THE BELL SYSTEM TECHNICAL JOURNAL

JANUARY 1983
VOL. 62, NO. 1 PART 1

**THE BELL SYSTEM TECHNICAL JOURNAL** (ISSN0005-8580) is published by the American Telephone and Telegraph Company, 195 Broadway, N.Y., N.Y. 10007, C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; T. O. Davis, Secretary.

The Journal is published in three parts. Part 1, general subjects, is published ten times each year. Part 2, Computing Science and Systems, and Part 3, single-subject issues, are published with Part 1 as the papers become available.

The subscription price includes all three parts. Subscriptions: United States—1 year $35; 2 years $63; 3 years $84; foreign—1 year $45; 2 years $73; 3 years $94. Subscriptions to Part 2 only are $10 ($11 Foreign). Single copies of the Journal are available at $5 ($6 foreign). 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 Bell System Technical Journal and sent to Bell Laboratories, Circulation Dept., Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.





Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room 1J-319, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

Printed in U.S.A. Second-class postage paid at Short Hills, N.J. 07078 and additional mailing offices. Postmaster: Send address changes to The Bell System Technical Journal, 101 J. F. Kennedy Parkway, Short Hills, N.J. 07078

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 62 January 1983 Number 1



# InGaAsP LEDs for 1.3-μm Optical Transmission

By H. TEMKIN, C. L. ZIPFEL, M. A. DIGIUSEPPE, A. K. CHIN, V. G. KERAMIDAS, and R. H. SAUL

(Manuscript received May 26, 1982)

*High-radiance InGaAsP double heterostructure light-emitting diodes (LEDs) that are fast and reliable have been developed for 1.3-μm optical transmission applications. These devices are prepared by liquid phase epitaxy on InP substrates. Improved growth procedures result in large-area wafers (up to 5.4 $cm^2$) with excellent uniformity and high device yields. Optical power of ~120 μW can be launched into 0.29-NA, 62.5-μm core fiber. On the basis of accelerated reliability tests the median life of these InGaAsP LEDs is expected to exceed 4 × $10^6$ h at 70°C and current densities up to 40 kA/$cm^2$. The projected failure rate owing to output power degradation (−1 dB) is below 1 FIT. These devices are expected to be useful in lightwave systems operating at data rates ≤140 Mb/s.*

## I. INTRODUCTION

The present generation of optical transmission systems employs GaAlAs laser emitters operating at 0.8 to 0.9 μm.[1] In this wavelength range fiber attenuation (about 4 dB/km) and especially chromatic dispersion limit the usefulness of light-emitting diodes (LEDs) to short data links of few kilometers. By shifting the LED emission wavelength close to 1.3 μm two properties of the fused silica fiber permit much longer distances and higher data rates. First, the fiber loss is greatly reduced at longer wavelength,[2] with cabled fiber loss below 1 dB/km possible. Second, pulse spreading owing to material dispersion is reduced as $d^2n/d\lambda^2$ of the refractive index approaches zero at 1.27 to 1.30 μm.[3] Thus, 1.3-μm LED-based transmission systems with high

data rates and repeater distances in excess of 10 km are feasible.[4] Indeed, transmission experiments have demonstrated repeaterless LED-based operation over a distance of 24 km at the DS3 rate (44.7 Mb/s) and 7 km at rates as high as DS4 (274 Mb/s).[5] Such LED systems are attractive alternatives to laser systems since light-emitting diodes operate over a wide temperature range with unequalled reliability and their drive electronics is very simple.

These advantages of longer wavelength operation can be realized for devices fabricated out of the quaternary alloy $In_{1-x}Ga_xAs_yP_{1-y}$.[6] These alloys can be grown exactly lattice-matched to InP substrates (for $y/x \approx 2.2$) with compositions ranging from 1.1 $\mu$m to 1.65 $\mu$m.[7] These quaternary LEDs are efficient, fast, and very reliable.[8–11] In addition, very sensitive InGaAsP detectors have been demonstrated,[12] offering a possibility of a common material technology for both emitting and receiving devices.

In this paper we describe the development of state of the art InGaAsP LEDs for optical transmission applications. The paper is divided into six sections. Following the introduction, Section II describes material preparation procedures leading to double-heterostructure wafers with a large area and high quality. Device fabrication is discussed in Section III, with particular attention devoted to large-scale fabrication techniques. In these two sections a description of general procedures is followed by more detailed descriptions of some of the more challenging problems. LED performance in terms of power launched into fibers, speed, and temperature characteristics is discussed in Section IV. Section V describes accelerated aging tests and reliability estimates. These results are discussed and summarized in Section VI.

## II. MATERIAL PREPARATION

Double heterostructure wafers, from 1 $cm^2$ to 5.4 $cm^2$ in area, are grown in a multiple-well, graphite boat using a two-phase growth technique.[7] InP substrates are polished to an optically smooth finish with a ~1-percent bromine-methanol solution. All the melts used are first equilibrated at 675°C for 16 hours in a $H_2$ ambient and then quenched. These prehomogenized melts, containing excess InP and Sn dopant where applicable, are then placed in the liquid phase epitaxy (LPE) boat and again equilibrated at 675°C for an additional hour, as schematically shown in Fig. 1. Zinc is added prior to this final equilibration step. During this warm-up and melt homogenization cycle the thermal decomposition of the substrate is minimized by either a cover piece or phosphorus vapor.[13]

Prior to the growth of the first epitaxial InP layer at 650°C, the substrate is subjected to a shallow In etch, which removes about 5 $\mu$m

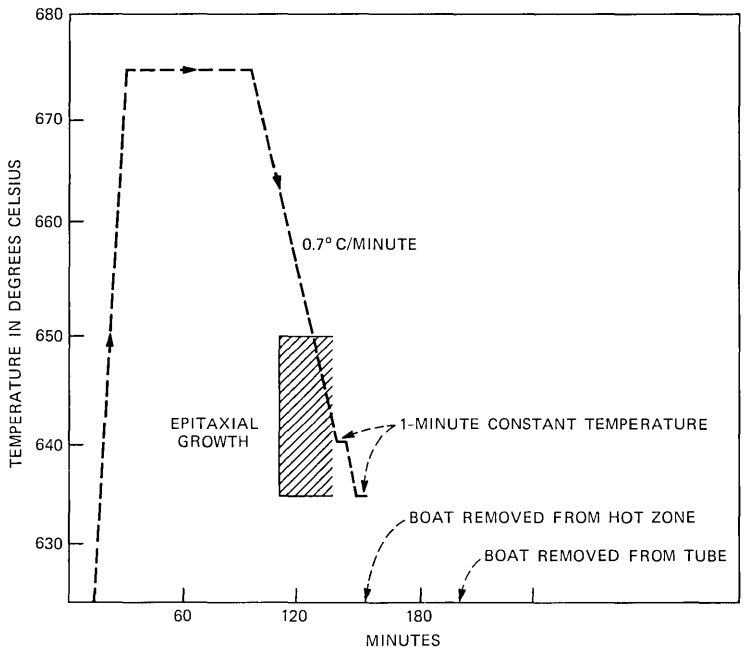



Fig. 1—Schematic diagram of melt homogenization and LPE growth cycle.

of the thermally degraded material. Following the growth of the Sn-doped InP buffer layer at a cooling rate of 0.7°C/min, a nominally undoped quaternary layer is grown isothermally at 640°C. Next, a Zn-doped InP confining layer is grown (again at 0.7°C/min), followed by a Zn-doped quaternary, or the InGaAs ternary, contact layer grown isothermally at 630°C. The InP buffer and confining layers are approximately 5 and 2 $\mu$m thick, respectively. The thickness of the InGaAsP active layer is varied between 0.5 and 1.4 $\mu$m, while the contact layer thickness is kept at ~0.5 $\mu$m.

Since the defects present in the substrate or epitaxial layer are reproduced in the next epitaxial layer,[14] substrate quality is of great importance. InP substrates used in this work have been cut in the ⟨100⟩ direction from the ⟨111⟩ pulled, twin-free, boules grown by the Liquid Encapsulated Czochralski (LEC) technique.[15] The substrates are doped $n$-type with either Sn or S up to the doping level of $\sim 5 \times 10^{18}$ $cm^{-3}$. The sulfur-doped material is preferred for reasons of superior macroscopic perfection and higher resistance to thermal damage.[16] The dislocation density in our S-doped wafers, as established by etch pitting and TCL,[17] is typically below $5 \times 10^3$ $cm^{-3}$, as much as an order of magnitude less than in Sn-doped wafers. This reduction has been attributed to lattice hardening.[18] Although the presence of threading

dislocations has been shown to greatly degrade reliability of GaAlAs LEDs,[19] we show in Section V that dislocation density has no significant effect on quaternary reliability. The influence of dislocations on device performance remains to be investigated.

The thermal damage to InP substrate prior to growth and the resulting damage to iso-epitaxial layers are by now well recognized.[20] The damage owing to loss of phosphorus ranges from the large dissociation pits to increased density of phosphorus vacancy-impurity complexes in InP layers,[13] and results in reduced device performance and yield. A number of techniques have been developed to control and eliminate the thermal decomposition, all attempting to provide phosphorus overpressure in the vicinity of the substrate. Commonly used are cover-piece protection,[8] In-Sn-P solution,[21] addition of $PH_3$,[22,23] or elemental phosphorus at elevated temperature to the growth ambient.[13] Deep, ~25-$\mu$m, In meltback also has been used to remove all the degraded material. We have evaluated these protection measures by means of spatially resolved photoluminescence and demonstrated that thermal degradation is present even on morphologically perfect surfaces.[20] A photoluminescence image of a thermally degraded substrate is shown in Fig. 2a. The damage manifests itself as spots of low luminescence efficiency. Iso-epitaxial layers grown on such a substrate show similar damaged regions and an increased dislocation density. This damage depends on the substrate dopant, and is greatly reduced, as indicated by photoluminescence (PL) intensity measurements, in S-doped InP, as compared to Sn-doped wafers. Elimination of thermal damage is achieved through a shallow, ~5-$\mu$m deep, In meltback of a

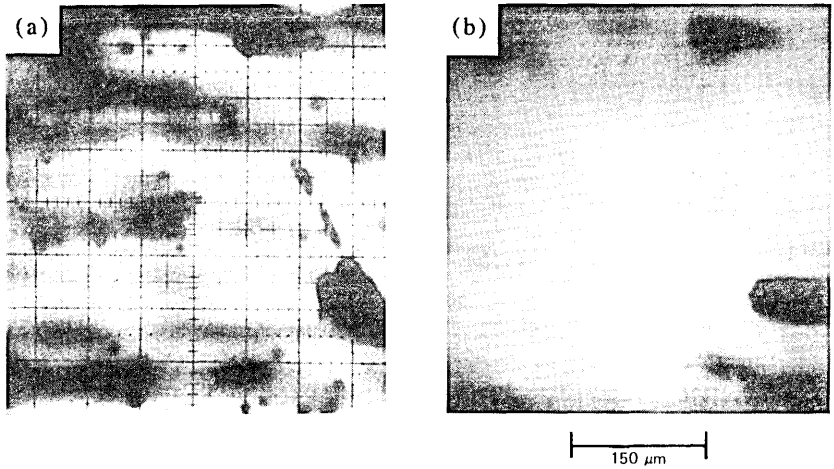



Fig. 2—Spatially resolved photoluminescence map of: (a) InP substrate subjected to a high temperature treatment, and (b) InP buffer layer grown on a well-protected substrate.

previously protected substrate, as shown by the uniform and bright PL scan of the InP buffer layer in Fig. 2b.

The second to grow is the the active layer of $In_{0.7}Ga_{0.3}As_{0.64}P_{0.36}$, a composition corresponding to the 1.3-$\mu$m emission wavelength. Particular attention is paid to the reproducible growth of layers with that composition, precisely lattice-matched to the InP substrate. In particular, a close lattice match, $\Delta a/a$ smaller than $6 \times 10^{-4}$, is required to prevent formation of misfit dislocations.[24] These act as nonradiative recombination centers and thus reduce luminescence efficiency. A transmission cathodoluminescence (TCL) image of misfit dislocations forming a characteristic criss-cross pattern on [100] oriented layer is shown in Fig. 3. This InGaAsP layer had a lattice mismatch of $1 \times 10^{-3}$.

In our experience, epitaxial layer morphology is not affected by lattice mismatch; layers of 1.3-$\mu$m InGaAsP with $\Delta a/a \approx 2.4 \times 10^{-3}$ have been grown with excellent morphology. Thus, it is necessary to monitor lattice match with X-ray diffractometry. The lattice constant can be well-controlled by precise weighing of the Ga and As components (up to $\pm 25$ $\mu$grams in a 10-gm melt) and careful reproduction of the growth temperature. Phosphorus loss in the equilibration cycle is compensated by the presence of excess InP. Accurate temperature control during the active layer growth results in reduced compositional

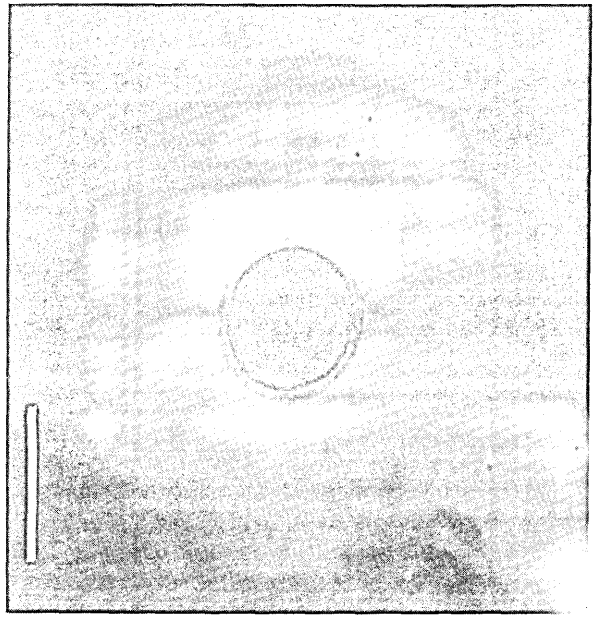


Fig. 3—Transmission cathodoluminescence (TCL) image of ⟨100⟩ misfit dislocations in a wafer with $\Delta a/a = 0.10$ percent.

grading. Compositional grading is especially pronounced for alloy compositions close to 1.3 $\mu$m.[25] The degree of compositional uniformity, obtained on a fairly routine basis, is demonstrated in a double X-ray spectrometer scan shown in Fig. 4. Double X-ray technique is used to fully resolve X-ray reflection width. The stronger peak originates in the InP buffer layer and the substrate, a much weaker and broader one, is due to the quaternary layer. The width of the InGaAsP reflection in this layer, ~60s of arc, is attributed to compositional grading in the direction of growth, i.e., along the thickness of the epitaxial layer. Larger grading results in broadening of the LED emission line width. In comparison, the width of the InP reflection is ~10s of arc, indicating the absence of any broadening because of thermal damage in the substrate.

The most common defects present in the epitaxial layers of InP and InGaAsP are inclusions of In. These originate at either meltback or wipe-off stages of the growth cycle and greatly reduce device yield and

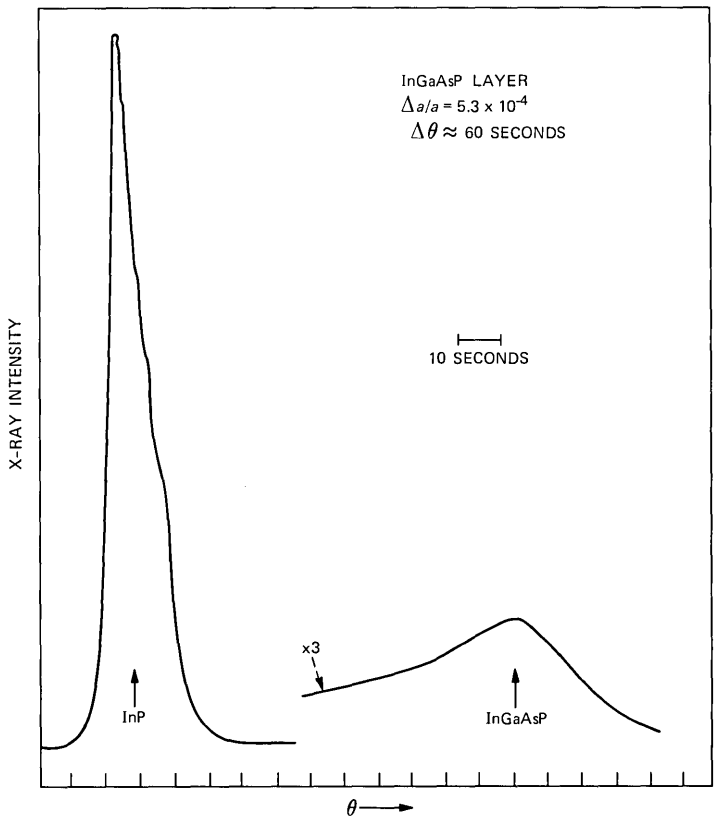



Fig. 4—Double X-ray diffraction scan of a device wafer. The width of the InGaAsP peak reveals a low degree of compositional grading.

reliability. A dramatic example of the wipe-off-related problem is shown in Fig. 5a. The InGaAsP active layer imaged in this Nomarski micrograph has been grown on a buffer layer that was not completely wiped off. The In inclusions left at the buffer surface locally prevent InGaAsP growth and produce holes in the active layer. This problem can be controlled by careful graphite boat design and maintenance of stringent wipe-off tolerances. Similarly, it is important to avoid melt contamination with particulate matter, such as carbon particles generated during prolonged equilibration baking. Figure 5b shows a more typical Nomarski image of the active layer. No defects were present throughout the entire 2.7-cm$^2$ wafer, resulting in excellent LED performance and uniformity. The horizontal lines visible in this image are growth lamellae characteristic of LPE-grown layers.

The most often used *p*-type dopant for the InP and InGaAsP layers is Zn. This dopant suffers from two well-known problems: high distribution coefficient and very high vapor pressure.[26] These can result in the contamination of the neighboring melts with concomitant junction misplacement into the buffer layer. This type of contamination is readily controlled by covering the melt wells and reducing the amount of Zn added. The undoped active layer is in fact grown as *n*-type material, and formation of the *p-n* junction is obtained by diffusion of Zn from the *p*-InP confining layer, as shown schematically in Fig. 6. To keep the *p-n* junction within the active layer, the amount of Zn in the InP confining layer must be kept low. However, to assure adequate electrical contact to the *p*-InP layer its doping level must be as high as

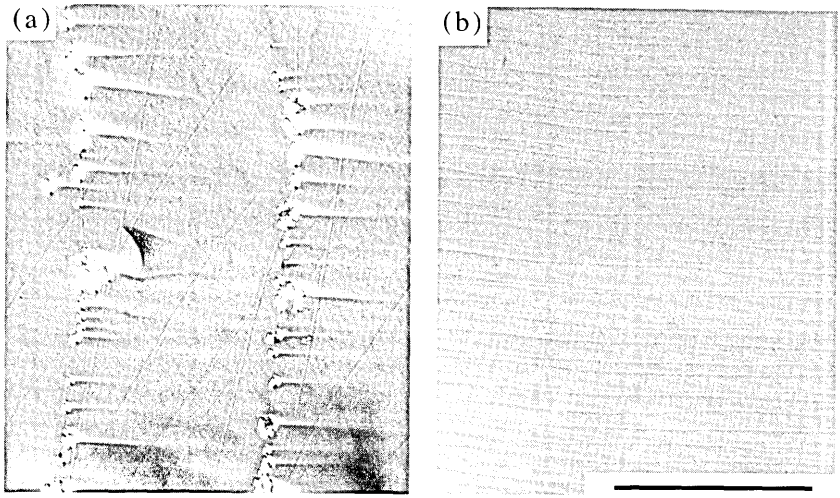



Fig. 5—A Nomarski micrograph (black bar = 100 μm) of the exposed InGaAsP active layer showing: (a) In inclusions owing to an incomplete wipe-off, and (b) inclusion-free surface of properly wiped-off layer.

possible. These conflicting requirements can be reconciled by reducing the doping level in the confining layer to $5 \times 10^{17}$ cm$^{-3}$ and adding a highly doped, $p \sim 2 \times 10^{19}$ cm$^{-3}$, InGaAsP contact layer. Thus, diffusion out of the contact and confining layers is used to form the *p-n* junction, the position of which can be well-controlled within the active layer. The influence of the junction position on LED performance is discussed in Section IV. In addition, contact resistance to the InGaAsP layer is found to be lower than that possible to the InP layer, as discussed in the LED fabrication section.

## III. DEVICE FABRICATION

To obtain high-power, reliable LEDs, a great deal of attention must be devoted to processing procedures. A sequence of processing steps developed for the large-scale manufacture of the 1.3-$\mu$m InGaAsP LED is described in this section, with particular attention given to stress reduction and yield improvement. A typical processing flow chart is shown in Fig. 7. Device fabrication starts with the thinning of the LPE wafer from ~400 to approximately 100 $\mu$m. This is done by chemical-mechanical polishing with a bromine-methanol solution of less than 1

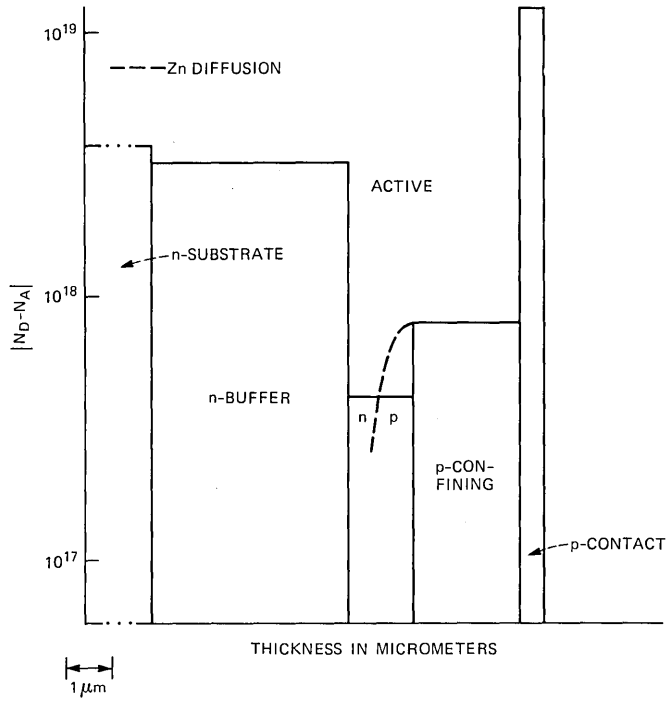



Fig. 6—A schematic diagram of the wafer structure indicating doping types and the Zn diffusion front.

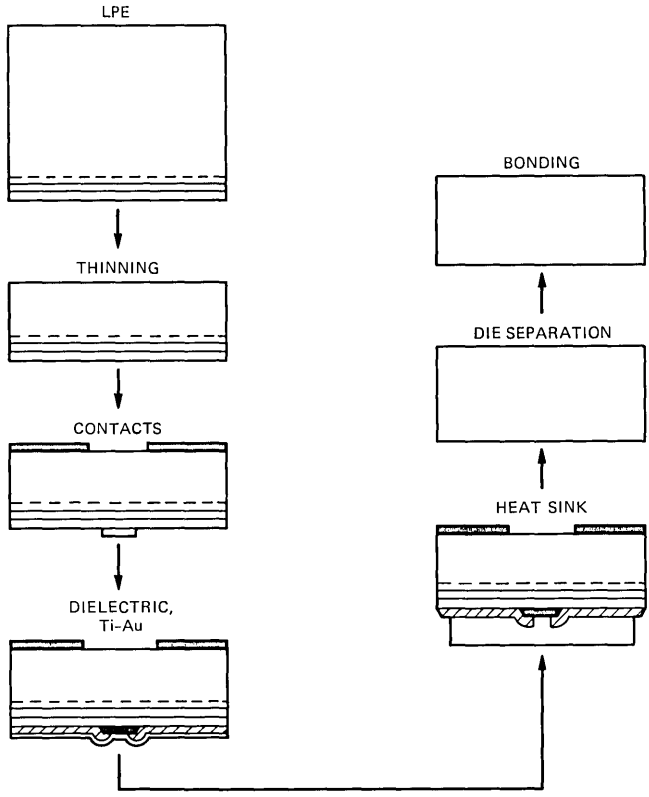



Fig. 7—Schematic processing diagram for a dielectric-isolated LED.

percent. Preliminary mechanical grinding is avoided to prevent formation of deep scratches and to minimize stresses. The finished substrate side should be mirror-like and free of polishing scratches since these affect light emission. After thinning, contacts on both sides are evaporated through a matched pair of shadow masks. Prior to contact deposition, careful cleaning with standard solvents followed by a 1-minute etch in 1:1 $HF/H_2O$ solution are performed. The last etch is important for removal of surface oxides to ensure contact adherence and reproducibility of contact resistance. All the contacts are deposited by *e*-gun evaporation in vacuum of better than $10^{-6}$ torr. The *p*-contact metallization consists of an 800Å-thick layer of 1 wt-percent Be in Au alloy, deposited from an alloy source of the same composition, followed by a 2100Å-thick layer of Au. The BeAu alloy is prepared in advance and maintains its composition owing to nearly congruent evaporation of these two metals. The *n*-contact consists of a "sandwich" of 2000Å of Au, 500Å of Sn, and another overlay of gold at least 10 kÅ thick. The metallizations are alloyed in a $N_2$-$H_2$ flow and a fixed alloying time

of 8 minutes at 420°C is used. The order of metal layers in the *n*-contact is essential for low contact resistance and good bondability. When the first layer of Au is omitted and Sn is deposited directly on the semiconductor, contact morphology and adherence degrade owing to the well-known Sn-Au interdiffusion.[27] In contrast, the layered metallization results in a smooth contact of excellent uniformity and bondability.

Following ohmic contact deposition a dielectric film is placed on the *p*-surface. This film of plasma-deposited $SiO_2$ is used to electrically isolate the exposed surface of the *p*-layer and restrict the current flow to the *p*-contact area. The $SiO_2$ film should be sufficiently thick to prevent pinhole formation and yet as thin as possible to minimize the stress on the semiconductor. A dramatic example of dielectric stress-induced ⟨110⟩ dark-line defects in the *p*-InP confining layer is shown in a TCL image of Fig. 8. While these DLDs do not propagate into the active layer, LEDs in which they have been induced degrade extremely fast. This problem does not arise when the $SiO_2$ layer thickness is reduced to ~1000Å and the deposition conditions are adjusted to minimize stress. Next, the $SiO_2$ layer above the contact is photolithographically removed from an area about half the diameter of the contact. For instance, for the 25-μm diameter *p*-contact, a dielectric opening of ~12 μm is used. The small opening results in the metal contact cushioning the stress of the dielectric edge. The $SiO_2$ and the photoresist are stripped from this area in an $O_2$ plasma. It is important to assure cleanliness of the dielectric openings to obtain stable, low-forward-voltage devices.

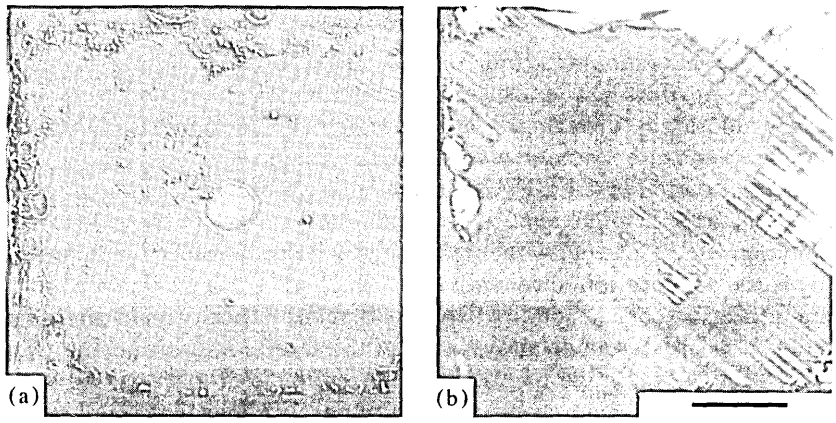



Fig. 8—Dark line defects induced by the dielectric stress in the *p*-InP confining layer: (a) Secondary electron image of the *p*-side of a rapidly degraded LED. The contact area is visible in the center. (b) TCL image showing ⟨110⟩ DLDs throughout the entire device.

A layer of Ti (1000Å) and a Au overlay (3000Å) are subsequently deposited over the full surface of the *p*-side, again by *e*-gun evaporation, and a 20-$\mu$m thick Au heatsink is electroplated. A Ti layer is used for good adhesion between $SiO_2$ and the heatsink. The heatsinks are formed in a photolithographically defined 20-mil-square pad pattern centered around the *p*-contact. The discontinuous Au film reduces wafer stress and makes diamond saw dicing easier. The "streets" in the heatsink pattern, similar to the *n*-type metallization squares, are aligned in the ⟨100⟩ direction. Misalignment tends to cause cleaving along ⟨110⟩ direction during the dicing operation and results in chip breakage. The dicing damage is removed by a ~30s etch in Br-methanol solution, which completes the processing sequence.

A recently described simplified processing sequence in which the dielectric layer and the subsequent photolithography are eliminated is also being used.[29] In this process, schematically shown in Fig. 9, current confinement is achieved through the formation of a Schottky barrier. The ohmic contacts are deposited and alloyed in the usual manner. Afterwards the entire *p*-surface, including the *p*-contact, is covered with a metal layer forming a highly resistive or non-ohmic contact (Schottky barrier) to the exposed semiconductor. This metal layer consists of 1000Å of Ti and 3000Å of Au. The initial layer of Ti adheres well to the semiconductor and forms the Schottky barrier. The processing sequence is again completed by electroplating heatsinks and dicing.

It is well known that the LED-fiber coupling efficiency improves as

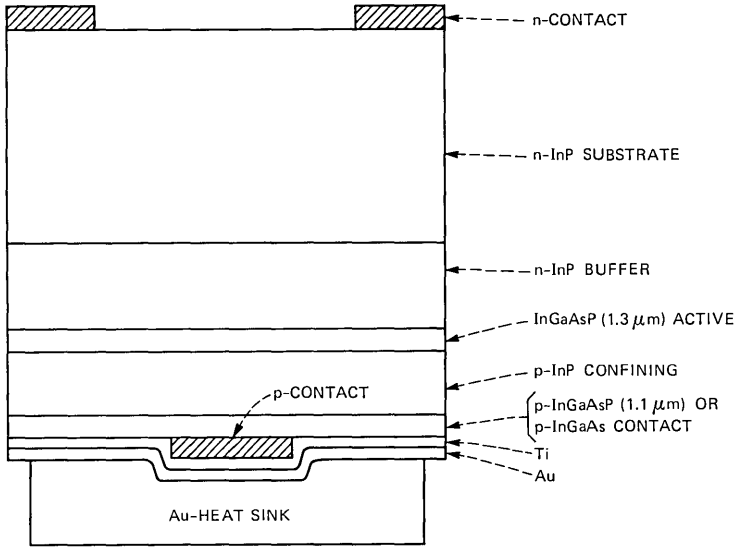



Fig. 9—Schematic structure of a Schottky-barrier isolated LED.

the light spot is made smaller than the fiber diameter. Similarly, the launched power can be increased by increasing the drive current. The resulting high-current density imposes stringent limits on the resistivity of the $p$-metallization. The resistance of the $p$-contact is strongly dependent on the material doping level and composition of the contact layer. In particular, contact resistance to both InP and InGaAsP layers decreases rapidly with increased doping level.[29] However, the low saturation limit for Zn incorporation in epitaxial InP ($\sim 4 \times 10^{18}$ cm$^{-3}$) limits the specific contact resistance to $r_c \sim 7.5 \times 10^{-5} \Omega$ cm$^2$ (see Ref. 30). This value of the specific contact resistance is sufficiently low only for $p$-contacts $\gtrsim 50$ $\mu$m in diameter. Furthermore, an unusual metallurgical reaction between Au and InP results in a spreading and enlargement of the $p$-contact upon metallization alloying, making formation of very small contacts difficult.[29] In contrast, the solubility of Zn in InGaAsP layers saturates at a doping level of $\sim 2 \times 10^{19}$ cm$^{-3}$, resulting in a specific contact resistance of $r_c \sim 2 \times 10^{-5} \Omega$ for the alloy composition corresponding to the 1.2-$\mu$m bandgap.

This composition of the contact layer is typically chosen to avoid absorption of the 1.3-$\mu$m light and allow back-reflection from the $p$-contact. However, the contact resistance continues to decrease as the composition of the InGaAsP layer is changed toward the ternary (1.65-$\mu$m) InGaAs alloy.[31] The results of specific contact resistance measurements as a function of alloy composition are shown in Fig. 10. The lowest $r_c \sim 7 \times 10^{-6} \Omega$ cm$^2$ was obtained on the InGaAs layer. This value is about a factor of two lower than the $r_c$ measured for the commonly used 1.2-$\mu$m InGaAsP and results in 25-$\mu$m-diameter $p$-

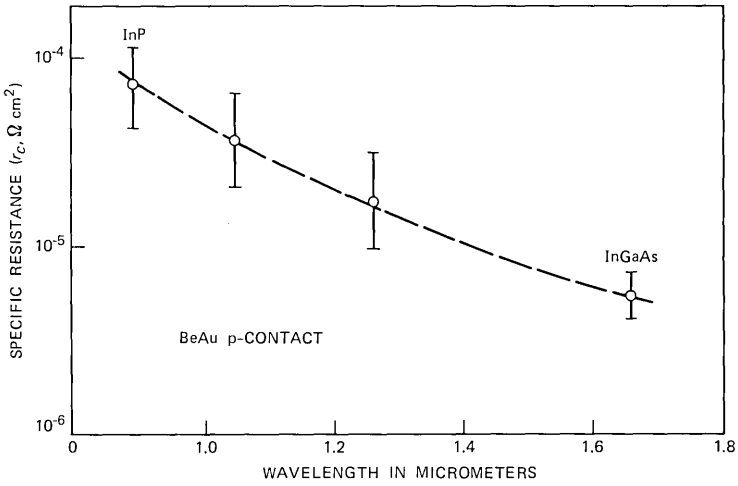



Fig. 10—Specific contact resistance as a function of alloy composition, from $p$-InP to $p$-InGaAs.

contact LEDs with a series resistance of less than 2Ω. LEDs incorporating such a ternary contact layer offer stable operation at very high current density without any appreciable reduction in the launched power, indicating that back-reflected light is not collected by the small acceptance angle of the fiber.

## IV. LED PERFORMANCE

A number of parameters should be considered in determining LED performance in a transmission application. For data rates below ~100 Mb/s, increasing launched power increases repeater distance (at higher data rates modal dispersion begins to be important). The launched power decreases with increasing temperature and this temperature dependence also affects the repeater spacing. The injection level, on which the optical power depends, and the operating temperature may strongly influence reliability considerations, as discussed in the next paragraph. Finally, modulation bandwidth of the device which determines the maximum data rate, and the optical power are related and can be traded for each other by proper adjustments of the doping density and active layer thickness.

The light emitted into the air is plotted as a function of bias current in Fig. 11a at the three ambient temperatures of 0°, 25°, and 70°C. The *L-I* curves are clearly sublinear above 100 mA. At room temperature, a 50-percent increase in the forward current, up to 150 mA, increases the light output by ~20 percent or 1 dB. It should be remembered, however, that for 1.3-μm operation a 1-dB improvement in power output could correspond to a ~1-km increase in repeater distance. Such high current densities, up to 30 kA/cm$^2$ at 150 mA bias for a 25-μm-diameter *p*-contact device, are possible only for devices with a very low series resistance, $r_s \lesssim 3\Omega$, as discussed in the preceding section. At 25°C nearly 3 mW of optical power are emitted by these high-radiance devices. A higher power output into the air can be obtained for LEDs with larger *p*-contact areas, where the ohmic heating is reduced. However, small light-emitting areas are necessary for the efficient coupling of light into optical fibers.

The temperature dependence of the light output between −40°C and +70°C is plotted in more detail in Fig. 11b. The light-output/temperature characteristic follows the empirical relationship $L = L_0 \exp(-T/T_1)$, where the characteristic temperature $T_1$ of our devices ranges between 180 and 220 degrees and is independent of current.[32] Similarly, the functional form of $L(I)$ is independent of temperature. Thus, a temperature rise of 50°C above room ambient results in a power output decline of approximately 1 dB. While this temperature sensitivity is somewhat greater than that of comparable GaAlAs LEDs,

where $T_1 \approx 300$–$350$°C has been observed, it is sufficiently small for LED operation without need for cooling devices.

Powers launched into fibers, as a function of fiber numerical aperture (NA) and core radius, are shown in Fig. 12. Two standard fibers, 0.23-NA, 50-$\mu$m core and 0.29-NA, 62.5-$\mu$m core are indicated by arrows. Closed circles denote butt-coupled powers; open circles and triangles refer to power launched using a lensing scheme in which a small sphere of high-index glass (~110 $\mu$m in diameter, $n \approx 1.9$) is mounted on the cleaved fiber end. All the devices have a 25-$\mu$m $p$-contact, active layer

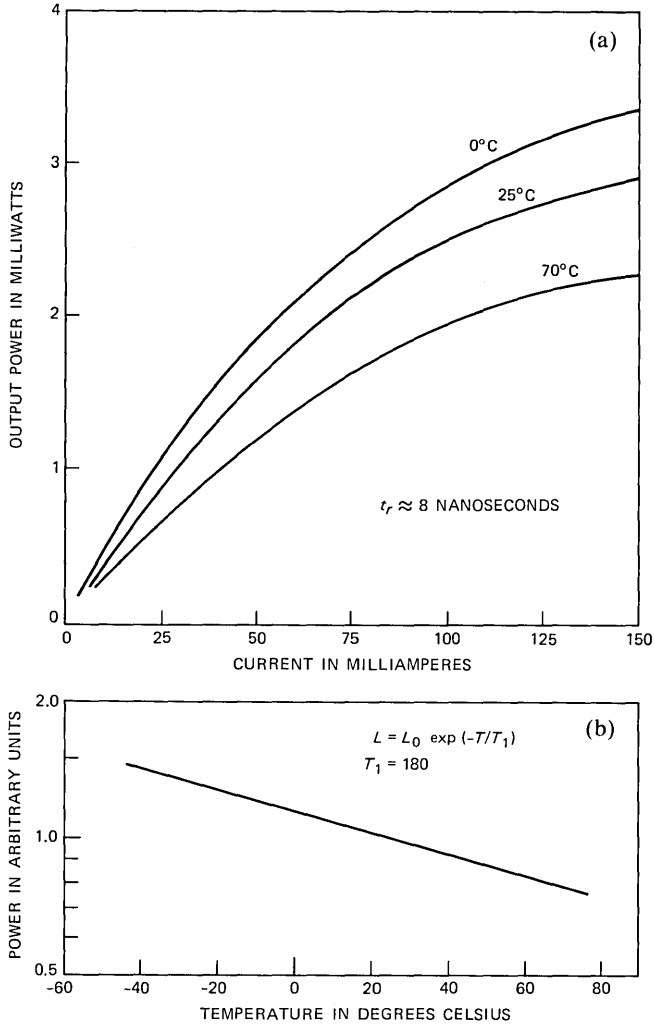



Fig. 11—(a) Light-current relation for a 25-$\mu$m diameter $p$-contact LED at three different ambient temperatures. (b) Temperature dependence of the light output at a constant current.

thickness of less than 0.7 $\mu$m, and a 10- to 90-percent risetimes of $\lesssim$4 ns. Because of current spreading the light spots are 30 to 35 $\mu$m in diameter, somewhat larger than the $p$-contact. No differences in current spreading or power launched have been found between dielectric and Schottky-isolated LEDs.

In the first approximation, the power coupled by an LED with a light-spot diameter $d_s$ to a graded index fiber with a core $d_f$, when $d_s \leq d_f$, is given by

$$P = 1/2(\mathrm{NA})^2 P_0 \text{ butt coupled,} \qquad \text{and}$$

$$P = 1/2(\mathrm{NA})^2 M^2 P_0 \text{ lens coupled,}$$

where NA is the numerical aperture of the fiber, $P_0$ is the power radiated from the LED surface, and $M$ is the magnification of the LED

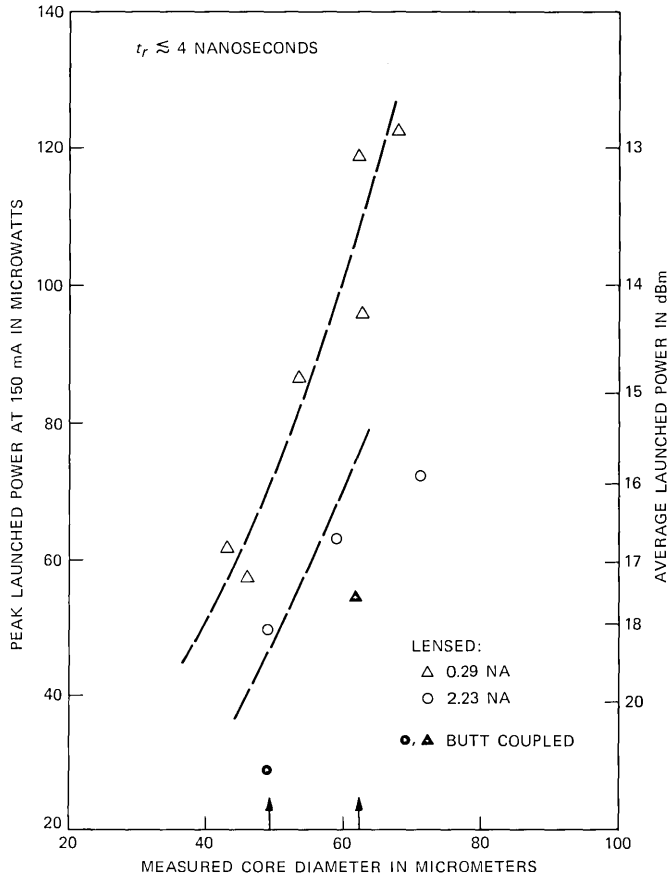



Fig. 12—Optical power launched into fiber as a function of NA and core radius. Two common fibers (0.23 NA, 50-$\mu$m core; 0.29 NA, 62.5-$\mu$m core) are indicated by arrows.

spot size by the lens. The latter has a maximum value when $M = d_f/d_s$, i.e., when the image just fills the fiber core. For a lensed system the coupled power, $P \approx (\mathrm{NA} \cdot d_f)^2$, as indicated by the dashed lines of Fig. 12. The data points and the dashed lines are in reasonably good agreement. The highest powers (peak) coupled by these fast LEDs were 50 and 120 $\mu$W into 50-$\mu$m and 62-$\mu$m core fibers, respectively. These powers are clearly a very strong function of the $d_f/d_s$ ratio, the numerical aperture, and the lensing scheme used. Addition of an antireflection coating results in as much as a 20-percent increase in the power launched.[33] A coupling experiment with a 100-$\mu$m core 0.25-NA step index fiber achieved 410 $\mu$W of optical power. Such fibers, however, are not useful for optical communications.

The 1.3-$\mu$m LEDs discussed in Fig. 12 are very fast. Using a square-current pulse of 100 mA and a dc offset of 6 mA, risetimes of less than 4 ns are measured. The fall times are typically 1 to 2 ns longer. The fastest devices show risetimes close to 2 ns. As indicated by electron beam induced current (EBIC) scans, these thin active-layer (0.5 to 0.7 $\mu$m) devices have a $p$-$n$ junction at the interface between the InGaAsP active layer and the $n$-InP buffer layer. The active layer is thus converted to $p$-type by Zn out diffusing from the confining $p$-InP layer. Hall measurements on InGaAsP layers grown under conditions simulating active-layer growth indicate doping level of $p \gtrsim 7 \times 10^{17}$ cm$^{-3}$. These LEDs have been successfully used in transmission experiments at 90-Mb/s rate.[34] On the basis of these risetime data, operation at 140 Mb/s is feasible without any speed-up circuitry.

The interplay of power and speed is well understood in GaAlAs LEDs.[35,36] The active layer of those devices can be doped directly with Ge and the resulting $p$-type doping level is well controlled. Thus, it is possible to change the doping level from $p \sim 1 \times 10^{17}$ cm$^{-3}$ to $p \sim 1 \times 10^{19}$ cm$^{-3}$ and the corresponding risetimes decrease from 25 to 4 ns. At the same time the power output is inversely proportional to the risetime, resulting in a constant power-bandwidth product in good agreement with theoretical models.[36] The results of a bandwidth-power study of 1.3-$\mu$m InGaAsP LEDs are summarized in Fig. 13. The 10- to 90-percent risetime (ns) is plotted on a log-log scale versus power launched into the larger fiber (62.5-$\mu$m core, 0.29 NA). Each data point represents an average of 10 devices from a single wafer. For the purpose of this study, the active-layer thickness is increased to 1.4 $\mu$m. It should be remembered that the $p$-$n$ junction is formed by Zn diffusion from the confining layer and increased active-layer thickness permits better control of the diffusion front depth. The risetime correlates well with the amount of Zn added to the confining layer. The heavily Zn-doped devices have risetimes on the order of 4 to 6 ns, and the $p$-$n$ junction is shifted all the way to the $n$-InP buffer interface.

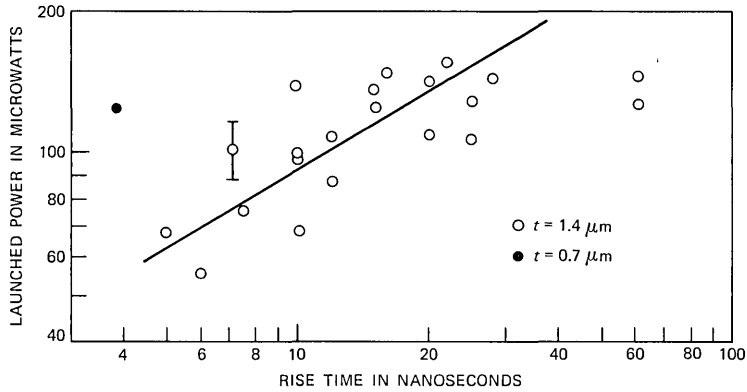



Fig. 13—Power versus risetime relationship for LEDs with a 1.4-μm thick active layer. Full dot demonstrates performance of 0.7-μm active-layer device.

As the amount of Zn is reduced, risetimes increase, up to 30 ns and output power increases by a factor of 2 to 3. However, in contrast to GaAlAs devices, the *p-n* junction displacement is gradually reduced, resulting in an *n*-type active layer. High powers of these LEDs demonstrate efficient hole injection into the InGaAsP layer. A simple model of a displaced junction gives a decay rate[37]

$$R = \frac{1}{\tau} = \omega \frac{1}{\tau_p} + (1 - \omega) \frac{1}{\tau_n},$$

where under high injection conditions

$$\omega = \left(1 + \frac{t_p}{t_n}\right)^{-1},$$

and $\tau_{n,p}$ refers to electron/hole lifetimes, respectively, $t_p$ is the thickness of the active-layer region converted to *p*-type, and $t_n = t_{\text{active}} - t_p$. Thus, the decay rate is equal to the volume-weighted average of both minority carrier lifetimes. In our devices $\tau_n \approx 4$ ns and $\tau_p \geq 30$ ns. This finding is in good agreement with a previous observation[9] of increased power output for increased $t_n/t_p$. However, the *n*-active layer LEDs are slower and may not be useful for high-data-rate applications.

The best power result on thin (0.7 μm), active-layer LEDs is shown for comparison as a full circle in Fig. 13. The slower devices launch as much as 150 μW into 62-μm core fiber and up to 177 μW with an antireflection coating. The relatively high power and high speed of thin active-layer LEDs are interpreted in terms of higher injected carrier density and the resulting bimolecular recombination, which shortens the recombination lifetime.[38] It should, however, be noted that the 1.4-μm active-layer preparation is more reproducible and device results more uniform.

## V. RELIABILITY

To estimate the reliability of LEDs for lightwave systems, life tests are performed under forward bias at elevated temperatures. The purpose is (*i*) to identify the failure modes relevant at operating temperatures, (*ii*) to find and eliminate the infant failure population, (*iii*) to determine the time, current, and temperature dependence of the degradation in light output to allow accurate extrapolations to end-of-life, and (*iv*) to determine the failure distribution so that failure rates can be calculated that reflect both the median life and the spread in the data. It is useful to begin by contrasting the failure modes in GaAlAs LEDs with what is known for quaternary LEDs.

The reliability of double heterostructure GaAs/GaAlAs has been studied extensively.[39,40] It is well known that ⟨100⟩ dark line defects (DLDs) form readily at grown-in dislocations in these devices. Their rate of growth is proportional to the square of current density ($J^2$), and independent of temperature. Depending upon the dislocation density in the epitaxial material, a certain fraction of devices (active area multiplied by the dislocation density) will fail catastrophically owing to DLD formation and must be screened out in a burn-in of about 100h. Under forward bias aging at 6 kA/cm$^2$ and $T_J$ up to 250°C, the surviving devices remain free of DLDs but change gradually in light output owing to two competing processes with different time dependences and different activation energies.[40] These processes can be separated, and the degradation process is found to have an activation energy of 0.65 eV. This degradation is assumed to be due to the introduction of nonradiative defects at the junction. The projected mean-time-to-failure (MTTF) at 70°C is $4 \times 10^6$h. A notable feature of GaAlAs LED degradation is the small spread in times-to-failure for samples of diodes.[40] The failure distribution can be described as lognormal with $\sigma \cong 0.5$. This small $\sigma$ coupled with the long MTTF gives very low estimated failure rates (e.g., a maximum of 10 FITs over a 20-year service life at $T_J = 70$°C).

