February 1985 Vol. 64 No. 2 Part 1

AT&T TECHNICAL JOURNAL

A JOURNAL OF THE AT&T COMPANIES

Three-Dimensional Motion

Homenet

Queueing

Markovian Models

Speech Recognition

EDITORIAL BOARD

EDITORIAL STAFF

B. G. KING, *Editor*
P. WHEELER, *Managing Editor*
L. S. GOLLER, *Assistant Editor*
A. M. SHARTS, *Assistant Editor*
B. VORCHHEIMER, *Circulation*

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

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

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

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

AT&T TECHNICAL JOURNAL

VOL. 64 FEBRUARY 1985 NO. 2, PART 1

AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

Algorithms for Estimation of Three-Dimensional Motion

By A. N. NETRAVALI and J. SALZ*

(Manuscript received March 1, 1984)

We derive robust algorithms for estimating parameters of motion of rigid bodies that are observed by a television camera. Motion may be three-dimensional, containing both translational and rotational components, but the observations using the television camera are two-dimensional, i.e., projections on the camera plane. Our algorithms do not require a priori knowledge of any corresponding points in three- and two-dimensional spaces. We give both recursive as well as nonrecursive algorithms that minimize the error in intensity by using the estimated motion parameters. Our theory has applications in interframe coding, computer vision, and computer animation. The efficacy of our methods and the quality of the estimation procedures must await experimental verification.

I. INTRODUCTION

One of the most important problems in machine analysis of image sequences captured by a television camera is estimating the motion of objects in the field of view.[1] We have previously given algorithms for estimating the displacement vector when the motion is restricted to translation in a plane perpendicular to the camera axis.[2,3] This was later extended to situations where the illumination in the scene is spatially nonuniform[4] and to computationally more complex algorithms with better properties.[5] In this paper, we propose a further extension by developing algorithms for estimating parameters of three-

*Authors are employees of AT&T Bell Laboratories.

dimensional motion. The motion thus may have both the translational as well as rotational components, and the translation may not be in a plane perpendicular to the camera axis. Of course, although the object is three-dimensional and it moves in a three-dimensional space, the observations made by the television camera are still in a two-dimensional space, i.e., the object is observed by being projected from the object space to the image plane. Thus, information is lost in going from the three-dimensional object space to the two-dimensional image plane; this is a major source of difficulty in such estimation problems. It leads to nonunique solutions or ambiguous situations, unless additional information is made available. One such example of additional information is the correspondence of points in two- and three-dimensional space. This example is often used to determine camera position[6] or to make motion estimation unique.[7,8] However, in many practical problems such correspondence is either difficult or impossible to establish.

Our contribution in this paper is twofold. First, we develop equations of motion by noting the fact that a television camera creates a frame every thirtieth of a second. Most rigid body motion, in such a small amount of time, tends to be small. Therefore we develop models of incremental motion that each use three parameters for translation and rotation. Our second contribution is to give robust recursive and nonrecursive algorithms for estimating these parameters. The algorithms minimize the error in observed intensity by using these estimated motion parameters. Also, since the estimation algorithm is based on linearizing the intensity function, it is applicable in situations where the motion parameters are small. We also give an extension based on successive linearization that will work even when the motion is substantially large.

Some of the limitations of our approach should be pointed out. First, we are considering rigid body motion, i.e., no deformation of the body is allowed as a function of time. Second, parameters are estimated to minimize the intensity estimation error, and therefore, they may not correspond exactly to the true motion parameters, particularly since the problem may not have a unique solution due to loss of information in transforming from three-dimensionality to two-dimensionality. However, we believe that in most reasonable cases, parameters estimated by our procedure will be those corresponding to motion. Third, traditional difficulties with dynamic scene analysis, such as occlusion, spatial nonuniformity of motion parameters and illumination, and lack of proper segmentation, are largely ignored at this stage. They will be considered in our future work. Last, and perhaps most important, we have no simulation results to evaluate the performance of our algorithms. We hope, however, that since motion estimation is important in such diverse fields as computer animation, computer vision, and interframe coding, these algorithms will be specialized to many of these applications and then evaluated.

II. MOTION MODEL

In this section, we develop a model of three-dimensional motion that includes translation and rotation. The only constraint we impose is that the body in motion stay rigid. Let us assume that the location of different points changes in the object space as a result of object motion and that only a two-dimensional projection on the image plane is observable using a camera. (See Fig. 1.) Let a point P designated by a vector $\mathbf{r}$ = col. (x, y, z) move to another point P' designated by vector $\mathbf{r}'$ = col. (x', y', z'). Since the body stays rigid,

$$\bar{\mathbf{r}} = \tilde{\mathbf{R}}\bar{\mathbf{r}} + \overline{\mathbf{T}}, \tag{1}$$

where $\tilde{\mathbf{R}}$ is a three-by-three rotation matrix and $\overline{\mathbf{T}}$ is a three-dimensional translation vector. $\tilde{\mathbf{R}}$ can be represented in terms of the Eulerian

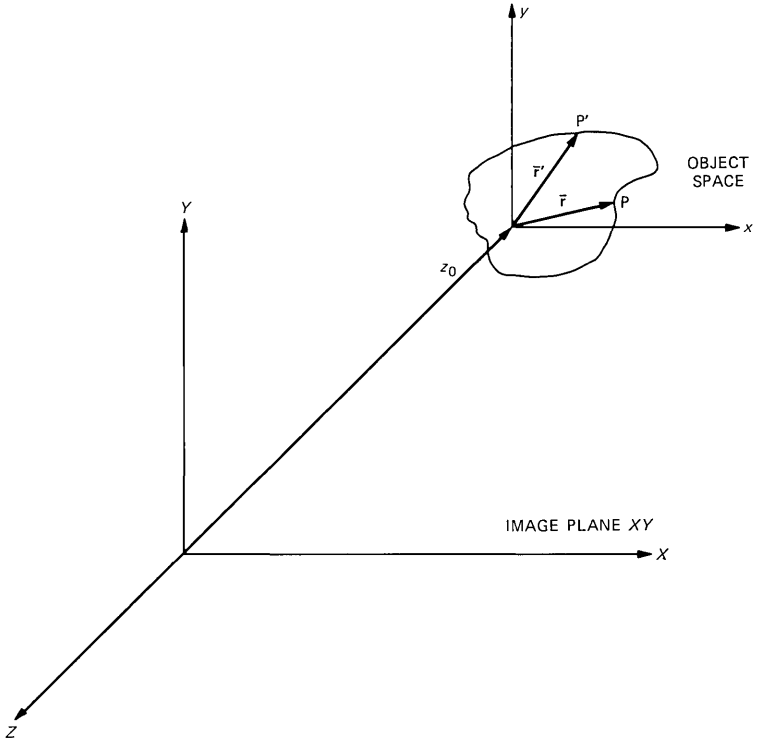


Fig. 1—Coordinate system showing object space and image plane.

angles ϕ, θ, and ψ as a product of three matrices, each corresponding to rotation about one axis. Thus

$$\tilde{\mathbf{R}} = \tilde{\mathbf{A}}\tilde{\mathbf{B}}\tilde{\mathbf{C}}, \tag{2}$$

where

$$\tilde{\mathbf{A}} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

$$\tilde{\mathbf{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \tag{4}$$

$$\tilde{\mathbf{C}} = \begin{pmatrix} \cos\psi & 0 & -\sin\psi \\ 0 & 1 & 0 \\ \sin\psi & 0 & \cos\psi \end{pmatrix}. \tag{5}$$

These equations then specify general rigid body transformations. In practice, a television camera observes a new scene every thirtieth of a second. During such a small time, changes in the parameters of motion (i.e., θ, ϕ, ψ, and $\overline{\mathbf{T}}$) will be small. We therefore specialize these equations to small or infinitesimal changes in motion parameters that have taken place within a frame time.

2.1 Infinitesimal motion

For infinitesimal motion, changes in Euler angles are small. If these are denoted by $\Delta\theta$, $\Delta\phi$, and $\Delta\psi$, then if we use approximations, $\cos\Delta\theta = 1$ and $\sin\Delta\theta = \Delta\theta$, eqs. (3), (4), and (5) become

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & \Delta\phi & 0 \\ -\Delta\phi & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6}$$

$$\tilde{\mathbf{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \Delta\theta \\ 0 & -\Delta\theta & 1 \end{pmatrix} \tag{7}$$

$$\tilde{\mathbf{C}} = \begin{pmatrix} 1 & 0 & -\Delta\psi \\ 0 & 1 & 0 \\ \Delta\psi & 0 & 1 \end{pmatrix}. \tag{8}$$

Therefore,

$$\begin{aligned}\tilde{\mathbf{R}} &= \tilde{\mathbf{A}}\tilde{\mathbf{B}}\tilde{\mathbf{C}} \\ &\cong \begin{pmatrix} 1 & \Delta\phi & -\Delta\psi \\ -\Delta\phi & 1 & \Delta\theta \\ \Delta\psi & -\Delta\theta & 1 \end{pmatrix} \\ &= \mathbf{I} + \begin{pmatrix} 0 & \omega_z & -\omega_y \\ -\omega_z & 0 & \omega_x \\ \omega_y & -\omega_x & 0 \end{pmatrix} \Delta t, \end{aligned} \tag{9}$$

where **I** is the identity matrix; and ω_x, ω_y, and ω_z are angular velocities about the x, y, and z axes, respectively. By substituting (9) into (1), we get

$$\bar{\mathbf{r}}' = \bar{\mathbf{r}}_{t+\Delta t} = \bar{\mathbf{r}}_t + \tilde{\mathbf{P}}\bar{\mathbf{r}}_t \Delta t + \tilde{\mathbf{T}}\Delta \mathrm{t}, \tag{10}$$

where

$$\tilde{\mathbf{P}} = \begin{pmatrix} 0 & \omega_z & -\omega_y \\ -\omega_z & 0 & \omega_x \\ \omega_y & -\omega_x & 0 \end{pmatrix} \tag{11}$$

and

$$\tilde{\mathbf{T}} = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}, \text{ vector of translational velocities.} \tag{12}$$

Next we denote the coordinates in the image plane by (X, Y). The transformation from object space to image plane then proceeds as follows:

$$X = \frac{z_0 x}{z} \tag{13}$$

$$Y = \frac{z_0 y}{z}, \tag{14}$$

where z_0 is the distance from the origin of the object space to the origin of the image plane (see Fig. 1). Now for small Δt, (10) becomes

$$x' = x + \omega_z y \Delta t - \omega_y z \Delta t + v_x \Delta t \tag{15}$$

$$y' = y - \omega_z x \Delta t + \omega_x z \Delta t + v_y \Delta t \tag{16}$$

$$z' = z + \omega_y x \Delta t - \omega_x y \Delta t + v_z \Delta t \tag{17}$$

and

$$\frac{x'}{z'} \cong \frac{x}{z} + \left(\omega_z \frac{y}{z} + \frac{v_x}{z} - \omega_y\right) \Delta t + \frac{x}{z}\left(-\omega_y \frac{x}{z} + \omega_x \frac{y}{z} - \frac{v_z}{z}\right) \Delta t. \tag{18}$$

Similarly,

$$\frac{y'}{z'} \cong \frac{y}{z} + \left(\omega_x + \frac{v_y}{z} - \omega_z \frac{x}{z}\right) \Delta t + \frac{y}{z}\left(-\omega_y \frac{x}{z} + \omega_x \frac{y}{z} - \frac{v_z}{z}\right) \Delta t. \quad (19)$$

By letting $z_0 = 1$ and using the definitions

$$X = \frac{x}{z}, \qquad X' = \frac{x'}{z'}$$

$$V_x = \frac{v_x}{z} \text{ and } V_y = \frac{v_y}{z}, \quad (20)$$

we obtain

$$X' = X + (\omega_z Y - \omega_y + V_x)\Delta t + \left(-X^2\omega_y + XY\omega_x - X\frac{v_z}{z}\right) \Delta t \quad (21)$$

$$Y' = Y + (-\omega_z X + \omega_x + V_y)\Delta t + \left(-XY\omega_y + Y^2\omega_x - Y\frac{v_z}{z}\right) \Delta t. \quad (22)$$

Let $a = v_z/z$ be the magnification parameter; then the differential movement of the coordinates in the image plane for infinitesimal motion is as follows:

$$\frac{dX}{dt} = \frac{X' - X}{\Delta t} = \omega_z Y - \omega_y(1 + X^2) + \omega_x XY + V_x - aX \quad (23)$$

$$\frac{dY}{dt} = \frac{Y' - Y}{\Delta t} = -\omega_z X + \omega_x(1 + Y^2) - \omega_y XY + V_y - aY. \quad (24)$$

Thus there are six unknowns (ω_x, ω_y, ω_z, V_x, V_y, a) that need to be evaluated to quantify motion. The only values that can be observed are the intensities of the image in the present and the previous frames. Several techniques can be formulated to estimate these parameters. In the following, two techniques are described in detail.

III. MOTION ESTIMATION

The first technique deals with situations where the intensity changes only slightly as a result of small changes in motion parameters, whereas the second technique does successive linearization and therefore can handle larger changes in intensity. Let $\mathbf{I}(X, Y, t)$ be the intensity function at time t. Then differential changes in intensity are expected by

$$\mathbf{I}(X, Y, t + \Delta t) = \mathbf{I}(X + \Delta_x \Delta t, Y + \Delta_y \Delta t, t), \quad (25)$$

where

$$\Delta_x = \omega_z Y - \omega_y(1 + X^2) + \omega_x XY - aX + V_x \tag{26}$$

$$\Delta_y = -\omega_z X + \omega_x(1 + Y^2) - \omega_y XY - aY + V_y. \tag{27}$$

Expanding the intensity function in power series in Δt yields

$$\begin{aligned}\mathbf{I}(X, Y, t + \Delta t) = \mathbf{I}(X, Y, t) &+ \frac{\partial}{\partial X}\mathbf{I}(X, Y, t)\\ &\cdot[\Delta_x]\Delta t + \frac{\partial}{\partial Y}\mathbf{I}(X, Y, t)\cdot[\Delta_y]\Delta t.\end{aligned} \tag{28}$$

Thus,

$$\begin{aligned}&\frac{\mathbf{I}(X, Y, t + \Delta t) - \mathbf{I}(X, Y, t)}{\Delta t}\\ &\quad = \omega_z[\mathbf{I}_x(X, Y, t)Y - \mathbf{I}_y(X, Y, t)X]\\ &\qquad - \omega_y[\mathbf{I}_x(X, Y, t)(1 + X^2) + \mathbf{I}_y(X, Y, t)XY]\\ &\qquad + \omega_x[\mathbf{I}_x(X, Y, t)XY + \mathbf{I}_y(X, Y, t)\cdot(1 + Y^2)]\\ &\qquad - a[\mathbf{I}_x(X, Y, t)X + \mathbf{I}_y(X, Y, t)Y]\\ &\qquad + V_x[\mathbf{I}_x(X, Y, t)] + V_y[\mathbf{I}_y(X, Y, t)].\end{aligned} \tag{29}$$

Let (X_i, Y_i) be a pel deemed to be from the set of "moving-area" pels, i.e., the frame difference at these locations is above a certain prespecified threshold. Then, for each such moving-area pel define the following six-dimensional vector

$$\boldsymbol{\phi}_1 = \begin{bmatrix} \mathbf{I}_x(X_i, Y_i, t)Y_i - \mathbf{I}_y(X_i, Y_i, t)X_i \\ -\mathbf{I}_x(X_i, Y_i, t)(1 + X_i^2) - \mathbf{I}_y(X_i, Y_i, t)X_iY_i \\ \mathbf{I}_x(X_i, Y_i, t)X_iY_i + \mathbf{I}_y(X_i, Y_i, t)(1 + Y_i)^2 \\ -\mathbf{I}_x(X_i, Y_i, t)X_i - \mathbf{I}_y(X_i, Y_i, t)Y_i \\ \mathbf{I}_x(X_i, Y_i, t) \\ \mathbf{I}_y(X_i, Y_i, t) \end{bmatrix}. \tag{30}$$

If

$$\mathbf{C} = \text{col.}\ (\omega_z, \omega_y, \omega_x, a, V_x, V_y)$$

denotes the six-dimensional parameter vector that needs to be estimated, then we can express the measured intensity difference, M_i,

$$M_i = \boldsymbol{\phi}_1^T\mathbf{C} + \text{noise}. \tag{31}$$

If the number of measurements is n, then the problem is to create a least-squares estimate of $\mathbf{C}$ (labeled $\hat{\mathbf{C}}_n$) that minimizes the following Mean-Squared Error (MSE) after these n measurements:

$$\text{MSE} = \min_{\mathbf{C}_n} \sum_{i=1}^{n} (M_i - \boldsymbol{\phi}_1^T \mathbf{C}_n)^2. \tag{32}$$

Carrying out the minimization, we get the set of equations

$$\sum_{i=1}^{n} \boldsymbol{\phi}_1 M_i = \left[\sum_{i=1}^{n} \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^T \right] \hat{\mathbf{C}}_n. \tag{33}$$

Thus, calculation of $\hat{\mathbf{C}}_n$ requires a matrix inversion at every step. The inversion can be carried out recursively as follows.
Let

$$\mathbf{A}_n = \sum_{i=1}^{n} \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^T \tag{34}$$

and

$$\eta_n = \sum_{i=1}^{n} \boldsymbol{\phi}_1 M_i. \tag{35}$$

Clearly,

$$\boldsymbol{\eta}_n = \boldsymbol{\eta}_{n-1} + \boldsymbol{\phi}_n M_n \tag{36}$$

and

$$\mathbf{A}_n = \mathbf{A}_{n-1} + \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T. \tag{37}$$

From the matrix inversion lemma in Ref. 9, we obtain

$$\mathbf{A}_n^{-1} = \mathbf{A}_{n-1}^{-1} - \frac{\mathbf{A}_{n-1}^{-1} \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{A}_{n-1}^{-1}}{1 + \boldsymbol{\phi}_n^T \mathbf{A}_{n-1}^{-1} \boldsymbol{\phi}_n}, \tag{38}$$

and when this is used in (33), we get the recursion

$$\hat{\mathbf{C}}_n = \hat{\mathbf{C}}_{n-1} - \frac{\mathbf{A}_{n-1}^{-1}}{1 + \boldsymbol{\phi}_n^T \mathbf{A}_{n-1}^{-1} \boldsymbol{\phi}_n} \cdot \boldsymbol{\phi}_n (\boldsymbol{\phi}_n^T \hat{\mathbf{C}}_{n-1} - M_n). \tag{39}$$

If

$$\frac{\mathbf{A}_{n-1}^{-1}}{1 + \boldsymbol{\phi}_n^T \mathbf{A}_{n-1}^{-1} \boldsymbol{\phi}_n} = \alpha = \text{constant}, \tag{40}$$

then the above reduces to a simple gradient algorithm.

If the motion was purely translational in the image plane and there was no zooming (i.e., $v_z = 0$), then

$$\omega_z = \omega_x = \omega_y = a = 0. \tag{41}$$

The estimation of motion parameters would then be analogous to our previous schemes. Matrix $\mathbf{A}_n$ would be two-by-two, and the vectors such as $\boldsymbol{\eta}_n$, $\mathbf{C}_n$ would be two-dimensional.

3.1 Successive linearization estimation

In the previous section we derived an iterative procedure that did not use the previous estimates in the linearization process. Improvement may be obtained if the intensity function is linearized at different locations in the previous frame based on the value of the previous estimates.[2] Thus, as before, let $\hat{\mathbf{C}}_{n-1}$ be the estimate of six parameters of motion made after observing a patch of $(n - 1)$ pels. We wish to revise this estimate to obtain $\hat{\mathbf{C}}_n$, which includes a patch of n pels obtained by adding a new pel to the previous patch of $(n - 1)$ pels. Define

$$\hat{\Delta}_x^{n-1} = \hat{\omega}_z^{n-1} Y - \hat{\omega}_y^{n-1}(1 + X^2) + \hat{\omega}_x^{n-1} XY - \hat{a}^{n-1} X + \hat{V}_x^{n-1} \tag{42}$$

$$\hat{\Delta}_y^{n-1} = -\hat{\omega}_z^{n-1} X + \hat{\omega}_x^{n-1}(1 + Y^2) - \hat{\omega}_y^{n-1} XY - \hat{a}^{n-1} Y + \hat{V}_y^{n-1}. \tag{43}$$

Also, define a new cost function DFD(·) to be*

$$\begin{aligned} &\mathrm{DFD}(X,\ Y,\ \hat{\Delta}_x^{n-1},\ \hat{\Delta}_y^{n-1},\ t) \\ &\qquad = \mathbf{I}(X,\ Y,\ t + \Delta t) - \mathbf{I}(X + \hat{\Delta}_x^{n-1}\Delta t,\ Y + \hat{\Delta}_y^{n-1}\Delta t,\ t). \end{aligned} \tag{44}$$

We note that, as defined above, DFD = 0 if

$$\begin{aligned} \hat{\Delta}_x^{k-1} &= \Delta_x, \quad \text{for any} \quad X,\ Y \\ \hat{\Delta}_y^{k-1} &= \Delta_y, \quad \text{for any} \quad X,\ Y, \end{aligned} \tag{45}$$

i.e., when our estimates of motion are equal to the true value of the parameters of motion. Also, for any given estimate of the motion parameters, DFD can be calculated if we know the intensities of two successive frames. As before, we can now expand DFD in Taylor's series. Thus

$$\begin{aligned} &\mathrm{DFD}(X,\ Y,\ \hat{\Delta}_x^{n-1},\ \hat{\Delta}_y^{n-1},\ t) \\ &\quad = \mathbf{I}(X,\ Y,\ t + \Delta t) - \mathbf{I}(X + \hat{\Delta}_x^{n-1}\Delta t,\ Y + \hat{\Delta}_y^{n-1}\Delta t,\ t) \\ &\quad = \mathbf{I}(X + \Delta_x\Delta t,\ Y + \Delta y\Delta t,\ t) - \mathbf{I}(X + \hat{\Delta}_x^{n-1}\Delta t,\ Y + \hat{\Delta}_y^{n-1}\Delta t,\ t) \\ &\quad = \mathbf{I}(X + \hat{\Delta}_x^{n-1}\Delta t + (\Delta_x - \hat{\Delta}_x^{n-1})\Delta t,\ Y + \hat{\Delta}_y^{n-1}\cdot\Delta t \\ &\qquad\quad + (\Delta_y - \hat{\Delta}_y^{n-1})\Delta t,\ t) - \mathbf{I}(X + \hat{\Delta}_x^{n-1}\cdot\Delta t,\ Y + \hat{\Delta}_y^{n-1}\Delta t,\ t). \end{aligned} \tag{46}$$

Let

$$\bar{\mathbf{I}} = \mathbf{I}(X + \hat{\Delta}_x^{n-1}\Delta t,\ Y + \hat{\Delta}_y^{n-1}\Delta t,\ t) \tag{47}$$

* As in Ref. 2, DFD stands for displaced frame difference.

$$
\begin{aligned}
&\frac{\mathrm{DFD}(X,\ Y,\ \hat{\Delta}_x^{n-1},\ \hat{\Delta}_y^{n-1},\ t)}{\Delta t}\\
&\quad\cong (\omega_z - \hat{\omega}_z^{n-1})[\overline{\mathbf{I}}_x Y - \overline{\mathbf{I}}_y X]\\
&\qquad - (\omega_y - \hat{\omega}_y^{n-1})[\overline{\mathbf{I}}_x(1 + X^2) + \overline{\mathbf{I}}_y XY]\\
&\qquad + (\omega_x - \hat{\omega}_x^{n-1})[\overline{\mathbf{I}}_x XY + \overline{\mathbf{I}}_y(1 + Y^2)]\\
&\qquad + (\hat{a}^{n-1} - a)[\overline{\mathbf{I}}_x X + \overline{\mathbf{I}}_y Y]\\
&\qquad + (V_x - \hat{V}_x^{n-1})[\overline{\mathbf{I}}_x]\\
&\qquad + (V_y - \hat{V}_y^{n-1})[\overline{\mathbf{I}}_y]. \qquad (48)
\end{aligned}
$$

Once again, define a vector of measured quantities

$$
\boldsymbol{\phi}_1^{n-1} = \begin{bmatrix} \overline{\mathbf{I}}_x Y_i - \overline{\mathbf{I}}_y X_i \\ -\overline{\mathbf{I}}_x(1 + X_i^2) - \overline{\mathbf{I}}_y X_i Y_i \\ \overline{\mathbf{I}}_x X_i Y_i + \overline{\mathbf{I}}_y(1 + Y_i^2) \\ -\overline{\mathbf{I}}_x X_i - \overline{\mathbf{I}}_y Y_i \\ \overline{\mathbf{I}}_x \\ \overline{\mathbf{I}}_y \end{bmatrix}. \qquad (49)
$$

Then

$$
\begin{aligned}
M_i^{n-1} &= \mathrm{DFD}(X,\ Y,\ \hat{\Delta}_x^{n-1},\ \hat{\Delta}_y^{n-1},\ t)/\Delta t\\
&= (\boldsymbol{\phi}_1^{n-1})^{\mathrm{T}}(\mathbf{C} - \hat{\mathbf{C}}_{n-1}) + \text{noise}. \qquad (50)
\end{aligned}
$$

The least-squares estimate, $\hat{\mathbf{C}}_n$, should minimize the following mean-squared error,

$$
\mathrm{MSE} = \min_{\mathbf{C}} \left\{ \sum_{i=1}^{n} [M_i^{n-1} - (\boldsymbol{\phi}_i^{n-1})^{\mathrm{T}}(\mathbf{C} - \hat{\mathbf{C}}_{n-1})]^2 \right. \qquad (51)
$$

for any given initial estimate $\hat{\mathbf{C}}_{n-1}$. As before, carrying out the minimization, we get

$$
\sum_{i=1}^{n} \boldsymbol{\phi}_i^{n-1} \cdot M_i^{n-1} = \left[\sum_{i=1}^{n} \boldsymbol{\phi}_i^{n-1} \boldsymbol{\phi}_i^{n-1^{\mathrm{T}}} \right] (\hat{\mathbf{C}}_n - \hat{\mathbf{C}}_{n-1}). \qquad (52)
$$

Then

$$
\hat{\mathbf{C}}_n = \hat{\mathbf{C}}_{n-1} + \left[\sum_{i=1}^{n} \boldsymbol{\phi}_i^{n-1} \boldsymbol{\phi}_i^{n-1^{\mathrm{T}}} \right]^{-1} \left[\sum_{i=1}^{n} \boldsymbol{\phi}_i^{n-1} M_i^{n-1} \right]. \qquad (53)
$$

As in the previous section, the matrix inversion lemma can be used to invert the matrix

$$\left[\sum_{i=1}^{n} \boldsymbol{\phi}_i^{n-1}\boldsymbol{\phi}_i^{n-1^{\mathrm{T}}}\right].$$

The real difference between this method and that of the previous section is that even if the motion is large (i.e., parameters of motion are somewhat large), the successive linearization, if it converges, gives more accurate estimates, since $(\mathbf{C} - \hat{\mathbf{C}}_{n-1})$ becomes smaller as iterations proceed.

IV. CONCLUSIONS

A mathematical theory that provides algorithms for robust estimation of a general set of motion parameters from frame sequences obtained from a television camera is now available. The theory was derived under mild assumptions. Both recursive and nonrecursive algorithms are provided. The efficacy of our algorithms has to be evaluated for each application. Potential applications are for interframe coding, computer vision, and computer animation.

REFERENCES

1. H. H. Nagel, "Overview of Image Sequence Analysis," in *Image Sequence Processing and Dynamic Scene Analysis*, T. S. Huang, Ed., New York: Springer-Verlag, 1983.
2. A. N. Netravali and J. D. Robbins, "Motion Compensated Television Coding," B.S.T.J., *58,* No. 3 (March 1979), pp. 629–68.
3. J. A. Stuller and A. N. Netravali, "Transform Domain Motion Estimation," B.S.T.J., *58,* No. 7 (September 1979), pp. 1673–782.
4. J. A. Stuller, A. N. Netravali, and J. D. Robbins, "Interframe Television Coding Using Gain and Displacement Compensation," B.S.T.J., *59,* No. 7 (September 1980), pp. 1227–40.
5. A. N. Netravali and J. D. Robbins, "Motion Compensated Television Coding: Some New Results," B.S.T.J., *59,* No. 9 (November 1980), pp. 1735–45.
6. S. K. Ganapathy, unpublished work.
7. R. Tsai and T. S. Huang, "Uniqueness and Estimation of Three Dimensional Motion Parameters of Rigid Objects with Curved Surfaces," CSL Technical Report R-921, University of Illinois, 1981.
8. B. Yen and T. S. Huang, "Determination of 3-D-Motion and Structure of a Rigid Body Using Spherical Projection," CSL Technical Report, University of Illinois, 1982.
9. B. D. O. Anderson and J. B. Moore, *Optimal Filtering, Information and System Science Series*, T. Kailath, Ed., Englewood Cliffs, NJ: Prentice-Hall, 1979.

AUTHORS

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, Ph.D. (Electrical Engineering), 1970, Rice University; Optimal Data Corporation, 1970–1972; AT&T Bell Laboratories, 1972—. Mr. Netravali has worked on problems related to filtering, guidance, and control for the space shuttle. At AT&T Bell Laboratories, he has worked on various aspects of digital processing and computing. He was a Visiting Professor in the Department of Electrical Engineering at Rutgers University. He is presently Director of the Computer Technology Research Laboratory.

Mr. Netravali holds over twenty patents and has had more than sixty papers published. He was the recipient of the Donald Fink Prize Award for the best review paper published in the Proceedings of the IEEE and the SMPTE journal award for the best paper published in 1982. Editorial board, Proceedings of the IEEE; Editor, IEEE Transactions on Communications; fellow, IEEE; member, Tau Beta Pi, Sigma Xi.

Jack Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; AT&T Bell Laboratories, 1961—. Mr. Salz first worked on the electronic switching system. Since 1968 he has supervised a group engaged in theoretical studies in data communications and is currently a member of the Communications Methods Research Department. During the academic year 1967–68, he was on leave as Professor of Electrical Engineering at the University of Florida. He was a visiting lecturer at Stanford University in Spring 1981 and a visiting MacKay Lecturer at the University of California, at Berkeley, in Spring 1983.

AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

Homenet: A Broadband Voice/Data/Video Network on CATV Systems

By M. HATAMIAN and E. G. BOWEN*

(Manuscript received April 3, 1984)

Homenet is a broadband distributed communication system that supports data, real-time digitized voice, and analog video on a single cable in a CATV type of network. The distance limitation problem encountered in local area networking schemes is eliminated by dividing the large CATV net into smaller "homenets." This feature makes the network suitable for a large number of users located in a relatively wide geographic scope. This paper describes the implementation of a small experimental version of this system in hardware. More attention is given to the protocol processing hardware, which implements a protocol based on collision detection called Movable Slot Time Division Multiplexing (MSTDM). The MSTDM protocol guarantees the continuity of voice signals received at the user station. Problems such as clock synchronization and confusion of data and voice packets are addressed, and solutions are given. Presently, an experimental network composed of five user nodes in two different frequency nets is operational. An interactive video retrieval service implemented in the network is described as an example of the type of user services (other than data/voice/one-way video) that can be offered at the main head end of the system.

I. INTRODUCTION

The concept of Local Area Networks (LANs) is a well-developed one in the field of computing and data communication.[1] These networks are used for sharing computing resources, and for communicat-

* Authors are employees of AT&T Bell Laboratories.

ing data among a number of users in a limited geographic scope. Adding voice and video capability to these networks makes them very attractive for the office information systems of the future. Furthermore, if the distance limitation and the constraint of limited geographic scope are removed, then such networks become potential candidates for the home information systems of the future, provided that the cost of user's terminal equipment is minimal. Such networks should more appropriately be called Metropolitan Area Networks (MANs) rather than LANs. Solving the distance limitation problem can also increase the attractiveness of the office information systems; office branches located at distant locations can become part of the network and can share information.

Homenet is a broadband data/voice/video communication system, first proposed in Ref. 1, which satisfies all the above requirements for MANs. The system combines frequency and time multiplexing, and supports the communication of data, real-time digitized voice, and analog TV signals on a single cable in a cable TV (CATV) type of network. This paper describes the hardware implementation of the homenet and some of its features. Presently, a fully working testbed composed of five user stations connected in two nets is in operation.

Sections II and III describe the homenet and the communication protocol. Section IV gives a detailed description of the hardware implementation of the system and its features. More attention is given to the protocol processing hardware, which is essentially the intelligent node of a distributed packet switching network.

II. WHAT IS A HOMENET?

Homenet is a broadband communication system based on a combination of frequency and time multiplexing, and distributed packet switching techniques. The system supports the communication of data, digitized voice, and one-way analog TV signals on a cable in a CATV type of network. Since it is basically a distributed switching network, all the switching functions are performed at the user's terminal equipment and there is no central switching involved. Reference 2 describes the system in detail.

A relatively large community of users is divided into small geographic regions and each region, called a homenet (or a net for short), is assigned a 6-MHz frequency band. Users within each frequency band can receive signals from any other net by tuning their receivers to the frequency of that net. This tuning is performed automatically by a signaling scheme. All users transmit their data and digitized voice signals on a single transmit frequency, F_0, which is then translated to its homenet frequency and propagated throughout the network in such a way that all the users in all nets are able to receive it. Before data are transmitted, users in each frequency band have to contend for the channel using the protocol described in the next section of this paper.

The overall operation of the homenet is shown in Fig. 1. Suppose that user No. N in net 3 wants to transmit a packet to user No. 1 in net 2. First, user No. N contends for the channel in net 3 and once access to the channel is gained, it transmits its data on frequency F_0 to the nominal head end of net 3 (H_3). At this point, frequency F_0 is translated to two different frequency bands, F_3 and Fr_3. Frequency F_3 is transmitted downstream to the users in net 3 and all the nets following net 3 (in this example there are no nets following net 3). Frequency Fr_3 is transmitted upstream to the main head end of the network (H_1) at net 1, where it is translated back to F_3 and sent to the users in net 1 and net 2. Now user No. 1 in net 2, with its tuner listening to frequency F_3, can receive the packet by demodulating the signal from F_3 to baseband and searching for its address in the address area of the packet. Other than the frequency translation operations, each nominal head end H_i is equipped with two notch filters—one for frequency band F_i, which stops the signal translated from Fr_i to F_i at the head end; and one for F_0, which allows all the nets to use the same transmit frequency F_0 without interfering with their adjacent nets.

The above scheme is certainly not the only possible way of using

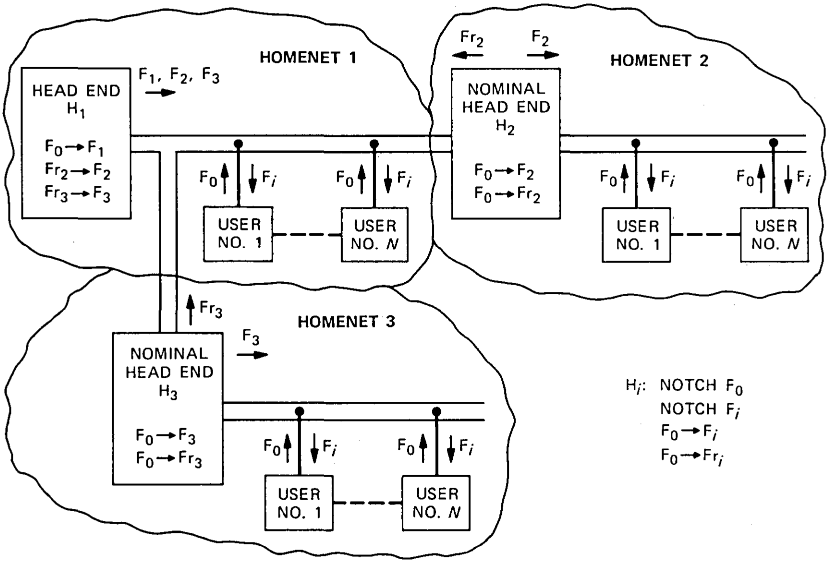


Fig. 1—Frequency assignment in homenet.

the frequency bands;[3] however, it is the one that requires the least amount of hardware for each user station and also is compatible with currently installed midsplit CATV networks.

Establishing a communication network like homenet for a large community of users of the size accommodated by homenet would not be possible by direct extension of LAN techniques (i.e., increasing the cable length and the bit rate). One of the major limitations of the local area networks is the cable length constraint, which forces the users to be located close to each other to prevent large propagation delays. *Ethernet,** one of the best-known local area networks,[4] is limited to a cable length of about 2.5 kilometers operating at 10 Mb/s; increasing the bit rate or the cable length results in an appreciable reduction in the efficiency of the system. In homenet this distance limitation problem is solved by using different frequency bands and grouping the users that are located close together into one frequency net.[5] This way, the users in each net have to contend for transmission rights only among themselves, and do not have to worry about transmitters in other frequency nets; they can still listen to all other frequency nets, so a complete connection between all users in all nets exists. This broadband technique increases the size of the network (i.e., the length of the cable) and at the same time reduces propagation delays, which are very important, especially for access strategies that use collision detection schemes such as CSMA/CD. Each net can operate at low bit rates but the total throughput of the network can go up to several hundred Mb/s, depending on the number of frequency bands used in the system.

III. COMMUNICATION PROTOCOL

The communication protocol used in each net of the homenet system is called Movable Slot Time Division Multiplexing (MSTDM)[6]—a variation of the CSMA/CD technique used in *Ethernet.* This protocol guarantees the continuity of the voice signals received at each user station, a task that no other currently available protocol based on collision detection can handle. This protocol is described in the following section.

