duplicated in the space of only three formant parameters. The method of orthogonal linear prediction can be viewed as a constraint technique for squeezing the original parameters into a smaller but informationally equivalent set of parameters. The informationally equivalent set is formed by the most significant orthogonal parameters (significance is measured in terms of the standard deviation, or eigenvalue, of the orthogonal parameters).

Experimental studies of a variety of speech utterances have shown that quite acceptable quality synthesis can be generated by transmitting only the six most significant orthogonal parameters, pitch, and power. The synthesis is performed by calculating the LPC parameters from the transmitted orthogonal parameters and *a priori* knowledge of the average values of the least significant orthogonal parameters. For acceptable quality synthesis, only 22 bits/frame are needed.

The allotment of bits was as follows:

Pitch: 3 bits/frame
Power: 3 bits/frame
Unvoiced-voice: 1 bit/frame
First orthogonal parameter: 4 bits/frame
Second orthogonal parameter: 3 bits/frame
Third orthogonal parameter: 3 bits/frame
Fourth orthogonal parameter: 2 bits/frame
Fifth orthogonal parameter: 2 bits/frame
Sixth orthogonal parameter: 1 bit/frame.

The total transmission storage requirement in this technique is 1100 bits/s for the synthesis parameters, 100 bits/s for the DPCM information, and an initial one-time investment of 240 bits* for the necessary eigenvector information. Figures 16 to 18 illustrate the synthetic speech spectrograms generated by this technique for the examples previously examined. Depending on the speaker and the utterance, the bit rate for the synthesis parameters can be reduced to between 600 and 1000 bits/s and still yield acceptable quality speech. The low bit rate required for orthogonal linear prediction is quite attractive, but unfortunately this method involves a complex eigenvector analysis and a delay in transmission to collect the statistical data necessary for the calculation of the eigenvectors.

VII. SUMMARY AND CONCLUSIONS

The goal of this paper was the development of a more efficient method of transmitting the LPC parameters. One proposed method involved the use of DPCM techniques. In DPCM transmission, we take advantage of the predictability of the parameter from its previous values to develop a more effective transmission scheme. Acceptable quality synthetic speech can be generated with DPCM by allotting between 1000 and 1500 bits/s. This rate of information transmission is significantly better than the bit rates necessary for the conventional PCM methods.

To enhance the practical application of the DPCM system, the methods of adaptive quantization and adaptive prediction were discussed. These methods allow the on-line calculation of the DPCM predictors and quantizer step size. To further decrease the storage re-

* Four bits for the average value of each of the six least significant parameters (24 bits) and three bits for each of the 12 coefficients required to compute each orthogonal parameter from the log-area coefficients ($36 \times 6 = 216$ bits).

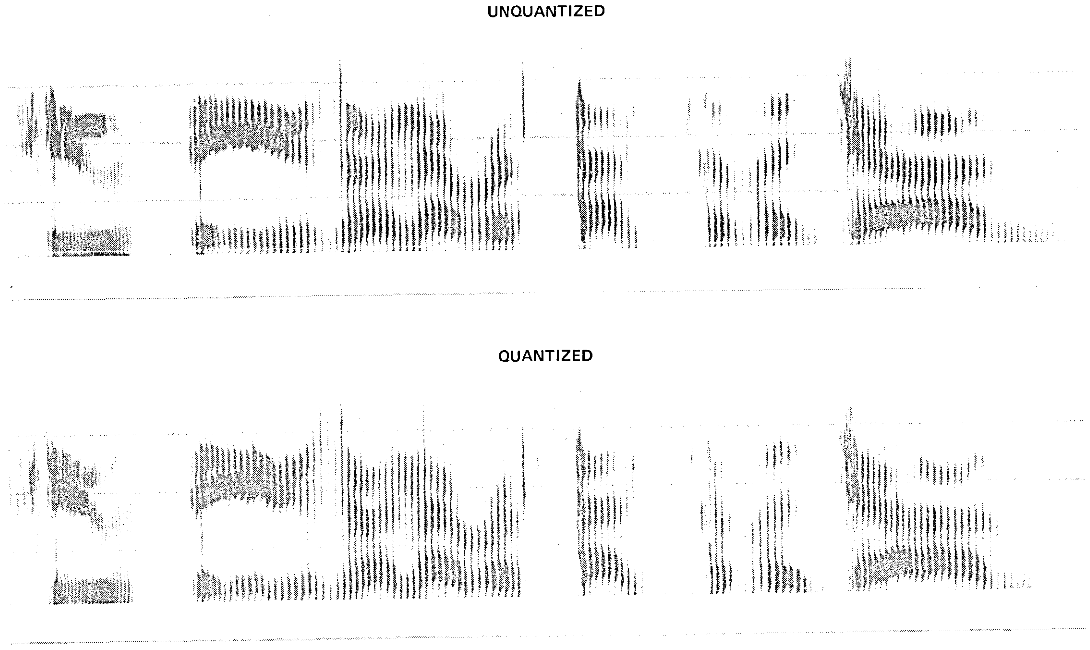


Fig. 16—Quantization of eigenvectors (1200 bits/s) of "Few thieves are never sent to the jug." (Male speaker, LG.)

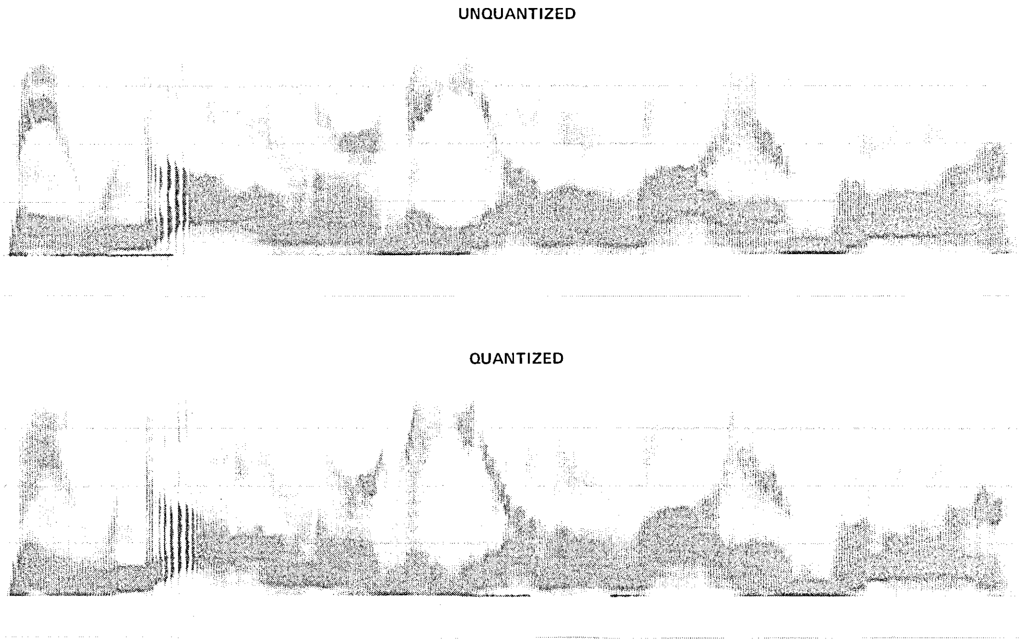
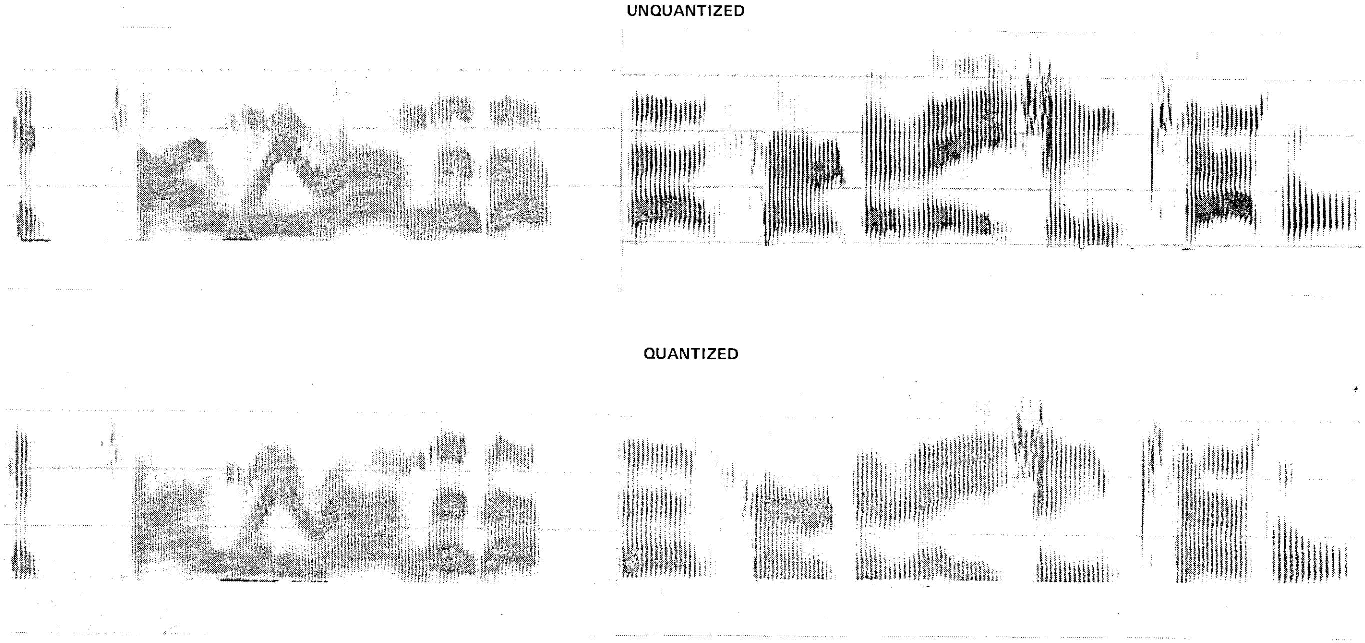


Fig. 17—Quantization of eigenvectors (1200 bits/s) of "May we all learn a yellow lion roar." (Female speaker, BM.)



UNQUANTIZED

QUANTIZED

Fig. 18—Quantization of eigenvectors (1200 bits/s) of "It's time we rounded up that herd of Asian cattle." (Male speaker, PB.)

quirements of the LPC vocoder, the redundancy of the log-area parameters was exploited. By transmitting only the most significant eigenvectors, a considerable saving in bit rate can be achieved.

The techniques discussed in this paper are not limited to the transmission of the LPC parameters, but can also be used in conjunction with other vocoder systems. For example, the bit rate of a formant vocoder[4] can be reduced using a DPCM scheme for transmitting the necessary information. These transmission techniques have wide application and can prove very beneficial in a variety of synthesis schemes.

REFERENCES

1. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., *50*, pp. 637–655, 1971.
2. F. Itakura et al., "An Audio Response Unit Based on Partial Autocorrelation," IEEE Trans. Comm., *COM-20*, No. 4 (August 1972), pp. 792–797.
3. J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method," Trans. Acoustics, Speech and Signal Processing, *ASSP-22*, April 1974, pp. 124–134.
4. J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear Prediction of Speech Theory and Practice," SCRL Monograph No. 10, Santa Barbara, Cal.: Speech Communications Research Lab, Inc., September 1973.
5. J. L. Flanagan et al., "Synthetic Voices for Computers," IEEE Spectrum, *7*, No. 10 (October 1970), pp. 22–45.
6. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," Cambridge, Mass.: Bolt, Beranek and Newman, Inc., BBN Report No. 2800, April 1974.
7. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers." Proc. IEEE, *62*, No. 5 (May 1974), pp. 611–632.
8. B. S. Atal, private communication.
9. M. Paez and T. Glisson, "Minimum Mean Squared Error Quantization in Speech PCM and DCPM Systems," IEEE Trans. Comm., *20* (April 1972), pp. 225–230.
10. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," to appear in Trans. Acoustics, Speech and Signal Processing.
11. P. Cummiskey, "Adaptive Differential Pulse-Code Modulation for Speech Processing," Ph.D. dissertation, Newark College of Engineering, Newark, N. J., 1973.
12. R. W. Stroh, "Optimum and Adaptive Differential PCM," Ph.D. dissertation, Polytechnic Institute of Brooklyn, Farmingdale, N. Y., 1970.
13. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
14. N. S. Jayant, "Adaptive Delta Modulation with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
15. A. E. Rosenberg and M. R. Sambur, "An Improved System for Automatic Speaker Verification," presented at the 86th meeting of the Acoustical Society of America, Los Angeles, October 30–November 2, 1973.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 10, December 1975
Printed in U.S.A.

On the Theory of Self-Resonant Grids

By I. ANDERSON

(Manuscript received May 27, 1975)

An approximate theory is developed to predict the frequency response of a self-resonant grid. The grid is comprised of capacitive and inductive elements and exhibits a band-stop resonance. The analysis is based upon the derivation, from physical considerations, of an equivalent circuit representation of the grid structure. Predicted results compare well with measured data.

I. INTRODUCTION

Arnaud and Pelow[1]* have recently described measurements of the transmission properties of several new types of self-resonant, metal grid structures. These grids, which are readily fabricated by photolithographic techniques, have applications as millimeter-wave quasi-optical filters, or diplexers, in communications satellite antennas and in beam waveguide systems. The grid elements are symmetrical such that the grids may be used with two orthogonal polarizations. In this paper, we derive theoretical expressions for the frequency response of the simplest of the new grids and compare the results with measured data.

The grid to be considered here is a periodic array of "Jerusalem" crosses as shown in Fig. 1a. We wish to determine the grid frequency response in terms of the dimensions of the elements when the planar transmitted wave is incident normally. On account of the complex geometry of the grid elements, an exact treatment as a boundary value problem would be prohibitively difficult. Computer-oriented, numerical techniques[2,3] have provided a powerful means of solution for grid structures in the form of arrays of rectangular, or circular, apertures. The successful application of these techniques requires,[4] however, considerable caution in approximating the unknown aperture fields. When the aperture geometry is complicated, as here, this aspect of the numerical approach poses a considerable difficulty.

Now, in general, the transmission properties of grid structures can be described[5] in terms of an equivalent impedance, together with a

* In eq. (2) of Ref. 1, λ should be replaced by $\lambda/2$.

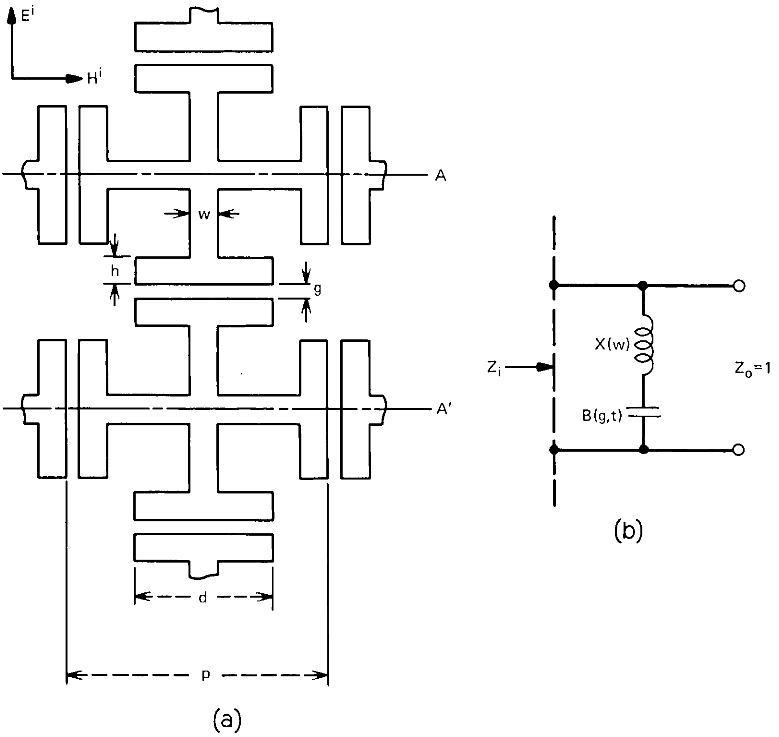


Fig. 1—Jerusalem-cross array and approximate equivalent circuit.

section of transmission line which represents propagation in free space. For example,[6] consider the transmission of a plane wave incident normally upon a grid of thin, perfectly conducting, parallel metal strips of period p. When $p \ll \lambda$, where λ is the wavelength, the equivalent impedance is a shunt inductance, or capacitance, depending upon whether the electric vector of the incident wave is parallel to, or perpendicular to, the edges of the strips. In the following section, an approximate circuit representation of the present grid is derived from physical considerations and from the known results for grids of parallel strips. This approach lends itself to a simple understanding of the grid transmission properties and, furthermore, leads to useful design formulae.

II. ANALYSIS

As shown in Fig. 1a, the period of the array is p, the width of the inductive strips is w, and the separation between adjacent crosses is g. The length and width of the capacitive segments of each cross are

d and h, respectively, and the thickness of the grid is t. It is assumed that

$$t \ll w \ll p, \qquad h \ll p < \lambda, \qquad \text{and} \qquad g \ll d \ll \lambda. \tag{1}$$

The electric field, E^i, is incident normally on the grid with the electric vector directed as shown. For purposes of discussion, we shall refer to this as the "vertical" direction; the incident magnetic field, H^i, is then in the horizontal direction. The effect of the vertical "dipoles," each of length d and width h, at the sides of the crosses is negligible for $d \ll \lambda$. It is therefore assumed that only current that flows parallel to E^i, along the vertical inductive strips and across the horizontal capacitive strips, is significant in determining the field scattered by the grid. On the basis of this assumption, we now consider the magnetic and electric fields in the vicinity of the grid.

Since $w \ll p$ and $h \ll p$, the magnetic field about the grid, due to current flowing along the vertical inductive strips, is approximately the same as that about a corresponding uniform inductive grid of period p and strip width w. Hence, the stored magnetic energy of the Jerusalem-cross grid may be represented approximately by the equivalent inductive reactance, $X(w)$, of this uniform grid, where[7]

$$X(w) = \frac{p}{\lambda}\left\{\ln\left[\operatorname{cosec}\left(\frac{\pi w}{2p}\right)\right] + F(\lambda, w)\right\} \tag{2}$$

and

$$F(\lambda, w) = \frac{Qc^2}{1 + Qs^2} + \left[\frac{pc}{4\lambda}(1 - 3s)\right]^2, \tag{3}$$

with

$$Q = \left[1 - \left(\frac{p}{\lambda}\right)^2\right]^{-\frac{1}{2}} - 1; \qquad c = \cos^2\left(\frac{\pi w}{2p}\right); \qquad s = 1 - c. \tag{4}$$

The reactance $X(w)$ is normalized with respect to the intrinsic impedance of free space. The first term in (2) can be derived[6] from magnetostatic considerations; the second term is a correction factor which is negligible when $p \ll \lambda$. Since $t \ll w$, the effect of thickness upon the inductive reactance is negligible.[8]

With regard to the distribution of electric field, it is noted, from symmetry considerations, that there is no component of electric field normal to the grid on the planes A and A' of Fig. 1a. Without disturbing the electric field we may, therefore, insert a pair of infinitely thin, perfectly conducting plates at A and A' which are perpendicular to the plane of the grid and distance p apart. In the quasi-static case, when $p \ll \lambda$, the electric flux about the grid elements within this parallel-plate transmission line is concentrated between the gaps of the horizontal capacitive segments. We assume this concentration to be maintained at all frequencies for which $p < \lambda$. Since $g \ll d$, the effect of fringing at the extremities of the segments is negligible and the electric flux, per unit width of the parallel-plate line, is d/p times that of a corresponding uniform capacitive grid of period p, gap width g, and thickness t. This implies that the stored electric energy, of the Jerusalem-cross grid, may be represented approximately by an equivalent capacitive susceptance

$$B(g, t) = \frac{d}{p} B_u(g, t), \tag{5}$$

where $B_u(g, t)$ is the (normalized) susceptance of the corresponding uniform grid. For the case $t = 0$ we have[9]

$$B_u(g, 0) = \frac{4p}{\lambda} \left\{ \ln \left[\operatorname{cosec} \left(\frac{\pi g}{2p} \right) \right] + F(\lambda, g) \right\}, \tag{6}$$

where $F(\lambda, g)$ is given by (3) with w replaced by g. The equivalent impedance of a uniform capacitive grid of thickness t includes[6] a segment of transmission line of length t. When $t < 0.5\lambda$, this transmission line may be represented by a Π-network of shunt capacitors and a series inductor. In the present case, $t \ll \lambda$, the series element may be neglected and the total susceptance is[10]*

$$B_u(g, t) = B_u(g, 0) + \frac{2\pi p t}{\lambda g}. \tag{7}$$

The second term in (7) may be derived equivalently by considering the additional (parallel-plate) capacitance introduced by the finite thickness of a capacitive diaphragm in a parallel-plate transmission line of height p. From (5), (6), and (7), the capacitive susceptance of the Jerusalem-cross grid is approximately

$$B(g, t) = \frac{4d}{\lambda} \left\{ \ln \left[\operatorname{cosec} \left(\frac{\pi g}{2p} \right) \right] + F(\lambda, g) + \frac{\pi t}{2g} \right\}. \tag{8}$$

We have obtained approximate values of reactances with which to describe the stored magnetic and electric energies of the grid and now consider the equivalent circuit representation. It has been assumed that only current that flows vertically along the inductive strips, and across the gaps of the horizontal capacitive segments, is significant in determining the transmission properties of the grid. This suggests that the Jerusalem-cross grid can be represented approximately by a

* In Ref. 10, the sign of the second term for B_1 in eq. (83) on p. 200 should be positive.

reactance, X_g, where

$$X_g = X(w) - \frac{1}{B(g, t)}, \tag{9}$$

shunted across a transmission line of impedance Z_o as shown in Fig. 1b. The impedances in the equivalent circuit are normalized with respect to the impedance (Z_o) of free space, and the impedance seen by a plane wave incident normally on the grid is Z_i. The grid transmission coefficient is now expressible in terms of the grid reactance X_g. The input impedance, Z_i, is

$$Z_i = \frac{jX_g}{1 + jX_g}. \tag{10}$$

The corresponding voltage reflection coefficient, R, is

$$R = \frac{Z_i - 1}{Z_i + 1} \tag{11}$$

and the grid power transmission, $|T|^2$, is

$$|T|^2 = 1 - |R|^2 = \frac{4X_g^2}{1 + 4X_g^2}. \tag{12}$$

Substituting (2) and (8) into (9) and (12) then gives the grid transmission response in terms of its dimensions. To the present order of approximation, $|T|^2$ is seen to be independent of h when $h \ll p$.