In contrast, quaternary LEDs do not form ⟨100⟩ DLDs at threading dislocations, permitting much larger, ~$10^4$/cm$^2$, dislocation densities in InP substrates. Diodes with several dislocations in the active area have been aged at 10 kA/cm$^2$ for $2 \times 10^4$h without the appearance of DLDs.[41,42] The only dislocations that have been observed in electroluminescence after aging are ⟨110⟩ misfit lines in some diodes where the lattice mismatch between the buffer and quaternary layers is greater than 0.06 percent. In some cases the appearance of ⟨110⟩ lines accompanies a rapid drop in light output. The only infant failures observed in lattice-matched InGaAsP LEDs are due to instabilities in *I-Vs* or in light output. These instabilities result from shunt paths across the junction caused by In inclusions. A burn-in of 100h at 150

mA is done to screen out unstable devices. However, the burn-in yield for good, i.e., inclusion-free, wafers is almost 100 percent.

We report here the results of two series of accelerated aging experiments on dielectric isolated devices. In the first series, diodes labeled FC291 were fabricated from a wafer grown early in the quaternary LED development program. These were three-layer structures, i.e., without a contact layer, with 50-$\mu$m diameter contacts (the active-layer thickness was 0.7 $\mu$m). The epitaxial material contained a number of In inclusions up to ~10 $\mu$m in diameter, as discussed in Section II, and would be considered poor by present standards. The elapsed time for this series of experiments is now $1.5 \times 10^4$h. The second series of experiments is on wafer-labelled LP137, with an active-layer thickness of 1.4 $\mu$m. These are four-layer structures with 25-$\mu$m contacts, so that the current density $J$ for a given device current is about four times that in FC291. These diodes were free of inclusions as shown by their electroluminescence patterns. The elapsed time for this series is $4 \times 10^3$h.

In FC291-type wafers, two modes of degradation have been identified: ⟨110⟩ DLD formation above 140°C and dark spot defect (DSD) formation at all temperatures. In the first mode, some diodes form ⟨110⟩ DLDs and drop rapidly in light output, even without bias. The DLDs originate at microscopic inclusion-like defects and are confined to the InP buffer layer. These ⟨110⟩ DLDs are similar to the dielectric stress-induced defects discussed in Section III. It is concluded that these DLDs are caused by thermal stress and that they grow by dislocation glide on the {111} slip planes. Because 140°C is apparently a threshold temperature for this effect, it will not be a relevant failure mode at operating temperatures.[43] This type of DLD has not been observed in recent wafers.

In analyzing the data for FC291, we eliminate all diodes that have ⟨110⟩ DLDs. In the remaining diodes, the light output falls linearly in time as shown in Fig. 14 and DSDs appear and grow. Because these LEDs do not fail abruptly, but rather drop gradually in light output, the definition of end-of-life should consider the overall system objective.[44] Since the repeater span is generally power-limited, it is desirable to allow as little power margin for end-of-life as possible, within system reliability requirements. For purposes of illustration, we define end-of-life as −1 dB change in output power. Figure 15 shows a typical failure distribution. It demonstrates that the failures are lognormally distributed with $\sigma = 0.6$. This $\sigma$ is similar to that of GaAlAs devices and demonstrates excellent sample homogeneity. Figure 16 is an Arrhenius plot for median life under the various aging conditions. The activation energy is 0.85 eV. This activation energy is significantly larger than that obtained for the degradation of GaAlAs LEDs.

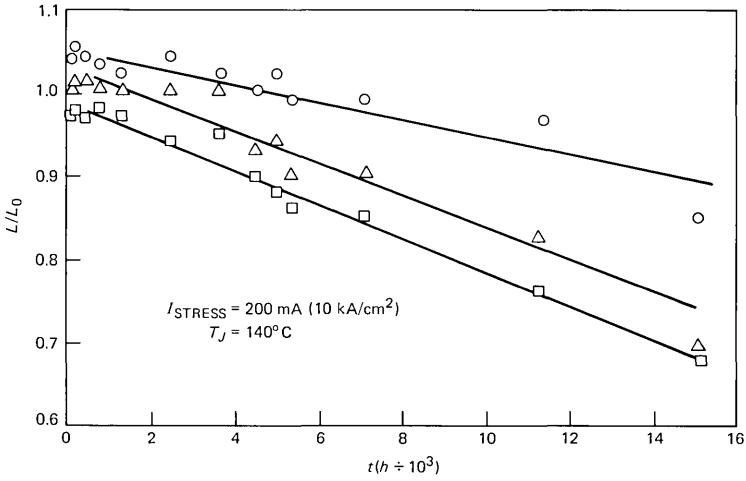



Fig. 14—Accelerated aging tests: light output versus time as a function of temperature for three typical diodes.

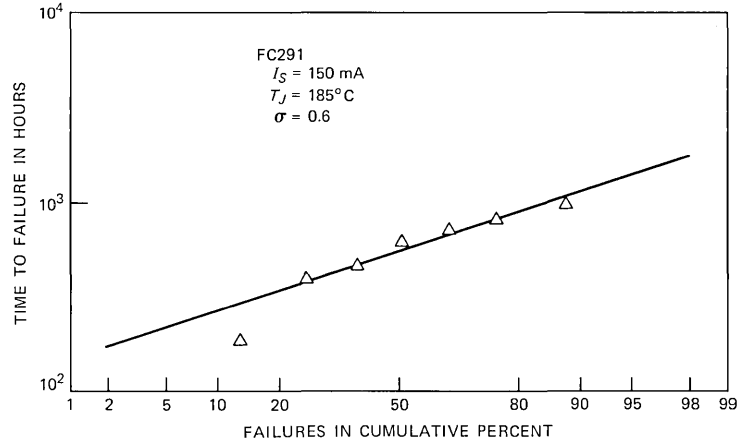



Fig. 15—Log-normal failure distribution typical of InGaAsP LED aging results.

Figure 16 also shows the results of experiments on wafer LP137. These diodes drop in light output linearly in time, but do *not* form DSDs. This is important, since DSD formation was believed to be a main degradation mechanism for InGaAsP LEDs. There is very little difference between median life (ML) for 20 kA/cm$^2$ and 40 kA/cm$^2$, suggesting that the failure mechanism (uniform degradation in light output) is not a strong function of current. Although the current density is four times higher in LP137 than in FC291, the ML has gone up by a factor of four. This probably reflects the improvement in the quality of the material.

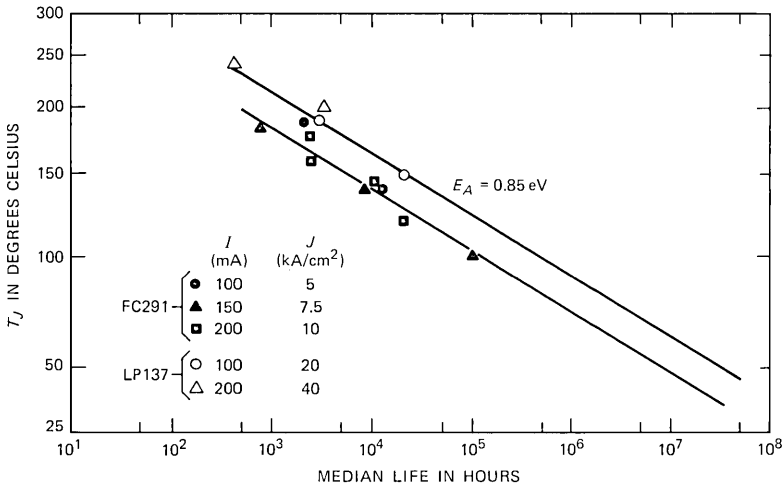



Fig. 16—Arrhenius plot of median life under various aging conditions. The activation energy is 0.85 eV.

The maximum operating temperature for these LEDs in most applications is 70°C. The 40 kA/cm$^2$ data projects median life of $4 \times 10^6$h at this temperature. For a service life of 20 years, using $\sigma = 0.6$, the projected failure rate is <1 FIT.[45] At 25°C median life of $3 \times 10^8$h is predicted. Higher reliability results if greater degradation is permitted. The system designer must trade off initial system performance for reliability objectives.

These reliability results are in good agreement with previous studies.[42,46] Thus, median life well in excess of $10^6$ hours has been predicted at 25°C and current densities below 20 kA/cm$^2$. Similarly, the activation energy of 0.85 eV is consistent, within experimental uncertainties, with the 1.0 eV obtained previously. However, absence of DSDs in the high-quality wafer LP137 suggests that this degradation mechanism may be extrinsic. Further progress in material preparation and processing techniques should result in even better reliability of InGaAsP LEDs.

## VI. SUMMARY

We have described the 1.3-$\mu$m InGaAsP LED developed for application in optical fiber transmission. This includes improvements and scale-up in the epitaxial growth procedures, careful design of the processing sequence, and optimization of the device structure. The resulting LED has a number of very attractive features. Its excellent optical powers allow repeater distances in the 10- to 20-km range for 62-$\mu$m fibers, depending on the cable loss. These long repeaterless spans can be realized at data rates of up to 90 Mb/s. This is accom-

plished with a relatively simple drive circuitry, as compared to lasers. In addition, the InGaAsP LED exhibits excellent thermal stability. As the ambient temperature is increased from 25°C to 70°C, the power output decreases by only ~1 dB. Finally, even at 70°C the median life exceeds $4 \times 10^6$h and the resulting failure rate is expected to be below 1 FIT.

## VII. ACKNOWLEDGMENTS

We thank T. P. Lee and M. A. Pollack for their generosity with materials and good advice, W. A. Bonner for all the excellent InP substrates, and S. Mahajan for defect characterization and many useful discussions. Many thanks to J. A. Lourenco, D. D. Roccasecca, I. Camlibel, R. H. Frahm, and G. Minneci for the superb technical assistance in crystal growth, processing, and device characterization; to F. Ermanis for uncounted SEM pictures; and to R. Caruso for X-ray diffraction measurements. Above all we would like to acknowledge A. A. Bergh, whose support has made this work possible.

## REFERENCES

1. A. A. Bergh, J. A. Copeland, and R. W. Dixon, "Optical Sources for Fiber Transmission Systems," IEEE Proc., *68,* No. 10 (October 1980), pp. 1240–47.
2. P. C. Schultz, Third Int. Conf. on Integrated Optics and Optical Fiber Comm., Technical Digest, April 1981, San Francisco; and D. Charlton and P. C. Schultz, "Progress in Optical Waveguide Processes, 1980," Electro-Optical System Design, December 1980, pp. 23–9.
3. J. W. Fleming, "Material Dispersion in Lightguide Glasses," Electron. Lett., *14,* No. 11 (May 1978), pp. 326–8.
4. J. Conradi, F. P. Kapron, and J. C. Dyment, "Fiber-Optical Transmission Between 0.8 and 1.4 $\mu$m," IEEE Trans. Electron. Dev., *ED-25,* No. 2 (February 1978), pp. 180–93.
5. D. Gloge, A. Albanese, C. A. Burrus, E. L. Chinnock, J. A. Copeland, A. G. Dentai, T. P. Lee, Tingye Li, and K. Ogawa, "High-Speed Digital Lightwave Communications Using LEDs and PIN Photodiodes at 1.3 $\mu$m," B.S.T.J., *59,* No. 8 (October 1980), pp. 1365–82.
6. G. A. Antypas, R. L. Moon, L. W. James, J. Edgecumbe, and R. L. Bell , "III-V Quaternary Alloys," Inst. Phys. Conf. Series. 1972 Symp. on GaAs, pp. 48–54.
7. M. A. Pollack, R. E. Nahory, J. C. DeWinter, and A. A. Ballman, "Liquid phase epitaxial $In_{1-x}Ga_xAs_yP_{1-y}$ lattice matched to ⟨100⟩ InP over the complete waveguide range $0.92 < \lambda < 1.65$ $\mu$m," Appl. Phys. Lett., *33,* No. 4 (August 1978), pp. 314–16.
8. P. D. Wright, Y. G. Chai, and G. A. Antypas, "InGaPAs-InP Double-Heterojunction High-Radiance LED's," IEEE Trans. Electron. Dev. *ED-26,* No. 8 (August 1975), pp. 1220.
9. O. Wada, S. Yamakoshi, M. Abe, Y. Nishitani, and T. Sakurai, "High Radiance InGaAsP/InP Lensed LED's for Optical Communication Systems at 1.2–1.3 $\mu$m," IEEE J. Quant. Electr., *QE-17,* No. 2 (February 1981), pp. 174–8.
10. R. C. Goodfellow, A. C. Carter, I. Griffith, and R. R. Bradley, "GaInAsP/InP, High-Radiance, 1.05–1.3 $\mu$m Wavelength LED's with Efficient Lens Coupling to Small Numerical Aperture Silica Optical Fibers," IEEE Trans. Electron. Dev., *ED-26,* No. 8 (August 1979), pp. 1215–20.
11. A. G. Dentai, T. P. Lee, and C. A. Burrus, "Small-Area High Radiance C.W. InGaAsP LEDs Emitting at 1.2 to 1.3 $\mu$m," Electr. Lett., *13* (1977) pp. 484–5.
12. R. F. Leheny, R. E. Nahory, and M. A. Pollack, "$In_{0.53}Ga_{0.47}As$ PIN Photodiodes for Long-Wavelength Fiber Optic Systems," Electr. Lett., *15* (1979), pp. 713–15.
13. V. G. Keramidas, H. Temkin, and W. A. Bonner, "Growth of InP and InGaAsP

layers by liquid phase epitaxy under phosphorus overpressure," Appl. Phys. Lett., *40* (1982), p. 731.
14. S. Mahajan, V. G. Keramidas, A. K. Chin, W. A. Bonner, and A. A. Ballman, "Perfection of homoepitaxial layers grown on (001) InP substrates," Appl. Phys. Lett., *38,* No. 4 (February 1981), pp. 255–8.
15. W. A. Bonner, "Reproducible Preparation of Twin-Free InP Crystals Using the LEC Technique," Mat. Res. Bull., *15* (1980), pp. 63–72.
16. M. A. DiGiuseppe, H. Temkin, and W. A. Bonner, "Large Area LPE Growth of InGaAsP/InP double heterostructures on InP preserved in a phosphorus ambient," J. Cryst. Growth, to be published.
17. A. K. Chin, H. Temkin, and S. Mahajan, "Transmission Cathodoluminescence," B.S.T..J., *60,* No. 9 (November 1981), pp. 2187–226.
18. Y. Seki, H. Watanabe, and J. Matsui, "Impurity effect on grown-in dislocation density of InP and GaAs crystals," J. Appl. Phys., *49* (February 1978), pp. 822–8.
19. A. K. Chin, V. G. Keramidas, W. D. Johnston, Jr., S. Mahajan, and D. D. Roccasecca, "Evaluation of defects and degradation in GaAs-GaAlAs wafer using transmission cathodoluminescence," J. Appl. Phys., *51,* No. 2 (February 1980), pp. 978–83.
20. H. Temkin, V. G. Keramidas, and S. Mahajan, "Thermal Decomposition of InP and Its Influence on Iso-Epitaxy," J. Electrochem. Soc., *128,* No. 5 (May 1981), pp. 1088–91.
21. G. A. Antypas, "Prevention of InP surface decomposition in liquid phase epitaxial growth," Appl. Phys. Lett., *37,* No. 1 (July 1980), pp. 64–5.
22. M. G. Astles, F. G. H. Smith, and E. W. Williams, "Indium Phosphide, Epitaxial Growth," J. Electrochem. Soc., *120* (1973), pp. 1750–7.
23. W. Y. Lum and A. R. Clawson. "Thermal Degradation of InP and Its Control in LPE Growth," J. Appl. Phys., *50,* No. 8 (August 1979), pp. 5296–301.
24. K. Nakajima, S. Yamazaki, S. Komiya, and K. Akita, "Misfit dislocation-free InGaAsP/InP heterostructure wafers grown by liquid phase epitaxy," J. Appl. Phys., *52,* No. 7 (July 1981), pp. 4575–82.
25. H. Temkin, V. G. Keramidas, M. A. Pollack, and W. R. Wagner, "Temperature dependence of photoluminescence of *n*-InGaAsP," J. Appl. Phys., *52,* No. 3 (March 1981), pp. 1574–8.
26. O. Wada, A. Majerfeld, and P. N. Robson, "Control of Zn-doping for growth of InP *pn* junction by liquid phase epitaxy," J. Electrochem. Soc., *127,* No. 10 (October 1980), pp. 2278–84.
27. S. Nakahara, R. J. McCoy, L. Buene, and J. M. Vandenberg, "Room temperature interdiffusion studies of Au/Sn thin film complexes," Thin Solid Films, *84* (1981), pp. 185–6.
28. A. K. Chin, C. L. Zipfel, B. V. Dutt, M. A. DiGiuseppe, K. B. Bauers, and D. D. Roccasecca, "New Restricted Contact LEDs Using a Schottky Barrier," Japn. J. Appl. Phys., *20,* No. 8 (August 1981), pp. 1487–91.
29. V. G. Keramidas, H. Temkin, and S. Mahajan, "Ohmic contacts to InP and InGaAsP," Inst. Phys. Conf. Ser., *56* (1981), pp. 293–9.
30. H. Temkin, R. J. McCoy, V. G. Keramidas, and W. A. Bonner, "Ohmic contacts to *p*-type InP using Be-Au metallization," Appl. Phys. Lett., *36,* No. 6 (March 1980), pp. 444–6.
31. H. Temkin, A. K. Chin, and M. A. DiGiuseppe, "$In_{0.53}Ga_{0.47}As$ contact layer for 1.3-$\mu$m light emitting diodes," Electron. Lett., *17,* No. 19 (September 1981), pp. 703–5.
32. H. Temkin, A. K. Chin, M. A. DiGiuseppe, and V. G. Keramidas, "Light-current characteristics of InGaAsP of light-emitting diodes," Appl. Phys. Lett., *39,* No. 5 (September 1981), pp. 405–7.
33. R. H. Burton, unpublished work.
34. G. T. Daryanani, unpublished work.
35. V. G. Keramidas, A. K. Chin, C. L. Zipfel, D. D. Roccasecca, and F. Ermanis, unpublished work.
36. T. P. Lee, and A. G. Dentai, "Power and modulation bandwidth of GaAs-AlGaAs high-radiance LED's for optical communication systems," IEEE J. Quant. Electr., *QE-14,* No. 3 (March 1978), pp. 150–9.
37. H. Temkin, W. B. Joyce, A. K. Chin, M. A. DiGiuseppe, and F. Ermanis, "Effect of p-n function position on the performance of InGaAsP light emitting diodes", Appl. Phys. Lett., *41,* No. 8 (October 1982), pp. 745–7.
38. J. S. Blakemore, *Semiconductor Statistics,* New York: Pergamon Press, 1962.
39. S. Yamakoshi, T. Sugahora, O. Hasegawa, Y. Toyama, and H. Takanashi, "Growth mechanism of ⟨100⟩ dark-line defects in high radiance GaAlAs LEDs," IEDM Proc., Washington, D. C. (1978), pp. 642–5.
40. C. L. Zipfel, R. H. Saul, A. K. Chin, and V. G. Keramidas, "Competing processes in

long-term accelerated aging of DH GaAlAs LEDs," J. Appl. Phys., *53,* No. 3, Part 1 (March 1982), pp. 1781–6.
41. S. Yamakoshi, M. Abe, S. Komiya, and Y. Toyama. "Degradation of high radiance InGaAsP/InP LEDs at 1.2–1.3 μm wavelength," IEDM Proc., Washington, D. C. (1979), pp. 122–5.
42. S. Yamakoshi, M. Abe, O. Wada, S. Komiya, and T. Sakurai, "Reliability of high radiance InGaAsP/InP LEDs operating in the 1.2–1.3 μm wavelength," IEEE J. Quant. Electr., *EQ-17,* No. 2 (February 1981), pp. 167–73.
43. H. Temkin, C. L. Zipfel, and V. G. Keramidas, "High-temperature degradation of InGaAsP/InP light emitting diodes," J. Appl. Phys., *52,* No. 8 (August 1981), pp. 5377–80.
44. R. H. Saul and C. L. Zipfel, unpublished work.
45. A. S. Jordan, "A comprehensive review of the lognormal failure distribution with application to LED reliability," *Microelectronics and Reliability, 18,* No. 3, Oxford: Pergamon Press, 1978, pp. 267–79.
46. R. Yeats, Y. G. Chai, T. D. Gibbs, and G. A. Antypas, "Performance Characteristics and Extended Lifetime Data for InGaAsP/InP LED's," IEEE Electr. Dev. Lett., *EDL-2,* No. 9 (September 1981), pp. 234–6.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# Analog Scramblers for Speech Based on Sequential Permutations in Time and Frequency

By N. S. JAYANT, R. V. COX, B. J. McDERMOTT, and A. M. QUINN

(Manuscript received July 20, 1982)

*Permutation of speech segments is frequently utilized in scramblers for analog speech privacy. This paper discusses a* "sequential" *permutation procedure that has better segment-separation properties than the well-known procedure of* "block" *permutation, where contiguous segments are arranged in blocks of appropriate size, and permuted within such blocks. It further proposes the application of the sequential procedure to a novel technique for simultaneous permutations in time and frequency. The paper also presents results of a subjective experiment where we measured residual speech intelligibility at the output of scramblers using permutations in time [time segment permutation (TSP)], or permutations in time* and *frequency [time-frequency segment permutation (TFSP)]. The experiment included examples of block TSP, sequential TSP, and sequential TFSP. We measured spoken-digit-intelligibility as a function of the communication delay introduced by the scrambling operation. We found that even with a delay of 512 ms, the residual intelligibility in a TSP scrambler is no lower than about 50 percent; however, a sequential TFSP scrambler can realize an average digit intelligibility in the order of 20 percent with a delay of 256 ms. A companion paper discusses the implementation of the sequential TFSP scrambler, and the quality of descrambled speech in the context of real-channel operation.*

## I. INTRODUCTION

Permutation of speech segments that are about 10 to 30 ms long is a bandwidth-preserving operation[1] that is frequently utilized in the design of scramblers for analog speech privacy.[2,3] The procedure,

known as Time Segment Permutation (TSP), lends itself to fairly robust real-channel operation[2] and efficient microprocessor implementation.[4] However, the reduction in speech intelligibility by this method is very small if the communication delay introduced by the scrambler and descrambler is constrained to be no longer than, say, 256 to 512 ms. This rather well-known deficiency was quantitatively demonstrated in a recent article[1] that discussed a segment scrambler, which will be referred to as "block" TSP in this paper (see Section 2.2). In this scrambler, contiguous speech segments are arranged in blocks of appropriate size, and permuted *within* such blocks. The entire block with permuted segments is then transmitted as scrambled speech, before proceeding to the next block. In this paper, we discuss another segment permutation procedure to be called "sequential" TSP (see Section 2.3). In this case, permutations are not constrained to be within blocks, and transmissions of scrambled speech are not constrained to be on a block-by-block basis. We will compare the two procedures in terms of how well they separate segments that are initially adjacent in the unscrambled speech. We will show that the sequential TSP has segment-separation properties that are much better than those of block TSP, but that, unfortunately, this is not accompanied by substantial gains in the residual intelligibility in scrambled speech, except for large values of communication delay (see Section 3.3).

To realize substantial reductions of intelligibility, it is imperative to use so-called two-dimensional approaches to scrambling.[1,3] One example of two-dimensional scrambling is the combination of block TSP and frequency inversion.[1] However, frequency inversion is a very straightforward, simple and time-invariant operation with only one possible input-output mapping, or "key." It therefore has no cryptanalytical strength. An important purpose of this paper is to propose and evaluate a two-dimensional procedure that offers a residual intelligibility very similar to that of block TSP plus frequency inversion, and a cryptanalytical strength that is much higher than in that method. This new procedure (Section IV) will be called Time-Frequency Segment Permutation (TFSP). In particular, we will be discussing a sequential version of this procedure, "sequential" TFSP.

A companion paper[5] discusses the implementation of the sequential TFSP scrambler, and the quality of descrambled speech in the context of a real-channel operation.

## II. ONE-DIMENSIONAL SCRAMBLERS: BLOCK TSP AND SEQUENTIAL TSP

This section describes a sequential approach to segment-permutation and shows that it has much better segment separation properties

than a non-sequential, or block, procedure such as that discussed in Ref. 1. The block and sequential approaches to be discussed below are sometimes referred to as "*hopping window*" and "*sliding window*" approaches.[6]

### 2.1 Temporal distance d

The purpose of permutation scrambling is to reduce intelligibility by altering the normal time order of speech segments. The greater the separation in scrambled speech between normally adjacent segments, the lower the intelligibility is expected to be. Conversely, adjacent samples in the scrambled speech should be well separated in normal speech. An important result of this paper is that average segment separation and intelligibility are generally, if not always, monotonically related. It is useful, therefore, to define the following objective measure of effectiveness for permutation scramblers as the *temporal distance between a pair of segments in normal speech that appear as adjacent segments in scrambled speech.* For simplicity, we will henceforth refer to this non-zero, positive parameter merely as the "temporal distance *d*." By definition,

$$d = 1 \quad \text{for adjacent segments in unpermuted speech}$$
$$d \geq 1 \quad \text{for adjacent segments in permuted speech.} \qquad (1)$$

Illustrations of temporal distance appear in Fig. 1. The scramblers used in this figure will be defined in the next two sections.

Briefly, block TSP is a procedure where an entire block of segments in scrambler memory is transmitted before proceeding to the next block. Segment selection involves a number of candidates that decreases from the initial number of segments in the block to 1, as a given block is processed. In sequential TSP, each stage of segment

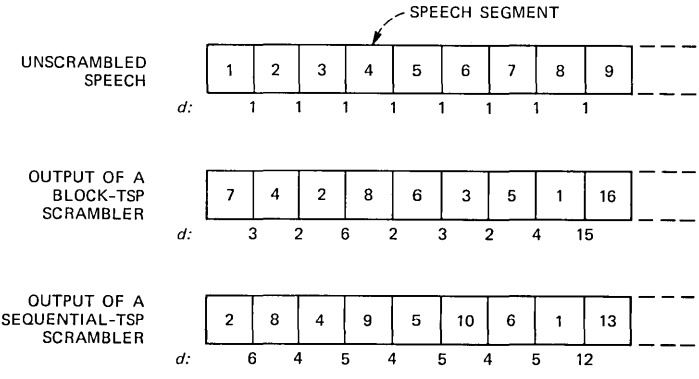



Fig. 1—Temporal distance *d*.

selection involves a constant number of candidates equal to the maximum number of segments that can be stored in scrambler memory.

### 2.2 Block TSP

Figure 2(a) defines the block approach to segment scrambling. In this approach, the scrambler memory consists of a block of $b'$ contiguous segments; one can identify in the scrambler output a corresponding block whose member segments are the same as the segments that comprise the input block. A succeeding block (block 2 in Fig. 2a), also comprising $b'$ segments, enters the scrambler memory after all the $b'$ segments of the preceding block (block 1 in Fig. 2a) have been processed and transmitted. Table Ia depicts a realization of the random processes in block TSP for the example of $b' = 8$ (Fig. 1). Shown are the successive contents of scrambler memory, the sequence of transmitted segments, and the temporal distance $d$ between adjacent segments in the scrambler output.

Note that transmitted segment $s$ and temporal distance $d$ are both random variables. In a practical implementation, the variable $s$ will be pseudorandom so that the intended receiver can invert the scrambler operation. What is significant is that the maximum value of $d$ in the table is 15. This is indeed a global maximum for a block TSP scrambler with memory $b' = 8$. This maximum separation is attained when the random permutation is such that the *last* segment of block $n$ [segment 16 from block $n = 2$ in Table Ia immediately follows the *first* segment of block $n - 1$ (segment 1 from block $n = 1$ in the table)]; and in general,

$$\max(d) = 2b' - 1 \text{ (in segments)}. \tag{2}$$

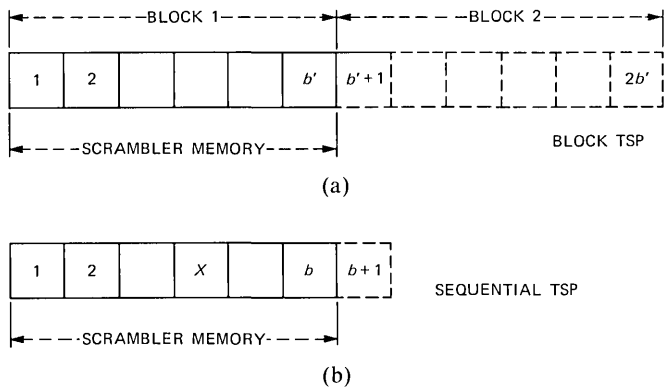



Fig. 2—Schematic representations of TSP scramblers using: (a) block and (b) sequential modes. In (a), an entire block of segments is transmitted before proceeding to the next block. Segment selection involves a number of candidates that decreases from $b'$ to 1, as a given block is processed. In the sequential case (b), each stage of segment selection involves a constant number of candidates equal to the scrambler memory $b$.

Table I—Illustration of two TSP algorithms where **m** is the contents of scrambler memory, $s$ is the transmitted segment, and $d$ is the temporal distance to previous transmitted segment

| (a) Block TSP ($b' = 8$) | | | | | | | | | | (b) Sequential TSP ($b = 8$, $t = 16$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **m** | | | | | | | | $s$ | $d$ | **m** | | | | | | | | $s$ | $d$ |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 7 | — | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 2 | — |
| 1 | 2 | 3 | 4 | 5 | 6 | 8 | | 4 | 3 | 1 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 8 | 6 |
| 1 | 2 | 3 | 5 | 6 | 8 | | | 2 | 2 | 1 | 3 | 4 | 5 | 6 | 7 | 9 | 10 | 4 | 4 |
| 1 | 3 | 5 | 6 | 8 | | | | 8 | 6 | 1 | 3 | 5 | 6 | 7 | 9 | 10 | 11 | 9 | 5 |
| 1 | 3 | 5 | 6 | | | | | 6 | 2 | 1 | 3 | 5 | 6 | 7 | 10 | 11 | 12 | 5 | 4 |
| 1 | 3 | 5 | | | | | | 3 | 3 | 1 | 3 | 6 | 7 | 10 | 11 | 12 | 13 | 10 | 5 |
| 1 | 5 | | | | | | | 5 | 2 | 1 | 3 | 6 | 7 | 11 | 12 | 13 | 14 | 6 | 4 |
| 1 | | | | | | | | 1 | 4 | 1 | 3 | 7 | 11 | 12 | 13 | 14 | 15 | 1 | 5 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 16 | 15 | 3 | 7 | 11 | 12 | 13 | 14 | 15 | 16 | 13 | 12 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 | | 12 | 4 | 3 | 7 | 11 | 12 | 14 | 15 | 16 | 17 | 3 | 10 |
| 9 | 10 | 11 | 13 | 14 | 15 | | | 10 | 2 | 7 | 11 | 12 | 14 | 15 | 16 | 17 | 18 | 18 | 15 |
| 9 | 11 | 13 | 14 | 15 | | | | 9 | 1 | 7 | 11 | 12 | 14 | 15 | 16 | 17 | 19 | 14 | 4 |
| . | . | . | | | | | | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | | | | | | . | . | . | . | . | . | . | . | . | . | . | . |

The communication delay $C$ in block TSP scrambling is the sum of two delays: ($i$) a delay of $b' - 1$ at the transmitter (the additional time for completion of the block subsequent to the arrival of segment 1), followed by ($ii$) an additional delay of $b'$ at the receiver (the maximum time for which the descrambler may have to wait before it has access to the permuted segment 1). As a result,

$$C = 2b' - 1 \text{ (in segments)}. \tag{3}$$

### *2.3 Sequential TSP*

Figure 2(b) defines a sequential approach to scrambling. In this approach, processing proceeds one segment at a time rather than one block at a time. The permutation process is constrained by two parameters: the scrambler memory $b$, and the maximum time $t$ (in multiples of segment duration) that a segment is allowed to stay in scrambler memory. When the first segment [such as $x$ in Fig. 2(b)] is pseudorandomly selected for transmission and released, the contents of the scrambler memory to the right of $x$ are shifted to the left by one unit, and the last position is *immediately* occupied by segment $b' + 1$. The next transmission is based on another random selection that, unlike in the block approach of Fig. 2(a), can involve the newest segment $b' + 1$, which was not a member of the original block. The above process continues on a segment-by-segment basis, with one constraint mentioned earlier: if a segment is retained in scrambler memory for $(t - 1)$ units of time, it is unconditionally released for transmission at time $t$, even if this means that the dictates of the random segment selector algorithm should be overridden. An important property of the sequential scrambler is that every segment has an

equal probability of spending the maximum allowed time in scrambler memory. To permit a meaningful comparison with block scrambling, we shall concentrate on the special case of

$$t = 2b \tag{4}$$

in most of the ensuing discussion. With the above specific design for $t$, the total encoding delay will be the same for both block and sequential scramblers for a given scrambler memory. This will be clear from subsequent discussion [see eq. (6)]. The design in (4) is also known to maximize the total number of unique permutations for a given delay and block length.[6] Results with a sequential TFSP scrambler indicate that (4) is also an optimal design from the viewpoint of residual intelligibility (see Section V).

Table Ib depicts a realization of the random process in sequential TSP for the example of $b = 8$ and $t = 16$ (Fig. 1). Shown once again are the successive contents of scrambler memory, the sequence of transmitted segments, and the temporal distance $d$ between adjacent segments in the scrambler output. Note in this example that segment 1 indeed stays in scrambler memory for the maximum of 16 time units. It stays in the extreme left-hand slot of scrambler memory for $t - b = 8$ time units; subsequent segments that succeed in reaching the extreme left slot tend to have maximum allowed staying times less than $t - b$ in that slot. Note also that, as in Table Ia, the temporal distance $d$ is a random variable with a maximum value of 15, equal to the maximum in block TSP. In fact this is a global maximum for the sequential design $b = t/2 = 8$; and in general, as in (2),

$$\max(d) = t - 1 = 2b - 1 \text{ (in segments)}. \tag{5}$$

In block TSP, the maximum separation (2) can be realized only in output segment pairs that involve the first and last segments of adjacent blocks (adjacent output segments 1 and 16 in Table Ia). In sequential TSP the maximum separation (4) can be realized in more general instances (for example, with adjacent output segments 3 and 18 in Table Ib).

The communication delay in sequential TSP is given by

$$C = t - 1 = 2b - 1 \text{ (in segments)}. \tag{6}$$

This is a delay inherent in the scrambling parameter $t$. There is no additional delay at the descrambler because of the absence of a block operation. After an initial waiting time of $t - 1$ segments, the descrambler has a guaranteed access to every consecutive segment needed to reconstitute the original input signal.

Table II provides a summary comparison of block and sequential approaches. The parameter $\min(d)$ in the last row will be explained

Table II—Summary comparison of block and sequential TSP

| | Block TSP | Sequential TSP (special case, $t = 2b$) | Sequential TSP (general case, $t > b$) |
|---|---|---|---|
| Scrambler memory | $b'$ | $b$ | $b$ |
| Total communication delay $C_t = C + 1$ | $2b'$ | $2b$ | $t$ |
| Maximum temporal distance $max(d)$ between adjacent output segments | $2b' - 1$ | $2b - 1$ | $t - 1$ |
| Minimum temporal distance $\min(d)$ that can be specified between adjacent output segments | 1 if $b' < 8$<br>2 if $b' \geq 8$ | $[(b + 1)/2]$<br>$[x]$: greatest integer $< x$ | |

presently. The total communication delay $C_t$ includes a delay of 1 unit that is inherent in buffer read-in and read-out, and hence equals $C + 1$ segments. If the duration of a time segment is $B$ ms, the delay in ms equals $C_tB$. In all of this paper, $B = 16$ ms.

### 2.4 min(d) and $\bar{d}$

The maximum temporal distance between adjacent output segments has been discussed above and shown to be a function of scrambler memory size. This section will provide further characterization of the random variable $d$, in particular, the minimum value of $d$ that can be specified a priori, the probability density function of $d$, and its average value $\bar{d}$.

An important property of sequential TSP—one that is not shared by block TSP—is that the selection of a segment for transmission always involves a constant number of candidate segments; this number is equal to the scrambler memory parameter $b$. A consequence of this property is that it is possible to specify in general a minimum distance $\min(d) > 1$ in the output of the sequential scrambler. The random number generator that dictates output segment selection is simply resampled repeatedly, if needed, until the output has a distance of at least $\min(d)$ from its predecessor in the output sequence. The intended descrambler is assumed to know the "key," or the random number sequence used by the scrambler. The descrambler does not need to know the value of $\min(d)$ implied by that key.

The maximum value of $\min(d)$ that will ensure a legitimate scrambler output depends both on scrambler memory $b$ and the contents thereof. A globally safe value (one that will not cause algorithm "hanging," as explained below) is

$$\min(d) = [(b + 1)/2], \tag{7}$$

where $[x]$ denotes the greatest integer less than $x$. The only occasion when (7) will have to be violated in scrambler operation is when a

segment has spent its full life span $t$ in scrambler memory, and it has to be released unconditionally without regard to its temporal distance from the most recent output.

Table III illustrates how the scrambling algorithm "hangs" when a $\min(d)$ greater than the value in (7) is specified a priori. In this example, $b = 8$, $t = 16$, $\min(d) = 6$, a value that exceeds the limit of 4 suggested by the formula (7).

As indicated in the last row of Table II, the $\min(d)$ values that can be specified a priori are much smaller in block TSP. For example, if $b' = 8$, after a possible output sequence of [3 5 7 4 6 8 2], segment 1 can never be transmitted (immediately after segment 2) unless $\min(d) = 1$. Even if the $\min(d)$ requirement is overridden in the case of the last transmitted segment of a block, the greatest value of $\min(d)$ that will not result in a "hanging" of the scrambler operation is a very slowly increasing function $b$. For example with $b = 8$, the $\min(d)$ value that can be specified a priori is no greater than 2. This should be compared with the value of 4 for sequential TSP with $b = 8$.

In both block and sequential TSP schemes, a priori insistence on a $\min(d)$ value greater than unity has attendant penalties in the cryptanalytical strength of the scrambler.[6] This is due of course to the fact that with $\min(d) > 1$, fewer random segment permutations are legal, in comparison with the totally random situation that obtains with $\min(d) = 1$.

An interesting property of a sequential scrambler with $\min(d) = 1$ and the constraint $t = 2b$ is that the fraction of segments that spends the maximum allowable time in scrambler memory is very nearly 20 percent for values of $b > 4$. An analytical demonstration of this property appears in the appendix.

Figure 3 shows histograms of the random variable $d$ in block TSP with $b' = 8$ and sequential TSP with $b = t/2 = 8$, and two values of $\min(d)$ in each case. The results are from a simulation involving a total of 132 segments (about 4 seconds of speech, with 32 ms segments). Note that, in general, the probability of $d$-values *less than* $\min(d)$ is very small. The probability is non-zero, however, because of occasional situations where a segment has spent the maximum lifespan of $t = 2b$

Table III—Illustration of an unrealizable $min(d) = 6$ in sequential TSP with $b = t/2 = 8$

| Contents of Scrambler Memory m | | | | | | | | Transmitted Segment $s$ | Temporal Distance $d$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 2 | — |
| 1 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 8 | 6 |
| 1 | 3 | 4 | 5 | 6 | 7 | 9 | 10 | 1 | 7 |
| 3 | 4 | 5 | 6 | 7 | 9 | 10 | 11 | 7 | 6 |
| 3 | 4 | 5 | 6 | 9 | 10 | 11 | 12 | | ≥6 |

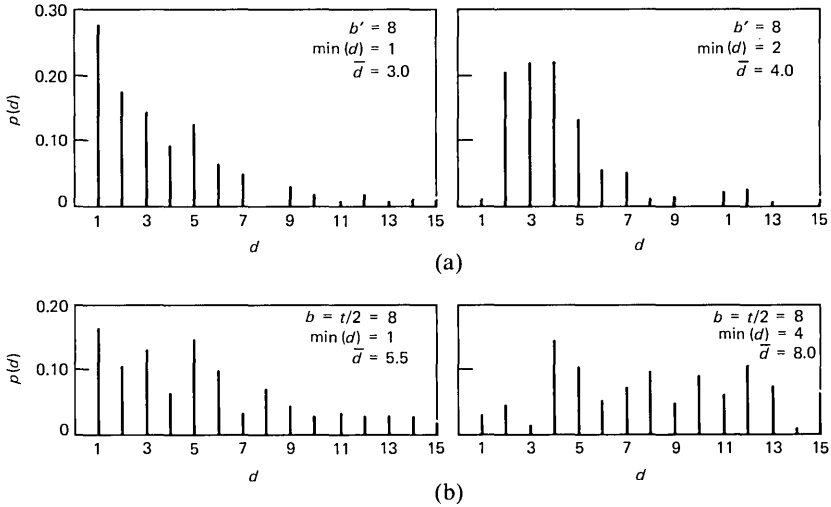



Fig. 3—Histograms of temporal distance $d$ in (a) block TSP ($b' = 8$) and (b) sequential TSP ($b = 8$, $t = 16$). The non-zero probability of $d < \min(d)$ is related to violations of $\min(d)$ because of unconditional releases of segments that have spent a full life-span = 16 in scrambler memory.

times in scrambler memory, and it becomes necessary to release it unconditionally without regard to the resulting value of $d$. Figure 3 also notes respective values of average separation $\bar{d}$. Note that this depends on scrambler type as well as on min $d$. It increases with $\min(d)$ for both types of scramblers, and it is greater for sequential TSP than for block TSP, for the case of $\min(d) = 1$.

Figure 4 compares $\bar{d}$ values for block and sequential TSP scramblers as a function of scrambler memory. Results for the sequential system correspond to the special case of $t = 2b$. Note that $\bar{d} \sim b$ in this case, one-half of $\max(d) = 2b$. The faster increase of $\bar{d}$ in sequential TSP is a result of the greater values of $\min(d)$ that can be specified in this system. Recall from Tables I and II that $\max(d)$ is the same for block and sequential systems for a given value of scrambler memory.

## III. RESIDUAL INTELLIGIBILITY IN BLOCK AND SEQUENTIAL TSP SCRAMBLERS

The residual digit intelligibility in a block TSP system has been the subject of a recent comprehensive article.[1] The emphasis in this section is on the digit intelligibility performance of sequential TSP, as evaluated in formal listening tests.

### *3.1 Test conditions*

The duration of speech segments was $B = 16$ ms for all schemes. This is a design that provides a useful compromise between the

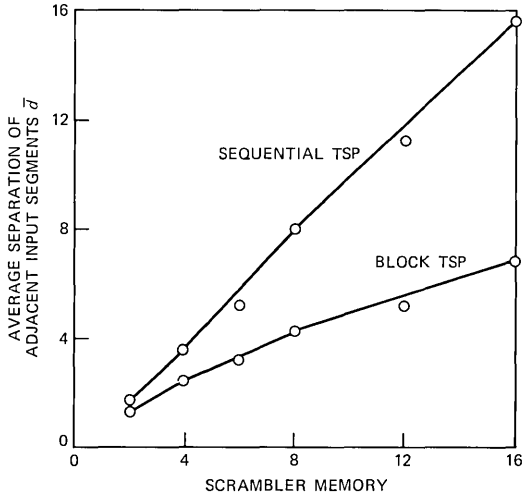



Fig. 4—Average temporal distance $\bar{d}$ as a function of scrambler memory in (*i*) block TSP and (*ii*) sequential TSP with $t = 2b$. Minimum specified distances are the min($d$) values in Table II.

conflicting requirements of encoding delay and bandwidth expansion.[1] Total communication delay ($C_t$) and minimal separation min($d$) were variables in the test. The three delay conditions that were tested were 256, 384, and 512 ms. The min($d$) values that were combined with each delay condition were the smallest and largest values that could be successfully assigned for each type of scrambler. Of course, the smallest value of min($d$) in all cases is 1, which would, at most, allow adjacent segments to be in reverse order when transmitted. However, the maximum value that min($d$) can assume depends upon the type of scrambler and the length of the memory buffer, as shown in Table II. The three delay conditions at each of two min($d$) values generated six test conditions for each type of scrambler, block and sequential.

One other condition was also included with the sequential scrambler using a delay of 512 ms. This included a $\mu$-law logarithmic compression at the scrambler output, with $\mu = 100$.

Before making the test tapes, we listened to recordings of four-digit numbers (as described in Section 3.2 below) as processed by each of the test conditions. Two sets of recordings were employed, one from a male talker and one from a female talker. The recorded speech bandwidth was 200 to 3200 Hz in each case. The consensus of opinion in this pilot listening session was that the intelligibility was higher for the recordings of the female talker, presumably because her speaking rate was slower. The average duration of the four-digit utterances was 2.34s for the female talker, as compared to 1.99s for the male talker. Recordings from both talkers were included for the final test.

The speech samples were 50 four-digit numbers such as 3860, spoken as: *three-eight-six-zero*. Each talker recorded a different list of fifty numbers. The list of numbers was balanced so that within every set of ten numbers, each of the ten digits occurred four times, and each occurrence of a given digit was at a different position in a number. Although the same list of 50 numbers was used for all the tests with a given talker, each subset of ten numbers was presented in a different random order in each test.

The use of digits rather than continuous speech as test inputs follows the procedure in previous tests.[1] Conversational speech has redundancies that make the task of an analog scrambler more difficult; sentence intelligibility, for example, would be higher than word intelligibility as a result of these redundancies. There are no such redundancies in the digit strings in our tests, these strings being sequences of randomly chosen digits. But still, our experience with analog scramblers indicates that digit intelligibility, measured as discussed, is a fairly critical test of scrambler performance. The fact that there is a limited stimulus vocabulary (of ten) makes the task of the analog scrambler quite difficult, perhaps more so than in the case of a continuous speech input, which has a much larger, albeit redundant vocabulary. In the context of scrambled speech,[3] as well as in the context of speech corrupted by additive noise,[7] there is clear evidence that digit intelligibility scores tend to be much higher than word intelligibility scores.

### 3.2 Test procedure

The subjects were employees of Bell Laboratories at Murray Hill, New Jersey. They were not formally trained listeners of scrambled speech. Each scrambler scheme was judged by 18 subjects, although each subject judged only six schemes, chosen to represent the range of expected intelligibility. The subjects listened to the recordings through earphones while seated in a sound-treated booth. They were told to listen to each number and write the four digits they heard on their answer sheets. They were also told that some of the numbers would be difficult to understand and, if they were uncertain, they were to write their best guess rather than to leave blanks.

### 3.3 Results

For each scrambler type, the mean and standard deviation of the percent-correct identification of the digits was computed for each digit position of the four digit numbers. These values confirmed two observations that we had made in the pilot test mentioned earlier.

The mean intelligibility scores confirmed our observation about the effect of the slower speaking rate (by the female talker) on TSP performance. The effect was most apparent in the scores for the two

voices with the sequential scrambler and $C_t = 512$ ms. While the mean intelligibility scores for the two middle digits of this condition were only 66 and 61 percent with the male talker, the same scores were 84 and 91 percent for the female talker.

In the preliminary listening session, we had also noticed that the first and last digits of the four-digit numbers seemed to be easier to identify than the two in the middle. To see whether this was also true in the test data, the percent-correct identification was computed for each digit position. The scores illustrated quite clearly that the digit position affects the residual intelligibility. At each time delay, the percent correct for the third-digit position has the lowest value and the fourth digit position the highest value. Digit positions one and four have less context than positions two and three in the sense that digit position one has no left neighbor and digit position four has no right neighbor. The consistently lower scores of the third digit position are very likely due to the strong influence of context. Indeed, the scores at this digit position are probably a better indication of the level of intelligibility that could be expected in continuous speech where pauses are less frequent.

Intelligibility scores showed that there is no general advantage in using $\min(d)$ values greater than one in sequential TSP. The result is surprising in view of the effect of $\min(d)$ on the average temporal separation $\bar{d}$ (see Fig. 3). One possible explanation of the result is the presence of some kind of a threshold effect in the perception of temporally scrambled speech.

The upper right-hand corner of Fig. 5 shows residual intelligibility in sequential TSP scrambling as a function of communication delay, for $\min(d) = 1$, and for input conditions most favorable to the scrambler—digit position three and male talker. The lower edge of the cross-hatched region refers to the same most favorable case, while the upper edge refers to the average, over both male and female talkers, and over all four digit positions.

Intelligibility scores for block TSP were significantly different from those of sequential TSP only in the case of a communication delay equal to 512 ms. The dashed-line block TSP characteristic in Fig. 5 refers to the most favorable case of male speaker and digit position three, and to the (only available) example of $\min(d) = 2$.

The results of Fig. 5 indicate that there is no intelligibility advantage in sequential TSP as compared with block TSP at low values of delay $C_t$. When $C_t$ is increased to 512 ms, the sequential approach produces a significant reduction of about 10 percent over the block approach. The condition involving $\mu$-law compression of speech reduces the residual intelligibility even more, as shown by the point marked $\mu = 100$. The refinement of $\mu$-law compression is simple to implement,

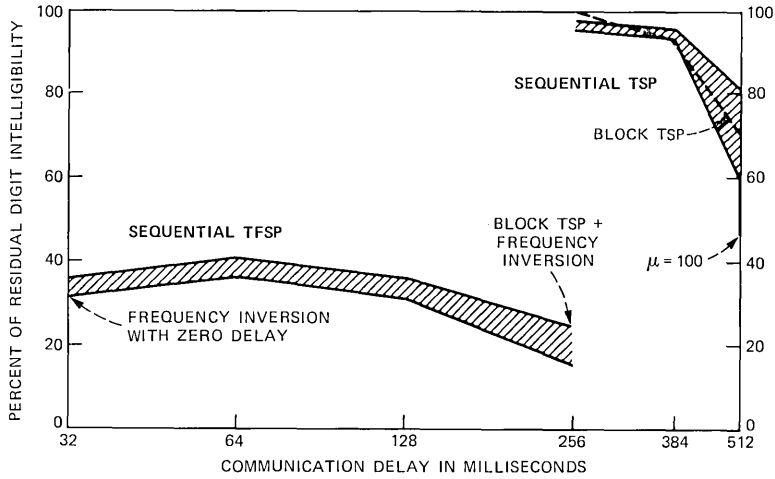



Fig. 5—Mean intelligibility of digits as a function of delay for sequential TSP sequential TFSP scramblers. Upper curves refer to averages over all four digit positions and over both male and female speakers. Lower curves refer to scores for the third digit position for the male speaker. The dashed lines refer to block TSP, digit position three, and male speaker.

although a penalty paid for reduced intelligibility in this case is increased bandwidth expansion and real-channel sensitivity, as compared with the conventional case of $\mu = 0$.