Integration of packetized data and voice in a local area network requires an upper limit on the voice packet delays to ensure that the voice receiver does not run out of samples before new voice samples arrive. This requirement in turn guarantees a glitch-free, continuous speech signal at the output of the voice receiver. None of the currently available protocols satisfy the above requirement and hence are not

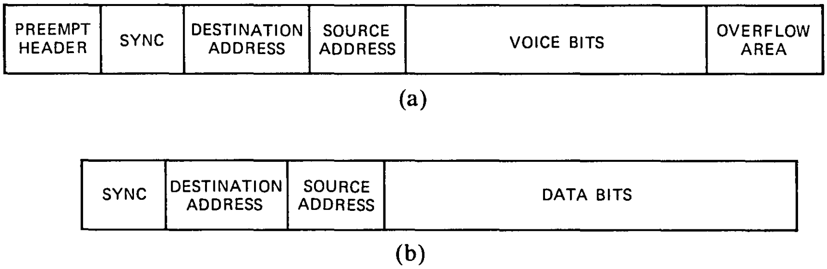
* Trademark of Xerox Corporation.

suitable for integration of data and digitized voice. MSTDM protocol places an upper bound on the voice packet delays; it guarantees the continuity of the reconstructed voice signal and also guarantees that once access to the channel is gained, no two voice sources can collide. MSTDM takes advantage of the periodicity of the voice packets; it also requires that the size of the data packets be smaller than the voice packets. A detailed treatment of this protocol can be found in Ref. 6. A description of its operation is given below.

In MSTDM a distinction is drawn between the first packet from a voice source and all the following voice packets from that source (called the secondary voice packets in this paper). The first voice packet and the data packets are treated the same way as in CSMA/CD. They check the channel busy signal to monitor the status of the channel, and once the channel is idle, they start transmitting. They listen to the channel while it is transmitting to make sure that no collision occurs, and if there is a collision, then the colliding sources stop their transmission and try to access the channel after a period of time defined by a retry strategy. This procedure stays the same for all data packets; however, for the voice sources the procedure is different. Once the first packet of voice successfully acquires the channel, then the following packets from that voice source get transmitted (when they are ready for transmission) as soon as the channel becomes idle; they do not listen to the channel for collision during transmission. If a collision occurs between these secondary voice packets and any other packet, the other transmitter is forced to stop its transmission and the secondary voice packets override.

Figures 2a and b show the voice and data packet formats, respectively. When there is a collision between a secondary voice packet and a data packet, the preempt portion of the voice packet, which does not contain any information, allows enough time for the data source to detect collision and stop its transmitter before the sync bits from the secondary voice packet appear on the channel.

After a voice source transmits a packet, it schedules its next trans-



(a)

(b)

Fig. 2—Packet formats for (a) voice packet, and (b) data packet.

mission for T seconds later. Assuming the rather unlikely situation where all the packets from a particular voice source find the channel idle when they are ready to be transmitted, then, for that particular voice source, the channel looks exactly like a TDM system with reserved time slots that are T seconds apart. However, it is quite likely that when the voice source is ready for its next transmission, a data source or another voice source is in the middle of transmitting its packet. In this case the voice transmitter has to wait for the channel to become idle. Obviously, in this situation the voice packet will be delayed and this delay causes the time slot for the voice source to move back in time. Therefore, for the voice sources the system looks like a TDM channel in which the time slots are not fixed in time and are free to move; hence the name movable slot TDM.

Reference 6 proves that if the voice packet delay is less than a packet transmission time (the upper bound on the voice packet delay in MSTDM to which we previously referred), then voice sources will never collide and no voice samples will be lost. To satisfy this requirement data packets are constrained to be shorter than voice packets.

The voice samples arriving during the voice packet delay time are stored in the overflow portion of the voice buffer (see Fig. 2a), and are transmitted along with the rest of the packet when the channel becomes available. The overflow area is always transmitted even if it does not contain any voice samples (i.e., the case when the voice packet is not delayed). This contributes to the proof of the fact that the voice sources never collide in MSTDM once they successfully transmit their first packet.[2] Obviously the overflow area of the voice buffer should be long enough to accommodate the voice samples that arrive during the voice packet delay, which is less than a packet transmission time. Since the transmission clock rate is much higher than the voice sampling rate, the size of the overflow area need not be larger than a few bits.

IV. HARDWARE IMPLEMENTATION

A small version of the homenet network described in previous sections has been built-in hardware for experimental purposes and feasibility studies. We currently have a fully working testbed composed of five user stations running in two frequency nets. Except for the protocol processing hardware, which is the most important part of a user station, all the components used in the system (such as frequency translators, taps, splitters, channel selectors, cable, etc.) are similar to those used by the CATV industry.

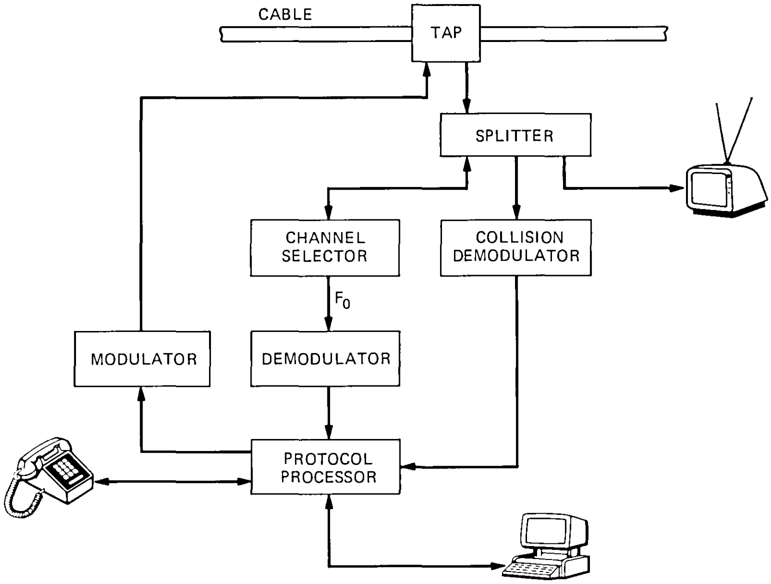


Fig. 3—Block diagram of user node hardware.

4.1 User node

Each user in homenet requires some hardware to interface his or her equipment to the communication cable. The user-node hardware can be thought of as a black box with one end connected to a cable and the other end to the user's voice source (normally a phone set), data source, and TV set. Figure 3 shows various components of the user-node hardware. The signal picked up from the cable through the tap is split and distributed in three ways. One line is connected directly to a TV set, another line feeds a collision demodulator whose function will be described later in this paper, and the third line is connected to a channel selector. A signaling scheme controls the channel selector and sets it to the frequency of the net that the user wants to hear. This frequency band is converted to a common Intermediate Frequency (IF) band F_0 and fed to a demodulator, which translates the information contained in band F_0 to a digital bit stream. This bit stream is then processed by the protocol processor and, if the information is destined for the user's address, it will be appropriately sent to either the data or the voice section. For the purpose of transmission, once the channel is accessed, the packets prepared by the protocol processor are modulated to the frequency band F_0 using a modulator and sent to the user's nominal head end (through the tap and over the cable) to be distributed throughout the network, as we described in Section II.

Except for the protocol processor, which is special-purpose hardware, the rest of the user node components are commercially available items. The modulators and demodulators are tuned to a center frequency of 43.4 MHz (F_0 in current system).

4.2 Protocol processor

The protocol processor is essentially the intelligent part of the user node hardware. It is responsible for digitization and packetization of voice signals, packetization of data, and most importantly, implementation of the MSTDM protocol.

The processor is divided into two main sections, transmitter and receiver. Each section has two separate circuits, one for voice and another for data. Following is a detailed description of the operation of these circuits.

4.2.1 Transmitter

The voice and data sections of the protocol processor's transmitter operate independently of each other, with their own dedicated buffers. In terms of the ordering of the sync and address fields, the packet formats are fixed and are as shown in Fig. 2. However, the position of the sync word, the number of sync words, the length of the voice preempt header, and the length of the packet can be arbitrarily set by the user to conform to a net standard. The packet length can be set to any number of bits fewer than 4096.

4.2.1.1 Voice. Before we describe the operation of the voice transmitter, we should discuss the structure of the voice buffer. As we mentioned before, the first voice packet is treated the same as data packets in terms of accessing the channel. The moment that the voice signal is activated, the voice transmitter starts filling the voice buffers and at the same time makes a request for transmission. When transmission right is granted, the number of collected voice samples may not be enough to fill the whole buffer, and as a result a number of empty locations (noise bits) will remain at the end of the buffer. Now, if the buffer is transmitted from beginning to end, then the empty area will cause a quiet interval (or a glitch) in the voice signal between the first and the second voice packets. In applications where a Time Assignment Speech Interpolation (TASI) mode of operation is desired, this effect can cause serious distortion in the reconstructed speech signal. However, if the empty portion of the first voice packet is transmitted before the actual voice bits, then the quiet interval will not be in the middle of the voice signal and will not create any difficulty. This procedure is illustrated in Fig. 4. The buffer is shown in this figure as

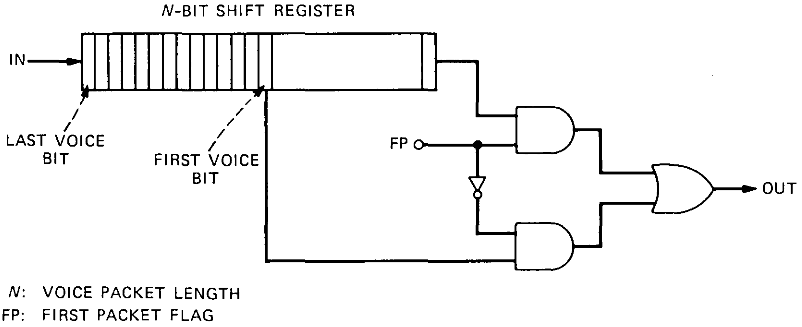


Fig. 4—Voice buffer operation.

a long First-In First-Out (FIFO) shift register; a First Packet flag signal (FP in Fig. 4) indicates whether the bits would be read out from the end of the FIFO register or from its head position (i.e., the first voice bit). In Very Large-Scale Integration (VLSI) design, the implementation of such a buffer is rather simple considering the regularity of its structure. In Transistor-Transistor Logic (TTL) design, we used RAMs as buffers, and counters and latch registers to keep track of the first voice bit position, last voice bit position, and the length of the unused portion of the buffer.

Figure 5 is a block diagram of the voice transmitter. To be able to handle the voice signal in real time, two buffers operating in ping-pong mode are needed; one buffer is being filled with voice bits while the other is being transmitted. The combination of RAM and counter blocks in Fig. 5 represent a FIFO register which, in conjunction with the first packet handler circuit, operates in a manner described above. The input voice signal from the voice source is digitized and converted to a serial bit stream by a 64-kb/s μ-law codec chip (8-kHz sampling rate, 8 b/sample). A multiplexer at the input of the ping-pong buffer controls the distribution of the input bits, clock signals (voice clock and transmission clock), and memory write pulses (R/W) to the buffers. The buffer that is supposed to be transmitted receives the transmission clock and no write pulses; the buffer that is being filled with the voice bits receives the voice clock (64 kHz), the voice bits, and the memory write pulses.

Before trying to establish a voice connection, the Timing and Control Circuit (TCC) first clears the buffers and then writes the header information (i.e., sync, destination and source addresses) into both ping and pong buffers through the input multiplexer. At the time the first packet of voice begins to be formed in the ping buffer, a request for transmission is made by TCC to the Transmit Request Circuit (XRC). Once XRC receives the request, it starts monitoring

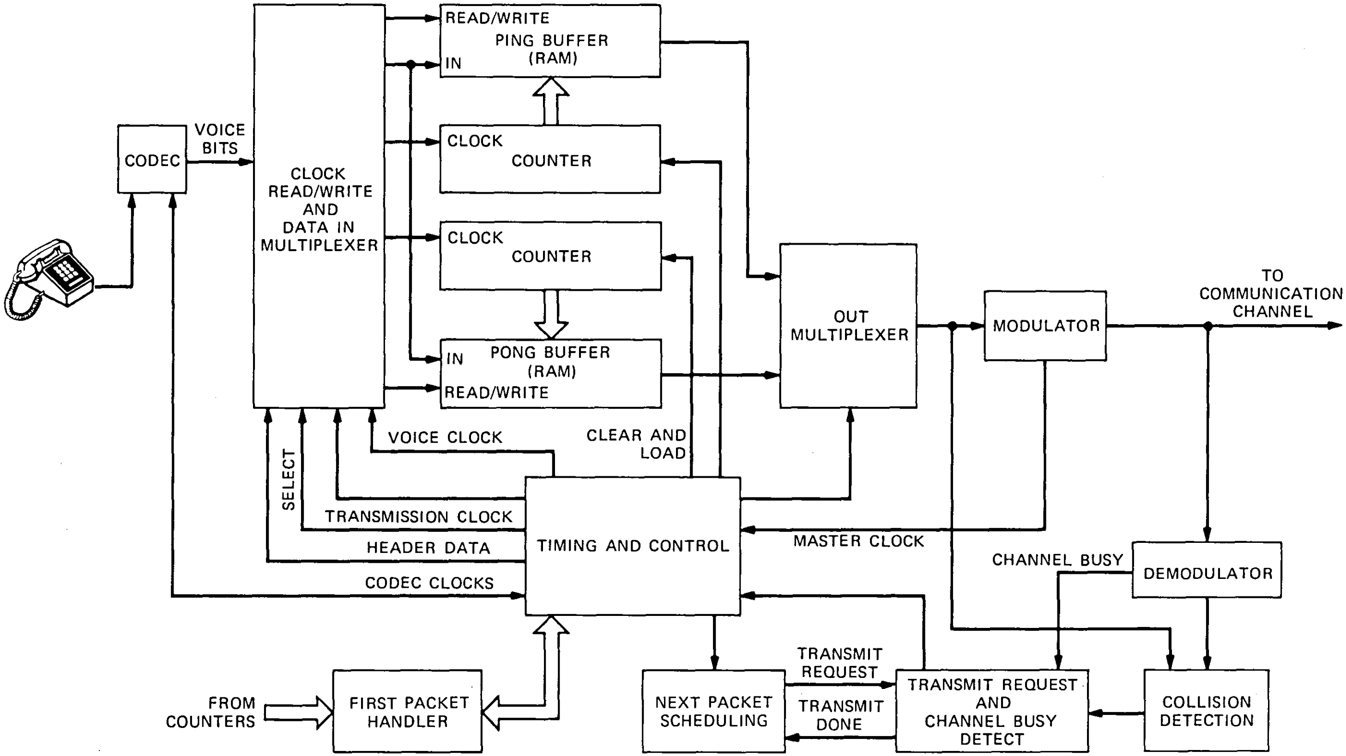


Fig. 5—Block diagram of the voice transmitter.

the channel busy signal provided by the demodulator, and the moment the channel becomes idle XRC sends an acknowledgment to TCC. Upon receiving this acknowledgment, TCC switches the ping-pong mode (by changing the multiplexer control signals) and sets a transmission flip-flop. Now the pong buffer starts receiving the voice bits and the ping buffer, which contains the first packet of voice, begins to be transmitted on the cable under the control of the first packet handler circuit, through the output multiplexer and the Frequency Shift Keying (FSK) modulator. If no collision is reported by the Collision Detection Circuit (CDC), then the transmission continues until the last bit of the voice packet is sent, at which time TCC clears the transmit flip-flop and sends a signal to the Next Packet Scheduling (NPS) circuit. NPS schedules a transmission request for T seconds later. T is switch selectable and is set by the user, depending on the length of the voice packet. During this T seconds, voice samples are being stored in the pong buffer, and the ping buffer, which has already transmitted its data, is waiting to be switched. After T seconds, NPS sends a transmission request to XRC and when transmission right is granted, TCC again sets the transmit flip-flop and switches the ping-pong mode. Now, the input voice samples are routed to the ping buffer and the secondary voice packet contained in the pong buffer begins its transmission. The first packet handler and the collision detection circuits are disabled at this time, because the first packet has been successfully transmitted and there is no need to check for collisions (see the description of MSTDM protocol in Section III). When the second packet is transmitted, NPS receives a transmit done signal and again schedules a request for T seconds later, and the procedure cycles until the user decides to end the voice connection.

If, during the transmission of the first packet, a collision is detected by CDC, then the transmit flip-flop is cleared, the buffer pointer is reset to the beginning, and TCC tries to make a request again. At present, the only retry strategy built into the protocol processor is on the basis of a random retry. The operation of the collision detection circuit is described in a separate section following the discussion of the data transmitter.

4.2.1.2 Data. Figure 6 is a block diagram of the data transmitter. This circuit is considerably less complex than the voice transmitter. It does not require the ping-pong buffering mode and there is no difference between the way the first packet and the following ones are handled; this eliminates the need for the first packet handler circuit.

Before establishing a data connection, the TCC writes the header information into the buffer, and then it sets the transmitter ready flip-flop. Now the transmitter buffer is ready to accept data bits. The data are written in the buffer and once the packet is formed and the

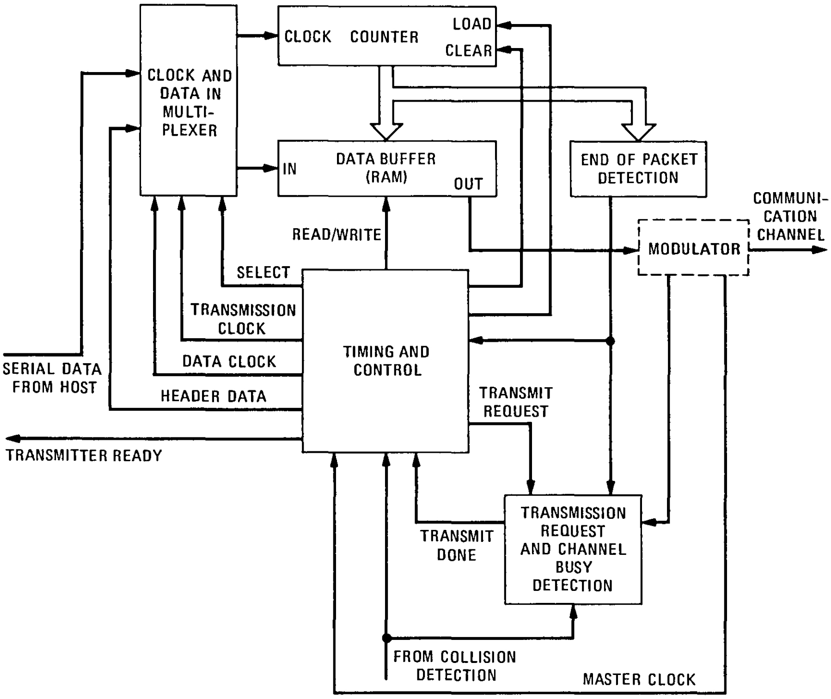


Fig. 6—Block diagram of the data transmitter.

buffer is full, the End of Packet Detection (EPD) circuit sends a signal to TCC, which resets the transmitter ready flip-flop. The transmitter is then ready to access the channel and data sources should not try to send any new data for packetization. At the same time, TCC sends a transmission request signal to the XRC. When the channel becomes idle, XRC sets a transmit flip-flop and the transmission of the packet begins. Once the last bit of the packet is transmitted, XRC generates a transmit done signal, which sets the transmitter ready flip-flop; the buffer then becomes available for packetizing data bits, and the cycle starts again.

In case of a collision with another packet, the same procedure used for the first packet of voice is used. The collision detection circuit is shared between the voice and data transmitter.

4.2.1.3 Collision detection. In homenet, collisions can be detected in the digital domain by comparing every single bit of a packet before and after its transmission. Relying on just the amplitude of the signal on the line at each transmitter for detecting interference from other transmitters can result in missing collisions that are caused by transmitters that are distant from each other. In homenet, however, all the transmitters in each net, say net H_i, send their signal, on frequency F_0, to the nominal head end. There it is translated to frequency band F_i and returned to all sites. By inserting proper attenuators on the transmitters, we arrange to have the same amplitude for all the signals received at the nominal head end from different transmitters. As a result, the signals that are returned in frequency F_i at each site have the same relative amplitude, and there is no chance of missing collisions between distant transmitters.

The collision detection circuit has a dedicated demodulator that is always listening to the frequency of the net that the transmitter is in (i.e., F_i for net H_i). This demodulator is referred to as the collision demodulator in the block diagram of Fig. 3. Notice that detecting collision in frequency F_0 rather than F_i does not have any advantage over baseband, and would not solve the amplitude problem.

It is obvious that the process of modulating the bits to frequency F_0, transmitting to the nominal head end, translating to frequency band F_i, and transmitting them back will introduce some location-dependent delay between the bits that leave the transmitter and the ones that are received by the collision demodulator. The collision detection circuit corrects for this delay by inserting an equal delay on the transmitted bits before they are compared. This is accomplished by an adjustable delay line on one input of the CDC, which is adjusted only once depending on the user's distance from the nominal head end.

4.2.2 Receiver

Much like the transmitter, the receiver is also divided into two almost independent sections, voice and data. In the transmitter, the collision detection circuit was shared between the voice and data sections. In the receiver, there is no need for collision detection; however, there is one circuit that is shared between the two sections, and that is the Sync and Address Detection (SAD) circuit.

4.2.2.1 Voice. Figure 7 shows the block diagram of the voice receiver. The operation of the ping and pong buffers in the receiver is similar to the voice transmitter. When one is receiving the input bit stream at the channel rate, the other is playing the previously received voice packet into a codec at the voice bit rate (64 kb/s). The switching of the ping-pong mode is done by the TCC. The input bit stream is first demodulated from band F_i to baseband and directly fed to the input of the buffers, the SAD, and the end of packet detection circuit. Right after the transition of the channel busy signal from an idle to a busy state, SAD starts looking for a sync word. The sync words for data and voice differ in their most significant bit. If the detected sync is a voice sync, then the receiver starts looking for either another sync

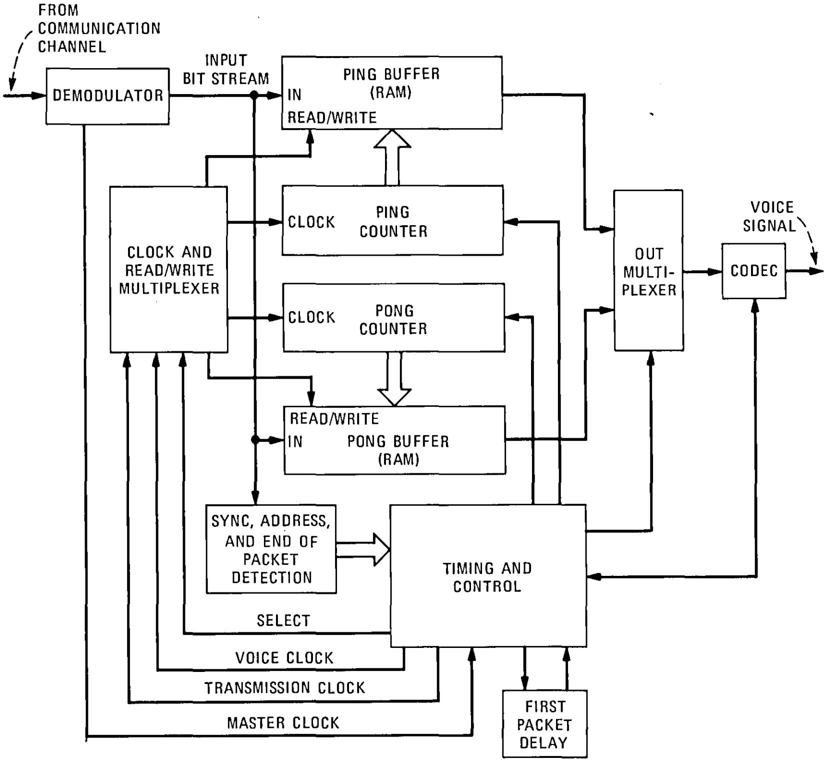


Fig. 7—Block diagram of the voice receiver.

word or its address. The sync words can be repeated consecutively in the header area of the packet as many times as desired. The destination address should always be immediately after a sync word.

If the sync word indicates a voice packet and the destination address field of the packet is matched with the receiver's address, then SAD first stores the following field (i.e., the source field) as the address of the source, and sends a signal to TCC indicating the beginning of the voice bits. TCC then sets a receive flip-flop and applies the transmission clock to the ping buffer. The voice bits begin to be stored in this buffer until an end of packet is detected by SAD, at which time TCC resets the receive flip-flop and switches the ping-pong mode. The end of packet detection is not done based on just the packet length. The voice transmitter always transmits a fixed number of bits for each packet, as required by MSTDM protocol. However, not all of these bits are actual voice bits; the bits in the overflow area may not be useful information. Therefore the end of the voice samples in the packet must be marked. This is done at the transmitter by placing a flag byte at the end of the voice samples with all eight bits set to "1." When the transmitter's ping-pong mode is switched, the value of the counter for the buffer that was receiving the voice bits is saved; when the transmission of this buffer begins, the control circuit monitors the buffer counter and when it is equal to the saved value, TCC forces the output bits to a high state until the buffer counters reach the packet length. Therefore the unused portion of the voice packet is transmitted as "all ones." At the receiver, SAD circuit searches for an "all one" flag byte and sends an end of packet signal to TCC, as mentioned above. The codec used for digitizing the voice signal does not use the "all one" level in its code. Therefore, no voice samples will be coded as all ones to cause a false end of packet detection at the receiver.

Once the end of packet is detected and the ping-pong mode is switched by TCC, the voice samples stored in, say, the ping buffer are played back to the codec to reconstruct the voice signal. The pong buffer is idle at this time and is waiting to receive the next voice packet, which will arrive sometime during the playback process.

In MSTDM protocol, the voice packets can exercise a bounded delay less than one packet transmission time. This delay may cause the receiver buffer to run out of voice samples before the next packet arrives, and may also cause a distortion in the speech signal. To alleviate this problem, a small delay is inserted before the beginning of the playback process for the first packet of voice (only the first packet). This task is accomplished by the First Packet Delay (FPD) circuit at the receiver (see Fig. 7). After the first packet is played back, TCC deactivates this circuit. The FPD circuit is also used for recovering from a distortion problem created by timing discrepancies between user nodes. This will be described in detail in a later section on clock synchronization.

4.2.2.2 Data. Figure 8 shows a block diagram of the data receiver. This circuit is the simplest section of the protocol processor. The sync and address detection circuit is shared between the voice and data sections. When a data sync pattern followed by the receiver's address is detected in the input bit stream, TCC first resets a data ready flip-flop, indicating that the receiver buffer is being filled with the incoming packet bits. The receive clock is then applied to the buffer counter and data bits are stored in the buffer until an end of packet signal is generated by the EPD circuit. The end of packet is simply detected by comparing the value of the buffer counter with the length of the data packet. This signal sets the data ready flip-flop, indicating that the receiver buffer contains valid data bits and is ready to transfer those to the host system. At this time, if the host ready signal is high, TCC applies the data clock to the buffer counter and the data bits are transferred to the host system. Once this transfer is made, the circuit

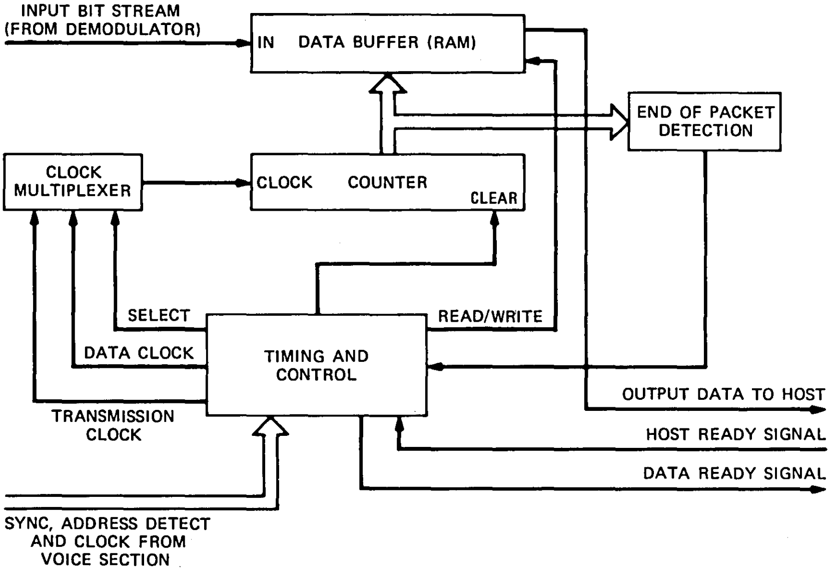


Fig. 8—Block diagram of the data receiver.

is reset and TCC waits for the next data sync detect signal, at which time the cycle starts again.

The performance of both the data receiver and transmitter can be somewhat improved using a ping-pong buffering scheme as in the voice section. However, since this is not a necessity, we did not choose to implement it in our experimental system. This option can always be easily incorporated into the system if needed.

4.2.3 Clock synchronization

When a communication link, either data or voice, is established between two sites, it is obvious that for proper operation the receiver's clock pulses should be synchronized with the incoming data (i.e., synchronized with the transmitter's clock). In homenet, each user station (user-node hardware) has its own crystal oscillator generating a 16-MHz master clock signal. All the clock pulses used by the protocol processor are derived from this master clock, and are synchronized with the incoming data using the transitions of the input bit stream. The synchronization circuit, which is part of the demodulator board, uses a very simple digital technique similar to the clock recovery circuits used with nonreturn to zero data streams.

For the purpose of synchronization, the receiver requires at least two or three transitions in the incoming bit stream before any useful data can be picked up from the line. For the voice packets the required transitions can be placed in the preempt portion of the packet; for the data packets a 4-bit preempt header is added to the beginning of the packet format shown in Fig. 2b.

Due to different operating conditions the frequency of the crystal oscillators at two ends of a communication link can be slightly different, thus creating a minor timing discrepancy. This small frequency difference does not create any difficulty in receiving the information bits because the bit values are read into the receiver buffer in the middle of clock pulses, and a small drift can be tolerated. However, due to the periodic nature of the voice sources and their real-time requirement, the timing discrepancy affects the voice section of the protocol processor in two ways, as described below.

First, we consider the effect of timing discrepancy on the next packet scheduling time. The situation is illustrated in the timing diagram shown in Fig. 9 for two arbitrary voice sources that have successfully transmitted their first packet and reserved a movable time slot on the channel. Voice source No. 1 schedules its next transmission for T seconds after it transmits the current packet (i.e., T seconds after the falling edge of TRANSMIT signal in Fig. 9). After T seconds, the NPS circuit generates a TRANSMIT REQUEST pulse and the voice source is guaranteed to have access to the channel within δ seconds, where δ is between zero and a maximum of one packet transmission time. Voice source No. 2 operates in exactly the same way except that, owing to the slight timing discrepancies between the two sources, the next packet scheduling time for this source will be $T + \epsilon$ rather than T. This can cause the TRANSMIT REQUEST pulse and the time slot for voice source No. 2 to drift very slowly in time with respect to voice source No. 1 (see Fig. 9). The drifting continues until the time slots for both sources are adjacent to each

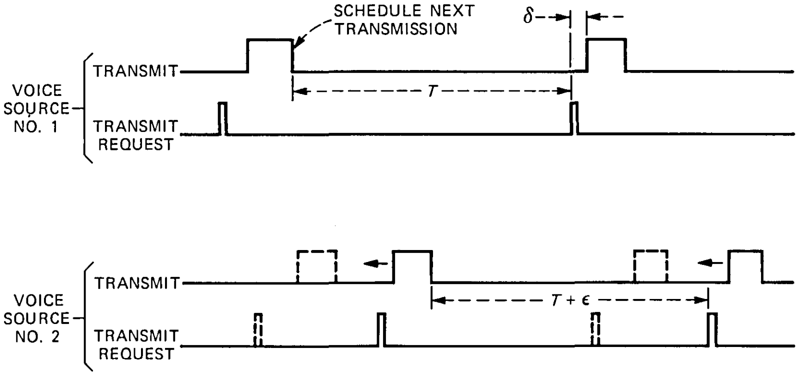


Fig. 9—Effect of timing discrepancy on next packet scheduling.

other. If there are a number of voice sources on the line, all the time slots gradually move until they are adjacent to each other. Notice that this gradual movement of the slots is different from the movement dictated by MSTDM protocol, which can vary depending on the data traffic. If the data traffic is light and the voice sources continue to stay on the channel for a long time, then, from the above discussion, all the time slots will eventually be packed next to each other. The timing discrepancy in this case does not introduce any difficulty; as a matter of fact, it creates a favorable situation.

With respect to the rate at which the voice samples are generated at the source and used at the destination, the timing discrepancy can create an undesirable situation. If the receiver clock is slower than the transmitter, then there will be a time when both ping and pong buffers at the receiver will be full when a new packet arrives. If the receiver clock is faster than the transmitter, then the receiver will eventually run of out of samples before a new packet arrives. These situations cause a distortion in the voice signal, which cannot be recovered from for some time. In the receiver hardware, both of the above situations cause the ping-pong switching command to be generated at the same time with transmission. To solve the problem, when such a condition is detected, TCC activates the first packet delay circuit (see Fig. 7), which inserts a delay in the playback of only the next packet and causes the ping-pong switching command not to overlap with the transmission. The consequence is that the current voice packet is lost, but the receiver goes back into a normal undistorted operation mode. In our system, which uses a conventional oscillator and crystal type, this situation occurs about every 20 minutes. In other words, if a conversation lasts for a long time, then every 20 minutes one packet of voice (i.e., a few milliseconds of voice signal) will be lost. There are more expensive solutions that totally eliminate the problem; however, losing a few milliseconds out of 20 minutes (12×10^5 ms) is of no significance at all, and more expensive solutions are not justified. Besides, using better crystal types can increase the 20-minute period to well over an hour.

4.2.4 Packet confusion

In the case of collision between a secondary voice packet and a data packet, a confusion between packets can occur at the receiver. Consider the following situation: a data source is transmitting when it detects a collision with a voice packet after the sync word has been transmitted (i.e., in the address area of the packet). The data transmitter stops, the voice transmitter continues with its transmission, and the line will look something like the following:

DATA SYNC	GARBAGE	VOICE SYNC	VOICE DESTINATION ADDRESS	VOICE SOURCE ADDRESS	VOICE BITS

In the above display the collision occurred in the garbage area, which includes the voice preempt header as well.

Let us consider what happens at the receivers. All the receivers detect the data sync and start looking for the destination address. One particular receiver whose address is equal to the first byte of the garbage area is going to detect a match, and it will erroneously start storing the following bits as a data packet. Obviously this situation is very undesirable. We can solve the problem by looking at the length of the received packet. A false data packet like the above will certainly be longer than a normal data packet. If, at the time that the end of packet pulse is detected (based on the packet length as described before) the receiver is still busy with the incoming bit stream, then the received data packet must be a false one and should therefore be discarded.

The receiver that is supposed to receive the voice packet first synchronizes on the data sync; however, since it does not find its address following the data sync, it resets itself and starts searching for a combination of sync and destination address again. It should be noted that this problem occurs during the preempt header so that the data transmitter stops and the voice sync and address are undisturbed. Therefore, we do not have to worry about the voice packet; it will get to where it is supposed to. Of course, in the highly unlikely situation of the voice destination address being exactly the same as what is found in the garbage area, the above assumption will not be true and the voice packet will be lost.

4.3 Head end

Aside from the frequency translation and filtering operations, a variety of user services can be incorporated in the homenet's main head end (head end H_1 in Fig. 1). Our experimental system presently supports two services, an interactive video disc and a service for establishing voice links with sources outside the network. Figure 10 shows the block diagram of the main head end. A user station identical to the ones used at the user nodes is dedicated to this head end.

A head-end processor, which, depending on the application and the network requirements, can be anything from a small microprocessor to a large computer system, controls all the services. A user can control a video disc located at the head end by sending commands to the head-end processor over the cable using its data transmitter. These com-

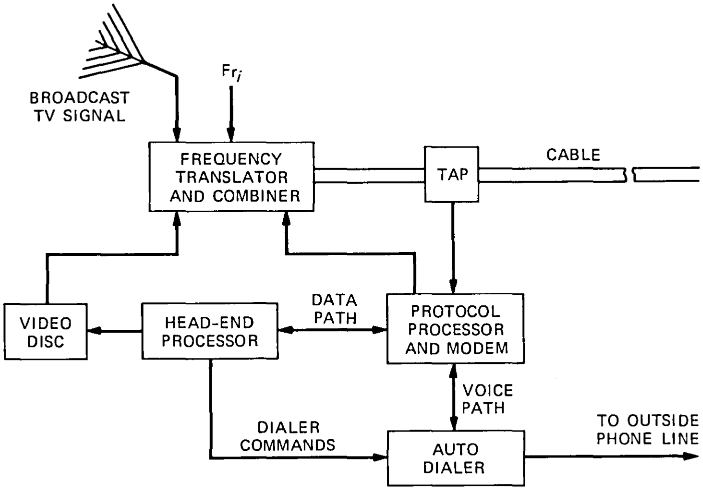


Fig. 10—Block diagram of homenet's main head end.

mands are interpreted by the head-end processor and the proper signals are sent to the video disc through a serial port.

To establish a voice link with sources outside the network, a user sends a data packet to the head-end processor giving the number to be called; the head-end processor then sends the proper commands to an autodialing system, which connects the voice path of the protocol processor to the outside line. Work on this autodialing option, as well as on other services such as call processing, file transfer, and higher-level communication protocols, is currently in progress.

V. CONCLUSION

This paper described the hardware implementation of a network for simultaneous communication of data, digitized voice, and analog video in a CATV system. The network uses a broadband approach to solve the distance limitation and delay problems suffered in local area networks. As a result, a considerably large community of users can be supported by the network. The problem of guaranteeing a continuous nondistorted speech signal at the receiver is solved using a variation of CSMA/CD protocol called movable slot time division multiplexing. This protocol places an upper bound on the maximum delay that can be experienced by a voice packet.

At the present time, an experimental network of five user nodes in two frequency bands is fully operational. The protocol processing hardware was described in detail. This processor is built with standard TTL components. It was shown that, owing to the method used by the protocol processor for scheduling packet transmission times and compensating for timing discrepancy, clock synchronization is not a major problem.

REFERENCES

1. W. R. Franta and I. Chlamtac, *Local Networks*, Lexington, MA: Lexington Books, 1981.
2. N. F. Maxemchuk and A. N. Netravali, "A Multifrequency Multiaccess System for Local Access," *ICC 83*, Boston, MA.
3. A. N. Netravali and N. F. Maxemchuk, private communication.
4. R. M. Metcalfe and D. R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Commun. ACM, *19*, No. 7 (July 1976), pp. 395–404.
5. A. N. Netravali and Z. L. Budrikis, unpublished work.
6. N. F. Maxemchuk, "A Variation on CSMA/CD That Yields Movable TDM Slots in Integrated Voice/Data Local Networks," B.S.T.J., *61*, No. 7 (September 1983), pp. 1527–50.