A first-order approximation for the rejection wavelength λ_r, defined by the equation $X_g = 0$, is readily found by assuming $p \ll \lambda$ so that the terms $F(\lambda, w)$ in (2) and $F(\lambda, g)$ in (8) may be neglected. Furthermore, if the effect of grid thickness is also neglected, by putting $t = 0$, and if the cosecants are replaced by the small argument forms, we find

$$X(w) \approx \frac{p}{\lambda} \ln \left(\frac{2p}{\pi w} \right), \qquad p \ll \lambda \tag{13}$$

$$B(g, t) \approx \frac{4d}{\lambda} \ln \left(\frac{2p}{\pi g} \right), \qquad p \ll \lambda, t = 0. \tag{14}$$

From (9), the wavelength, λ_r, at the rejection resonance is then

$$\lambda_r \approx 2 \sqrt{dp \ln \left(\frac{2p}{\pi w} \right) \ln \left(\frac{2p}{\pi g} \right)}. \tag{15}$$

This result provides an approximate functional dependence of the resonant wavelength upon the grid geometry.

The effect of the dielectric sheet which supports the grid has been neglected in the preceding analysis. In general, the presence of an adjacent, low-loss, dielectric layer will increase the grid susceptance,

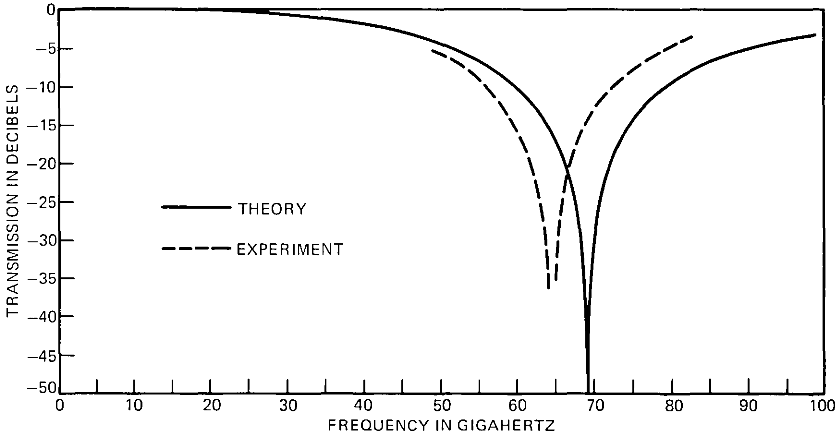


Fig. 2—Transmission response of Jerusalem-cross array.

$B(g, t)$, by modifying the electric field in the vicinity of the capacitive gaps. In the case examined by Arnaud and Pelow,[1] however, the sheet, which has a relative permittivity of about 2.5, is thinner than the grid itself and, as such, is not expected to modify the grid transmission to the present order of approximation.

III. COMPARISON WITH MEASUREMENTS

The grid measurements of Arnaud and Pelow[1] were conducted under approximately plane wave conditions and for a range of incidence angles from 5 to 45 degrees.* It was found that the frequency of the rejection resonance, and the shape of the transmission response, were practically independent of the angle of incidence within this range. Figure 2 shows the predicted frequency response for normal incidence, as obtained from (12), compared with measured data for an incidence angle of 5 degrees. The experimental curve is from Fig. 3 of Arnaud and Pelow's paper and is for a grid of dimensions $p = 1.400$ mm, $d = 0.750$ mm, $w = h = 0.180$ mm, $g = 0.090$ mm, and $t = 0.018$ mm. The shape of the transmission response is predicted well by the theory; the error in the prediction of the rejection frequency is 7 percent. The first-order expression (15) for the rejection wavelength is within 10 percent of the value obtained from (12).

IV. CONCLUSIONS

We have examined the transmission properties of a self-resonant grid that is comprised of capacitive and inductive elements. An

* No measurements were taken at exactly normal incidence to avoid multiple reflections within the measuring system.

approximate theory has been developed to predict the frequency response of the grid when illuminated by a plane wave at normal incidence. The theory is based upon the construction of an appropriate equivalent circuit in which the values of the reactances are obtained by modification of known solutions for simple, parallel strip grids. A comparison of results with measured data shows an error of 7 percent in the prediction of the grid rejection resonance. By way of comparison, the corresponding approximate expressions (2) and (6), for parallel strip grids as obtained from rigorous analyses,[7] can be in error by about 1 to 5 percent over the range of frequencies considered here.

V. ACKNOWLEDGMENTS

Discussions with J. A. Arnaud and A. A. M. Saleh, together with the valuable comments of J. W. Carlin, are appreciated.

REFERENCES

1. J. A. Arnaud and F. A. Pelow, "Resonant-Grid Quasi-Optical Diplexers," B.S.T.J., *55*, No. 2 (February 1975), pp. 263–283.
2. C. C. Chen, "Transmission of Microwave Through Perforated Flat Plates of Finite Thickness," IEEE Trans. Microwave Theory and Techniques, *MTT-21*, No. 1 (January 1973), pp. 1–6. (Also other papers cited as references.)
3. R. F. Harrington, *Field Computation by Moment Methods*, New York: MacMillan, 1968.
4. R. Mittra, T. Itoh, T. S. Li, "Analytical and Numerical Studies of the Relative Convergence Problem Arising in the Solution of an Integral Equation by the Moment Method," IEEE Trans. Microwave Theory and Techniques, *MTT-20*, No. 2 (February 1972), pp. 96–104.
5. H. G. Booker, "The Elements of Wave Propagation Using the Impedance Concept," J. IEE, Pt. 3, *94* (May 1947), pp. 171–202.
6. G. G. MacFarlane, "Quasi-Stationary Field Theory and its Application to Diaphragms and Junctions in Transmission Lines and Wave Guides," J. IEE, Pt. 3A, *93*, 1946, pp. 703–719.
7. N. Marcuvitz, *Waveguide Handbook*, M.I.T. Rad. Lab. Ser., No. 10, New York: McGraw-Hill, 1951, pp. 218 and 284–285.
8. C. G. Montgomery, R. H. Dicke, and E. M. Purcell, *Principles of Microwave Circuits*, M.I.T. Rad. Lab. Ser., No. 8, New York: McGraw-Hill, 1948, pp. 166–167.
9. Ref. 7, pp. 218 and 280.
10. Ref. 8, pp. 188 and 200. See also Fig. 14 of Ref. 6.

Copyright © 1975 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 10, December 1975
Printed in U.S.A.

An Approximate Method for Calculating Delays for a Family of Cyclic-Type Queues

By S. HALFIN

(Manuscript received May 22, 1975)

A study of the marker-register dial-tone delay problem in No. 5 crossbar switching machines led to a special type of cyclic queuinq model. In this paper, we present a method for calculating approximately the steady-state delays of an arriving customer. When applied to the marker-register problem, the model emphasizes the order in which markers are assigned to waiting calls and the fact that part of the markers' time is unproductive when an "all registers busy" condition occurs. Some numerical results are presented, which agree with the observed phenomenon that, for a constant marker load, the delays of Touch-Tone® *calls are influenced by the load on the dial-pulse originating registers, and vice versa. The results are compared to those of a simulation of the same problem. The numerical results compare favorably in the range of loads that produce a dial-tone speed of between 0.05 and 0.15.*

I. INTRODUCTION

The queuing model described in this paper resulted from a study of the marker-register dial-tone delay problem in No. 5 crossbar switching machines. A number of queues with Poisson arrivals of equal rates are served in a cyclic order by a server with constant service time. Upon arriving at a nonempty queue, the server chooses a customer from the queue at random. After one service time, the customer either leaves the system with a certain predetermined probability or rejoins his queue. In both cases, the server uses a fixed amount of time and moves to the next queue; thus, at most one customer leaves the system following each arrival of the server at a queue.

Related models were treated by Cooper,[1] Cooper and Murray,[2] and Eisenberg.[3] In Refs. 1 and 2, the server either empties the queue being served or serves all those present at the queue in its arrival epoch. The case of two queues with different arrival rates is treated in Ref. 3. In these papers, the Laplace-Stieltjes transforms of the waiting time distributions were obtained. Attempts to obtain the distributions for similar models by approximate methods were made by Leibowitz[4] and Schay, Jr.[5]

The method presented here approximately calculates the steady-state delays of an arriving customer. The approximation is carried out by modifying the model to achieve a manageable state space. Next, the method is applied to the problem of dial-tone delay in the No. 5 crossbar switching machine. The order in which markers are assigned to waiting calls and the fact that part of the markers' time is unproductive when an "all registers busy" condition occurs are emphasized in the model. The numerical results presented agree with the observed phenomenon that, for a constant marker load, the delays of *Touch-Tone®* calls are influenced by the load on the dial-pulse originating registers, and vice versa. The results are compared to those of a simulation of the same problem.

Several features of the No. 5 crossbar machine, which may have an influence on the dial-tone speed, were excluded. Some of these features were investigated in subsequent work by H. A. Guess[6] and are described in more detail in Section XVI of this paper.

II. THE MODEL

Let N queues $E_1, E_2, \cdots, E_N$, $N \geqq 2$, be given. Customer arrivals to each queue constitute independent Poisson processes, all having the same rate λ. The queues are served by a single server in the following way: At each point in time, the server is at some queue. Transitions in its position occur at discrete time epochs which are equally spaced with periods of duration T. At such time epochs, the server moves instantaneously to the next nonempty queue in a circular order, chooses a waiting customer in that queue at random, and stays with this customer until the end of the period. The served customer then leaves the queue with probability p^*, or remains in the queue with probability $q^* = 1 - p^*$. If all the queues are empty at a transition point, then no change occurs in the position of the server.

We assume further that T is small with respect to the accuracy with which we want to know the delays, and thus all the arrivals can be assumed to occur at the transition epochs and the queuing process can be considered in discrete time. Thus, if $X_{i,k}$ is the number of arrivals to E_i at the kth time epoch, then all $X_{i,k}$, $1 \leqq i \leqq N, k \geqq 0$, are independent identically distributed random variables, all having a Poisson distribution with mean λT.

III. THE FULL STATE SPACE

The state of the queuing system at any time epoch is defined as the $(N + 1)$tuple $(m_1, m_2, \cdots, m_N, n)$, where m_i is the number of waiting customers at E_i, and n is the position of the server *before* the transition. We call the set of all such states the full state space. It is then clear from the discussion in the previous section that our queuing system, with the full state space, is a stationary Markov chain. Thus, in principle, one can calculate transient and steady-state probabilities. However, the full state space is much too large for practical computational purposes.

Consider, for instance, the dial-tone delay problem where a typical number of N is 15. Then, even if the system is so underloaded that each m_i can be restricted to be either 0 or 1, we have $2^{15} \times 15 \sim 500{,}000$ states. A natural approach, which we follow in the remainder of the paper, is to approximate the behavior of our system with systems having a smaller size state space.

IV. THE "BLACK BOX" APPROACH

Some important facts about the system can be deduced by considering only the total number of customers in all the queues $s = m_1 + m_2 + \cdots + m_N$. It is clear that the system with this single state is again a stationary Markov chain. It is, in fact, a discrete version of an $M/D/1$ queue, with the added feature that a customer who is held by the server returns to the queue with probability q^*. Alternatively, the service time measured in units of T may be considered as having a geometrical distribution with mean $1/p^*$. The traffic intensity of the system is the $\rho = \lambda NT/p^*$; thus, we have Theorem 1.

Theorem 1: A necessary and sufficient condition for the nonsaturation of the system is $\lambda NT < p^*$.

Note that, because of the symmetry, a particular queue in the system is saturated if and only if the system as a whole is saturated; hence, Theorem 1 provides a saturation condition for all the individual queues.

Let us denote by A_i^* the probability that a Poisson-distributed random variable with mean λNT attains the value i. The system has the following transition probabilities. For $s > 0$,

$$\Pr\,(s \rightarrow s') = A^*_{s'-s+1}p^* + A^*_{s'-s}q^*$$

and

$$\Pr\,(0 \rightarrow s') = A_{s'}.$$

The equations for the steady-state probabilities P_s are then

$$P_{s'} = A^*_{s'}P_0 + \sum_{s=1}^{s'+1} (A^*_{s'-s+1}p^* + A^*_{s'-s}q^*)P_s \qquad s' = 0, 1, \cdots. \quad (1)$$

Equations (1) can be solved recursively, starting with $P_0 = 1 - \rho$.

One can also calculate the generating function

$$\hat{P}(u) = \sum_{s=0}^{\infty} P_s u^s$$

$$= \frac{(p^* - \alpha)(1 - u)}{uq^* + p^* - ue^{\alpha(1-u)}},$$

where $\alpha = \lambda NT$. By evaluating $\hat{P}'(1)$, the expected value of the total number of customers in the system $\bar{s}$ is

$$\bar{s} = \frac{2\alpha - \alpha^2}{2(p^* - \alpha)}. \tag{2}$$

V. THE MODIFIED MODEL

We now consider our original system with a new state space. A state will now consist of a triplet (m, M, n), where $m = m_1$, $M = m_2 + m_3 + \cdots + m_N$, and n is the same as previously, namely the position of the server before a transition. It is clear that the new state space is much smaller than the full state space; however, the Markovian property is lost. This can be seen by the following argument: If M was positive at time $k - 1$, and if in the transition between $k - 1$ to k the server skipped a large number of queues, then those M customers were concentrated in the remaining queues; thus, it is probable that, in the k to $k + 1$ transition, a small number of queues will be skipped.

At this point, we modify our model to make it a Markov chain with respect to the new state space. To do this, we need to define one-step transition probabilities so that the behavior of the modified model will approximate that of the original model. Let the position of a customer be the queue number where he waits. Our key assumption concerns the probability distribution of the positions of the M customers, given the state (m, M, n).

For the remainder of this section, let us enumerate the queues $E_2, \cdots, E_N$ by starting with the queue following the position of the server and observing the cyclic order, skipping E_1. Next, we make the following assumptions: Let $M > 0$; then

(*i*) The positions of the M customers are independent, identically distributed, random variables.

(*ii*) The probability that any one of the M customers will be in the ith queue (in the new order) is $\pi_i(M)$, where

$$\pi_i(M) = b(M) + \frac{(N - i - 1)R(M)}{M(N - 2)} \qquad i = 1, 2, \cdots, N - 1,$$

where $b(M) = 1/(N - 1) - [R(M)/2M]$ and determination of $R(M)$ is described below.

The rationale behind assumption (*ii*) is that a customer is less likely to be in a queue that has just recently been visited. The average difference between the number of customers in the first and last queues is, according to assumption (*ii*), $M[\pi_1(M) - \pi_{N-1}(M)] = R(M)$, which should be approximately equal to the expected number of arrivals during one full cycle of the server. Thus,

$$R(M) \sim \lambda T R_0(M),$$

where $R_0(M)$ is the expected number of nonempty queues, given M. Finally, we approximate $R_0(M)$ by

$$R_0(M) = (N-1)\left[1 - \left(\frac{N-2}{N-1}\right)^M\right],$$

where the right-hand side is the expected number of nonempty queues among $E_2, \cdots, E_N$, if the M customers are uniformly distributed. We conclude this section by using the new assumptions to calculate some probabilities that will be needed later.

Let J denote the number of successive empty queues following the position of the server (when E_1 is disregarded), $J = 0, 1, \cdots, N-1$.

The distribution of J depends on M. Let

$$\begin{aligned} Q_j(M) &= \Pr(J \geqq j-1) \\ q_j(M) &= \Pr(J = j-1) \\ j &= 1, 2, \cdots, N. \end{aligned}$$

Using assumptions (*i*) and (*ii*), we have

$$Q_j(M) = \left[\sum_{i=j}^{N-1} \pi_i(M)\right]^M \qquad j = 1, \cdots, N-1$$

$$Q_N(M) = \begin{cases} 0 & \text{if} \quad M > 0 \\ 1 & \text{if} \quad M = 0, \end{cases}$$

and

$$\begin{aligned} q_j(M) &= Q_j(M) - Q_{j+1}(M) \qquad j = 1, \cdots, N-1 \\ q_N(M) &= Q_N(M). \end{aligned}$$

VI. TRANSITION PROBABILITIES FOR THE MODIFIED SYSTEM

Given a state (m, M, n), the transition probabilities of the position of the server can be expressed in terms of the $Q_j(M)$ and $q_j(M)$'s as follows.

For $m > 0$,

$$\Pr\,(n \rightarrow n') = \begin{cases} q_{n'-n}(M) & \text{if} \quad n' > n \\ Q_{N-n+1}(M) & \text{if} \quad n' = 1 \\ 0 & \text{if} \quad 1 < n' \leqq n. \end{cases}$$

For $m = 0$, $M > 0$,

$$\Pr\,(n \rightarrow n') = \begin{cases} q_{n'-n}(M) & \text{if} \quad n' > n \\ 0 & \text{if} \quad n' = 1 \\ q_{N+n'-n-1} & \text{if} \quad 1 < n' \leqq n, \end{cases} \tag{3}$$

and, for $m = 0$, $M = 0$,

$$\Pr\,(n \rightarrow n') = \begin{cases} 1 & \text{if} \quad n' = n \\ 0 & \text{if} \quad n' \neq n. \end{cases}$$

Let us denote by A_i and a_i the probability that Poisson-distributed random variables with means $\lambda(N - 1)T$ and λT, respectively, attain the value i. We can now write the state transition probabilities.

For $m > 0$, $M > 0$:

$$\Pr\,[(m, M, n) \rightarrow (m', M', n')] = \begin{cases} a_{m'-m}q_{n'-n}(M)[A_{M'-M+1}p^* + A_{M'-M}q^*] & \text{if} \quad n' > n \\ A_{M'-M}Q_{N+1-n}(M)[a_{m'-m+1}p^* + a_{m'-m}q^*] & \text{if} \quad n' = 1, \\ 0 & \text{otherwise.} \end{cases}$$

For $m = 0$, $M > 0$:

$$\Pr\,[(0, M, n) \rightarrow (m', M', n')] = a_{m'}[A_{M'-M+1}p^* + A_{M'-M}q^*] \times \begin{cases} q_{n'-n}(M) & \text{if} \quad n' > n \\ 0 & \text{if} \quad n' = 1 \\ q_{N+n'-n-1}(M) & \text{if} \quad 1 < n' \leqq n. \end{cases}$$

For $m > 0$, $M = 0$:

$$\Pr\,[(m, 0, n) \rightarrow (m', M', n')] = \begin{cases} A_{M'}[a_{m'-m+1}p^* + a_{m'-m}q^*] & \text{if} \quad n' = 1, \\ 0 & \text{if} \quad n' \neq 1, \end{cases}$$

and finally, for $m = M = 0$,

$$\Pr\,[(0, 0, n) \rightarrow (m', M', n')] = \begin{cases} a_{m'}\,A_{M'} & \text{if} \quad n' = n \\ 0 & \text{if} \quad n' \neq n. \end{cases} \tag{4}$$

VII. STEADY-STATE EQUATIONS FOR THE MODIFIED SYSTEM

If we consider the total number of customers in the system $s = m + M$, then it is clear that all the results of Section IV remain valid for the modified system. In particular, $\lambda NT < p^*$ is the necessary and sufficient condition for the existence of a steady state. In the following discussion, we assume that this condition is satisfied. Let $P(m, M, n)$ be the steady-state probability of the state (m, M, n) satisfying the following equations.

For $n' > 1$, $m' \geqq 0$, $M' \geqq 0$:

$$\begin{aligned} P(m', M', n') &= \sum_{n=1}^{n'-1} \sum_{m=0}^{m'} \sum_{M=1}^{M'+1} a_{m'-m} q_{n'-n}(M)[A_{M'-M+1}p^* + A_{M'-M}q^*] \\ &\quad \times P(m, M, n) + \sum_{n=n'}^{N} \sum_{M=1}^{M'+1} a_{m'}[A_{M'-M+1}p^* + A_{M'-M}q^*] \\ &\quad \times q_{N+n'-n-1}(M) P(0, M, n) + a_{m'} A_{M'} P(0, 0, n'). \qquad (5) \end{aligned}$$

For $n' = 1$, $m' \geqq 0$, $M' \geqq 0$:

$$\begin{aligned} P(m', M', 1) &= \sum_{n=1}^{N} \sum_{m=1}^{m'+1} \sum_{M=0}^{M'} A_{M'-M} Q_{N+1-n}(M)[a_{m'-m+1}p^* + a_{m'-m}q^*] \\ &\quad \times P(m, M, n) + a_{m'} A_{M'} P(0, 0, 1). \end{aligned}$$

VIII. NUMERICAL SOLUTION OF THE STEADY-STATE EQUATIONS

Equations (5) can be solved either as a system of linear equations with the auxiliary equation

$$\sum_{m,M,n} P(m, M, n) = 1$$

or by starting with any initial distribution and iterating it through (4) until a desired degree of convergence is obtained.

It seems adequate to adopt the second method since the steady state of the modified system is of interest to us only as an approximation to the steady state of the original system. Thus, an extensive computational effort to obtain an accurate solution to (4) is not warranted. A good initial distribution to start the iterations can be obtained in the following way. We have

$$\sum_{m+M=s} \sum_{n=1}^{N} P(m, M, n) = P_s \qquad s = 0, 1, \cdots,$$

where the P_s were computed by the method outlined in Section IV. If we divide the s customers uniformly among the queues and make the position of the server random, we get the following distribution:

$$P_0(m, s - m, n) = \frac{1}{N} B\left(m, \frac{1}{n}, s\right) P_s$$

$$s = 0, 1, \cdots \qquad m = 0, \cdots, s \qquad n = 1, \cdots, N,$$

where $B(m, 1/N, s)$ is the binomial probability of m successes out of s trials, with probability of success $= 1/N$.