### *3.4 Discussion*

It is interesting that sequential TSP, in spite of its better segment separation properties, provides an insignificant reduction of intelligibility as compared with block TSP, at low values of delay $C_t$. However, this result can be reconciled with the objective results in Fig. 4, which show that the increase of average speech segment separation owing to sequential scrambling is an increasing function of scrambler memory. In the "one-dimensional" procedure of time segment permutation, the only way of providing greater scrambling memory is by increasing delay and, apparently, the objective separation gain becomes perceptually significant only at $C_t$ values in the order of 512 ms.

In the task of identifying one of ten possible digits, the lowest meaningful intelligibility score is 10 percent, corresponding to purely random guessing. This lower bound is particularly meaningful if listeners do not use complex cues and indeed perform decision tasks with ten alternatives. Clearly, none of the TSP scramblers in this study approaches a score in the order of 10 percent. This reinforces our earlier stand[1] that time permutation is best used in conjunction with frequency manipulations such as frequency inversion or frequency band permutations to provide practical and useful values of residual

intelligibility. The much lower residual intelligibility of TSP with frequency inversion is illustrated by the 25-percent result in Fig. 5, for the case of block TSP and $C_t = 256$ ms, a condition tested in earlier work.[1] In that study, frequency inversion alone provided a residual intelligibility of 30 percent. Unfortunately, however, frequency-inverted speech has identifiable characteristics that can be learned,[8] and frequency inversion is also very easy to undo. The next section describes another "two-dimensional" procedure for analog scrambling; it employs sequential permutations of a time-frequency speech matrix using sub-band partitions of 16 ms time segments. This two-dimensional procedure also realizes a residual intelligibility on the order of 15 to 25 percent. In addition, it has better cryptanalytical properties than TSP with frequency inversion.

## IV. PERMUTATIONS WITH A TIME-FREQUENCY MATRIX: TIME-FREQUENCY SEGMENT PERMUTATION (TFSP)

In this section, we propose a scrambler that provides a simultaneous and fully two-dimensional manipulation of both time and frequency information in speech. The permutations of time-frequency segments are based on the sequential permutation approach described in Section 2.3.

The basic principle of the proposed scrambler can be explained with reference to Fig. 6. The $(f \times b)$ matrix depicts a total of $fb$ time-frequency segments. These belong to $b$ contiguous time segments of speech, each of which is split into $f$ contiguous frequency sub-bands or segments. The scrambler memory is considered to be equal to the product $fb$. For subsequent discussions, the contents of this memory can be considered to be a one-dimensional array of $fb$ time-frequency segments. A random number algorithm picks one of these $fb$ segments for transmission. When this $p$-th segment is transmitted, the contents of all $q$-th cells $(q > p)$ are promoted by one position in the memory, and the $fb$-th cell is filled by an incoming $(fb + 1)$-th time-frequency

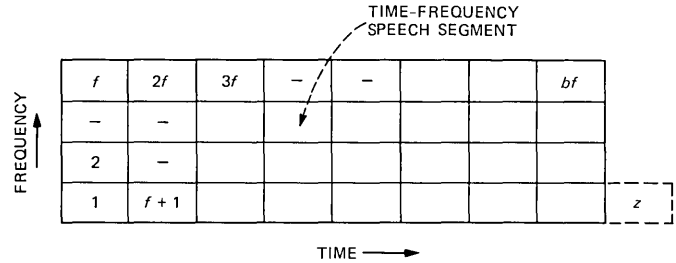



Fig. 6—Sequential permutations of a time-frequency speech matrix: sequential TFSP. Time-frequency segment $z$ enters scrambler memory as soon as $p$-th segment in memory $(1 < p \leq bf)$ is randomly selected and transmitted.

segment. This random scrambling procedure is repeated with two constraints previously described in the discussion of sequential TSP (see Section II):

(*i*) No time-frequency segment stays in the scrambler memory matrix for a number of stages greater than $t_f = 2bf$. This corresponds to the special case $t = 2b$ in the one-dimensional case [Section 2.3; eq. (4) and Table II]. It implies a total communication delay of $C_t = 2fb$ segments. However, since $f$ successive frequency segments are combined to produce one time segment, the total communication delay is independent of $f$, and given as in the one-dimensional case, by

$$C_t \text{ (in ms)} = 2bB,$$

where $B$ is the duration (in ms) of a time segment. As in Sections II and III, $B = 16$ ms.

(*ii*) Contiguous time-frequency segments in scrambler output can be arranged to have a separation which, with a high probability, exceeds $\min(d_f)$; $1 < min(d_f) \leq fb/2$. The parameter $\min(d_f)$ provides a trade-off between average segment separation in scrambler output, which is an increasing function of $\min(d_f)$, and total number of possible permutations, which is a decreasing function of $\min(d_f)$.

Every set of $f$ successive outputs from the scrambler matrix is reconstituted into a pseudo-speech time segment for transmission. One-dimensional sequential TSP is the special case of $f = 1$ in Fig. 6.

## V. RESIDUAL INTELLIGIBILITY IN A SEQUENTIAL TFSP SCRAMBLER

Residual intelligibility tests were conducted following the same general procedures used for the block and sequential TSP tests (see Sections 3.1 and 3.2). The same original recordings of four-digit numbers spoken by a male and female talker were used as speech inputs. Eighteen subjects listened to the recorded digits as processed by each test scheme.

### *5.1 Test conditions*

A total of seven conditions were tested, with the simple design of $\min(d_f) = 1$ in each case. The first four conditions correspond to the $t_f = 2bf$ design, with $f = 4$ and $b = 1$, 2, 4, and 8. Corresponding memory sizes are 4, 8, 16, and 32 time-frequency segments. Corresponding communication delays are $2bB = 32b = 32$, 64, 128, and 256 ms. The other three conditions tested used $b = 8$ also, but values of $t_f \neq 2bf$; specifically, $t_f = 1.5bf$, $3bf$, and $4bf$. In other words, the memory size is fixed at 32 segments for these three cases, but the maximum age in memory varies from 1.5 to 4 times the memory length. The communication delays corresponding to these three conditions are 192, 384, and 512 ms.

### 5.2 Results

The mean and standard deviation of the percent-correct identification was computed for each digit position of each test condition. In general, the subject variability was not large. The average standard deviation was about 6 percent for the male voice and slightly higher, 9 percent, for the female voice. However, there was one subject whose scores were consistently very high, 1.5 to 2.25 standard deviations above the mean. (On one test he correctly identified 72 percent of the digits in the third position while the average for the remaining subjects was 37 percent.) When questioned, he claimed that he did not have any special strategy. However, a closer examination of his data showed a pattern that was evident to some extent in the scores of other better-than-average listeners. The data suggest that these listeners would focus their attention on one of the four digit positions, even when the pauses between digits were difficult to detect.

Since the four digit numbers were balanced and presented in groups of ten, their scores at each digit position for each group could be compared. For instance, for one type of scrambler, the unusual subject mentioned above had scores of 0.20, 0.40, 0.10, and 0.60 for the four digit positions of one group of ten numbers. For the next group of ten numbers, his scores were 0.60, 0.50, 0.30, and 0.10, suggesting that he had shifted his attention to the first two digit positions while listening to the second group of ten numbers. Because of the difference owing to the speaking rate of the two talkers, the mean intelligibility (across subjects and digits) was compared for the two talkers. Even though the differences in the average scores were small—on the order of 5 to 8 percent—the scores for the female voice were consistently higher and, in all cases but one, the difference was statistically significant.

In general, the scores for different digit positions did not display the context effect mentioned for one-dimensional TSP. On the whole, about 90 to 95 percent of the subjects' scores were not significantly different for different digit positions. The remaining 5 to 10 percent of the subjects, who had significantly different scores owing to digit position, were generally the listeners whose overall scores were higher than the average. Although the effects of talker and digit position were not as strong as in the TSP experiment, to be consistent with those results, the scores of the third digit position with the male voice were again evaluated separately, and considered as the closest approximation to scores with continuous speech. These values are indicated by the lower edge of the cross-hatched TFSP region in the lower part of Fig. 5. The upper edge of this region refers, once again, to averages over both male and female talkers, and over all four digit positions.

Four observations are worth noting in the TFSP characteristic of Fig. 5: (*i*) the intelligibility with the smallest communication delay (32

ms) is close to that in frequency inversion (which has a zero communication delay); (*ii*) the intelligibility with a delay of 256 ms is close to that in block TSP (with the same delay) plus frequency inversion; (*iii*) there is a significant drop in intelligibility when the delay exceeds 128 ms (corresponding to a 4 × 4 time-frequency matrix in Fig. 6), and, finally, (*iv*) the intelligibility with a 256-ms delay is significantly higher than the expected lower bound of 10 percent; the characteristic, however, shows a tendency to drop further at delays greater than 256 ms, and with the suggested design of $t_f = 2bf$.

The differences among the scores with the design $t_f = 2bf$ and delays less than 256 ms are not statistically significant, but these scores are all significantly different from the score of 15 percent at a delay of 256 ms. For the case of $b = 8$, the scores at the other three values of $t_f$ ($1.5bf$, $3bf$, and $4bf$) were also not significantly different from each other, but they were all significantly higher than the score when $t_f = 2bf$ (the 256-ms point in Fig. 5). This result suggests that the maximum staying time of $2bf$ may represent an optimal design that minimizes identification. This result is very interesting because the last two of the three $t_f$ conditions above involve communication delays that are 50 and 100 percent greater than the delay in the $t_f = 2bf$ design.

A significant property of all the analog scramblers in this paper is that intelligibility is digit-dependent. This is shown by the illustrative confusion matrices of Table IV. As seen from the diagonal terms in these matrices, digits six and five are the most difficult to scramble. The very high residual intelligibilities for these digits expose what may be an inherent limitation of analog scramblers, at least those based on permutations, as opposed to digital scramblers that transform any given input to an output that sounds like white noise. The fact that the output of the analog scramblers discussed is not the white-noise type is well illustrated by the spectrograms of Fig. 7. Because of the residual structure in these spectrograms, they can be used as the starting point for non-real-time descrambling by a trained eavesdropper. This is especially the case with the TSP spectrogram of Fig. 7b.

The matrices in Table IV also show that the male speaker was easier to scramble than the female speaker. As stated earlier, we feel that this is due to the slower speaking of the female speaker.

### *5.3 Interpretation of digit-intelligibility scores*

An important consideration in interpreting the results of the tests described in this paper is the use of spoken digits as speech input. The lower bound of 10 percent is particularly meaningful if the information available to the listener is limited only to the possibility of the ten digits. Actually, the subjects are trying to recognize phonemes and the phonemes of the spoken words for each digit are not a balanced sample

Table IV—Digit confusion matrices in a TFSP scrambler with communication delay = 256 ms (results are averages over 18 listeners)

| | | LISTENERS' RESPONSE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| | 0 | *0.30* | 0.08 | 0.05 | 0.05 | 0.10 | 0.09 | 0.05 | 0.11 | 0.05 | 0.11 | |
| S | 1 | 0.09 | *0.17* | 0.05 | 0.05 | 0.06 | 0.10 | 0.03 | 0.10 | 0.07 | 0.27 | |
| C | 2 | 0.14 | 0.05 | *0.18* | 0.19 | 0.02 | 0.07 | 0.18 | 0.07 | 0.05 | 0.05 | |
| R I | 3 | 0.14 | 0.08 | 0.15 | *0.14* | 0.11 | 0.07 | 0.15 | 0.07 | 0.05 | 0.03 | |
| A N | 4 | 0.08 | 0.12 | 0.04 | 0.05 | *0.23* | 0.18 | 0.10 | 0.06 | 0.08 | 0.05 | FEMALE |
| M P | 5 | 0.03 | 0.06 | 0.02 | 0.04 | 0.07 | *0.57* | 0.03 | 0.06 | 0.07 | 0.07 | SPEAKER |
| B U | 6 | 0.03 | 0.05 | 0.03 | 0.04 | 0.08 | 0.09 | *0.53* | 0.09 | 0.04 | 0.03 | |
| L T | 7 | 0.06 | 0.11 | 0.03 | 0.07 | 0.07 | 0.15 | 0.10 | *0.28* | 0.07 | 0.07 | |
| E | 8 | 0.06 | 0.13 | 0.07 | 0.06 | 0.06 | 0.14 | 0.15 | 0.07 | *0.23* | 0.03 | |
| R | 9 | 0.08 | 0.10 | 0.04 | 0.06 | 0.04 | 0.22 | 0.02 | 0.06 | 0.06 | *0.32* | |

| | | LISTENERS' RESPONSE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| | 0 | *0.24* | 0.12 | 0.08 | 0.05 | 0.08 | 0.11 | 0.09 | 0.08 | 0.09 | 0.06 | |
| S | 1 | 0.08 | *0.17* | 0.06 | 0.08 | 0.09 | 0.19 | 0.06 | 0.09 | 0.09 | 0.09 | |
| C | 2 | 0.22 | 0.10 | *0.13* | 0.08 | 0.10 | 0.05 | 0.18 | 0.08 | 0.04 | 0.03 | |
| R I | 3 | 0.14 | 0.11 | 0.09 | *0.20* | 0.10 | 0.06 | 0.10 | 0.10 | 0.04 | 0.06 | |
| A N | 4 | 0.08 | 0.13 | 0.07 | 0.07 | *0.18* | 0.15 | 0.11 | 0.10 | 0.03 | 0.07 | MALE |
| M P | 5 | 0.09 | 0.11 | 0.02 | 0.03 | 0.06 | *0.40* | 0.04 | 0.11 | 0.08 | 0.08 | SPEAKER |
| B U | 6 | 0.11 | 0.07 | 0.09 | 0.09 | 0.07 | 0.05 | *0.35* | 0.08 | 0.06 | 0.04 | |
| L T | 7 | 0.15 | 0.12 | 0.06 | 0.06 | 0.08 | 0.07 | 0.16 | *0.16* | 0.09 | 0.05 | |
| E | 8 | 0.09 | 0.13 | 0.09 | 0.07 | 0.09 | 0.09 | 0.12 | 0.12 | *0.13* | 0.08 | |
| R | 9 | 0.18 | 0.11 | 0.04 | 0.09 | 0.09 | 0.12 | 0.08 | 0.10 | 0.10 | *0.08* | |

of the English language. For instance, eight is the only spoken digit with an initial vowel, seven and zero are the only words with two syllables, and five of the ten spoken digits have unvoiced fricatives as the initial phoneme (three, four, five, six, seven). Thus, if an astute listener recognized the first phoneme as an unvoiced fricative, then the probability of being correct by guessing is 20 percent rather than 10 percent. If, in addition, the word were recognized as having only one syllable (eliminating seven), the probability would be 25 percent.

In the first experiment of this series (Section III), the unusually high scores for the spoken digit "six" were observed, but a more detailed analysis was not done. In the analysis of the TFSP data, the scores for each of the spoken digits were computed and compared as shown in Fig. 8. The mean intelligibility of each spoken digit (with digit position disregarded) is indicated for both the male ($M$) and female talker ($F$), for scramblers with $t_f = 2fb$ and $b = 1$, 2, 4, and 8. Labels 1, 2, 4, and 8 refer to these values of $b$; they also indicate respective communication delays of 32, 64, 128, and 256 ms. The generally higher scores for the female talker are apparent. More important, these plots indicate that some of the linguistic cues were affecting the listeners' judgments. The scores for the two-syllable words, zero and seven, are elevated and the scores for five and six are extremely high. The range of scores indicate

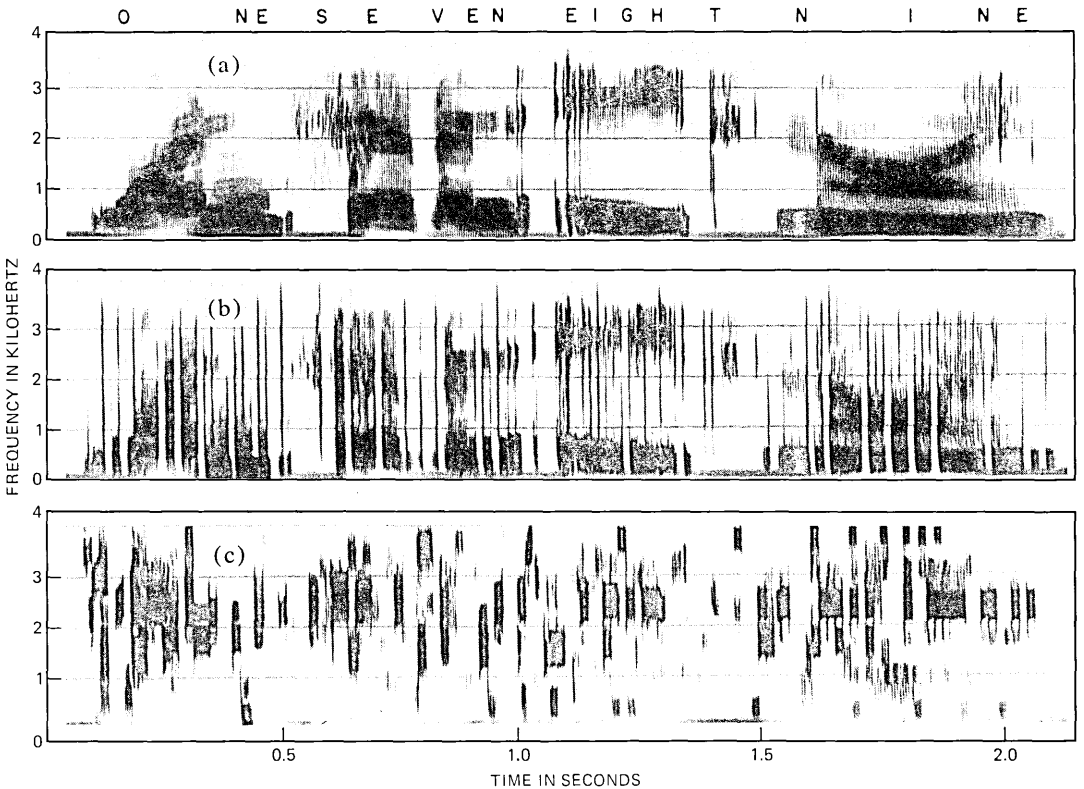



Fig. 7—Speech spectrograms: (*i*) unscrambled speech (female speaker, digit sequence "1789"), (*ii*) output of sequential TSP scrambler with communication delay of 128 ms, and (*iii*) output of sequential TFSP scrambler with the same communication delay of 128 ms.

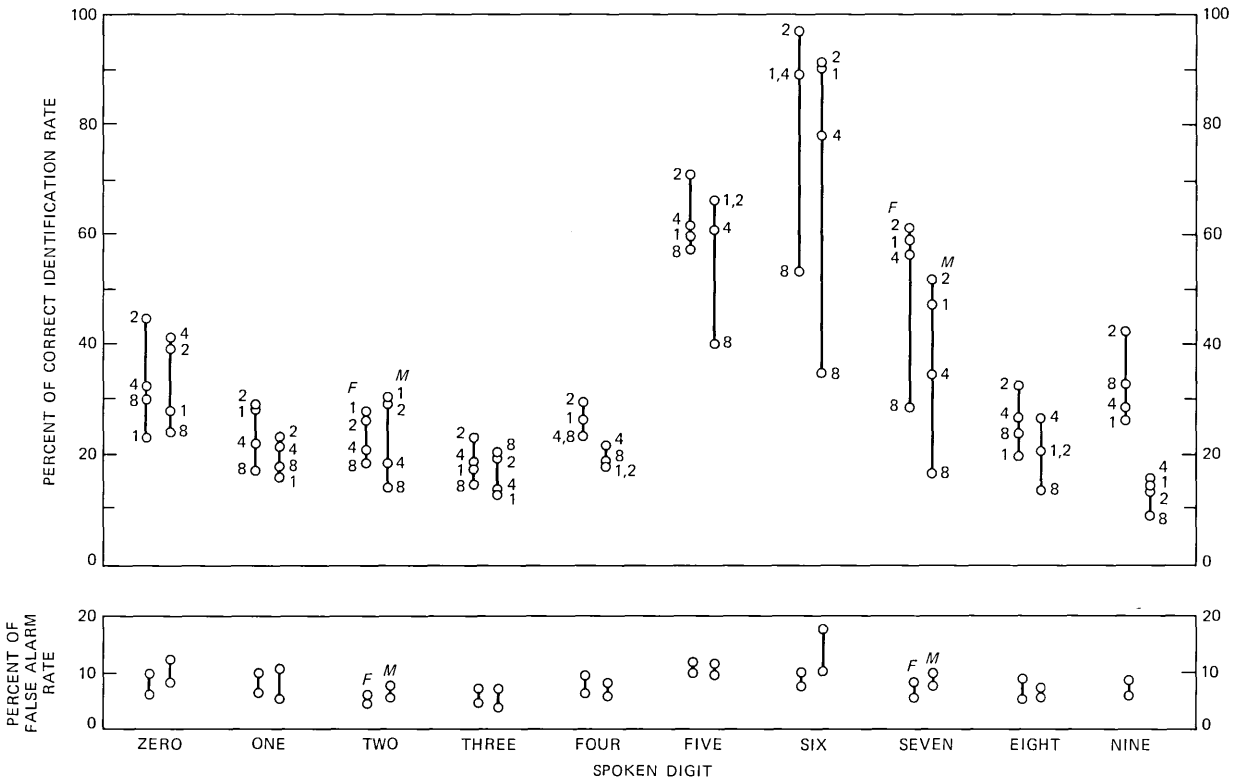



Fig. 8—Correct identification and false alarm rate scores as a function of spoken digit. For each digit, there are two sets of scores, one for female speaker (left vertical bar) and one for male speaker (right vertical bar). These are distinguished by labels *F* and *M* for the digit examples TWO and SEVEN. Each set of scores has four values, corresponding to four values of total communication delay. These four values are identified by numbers that represent multiples of 32 ms. Numbers 1, 2, 4, and 8 therefore indicate delays of 32, 64, 128, and 256 ms.

that the reduced intelligibility with a delay of 256 ms is largely the result of disproportionately lowering the identification of "five," "six," and "seven."

The lower plot in Fig. 8 shows the false alarm rate, i.e., the percent of time that a digit was incorrectly identified as the one labelled in the figure. For simplicity, these plots show only the maximum and minimum values of false alarm rates, as a function of $b$, rather than values for all four values of ($b$), as in the upper plot of the figure. If subjects were really guessing among the ten possible digits when they were uncertain, these values should be about 10 percent. The somewhat higher false alarm rate for the digit six indicates that it was probably being used as a default response by some of the subjects.

The observations of this section reinforce our earlier stand that the intelligibility of connected speech where the words are less predictable may, indeed, be much lower than the scores shown for the spoken digits in Fig. 5.

In Fig. 8, the condition of 256-ms delay (labeled 8 in the figure) represents the least intelligibility in only 14 out of 20 conditions shown, and the condition of 32-ms delay (labeled 1 in the figure) represents the greatest intelligibility in only 3 out of 20 conditions. A good example of monotonic behavior is the digit "two." Good examples of non-monotonic behavior are the female utterances of "zero" and "nine." In the latter two examples, the intelligibility difference between delays of 32 and 64 ms (points labeled 1 and 2) were in fact tested to be statistically significant. We do not have an adequate explanation of why an increase of delay from 32 to 64 ms (and in some cases, to 128 and 256 ms) should actually *increase* residual intelligibility, a feature that is counter to the general trends of the average scores in Fig. 5. (Recall that in these averages, the differences between scores at 32, 64, and 128 ms were stated to be statistically insignificant). The phenomenon of significant intelligibility increases owing to delay seems however to be peculiar to digits with sustained sounds such as the *o* in zero and the *n* in nine. In these cases, certain increases of communication delay, or equivalently, certain increases of staying time in memory, may increase the probability that at least one of many fragments of the long sustained sound gets outputted in an "interference-free" context such as an interdigit silence, causing an increase of intelligibility. On the average, however, separation of intra-digit fragments *increases* as a function of delay, and this causes a *decrease* of intelligibility. This was indeed noted in the average scores of Fig. 5.

## REFERENCES

1. N. S. Jayant, B. J. McDermott, S. W. Christensen, and A. M. S. Quinn, "A Comparison of Four Methods for Analog Speech Privacy," IEEE Trans. Commun., *COM-29* (January 1981), pp. 18–23.

2. R. C. French, "Speech scrambling and synchronization," Philips Res. Rep., No. 9 (1973), pp. 1–115.
3. E. R. Brunner, "Efficient scrambling techniques for speech signals," Proc. Int. Conf. Commun., Seattle, WA, June 1980, pp. 16.1.1–6.
4. S. Udalov, "Microprocessor-based techniques for narrow-band scrambling," in Proc. Int. Conf. Commun., Seattle, WA, June 1980, pp. 16.4.1–5.
5. R. V. Cox and J. M. Tribolet, "Analog Voice Privacy Systems Using TFSP Scrambling: Full Duplex and Half Duplex," B.S.T.J., this issue.
6. S. T. Hong and W. Kuebler, "An Analysis of Time Segment Permutation Methods in Analog Voice Privacy Systems", Proc. 1981 Carnahan Conference on Crime Countermeasures, Univ. of Kentucky, Lexington, KY, May 1981.
7. G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of test materials," J. Exp. Psychol., *41*, (1951), pp. 329–35.
8. B. Blesser, "Speech perception under conditions of spectral transformation: I. Phonetic characteristics," J. Speech and Hearing, *15*, (1972), pp. 5–41.

## APPENDIX

Let the probability of "forced-exit" (i.e., the probability that a segment spends the maximum allowable time in scrambler memory) be $P$. The probability that a segment exits as a result of random selection is therefore $1 - P$. With a buffer size $b$, and a uniform pdf for random segment selections in buffer positions 1 to $b$, the probability of unforced exit from any given buffer stage is therefore $(1 - P)/b$, and the corresponding probability of non-exit is $P_n = [1 - (1 - P)/b]$. With the $t = 2b$ design, the probability of non-exit for all possible ages $s$, $s = 1, 2, \cdots, 2b$ of a given segment is $(P_n)^{2b}$. By definition, this should equal the forced-exit probability $P$. Therefore,

$$P = P_n^{2b} = \left(1 - \frac{1 - P}{b}\right)^{2b}. \tag{8}$$

Taking logarithms and using $\ln(1 - x) \sim -x$ for $x \ll 1$,

$$\ln P = 2(P - 1); \qquad P \sim 0.203$$

for large $b$. Numerical solution of (8) shows that $P$ is extremely close to the above asymptotic value of 20 percent for values of $b > 4$. The value at $b = 4$ is about 0.210.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# Analog Voice Privacy Systems Using TFSP Scrambling: Full Duplex and Half Duplex

By R. V. COX and J. M. TRIBOLET*

(Manuscript received July 20, 1982)

*In this paper we present possible full-duplex and half-duplex analog voice privacy systems that have been simulated over real channels. Previous papers have been concerned primarily with the issues of the strength of a system (i.e., unintelligibility to the casual eavesdropper and relative cryptanalytical strength for the sophisticated eavesdropper) and the amount of delay of a system. Well-known but not addressed have been the problems of decoding the scrambled signal in a real-channel environment. At the heart of the encryption systems proposed here is the sequential time and frequency segment permutation structure proposed by Jayant and Cox. This structure relies on digital processing to divide the signal into sub-bands and then to permute these bands in both time and frequency simultaneously to synthesize the scrambled analog signal. In discussing the decoding we address the issues of compensating for the properties of the channel, re-sampling the analog signal, and establishing and maintaining synchronization between the de-scrambler and the scrambler.*

## I. INTRODUCTION

In this paper we discuss a possible analog voice encryption scheme. Although the channel signal is analog, all of the signal processing is done digitally. This gives us a flexibility that analog processing cannot give, and thereby allows us to do simultaneous time and frequency permutation. It also allows us to perform synchronization and equalization at the receiver. With the advent of the Digital Signal Processor (DSP),[1] digital processing becomes economically competitive as well as feasible.

---

* This work was performed while Mr Tribolet was on sabbatical leave from the Instituto Superior Tecnico, Lisbon, Portugal.

The systems that we will discuss in this paper are a full-duplex system and a half-duplex system. We have assumed either system would be used over all standard telephone channels with a bandwidth of 200 to 3500 Hz. Because of the delays involved in the signal processing, echoes could become a problem, so we chose to look at solutions that would avoid the echo problem. Alternatively, echoes might be cancelled using an appropriate algorithm, but this problem was not addressed in this study. One such solution to the echo problem is to choose a half-duplex system. Alternatively, for a full-duplex system we have divided the telephone band into three segments: 200 to 500 Hz is used for signalling and equalization by both users, 500 to 2000 Hz is allocated to the first user, and 2000 to 3500 Hz is allocated to the second user. In this way faulty echo cancellation need not impair the quality of the speech, since the two speakers use disjoint bands. In a half-duplex system only one user could speak at a time and would be allocated the entire channel except for the frequencies used for equalization.

In this study we have relied almost entirely on computer simulation of the algorithms involved. However, the algorithms were designed with future implementation on the Bell Laboratories DSP in mind.[1] The DSP is a powerful, single-chip, programmable processor that is especially suited for performing digital signal processing functions. It has an 800-ns machine cycle that is established by a 5-MHz clock. It contains provision for a 1024- × 16-bit read-only memory (ROM) for storage of the program, tables, and coefficients. A 128- × 20-bit random-access memory (RAM) is available for the storage of dynamic data and state variables. Our experience in working with the DSP has shown that the three constraints of RAM, ROM, and execution speed all come into play in the final design of the algorithm.

In Section II we give an overall description of the system. Subsequent sections deal with the major components of the system and the problems we addressed and solved for these components. Section VIII presents conclusions.

## II. THE OVERALL SYSTEM

Figures 1a and 1b are overall block diagrams for the proposed full-duplex system's transmitter and receiver. At the transmitter an analog speech signal $s(t)$ is first filtered and sampled at 8 kHz by a standard analog/digital (A/D) converter, such as the M7062 coder-decoder (CODEC). The resulting digital signal is denoted as $s(n)$. This signal is then fed to a filter bank that splits $s(n)$ into three frequency components: $s_1(n)$ is the 0- to 500-Hz band, $s_2(n)$ is the 500- to 1000-Hz band, and $s_3(n)$ is the 2000- to 2500-Hz band. Each of these signals has its sampling rate reduced from 8 kHz to 1 kHz as part of the filtering

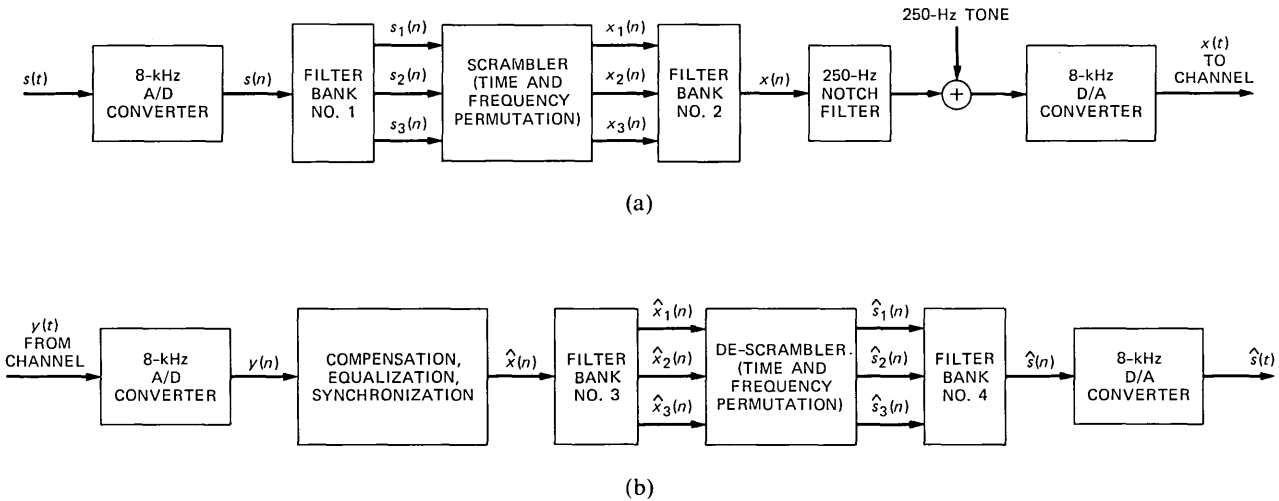



Fig. 1—Block diagram of the proposed full-duplex system. (a) Transmitter. (b) Receiver.

process. For the half-duplex system there are five bands, including these three and also the bands 1000 to 1500 and 1500 to 2000 Hz. The sub-band signals are then fed to a time-and-frequency block scrambler. We have been using a block size of 5 ms. The scrambler uses a pseudo-random key to output three or five blocks from its buffer. These output signals, $x_i(n)$, are fed to a second filter bank. The second filter bank is a synthesis filter bank using quadrature mirror filters. For the full-duplex system it combines the three signals into either a 500-Hz to 2-kHz band or a 2-kHz to 3.5-kHz band, and each user gets one of these bands. In the half-duplex system the five bands are combined into a 500- to 3000-Hz band. As mentioned previously, this division of the available bandwidth can be used to prevent echoes.

At start-up and as required subsequently an impulse is transmitted to equalize the channel at the receiver. A 250-Hz tone is also transmitted continuously to compensate for frequency shift. For this reason $x(n)$ is filtered by a 250-Hz notch filter before adding the 250-Hz tone. Then the synthesized signal is fed to a standard digital/analog (D/A) converter and transmitted as $x(t)$.

The received signal is denoted as $y(t)$ because a number of changes to $x(t)$ may have taken place. The job of the synchronization and equalization subsystem is to produce an output $\hat{x}(n)$ that is as close to $x(n)$ as possible. To produce $\hat{s}(t)$ with as little noise as possible, $\hat{x}(n)$ must have sample-to-sample integrity with $x(n)$. A small slippage of the sampling instants of the receiver relative to the transmitter will result in an error at the ends of the blocks, thus producing a frame-rate-type noise. Because the filters are long the time duration of the noise is much longer than the relative slippage. More details are given about this subsystem in the section on synchronization and equalization. Next, the signal $\hat{x}(n)$ is fed to the receiver analysis filter bank (filter bank #3). This filter bank uses the same quadrature mirror filter (QMF) structure as filter bank #2 (a synthesis filter bank). Because the same QMFs are used, $x_i(n)$ can be almost exactly recovered in the absence of noise and other distortion in the channel. Only QMFs have this property.[2] This is because they are bandwidth preserving. The scrambling process increases the bandwidth of the signal. The individual bands can only be recovered if the increased bandwidth is fully preserved. The de-scrambler delays the time-and-frequency-permuted blocks by an amount complementary to their delay in the transmitter. In this way all blocks experience the same total delay and can be output in the correct order—oldest blocks first. Filter bank #4 is the synthesis counterpart to filter bank #1. Presumably, the $\hat{s}_i$ have the same bandwidths as the $s_i$ and so QMFs are not necessary. However, in order that as much as possible of the 0- to 1000-Hz bandwidth be preserved in the full-duplex system, it is recommended that a QMF be

used to split the 0- to 1000-Hz band and subsequently to reconstruct it. The signal $\hat{s}(n)$ is then fed to a standard D/A converter to form the decrypted output signal $\hat{s}(t)$. For the full-duplex system $\hat{s}(t)$ is not a full-bandwidth signal. Its frequency content is limited to 0 to 1 kHz and 2 to 2.5 kHz. As a result it sounds somewhat muffled. Nevertheless, it seems quite intelligible.

In the following sections details will be presented on the major subsystems: the filter banks, the scrambler and de-scrambler algorithms, and the synchronization-equalization system.

## III. THE FILTER BANKS FOR THE FULL-DUPLEX SYSTEM

Considerable time and effort went into the design of the filter banks used here. Our primary constraint was to come up with a set of filter banks such that each filter bank would occupy only one DSP. At the same time the overall structure must have fairly sharp filters to preserve as much bandwidth as possible of the original speech, since this in turn preserves intelligibility. Also, aliasing between speakers had to be prevented.

Figure 2 shows the structure of filter bank #1. The signal $s(n)$ is first split into two bands (0 to 1 kHz and 2 to 3 kHz). A 95-tap finite impulse response (FIR) filter is used for each band. Since both filters use the same input data the memory required can be shared by both. That way only 95 RAM locations are used. The output of each filter is decimated by four since each filter accomplishes a 4:1 bandwidth reduction on the original signal. This also reduces the computation load on the processor. The output of the low-band filter is fed to a QMF pair to split the signal by another factor of two, producing $s_1$ and $s_2$. These two filters can also share RAM memory, thus using 16 additional locations. The high-band signal is low-pass filtered using the same QMF to produce $s_3$. This requires 16 more RAM locations in the DSP. In all we have used 127 DSP RAM locations. One way of

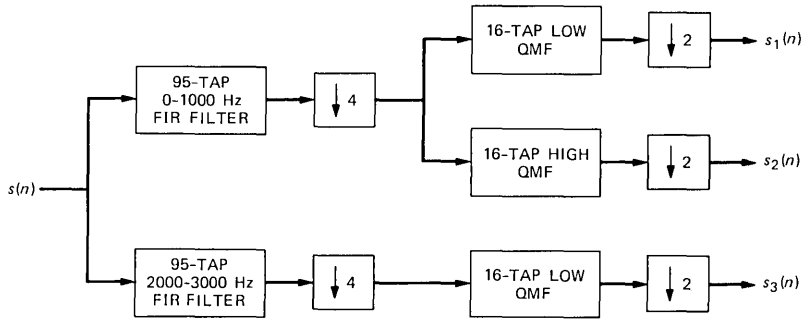



Fig. 2—Structure of filter bank #1 for the full-duplex system.

viewing this structure is that for every eight input samples there will be three output samples. Thus, there is ample computation time for the DSP.

Filter bank #4 uses the same filters. Its structure is shown in Fig. 3. Each QMF requires only eight RAM locations since the input signals are being interpolated 2:1. This accounts for 24 RAM locations. The 95 tap filters each require 24 RAM locations since they are interpolating 4:1. Thus, only 72 RAM locations are used. Larger filters might have been used but they do not add appreciably to the quality. The number of multiply-accumulates per output point is also 72, since each RAM location must be accessed once per output sample. This is also ample time for the DSP.

Figure 4 shows the structure of filter bank #2. The three outputs from the scrambler are first interpolated by 16-tap QMFs. This requires eight RAM locations for each filter, or 24 total. The two outputs from $x_1$ and $x_2$ are combined using a QMF pair before the next stage. The two outputs from the first stage are combined using a 32-tap QMF pair in the second stage. This produces a 4-kHz sampled signal with a bandwidth of 500 to 2000 Hz. Depending on whether the user is assigned to the upper or lower band, the next QMF is either high or low. The second stage required 32 RAM locations as does the third stage. This gives a total RAM requirement of 88 and also means 88 multiply accumulates are performed per output point. The 64-point QMF is a fairly sharp filter with a large rejection band. It is quite adequate to keep separate the encrypted speech of the two talkers. All of the QMFs were designed by Johnston.[3]

Filter bank #3 uses the same QMFs for analysis and its structure is shown in Fig. 5. The encrypted speech is first filtered by a 64-tap QMF. This is a high- or low-pass filter, depending on the band assigned. The output is decimated 2:1 to put the signal on the 0- to 2000-Hz band. A 32-tap QMF pair splits this band and then 16-tap QMF filters are used to do the final splitting. Stage 1 requires 64 RAM locations and stages

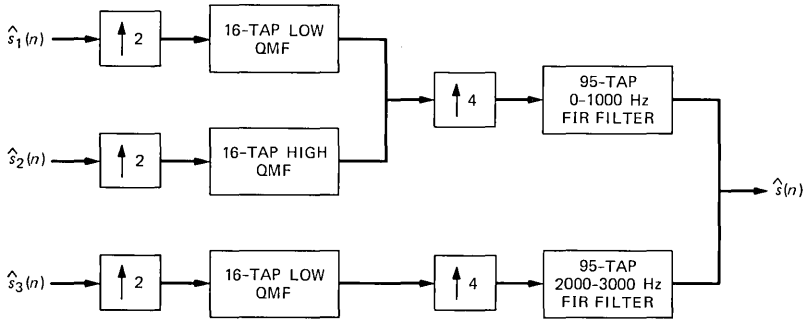



Fig. 3—Structure of filter bank #4 for the full-duplex system.

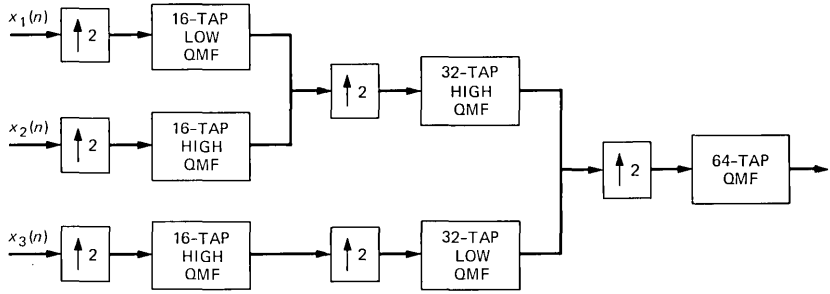



Fig. 4—Structure of filter bank #2 for the full-duplex system.

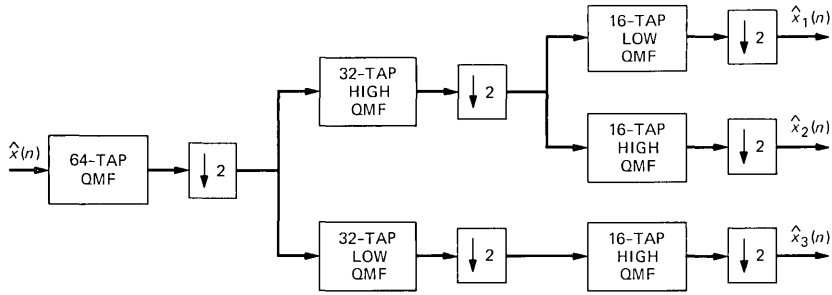



Fig. 5—Structure of filter bank #3 for the full-duplex system.

2 and 3 require 32 each, thus using all 128 RAM locations in a DSP. As in the other analysis filter, eight input samples produce three output samples, thus allowing adequate time for the computations.

The reason for using quadrature mirror filters in filter banks 2 and 3 is worth reiterating. Since we do segment permutation both within and between bands, the bandwidth of the signals $x_1$, $x_2$, and $x_3$ will be the full 500 Hz. If we chose a filter that did not preserve the full bandwidth, these signals would be altered. The result would sound like frame rate noise or "burble." By using the QMF structure the bands will be preserved to within 30 dB with no loss of bandwidth. Conversely, if we chose a full bandwidth filter other than a QMF, then when the bands were combined they would alias with each other. This also happens with a QMF, but when they are separated by the same QMF in filter bank #3 the aliasing is cancelled. This is not the case with an ordinary full-bandwidth filter. This is why the QMF structure is essential.

## IV. THE FILTER BANKS FOR THE HALF-DUPLEX SYSTEM

For the half-duplex system quadrature mirror filters were used in all the filter banks. The five bands of the input speech that are used are

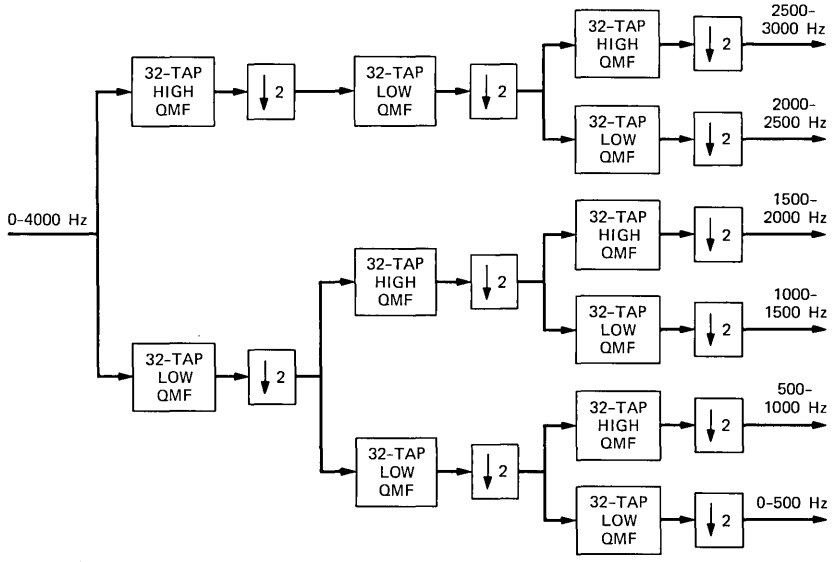



Fig. 6—Analysis filter bank for the half-duplex system.

0 to 500, 500 to 1000, 1000 to 1500, 1500 to 2000, and 2000 to 2500 Hz. A three-stage filter bank structure was devised as shown in Fig. 6. In the first split a 32-tap filter divides the 0- to 2000- and 2000- to 4000-Hz bands. In the next stage the same 32-tap filter was used to divide these two bands into four bands: 0 to 1000, 1000 to 2000, 2000 to 3000, and 3000 to 4000. The highest band was discarded. In the third stage each of these three bands is again split using the 32-tap QMF. After each stage of filtering 2:1 decimation takes place. Thus, each band is sampled at 1000 Hz. For the transmitter the 2500- to 3000-Hz band is discarded. However, for the receiver it is the 0- to 500-Hz band that is discarded.

The synthesis filter bank counterpart is shown in Fig. 7. For the transmitter nothing is put in the 0- to 500-Hz band, since it is used later for the phase-roll compensation. For the receiver nothing is put in the 2500- to 3000-Hz band.

These filter banks require more RAM memory than the full-duplex banks did. The analysis filter bank requires 192 memory locations while the synthesis filter bank requires 160 memory locations. No structure could be found that used less than 128 locations and also gave good performance.

## V. THE SCRAMBLER AND DE-SCRAMBLER

The scrambler and de-scrambler are based on Jayant's idea of a rolling code scrambler[4] with one new twist. Figure 8 shows a memory

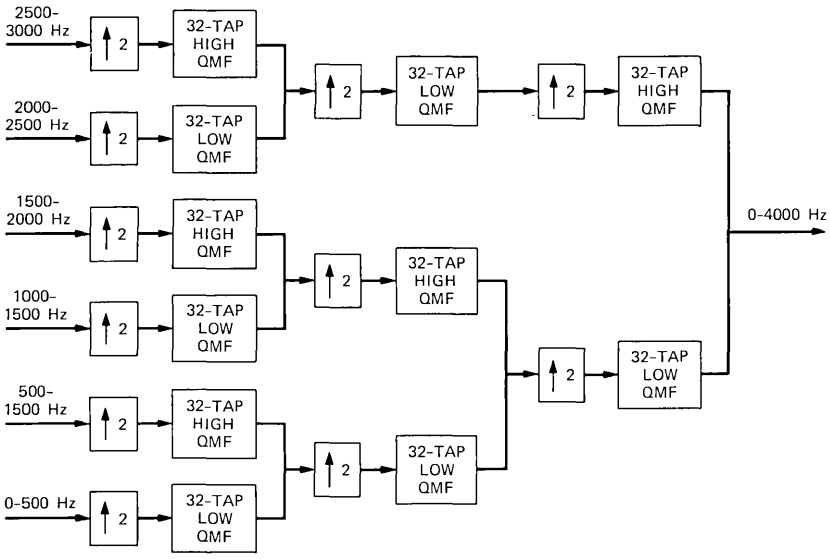



Fig. 7—Synthesis filter bank for the half-duplex system.

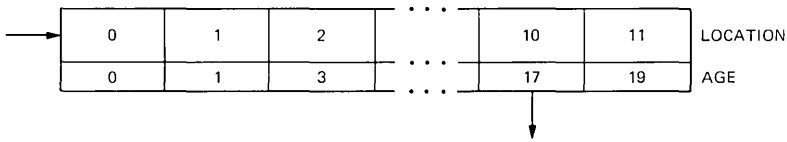



Fig. 8—Scrambler buffer.

buffer. The locations actually correspond to blocks of memory. The age is the age of the block in that location. At any instant data is always being entered into the leftmost block; hence its age is always zero. One of the blocks is also being emptied from the buffer. In this example it is block 10. The algorithm for choosing the block to output is derived as follows. First, the rightmost block (block 11, here) is checked to see if it has reached its maximum age. It if has, then it must be transmitted. Otherwise a random number generator is used to pick a random integer from 0 to 11, and that number determines the block chosen. Once a block is emptied, all blocks to its left are shifted right one location, thus freeing up block 0 for a new input.

The de-scrambler buffer shown in Fig. 9 is similar to the scrambler buffer. Here we see block 10 from Fig. 8 being entered into the buffer. Block 23 is always the oldest block in the buffer and so it is chosen for emptying. Other blocks in this buffer are kept in order of increasing age. So we see that the structure for the two buffers is identical. The method chosen for selecting the output block from the scrambler

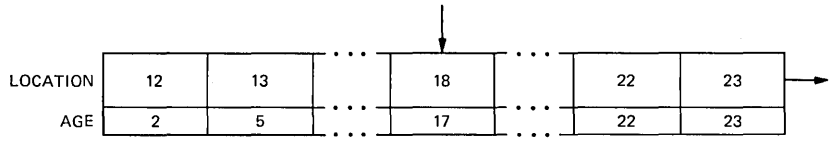



Fig. 9—De-scrambler buffer.

buffer is equivalent to determining where to place the input block in the de-scrambling buffer. Note that the total delay for all blocks will be the same. All of this is exactly the same as Jayant's ideas.