AUTHORS

Mehdi Hatamian, B.S.E.E., 1977, Ary-Mehr University of Technology, Tehran; M.S. and Ph.D. (Electrical Engineering), University of Michigan, Ann Arbor, in 1978 and 1982, respectively; AT&T Bell Laboratories, 1982—. From 1979 to 1982 Mr. Hatamian was a Research Assistant on a NASA project working on the design and development of hardware and software for one of the Space Shuttle experiments. His research interests are in real-time signal processing, circuit design, image processing, and VLSI design. Member, IEEE, Sigma Xi.

Edward G. Bowen, A.S. (Electrical Engineering), 1963, Vermont Technical College; B.S. (Electrical Engineering), 1971, Newark College of Engineering; AT&T Bell Laboratories, 1963—. At AT&T Bell Laboratories, Mr. Bowen is an Associate Member of the Technical Staff. His principal interests are coding of video signals.

AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

Analysis of a Multistage Queue

By B. T. DOSHI and K. M. REGE*

(Manuscript received June 12, 1984)

Multistage queueing mechanisms with quantum service are suitable in various computer and communication systems to guarantee small delays to short jobs without first knowing the service requirement of any job. In this paper we analyze the efficacy of one such scheme—a two-stage First-In First-Out (FIFO) and Round Robin (RR)—in discriminating between short and long jobs. We obtain the distribution of the delay for short jobs, the cycle time in the RR queue for long jobs, and the number of messages in the FIFO and the RR queues. For the specific parameters used in our numerical results, the two-queue scheme seems to discriminate effectively between the long and short jobs.

I. INTRODUCTION

In computer systems as well as data communication systems, it is frequently desirable to guarantee that short jobs see small delay even under a high load. This may be done at the expense of long jobs. It is also true that in many of these systems the time required to do a job is not known beforehand. Thus simple priority schemes based on the service requirements of jobs are not possible. If the jobs are served in order of arrival, First-In First-Out (FIFO), then all jobs will see long delays at high load. To discriminate between short and long jobs without knowing the type of a job beforehand, various schemes based on quantum service are used. The simplest of these is a Round Robin (RR) scheme. Here, when a job arrives, it is put behind all the waiting

* Authors are employees of AT&T Bell Laboratories.

jobs. When it reaches the server it gets at most Δ time units of service. If its service requirement is smaller than Δ, then the job leaves before the quantum Δ expires. Otherwise, after getting Δ units of service, it is put at the back of the queue and waits for the next pass at the server. Since a shorter job requires fewer passes, its delay is smaller. For this scheme, Wolff[1] obtained the mean delay conditioned on the service requirement as the solution of an infinite system of linear equations. Other schemes are possible if more discrimination is desired between short and long jobs. At one extreme we have a scheme based on infinite number of queues (IQ). In this scheme the server keeps an infinite number of queues numbered 1, 2, ... On arrival a job is placed at the back of the queue numbered "1." When the server completes a service, it takes the first job from the lowest numbered nonempty queue. If this job is from queue n, then it gets at most Δ_n time units of service. If its service is not complete by then, it is put at the back of the queue numbered $n + 1$. Schrage analyzed this scheme and derived the mean and Laplace-Stieltjes transform of the delay conditioned on the service requirement.[2] Somewhere between RR and IQ schemes are the schemes based on a finite number $(N + 1)$ of queues. The first N queues behave as they do in the IQ scheme, while the last one can be either FIFO or RR. In this paper we study one such scheme. In particular, we consider the case where $N = 1$ and the second queue is served round robin. We call this queueing system a FIFO-RR system. The analysis for general N is almost identical but the resulting expressions and notation are more complex.

Fraser and Morgan[3] have analyzed this FIFO-RR discipline as the model of the trunk service discipline in *Datakit*™ Virtual Circuit Switch (VCS) (see Ref. 3 for details of the trunk module operation in *Datakit* VCS). They obtain the mean delay for various classes of jobs under fairly general assumptions, essentially by extending the results in Wolff[1] to the FIFO-RR system. They also use simulation to obtain the percentiles of the delay distributions. In this paper we focus on analytical methods to obtain information about the delay distributions. In particular, we derive a simple expression for the transform of the delay distribution for short jobs under the assumption of Poisson arrivals and general service time distribution. This transform is inverted numerically to obtain the delay distribution. This enables us to get the delay distribution for one-character typed messages and short control messages in communication applications. For jobs long enough to require service in both FIFO and RR queues, the analysis is more difficult. Under more restrictive assumptions, we get the marginal generating function of the number of jobs in the FIFO and RR queues, the transform of the cycle time in the RR queue, and the mean sojourn time in essentially closed form. We illustrate our analysis with numerical results from a data communication application such as the trunk service in *Datakit* VCS. In particular, we show that extremely short jobs see very short delay even under very high overall load. We also discuss how our model may differ from the actual service discipline in data communication applications and the performance implications of these differences.

The analysis presented here for the RR queue uses busy cycle analysis to derive quantities of interest. Recently, Ramaswami showed that some of these quantities can also be derived using matrix methods.[4]

This paper is organized as follows: In Section II we define the model formally and introduce the notation. The delay in the FIFO queue is analyzed in Section III. In Section IV we derive the performance measures for the RR queue. Finally, in Section V we illustrate our results with an application from communication over a 56-kb/s link.

II. MODEL

In this section we formally define the model of the FIFO-RR queues, which we will analyze in Sections III and IV. The analysis of Section III is for the FIFO queue and thus will give the delay distribution for the jobs with service time less than or equal to the quantum size in the FIFO queue. This will be done under fairly weak assumptions. In Section IV we will analyze the RR queue under more restrictive assumptions.

Assume that the arrival process of the jobs is Poisson at rate λ. Let H be the distribution function of the service time. Let Δ_1 and Δ_2 denote, respectively, the quanta of service in the FIFO and the RR queue.

In Section III we will let H be general. In Section IV we will assume that there are two types of jobs. A fraction p of the jobs are short enough to be completed within one quantum in the FIFO queue. Thus, if H_1 is the distribution function of the service time of the short jobs, then

$$H_1(\Delta_1) = 1. \tag{1}$$

The other fraction, $(1 - p)$, of jobs may be long and has distribution function

$$H_2(x) = 1 - e^{-\mu x} \qquad 0 \le x < \infty \tag{2}$$

for some $\mu > 0$. Thus

$$H(x) = pH_1(x) + (1 - p)(1 - e^{-\mu x}), \qquad 0 \le x < \infty. \tag{3}$$

Let h_{i1} and h_{i2} denote the first two moments of H_i, $i = 1, 2$. Of course, $h_{21} = 1/\mu$ and $h_{22} = 2/\mu^2$.

When a job arrives it is put at the back of the FIFO queue. When its turn arrives it gets up to Δ_1 units of service. If the complete service is not rendered by then, the job moves to the RR queue. The RR queue is served in a round robin way with quantum size Δ_2. The FIFO queue has priority over the RR queue to the extent that after each quantum of service, the next service is from the FIFO queue as long as there is work in the FIFO queue.

III. ANALYSIS OF THE FIFO QUEUE

Let

$$q_1 = H(\Delta_1), \tag{4}$$

and for $i \geq 1$,

$$r_i = \frac{H(\Delta_1 + i\Delta_2) - H[\Delta_1 + (i-1)\Delta_2]}{1 - q_1}. \tag{5}$$

Let

$$\bar{N}_2 = \sum_{i=1}^{\infty} i r_i. \tag{6}$$

Thus $\bar{N}_2$ is the expected number of passes at the server in the RR queue given that a job enters the RR queue. Let

$$Q_1(t) = H(t) \qquad 0 \leq t < \Delta_1, \tag{7}$$

and

$$Q_2(t) = \sum_{i=1}^{\infty} \frac{\{H[\Delta_1 + (i-1)\Delta_2 + t] - H[\Delta_1 + (i-1)\Delta_2]\}}{1 - q_1}, \qquad 0 \leq t < \Delta_2. \tag{8}$$

Then the rate of service completions in the FIFO queue is $\lambda_1 = \lambda$, and the distribution of the service time, X_1, in the FIFO queue is given by

$$P\{X_1 \leq t\} = F_1(t) = Q_1(t) = H(t), \qquad 0 \leq t < \Delta_1, \tag{9}$$

and

$$P\{X_1 = \Delta_1\} = F_1(\Delta_1) - F_1(\Delta_1^-) = 1 - H(\Delta_1^-). \tag{10}$$

The rate of service completions in the RR queue is

$$\lambda_2 = \lambda \bar{N}_2 (1 - q_1) \tag{11}$$

and the distribution of the amount of service, X_2, in a typical service in the RR queue is

$$P\{X_2 \leq t\} = F_2(t) = Q_2(t)/\bar{N}_2, \qquad 0 \leq t < \Delta_2, \tag{12}$$

$$P\{X_2 = \Delta_2\} = F_2(\Delta_2) - F_2(\Delta_2^-) = \frac{\bar{N}_2 - 1}{\bar{N}_2} + \frac{1 - Q_2(\Delta_2^-)}{\bar{N}_2}. \tag{13}$$

Now consider a nonpreemptive priority queueing system with two FIFO queues and one server. The arrival rate and the service time distribution in queue i and λ_i are F_i, respectively, $i = 1, 2$. It can be shown using level-crossing arguments (see Refs. 5 and 6) that the distribution of the waiting time in the high-priority queue does not depend on the actual dynamics of arrivals in the low-priority queue. Thus, the waiting time distribution for an arbitrary arrival in queue 1 for this system is the same as that for an arbitrary arrival in the FIFO queue in the original FIFO-RR system. Thus, let $\tilde{f}_1$ and $\tilde{f}_2$ be the Laplace-Stieltjes transforms of F_1 and F_2, respectively, and let $\tilde{W}_1$ be the Laplace-Stieltjes transform of the waiting time in the FIFO queue. Let

$$\zeta_1 = \lambda_1 \int_0^{\Delta_1^+} t dF_1(t), \tag{14}$$

$$\zeta_2 = \lambda_2 \int_0^{\Delta_2^+} t dF_2(t). \tag{15}$$

Then, from Ref. 7,

$$\tilde{W}_1(s) = \begin{cases} \dfrac{s(1 - \zeta_1 - \zeta_2) + \lambda_2[1 - \tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1 \tilde{f}_1(s)} & \text{if } \zeta_1 + \zeta_2 < 1 \\[2ex] \dfrac{1 - \zeta_1}{s - \lambda_1 + \lambda_1 \tilde{f}_1(s)} \cdot \dfrac{\lambda_2[1 - \tilde{f}_2(s)]}{\zeta_2} & \text{if } \zeta_1 + \zeta_2 \geq 1, \quad \zeta_1 < 1. \end{cases} \tag{16}$$

Equation (16) can be inverted using a method of Jagerman[8] to obtain the waiting time distributions numerically.

Let us now consider the total sojourn time (waiting time + service time) for a job in the FIFO queue. Its Laplace-Stieltjes transform is given by

$$\begin{aligned} \tilde{D}_1(s) &= \tilde{W}_1(s) \tilde{f}_1(s) \\ &= \frac{s(1 - \zeta_1 - \zeta_2) + \lambda_2[1 - \tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1 \tilde{f}_1(s)} \cdot \tilde{f}_1(s). \end{aligned} \tag{17}$$

Also, the transform of the total time in the system for a job that has service requirement $x \leq \Delta_1$ is given by

$$\tilde{D}_{1,x}(s) = \tilde{W}_1(s) e^{-sx}. \tag{18}$$

The FIFO queue is essentially an M/G/1 queue with arrival rate λ and service time distribution F_1. Thus, from Ref. 9, the distribution of the number in the system at an arbitrary instant is the same as that at an arbitrary arrival epoch and is the same as that seen by a random departure from the FIFO queue (either exiting the system or going to the RR queue). Also, the distribution $\{P_{1,K}\}$ of the number in the FIFO queue at a random departure epoch is related to the sojourn time distribution by

$$\hat{P}_1(z) = \sum P_{1,K} z^K = \tilde{D}_1[\lambda(1 - z)]. \tag{19}$$

Thus the generating function of the number in the FIFO queue at an arbitrary instant is given by

$$\hat{P}_1(z) = \tilde{D}_1[\lambda(1 - z)], \tag{20}$$

where $\tilde{D}_1$ is given by eq. (17).

IV. ANALYSIS OF THE RR QUEUE

In this section we mainly derive the expressions for various quantities of interest for the RR queue. However, in that process we also obtain some additional quantities related to the FIFO queue. As mentioned earlier, in this section we will use the following distribution function of the service time:

$$H(x) = pH_1(x) + (1 - p)(1 - e^{-\mu x}), \qquad 0 \le x < \infty,$$

with

$$H_1(\Delta_1) = 1.$$

As in Section III, let X_1 and X_2 denote the service times in typical chunks of service in the FIFO and the RR queues, respectively. Let F_1 and F_2 be the distribution functions of X_1 and X_2, respectively. Then, from Section III, we have

$$F_1(x) = \begin{cases} pH_1(x) + (1 - p)(1 - e^{-\mu x}) & 0 \le x < \Delta_1 \\ 1 & x \ge \Delta_1, \end{cases}$$

$$F_2(x) = \begin{cases} 1 - e^{-\mu x} & 0 \le x < \Delta_2 \\ 1 & x \ge \Delta_2, \end{cases}$$

$$\tilde{f}_1(s) = p\tilde{h}_1(s) + \frac{(1 - p)}{\mu + s} \{\mu + se^{-(s+\mu)\Delta_1}\},$$

$$\tilde{f}_2(s) = \frac{1}{\mu + s} \{\mu + se^{-(\mu+s)\Delta_2}\},$$

$$\zeta_1 = \lambda \left[ph_1 + \frac{(1 - p)(1 - e^{-\mu\Delta_1})}{\mu} \right],$$

and

$$\zeta_2 = \frac{\lambda(1 - p)e^{-\mu\Delta_1}}{\mu}.$$

We begin by defining and analyzing various busy periods and cycles associated with the FIFO-RR system. These will be used subsequently to derive quantities of interest. The system is said to be busy as long as a job is being processed at high or low priority. The continuous interval of time during which the system is busy is called a system-busy period. A 1-busy period is started by a job arriving at the system while the server is idle and lasts until no job is left in the FIFO queue (so that the server moves to the RR queue). A 2-busy period is started by a service quantum in the RR queue and lasts until the end of this quantum and the time required to empty the FIFO queue. Note that each service quantum in the RR queue generates a 2-busy period and that a system-busy period consists of exactly one 1-busy period, which triggers off the system-busy period and is followed by zero or more 2-busy periods.

Let $\beta(x, k)$ denote the joint probability that the length of the system-busy period is less than or equal to x and that during this busy period exactly k jobs get routed to the RR queue after completing their service quanta in the FIFO queue. Let

$$\beta(s, z) = \int_{0^-}^{\infty} \sum_{k=0}^{\infty} e^{-sx} z^k dB(x, k) \tag{21}$$

denote the joint transform of $B(x, k)$.

Similarly, let $B_1(x, k)[B_2(x, k)]$ denote the joint probability that the length of a 1-busy period (2-busy period) is less than or equal to x and that during this busy period exactly k jobs are moved to the back of the RR queue after receiving one service quantum during that cycle. In the case of a 2-busy period, k includes the job in the RR queue that started this busy period if it was routed to the back of the RR queue. Let $\beta_1(s, z)$ and $\beta_2(s, z)$ denote the joint transforms of $B_1(x, k)$ and $B_2(x, k)$, respectively:

$$\beta_i(s, z) = \int_0^{\infty} e^{-sx} \sum_{k=0}^{\infty} z^k dB_i(x, k), \qquad i = 1, 2. \tag{22}$$

In the Appendix we obtain expressions for these quantities in the form of functional equations.

We will now derive the expressions for the cycle time, the distribution of the number in the RR queue at an arbitrary instant, and the mean sojourn time in the RR queue. The actual distribution function of the sojourn time does not seem to lead to a simple form.

First we consider the number in the RR queue at special time points. We look at the points in time when a service quantum has just completed and the FIFO queue is empty. The interarrival times of the new arrivals and the remaining service requirements of the jobs in the RR queue are independent random variables with exponential distributions. Thus the number in the RR queue at these imbedded instants forms a Markov chain. Let n_k denote the number in the RR queue at the kth such instant. Then we have the following transition mechanism:

$$n_{k+1} = \begin{cases} n_k - 1 + \ell_k & \text{if } n_k \geq 1 \\ \ell_k' & \text{if } n_k = 0, \end{cases} \tag{23}$$

where ℓ_k denotes the number of jobs sent to the back of the RR queue during a 2-busy period (including the message in the RR queue that started this 2-busy period if it gets sent to the back of the RR queue after completing its service quantum), and ℓ_k' denotes the number sent to the RR queue during a 1-busy period.

Let $\Psi_k(z)$ be the generating function of n_k. Then eq. (23) can be rewritten as

$$\Psi_{k+1}(z) = \frac{[\Psi_k(z) - \Psi_k(0)]\beta_2(0, z)}{z} + \Psi_k(0)\beta_1(0, z), \tag{24}$$

and the equilibrium generating function $\Psi(z) = \lim_{k \to \infty} \Psi_k(z)$ is given by

$$\Psi(z) = \frac{\Psi(0)[z\beta_1(0, z) - \beta_2(0, z)]}{z - \beta_2(0, z)}. \tag{25}$$

Equating $\Psi(1)$ with 1, we get the unknown $\Psi(0)$ as

$$\Psi(0) = \frac{1 - \tilde{b}_2}{1 + \tilde{b}_1 - \tilde{b}_2},$$

where

$$\tilde{b}_i = \left. \frac{d\beta_i(0, z)}{dz} \right|_{z=1} \qquad i = 1, 2. \tag{26}$$

We now evaluate the Laplace-Stieltjes transform of the cycle time defined as the time interval between two successive passes through the server by a job in the RR queue. Let t_1 and t_2 be the instants at which the server begins to provide two successive service quanta to a tagged job in the RR queue. (In case the tagged job leaves the system after receiving the first quantum, t_2 is the instant at which the job would have begun to receive the second quantum had it still been in the system.) Then $t_2 - t_1$ is the cycle time.

Let m denote the number of messages in the RR queue at time t_1. Then the generating function of m is given by

$$E[z^m] = E[z^n \mid n \geq 1], \tag{27}$$

where n is the number in the RR queue at the imbedded instants discussed above. Thus

$$E[z^m] = \frac{\Psi(z) - \Psi(0)}{1 - \Psi(0)}, \tag{28}$$

where Ψ is as given by eq. (25). Now, because of the memoryless property of the service requirements of jobs in the RR queue, the cycle time $t_2 - t_1$ is the sum of m independent and identically distributed 2-busy periods. Thus

$$\begin{aligned} \chi(s) = E[s^{(t_2-t_1)}] &= E[\beta_2(s, 1)^m] \\ &= \frac{\Psi(\beta_2(s, 1)) - \Psi(0)}{1 - \Psi(0)}. \end{aligned} \tag{29}$$

We can now obtain an expression for the generating function of the number in the RR queue at an arbitrary instant.

Let $\tilde{n}$ and $\hat{n}$ denote, respectively, the number of jobs in the RR queue just after an arbitrary departure from and just before an arbitrary arrival to the RR queue. Then $\tilde{n}$ and $\hat{n}$ have the same distribution. Let n denote, as before, the number in the RR queue at an instant when a service quantum has just completed and the FIFO queue is empty. Then

$$\begin{aligned} P\{\hat{n} = k\} &= P\{\tilde{n} = k\} \\ &= \frac{\begin{array}{c} P\{n = k + 1 \text{ and a departure occurs at} \\ \text{the end of this service quantum}\} \end{array}}{\begin{array}{c} P\{n \geq 1 \text{ and a departure occurs at the} \\ \text{end of this service quantum}\} \end{array}} \\ &= \frac{P\{n = k + 1\}(1 - e^{-\mu\Delta_2})}{P\{n \geq 1\}(1 - e^{-\mu\Delta_2})} \\ &= \frac{P\{n = k + 1\}}{P\{n \geq 1\}}. \end{aligned} \tag{30}$$

Now, the number of jobs in the RR queue just before a randomly selected arrival to that queue is the same as the number in the RR queue when this tagged job began to receive its first (and only) service quantum in the FIFO queue. This number is a function only of the arrivals prior to the arrival of the tagged job. Also, this number is independent of the tagged job's service time in the FIFO queue and, in particular, is independent of whether or not the tagged job enters the RR queue. Therefore, the number in the RR queue when an arbitrary job completes its service in the FIFO queue has the same distribution as $\hat{n}$. Its generating function is given by

$$\begin{aligned}\xi(z) &= \sum_{K=0}^{\infty} P(\hat{n} = k)z^K \\ &= \sum_{K=0}^{\infty} \frac{P(n = k + 1)z^K}{P(n \geq 1)} \\ &= \frac{(\Psi(z) - \Psi(0))}{z(1 - \Psi(0))}. \end{aligned} \tag{31}$$

We can now derive the generating function of the number in the RR queue at an arbitrary instant. If the observation instant lies in an interval of time during which the server is serving the FIFO queue, the number of jobs in the RR queue is the same as when the job being served finishes its service quantum in the FIFO queue, that is, it has the generating function $\xi(z)$. If the server is working on a job in the RR queue, then the number in the RR queue has the same distribution as the variable m defined above, that is, it has the generating function

$$\frac{\Psi(z) - \Psi(0)}{1 - \Psi(0)}.$$

Finally, if at the observation instant the system is empty, the generating function of the number in the RR queue is 1. Thus the generating function of the number in the RR queue at an arbitrary instant is given by

$$\hat{P}_2(z) = \zeta_1\xi(z) + \zeta_2 \frac{\Psi(z) - \Psi(0)}{1 - \Psi(0)} + (1 - \zeta_1 - \zeta_2). \tag{32}$$

The average number in the RR queue at an arbitrary instant is given by

$$L_2 = \left.\frac{d\hat{P}_2(z)}{dz}\right|_{z=1}. \tag{33}$$

Finally, the mean sojourn time in the RR queue can be obtained by using Little's law for that queue:

$$\bar{S}_2 = \frac{L_2}{\lambda_2} = \frac{\left.\dfrac{d\hat{P}_2(z)}{dz}\right|_{z=1}}{\lambda(1 - \rho)e^{-\mu\Delta_1}}. \tag{34}$$

V. SPECIAL CASES AND NUMERICAL EXAMPLES

We now consider two special cases of the general model analyzed in Sections III and IV. These examples are typical of some communication applications. The service times here will correspond to the number of characters in the message.

The first case corresponds to three types of jobs: one time unit long, Δ_1 time units long, and $\Delta_1 + n\Delta_2$ time units long ($n \geq 1$). Let λ be the total arrival rate. Let X denote the service time of a job. Let

$$q_{11} = P\{X = 1\}, \tag{35}$$

$$q_{12} = P\{X = \Delta_1\}, \tag{36}$$

$$q_1 = q_{11} + q_{12}, \tag{37}$$

$$r_n = \frac{P\{X = \Delta_1 + n\Delta_2\}}{1 - q_1}, \quad n \geq 1. \tag{38}$$

Then

$$\bar{N}_2 = \sum_{n=1}^{\infty} nr_n \tag{39}$$

$$\begin{aligned} \tilde{f}_1(s) &= q_{11}e^{-s} + q_{12}e^{-s\Delta_1} + (1 - q_1)e^{-s\Delta_1} \\ &= q_{11}e^{-s} + (1 - q_{11})e^{-s\Delta_1}, \end{aligned} \tag{40}$$

$$\zeta_1 = \lambda[q_{11} + \Delta_1(1 - q_{11})] \tag{41}$$

$$\tilde{f}_2(s) = e^{-s\Delta_2}, \tag{42}$$

$$\lambda_2 = \lambda\bar{N}_2(1 - q_1), \tag{43}$$

and

$$\zeta_2 = \lambda_2\Delta_2. \tag{44}$$

Thus,

$$\begin{aligned} \tilde{W}_1(s) &= \frac{s(1 - \zeta_1 - \zeta_2) + \lambda_2[1 - \tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1\tilde{f}_1(s)} \\ &= \frac{(s\{1 - \lambda[q_{11} + (1 - q_{11})\Delta_1] - \lambda\bar{N}_2\Delta_2(1 - q_1)\} + \lambda\bar{N}_2(1 - q_1)(1 - e^{-s\Delta_2}))}{s - \lambda + \lambda(q_{11}e^{-s} + (1 - q_{11})e^{-s\Delta_1})}. \end{aligned} \tag{45}$$

The second example corresponds to the traffic mixture assumed in eqs. (1) and (2), that is, a proportion p of the jobs have service time less than or equal to Δ_1 and others have service time exponentially distributed with mean $1/\mu$. Thus,

$$q_1 = p + (1-p)(1 - e^{-\mu\Delta_1}), \tag{46}$$

$$1 - q_1 = (1-p)e^{-\mu\Delta_1}, \tag{47}$$

$$Q_1(t) = pH_1(t) + (1-p)(1 - e^{-\mu t}), \quad 0 \le t < \Delta_1, \tag{48}$$

$$F_1(t) = Q_1(t) = pH_1(t) + (1-p)(1 - e^{-\mu t}), \quad 0 \le t < \Delta_1, \tag{49}$$

and

$$F_1(\Delta_1) - F_1(\Delta_1^-) = (1-p)(1 - e^{-\mu\Delta_1}). \tag{50}$$

Let $\tilde{h}_1$ be the Laplace-Stieltjes transform of H_1. Then, from eqs. (1) and (2), we get

$$\begin{aligned}\tilde{f}_1(s) &= p\tilde{h}_1(s) + (1-p)\int_0^{\Delta_1} e^{-si}\mu e^{-\mu t}dt + (1-p)e^{-s\Delta_1-\mu\Delta_1} \\ &= p\tilde{h}_1(s) + (1-p)\left[\frac{\mu}{s+\mu}(1 - e^{-\Delta_1(s+\mu)}) + e^{-(s+\mu)\Delta_1}\right] \\ &= p\tilde{h}_1(s) + \frac{(1-p)}{s+\mu}(\mu + se^{-(s+\mu)\Delta_1}).\end{aligned} \tag{51}$$

Also,

$$\begin{aligned}r_i &= \frac{(e^{-\mu(\Delta_1+(i-1)\Delta_2)} - e^{-\mu(\Delta_1+i\Delta_2)})[1-p]}{1-q_1} \\ &= \frac{e^{-\mu\Delta_2(i-1)}e^{-\mu\Delta_1}(1 - e^{-\mu\Delta_2})(1-p)}{1-q_1}.\end{aligned} \tag{52}$$

Thus

$$\bar{N}_2 = \frac{e^{-\mu\Delta_1}(1-p)}{(1-q_1)(1-e^{-\mu\Delta_2})} = \frac{1}{(1-e^{-\mu\Delta_2})}, \tag{53}$$

$$\lambda_2 = \frac{\lambda e^{-\mu\Delta_1}(1-p)}{(1-e^{-\mu\Delta_2})}. \tag{54}$$

Finally,

$$F_2(t) = 1 - e^{-\mu t}, \qquad 0 \le t < \Delta_2, \tag{55}$$

and

$$F_2(\Delta_2) - F_2(\Delta_2^-) = e^{-\mu\Delta_2}. \tag{56}$$

Thus,

$$\begin{aligned}\tilde{f}_2(s) &= \int_0^{\Delta_2} \mu e^{-\mu t} e^{-st} dt + e^{-\mu\Delta_2} e^{-s\Delta_2} \\ &= \frac{\mu}{s+\mu}\,(1 - e^{-(s+\mu)\Delta_2}) + e^{-(s+\mu)\Delta_2} \\ &= \frac{1}{s+\mu}\,(\mu + s e^{-(s+\mu)\Delta_2}). \end{aligned} \tag{57}$$

For ζ_1 and ζ_2, we get

$$\zeta_1 = \lambda \left(p h_1 + \frac{(1-p)(1-e^{-\mu\Delta_1})}{\mu} \right), \tag{58}$$

and

$$\begin{aligned}\zeta_2 &= \frac{\lambda e^{-\mu\Delta_1}(1-p)}{(1-e^{-\mu\Delta_2})} \, \frac{1-e^{-\mu\Delta_2}}{\mu} \\ &= \frac{\lambda(1-p)e^{-\mu\Delta_1}}{\mu}. \end{aligned} \tag{59}$$

Thus, from (16) we get

$$\begin{aligned}\tilde{W}_1(s) &= \frac{s(1-\zeta_1-\zeta_2) + \lambda_2[1-\tilde{f}_2(s)]}{s - \lambda_1 + \lambda_1 \tilde{f}_1(s)} \\ &= \frac{s\left(1 - \lambda p h_1 - \lambda \dfrac{(1-p)}{\mu}\right) + \dfrac{\lambda_2 s}{s+\mu}\,(1 - e^{-(s+\mu)\Delta_2})}{s - \lambda p(1 - \tilde{h}_1(s)) - \dfrac{\lambda(1-p)s}{\mu+s}\,(1 - e^{-(s+\mu)\Delta_1})}. \end{aligned} \tag{60}$$

In communication applications there is usually an overhead associated with each segment of transmitted data. Thus, with a typical segment of X_1 characters transmitted from the FIFO queue, δ_1 overhead characters are added. Similarly, δ_2 characters are added to each segment of data transmitted from the RR queue. The effective numbers of characters sent in a typical service segment from the FIFO and the RR queue are then $X_1' = X_1 + \delta_1$ and $X_2' = X_2 + \delta_2$, respectively. The corresponding transforms are then $\tilde{f}_i'(s) = e^{-\delta_i s}\tilde{f}_i(s)$. If we replace $\tilde{f}_1$ and $\tilde{f}_2$ by $\tilde{f}_1'$ and f_2' and adjust the occupancy numbers accordingly in all the waiting time transforms, the resulting expressions will give transforms of the waiting time in the presence of the overhead characters. Besides these overhead characters associated with each service segment, there are usually overhead characters associated with frames, that is, data from various service segments are combined into frames of some maximum size and, at the end of each frame, framing protocol characters are added. The exact analysis of the waiting time in presence of the framing overhead is not easy, but good approximations can be obtained by distributing the framing overhead over all transmitted characters. We have chosen to exclude the framing overhead in our numerical calculations.

Next we numerically evaluate performance measures for short and long messages for a few traffic mixes. We assume that the communication link runs at 56 kb/s. Thus, each character corresponds to 1/7 ms of delay. We use $\delta_1 = \delta_2 = 2$ in all the cases described below.

First consider short messages ($\leq \Delta_1$ characters). The Laplace-Stieltjes transform of the waiting time is given by eq. (16) with appropriate modifications to account for the overhead characters. We numerically inverted this transform using the inversion algorithm of Jagerman[8] for the following traffic mixes and quanta sizes (these traffic mixes are selected to give the same mean number of characters per message):

1. $P(X = 1) = 100/111$, $P(X = \Delta_1) = 10/111$, $P(X = \Delta_1 + n\Delta_2) = r_n \times 1/111$, where $\sum_{n=1}^{\infty} r_n = 1$ and $\bar{N}_2 = \sum_{n=1}^{\infty} nr_n = 99/6$. Also, $\Delta_1 = 16$, $\Delta_2 = 48$. This will be called the traffic mix M_1. Note that $\bar{N}_2$ uniquely defines the delay distribution irrespective of the individual values of r_ns.

2. $P(X = 1) = 100/111$ and with probability 11/111, X is exponentially distributed with mean 912/11. $\Delta_1 = 16$, $\Delta_2 = 48$. This will be called the traffic mix M_2.

3. For the third traffic mix, M_3, we assume that $P\{X = 1\} = 100/111$, $P\{$is exponentially distributed with mean $40\} = 10/111$ and $P\{X$ is exponentially distributed with mean $512\} = 1/111$. $\Delta_1 = 16$ and $\Delta_2 = 48$.

4. The traffic mix here is the same as in M_3 but we use $\Delta_1 = \Delta_2 = 16$ and $\Delta_1 = \Delta_2 = 64$. These cases will be denoted by M_3' and M_3'', respectively. With these sizes of the quanta the cases M_3' and M_e'' correspond to the cases studied in Refs. 3 and 10 with and without the framing overhead, respectively. We will use the results in Ref. 10 to cross check our calculations.

Figures 1 through 3 show the tails of the delay distribution for short messages under the traffic mixes M_1, M_2, and M_3, respectively. Two occupancy numbers are given for each curve in these figures. The lower number is the raw occupancy, while the larger number indicates the total occupancy including the overhead characters. A number larger than one indicates the saturation of the RR queue and an indication that not all the offered messages will be completely trans-

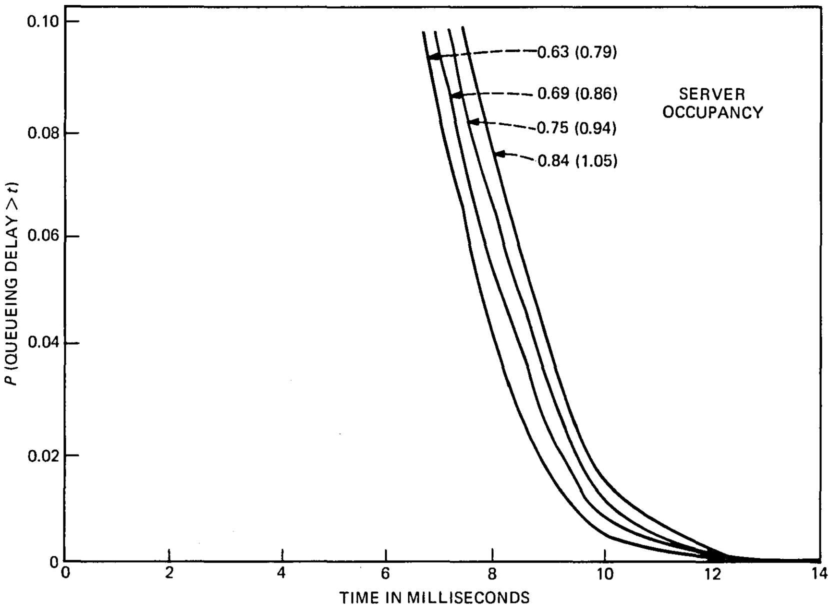


Fig. 1—Delay distribution for traffic mix $M1$: short messages.

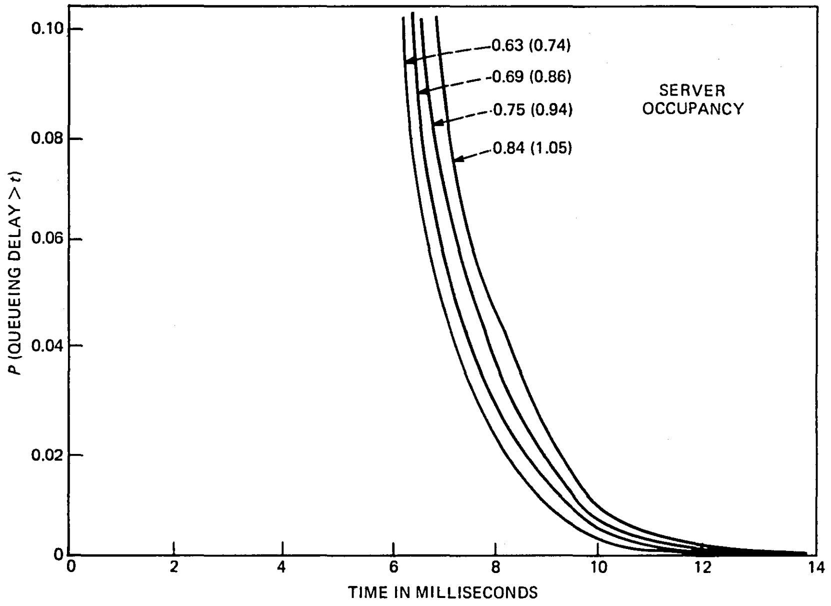


Fig. 2—Delay distribution for traffic mix $M2$: short messages.

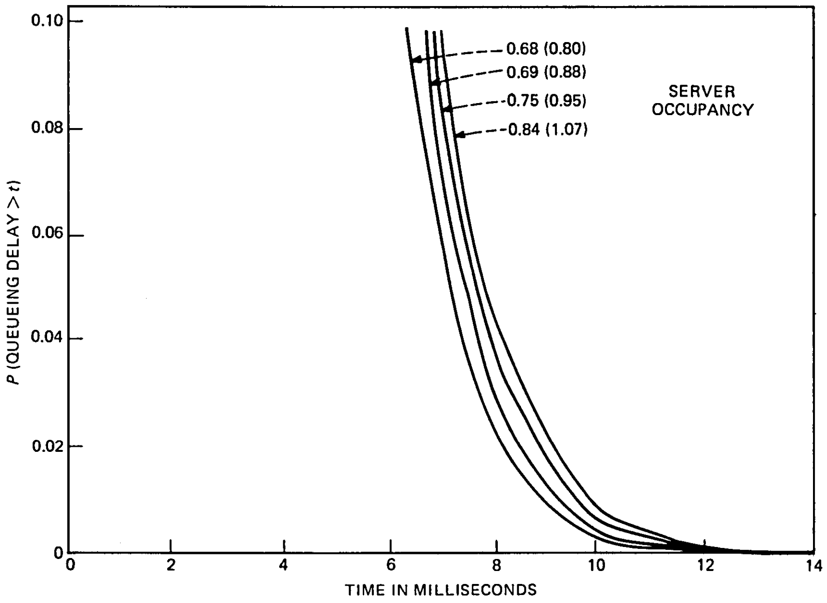


Fig. 3—Delay distribution for traffic mix $M3$: short messages.

mitted. Note, however, that the occupancy of the server due to the FIFO queue is still well below one for these curves.

It is clear from these figures that, for the parameters chosen, the short messages will see very short delays due to queueing even when the long messages see essentially infinite delays. Of course, if the occupancy of the server due to the FIFO is close to one, the short messages will also see long delays.