IX. EVALUATION OF THE DELAYS FOR THE MODIFIED SYSTEM

Let $f(m, M, n, k)$ be the probability that a customer arriving at E_1 will wait k service periods before leaving the system, given that on his arrival the system went into the state (m, M, n). The customer stays for at least one service period, so $k \geqq 1$. Notice also that $m \geqq 1$.

Theorem 2: The delay probabilities satisfy the following recursive equations:

$$f(m, M, n, 1) = \frac{p^*}{m} Q_{N-n+1}(M)$$

and

$$f(m, M, n, k+1) = \sigma_1 + \sigma_2$$

$$m \geqq 1, \quad M \geqq 0, \quad 1 \leqq n \leqq N; \qquad k \geqq 1,$$

where

$$\sigma_1 = \begin{cases} \displaystyle\sum_{n'=n+1}^{n} \sum_{M'=M-1}^{\infty} \sum_{m'=m}^{\infty} a_{m'-m} q_{n'-n}(M)[A_{M'-M+1}p^* + A_{M'-M}q^*] \\ \qquad\qquad \times f(m', M', n', k) \qquad \text{if} \quad M > 0,\ n < N \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{if} \quad M = 0 \text{ or } n = N \end{cases}$$

$$\sigma_2 = \sum_{M'=M}^{\infty} \sum_{m'=m-1}^{\infty} A_{M'-M} Q_{N+1-n}(M) \left[\frac{m-1}{m} a_{m'-m+1} p^* + a_{m'-m} q^* \right] \times f(m', M', 1, k).$$

Proof: $f(m, M, n, 1)$ equals the probability that the server moves to E_1, that the particular customer is selected for service, and that he leaves the system after the service period. Hence, the formula for $k = 1$.

For $k \geqq 1$, we have

$$f(m, M, n, k+1) = \sum_{m', M', n'} \Pr[(m, M, n) \to (m', M', n') \cap \text{the customer stays in } E_1] f(m', M', n', k).$$

Using eq. (4), we get that σ_1 is the part of the right-hand side corresponding to $n' \neq 1$ and σ_2 is the part of the right-hand side corresponding to $n' = 1$.

Theorem 2 provides a method for calculating the delays conditional on the state. Let $f^*(k)$ be the probability that a customer arriving at E_1 will wait k service periods before leaving the system, given that before his arrival the system was in the steady state.

Theorem 3:

$$f^*(k) = \sum_{m'=1}^{\infty} \sum_{M'=0}^{\infty} \sum_{n'=1}^{N} f(m', M', n', k) P^*(m', M', n') \qquad k = 1, 2, \cdots,$$

where

$$P^*(m', M', n') = \frac{1}{1 - a_0} \Bigg[P(m', M', n') - a_0 \sum_{n=1}^{n'-1} \sum_{M=1}^{M'+1} q_{n'-n}(M) \cdot (A_{M'-M+1}p^* + A_{M'-M}q^*) P(m', M, n) \Bigg]$$

and

$$P^*(m', M', 1) = \frac{1}{1 - a_0} \Bigg[P(m', M', 1) - a_0 \sum_{n=1}^{N} \sum_{M=0}^{M'} Q_{N+1-n}(M) \cdot [p^*P(m' + 1, M, n) + q^*P(m', M, n)] \Bigg]$$

for $m' \geqq 1$, $M' \geqq 0$, $2 \leqq n' \leqq N$.

Proof: The theorem is valid if it is shown that $P^*(m', M', n')$ is the probability of the state (m', M', n') at the point of arrival of the customer, say, point u, for all $m' \geqq 1$, $M' \geqq 0$, $1 \leqq n' \leqq N$. The probabilities of states with $m' = 0$ is zero, since there is at least one customer in E_1. We know that at $u - 1$ the system was in a steady state. The transition probabilities between $u - 1$ and u, conditional that at least one customer arrives at E_1, are obtained from (4) by replacing a_i with a_i^*, where

$$a_i^* = \begin{cases} \dfrac{1}{1 - a_0} a_i & i = 1, 2, \cdots \\ 0 & i = 0. \end{cases}$$

The expressions for $P^*(m', M', n')$ can now be calculated by operating on the steady-state probabilities with the modified transition probabilities and using the fact that the steady-state probabilities satisfy eqs. (5).

X. VALIDITY OF THE MODIFIED SYSTEM

The difference between the original and the modified systems is in the rules of the server movement. In the modified system, the server does not follow the cyclic order. However, to calculate the delay distribution, we are interested only in the pattern of the time points when the server is in E_1, and that, hopefully, is similar to the corresponding pattern in the original system. The degree of similarity is difficult to check, except by simulating the original system. Verification of the "reasonability" of the modified system may be made by checking whether each E_n receives the same number of visits by the server, and if E_1 gets the right amount of expected number of waiting customers, i.e.,

$$\sum_{m=0}^{\infty} \sum_{M=0}^{\infty} P(m, M, n) \sim \frac{1}{N} \qquad n = 1, \cdots, N$$

and

$$\sum_{m=0}^{\infty} m \sum_{M=0}^{\infty} \sum_{n=1}^{N} P(m, M, n) \sim \frac{\bar{s}}{N},$$

where $\bar{s}$ is given by eq. (2).

XI. THE MARKER-REGISTER SYSTEM

In the rest of the paper, we apply the model to the dial-tone delay problem in the No. 5 crossbar switching machine. Following are some operational features of the No. 5 crossbar switching machines which are relevant to the dial-tone delay distribution.

Calls appear on line link frames (LLF). The dial tone markers (DTM) which are not busy are paired to the waiting calls. Under "normal" operation, i.e., when several DTMs are free, the LLFs look for available DTMs according to a fixed preference order. When all the DTMs become busy, a gate is closed and the DTMs serve first those LLFs that contain waiting calls at that moment. If an LLF has more than one call waiting, only one call will be served during the gating period.

When a DTM becomes idle following the "all markers busy" condition, it looks for a waiting call according to the following scheme. Each DTM has its own order in which it scans the LLFs. Thus, for example, when there are four DTMs and 60 LLFs, the first DTM will scan the LLFs in the natural order from 0 to 59, the second DTM will start at LLF 15, go to 59, and then come back to 0 to 14, etc.*

When a DTM locates an LLF with waiting calls, it chooses one of those calls and proceeds to look for an originating register (OR) for the call. The above choice may be considered random for all practical purposes.† If the DTM finds a vacant OR, then it connects the call to the OR, and the calling customer gets a dial tone. If no OR is available, then the DTM releases the call, and it continues to wait in its LLF and to bid for a DTM. In both cases, the holding time of the DTM is constant and approximately equal. We denote this time by T, where T is approximately 0.25 second. In fact, this time is approximately 0.21 second in

* This is the recommended arrangement, although not all No. 5 crossbar entities observe it.

† We omit consideration of the systematic preference for serving calls in vertical group 2.

the case where no OR can be found, but we ignore this difference to simplify our model.

We also assume that the distribution of holding times of the ORs is negative exponential for a conservative estimate of the delay distribution. The arrival of calls to each LLF is assumed to be Poisson, with the rate being equal for all LLFs. We denote the rate for a single LLF by λ. Finally, in many cases there are two types of calls, dial-pulse and *Touch-Tone*, where both types are served by the same DTMs but require different ORs. When there are two types of calls, the ratio between their arrival rates is assumed to be the same in every LLF.

XII. A QUEUING MODEL FOR ONE TYPE OF CALL

The system described in the previous section is quite complicated, and it appears that, to model such a system and be able to derive numerical results from the model, some simplifying assumptions are inevitable. One such model was proposed by W. S. Hayward.[7] Its basic assumption is that, in order to be served, a call must find both a marker and a register idle. Once the marker and register start processing a call, they act independently of each other, each having exponentially distributed holding times. To solve the resulting state equations, Hayward introduced a system with one type of server, which approximates the behavior of his model.

The present queuing model emphasizes the order in which the markers are assigned to waiting calls, and takes into account the fact that the time a marker spends serving a call is nearly the same, whether or not it found a free register.

First we assume that each DTM serves only those LLFs which are of high priority on its list. Thus, in the example of the previous section, the first DTM will serve only the first 15 LLFs, the second DTM will serve only the next 15 LLFs, etc. Such an assumption is justified under heavy load conditions. We denote by N the number of LLFs which are served by one DTM.

Next, we assume that each DTM serves its LLFs in a cyclic order and that, whenever it finds a LLF with waiting calls, it serves exactly one call. This assumption is asymptotically valid under heavy traffic loads, because of the gating procedure described in the previous section.

Finally, we assume that whenever a DTM serves a call there is a fixed probability p^* that an OR will be available and thus that the waiting time of the call will end (i.e., the customer gets a dial tone). This assumption would hold if the availability of the ORs is independent of the number of waiting calls, which is clearly not the case. This assumption will cause our model to somewhat underestimate the delays, while the first assumption tends to overestimate them. The value of p^* can be taken approximately as the delay probability of a call, given that the arrival process of the calls to the ORs is Poisson, and therefore it can be computed by the Erlang C formula.

Thus, we arrive at the model which was described in Section II, with the server the DTM. The server chooses a customer (call) from the queue at random. After one service time, the customer either leaves the system with probability p^* or rejoins his queue (that is, waits in his LLF). The server then moves to the next LLF having a waiting call.

XIII. SOME THEORETICAL AND NUMERICAL RESULTS

It was shown in Section IV that the occupancy of the DTM is

$$\rho = \frac{\lambda NT}{p^*}.$$

Hence, a necessary and sufficient condition for nonsaturation of the system is $\lambda NT < p^*$. Also, the expected total number of waiting calls in the LLFs which are served by the DTM is

$$\bar{s} = \frac{\rho}{2(1-\rho)}(2-\rho p^*).$$

Thus, for a fixed occupancy of the DTM, the expected total number of waiting calls is a monotone decreasing function of p^*. The same is true for the expected delay, $\bar{W}$, since by Little's formula

$$\bar{W} = \frac{\bar{s}}{\lambda TN} = \frac{\bar{s}}{\rho p^*},$$

and so

$$\bar{W} = \frac{1}{2(1-\rho)}\left(\frac{2}{p^*}-\rho\right).$$

A standard measure for the quality of service is the dial-tone speed (DTS), which is the probability that the call will have to wait three seconds or more for a dial tone. Figure 1 presents some computed values of the DTS for various values of p^* and λ with ρ held constant at three different values. It is seen that DTS is also a monotone decreasing function of p^*, for a fixed DTM occupancy.

XIV. A MODEL FOR TWO TYPES OF CALLS

Consider now the case of a system having dial-pulse and *Touch-Tone* calls; this is the usual situation in No. 5 crossbar offices today. Let the arrival rates from each LLF be λ_1 and λ_2, and let the probabilities of finding available registers be p_1^* and p_2^* for dial-pulse and *Touch-Tone*

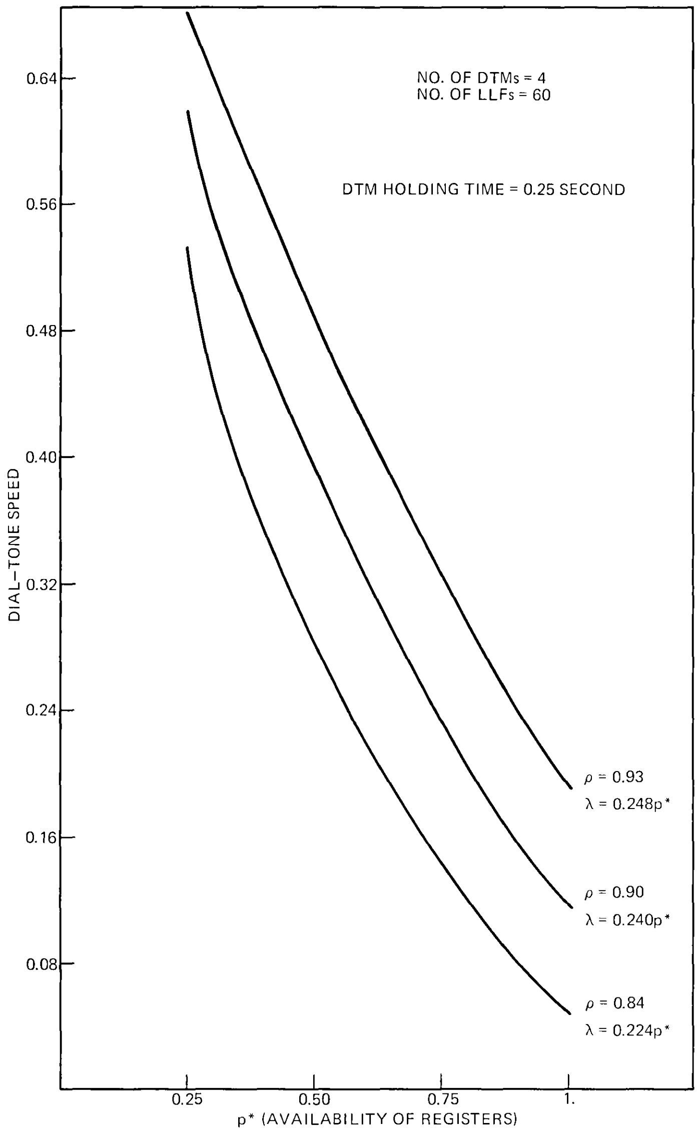


Fig. 1—DTS as a function of p^* and λ for constant DTM occupancy ρ (queuing model).

calls, respectively. The loads on the DTM due to the two types of calls are

$$\rho_1 = \frac{\lambda_1 NT}{p_1^*} \quad \text{and} \quad \rho_2 = \frac{\lambda_2 NT}{p_2^*}.$$

To approximate the delays of this system, we consider a system with one type of call, for which $\lambda = \lambda_1 + \lambda_2$ and $\rho = \rho_1 + \rho_2$. The appropriate p^* of this system satisfies

$$p^* = \frac{\lambda_1 + \lambda_2}{\lambda_1/p_1^* + \lambda_2/p_2^*}.$$

Now the state probabilities can be computed by the approximate method; however, it is necessary to modify the formulas for the computation of the delay distributions to obtain those distributions conditional on the type of call.

Let $f_i(m, M, n, k)$ be the probability that a call of type i $(i = 1, 2)$ arriving at LLF No. 1 will wait k service periods before leaving the system, given that on its arrival the system went into the state (m, M, n). The recursive formulas of Theorem 2 have to be modified to

$$f_i(m, M, n, 1) = \frac{p_i^*}{m} Q_{N-n+1}(M)$$

and

$$f_i(m, M, n, k+1) = \sigma_1 + \sigma_2$$
$$m \geqq 1, \qquad M \geqq 0, \qquad 1 \leqq n \leqq N, \qquad k \geqq 1,$$

where σ_1 is as in Section IX except that f is replaced by f_i and

$$\sigma_2 = \sum_{m'=M}^{\infty} \sum_{m'=m-1}^{\infty} A_{M'-M} Q_{N+1-n}(M) \left[\frac{m-1}{m} a_{m'-m+1} p^* + a_{m'-m} \left(\frac{m-1}{m} q^* + \frac{q_i^*}{m} \right) \right] [f_i(m', M', 1, k)].$$

The proof of the validity of the modified formulas is along the same lines as the proof of Theorem 2.

XV. NUMERICAL RESULTS

Several computer runs were made for a typical large system with 60 LLFs, 4 DTMs ($N = 15$), 100 dial-pulse ORs and 50 *Touch-Tone* ORs, with both dial-pulse ORs and *Touch-Tone* ORs having a mean holding time of 13 seconds. T was taken to be 0.25 second in all runs. The parameters varied were $\lambda = \lambda_1 + \lambda_2$, the total input rate per LLF, and $\alpha = \lambda_1/\lambda_2$ (the ratio of the rates of the two types of calls).

Figures 2 and 3 describe the results for $\alpha = 2$, that is, when the ratio of the loads is the same as that of the ORs. Figure 2 describes the behavior of the occupancies of the ORs and the DTM, while in Fig. 3 the DTS is plotted as a function of λ. Figures 4 and 5 present the corresponding results for $\alpha = 3$. The values of λ were chosen to be near the point of saturation, i.e., where the occupancy of the DTM ap-

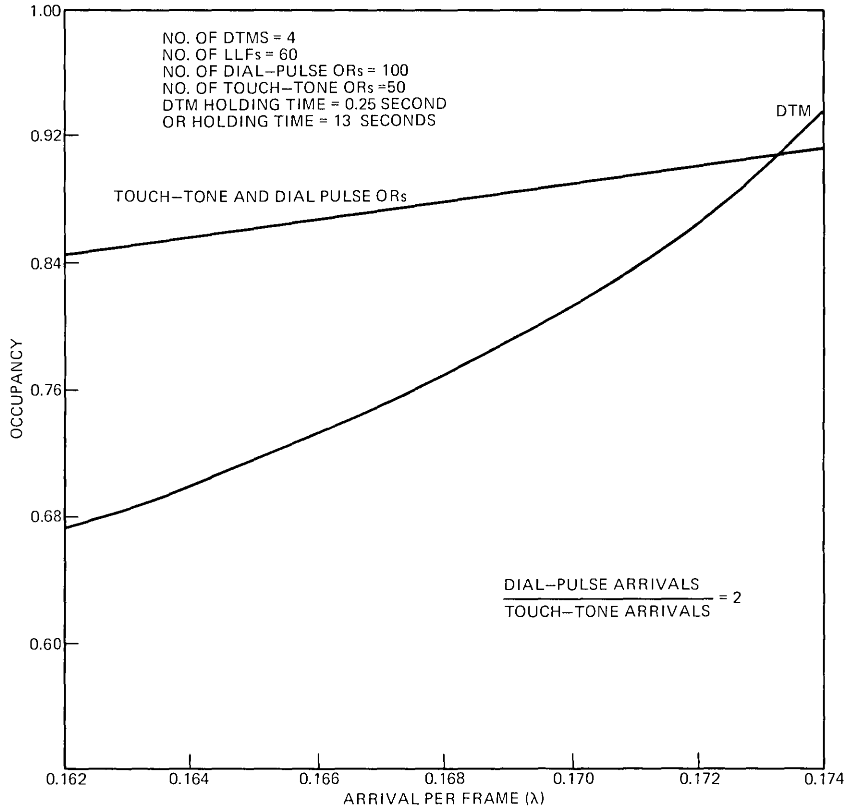


Fig. 2—Occupancies of the DTM and ORs as a function of λ (queuing model).

proaches 1. In this range, the DTS is sensitive to small perturbations in λ.

Figure 6 shows the dependence of the DTS on α for a fixed λ ($\lambda = 0.158$ per second). It can be seen that the quality of service deteriorates as α diverges from the neighborhood of the ratio of the number of dial-pulse ORs/number of *Touch-Tone* ORs (which equals 2 in our case). This is consistent with the observation that *Touch-Tone* delays are significantly influenced by the dial-pulse OR load for a constant offered load to the DTM.

The results were compared with those of a simulation model which we constructed for the system described in Section XI. Tables I and II present a comparison between the results of the queuing model and the simulation. We compare the intensities of input for which levels of the DTS are reached between 0.05 and 0.25 for a balanced system and between 0.05 and 0.15 for an unbalanced system. Examining those

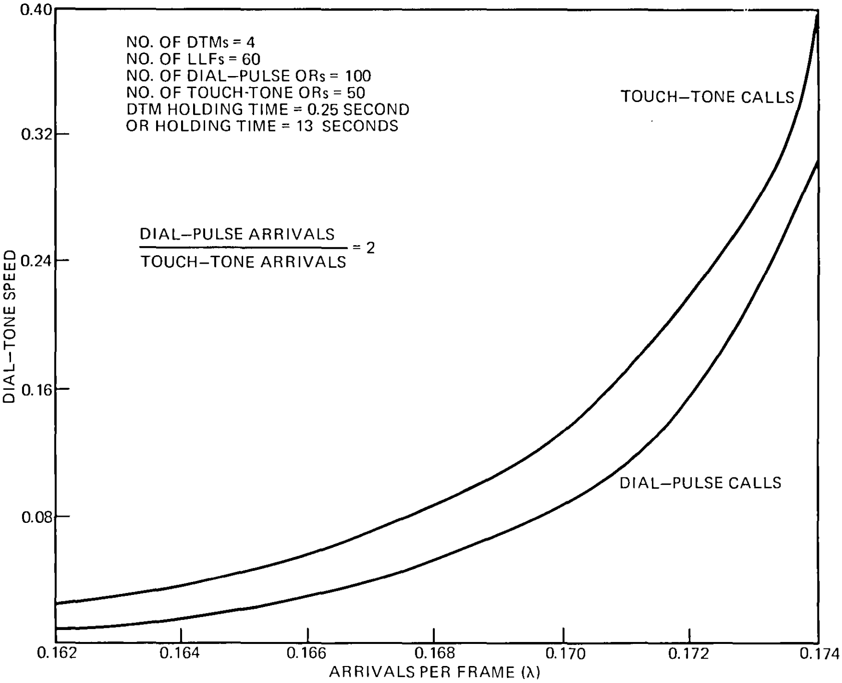


Fig. 3—DTS as functions of λ (queuing model).