In any real system provision must be made for resynchronizing the de-scrambler if an error occurs at the receiver. This is the basis for modifying Jayant's ideas. The new idea presented here is to periodically reset the transmitter or scrambler buffer to its initial state. In its initial state the age of each block corresponds to its location, i.e., block 11 is only age 11, etc. Suppose we want the period between re-initialization to be $N$ and we have a buffer with $M$ blocks in it. At times 1 through $N - M$ we use Jayant's ideas as just described to select the output block. At time $N - M + 1$ we eliminate block 0 as a candidate for transmission (we still have $M - 1$ blocks to choose from). At time $N - M + 2$ we eliminate both blocks 0 and 1. We continue in this manner, restricting ourselves to a smaller set of blocks each time. At time $N - 2$ we can only choose block $M - 2$ or block $M - 1$ and at time $N - 1$ we can only choose block $M - 1$. At that point the buffer will have been re-initialized, i.e., at time $N$ the locations and ages of the blocks will be in correspondence, just as they were at time 1. At time $N$ we substitute a synchronization pulse for block $M - 1$. Note that the eavesdropper without the key will still be unable to tell what block is coming out at any given time, even though one can tell when the key ends by detecting the sync pulse.

In our proposed system each block represents a frequency band as well as a time. At any given instant three or five blocks are being entered rather than one. This easily can be accounted for without modifying the algorithm very much. At any time we have three or five output pointers instead of one. We just read the key once for each band. The 3- or 5-band structure can be used to our advantage in doing the re-initialization. Suppose we let $N$ and $M$ be multiples of three for the full-duplex case or five for the half-duplex case. The $N$-th entry in the key will correspond to the output for the highest band. At time $N$ we are supposed to transmit the known sync signal. Because of the way we constructed the key, the omitted data in block $M - 1$ will also be from the highest band and therefore its loss will cause less degradation to the speech than if it were from a lower band. At the receiver the reception of this known signal indicates that the scrambler buffer

has been re-initialized and so the de-scrambler buffer should also be re-initialized. While this procedure will allow a potential eavesdropper to know the period of the key, it will not give away the key or any more information than before. At the receiver when the known sync signal is received, zeros are substituted so that the sync signal is not heard as part of the decrypted signal. The momentary loss of bandwidth in the decrypted signal does not affect intelligiblity while the signaling "chirps" that we send are quite noticeable in the encrypted signal.

We note that for the buffer sizes shown in Figs. 8 and 9, here, namely 12 blocks with five samples per block, a single DSP is more than adequate for use as the scrambler buffer. A DSP would be sufficient for a buffer twice this size.

## VI. SYNCHRONIZATION AND EQUALIZATION

Our system uses the same signals for both synchronization and equalization and therefore we discuss them together. First we discuss what we mean by these terms. Then we address how the problems are solved.

As mentioned previously sample-to-sample integrity is essential if a good quality $\hat{s}(t)$ is to be recovered from $y(t)$. This means both that the input samples $\hat{x}(n)$ should be as close as possible to $x(n)$ and that the state of the de-scrambler should exactly match that of the scrambler. Thus, there are two types of synchronization to perform. We must synchronize the de-scrambler state with the scrambler state and match $\hat{x}(n)$ with $x(n)$.

The signaling chirp indicating the re-initialization of the transmitter is used to synchronize the state of the receiver with the transmitter. This chirp can be detected to one of two ways. One way is to use a matched filter to detect the chirp. The advantage of this method is that it can be used on the full-band, 8000-Hz sampled signal. The alternate method is to use the filter bank structure and apply a match filter to $\hat{x}_3$ or $\hat{x}_5$. This method will work provided that the decimation cycles of the filter bank structure are in the proper phase. In general the odds of randomly selecting the right phase are 7:1 against. For this reason we use the first method for start-up. After sync is established the second method can be used to check sync.

To match $\hat{x}(n)$ with $x(n)$ we need to do channel equalization so that the two sets of samples are as close as possible. In turn this means looking at all the possible channel degradations and being prepared to handle each one. Figure 10 shows some of the things that can happen to the signal $x(n)$. First the signal is passed through an anti-aliasing filter. Next it is modulated up by a frequency $f$ and passed through the channel. We can view the passage through the channel as more filtering

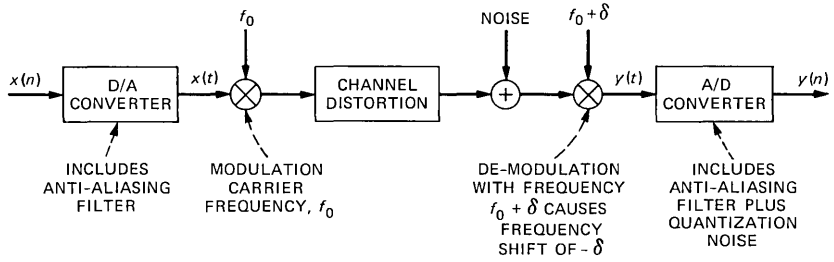



Fig. 10—Channel model.

followed by additive noise. The signal is then demodulated by a frequency $f + \delta$, which causes a frequency shift, or "phase roll," of $-\delta$. Finally, in the A/D process it is filtered once more and quantization noise is added. Fortunately, all of these degradations are independent and can be attacked separately.

We first argue that the noise will have the same effect as white noise of the same power, since the speech bands are constantly being permuted. Other than that the noise is essentially left unchanged, since we have no way to remove it.

We next attack the frequency shift and remove it. This frequency shift is also referred to as "phase roll" in the literature. A small shift in frequency of normal speech would probably not be detectable. However, for our work it is catastrophic. This is because the QMF filters cannot tolerate any frequency shift if they are to properly cancel out aliasing.

To compensate for the phase roll we propose that a 250-Hz sidetone be transmitted. This tone will be received with the same frequency shift as the rest of the signal. Figure 11 shows a novel phase-roll compensation scheme we have devised that works quite well. In Figure 11 the input signal branches immediately. The top branch goes first through a sharp bandpass filter (BPF #1), which passes only 240 to 260 Hz. The purpose is to immediately isolate the 250-Hz tone (which may have been slightly shifted). Next the tone is frequency inverted, making it approximately 3750 Hz. Then it is interpolated 2:1 using a second bandpass filter as the interpolation filter (BPF #2) with a center frequency of 3750 Hz. This signal is now modulated by a 250-Hz tone and then bandpass-filtered again (BPF #3). This produces a modulation signal of about 4000 Hz. If the original signal was shifted up $\delta$ Hz, the modulation signal will be 4000-$\delta$ Hz. In the second branch the remainder of the signal is interpolated 2:1 and then modulated up by 4000-$\delta$. It is then decimated and frequency inverted. In this way the received signal $s(n)$ is modulated by exactly the right amount to compensate for the phase roll. One might even say the phase roll is

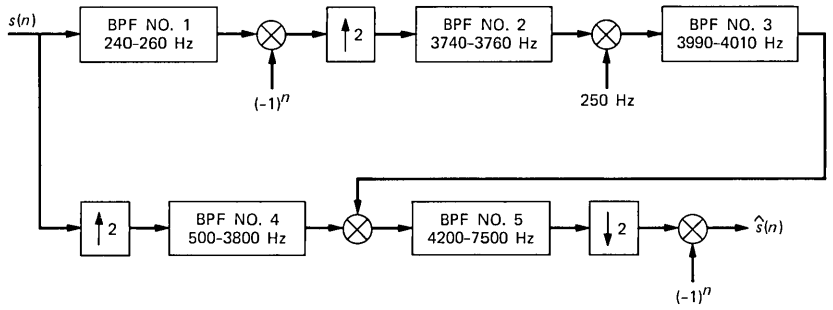



Fig. 11—Phase-roll compensation network.

made to compensate for itself. The interpolation filter in the second branch is actually a bandpass filter that removes the 250-Hz sinewave. Thus, the final output signal does not contain the 250-Hz sinewave.

The phase-roll compensation scheme was tested using computer simulations. The received signal was passed through a computer program that introduced an unknown amount of phase roll. The output from this program was then passed through the phase-roll compensation system. The compensation system completely cancelled the phase roll and its effects, but introduced a small amount of background noise. The noise appeared to be independent of the amount of phase roll in the signal. It was such a low-level noise that a listener needed to be in a quiet room to hear it. On this basis the compensation scheme was judged to work well.

Next, we can measure $h(n)$ by transmitting an impulse and measuring the response at the receiver. If noise is a problem, the measurement can be done several times. Given $h(n)$, $n = 0, 1, \cdots, N$ we form $g(n) = h(N - n)$, $n = 0, 1, \cdots, N$. If we convolve $h(n)$ with $g(n)$, the result will be symmetric sequence $f(n)$, $n = 0, 1, \cdots, 2N$. Because $f(n)$ is symmetric it is linear phase. This suggests that if we use $g(n)$ to filter all the data, $y(n)$, the output will be an integer-delayed version of the signal $x(n)$ with some amplitude modification in the different frequency bands. If we feed $f(n)$ into filter bank #3, we can find the gain to apply to the signals in each band to correct for the amplitude modification. In general the frequency response of $h(n)$ is fairly flat in amplitude for the frequency bands of interest. We have found that an impulse response length of 64 samples centered around the maximum value of the impulse response works well.

## VII. SIMULATION RESULTS TO DATE

Simulation results to date have been extremely encouraging. In the transmitter program some header information is first generated. This

consists of a sine wave that has one of two frequencies, 1000 Hz or 3000 Hz, and thereby determines whether the receiver will use the high channel or low channel. This sine wave is 32 ms long. After a 16-ms wait it transmits an impulse for measuring the channel's impulse response. After 16 ms more it transmits the sine wave again for 32 ms. It then begins the encryption process by sending the "chirp" sync pulse in $x_3$ (or $x_5$) to indicate that the buffer is in initialized status. Superimposed on the signal throughout is a 250-Hz sine wave for the phase-roll compensation. The output of this program is the digital file $x(n)$.

The first channel we used for transmission was just to output $x(n)$ to the computer's D/A and loop right back through the computer's A/D. This "channel" is noise free and has no phase roll but is not perfectly flat in frequency response. It was a good test of our synchronization and equalization algorithms. Our receiver program is able to automatically synchronize itself and do the adaptive equalization needed for this channel. We found that using 5-ms block lengths gave the best results. For block lengths this short, any frame-rate noise sounds more like quantization noise. For larger block lengths, such as 16 ms, the frame-rate noise sounds like burble. Since some frame-rate noise always creeps in, we chose the 5-ms block length as subjectively better.

The next channel we tried was the telephone system at Murray Hill. Using two data sets connected to the D/A and A/D of the computer we were able to loop through the Murray Hill switch. Again the results were good. We then tried introducing a phase roll to the received signal and using our phase-roll compensation system in the same way as described in the previous section. As reported above, this also worked well.

Thus far we have not encountered a noisy channel and so we were forced to simulate one, as we did for the phase roll. The result confirmed our prediction that the noise would sound the same as white noise added to clear speech.

Another advantage we have enjoyed thus far is that the A/D and D/A have the same clock. Therefore, there is no relative drift between them. Such a drift for real hardware would cause a problem if it were at all large. In practice, if the system is regularly resynchronized, the receiver clock can be tied to the resynchronization to compensate for any drift. When we resynchronize the buffers, the chirp could also be used to resynchronize the receiver clock with the transmitter clock.

## VIII. CONCLUSIONS

We have proposed a novel time-and-frequency scrambling system. To avoid the echo cancellation problem we have resorted to one of

two solutions. For a full-duplex system we used a narrow bandwidth, which reduced the quality of the speech but permitted simultaneous talking by both speakers. The narrow bandwidth may have also caused some small loss in intelligibility. For a half-duplex system this is not the case. We used a wider bandwidth system. One can even conceive of an automatic system in which each user receives half bandwidth when both speak but when one is silent the other receives the full bandwidth.

What makes our novel system feasible are the flexibility of digital signal processing and the equalization and synchronization schemes that were devised using digital processing. The digital processing also permits a full time-and-frequency permutation.

## REFERENCES


1. "Digital Signal Processor—A Programmable Integrated Circuit for Signal Processing," B.S.T.J., *60*, No. 7, Part 2 (September 1981).
2. D. Esteban and C. Galand, "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes," Proc. of 1977 ICASSP (May 1977), pp. 191–5.
3. J. D. Johnston, "A Filter Family Designed For Use in Quadrature Mirror Filter Banks," Proc. of 1980 ICASSP (April 1980), pp. 291–4.
4. N. S. Jayant, R. V. Cox, B. J. McDermott, and A. M. Quinn, "Analog Scramblers for Speech based on Sequential Permutations in Time and Frequency," B.S.T.J., this issue.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# Maximum-Power and Amplitude-Equalizing Algorithms for Phase Control in Space Diversity Combining

By P. D. KARABINIS

(Manuscript received October 12, 1981)

*Space diversity receivers equipped with continuous combiners have, in recent years, found wide use in terrestrial microwave radio systems as a means of mitigating the effects of multipath fading. An analysis of two phase-control algorithms for space diversity combining, comparing the performance capabilities of a maximum-power and an amplitude-equalizing algorithm, is presented in this paper. The extended capabilities of the equalizing algorithm in mitigating linear and quadratic channel distortions are investigated through computer simulations. A reduction in digital radio outage time afforded by the equalizing algorithm (relative to that of the maximum-power algorithm) is approximated for over-water-paths where it is assumed that fading can be described in terms of a two-ray fading model. For the model chosen, we show that the additional outage reduction owing to the equalizing phase-control algorithm is highly dependent on the statistics of the multipath delay parameter, $\tau$, and can vary from unity to a factor of five for different probability-density functions of $\tau$.*

## I. INTRODUCTION AND SUMMARY

High-speed digital radio systems operating in the 4-, 6-, and 11-GHz common-carrier bands experience severe performance degradations during periods of multipath fading.[1-7] These degradations are mainly due to intersymbol interference resulting from dispersive channel characteristics (amplitude and group-delay distortions) that typically accompany multipath fading.

Space diversity reception with continuous combining of the received signals has proven very effective in mitigating the amplitude and group-delay distortions caused by multipath fading.[4,5] Traditionally, maximum combined-signal power has been used as the control crite-

rion for combining the received signals.[8] Although maximum-power combining does not react to inband dispersion directly, the method does provide a substantial statistical improvement in this respect. In general, however, the output signal of a maximum-power combiner during dispersive fading will contain, to some degree, linear and quadratic amplitude distortion, with the linear component usually being the predominant one.[4,8] Because of this, the use of an amplitude slope equalizer following a maximum-power combiner has proven very effective in reducing the multipath outage of high-speed digital radio signals to levels compatible with long-haul outage objectives.[4]

With the advent of higher-level digital-modulation formats, equalization requirements of microwave radio channels are becoming increasingly stringent.[9] To optimize the performance of space diversity combiners used in modern digital radio systems, alternate phase-control algorithms have recently been proposed that depart from the traditional maximum-power algorithm in that, instead of maximum combined-signal power, minimum combined-signal dispersion becomes the controlling criterion.[10,11] With these equalizing phase-control algorithms, space diversity combiners have been shown to provide reduced levels of signal dispersion during periods of multipath fading[7] (as compared with maximum-power combiners), thus further reducing the multipath-related outage of high-speed digital signals. One such equalizing algorithm is proposed, analyzed, and compared with a maximum-power algorithm in this paper.

In Section II, the conventional maximum-power combiner is reviewed first, with a particular implementation of the maximum-power phase-control algorithm described in detail. Then we show how the maximum-power algorithm can be converted into an equalizing algorithm by judiciously extracting a phase-control signal from selected portions of the combined-signal spectrum. The analysis reveals that the maximum-power algorithm is only a special case of a more general algorithm that can adaptively maximize the combined-signal power over any frequency interval in the channel. The particular equalizing phase-control algorithm described minimizes the power difference between two spectral samples derived symmetrically from the upper and lower portions of the combined-signal spectrum. To achieve this, the algorithm maximizes the power of the upper or lower spectral sample (as determined by a control criterion) and attempts to bring the two frequency subbands in power equilibrium. As a consequence, we show that the linear and quadratic amplitude distortions of the combined-signal spectrum are substantially reduced for many dispersive channel conditions.

Section III presents results of computer simulations illustrating the response of the maximum-power and equalizing combiners to disper-

sive channel conditions. An improvement factor of the equalizing combiner over the maximum-power combiner is calculated for several assumed probability-density functions of the parameters of a two-ray fading model. Since a two-ray fading model is more likely to represent fading accurately for microwave radio paths over water (where a strong reflective component is typically present during fading) our results too should be viewed in that context. For microwave radio paths over land, it has been shown by Rummler[12] that a three-ray model is required to accurately describe multipath fading. For the two-ray model chosen here, we show that the equalizing combiner improvement factor is strongly dependent on the assumed probability-density function of the multipath delay parameter $\tau$ and varies from approximately unity (no improvement) to a factor of five, for different probability-density functions of $\tau$ and expected values of $\tau$.

## II. MAXIMUM-POWER AND AMPLITUDE-EQUALIZING PHASE-CONTROL ALGORITHMS

A simplified block diagram of a space diversity receiver providing two signal inputs to a continuous combiner is shown in Fig. 1. Following a fixed delay adjustment of the diversity-signal path (to equalize the

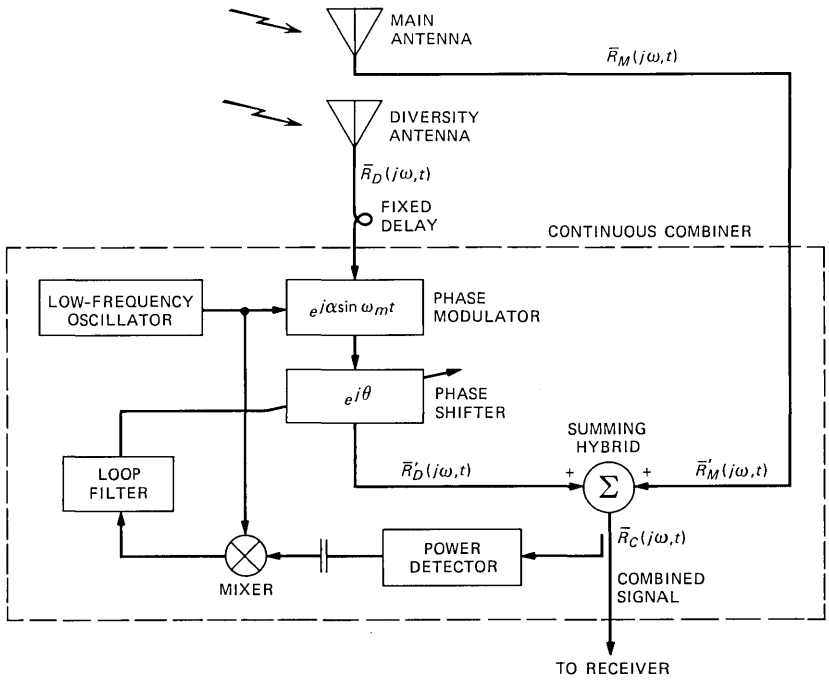



Fig. 1—Block diagram of a space diversity continuous combiner.

electrical path lengths leading to the combiner inputs), the phase of the diversity signal undergoes a dynamic adjustment so as to always maintain the proper relationship with respect to the main-antenna signal phase at the summing hybrid. For the purpose of generating a phase-control signal, the phase of the diversity antenna signal is perturbed sinusoidally, resulting in a periodic modulation of the combined-signal power.[13] The fundamental component of the combined-signal power modulation is detected and used in a feedback arrangement to control the phase-shifter value.[14] Depending on the phase-control algorithm used, the phase correction can be chosen to maximize the average combined-signal power or minimize the dispersion of the combined signal.

In the appendix, the combined-signal power modulation resulting from the sinusoidal phase modulation of the diversity antenna signal is calculated in terms of the spectral densities and phase misalignment of the antenna signals at the inputs of the summing hybrid. In Section 2.1 below we show how the fundamental component of the combined-signal power modulation can be used as the control signal in a simple feedback loop to form a maximum-power combining algorithm. Section 2.2 describes a method of converting the maximum-power combiner of Section 2.1 into an equalizing combiner by selectively extracting the combined-signal power modulation from two frequency bands symmetrically located to the left and right of the channel center frequency.

### *2.1 Maximum-power phase-control algorithm*

The average combined-signal power over the entire channel can be expressed, using eqs. (44) through (46) of the appendix, as:

$$p_0(t, \theta) = \frac{1}{2\pi}\int_{\omega_1}^{\omega_2} \delta^2(\omega, t) V^2(\omega, t) d\omega + \frac{1}{2\pi}\int_{\omega_1}^{\omega_2} V^2(\omega, t) d\omega + \frac{1}{\pi} J_0(\alpha) \int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \cos\{\psi(\omega, t) + \theta\} d\omega, \quad (1)$$

where the frequency interval $\omega_1 \leq \omega \leq \omega_2$ is assumed to span the channel bandwidth, $\delta^2(\omega, t)V^2(\omega, t)$ and $V^2(\omega, t)$ denote the diversity and main antenna power-spectral densities, respectively, $\psi(\omega, t)$ is the phase of the diversity antenna spectrum relative to the phase of the main antenna spectrum before any phase correction, $\theta$ represents the frequency-independent phase correction introduced in the diversity antenna spectrum by the phase shifter (see Fig. 1), and $J_0(\alpha)$ denotes a Bessel function of the first kind whose argument is the magnitude of the sinusoidal phase modulation introduced in the diversity signal. The explicit dependence of the variables appearing in eq. (1) on time

accounts for slow variations of these parameters caused by changes in the transmission medium. However, these variations are assumed quasi-static with respect to the transmission rate of the microwave radio channel.

To find $\theta_m$, the value of $\theta$ that maximizes $p_0(t, \theta)$, we set the first partial derivative of $p_0(t, \theta)$ to zero and require the second partial derivative to be negative. That is,

$$\left.\frac{\partial p_0(t, \theta)}{\partial \theta}\right|_{\theta=\theta_m} = -\frac{1}{\pi} J_0(\alpha) \int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \sin\{\psi(\omega, t) + \theta_m\} d\omega = 0 \quad (2)$$

and

$$\left.\frac{\partial^2 p_0(t, \theta)}{\partial \theta^2}\right|_{\theta=\theta_m} = -\frac{1}{\sqrt{2}\pi} J_0(\alpha) \int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \cos\{\psi(\omega, t) + \theta_m\} d\omega < 0. \quad (3)$$

Having established the conditions for maximum-power combining, we will now show that the phase-control loop of Fig. 2 finds a value of $\theta$ that satisfies these conditions.

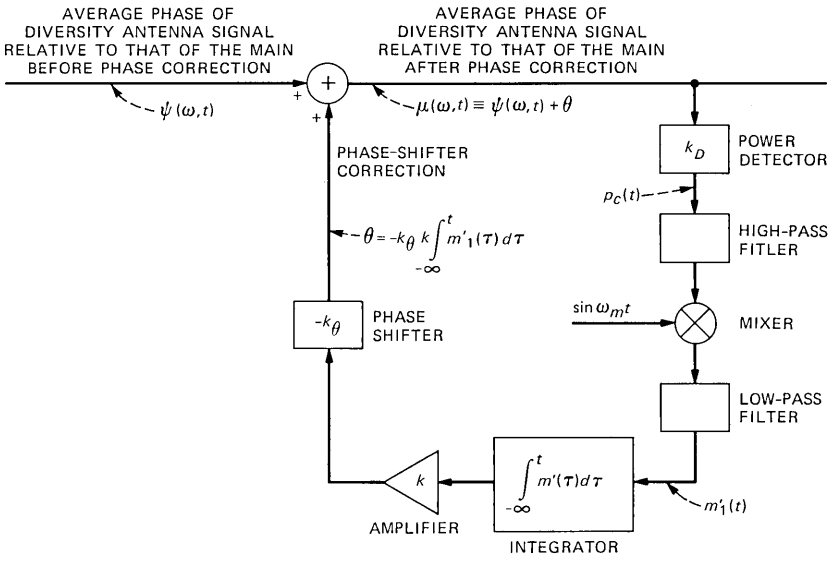



Fig. 2—Block diagram of maximum-power combiner phase-control loop.

From the assumption of a slowly varying transmission medium, the fundamental component of the combined-signal power modulation, $m_1'(t)$, [derived in the appendix, eq. (55)] will be a narrowband process. Choosing the corner frequencies of the high-pass and low-pass filters of Fig. 2 well below and above the highest spectral component of $m_1'(t)$, respectively, the function $m_1'(t)$ will be unaffected by these filters.* Therefore, from an examination of the phase-control loop of Fig. 2 we can write:

$$\theta = -k_\theta k \int_{-\infty}^{t} m_1'(\tau)d\tau, \tag{4}$$

where $k$ and $-k_\theta$ denote component gains as shown in Fig. 2. The rate of change of $\theta$ with respect to time can be expressed from eq. (4) and eqs. (55) and (49) of the appendix as

$$\frac{d\theta}{dt} = -k_\theta k m_1'(t)$$

$$= -4J_1(\alpha)k_\theta k \frac{\int_{\omega_1}^{\omega_2} \delta(\omega, t)V^2(\omega, t)\sin\{\psi(\omega, t) + \theta\}d\omega}{\int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t)d\omega}. \tag{5}$$

Observe from eq. (5) that a nonchanging value of $\theta$ can exist if and only if

$$m_1'(t) = \frac{4J_1(\alpha)\int_{\omega_1}^{\omega_2} \delta(\omega, t)V^2(\omega, t)\sin\{\psi(\omega, t) + \theta\}d\omega}{\int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t)d\omega} = 0, \tag{6}$$

which implies that

$$\int_{\omega_1}^{\omega_2} \delta(\omega, t)V^2(\omega, t)\sin\{\psi(\omega, t) + \theta\}d\omega = 0. \tag{7}$$

Therefore, if $m_1'(t) \neq 0$ the phase shifter correction, $\theta$, will continue to change until a value satisfying eq. (7) is found. In satisfying eq. (7), eq. (2) and, therefore, the first condition for maximum-power combining is simultaneously satisfied. It remains to be shown that this value of

* It is also assumed that the modulation frequency $\omega_m$ is large compared with the bandwidth of $m_1'(t)$. For example, if the bandwidth of $m_1'(t)$ is limited to 1 Hz, then we choose $f_m = (\omega_m/2\pi) \geq 10$ Hz.

$\theta$ also satisfies the second condition for maximum-power combining as expressed by inequality (3).

Assume that a value of $\theta$, $\theta = \theta_0$, satisfying eq. (7) is found by the phase-control loop. That is, assume an initial condition of $\theta = \theta_0 \pm \Delta\theta$, $\Delta\theta > 0$, and a final condition of $\theta = \theta_0$. With $\theta = \theta_0 \pm \Delta\theta$ we have

$$\int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \sin\{\psi(\omega, t) + \theta_0 \pm \Delta\theta\} d\omega \neq 0. \tag{8}$$

Defining,

$$\begin{aligned} \xi &\equiv \int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \sin\{\psi(\omega, t) + \theta_0 \pm \Delta\theta\} d\omega \\ &= \frac{\int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t) d\omega}{4J_1(\alpha)} m_1'(t), \end{aligned} \tag{9}$$

$\xi$ may be expressed as

$$\begin{aligned} \xi = \cos(\pm\Delta\theta) &\int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \sin\{\psi(\omega, t) + \theta_0\} d\omega \\ + \sin(\pm\Delta\theta) &\int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \cos\{\psi(\omega, t) + \theta_0\} d\omega. \end{aligned} \tag{10}$$

Since by assumption, the first term of eq. (10) is zero we are left with

$$\begin{aligned} \xi &= \sin(\pm\Delta\theta) \int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \cos\{\psi(\omega, t) + \theta_0\} d\omega \\ &= \frac{\int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t) d\omega}{4J_1(\alpha)} m_1'(t). \end{aligned} \tag{11}$$

If the assumed $\theta$ increment is taken positive ($+\Delta\theta$) the integral of eq. (11), $\int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \cos\{\psi(\omega, t) + \theta_0\} d\omega$, must be positive to make $m_1'(t)$ positive. A positive $m_1'(t)$ will generate a negative $d\theta/dt$ [see eq. (5)] as is required to force $\theta = \theta_0 + \Delta\theta$ to $\theta = \theta_0$, as per assumption. Otherwise, $\theta = \theta_0 + \Delta\theta$ will migrate away from $\theta = \theta_0$, contradicting the initial assumption. Similarly, if we assume a negative $\theta$ increment ($-\Delta\theta$) the integral $\int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \cos\{\psi(\omega, t) + \theta_0\} d\omega$ must again be positive in order for a positive $d\theta/dt$ to be generated as is required to force $\theta = \theta_0 - \Delta\theta$ to $\theta = \theta_0$.

Having shown that $d\theta/dt = 0$ if and only if $m_1'(t) = 0$ (which requires

that $\int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \sin\{\psi(\omega, t) + \theta_0\} d\omega = 0$), in conjunction with $d\theta/dt = 0$ if and only if $\int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \cos\{\psi(\omega, t) + \theta_0\} d\omega > 0$, we have shown that the phase-control loop of Fig. 2 satisfies both conditions for maximum-power combining.

### 2.2 Amplitude-equalizing phase-control algorithm

In the previous section we showed that the phase-control loop of Fig. 2 finds a phase shifter value that forces the fundamental component of the combined-signal power modulation, $m_1'(t)$, to zero over the frequency interval $\omega_1 \leq \omega \leq \omega_2$, and in doing so, also maximizes the average combined-signal power over the same frequency interval. Therefore, when $\omega_1 \leq \omega \leq \omega_2$ spans the entire channel bandwidth, as is the case in Fig. 2 (where the combined-signal spectrum is allowed to be seen in its entirety by the power detector), the combined-signal power will be maximized over the entire channel. However, if a narrowband filter is inserted before the power detector of Fig. 2, allowing only a fraction of the combined-signal spectrum to be seen by the power detector, the combined-signal power will be maximized over the frequency band seen by the power detector and not necessarily over the entire channel bandwidth, since the phase difference between the received spectra will, in general, be frequency dependent during periods of multipath fading.

In this section we will examine a more general version of the phase-control loop of Fig. 2, where the signal at the power-detector input is a filtered version of the combined-signal spectrum. Specifically, we will show that an equalizing phase-control algorithm, which minimizes the amplitude slope across the combined-signal spectrum, can be derived from the phase-control loop of Fig. 2 by requiring that the power detector input signal be a weighted sum of spectral components derived symmetrically from the left and right portions of the combined-signal spectrum.

Consider Fig. 3 where two narrowband portions of the combined-signal spectrum, each occupying a bandwidth $2\,\Delta\omega$ and centered at $\omega_l$ and $\omega_r$, respectively, are extracted by the two narrowband filters. The center frequencies of the narrowband filters, $\omega_l$ and $\omega_r$, are equally spaced from the center of the channel and are located to the left and right of the channel center-frequency, respectively. From the output signal of each narrowband filter, a power detector generates a voltage proportional to the combined-signal power found within the corresponding frequency interval, $\omega_l - \Delta\omega \leq \omega \leq \omega_l + \Delta\omega$ and $\omega_r - \Delta\omega \leq \omega \leq \omega_r + \Delta\omega$. Denoting the power detector output voltages by $p_l(t)$ and $p_r(t)$, corresponding to frequency intervals $\omega_l - \Delta\omega \leq \omega \leq \omega_l + \Delta\omega$ and $\omega_r - \Delta\omega \leq \omega \leq \omega_r + \Delta\omega$, respectively, the development in the appendix can be used to express $p_l(t)$ and $p_r(t)$ as

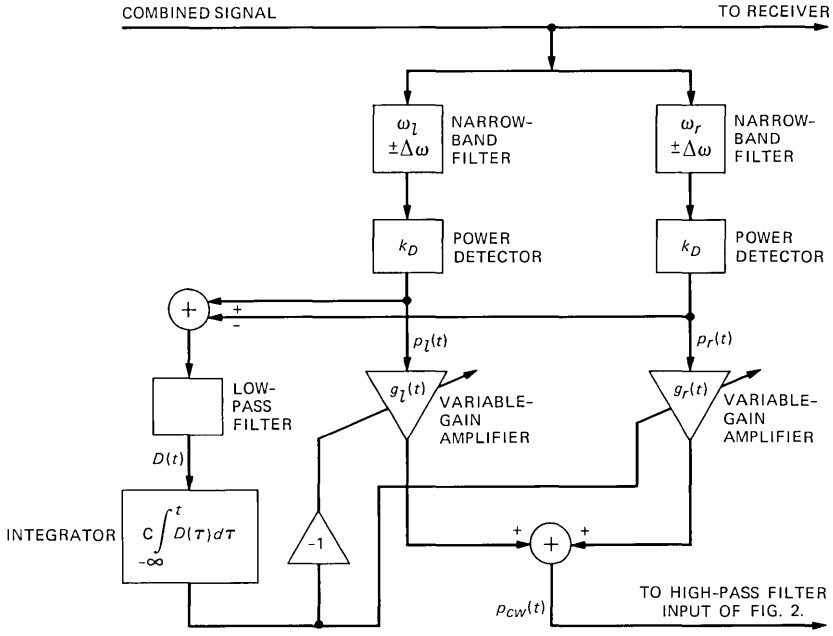



Fig. 3—Power detector portion of equalizing-combiner phase-control loop.

$$p_l(t) = k_D p_{0l}(t)[1 - m'_{1l}(t)\sin \omega_m t + m'_{2l}(t)\cos 2\omega_m t - \cdots] \quad (12)$$

and

$$p_r(t) = k_D p_{0r}(t)[1 - m'_{1r}(t)\sin \omega_m t + m'_{2r}(t)\cos 2\omega_m t - \cdots], \quad (13)$$

where $k_D$ is a proportionality constant associated with each power detector (two identical power detectors are assumed) and with all other variables, $p_{0l}(t)$, $p_{0r}(t)$, $m'_{1l}(t)$, $m'_{1r}(t)$, $\cdots$, conforming to eqs. (52) through (54) of the appendix.

The power detector outputs, $p_l(t)$ and $p_r(t)$, are weighted by variable-gain amplifiers and then summed. Thus, a voltage, $p_{cw}(t)$, representing a weighted combined-signal power over the channel is derived (see Fig. 3), and can be expressed as

$$p_{cw}(t) = g_l(t)p_l(t) + g_r(t)p_r(t). \quad (14)$$

The weighting gains, $g_l(t)$ and $g_r(t)$, are determined by the integral of a difference function, $D(t)$, which is proportional to the average power difference between the frequency bands $\omega_l - \Delta\omega \leq \omega \leq \omega_l + \Delta\omega$ and $\omega_r - \Delta\omega_r \leq \omega \leq \omega_r + \Delta\omega$. That is, we let

$$D(t) \equiv k_D[p_{0l}(t) - p_{0r}(t)], \quad (15)$$

$$g_l(t) \equiv -k_g C \int_{-\infty}^{t} D(\tau)d\tau, \tag{16}$$

and

$$g_r(t) \equiv k_g C \int_{-\infty}^{t} D(\tau)d\tau, \tag{17}$$

where $k_g$ is a proportionality constant associated with the variable-gain amplifiers and C is an integration constant. In addition, we require that the gains $g_l(t)$ and $g_r(t)$ be bounded

$$\left.\begin{aligned} 0 \le g_l(t) \le g_{\max} \\ 0 \le g_r(t) \le g_{\max} \end{aligned}\right\} \text{for all } t. \tag{18}$$

It can now be seen from eqs. (15) through (17) that if a positive amplitude slope exists across the combined-signal spectrum, $p_{0r}(t) > p_{0l}(t)$, $D(t)$ will be negative causing $g_l(t)$ to increase [or remain fixed at $g_{\max}$ if, initially, $g_l(t) = g_{\max}$] and $g_r(t)$ to decrease [or remain fixed at zero if, initially, $g_r(t) = 0$]. That is,

$$\begin{aligned} &\text{if } p_{0r}(t) > p_{0l}(t) \\ &\text{(positive amplitude slope)} \end{aligned} \qquad \text{then:} \quad \begin{cases} \dfrac{dg_l(t)}{dt} \ge 0 \\ \dfrac{dg_r(t)}{dt} \le 0. \end{cases} \tag{19}$$

Similarly, if a negative slope exists across the combined-signal spectrum, $p_{0r}(t) < p_{0l}(t)$, $D(t)$ will be positive causing $g_l(t)$ to decrease (or remain fixed at zero) and $g_r(t)$ to increase (or remain fixed at $g_{\max}$). Therefore:

$$\begin{aligned} &\text{if } p_{0r}(t) < p_{0l}(t) \\ &\text{(negative amplitude slope)} \end{aligned} \qquad \text{then:} \quad \begin{cases} \dfrac{dg_l(t)}{dt} \le 0 \\ \dfrac{dg_r(t)}{dt} \ge 0. \end{cases} \tag{20}$$

As the integration constant C [appearing in eqs. (16) and (17)] increases, the response of $p_{cw}(t)$ to a changing amplitude slope becomes faster. In the limit, as C approaches infinity, $p_{cw}(t)$ can assume only two values: either $g_{\max} p_l(t)$ or $g_{\max} p_r(t)$ corresponding to $D(t) < 0$ and $D(t) > 0$, respectively. That is, $p_{cw}(t)$ will instantaneously reflect the power-detector voltage corresponding to the weaker side of the combined-signal spectrum. Thus, with C very large, substituting the block diagram of Fig. 3 in place of the power detector of Fig. 2 results in a

phase-control loop that will always maximize the power of the weaker portion of the combined-signal spectrum and, therefore, minimize the amplitude slope across the combined-signal spectrum.

In the absence of dispersive channel conditions, the amplitude slope of the combined-signal spectrum will be zero. During such periods, random noise effects will dictate the value of $D(t)$. This, however, is of no consequence since in the absence of channel dispersion the phase difference between the received spectral densities is frequency independent and, therefore, maximizing the combined-signal power over any frequency interval within the channel necessarily maximizes the combined-signal power over the entire channel. Consequently, during dispersion-free periods, the equalizing phase-control algorithm described above becomes indistinguishable from the maximum-power phase-control algorithm.

In the following section, results of computer simulations, illustrating the relative performance capabilities of the maximum-power and amplitude-equalizing phase-control algorithms during dispersive channel conditions, are presented.

## III. RELATIVE PERFORMANCE OF MAXIMUM-POWER AND AMPLITUDE-EQUALIZING PHASE-CONTROL ALGORITHMS

The algorithms described above have been stressed by computer simulations using a two-ray fading model. The fade parameters of the model, relative amplitude and delay, are treated as independent random variables and are assigned physical significance. That is, the model adopted here differs from that of Rummler's[12] where the delay parameter $\tau$ is assigned mathematical significance only and treated as a constant. Following the approach of Jakes[3] and Greenstein and Prabhu[15] we assume an exponential form for the probability-density function of the delay parameter $\tau$. For comparison purposes we also examine other forms of the probability-density function of $\tau$, including a Gaussian probability-density function.[16] The probability-density function of the relative amplitude of the delayed wavefront is assumed constant for deep fades.[3,15,16]

To facilitate the simulation of the space diversity arrangement of Fig. 1, we adopt a ray-optical view of wave propagation,[17-19] and assume a circular trajectory for the delayed wavefront. These assumptions enable one to calculate the fade on the diversity antenna given the fade on the main antenna. However, for a given fade-notch location on the main antenna, the fade-notch location on the diversity antenna will in practice be distributed over a wide frequency range since there are many physical $\tau$'s that can generate a particular fade-notch location on the main antenna. Each one of these $\tau$'s will generate a fade notch on the diversity antenna at a different frequency (because of the

geometry of the space diversity arrangement), thus giving the appearance of independent fading between the two antennas.

The receiving antennas are assumed identical and spaced vertically by 30 feet. The channel bandwidth is set at 30 MHz, centered at an RF frequency of 6 GHz, and propagation is assumed over a hop length of 26.4 miles. Finally, all fades seen by the two antennas are assumed, for simplicity, to be of the minimum-phase type.

The space diversity channel model resulting from the above assumptions provides a tractable means of generating a set of stressing simulated fade conditions. The results generated by these simulations are to be viewed as first-order approximations only, until such time as other analytical methods are developed allowing for alternate, perhaps more accurate, channel descriptions.

### 3.1 Amplitude responses of maximum-power and amplitude-equalizing combiners during dispersive fading

The transfer function of the multipath medium as seen by the main and diversity antennas is expressed as:

$$\bar{H}_M(j\omega, t) = 1 + b(t)e^{-j\omega\tau(t)} \tag{21}$$

and

$$\bar{H}_D(j\omega, t) = 1 + b(t)e^{-j\omega\{\tau(t)+\tau'(t)\}}, \tag{22}$$

where the subscripts $M$ and $D$ refer to the main and diversity antennas, respectively. Assuming a circular trajectory for the delayed wavefront, $\tau'(t)$ is approximated by:

$$\tau'(t) \approx h\left\{\frac{6\tau(t)}{cd}\right\}^{1/2}, \tag{23}$$

where $h$, $d$, and $c$ denote antenna separation, hop length, and speed-of-light in air, respectively. Having calculated the state of the channel from eqs. (21) through (23), for given $b(t)$ and $\tau(t)$ values, the computer program then computes $V(\omega, t)$, $\delta(\omega, t)$, and $\psi(\omega, t)$, as defined by eqs. (37) through (39) of the appendix. The response of the phase-control loop of Fig. 2 is then evaluated in both the maximum-power and equalizing phase-control modes. In the equalizing mode, the narrowband filters (see Fig. 3) are assumed ideal, with $\omega_l$, $\omega_r$, and $\Delta\omega$ set at $\omega_c - B/4$, $\omega_c + B/4$, and $B/4$, respectively, with $B/2\pi$ denoting the channel bandwidth (in Hz) and $\omega_c$ the channel center frequency. With these filter settings, the equalizing-combiner algorithm performs maximum-power combining over the entire upper or lower half of the channel.

The computer simulations were repeated with other combinations of $\Delta\omega$, $\omega_l$, and $\omega_r$ to parametrically investigate the effect of these filter

settings on the equalizing-combiner response. For some particular dispersive channel conditions, the equalizing-combiner response varied for different $\Delta\omega$, $\omega_l$, and $\omega_r$ values. On an overall basis, however, the improvement afforded by the equalizing combiner was found insensitive to these filter parameters as long as $\omega_l$ and $\omega_r$ were chosen reasonably close to the channel edges.

For each simulated channel condition, two phase shifter corrections, $\theta_M(t)$ and $\theta_E(t)$, corresponding to the maximum-power and equalizing phase-control loops, respectively, were calculated. Then, effective transfer functions were formed, relating the maximum-power and equalizing-combiner output signals to the transmitted signal. The effective transfer functions generated by the maximum-power and equalizing combiners are expressed by

$$\bar{E}_M(j\omega, t) \equiv \frac{1}{\sqrt{2}} [\bar{H}_M(j\omega, t) + \bar{H}_D(j\omega, t)e^{j\theta_M(t)}] \tag{24}$$

and

$$\bar{E}_E(j\omega, t) \equiv \frac{1}{\sqrt{2}} [\bar{H}_M(j\omega, t) + \bar{H}_D(j\omega, t)e^{j\theta_E(t)}], \tag{25}$$

respectively, where the factor $1/\sqrt{2}$ is required to satisfy power conservation constraints at the output of the summing hybrid of Fig. 1.

Representative combiner responses, $\bar{E}_M(j\omega)$ and $\bar{E}_E(j\omega)$, are shown in Figs. 4 and 5. In Fig. 4, both transfer functions, $\bar{H}_M(j\omega)$ and $\bar{H}_D(j\omega)$, as seen by the main and diversity antennas respectively, contain a 40-dB fade notch within the channel bandwidth. The response of the maximum-power combiner, as shown, contains a substantial amount of linear and quadratic amplitude distortion across the channel. Its group-delay response (not shown) is also severely distorted. Unlike the maximum-power combiner, the equalizing-combiner amplitude response is flat (to within 1 dB) across the entire channel. The group-delay response of the equalizing combiner (not shown) is also flat across the channel.

Figure 5 shows a channel condition where only one antenna sees a transfer function with an inband notch. As with the situation of Fig. 4, here too, the maximum-power combiner suffers from linear and quadratic amplitude distortions. Note again the almost ideally flat response of the equalizing combiner. Figures 4 and 5 are typical of most simulated channel conditions where one or both antennas experience fade notches well within the channel bandwidth. There are channel conditions, however, where both antenna signals are affected by fade notches that fall outside the channel bandwidth, both below or above the channel-center frequency. These channel conditions produce a predominantly linear amplitude distortion in the channel, accompa-

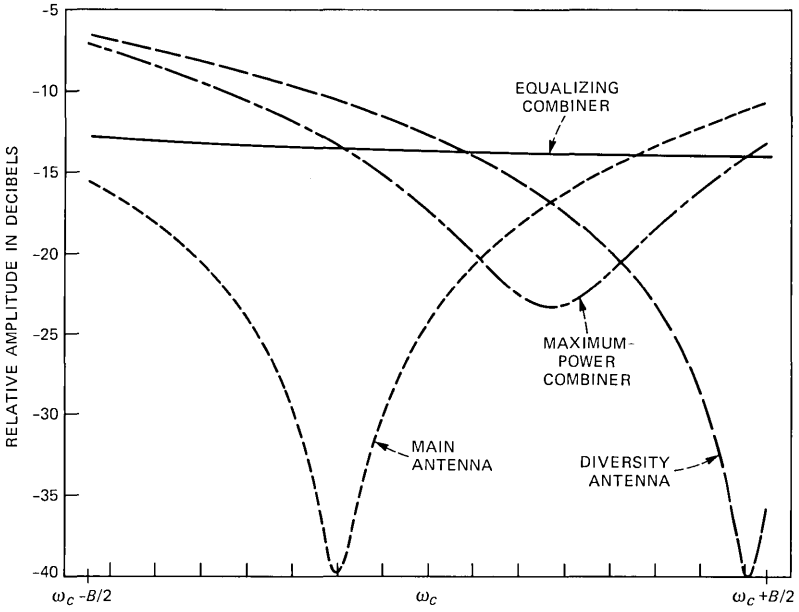



Fig. 4—Combiner response to a dispersive channel condition generated by $b = 0.99$ and $\tau = 2.4183$ ns. The channel bandwidth = $B/2\pi = 30$ MHz.

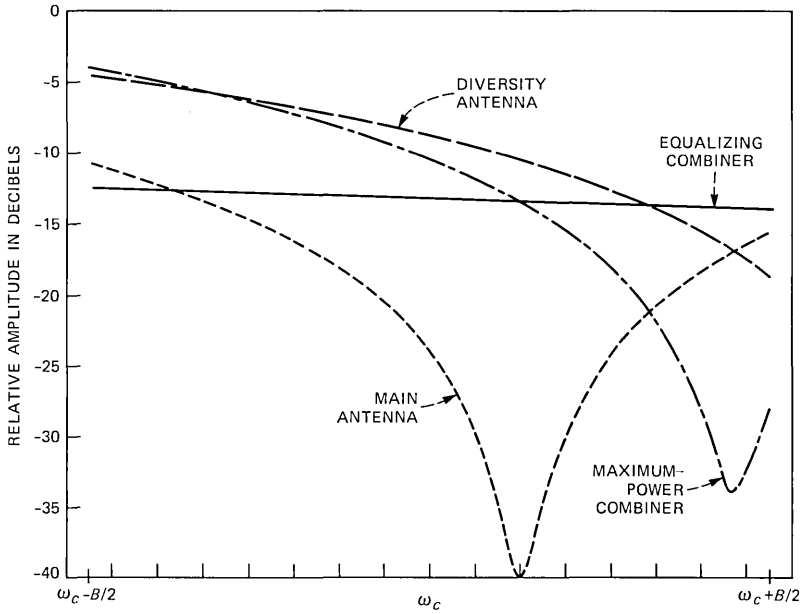



Fig. 5—Combiner response to a dispersive channel condition generated by $b = 0.99$ and $\tau = 2.4151$ ns. The channel bandwidth = $B/2\pi = 30$ MHz.

nied by negligible phase dispersion, since the phase dispersion produced by a fade is concentrated in the vicinity of the fade notch. In response to such a channel condition, the equalizing-combiner output signal can be only marginally better than that of the maximum-power combiner since the inband phase dispersion, required by the equalizing combiner to achieve amplitude equalization, does not exist. Unfortunately, the channel condition described above has a higher probability of occurrence than those depicted in Figs. 4 and 5 (where at least one antenna signal contains a fade notch within the channel) since the channel bandwidth is typically a small fraction of the period $1/\tau$, or $1/(\tau + \tau')$, with which the fade notches repeat. As a result, the reduction in outage time afforded by the equalizing combiner, relative to that of the maximum-power combiner, is limited to relatively small values, as we will see in the following section.

### *3.2 Relative improvement factor of amplitude-equalizing combiner*

In this section, a method of calculating regions on the $\lambda - \tau$ plane ($\lambda \equiv 1 - b$, denoting the fade depth at the fade notch) corresponding to system outage is described. Several outage regions corresponding to the maximum-power and equalizing combiners are calculated. Integration over these regions (having assumed some joint probability-density function for the variables $\lambda$, $\tau$) allows us to compare the performance of the two combiner algorithms and associate a relative improvement factor with the equalizing-combiner algorithm. The relative-improvement factor of the equalizing combiner is shown to be highly dependent on the assumed probability-density function associated with the multipath delay $\tau$, and on the expected value of the multipath delay, $\tau_0$. The relative improvement factor of the equalizing combiner is calculated and plotted as a function of $\tau_0$ for two assumed probability-density functions of $\tau$.