As mentioned earlier, cases M_3' and M_3'' were selected to match the traffic mix and the quanta sizes of those in Refs. 3 and 10 so that a cross check can be carried out. In these references Fraser and Morgan get the delay distribution (in particular, the 95th percentile of the delay distribution) for short messages via simulation. Since the results in Ref. 3 are obtained in presence of the framing overhead, they cannot be compared directly with our results. However, in their unpublished work Fraser and Morgan[10] obtain the 95th percentile of the delay distribution without the framing overhead. In Fig. 4 we plot the 95th percentile of the delay as a function of the raw occupancy of the server for M_3' and M_3''. The simulation points from Ref. 10 are superimposed on these curves and the agreement looks very good.

We next look at other performance measures studied in Section IV.

We have chosen two different traffic mixes to illuminate the effects of various load parameters on the performance measures relevant to queue 2. The traffic mixes are:

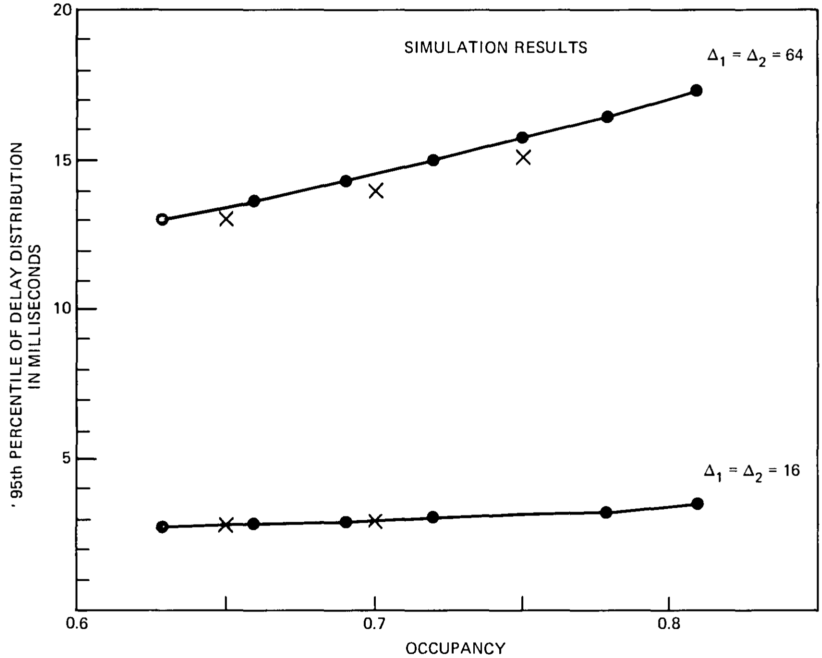


Fig. 4—95th percentile of the distribution for short messages.

1. A message has a single character in it with probability 0.98, and with probability 0.02 it is exponentially distributed with a mean length of five hundred characters. We call this traffic mix M_4.

2. The probability of a message being a single character one is 0.9, and with probability 0.1 it is exponentially distributed with a mean length of one hundred characters. This traffic mix will be referred to as M_5.

In both cases we assume that the "quantum overhead" (i.e., $\delta_1 = \delta_2$) is two characters. The quantum size Δ_1 is assumed to be 16 characters and Δ_2 to be 48 characters. The performance measures relevant to queue 2 include the mean cycle time, the mean sojourn time (for messages entering queue 2), and the mean queue length. These are plotted as functions of the overall occupancy (or, equivalently, the arrival rate) for both traffic mixes in Figs. 5 and 6.

From Figs. 5 and 6 it can be seen that the mean sojourn time for messages entering queue 2 varies essentially in direct proportion to the mean message length of type 2 jobs and in inverse proportion to $(1 - \rho)$, where ρ is the overall utilization of the server. The average cycle time also appears to vary inversely in proportion to $(1 - \rho)$; moreover, it is insensitive to the mean length of type 2 messages as long as it is large compared to Δ_2. The mean queue length also displays

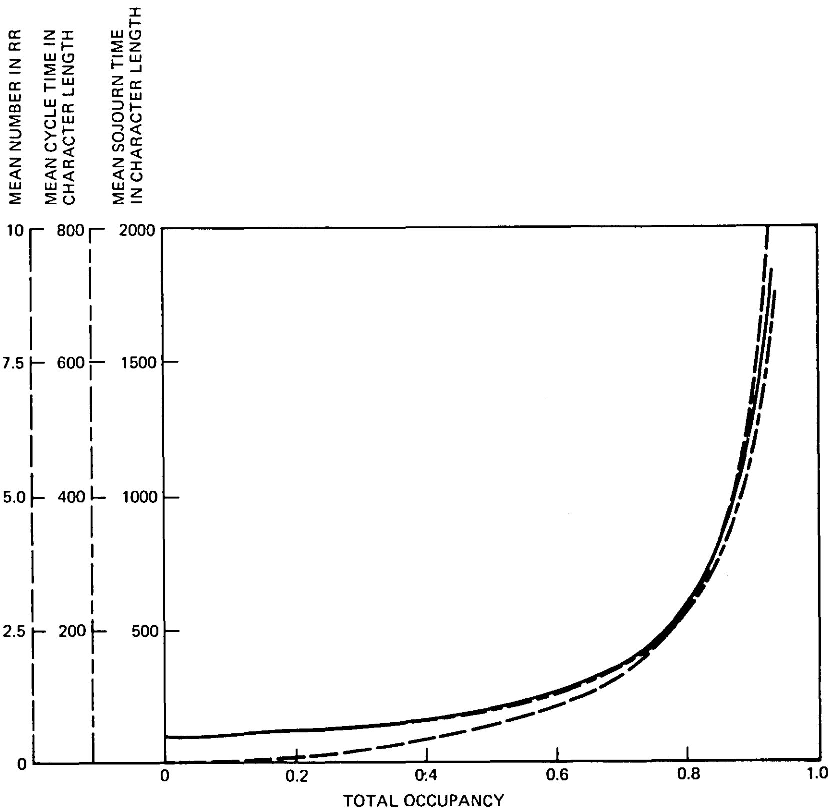


Fig. 5—Some RR performance curves for traffic mix $M4$.

a similar behavior (α $\rho/(1 - \rho)$). The assumption of exponentially distributed message lengths for jobs in queue 2 certainly plays an important role in this behavior. However, it is heartening that the mean cycle time should depend upon the server occupancy alone and be insensitive to the mean message length.

VI. REMARK

In data communication systems providing virtual circuit service it is necessary to move virtual circuits rather than individual messages from the FIFO to the RR queue and vice-versa. That is, once Δ_1 characters are removed for a virtual circuit in the FIFO queue, it is moved to the RR queue. On successive turns Δ_2 characters are removed from this virtual circuit until there are no data to be transmitted on the virtual circuit. At that time the virtual circuit is moved back to the FIFO queue so that the first part of the next message is served

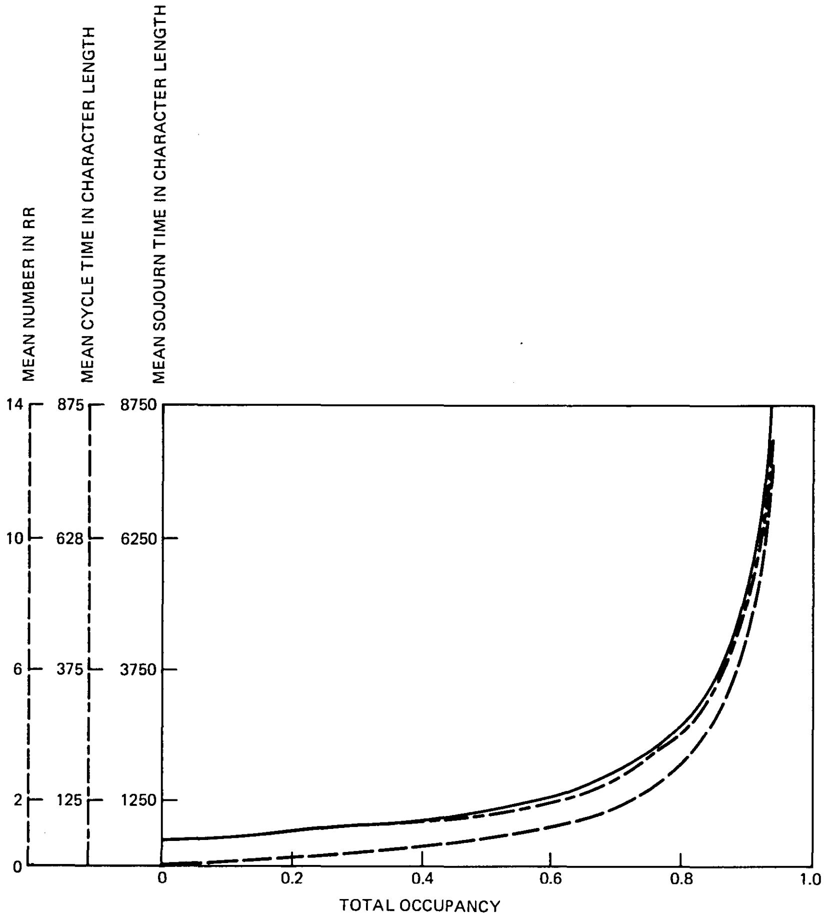


Fig. 6—Some RR performance curves for traffic mix $M5$.

from the FIFO queue. This, of course, implies that a short message immediately following a long message may see considerably longer delay than predicted by our analysis. This is unavoidable but not likely to happen often in practice.

Next, the access lines bringing data to the node that serves the link under consideration may be running slower than the 56-kb/s link. In that case a long message will be seen as a number of short messages by the node using FIFO-RR discipline. This will tend to increase the delay for the short messages. The effects of slower access lines are studied in Refs. 9 and 11. Also, under a heavy load, flow control may force a long message to be transmitted as a number of shorter messages, thus increasing the utilization in the FIFO queue. The effect is similar to that of the slower access lines. Essentially, this breaking up of long messages allows the same message to reappear in the FIFO queue every so often. This could be discouraged by forcing each virtual circuit to pass through the RR queue for at least one cycle after every service in the FIFO queue. Only when a virtual circuit in the RR queue is found empty, it is moved back to the FIFO queue. Genuinely short messages could be exempted from this requirement by making the decision to move the virtual circuit from the FIFO to the RR queue depend on whether Δ_1, or fewer than Δ_1, characters were transmitted.

VII. ACKNOWLEDGMENT

The authors would like to thank S. Morgan for helpful suggestions on an earlier draft and for providing simulation results in Fig. 4.

REFERENCES

1. R. W. Wolff, "Time Sharing With Priorities," SIAM, J. Appl. Math., *19*, No. 3 (November 1970), pp. 566–74.
2. L. E. Schrage, "The Queue M/G/1 With Feedback to Lower Priority Queues," Manage. Sci., *13* (May–June 1967), pp. 466–74.
3. A. G. Fraser and S. P. Morgan, "Queueing and Framing Disciplines for a Mixture of Data Traffic Types," AT&T Bell. Lab. Tech. J., *63*, No. 6, Part 2 (July–August 1984), pp. 1061–87.
4. V. Ramaswami, "Explicit Matrix Geometric Solutions for a Class of Markov Processes," private communication.
5. P. H. Brill and M. J. M. Posner, "Level Crossings in Point Processes Applied to Queues: Single Server Case," Oper. Res., *25*, No. 4 (July–August 1977), pp. 662–73.
6. B. T. Doshi, "An M/G/1 Queue With a Hybrid Discipline," B.S.T.J., *62*, No. 5 (May–June 1983), pp. 1251–71.
7. L. Kleinrock, *Queueing Systems, Vol. II*, New York: Wiley, 1975.
8. D. L. Jagerman, "An Inversion Technique for the Laplace Transform With Application to Approximation," B.S.T.J., *57*, No. 3 (March 1978), pp. 669–710.
9. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30*, No. 2 (March–April 1982), pp. 223–31.
10. A. G. Fraser and S. P. Morgan, unpublished work.
11. H. Rudin, Jr., "Buffered Packet Switching: A Queue With Clustered Arrivals," Int. Switching Symp. Rec., MIT, 1972, pp. 259–65.

APPENDIX

Analysis of the Busy Periods

We now derive expressions for the joint transforms of $\beta_1(x, k)$ and $\beta_2(x, k)$ and the Laplace-Stieltjes transform of the system-busy period.

Recall that

$$\beta_i(s, z) = E[e^{-sb_i}z^K]$$
$$= \int_{0-}^{\infty} \sum_{k=0}^{\infty} e^{-sx}z^k dB_i(x, k), \qquad i = 1, 2, \tag{61}$$

where b_i denotes the length of an i-busy-period and K the number of jobs moving to the back of the RR queue during this busy period.

Let X denote the total service time of a job, X_1 the portion that gets served in the FIFO queue, and N_1 the number of new arrivals during the time X_1. Then

$$\beta_1(s, z) = E[E[e^{-sb_1}z^K \mid X, N_1]], \tag{62}$$

where

$$E[e^{-sb_1}z^K \mid X, N_1] = \begin{cases} e^{-sX}\beta_1^{N_1}(s, z) & 0 \le X \le \Delta_1 \\ e^{-s\Delta_1}z\beta_1^{N_1}(s, z) & X > \Delta_1. \end{cases} \tag{63}$$

Thus

$$\begin{aligned} \beta_1(s, z) = {} & p\int_0^{\Delta_1^+} e^{-sx}e^{-\lambda x[1-\beta_1(s,z)]}dH_1(x) \\ & + (1-p)\int_0^{\Delta_1} \mu e^{-\mu x}e^{-sx}e^{-\lambda x[1-\beta_1(s,z)]}dx \\ & + (1-p)\int_{\Delta_1}^{\infty} e^{-s\Delta_1}ze^{-\lambda\Delta_1[1-\beta_1(s,z)]}\mu e^{-\mu x}dx \\ = {} & p\tilde{h}_1[s + \lambda(1 - \beta_1(s, z))] + (1 - p) \\ & \cdot \frac{\mu}{\mu + s + \lambda(1 - \beta_1(s, z)} [1 - e^{-\Delta_1(\mu+s+\lambda(1-\beta_1(s,z)))}] \\ & + (1 - p)ze^{-(\mu+s+\lambda(1-\beta_1(s,z)))\Delta_1} \\ = {} & \tilde{f}_1(s + \lambda(1 - \beta_1(s, z))) \\ & + (1 - p)(z - 1)e^{\Delta_1(\mu+s+\lambda(1-\beta_1(s,z)))}, \end{aligned} \tag{64}$$

where $\tilde{f}_1$ is as in Section IV.

Similarly, let X_2 denote the length of a typical service in the RR queue. Then

$$\begin{aligned} E[e^{-sb_2}z^K \mid X_2] &= e^{-sX_2}e^{-\lambda X_2(1-\beta_1(s,z))}, & 0 \le X_2 < \Delta_2, \\ &= e^{-sX_2}ze^{-\lambda X_2(1-\beta_1(s,z))}, & X_2 = \Delta_2. \end{aligned} \tag{65}$$

Thus

$$\begin{aligned}\beta_2(s, z) &= E[E[e^{-sb_2}z^K \,|\, X_2]] \\ &= \int_0^{\Delta_2} \mu e^{-\mu x} e^{-x(s+\lambda(1-\beta_1(s,z)))} dx \\ &\quad + z e^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))} \\ &= \frac{\mu}{\mu + s + \lambda(1 - \beta_1(s, z))} (1 - e^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))}) \\ &\quad + z e^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))} \\ &= \tilde{f}_2(s + \lambda(1 - \beta_1(s, z))) + (z - 1)e^{-\Delta_2(\mu+s+\lambda(1-\beta_1(s,z)))}. \qquad (66)\end{aligned}$$

Finally, since the system is work conserving, the system-busy period is the same as that in an ordinary FIFO system. Thus, its Laplace-Stieltjes transform β is given by

$$\beta(s) = \tilde{h}[s + \lambda(1 - \beta(s))], \qquad (67)$$

where

$$\tilde{h}(s) = p\tilde{h}_1(s) + (1 - p)\frac{\mu}{\mu + s}. \qquad (68)$$

In the presence of "chunk overheads," analytic expressions for the busy-period transforms can be obtained by making proper substitutions for $\tilde{h}(\cdot)$, $\tilde{f}_1(\cdot)$, $\tilde{f}_2(\cdot)$, etc., in eqs. (64), (66), and (67).

AUTHORS

Bharat T. Doshi, B. Tech. (Mechanical Engineering), 1970, I.I.T. Bombay; Ph.D. (Operations Research), 1974, Cornell University; AT&T Bell Laboratories, 1979—. Before joining AT&T Bell Laboratories, Mr. Doshi was an assistant Professor at Rutgers University. At AT&T Bell Laboratories his technical work includes modeling and analysis of processor schedules, communication network performance, and overload control. His research interests include queueing and scheduling theory applied to performance analysis of computer, communication, and production systems. Member, IEEE, ORSA; Associate Editor, OR Letters.

Kiran M. Rege, B. Tech. (Electrical Engineering), 1977, I.I.T., Bombay; Ph.D. (Electrical Engineering), 1981, University of Hawaii; AT&T Bell Laboratories, 1982—. Mr. Rege spent 1984 on leave of absence, teaching at I.I.T. Bombay in the Department of Electrical Engineering. His technical work at AT&T Bell Laboratories includes modeling and analysis of switching, computer, and communication systems. His research interests include communication theory, queueing theory, and performance analysis of computer and communication systems.

AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

A Probabilistic Distance Measure for Hidden Markov Models

By B.-H. JUANG and L. R. RABINER*

(Manuscript received July 31, 1984)

We propose a probabilistic distance measure for measuring the dissimilarity between pairs of hidden Markov models with arbitrary observation densities. The measure is based on the Kullback-Leibler number and is consistent with the reestimation technique for hidden Markov models. Numerical examples that demonstrate the utility of the proposed distance measure are given for hidden Markov models with discrete densities. We also discuss the effects of various parameter deviations in the Markov models on the resulting distance, and study the relationships among parameter estimates (obtained from reestimation), initial guesses of parameter values, and observation duration through the use of the measure.

I. INTRODUCTION

Consider two N-state first-order hidden Markov models specified by the parameter sets $\lambda_i = (\mathbf{u}^{(i)}, \mathbf{A}^{(i)}, \mathbf{B}^{(i)})$, $i = 1, 2$, where $\mathbf{u}^{(i)}$ is the initial state probability vector, $\mathbf{A}^{(i)}$ is the state transition probability matrix, and $\mathbf{B}^{(i)}$ is either an $N \times M$ stochastic matrix (if the observations are discrete) or a set of N continuous density functions (if the observations are continuous).[1] Our interest in this paper is to define a distance for every such pair of hidden Markov models (λ_1, λ_2) so we can measure the dissimilarity between them. Another goal is to study the properties of hidden Markov models, using the distance measure, in order to understand the model sensitivities.

* Authors are employees of AT&T Bell Laboratories.

The need of such a distance measure arises mainly in estimation and classification problems involving Hidden Markov Models (HMMs). For example, in using a reestimation algorithm to iteratively estimate the model parameters,[2] a distance measure is necessary not only to monitor the behavior of the reestimation procedure, but to indicate the expected performance of the resulting HMM. In classification, a good distance measure would greatly facilitate the nearest-neighbor search, defining Voronoi regions[3] or applying the generalized Lloyd algorithm[4] for hidden Markov model clustering.

The only measure for comparing pairs of HMMs that has appeared previously in the literature is the one proposed by Levinson et al. for discrete-observation density hidden Markov models.[1] The distance, which is a Euclidean distance on the state-observation probability matrices, is defined as

$$d(\lambda_1, \lambda_2) \triangleq \|\mathbf{B}^{(1)} - \mathbf{B}^{(2)}\| \triangleq \left\{\frac{1}{MN}\sum_{j=1}^{N}\sum_{k=1}^{M}[b_{jk}^{(1)} - b_{p(j)k}^{(2)}]^2\right\}^{1/2}, \quad (1)$$

where $\mathbf{B}^{(i)} = [b_{jk}^{(i)}]$ is the state-observation probability matrix in model λ_i and $p(j)$ is the state permutation that minimizes the measure of eq. (1). The metric of eq. (1) was called a "measure of estimation error" in Ref. 1 and was used to characterize the estimation error occurring in the reestimation process. Minimum bipartite matching was used to determine the optimum state permutation for aligning the states of the two models. The measure of eq. (1) did not depend at all on estimates of $\mathbf{u}$ or $\mathbf{A}$, since it is generally agreed that the $\mathbf{B}$ matrix is, in most cases, a more sensitive set of parameters related to the closeness of HMMs than the $\mathbf{u}$ vector or the $\mathbf{A}$ matrix.

The distance measure of eq. (1) is inadequate for the following reasons: (1) it does not take into account the deviations in all the parameters of the HMM; (2) its evaluation requires a great deal of computation in the discrete case and probably would become intractable when dealing with continuous-observation hidden Markov models; and (3) it is unreliable when comparing HMMs with highly skewed densities. Hence, our aim is to find a distance measure that truly measures the dissimilarity between pairs of hidden Markov models, can be easily evaluated, is reliable for any pair of Markov models, and is meaningful in the probabilistic framework of the HMM itself.

In this paper, we propose such a distance measure for comparing pairs of HMMs that follows the concept of divergence,[5] cross entropy, or discrimination information.[6] The distance measure, denoted by $D(\lambda_1, \lambda_2)$, has the form

$$\log \Pr(\mathbf{O}_T|\lambda_1) - \log \Pr(\mathbf{O}_T|\lambda_2),$$

where $\mathbf{O}_T$ symbolizes an observation sequence of T observations. Because the distance measure is the difference in log probabilities of the observation sequence conditioned on the models being compared, it will sometimes be referred to as "divergence distance" or "directed divergence measure." In the next section, we formally define the distance measure, and we discuss Petrie's results[7] that further give the distance measure theoretical justification. In Section III, numerical examples related to discrete-observation hidden Markov models are given. We show the effects of individual parameter deviations upon the distance measure and demonstrate several interesting properties of discrete-observation models that are made explicit through the use of the proposed distance measure. A discussion of the use of such a distance measure in continuous-observation models is given in Ref. 8, where hidden Markov models with continuous mixture densities are discussed.

II. DEFINITION OF THE PROPOSED HMM DISTANCE MEASURE

In this section, we define the distance measure for any pair of Markov models, discuss Petrie's Limit Theorem and statistical analysis of probabilistic functions of Markov chains,[7] and then give the proposed distance measure an interpretation from the Kullback-Liebler statistic point of view. The presentation is explicit for discrete-observation models but can easily be extended to continuous-observation cases.

Let $\mathscr{A}_s = \{1, 2, \cdots, N\}$ be a state alphabet, and let $\mathscr{A}_0 = \{y_1, y_2, \cdots, y_M\}$ be an observation alphabet. The Cartesian product $\mathscr{O}_\infty = \prod_{t=1}^{\infty} \mathscr{A}_{0t}$, $\mathscr{A}_{0t} = \mathscr{A}_0$, for all t, forms an observation space in which every point $\mathbf{O}$ has coordinate $\mathbf{o}_t \in \mathscr{A}_{0t} = \mathscr{A}_0$. We are concerned about a class of stochastic processes generated by a hidden Markov source defined by an $N \times N$ ergodic stochastic matrix $\mathbf{A} = [a_{ij}]$ and by an $N \times M$ stochastic matrix $\mathbf{B} = [b_{jk}]$. Matrix $\mathbf{A}$, the state transition probability matrix, generates a stationary Markov process $\mathbf{S} = \cdots \mathbf{s}_{t-1}\mathbf{s}_t\mathbf{s}_{t+1} \cdots$ according to $a_{ij} = \Pr\{\mathbf{s}_{t+1} = j \mid \mathbf{s}_t = i\}$. Based upon $\mathbf{S}$, $\mathbf{B}$ generates $\mathbf{o}_t$ according to $b_{jk} = \Pr\{\mathbf{o}_t = \mathrm{y}_k \mid \mathbf{s}_t = \mathrm{j}\}$. Let $\mathbf{a}^* = [a_1, a_2, \cdots, a_N]$ be the stationary absolute distribution vector for $\mathbf{A}$, i.e., $\mathbf{a}^* \mathbf{A} = \mathbf{a}^*$, where * denotes the transpose. Then, matrices $\mathbf{A}$ and $\mathbf{B}$ define a measure, denoted by $\mu(\cdot \mid \boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B})$, on $\mathscr{O}_\infty$ by

$$\mu(\mathbf{O}_T \mid \boldsymbol{\lambda}) = \sum_{\text{all } \mathbf{S}_T} a_{\mathbf{s}_0} \prod_{t=1}^{T} a_{\mathbf{s}_{t-1}\mathbf{s}_t} b_{\mathbf{s}_t I(\mathbf{o}_t)}, \tag{2}$$

where $\mathbf{O}_T = (\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T)$ is the observed sequence up to time T (i.e., a truncated $\mathbf{O}$), $\mathbf{S}_T = (\mathbf{s}_0, \mathbf{s}_1, \cdots, \mathbf{s}_T)$ is the corresponding unobserved state sequence, and $I(\cdot)$ is the index function

$$I(\mathbf{o}_t) = k \quad \text{if} \quad \mathbf{o}_t = y_k.$$

Let $\mathbf{\Lambda}_a$ be the space of $N \times N$ *ergodic* stochastic matrices, $\mathbf{\Lambda}_b$ be the space of $N \times M$ stochastic matrices, and $\mathbf{\Lambda} = \mathbf{\Lambda}_a \times \mathbf{\Lambda}_b$. Clearly, $\lambda \in \mathbf{\Lambda}$, and for every point in $\mathbf{\Lambda}$ there is a stationary measure $\mu(\cdot \mid \lambda)$ associated with it.

Now consider a probability space $(\mathscr{O}_\infty, \mu(\cdot \mid \lambda_0))$, which will be abbreviated as $(\mathscr{O}_\infty, \lambda_0)$ in the following without ambiguity. Let an observation sequence $\mathbf{O}_T$ be generated according to the distribution $\mu(\cdot \mid \lambda_0)$.

For each T and each $\mathbf{O} \in \mathscr{O}_\infty$, define the function $H_T(\mathbf{O}, \lambda)$ on $\mathbf{\Lambda}$ by

$$H_T(\mathbf{O}, \lambda) = \frac{1}{T} \log \mu(\mathbf{O}_T \mid \lambda). \tag{3}$$

Each $H_T(\cdot, \lambda)$ is thus a random variable on the probability space $(\mathscr{O}_\infty, \lambda_0)$. Also, for a given fixed observation $\mathbf{O}_T$, $H_T(\mathbf{O}, \lambda)$ is a function on $\mathbf{\Lambda}$. Petrie[7] proved (limit theorem) that for each λ in $\mathbf{\Lambda}$,

$$\begin{aligned} \lim_{T \to \infty} H_T(\mathbf{O}, \lambda) &= \lim_{T \to \infty} \frac{1}{T} \log \mu(\mathbf{O}_T \mid \lambda) \\ &= H(\lambda_0, \lambda) \end{aligned} \tag{4}$$

exists almost everywhere $\mu(\cdot \mid \lambda_0)$. Furthermore,

$$H(\lambda_0, \lambda_0) \geqslant H(\lambda_0, \lambda) \tag{5}$$

with equality if and only if $\lambda \in G(\lambda_0) = \{\lambda \in \mathbf{\Lambda} \mid \mu(\cdot \mid \lambda) = \mu(\cdot \mid \lambda_0)$ as measures on $\mathscr{O}_\infty\}$. Define $\mathbf{\Lambda}_T(\mathbf{O}) = \{\lambda' \in \mathbf{\Lambda} \mid H_T(\mathbf{O}, \lambda)$ is maximized at $\lambda'\}$. Then, $\mathbf{\Lambda}_T(\mathbf{O}) \to G(\lambda_0)$ almost everywhere $\mu(\cdot \mid \lambda_0)$ (see Ref. 7, Theorem 2.8). The results give further justification to the well-known reestimation procedure[9] for Markov modeling.

With the above background, we define a distance measure $D(\lambda_0, \lambda)$ between two Markov sources λ_0 and λ by

$$\begin{aligned} D(\lambda_0, \lambda) &= H(\lambda_0, \lambda_0) - H(\lambda_0, \lambda) \\ &= \lim_{T \to \infty} \frac{1}{T} [\log \mu(\mathbf{O}_T \mid \lambda_0) - \log \mu(\mathbf{O}_T \mid \lambda)]. \end{aligned} \tag{6}$$

The aforementioned limit theorem guarantees the existence of such a distance measure and eq. (5) ensures that $D(\lambda_0, \lambda)$ is nonnegative. $D(\lambda_0, \lambda) = 0$ if and only if $\lambda \in G(\lambda_0)$, a point that is indistinguishable by the associated probability measure.

By invoking ergodicity,[10] we see that the distance is in fact the Kullback-Leibler number[6] between measures $\mu(\cdot \mid \lambda_0)$ and $\mu(\cdot \mid \lambda)$. If $\mathscr{H}_0$ and $\mathscr{H}_1$ are the hypotheses that $\mathbf{O}_T$ is from the statistical population with measure $\mu(\cdot \mid \lambda_0)$ and $\mu(\cdot \mid \lambda_1)$, respectively, $D(\lambda_0, \lambda)$ is then the average information per observation sample in $\mathbf{O}_T$ for discrimination in favor of $\mathscr{H}_0$ against $\mathscr{H}_1$. Since $\mathbf{O}_T$ is generated according to $\mu(\cdot \mid \lambda_0)$, $\lim_{T\to\infty}(1/T)\log \mu(\mathbf{O}_T \mid \lambda_0)$ should be a maximum over Λ, and $D(\lambda_0, \lambda)$ is a measure of directed divergence, from λ_0 to λ, manifested by the observation $\mathbf{O}_T$.

The distance measure of eq. (6) is clearly nonsymmetric. A natural extension of this measure is the symmetrized version of eq. (6), i.e.,

$$D_s(\lambda_0, \lambda) = \frac{1}{2}[D(\lambda_0, \lambda) + D(\lambda, \lambda_0)], \tag{7}$$

which is the average of the two nonsymmetric distances. $D_s(\lambda_0, \lambda)$ is symmetric with respect to λ_0 and λ and represents a measure of the difficulty (or ease) of discriminating between $\mu(\cdot \mid \lambda_0)$ and $\mu(\cdot \mid \lambda)$, or equivalently, λ_0 and λ. For our purpose, however, there is no particular requirement that the distance be symmetric, and our study will mainly concentrate on the definition of eq. (6).

III. DISCRETE-OBSERVATION HIDDEN MARKOV MODELS

Using the distance measure of eq. (6), we have studied the behavior of several discrete-observation hidden Markov models. In this section, we present some results on the sensitivities of the reestimation procedure to observation sequence length, initial parameter estimates, etc. We begin with a discussion of the evaluation of such a distance measure.

3.1 Evaluation of the distance measure

Evaluation of the distance of eq. (6) is rather straightforward. A standard Monte Carlo simulation procedure based upon a good random number generator is used to generate the required observation sequence $\mathbf{O}_T$ according to the given distribution $\mu(\cdot \mid \lambda_0)$. The probabilities of observing the generated sequence from models λ_0 and λ are then calculated respectively. By way of example, Fig. 1a shows the logarithm of $\mu(\mathbf{O}_T \mid \lambda_0)$ and $\mu(\mathbf{O}_T \mid \lambda)$, respectively, as a function of the observation duration T. The resulting distance $D(\lambda_0, \lambda)$ is then plotted in Fig. 1b. For this example, $\lambda_0 = (\mathbf{A}_0, \mathbf{B}_0)$, $\lambda = (\mathbf{A}, \mathbf{B})$, $N = M = 4$, where

$$\mathbf{A}_0 = \begin{bmatrix} 0.8 & 0.15 & 0.05 & 0 \\ 0.07 & 0.75 & 0.12 & 0.06 \\ 0.05 & 0.14 & 0.8 & 0.01 \\ 0.001 & 0.089 & 0.11 & 0.8 \end{bmatrix} \quad \mathbf{B}_0 = \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix}$$

and

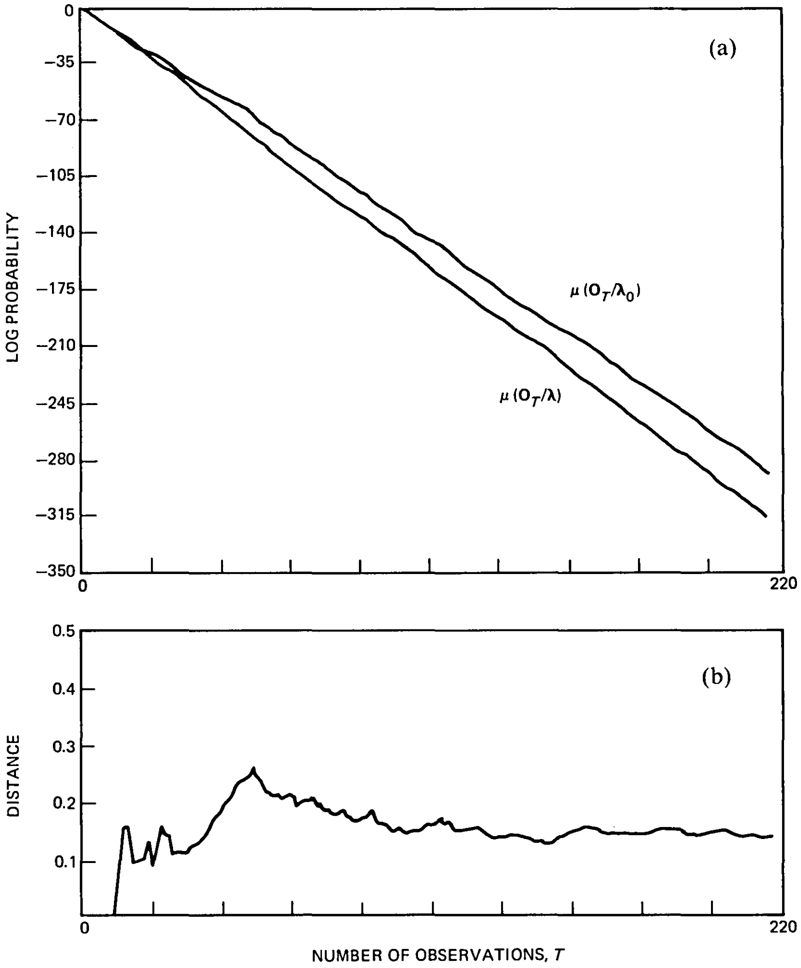


Fig. 1—(a) Log probabilities $\mu(\mathbf{O}_T \mid \lambda_0)$ and $\mu(\mathbf{O}_T \mid \lambda)$ versus the number of observations for a pair of models that are close in distance. (b) Distance $D(\lambda_0, \lambda)$ versus the number of observations for the same pair of models.

$$\mathbf{A} = \begin{bmatrix} 0.4 & 0.25 & 0.15 & 0.2 \\ 0.27 & 0.45 & 0.22 & 0.06 \\ 0.35 & 0.14 & 0.4 & 0.11 \\ 0.111 & 0.119 & 0.23 & 0.54 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.1 & 0.15 & 0.65 & 0.1 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.15 & 0.25 & 0.4 & 0.2 \end{bmatrix}.$$

We can see from Fig. 1 that, for this example, it takes around 150 observation samples to converge to a distance of 0.14 (to within statistical fluctuations). It is readily shown that the number of observations needed for convergence of the distance to a fixed value is dependent on N and M.

Although the definition of the distance of eq. (6) requires that the pair of models being compared both be ergodic and that there exist a stationary absolute distribution vector **a** such that $\mathbf{a}^* \mathbf{A} = \mathbf{a}^*$, practical evaluation of the distance can still be performed for other types of Markov models. We often define the distance measure by replacing the stationary equilibrium distribution vector with the initial state probability vector. In the case of left-to-right models,[1] we use a series of restarted sequences as the generated sequence for distance evaluation, because of the trap state in left-to-right models. In fact, except for some possible minor theoretical discrepancies (which might be traced back to the problem of nonergodic model estimation), the proposed distance measure appears to work quite reliably for any pair of such HMMs. Particularly, in the previous example, the initial state probability vectors associated with models λ_0 and λ were $\mathbf{u}_0^* = [0.75\ 0.15\ 0.05\ 0.05]$ and $\mathbf{u}^* = [0.4\ 0.25\ 0.15\ 0.2]$, respectively.

3.2 Effects of parameter deviations on the distance

We are interested in studying the relationship between parameter deviation and model distance, as well as the relative sensitivity of the distance to different parameter sets that define the HMMs. To illustrate such parameter sensitivities, we have studied HMMs whose parameters are related to the matrices $\mathbf{W}_1$ and $\mathbf{W}_2$,

$$\mathbf{W}_1 = \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 0.05 & 0.1 & 0.65 & 0.25 \\ 0.1 & 0.05 & 0.75 & 0.1 \\ 0.45 & 0.45 & 0.05 & 0.05 \\ 0.05 & 0.1 & 0.65 & 0.2 \end{bmatrix},$$

and to the vector $\mathbf{v}^* = [0.75\ 0.15\ 0.05\ 0.05]$. In particular, model λ_0 is defined by $(\mathbf{u}_0, \mathbf{A}_0, \mathbf{B}_0)$, where $\mathbf{u}_0 = \mathbf{v}$ and $\mathbf{A}_0 = \mathbf{B}_0 = \mathbf{W}_1$. We chose $\mathbf{A}_0 = \mathbf{B}_0$ to avoid a priori numerical difference in different parameter sets. The alternate model, $\lambda = (\mathbf{u}, \mathbf{A}, \mathbf{B})$, is varied from λ_0 by modifying, in turn, either **A** or **B**.