Table I — Values of λ for which given levels of DTS are reached, for $\lambda_1/\lambda_2 = 2$

(a) Dial-Pulse Calls

DTS	Queuing Model	Simulation
0.05	0.1680	0.1695
0.10	0.1705	0.1735
0.15	0.1720	0.1760
0.20	0.1730	0.1775
0.25	0.1735	0.1790

(b) *Touch-Tone* Calls

DTS	Queuing Model	Simulation
0.10	0.1690	0.1690
0.15	0.1705	0.1720
0.20	0.1720	0.1740
0.25	0.1725	0.1760

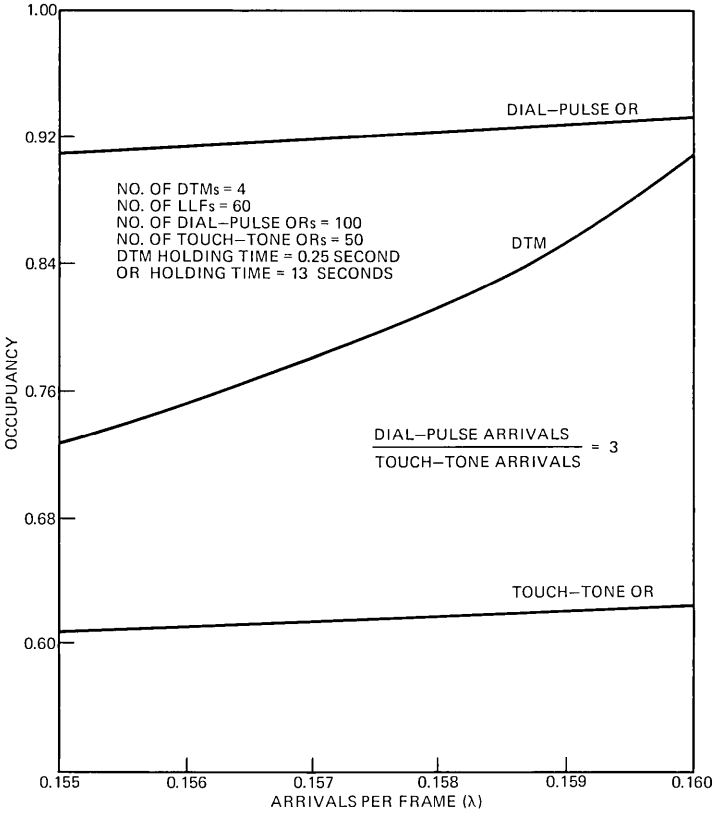


Fig. 4—Occupancies of the DTM and ORs as a function of λ (queuing model).

Table II — Values of λ for which given levels of DTS are reached, for $\lambda_1/\lambda_2 = 3$

(a) Dial-Pulse Calls

DTS	Queuing Model	Simulation
0.05	0.1540	0.1550
0.10	0.1565	0.1590
0.15	0.1580	0.1610
0.20	0.1585	0.1620

(b) *Touch-Tone* Calls

DTS	Queuing Model	Simulation
0.05	0.1575	0.1600
0.10	0.1590	0.1625
0.15	0.1600	0.1640

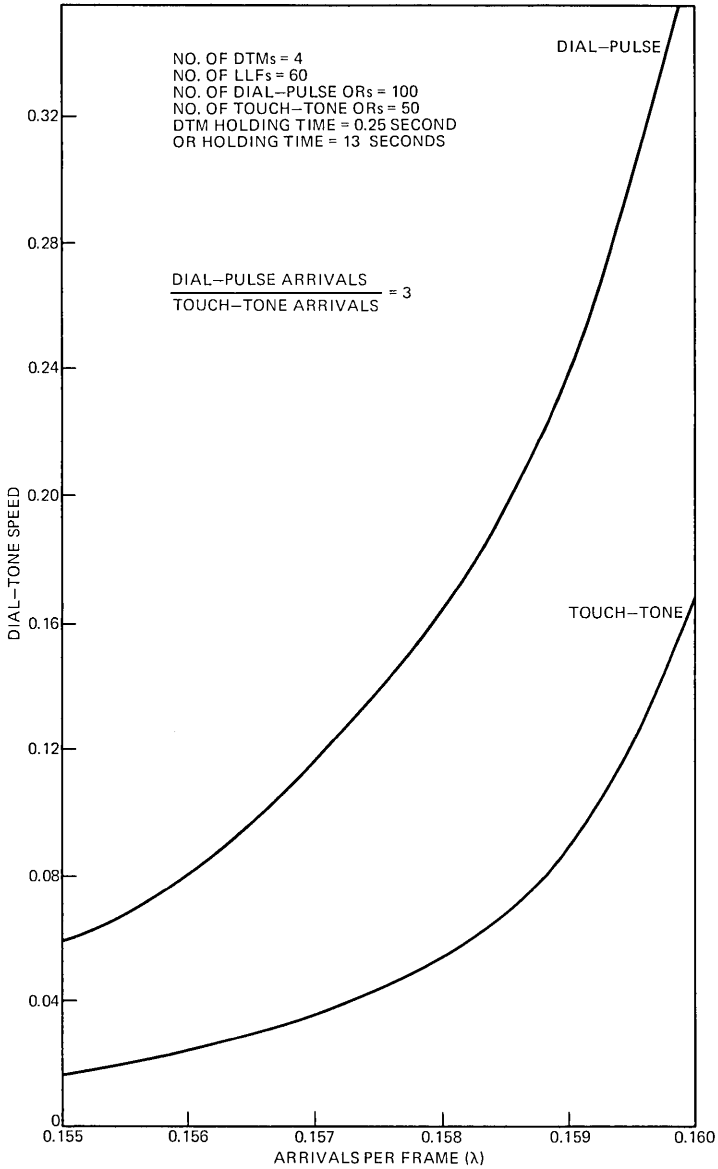


Fig. 5—DTS as functions of λ (queuing model).

tables, we observe that the differences in the corresponding figures for the two models in these regions are less than 4 percent of the total input. Also, it can be observed that the computed DTS grows faster in the queuing model than in the simulation. The reason is that the

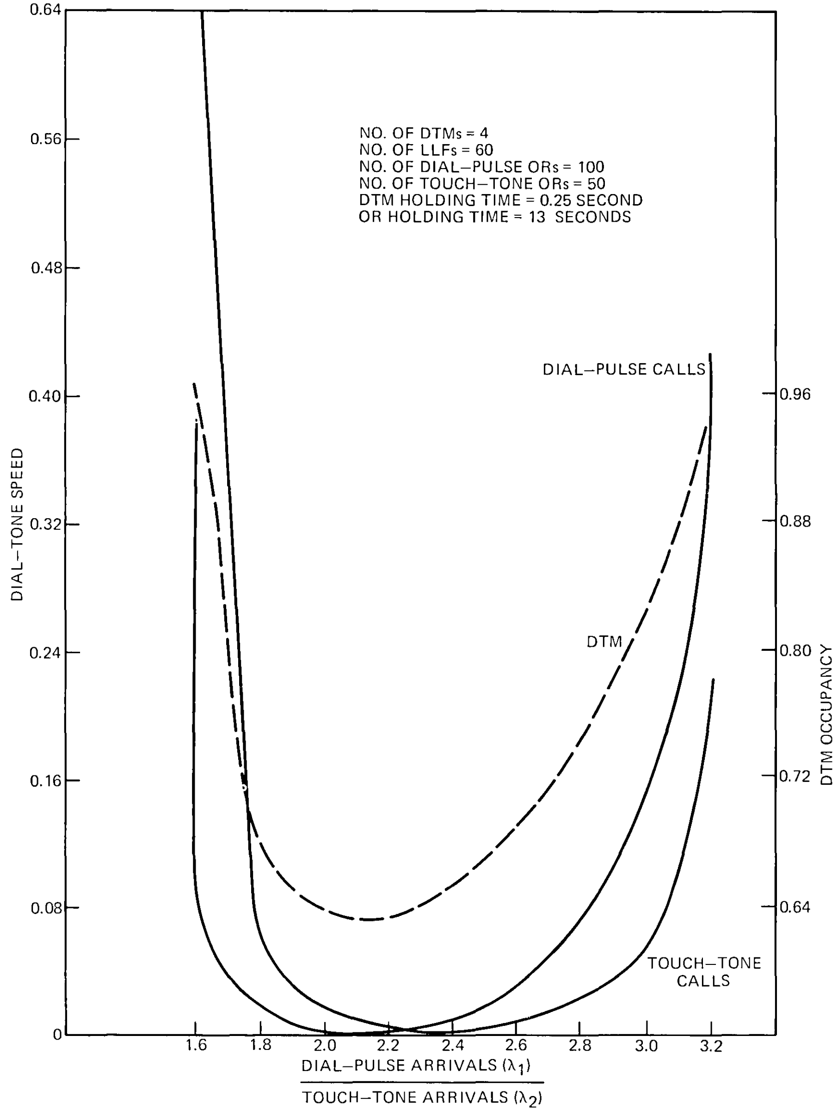


Fig. 6—DTS and DTM occupancy as functions of λ_1/λ_2 for a fixed λ, $\lambda = 0.158$.

assumption made in Section XII, that the probability of finding all registers busy can be computed by the Erlang C formula, is incorrect in this region. Figure 7 compares the DTS as computed by the two models. Again we conclude that the fit is fair in the "critical" region. It would have been desirable to validate the results by performing a field trial. Such a trial should consist of measuring the DTS for a No. 5

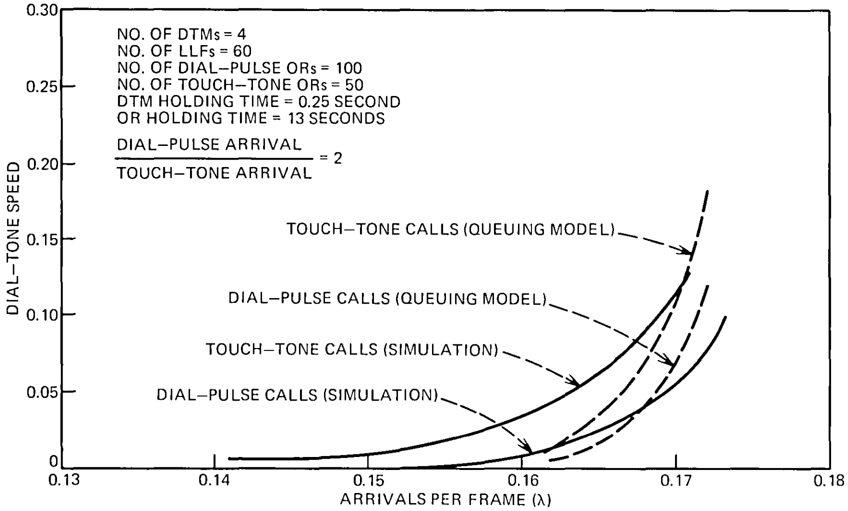


Fig. 7—DTS as function of λ (queuing model and simulation).

crossbar machine with high constant loads. However, observing Fig. 8, where the cumulative distributions of the hourly DTS from the simulation were plotted, one sees that these distributions have long tails. This implies that, to get a reasonably accurate estimate for the DTS, say, with a standard error of 1 percent, we would have to run the trials for around 25 hours while keeping the load constant. This seems to be a difficult task.

XVI. SUMMARY AND CONCLUSIONS

We presented a method for modifying the original model, as presented in Section II, to a model which has a much smaller state space. Methods were described for calculating the steady-state distributions of the states and of the delays in the modified model. The model was applied to the problem of calculating dial-tone delays in the No. 5 crossbar switching machine. This was accomplished by making some simplifying assumptions about the order of service of the waiting calls by the markers. The numerical results were compared to those of a simulation, and found to be close on an important range of the DTS. This gives us a certain amount of confidence that both models are valid, which is especially important because of the difficulty in validating the models by experimental data, as discussed in Section XV. However, the reader should be aware that several features of the No. 5 crossbar machine, which may have an influence on the DTS, were excluded.

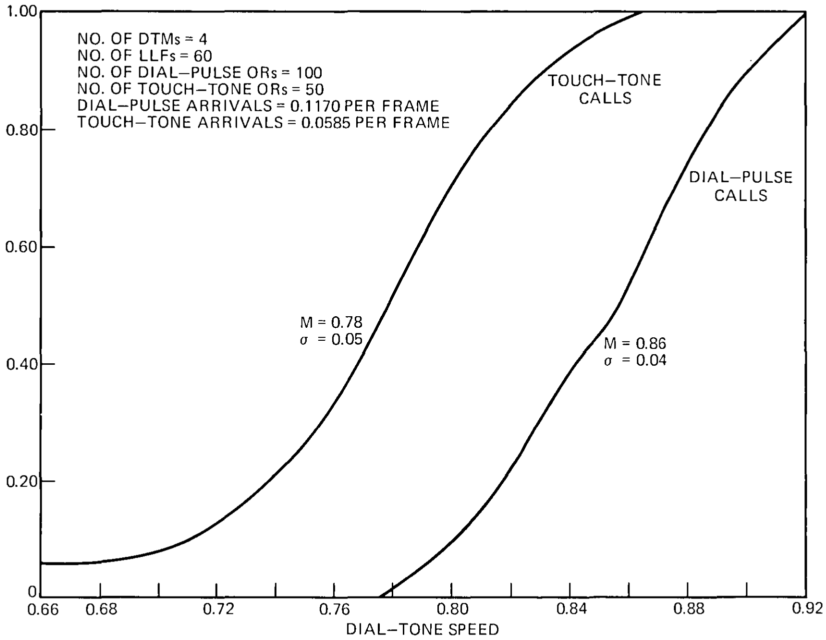


Fig. 8—Cumulative distributions of the hourly dial-tone speed (simulation).

A major effect not covered by the present model is the effect of horizontal group blocking on dial-tone speed and on dial-tone marker waste usage. Recently obtained field data and theoretical studies reported in a subsequent paper by H. A. Guess[6] have shown that dial-tone speed and dial-tone marker occupancy can be appreciably increased by horizontal group blocking caused by high average line link frame loads and also by poor load balance. Consequently, the dial-tone speeds associated with a given call origination rate in an actual No. 5 crossbar office may be higher than would be predicted by our model.

XVII. ACKNOWLEDGMENTS

I wish to thank J. G. Kappel and M. F. Morse for introducing me to the dial-tone delay problem.

REFERENCES

1. R. B. Cooper, "Queues Served in Cyclic Order: Waiting Times," B.S.T.J., *49*, No. 3 (March 1970), pp. 399–414.
2. R. B. Cooper and G. Murray, "Queues Served in Cyclic Order," B.S.T.J., *48*, No. 3 (March 1969), pp. 675–689.
3. M. Eisenberg, "Two Queues with Alternating Service," unpublished work.

4. M. A. Leibowitz, "An Approximate Method for Treating a Class of Multiqueue Problems," IBM Journal, *5*, July 1961, pp. 204–209.
5. G. Schay, Jr., "Approximate Methods for a Multiqueuing Problem," IBM Journal, *6*, April 1962, pp. 246–249.
6. H. A. Guess, "The Effect of Frame Load and Balance on Dial-Tone Delay ir No. 5 Crossbar," B.S.T.J., this issue, pp. 1755–1793.
7. W. S. Hayward, "Capacities of Interacting Server Groups in Common Control Switching Systems," Proceedings of the Sixth International Teletraffic Congress, Munich, September 9–15, 1970, pp. 325/1–6.

Copyright © 1975 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 10, December 1975
Printed in U.S.A.

The Effect of Frame Load and Balance on Dial-Tone Delay in No. 5 Crossbar

By H. A. GUESS

(Manuscript received May 22, 1975)

Recently obtained field data and theoretical studies show that, for a fixed subscriber calling rate, dial-tone delay in No. 5 crossbar can be appreciably increased by high average line-link frame loads and also by poor load balance. The increased delay is caused by waste dial-tone-marker usage generated by a small number of calls that encounter horizontal group blocking in obtaining a dialing connection. This paper discusses an analytical model to predict the time from receiver off-hook to receipt of dial tone under various service conditions in No. 5 crossbar.

I. INTRODUCTION

1.1 Description of the dial-tone connection process

In a No. 5 crossbar switching machine, dial tone is provided to a subscriber line, terminating on a line-link frame (LLF), by an originating register (OR), terminating on a trunk-link frame (TLF), via a series of three network links: line links, junctors, and trunk links. The dialing connections are set up by dial-tone markers (DTMs), which are common control devices. Each line-link frame contains a number of crossbar switches that are used to establish connections between subscriber lines and trunks, or between subscriber lines and service circuits, such as originating registers. The crossbar switches that form line concentrators on which groups of subscriber lines terminate are called horizontal groups. Maximum size offices typically contain from 4 to 6 DTMs, 40 to 60 LLFs, 20 to 30 TLFs, and up to 140 ORs. Each LLF contains 10 horizontal groups and each horizontal group is a concentrator containing between 29 and 59 subscriber lines on the input side of the switch and 10 line links on the output side of the switch.

To provide dial tone to a subscriber, the off-hook signal from the subscriber line initiates a bid for a DTM through a connector circuit unique to each LLF. As soon as a DTM becomes available, it locates an unoccupied OR and then attempts to find a dialing path (consisting of a line link, a junctor, and a trunk link) connecting the OR with the subscriber line. As soon as the connection is established, the DTM releases and proceeds to serve other calls waiting for dial tone. The OR provides dial tone, receives the dialed digits, obtains a completing marker, transmits information to it, and releases. The completing marker then establishes the connection between the calling subscriber and an outgoing or intraoffice trunk.

If all ORs are busy, the DTM releases and the call rejoins what is effectively a queue of calls waiting for a DTM. If an OR is available, but no (unoccupied) dialing path connecting the OR with the subscriber line can be found, the DTM will release the OR, obtain a second (usually different) OR, and try to find a dialing path between that OR and the subscriber line. When a DTM is unable to find a dialing path between a given OR and a given subscriber line, a matching failure is said to occur. If, on the second try, the DTM cannot find a dialing path, the DTM releases and the call rejoins the queue of calls waiting for a DTM. In such a case, a DTM second-failure-to-match (DTM2FTM) is said to occur.

The method of assigning DTMs to waiting calls is controlled by a type of "gating" circuitry designed to equalize service and to reduce the incidence of long delays in obtaining dial tone. When one or more DTMs are free (light traffic operation), the LLFs look for DTMs according to a fixed preference order. When all DTMs become busy and a request for a DTM occurs, a gate is closed and only the LLFs containing calls waiting for dial tone at that moment are put inside the gate. The DTMs then proceed to serve the LLFs inside the gate. If more than one call is waiting on an LLF, only one call will be served during the gating period. Once a call on an LLF is served during a given gating period, that LLF is put out of the gate, whether or not the DTM is successful in establishing a dialing connection for the call. When all LLFs with requests at the start of the gating cycle have been put out of the gate, the gate opens; if there are sufficient waiting calls to cause all of the DTMs to become busy again, a new gating cycle will be started; otherwise, light traffic operation will resume. During a gating period, a DTM that becomes idle scans the LLFs in cyclic order. Each DTM has a different starting LLF for the scan so as to equalize service. (The description of the gating procedure is taken from Refs. 1 and 2.)

1.2 Effects of matching failures on dial-tone delay

Repetitive matching failures can occur in establishing the dialing connection essentially because the first-stage crossbar switch on which the subscriber line terminates (the horizontal group) is a concentrator whose output links (line links) have holding times that are much longer than the holding times of the common control devices that set up the dialing connections (the DTMs). The average line-link holding times, being largely determined by conversation holding times (with allowance for ineffective attempts), can be on the order of 150 seconds or more, while the DTM holding times are typically on the order of 0.25–0.40 second.

Although matching failures (blocking) can occur when one or more of the 10 line links on a horizontal group are unoccupied, *repetitive* matching failures under such conditions are rare because a DTM is quite likely to be successful in establishing a dialing connection after a few attempts. The expected holding time in a blocked condition of a call that finds all 10 line links busy is the lesser of the length of time for one of the 10 line links to become free (with a short added time for DTM uses and matching failures in setting up the dialing path) and the length of time that a subscriber is willing to wait.

Since, at average busy-hour load levels, less than about one percent of all originating calls are predicted to encounter an all-10-line-links-busy condition, previous dial-tone-delay studies have assumed the effect of matching failures on dial-tone delay to be small. However, the fact that line-link holding times are much longer than DTM holding times means that a call which finds all 10 line links busy can remain blocked for long enough to consume a large number of DTM uses (except at calling rates sufficiently high that very little of the offered blocked-call load is carried by the DTMs). Thus, it is possible for a small number of calls experiencing blocking to generate a disproportionate number of waste DTM uses, increase DTM occupancy, and thereby increase dial-tone delay for all other calls in the office.

II. ANALYTICAL MODEL (LIMITING FORM)

2.1 Assumptions

Since the main effect of matching failures on dial-tone delay is caused by the resulting waste DTM use, and since the dial-tone delay distribution of the small proportion of calls experiencing repetitive matching failures can be calculated approximately (see Section 5.2), the effect of matching failures on the dial-tone delay distribution of calls that do not experience repetitive matching failures has been represented in terms of a queuing model with two classes of calls, good calls and bad calls, defined as follows:

(*i*) A *good call* experiences no matching failures but is subject to delay caused by DTMs and ORs.

(*ii*) A *bad call* experiences total network blocking (no dialing connection available) and defects from the system after an exponential waiting time.

We will first describe the mathematical structure of the analytical model used for dial-tone delay calculations. We will refer to this model as *the limiting model.* Next, we derive the DTM saturation load* and prove that it is not changed by the presence of bad calls.

In Appendix B, we prove that the equilibrium queue length and waiting time distributions of the limiting model are the limits in distribution of a sequence of equilibrium distributions arising from a model in which the expected bad-call arrival rate approaches zero and the expected bad-call waiting time (until defection) approaches infinity in such a way that their product, total erlangs of bad calls, is constant.

In the limiting model, a good call arrives and finds a random (truncated) Poisson-distributed (but time-independent) number of bad calls permanently present in the system.† The queue discipline is characterized by random order of service. The DTMs cannot distinguish between good and bad calls when choosing a call to be served. This corresponds to the fact that, in the No. 5 crossbar switching machine, a DTM cannot recognize that a 10-line-links-busy condition exists on a particular horizontal group and, hence, cannot avoid wasting time serving calls for which no dialing path exists. Equilibrium good-call queue length distributions are computed conditional upon the number of bad calls present in the system. Calculating the expectation of the conditional distribution over the distribution of bad calls gives the unconditional dial-tone-delay distribution for good calls. The conditional distributions depend on the total office calling rate and on the number of LLFs, DTMs, and ORs but do not depend on the horizontal group load or the load variation. Hence, delay distributions for a range of frame load and balance effects can be calculated using the same set of conditional distributions, thereby greatly reducing the computer time needed for parametric studies.