The computer simulation described in the previous section was extended to statistically evaluate the relative performance of the maximum-power and amplitude-equalizing combiners with respect to dispersive channel conditions of at least 6-dB peak-to-peak amplitude distortion in the channel. The threshold value of 6-dB peak-to-peak amplitude distortion was chosen for simplicity, since the accumulated time at this, or greater dispersive levels, has been found strongly correlated with the total outage time of high-speed digital signals.[2,8]

Starting from a value of $\tau$ that produces a fade notch at the center of the main antenna spectrum, as determined by

$$\tau = \frac{2n + 1}{2f_c}, \; f_c = \frac{\omega_c}{2\pi}, \; n = 0, 1, 2, \cdots,$$

values of $\lambda$ satisfying the outage condition at the maximum-power

and/or equalizing-combiner outputs are determined. Then, by perturbing each initial value of $\tau$, the fade notch is moved by a small amount (2-MHz increment) and $b$ is varied again until all $\lambda$ values satisfying the outage condition at the new setting are found. Perturbations in $\tau$ continue until a frequency band of $\pm 100$ MHz around the band center has been examined. Following this procedure, the first five outage regions, corresponding to $n = 13$ through 17, were evaluated for the maximum-power and amplitude-equalizing combiners. No outage regions were found for $n = 0$ through 12 and $n = 18$ through 25. The effect of outage regions corresponding to values of $n$ greater than 25 were neglected in this study. The first five outage regions that were evaluated are shown in Figs. 6 through 10. On each figure, the value of $n$ and the value of $\tau$ generating the fade notch at the center of the

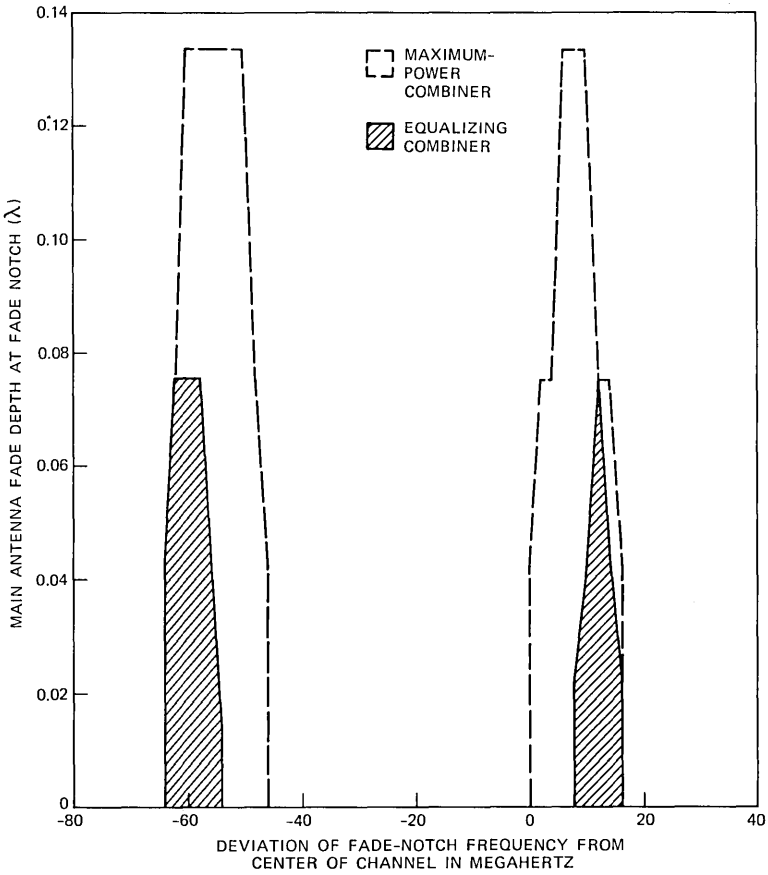



Fig. 6—Maximum-power and equalizing-combiner outage regions for $n = 13$ and $\tau = 2.25$ ns.

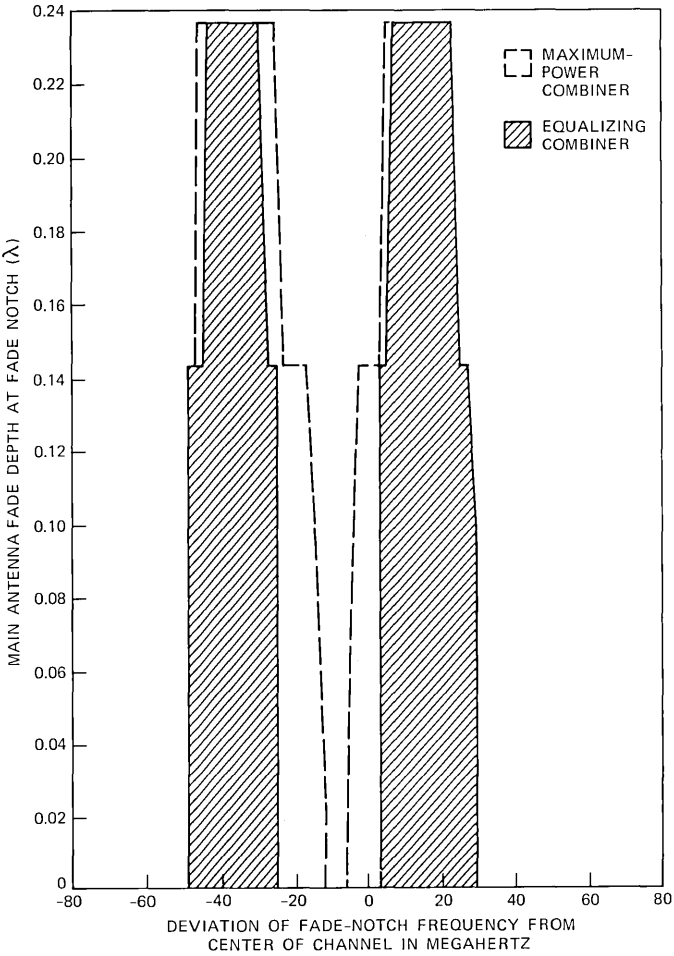



Fig. 7—Maximum-power and equalizing-combiner outage regions for $n = 14$ and $\tau = 2.4167$ ns.

main antenna spectrum are shown. The curves show fade-notch depth as a function of fade-notch location for the main antenna, which results in system outage as defined by the 6-dB amplitude distortion on the combiner outputs. Note that in every case the equalizing-combiner outage region (the shaded area) is totally contained within the maximum-power combiner outage region, which is bounded by the dashed curve. From this simple observation we conclude that the equalizing combiner is outperforming the maximum-power combiner in maintaining reduced levels of channel distortion. The performance improvements, however, are not large. As we can see in Figs. 7, 8, and 9, which contain the larger maximum-power combiner outage regions, the

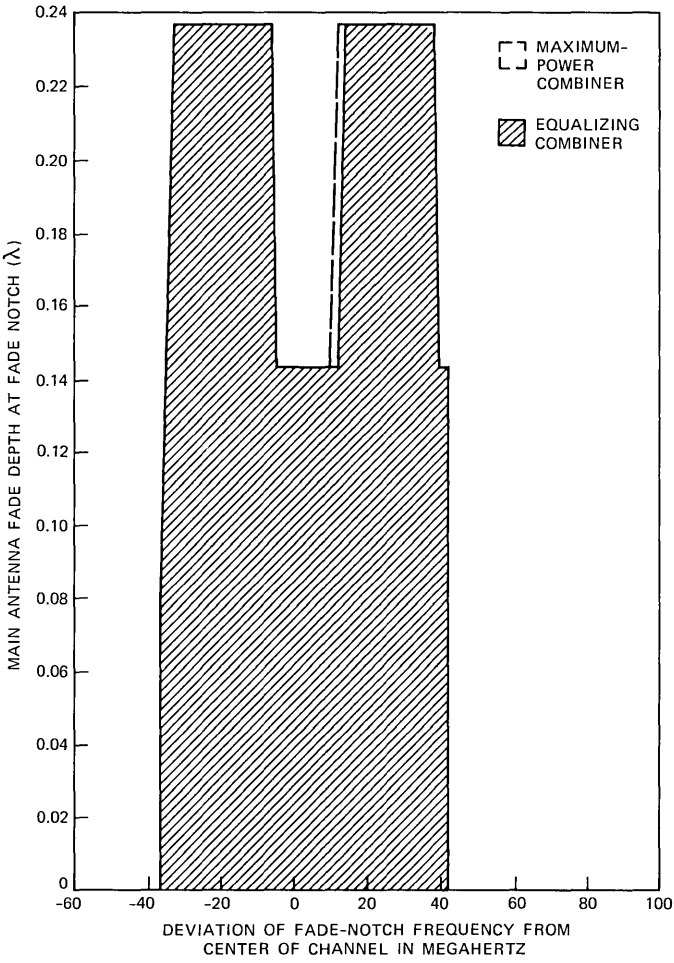



Fig. 8—Maximum-power and equalizing-combiner outage regions for $n = 15$ and $\tau = 2.5833$ ns.

equalizing-combiner outage regions are comparable to those of the maximum-power combiner. Specifically, Fig. 8 shows almost identical outage regions for the two combiners. Close examination of the channel conditions responsible for the outage regions of Fig. 8 reveals very closely correlated fading between the two antenna spectra (i.e., almost simultaneous frequency selective fading). On all other figures, the equalizing combiner outage regions are due to channel conditions that contain little or no inband phase distortion (i.e., fade notches very close to, or completely outside, the channel limits).

To calculate an outage probability for the maximum-power and

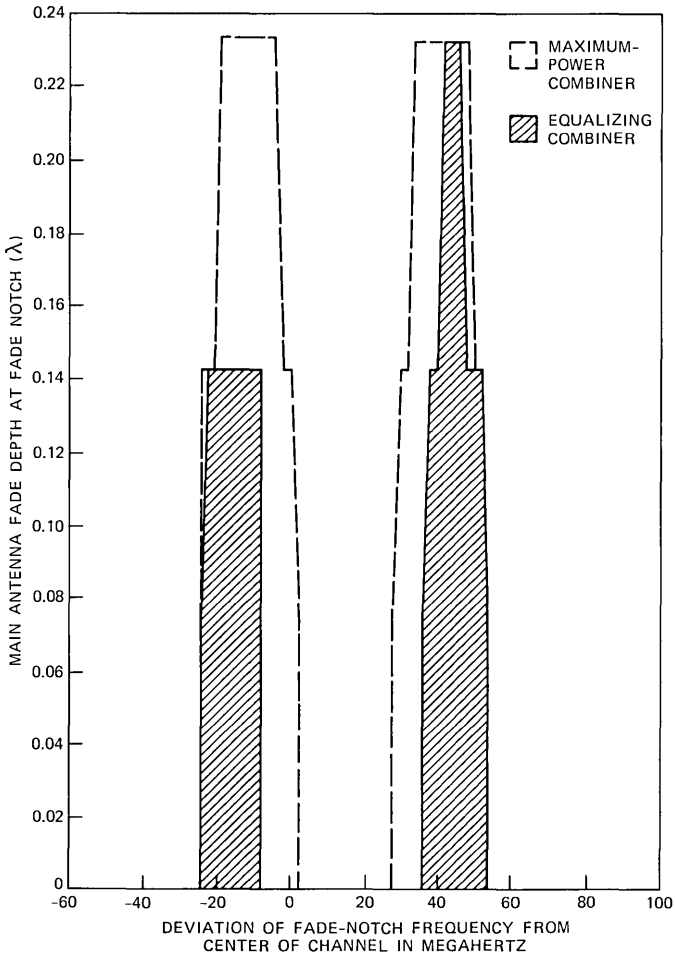



Fig. 9—Maximum-power and equalizing-combiner outage regions for $n = 16$ and $\tau = 2.75$ ns.

equalizing combiners, their corresponding outage regions must be weighted by the joint probability-density function $p_{\tau,\lambda}(\tau, \lambda)$, followed by integration over the appropriate limits and summation of all weighted contributions from all outage regions. Accordingly, the probability of outage of the maximum-power combiner can be expressed as:

$$P_M = \sum_{n=13}^{17} \int_{\tau_M^-(n)}^{\tau_M^+(n)} \int_0^{\lambda_M(n,\tau)} p_{\tau,\lambda}(\tau, \lambda) d\lambda d\tau, \tag{26}$$

where the subscript $M$ stands for maximum-power combiner, the limits

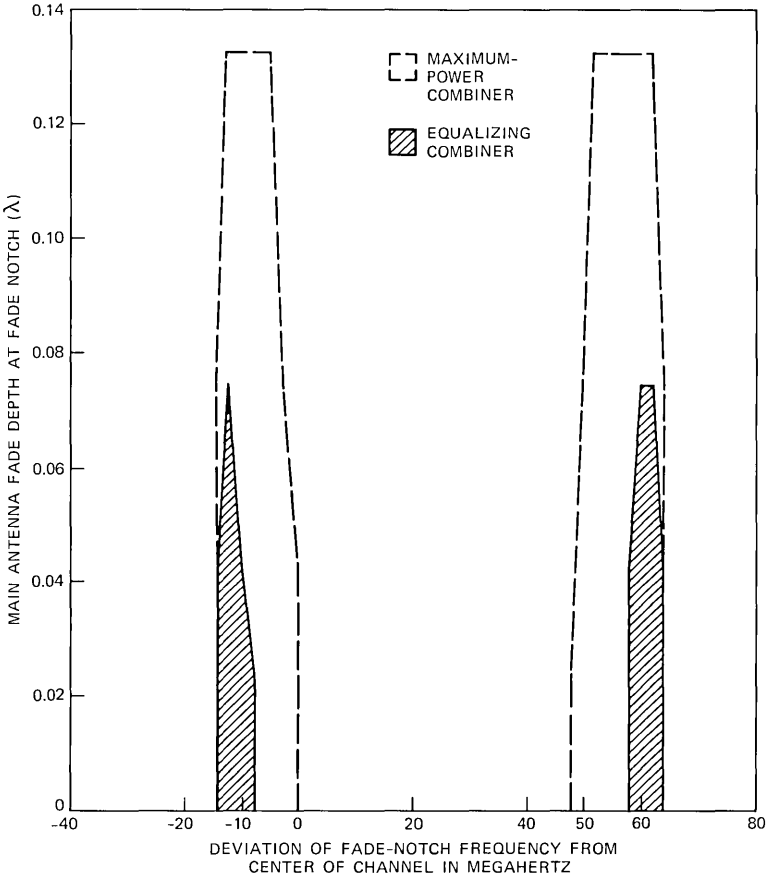



Fig. 10—Maximum-power and equalizing-combiner outage regions for $n = 17$ and $\tau = 2.9167$ ns.

$\tau_M^-(n)$ and $\tau_M^+(n)$ define the range of variation of $\tau$ (for a particular value of $n$) inside which the outage region lies, and $\lambda_M(n, \tau)$ denotes the minimum fade-depth resulting in outage as a function of $\tau$, for a specific value of $n$. Similarly, the probability of outage of the equalizing combiner can be evaluated from:

$$P_E = \sum_{n=13}^{17} \int_{\tau_E^-(n)}^{\tau_E^+(n)} \int_0^{\lambda_E(n,\tau)} p_{\tau,\lambda}(\tau, \lambda) d\lambda d\tau. \tag{27}$$

Using eqs. (26) and (27) we define an equalizing-combiner improvement factor, $I$, over the maximum-power combiner,

$$I \equiv \frac{P_M}{P_E}. \tag{28}$$

To simplify the calculations in evaluating eq. (28) we assume statistical independence between the fade depth $\lambda$ and delay $\tau$, allowing the joint probability-density function to be written as

$$p_{\tau,\lambda}(\tau, \lambda) = p_\tau(\tau) p_\lambda(\lambda). \tag{29}$$

Two probability-density functions of $\tau$ are considered.[3,15,16] The exponentially distributed

$$p_\tau^{(e)}(\tau) = \frac{1}{\tau_0} e^{-\frac{\tau}{\tau_0}}, \qquad \tau \geq 0, \tag{30}$$

and the Gaussian distributed

$$p_\tau^{(G)}(\tau) = \frac{2}{\pi\tau_0} e^{-\frac{\tau^2}{\pi\tau_0^2}}, \qquad \tau \geq 0, \tag{31}$$

where in both cases $\tau_0$ denotes the expected value of $\tau$. The probability-density function of $\lambda$ is assumed of the form[3,15]

$$p_\lambda(\lambda) = \frac{\rho}{1 - e^{-\rho}} e^{-\rho\lambda}, \qquad 0 \leq \lambda \leq 1, \tag{32}$$

with $\rho = 10$.

The equalizing-combiner improvement factor, $I$, is evaluated for the two joint probability-density functions, $p_\lambda(\lambda)p_\tau^{(e)}(\tau)$ and $p_\lambda(\lambda)p_\tau^{(G)}(\tau)$, as a function of the expected delay difference $\tau_0$. The results are shown on Fig. 11. Note the strong dependence of $I$ on $\tau_0$ for the curve corresponding to $p_\tau(\tau) = p_\tau^{(G)}(\tau)$. This is due to the Guassian nature of $p_\tau^{(G)}(\tau)$ which, for small values of $\tau_0$, provides smaller weighting factors for the large outage regions (of Fig. 7, 8, and 9) than does $p_\tau^{(e)}(\tau)$. Also,

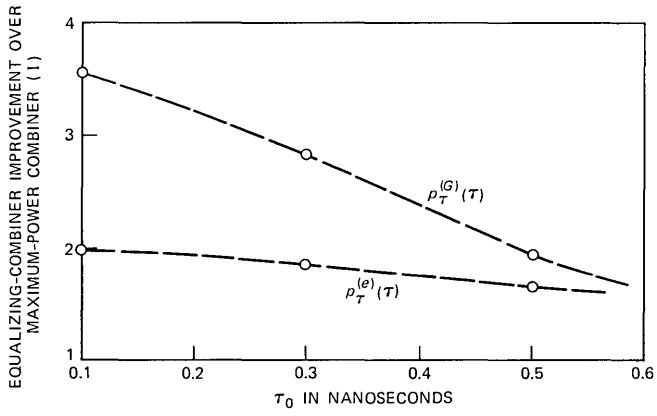



Fig. 11—Equalizing-combiner improvement factor as a function of the expected value of $\tau$.

note the change in improvement factor, for a given value of $\tau_0$, as a function of the two assumed probability-density functions of $\tau$. This is to be expected since the ratio of the outage areas of the maximum-power combiner to those of the equalizing combiner varies considerably for Figs. 6 through 10. Thus, the particular weighting given to the areas of each figure, as determined by the assumed probability-density function of $\tau$, will control the outcome of $I$. To illustrate this point further, we consider a delay difference $\tau$, which is uniformly distributed between $\tau = 2.8$ and 3.1 ns, and zero outside this range. That is, we assume:

$$p_\tau(\tau) \equiv p_\tau^{(u)}(\tau) = \begin{cases} \text{constant}, & \text{for } 2.8 \leq \tau \leq 3.1 \text{ ns}, \\ 0, & \text{otherwise}. \end{cases} \tag{33}$$

With this probability-density function we calculate an equalizing-combiner improvement factor of five relative to that of the maximum-power combiner, since only the outage region of Fig. 10 contributes (equalizing-combiner improvement factors of this magnitude have recently been observed on a reflective microwave radio path over water[7]). By a similar argument, if $\tau$ is assumed uniformly distributed between $\tau = 2.4$ and 2.7 ns, with zero probability of assuming a value outside this range, the equalizing-combiner improvement factor reduces to approximately unity.

From the above considerations it becomes apparent that the delay-difference statistics used in the two-ray model will play a major role in determining the equalizing-combiner performance relative to that of the maximum-power combiner. Regardless of the delay-difference statistics, however, our analysis predicts only moderate outage reductions to be gained from converting a maximum-power combiner to an equalizing combiner. This suggests that an amplitude-slope equalizer may still be required following the equalizing combiner to bring the outage time of high-speed digital signals to levels below long-haul outage objectives. Nevertheless, the equalizing combiner's ability to mitigate quadratic as well as linear amplitude distortions, during highly dispersive channel conditions, will contribute to enhancing the reliability and robustness of high-speed digital radio systems.

## IV. ACKNOWLEDGMENTS

I am indebted to L. A. Palazzo for the great deal of support he provided in generating the computer simulations and to T. L. Osborne for his constant encouragement. I am also grateful to W. D. Rummler for his many useful suggestions.

## REFERENCES

1. A. J. Giger and W. T. Barnett, "Effects of Multipath Propagation on Digital Radio," Paper 46.7, Int. Conf. on Commun., June 14–18, 1981, Denver, Colorado.

2. C. W. Lundgren and W. D. Rummler, "Digital Radio Outage Due to Selective Fading—Observation vs. Prediction from Laboratory Simulation," B.S.T.J., *58,* No. 5 (May–June 1979), pp. 1073–100.
3. W. C. Jakes, Jr., "An Approximate Method to Estimate an Upper Bound on the Effect of Multipath Delay Distortion on Digital Transmission," IEEE Trans. on Commun., *COM-27,* No. 1 (January 1979), pp. 76–81.
4. C. W. Anderson, S. G. Barber, and R. N. Patel, "The Effect of Selective Fading on Digital Radio," IEEE Trans. on Commun., *COM-27,* No. 12 (December 1979), pp. 1870–5.
5. W. T. Barnett, "Multipath Fading Effects on Digital Radio," IEEE Trans. on Commun., *COM-27,* No. 12 (December 1979), pp. 1842–8.
6. S. Komaki, I. Horikawa, K. Morita, and Y. Okamoto, "Characteristics of a High Capacity 16 QAM Digital Radio System in Multipath Fading," IEEE Trans. on Commun., *COM-27,* No. 12 (December 1979), pp. 1854–61.
7. T. Murase, K. Morita, and S. Komaki, "200 Mb/s 16-QAM Digital Radio System with New Countermeasure Techniques for Multipath Fading," Paper 46.1, Int. Conf. on Commun., June 14–18, 1981, Denver, Colorado.
8. Y. Y. Wang, "Simulation and Measured Performance of a Space Diversity Combiner for 6 GHz Digital Radio," IEEE Trans. on Commun., *COM-27,* No. 12 (December 1979), pp. 1896–907.
9. P. Hartmann and B. Bynum, "Design Considerations for an Extended Range Adaptive Equalizer," Paper 46.5, Int. Conf. on Commun., June 14–18, 1981, Denver, Colorado.
10. W. T. Barnett, C. W. Lundgren, Jr., W. D. Rummler, and Y. Y. Wang, "Equalizing Signal Combiner," U. S. Patent 4,261,056, applied for July 16, 1979, issued April 7, 1981.
11. S. Komaki, Y. Okamoto, and K. Tajima, "Performance of 16-QAM Digital Radio System Using New Space Diversity," Paper 52.2, Int. Conf. on Commun., June 8–12, 1980, Seattle, Washington.
12. W. D. Rummler, "A New Selective Fading Model: Application to Propagation Data," B.S.T.J., *58,* No. 5 (May–June 1979), pp. 1037–71.
13. H. Miedema, unpublished work.
14. K. L. Seastrand, Jr., "Space Diversity Receiver with Combined Step and Continuous Pase Control," U. S. Patent 4,160,952, applied for May 12, 1978, issued July 10, 1979.
15. L. J. Greenstein and V. K. Prabhu, "Analysis of Multipath Outage with Applications to 90 Mbit/s PSK Systems at 6 and 11 GHz," IEEE Trans. on Commun., *COM-27,* No. 1 (January 1979), pp. 68–75.
16. M. Emshwiller, "Characterization of the Performance of PSK Digital Radio Transmission in the Presence of Multipath Fading," Int. Conf. on Commun., ICC 1978.
17. L. W. Pickering and J. K. DeRosa, "Refractive Multipath Model for Line-of-Sight Microwave Relay Links," IEEE Trans. on Commun., *COM-27,* No. 8 (August 1979), pp. 1174–82.
18. O. Sasaki and T. Akiyama, "Multipath Delay Characteristics on Line-of-Sight Microwave Radio System," IEEE Trans. on Commun., *COM-27,* No. 12 (December 1979), pp. 1876–86.
19. M. Ramadan, "Availability Prediction of 8 PSK Digital Microwave Systems During Multipath Propagation," IEEE Trans. on Commun., *COM-27,* No. 12 (December 1979), pp. 1862–9.

## APPENDIX

### *Modulation of Combined-Signal Power*

By modulating the phase of one antenna signal with a low-frequency sinsuoid, the combined-signal power over some arbitrary frequency interval, $\omega_1 \leq \omega \leq \omega_2$ (spanning all or some portion of the channel) will be modulated periodically. In this appendix, we expand the periodic modulation of the combined-signal power in a Fourier series and show that, in the absence of channel dispersion, the fundamental component of the expansion is proportional to the sine of the phase difference between the antenna signals at the inputs of the summing hybrid (see

Fig. 1). During dispersive channel conditions, we show that the fundamental component of the expansion is a weighted average of the sine of the phase difference between the antenna signals at the inputs of the summing hybrid. We thus show that the fundamental component of the combined-signal power modulation can be used as a measure of phase-lead or phase-lag of one antenna signal relative to the other.[13]

Let $\bar{R}_M(j\omega, t)$ and $\bar{R}_D(j\omega, t)$ denote the complex spectral densities received by the main and diversity antennas, respectively, in response to a transmitted signal $s(t)$, with complex spectral density $\bar{S}(j\omega)$. Denoting the transmission-medium transfer function by $\bar{H}_M(j\omega, t)$ and $\bar{H}_D(j\omega, t)$, as seen by the main and diversity antennas, respectively, we can write

$$\bar{R}_M(j\omega, t) = \bar{H}_M(j\omega, t)\bar{S}(j\omega)e^{-j\omega T_0} \tag{34}$$

and

$$\bar{R}_D(j\omega, t) = \bar{H}_D(j\omega, t)\bar{S}(j\omega)e^{-j\omega T_0}, \tag{35}$$

where the delay $T_0$ denotes the line-of-sight propagation delay of the radio link. The dependence of $\bar{H}_M$ and $\bar{H}_D$ on $t$ allows for variations in the transmission-medium transfer function during anomalous propagation conditions (i.e., during multipath fading). However, in writing eqs. (34) and (35) the transmission-medium transfer functions are assumed quasi-static with respect to the rate-of-change of $s(t)$. Defining,

$$\bar{S}(j\omega) \equiv S_0(\omega)e^{j\phi(\omega)}, \tag{36}$$

$$\bar{H}_M(j\omega, t) \equiv H(\omega, t)e^{j\eta(\omega,t)}, \tag{37}$$

$$\bar{H}_D(j\omega, t) \equiv \delta(\omega, t)H(\omega, t)e^{j\{\eta(\omega,t)+\psi(\omega,t)\}}, \tag{38}$$

and

$$V(\omega, t) \equiv H(\omega, t)S_0(\omega), \tag{39}$$

eqs. (34) and (35) can be rewritten as

$$\bar{R}_M(j\omega, t) = V(\omega, t)e^{j\{\phi(\omega)+\eta(\omega,t)-\omega T_0\}} \tag{40}$$

and

$$\bar{R}_D(j\omega, t) = \delta(\omega, t)V(\omega, t)e^{j\{\phi(\omega)+\eta(\omega,t)+\psi(\omega,t)-\omega T_0\}}. \tag{41}$$

The quantities $H(\omega, t)$ and $\eta(\omega, t)$ appearing in eq. (37) represent the amplitude and phase transfer functions associated with the dynamic transmission medium. The parameter $\delta(\omega, t)$ of eq. (38) denotes the voltage ratio of the diversity to the main antenna signal as a function of frequency and time, and $\psi(\omega, t)$ describes the phase of the spectrum

received by the diversity antenna referenced to the phase of the spectrum received by the main antenna. Both parameters $\delta(\omega, t)$ and $\psi(\omega, t)$ must be frequency- as well as time-dependent since during anomalous propagation periods the spectral densities received by the two antennas will, in general, differ by a complex, frequency-dependent, time-varying factor.

Assuming ideal waveguide systems and ideal phase-shifting elements, the spectral densities at the inputs of the summing hybrid (Fig. 1) can be expressed as:

$$\bar{R}'_M(j\omega, t) = \bar{R}_M(j\omega, t)e^{-j\omega\tau_{WG}} \tag{42}$$

and

$$\bar{R}'_D(j\omega, t) = \bar{R}_D(j\omega, t)e^{j\{\theta+\alpha\sin\omega_m t-\omega\tau_{WG}\}}, \tag{43}$$

where $\tau_{WG}$ denotes the waveguide propagation delay (assumed the same for both equalized waveguide systems), and $\theta + \alpha \sin \omega_m t$ represents the instantaneous frequency-independent phase shift introduced into the diversity signal by the phase modulator/phase shifter combination. It can be seen from eqs. (40) through (43) that at the inputs of the summing hybrid the instantaneous phase difference between the diversity antenna spectrum and the main antenna spectrum is $\psi(\omega, t) + \theta + \alpha \sin \omega_m t$. The combined signal spectrum can now be expressed as

$$\begin{aligned}\bar{R}_C(j\omega, t) \equiv V_C(\omega, t)e^{j\phi_C(\omega,t)} &= \frac{1}{\sqrt{2}}[\bar{R}'_M(j\omega, t) + \bar{R}'_D(j\omega, t)] \\ &= \frac{1}{\sqrt{2}}[1 + \delta(\omega, t)e^{j\{\psi(\omega,t)+\theta+\alpha\sin\omega_m t\}}] \\ &\quad \cdot V(\omega, t)e^{j\{\phi(\omega)+\eta(\omega,t)-\omega(T_0+\tau_{WG})\}}.\end{aligned} \tag{44}$$

Expanding the exponential function inside the square brackets in terms of sine and cosine functions, followed by an expansion of $\sin\{\alpha \sin \omega_m t\}$ and $\cos\{\alpha \sin \omega_m t\}$ in terms of Bessel functions, yields (after some algebraic manipulation),

$$\begin{aligned}&V_C^2(\omega, t) \\ &\quad = \frac{1}{2} V_{co}^2(\omega, t)[1 - m_1(\omega, t)\sin \omega_m t + m_2(\omega, t)\cos 2\omega_m t \cdots],\end{aligned} \tag{45}$$

where

$$V_{co}^2(\omega, t) \equiv \delta(\omega, t)V^2(\omega, t)[\delta(\omega, t) + \frac{1}{\delta(\omega, t)} + 2J_0(\alpha)\cos \mu(\omega, t)], \tag{46}$$

$$m_1(\omega, t) \equiv \frac{4J_1(\alpha)\sin \mu(\omega, t)}{\delta(\omega, t) + \dfrac{1}{\delta(\omega, t)} + 2J_0(\alpha)\cos \mu(\omega, t)}, \tag{47}$$

$$m_2(\omega, t) \equiv \frac{4J_2(\alpha)\cos \mu(\omega, t)}{\delta(\omega, t) + \dfrac{1}{\delta(\omega, t)} + 2J_0(\alpha)\cos \mu(\omega, t)}, \cdots, \tag{48}$$

$$\mu(\omega, t) \equiv \psi(\omega, t) + \theta, \tag{49}$$

and $J_0$, $J_1$, $J_2$ denote Bessel functions of the first kind. Note that the modulation component, $m_1(\omega, t)$, at the fundamental frequency of phase modulation, $\omega_m$, is proportional to the sine of the average phase difference $\mu(\omega, t)$. [The average phase difference is obtained by averaging the instantaneous phase difference, $\psi(\omega, t) + \theta + \alpha \sin \omega_m t$, over one period of $\alpha \sin \omega_m t$. Over this time interval the term $\psi(\omega, t) + \theta$ is assumed fixed.] Furthermore, since $\alpha$ (the magnitude of phase modulation) is typically kept small, $0 < \alpha \ll \pi/2$, $J_0(\alpha)$ will be positive and less than unity. Also, the denominator term of $m_1(\omega, t)$ will always be positive since $-2 < 2J_0(\alpha)\cos \mu(\omega, t) < 2$ for all $\mu(\omega, t)$, given that $0 < \alpha \ll \pi/2$, and $\delta(\omega, t) + 1/\delta(\omega, t) \geq 2$ for all $\delta(\omega, t)$, $0 \leq \delta(\omega, t_0) \leq \infty$. Therefore, the sign of $m_1(\omega, t)$ will indicate whether, at a particular frequency and point in time, the phase of the diversity spectrum leads or lags the phase of the main antenna spectrum.

The combined-signal power contained in the frequency interval $\omega_1 \leq \omega \leq \omega_2$ can be expressed as:

$$p_C(t) = \frac{1}{\pi} \int_{\omega_1}^{\omega_2} V_C^2(\omega, t) d\omega. \tag{50}$$

Substituting eq. (45) into eq. (50) and manipulating terms results in:

$$p_C(t) = p_0(t)[1 - m_1'(t)\sin \omega_m t + m_2'(t)\cos 2\omega_m t \cdots] \tag{51}$$

where

$$p_0(t) \equiv \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t) d\omega, \tag{52}$$

$$m_1'(t) \equiv \frac{\displaystyle\int_{\omega_1}^{\omega_2} m_1(\omega, t) V_{co}^2(\omega, t) d\omega}{\displaystyle\int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t) d\omega}, \tag{53}$$

$$m_2'(t) \equiv \frac{\int_{\omega_1}^{\omega_2} m_2(\omega, t) V_{co}^2(\omega, t) d\omega}{\int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t) d\omega}, \quad \cdots . \tag{54}$$

Note that $m_1'(t)$ represents a weighted average of $m_1(\omega, t)$ over the interval $\omega_1 \leq \omega \leq \omega_2$, where the weighting function $V_{co}^2(\omega, t)$ is proportional to the average combined-signal power spectral density.

In the absence of channel dispersion (i.e., flat fading, or no fading at all) $\mu$ and $m_1$ become frequency independent, thus making $m_1'$ identically equal to $m_1$ [see eq. (53)]. Since $m_1$ is proportional to sin $\mu$, the sign of $m_1'$ can be used to determine whether $\mu$ represents a positive or negative angle (phase lead or phase lag of the diversity signal with respect to the main antenna signal) during dispersion-free periods. When the channel is dispersive, however (i.e., during periods of frequency-selective fading), the phase difference between the two spectra becomes frequency-dependent and $m_1'$ will reflect a weighted average of sin $\mu$ over the interval $\omega_1 \leq \omega \leq \omega_2$. This can be seen by substituting eqs. (46) and (47) into eq. (53). Doing so, we find

$$m_1'(t) = \frac{4J_1(\alpha) \int_{\omega_1}^{\omega_2} \delta(\omega, t) V^2(\omega, t) \sin \mu(\omega, t) d\omega}{\int_{\omega_1}^{\omega_2} V_{co}^2(\omega, t) d\omega}, \tag{55}$$

where the weighting function $\delta(\omega, t) V^2(\omega, t)$ is the magnitude of the product of the received spectral densities. In Section 2.1 we show that when $\omega_1 \leq \omega \leq \omega_2$ covers the entire channel bandwidth and $m_1'(t)$ is used in a feedback arrangement to control the phase-shifter value $\theta$, maximum-power combining of the received signals results for all channel conditions.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# Stochastic Analysis of Mechanizing Transaction Data Bases

By J. A. MORRISON and W. W. YALE

(Manuscript received February 16, 1982)

*In this paper we consider the transfer of records from a manual file system (MFS) to a mechanized data base management system when the conversion takes place through two processes. In the going-forward process a transfer is made by the regular work force when an order is received to change or delete a record in the MFS. In addition, the transfer of records is carried out by a crash force, working at a fixed rate. We investigate the expected number of records remaining in the MFS at future times, and the expected number of records removed from the MFS during the corresponding periods by the going-forward work force and by the crash force. We also derive the expected time taken to go from a prescribed number of records in the MFS to a smaller number. We give simple approximations for all of these quantities. The results have been used elsewhere in the construction of an economic model used to estimate cost/benefits and labor force levels during the mechanization of transaction data bases. The numerical results presented graphically are typical of two Bell System applications.*

## I. INTRODUCTION

In this paper, we analyze the generic problem of implementing a system that requires manual effort for converting paper records in a manual file system (MFS) to mechanized computer records in a data base management system (DBMS). We can achieve conversion to a mechanized environment in a number of ways: by going-forward on an activity basis, by crash-conversion utilizing an augmented work force, or by some mix of the two dictated by available resources such as money, people, space, and time objectives.

A paper record is mechanized using the going-forward approach if the record is transferred from the MFS to the DBMS by regular record

maintenance employees when a change or deletion activity is performed upon that record. If a randomly chosen record is transferred by a separate temporary work force, the record is said to be transferred using the crash approach. The paper record could contain all of the information that is intended to be mechanized or an enhancement to an already existing mechanized record.

The arrival times of activities (change or deletion orders) for a particular record are assumed to form a Poisson process. We further assume that the distributions of the activity arrival times for every record are independent and identical. Hence, the sequence of times at which these change and deletion activities occur forms a nonstationary Poisson process. In addition, the transfer of records is carried out by a special task force with constant total mean rate, the amount of work required for a record to be transferred being exponentially distributed. Hence, there are two different stochastic processes that are concurrently operating against the MFS. Each record in the MFS is transferred if an activity occurs upon it, while at the same time each record in the MFS has a chance of randomly being selected by a crash person when he/she completes a conversion.

When an activity occurs against a record in the MFS, it will be assumed that a going-forward person will immediately attend to transferring and processing the record. Therefore, if substantially more than the average number of records concurrently experience activity, it will be assumed that management will borrow going-forward people from other areas or have the normal going-forward force work overtime. However, the time it takes for a crash-conversion person to perform the conversion is pertinent since the force is assumed to be fixed. In this case, the average number of records that the entire crash conversion force can convert in a specified period determines the mean of the stochastic process describing the crash effort.

In this paper we highlight the derivations of formulas that are used in the construction of a model[1] used to estimate cost/benefits and labor force levels during the mechanization of transaction data bases. In the next section, we investigate the expected number of records remaining in the MFS, removed by the going-forward people, and removed by the crash force people. In particular, we point out a simple approximation to the expected number of records remaining in the MFS, which is valid over a significant part of the range of interest. We derive the expected passage time from $l$ to $m$ records in the MFS ($l > m$) in Section III, and obtain a simple asymptotic approximation for the expected time at which the MFS becomes empty. Finally, in Section IV, we present some conclusions, and discuss the relevance of the results to Bell System applications.

## II. EXPECTED NUMBER OF RECORDS

We will now consider a stochastic model for the transfer of records from a MFS to a DBMS. We assume that there are $n$ crash force people, and that the conversion times of the records are independent and exponentially distributed with mean rate $\rho$. At time $\tau = 0-$ there are $S$ records in the MFS. At time $\tau = 0$ there are $n$ of these records removed from the MFS by the crash force. When the conversion of a record is completed, that record is instantaneously transferred to the DBMS, and another record is removed from the MFS by the crash person involved. A record is also removed from the MFS, by a going-forward person, when a change or deletion order is received; that record is transferred to the DBMS at some later time. We assume that the sequence of times at which change and deletion orders are received forms a nonstationary Poisson process with rate $i\mu$ where $i$ is the number of records in the MFS. (The mean deletion and change order rates are summed to get $\mu$, since it is a well-known fact that if two independent Poisson processes are affecting the records, then the composite of these two processes can be described by a Poisson process where the parameters are summed.) If a change or deletion order is received for a record that is in the process of being converted by a crash person, the change or deletion order is not executed until after the record has been transferred to the DBMS.

The number of records in the MFS at time $\tau = 0+$ is

$$M = S - n. \tag{1}$$

For convenience, we let

$$n\rho = \nu\mu. \tag{2}$$

This is the mean rate at which the crash force converts records. We also let $P_i(\tau)$ denote the probability that there are $i$ records in the MFS at time $\tau$. But, the number of records in the MFS is a pure death process,[2] with death rate

$$\mu_i = n\rho(1 - \delta_{i0}) + i\mu = \mu[\nu(1 - \delta_{i0}) + i], \qquad i = 0, \cdots, M, \tag{3}$$

where $\delta_{ij}$ denotes the Kronecker delta; $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$. Hence,

$$\frac{dP_i}{d\tau} = -\mu[\nu(1 - \delta_{i0}) + i]P_i + \mu(1 - \delta_{iM})(\nu + i + 1)P_{i+1}, \tag{4}$$

for $\tau > 0$ and $i = 0, \cdots, M$. The initial condition at time $\tau = 0+$ is

$$P_i(0+) = \delta_{iM}, \qquad i = 0, \cdots, M. \tag{5}$$

The expected number of records in the MFS at time $\tau$ is

$$N(\tau) = \sum_{i=1}^{M} iP_i(\tau). \tag{6}$$

It follows from (4) through (6) that

$$\frac{dN}{d\tau} + \mu N = -\nu\mu[1 - P_0(\tau)]; \qquad N(0+) = M. \tag{7}$$

It is straightforward to calculate the Laplace transforms of $P_i$, $i = 0, \cdots, M$, from (4) and (5), and thence that of $N$ from (7). By inverting the latter transform, Morrison[3] obtained an explicit expression for $N(\tau)$ in terms of an incomplete beta function.[4] Asymptotic approximations to $N(\tau)$, involving the complementary error function,[4] were derived when $M \gg 1$ and $\nu \gg 1$.

Since $0 \le P_0(\tau) \le 1$, it follows from (7) that

$$L(\tau) \equiv (M + \nu)e^{-\mu\tau} - \nu \le N(\tau) \le Me^{-\mu\tau}. \tag{8}$$

The lower bound is, of course, superfluous if $L(\tau) \le 0$. When $M \gg 1$ and $\nu \ll M$, these bounds yield an accurate approximation to $N(\tau)$ as long as $Me^{-\mu\tau} \gg \nu$. An upper bound for $N(\tau)$, which is valid only for $L(\tau) \ge 0$, and exact for $L(\tau) = 0$, was derived from the exact result.[3] It was shown that

$$0 \le N(\tau) - L(\tau) \le \frac{\Gamma(M + \nu)}{\Gamma(M)\Gamma(\nu)} e^{-\nu\mu\tau}(1 - e^{-\mu\tau})^M \quad \text{for} \quad L(\tau) \ge 0. \tag{9}$$

From (9) we deduce that

$$N(\tau) \approx L(\tau) \quad \text{for} \quad M \gg 1, \qquad L(\tau) \gg \min(\sqrt{M}, \sqrt{\nu}) \frac{\nu^\nu e^{-\nu}}{\sqrt{\nu}\Gamma(\nu)}, \tag{10}$$

where, from the properties of the gamma function,[4]

$$\frac{\nu^\nu e^{-\nu}}{\Gamma(\nu)} < \nu \quad \text{for} \quad \nu > 0; \qquad \frac{\nu^\nu e^{-\nu}}{\sqrt{\nu}\Gamma(\nu)} \sim \frac{1}{\sqrt{2\pi}} \quad \text{for} \quad \nu \gg 1. \tag{11}$$

Hence, (10) gives a larger range of validity for the approximation $N(\tau) \approx L(\tau)$ than does (8), and significantly so if $\nu \gg 1$, as it is in cases of interest. We remark that intuitively we expect $P_0(\tau)$, the probability that there are no records in the MFS at time $\tau$, to be extremely small for a significant time when $M \gg 1$, and in that time interval we deduce from (7) that $N(\tau) \approx L(\tau)$. The intuitive derivation, however, does not give the range of validity of the approximation.

Typical numerical results are illustrated in Figs. 1 and 2 for the case $S = 10^5$, $n = 10$, $\mu = 0.001608$ per day, and $\nu = 18657$. Only $N(\tau)$ and the lower bound $L(\tau)$ are shown in Fig. 1. The upper bound given by (9) is also shown in Fig. 2. This figure depicts the tail region, in which the expected number of records in the MFS is considerably fewer than

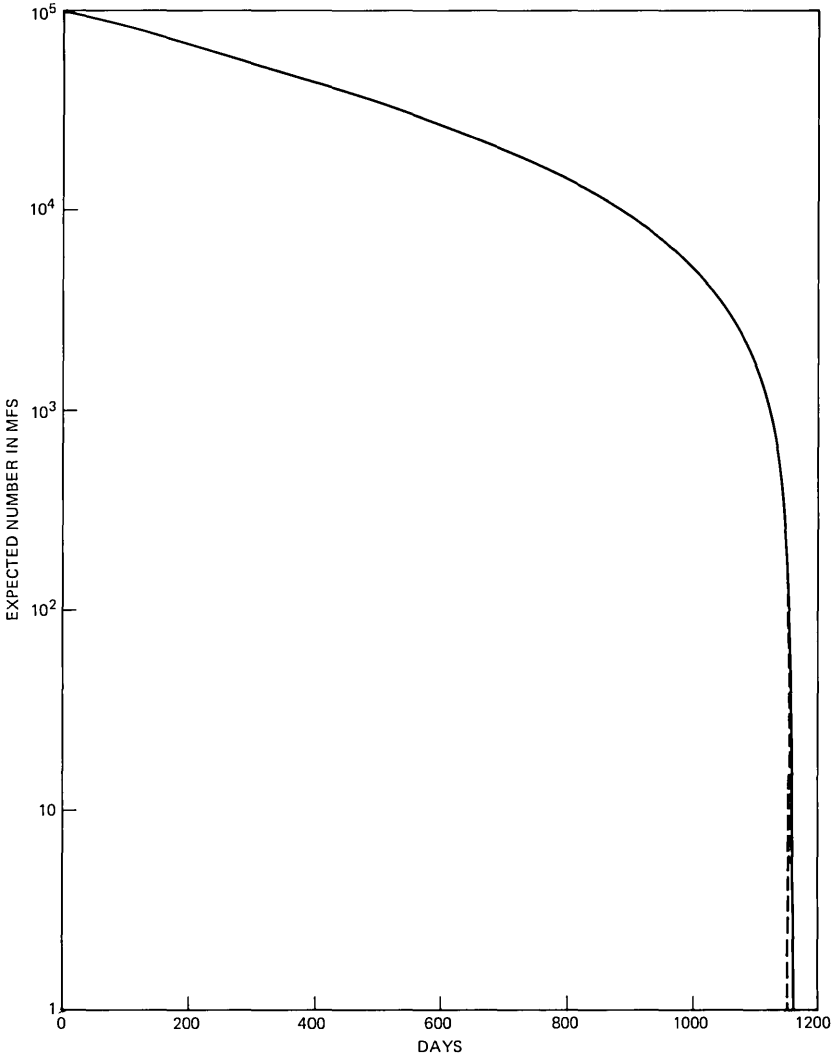



Fig. 1—Expected number of records in the MFS, and lower bound, as a function of time, for $S = 10^5$, $n = 10$, $\mu = 0.001608$ per day, and $\nu = 18657$.

the initial number. It is seen that $N(\tau)$ is still quite close to $L(\tau)$ for values of $L(\tau)$ just a few times greater than $\min(\sqrt{M}, \sqrt{\nu})/\sqrt{2\pi} \approx 54.5$.

Another quantity of interest is the expected number of records $F(\tau)$ removed from the MFS by the going-forward people in the time interval $(0, \tau]$. Let $P_{ik}(\tau)$ denote the probability that there are $i$ records in the MFS at time $\tau$, and that $k$ records have been removed from the MFS by the going-forward people in $(0, \tau]$. Then

$$F(\tau) = \sum_{i=0}^{M-1} \sum_{k=1}^{M-i} kP_{ik}(\tau). \tag{12}$$

But,

$$\frac{dP_{ik}}{d\tau} = -\mu[\nu(1 - \delta_{i0}) + i]P_{ik} + \mu\nu(1 - \delta_{k,M-i})P_{i+1,k} + \mu(1 - \delta_{k0})(i + 1)P_{i+1,k-1}, \tag{13}$$

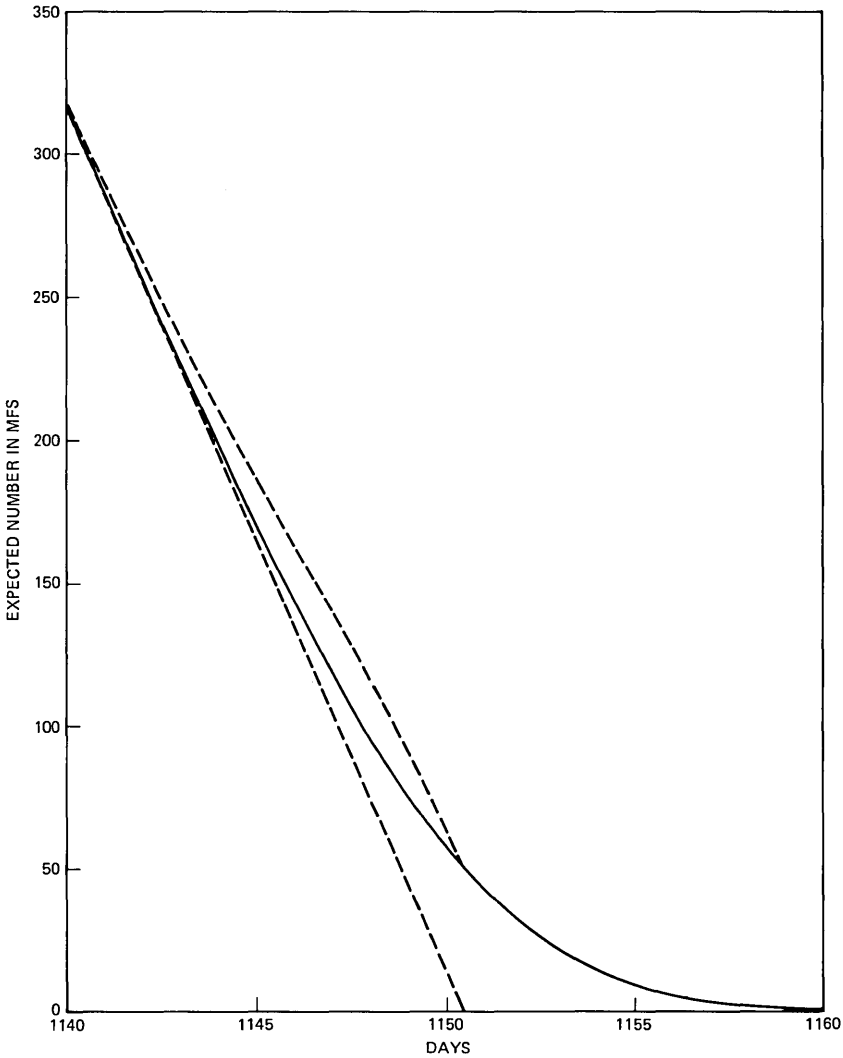



Fig. 2—Expected number of records in the MFS, and upper and lower bounds, as a function of time, for $S = 10^5$, $n = 10$, $\mu = 0.001608$ per day, and $\nu = 18657$.

for $\tau > 0$, $k = 0, \cdots, M - i$, and $i = 0, \cdots, M$. The initial condition at time $\tau = 0+$ is

$$P_{ik}(0+) = \delta_{iM}, \qquad k = 0, \cdots, M - i, \qquad \text{and} \qquad i = 0, \cdots, M. \tag{14}$$

We note that $i = M$ implies that $k = 0$. Also,

$$P_i(\tau) = \sum_{k=0}^{M-i} P_{ik}(\tau), \tag{15}$$

and if we sum (13) and (14) from $k = 0$ to $M - i$ we obtain (4) and (5).