We first study the effect of changes in only the state transition probability matrix on the computed distance. We form a sequence of models $\lambda = (\mathbf{u}, \mathbf{A}, \mathbf{B})$, where $\mathbf{u} = \mathbf{u}_0 = \mathbf{v}$, $\mathbf{B} = \mathbf{B}_0 = \mathbf{W}_1$ and

$$\mathbf{A} = \left(\frac{1}{1+\delta}\right)\mathbf{W}_1 + \left(\frac{\delta}{1+\delta}\right)\mathbf{W}_2, \tag{8}$$

with δ varying from 0.001 to 0.991 in 99 equal steps. For each pair (λ_0, λ), $D(\lambda_0, \lambda)$ is then evaluated. The bottom curve in Fig. 2a shows a plot of $D(\lambda_0, \lambda)$ as a function of the deviation factor δ. Furthermore, for potential geometric interpretations, we calculate the signal-to-noise ratio γ_A, defined by

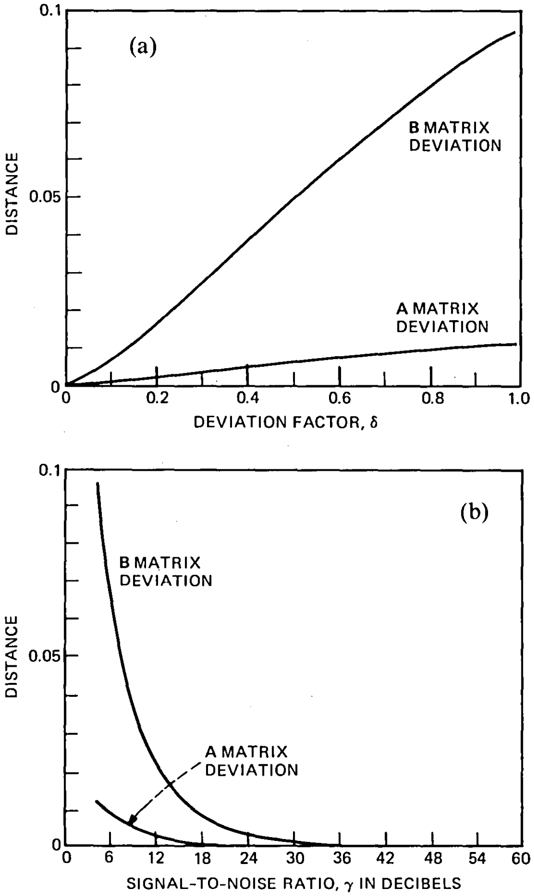


Fig. 2—(a) Relationship between the probabilistic distance and the model deviation factor δ of the **A** and **B** parameters for a pair of HMMs. (b) Relationship between the probability distance and measured parameter deviations of the **A** and **B** parameters for a pair of HMMs.

$$\gamma_{\mathrm{A}} = 10 \log_{10} \frac{||A_0||^2}{||A_0 - A||^2}, \tag{9}$$

where $||\cdot||$ denotes matrix norm ($||\mathbf{A}||^2 = \sum_i \sum_j a_{ij}^2$ for $\mathbf{A} = [a_{ij}]$). For small δ, $\mathbf{A}$ is very close to $\mathbf{A}_0$ and γ_{A} is large. Accordingly, the distance $D(\lambda_0, \lambda)$ as a function of γ_{A} is plotted in Fig. 2b. (Note that small values of δ in Fig. 2a correspond to large values of γ_{A} in Fig. 2b—i.e., the direction of the curves is reversed.)

Similarly, we study the effect of changes in only the observation probability matrix **B**. The sequence of models $\lambda = (\mathbf{u}, \mathbf{A}, \mathbf{B})$ for comparison is formed by setting $\mathbf{u} = \mathbf{u}_0 = \mathbf{v}$, $\mathbf{A} = \mathbf{A}_0 = \mathbf{W}_1$ and

$$\mathbf{B} = \left(\frac{1}{1+\delta}\right)\mathbf{W}_1 + \left(\frac{\delta}{1+\delta}\right)\mathbf{W}_2, \tag{10}$$

again, with δ varying from 0.001 to 0.991. The relationships, $D(\lambda_0, \lambda)$ versus δ and $D(\lambda_0, \lambda)$ versus $\gamma_{\mathbf{B}}$,

$$\gamma_{\mathbf{B}} = 10 \log_{10} \frac{||\mathbf{B}_0||^2}{||\mathbf{B}_0 - \mathbf{B}||^2}, \tag{11}$$

are shown as the upper curves in Figs. 2a and b, respectively.

Both curves of Fig. 2b show a simple monotonic exponential trend for the example studied. This exponential trend may be intuitively anticipated from eq. (2), which shows that μ is in the form of a product. This monotonic relationship is, in general, true when the signal-to-noise ratio is adequately high, i.e., models are close enough in the Euclidean distance sense. This result is consistent with Theorem 3.19 in Ref. 7, which gives the set $\mathbf{G}(\lambda_0)$ a geometric interpretation. For more complicated models or other types of deviations than those of eqs. (8) and (10), however, the simple monotonic exponential relationship of the type shown in Fig. 2b may not be observed in low signal-to-noise ratio regions.

Another important property of the distance measure, as seen in Fig. 2, is that deviations in the observation probability matrix **B** give, in general, larger distance scores than similar deviations in the state-transition matrix **A**. Thus, the **B** matrix appears to be numerically more important than the **A** matrix in specifying a hidden Markov model. It is our opinion that this may be a desirable inherent property of hidden Markov models for speech recognition applications.

3.3 Examples of the use of the distance in model estimation

3.3.1 Ergodic models

Consider the following models:

1. λ_a: $N = M = 4$, balanced model

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix};$$

2. λ_b: $N = M = 4$, skewed model

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0.25 & 0.75 \\ 0.15 & 0 & 0 & 0.85 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0.22 & 0.78 & 0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.25 & 0.75 & 0 & 0 \\ 0 & 0.15 & 0.85 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0.2 & 0 & 0 & 0.8 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix};$$

3. λ_c: $N = M = 5$, deterministic observation

$$\mathbf{A} = \begin{bmatrix} 0 & 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \\ 0.4 & 0 & 0.2 & 0 & 0.4 \\ 0.3 & 0.2 & 0.1 & 0.4 & 0 \\ 0.2 & 0.1 & 0.2 & 0.4 & 0.1 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

It should be pointed out that λ_a is a balanced model in which transitions as well as observations are equiprobable within the structural constraints, while λ_b is a skewed model with the same Markov chain structure as λ_a. Model λ_c has a unique observation probability matrix, namely the identity matrix, which links observations to distinct model states.

A number of observation sequences, $\mathbf{O}_T$, of different duration were generated from these models. Then, for each $\mathbf{O}_T$ sequence, a model estimate, generically denoted as λ_a', λ_b', or λ_c', was obtained using the reestimation algorithm, which, starting from an arbitrary guess, iterated until a certain convergence criterion was met.[1]

Each sequence $\mathbf{O}_T$ of duration T thus corresponds to a model estimate for which the divergence distance can be evaluated from the generating model. Figures 3a, b, and c are plots of $D(\lambda_a, \lambda_a')$, $D(\lambda_b, \lambda_b')$, and $D(\lambda_c, \lambda_c')$ respectively, as a function of the duration T. These figures display typical simulation results of the statistical reestimation technique. Important considerations behind the simulation process include: (1) characteristics of the generating source, such as λ being a balanced or skewed model; (2) effectiveness of the estimation technique; and (3) the number of observations needed for a good estimate. Here we provide qualitative discussions of the plotted results.

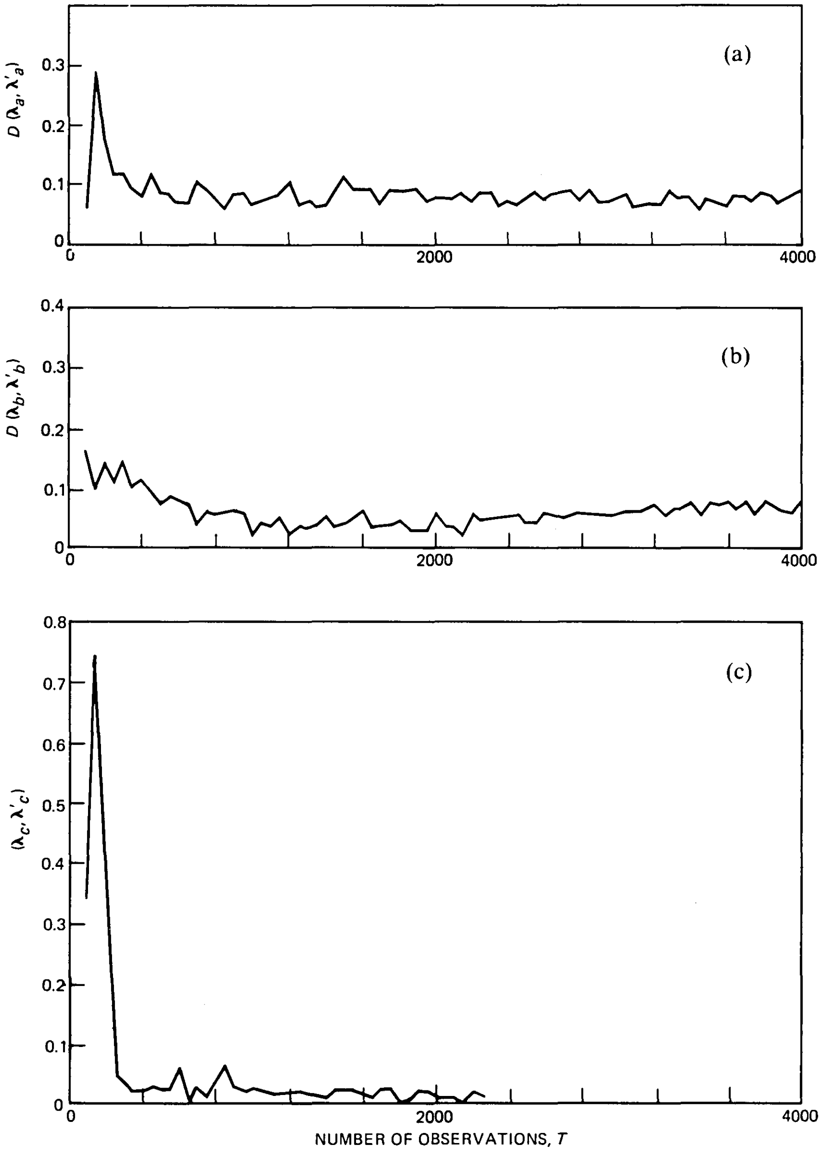


Fig. 3—Distance performance of reestimated ergodic models as a function of the observation duration: (a) λ_a, balanced model; (b) λ_b, skewed model; (c) λ_c, model with deterministic observation.

Figure 3a indicates that the distance between λ_a and λ_a' stabilizes after T grows beyond about five hundred samples. The distance for $T > 500$ is small (about 0.085), with a range of statistical variation between ±0.025. The distance scores of Fig. 3b do not seem to be as well behaved as those of Fig. 3a. Although the estimate λ_b' for λ_b may be as good as λ_a', judging from the distance, λ_b' appears to be more data dependent. A slow drifting from $D \simeq 0.04$ at $T = 1000 \sim 2000$ region to $D \simeq 0.07$ at $T = 3000 \sim 4000$ region is seen. This can be attributed to the fact that λ_b is a skewed model and the associated measure $\mu(\cdot \mid \lambda_b)$ has a slightly wider dynamic range than $\mu(\cdot \mid \lambda_a)$; hence deviations in D are manifested over a broad range of values of T. Those generated sequences, $\mathbf{O}_T$, of high $\mu(\mathbf{O}_T \mid \lambda_b)$ will result in a close estimate λ_b', and the wide dynamic range in $\mu(\cdot \mid \lambda_b)$ will directly translate into the observed variations in $D(\lambda_b, \lambda_b')$ for long observation sequences. This long-term drifting of D is reminiscent of the residual difference between uncorrelated and highly correlated sources in statistical data analysis.

The results of Fig. 3c indicate that when the generating source involves only a Markov chain and does not have variations in the observation density, very good estimates can be obtained with a small amount of data. Also, the **B** matrix, because it is an identity matrix, greatly narrows the range of $\mu(\cdot \mid \lambda_c)$, resulting in negligible variations in $D(\lambda_c, \lambda_c')$ when $\mathbf{O}_T$ is sufficiently long.

3.3.2 Left-to-right models

Another series of simulations dealt with nonergodic models of the types shown in Fig. 4. These models are identical to the three models SRC195, SRC295, and SRC395 studied in Ref. 1. We denote these models by λ_{195}, λ_{295}, and λ_{395} as in Fig. 4. For these models, $N = 5$, $M = 9$, $\mathbf{u}^* = [1\ 0\ 0\ 0\ 0]$, and

$$\mathbf{B} = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.3 & 0.7 \end{bmatrix}.$$

An additional model, λ_{595}, which had the same state transition probability and initial state probability as λ_{295} but with the following **B** matrix

$$\mathbf{B} = \begin{bmatrix} 0.8 & 0.1 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0.8 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.8 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0.8 & 0.1 \end{bmatrix},$$

was also studied.

The observation matrix **B** for λ_{195} through λ_{395} is *non-overlapping*; observations generated during one state cannot appear during another state. As was the case for model λ_c in the previous section, these

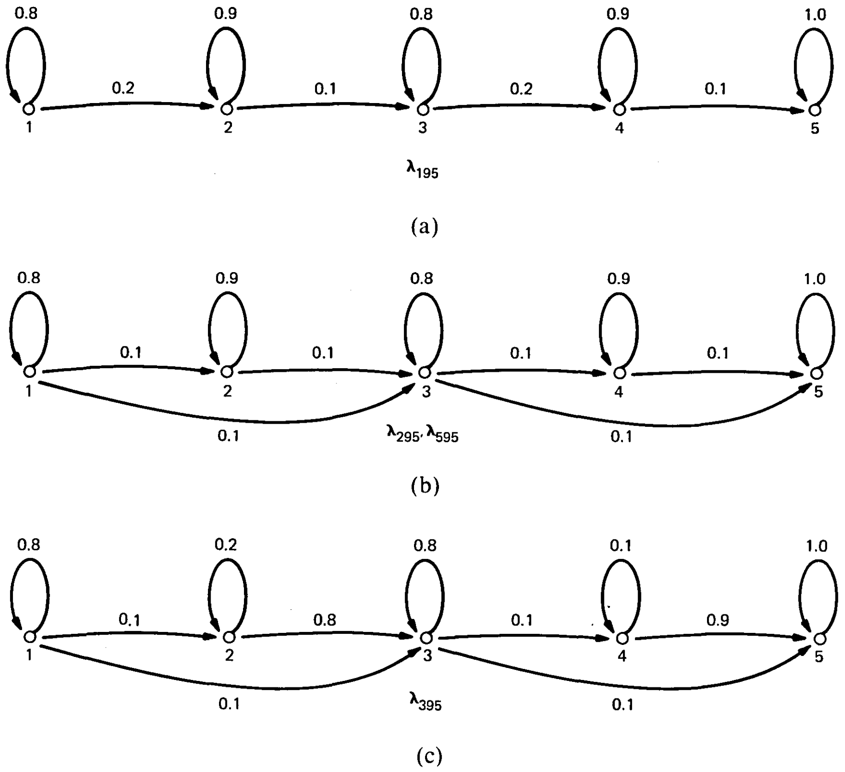


Fig. 4—Left-to-right hidden Markov models: (a) λ_{195}, (b) λ_{295} and λ_{595}, and (c) λ_{395}, used in the study of model parameter estimation sensitivity.

models show a rigid correspondence between states and observations. In this case, we can observe some particular effects of the coupling between matrices **A** and **B** upon model estimates as well as the distance.

The same simulation procedure as above was followed: (1) observation sequences were first generated; (2) model estimates were then obtained using the reestimation algorithm with different initial guesses; and (3) distances between the generating model and the estimated model were calculated and plotted as a function of T, the *total* sequence duration. (Note that because of the trap state in these models, the measurement sequence is a series of restarted sequences and T is the total duration.) Four kinds of initial guesses of model parameters were used. Type 1 is a totally random guess (except for the necessary stochastic constraints). Type 2 is a random guess with known state-transition constraints; that is, elements in **A** corresponding to prohibited transitions are initially set to null, while others are randomly chosen with stochastic constraints. Type 3 is a random guess

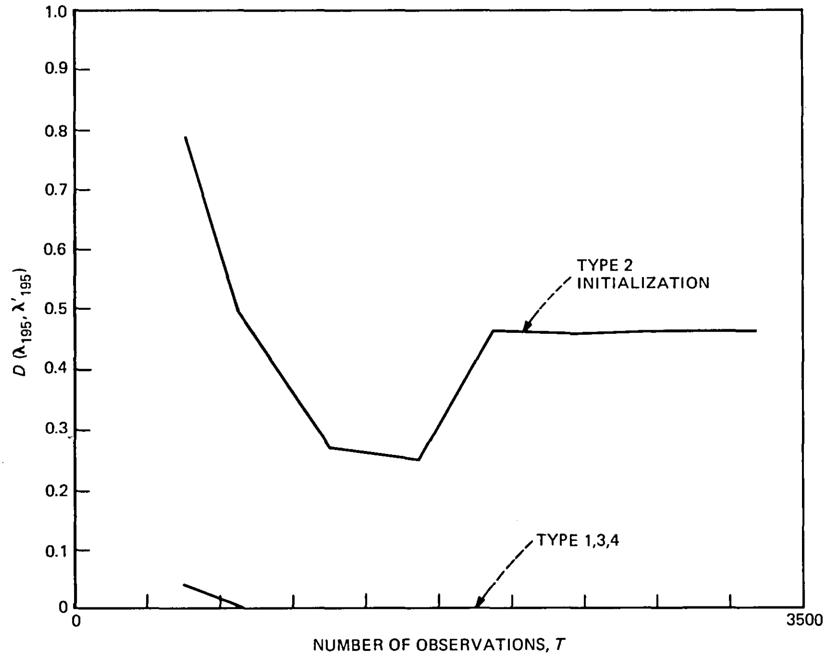


Fig. 5—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{195}.

with both known state-transition constraints and known state-observation constraints, so in the initial matrices **A** and **B**, those elements corresponding to prohibited transitions and impossible observations are set to null. Type 4 is the generating model itself. Type 4 is useful for studying the convergence properties of the reestimate algorithm itself, since the sequence is unlikely to display complications often observed in sequences that converge to a local optimum.

A set of curves showing the measured distances versus the number of observations, for the four types of initial guess, are plotted in Figs. 5, 6, 7, and 8, corresponding to λ_{195}, λ_{295}, λ_{395}, and λ_{595}, respectively. Figure 5 shows that for model λ_{195}, the model estimates with Type 1, 3, and 4 initial guesses quickly converge to the generating model λ_{195}, i.e., the distances became essentially 0. Note that λ_{195} has a highly constrained structure with high probability of staying in the current state. This, combined with the fact that **B** is non-overlapping, says that this source would most probably produce observation sequences in which the corresponding state sequence is well defined, and the duration of each state (as determined from the **A** matrix) is unlikely to differ from one another dramatically. Type 2 initial guesses maintain the same Markov chain structure, but with random transition

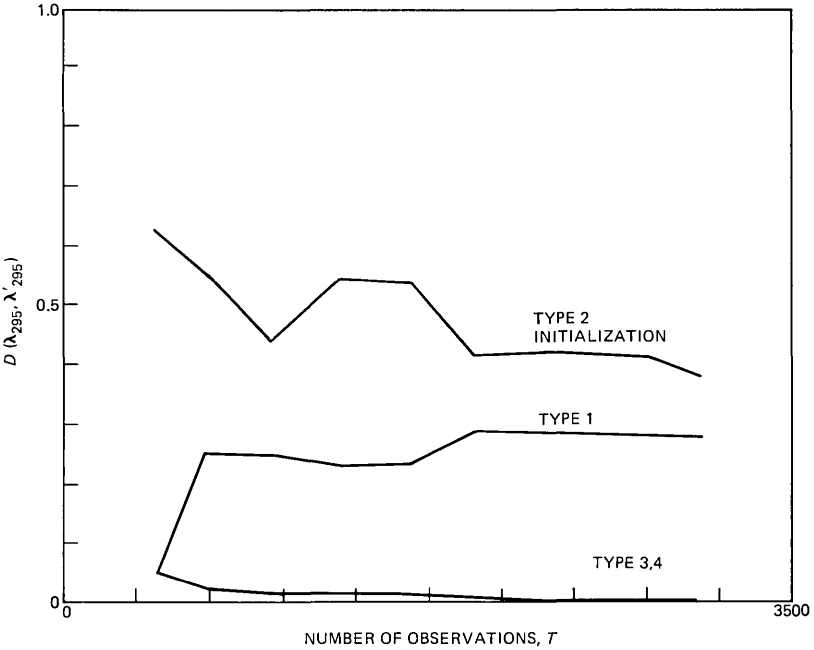


Fig. 6—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{295}.

probabilities different from the generating source. In fitting the observation sequence to a Type 2 initial estimate, the constrained **A** matrix is changed very little in the reestimation process, which instead mainly extends and modifies the **B** matrix to include, in one state, different observations that originally occurred in different states. Depending on the initial guess values, the resultant **B** matrix may be significantly overlapped, thereby leading to a significant distance from the generating source. Figure 5 shows that this analysis is indeed the case for model λ_{195}. With Type 3 initial guesses, the initial constraints in matrices **A** and **B** are retained through the reestimation process, and optimization of the **A** matrix is independent of that of the **B** matrix. Furthermore, optimization of the **B** matrix, in the current case, is carried out independently for each state, because with the initial constrained **B** matrix, the underlying state sequence is immediately known. Therefore, the results from using Type 4 initial guesses are virtually identical to the results from using Type 3 initial guesses, as shown in Figs. 5 through 8, for all models that were studied.

For λ_{295}, trends similar to those of Fig. 5 are observed and shown in Fig. 6, but some problems due to the allowed state-skipping transitions are observed for Type 1 initial parameter estimates. As explained above, estimated models based upon Type 2 initial guesses are at a

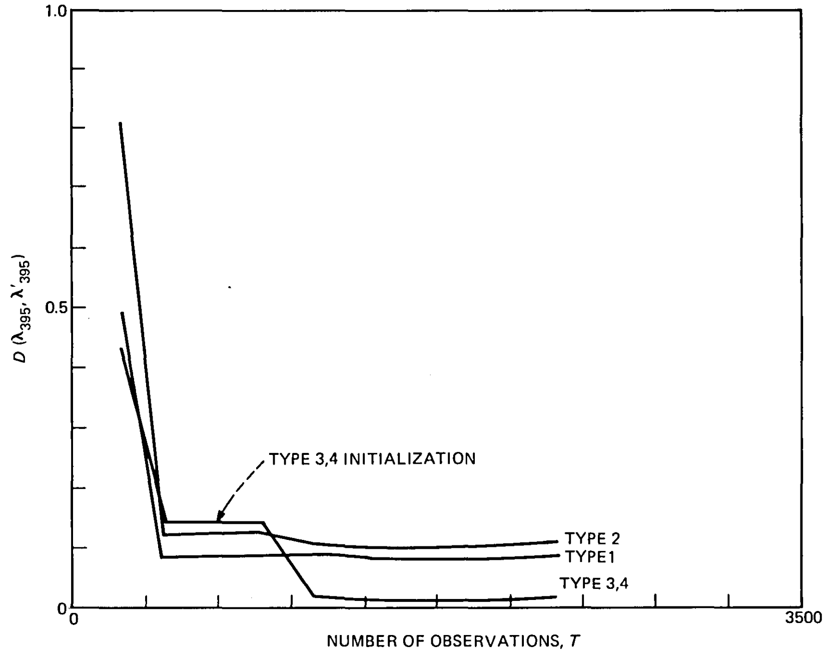


Fig. 7—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{395}.

significant distance from the generating source. With Type 1 initial guesses, the converged results are only slightly better than those with Type 2 guesses. Type 3 and 4 initial guesses lead to virtually the optimal estimate, i.e., models with essentially zero distance from the generator.

The results using model λ_{395} are shown in Fig. 7. Model λ_{395} has a state-transition structure similar to that of λ_{295}, but with different transition probabilities. The fact that $a_{22} = 0.2$ and $a_{44} = 0.1$ in λ_{395} makes it essentially a three-state model (i.e., two of the states are highly transient). Again, the dependence of model estimates upon the type of initial guess is similar to what is mentioned above, except the distances now are smaller than those obtained using λ_{295}. However, as seen in Fig. 7, when the total duration is small (i.e., 244 samples), the estimated models are at a significant distance from the generating source, regardless of the initial guess. This is because states 2 and 4 are not well represented in the observation sequences. The sudden drop of distance for Type 3 and 4 initial guesses at $T \simeq 1150$ samples indicates that the transient states of the generating model are sufficiently well represented for $T \geq 1150$, and with proper initial guesses, an estimate virtually identical to the generating source can be obtained.

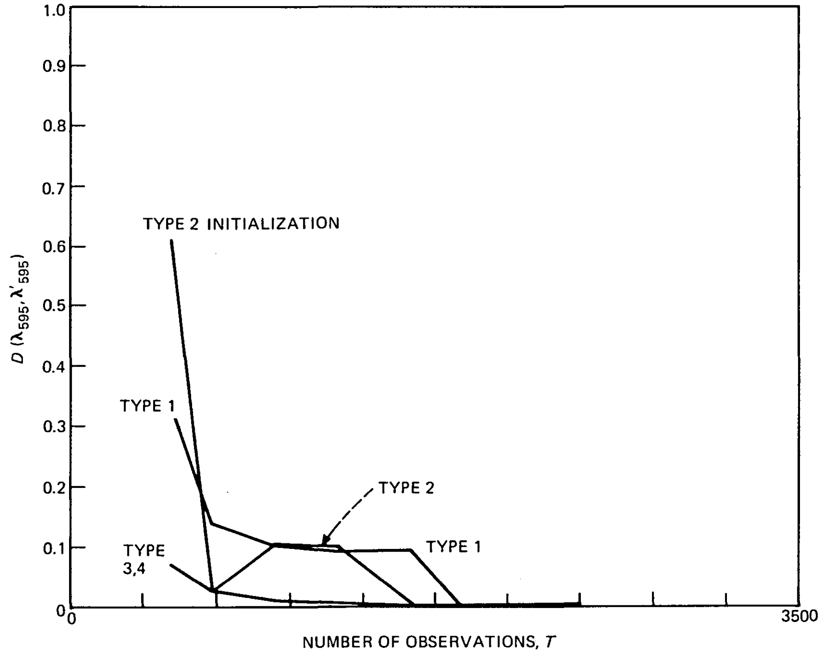


Fig. 8—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{595}.

The results for model λ_{595} are given in Fig. 8. Since source λ_{595} has an overlapping observation matrix **B**, many of the phenomena that occurred in λ_{195}, λ_{295}, and λ_{395} no longer appear. Indeed, as shown in Fig. 8, estimated models of nearly zero distance from the generating source have been obtained regardless of the initial guess, provided the observation sequences are sufficient in duration. The effects of initial guess are manifested only in the way the estimate converges as T grows.

Figures 5, 6, 7, and 8 not only provide results pertaining to the performance of model estimates and its relationship to model initialization as well as observation length, but also show the effectiveness of the distance measure of eq. (6) in measuring the dissimilarity between any pair of hidden Markov models.

IV. CONCLUSION

We have defined a probabilistic distance measure for hidden Markov models. The measure is consistent with the probabilistic modeling technique and can be efficiently evaluated through Monte Carlo procedures. The distance measure was employed in the study of relative parameter sensitivities as well as the relationship among model estimate, initial guess for the reestimation algorithm, and the observation sequence duration for discrete density hidden Markov models. Much of the behavior of hidden Markov models and the reestimated results have been observed through the use of such a distance measure. The study in turn confirms the effectiveness and reliability of the distance measure. Potential applications of the distance measure may include hidden Markov model selection as well as clustering.

REFERENCES

1. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," B.S.T.J., *62*, No. 4 (April 1983), pp. 1035–74.
2. L. E. Baum et al., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," Ann. Math. Statist., *41* (1970), pp. 164–71.
3. J. H. Conway and N. J. A. Sloane, "Voronoi Regions of Lattices, Second Moments of Polytypes, and Quantization," IEEE Trans. Inform. Theory, *IT-28* (March 1982), pp. 227–32.
4. Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. Commun., *COM-28* (January 1980), pp. 84–95.
5. S. Kullback, *Information Theory and Statistics*, New York: Wiley, 1958.
6. R. M. Gray et al., "Rate-Distortion Speech Coding With a Minimum Discrimination Information Distortion Measure," IEEE Trans. Inform. Theory, *IT-27*, No. 6 (November 1981), pp. 708–21.
7. T. Petrie, "Probabilistic Functions of Finite State Markov Chains," Ann. Math. Statist., *40*, No. 1 (1969), pp. 97–115.
8. L. R. Rabiner et al., unpublished work.
9. L. E. Baum and J. A. Eagon, "An Inequality With Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology," Bull. AMS, *73* (1967), pp. 360–3.
10. P. Billingsley, *Ergodic Theory and Information*, New York: Wiley, 1965.

AUTHORS

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

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D., 1967 (Electrical Engineering), The Massachusetts Institute of Technology; AT&T Bell Laboratories, 1962—. Presently, Mr. Rabiner is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983). Member, National Academy of Engineering, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.

AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

A Conditional Response Time of the M/M/1 Processor-Sharing Queue

By B. SENGUPTA and D. L. JAGERMAN*

(Manuscript received April 24, 1984)

In this paper we examine the distribution of the response time of an arriving customer conditioned on the number of customers present in an M/M/1 processor-sharing queue. We show that the *r*th moment of the distribution is a polynomial in the number of customers present, and obtain a recursion for its determination. The Laplace-Stieltjes transform of the conditional distribution is also obtained. We find an expansion in powers of the arrival rate, which permits accurate computation when the utilization is not too close to one, and also give an asymptotic expansion when the number of customers in the system is large. This permits assessment of response time in a heavily loaded system. We present numerical results using these methods and discuss their relative merits.

I. INTRODUCTION

The behavior of many computer systems can be approximated by the processor-sharing discipline. In this discipline, the server (CPU) operates at a rate of μ, and whenever i customers are present, each customer receives service at a rate of μ/i. We assume that the arrivals occur according to a Poisson process at a rate of λ and that each customer's service requirement is exponentially distributed.

Of interest to us is the response time conditioned on the number of customers seen by an arriving customer, including itself. The response time is the elapsed time between arrival and departure for a customer. In this paper, we show that the *r*th moment of the conditional response

* Authors are employees of AT&T Bell Laboratories.

time is a polynomial of degree r in the number of customers seen by the arrival, including itself. Further, we show several methods of obtaining the distribution of the conditional response time in the form of a Laplace-Stieltjes transform.

The results presented in this paper may be useful in the design of a computer or switching system. For example, the designer of such a system may ask, "What is the minimum number of customers that must be present in the system before the 95th percentile of the response time exceeds some prespecified threshold?" The answer to this question gives some indication of the performance of the system under overload. It may be instrumental in deciding whether any special overload control mechanism is needed. If an overload control mechanism is needed (i.e., the arrivals are blocked if a certain number of customers are in the system), the analysis of such a queue is straightforward and will not be discussed in this paper.

This problem has been studied in greater generality by Coffman et al.,[1] who obtain the waiting time distribution conditioned on the number seen by an arrival and the amount of service required by the arriving customer. However, it is difficult to obtain all the results of this paper directly from those in Ref. 1. Related work in this area was done by Sakata et al.,[2] who characterized a solution for the M/G/1 processor-sharing model. Recently, Ott[3] and Ramaswami[4] have provided methods for characterizing the unconditional distributions of the response time for the M/G/1 and GI/M/1 queues, respectively. The unconditional distribution of the response time for the M/M/1 queue has also been solved by Morrison.[5] Rege and Sengupta[6] have solved a version of the M/M/1 processor-sharing model, which includes multiprogramming.

In Section II of this paper we derive the conditional moments of the response time. In Section III, we obtain the Laplace-Stieltjes transform of the distribution of conditional response time. In Section IV, we obtain an expansion for the transform in powers of λ, which is useful for computation when λ is not large. In Section V, the asymptotic expansion is given for $x \to \infty$. This is especially useful for computation in a heavily loaded system. In Section VI, we discuss the numerical issues of this problem.

II. MOMENTS OF THE CONDITIONAL RESPONSE TIME

Let X denote the number of customers seen by an arrival, including itself, and let T be the response time of this arrival. Let $u(x, s)$ be the Laplace-Stieltjes transform of the distribution of the conditional response time, i.e., $u(x, s) = E(e^{-sT} | X = x)$ for $x = 1, 2, \cdots$. By conditioning on the first event (defined as the next arrival or departure, whichever occurs first), we obtain

$$u(x+1, s) = \left(\frac{\lambda+\mu}{\lambda+\mu+s}\right)\left(\frac{\lambda}{\lambda+\mu}u(x+2, s) + \frac{\mu}{\lambda+\mu}\,\frac{x}{x+1}u(x, s) + \frac{\mu}{\lambda+\mu}\,\frac{1}{x+1}\right) \quad (1)$$

and

$$u(1, s) = \left(\frac{\lambda+\mu}{\lambda+\mu+s}\right)\left(\frac{\lambda}{\lambda+\mu}u(2, s) + \frac{\mu}{\lambda+\mu}\right). \quad (2)$$

Let $\mu_r(x)$ be the rth moment of T conditioned on $X = x$, i.e., $\mu_r(x) = E(T^r \mid X = x)$. Then by taking the rth derivative of (1) and (2) with respect to s, multiplying by $(-1)^r$ and setting s to zero, we obtain

$$(x+1)\mu_r(x+2) - (1+b)(x+1)\mu_r(x+1) + bx\mu_r(x) = -r(x+1)\mu_{r-1}(x+1)/\lambda \quad (3)$$

for $x = 1, 2 \cdots$

$$\mu_r(2) - (1+b)\mu_r(1) = -r\mu_{r-1}(1)/\lambda \quad (4)$$

and

$$\mu_0(\cdot) = 1.$$

In eqs. (3) and (4), $b = \mu/\lambda$.

Theorem 1: The solution of (3) *and* (4) *has the form:*

$$\mu_r(x) = \sum_{j=0}^{r} a_{rj}x^j \quad \text{for} \quad x = 1, 2, \cdots; \qquad r = 1, 2, \cdots$$

in which the coefficients a_{rj} satisfy:

$$a_{rj} = \left\{\sum_{k=j}^{r-1} (-\, ra_{r-1,k}b_{k,j}/\lambda - a_{r,k+1}c_{k+1,j}) - ra_{r-1,j-1}/\lambda\right\} c_{jj}^{-1}$$

$$\textit{for} \quad j = r-1, \cdots, 1; \qquad r = 1, 2, \cdots$$

$$a_{r0} = \left\{\sum_{k=0}^{r-1} (ra_{r-1,k}/\lambda + (a_{r,k+1}(2^{k+1} - 1 - b))\right\} b^{-1}$$

$$\textit{for} \quad r = 1, 2, \cdots$$

$$a_{00} = 1$$

and

$$b_{kj} = \frac{(k+1)!}{j!(k-j+1)!},$$

$$c_{kj} = b_{kj}\left\{\left(\frac{k+j+1}{k+1}\right)2^{k-j} - (1+b)\right\}$$

$$\textit{for} \quad j = 1, \cdots, r; \qquad k = j, \cdots, r; \qquad r = 1, 2 \cdots.$$

Proof: We first write $\mu_r(x)$ as a power series of the form

$$\mu_r(x) = \sum_{j=0}^{\infty} a_{rj}x^j$$

and note that

$$\mu_r(x+m) = \sum_{j=0}^{\infty} x^j \sum_{k=j}^{\infty} a_{rk}k!m^{k-j}/(j!(k-j)!).$$

We substitute this in (3) and arrange the left- and right-hand sides as a power series in x. For $j > r$, we find that the terms containing x^j on the left-hand side vanish. For any $j > r$, the coefficient of x^j contains a_{rk} for $k \geq j$, and these coefficients can be chosen arbitrarily. It is, therefore, not inconsistent to choose $a_{rj} = 0$ whenever $j > r$. The rest of the theorem follows by equating coefficients of x^j on the left- and right-hand sides of (3) for $j = r, \cdots, 0$ and verifying that the boundary condition is satisfied. □

Corollary 1: For $r = 1$ and 2,

$$\mu_1(x) = \frac{(x+1)}{\lambda(2b-1)}$$

and

$$\mu_2(x) = \frac{2(x+1)\left(x + \dfrac{4b}{2b-1}\right)}{\lambda^2(2b-1)(3b-2)}$$

for

$$x = 1, 2, \cdots.$$

We observe that the mean conditional response time is a linear function of x. This particular result is true for the M/M/1 First Come First Served (FCFS) queue as well, for which $\mu_1(x) = x/\mu$. It is readily seen that the mean conditional response time for the processor-sharing queue is smaller than that for the FCFS queue if and only if

$$x - 1 > L,$$

where L is the mean number in the system. Thus, the processor-sharing queue has a smaller mean response time whenever the number of customers seen by an arrival (excluding itself) is *greater* than the mean number in the system.

This result parallels the known result about the mean response time conditioned on the service requirement (see Ref. 7). There, the processor-sharing queue has a smaller mean response time whenever the service requirement is *less* than the mean service time.

III. DISTRIBUTION OF THE CONDITIONAL RESPONSE TIME

The difference equation for the Laplace-Stieltjes transform of the response time distribution, $u(s, x)$, is repeated here for convenience:

$$(x-1)u(x) - a(x-1)u(x-1) + b(x-2)u(x-2) = -b, \qquad x \geq 2,$$

$$b = \frac{\mu}{\lambda}, \qquad a = b + 1 + \frac{s}{\lambda}. \tag{5}$$

The explicit indication of the dependence of $u(s, x)$ on s has been suppressed. An exact solution for the Laplace transform suitable for numerical inversion will be obtained.[8] Only one boundary condition is given explicitly in (5), namely, for $x = 2$, one has

$$u(2) - au(1) = -b; \tag{6}$$

however, another boundary condition is available through the requirement that $u(s, x)$ be a transform as a function of s. This, of course, will be satisfied by the solution to be obtained.