Since DTM holding times are approximately constant (for a given set of office parameters and traffic characteristics) and since these holding times are small with respect to the accuracy with which it is necessary to be able to predict delays, we treat the dial-tone delay process as a discrete time queue with a constant service time of T seconds. Good calls are assumed to arrive in batches according to a Poisson process at times kT, for $k = 1, 2, \cdots$. Immediately upon arrival, the good calls join the queue of good and bad calls waiting for dial tone. Calls are chosen at random from the queue for service

* The DTM saturation load is defined to be the good-call originating load below which a steady state good-call-queue-length distribution exists and above which such a steady state distribution does not exist.

† When the bad-call input is assumed to be peaked, the number of bad calls in the system has a (truncated) negative binomial distribution. This is discussed in Section IV.

by the DTMs with each call—whether a good call or a bad call—having an equal probability of being chosen. A good call served on the DTM cycle beginning at time kT will either acquire an OR, receive dial tone, and thereupon exit the system at time $(k + 1)T$, or else will fail to acquire an OR and will return to the queue of calls waiting for dial tone at time $(k + 1)T$. A bad call served on the DTM cycle beginning at time kT will return to the queue of calls waiting for dial tone at time $(k + 1)T$. Note that since new arrivals occur only at the time points kT, a DTM that is idle on the cycle beginning at time kT will remain idle at least until the start of the cycle beginning at time $(k + 1)T$. The dial-tone delay for a call is the length of time from the moment when the call arrives to when it obtains dial tone and (simultaneously) leaves the dial-tone queue. Thus, the minimum possible dial-tone delay in the model is T seconds.

In the analytical model, two DTMs serving N line-link frames are used to represent four DTMs serving $2N$ line-link frames. A single DTM holding time equal to the office average DTM holding time is used for both good and bad calls. In actuality and in the simulation models, bad calls have somewhat longer DTM holding times than good calls and good calls that encounter a condition of all-ORs-busy have somewhat shorter DTM holding times than good calls that do not encounter a condition of all-ORs-busy. Comparison of results from the analytical model with those of the gating simulation model indicates that these simplifications tend to offset each other in the range of interest.

A further simplification in the analytical model concerns the manner in which availability of ORs is treated. Since No. 5 crossbar offices typically contain over 100 ORs and since the ORs are in tandem with the DTMs, it is presently not practical to keep track of the number of occupied ORs directly in an analytical model. Availability of ORs is treated by assuming that at each time point kT, all ORs are busy with probability q and two or more ORs are free with probability $p = 1 - q$. The calculated probability that exactly one OR is free and that a dialing connection is available between this one OR and the given subscriber line is sufficiently small, in the occupancy range of interest, so as not to warrant the additional complexity caused by introducing this effect.

The idea of using a discrete time model and the method of treating OR availability through use of a fixed probability of all ORs busy are due to Halfin.[3] Following Halfin, the probability q is taken to be the erlang C probability of all ORs busy. At OR occupancies above about 0.90 with frame loads low enough that few second failures to match occur, this method of treating OR availability somewhat underpredicts the delay caused by an all-registers-busy condition (based on comparison with simulation results). The underprediction arises because, at higher OR occupancies, once all ORs become busy they tend to remain busy for a time period equal to several DTM cycles, as one would expect from the erlang *C* delay formula.

Preliminary studies were made using an analytical single-server cyclic queuing model, developed by S. Halfin,[3] which represented the No. 5 crossbar gating process in considerable detail but did not take into account the effects of horizontal group blocking. These studies showed that the delays predicted by the cyclic queuing model do not differ appreciably from those of a discrete-time M/D/1 queue with feedback and random order of service. The latter model requires about 1/30th of the computer time required by the former. Both models overpredicted simulations. For these reasons, explicit representation of the gating process was not attempted and a discrete-time M/D/2 queue with feedback and random order of service was taken as the starting point for developing an analytical dial-tone-delay model to include effects of horizontal group blocking.

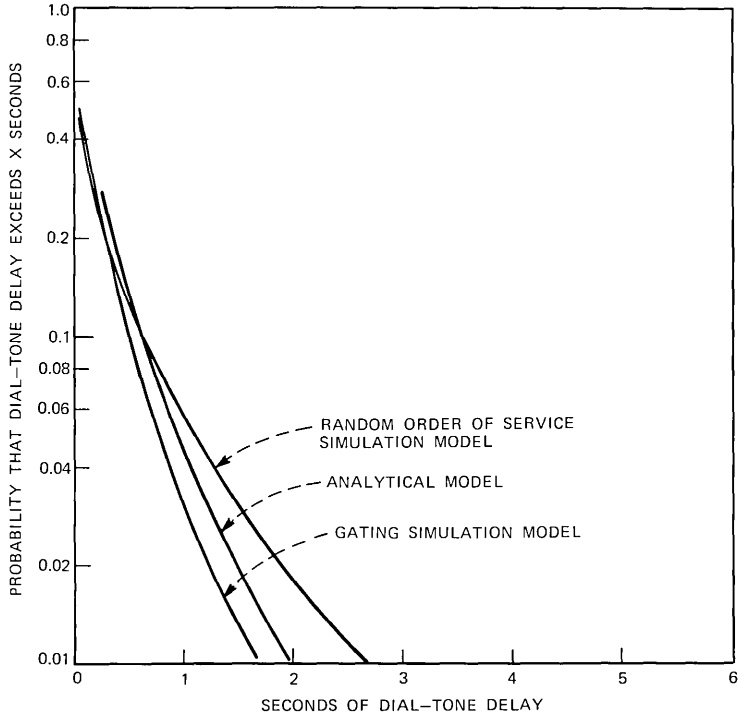


Fig. 1—Comparison of analytical and simulation models. Dial-tone delays are at 1400 CCS/LLF and 0.70 DTM OCC (excluding DTM2FTMS).

Comparison of simulation results with predictions using the limiting analytical model (based on two markers and random order of service) shows that the predicted delays typically fall somewhat above or very close to simulated delays based on four markers with gating order of service and somewhat below simulated delays based on four markers with random order of service. In the light of these results and because of the large scatter in actual measured No. 5 crossbar dial-tone delays, it did not seem worth the considerable added complexity to include explicit representation of the gating process in the analytical model. Typical results are shown in Figs. 1 and 2.

2.2 Queue length equations for the limiting model

The queue-length process for good calls in the limiting model is a discrete-time Markov chain with finitely many irreducible classes $\{C_k\}$ that are noncommunicating in the sense that no transition between different classes is possible. Hence, each class is itself an irreducible Markov chain. The kth class is the queue-length Markov chain for a discrete-time M/D/2 queue with random order of service,

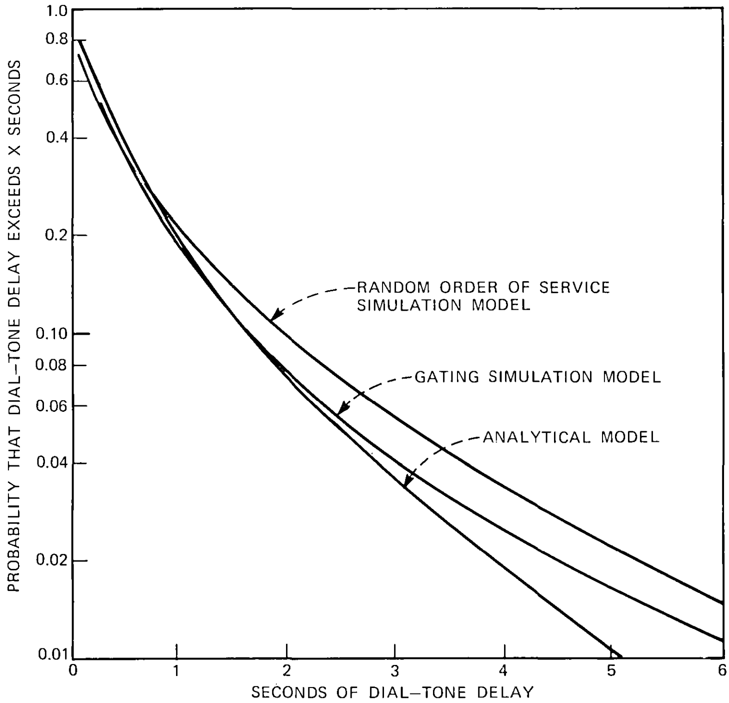


Fig. 2—Dial-tone delays are at 1600 CCS/LLF and 0.80 DTM OCC (excluding DTM2FTMS).

"feedback" (occurring when an all-ORs-busy condition is encountered), and k blocked calls residing permanently in the system. The DTMs cannot distinguish these k permanently blocked bad calls from the good calls in the system, so each of the k bad calls and each of the good calls in the system compete on an equal basis for DTMs. Upon completion of service by a DTM, a bad call rejoins the queue. With probability q, the good calls served at any given time period rejoin the queue and with probability p they leave it. We will show that each of these queuing systems has the same saturation load and we will describe how the equilibrium-queue-length distribution and the equilibrium-waiting-time distribution for each system is calculated.

Viewing the whole process again from the standpoint of a Markov chain with finitely many noncommunicating classes, the queue-length and waiting-time distributions referred to above may be regarded as being conditional on k, the number of blocked calls (permanently) present in the system. Now let k be a random variable with a truncated Poisson distribution of mean x, or, equivalently, regard the Markov chain as having any initial distribution $\pi(k, i)$, where i denotes the number of good calls in the system and where the marginal distribution $\pi(k, \cdot)$ of the number of bad calls in the system is truncated Poisson with mean x. Then the (unconditional) equilibrium-queue-length and equilibrium-waiting-time distributions for this system may be computed by taking the expectation (with respect to the truncated Poisson distribution of k) of the conditional queue-length and waiting-time distributions for each of the individual systems represented by the classes C_k.

Let X_n denote the number of good calls in the queue at time nT and let Y denote the number of bad calls (permanently) present in the queuing system. Let

$$P_k(i, j) = \Pr [X_{n+1} = j | X_n = i \text{ and } Y = k]$$

and let $P_k(i)$ be the equilibrium-queue-length distribution for the case of k bad calls. Let $A(n)$ be the probability that n good calls arrive in one service time interval (of length T). Then, by assumption,

$$\begin{aligned} A(n) &= \frac{e^{-\lambda}\lambda^n}{n!} \qquad && n = 0, 1, 2, \cdots \\ &= 0 && \text{otherwise,} \end{aligned}$$

where

$\lambda = \lambda_1 NT$

$\lambda_1 =$ LLF originating call rate (calls per second on one LLF),

$N =$ number of LLFs served by the two DTMs.

The transition functions $P_k(\cdot, \cdot)$ are given by

$$\begin{aligned} P_0(0, j) &= A(j) \\ P_0(1, j) &= pA(j) + qA(j-1) \\ P_0(i, j) &= pA(j-i+2) + qA(j-i) \qquad \text{for } i \geqq 2, \end{aligned} \tag{1}$$

and for $k \geqq 1$,

$$\begin{aligned} &P_k(0, j) = A(j) \\ &P_k(i, j) \\ &\quad = p[k(k-1)A(j-i)+2ikA(j-i+1)+i(i-1)A(j-i+2)]/ \\ &\qquad\qquad [(k+i)(k+i-1)] + qA(j-i) \qquad \text{for} \qquad i \geqq 1. \end{aligned} \tag{2}$$

Note that $P_k(i, j) = 0$ for $i > j + 2$ and that $P_k(j + 2, j) > 0$.

If the equilibrium distributions $P_k(\cdot)$ exist, then they satisfy the equation

$$P_k(j) = \sum_{i=0}^{j+2} P_k(i)P_k(i, j). \tag{3}$$

Let f_k be the generating function for the kth queue length process; then

$$f_k(z) = \sum_{i=0}^{\infty} P_k(i)z^i.$$

The generating function f_0 is easily seen to be given by

$$f_0(z) = \frac{p(z-1)[(z+1)P_0(0) + zP_0(1)]}{z^2e^{\lambda(1-z)} - (p + qz^2)}. \tag{4}$$

If $P_0(\cdot)$ exists, then $\lim_{z\uparrow 1} f_0(z) = 1$, and hence

$$P_0(0) + \tfrac{1}{2}P_0(1) = 1 - \frac{\lambda}{2p}. \tag{5}$$

Let $S_k^{(2)}(p)$ denote the saturation load of the kth process with two DTMs, as discussed above, and let $S_k^{(1)}(p)$ denote the saturation load of the analogous one-DTM system with k blocked calls. It can be shown that $S_k^{(1)}(p) = p$ for all $k = 0, 1, 2, \cdots$, i.e., a necessary and sufficient condition that a steady-state queue length distribution exists in the one-DTM system is that $\lambda < p$.* Using this result, we will show that $S_k^{(2)}(p) = 2p$ for all $k = 0, 1, \cdots$.

Since the Markov chain for the kth process above is irreducible, $P_0(0) > 0$ and $P_0(1) > 0$ whenever the stationary distribution $P_0(\cdot)$ exists. Hence, it follows from eq. (5) that $S_0^{(2)}(p) \leqq 2p$. Since $S_{k_2}^{(2)}(p)$

* This result can be proved using a theorem of Kushner (Ref. 4) and some properties of generating functions.

$\leqq S_{k_1}^{(2)}(p) \leqq S_0^{(2)}(p) \leqq 2p$ for $k_2 \geqq k_1 \geqq 0$, it suffices to show that $S_k^{(2)}(p) \geqq 2p$ for all $k \geqq 0$.

Consider a two-DTM system with $2k$ blocked calls, which differs from the $2k$th process defined above only in that (*i*) each DTM maintains a separate queue with half the originating traffic being assigned to one of the DTMs and the other half being assigned to the other DTM and (*ii*) each DTM serves k blocked calls. Let $\tilde{S}_{2k}(p)$ denote the saturation load of this modified system. Then clearly, $\tilde{S}_{2k}(p) \leqq S_{2k}^{(2)}(p)$ and, by symmetry, $\tilde{S}_{2k}(p) = 2S_k^{(1)}(p) = 2p$, because the modified system simply consists of two identical one-marker systems working separately, with each being assigned half the incoming traffic. Hence, $S_{2k}^{(2)}(p) \geqq 2p$ and so $S_k^{(2)}(p) = 2p$ for all $k \geqq 0$.

Thus, for $\lambda/2p < 1$, a stationary distribution exists for the kth process for each $k \geqq 0$. In the sequel, we will confine ourselves to the case where $\lambda/2p < 1$. In this case, f_0 exists and is analytic for $|z| < 1$ and continuous for $|z| = 1$, so the numerator of the expression on the right-hand side of eq. (4) must vanish for any $|z| \leqq 1$ for which the denominator vanishes. It is easy to see that the denominator has a single real root on the open interval $(-1, 0)$ and that $z = 1$ is a root. By applying Rouche's theorem to the functions $z^2 \exp[\lambda(1 - z)]$ and $p + qz^2$ along the circle $|z| = 1 + \epsilon$, one can show that for $\lambda/2p < 1$ and for $\epsilon > 0$ sufficiently small (depending on λ and p), the expression $z^2 \exp[\lambda(1 - z)] - (p + qz^2) = 0$ has exactly two roots in the open disc $|z| < 1 + \epsilon$. Thus, the root on $(-1, 0)$ and the root at 1 are the only roots within the closed unit disc and

$$(\xi + 1)P_0(0) + \xi P_0(1) = 0, \tag{6}$$

where ξ is the unique root of the expression

$$z^2 e^{\lambda(1-z)} - (p + qz^2) = 0, \qquad -1 < z < 0. \tag{7}$$

Solving eqs. (5) and (6) for $P_0(0)$ and $P_0(1)$ yields the unique solution

$$P_0(0) = \frac{-\xi(2p - \lambda)}{(1 - \xi)p}$$

$$P_0(1) = \frac{(1 + \xi)(2p - \lambda)}{(1 - \xi)p}, \tag{8}$$

where ξ is defined by (7). The values $P_0(j)$ for $j \geqq 2$ can now be computed recursively from (3) using (8).

When $k > 0$, computation of $P_k(\cdot)$ by the method of generating functions involves numerical solution of some rather unwieldy differential equations. The following approach, which makes use of properties of recurrent Markov chains, is easier to implement computationally.

In Appendix A we show that, for $\lambda/2p < 1$, $P_k(\cdot)$ is given by $c(\boldsymbol{\alpha} + r\boldsymbol{\beta})$, where c and r are constants and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are the two (particular) eigenvectors, defined by (9) below, of the transition matrix $P_k(i, j)$.* We further show that the condition, $\alpha_j + r\beta_j \geqq 0$ for each j, determines r uniquely and provides an easy way to calculate r. Once r is obtained, the constant c is uniquely determined by the condition that $\sum_{j=0}^{\infty} P_k(j) = 1$. Since the quantities α_j and β_j can be computed recursively, the above observations lead to a simple algorithm for computing $P_k(\cdot)$. The results are given below, the proof is in Appendix A. Let

$$\alpha_0 = 1$$

$$\alpha_1 = 0$$

$$\alpha_{n+2} = \frac{\alpha_n - \sum_{j=0}^{n+1} \alpha_j P_k(j, n)}{P_k(n+2, n)} \qquad n \geqq 0 \tag{9a}$$

$$\beta_0 = 0$$

$$\beta_1 = 1$$

$$\beta_{n+2} = \frac{\beta_n - \sum_{j=0}^{n+1} \beta_j P_k(j, n)}{P_k(n+2, n)} \qquad n \geqq 0. \tag{9b}$$

It is shown in Appendix A that there is a unique constant r satisfying

$$m_n \leqq r \leqq M_n \qquad \text{all } n \geqq 3 \tag{10a}$$

$$\lim_{n\to\infty} m_n = r = \lim_{n\to\infty} M_n, \tag{10b}$$

where, for $n \geqq 3$, the increasing sequence m_n and the decreasing sequence M_n are defined by

$$\begin{aligned} m_n &= \max_{j \leqq n} \left[-\frac{\alpha_j}{\beta_j} : \alpha_j < 0, \beta_j > 0 \right] \\ M_n &= \min_{j \leqq n} \left[-\frac{\alpha_j}{\beta_j} : \alpha_j > 0, \beta_j < 0 \right] \cdot \end{aligned} \tag{11}$$

Since the quantities m_n and M_n can be computed recursively, eqs. (10a), (10b), and (11) yield a well-defined algorithm for computing r.

Once r has been computed, $P_k(j)$ can be calculated from (3), setting $P_k(0) = c$, $P_k(1) = rc$ and determining c by the requirement that $\sum_{j=0}^{\infty} P_k(j) = 1$.

* Each of c, r, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ depends on k. For each integer $j \geqq 0$, α_j and β_j, denote the jth components of the eigenvectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

Since recursive calculation of $P_k(j)$ by (3) involves successive subtractions, it is essential to perform all computations in double precision.

In a No. 5 crossbar switching machine, the measured DTM occupancy during an hour is defined to be the total work time (in seconds) of all the DTMs divided by the product of the number of DTMs and the number of seconds in an hour. Thus, the measured occupancy is the average fraction of time that a given DTM is occupied. In the model the (steady-state) DTM occupancy is defined to be the limit, as n approaches infinity, of the expected fraction of time that a given DTM is occupied during the n work cycles of length T beginning at times $T, 2T, \cdots, nT$. This limit is equal to the probability that a given DTM is occupied on a work cycle beginning with the system in steady state. Hence,

$$\begin{aligned}\text{DTM occupancy} &= P\text{ (two or more calls are waiting for dial tone)}\\ &\qquad + \tfrac{1}{2}P\text{ (exactly one call is waiting for dial tone)}\\ &= 1 - P\text{ (no calls are waiting for dial tone)}\\ &\qquad - \tfrac{1}{2}P\text{ (exactly one call is waiting for dial tone).}\end{aligned}$$

As discussed earlier, the number of bad (blocked) calls waiting for dial tone at any given time has a truncated Poisson distribution in the model. Let

x = total erlangs of blocked calls on two DTMs

K = maximum possible number of blocked calls that can be in the system (waiting for dial tone) at any time*

$$c_{xK} = \sum_{k=0}^{K} \frac{x^k}{k!}.$$

Then the probability that k blocked calls are in the system is given by $(x^k/k\,!)c_{xK}^{-1}$. Let ρ_u denote the (steady state) DTM occupancy for the case of no blocked calls in the system and let ρ_b denote the (steady state) DTM occupancy including the effect of blocked calls. It follows from (5) that

$$\rho_u = \lambda/2p. \tag{12a}$$

Hence,

$$\begin{aligned}\rho_b &= 1 - c_{xK}^{-1}P_0(0) - \tfrac{1}{2}\left[c_{xK}^{-1}P_0(1) + xc_{xK}^{-1}P_1(0)\right]\\ &= 1 - c_{xK}^{-1}(1-\rho_u) - (x/2)c_{xK}^{-1}P_1(0).\end{aligned} \tag{12b}$$

2.3 Delay equations for the limiting model

In this section, we calculate the dial-tone delay probabilities for a call that arrives when the system (just prior to the arrival of the call)

[*] The truncation parameter K determines the number of conditional delay distributions to be calculated, so K should be taken to be no larger than needed to retain sufficient accuracy in the calculations. For values of x in the range of interest $e^{-x}c_{xK}$ differs from 1 by less than about 10^{-3} for $K = 5$; thus, for computational purposes, K may be taken to be 5.

is in steady state. Recall that the queuing model is in discrete time with arrivals at times nT and DTMs operating at times nT. Calls served on the DTM cycle beginning at nT either exit the system at time $(n + 1)T$ or else return to the queue at time $(n + 1)T$.

Thus, the steady-state probability that a call experiences a dial-tone delay of $(m + 1)T$ seconds can be computed recursively in terms of the probability of a delay of mT seconds by straightforward conditional probability calculations involving matrix multiplications.