It follows directly from (12), (13), and (15) that

$$\frac{dF}{d\tau} = \mu \sum_{i=1}^{M} iP_i(\tau) = \mu N(\tau). \tag{16}$$

This implies that the rate of change of the expected number of records removed from the MFS by the going-forward people is equal to the expected value at time $\tau$ of the rate at which change and deletion orders are received. If we let $C(\tau)$ denote the expected number of records removed from the MFS by the crash people in $(0, \tau]$, then

$$C(\tau) = M - N(\tau) - F(\tau). \tag{17}$$

From (7), (16), and (17), we obtain

$$\frac{dC}{d\tau} = \mu\nu[1 - P_0(\tau)] = \mu\nu \sum_{i=0}^{M} (1 - \delta_{i0})P_i(\tau), \tag{18}$$

which has an interpretation analogous to that of (16). We note that $dC/d\tau \approx \mu\nu$ in the region where $P_0(\tau)$ is small. Also, corresponding to the approximation $N(\tau) \approx L(\tau)$, where $L(\tau)$ is defined in (8), it follows from (16) that

$$F(\tau) \approx (M + \nu)(1 - e^{-\mu\tau}) - \mu\nu\tau, \tag{19}$$

since $F(0+) = 0$.

## III. EXPECTED PASSAGE TIMES

We next turn to the calculation of the expected passage time from $l$ to $m$ records in the MFS. Now,[5] for a pure death process, with death rate $\mu_i$, the passage time $\tau_{i,i-1}$ from state $i$ to state $i - 1$ is exponentially distributed with density $\mu_i \exp(-\mu_i\tau)$. Hence, in particular, the expected value of $\tau_{i,i-1}$ is

$$E\tau_{i,i-1} = \frac{1}{\mu_i}. \tag{20}$$

The passage time from $l$ to $m$ records in the MFS is

$$\tau_{l,m} = \sum_{i=m+1}^{l} \tau_{i,i-1}, \tag{21}$$

and the random variables in the sum are independent. From (3), (20), and (21), we obtain

$$\mu E\tau_{l,m} = \sum_{i=m+1}^{l} \frac{1}{(\nu + i)} = \psi(l + \nu + 1) - \psi(m + \nu + 1), \qquad (22)$$

where $\psi$ denotes the logarithmic derivative of the gamma function.[4]

We will make use of the asymptotic result

$$\psi(x) = \log x + O\left(\frac{1}{x}\right) \qquad \text{for} \quad x \gg 1. \qquad (23)$$

Hence, from (22), the expected time at which the MFS becomes empty is given by

$$\begin{aligned} \mu E\tau_{M,0} &= \psi(M + \nu + 1) - \psi(\nu + 1) \\ &\sim \log\left(\frac{M + \nu}{\nu}\right) \qquad \text{for} \quad \nu \gg 1. \end{aligned} \qquad (24)$$

Also, the expected time starting from $l$ records until the MFS becomes empty is given by

$$\mu E\tau_{l,0} = \psi(l + \nu + 1) - \psi(\nu + 1) \sim \log\left(\frac{l + \nu}{\nu}\right) \qquad \text{for} \quad \nu \gg 1$$

$$\sim \frac{1}{\nu} \qquad \text{for} \quad \nu \gg 1, \qquad l \ll \nu. \qquad (25)$$

Next, the expected time at which the number of records in the MFS first reaches $m$ is given by

$$\begin{aligned} \mu E\tau_{M,m} &= \psi(M + \nu + 1) - \psi(m + \nu + 1) \\ &\sim \log\left(\frac{M + \nu}{m + \nu}\right) \qquad \text{for} \quad \nu \gg 1. \end{aligned} \qquad (26)$$

But, from (8) to (10), the time $\tau_m$ at which the expected number of records in the MFS is equal to $m$ satisfies

$$\mu\tau_m \sim \log\left(\frac{M + \nu}{m + \nu}\right)$$

$$\text{for} \quad M \gg 1, \qquad \nu \gg 1, \qquad m \gg \min(\sqrt{M}, \sqrt{\nu})/\sqrt{2\pi}. \qquad (27)$$

Hence, under the restrictions in (27), we have $E\tau_{M,m} \sim \tau_m$.

We now consider the final stage of the conversion by the crash force, starting from the time $\tau_{M,0}$ when the MFS becomes empty. Each of the $n$ crash people is then busy with a record. We first assume that a person is removed from the crash force when he/she completes the conversion of the record on which he/she has been working. If $j$ is the number of records remaining to be converted by the crash force, we then have a pure death process with $\mu_j = j\rho, j = 0, \cdots, n$. If $\tau_c$ denotes

the time at which the conversion by the crash force is completed, then, with the help of (2),

$$\mu E(\tau_c - \tau_{M,0}) = \mu \sum_{j=1}^{n} \frac{1}{\mu_j} = \frac{n}{\nu} \sum_{j=1}^{n} \frac{1}{j} = \frac{n}{\nu} [\psi(n+1) - \psi(1)]. \quad (28)$$

At the other extreme, we assume that the entire crash force works jointly on the remaining records, until the conversion of the last record is completed. This will give a lower bound on the expected time to complete the conversion of the final $n$ records, since the crash force continues to work at its maximum rate. In this case the death rate is $\hat{\mu}_j = n\rho = \nu\mu$, $j = 1, \cdots, n$. If $\hat{\tau}_c$ denotes the corresponding time at which the conversion by the crash force is completed, then

$$\mu E(\hat{\tau}_c - \tau_{M,0}) = \mu \sum_{j=1}^{n} \frac{1}{\hat{\mu}_j} = \frac{n}{\nu}. \quad (29)$$

## IV. CONCLUSION

In this paper we presented formulas that are employed in an economic model[1] used to estimate cost/benefits and labor force levels associated with the mechanization of transaction data bases. Because the formulas can be computed very rapidly on any modern computer, the model can be used as an interactive managerial decision tool to perform what-if studies in real time. Numerical results presented graphically are typical of two applications, namely, (*i*) the conversion of manual records in a service order processing operation of a Bell Operating Company business office, and (*ii*) the mechanization of equipment inventory records during the implementation of Trunks Integrated Record Keeping System.[1] We do not presume that the stochastic model is appropriate for all possible data base conversions, but evidence from field trials indicates that it does adequately model a certain class of Bell System mechanization problems.

## V. ACKNOWLEDGMENTS

We would like to thank T. A. Bottomley for many helpful contributions concerning the definition of the problem, and J. B. Seery for writing the programs to obtain the numerical results presented in Figs. 1 and 2.

## REFERENCES

1. W. W. Yale, unpublished work.
2. L. Kleinrock, *Queueing Systems, Volume I: Theory*, New York: John Wiley, 1975.
3. J. A. Morrison, unpublished work.
4. W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, New York: Springer-Verlag, 1966.
5. J. Keilson, *Markov Chain Models—Rarity and Exponentiality*, Applied Mathematical Sciences, *28*, New York: Springer-Verlag, 1979, p. 21.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# Smoothing With Periodic Cubic Splines

By N. Y. GRAHAM

(Manuscript received May 7, 1982)

*In this paper we present a mathematical algorithm for constructing a smoothing cubic spline with periodic end conditions and a predetermined 'closeness of fit' to a given set of points in the plane. In addition to providing a mathematical tool for smoothing raw data in which the underlying function is known to be periodic, this algorithm has special significance in computer graphics, because the use of smoothing functions with periodic end conditions is essential for producing visually acceptable, smooth, closed curves. Sample plots are included to illustrate the power and flexibility of this algorithm.*

## I. INTRODUCTION

Although "natural" splines are used extensively and are quite appropriate for smoothing many types of data, they often produce less than satisfactory results when used to smooth data points that belong to a periodic function. The inappropriateness of using "natural" splines to approximate periodic data is especially evident in graphics applications. In particular, when the data points represent a closed curve, smoothing (parametrically) with "natural" splines will lead to unacceptable results because the "natural" end conditions will cause the curve either to close up with a noticeable discontinuity, or to not close up at all (see Fig. 1).

Existing methods for constructing smoothing splines with a predetermined closeness of fit all lead to splines with "natural" end conditions.[1,2] A method developed by Spath[3] produces a smoothing spline with periodic end conditions, but the closeness of fit cannot be determined in advance. In this paper we will describe a method for constructing a smoothing cubic spline that has periodic end conditions and that also satisfies a predetermined closeness of fit to a given set of data points.

This algorithm has potentially wide applicability, especially in the realm of interactive graphics. It makes possible the computer genera-

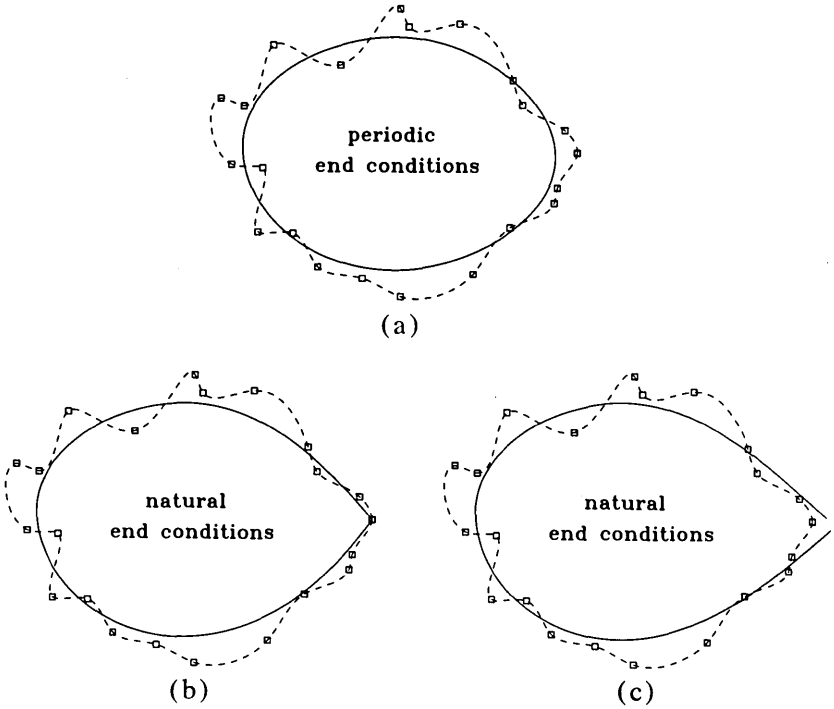



Fig. 1—Comparison of periodic versus natural spline smoothing. (a) Periodic cubic spline smoothing with uniform weights. (b) Natural cubic spline smoothing with small weights at end points. (c) Natural cubic spline smoothing with uniform weights.

tion of free-form, smooth, closed curves by merely specifying the approximate locations of as few as three distinct points. The shape of the curve can be controlled easily by moving one or more points, or by adjusting the weighting factors associated with some or all of the points.

An efficient program based on this new algorithm has been written and tested. Sample plots illustrating this method are included.

## II. TERMINOLOGY

Let $P_k = (x_k, y_k)$, $k = 1, n$, be $n$ points in the plane. A "cubic spline" on $[x_1, x_n]$ with knots at $x_1, \cdots, x_n$, is a function $f$ that coincides with a third-order polynomial $f_k$ on each sub-interval $[x_k, x_{k+1}]$, $k = 1, n - 1$, and such that $f$ is continuous and has continuous first and second derivatives over the entire interval $[x_1, x_n]$.

In other words, $f$ is a cubic spline on $[x_1, x_n]$ if, for each $k = 1, n - 1$ there exist real numbers $a_k$, $b_k$, $c_k$, $d_k$ (the "spline coefficients" of $f$) such that for every $x$ in $[x_k, x_{k+1}]$,

$$f(x) = f_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3. \quad (1)$$

Furthermore, the continuity of $f$, $f'$, and $f''$ on $[x_1, x_n]$ implies that at each interior knot $x_k$, $k = 2, n - 1$,

$$f_{k-1}(x_k) = f_k(x_k), \quad (2)$$

$$f'_{k-1}(x_k) = f'_k(x_k), \quad (3)$$

$$f''_{k-1}(x_k) = f''_k(x_k). \quad (4)$$

The cubic spline $f$ is said to be "periodic" if it satisfies the following additional conditions (known as "periodic end conditions"):

$$f(x_n) = f(x_1), \quad (5)$$

$$f'(x_n) = f'(x_1), \quad (6)$$

$$f''(x_n) = f''(x_1). \quad (7)$$

A "natural" cubic spline differs from a "periodic" cubic spline in that it satisfies the so-called "natural end conditions": $f''(x_1) = f''(x_n) = 0$.

### III. FORMAL STATEMENT OF THE PROBLEM

Let $P_k = (x_k, y_k)$, $k = 1, n$, be $n$ points in the plane, with $x_1 < x_2 < \cdots < x_n$. Let $w_k$, $k = 1, n$, be positive real numbers ("weighting factors") associated with $P_k$, $k = 1, n$, respectively. (Assume $y_n = y_1$, $w_n = w_1$.) Given an arbitrary constant, $M > 0$, the problem is to determine the set of $4(n - 1)$ coefficients of the periodic cubic spline $f$ on $[x_1, x_n]$ with knots at $x_1, \cdots, x_n$, such that

(*i*) $f$ has minimal "total curvature" $G(f) = \int_{x_1}^{x_n} f''(x)^2 dx$, and

(*ii*) $f$ satisfies the following weighted, distance-squared constraint, or "closeness of fit," with respect to the given points:

$$H(f) = \sum_{k=1}^{n} \left[ \frac{f(x_k) - y_k}{w_k} \right]^2 \leq M.$$

Note that the weighting factors give inverse importance to the points. Note also that $M$ controls the degree of smoothness, so that increasing the value of $M$ for a fixed set of weighting factors will lead to a smoother, or flatter, spline. Conversely, choosing a sufficiently small value of $M$ will lead to a spline that closely approximates an interpolating spline. In general, an appropriate choice for the value of $M$ will depend on the values chosen for the weighting factors. If, for example, the weighting factors are chosen so that $w_k$ is the standard deviation at $x_k$, then a suitable choice for $M$ would be some value between the confidence limits, $n - \sqrt{2n}$ and $n + \sqrt{2n}$.

## IV. GENERAL APPROACH TO THE SOLUTION

To minimize $G(f)$ subject to the constraint $H(f) \leq M$, we introduce an auxiliary variable $z$ and a non-negative Lagrange multiplier $p$ and minimize

$$F(f) = G(f) + p[H(f) + z^2 - M].$$

This is the approach used by Reinsch in Ref. 2 to solve the analogous problem for "natural" cubic splines. However, Reinsch minimizes $F$ over the class of all continuous functions with continuous first and second derivatives on $[x_1, x_n]$, which leads to a "natural" cubic spline as the solution. We will restrict the class of admissible functions in the minimization of $F$ to periodic cubic spline splines on $[x_1, x_n]$ with knots at $x_1, \cdots, x_n$, obtaining a direct solution to our problem.

## V. LINEAR SYSTEM RESULTING FROM CONTINUITY AND PERIODICITY CONDITIONS

Let $f$ be an arbitrary periodic cubic spline on $[x_1, x_n]$, with spline coefficients $a_k$, $b_k$, $c_k$, $d_k$, $k = 1, n-1$. Then, for $x$ in $[x_k, x_{k+1}]$, $f(x)$ is given by (1) above, and $f'(x)$ and $f''(x)$ are given below:

$$f'(x) = f_k'(x) = b_k + 2c_k(x - x_k) + 3d_k(x - x_k)^2, \tag{8}$$

$$f''(x) = f_k''(x) = 2c_k + 6d_k(x - x_k). \tag{9}$$

Expressing $f$, $f'$, and $f''$ explicitly in eqs. (1), (8), and (9), respectively, allows us to derive linear relationships among the spline coefficients. From the continuity of $f'$ at the interior knots of (3) and the periodicity of $f'$ on $[x_1, x_n]$ in (6), it follows that:

$$2c_k h_k = b_{k+1} - b_k - 3d_k h_k^2, \qquad \text{for} \quad k = 1, n-1, \tag{10}$$

where $h_k = x_{k+1} - x_k$ for $k = 1, n-1$, and $b_n$ denotes $b_1$.

From the continuity of $f$ at the interior knots of (2) and the periodicity of $f$ on $[x_1, x_n]$ in (5), the first-order coefficients may be expressed in terms of the constant, second-order, and third-order coefficients:

$$b_k = (a_{k+1} - a_k)/h_k - c_k h_k - d_k h_k^2, \qquad \text{for} \quad k = 1, n-1, \tag{11}$$

where $a_n$ denotes $a_1$.

From the continuity of $f''$ at the interior knots of (4) and the periodicity of $f''$ on $[x_1, x_n]$ in (7), the third-order coefficients may be expressed as a function of the second-order coefficients:

$$d_k = (c_{k+1} - c_k)/3h_k, \qquad \text{for} \quad k = 1, n-1, \tag{12}$$

where $c_n$ denotes $c_1$.

Using (11) and (12) to eliminate the $b_k$'s and $d_k$'s from (10) leads to a system of linear equations in the $a_k$'s and $c_k$'s, given in matrix notation as follows:

$$\mathbf{Sc} = 3\mathbf{Qa}, \tag{13}$$

where $\mathbf{S}$ and $\mathbf{Q}$ are symmetric, cyclic-tridiagonal matrices of order $n - 1$, and $\mathbf{c}$, $\mathbf{a}$ are the column vectors $(c_1, \cdots, c_{n-1})^T$, $(a_1, \cdots, a_{n-1})^T$, respectively.

The non-zero entries of $\mathbf{S}$ and $\mathbf{Q}$ are expressed in terms of the distances between successive knots:

$$\begin{aligned}
\mathbf{S}(k, k) &= 2(h_{k-1} + h_k) \qquad \text{for} \quad k = 1, n - 1 \\
\mathbf{S}(k, k+1) &= \mathbf{S}(k+1, k) = h_k \qquad \text{for} \quad k = 1, n - 2 \\
\mathbf{S}(1, n-1) &= \mathbf{S}(n-1, 1) = h_{n-1} \\
\mathbf{Q}(k, k) &= -1/h_{k-1} - 1/h_k \qquad \text{for} \quad k = 1, n - 1 \\
\mathbf{Q}(k, k+1) &= \mathbf{Q}(k+1, k) = 1/h_k \qquad \text{for} \quad k = 1, n - 2 \\
\mathbf{Q}(1, n-1) &= \mathbf{Q}(n-1, 1) = 1/h_{n-1},
\end{aligned}$$

where $h_0$ denotes $h_{n-1}$. By Gershgorin's Theorem[4] it can be shown that $\mathbf{S}$ is positive definite (and therefore non-singular), while $\mathbf{Q}$ is positive semi-definite and singular with rank $n - 2$.

## VI. LINEAR SYSTEM RESULTING FROM MINIMIZING $F$ WITH RESPECT TO THE CONSTANT COEFFICIENTS

Note that (13) is a system of $n - 1$ linear equations in $2(n - 1)$ unknowns: the constant coefficients and the second-order coefficients. We shall derive a second system of $n - 1$ linear equations in these unknowns. We proceed by first showing that $H$ and $G$ can be expressed as functions of the constant coefficients only.

From the spline representation of $f$ in (1), it follows immediately that

$$H = \sum_{k=1}^{n} \left[\frac{f(x_k) - y_k}{w_k}\right]^2 = \sum_{k=1}^{n} \left(\frac{a_k - y_k}{w_k}\right)^2$$

is a function of $a_1, \cdots, a_n$. And since the periodicity of $f$ implies

$$a_n = f(x_n) = f(x_1) = a_1,$$

$H$ is a function of the constant spline coefficients $a_1, \cdots, a_{n-1}$.

From the explicit representation of $f''$ in (9), the "total curvature" $G$ can be expressed as

$$\begin{aligned}
G = \int_{x_1}^{x_n} f''(x)^2 dx &= \sum_{k=1}^{n-1} \int_{x_k}^{x_{k+1}} f''(x)^2 dx \\
&= \sum_{k=1}^{n-1} \int_{x_k}^{x_{k+1}} [2c_k + 6d_k(x - x_k)^2 dx.
\end{aligned}$$

Evaluating the integrals over each sub-interval directly and eliminating the $d_k$'s with (12) lead to the following:

$$G = \sum_{k=1}^{n-1} \frac{4}{3} h_k (c_k^2 + c_k c_{k+1} + c_{k+1}^2).$$

Rewriting in matrix notation and applying (13), we have:

$$G = (\tfrac{2}{3})\mathbf{c}^T\mathbf{S}\mathbf{c} = (\tfrac{2}{3})(3\mathbf{S}^{-1}\mathbf{Q}\mathbf{a})\mathbf{S}(3\mathbf{S}^{-1}\mathbf{Q}\mathbf{a}) = 6\mathbf{a}^T\mathbf{Q}\mathbf{S}^{-1}\mathbf{Q}\mathbf{a}.$$

Thus, $G$ also can be expressed in terms of the constant spline coefficients. Note that from this representation of $G$,

$$\frac{\partial G}{\partial a_k} = 12\mathbf{Q}_k\mathbf{S}^{-1}\mathbf{Q}\mathbf{a}$$

for $k = 1, n-1$, where $\mathbf{Q}_k$ is the $k$th row of $\mathbf{Q}$.

Since $G$ and $H$ are functions of $a_1, \cdots, a_{n-1}$, then $F$ is a function of the independent variables $a_1, \cdots, a_{n-1}, p$, and $z$. In order for $F$ to be minimized, the partial derivative of $F$ with respect to each of its independent variables must vanish. Thus, for each $k = 1, n-1$, differentiating $F$ with respect to $a_k$ yields:

$$\frac{\partial F}{\partial a_k} = \frac{\partial G}{\partial a_k} + p\frac{\partial H}{\partial a_k} = 12\mathbf{Q}_k\mathbf{S}^{-1}\mathbf{Q}\mathbf{a} + 2p\left(\frac{a_k - y_k}{w_k^2}\right) = 0.$$

Rewriting this set of $n-1$ linear equations in matrix notation and using (13) to replace $\mathbf{S}^{-1}\mathbf{Q}\mathbf{a}$ with $\mathbf{c}/3$ lead to:

$$4\mathbf{Q}\mathbf{c} + 2p\mathbf{W}^{-2}(\mathbf{a} - \mathbf{y}) = \mathbf{0}, \tag{14}$$

where $\mathbf{y}$ is the column vector $(y_1, \cdots, y_{n-1})^T$. Combining (13) and (14), we have the following linear system in $\mathbf{c}$:

$$(p\mathbf{S} + 6\mathbf{Q}\mathbf{W}^2\mathbf{Q})\mathbf{c} = 3p\mathbf{Q}\mathbf{y}. \tag{15}$$

The matrix $\mathbf{A}_p = (p\mathbf{S} + 6\mathbf{Q}\mathbf{W}^2\mathbf{Q})$ is symmetric, five-banded with three non-zero entries in the upper right and lower left corners. It can be shown that, for all positive values of $p$, $\mathbf{A}_p$ is positive definite. Thus, for each $p > 0$, (15) has a unique solution in $\mathbf{c}$.

Note that for each non-zero value of $p$, $\mathbf{a} = \mathbf{y} - (2/p)\mathbf{W}^2\mathbf{Q}\mathbf{c}$ from (14). Note also that from (12) $\mathbf{d}$ is uniquely determined by $\mathbf{c}$, and from (11) $\mathbf{b}$ is uniquely determined by $\mathbf{a}$, $\mathbf{c}$, and $\mathbf{d}$. Thus, to each positive value of $p$ corresponds a unique periodic cubic spline on $[x_1, x_n]$, whose coefficients are given in the vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$.

## VII. CONSEQUENCES OF MINIMIZING $F$ WITH RESPECT TO $p$ AND $z$

Minimizing $F = G + p(H + z^2 - M)$ with respect to the Lagrange multiplier $p$ leads to:

$$\frac{\partial F}{\partial p} = H + z^2 - M = 0,$$

which merely states that the distance constraint on $H$ (expressed as an equality in terms of the auxiliary variable $z$) must be satisfied when the minimal value of $F$ is attained.

On the other hand, minimizing $F$ with respect to $z$ yields:

$$\frac{\partial F}{\partial z} = 2pz = 0,$$

which implies that at least one of the two variables, $p$ or $z$, must be equal to 0 when the minimal value of $F$ is attained.

Note that if $p = 0$ when $F$ is minimized, then $F = G$. Since $G = (2/3)\mathbf{c}^T\mathbf{S}\mathbf{c}$ and $\mathbf{S}$ is positive definite, then $G$ is minimized when $\mathbf{c} = \mathbf{0}$. This in turn implies $\mathbf{d} = \mathbf{0}$, so that the second- and third-order coefficients vanish, resulting in a piecewise linear minimizing spline. The properties of being piecewise linear and having a continuous first derivative together imply that the minimizing spline is a straight line. Furthermore, periodicity of the spline implies that the straight line is in fact horizontal.

On the other hand, if $p > 0$ when $F$ is minimized, then $z = 0$, so that $H = M$. Since $\mathbf{a} = \mathbf{y} - (2/p)\mathbf{W}^2\mathbf{Q}\mathbf{c}$ from (14), and $\mathbf{c} = 3p\mathbf{A}_p^{-1}\mathbf{Q}\mathbf{y}$ from (15), $H$ can be expressed as a function of $p$. Thus, if minimization of $F$ occurs for a positive value of $p$, it remains to determine the value of $p$ for which $H(p) = M$.

## VIII. PROPERTIES OF *H* AS A FUNCTION OF *p*

The following facts can be established: for all positive values of $p$, $H(p)$ is a continuous, convex function of $p$ with negative slope. Furthermore, as $p$ approaches zero from the right, $H(p)$ becomes arbitrarily large.

## IX. ALGORITHM FOR DETERMINING SPLINE COEFFICIENTS

We can now state the following algorithm for determining the minimizing spline. Compute the equation of the horizontal straight line with the least-squares fit to the given data points:

$$f(x) = \left(\sum_{k=1}^{n} \frac{y_k}{w_k^2}\right) \Big/ \left(\sum_{k=1}^{n} \frac{1}{w_k^2}\right).$$

Determine if this line satisfies the distance constraint on $H(f)$. If it does, we are done. If it does not satisfy the distance constraint, start with some positive value of $p$ and search for the value of $p$ for which $H(p) = M$, using a combination of Newton's method when moving to the right and a binary search when moving to the left (or any applicable

search). Insert this value of $p$ in (15), solve the linear system for $\mathbf{c}$, and compute the related values of $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{d}$ using (14), (11), and (12). The periodic cubic spline associated with this value of $p$ will be our solution.

## X. PRACTICAL CONSIDERATIONS IN SOLVING THE LINEAR SYSTEM

Since the matrix $\mathbf{A}_p$ is positive definite for each $p > 0$, it can be decomposed into the product of a lower triangular matrix $\mathbf{R}$ and its transpose, using the square-root (Cholesky's) method.[5] The linear system $\mathbf{A}_p\mathbf{c} = 3p\mathbf{Q}\mathbf{y}$ can then be solved efficiently in two steps, by applying forward substitution to the lower triangular system $\mathbf{R}\mathbf{v} = 3p\mathbf{Q}\mathbf{y}$, followed by backward substitution to the upper triangular system $\mathbf{R}^T\mathbf{c} = \mathbf{v}$.

Furthermore, since $\mathbf{A}_p$ is symmetric and five-banded with three non-zero entries in two corners, its decomposition $\mathbf{R}$ will consist of three non-zero bands (the main diagonal and the two diagonals below it) and two non-zero rows along the bottom, so that $\mathbf{R}$ can be stored in fewer than $5n$ locations. (The entries of $\mathbf{A}_p$ need not be stored; they may be computed as needed.)

The sparseness of the matrix $\mathbf{R}$ and its structure described above lead not only to its compact storage, but also to the linear time solution of the upper and lower triangular systems, and hence to the linear time solution of the system $\mathbf{A}_p\mathbf{c} = 3p\mathbf{Q}\mathbf{y}$, for each non-zero value of $p$.

It should be pointed out that, unless the number of data points to be smoothed is rather limited (approximately 30 or fewer), the straightforward application of Cholesky's method to decompose $\mathbf{A}_p$ will encounter underflow problems. This is due to the fact that, as the dimension of $\mathbf{A}_p$ increases, entries with exponentially decreasing magnitudes will appear in its decomposition $\mathbf{R}$. This difficulty can be circumvented by truncating sufficiently small values to zero, while still retaining single-precision accuracy in the solution of the triangular systems. (Truncation has the additional advantage of significantly reducing the number of arithmetic operations required in computing the entries of $\mathbf{R}$ when the number of data points is large.)

The program implementing this algorithm has been tested on an IBM-370 computer using single-precision arithmetic, and has successfully smoothed up to 250 points before encountering detectable round-off errors. On the average, the Newton and binary search converged after six to eight iterations, independent of the number of data points.

## XI. SAMPLE PLOTS

The data points in Figs. 1 and 2 were generated by adding random noise to an ellipse. The dotted curves represent cubic spline interpo-

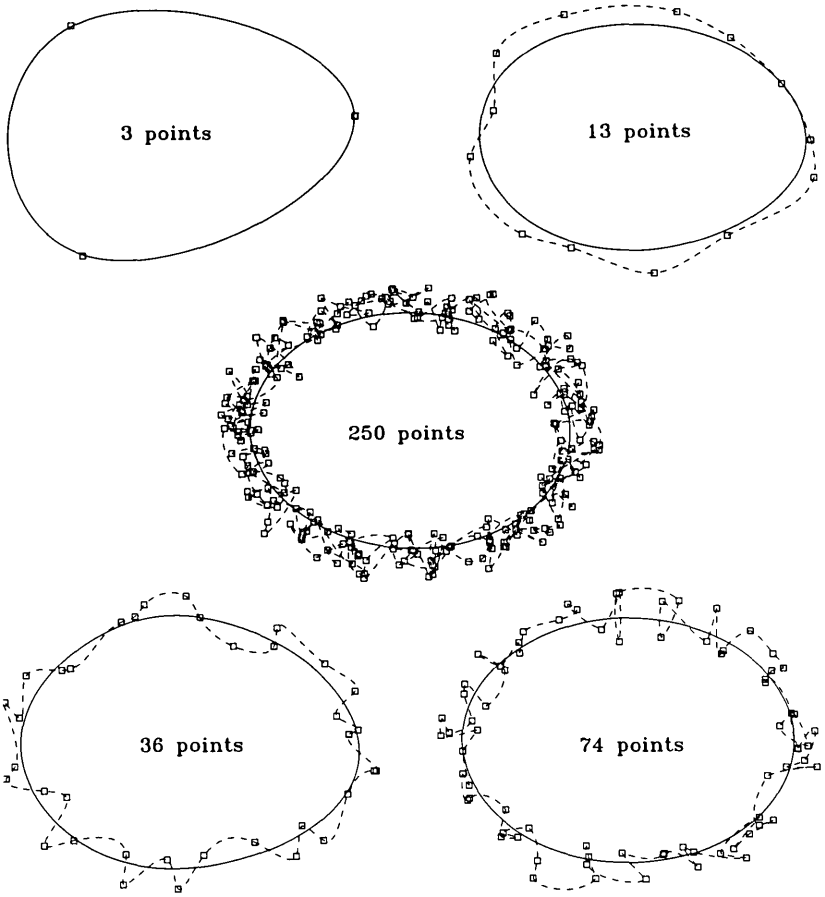



Fig. 2—Periodic cubic spline smoothing for a varying number of data points with uniform weights.

lation of the data (using periodic splines), while the solid curves represent cubic spline smoothing of the same data. In each case a parameter was introduced so that the curve could be represented as two separate single-valued functions of the parameter. Then smoothing was performed twice, once with the $x$ values as a function of the parameter and then again with the $y$ values as a function of the parameter. The smoothed $x$ and smoothed $y$ values were then plotted against the parameter to produce the closed curve. Figure 2 illustrates the algorithm with a varying number of data points, from only three distinct points to 250 points. In each case, uniform weights were used. A "tight" fit was chosen in the example with three points to show how the method can be used to simulate periodic interpolation of the points.

## REFERENCES


1. Carl de Boor, *A Practical Guide to Splines*, New York: Springer-Verlag, 1978.
2. Christian H. Reinsch, "Smoothing by Spline Functions," Numerische Mathematik, *10* (1967), pp. 177–83.
3. H. Spath, "Spline Algorithms for Curves and Surfaces," Winnepeg: Utilitas Mathematica Publishing Inc., 1974.
4. David I. Steinberg, *Computational Matrix Algebra*, New York: McGraw-Hill, Inc., 1974, p. 251.
5. V. N. Faddeeva, *Computational Methods of Linear Algebra*, New York: Dover Publications, Inc., 1959.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# Hybrid-Mode, Shielded, Offset Parabolic Antenna

By R. A. SEMPLAK

(Manuscript received September 24, 1981)

*In this paper measurements are presented for both a hybrid-mode feed having corrugations in two opposite walls and a modified scale-model pyramidal horn-reflector antenna using this feed. Comparisons of the measured and theoretical data for the hybrid-mode shielded offset parabolic antenna with the theoretical data for the standard pyramidal horn-reflector antenna show that a net improvement can be obtained in sidelobe level of about 4 dB in the transverse polarization in the transverse plane and 4 to 13 dB in the longitudinal polarization in the longitudinal plane.*

## I. INTRODUCTION

In most respects, the standard pyramidal horn-reflector is an excellent antenna. It is a combination of a square pyramidal horn and a reflector that is a section of a paraboloid of revolution whose focus coincides with the apex of the square horn. Its geometry provides a shielded offset parabolic antenna that is broadband with good front-to-back discrimination and good return loss. However, inherent in the design of the pyramidal horn-reflector antenna is a problem that results from illuminating the reflector with a dominant waveguide mode.[1] The theoretically obtainable off-axis radiation levels for transverse polarization in the transverse plane and longitudinal polarization in the longitudinal plane are considerably higher than those obtained for longitudinal polarization in the transverse plane and transverse polarization in the longitudinal plane, i.e., the former are essentially the equivalent of an aperture with constant illumination, whereas the latter aperture field distributions are tapered to zero at the edges.

In the discussion of the horn-reflector antenna, it should be remembered that longitudinal polarization and longitudinal plane indicate that the electric field in the aperture and the plane of antenna rotation (for radiation measurements) are aligned with the pyramidal horn axis,

whereas transverse polarizations and transverse plane indicate that the electric field in the aperture and plane of antenna rotation are perpendicular to the horn axis.

As used in microwave radio-relay systems, the horn-reflector antenna is mounted with the axis of the horn normal to the earth's surface. Hence, longitudinal and transverse polarizations could be called vertical and horizontal, respectively. However, the aperture field distribution for each polarization is different. Moreover, when the antenna is used as an earth station antenna for satellite communications, or as a radiometer, or simply to obtain radiation characteristics in the longitudinal plane, the antenna is mounted on its side, i.e., the longitudinal axis of the horn is parallel with the earth's surface, and aperture field distributions for so-called vertical and horizontal polarizations are now interchanged. To avoid this ambiguity, longitudinal and transverse polarizations are referred to the axes of the horn.

Radio interference from adjacent paths limits the number of converging routes of a common-carrier microwave radio system. In recent years, demands have been made to improve the sidelobe performance of the pyramidal horn-reflector antenna. The use of blinders[2] (extensions to the side walls of the antenna aperture) provides a degree of far sidelobe reduction, i.e., lobes beyond 35 degrees from the axis of the main beam. A method now exists for eliminating the troublesome reflections from the flat weather cover of the horn-reflector antenna by using a focused weather cover.[3]

As used in terrestrial microwave radio systems, the horn-reflector antenna is mounted with the axis of the pyramidal horn normal to the earth's surface. Therefore, to address the problem of obtaining improvement in the sidelobe structure, primary consideration should be given to those sidelobes produced by transverse polarization in the transverse plane. Hence, one needs to consider those special feeds[4,5] that would decrease the field intensity at the two side walls.

In Section II a discussion of the design and radiation measurements of a 30-GHz, two-wall, hybrid-mode feed is presented; in Section III the radiation measurements obtained in both the transverse and longitudinal planes when this feed is used on a 30-GHz, scale-model, horn-reflector antenna are discussed;[6] and in Section IV the conclusion is made that the antenna as modified is no longer the familiar horn-reflector antenna, but a shielded, hybrid-mode-fed, offset paraboloid.

## II. TWO-WALL HYBRID-MODE FEED

The hybrid-mode feed fabricated for this experiment was designed at a frequency of 30 GHz as a 6-inch-square aperture with two opposite walls corrugated. The design of the corrugations at this frequency requires a tooth and groove width of 0.020 inches (0.508 mm) and a

tooth height or groove depth of 0.0997 inches (2.532 mm). These dimensions present a formidable machining task. Hence, other methods for fabrication were considered. As shown in Fig. 1, the approach used for this feed consisted of stacking alternating 0.020-inch- (0.508-mm) thick strips of brass and aluminum, with the height of the brass strips twice that of the aluminum strips. The assembly of alternating strips was electroformed and then machined to obtain the proper tooth height. The aluminum was chemically removed to provide the grooves. The rest of the fabrication proceeded along normal lines.

The measured vertical and horizontal polarization radiation characteristics for the two-wall hybrid-mode feed are shown in Figs. 2 and 3, respectively. The position of the two corrugated walls with respect to the electric field is indicated on the insert in each figure. For vertical polarization (Fig. 2), the corrugated walls are parallel to the electric field, whereas for horizontal polarization (Fig. 3) the two corrugated walls are normal to the electric field. An examination of these two figures indicates that, except for a few minor differences, the two patterns are essentially equal and symmetric. The dashed lines on these two figures indicate the respective cross-polarized response of the feed. From Fig. 3, we can conclude that the two-wall hybrid-mode feed provides a tapered field distribution for the transverse polarization in the transverse plane of the horn-reflector antenna. By rotating the hybrid-mode feed 90 degrees the same tapered field distribution can be provided for the longitudinal polarization in the longitudinal plane.

## III. MODIFIED HORN-REFLECTOR ANTENNA

### *3.1 Transverse polarization, transverse plane*

Using the radio range facilities at Holmdel, New Jersey, measurements were made of the far-field radiation characteristics of a 30-GHz,

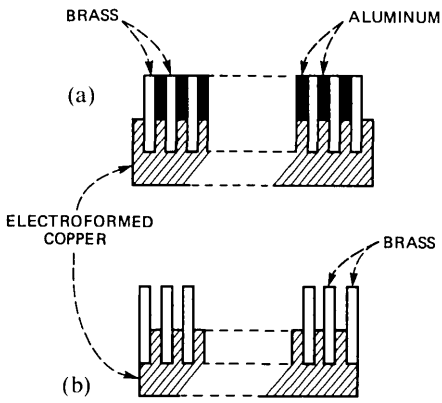



Fig. 1—(a) An assembly of alternating brass and aluminum strips used to achieve a corrugated wall. (b) The finished wall.

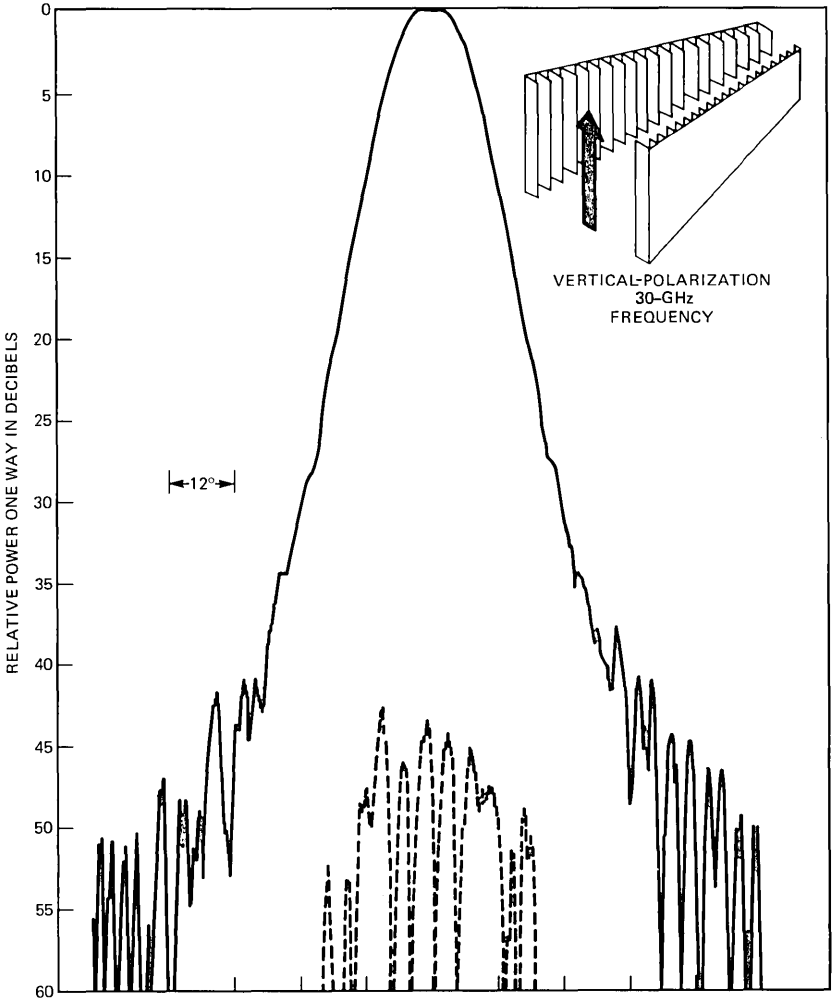



Fig. 2—The polarized radiation characteristics of the two-wall corrugated feed measured vertically are shown by the solid curve. The dashed curve shows the cross-polarized response. As depicted in the insert, the corrugated walls are parallel to the electric field.

scale-model, horn-reflector antenna[6] illuminated by the two-wall hybrid-mode feed discussed above. The solid line of Fig. 4 is the far-field radiation characteristic of this antenna for transverse polarization in the transverse plane. Recall that the two corrugated walls of the feed are normal to the transverse field. In Fig. 4, the dashed curve indicates the calculated envelope of peaks for the hybrid-mode antenna with ideal dual-mode excitation. Comparison of these two curves reveals good agreement over the main beam and first sidelobes but increasing disagreement for the sidelobes that are further out.

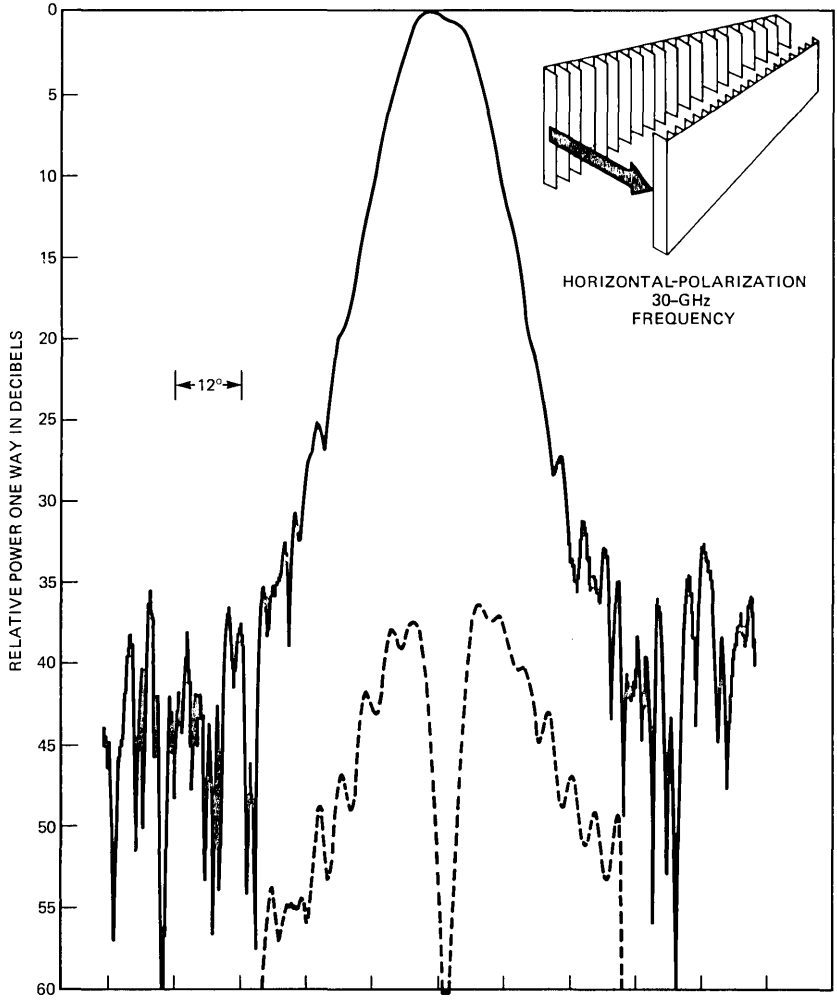



Fig. 3—The polarized radiation characteristics of the two-wall corrugated feed measured horizontally are shown by the solid curve. The dashed curve shows the cross-polarized response. As depicted in the insert, the corrugated walls are normal to the electric field.

A more extensive comparison is presented in Fig. 5. In this figure, the curves represent the characteristics obtained for transverse polarization in the transverse plane for the following: the broken curve is the theoretical response for the standard pyramidal horn-reflector antenna;[6] the dashed curve represents the theoretical response for the hybrid-mode antenna with ideal dual-mode excitation; and the solid curve is the measured data for the hybrid-mode antenna. An examination of the two theoretical curves indicates the possible improvement

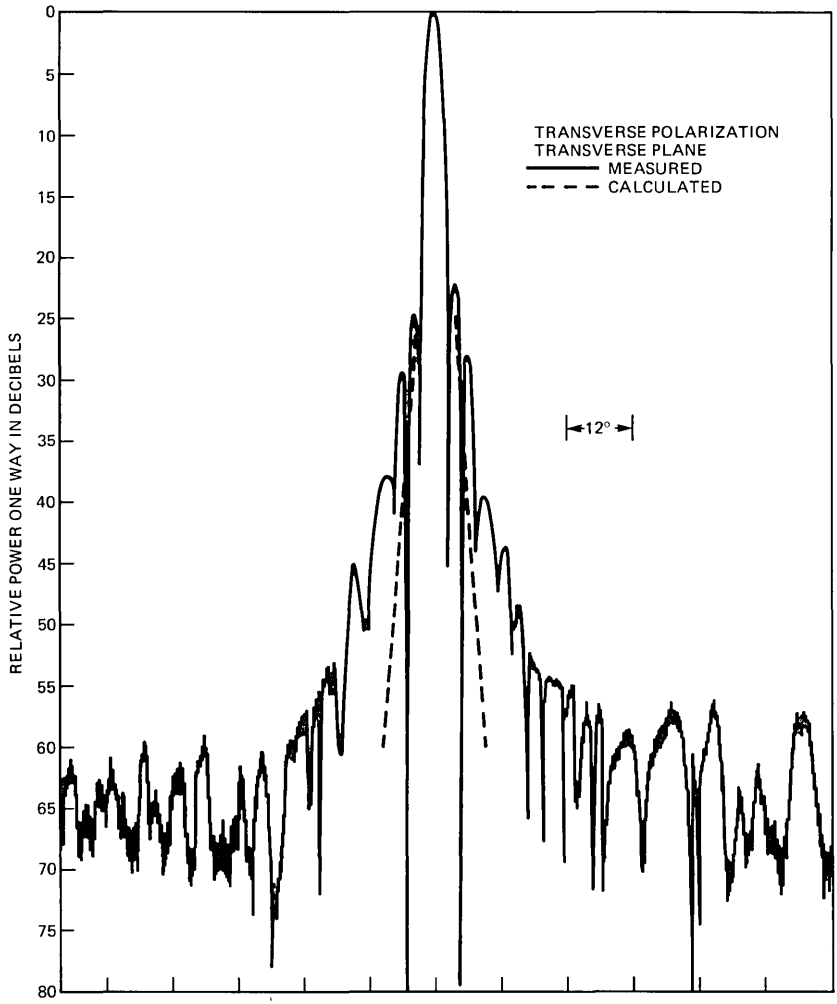



Fig. 4—The measured (solid curve) and calculated (dashed curve) radiation characteristics of the modified horn-reflector antenna for transverse polarization in the transverse plane.

in sidelobe structure that could be obtained by using the hybrid-mode feed instead of the standard horn feed of the standard horn-reflector. From comparison to the measured response (solid curve) one can see the actual improvement over that of the standard pyramidal horn-reflector antenna. However, the measured response is not quite as low as predicted by theory. The observed departure leads one to suspect that the two-wall corrugated structure used in these experiments may be flawed. By examining Figs. 2 and 3 more carefully, one observes regular small undulations in the radiation characteristics of the two-

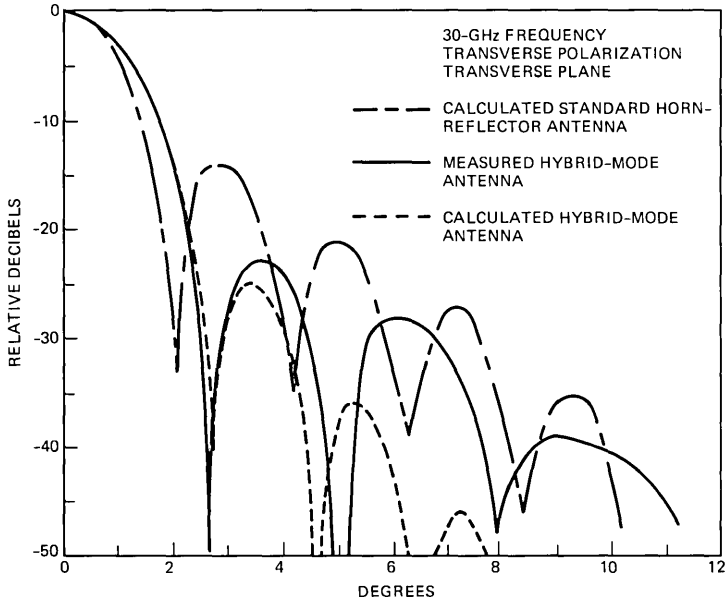



Fig. 5—Comparison of the measured response of the modified horn-reflector antenna (solid curve) with the theoretical response of the unmodified standard horn-reflector antenna (broken curve) and the theoretical response of the modified antenna (dashed curve).

wall corrugated feed that could indeed be indicative of the presence of modes other than the desired $HE_{11}$ mode.