Designate by Lu the left-hand side of (5) so that the difference equation to be solved is

$$Lu = -b; \tag{7}$$

this will be referred to as the *complete* equation. A function $v(x)$ will now be found satisfying

$$Lv = 0. \tag{8}$$

Equation (8) is referred to as the *homogeneous* equation and $v(x)$ as a complementary solution. Laplace's method will be used to solve (8).[9]

Assume a representation of $v(x)$ of the form

$$v(x) = \int \tau^{x-1} g(\tau) d\tau \tag{9}$$

in which the function $g(\tau)$ and a path of integration in the complex τ-

plane are to be chosen. Substitution into (8) and subsequent integration by parts yields

$$Lv = (\tau^2 - a\tau + b)\tau^x g(\tau)\,| \\ + \int \tau^{x-1}[(\tau^3 - a\tau^2 + b\tau)\dot{g}(\tau) + \tau^2 g(\tau)]d\tau. \quad (10)$$

The vertical bar designates evaluation of the concomitant on the path to be chosen, and the dot indicates $d/d\tau$. To satisfy (8) the differential equation

$$(\tau^2 - a\tau + b)\dot{g}(\tau) + \tau g(\tau) = 0 \quad (11)$$

is to be solved for $g(\tau)$. Let the roots of

$$\tau^2 - a\tau + b = 0 \quad (12)$$

be

$$\gamma = \frac{a - \sqrt{a^2 - 4b}}{2}, \qquad \gamma_1 = \frac{a + \sqrt{a^2 - 4b}}{2} \quad (13)$$

and let

$$\alpha = \frac{\gamma_1}{\gamma_1 - \gamma} = \frac{1}{2}\left(1 + \frac{a}{\sqrt{a^2 - 4b}}\right); \quad (14)$$

then the solution of (11) may be written as

$$g(\tau) = (\gamma - \tau)^{\alpha-1}(\gamma_1 - \tau)^{-\alpha}. \quad (15)$$

Since, for real s, $0 < \gamma < \gamma_1$, the path of integration in (9) is chosen as the segment $(0, \gamma)$ of the real axis; for this choice, the concomitant term of (10) vanishes and, hence, (8) is satisfied. Thus, one has

$$v(x) = \int_0^\gamma \tau^{x-1}(\gamma - \tau)^{\alpha-1}(\gamma_1 - \tau)^{-\alpha}d\tau. \quad (16)$$

Setting $z = \gamma^2/b = \gamma/\gamma_1$ and changing the variable of integration allows $v(x)$ to be written in the form

$$v(x) = \gamma^{x-1}z^\alpha \int_0^1 \tau^{x-1}(1 - \tau)^{\alpha-1}(1 - \tau z)^{-\alpha}d\tau. \quad (17)$$

Comparison of (17) with the integral form of the hypergeometric function, $F(a, b; c; z)$ shows that (see Ref. 10)

$$v(x) = \gamma^{x-1}\frac{\Gamma(\alpha)\Gamma(x)}{\Gamma(\alpha + x)}z^\alpha F(\alpha, x; \alpha + x; z). \quad (18)$$

This is particularly advantageous since one has

$$F(\alpha, x; \alpha + x; z) = \sum_{j=0}^{\infty} \frac{(\alpha)_j (x)_j z^j}{(\alpha + x)_j j!}, \tag{19}$$

in which

$$(a)_0 = 1, \qquad (a)_j = a(a + 1) \cdots (a + j - 1), \qquad j \geq 1. \tag{20}$$

It may be observed that the series of (19) converges absolutely for $|z| < 1$, that is, for $\gamma^2 < b$ which holds.

The computation of F in (19) may be simply carried out by the recursions

$$\begin{aligned} F_0 &= 1, \qquad F_j = F_{j-1} + T_j, \\ T_0 &= 1, \qquad T_j = \frac{(\alpha + j - 1)(x + j - 1)}{(\alpha + x + j - 1)j} z T_{j-1}. \end{aligned} \tag{21}$$

Similarly, since

$$\frac{\Gamma(\alpha)\Gamma(x)}{\Gamma(\alpha + x)} = \frac{(x - 1)!}{(\alpha)_x}, \tag{22}$$

this too may be easily computed recursively. It may be considered, therefore, that the computations that will be needed in the inversion procedure to be used may be conveniently performed for the function $v(x)$.

To solve the complete eq. (7), its order will be depressed. It is convenient to write it in the form

$$Lu = (x + 1)u(x + 2) - a(x + 1)u(x + 1) + bxu(x) = -b. \tag{23}$$

Let

$$u(x) = v(x)\tau(x); \tag{24}$$

then

$$Lu = L(v\tau) = \tag{25}$$

$$(x + 1)v(x + 2)\tau(x + 2) - a(x + 1)v(x + 1)\tau(x + 1) + bxv(x)\tau(x).$$

Use of the formulae

$$\begin{aligned} \tau(x + 1) &= \tau(x) + \Delta\tau(x), \\ \tau(x + 2) &= \tau(x) + 2\Delta\tau(x) + \Delta^2\tau(x) \end{aligned} \tag{26}$$

in (25) yields

$$\begin{aligned} Lu = {} & (x + 1)v(x + 2)\Delta^2\tau \\ & + [2(x + 1)v(x + 2) - a(x + 1)v(x + 1)]\Delta\tau. \end{aligned} \tag{27}$$

Setting

$$w(x) = \Delta\tau(x) \tag{28}$$

in (27) now provides the first-order equation

$$w(x+1) - R(x)w(x) = -\frac{b}{(x+1)v(x+2)},$$

$$R(x) = a\frac{v(x+1)}{v(x+2)} - 1. \tag{29}$$

Direct substitution verifies that

$$w(x) = \sum_{j=1}^{\infty} \frac{b}{(x+j)v(x+j+1)R(x)R(x+1)\cdots R(x+j-1)} \tag{30}$$

is a solution of (29).

To show that $w(x)$ converges, one has

$$(1-\tau z)^{-\alpha} \sim (1-z)^{-\alpha}, \qquad \tau \to 1-; \tag{31}$$

hence, from (17),

$$v(x) \sim \gamma^{x-1}\left(\frac{z}{1-z}\right)^{\alpha} \int_0^1 \tau^{x-1}(1-\tau)^{\alpha-1}d\tau, \tag{32}$$

$$v(x) \sim \gamma^{x-1}\left(\frac{z}{1-z}\right)^{\alpha} \frac{\Gamma(\alpha)\Gamma(x)}{\Gamma(\alpha+x)}, \qquad x \to \infty. \tag{33}$$

Further, since

$$\frac{\Gamma(x)}{\Gamma(\alpha+x)} \sim x^{-\alpha}, \qquad x \to \infty, \tag{34}$$

one has

$$v(x) \sim \gamma^{x-1}\left(\frac{z}{1-z}\right)^{\alpha} \Gamma(\alpha)x^{-\alpha}, \qquad x \to \infty; \tag{35}$$

thus

$$R(x) \to \frac{a}{\gamma} - 1, \qquad x \to \infty. \tag{36}$$

Evaluation of the jth term comprising $w(x)$ now shows that it is of the order

$$j^{\alpha-1}\left(\frac{\gamma}{b}\right)^{j}. \tag{37}$$

Since $\gamma\gamma_1 = b$ and $\gamma_1 > \gamma$, one has $\gamma < \sqrt{b}$, thus $\gamma < b$ if $b > 1$, that is, if the offered load $b^{-1} = \lambda/\mu < 1$. In this case the series for $w(x)$ converges.

A one-parameter family of solutions of (7) obtained from (28) and (24) is

$$u(x) = v(x)\left[\mathrm{D} + \sum_{j=1}^{x-1} w(j)\right], \tag{38}$$

in which D is a constant that is determined by use of the boundary condition (6). The final result is

$$u(x) = v(x)\left[\sum_{j=1}^{x-1} w(j) - \frac{b + w(1)v(2)}{v(2) - av(1)}\right]. \tag{39}$$

In order that $u(x)/s$ be a Laplace transform, one must have

$$\lim_{s\to\infty} \frac{u(x)}{s} = 0, \tag{40}$$

which is verified by (39). Accordingly, (40) is the second boundary condition required to specify a unique solution of (7). That this is true follows from an examination of the complete solution of (7), which will not be done here.

IV. PERTURBATION IN λ

For this purpose (5) is written in the form

$$\lambda(x + 1)u(x + 2) - (\mu + s + \lambda)(x + 1)u(x + 1) + \mu x u(x) = -\mu, \tag{41}$$

and the boundary condition in (6) takes the form

$$\lambda u(2) - (\mu + s + \lambda)u(1) = -\mu. \tag{42}$$

Fortunately this constitutes a singular perturbation. Writing $u(x)$ in the form

$$u(x) = \sum_{j=0}^{\infty} u_j(x)\lambda^j \tag{43}$$

and substituting into (41), (42) yields the equations

$$(x + 1)u_0(x + 1) - \left(1 + \frac{s}{\mu}\right)^{-1} x u_0(x) = \left(1 + \frac{s}{\mu}\right)^{-1},$$

$$u_0(1) = \left(1 + \frac{s}{\mu}\right)^{-1}, \tag{44}$$

and

$$(x+1)u_j(x+1) - \left(1+\frac{s}{\mu}\right)^{-1} xu_j(x)$$

$$= \frac{1}{\mu}\left(1+\frac{s}{\mu}\right)^{-1} (x+1)\Delta u_{j-1}(x+1),$$

$$u_j(1) = \frac{1}{\mu}\left(1+\frac{s}{\mu}\right)^{-1} \Delta u_{j-1}(1). \tag{45}$$

These are first-order equations for $u_0(x)$ and $u_j(x)$, which present no difficulty of solution. One obtains

$$u_0(x) = \frac{\mu}{sx}\left[1 - \left(1+\frac{s}{\mu}\right)^{-x}\right],$$

$$u_j(x) = \frac{1}{\mu x}\sum_{\ell=1}^{x}\left(1+\frac{s}{\mu}\right)^{\ell-x-1} \ell\Delta u_{j-1}(\ell), \qquad j \geq 1. \tag{46}$$

V. ASYMPTOTICS FOR $x \rightarrow \infty$

The derivation of the asymptotic expansion of $u(x)$ depends on the operational method of Boole,[11] Jagerman,[12] and Milne Thomson (see Ref. 9); it is somewhat involved, so the details will be omitted but the results may be easily stated. One may write

$$u(x) \sim \sum_{j=1}^{\infty} \frac{\alpha_j}{x(x+1)\cdots(x+j-1)}, \tag{47}$$

in which the coefficients are given by

$$\alpha_1 = \frac{\lambda b}{s}, \qquad \alpha_2 = -\frac{\lambda^2 b}{s^2},$$

$$\alpha_3 = -\frac{\lambda^3 b(b-2)}{s^3}, \qquad \alpha_4 = -\frac{\lambda^4 b(b-2)(2b-3)}{s^4} - \frac{2\lambda^3 b^2}{s^3}, \tag{48}$$

and, in general, by the recursion

$$\alpha_j = -\frac{\lambda}{s}((j-2)(2-\alpha)+1)\alpha_{j-1} + \frac{\lambda}{s}(j-2)^2\alpha_{j-2}, \qquad j \geq 2. \tag{49}$$

Let $F(\tau, x)$ be the complementary distribution corresponding to $u(x)$; then

$$F(\tau, x) \sim 1 - \frac{\mu\tau}{x}\left[1 - \frac{1}{2}\frac{\lambda\tau}{x+1} - \frac{1}{6}\frac{\lambda^2\tau^2(b-2)}{(x+1)(x+2)} - \frac{\frac{1}{24}\lambda^3\tau^3(b-2)(2b-3) + \frac{1}{3}\lambda^2\tau^2 b}{(x+1)(x+2)(x+3)}\right], \qquad x \to \infty. \qquad (50)$$

VI. NUMERICAL RESULTS

Our paper was concerned with exact answers for the moments (Section II), answers in the form of Laplace-Stieltjes transforms (Sections III and IV) for the distribution function, and asymptotic answers (Section V) for large x. One issue that we were concerned with was the applicability of these methods for ranges of parameter values of the problem.

In Table I, we present the first and second moments of the sojourn time calculated by three different methods (Sections II, III, and IV). The value of μ was chosen to be 1, x was taken to be 3, and λ was varied from 0.2 to 0.9. We numerically inverted the transforms obtained in Sections III and IV by the method of Ref. 8 and computed the moments from the distribution. As can be seen from the results, the results from the two distributions are accurate for the first moment for moderate and low value of λ. The method of Section III seems to be slightly better than the method of perturbations for a large value of λ. For the second moment, the results seem to be slightly less accurate. We state that we calculate the moments by numerically integrating the complementary distribution. Thus, truncation will cause the calculated moments to be underestimated. This fact is borne out in all the examples. In Fig. 1, we show the complementary distribution of the sojourn time by inverting the transforms from eq. (39).

In Table II, we show the accuracy of the asymptotic results presented in Section V. Here we calculate the complementary distribution when $\lambda = 0.5$, $\mu = 1$, and $x = 10$. As can be seen, there is good correspondence between the perturbation method (which uses the numerical inversion technique of Ref. 8) and the asymptotic solution.

Table I—Numerical results for the first and second moments ($x = 3$, $\mu = 1$)

	Mean = $\mu_1(x)$			Second Moment = $\mu_2(x)$		
λ	Exact (Section II)	From eq. (39)	From eq. (46)	Exact (Section II)	From eq. (39)	From eq. (46)
0.2	2.222	2.216	2.216	8.927	8.659	8.659
0.5	2.667	2.636	2.636	15.111	13.758	13.766
0.9	3.636	3.619	3.384	40.220	35.664	25.243

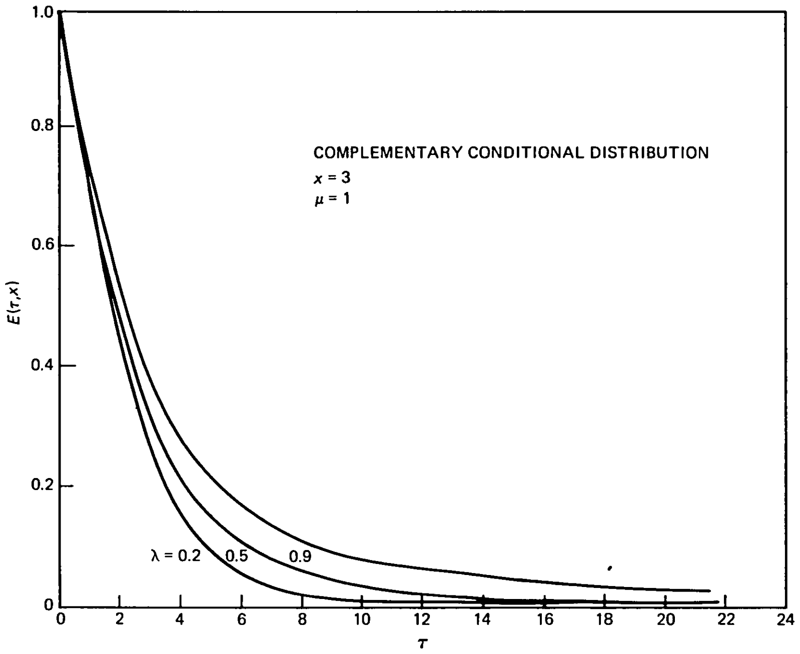


Fig. 1—Complementary conditional distribution of the sojourn time.

Table II—Numerical results for the asymptotic solution ($\lambda = 0.5$, $\mu = 1$, $x = 10$)

	$F(\tau, x) = P(T > \tau \mid X = x)$	
τ	From Asymptotic Solution (50)	From eq. (Perturbation) (46)
1.0	0.9023	0.9023
2.0	0.8092	0.8092
3.0	0.7207	0.7209
4.0	0.6370	0.6376
5.0	0.5580	0.5596
6.0	0.4839	0.4874
7.0	0.4147	0.4215
8.0	0.3504	0.3621
9.0	0.2912	0.3094
10.0	0.2370	0.2630

In conclusion, we would recommend the method of Section III for small values of x; the asymptotic solution of Section V for large values of x; and the perturbation method of Section IV, where x takes a wide range of values and when λ is not large.

REFERENCES

1. E. G. Coffman, R. R. Muntz, and H. Trotter, "Waiting Time Distributions for Processor Sharing Systems," JACM, *17,* No. 1 (January 1970), pp. 123–30.
2. M. Sakata, S. Noguchi, and J. Oizumi, "Analysis of a Processor Shared Queueing Model for Time Sharing Systems," Proc. Second Hawaii Int. Conf. on Systems Sciences, University of Hawaii (1969), pp. 625–7.
3. T. J. Ott, "The Sojourn Time in an M/G/1 Queue With Processor Sharing," J. App. Prob., *21* (June 1984), pp. 360–78.
4. V. Ramaswami, "Sojourn Time in the GI/M/1 Queue with Processor Sharing," J. App. Prob., *21* (June 1984), pp. 445–50.
5. J. A. Morrison, "Response Time Distribution for a Processor-Sharing System," SIAM J. App. Math.
6. K. M. Rege and B. Sengupta, "Sojourn Time Distribution in a Multiprogrammed Computer System," AT&T Tech. J., *64,* No. 5 (May–June 1985).
7. L. Kleinrock, *Queueing Systems, Vol. II: Computer Applications,* New York: Wiley Interscience, 1976.
8. D. L. Jagerman, "An Inversion Technique for the Laplace Transform," B.S.T.J., *61,* No. 8 (October 1982), pp. 1995–2002.
9. L. M. Milne Thomson, *Calculus of Finite Differences,* London: MacMillan and Co., 1933.
10. *Bateman Manuscript Project, Vol. 1,* New York: McGraw Hill, 1953.
11. G. Boole, "A Treatise on the Calculus of Finite Differences," New York: G. E. Stechert & Co., 1931.
12. D. L. Jagerman, *Difference Equations With Applications to Stochastic Models,* New York: Marcel Dekker (to be published).

AUTHORS

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

Bhaskar Sengupta, B. Tech. (Electrical Engineering), 1965, I.I.T. Kharagpur, Eng. Sc.D. (Operations Research), 1976, Columbia University; AT&T Bell Laboratories, 1981—. Mr. Sengupta has worked in IBM and Service Bureau Company and was an Assistant Professor at the State University of New York at Stony Brook. He was also a consultant to Turner Construction Company in New York. At AT&T Bell Laboratories he works on performance problems for communication, computer, and manufacturing systems.

AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

A Study on the Ability to Automatically Recognize Telephone-Quality Speech From Large Customer Populations

By J. G. WILPON*

(Manuscript received August 7, 1984)

To ascertain whether a speaker-independent word recognition system, using current technology, could function in normal telephone environments, it was necessary to conduct a study under such real-world conditions. Such an experiment was described by Wilpon and Rabiner (1983), in which telephone customers, speaking under ordinary telephone conditions, in Portland, Maine, were asked to speak their telephone number as a sequence of isolated digits. For each customer a maximum of four digits were obtained. The results from that study were very encouraging and led to further improvements in our recognition systems. To further study the feasibility of implementing speech recognition systems for general use over the telephone network, another field study was initiated. In this test, spoken seven-digit telephone numbers were obtained from a large number of telephone customers over a variety of transmission facilities in Baton Rouge, Louisiana. This paper presents the results of several recognition experiments performed on this database. Experiments were also carried out quantifying the robustness of template sets created in Portland, Baton Rouge, and under laboratory conditions in our Murray Hill, New Jersey, laboratory. Finally, a recognition system that incorporates syntactic information available in a seven-digit telephone is discussed. Our tests indicate a number of distinct real-world problems that must be considered when implementing a speech recognition system for widespread use. A discussion of the overall results and the implications for future research will be given.

*AT&T Bell Laboratories.

I. INTRODUCTION

The development of a speaker-independent speech recognition system that performs well over dialed-up telephone lines has been a goal of AT&T Bell Laboratories for close to a decade.[1-8] However, until recently all evaluations of our recognition systems have been based on laboratory recording conditions. These conditions typically consisted of cooperative subjects using local dialed-up lines over a Private Branch Exchange (PBX). Peak signal-to-average noise ratios under these conditions generally ranged from 40 to 60 dB. Using such local switched lines, the performance of the speech recognition algorithms tested was found to be quite good for a wide range of vocabulary sizes and complexities and for a wide range of talkers.

An earlier effort was made to test the viability of our speaker-independent, isolated word recognition systems on a very large telephone customer population.[8] The task was conducted under "real world" conditions, i.e., asking telephone customers to speak their telephone numbers in a home environment over randomly dialed-up lines in Portland, Maine (PO). Under these conditions, signal-to-noise ratios (s/n) of between 8 dB and 60 dB were encountered. The results of this study yielded a recognition accuracy of 93.1 percent. While this was not as high an accuracy as was achieved in the laboratory,[2] given the transmission medium and the problems associated with obtaining isolated digit strings over standard telephone lines, these results were extremely encouraging.

There were several other shortcomings associated with our previous study. First, the wide variety of transmission and switching conditions made it very difficult to detect the spoken words automatically. Second, for privacy, our database consisted of at most the last four digits of the customer's telephone number. Because of this, parts of the first digit recorded were sometimes deleted. (The digitization of the input speech had to be initiated by a site observer after the first three digits were spoken. In some cases the observer was not quick enough to start the recording procedure before the fourth digit was spoken.) Third, about 50 percent of the speech data available from recording was thrown away, either because it contained some connected digit strings or the background conditions were too severe.

Another problem that existed in our earlier study was getting casual telephone customers to speak their phone number as a sequence of isolated digits. This was related to human factors issues, that is, people do not normally speak in an isolated word format.

As a result of the problems that were encountered during our initial exercise in the "real world," we found that we were testing our recognition systems on only a small percentage of all the speech data to which we had access. The purpose of this paper is to describe a new data collection exercise that was carried out to more accurately determine our speech recognition system's capabilities over randomly dialed-up telephone lines. Over a two-week period we recorded all customer information from approximately 7400 callers. No calls were eliminated, and all seven digits were recorded.

The database was collected over randomly dialed-up telephone lines at an AT&T centralized switching office in Baton Rouge, Louisiana (BR). The customers that participated would normally speak their telephone number to an operator. That is, the subjects were performing a task that they had done before, except that now input was to be given in an isolated fashion. Special-purpose hardware[9] was attached to one operator console, which automatically answered a call and asked the customer to speak his phone number as a series of isolated digits. The hardware also cataloged the caller's transaction, and digitized and stored the customer's speech on magnetic tape.

There are several very important issues that need to be studied before speech recognition can be made available to large telephone user populations. The most important issue is end-to-end system recognition accuracy. That is, if over time N calls are received by the system and must be handled, what is the percentage of the N calls that will be able to go through the system automatically without any failures? Such failures include the caller hanging up, word endpoint problems, isolated input problems, and the possibility that a human operator would have to intervene during the course of the transaction (e.g., if the customer misunderstood the instructions). These issues are examined in detail within the text of this paper.

Another issue that will be discussed is the robustness of speaker-independent templates created in one recording environment, using one set of talkers, and tested under different conditions with new sets of talkers. In past recognition studies, training data and testing data were collected under laboratory conditions in our Murray Hill, N.J., (MH) laboratory.[1-7] The subjects for these studies were all native speakers of American English mostly from the New York metropolitan area. In our Portland study, the speech data obtained were tested against a speaker-independent template set created from laboratory speech data. The results indicated that the MH template set was inadequate for recognizing speech from Portland customers. With the addition of the Baton Rouge database, more experiments were carried out using speech data from BR, PO and MH. All possible combinations of template sets and testing conditions were tried and the results show the Baton Rouge template set to be quite robust for a wide range of talkers and over a wide range of transmission mediums.

Although past research has shown that isolated word recognition systems perform adequately, the power of speech recognition lies in its ability to perform a given task reliably, i.e., the word recognizer should be embedded within a larger system. The task can usually be specified as a set of simple rules that define the task syntax. The syntax is able to limit the possible recognition sequences at each point in the transaction. Several task-oriented systems have been described in early work, for example, a voice-controlled repertory dialer system[10] and a directory listing retrieval system.[11] For each of these systems the addition of syntactic constraints greatly increased recognition performance.

Since past studies have shown the additional syntactic information to be useful, a system was constructed that incorporated knowledge about our task, i.e., the speaking of a seven-digit telephone number as a series of isolated digits. A full description of the system syntax and results will be presented.

In Section II, we briefly review the results obtained from the Portland study. Section III gives a description of the recording procedure used to obtain data in Baton Rouge. Section IV discusses the composition of the BR database. In Section V, we review some recent advances in speech recognition, which apply to our study. In Section VI we present the results from a series of recognition experiments performed on the BR database. The issue of template robustness is discussed in Section VII. A discussion of the overall results and their implications is given in Section VIII.

II. REVIEW OF PORTLAND DATA COLLECTION EXPERIMENT

Recordings were made at an AT&T switching office in Portland, Maine.[8] A prerecorded spoken message (a prompt) was given to each customer requesting that he speak his telephone number as a sequence of isolated digits. For reasons of customer privacy we recorded only the last four digits of the telephone number. As each of the digits was spoken, a site observer entered the digits on a keyboard. The observer determined whether the digit sequence was spoken in an isolated format (i.e., spoken with sufficient pauses between words). If not, the observer initiated another prerecorded spoken message (a reprompt) requesting the user to repeat his number with a longer pause between digits. If the observer decided that the final speech was unacceptable (either because it was spoken in a connected manner or because of unacceptably poor telephone line conditions), a reject code was entered and the entire procedure was terminated for the current call.

The recordings were bandpass filtered from 100 Hz to 3200 Hz, sampled at a 6.67-kHz rate, and then digitally transmitted to our laboratory in Murray Hill, N.J., for analysis. The log energy of the waveform was displayed to another observer, along with the automatically determined sets of endpoints indicating where in the recording interval the isolated words could be found. At this point the second observer had the option of modifying any or all sets of endpoints computed or eliminating any digit from the string. The segmented speech was then entered into a database for later examination. Using this procedure 11,035 digits from 3100 customers were recorded over a 23-day period.

Using a 3900 token subset of the PO speech data to train the recognizer, a 30-template-per-word reference set was created. (Several different template sets were tested in the PO study. The results presented here are for that template set that yielded the highest recognition accuracy.) When this set was tested against the full 11,035-digit database, a recognition accuracy of 93.1 percent was obtained.

There were several problems that occurred during the recording phase. These were classified as being in one of two groups. The first group consisted of problems associated with the telephone transmission conditions, e.g., loud static noises—probably caused by atmospheric disturbances, pops and/or clicks (switching transients), loud tones (mostly carrier frequency tones at 2600 Hz), and loud broadband "humming" noises (probably caused by a missing ground connection somewhere in the transmission path). Resulting peak signal-to-noise ratios varied from as little as 8 dB to as much as 60 dB. The second group consisted of problems related to the talker and the environment in which he or she spoke. These included nonisolation of speech (i.e., the digits were connected) and the presence of extraneous background speech. Most of these failures were severe enough to warrant elimination of the customer's speech from the database. This occurred for 47 percent of all calls available for processing.

As a consequence of the Portland study, several areas for improving recognition performance were discovered. Subsequently, additional research in speech endpoint detection algorithms[12] and clustering algorithms[13] (i.e., template generation procedures) was carried out. Results from this research have been applied throughout our BR study (see Section V).

III. RECORDING PROCEDURE USED IN BATON ROUGE

Figure 1 shows a block diagram of the overall recording setup used in this study. All recordings were made at an AT&T switching office in Baton Rouge, La. To record customer data in an efficient manner special-purpose hardware and control software were required. The hardware included an MC68000 controller, a terminal, a 7-1/2 inch magnetic tape unit, A/D and D/A converters, a cartridge tape unit, and signal conditioning circuitry. This hardware was attached to a dedicated operator console, a full description of which is given in Pirz

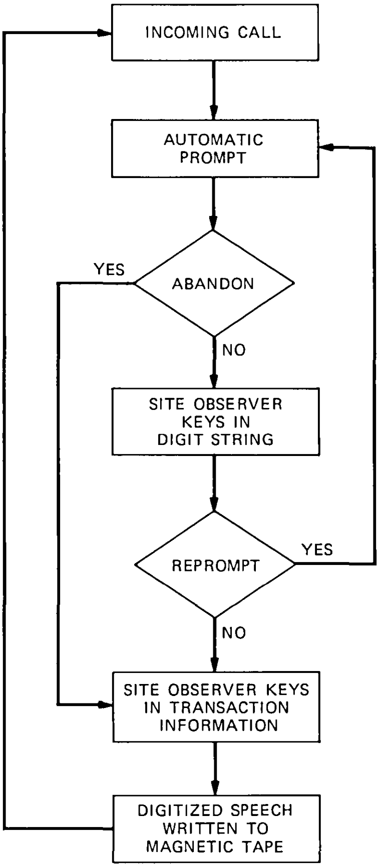

INCOMING CALL
AUTOMATIC PROMPT
YES
ABANDON
NO
SITE OBSERVER KEYS IN DIGIT STRING
REPROMPT
YES
NO
SITE OBSERVER KEYS IN TRANSACTION INFORMATION
DIGITIZED SPEECH WRITTEN TO MAGNETIC TAPE

Fig. 1—Block diagram of overall recording system used in the Baton Rouge study.

and Bauer.[9] The sequence of events to record a single customer's speech input was as follows:

1. An incoming customer call was automatically answered by the Special-Purpose Hardware (SPH). A prerecorded prompt was then played to the customer requesting that he speak his seven-digit telephone number as a sequence of isolated digits. As the digits were spoken, a Site Observer (SO) keyed in the identity of the spoken digit string. Also, as the customer spoke, the SPH digitized the speech at a 6.67-kHz rate with appropriate filtering applied.

2. Once the customer finished speaking, the SO made a judgment as to whether the speech was spoken in an isolated format (i.e., with sufficient pauses between words). If not, the SO would initiate a reprompt requesting the customer to repeat his number with longer pauses between words.

3. After the customer completed his task, he or she was given a prerecorded "Thank you" message. The SO then entered an ASCII character string indicating any comments about the talker, such as sex, ability to follow instructions, etc.

4. After the above steps were completed, the digitized speech was written out to the magnetic tape unit. Appropriate header information containing the identity of the input speech string, date and time of utterance, and any comments entered by the SO was also recorded on tape. If a reprompt had to be made, both the original and reprompted speech were saved.

It should be noted that a significant number of customers abandoned their call without speaking their phone number. All abandoned calls were cataloged and will be discussed later.

Using this procedure we recorded data from 7373 subjects (on 33 magnetic tapes) over a two-week period. After the data collection was completed, the speech was read from the magnetic tapes into a Data General MV8000 minicomputer where all further analysis was performed.

IV. COMPOSITION OF FINAL DATABASE

Recordings were made for an average of six hours a day, five days a week, for two weeks. Tables I and II show a detailed analysis of the final telephone customer database. Data were collected from a total of

Table I—Statistics on total number of calls handled in BR study

	Number of Callers	Percent
(1) Total callers	7373	100
(2) Abandoned calls	1468	20
(3) Net total calls (1–2)	5905	80
(4) Operator intervention	2301	31
(5) Unidentifiable calls	518	7
(6) > 7 digits spoken	269	4
(7) Processable calls (3–4–5–6)	2817	38

Table II—Sex makeup of processed calls from BR data

	Number of Calls	Percent
Net total calls	5905	100
Adult male	2137	36
Adult female	3524	60
Children	168	3
Unclassifiable	76	1

7373 callers. Of this total, 1468, or 20 percent were callers that abandoned their calls, and therefore did not enter any speech data. In these cases the caller hung up before beginning the recording task. This leaves a net total of 5905 calls that yielded some speech output. Of the remaining callers, 2137 (36 percent) were adult males, 3524 (60 percent) were adult females, 168 (3 percent) were children, and 76 (1 percent) were unclassifiable. Of the 5905 useful calls, 2301 or 31 percent required the telephone operator to cut in during the middle of the recording transaction. In these cases, the user got confused about the task he was to perform or simply did not want to cooperate. Since the caller had to supply his telephone number to complete his telephone call, the operator had to intervene. Generally, the user gave his phone number in a continuous fashion to the operator. Therefore, no useful isolated data could be extracted from these callers. There were several calls (518, or 7 percent) where the SO and later another observer could not understand one or more of the words that were spoken and therefore could not tag them correctly. These were caused either by very bad transmission noises or a very pronounced accent. For another 269 calls (4 percent), the customer spoke more than seven digits.

If we take into account all the calls that had some problems associated with them we would be left with a total of 2817 calls (38 percent) that were "processable," i.e., these calls contained only spoken digits. Therefore, an automatic procedure could be devised to first find the spoken words (endpointing) and second perform recognition on those words. All further discussion of the BR database will refer to this data set.

One problem that existed in our earlier recognition experiment of telephone-quality speech was the inability of the prompts to get the callers to speak in an isolated format.[8] This problem still existed in the BR study. Therefore, we decided to segment the database into two sets—one where all the digits were spoken in isolation, and another that contained those calls with any connected digits. Of the 2817 processable calls, 1837 contained only isolated digits, and 980 contained some connected digits.

The recording hardware had memory for at most a 15-second utterance. Some callers paused so long in between digits that they simply ran out of time. Therefore, we further classified the calls on the basis of whether all seven digits were present. The reason for this classification was that if it were known a priori that exactly seven digits were present, we could devise a procedure that recognizes them more accurately. Of the 1837 calls containing only isolated digits, 1634 (89 percent) consisted of all seven digits. Of the 980 calls containing some connected speech, only three calls had fewer than seven digits.

V. REVIEW OF RECOGNITION SYSTEM IMPROVEMENTS

5.1 Endpoint detector improvements

Before evaluating this telephone-quality speech database, several issues had to be addressed. One issue was the detection of speech within some time interval. In our previous study,[8] we used an endpointing algorithm developed by Lamel et al.[14] This technique had proved quite robust in detecting speech over local dialed-up telephone lines. However, endpoint detection becomes a much more difficult problem when the transmission system is corrupted by the many noises found on standard dialed-up telephone lines (such as those received at TSPS offices). Such factors as popping sounds, crackling noises, background speech, carrier frequency tones, and other nonstationary noises make it very hard to detect word boundaries accurately.

In Wilpon et al. it was shown that only 69 percent of all words were detected by the Lamel approach when tested on a large random subset of the PO database.[12] Among these, the recognizer accurately classified 85 percent. This yielded an overall recognition system accuracy of only 59 percent. Because of these results, it was decided to try and improve the endpoint algorithm before proceeding further. This led to the development of a new word detection algorithm, called a top-down design.[12] The new approach makes the assumption that if speech is present in some time interval its energy level will be above that of any noise also present. Simply put, the new algorithm searches for strong (vowel-like) peaks in the energy contour of a speech utterance and processes the speech around the peaks to find potential beginning and ending points. Several rules involving duration, onset, and decay times are then used to refine the endpoint estimates.

Applying this new endpoint algorithm to the same data set as was tested with the Lamel algorithm (i.e., a subset of the PO database) yielded a word detection rate of 98 percent and a recognition accuracy of 90 percent for an overall system accuracy of 89 percent. Clearly, from the results obtained, the top-down endpointing algorithm is superior to the Lamel approach. As a result of this research, the top-down algorithm was used in all studies of the BR database. Whereas in the PO database study over 50 percent of the database had to have manual corrections made to the endpoints (because of endpoint algorithm failures) *no* manual correction of endpoints was performed in the BR study.

5.2 Clustering analysis improvements

In Wilpon and Rabiner[13] a new clustering algorithm was presented that uses the best features of several previously used algorithms—i.e., ISODATA[2,15], *K*-means[2,15], and UWA[6]. This algorithm is called the

Modified K-Means (MKM) clustering algorithm. Its main advantage over other algorithms is that it is completely automatic, and requires no user input (other than a similarity matrix). This algorithm was tested extensively on the BR database and was shown to yield recognition results as good as previously used algorithms. In the experiments to follow, all template sets created from subsets of the BR speech data will have been created using the MKM algorithm.

VI. RECOGNITION RESULTS ON THE BR DATABASE

6.1 Isolated word recognition results

For all isolated word recognition experiments performed on the BR database, the isolated database was divided into two disjoint subsets, one to train the recognizer and the other to test the system. A total of 4783 tokens were used for training and another 7973 tokens for testing. Table III shows the distribution of the training and testing tokens among the ten digits.

In the PO study, template sets were created both from a random subset of the speech data and from the "cleanest" speech data (i.e., as judged by a human to be close to laboratory-quality data). Similar recognition results were obtained using both template sets. Therefore, in the BR study template sets were created using only a random subset of speech data.

The recognition system used in all evaluations was the Linear Predictive Coding (LPC)-based isolated word recognition system developed and tested extensively at AT&T Bell Laboratories.[1-7] As we stated in Section 5.2 the MKM clustering algorithm was used to create several sizes of template sets. Table IV shows the recognition results for seven different clustering configurations: 3, 6, 12, 20, 30, 50, and 75 clusters per word. Shown is the per-digit accuracy, average digit

Table III—Number of tokens for each digit used in training and testing for evaluating the BR database

	Training Set	Testing Set
0	271	312
1	259	405
2	675	1145
3	606	943
4	580	970
5	489	901
6	592	1070
7	443	750
8	454	834
9	414	643
Total	4783	7973

Table IV—Recognition results in percent using BR speech data for training and for testing

	Number of Templates per Word						
Digit	3	6	12	20	30	50	75
0	60.5	71.0	70.7	72.8	69.6	72.8	72.8
1	85.6	86.1	85.6	88.0	88.2	88.5	87.7
2	64.2	67.5	74.5	76.8	78.2	81.5	82.6
3	74.2	84.0	87.5	89.6	91.3	89.9	90.4
4	88.4	92.4	93.5	92.6	95.4	94.2	92.9
5	78.9	78.9	86.4	85.0	87.8	89.1	89.5
6	63.3	75.1	79.8	80.9	79.3	87.7	84.9
7	62.8	74.7	80.2	84.9	88.4	88.4	90.1
8	71.4	70.3	75.8	80.1	81.2	83.3	83.3
9	58.8	69.8	63.4	70.9	74.9	75.6	80.6
Average	71.0	77.0	80.5	82.7	84.4	86.1	86.3
String rate	16.3	24.2	29.5	34.1	36.0	42.2	43.5

accuracy over all digits, and string accuracy, where a string is nominally seven digits long. We see that for all template sizes the digits zero (or oh) and nine have the highest error rates. The major confusion for the digit zero (oh) was the digit four. In the BR and PO studies about one-half of the talkers pronounced that word four as /foe/ rather than /fawr/ and used the word oh instead of zero more than 75 percent of the time. A possible explanation for the confusion could be that endpoint detector included too much background noise when determining the beginning point for some of the pronunciations of the word oh, thereby making the word oh like a /foe/ and misrecognizing it. Alternatively, since the frication at the beginning of the word four closely resembles typical background noise encountered in our testing environment, it would be easy for the speech endpoint detector to misplace the beginning marker for this word, thus totally eliminating the fricative sound. Since templates for the word four are created from this type of data, a spoken digit oh could be misrecognized. Low accuracy for the digit nine was also obtained. A possible reason for such low accuracy is that the nasal sound is being masked by the various noises on the telephone line. Figure 2 shows a plot of digit recognition accuracy as a function of the number of templates (or clusters) created for each word. We can see that as the number of templates per word increases, the recognition accuracy increases asymptotically, with the best accuracy (86.3 percent) occurring with a 75-template-per-word set.