Let

Y = number of bad calls (permanently) present in the system
X_n = number of good calls in the queue at time nT
$\hat{X}$ = number of good calls in the queue immediately after an arrival when, just prior to the arrival, the system was in steady state.

Thus, $\hat{X}$ is the total queue length just after arrival of a call when the system is in steady state. Let

$$\hat{P}_k(j) = P\left\{X_n = j \,\middle|\, \begin{array}{l}\text{Queue is in steady state at time } (n-1)T \\ \quad\text{with } k \text{ bad calls permanently present in} \\ \quad\text{the system, and at least one good call} \\ \quad\text{arrives at time } nT\end{array}\right\}.$$

Then

$$P[\hat{X} = j \,|\, Y = k] = \hat{P}_k(j).$$

Let

$$\hat{P}_k(i, j) = P\left[X_{n+1} = j \,\middle|\, \begin{array}{l}X_n = i,\ Y = k, \text{ and at least one good} \\ \quad\text{call arrives at time } nT\end{array}\right]$$

and

$$\begin{aligned}\hat{A}(n) &= \frac{A(n)}{1 - A(0)} \qquad && n \geqq 1 \\ &= 0 && \text{otherwise.}\end{aligned}$$

Then $\hat{A}(n)$ is the conditional probability that n calls arrive at time nT given that at least one call arrives at time nT and $\hat{P}_k(i, j)$ is defined by eqs. (1) and (2) with $\hat{A}$ used in place of A. Also,

$$\hat{P}_k(j) = \sum_{i=0}^{j+2} P_k(i)\hat{P}_k(i, j).$$

Let $W_n(i, k)$ be the conditional probability that a call arriving in steady state has a dial-tone delay of nT seconds, given that the queue length of good calls upon arrival of the call is i and that the number of bad calls permanently present in the system is k. Thus, letting D

denote the dial-tone delay, we have

$$W_n(i, k) = P[D = nT \,|\, \hat{X} = i \text{ and } Y = k] \qquad i \geqq 1.$$

Then

$$W_1(i, k) = \min\left[1, \frac{2}{(k+i)}\right] p \qquad i \geqq 1,$$

and $W_{n+1}(\cdot, k)$ can be computed recursively from $W_n(\cdot, k)$ as described below.

Let $X_n = i \geqq 1$ and consider any one given call out of the i calls present at time nT. Let

$$\mathcal{P}_k(i, j) = P\left[\begin{array}{l|l} X_{n+1} = j \text{ and the given call does} & \\ \text{not leave the system on the DTM} & X_n = i \text{ and } Y = k \\ \text{cycle beginning at time } nT & \end{array}\right].$$

Then

$$W_{n+1}(i, k) = \sum_{j=0}^{\infty} \mathcal{P}_k(i, j) W_n(j, k),$$

and, for $i \geqq 1$, $\mathcal{P}_k(i, j)$ is given by*

$$\mathcal{P}_0(i, j) = \left[1 - \frac{\min(2, i)}{i}\right][pA(j - i + 2) + qA(j - i)] + \left[\frac{\min(2, i)}{i}\right] qA(j - i)$$

$$\mathcal{P}_1(i, j) = \left[\frac{i-1}{i+1}\right]\left[\frac{p}{i}[(i-2)A(j - i + 2) + 2A(j - i + 1)] + qA(j - i)\right] + \left[\frac{2}{1+i}\right] qA(j - i)$$

and, for $k \geqq 2$,

$$\mathcal{P}_k(i, j) = \left[1 - \frac{2}{k+i}\right]\left\{p \sum_{l=0}^{2} \frac{\binom{i-1}{l}\binom{k}{2-l}}{\binom{k+i-1}{2}} A(j - i + l) + qA(j - i)\right\} + \left(\frac{2}{k+i}\right) qA(j - i).$$

The second term on the right-hand side of the above equation is the probability of the event that: (*i*) the given call *is* selected for

* The derivation of $\mathcal{P}_k(i, j)$ for $k \geqq 2$ is given below. The derivations of $\mathcal{P}_k(i, j)$ for $k = 0$ and $k = 1$ are analogous and are omitted.

service by the DTMs, (*ii*) all ORs are busy, and (*iii*) $X_{n+1} = j$. (Recall that, in the model, either all ORs are busy, with probability q, or else two or more ORs are available, with probability $p = 1 - q$). The first term is the probability that: (*i*) the given call is *not* selected for service and (*ii*) $X_{n+1} = j$. In the first term, the expression in large brackets is the conditional probability that $X_{n+1} = j$ given that the call is not selected for service. This conditional probability is itself composed of the probabilities of two mutually exclusive events. The second term within the large brackets is the (conditional) probability of the event that all ORs are busy and $X_{n+1} = j$. The first term within the large brackets is the (conditional) probability that two or more ORs are free and $X_{n+1} = j$. The sum from $l = 0$ to $l = 2$ in this term pertains to the cases where 0, 1, or 2 good calls, respectively, are selected for service by the DTMs. Since this sum is part of a probability conditional upon a given good call not having been selected by the DTMs, the available population from which the DTMs may select calls consists of k bad calls ($k \geqq 2$) and $i - 1$ good calls ($i \geqq 1$). Since the selection is without replacement, the selection probabilities have the hypergeometric form shown above.

By the law of total probabilities,

$$P[D = nT \mid Y = k] = \sum_{i=1}^{\infty} W_n(i, k) P[\hat{X} = i \mid Y = k],$$

$$P[D > nT \mid Y = k] = 1 - \sum_{m=1}^{n} P[D = mT \mid Y = k],$$

and

$$P[D > nT] = \sum_{k=0}^{K} \frac{c_{xK}^{-1} x^k}{k!} P[D > nT \mid Y = k], \tag{13}$$

where

$$c_{xK} = \sum_{k=0}^{K} x^k / k!.$$

III. CALCULATION OF OFFERED BLOCKED-CALL LOAD

In the limiting analytical model, the number of blocked calls in the system has a time-independent truncated Poisson distribution with mean x. This section describes a method for computing x, the mean offered erlangs of bad calls, in terms of the distribution of carried load among horizontal groups in the office. Using the limiting analytical model in conjunction with these methods for calculating the offered blocked-call load, we can calculate the No. 5 crossbar dial-tone delay distribution in terms of the calling rate and the distribution of carried load among horizontal groups in the office. Thus, we can predict the effect of frame load and balance on dial-tone delay in No. 5 crossbar offices.

We first construct a model for an individual horizontal group and express the expected bad-call load from one horizontal group as a function of the carried horizontal group load. To get the total expected bad-call offered load for the office, the expected contribution from an individual horizontal group is integrated (numerically) over the office horizontal group load distribution. The model, described below, for representing an individual horizontal group may be called a modified finite source Palm delay model.

We assume that the input to a horizontal group is from a finite number of sources with equal calling rates and that call-holding times are exponential. The number of sources N is taken to be 35 although in actuality most horizontal groups have 49 or 59 subscriber lines. The reason for the use of the lower number of sources is that an earlier study by W. S. Hayward, Jr.[5] showed that blocking on concentrators with unequal line occupancies can be approximated by blocking calculations based on equal calling rates and a lower number of sources. The use of 35 sources was suggested by J. G. Kappel.[6]

The calculations take into account the fact that an incoming call cannot occupy a subscriber line when all ten line links are occupied. Calls that find all line links busy will either defect or will eventually obtain a line link. While waiting for a line link to become available, a call is assumed to have an exponential waiting time until defection, with a mean of 30 s.*

In the case of ideal load balance, each horizontal group in the office is assumed to have a true carried load of $\bar{z}$ erlangs. In the case of less than ideal load balance, the distribution of true carried load among the individual horizontal groups is assumed to be normal with mean $\bar{z}$ and coefficient of variation c_g, where $\bar{z}$ is the office average carried horizontal group load and where c_g is the group-to-group coefficient of load variation for the office. The term c_g may be inferred from office load balance data either by using analysis of variance or, more commonly, by subtracting a standard value of the residual variance from the total measured variance of the office horizontal group load distribution.

* This is the same value that was used in step-by-step dial-tone-delay calculations. (See Ref. 7.) These calculations are based on the Palm delay model[8] using an assumed mean call holding time of 150 s and an assumed j factor of 5. In the notation of Ref. 9, this value of j corresponds to a mean-time-to-defection of 30 s. In Ref. 9, it is also stated that this value was found to be slightly conservative for most applications, based on review of panel office data reported in Ref. 9 and other (unpublished) step-by-step data.

To obtain the total offered blocked-call load for the office, we write the offered blocked-call load for an individual horizontal group $x(z)$ as a function of the individual horizontal group carried load z and integrate the function over the office distribution of carried load. We now describe how $x(z)$ is calculated.

Let N denote the number of subscriber lines per horizontal group. As discussed above, N has been taken to be 35 in all computations. Let λ_n denote the combined originating and terminating rate when n subscriber lines are occupied and let μ_n denote the subscriber line hang-up rate when n subscriber lines are occupied. Then

$$\begin{aligned} \lambda_n &= \lambda(N-n) \qquad && 0 \leqq n \leqq 9 \\ &= \frac{\lambda}{2}(N-n) \qquad && 10 \leqq n \leqq N-1 \end{aligned}$$

and

$$\begin{aligned} \mu_n &= n/H \qquad && 0 \leqq n \leqq 10 \\ &= 10/H + (n-10)/H_b \qquad && 11 \leqq n \leqq N, \end{aligned}$$

where H denotes the office average line-link holding time and H_b denotes the reciprocal of the defection rate for a call that is waiting for a line link to become available. As discussed above, H_b is taken to be 30 s based on results in Ref. 9.

The parameter λ (combined originating and terminating rate per unoccupied subscriber line) is an unknown whose value will be obtained from the horizontal group-carried load z. The factor of $\frac{1}{2}$ appearing in the definition of λ_n for $10 \leqq n \leqq N-1$ reflects the fact that, when all 10 line links are busy, an incoming call cannot cause a subscriber line to be occupied. The definition of the hang-up rate μ_n for $11 \leqq n \leqq N$ reflects the assumption that, when all 10 line links are occupied, the holding times of the subscriber lines for which no line links are available should be shorter than full call holding times.

For a horizontal group with carried load z, let

π_n = steady-state probability that n (out of N) subscriber lines are occupied.

Then, using standard methods for computing the steady-state distribution of a birth-and-death process,[10]

$$\pi_0 = c \text{ and } \pi_n = c\left(\frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n}\right) \quad \text{for } n \geqq 1,$$

where the constant c is determined by the requirement that

$$\sum_{n=0}^{N} \pi_n = 1.$$

Then the carried horizontal group load z is given by

$$z = \sum_{n=1}^{9} n \cdot \pi_n + 10 \sum_{n=10}^{N} \pi_n. \tag{14}$$

It is not hard to show that the carried horizontal group load z is a strictly increasing function of the subscriber line occupancy rate λ and that $z \leqq \min [10, \lambda NH]$. This makes it easy to determine numerically the unique value of λ corresponding to a given carried load z.

Once λ has been determined, the quantities λ_n, μ_n, and π_n are used to compute the offered blocked-call load contributed by the horizontal group. We take $x(z)$ to be the expected number of occupied subscriber lines for which no line links are available. This yields the value

$$x(z) = \sum_{n=11}^{N} (n - 10)\pi_n. \tag{15}$$

IV. PEAKEDNESS OF THE BLOCKED-CALL STREAM

In the model discussed in Section III, blocked calls arrive according to a Poisson process and defect after an exponential waiting time. Since blocked calls constitute an overflow stream and since it is well known[11] that overflow traffic usually has a peakedness* greater than 1 and hence is not Poisson, some discussion of the peakedness of the blocked-call stream is in order.

The blocked-call stream is the superposition of overflow traffic from all of the (typically 400 to 600) horizontal groups in an office. In the case of an office with ideal load balance (identical horizontal group loads), all horizontal groups would have equal expected contributions to the blocked-call stream and standard limit theorems would suggest that the blocked-call stream should be approximately Poisson.

Comparisons of calculated and observed dial-tone delays discussed in Section V (covering measured DTM occupancies up to about 0.84) indicate that, when the blocked-call stream is assumed to be Poisson, the calculated delays generally fall in the midrange of applicable data. However, there is a large variability in observed dial-tone delays measured in the same office under nearly identical levels of DTM occupancy, second-failures-to-match, and percent all-ORs-busy. The presence of this variability may be regarded as evidence that, in some busy hours, the blocked-call stream may be peaked in nature. Peakedness of the blocked-call stream is capable of accounting for substanti-

* The peakedness of a stream of calls is, by definition, the variance-to-mean ratio of the number of busy servers when the stream is offered to an infinite group of servers with independent identically distributed exponential holding times.

ally higher calculated dial-tone delay at a given level of DTM occupancy, second-failures-to-match, and percent all-ORs-busy than would be calculated under the assumption that the blocked-call stream is Poisson.

One would expect the blocked-call stream to be peaked whenever most of the blocked calls are contributed by a small number of highly overloaded horizontal groups. Data on the horizontal group load distributions during individual busy hours are not available for the test discussed in Section V and would be impractical to obtain on an ongoing basis in any office. In the absence of data from which one could deduce directly the blocked-call peakedness, a treatment has been made using some results which R. I. Wilkinson obtained in the course of formulating his "equivalent random" method of characterizing overflow traffic.[11]

Wilkinson[11] assumes that traffic arrives and departs according to a birth-and-death process in which the arrival rate is increased whenever the number of calls in the system exceeds a nominal number and is decreased whenever the number of calls in the system is less than this nominal number. The equilibrium distribution of the number of calls in the system resulting from these assumptions is shown to be negative binomial and, hence, this distribution is completely determined by its mean and variance (or by its mean and peakedness). Wilkinson then shows that negative binomial distributions rather closely approximate true overflow distributions under a number of different conditions.

The effects of peakedness of the blocked-call stream on dial-tone delay were explored using the limiting analytical model of Section II by replacing the (truncated) Poisson distribution of blocked calls by a (truncated) negative binomial distribution of blocked calls. (Note that this does not require recomputing the conditional delay distributions.) In this manner, dial-tone delay distributions were calculated and compared with the observed dial-tone-delay distributions discussed in Section V, assuming peakedness values of 2 and 4. The resulting delay curves approximated the higher delays observed for given occupancy parameters. These calculations indicate that much of the observed variability in dial-tone delays under nearly identical load can be explained by peakedness of the blocked-call load.

It is not difficult to modify the birth-and-death model discussed in Appendix B so as to accommodate peaked blocked-call input using Wilkinson's method of representing this input.[12]

Thus, when a negative binomial blocked-call distribution is used in the limiting model, the resulting distribution of the good-call dial-tone delay may be regarded as the limit of a sequence of good-call-delay distributions corresponding to models with blocked-call input having the peaked form suggested by Wilkinson.[11] The limit is taken as the bad-call-arrival rates approach zero and the mean bad-call-waiting times approach infinity with their products approaching positive constants. Incorporation of the negative binomial distribution in the limiting model is accomplished simply by replacing the terms of the truncated Poisson distribution in eq. (13) with the corresponding terms of the negative binomial distribution,[13] using any convenient truncation.

V. COMPARISON OF CALCULATED AND OBSERVED DIAL-TONE DELAY DISTRIBUTIONS

5.1 Summary

Theoretical dial-tone-delay distributions, calculated using the analytical model discussed in Section II, were compared with dial-tone-delay measurements made in a field test. The test was conducted in a No. 5 crossbar office with 60 LLFs, 4 DTMs, 68 dial-pulse (DP) ORs, and 68 multifrequency (MF) ORs. The data discussed in this section are from the time period February through April, 1974.

The main conclusions of this study are that (*i*) the calculated delays generally fall in the midrange of applicable data, (*ii*) there is a large variability in observed dial-tone delays measured under nearly identical levels of DTM occupancy, second-failures-to-match, and percent of all-ORs-busy, and (*iii*) the field data show a clear increase in the ratio of waste DTM usage to total DTM usage as frame load increases. The observed amount of increase in waste DTM usage agrees with theoretical predictions.

In this section, the manner in which the dial-tone-delay measurements were taken is discussed and the method used to obtain the calculated delays is outlined. Next, some of the sources of variability in No. 5 crossbar dial-tone delays are identified and an explanation is given as to why a large variability in observed delays should be expected. Two data plots are given indicating, respectively, the effects of frame load on waste DTM usage due to second-failures-to-match and the effects of frame load on incoming-first-failures-to-match. Finally, the conclusions of the study are discussed.

5.2 Methods of obtaining the calculated and observed delay distributions

Hourly dial-tone-delay measurements were made by placing approximately 900 test calls per hour, using a standard 3-s dial-tone-delay testing machine which had been modified to record the proportion of test calls with delays exceeding X seconds, for $X = 0.5$, 1.0, 1.5, 2.0, 2.5, and 3.0.

The *observed actual* DTM *occupancy* $\hat{\rho}_b$ for a single hour is expressed in terms of the measured parameters by the formula:

$$\hat{\rho}_b = \frac{(\text{Total DTM peg count})(0.015) + (\text{Measured seconds of DTM usage})}{(4)(3600)}. \tag{16}$$

The term in eq. (16) involving the DTM peg count is to adjust for the seconds of DTM usage which DTM usage measuring devices do not record.

The (adjusted) DTM holding time for each hour was computed by A. R. Thorne from the DTM peg count, the all-ORs-busy peg count, the adjusted DTM usage [the latter of which comprises the numerator of eq. (16)], and an additional "light-traffic adjustment" (used whenever the observed actual DTM occupancy is below 0.80). Most of the adjusted DTM holding times are between 0.28 s and 0.31 s. A DTM holding time of 0.30 s is assumed in the theoretical dial-tone-delay distributions discussed in this section.

To compare predicted dial-tone delays with measured dial-tone delays, it was necessary to infer the observed increment in DTM occupancy due to DTM second-failures-to-match (DTM2FTM). This increment, denoted by Δ, is taken to be

$$\Delta = \frac{(\text{Total DTM2FTM peg count})(HBC + 0.015)}{(4)(3600)}, \tag{17}$$

where

HBC = DTM holding time during a second-failure-to-match (seconds).

A value of 0.40 s is used for HBC, based on data obtained during an earlier dial-tone-delay field test.[6] The *observed good-call* DTM *occupancy* $\hat{\rho}_u$ is defined to be

$$\hat{\rho}_u = \hat{\rho}_b - \Delta. \tag{18}$$

Dial-tone-delay distributions were calculated, using the analytical model, for a range of values of actual DTM occupancy ρ_b and good-call DTM occupancy ρ_u, where the parameters ρ_b and ρ_u are defined in Section II. The specific manner in which the distributions were calculated is discussed below. The results of these calculations were tabulated into a set of dial-tone-delay distributions, indexed by the pairs (ρ_b, ρ_u). To compare the calculated and observed dial-tone delays, measured dial-tone-delay distributions from data-collection hours with similar values of $\hat{\rho}_b$ and $\hat{\rho}_u$ were plotted on a graph along with one theoretical dial-tone-delay distribution selected from the tabulation discussed above, such that $\rho_b \approx \hat{\rho}_b$ and $\rho_u \approx \hat{\rho}_u$ hold for the values $\hat{\rho}_b$ and $\hat{\rho}_u$ corresponding to the observed delays shown on the graph.

These graphs are shown in Figs. 3, 4, and 5. Table I lists data pertaining to each of the observed delay curves on each of the graphs. At the top of each graph are listed the actual DTM occupancy ρ_b and the good-call DTM occupancy ρ_u used for the theoretical dial-tone-delay curve (the solid line) on the graph. Also listed are the ranges of the $\hat{\rho}_b$ and $\hat{\rho}_u$ values corresponding to the observed dial-tone-delay curves on the graph. The plotting symbols on the graphs indicate measured dial-tone delays. The dotted lines are smoothing curves fitted to the measured delays by the computer plotting routine used to draw the graphs.

The predicted dial-tone-delay distributions were obtained in several steps. First, values of the good-call origination rate per LLF λ_1 were

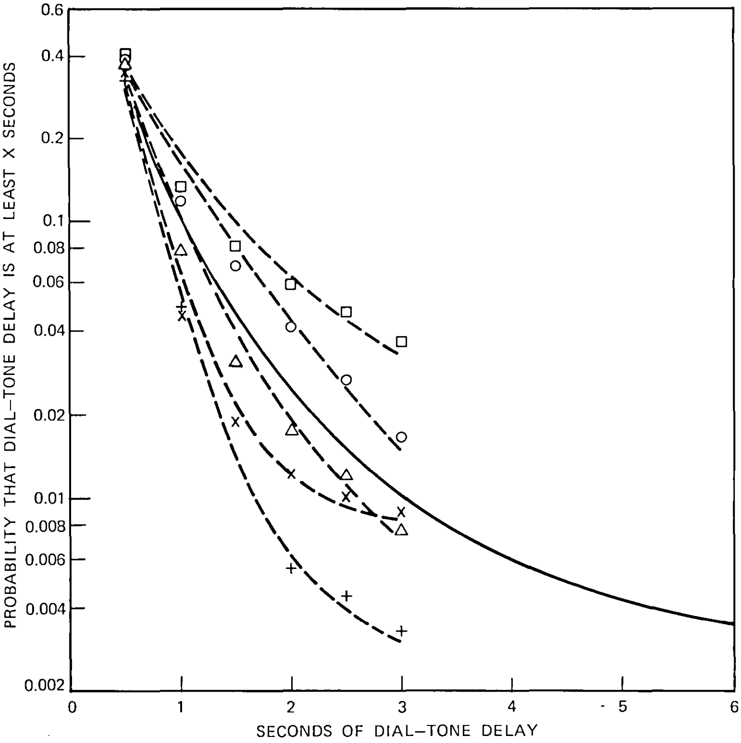


Fig. 3—Calculated and observed dial-tone delays. Actual occupancy in calculations = 0.75 (data: 0.740 to 0.758). Good-call occupancy in calculations = 0.70 (data: 0.684 to 0.697).