A comparison of the hybride-mode antenna measurements made for longitudinal polarization in the transverse plane with those of Ref. 6 indicate only minor differences and therefore are not included here.

### *3.2 Longitudinal polarization, longitudinal plane*

To examine the possibility of improving radiation characteristics in the longitudinal plane, the hybrid-mode antenna is placed on the antenna positioner with its longitudinal axis parallel to the earth's surface. The two-wall hybrid-mode feed is rotated 90 degrees so that the corrugations are now on the front and back walls of the pyramidal portion of the antenna, i.e., the corrugations are now normal to the longitudinal field. The measured data obtained for this configuration are shown by the solid line in Fig. 6. The dashed line in this figure represents the calculated theoretical response for the hybrid-mode antenna with the ideal dual-mode excitation. Comparison of these two curves indicates good agreement.

A more extensive comparison is presented in Fig. 7. In this figure, the set of curves represents the characteristics obtained for longitudinal polarization in the longitudinal plane for the following: the broken

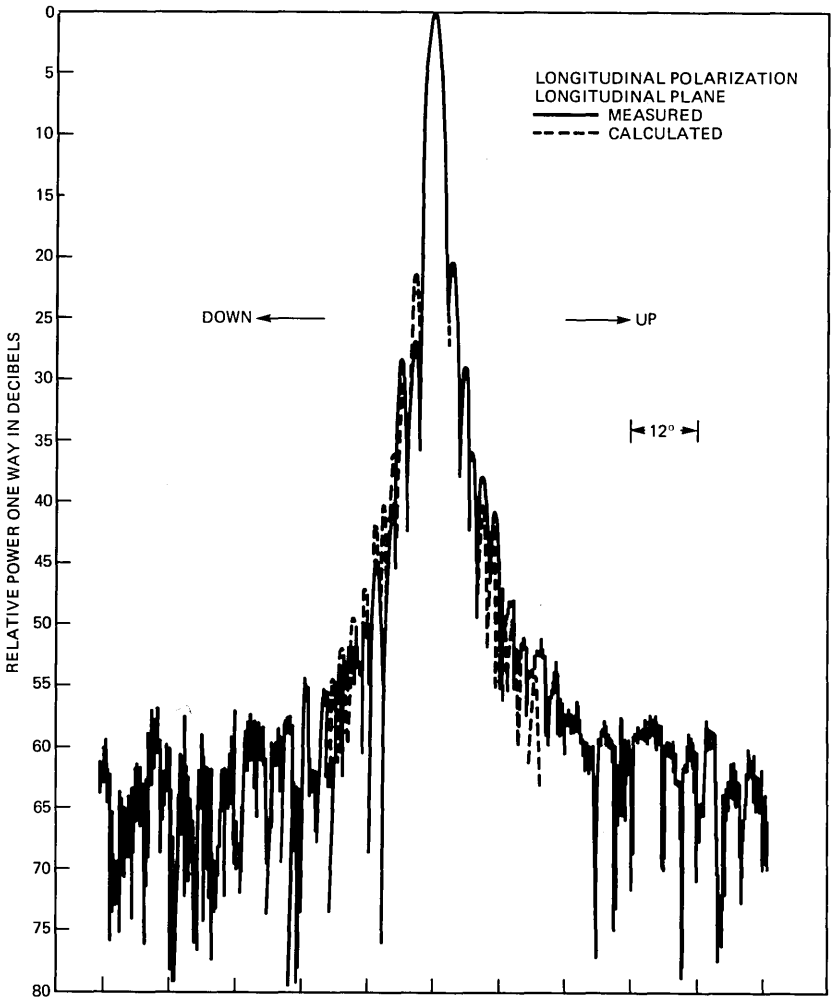



Fig. 6—The measured (solid curve) and calculated (dashed curve) radiation characteristics of the modified horn-reflector antenna for longitudinal polarization in the longitudinal plane.

curve is the theoretical response for the standard pyramidal horn-reflector antenna;[6] the dashed curve represents the theoretical response for the hybrid-mode antenna with ideal dual-mode excitation; and the solid curve is the measured data for the hybrid-mode antenna. Comparison of the two theoretical curves indicates the possible reduction in sidelobes that could be obtained by using a hybrid-mode-fed antenna. From the measured response (solid curve) one can observe the actual improvement over that of a standard pyramidal horn-reflector antenna and the good agreement with the theoretical values

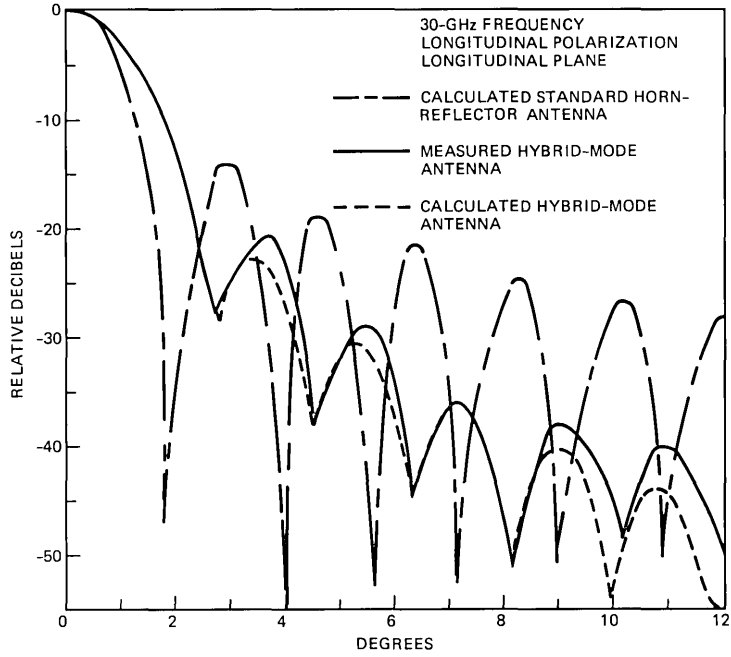



Fig. 7—Comparisons of the measured response of the modified horn-reflector antenna (solid curve) with the theoretical response of the unmodified standard horn-reflector antenna (broken curve) and the theoretical response of the modified antenna (dashed curve).

for the hybrid-mode antenna. The agreement here is better than that shown in Fig. 5, but one should recall that the pyramidal sidewalls contribute more to sidelobe structure than do the front and back walls. However, when compared to the transverse plane, the sidelobe structure in the longitudinal plane is, in general, higher. This can be attributed to the nearly parallel top and bottom edges of the antenna aperture.

A comparison of the hybrid-mode antenna radiation characteristics for transverse polarization in the longitudinal plane with those of Ref. 6 indicate only minor differences and therefore are not included here.

One should recall that, at present, the bandwidth of corrugated feeds is limited to less than two-to-one. This bandwidth limitation would tend to restrict the application of a two-wall corrugated feed to those system antennas where only two frequencies, i.e., 4 and 6 GHz, or 6 and 11 GHz were used. It should be noted that a two-wall hybrid-mode feed as discussed here will only reduce sidelobes in one plane; therefore one has the option when modifying existing antennas of applying the correction to that particular plane where interference in the near-in sidelobes is a problem. Of course, the use of a four-wall hybrid-mode feed reduces the sidelobes in both planes.

## IV. CONCLUSIONS

Measurements for both a two-wall hybrid-mode feed and a modified scale-model horn-reflector using the hybrid-mode feed (a hybrid-mode, shielded, offset parabolic antenna) are presented and discussed. The high-sidelobe characteristics of a standard pyramidal horn-reflector antenna displayed in the transverse plane for transverse polarization and in the longitudinal plane for longitudinal polarization can be improved by modifying the antenna with a two-wall or four-wall hybrid-mode feed.

## REFERENCES


1. A. B. Crawford, D. C. Hogg, and L. E. Hunt, "A Horn-Reflector Antenna for Space Communications," B.S.T.J., *40,* No. 4 (July 1961), pp. 1095–116.
2. D. T. Thomas, "Analysis and Design of Elementary Blinders for Larger Horn-Reflector Antennas," B.S.T.J., *50,* No. 9 (November 1971), pp. 2979–95.
3. R. A. Semplak, "Horn-Reflector Antenna—Eliminating Weather-Cover Reflections," B.S.T.J., *59,* No. 8 (October 1980), pp. 1333–42.
4. C. Dragone, "Attenuation and Radiation Characteristics of $HE_{11}$-Mode," IEEE Trans. Microwave Theory and Techniques, *MTT-28,* No. 7 (July 1980), pp. 704–10.
5. C. Dragone, "High-Frequency Behavior of Waveguides with Finite Surface Impedances," B.S.T.J., *60,* No. 1 (January 1981), pp. 89–116.
6. R. A. Semplak, "A 30-GHz Scale-Model, Pyramidal, Horn-Reflector Antenna," B.S.T.J., *58,* No. 6 (July–August 1979), pp. 1551–6.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# An Experimental Teleterminal—The Software Strategy

By D. L. BAYER and R. A. THOMPSON

(Manuscript received September 29, 1982)

*A research model of a synergism between a telephone and a computer terminal is described in this paper and its companion. The hardware design of this "teleterminal," described in the companion paper, includes an internal microprocessor and a data connection to a host computer. This paper describes the software in these two machines. The software resident in the teleterminal's internal microprocessor addresses internal issues like cursor control, tab expansion, and data modes. The software resident in the host controls screen content and menu selection. The teleterminal is a research tool for the investigation of the human interface for the access to experimental services. Such services include calling by name, directory retrieval, a variety of information (e.g., yellow pages, department store catalogs, community bulletin boards, newspapers, and libraries) and a variety of personal services such as mail, personal calendar, entertainment, and shopping. The user-interface is a conceptual tree-like structure that can be customized by the user.*

## I. INTRODUCTION

A *teleterminal* is defined to be a piece of equipment that merges the functionality of the traditional *tele*phone with that of a computer *terminal.* It is characterized by:

(*i*) A traditional telephone facility
(*ii*) Internal intelligence
(*iii*) A data communication facility
(*iv*) A general-purpose display
(*v*) Dynamic labeling of buttons (soft keys).

This merger of functionality has been seen to be synergistic. One such teleterminal, whose hardware is described in Ref. 1 and in the companion paper,[2] is illustrated in Fig. 1.

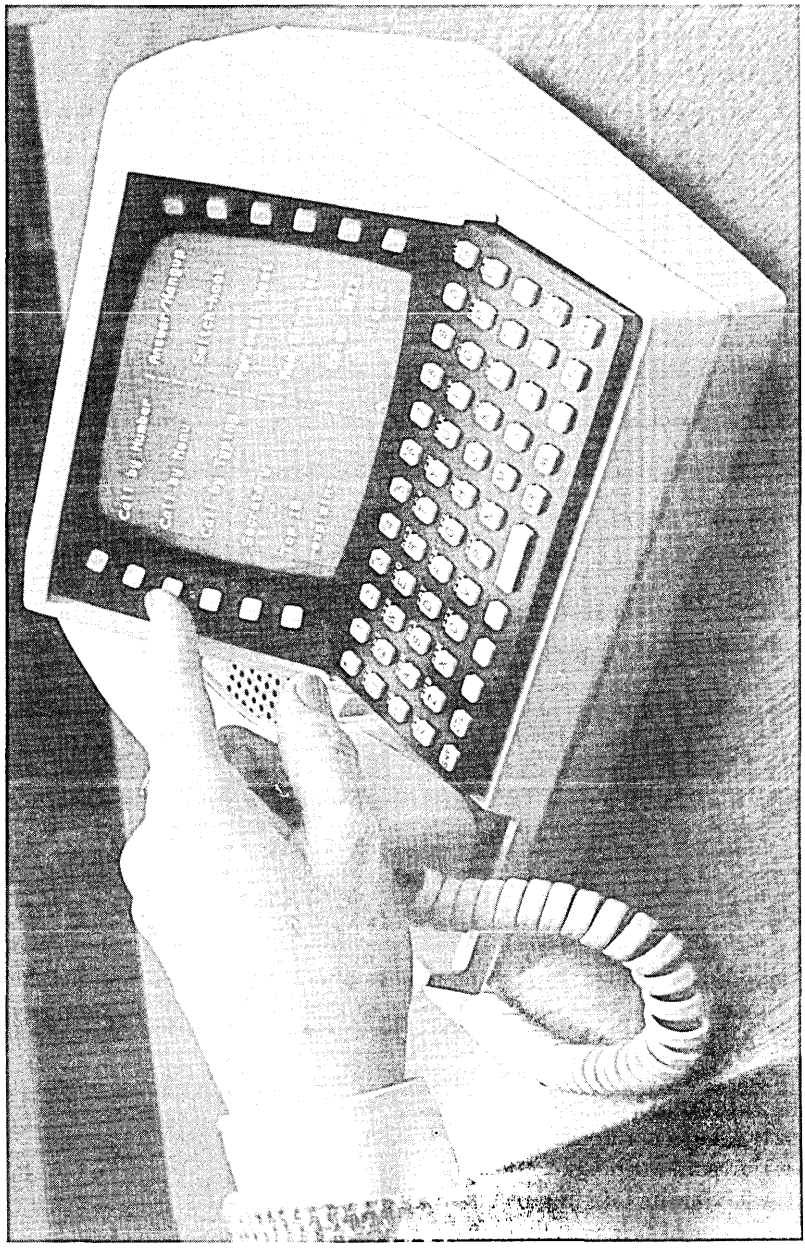


Fig. 1—A teleterminal.

The construction of and, more significantly, the experimentation with such a teleterminal is part of a large-scale, long-range research investigation into the systems, software, and applications aspects of telecommunications. Cognitive and social implications are a significant part of the study. An evolving futuristic "test-bed" environment[3] currently consists of a highly reliable host computer (a three-processor Tandem-16), an "intelligent" digital switching office,[4] and a collection of teleterminals. A number of systems-level principles have been uncovered in the course of this research and appropriate papers have been submitted to and published in appropriate journals. It is the purpose of this paper and its companion to describe the equipment used in this research and thereby serve as a common reference for those other papers of a more general nature.

Compatibility with today's environment would be of major importance if a real product were being described, but it is of little consequence in a research environment, except that it simplifies the dissemination of teleterminals to interested people. There are about three dozen of these teleterminals in an active user community. It must be emphasized that the teleterminal is *not* a proposed product; it is a research tool. Design decisions, and discussions in this paper about it, are based on this important point. If a product were anticipated, many decisions would have been made differently and the reasons discussed would change dramatically. Furthermore, neither this research nor this paper is intended to describe nor imply the future direction of the Bell System as it relates to advanced communications systems and services, nor is it intended to endorse any hardware or software offerings by any vendor.

The next section of this paper presents the research objectives of the project that have an impact on software. They include assumptions about the system environment and the impact on the user. In Section III, intelligence distribution is discussed and the internal and host programs are described. In Section IV, a scenario is presented to illustrate program execution, the user interface, and the kinds of capabilities provided. Section V contains a discussion of lessons learned and future directions.

## II. OBJECTIVES

Objectives that have software implications are emphasized. These are presented from the viewpoint of both the system and the user.

### *2.1 System environment*

In the current system configuration, each teleterminal has a standard terminal connection to a host computer and a standard telephone connection to a telephone office. The only assumed commonality

occurs at the teleterminal itself. Compatibility with the "POTS" (plain old telephone service) world requires that old-fashioned telephone hardware like the switchhook and *Touch-Tone** dialing equipment be necessary parts of the teleterminal.

Such a system configuration impacts software in that this required POTS hardware reduces available space for microprocessor memory inside the teleterminal and that special microprocessor software is needed to interact with this POTS hardware. Furthermore, the host software must transmit phone numbers to the teleterminal (with a special data mode) instead of directly to the telephone office controller and it must be concerned with such POTS functions as the switchhook state and dial tone. "POTS" functions, like "busy" or "network blocking", could be recognized by including certain tone-detecting filter circuits in the hardware, but to conserve space this was not done. As a consequence, the software is limited to assuming the presence of dial tone after a fixed delay and the software can make no branches on whether or not a call completes.

A second system aspect is our desire to use the teleterminals in a "test-bed" environment for the ongoing investigation of the human interface and the study of experimental services. This suggests that changes will be made to the software frequently and reflection of this in the initial design of the software architecture was felt to be wise. A distribution of intelligence was selected wherein the functional software is located in a centralized general-purpose host computer. Furthermore, this "access program" can be structured to simplify anticipated changes. Such a "host-centered" distribution of intelligence may not be ideal for a marketed product where the design goals would be different and the interface would be more stable. These kinds of items are discussed in Section V.

### 2.2 User impact

After system environment, the second area of objectives with software ramifications is the impact on the user. There are two such impacts: the physical cosmetics of the teleterminal and the conceptual interface to the software. Cosmetic features that affect software are the small size, the incorporation of both telephone and terminal characteristics, and the inclusion of function keys.

For the experiments with new applications, the teleterminal replaces a conventional business telephone. It is important that the set occupy approximately the same desk space as a conventional business telephone. Furthermore, the set is "real" in the sense that all electronics, except power supplies, reside within the set. This results in a limited

* Registered service mark of AT&T.

physical space for internal memory and thus restricts the size of the internal software. The placement of the function keys adjacent to the cathode ray tube (CRT) screen has major impact on the software. At the "low" level, internally executed "primitives" are required for labeling buttons and sensing button-pushes. At the "high" level, these buttons make possible the use of a tree-structured user interface to access various functions.[5]

From the user view, three general requirements of this access method are simplicity, convenience, and customizability. The interface must be as conceptually simple as possible, even to the most casual, unsophisticated user. The required convenience of the interface is attained by enhancing the access to often-used functions and the names of often-called people and by providing a "translation" capability between the user's name for some function and the "telephonese" or "computerese" for accessing it. Whether customization is done by the vendor or by the user, the need for real simplicity in the underlying data structure, and not just apparent simplicity in the user's view of it, is reinforced.

The structure of the user interface is tree-like in a deliberate attempt at "congruency"[5] with simple models of human cognition.[6] Similar "access trees" have appeared in the literature. References 7 and 8 are representative. In the marketplace, Hewlett-Packard's new line of terminals and the British PRESTEL are examples. In the mathematical sense, a "tree" is an acyclic, connected graph;[9] but the concept of tree to be presented is more like the "lay" notion of a large woody plant.

Informally, the parts of a tree include the root, branches, intermediate nodes, and leaves. Let a screen of button labels correspond to a node. Let the dynamic label of a button correspond to a branch or leaf, depending on whether its selection causes a traversal to a new node or the actuation of some function, respectively. The structure is not a mathematical tree because multiple branches to the same node are permitted, as are "backup" and "restart" branches (which make the "graph" cyclic). Analogous to the root, or initial node, of the tree is an initial labeling of the buttons by which the user perceives the "entering" of his structure of functions and directories. The root of an example access tree is illustrated in Fig. 2.

The organization of this access tree is intended to be defined or selected by the user. A *functional* organization would show a root menu with branches like **Telephone Functions,*** **Computer Functions**, **Calendar**, **Mail**, etc. A utilitarian organization would attempt to place highly used functions near the root and seldom used functions

* Labels identifying function keys are set in boldface type.

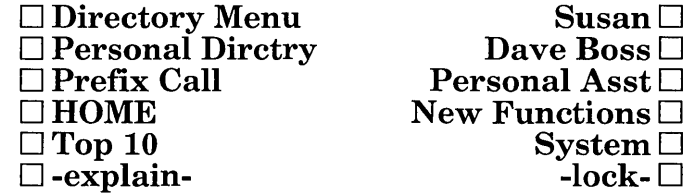



Fig. 2—A sample access tree.

far from the root. Some compromise of the two organizations is probably most appropriate, depending on the user's level of sophistication and the frequency of use.

## III. INTELLIGENCE DISTRIBUTION

The user interface to system capabilities is determined by the properties of the terminal and the network resources that the user can access through the terminal. The capabilities of the terminal are constrained by the computing power of the internal processor and by the data switching capabilities of the communications system. Currently available microprocessors span at least an order of magnitude in speed and several orders of magnitude in memory size. In the case of the teleterminal, power requirements and chip count were important considerations in the selection of the internal microprocessor. Many of the design considerations described in the previous section constrained the type of internal computations and thus focused our efforts on software issues relating to telephone applications vis-a-vis general-purpose word-processing systems.

In the next subsection, the capabilities of the processor within the teleterminal are presented. A list of functions performed by the instrument is given and the impact of this processing power on feature software is discussed. The limitations of, and alternatives to, the current design are presented in Section V.

### *3.1 Internal processing*

The teleterminal contains an Intel 8748 microprocessor with 2048 bytes of programmable read-only memory (PROM) for program storage and 288 bytes of random access memory (RAM) for data storage. The processor executes instructions in from four to eight microseconds. A common operation such as moving a byte from one location to another in RAM requires four instructions, occupies six bytes of program memory, and takes 32 microseconds. A single vectored interrupt is available for performing time-critical functions.

The internal program is written in assembly language and occupies most of the available program memory. The function of the internal program is to control the microprocessor's peripherals: a Matrox

display processor, a tone generator with muting relay, an ASCII keyboard with clicker relay, and a universal asynchronous receiver/transmitter (UART) for interface to the host computer.

The Matrox display processor cycles through a 512-byte dual-ported buffer memory displaying its contents as 16 rows of 32 characters on a CRT. In addition to character generation, the display processor provides character blinking at a half-hertz rate. The internal processor manages the contents of the buffer memory. Scrolling is implemented by moving the entire contents of the display buffer. A cursor is implemented by displaying an underscore character (_) when a space character lies at the cursor coordinates or by blinking the character at the cursor coordinates (inverse video would be more attractive but is not available in the current hardware). Button labels are cleared, positioned, and formatted by the internal processor.

The tone generator is used to dial over the telephone line associated with the teleterminal. The state of the switchhook is examined to determine if the receiver is off-hook. After a short delay to allow for dial tone, digits are dialed by activating the muting relay, activating the tone generator for approximately 100 milliseconds, then deactivating the muting relay, and waiting approximately 60 milliseconds before repeating.

The keyboard peripheral consists of three types of keys: standard keys that correspond to ASCII codes, special control characters within the control code set that correspond to the buttons adjacent to the screen, and polled control buttons. The first two types of keys are scanned, encoded into ASCII character codes, and latched by the keyboard circuitry.[2] The processor examines the latch approximately 60 times per second. When a character has been latched by the hardware, the processor reads the character, clears the latch, and activates the clicker relay to give the user audio feedback. Normal character codes are delivered to the output UART for transmission to the host. For codes corresponding to screen buttons, the processor emits a unique three-character sequence for each of the buttons.

Control buttons are wired directly to individual bits on one of the processor's input/output (I/O) ports. After debouncing, the buttons are used to (also see Section 3.3):

(*i*) Freeze the screen

(*ii*) Emit a special host-specified "delete previous character" code

(*iii*) Emit a special host-specified "delete current line" code

(*iv*) Emit a special host-specified "interrupt" code. The internal buffer of input characters from the host is also purged.

The input characters from the UART (data from the host) are read at interrupt level and collected in a 255-byte circular buffer. In full-duplex operation, flow control is implemented by emitting a single

symbol when the input buffer approaches overflow. This symbol can be set by the host with a control message. The characters in the circular buffer are processed by the base-level part of the control program. Data is separated into three classes: control messages, response messages, and ordinary text. Control messages change the "state" of the set and determine how ordinary text is to be processed. The following "states" exist:

(*i*) Label button—Data is positioned adjacent to the button, left or right justified, and constrained to fifteen characters per line.

(*ii*) No scroll—Data is constrained to the bottom line of the screen.

(*iii*) Dial—Data is interpreted as dial control: that is, a digit to dial, a request for dial delay (usually for second dial tone), a request to time and display call duration, or a request for current call status.

(*iv*) Terminal—Data is displayed much the same as on any simple CRT terminal.

Response messages are character strings generated by the teleterminal in response to either a button push or a host request for status. These messages appear in the teleterminal's input stream because the host computer, in full-duplex mode, echos all characters transmitted from the set. Status response messages are ignored by the set. Button-push response messages cause the set to start flashing the first character of the associated button label, providing a feedback cue equivalent to echoing ordinary characters.

In summary, the internal processing capabilities are used to control the CRT display and perform dialing functions. Support for telephone-related applications is manifested in functions that dial, label buttons, flash-button labels, and restrict output to the bottom line on the CRT.

### 3.2 Host processing

The description of the host software is contained in the next three subsections. First the host's software environment is discussed, then the structure of the user's data is defined, and finally the "access program" is described.

#### 3.2.1 Host software environment

At the system level, the host software environment is that of a transaction-oriented, time-sharing, general-purpose computer. The current implementation, under the *UNIX** operating system,[10] provides a convenient machinery for interfacing the access program not only to the standard commands, system calls, and supported programs of the *UNIX* operating system, such as those providing computer mail,

* Trademark of Bell Laboratories.

calendar functions, and various games, but also to current and future specialized application programs, such as have been written for mail access, call-memos, and personal calendar maintenance. The access program and the peripheral programs have been written in the C programming language[11] by a multiplicity of authors with program interfaces provided by the standard system calls and interprocess messages of the *UNIX* operating system. The complete environment resides in a Tandem computer system.

The user's tree-structured data are isolated from the access program that manages those data. This isolation is accomplished by keeping user-specific data in a separate file (per-user) and out of the program. This isolation permits the existence of a single copy of the program in a computer that supports multiple users, each with a unique and customized access structure. Furthermore, this software practice permits changes in directories or function procedures and similar modifications and customizations without requiring the recompiling of a program.

A data change to the access tree data file may be as trivial as an updated telephone number or name of an application program, or it may be purely cosmetic like renaming a button label, moving a familiar function to a new position in the tree, or placing a new label on some vacant button. Or, it may be more significant, like pruning the tree, attaching a new node to the tree, grafting an entire subtree to a new position, or adopting an entirely new access tree. Currently, these changes are made through the *UNIX* text editor to the data only and the program is not recompiled. An experimental editor with a potentially improved human interface is described in Ref. 12.

### *3.2.2 Data description*

The access tree is logically organized as a hierarchical data structure. The current implementation uses a magnetic disk medium and so the physical organization is linearized into a serial file of records. The data consist of an iteration of "record-groups" (each corresponds to a node of the tree or screen on the teleterminal), each in turn made up of ordered "records" (each corresponds to a button label). The actual data file read by the access program is obtained by "compiling" the file to be described here. A structure diagram of the data, in the sense of Michael Jackson,[13] is shown in Fig. 3.

The label is the text that appears adjacent to the corresponding button when the parent-node is active. The successor-node identifies the node to which the tree traverses when the corresponding branch is selected. The button-number is usually omitted but, if specified, is an integer between zero and eleven. The default is the "next" button in logical order. The command is executed by the program as described

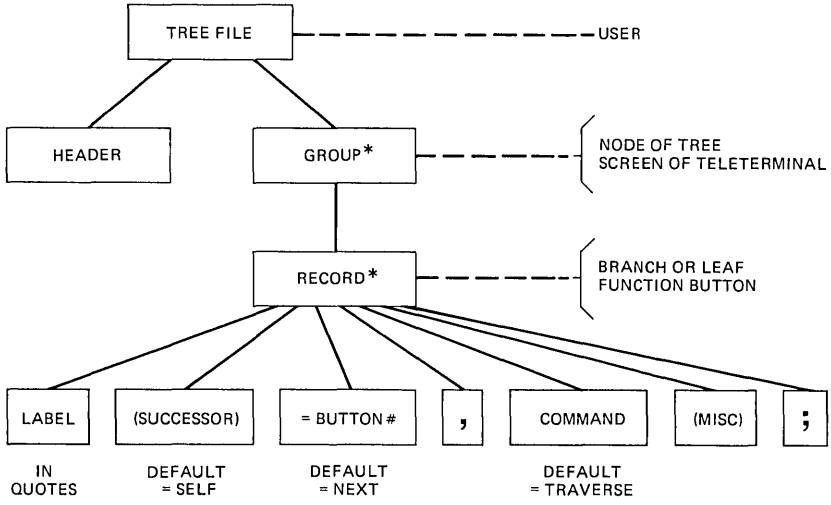



Fig. 3—Structure diagram of access tree data.

in the next subsection. The miscellaneous field contains data pertinent to the command, such as a telephone number or program name.

Data files in the format of such access trees also require "headers". Since each user is assumed to possess a personal tree file, it is appropriate to place user information in the header such as: a list of telephone numbers from which one is selected for "call-return" messages, the password for allowing access to "private" parts of the tree, and the default-assumed location used for adjusting call prefixes. The first several lines of a typical user's tree file are illustrated here in Fig. 4.

Record-groups are enclosed in brackets,"{ }", and symbolically named. The records are assigned to buttons in consecutive order within a record-group by default. The default successor-node is the current node and the default command is TRAVERSE. The correlation between a record-group and the user's perception on the CRT screen is seen by comparing the record-group named ROOT from the partial file of Fig. 4 with the "Root node" illustrated in Fig. 2. In the node corresponding to the PREROOT record-group of Fig. 4, buttons numbered one, seven, nine, and ten are left unlabeled.

From the header of the file in Fig. 4 it is seen that this tree belongs to a user whose Bell Laboratories phone number is 6170 at Murray Hill (represented by the mnemonic prefix "MH") and whose home phone number is 4649999 in dial area 201 (represented by the mne-

```
TREE heading
Numbers: murray-hill: MH6170, home: N4649999
Password: joesentme
Location: MH
%%
PREROOT
{
"introduction", CATFILE(expl/e.0,trav);
"Other User" =2; CATFILE(/get/dump/othus,exit);
"Other Phone", OTHERPHONE;
"Other Location", OTHERLOC;
"-explain-", EXPLAIN;
"open"(ROOT), CHANGELOCK;
"Bell logo" =8, CATFILE(/get/dump/bell,trav);
"-exit-" =11, CATFILE(none,exit);
}
ROOT
{
"Directory Menu" (DTORYMENU);
"Personal Dirctry" (PERSDIRY);
"Prefix Call" (PREFIX);
"HOME", CALL(N4649999,zhome);
"Top 10" (TOP10);
"-explain-", EXPLAIN;
"Susan"(PERSASST), CALL(MH4236,sst);
"Dave Boss", CALL(MH4235,db);
"Personal Asst"(PERSASST);
"New Functions"(NEWSVC);
"System", RUNSCROLL(/bin/sh,sh);
"-lock-"(PREROOT), CHANGELOCK;
}
PERSASST
{
"Today's Appnts", RUNSCROLL(/usr/lbin/caltoday,caltoday);
"Other Appnts", RUNSCROLL(/usr/lbin/calexam,calexam);
"Make Appnts", RUNSCROLL(/usr/lbin/calenter,calenter);
"Set Reminder", REMINDER;
"Time & Date", RUNBOTTOMLINE(/bin/date,date);
"2-month Cal", CAL2MONTH;
"Read Mail"(PREFIX), READMAIL;
"Send Mail", SENDMAIL;
"Send Call-Memo", SENDCALLMEMO;
"Std. Call-Memo", STDCALLMEMO;
"-backup-"(ROOT), TEMPBACKUP;
"-restart-"(ROOT);
}
DTORYMENU
{
"Emergency"(EMERGDTORY);
  •
  •
  •
```

Fig. 4—Example of a partial tree file.

monic prefix "N"). This user's password is "joesentme" and dial digits will be prefixed assuming (in the default) they are dialed from Bell Laboratories at MH. These default attributes may be changed by the user at the node called the PREROOT (see Fig. 4), first encountered when the file is initially opened at the beginning of the access program. Functions at the PREROOT allow the user to change his default location or add another phone number temporarily to those listed in the header, or allow a new user to identify himself.

Each user is assigned a computer user-identification and a corresponding file system "home" directory. The user's private access tree is expected to be found in this directory along with other pertinent files like a personal directory, a mailbox, and a personal calendar. Besides each private tree-file, there is a collection of "public" tree files, any of which may be accessed by any user. The real users have access structures in which only real or imminent functions are accessible. The running example throughout this paper is such a tree. One pseudo-user, called "demo", has indicated access to named functions that are not implemented but from which provocative demonstrations have been given.

#### *3.2.3 Program description*

The structure (also in the sense of Michael Jackson) of the host-resident access program is illustrated in Fig. 5. The uppermost level shows the initial processing of the tree-file, the initial labeling of the buttons, and an iteration of processing "button pushes". Each button push is processed by reading the button identification message from the teleterminal, checking the validity of the message and its "type," executing the corresponding command with respect to current "modes," traversing the tree to the successor-node, and displaying the button labels pertaining to that node. The formats of various messages to and from the teleterminal are discussed in the next subsection. Command "modes" and "types" are briefly described.

The lock mode and the explain mode are program states that apply to the imminent button push but which are manipulated by a previous button push. The lock mode disables most commands and is a means for providing privacy protection: it is toggled by the CHANGELOCK command (see Fig. 4), which prompts for a password when "unlocking." The explain mode is enabled by the EXPLAIN command (see Fig. 4) and causes the program to substitute an explanation of the subsequent button push for the regular action of that button push.

A very brief description of each of the currently implemented commands is tabularized in Fig. 6. These commands are usually requested by the user through the operation of an appropriately labeled button. However, as another "type" of request, unprompted

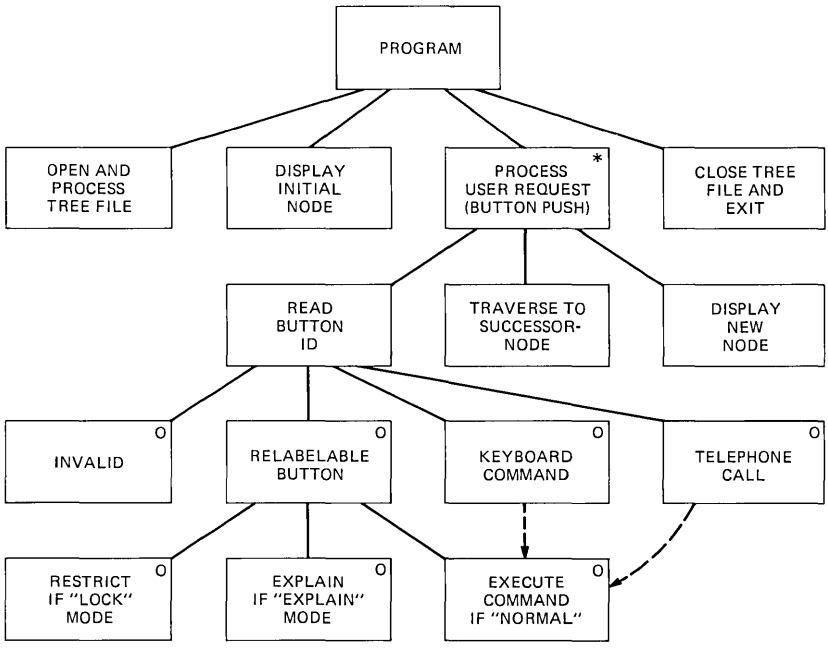



Fig. 5—Structure diagram of the access program.

keyboard entries are also interpreted as user-defined requests for specific commands. For example, unprompted numeric key entries are interpreted as telephone numbers.

Examination of Figs. 2, 4, and 6 will aid the reader in correlating the data file with the program. Several button pushes will be illustrated here, with more in the scenario presented in Section IV. In Fig. 4, the line beginning with "Personal Asst" (in the ROOT record-group) illustrates that **Personal Asst** is the label for button number eight on the root node (see Fig. 2). It is further seen from that record that when this button is selected by the user, the default TRAVERSE command is executed by the program and the tree traverses to the node whose record group is named PERSASST. Every command concludes with a traversal to the successor node; the TRAVERSE command does nothing else before that. The TRAVERSE command (see Fig. 6) requires no other data, and so none are present in the miscellaneous field of the record. The PERSASST record group is included in Fig. 4 and it corresponds to the button labels on the "Personal Assistant" node shown here in Fig. 7.

The line beginning with "System" on Fig. 4 illustrates that **System** is the label for button number ten on the root node (Fig. 2). When this button is selected by the user, the RUNSCROLL command is executed

| COMMAND | DESCRIPTION |
|---|---|
| **BACKRESTART** | backup or restart in directory |
| **CAL2MONTH** | special interface for two-month calendar |
| **CALENDAR** | special interface for general calendar |
| **CALL** | call from button label |
| **CALL2LINE** | call from two-line button label |
| **CALLLOCAL** | call with local prefix |
| **CALLPOTS** | "POTS" call |
| **CALLPREV** | call party previously called |
| **CALLPRFX** | call with labeled tie-line prefix |
| **CALLTRAV** | "POTS" call or traverse in directory |
| **CALLTREE** | call from temporary tree |
| **CATFILE** | cat indicated file and exit or traverse |
| **CHANGECALLTRAV** | toggle "call/traverse" mode in directory |
| **CHANGELOCK** | toggle "lock" mode |
| **DIRCATEGORY** | select directory category |
| **DIRINSTALL** | proceed with personal directory installation |
| **DIRTREE** | make temporary tree from temporary directory |
| **DUMMYCALL** | dummy version of CALL |
| **EXITENTRY** | traverse to TREE with selected directory entry |
| **EXITROOT** | OTHERFILEEXIT or OTHERFILEROOT via mode |
| **EXPLAIN** | toggle "explain" mode |
| **INSERTLABEL** | insert button label into textual files |
| **LEFTRIGHT** | left/right traversal in directory |
| **MAILRETURN** | "detour" return to mail program |
| **OTHERFILE** | use other file instead of TREE |
| **OTHERFILEEXIT** | leave other file to exit point in TREE |
| **OTHERFILEROOT** | leave other file to TREE root |
| **OTHERLOC** | change default location |
| **OTHERPHONE** | enter "other number" for call-memos |
| **READMAIL** | special interface to read mail |
| **REMINDER** | special interface for reminder service |
| **RUNBOTTOMLINE** | use bottom line, create child, continue |
| **RUNRETURN** | clear & scroll, create child, label return button |
| **RUNSCROLL** | clear & scroll, create child, wait |
| **SENDCALLMEMO** | special interface to send a call-memo |
| **SENDMAIL** | send mail via *UNIX* software |
| **STDCALLMEMO** | special interface for standard call-memo |
| **TEMPBACKUP** | temporal backup |
| **TEMPLATEBEGIN** | initiate template creation |
| **TEMPLATEFILL** | fill in directory template |
| **TEMPLATEINIT** | initialize template from previous call |
| **TEMPLATEMATCH** | count or list template matches in directory |
| **TRAVERSE** | simple tree traversal to successor |

Fig. 6—Current commands.

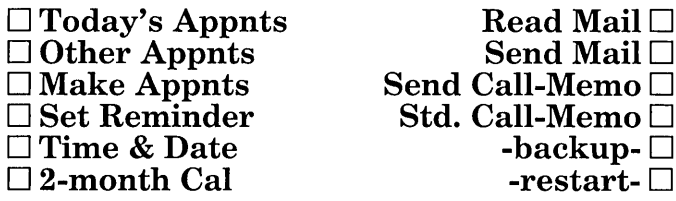



Fig. 7—The **Personal Assistant** node.

by the program and afterwards the tree traverses (by default) to the current node. In executing the RUNSCROLL command (see Fig. 6), the program clears the teleterminal of button labels and causes the indicated program (in this case, a *UNIX* program called the "shell"[14]) to execute from the teleterminal as a computer terminal. The screen now looks like that of a traditional terminal to the user, who is "talking" to the command interpreter of a supposedly familiar operating system. Upon termination of the shell, the access program gets control back and it executes the default traversal to the current, or root, node. In the "jargon" of the *UNIX* operating system, the activity is that the access program puts itself to sleep, and requests the spawning of a child process. The program expects to find the path to the particular child in the miscellaneous field of the record. When the user terminates the session with the child, it is killed and the parent (the access program) is awakened. The entire operation is perceived by the user as a leaf in that a function was requested, granted, and no traversal took place.

Changes to the data were discussed in the previous subsection. But not all changes to the access method can be implemented as simple "edits" of the access file. A new or changed command requires a program change and subsequent recompilation. However, the program has been structured specifically to simplify such changes. Command execution is implemented with a huge multiple branch. The C code in each branch is functionally decomposed. It should be relatively easy for a neophyte programmer (not our unsophisticated user, however) to modify this part of the program. A third kind of change is the unforeseen "nasty" kind of program change that requires a higher level of skill in the person making the change.

The user is expected to "log in" to his own account on the host from his teleterminal, and then to cause the access program to execute from his home directory. The program runs as a "shell" through which other programs, including the shell and other application programs, are invoked. It is planned for the invocation of the access tree program to be placed in the user's profile to facilitate "log in."

### 3.3 System considerations

The internal firmware was designed so that communication between the user and the application software can proceed with a minimum of teleterminal-related details in the application programs. The areas of concern were system conventions imposed by the operating system, which impact the human interface to the application software and the potential for application software dependence on the detailed characteristics of the teleterminal. While the software was specifically implemented in the environment of the *UNIX* operating system, and the style of the paper assumes that, steps were taken in the design of the internal firmware to avoid assumptions about the specific operating system.

Operating systems establish conventions to control terminal I/O. Most systems usurp characters from the ASCII set for line deletion, character deletion, and interruption of program execution. In the *UNIX* operating system, the default symbols are: '@' to erase a line, '#' to erase a character, and DEL to interrupt program execution. Because our goal was to provide a consistent human interface independent of any operating system idiosyncrasies, the teleterminal allows the application to down-load the codes that should be used for these functions. Extra keys on the keyboard are provided for the erase-character, erase-line, and program-interrupt functions. Since conditions arise when a user may want to enter one of the special characters, the teleterminal will automatically prefix these characters with a host-specified escape-character (back-slash, by default, in the *UNIX* operating system).

Minimizing software dependence on the details of the teleterminal is achieved by defining a high-level interface to the functions:

(*i*) Labeling re-labelable buttons

(*ii*) Dialing

(*iii*) Forcing output to be placed on a single line of the display device.

By and large these functions are sufficient to allow all the tree-oriented software to execute on terminals that provide some form of relabelable buttons. Besides the use of a CRT to label buttons adjacent to the screen, alternatives are a button with embedded light-emitting diodes (LEDs) or a CRT with a touch-sensitive screen. Control of the screen is typically achieved by the host software through print statements. The teleterminal-specific details, such as the actual character strings defining control functions, can be isolated in one file. Using the macro facility of the C language, these control messages can be made to look like function calls. For example, a macro for labeling buttons may be defined such that the statement:

```
button_label(btn_num,"Prefix Call");
```

would expand to the C statement:

```
printf("%c%c%s\n",ESC,'a'-1+btn_num,"Prefix Call");
```

When executed (assuming btn_num is equal to 2) this statement results in the string

ESC b Prefix Call CR NL

being transmitted to the teleterminal, which causes the second button on the left-hand side of the screen to be cleared, then labeled with "Prefix Call". Similar macros may be defined for other primitives.

## IV. A SCENARIO

The functionality of the tree-like access method is demonstrated by a scenario wherein **A** calls **B**, but **B** is either busy or not home: **A** leaves **B** a message to return the call; then **B** becomes available, reads the message, and returns the call to **A**. We (a user named "us") are **A**, Susan is **B**, and the scenario begins at the root node (Fig. 2). A monologue for introducing a new user to the teleterminal, its services, and tree-like access is appropriate for the unsophisticated user before presenting this scenario. Such a monologue is illustrated in Ref. 5.

We select the **Susan** button and place a call to Susan, the department secretary. Customization of the tree-file is demonstrated by the presence of the **Susan** button in the root, indicating that this call is frequently made; and also by the button text itself: **Susan** is a "friendlier" label than initials and last name (like a typical directory entry) or **secretary**. As seen in Fig. 4, the miscellaneous field of the "Susan" record in the tree-file contains her telephone number and her computer user identification. The CALL command (see Fig. 6) causes the telephone number to be dialed and the id to be stored for possible subsequent use. The tree is traversed to the **Personal Assistant** (Fig. 7).

As an aside, this last traversal exemplifies two points already made. The user, in customizing his own tree-file, determines the successor node of *branches* like **Susan**. There are three appropriate choices:

(*i*) Traversal to the current node emphasizes the leaf nature of such a botton and, often, if such a call fails to complete, a "related" call (that is, to someone on the same node) may be the logical subsequent event.

(*ii*) Traversal to the root gives the perception of "resetting" the teleterminal for subsequent use.

(*iii*) Traversal to the **Personal Assistant** is appropriate if calendar and message functions are likely to be subsequently used.

With the use of **restart** and **backup** buttons, the decision is not all that critical. The second point that is exemplified is the fact that the access tree is not, strictly speaking, a tree: two branches to the **Personal Assistant** have already been demonstrated.

Continuing with the scenario, let **Susan** be busy and let us send her a "Call-Memo" requesting that she call us back. There are two "Call-Memo" buttons on the **Personal Asst** node (Fig. 7): **Send Call-Memo** prompts the user through the creation of a general call-memo and **Std Call-Memo** immediately sends a standard "I called you—Please call me back" call-memo. This standard message is created once by, or for, the user and is available in the user's directory along with a host of other handy files, like a personal directory and a personal calendar. We select **Std Call-Memo** and we are prompted on the bottom line of our teleterminal as shown here in Fig. 8.

The program can't be sure that the call-memo is intended for the person last called, but that is the default operation. If the user wishes to change the default, he identifies the desired user (typing is echoed after the colon on the bottom line). We simply type the "CR" key to continue the default operation and the scenario proceeds. The program responds on the bottom line as shown here in Fig. 9.

Proceeding to the second half of the scenario, **Susan** is tracked as she returns the call. Her first action is to read her mail. From the root (assuming that her tree-file resembles ours), she selects **Personal Asst** and then **Read Mail**. A special mail program is invoked that interfaces the mail program provided in the *UNIX* operating system

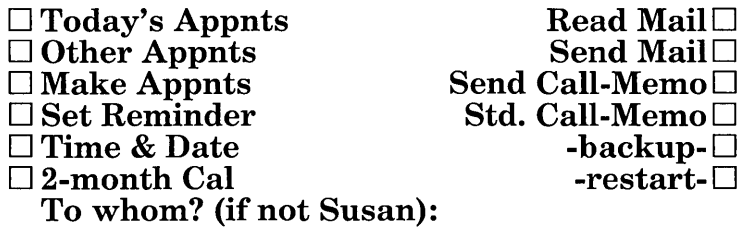



Fig. 8—The **Personal Assistant** node with **Std Call-Memo** prompt.

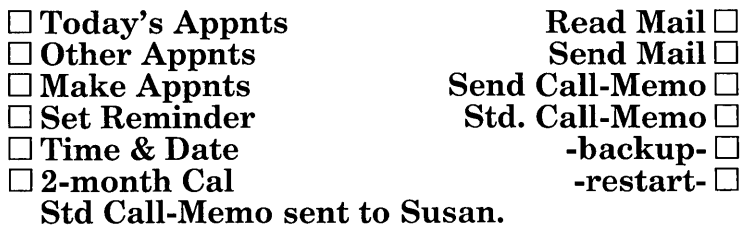



Fig. 9—The **Personal Assistant** node with **Std Call-Memo** response.

to the teleterminal. If our call-memo is her only new message, the screen appears as shown here in Fig. 10.
After selecting our message, the call-memo is scrolled onto the screen as shown here in Fig. 11.

There are a number of different responses to such a message: the choice of action is made after pressing the **process msg** button, as shown in Fig. 12. In our scenario, Susan elects to return the call and selects **Return Call.** Our telephone number is extracted from the call-memo by the program, the mnemonic prefix **"MH"** is appropriately translated into the correct dial prefix, depending on the location of Susan's teleterminal, and the call is made. The screen returns to the call-memo (Fig. 11) from which point the message would probably be **Thrown Out**, ending the scenario.

User perception and the importance of cosmetics cannot be overemphasized. The reader **must** be aware that the demonstration of this very scenario on an actual teleterminal is many times more appealing than a description of the demonstration in such a paper as this. Many

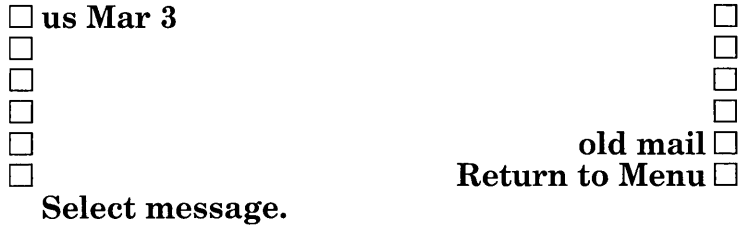



Fig. 10—Root → **Personal Assistant** → **Read Mail.**

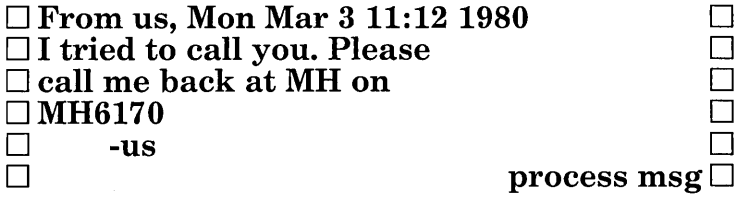



Fig. 11—The Standard Call-Memo.

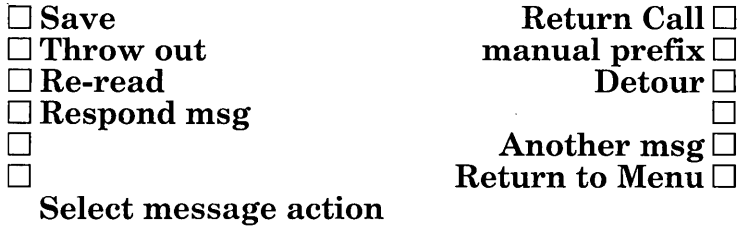



Fig. 12—The Message Action menu.

other functions have been implemented but the interest of brevity prevents any description. These include a personal calendar (see the first three buttons on Fig. 7), organizational and alphabetic directories, special-purpose directories (such as emergency numbers, dial-a-"x", and assistance), a personal telephone directory, manual selection of dial prefixes, and a food applications demonstration. In addition, the tree provides access to standard *UNIX* programs supplying facilities like a monthly calendar, time and date, a desk calculator simulation, and a large selection of computer games. Furthermore, accesses to a voice storage system and a dial-up dictation system have been implemented that present a highly satisfactory human interface.