The average string length was seven digits. Therefore, theoretically the average string accuracy is the average per-digit accuracy raised to the seventh power (since all single digit recognitions are independent of one another). However, for all template set sizes the actual string accuracy was greater than the theoretical result, that is, the error rate

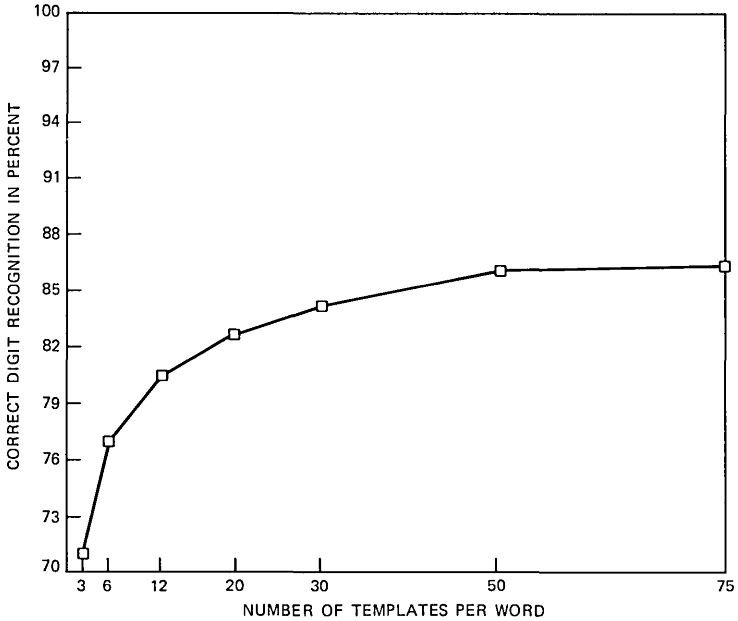


Fig. 2—Recognition accuracy for training and testing on BR data as a function of the number of templates used per word.

was not uniform over all talkers nor independent of the talker. A similar result on another speech database was obtained by Rosenberg and Shipley.[16]

Figure 3 demonstrates the effects of imposing a rejection threshold on the recognition system. A recognition distance score above the threshold would result in a no decision choice by the recognizer. Shown in this plot is the percent of no decisions versus the percent of error rate. We see that if only a 1-percent error rate could be tolerated by a task using this recognizer under these recording conditions, then a 60-percent no decision rate must also be accepted. However, a 10-percent probability of error was attained with only a 9-percent no decision rate.

If we compare these results (using a 30-template-per-word solution for comparison) with those obtained from the PO database study, the results from the BR study seem to be worse (84.4 percent for BR versus 93.1 percent for PO). However, in the PO study 50 percent of the speech data available for testing was eliminated from the database because of noise conditions, connected rather than isolated input, and hardware failures. Also, the automatic endpoint detector[14] was overruled by human intervention about 50 percent of the time.[8] In contrast, in the BR study all the data available were used and automatically

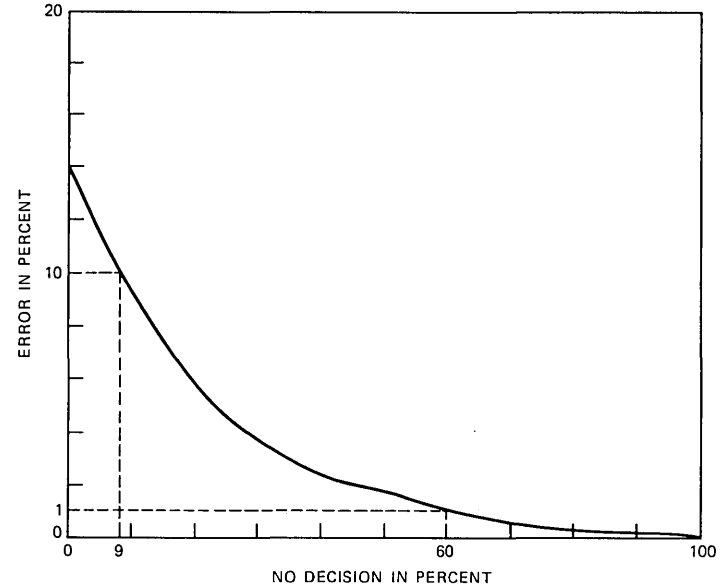


Fig. 3—Plot showing recognition error rate versus no decision choice—training and testing data from BR.

detected before recognition was performed.[12] For this reason it is felt that the BR results are very encouraging.

6.2 Connected word recognition results

Recognition was performed on the 980 call subset of the processable calls, which contained some connected digit sequences using the level-building Dynamic Time Warp (DTW) algorithm of Myers et al.[17] Testing was carried out with and without augmenting the Itakura log likelihood distance with an energy distance as described in Rabiner.[18] The template set used was the 30-template-per-word set created from isolated tokens from the BR database. No embedded training, as described in Rabiner et al.[19], was used.

Since there was not an abundance of connected digit strings within the 980 strings (only 1790), I will just present the results and not make any categorical remarks on connected digits input over noisy channels. Table V shows the results of these experiments. The results indicate that using energy information in the distance computation improved the string accuracy for all string lengths from 45.8 to 61.0 percent if the string length is unknown, and from 63.6 to 66.8 percent if the string length is known. It is expected that the use of embedded training would greatly improve these results.

Table V—Recognition results in percent from connected digit sequences from BR speech data

No. of Digits in String	No. of Occurrences	Percent Correct Recognition Without Energy Known Length	 Unknown Length	 With Energy Known Length	 Unknown Length
2	1441	68.5	49.4	71.1	65.3
3	277	46.6	33.6	51.6	45.9
4	67	32.8	22.4	38.8	32.8
5	5	20.0	0.0	60.0	40.0
Total	1790	63.6	45.8	66.8	61.0

6.3 A syntax-directed recognition system based on isolated digit input

The results described previously assume that all single digit recognitions are independent of each other. But in fact that is not the case for this database, as customers were asked to speak their seven-digit telephone number. For this well-defined task there is some syntactic information that can be used to help guide the recognition system. For example, the first three digits of the seven-digit input define the local exchange. In general, there are significantly fewer than the 1000 exchanges within an area-code region. However, the last four digits are usually distributed uniformly over the 10,000 possible sequences.

A recognition system was assembled to make use of the syntactic structure of telephone numbers. First, the database was searched to find all valid exchanges. This yielded a total of 86 valid exchanges out of a possible 1000. The recognition system was then programmed to do the following task. For each customer, digit recognition was performed on the exchange. The output of the recognition system was a set of similarity scores[1] for each digit, for all digits in the exchange. Next, the customer's actual utterances for the exchange were tagged as being that valid exchange with the lowest total distance (i.e., the sum of the individual digit scores). The utterances were then converted into speaker-dependent templates and added to the previously created speaker-independent template set. This new template set was then used to recognize the last four digits in the telephone number. This procedure was done on a per-talker basis. If in the last four digits the customer spoke any of the digits that were in the exchange, having a template of those words created by the user should increase the probability that the recognizer would correctly recognize those words.

Whereas the recognition accuracy (for a 30-template-per-word reference set) yielded a digit accuracy of 84.4 percent when the above system was implemented, the digit accuracy increased to 87.2 percent. The string accuracy without syntax, as tested on 984 seven-digit

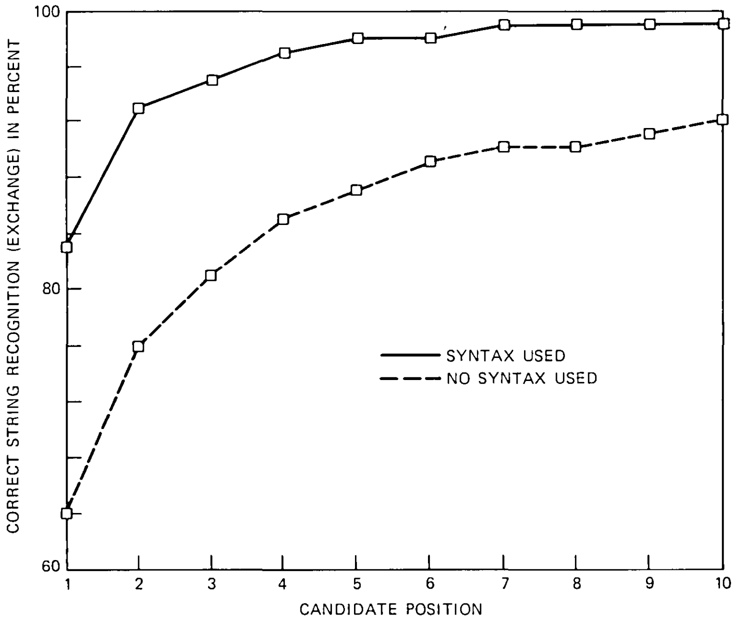


Fig. 4—Recognition accuracy on the exchange string (three digits) from the BR database as a function of candidate position (solid line is with syntax and dashed line is without syntax).

strings, was 38.2 percent. This increased to 47.8 percent when syntax was added.

Figure 4 shows a plot of exchange recognition accuracy as a function of whether the correct exchange was within the top ten candidates. The dashed line shows the results when no syntax is used, and the solid line shows the results when syntax is used. We see that the correct exchange (all three digits) has been recognized correctly in the system with no syntax 64 percent of the time and 87 percent within the top five candidates, whereas in the syntax-directed system the results are 83 percent and 98 percent, respectively. Figure 5 shows a plot of the number of times the correct exchange is within a distance Δ from the minimum possible exchange score (over the 1000 possible exchanges). For the Itakura log-likelihood ratio distance the mean distance for a correct recognition is about 0.30 and for an incorrect recognition about 0.45.[2,8] We see that within a distance of 0.25 from the minimum the correct exchange (three digits) is always present. Figure 6 shows the average number of possible exchanges within a Δ region over all strings. It shows that using syntax greatly reduces the number of recognition candidates (e.g., from 80 to 10 for a $\Delta = 0.20$).

Figure 7 shows a plot of the string accuracy of the last four digits as

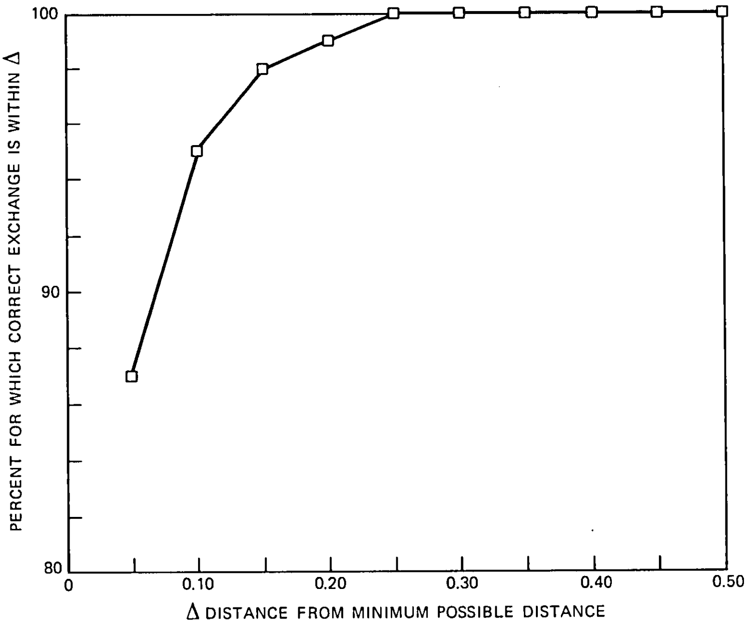


Fig. 5—Recognition accuracy on the exchange string (three digits) as a function of whether the correct exchange is within a Δ distance from the minimum possible exchange score.

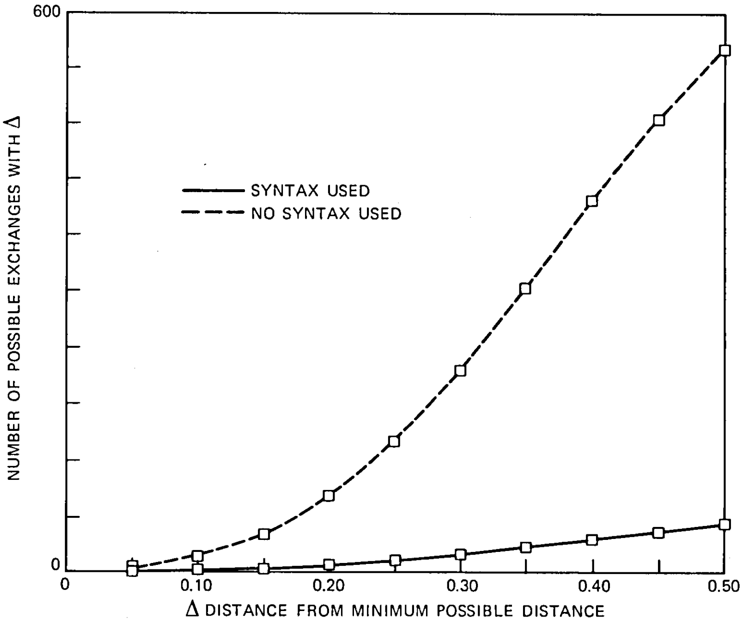


Fig. 6—Plot showing the average number of exchange candidates within a Δ region.

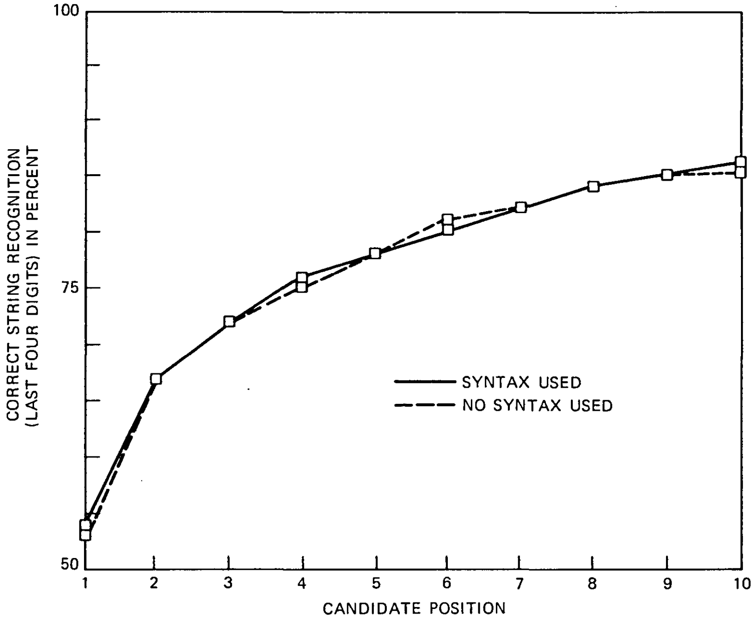


Fig. 7—Recognition accuracy on the last four digits from the BR database as a function of candidate position. The dashed line indicates no syntax was used and the solid line indicates syntax was used.

a function of whether the correct string was within the top ten candidates. The dashed line shows the results without using the additional templates generated by applying the syntactic rules to the first three digits. The solid line shows the recognition results when the original speaker-independent template set was augmented with the speaker-dependent templates determined through the syntactic rules on the first three digits. With syntax the string accuracy only improved from 53.3 to 54.5 percent and was within the top five candidates 78 percent of the time.

These results show that adding task information improved the overall system recognition accuracy. Augmenting the template set with the extra exchange utterance templates slightly improved performance on the last four digits. However, the main contribution of the syntax was to recognize the exchange more accurately.

VII. ROBUSTNESS OF SPEAKER-INDEPENDENT TEMPLATES

One of the goals of this study was to examine the robustness of speaker-independent templates created using one population of talkers under one set of transmission conditions for different populations and transmission conditions. An experiment was carried out in which template sets created from each of the three regional databases (Portland, Baton Rouge, and Murray Hill) were tested on speech data from all three databases.

Initially, a template set was created from a Murray Hill speech database that consisted of 100 talkers, 50 male and 50 female. The data were collected under laboratory conditions over local Private Branch Exchanges (PBXs). A clustering analysis was performed and a set of 12 speaker-independent templates per word was created. This template set has been tested extensively in other experiments (see Refs. 2, 8, 11, 14, 16, and 17). For testing purposes, another group of 100 talkers (disjoint from the training population) each provided one replication of the digits vocabulary.

The template set used to represent Portland data was a 30-template-per-word set created from the "cleanest" speech obtained in the PO study.[8] For testing the entire 11,035-digit database was used. For comparison purposes a 30-template-per-word reference set was used to model the Baton Rouge database. For testing purposes the entire 7973-digit testing set was used.

Table VI shows the results for all cross recognition tests. The symbol <AVG> stands for the averaging of recognition results over all three databases given a particular training or testing dataset. In order not to distort the averages (since each database had a different number of tokens), a simple nonweighted averaging was performed. It is felt that there was sufficient data in each regional database to make this result meaningful.

Figure 8 (a graphical form of Table VI) shows the results when

Table VI—Cross template and testing set recognition accuracy

Training Set	Testing Set	Recognition Accuracy in Percent
BR	BR	84.4
BR	PO	85.8
BR	MH	92.3
PO	PO	93.1
PO	BR	76.8
PO	MH	91.6
MH	MH	98.4
MH	PO	77.4
MH	BR	62.3
BR	<AVG>	87.4
PO	<AVG>	87.1
MH	<AVG>	79.3
<AVG>*	BR	74.3
<AVG>	PO	85.4
<AVG>	MH	94.1

* <AVG> = averaged over all three databases

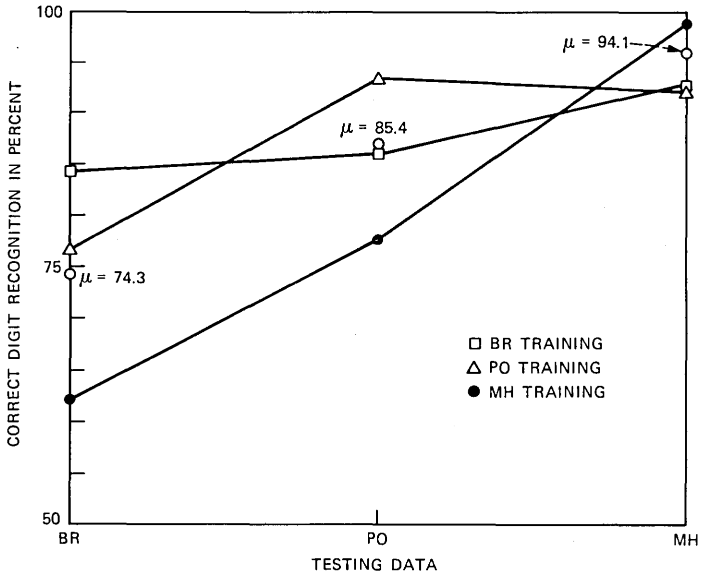


Fig. 8—Recognition results for each regional template set as a function of the testing set.

testing each of the template sets against each of the testing datasets, where μ is the average recognition accuracy over all three testing sets given a training set. Shown is recognition accuracy as a function of which testing set was used. This figure shows that the best recognition results for each of the test sets occurred, not surprisingly, using the template set also created from data in the same region. The recognition performance was best with MH data, then with PO data, and last with BR data. Also, notice the greater variation in recognition accuracies as the self-recognition scores decline. For example, the PO and BR templates performed about the same against MH testing data and about 7 percent worse than MH templates, whereas the PO and MH template sets performed, respectively, 8 and 21 percent worse than the BR template set when tested with BR data.

In Fig. 9, the results are shown as a function of individual digits. As was the case in the earlier PO study, we see that the MH templates do not adequately represent the speaking style or noise conditions present in the PO or BR testing data. For most of the digits in the BR and PO testing population the templates from PO and BR yielded much better results.

Figure 10 shows the results in a different context. Shown is recognition accuracy as a function of the template set used. Interestingly, the BR template set performed the best over all three testing conditions. (Even though the same average accuracy was obtained with the

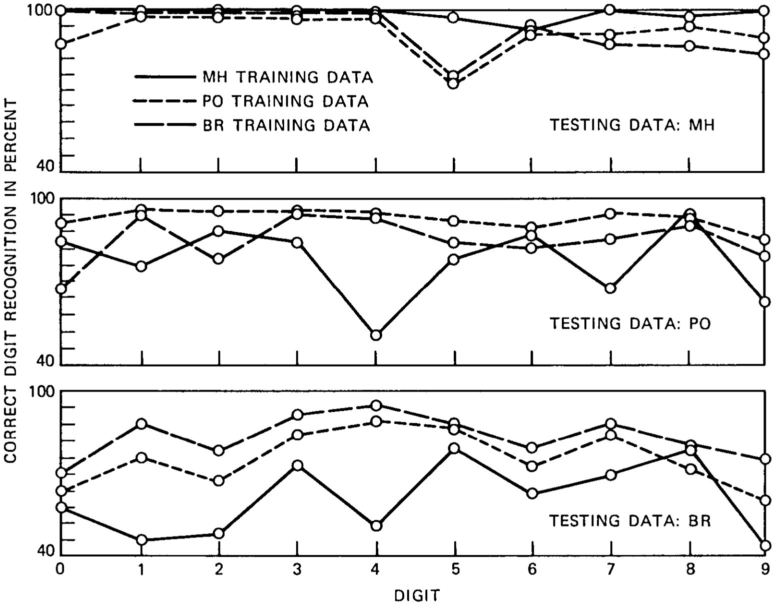


Fig. 9—Recognition results for each regional template set as a function of testing set, on a per-digit basis.

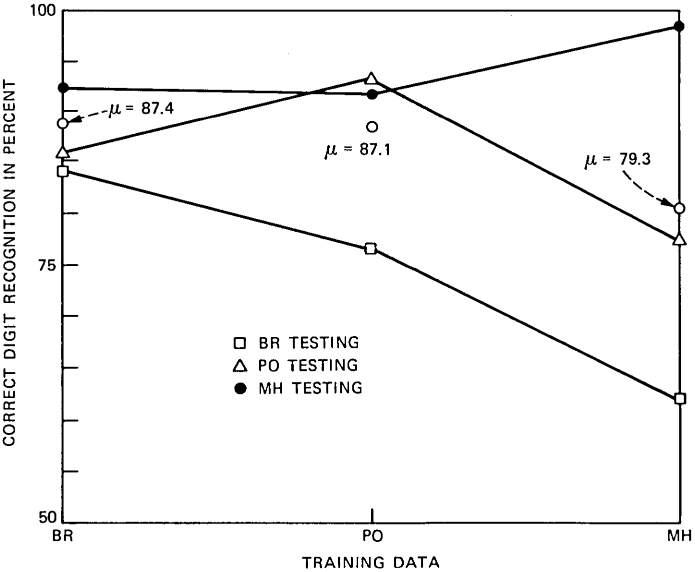


Fig. 10—Recognition accuracy for each regional testing set as a function of the template set used.

BR and PO templates, the BR templates yielded a standard deviation of 4.3 percent compared with 9.1 percent for the PO template set.) The MH template set performed significantly worse than either the BR or PO template sets, with an average recognition accuracy of 79.3 percent and a standard deviation of 18.1 percent.

Figure 11 shows the per-digit recognition results for each testing data set given a particular template set. For most digits in the MH testing set, the template sets created from BR and PO data yielded as good a recognition accuracy as did the templates created from MH data. However, again we see the converse not to be true, that is, the MH templates yielded significantly poorer recognition results when tested against PO and BR testing data than did template sets created from those regions.

Tables VII through IX show confusion matrices generated from each of the above recognition experiments. Shown are only those confusions that occurred more than 3 percent of the time. In Table VII, results are shown for each testing set when recognition was performed using the BR template set. In each of the BR and PO testing sets the biggest error was the spoken word oh being confused for four. Possible explanations for the confusions have been given in Section 6.1. This problem did not occur in the MH data as all talkers

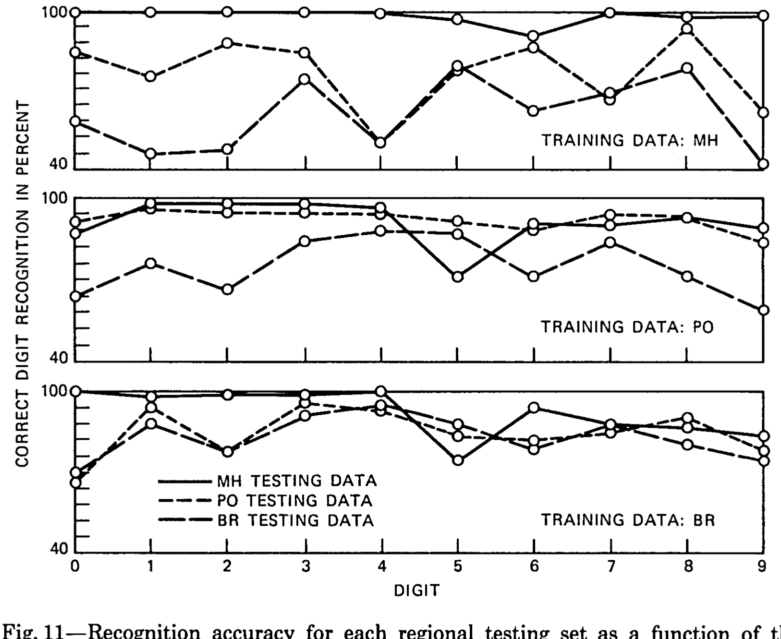


Fig. 11—Recognition accuracy for each regional testing set as a function of the template set used, on a per-digit basis.

Table VII—Confusion matrix for each testing set when recognition was performed using the BR template set

Spoken Digit	Recognized Digit									
	0	1	2	3	4	5	6	7	8	9
					(a) BR Testing Data					
0	69.6				15.6			5.1		
1		88.2			5.3					
2	4.6		78.2				4.7	5.4		
3				91.3			3.0			
4					95.4					
5						87.8				6.9
6				4.2			79.3		11.6	
7								88.4		
8				7.5			5.6		81.2	
9		3.1				12.8		4.6		74.9
					(b) PO Testing Data					
0	66.3				27.7					
1		94.2								
2	8.1		77.8		8.1			4.5		
3				95.7						
4		3.3			93.6					
5		6.8				84.0				5.5
6							82.4	6.8	5.1	
7								85.5		5.0
8							5.8		90.3	
9		4.4		4.1		8.9				79.0
					(c) MH Testing Data					
0	100.0									
1		98.0								
2			99.0							
3				99.0						
4					100.0					
5		4.0			5.0	75.0		4.0		11.0
6			3.0				94.0			
7			10.0					87.0		
8	3.0		3.0	3.0			4.0		87.0	
9		9.0		5.0						84.0

used the word zero. Notice that the reverse confusion (i.e., four misrecognized as zero or oh) did not occur.

Table VIII shows the confusion matrices when using the template set generated from MH data. The only confusion when tested against MH testing data was six versus seven. Notice when testing against BR data that all digits are misrecognized as seven a large percent of the time. For this testing data there are many major confusions.

Table IX shows the confusions generated from training with PO data. As with the other template sets the five-nine confusion is prominent. Also, the MH testing set produced a large confusion between the digits nine and one. In examining Tables VII through IX the template set generated from BR speech data yielded fewer major

Table VIII—Confusion matrix using template set generated from MH data

Spoken	Recognized Digit									
Digit	0	1	2	3	4	5	6	7	8	9
(a) BR Testing Data										
0	58.4		11.8				4.7	11.8	6.5	
1	7.8	46.3			6.4	20.1		4.0		8.3
2			47.9	3.3			7.0	19.4	19.9	
3				74.2			7.0	5.4	7.5	
4	28.6	3.1			50.6	8.5		3.8		
5	3.3					79.6	7.4	4.2		3.9
6				4.0			62.6	5.0	24.9	
7			4.6			3.0	12.3	69.4	5.1	3.0
8				7.4				4.7	79.2	
9						26.7	14.7	10.2		43.2
(b) PO Testing Data										
0	84.3		5.6					3.3		
1	6.3	72.9				7.8			4.3	6.3
2			86.4					6.0		
3				87.7					5.4	
4	34.0				48.7	8.8				
5	3.2					80.4	5.6			6.5
6							85.8		7.5	
7						3.2	9.2	74.8	3.0	3.2
8									95.3	
9				4.4		12.6	12.3	3.9		64.3
(c) MH Testing Data										
0	100.0									
1		100.0								
2			100.0							
3				100.0						
4					99.0					
5						97.0				
6							91.0	6.0		
7								100.0		
8									98.0	
9										99.0

confusions (over all three test sets) than either the PO or MH template sets.

Summarizing this experiment, the template set created from a subset of BR speech data was quite robust over different populations and noise conditions, yielding an average recognition accuracy of 87.4 percent. Additionally, the MH template set, which was created under laboratory conditions, provided poor recognition results when tested under "real-word" recording conditions.

In a final experiment the template sets created from the PO, BR, and MH data were combined together to form one large template set with 72 templates per word (i.e., 30 templates per word from each of the PO and BR sets, and 12 templates per word from the MH template set). The testing set for this experiment was the combined testing sets from PO, BR, and MH.

Table IX—Confusion matrix generated from training with PO data

Spoken	Recognized Digit									
Digit	0	1	2	3	4	5	6	7	8	9
					(a) BR Testing Data					
0	63.8		12.5		5.7		3.6	6.8		
1		76.2			13.4	5.6				
2			66.8	5.4			12.7	4.2	8.0	
3				84.6			3.8			
4	5.0		3.4		88.5					
5						87.7		4.7		3.1
6				5.8			72.1	3.5	15.5	
7							5.8	84.9		
8				10.0			7.9	5.6	72.5	
9						19.4		12.5		60.1
					(b) PO Testing Data					
0	90.9									
1		95.2								
2			95.0							
3				94.2						
4	3.8				93.4					
5						93.1				3.0
6							87.0	3.2	4.2	
7								93.3		
8									95.3	
9						9.2				85.0
					(c) MH Testing Data					
0	87.0		9.0		4.0					
1		98.0								
2			98.0							
3				98.0						
4			3.0		97.0					
5					8.0	72.0		3.0		14.0
6			4.0				91.0	5.0		
7			7.0					91.0		
8				3.0					94.0	
9		7.0								90.0

A recognition accuracy of 90.9 percent was achieved under these conditions. In examining a histogram of template usage, several templates were used more often for incorrect recognitions than for correct recognitions. A test was carried out in which template sets were created as subsets of the full 72-template-per-word set. These subsets were chosen such that the Net Percent Correct Recognition (NPCR) per template, as defined over all three testing sets (i.e., the percentage of the time that the template was used for a correct score minus that when used incorrectly), was greater than a threshold. Figure 12 shows a plot of the total number of templates used (dashed line) and the recognition accuracy (solid line) as a function of the threshold. As indicated by the results, 20 templates of the original 720-template set yielded only incorrect recognitions. Also, most templates yielded a NPCR of 50 percent. As we look for a NPCR of greater than 50 percent, the number of templates that qualify goes down exponentially. Only 144 of the original 720 templates yielded an NPCR of 100 percent.

The recognition curve (solid line) shows a similar shape. We see that recognition accuracy stays constant for NPCRs of less than 80 percent, then falls rapidly to 68 percent for an NPCR of 100 percent. These two curves show that the total number of templates can be reduced by 22 percent from 720 to 560 (or an average of 56 templates per word) without reducing the overall recognition accuracy.

Table X shows a confusion matrix for the combined template set (only entries greater than 3 percent are shown). The results indicate that this template set yielded fewer confusions than did either of the individual template sets, with the major confusions being zero (oh)-four, nine-five, and six-eight.

VIII. DISCUSSION

The results described in earlier sections show that:

1. Based on the collection of speech data in Portland and Baton Rouge, it is clear that significant problems exist prompting casual telephone customers to speak digit strings in an isolated format.

2. One problem that existed in the Portland study was the inability to detect words automatically in nonideal environments. With the use of the top-down endpoint detection algorithm,[12] this problem was greatly reduced in the Baton Rouge tests.

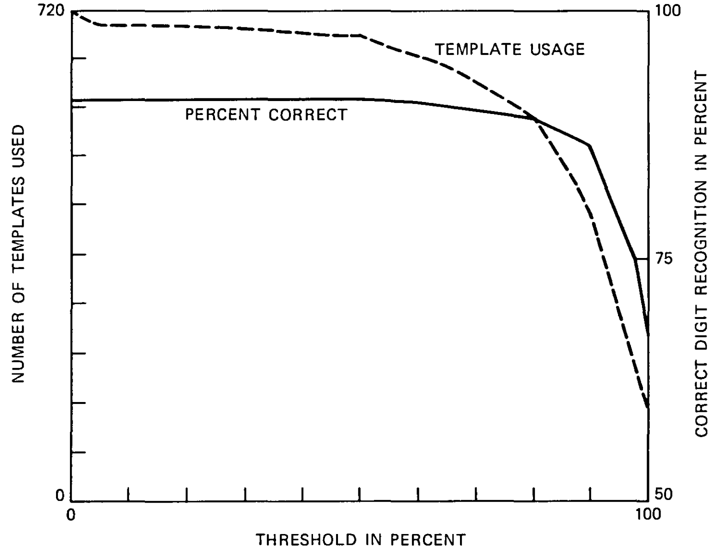


Fig. 12—Plot showing recognition accuracy and number of templates used as a function of template use threshold.

Table X—Confusion matrix—combined template set versus all testing data

Spoken	Recognized Digit									
Digit	0	1	2	3	4	5	6	7	8	9
0	86.5				5.6					
1		94.5								
2			90.7							
3				93.9						
4					94.4					
5						92.5				3.4
6							85.7		7.1	
7								92.0		
8							3.3		90.5	
9						7.5				84.2

3. The recognition results obtained from the BR tests were worse than those obtained in the Portland study (84.2 and 93.1 percent, respectively, on comparable sized reference sets). Since in the Portland experiment 50 percent of all data was eliminated from testing and 50 percent of the remaining data needed human interaction to correct endpoint failures, we feel the results from the Baton Rouge study, which eliminated no data and did not allow for endpoint corrections, more accurately demonstrate our current capabilities.

4. The addition of syntactic constraints on the isolated word recognizers output increased the overall recognition system accuracy—from 84.4- to 87.2-percent digit accuracy and from 38.2- to 47.8-percent string accuracy (i.e., seven-digit telephone number).

5. The template set created from a subset of BR data is quite robust over different populations and noise conditions, averaging 87.4 percent over the three regional data sets. By creating a combined template set based on the templates generated from speech data from Portland, Baton Rouge, and Murray Hill, a recognition accuracy of 91 percent was obtained when tested on 20,000 tokens of PO, BR, and MH data.

These results are very encouraging, as they indicate that regional "speaker-independent" template sets may not be required to obtain the highest recognition accuracy possible over all regions. However, since for each regional database the best recognition scores occurred using training and testing data from that region, having regional templates will improve accuracies in the individual regions.

To compute the end-to-end recognition system performance number, several intermediate results must be combined together. Figure 13 shows the combination of all steps in the recognition system. Starting with all calls that were handled by our recognition system, initially 20 percent abandoned the transaction. Of the 80 percent of calls remaining, 52 percent required some form of operator assistance to complete the call and 17 percent contained some connected input.

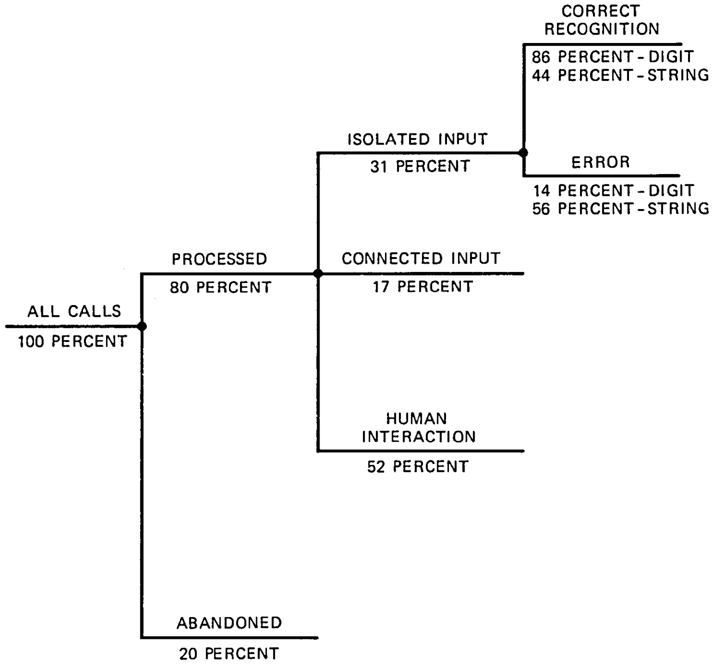


Fig. 13—End-to-end recognition system performance from isolated digit input.

This left only 24 percent of the calls consisting solely of isolated digit input. Therefore, we can see that the end-to-end system performance was 21.3 percent on digits and 11 percent on strings (where a string consisted nominally of seven isolated digits). The full recognition system was able to handle automatically 11 percent of *all* calls received. If such a system were to be implemented, hopefully over a period of time customers would learn the required task. This would greatly reduce the number of transactions needing manual assistance and the number of calls containing connected input. Once connected digit recognition has achieved the same performance as isolated speech, the restriction of isolated input can be relaxed. These improvements should greatly increase end-to-end system performance.

IX. SUMMARY

Results have been presented from a series of speech recognition experiments on a speech database obtained from 7373 telephone customers speaking in an actual telephone environment in Baton Rouge, Louisiana. The best performance of 86.3-percent correct digit recognition was obtained when a set of speaker-independent templates was created from a subset of the data and tested on the remaining data.

We described a syntax-directed recognition system that incorporates information about a seven-digit telephone number task. System accuracy was shown to improve by 9.4 percent.

Finally, a series of recognition tests was performed to quantify the robustness of speaker-independent templates created under one set of recording conditions and tested under another. Additionally, a template set was created from a subset of each of the regional templates sets. A recognition accuracy of 91 percent was obtained when tested against 20,000 isolated tokens from PO, BR, and MH data.