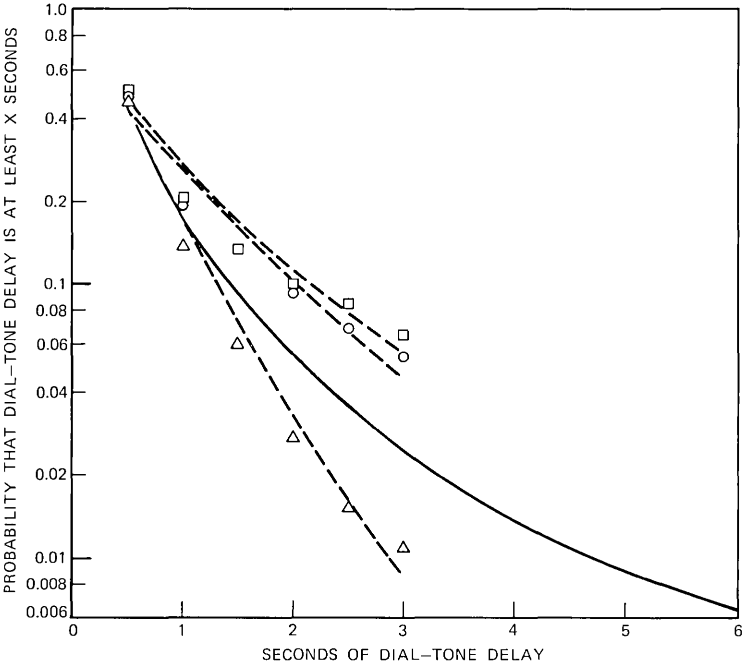


Fig. 4—Calculated and observed dial-tone delays. Actual occupancy in calculations = 0.82 (data: 0.795 to 0.837). Good-call occupancy in calculations = 0.76 (data: 0.744 to 0.782).

computed so as to produce good-call DTM occupancies of $\rho_u = 0.60$, 0.66, 0.68, 0.70, 0.72, 0.74, 0.76, 0.78, and 0.80. The values of λ_1 were obtained numerically from the formula $\rho_u = \lambda_1 NT/2p$ given in Section III. Note that p, the erlang C probability of all-ORs-busy, is a function of λ_1, whereas N and T are constants. For the office in which the test was conducted, $N = 30$ and, as discussed above, $T = 0.30$. In computing p, an average OR holding time of 10.25 s was assumed based on data from the test. In all cases, the calculated values of p were greater than 0.99, so the all-ORs-busy condition has a calculated probability of less than 0.01 under the conditions to which these distributions apply. (The observed fractions of all-ORs-busy were also below 0.01 during most of the hours of the test. Hence, ORs do not appear to have caused much of the dial-tone delay observed in the test.)

Next, the conditional delay distributions corresponding to these parameters were computed using the analytical model. Using eq. (12), values of x (total erlangs of blocked calls) corresponding to a range

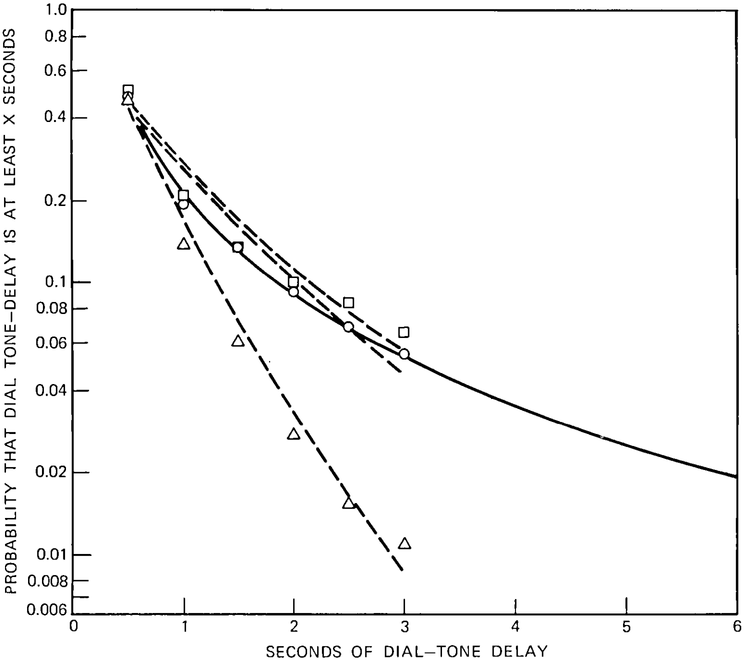


Fig. 5—Calculated and observed dial-tone delays. Actual occupancy in calculations = 0.82 (data: 0.795 to 0.837). Good-call occupancy in calculations = 0.76 (data: 0.744 to 0.782). (Note that a blocked-call peakedness of 4 is assumed in the calculations shown above. Figures 3 and 4 are based on an assumed blocked-call peakedness of 1.)

of values of ρ_b were computed for each fixed value of ρ_u. For each such pair of ρ_b and ρ_u, the good-call dial-tone-delay distributions were then computed by (13) using the value of x, obtained as discussed above, and the conditional delay distributions corresponding to ρ_u. The result of these calculations is a set of good-call dial-tone-delay distributions indexed by the pairs (ρ_b, ρ_u).

These good-call dial-tone-delay distributions include the *indirect* effect of bad calls in that they reflect the increased DTM congestion produced by the bad calls. As discussed earlier, the *direct* effect of bad calls is expected to be small and therefore may be accounted for in a somewhat approximate manner. To reduce the number of variables that need to be considered, a single blocking probability is used in lieu of integrating the blocking probability formula over an assumed distribution of expected horizontal group loads. For the purpose of computing an average blocking probability PB, the average waiting-

Table I — Dial-tone-delay data used for comparison with calculated delays

Fig. No.	Date	Time	% Actual DTM Occupancy	% Good-Call Occupancy	No. DTS Test Calls	% AORB	CCS/LLF	0.5 s*	1.0 s	1.5 s	2.0 s	2.5 s	3.0 s
3	3/4/74	1500–1600	74.0	69.7	900	0.33	N.A.	370	120	73	53	42	33
	3/25/74	1500–1600	74.5	68.4	900	0.30	1344	354	106	62	37	24	15
	4/9/74	1530–1630	75.5	69.4	900	0.00	1315	298	44	9	5	4	3
	2/6/74	1500–1600	75.8	69.7	900	0.01	1438	314	41	17	11	9	8
	2/22/74	1600–1700	74.7	69.4	900	0.06	1351	341	71	28	16	11	7
4 and 5	2/6/74	1600–1700	80.9	74.4	900	0.61	1480	461	188	122	90	76	59
	2/15/74	1600–1700	80.4	76.1	900	0.92	1390	430	175	121	83	62	49
	4/8/74	1530–1630	83.7	78.2	900	0.25	1384	426	126	53	25	14	10

* For each hour listed, the number in this column is the observed number of dial-tone speed (DTS) test calls whose dial-tone delay exceeds 0.5 s. The columns labeled "1.0 s" through "3.0 s" are defined analogously.

time-until-defection of a bad call (BCWT) is assumed to be 10 s.* A consequence of this assumption is that the total bad-call origination rate corresponding to x total erlangs of bad calls on two DTMs is $x/10$ bad calls per second. The total origination rate of good calls on all 30 LLFs is $30\lambda_1$. The average blocking probability PB corresponding to the occupancy pair (ρ_b, ρ_u) is then defined to be the ratio of the bad-call origination rate to the total origination rate of good and bad calls. Thus,

$$\text{PB} = \frac{x/10}{x/(10) + 30\lambda_1}, \tag{19}$$

where x and λ_1 are the values corresponding to the occupancy pair (ρ_b, ρ_u).

The conditional delay distribution of bad calls should be nearly exponential with mean delay $\text{HT}/(10 - a)$, where HT is the average call holding time and a is the carried horizontal group load in erlangs.† For each occupancy pair (ρ_b, ρ_u), the theoretical dial-tone-delay distribution for all calls is then given by

$$P(D > t) = (1 - \text{PB})P_G(D > t) + \text{PB}\left\{\exp\left[-\frac{(10 - a)\,t}{\text{HT}}\right]\right\}, \tag{20}$$

where $P_G(D > t)$ is the good-call dial-tone-delay distribution corresponding to the occupancy pair (ρ_b, ρ_u).

In the theoretical dial-tone-delay curves shown in Figs. 3, 4, and 5, $a = 0$ is used in eq. (20). The effect on $P(D > 3\text{ s})$ of setting $a = 0$, rather than using a more nearly correct value for each graph, is less

* The value of BCWT = 10 s used in calculating PB is consistent with the mean-time-to-defection $H_b = 30$ s used in the horizontal group blocking calculations in Section III. The difference between the numerical values arises because these two parameters are defined differently. For the purpose of calculating PB, it is assumed that a bad call arrives, remains waiting for dial tone an exponential length of time with mean BCWT, and then defects. In the horizontal group blocking model discussed in Section III, a call which finds all 10 line links busy may either defect or may eventually obtain an idle line link. The call contributes to the bad-call load during the time that it remains waiting for one of the line links to become available. Thus, the quantity in the horizontal group blocking model corresponding most nearly to the parameter BCWT is the mean time until a call that finds all 10 line links busy either obtains a line link or defects. Based on delay calculations for the horizontal group blocking model discussed in Section III, and assuming a mean-time-to-defection of 30 s for calls that do not obtain a line link and a mean line-link holding time of 150 s, the mean time until a call that finds all 10 line links busy either obtains a line link or else defects is calculated to be between about 10 and 12 s for horizontal group carried loads in the range of interest. Thus, it is reasonable to take BCWT = 10 s.

† This expression for the conditional delay distribution of bad calls should overestimate their delays somewhat because the expression does not account for finite-source effects. Since the fraction of calls experiencing these delays (i.e., the fraction of bad calls) is typically less than 0.01, the effect on 3.0-s dial-tone-delay probabilities of neglecting finite-source effects is typically less than 5×10^{-4}. For this reason, these effects are neglected in eq. (20).

than 8×10^{-4} in all cases. This error is in the opposite direction from the error, of comparable magnitude, resulting from not including finite-source effects in the bad-call delay distribution.

Since the analytical model is a discrete time model with a time step of T seconds, $P(D > kT) = P[D \geqq (k+1)T]$ for $k = 1, 2, \cdots$. In Figs. 3, 4, and 5, $T = 0.30$. Because of the way that dial-tone-delay measurements are taken, a call which is recorded as having a delay greater than t seconds may actually receive dial tone within a few milliseconds after time t. Thus, in comparing the theoretical dial-tone delays with the observed dial-tone delays, the observed fraction of dial-tone delays *greater than t* seconds is taken as representing the observed fraction of dial-tone delays *greater than or equal to t* seconds. The theoretical delay curve plotted is the curve $P(D \geqq kT)$ for $k = 1, 2, \cdots$, interpolated so as to produce a smooth curve.

5.3 Sources of variation in observed No. 5 crossbar dial-tone delays

Dial-tone delay in No. 5 crossbar is influenced by a number of factors capable of producing a large variation in delays measured in different hours within the same office under very similar conditions of DTM occupancy, percent all-ORs-busy, and second-failures-to-match. As discussed in Section IV, much of the variability in dial-tone delay measured under very similar load conditions can be explained by differences in peakedness of the blocked call stream. Whenever most of the blocked-call load comes from a small number of extremely overloaded horizontal groups, the blocked-call stream should have a peakedness greater than one. When a large number of moderately overloaded groups contribute to the blocked-call load, the blocked-call stream should be approximately Poisson (peakedness equal to one). Thus, differences in the individual busy-hour-load balance would be expected to produce different amounts of blocked-call peakedness, which in turn can account for appreciable differences in dial-tone delay measured under nearly identical average load conditions. For example, Fig. 5 shows dial-tone delays calculated assuming a blocked-call peakedness of 4 for the same conditions shown in Fig. 4, which is based on a blocked-call peakedness of 1.* The calculated delay curve in Fig. 5 fits the top two observed delay distributions rather closely.

Some additional identifiable sources of dial-tone-delay variation under similar load conditions are: (*i*) within-hour trends in traffic, (*ii*) nonstandard (and possibly erratic) gating caused by improper functioning of the master traffic controller circuitry, (*iii*) DTM prefer-

* The method by which blocked-call peakedness is treated in the model is discussed in Section IV.

ence for calls from a small subset of lines on each horizontal group, (*iv*) variation in DTM first-failures-to-match, and (*v*) competition between DTMs and completing markers for line-link connectors.

The first of these sources should produce effects similar to those of blocked-call peakedness. The second source may cause nonuniform congestion. The third source is predicted to result in a slight outward shift in the delay curve. The fourth and fifth sources should be reflected in increased measured DTM holding time and in increased DTM occupancy. Although approximate allowances can be made for the average congestion increase produced by some of these phenomena, no quantitative estimate is available for their total contribution to hourly *variation* in dial-tone delay.

In addition to identifiable sources of dial-tone-delay variation, simulation studies indicate that there can be an appreciable residual variation in simulated hourly No. 5 crossbar dial-tone delays obtained in different runs with identical input parameters (and, hence, with identical expected load conditions).*

Figure 6 shows four dial-tone-delay distributions obtained using the gating simulation model. In this model, the blocked-call stream is Poisson. These distributions were produced by simulating four individual hours, using identical input parameters. The delay distributions shown are for the calls that did not encounter horizontal group blocking and are based on the total number of such calls processed during the hour. The set of four 3-s dial-tone delays has a coefficient of variation of 0.28 and a mean of 0.069. (The coefficient of variation is the ratio of the standard deviation to the mean.) Plotted on the graph along with the delay curves are error bars indicating the 2-sigma limits of 0.034 and 0.107 associated with the above mean and coefficient of variation.

Actual dial-tone-delay measurements are based on test calls. During a given busy hour in a typical No. 5 crossbar office, approximately 900 test calls are made on a fixed set of 60 (out of 600) horizontal groups. The use of test calls introduces sampling error, which is not represented in the distributions shown in Fig. 6 and which would have an associated coefficient of variation of about 0.12 for the parameters applicable to Fig. 6.

5.4 Data on frame load effects

Figure 7 is a data plot of line-link frame load versus waste DTM usage due to second-failures-to-match based on data from the test. The quantity "DTM usage fraction due to 2FTMs" shown on the ordinate is

* This result is one of the main conclusions of an earlier No. 5 crossbar dial-tone-delay simulation study conducted by S. Halfin.[3]

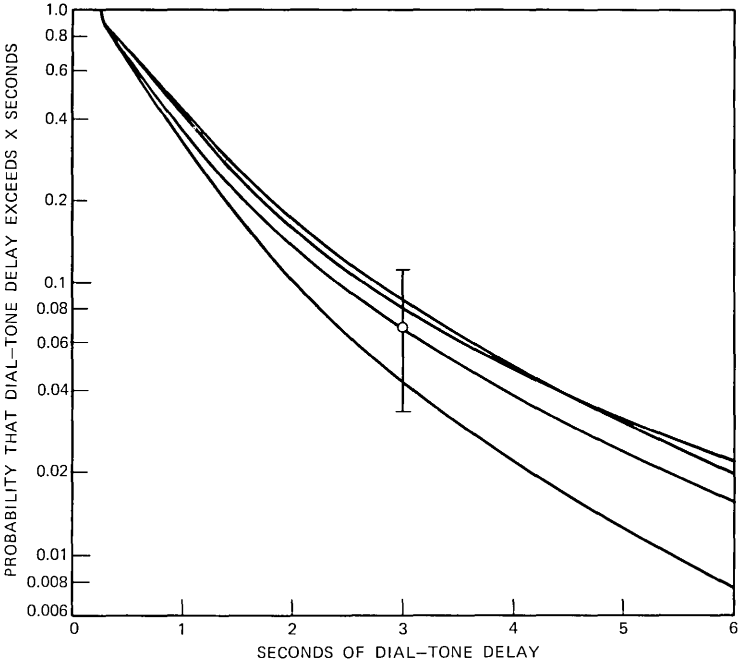


Fig. 6—Simulated No. 5 crossbar dial-tone delays. Distributions are based on simulation of four individual hours with identical inputs.

defined by

$$\text{DTM usage fraction due to 2FTMs} = \frac{\text{Total DTM usage due to 2FTMs}}{\text{Total DTM usage}}$$

$$= \frac{\Delta}{\hat{\rho}_b}.$$

Figure 8 is a data plot of line-link frame load versus incoming-first-failure-to-match (IFFM) based on data from the test. In each figure, each of the data points represents data from 1 h. The average actual DTM occupancy (averaged over all the data-collection hours with measured line-link frame loads of 1100 CCS/LLF or more) is 0.54; the DTM occupancy range is from less than 0.40 to 0.84.

Existing theory indicates that, for a given office configuration (including a given junctor pattern), IFFM is directly and primarily dependent on frame load. This conclusion is borne out by Fig. 8, which demonstrates a well-defined trend (with a moderate amount of data scatter) of increasing IFFM with increasing frame load.

The extent to which frame load affects DTM usage and dial-tone delay in any given hour depends, in an indirect way, on several differ-

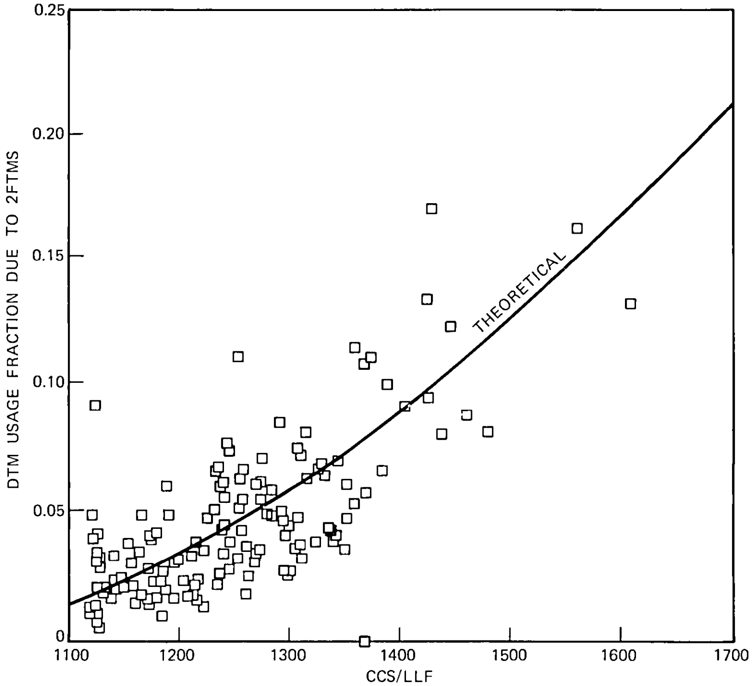


Fig. 7—Effect of frame load on increased DTM usage due to 2FTMs. The assumed DTM holding time during a 2FTM = 0.40 s.

ent variables, including DTM occupancy and the distribution of carried load among horizontal groups. Because of this dependence, the nature of which has been discussed more fully in previous sections, the data scatter in the plot of waste DTM usage fraction versus frame load (Fig. 7) is much larger than in the plot of IFFM versus frame load (Fig. 8).

The curve labeled "THEORETICAL" in Fig. 7 was calculated using the limiting analytical model [eq. (12)] and the horizontal group blocking model discussed in Section III. In these calculations, the blocked-call load was assumed to be Poisson (peakedness equal to 1).

To obtain the theoretical curve, it was first necessary to determine what calling rates should be assumed in the calculations. The calling rates were inferred from the data upon which Fig. 7 is based by first using linear regression (least squares) to express the observed good-call DTM occupancy, $\hat{\rho}_u$, as an empirical function of the observed frame load, CCS/LLF. For each value of the frame load, the calling rate was taken to be the particular calling rate corresponding to the least squares value of $\hat{\rho}_u$, assuming an average DTM holding time of 0.30 s and a

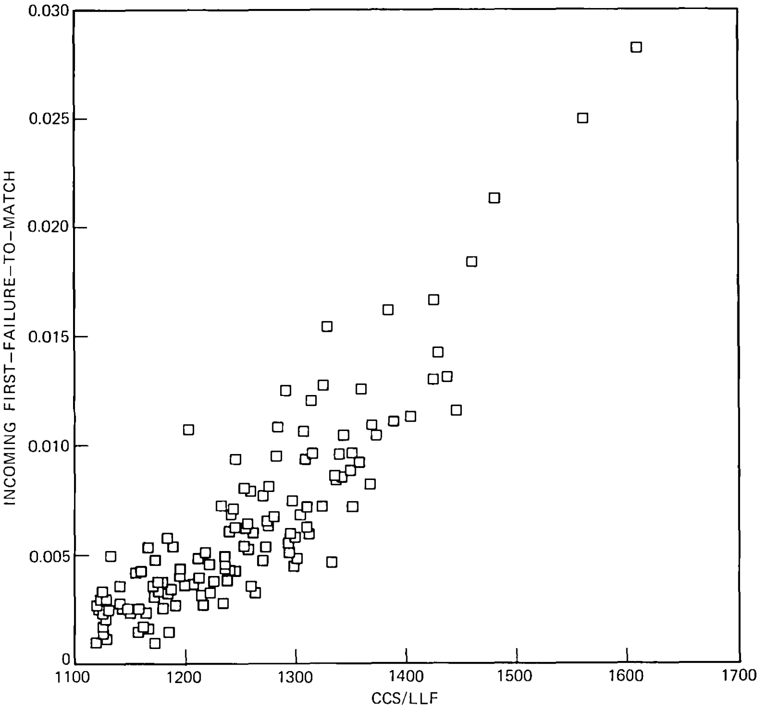


Fig. 8—Effect of frame load on incoming first-failure-to-match.

zero probability of all-ORs-busy. The justification for the assumptions regarding the DTM holding time and the probability of all-ORs-busy is given earlier in this section.

Next, the horizontal group blocking model discussed in Section III was used to compute the expected blocked-call offered load corresponding to each value of frame load and (empirically associated) calling rate. The office distribution of carried horizontal group load was assumed to be normal with group-to-group coefficient of variation inferred from the office horizontal group load distribution using the method discussed earlier. The calculated values of offered blocked-call load were then used in eq. (12) to compute the theoretically predicted fraction of waste DTM use, $(\rho_b - \rho_u)/\rho_b$, corresponding to these frame loads and calling rates.

Since both the theoretical and the observed values of waste DTM use depend not only on frame load but also on the calling rate, neither the data plotted in Fig. 7 nor the theoretical curve shown on the figure should be regarded as being applicable to calling rates or load balance conditions other than those upon which this figure is based.

The purpose of Fig. 7 is to illustrate—for the conditions of calling rate, frame load, and load balance represented in this study—an empirical relationship between frame load and the ratio of waste DTM usage to total DTM usage with increasing frame load and to show that this empirical relationship is consistent with predictions of the theoretical models.

5.5 Conclusions

The conclusions which the author has drawn from this comparison of theoretical and observed dial-tone-delay distributions are as follows:

(*i*) Where two or more observed delay distributions appear on the same graph, the theoretical delay distributions usually fall approximately midway between the maximum and minimum observed delay distributions. This indicates that, in the (actual) DTM occupancy range spanned by these data (DTM occupancies up to about 0.84), the analytical model shows good agreement with the data.

(*ii*) Much of the large variability in observed 3-s dial-tone delays measured under nearly equal conditions of DTM occupancy, second-failures-to-match, and percent all-ORs-busy can be explained by assuming the blocked-call stream to have different peakedness values (ranging from 1 to about 4) in different hours. High blocked-call peakedness, illustrated by Fig. 5 in which the peakedness is taken to be four, would be expected whenever most of the blocked-call load comes from a small number of extremely overloaded horizontal groups. Low blocked-call peakedness, illustrated by Fig. 4 in which the peakedness is taken to be 1, would be expected whenever a large number of moderately overloaded horizontal groups contribute more or less equally to the blocked-call load.

(*iii*) Simulation results indicate that there should be a large residual variability in 3-s dial-tone delays measured in different (simulated) hours under identical expected load conditions. This variability, illustrated in Fig. 6, is in addition to the variability due to blocked-call peakedness discussed above. A third nonnegligible source of variability in observed dial-tone delays is the sampling variability associated with the use of test calls to measure delays.

(*iv*) For the conditions of calling rate, frame load, and load balance represented in this study, the observed increase in the ratio of waste DTM usage to total DTM usage as frame load increases is consistent with theoretical predictions. This is illustrated in Fig. 7 and discussed in Section 5.4.

VI. ACKNOWLEDGMENT

The author would especially like to thank J. G. Kappel for guiding this work and contributing numerous suggestions. In addition, the author appreciates useful comments, results, and suggestions contributed by A. L. Dudick, S. Halfin, W. S. Hayward, Jr., N. A. Marlow, A. R. Thorne, and F. R. Wallace of Bell Laboratories; H. S. McKane and H. A. Reading of AT&T; and J. C. Warner of Michigan Bell Telephone Company.

APPENDIX A

Computation of the Two-Marker Stationary Distribution for $k \geqq 1$

The purpose of this appendix is to prove that the limit r defined by (10) exists and that the quantities $v_j = \alpha_j + r\beta_j$, satisfy (21), where α_j and β_j are defined by (9).

$$\begin{aligned} v_j &\geqq 0, && \text{all } j \geqq 0 \\ v_0 &= 1; \\ v_j &= \sum_{i=0}^{j+2} v_i P_k(i, j), && \text{all } j \geqq 0 \\ &\sum_{j=0}^{\infty} v_j < \infty. \end{aligned} \tag{21}$$

We first show that there exists exactly one positive number r such that $v_j = \alpha_j + r\beta_j$ satisfies (21). Then we show that the limit in (10) exists and is equal to this number.

When $\lambda < 2p$, we know that the stationary distribution P_k exists and $P_k(j) > 0$ for each j. Let $\pi_j = P_k(j)/P_k(0)$. Then for $r = P_k(1)/P_k(0)$, we have

$$\begin{aligned} \pi_0 &= 1, \\ \pi_1 &= r, \\ \pi_j &= \sum_{i=0}^{j+2} \pi_i P_k(i, j), && \text{all } j \geqq 0. \end{aligned}$$

Now let $v_j = \alpha_j + r\beta_j$, with r as defined above. Then, it follows from (9) that

$$\begin{aligned} v_0 &= 1, \\ v_1 &= r, \\ v_j &= \sum_{i=0}^{j+2} v_i P_k(i, j), && \text{all } j \geqq 0. \end{aligned} \tag{22}$$

Since $v_0 = \pi_0$ and $v_1 = \pi_1$ and since the vectors $\mathbf{v}$ and $\boldsymbol{\pi}$ both satisfy the same recurrence relation, which has the property that the jth term for each $j \geqq 2$ is uniquely determined by the 0th and 1st terms, we see that $v_j = \pi_j > 0$ for all $j \geqq 0$. Since the Markov chain is positive recurrent for $\lambda/2p < 1$, it follows from (22) and Karlin[14] that $\sum_{j=0}^{\infty} v_j < \infty$. Hence, v satisfies (21). Thus, there exists at least one value of r for which $v_j = \alpha_j + r\beta_j$ satisfies (21). If r' is any number such that $v_j' = \alpha_j + r'\beta_j$ satisfies (21), then by uniqueness of the stationary distribution, $r' = v_1'/v_0' = P_k(1)/P_k(0)$, so $r' = r$. Thus, exactly one such r exists.

Since $\alpha_j + r\beta_j > 0$ for each j, it follows that (10a) holds. Hence, the increasing sequence m_n is bounded above by r and the decreasing sequence M_n is bounded below by r. So both $m \equiv \lim_n m_n$ and $M \equiv \lim_n M_n$ exist and satisfy

$$m \leqq r \leqq M, \tag{23}$$

$$\begin{aligned} m_n &\leqq m, \\ M &\leqq M_n. \end{aligned} \tag{24}$$

If either of the inequalities in (23) were strict, then $m < M$ and, in view of (24), each x in the interval $m < x < M$ would satisfy

$$m_n \leqq x \leqq M_n$$

for each n, from which it follows that $v_j = \alpha_j + x\beta_j$ satisfies (21). This contradicts the uniqueness of r and proves that $m = M = r$.

It is easy to extend the above result so as to give a method for computing the stationary distribution of any irreducible positive recurrent Markov chain on the nonnegative integers such that a fixed positive integer n_0 exists for which $P(j + n_0, j) > 0$ and $P(j + n, j) = 0$ for all j and all $n > n_0$.

APPENDIX B

Convergence in Distribution to the Limiting Model

In the limiting model, the number of bad (i.e., blocked) calls in the system is a time-independent, truncated-Poisson-distributed random variable. We will show that the steady-state distributions of good- and bad-call queue lengths and the steady-state good-call dial-tone-delay distribution in the limiting model are the limits of the corresponding distributions for a sequence of models in which the bad call queue lengths form birth-and-death processes.

In the nth model, the bad-call arrival rate, denoted by $\lambda^{(n)}$, approaches zero and the bad-call mean waiting time until defection, denoted by $H^{(n)}$, approaches infinity in such a way that their product (total expected erlangs of bad calls) approaches a finite positive constant x.

The physical motivation for considering the limiting model is that the actual situation is one in which blocked calls appear infrequently (relative to total call arrival rates) and tend to remain in the system for a long time (relative to DTM holding times). The mathematical motivation is that the limiting model is much easier to analyze than a model in which the bad-call arrival and departure processes are represented explicitly.

Preliminary computations using a single marker version of the analytical model, which represented the bad-call arrival and departure processes explicitly, showed that different bad-call arrival rates and waiting times had little effect on dial-tone-delay distributions as long as the product of the bad-call arrival rate and bad-call waiting time remained constant. In addition, these distributions were all quite close to those obtained from the limiting form of the model. These results make sense intuitively because bad calls simply cycle through the system, absorbing some DTM uses while present, and eventually defect; thus, what should matter is mainly the distribution of the number of bad calls in the system at any time. In the version of the analytical model discussed in this appendix, the blocked call queue length distribution is shown to be a truncated Poisson with mean equal to the total erlangs of bad calls.

Let K be the maximum possible number of bad calls that can be in the system at any time. In all computations using realistic No. 5 crossbar busy-hour input parameters, the expected erlangs of bad calls have been low enough that the Poisson probability of more than five blocked calls being present in a two-marker system has been less than 10^{-3}. Hence, K may be taken to be 5. (Note that in an actual No. 5 crossbar office, the maximum number of bad calls that can be in the system at any time is trivially bounded above by the number of subscriber line terminations in the office.)

Let $\hat{S}_n(p)$ denote the saturation load of the nth model. Since no more than K blocked calls can be in the queue at any time, in the nth model, we have $2p = S_0^{(2)}(p) \geqq \hat{S}_n(p) \geqq S_K^{(2)}(p) = 2p$ for all n; hence, $\hat{S}_n(p) = 2p$. Thus, the saturation load of the nth model is the same as the saturation load of the limiting model. For the nth model, let

$$\begin{aligned} X_m^{(n)} &= \text{number of good calls in the queue at time } mT, \\ Y^{(n)}(mT) &= \text{number of bad calls in the queue at time } mT, \\ p_n(k, l) &= P\{Y^{(n)}[(m+1)T] = l \,|\, Y^{(n)}(mT) = k\}. \end{aligned}$$

In the nth model, $Y^{(n)}(t)$, the bad-call population size at time t, is a finite-state space birth-and-death process with birth-and-death rates

$$\begin{aligned} \lambda_m^{(n)} &= \lambda^{(n)} && 0 \leqq m \leqq K-1 \\ &= 0 && \text{otherwise} \\ \mu_m^{(n)} &= \frac{m}{H^{(n)}} && 0 \leqq m \leqq K \\ &= 0 && \text{otherwise.} \end{aligned}$$

Since the queuing model is in discrete time, only the values of $Y^{(n)}(t)$ at $t = mT$ are of interest. The transition matrix $p_n(k, l)$ is given by

$$\mathbf{p} = e^{T\mathbf{A}_n},$$

where $\mathbf{A}_n$ is the infinitesimal generator matrix of the birth-and-death process $Y^{(n)}$. (See Ref. 15.) Since $\lim_{n\to\infty} \lambda^{(n)} = 0$ and $\lim_{n\to\infty} H^{(n)} = \infty$, it follows that $\lim_{n\to\infty} \mathbf{A}_n = \mathbf{0}$. Hence,

$$\lim_{n\to\infty} p_n(k, l) = \delta_{kl} \tag{25}$$

for each $k, l = 0, 1, \cdots, K$, where $\delta_{kl} = 0$ for $k \neq l$ and $\delta_{kk} = 1$.

We also have

$$\begin{aligned} P\{X_{m+1}^{(n)} = j,\ Y^{(n)}[(m+1)T] = l \,|\, X_m^{(n)} = i,\ Y_{(mT)}^{(n)} = k\} \\ = p_n(k, l)P_k(i, j), \end{aligned}$$

where $P_k(i, j)$ is given by (1) and (2). Since the nth model has saturation load $2p$, we know that a stationary queue length distribution $\pi_n(j, l)$ exists for $\lambda/2p < 1$ and satisfies the equation

$$\pi_n(j, l) = \sum_{k=0}^{K} \sum_{i=0}^{j+2} \pi_n(i, k)p_n(k, l)P_k(i, j). \tag{26}$$

Let $\pi_n(\cdot, l)$ be the marginal equilibrium distribution of the number of bad calls in the system. Then

$$\pi_n(\cdot, l) = \sum_{j=0}^{\infty} \pi_n(j, l) \tag{27}$$

and by (26)

$$\begin{aligned} \pi_n(\cdot, l) &= \sum_{k=0}^{K} \sum_{i=0}^{\infty} \pi_n(i, k)p_n(k, l) \sum_{j=0}^{\infty} P_k(i, j) \\ &= \sum_{k=0}^{K} \pi_n(\cdot, k)p_n(k, l). \end{aligned} \tag{28}$$

Hence, $\pi_n(\cdot, k)$ is the stationary distribution of a Markov chain whose transition matrix is $p_n(k, l)$. Using standard theorems on birth-and-death processes (Ref. 10) it follows that, for all values of $l = 0, 1, \cdots, K$,

$$\begin{aligned} \pi_n(\cdot, k) &= \lim_{t \to \infty} P[Y^{(n)}(t) = k \,|\, Y^{(n)}(0) = l] \\ &= \frac{c_{x_nK}^{-1} x_n^k}{k!} \qquad 0 \leqq k \leqq K \\ &= 0 \qquad\qquad \text{otherwise,} \end{aligned} \tag{29}$$

where

$$x_n = \lambda^{(n)} H^{(n)}$$

and

$$c_{x_nK} = \sum_{k=0}^{K} \frac{x_n^k}{k!}.$$

We will now show that

$$\lim_{n \to \infty} \pi_n(j, k) = \frac{c_{xK}^{-1} x^k}{k!} P_k(j) \tag{30}$$

for all $k = 0, 1, \cdots, K$ and all $j = 0, 1, \cdots,$ where P_k is the stationary distribution of good-call queue length in the model with k bad calls permanently present in the system. The right-hand side of (30) is the queue-length distribution for good and bad calls in the limiting model.

Let $X^{(n)}$ denote the number of good calls in the queue for the nth model in equilibrium. Since the number of bad calls present in the nth model is always less than or equal to K for all n, it is clear on intuitive grounds and can be proved rigorously using stochastic ordering that

$$P[X^{(n)} \geqq j] \leqq \sum_{i=j}^{\infty} P_K(i) \tag{31}$$

for each j.

To prove (30), it suffices to show that if $\pi_{n'}$ is any subsequence of π_n for which $\lim_{n'} \pi_{n'}(j, k) \equiv \pi(j, k)$ exists for each j and k, then

$$(i) \quad \pi(j, k) = \sum_{i=0}^{j+2} \pi(i, k) P_k(i, j)$$

and

$$(ii) \quad \sum_{j=0}^{\infty} \pi(j, k) = \frac{c_{xK}^{-1} x^k}{k!}.$$

To see that (i) and (ii) are sufficient, note that, by uniqueness of the stationary distribution, P_k, (i) and (ii) imply

$$\pi(j, k) = \frac{c_{xK}^{-1} x^k}{k!} P_k(j). \tag{32}$$

Since $0 \leqq \pi_n(j, k) \leqq 1$ for all j and k, every subsequence $\pi_{n'}$ contains a further subsequence $\pi_{n''}$ for which the limit $\lim_{n''} \pi_{n''}(j, k)$ exists for all j and k. In view of the above discussion, all of these subsequences have the same limit, namely the right-hand side of (32). Equation (30) follows.

We now prove (*i*) and (*ii*). Condition (*i*) follows immediately from (25) and (26). To see that (*ii*) holds, we proceed in two steps. First, note that by (27), (29), and Fatou's lemma,

$$\sum_{j=0}^{\infty} \pi(j, k) \leqq \frac{c_{xK}^{-1} x^k}{k!}$$

for each k. If any of the above inequalities were strict, then

$$\sum_{k=0}^{K} \sum_{j=0}^{\infty} \pi(j, k) < 1.$$

Hence, to prove (*ii*) it suffices to show

$$\sum_{i=0}^{\infty} \sum_{k=0}^{K} \pi(i, k) \geqq 1,$$

i.e., we must show that, in the limit, no probability mass escapes to infinity. Let $\epsilon > 0$ and choose j_0 such that

$$\sum_{j=j_0}^{\infty} P_K(j) < \epsilon.$$

Then, using eq. (31),

$$\begin{aligned}
\sum_{i=0}^{\infty} \sum_{k=0}^{K} \pi(i, k) &\geqq \sum_{i=0}^{j_0} \sum_{k=0}^{K} \pi(i, k) \\
&= \lim_{n''} \sum_{i=0}^{j_0} \sum_{k=0}^{K} \pi_{n''}(i, k) \\
&= \lim_{n''} P[X^{(n'')} \leqq j_0] \\
&\geqq \sum_{i=0}^{j_0} P_K(i) \\
&\geqq 1 - \epsilon.
\end{aligned}$$

Thus, (*ii*) holds and eq. (30) follows.

It can be readily shown, using the above results together with standard theorems, that the delay probabilities in the nth model converge to those in the limiting model. (See Refs. 16 and 17.)

REFERENCES

1. S. Halfin, unpublished work.
2. F. R. Wallace, personal communication.
3. S. Halfin, "An Approximate Method for Calculating Delays for a Family of Cyclic Type Queues," B.S.T.J., this issue, pp. 1733–1754.
4. H. Kushner, *Introduction to Stochastic Control,* New York: Holt, Rinehart, and Winston, 1971, p. 211.
5. W. S. Hayward, Jr., "Traffic Engineering and Administration of Line Concentrators," Congress Record, Paper No. 23, Second International Teletraffic Congress, The Hague, 1958.
6. J. G. Kappel, personal communication.
7. C. Clos and R. I. Wilkinson, "Dialing Habits of Telephone Customers," B.S.T.J., *31,* No. 1 (January 1952), p. 46, last paragraph.
8. J. Riordan, *Stochastic Service Systems,* New York: John Wiley, 1962, pp. 109–112.
9. Ref. 7, pp. 32–67.
10. S. Karlin, *A First Course in Stochastic Processes,* New York: Academic Press, 1968, p. 194.
11. R. I. Wilkinson, "Theories for Toll Traffic Engineering in the USA," B.S.T.J., *35,* No. 2 (March 1956), pp. 421–514.
12. Ref. 11, p. 453.
13. Ref. 11, p. 454, eq. (19).
14. Ref. 10, theorem 3.2.
15. Ref. 10, pp. 206–208.
16. P. Billingsley, *Convergence of Probability Measures,* New York: John Wiley, 1968, p. 224.
17. H. Royden, *Real Analysis,* 2nd Edition, New York: MacMillan, 1968, p. 232.

Contributors to This Issue

Iain Anderson, A.H.-W.C., 1964, Heriot-Watt College, Edinburgh, Scotland; M.Sc., 1966, University College, London, England; Ph.D., D.I.C., 1969, Imperial College, London, England; Bell Laboratories, 1970–1975. Mr. Anderson has studied topics in diffraction theory, antenna analysis, and radome design. Associate, IEE.

Ta-Shing Chu, B.S., 1955, The National Taiwan University; M.S., 1957, and Ph.D., 1960, Ohio State University; Bell Laboratories, 1963—. Mr. Chu has been engaged in research on tropospheric wave propagation and microwave antennas. He is currently concerned with electromagnetic problems in the area of satellite communications. Member, IEEE, Commission II of URSI, Sigma Xi, Pi Mu Epsilon.

Michael J. Gans, B.S. (E.E.), 1957, Notre Dame University, M.S., 1961, Ph.D. (E.E.), 1965, University of California at Berkeley; Bell Laboratories, 1966—. At Bell Laboratories, Mr Gans has been engaged in research on antennas for mobile radio and satellite communications.

Harry A. Guess, B.S. and M.S. (Applied Mathematics), 1964, Georgia Institute of Technology; U. S. Navy and Division of Naval Reactors, USAEC, 1964–1969; National Science Foundation Fellowship, 1969–1972; M.S. (Operations Research) and Ph.D. (Mathematics), 1972, Stanford University; Assistant Professor, University of Rochester, 1972–1973; Bell Laboratories, 1973—. Mr. Guess has worked principally in traffic engineering and network performance analysis.

Shlomo Halfin, M.Sc., 1958, and Ph.D. (Mathematics), 1962, The Hebrew University of Jerusalem (Israel); Bell Laboratories, 1968—. Mr. Halfin's work is in the field of operations research theory and its applications. Member, Operations Research Society of America, Society for Industrial and Applied Mathematics.

W. E. Legg, Rutgers University, 1945–1949; Bell Laboratories, 1945—. Mr. Legg has worked on dielectric lenses and microwave antennas for radio relay systems. He was involved in development of project Echo and Telstar tracking equipment and participated in mobile radio telephone experiments. He is presently engaged in antenna measurements for satellite communication in the Radio Research Laboratory.

John B. MacChesney, B.A. (Chemistry), 1951, Bowdoin College; Ph.D. (Geochemistry), 1959, Pennsylvania State University; Bell Laboratories, 1959—. Mr. MacChesney has worked on various materials-related topics. His current interests include materials and processes for the preparation of optical fibers. Member, Sigma Xi, American Ceramic Society.

Paul B. O'Connor, Newark College of Engineering; Bell Laboratories, 1959—. Mr. O'Connor has worked on ferrite preparation and single crystal growth, and has done research on ceramic glazes. He is co-inventor of the iron oxide see-through photomask and has been doing research on optical waveguides for the past two years.

Herman M. Presby, B.A., 1962, and Ph.D., 1966, Yeshiva University; Research Scientist, Columbia University, 1966–1968; Asst. Prof. Physics, Belfer Graduate School of Science, Yeshiva University, 1968–1972; Bell Laboratories, 1972—. Mr. Presby is engaged in studies on the properties of optical fiber waveguides.

Marvin R. Sambur, B.E.E., City College of New York, 1968; S.M., 1969, and Ph.D., 1972, Massachusetts Institute of Technology; Bell Laboratories, 1968—. At present, Mr. Sambur is engaged in automatic speaker verification and automatic speech recognition research in the Acoustics Research group. Member, MPA-TC subcommittee on Speech Recognition and Understanding, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

R. D. Standley, B.S., 1957, University of Illinois; M.S., 1960, Rutgers University; Ph.D., 1966, Illinois Institute of Technology; USASRDL, Ft. Monmouth, N.J., 1957–1960; IIT Research Institute, Chicago, 1960–1966; Bell Laboratories, 1966—. Mr. Standley has been engaged in research projects involving microwave, millimeter wave, and optical components. He is presently concerned with electron beam lithography as applied to fabrication of integrated optic devices. Member, IEEE, Sigma Tau, Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering. Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Computer & Control Abstracts, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electrical & Electronic Abstracts, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Metals Abstracts, Science Abstracts,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Bell System