REFERENCES

1. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 67–72.
2. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27*, No. 4 (August 1979), pp. 336–49.
3. L. R. Rabiner and J. G. Wilpon, "Speaker Independent, Isolated Word Recognition for a Moderate Size 54 Word Vocabulary," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27*, No. 6 (December 1979), pp. 583–7.
4. J. G. Wilpon, L. R. Rabiner, and A. F. Bergh, "Speaker Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary," J. Acoust. Soc. Amer., *72*, No. 2 (August 1982), pp. 390–6.
5. L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68*, No. 5 (October 1980), pp. 1069–70.
6. L. R. Rabiner and J. G. Wilpon, "Considerations in Applying Clustering Techniques to Speaker Independent Word Recognition," J. Acoust. Soc. Amer., *66*, No. 3 (September 1979), pp. 663–73.
7. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated Word Recognition for Large Vocabularies," B.S.T.J., *61*, No. 10, Part 1 (December 1982), pp. 2989–3005.
8. J. G. Wilpon and L. R. Rabiner, "On the Recognition of Isolated Digits From a Large Telephone Customer Population," B.S.T.J., *62*, No. 7 (September 1983), pp. 1977–2000.
9. F. Pirz and K. Bauer, unpublished work.
10. L. R. Rabiner, J. G. Wilpon, and A. E. Rosenberg, "A Voice Controlled Repertory Dialer System," B.S.T.J., *59*, No. 7 (April 1980), pp. 571–92.
11. B. Aldefeld, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Automated Directory Listing Retrieval System Based on Isolated Word Recognition," J. Acoust. Soc. Am., *68*, No. 5 (November 1980), pp. 1271–6.
12. J. G. Wilpon, L. R. Rabiner, and T. Martin, "An Improved Word-Detector Algorithm for Telephone-Quality Speech Incorporating Both Syntactic and Semantic Constraints," AT&T Bell Lab. Tech. J., *63*, No. 3 (March 1984), pp 479–98.
13. J. G. Wilpon and L. R. Rabiner, "A Modified *K*-Means Clustering Algorithm for Use in Speaker Independent Isolated Work Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-33*, No. 3 (June 1985).
14. L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-29* (August 1981), pp. 777–85.
15. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg and J. G. Wilpon, "Interactive Clustering Techniques for Selecting Speaker Independent Reference Templates for Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27*, No. 2 (April 1979), pp. 134–41.
16. A. E. Rosenberg and K. L. Shipley, "Evaluation of An Isolated Word Recognizer in Talker-Dependent and Talker-Independent Modes Using a Large Telephone-

Band Database," Conf. Rec., 1984 IEEE Int. Conf. Acoust., Speech, Signal Processing, March 1984.

17. C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-29*, No. 2 (April 1981), pp. 284–97.
18. L. R. Rabiner, "On the Applications of Energy Contours to the Recognition of Connected Word Sequences," AT&T Bell Lab. Tech. J., *63* (December 1984), pp. 1981–95.
19. L. R. Rabiner, A. F. Bergh, and J. G. Wilpon, "An Improved Training Procedure for Connected-Digit Recognition," B.S.T.J., *61* (July–August 1982), pp. 981–1001.

AUTHOR

Jay G. Wilpon, B.S., A.B. (cum laude in Mathematics and Economics, respectively), 1977, Lafayette College, Easton, Pa.; M.S. (Electrical Engineering/Computer Science), 1982, Stevens Institute of Technology, Hoboken, N.J.; AT&T Bell Laboratories, 1977—. Since June 1977 Mr. Wilpon has been with the Acoustics Research Department at AT&T Bell Laboratories, Murray Hill, N.J., where he is a Member of the Technical Staff. He has been engaged in speech communications research and presently is concentrating on problems of speech recognition. He has published extensively in this field and has been awarded several patents. His current interests lie in training procedures, speech detection algorithms, and determining the viability of implementing speech recognition systems for general usage.

PAPERS BY AT&T BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Cleveland W. S., McGill R., **Graphical Perception—Theory, Experimentation, and Application to the Development of Graphical Methods.** J Am Stat A 79(387):531–554, Sep 1984.

DeTreville J., **Phoan: An Intelligent System for Distributed Control Synthesis.** SIGPLAN Not 19(5):96–103, May 1984.

Flajolet P., Odlyzko A. M., **Limit Distributions for Coefficients of Iterates of Polynomials With Applications to Combinatorial Enumerations.** Math Proc C 96(Sep):237–253, Sep 1984.

Kapadia A. S., Kazmi M. F., Mitchell A. C., **Analysis of a Finite-Capacity Nonpreemptive Priority Queue.** Comput Oper 11(3):337–343, 1984.

Kernighan B. W., **The *UNIX* System and Software Reusability.** IEEE Soft E 10(5):513–518, Sep 1984.

Lloyd S. P., **Optimal Dual Output Extended Hydrophone.** SIAM J A Ma 44(5):1031–1040, Oct 1984.

Luss H., **Capacity Expansion Planning for a Single Facility Product Line.** Eur J Oper 18(1):27–34, Oct 1984.

Morton M. J., Gray H. L., **The G-Spectral Estimator.** J Am Stat A 79(387):692–701, Sep 1984.

Topkis D. M., **Adjacency on Polymatroids.** Math Progr 30(2):229–237, Oct 1984.

Tukey P. A., Wachter K. W., **A Demographic Analogy for Shareowner Accounts.** J Am Stat A 79(387):525–530, Sep 1984.

ENGINEERING

Agrawal G. P., Olsson N. A., Dutta N. K., **Effect of Fiber-Far-End Reflections on Intensity and Phase Noise in InGaAsP Semiconductor Lasers.** Appl Phys L 45(6):597–599, Sep 15, 1984.

Alferness R. C., **A Strong Potential for Integrated-Optics Applications.** Laser Foc E 20(10):186–188, Oct 1984.

Ashkin A., **Stable Radiation-Pressure Particle Traps Using Alternating Light Beams.** Optics Lett 9(10):454–456, Oct 1984.

Banu M., Tsividis Y., **Detailed Analysis of Nonidealities in MOS Fully Integrated Active RC Filters Based on Balanced Networks.** IEE Proc-G 131(5):190–196, Oct 1984.

Baumert R. J., Cameron L. E., Wilson R. A., **A Mixed EFL I^2L Digital Telecommunication Integrated Circuit.** IEEE J Sol I 19(1):26–31, 1984.

Chu S., Burrus C. S., **Multirate Filter Designs Using Comb Filters.** IEEE Circ S 31(11):913–924, Nov 1984.

Dragone C., **Scattering at a Junction of Two Waveguides With Different Surface Impedances.** IEEE Micr T 32(10):1319–1328, Oct 1984.

Feldman L. C., **Rutherford Scattering-Channeling Analysis of Semiconductor Structures.** P Soc Photo 452:192–201, Nov 9–10, 1983.

Harris T. D., Williams A. M., **Low Absorbance Measurements.** P Soc Photo 426:110–115, Aug 23–24, 1983.

Hasegawa A., **Numerical Study of Optical Soliton Transmission Amplified Periodically by the Stimulated Raman Process.** Appl Optics 23(19):3302–3309, Oct 1, 1984.

Heffron G., **Teleconferencing Comes of Age.** IEEE Spectr 21(10):61–66, Oct 1984.

Heiman A., Barness Y., **Optimal Design of PLL With Two Separate Phase Detectors—Reply (Letter).** IEEE Commun 32(9):1058–1060, Sep 1984.

Henry P. S. et al., **New Technique for Timing Recovery.** Electr Lett 20(20):815–816, Sep 27, 1984.

Kaiser P., **Optical-Technology Used in the Atlanta Single-Mode Experiment.** P Soc Photo 425:127–133, 1983.

Kelso S. M., Aspnes D. E., Olson C. G., Lynch D. W., Bachmann K. J., **Determination of Indirect Conduction-Band Minima in Semiconductors by Core-Level Reflectance Spectroscopy.** P Soc Photo 452:100–109, Nov 9–10, 1983.

Liao P. F., Glass A. M., Johnson A. M., Olson D. H., Humphrey L. M., Stern M. B., **Enhancement of Optical-Detector Response Via Microstructured Electrodes.** P Soc Photo 439:197–201, 1983.

Lin C. L., Eisenstein G., Burrus C. A., Tucker R. S., **Temporal and Spectral Characteristics of Single-Longitudinal-Mode Short-Coupled-Cavity (SCC) InGaAsP Lasers Under Multigigahertz Direct Modulation.** Electr Lett 20(20):842–844, Sep 27, 1984.

Linke R. A., Gnauck A. H., **High Speed Laser Driving Circuit and Gigabit Modulation of Injection Lasers.** P Soc Photo 425:123–125, 1983.

Marcuse D., **Computer Simulation of Laser Photon Fluctuations—Single-Cavity Laser Results.** IEEE J Q El 20(10):1148–1155, Oct 1984.

Marcuse D., **Computer Simulation of Laser Photon Fluctuations—Theory of Single-Cavity Laser.** IEEE J Q El 20(10):1139–1148, Oct 1984.

Miller S. E., Marcuse D., **On Fluctuations and Transients in Injection Lasers.** IEEE J Q El 20(9):1032–1044, Sep 1984.

Nordland W. A., Kazarinov R. F., Merritt F. R., Savage A., Bonner W. A., **Modified Single-Phase LPE Technique for $In_{1-x}Ga_xAs_{1-y}P_y$ Laser Structures.** Electr Lett 20(20):806–808, Sep 27, 1984.

O'Connor P., Flahive P. G., Clemetson W., Panock R. L., Wemple S. H., Shunk S. C., Takahashi D. P., **A Monolithic Multigigabit/Second DCFL GaAs Decision Circuit.** IEEE Elec D 5(7):226–227, Jul 1984.

Olson J. W., Schepis A. J., **Description and Application of the Fiber *SLC*™ Carrier System.** J Lightw T 2(3):317–322, Jun 1984.

Petrou A., Perry C. H., Smith M. C., Worlock J. M., Aggarwal R. L., Gossard A. C., Wiegmann W., **Magneto-Raman and Magneto-Photoluminescence Characterization of MQW Heterostructures.** P Soc Photo 452:51–58, Nov 9–10, 1983.

Reichmanis E., Smolinsky G., **Deep UV Positive Resists for Two-Level Photoresist Processes.** P Soc Photo 469:38–44, Mar 12–13, 1984.

Schulte H. J., **Transmission Tests During the SL Lightwave Submarine Cable System Sea Trial.** P Soc Photo 425:142–148, 1983.

Sears F. M., Cohen L. G., Stone J., **Measurements of the Axial Uniformity of Dispersion Spectra in Single-Mode Fibers.** P Soc Photo 425:56–62, 1983.

Stormer H. L., Baldwin K., Gossard A. C., Wiegmann W., **Modulation-Doped Field-Effect Transistor Based on a Two-Dimensional Hole Gas.** Appl Phys L 44(11):1062–1064, Jun 1, 1984.

Wiesenfeld J. M., Stone J., Marcuse D., **Chirp in Picosecond Semiconductor Film Lasers and Passive Pulse Compression in Optical Fibers.** P Soc Photo 439:71–78, 1983.

Yen R., Shank C. V., Fork R.L., **Femtosecond Optical Pulses and Technology.** P Soc Photo 439:2–5, 1983.

MANAGEMENT/ECONOMICS

Weiss A., Landau H. J., **Wages, Hiring Standards, and Firm Size.** J Labor Ec 2(4):477–499, Oct 1984.

PHYSICAL SCIENCES

Abrahams S. C., Ravez J., Canovet S., Grannec J., Loiacono G. M., **Phase Transitions and Ferroelectric Behavior in the $Pb_3(MF_6)_2$ Family (M = Ti, V, Cr, Fe, Ga).** J Appl Phys 55(8):3056–3060, 1984.

Alferness R. C., Divino M. D., **Efficient Fiber to X-Cut Ti: $LiNbO_3$ Waveguide Coupling for $\lambda = 1.32$ μm.** Electr Lett 20(11):465–466, May 24, 1984.

Auston D. H., Cheung K. P., Valdmanis J. A., Kleinman D. A., **Cherenkov Radiation From Femtosecond Optical Pulses in Electro-Optic Media.** Phys Rev L 53(16):1555–1558, Oct 15, 1984.

Baiocchi F. A., Wetzel R. C., Freund R. S., **Electron-Impact Ionization and Dissociative Ionization of the CD_3 and CD_2 Free Radicals.** Phys Rev L 53(8):771–774, Aug 20, 1984.

Belfiore L. A., Schilling F. C., Tonelli A. E., Lovinger A. J., Bovey F. A., **Magic Angle Spinning Carbon-13 NMR Spectroscopy of Three Crystalline Forms of Polybutene-1.** Polym Prepr 25(1):351–353, 1984.

Berreman D. W., Meiboom S., **Tensor Representation of Oseen-Frank Strain-Energy in Uniaxial Cholesterics.** Phys Rev A 30(4):1955–1959, Oct 1984.

Bishop D. J. Varma C. M., Batlogg B., Bucher E., Fisk Z., Smith J. L., **Ultrasonic Attenuation in UPt_3.** Phys Rev L 53(10):1009–1011, Sep 3, 1984.

Bondybey V. E., Haddon R. C., English J. H., **Fluorescence and Phosphorescence of 9-Hydroxyphenalenone in Solid Neon and Its Hydrogen Tunneling Potential Function.** J Chem Phys 80(11):5432–5437, Jun 1 1984.

Bondybey V. E., English J. H., **Structure of the CuO_2 and Its Photochemistry in Rare Gas Matrices.** J Phys Chem 88(11):2247–2250, May 24, 1984.

Buene L., Kaufmann E. N., Hamm R., Marra W. C., McDonald M. L., **Metastable Alloys of Beryllium Prepared by Ion Implantation.** Metall T-A 15(10):1787–1805, Oct 1984.

Capasso F., Schwartz B., Logan R. A., **Self-Aligned InP p-n Junction Diodes Fabricated With $^3He^+$ Bombardment.** IEEE Elec D 5(4):121–122, 1984.

Cava R. J., Fleming R. M., Rietman E. A., Dunn R. G., Schneemeyer L. F., **Thermally Stimulated Depolarization of the Charge-Density Wave in $K_{0.3}MoO_3$.** Phys Rev L 53(17):1677–1680, Oct 22, 1984.

Chabal Y. J., Patel C. K. N., **Infrared-Absorption in a-Si:H: First Observation of Gaseous Molecular H_2 and Si-H Overtone.** Phys Rev L 53(2):210–213, Jul 9, 1984.

Chen C.Y., Cox H. M., Garbinski P. A., Hummel S. G., **Back-Side Illuminated $Ga_{0.47}In_{0.53}As$ Photoconductive Detectors and Associated Dark Zones.** Appl Phys L 45(8):867–869, Oct 15, 1984.

Chen R. W., Nair V. N., Odlyzko A. M., Shepp L. A., Vardi Y., **Optimal Sequential Selection of N Random Variables Under a Constraint.** J Appl Prob 21(3):537–547, Sep 1984.

Chin B. H., Frahm R. E., **Germanium Doping of InP Films Grown by Liquid-Phase Epitaxy.** J Elchem So 131(10):2359–2360, Oct 1984.

Chraplyvy A. R., Stone J., **Synchronously Pumped D_2 Gas-in-Glass Fiber Raman Laser Operating at 1.56 μm.** Optics Lett 9(6):241–242, Jun 1984.

Connor J. A., Hockberger P., **A Novel Membrane Sodium Current Induced by Injection of Cyclic-Nucleotides Into Gastropod Neurons.** J Physl Lon 354(Sep):139–162, Sep 1984.

Connor J. A., Hockberger P., **Intracellular PH Changes Induced by Injection of Cyclic Nucleotides Into Gastropod Neurons.** J Physl Lon 354(Sep):163–172, Sep 1984.

Cooke W. E., Jopson R. M., Bloomfield L. A., Freeman R. R., Bokor J., **Correlations in Highly Excited Two-Electron Atoms—Planetary Behavior.** AIP Conf Pr (119):91–100, 1984.

Cullen P., Harbison J. P., Lang D. V., Adler D., **A DLTS Study of the Effects of Boron Counterdoping on the Gap States in n-Type Hydrogenated Amorphous Silicon.** Sol St Comm 50(11):991–994, Jun 1984.

Dahbura A. T., Masson G. M., **An $O(n^{2.5})$ Fault Identification Algorithm for Diagnosable Systems.** IEEE Comput 33(6):486–492, Jun 1984.

Downey P. M., Schwartz B., **Picosecond Photoconductivity in $^3He^+$ Bombarded InP.** P Soc Photo 439:30–39, 1983.

Dubois L. H., Schwartz G. P., **A High-Resolution EELS Study of Free-Carrier Variations in H_2^+/H^+ Bombarded (100)GaAs.** J Vac Sci B 2(2):101–106, Apr–Jun 1984.

Duguay M. A., Damen T. C., **Semiconductor Lasers Optically Pumped by Injection Lasers.** P Soc Photo 439:56–59, 1983.

Duncan T. M., **The Distribution of Carbon in Boron Carbide: A ^{13}C Nuclear Magnetic Resonance Study.** J Am Chem S 106(8):2270–2275, 1984.

Duncan T. M., Karlicek R. F., Bonner W. A., Thiel F. A., **A ^{31}P Nuclear Magnetic Resonance Study of InP, GaP and InGaP.** J Phys Ch S 45(4):389–391, 1984.

Dutta N. K., Olsson N. A., Tsang W. T., **Carrier Induced Refractive-Index Change in AlGaAs Quantum Well Lasers.** Appl Phys L 45(8):836–837, Oct 15, 1984.

Eaves L., Guimaraes P. S., Portal J. C., Pearsall T. P., Hill G., **High-Field Resonant Magnetotransport Measurements in Small n^{+}nn^{+} GaAs Structures: Evidence for Electric-Field-Induced Elastic Inter-Landau-Level Scattering.** Phys Rev L 53(6):608–611, Aug 6, 1984.

Fleming R. M., Muncton D. E., Axe J. D., Brown G. S., **High-*Q*-Resolution Scattering Using Synchrotron *x* Radiation: 2H-TaSe$_2$ and NbSe$_3$.** Phys Rev B 30(4):1877–1883, Aug 15, 1984.

Freeman R. R., Kincaid B. M., **Production of Coherent XUV and Soft X-Rays Using a Transverse Optical Klystron.** AIP Conf Pr (119):278–292, 1984.

Gieren A. et al., **(GE) Comparison of Naphthothiadiazines and Isoelectronic Naphthotriazines by X-Ray Structural-Analysis—X-Ray Structural Analyses of Acenaphthyleno [5.6-cd] [1.2.6] Thiadiazine, 2-Methyl-2H-Acenaphthyleno [5.6-de]-1.2.3-Triazine and Naphtho[1.8-de] Thiadiazine.** Z Naturfo B 39(7):975–984, Jul 1984.

Gilroy H. M., Chan M. G., **Effect of Pigments on the Aging Characteristics of Polyolefins.** Polym Sci T 26:273–287, 1984.

Graedel T. E., **Effects of Below-Cloud Gas Scavenging on Raindrop Chemistry Over Remote Ocean Regions.** Atmos Envir 18(9):1835–1842, 1984.

Greenblatt M., Nair K. R., McCarroll W. H., Waszczak J. V., **Electrical Conductivity and Magnetic Susceptibility of Rutile Type CrVNbO$_6$, FeVNbO$_6$ and Ni-V$_2$Nb$_2$O$_{10}$.** Mater Res B 19(6):777–782, Jun 1984.

Greywall D. S., **Thermal Conductivity of Normal Liquid ^{3}He.** Phys Rev B 29(9):4933–4945, 1984.

Gross B., Von Seggern H., West J. E., **Positive Charging of Fluorinated Ethylene Propylene Copolymer (Teflon) by Irradiation With Low-Energy Electrons.** J Appl Phys 56(8):2333–2336, Oct 15, 1984.

Hauser J. J. et al., **Structure-Sensitive Magnetic Properties of Ni-Mn Alloys.** Phys Rev B 30(7):3803–3807, Oct 1, 1984.

Heritage J. P., Sermage B., Martinez D. E., **Photoexcited Carrier Lifetime and Auger Recombination in 1.3 Micron Bandgap InGaAsP.** P Soc Photo 439:14–17, 1983.

Huse D. A., **Exact Exponents for Infinitely Many New Multicritical Points.** Phys Rev B 30(7):3908–3915, Oct 1, 1984.

Inoue A., Okamoto S., Masumoto T., Chen H. S., **Effect of Cold Rolling on the Superconducting and Electronic Properties of Two Amorphous Alloys; Nb$_{50}$Zr$_{35}$Si$_{15}$ and Nb$_{70}$Zr$_{15}$Si$_{15}$.** J Mater Sci 19(4):1251–1260, 1984.

Jackel L. D., Howard R. E., Mankiewich P. M., Craighead H. G., Epworth R. W., **Beam Energy Effects in Electron-Beam Lithography—The Range and Intensity of Backscattered Exposure.** Appl Phys L 45(6):698–700, Sep 15, 1984.

Jackson D. P., Buck T. M., Wheatley G. H., **Atom Layer Effects in the Scattering of keV Ne From Cu$_3$Au(100).** Nucl Inst B 230(1–3):440–443, 1984.

Jackson S. A., Peeters F. M., **Magnetic-Field Detrapping of Polaronic Electrons on Films of Liquid Helium.** Phys Rev B 30(8):4196–4202, Oct 15, 1984.

Jelinski L. W., Dumais J. J., Engel A. K., **Solid State ^{2}H NMR Studies of Molecular Motion—Poly(butylene terephthalate) and Poly(butylene terephthalate)-Containing Segmented Copolymers.** ACS Symp S (247):55–65, 1984.

Jelinski L. W., Dumais J. J., Luongo J. P., Cholli A.L., **Thermal-Oxidation and Its Analysis at Low Levels in Polyethylene.** Macromolec 17(9):1650–1655, Sep 1984.

Jopson R. M., Freeman R. R., Cooke W. E., Bokor J., **Two-Photon Spectroscopy of 7 sn'd Autoionizing States of Barium.** Phys Rev A 29(6):3154–3158, Jun 1984.

Joy D. C., **A Parametric Partial Cross Section for ELS.** J Microsc D 134(Apr):89–92, 1984.

Kamgar A., Fichtner W., Sheng T. T., Jacobson D. C., **Junction Leakage Studies in Rapid Thermal Annealed Diodes.** Appl Phys L 45(7):754–756, Oct 1, 1984.

Kaminow I. P., **Polarization-Maintaining Fibers.** Appl Sci Re 41(3–4):257–270, 1984.

Karp B. C., Ludwig E. J., Thompson W. J., **Alpha-Particle D-State Components From (D, Alpha) Analyzing Powers.** Phys Rev L 53(17):1619–1622, Oct 22, 1984.

Kelber J. A. et al., **Photon-Stimulated Desorption of Solid Neopentane.** Phys Rev B 30(8):4748–4752, Oct 15, 1984.

Kuck V., **Critical Temperature for Solubility of a Phenolic Antioxidant.** Polym Sci T 26:103–110, 1984.

Lagowski J., Lin D. G., Gatos H. C., Parsey J. M., Kaminska M., **Real and Apparent Effects of Strong Electric Fields on the Electron Emission From Midgap Levels EL2 and EL0 in GaAs.** Appl Phys L 45(1):89–91, Jul 1, 1984.

Lake G., Schommer R. A., **A Successful Survey of H-I in Low-Luminosity Elliptical Galaxies.** Astrophys J 280(1):107–116, May 1, 1984.

Lambert W. R., Trevor P. L., Doak R. B., Cardillo M. J., **Inelastic Helium Scattering From Ag(001) and Ag(001)*c* (2 × 2) Cl.** J Vac Sci A 2(2):1066–1068, Apr–Jun 1984.

Larson R. G., **A Constitutive Equation for Polymer Melts Based on Partially Extending Strand Convection.** J Rheol 28(5):545–571, Oct 1984.

Lawrence J. M., Thompson J. D., Fisk Z., Smith J. L., Batlogg B., **x-P-T Phase Diagram for the γ-α Transition in $Ce_{0.9\text{-}x}La_xTh_{0.1}$ Alloys.** Phys Rev B 29(7):4017–4025, 1984.

Lin B. J. F., Tsui D. C., Paalanen M. A., Gossard A. C., **Mobility of the Two-Dimensional Electron Gas in GaAs-$Al_xGa_{1\text{-}x}As$ Heterostructures.** Appl Phys L 45(6):695–697, Sep 15, 1984.

Liou K. Y., Burrus C. A., Linke R. A., Kaminow I. P., Granlund S. W., Swan C. B., Besomi P., **Single-Longitudinal-Mode Stabilized Graded-Index-Rod External Coupled-Cavity Laser.** Appl Phys L 45(7):729–731, Oct 1, 1984.

Lovinger A. J., Johnson G. E. Bair H. E., Anderson E. W., **Stuctural, Dielectric, and Thermal Investigation of the Curie Transition in a Tetrafluoroethylene Copolymer of Vinylidene Fluoride.** J Appl Phys 56(9):2412–2418, Nov 1, 1984.

Lucchese R. R., Tully J. C., **Trajectory Studies of Vibrational Energy Transfer in Gas-Surface Collisions.** J Chem Phys 80(7):3451–3462, 1984.

Lyons K. B., Sturge M. D., Greenblatt M., **Low-Frequency Raman-Spectrum of $ZrSiO_4$: V^{4+}: An Impurity-Induced Dynamical Distortion.** Phys Rev B 30(4):2127–2132, Aug 15, 1984.

Marcucella H., Munro I., MacDonald J. S., **Patterns of Ethanol Consumption as a Function of the Schedule of Ethanol Access.** J Pharm Exp 230(3):658–664, Sep 1984.

Martinez O. E., Gordon J. P., Fork R. L., **Negative Group-Velocity Dispersion Using Refraction.** J Opt Soc A 1(10):1003–1006, Oct 1984.

Masnovi J. M., Huffman J. C., Kochi J. K., Hilinski E. F., Rentzepis P. M., **Picosecond Spectroscopy of Charge-Transfer Processes. Photochemistry of Anthracene-Tetranitromethane EDA Complexes.** Chem P Lett 106(1–2):20–25, 1984.

McAfee K. B., Walker K. L., Laudise R. A., Hozack R. S., **Dependence of Equilibria in the Modified Chemical Vapor-Deposition Process on $SiCl_4$, $GeCl_4$, and O_2.** J Am Ceram 67(6):420–424, Jun 1984.

McCall D. W., Douglass D. C., Blyler L. L., Johnson G. E., Jelinski L. W., Bair H. E., **Solubility and Diffusion of Water in Low-Density Polyethylene.** Macromolec 17(9):1644–1649, Sep 1984.

Miller B., **Charge-Transfer and Corrosion Processes at III-V Semiconductor Electrolyte Interfaces.** J Elec Chem 168(1–2):91–100, Jun 25, 1984.

Miller B., Rosamilia J. M., **Maximum Power Spectroscopy for Photovoltaic Devices.** J Elchem So 131(10):2266–2271, Oct 1984.

Miller T. A., Suzuki T., Hirota E., **High-Resolution, cw Laser Induced Fluorescence Study of the $A\ {}^2\pi_u - X\ {}^2\Sigma_g^+$ System of N_2^+.** J Chem Phys 80(10):4671–4678, May 15, 1984.

Murarka S. P., **Phosphorus Out-Diffusion During High-Temperature Anneal of Phosphorus-Doped Polycrystalline Silicon and SiO_2.** J Appl Phys 56(8):2225–2230, Oct 15, 1984.

Nakahara S., **Detection of Gas-Bubbles and Organic-Molecules Included in**

Electrodeposited Films by Transmission Electron Microscopy. J Elchem Sc 131(10):2246–2250, Oct 1984.

Nakahara S., Felder E. C., Temkin H., **Characterization of Near-Surface Line Defects Formed During High-Temperature Annealing of Gold-Metallized III-V-Compound Semiconductors (InP and GaAs).** J Elchem So 131(8):1917–1920, Aug 1984.

Ogawa S., Lee T. M., **The Relation Between the Internal Phosphorylation Potential and the Proton Motive Force in Mitochondria During ATP Synthesis and Hydrolysis.** J Biol Chem 259(16):1004–1011, Aug 25, 1984.

Olsson N. A., Dutta N. K., Besomi P., Shen T. M., Nelson R.J., Linke R. A., Tucker R. S., **Two Gbit/s Operation of Single-Longitudinal-Mode 1.5 μm Double-Channel Planar Buried-Heterostructure C^3 Lasers.** Electr Lett 20(10):395–397, May 10, 1984.

Paalanen M. A., Hebard A. F., **A Criterion for the Determination of Upper Critical Fields in Highly Disordered Thin-Film Superconductors.** Appl Phys L 45(7):794–796, Oct 1, 1984.

Panish M. B., Sumski S., **Gas Source Molecular Beam Epitaxy of $Ga_xIn_{1-x}P_yAs_{1-y}$.** J Appl Phys 55(10):3571–3576, 1984.

Panish M. B., Temkin H., **GaInAsP InP Heterostructure Lasers Emitting at 1.5-μm and Grown by Gas Source Molecular Beam Epitaxy.** Appl Phys L 44(8):785–787, 1984.

Patel D. J., Kozlowski S. A., Ikuta S., Itakura K., **Deoxyadenosine-Deoxycytidine Pairing in the d(C-G-C-G-A-A-T-T-C-A-C-G) Duplex: Conformation and Dynamics at and Adjacent to the dA.dC Mismatch Site.** Biochem 23(14):3218–3226, Jul 3,1984.

Patel D. J., Kozlowski S. A., Ikuta S., Itakura K., **Deoxyguanosine-Deoxyadenosine Pairing in the d(C-G-A-G-A-A-T-T-C-G-C-G) Duplex: Conformation and Dynamics at and Adjacent to the dG.dA Mismatch Site.** Biochem 23(14):3207–3217, Jul 3,1984.

Patel D. J., Kozlowski S. A., Ikuta S., Itakura K., **Dynamics of DNA Duplexes Containing Internal G·T, G·A, A·C, and T·C Pairs: Hydrogen Exchange at and Adjacent to Mismatch Sites.** Fed Proc 43(11):2663–2670, Aug 1984.

Petroff P. M., Gossard A. C., Wiegmann W., **Structure of AlAs-GaAs Interfaces Grown on (100) Vicinal Surfaces by Molecular-Beam Epitaxy.** Appl Phys L 45(6):620–622, Sep 15, 1984.

Raghavachari K. et al., **Theoretical Study of Silylene Insertion Into N-H, O-H, F-H, P-H, S-H, and CL-H Bonds.** J Am Chem S 106(20):5853–5859, Oct 3, 1984.

Rowe H. E., **Waves With Random Coupling and Random Propagation Constants.** Appl Sci Re 41(3-4):237–255, 1984.

Schilling F. C., Bovey F. A., Tonelli A. E., Tseng S., Woodward A. E., **Solid-State Carbon-13 NMR Study of the Fold Surface of Solution-Grown 1,4-Trans-Polybutadiene Crystals.** Macromolec 17(4):728–733, 1984.

Schneemeyer L. F., Disalvo F. J., Spengler S. E., Waszczak J. V., **Dramatic Impurity Effects on the Charge-Density Wave in Potassium Molybdenum Bronze.** Phys Rev B 30(8):4297–4301, Oct 15, 1984.

Schwenk H., Hess E., Andres K., Wudl F., Aharon-Shalom E., **Isotope Effect in the Organic Superconductor $(TMTSF)_2ClO_4$.** Phys Lett A 102(1-2):57–60, Apr 30, 1984.

Silfvast W. T., Wood O. R., Macklin J. J., Lundberg H., **Photoionization Lasers Pumped by Broad-Band Soft-X-Ray Radiation From Laser-Produced Plasmas.** AIP Conf Pr (119):427–436, 1984.

Starnes W. H. et al., **Mechanism of Polyvinyl-Chloride, Fire Retardance by Molybdenum(VI) Oxide—Further Evidence in Favor of the Lewis Acid Theory.** Polym Sci T 26:237–248, 1984.

Stern M. B., Harrison T. R., Archer V. D., Liao P. F., Bean J. C., **Raman-Spectroscopic Analysis of the CaF_2-Si Heterostructure Interface.** Sol St Comm 51(4):221–224, Jul 1984.

Tarascon J. M., Hull G. W., Disalvo F. J., **A Facile Synthesis of Pseudo One-Monodimensional Ternary Molybdenum Chalcogenides $M_2Mo_6X_6$ (X = Se, Te; M = Li, Na..Cs).** Mater Res B 19(7):915–924, Jul 1984.

Taylor G. N., Stillwagon L. E., Venkatesan T., **Gas-Phase-Functionalized Plasma-**

Developed Resists: Initial Concepts and Results for Electron-Beam Exposure. J Elchem So 131(7):1658–1664, Jul 1984.

Tewksbury S. K., Biazzo M. R., Lindstrom T. L., Tennant D. M., **Depletion Layer Formation Rate at T < 30 °K in Buried Channel, Metal-Oxide-Semiconductor Transistors.** J Appl Phys 56(2):517–521, Jul 15, 1984.

Tinubu S. O., Gupta K. C., **Optimal Synthesis of Function Generators Without the Branch Defect.** J Mech Tran 106(3):348–354, Sep 1984.

Tonelli A. E., **^{13}C-NMR Chemical Shifts and the Conformations of Rigid Polypeptides.** Biopolymers 23(4):819–829, 1984.

Tonelli A. E., Schilling F. C., **Carbon-13 NMR Chemical Shifts and the Microstructure of Propylene-Vinyl Chloride Copolymer With Low Propylene Content.** Polym Prepr 25(1):334–335, 1984.

Van Saarloos W., Weeks J. D., **Surface Undulations in Explosive Crystallization—A Nonlinear Analysis of a Thermal Instability.** Physica D 12(1–3):279–294, Jul 1984.

Venkatesan T., Brown W. L., Wilkens B. J., Reimann C. T., **Secondary Electron, Ion and Photon Emission During Ion Beam Irradiation of Polymer and Condensed Gas Films.** Nucl Inst B 229(2–3):605–609, 1984.

Vieira N. D., Mollenauer L. F., Szeto L. H., **Optical Properties of the In^0(1) Center in KCl.** Sol St Comm 50(12):1037–1041, Jun 1984.

Vonseggern H., Wang T. T., **Polarization Behavior During High-Field Poling of Poly(vinylidene Fluoride).** J Appl Phys 56(9):2448–2452, Nov 1, 1984.

Warren W. W., Sotier S., Brennert G. F., **Localization of Electrons in Ionic Liquids: Nuclear Magnetic Resonance in Cs-CsI and CsI-I Solutions.** Phys Rev B 30(1):65–77, Jul 1, 1984.

Weiss M. A., Patel D. J., Sauer R. T., Karplus M., **^{1}H-NMR Study of the λ Operator Site O_L1: Assignment of the Imino and Adenine H2 Resonances.** Nucl Acid R 12(9):4035–4047, May 11, 1984.

Wertheim G. K. et al., **Line-Shapes in Surface-Atom Core-Level Photoemission From Ta(111), W(111), and W(100).** Phys Rev B 30(8):4343–4347, Oct 15, 1984.

White J. C., Craighead H. G., Howard R. E., Jackel L. D., Wood O. R., **VUV Laser Photolithography.** AIP Conf Pr (119):324–329, 1984.

Whitt W., **Minimizing Delays in the *GI*/*G*/1 Queue.** Oper Res 32(1):45–51, Jan–Feb 1984.

Wolf T. M., Taylor G. N., Venkatesan T., Kraetsch R. T., **The Scope and Mechanism of New Positive Tone Gas-Phase-Functionalized Plasma-Developed Resists.** J Elchem So 131(7):1664–1670, Jul 1984.

Yin M. T., **Si-III (BC-8) Crystal Phase of Si and C: Structural Properties, Phase Stabilities, and Phase Transitions.** Phys Rev B 30(4):1773–1776, Aug 15, 1984.

Yu C. C., Anderson P. W., **Local-Phonon Model of Strong Electron Phonon Interactions in *A*15 Compounds and Other Strong-Coupling Superconductors.** Phys Rev B 29(11):6165–6186, Jun 1, 1984.

SOCIAL AND LIFE SCIENCES

Poltrock S. E., Brown P., **Individual Differences in Visual Imagery and Spatial Ability.** Intelligenc 8(2):93–138, Apr–Jun 1984.

SPEECH/ACOUSTICS

Tartter V. C., **Laterality Differences in Speaker and Consonant Identification in Dichotic Listening.** Brain Lang 23(1):74–85, Sep 1984.

Cohen L. G., Pearson A. D., **A Systematic Approach to Fabricating Single-Mode Lightguides.** P Soc Photo 425:28–32, 1983.

CONTENTS, MARCH 1985

ASSURING HIGH RELIABILITY OF LASERS AND PHOTODETECTORS FOR SUBMARINE LIGHTWAVE CABLE SYSTEMS

ERRATA

AT&T Bell Laboratories Technical Journal, Vol. 63, No. 8, Part 2, “The *UNIX* System.”

Page	*Line*	*Correction*
1597	40	Footnote: "Systems III and V are . . ."
1599	17	`cat 'cat filelist'`
1600	8	″ should be "
1617	38	`watch fig1.pic|pic|troff|proof`
1620	44	Last character on page should be part of first word on next
1630	44	9. D. M. Ritchie, “The *UNIX* System: A Stream Input-Output System,” AT&T Bell Lab. Tech. J., this issue.
1655	11	`"rxw"` => `"rwx"`
1662	37	Delete `"remote"`
1671	27	7. R. T. Morris, “Another Try at Uucp,” AT&T Bell Laboratories Computing Science Technical Report No. 111.
1710	14, 15	Add */ to lines 4, 5 of "class date" display
1715	25	`"node *1"` (one) => `"node *l"` (ell)
1719	25	Add space after `T&`
1756	11	Add "basis" at end of first sentence
1814	7	"Steven" => "Stephen"
1836	Figure caption	"32 elements" => "3 elements"
1840	1	Space before, but not after + and − signs
1842	24, 25	`awk −F'b̸' '{printf("%s... )}'`

AT&T TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development*, and *Solid State Abstracts Journal*. Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

AT&T