## V. FUTURE DIRECTIONS

The functional split between the teleterminal and the host has provided a flexible and efficient interface for experimenting with new applications. We have "ported" the application software from a Digital Equipment PDP 11/45 system using the MERT[15] operating system to a Tandem computer using the Guardian[16] operating system. The human interface remained the same even though the I/O characteristics of the two systems are quite different.

Our experience exposed several areas where the current system is deficient. In the human interface, the teleterminal provides special keys for delete line, delete character, and program interrupt. Two additional special keys labeled "end of file" and "program abort" are needed to provide a completely operating-system-independent interface. Keys with the permanent labels "explain," "backup," and "restart," would improve the human interface to the access tree by freeing three of the relabelable buttons for other functions. An extension of the screen management allowing an arbitrary screen split between button labels and scrolled text would improve the human interface by allowing communications in excess of 32 characters.

The interface between the teleterminal and application programs could be improved by having the teleterminal emit a special button status message when the telephone goes off-hook. This would allow applications to respond directly to dial requests, eliminating a button push under some circumstances. A button label frame buffer capable of retaining about 30 button labels coupled with a special "more" key would remove the teleterminal-specific restriction that no more than twelve buttons can be labeled at one time.

The current teleterminal interface to the host supports a single character-oriented data stream. Application software must take special care to ensure that a message signaling the occurrence of an asynchronous event does not get interspersed with other messages. For example, if an electronic mail application wanted to inform the user of the

arrival of a message, the application software would have to coordinate an attempt to write the message "you have mail" with all other programs that might have reason to write to the instrument. This responsibility belongs in the network and in the teleterminal.

The solution requires the network to support some form of data multiplexing. For teleterminal applications, the network must support a control channel and several data channels. The control channel is a message-oriented channel used to establish the switch between data channels. The network must ensure that control messages are atomic, delivered error free, and are flow controlled. In order to ensure that control messages remain synchronized with the data channels, it is advisable that the control channel be implemented as a subchannel on a multiplexed data stream. The teleterminal would require enhancements to conform to the message protocol of the control channel and to route data over the appropriate data channel. It is sufficient that the teleterminal be able to maintain several data connections but only be able to receive (and send) data over one channel at a time.

For a network with this form of data multiplexing, asynchronous events, such as the arrival of electronic mail, could be implemented by sending an "alert" message to the teleterminal over the control channel. The teleterminal would respond by flashing a button with the **alert** label. The user would eventually respond to the alert by pressing the **alert** button, causing the teleterminal to send a "hold" status message over the current active data channel, switch to the mail data channel, and send a "proceed" status message over the mail channel. The mail program could then take control of the screen. To return to the program interrupted by the mail function, either the user would press a permanently-labeled **resume** buttom or the mail program would automatically return by sending an appropriate control message to the teleterminal. In either case, the teleterminal would send a status message over the previous data channel. The application program that had been put on hold would take over the screen by rewriting the last frame before the interruption.

A new design of the teleterminal has been recently completed and it incorporates many of the features described. Cosmetically, the screen is logically the same size as that of a conventional computer terminal and the keyboard is large enough for touch-typing (see Fig. 13). Thus, the set of services appropriately addressed is expanded from "enhanced telephony" to include information management and office automation.[17] Functionally, the new teleterminal provides a telephone interface like that of a traditional 6-button key telephone, thus allowing easy interface to call directors and speakerphones. The internal processor is a "microcomputer system" running under a *UNIX*-like operating system.[18] At this time it simply emulates the functions of the old

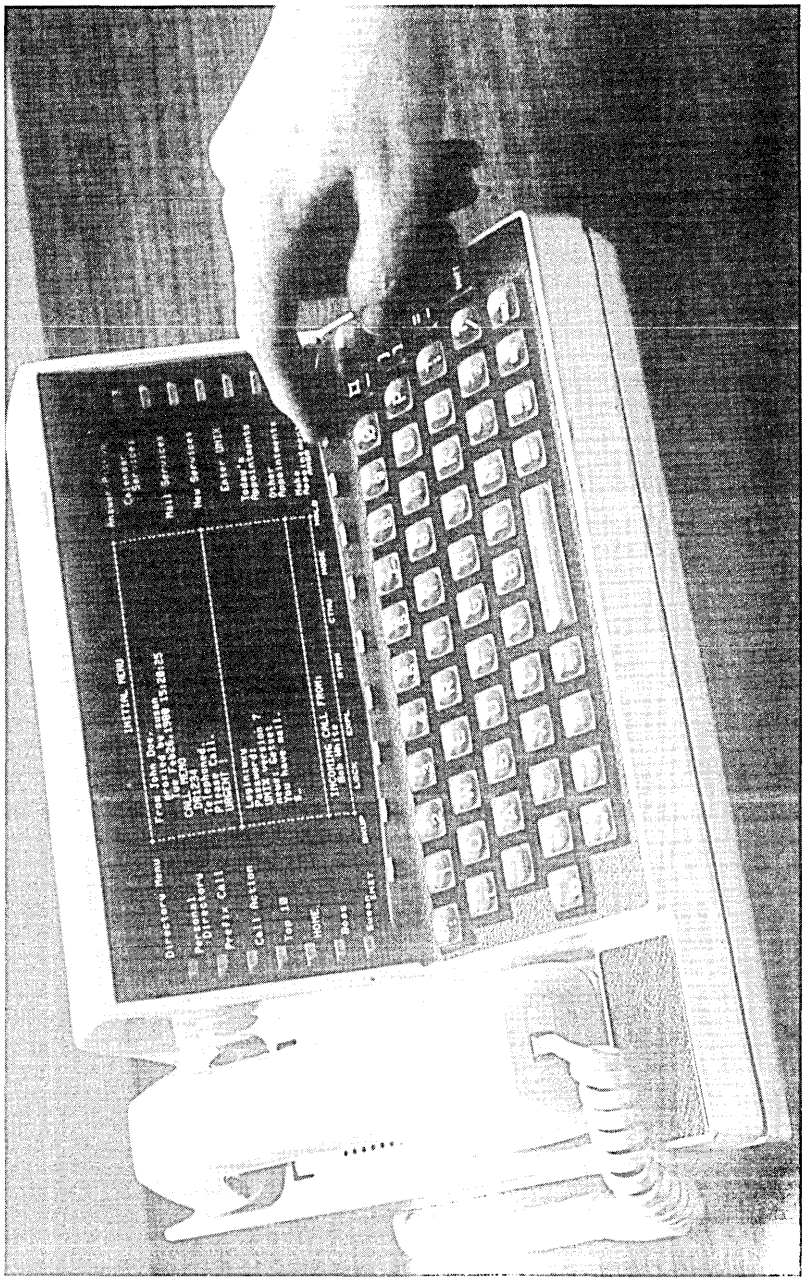


Fig. 13—Redesigned teleterminal.

internal software; however, many more elaborate functions are planned.

One of these is the capacity for a multiplicity of message "windows," each supporting one of the independent data connections mentioned above. Investigations into combining voice and data onto one digital channel are proceeding.[19] New applications become feasible once the teleterminal is connected to an intelligent switching office: for example, displaying the number of a calling party and a variety of transfer and screening functions. Many new capabilities, and improvements to existing capabilities, have been suggested by the teleterminal users and these are gradually being added.

## VI. CONCLUSION

This paper has described the software strategy behind the construction of a research tool called a teleterminal. The realization of this strategy lies in a software architecture whereby intelligence is distributed between the teleterminal's internal processor and a host computer. The programs resident in each of these locations, and their intercommunication, have been discussed.

More in the abstract, another part of this strategy is an access method whereby the user of a future teleterminal can conveniently interface to the abundance of applications that could be available. General concepts have been presented and a first implementation has been described from the user's viewpoint. The proposed access method has been found highly acceptable by colleagues and acquaintances but formal testing[20,21] on the general public has only just begun. This is ongoing work and this paper represents a "snapshot in time" that is already slightly out of date. The first working teleterminal was demonstrated internally in September of 1978. Since that time three dozen teleterminals have been constructed and a user group assembled. The software has undergone innumerable changes, and the newer "model" was designed and constructed.

Acknowledgments are owed to the following colleagues for their contributions to the software effort of the project: Bob Anderson for assistance with the internal software, Martin Sturzenbecker for the special mail program and a filter that enables the access program to work at a regular terminal, Ron Gordon for the tree "compiler" and the definition of the "high-level" data language illustrated in Fig. 4, Ron and Martin for a "Food Applications" program that clearly illustrates the value of organizing "atomic functions" into "generic capabilities," Misha Buric for the personal calendar program, Al Usas for support with the "*UNIX*-Tandem" environment, and Bob Allen and Donna Zanolla for their human factors testing.

## REFERENCES


1. D. W. Hagelbarger, R. V. Anderson, and P. S. Kubik, "Experiments with Teleterminals," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.1.1-5.
2. D. W. Hagelbarger, "Experiments with Teleterminals," *B.S.T.J.*, this issue.
3. G. D. Bergland, "An Experimental Telecommunications Test Bed," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.3.1-5.
4. R. W. Lucky, "A Flexible Experimental Digital Switching Office," *Proc. 1978 Int. Zurich Seminar on Digital Commun.*, Zurich, Switzerland, 1978.
5. R. A. Thompson, "Accessing Experimental Telecommunications Services," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.2.1-5.
6. G. Cohen, *The Psychology of Cognition*, New York: Academic Press, 1977.
7. G. T. Uber, et al., "The Organization and Formating of Hierarchical Displays for the Online Input of Data," *Proc. Fall Joint Computer Conf.*, 1968.
8. D. L. McCracken and G. G. Robertson, "Editing Tools for ZOG, a Highly Interactive Man-Machine Interface," *Proc. Int. Conf. Commun.*, Boston, MA, 1979.
9. M. Behzad and G. Chartrand, *Introduction to the Theory of Graphs*, Boston, MA: Allyn and Bacon, 1971.
10. "The UNIX Time-Sharing System," *57,* Part 2, a dedicated special issue of the *B.S.T.J.* (July-August 1978).
11. B. W. Kernighan and D. M. Ritchie, *The C Programming Language*, Englewood Cliffs, NJ: Prentice-Hall, 1978.
12. R. D. Gordon and D. L. Smith, "An Access Tree Editor," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.7.1-5.
13. M. A. Jackson, *Principles of Program Design*, New York: Academic Press, 1975.
14. S. R. Bourne, "The UNIX Shell," *B.S.T.J., 57* (July-August 1978), pp. 1971–90.
15. H. Lycklama and D. L. Bayer, "The MERT Operating System," *B.S.T.J., 57* (July–August 1978), pp. 2049–86.
16. *Tandem 16, Guardian™ Operating Manual.* Tandem Computers Inc., 1980.
17. R. N. Klapman, "Enhanced Communications in an Executive Office," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.4.1-5.
18. W. M. Schell, "Control Software for an Experimental Teleterminal," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.8.1-5.
19. R. A. Thompson, "An Experimental User-Resident Communications Controller Supporting Sub-Rate Circuit-Switched Service," *Proc. Int. Symp. on Subscriber Loops and Services*, Munich, Germany, 1980, pp. 68–71, and *the IEEE Trans. Commun., COM-30*, Number 6 (June 1982), pp. 1399–408.
20. R. B. Allen, "Cognitive Factors in the Use of Menus and Trees: An Experiment," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.5.1-5.
21. R. A. Thompson, "User's Perceptions with Experimental Services and Terminals," *Proc. Nat. Telecommun. Conf.*, New Orleans, LA, 1981, pp. F2.6.1-5.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 1, January 1983
*Printed in U.S.A.*

# Experimental Teleterminals—Hardware

BY D. W. HAGELBARGER, R. V. ANDERSON, and P. S. KUBIK

(Manuscript received May 7, 1982)

*We have designed and built a series of experimental telephones that have facilities for both voice and data communication. The first of these has a 5-inch cathode ray tube character display and a small, full ASCII keyboard. Next to the display are twelve buttons whose labels are part of the display. The set is only a little larger than a standard telephone and can sit permanently on your desk without dominating it. It is a combination of an abbreviated computer terminal and a telephone and uses a microprocessor. Favorable experience with this first model led to several variations with larger displays and keyboards. We are still conducting experiments exploring the uses of these teleterminals.*

## I. INTRODUCTION

We have felt for some time now that a telephone set capable of providing integrated voice and data communications is desirable. A series of experiments to explore the possible uses of such a phone is under way. The key word here is "experiments"; we are not describing a development project and in fact one would not build a commercial product the way these experiments were built. The sets described here were internally referred to as GETSETS, an acronym for General-purpose Electronic Telephone SETS.

We have made a telephone with a microprocessor, a small keyboard, and a 5-inch cathode ray tube (CRT) display (see Fig. 1). The phone is small enough to sit permanently on your desk and measures about 9 inches wide, 10 inches deep, and 6 inches high. The keyboard has full American Standard Code for Information Interchange (ASCII) capability and the display has a capacity of 16 lines of 32 characters each. There are 12 buttons next to the display that can be labeled by writing text on the screen adjacent to the buttons. These buttons permit the user to perform any of a large number of complicated operations. The usual problem with having a large repertory of complicated operations


Fig. 1—Set 32.

is that it is difficult to remember the exact form of the commands if they are only used occasionally. It is much easier to recognize the command when it is presented as a button label, rather than having to recall it.

Even though the GETSET can function as a computer terminal, that is not our intent; the emphasis is on using the microprocessor to provide new and improved telephone functions. Modern electronic private branch exchanges, for example, have a processor powerful enough to provide abbreviated dialing, message forwarding, incoming party announcing, appointment calendar, and similar services. To access these functions, the telephone set needs a display and an alphanumeric keyboard. Since the set interacts with a host computer the keyboard needs full ASCII capability. All the ASCII characters, no matter how obscure, must be available. (It has been our experience that any apparently unneeded key, such as '$' in scientific computing, is given a critical function.) The display is a compromise between the desire for many characters and the limits of desk space.

In the future it will no doubt be possible to send mixed voice and data over a digital network; however, to experiment today we must use the existing analog network. In today's network there is a digital data path to the host computer and a separate voice line used for placing calls. The voice line uses the standard tone calling, dial tone, ringing, busy, etc.

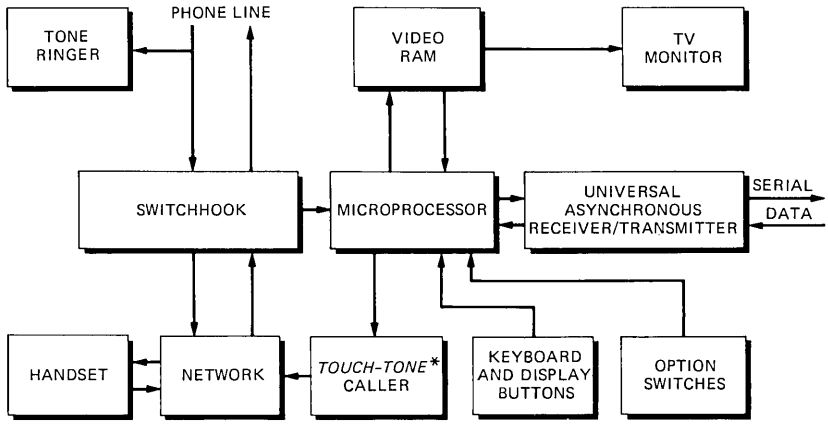
## II. MECHANICAL CONSTRUCTION

We started with a 10-button Multibutton Electronic Telephone (MET) set and removed the flat plate that contains the lamps, keys, and tone caller from the right side. This was replaced with the keyboard and a hood covering the CRT monitor. The phone network, speaker for the tone ringer, and switchhook are, as before, under the handset; however, it was necessary to repackage the network with the hybrid transformer shielded with mu-metal and rotated approximately 60 degrees about a vertical axis in order to eliminate magnetic pickup from the vertical oscillator coil and deflection yoke of the monitor.

All the electronics except for the power supply are in the set, under the keyboard and CRT tube. In so far as possible all parts are mounted to the sturdy grey plastic frame of the MET set. This allows the set to be operated with the top and bottom covers removed. There is a small "dog house" on the back of the set that covers the neck of the CRT tube, which is about 1½ inches too long.

## III. FUNCTIONAL DESCRIPTION

Figure 2 shows a block diagram of the set. The keyboard chip, Standard Microsystems Corp. (SMC) KR2376-ST, converts key switch closures to 7-bit parallel ASCII codes, which are sent to the microprocessor. The buttons adjacent to the display are connected to this chip and are assigned control codes. The microprocessor, Intel 8748 with 8155 random access memory (RAM) and 8755A erasable programmable read only memory (EPROM), communicates with the host computer over a pair of serial data lines (typically 1200 baud) using a universal asynchronous receiver/transmitter (UART), Harris 6402-9.




*SERVICE MARK OF AT&T

Fig. 2—Block diagram of SET32.

The microprocessor controls the display with a Matrox MTX 1632A Video RAM. To the microprocessor this looks like a 512 × 8-bit random access memory. Writing an ASCII code in location 000 causes the corresponding character to appear at the left end of the top line of the display. Location 001 comes next, and so on, with location 511 appearing at the right end of the bottom line. The Video RAM contains the character generator, refresh memory, and video and sync generators to drive the monitor.

The microprocessor also controls the tone caller, selecting the frequencies and closing a relay that mutes the phone transmitter, attenuates the receiver, and connects the caller through an isolation transformer to the network. The microprocessor can sense the state of the switchhook and make the clicker give audio feedback for the keyboard.

The monitor is a Ball Corp. TV-50 Data Display Monitor. We had to remake the printed circuit card in order to pack it into the set. This required mounting the components cordwood style and arranging the tall components to miss the neck of the CRT tube.

The tone ringer uses the American Microsystems Inc. (AMI) S2561 chip. It monitors the phone line for ringing current and converts it to tones driving the small standard speaker used in the MET set. It has a feature of ringing louder if the phone is not answered promptly. We have kept the rheostat from the MET set, which allows one to control the loudness of ringing.

Four of the seven option switches, which are mounted at the back of the set, control the baud rate, which can be set to any of the following speeds: 50, 75, 110, 134.5, 150, 200, 300, 600, 1200, 1800, 2400, 4800, or 9600. The other switches are: local or on-line, half or full duplex, and RS 232 or current loop. There is also a reset push button that clears the screen and restarts the microprocessor program. Another control sets display brightness.

## IV. POWER

The set draws about 0.7 ampere at +5 volts, about 0.8 ampere at +12 volts, and about 25 milliamperes at −12 volts for a total of about 13 watts. We have added slots in the rear top cover below the hood. These allow air flow, which keeps the set from overheating. The −12 volts is used only by the keyboard chip and the Electronic Industries Association (EIA) interface. The power at +12 volts is dependent on the display adjustment; making the raster wider, for instance, increases the current needed.

## V. FIRMWARE

The program for the microprocessor is discussed in the companion paper.[1] Tables I and II of Ref. 1 summarize the command codes that the host uses to control the teleterminal.

## VI. FURTHER DETAILS

### *6.1 The keyboard*

The keyboard is based on a calculator keyboard. It has a molded silastic membrane with a hollow conical bump for each key. Inside each bump is a piece of conducting rubber; pressing on a button flattens the bump and pushes the conductive part against a printed circuit board where it electrically connects a pair of intermeshed fingers. The shape is such that the cone "snaps through" and gives a good feel. The keys are 12 mm on centers and the buttons are 5.5 by 6.5 mm. Making the keys relatively small allows even a very large finger to push a button without hitting the ones beside it. The buttons are too close together to permit touch-typing; it is really a two-finger keyboard.

The key array is 5 keys high and 12 wide with the space "bar" taking up two key positions. The top row is used for placing calls. In addition to the ten digits, the leftmost key has "*" and the rightmost is labeled with a modified "square" as well as "underline." (In ASCII use, the key produces "underline" in both upper and lower case.)

The clicker, which gives acoustic feedback, is a Western Electric MA5A relay with the contact stack removed. It is mounted so the armature strikes the frame to make the click. The microprocessor drives it with a 14-millisecond pulse.

The six buttons on each side of the display are also mounted 12 mm on centers. The vertical centering and height of the display are adjusted so that the buttons line up with the first line and every third line after that.

### *6.2 CRT monitor*

Since telephones are used in many different ambient light situations, it is desirable to improve the contrast of the screen. We have tried tubes with dark-colored phosphor and also bonded, neutral-grey filter faceplates. Both of these decrease the ambient light reflected from the screen. Current sets have the dark-colored phosphor and an etched front surface on the faceplate.

### *6.3 Microprocessor*

We started with just an Intel 8748; however, it soon became clear that it was not fast enough to keep up with incoming display data at 1200 baud. We therefore added the 8155 RAM, which has 256 bytes, as a buffer. This is enough for half of a solid screenful and seems to be adequate. It also gave us some more input/output (I/O) ports. Next we ran out of program space so we added the 8755A EPROM. This added 2K bytes, about half of which are used, and more I/O ports, which we are also using.

## VII. STATUS AND FURTHER EXPERIMENTS

We have built about three dozen of these sets. Approximately half of them are being used by members of our center in an experiment to test their use by a community of users and to develop new telephone functions. Others are mainly used for demonstrating the concept of a teleterminal. To test various proposed improvements, we have built three "breadboard" models. These models contain only keyboards and displays. We connect them to a Cromemco Z80 microprocessor. This has the advantage of allowing us to program in C language rather than assembly language.

Figure 3 shows the color set. It has the same keyboard as the original set. The display is a 5-inch color tube from a Sony color TV monitor. Since the color tube is larger than the black and white one, this makes the case deeper and boxier than the first set.

Figure 4 shows an intermediate-size set. The display is the same height but 1-1/2 inches wider than the first set. It has 16 lines of 64 characters each. We have added nine new buttons across the bottom of the screen that can be labeled with the last line of the display. We have used the same membrane keys but spread them sideways to take up the extra 1-1/2 inches. You can place your fingers on adjacent keys (with no extra space). It may be possible to touch-type on this keyboard.

Figure 5 shows the largest of the three models. It has almost a full-

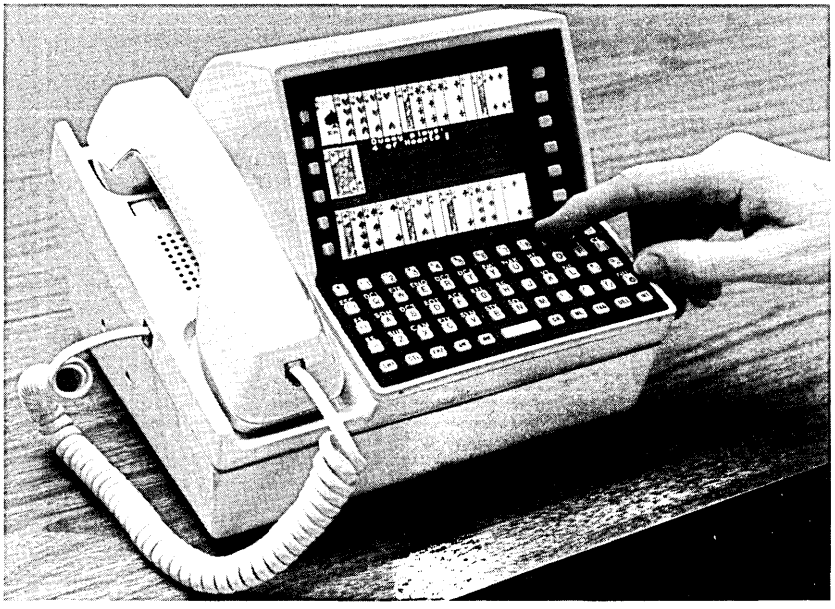


Fig. 3—The color set.

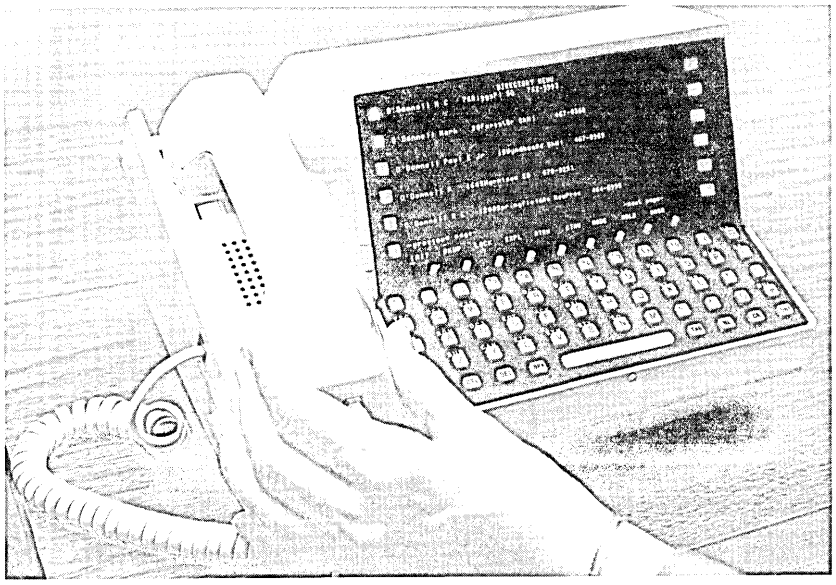


Fig. 4—Set 64.

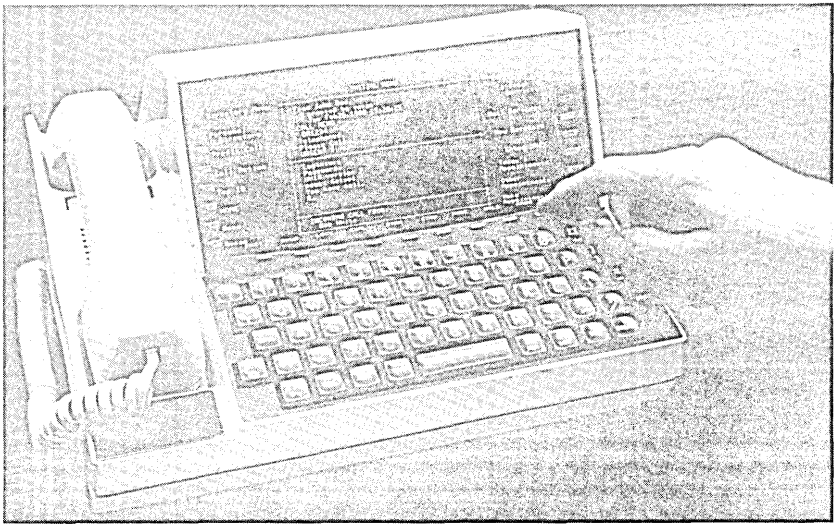


Fig. 5—Set 80.

sized keyboard. The display has 24 lines of 80 characters, the same as most computer terminals. It is a low-profile tube so the display is about 3-1/2 inches high and 7-1/4 wide; however, the definition is good and it is easy to read. The characters are about the same size as

ordinary typewriter output. There are eight buttons on each side and nine across the bottom of the display. With the addition of a speakerphone and a telephone interface that imitates a six-button keyset, this set is being used in the Executive Planning Information and Communication (EPIC) trial.[2]

## VIII. CONCLUSION

This paper has described a series of experimental teleterminals. These have been used to provide a new, computer-enhanced telephone and other services. Probably no one teleterminal suits all users' needs. Rather there should be a variety available, so that users can choose the keyboard display and other features that are most suited to their needs.

## IX. ACKNOWLEDGMENTS

Many people have made contributions to this project. We owe inspiration to H. S. McDonald for his notion of a "Superphone." We are indebted to D. L. Bayer for guidance and help in programming the microprocessor. R. A. Payne built the cases for the color set and the intermediate-size set; W. Kaminski designed the keyboards and displays for all the breadboard sets, and J. M. Gaughran contributed the layout for more than a dozen different printed circuit cards. He also expedited their production, often under a tight schedule. We particularly want to thank R. A. Thompson for the programs in the host computer, which make the teleterminals do something useful.

## REFERENCES

1. D. L. Bayer and R. A. Thompson, "An Experimental Teleterminal—The Software Strategy," B.S.T.J., this issue.
2. Richard N. Klapman, "Enhanced Communications in an Executive Office," Nat. Telecommun. Conf. Rec., *3,* 1981, pp. F2.4.1–5.

## CONTRIBUTORS TO THIS ISSUE

**Robert V. Anderson,** Bell Laboratories, 1954—. Mr. Anderson began his career as a Technical Aide. He has worked on a wide variety of projects including error-correcting code demonstrator, seismic encoder, world's fair exhibits, aids for the handicapped, computer control of the Murray Hill exhibit area, and computer terminals. He is currently a Member of Technical Staff.

**Douglas L. Bayer,** B.A., 1966, Knox College; M.S. (Physics), 1968, and Ph.D. (Nuclear Physics), 1970, Michigan State University; Rutgers University, 1971-1973; Bell Laboratories, 1973—. At Rutgers University, Mr. Bayer was involved in low-energy nuclear scattering experiments. At Bell Laboratories, he has been involved in computer operating systems research on mini- and microcomputers. Member, ACM.

**Aland K. Chin,** B.A., 1972, Brandeis University; M.S., 1975, Ph.D., 1977, Cornell University; Senior Research Engineer, Honeywell Electro-Optics Center, 1977–1978; Bell Laboratories, 1978—. Mr. Chin is involved in the design, processing, and characterization of light-emitting diodes for optical communication systems. Member, American Physical Society, American Association for the Advancement of Science, Phi Beta Kappa, New York Academy of Science, Electrochemical Society.

**Richard V. Cox,** B.S. (Electrical Engineering), 1970, Rutgers University; M.A., 1972, Ph.D., 1974 (Electrical Engineering), Princeton University; The Aerospace Corporation, 1973–1977; Assistant Professor, Rutgers University, 1977–1979; Bell Laboratories, 1979—. Mr. Cox is a member of the Acoustics Research Department. His current research interests are in digital speech coding, analog speech scrambling, and real-time speech processing systems.

**Michael A. DiGiuseppe,** B.S. (Chemistry), 1968, Polytechnic Institute of New York; Ph.D. (Chemistry), 1975, Brown University; Allied Corporation, 1973–1980, Bell Laboratories, 1980—. At Allied Corporation Mr. DiGiuseppe was engaged in the crystal growth of garnets for use as laser-hosts and substrates for magnetic films. Research efforts focused on high temperature phase equilibria in oxide melts. At Bell Laboratories Mr. DiGiuseppe is engaged in the crystal growth research of alloys for lightwave applications. Member, AACG (current AACG-NJ chairman), ACS, AAAS, Sigma Xi.

**Nancy Y. Graham,** A.B., 1959, M.A., 1962 (Mathematics), University of California, Berkeley; Bell Laboratories, 1970–1975, 1979—. Ms. Graham was a part-time employee in the Acoustics Research Department from 1970 to 1975, where she became interested in computer graphics. Ms. Graham returned to Bell Laboratories in 1979 and is a Member of Technical Staff in the Computer Graphics Group.

**David W. Hagelbarger,** A. B. (Chemistry, Mathematics, Physics), 1942, Hiram College; Ph.D. (Physics), 1947, California Institute of Technology; Bell Laboratories, 1949—. Mr. Hagelbarger has done research in a variety of fields including learning machines, stability of molten zones, magnet design, error-correcting codes, information retrieval, seismic monitoring of nuclear test bans, world's fair exhibits, switching networks, aids for the handicapped, and teaching aids. He is currently interested in improving the interface between computers and people.

**Nuggehally S. Jayant,** B.Sc. (Physics and Mathematics), 1962, Mysore University; B.E., 1965, and Ph.D. (Electrical Communication Engineering), 1970, Indian Institute of Science, Bangalore; Research Associate at Stanford University, 1967–1968; Bell Laboratories, 1968—. Mr. Jayant was a visiting scientist at the Indian Institute of Science in 1972 and 1975. He has worked in the field of digital coding and transmission of waveforms, with special reference to robust speech communications. He is also editor of the IEEE Reprint Book, *Waveform Quantization and Coding.*

**Peter D. Karabinis,** B.E.(E.E.), 1974, M.E.(E.E.), 1976, The City College of New York; Bell Laboratories, 1976—. Mr. Karabinis is a member of the Exploratory Radio Systems Department of the Radio Transmission Laboratory. His interests lie in wideband digital transmission over dispersive channels, optimum transmitter-receiver design techniques, and digital signal processing. He has worked on space diversity combiners and intermediate frequency equalizers, and is currently involved in design and development of components for high-speed digital radio systems. Member, Eta Kappa Nu.

**Vassilis G. Keramidas,** Ph.D. (Solid State Science), 1973, Materials Research Laboratory, Pennsylvania State University; Bell Laboratories, 1973—. Mr. Keramidas has worked on LEDs for displays and optoelectronics, on ohmic contacts to compound semiconductors and on the crystal growth, by liquid phase epitaxy, and characterization of

materials for LEDs for lightwave communications. He is currently Supervisor of a Special Materials Group. Member of American Physical Society, Electrochemical Society, American Association for Crystal Growth.

**Peter S. Kubik,** Western Electric, 1942–1947; Bell Laboratories, 1947—. Mr. Kubik started his career as a screw machine operator at Western Electric. He attended evening classes in mechanical design at Stevens Institute and Newark College of Engineering. Mr. Kubik did the physical design of two experimental switching offices, ESSEX and XDS. He has worked on a variety of projects including magnetostrictive delay lines, the first successful gas laser, world's fair exhibits, educational aids, aids for the handicapped, and computer terminals. He is currently a Member of Technical Staff.

**Barbara J. McDermott,** B.A., 1949, University of Michigan, Ann Arbor; M.A., 1962, Columbia University, New York; Haskins Laboratories, 1950–1959; Bell Laboratories, 1959—. From 1950–1959 Ms. McDermott was a research assistant at Haskins Laboratories, New York. Since 1959, she has been a Member of the Speech Research Department, Bell Laboratories, Murray Hill, NJ, where her principal research interest has been the perception of transmitted speech.

**John A. Morrison,** B.Sc., 1952, King's College, University of London; Sc.M., 1954 and Ph.D., 1956, Brown University; Bell Laboratories, 1956—. Mr. Morrison has done research in a number of different areas of applied mathematics and mathematical physics. He has recently been interested in probability theory, and various queueing problems in particular. He was a Visiting Professor of Mechanics at Lehigh University during the Fall semester, 1968. He is currently the Managing Editor of SIAM Review. Member, American Mathematical Society, SIAM, IEEE, Sigma Xi.

**Ann Marie S. Quinn,** B.S. (Linguistics and Speech Science), Rutgers University, New Brunswick, NJ; Bell Laboratories, 1969—. Ms. Quinn is presently working in the Acoustics Research Department.

**Robert H. Saul,** Ph.D. (Metallurgy and Materials Science), 1967, Carnegie-Mellon University; Bell Laboratories, 1967—. Initially Mr. Saul was engaged in characterization and development of epitaxial growth techniques for opto-electronic materials and devices. In 1972

he became Supervisor of a group responsible for developing visible light-emitting diodes. That work pioneered the use of multi-slice epitaxial techniques for achieving state of the art performance in a manufacturable growth system. Since 1975 Mr. Saul has supervised a group responsible for developing a variety of infrared light-emitting diodes that are used in optical-isolators and fiber optic systems. His current interests include development of long wavelength sources for lightwave transmission systems and reliability of detectors. Mr. Saul holds eight patents in the area of materials growth and opto-electronic devices. Senior member, IEEE; member, American Physical Society, Sigma Xi, Tau Beta Pi. Chairman, 1982 IEEE Specialist Conference on Light-Emitting Diodes and Photodetectors.

**R. A. Semplak,** B.S. (Physics), 1961, Monmouth College; Bell Laboratories, 1955—. Mr. Semplak's main research interest is in studies of atmospheric effects on micro- and millimeter-wave propagation. Currently, he is a member of the Radio Communications Research Department. Member, Sigma Xi and Commission F of the International Union of Radio Science (URSI/USNC).

**Henryk Temkin,** Ph.D. (Physics), 1975, Stevens Institute of Technology; Cornell University, 1975–1977; Bell Laboratories, 1977—. Mr. Temkin is involved in the development of materials and devices for fiber communications. He is currently a member of the Semiconductor Electronics Research Department. Member, American Physical Society, Electrochemical Society.

**Richard A. Thompson,** B.S., 1964 (Electrical Engineering), Lafayette College; M.S., 1966 (Electrical Engineering), Columbia University; Ph.D., 1971 (Computer Science), University of Connecticut; Bell Laboratories, 1963–1968, 1977—. Mr. Thompson is currently in the Digital Systems Research Department at Murray Hill, NJ. He was a member of the Electrical Engineering Department at Virginia Polytechnic Institute and State University from 1971–1977, achieving the rank of Associate Professor. His research interests include probabilistic formal languages, fault tolerance and cellular automata, the human-machine interface, and communications systems. Member, IEEE, active participant in Computer and Communications Societies.

**José M. Tribolet,** B.S. (Electrical Engineering), 1972, Instituto Superior Técnico, Lisbon, Portugal; M.S., E.E., and Sc.D degrees (Electrical Engineering) from the Massachusetts Institute of Technol-

ogy, Cambridge, in 1974, 1975, and 1977, respectively. From 1972 to 1977 Mr. Tribolet was a member of the Massachusetts Institute of Technology Research Laboratory of Electronics, where his research activities involved the application of homomorphic signal processing to speech and seismic data analysis. From 1977 to 1978 he was with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ, as a post-doctoral fellow, working on adaptive transform coding of speech. He is presently Full Professor of Electrical Engineering and Computer Science at the Instituto Superior Técnico, Lisbon, Portugal, where he directs the Research Institute in Systems Engineering and Computer Science (INESC). Mr. Tribolet was recently on sabbatical leave at the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ, from July through December 1981, where he worked on speech recognition, coding, and scrambling. Member, Sigma Xi.

**Wilson W. Yale,** B.S. (Mathematics), 1973, M.S. (Mgt. Science), 1976, Ph.D. (Industrial Engineering), 1978, Lehigh University; Bell Laboratories, 1978—. Mr. Yale has done research in non-linear programming, cutting tool engineering, finance, decision support systems, and customer decision modeling. More recently he has been interested in stochastic problems arising from office automation and from estimating the demand for virtual private-line networks. He was a Visiting Professor of Industrial Engineering at Lehigh University during the spring semester of 1979. Member, ORSA, TIMS.

**Christie L. Zipfel,** A.B. (Physics), 1963, Vassar College; M.S., 1965, and Ph.D., 1969 (Physics), University of Michigan; Bell Laboratories, 1974—. Mrs. Zipfel is a member of the Lightwave LED Group.

# PAPERS BY BELL LABORATORIES AUTHORS

## COMPUTING/MATHEMATICS

Fishburn P. C., Farquhar P. H., **Finite-Degree Utility Independence.** Math Oper R 7(3):348–353, 1982.

Gehrlein W. V., Gopinath B., Lagarias, J. C., Fishburn P. C., **Optimal Pairs of Score Vectors for Positional Scoring Rules.** Appl Math O 8(4):309–324, 1982.

Graham R. L., Foulds L. R., **Unlikelihood That Minimal Phylogenies for a Realistic Biological Study Can Be Constructed in Reasonable Computational Time.** Math Biosci 60(2):133–142, 1982.

Harrison M. L., Davidson R. P., Wadsack R. L., **Bellmactm-32 - A Testable 32-Bit Microprocessor.** J Digit Sys 6(2–3):131–138, 1982.

Jagpal H. S., Sudit E. F., Vinod H. D., **Measuring Dynamic Marketing Mix Interactions Using Translog Functions.** J Bus 55(3):401–415, 1982.

Kantor W. M., **Spreads, Translation-Planes and Kerdock Sets.** Siam J Alg 3(3):308–318, 1982.

Lagarias J. C., **Best Simultaneous Diophantine Approximations. 1. Growth-Rates of Best Approximation Denominators.** T Am Math S 272(2):545–554, 1982.

Radner R., Rosenthal R. W., **Private Information and Pure-Strategy Equilibria.** Math Oper R 7(3):401–409, 1982.

Shepp L. A., **The XYZ Conjecture and the FKG Inequality.** Ann Probab 10(3):824–827, 1982.

Sobotka L. J., **Special Issue—International Test Conference (Editorial).** J Digit Sys 6(2–3):103–105, 1982.

Vardi Y., **Nonparametric-Estimation in Renewal Processes.** Ann Statist 10(3):772–785, 1982.

Walley P., Fine T. L., **Towards A Frequentist Theory of Upper and Lower Probability.** Ann Statist 10(3):741–761, 1982.

Wei M. C., Sholl H. A., **An Expression Model for Extraction and Evaluation of Parallelism in Control-Structures.** IEEE Comput 31(9):851–863, 1982.

Yannakak M., **The Complexity of the Partial Order Dimension Problem.** Siam J Alg 3(3):351–358, 1982.

## ENGINEERING

Capasso F., Logan R. A., Tsang W. T., **Interdigitated pn Junction Device With Novel Capacitance Voltage Characteristic, Ultralow Capacitance and Low Punch-Through Voltage.** Electr Lett 18(18):760–761, 1982.

Cheng C. L., Coldren L. A., Miller B. I., Rentschl J. A., Shen C. C., **Low-Resistance Ohmic Contacts to P-InP.** Electr Lett 18(17):755–756, 1982.

Kamgar A., **Miniaturization of Si MOSFET at 77-K.** IEEE Device 29(8):1226–1228, 1982.

Kartalopoulos S. V., **The Minimization of Logic Functions Utilizing 2-Dimensional Representations of Hypercubes.** Int J Elect 53(3):233–245, 1982.

Kenney J. J., **Digital Radio for Transmission at 6 and 11 GHz.** Microwave J 25(8):71+, 1982.

Levine B. F., et al., **High-Quantum-Efficiency Low-Threshold Microcleaved $Al_xGa_{1-x}As$ Lasers.** Electr Lett 18(16):690–691, 1982.

Lin C., Glodis P. F., **Tunable Fiber Raman Oscillator in the 1.32–1.41 $\mu$m Spectral Region Using a Low-Loss, Low OH-Single-Mode Fiber.** Electr Lett 18(16):696–697, 1982.

Logan, R. A., et al., **Low-Threshold GaInAsP-InP Mesa Lasers.** Electr Lett 18(18):782–783, 1982.

Ng K. K., Taylor G. W., Sinha A. K., **Instability of MOSFET due to Redistribution of Oxide Charges.** IEEE Device 29(8):1323–1330, 1982.

Stolen R. H., Howard R. E., Pleibel W., **Substrate-Tube Lithography for Optical Fibers.** Electr Lett 18(18):764–765, 1982.

Stone J., Cohen L. G., **Minimum-Dispersion Spectra of Single-Mode Fibers Measured With Subpicosecond Resolution by White-Light Cross-Correlation.** Electr Lett 18(16):716–718, 1982.

Temkin H., Chin A. K., Dutt B. V., **Schottky-Barrier Restricted GaAlAs Laser.** Electr Lett 18(16):701–703, 1982.

Tsai Y. K. et al., **Structure-Preserving Model-Reduction With Applications to Power-System Dynamic Equivalents.** IEEE Circ S 29(8):525–535, 1982.

Vanderziel J. P., Logan R. A., **Generation of Short Optical Pulses in Semiconductor-Lasers by Combined dc and Microwave Current Injection.** IEEE J Q EL 18(9):1340–1350, 1982.

## MANAGEMENT/ECONOMICS

Hsiao C., **Autoregressive Modeling and Causal Ordering of Economic Variables.** J Econ Dyn 4(3):243–259, 1982.

Rosenthal R. W., **A Dynamic Oligopoly Game With Lags in Demand—More on the Monotonicity of Price in the Number of Sellers.** Int Econ R 23(2):353–360, 1982.

## PHYSICAL SCIENCES

Anderson P. W., **Localization—A Growth Industry (Editorial).** Physica B&C 110(1–3):1830–1836, 1982.

Archer V. D., **Methods for Defect Evaluation of Thin (100) Oriented Silicon Epitaxial Layers Using a Wet Chemical Etch.** J Elchem So 129(9):2074–2076, 1982.

Banavar J. R., Cieplak M., Cieplak M. Z., **Influence of Boundary-Conditions on Random Unfrustrated Magnetic Systems.** Phys Rev B 26(5):2482–2489, 1982.

Beck S. M., Brus L. E., **Transient Raman-Scattering Study of the Initial Semi-Quinone Radical Kinetics Following Photolysis of Aqueous Benzoquinone and Hydroquinone.** J Am Chem S 104(18):4789–4792, 1982.

Beni G., Scrosatti B., **(IT) Recent Advances in Electrochemical Energetics. 4. Electrochromic Displays.** Chim Ind M 64(7–8):487–491, 1982.

Benton J. L., Kimerling L. C., **Capacitance Transient Spectroscopy of Trace Contamination in Silicon.** J Elchem So 129(9):2098–2102, 1982.

Bhatt R. N., Ramakrishnan T. V., **Effect of Mass Anisotropy on the Low-Temperature Conductivity of Disordered-Systems in Two Dimensions.** Physica B&C 110(1–3):2078–2080, 1982.

Bhatt R. N., **Frustration in Highly Random Magnetic Systems.** Physica B&C 110(1–3):2145–2147, 1982.

Bondybey V. E., **Laser Vaporization of Silicon-Carbide—Lifetime and Spectroscopy of $SiC_2$.** J Phys Chem 86(17):3396–3399, 1982.

Bondybey V. E., English J. H., Miller T. A., **The Absorption-Spectrum of the Perfluoropyridine Cation in a Ne Matrix.** Chem P Lett 90(5):394–396, 1982.

Bosch M. A., Dayem A. H., Harrison T. R., Lemons R. A., **Lamellar Solidification of Laser Melted Co-Si Films.** Appl Phys L 41(4):363–364, 1982.

Bowmer T. N., Reichmanis E., Wilkins C. W., Hellman M. Y., **Radiation Degradation of Copolymers of Methyl-Methacrylate and 3-Oximino-2-Butanone Methacrylates.** J Pol Sc Pc 20(9):2661–2668, 1982.

Brown B. L., Mills A. P., Tyson J. A., **Results of a 440-Day Search for Gravitational Radiation.** Phys Rev D 26(6):1209–1218, 1982.

Bruckenstein S., Miller B., **Current-Voltage Analysis of Photo-Electrochemical Cells Under Mass and Light-Flux Variation.** J Elchem So 129(9):2029–2034, 1982.

Bucksbaum P. H., Boker J., Storz R. H., White J. C., **Amplification of Ultrashort Pulses in Krypton Fluoride at 248-NM.** Optics Lett 7(9):399–401, 1982.

Cais R. E., Kometani J. M., **Ethylene Vinyl Bromide Copolymers by Reductive Debromination of Polyvinyl Bromide—A C-13 NMR-Study.** Macromolec 15(4):954–960, 1982.

Capasso F., Tsang W. T., Hutchinson A., Williams G. F., **Enhancement of Electron-Impact Ionization in Super-Lattices—A New Avalanche Photo-Diode With A Large Ionization Rate Ratio.** Inst Phys C1982(63):569–570, 1982.

Capasso F., Tsang W. T., Hutchinson A. L., Foy P. W., **The Graded Bandgap Avalanche-Diode—A New Molecular-Beam Epitaxial Structure With A Large Ionization Rate Ratio.** Inst Phys C1982(63):473–478, 1982.

Chen C. Y., Cho A. Y., Gossard A. C., Garbinski P. A., **Offset Channel Insulated Gate Field-Effect Transistors.** Appl Phys L 41(4):360–362, 1982.

Chen L. J., Mayer J. W., Tu K. N., Sheng T. T., **Lattice Imaging of Silicide Silicon Interfaces.** Thin Sol Fi 93(1–2):91–97, 1982.

Chu S. N. G., **A Simple Damage-Free Grooving Method for Revealing the Quality of InP/InGaAsP Multilayer Structure.** J Elchem So 129(9);2082–2085, 1982.

Davies J. H., Lee P. A., Rice T. M., **Electron Glass.** Phys Rev L 49(10):758–761, 1982.

Disalvo F. J., Waszczak J. V., **Magnetic-Properties of Copper Chalcogenide Spinels.** Phys Rev B 26(5):2501–2506, 1982.

Dubois L. H., **Oxygen-Chemisorption and Cuprous-Oxide Formation on Cu(111)—A High-Resolution EELS Study.** Surf Sci 119(2–3):399–410, 1982.

Duncan A., Roskies R., Vaidya H., **Monte-Carlo Study of Long-Range Chiral Structure in QCD.** Phys Lett B 114(6):439–444, 1982.

Dynes R. C., **Localization and Correlation—Effects in Metals and Semiconductors Experiment.** Physica B&C 110(1–3):1857–1865, 1982.

Espinosa G. P., Cooper A. S., Barz H., **Isomorphs of the Superconducting—Magnetic Ternary Stannides.** Mater Res B 17(8):963–969, 1982.

Frankenthal R. P., Siconolfi D. J., **The Equilibrium Surface—Composition of Tin Lead Alloys.** Surf Sci 119(2–3):331–348, 1982.

Girvin S. M., Jonson M., Lee P. A., **Interaction Effects in Disordered Landau-Level Systems in Two Dimensions.** Phys Rev B 26(4):1651–1659, 1982.

Golding B., Fox D. L., Haemmerle W. H., **Dephasing of Extrinsic Tunneling Systems in Silica Glass.** Physica B&C 110(1–3):2039–2040, 1982.

Goldsmith P. F., Snell R. L., DeGuchi S., Kortkov R., Linke R. A., **Vibrationally Excited Cyanoacetylene in the Orion Molecular Cloud.** Astrophys J 260(1):147–158, 1982.

Golovchenko J. A., Cox D. E., Goland A. N., **Critical Analysis of the Charge-State Dependence of the Energy-Loss of Channeled Ions.** Phys Rev B 26(5):2335–2340, 1982.

Greenside H. S., Coughran W. M., Schryer N. L., **Non-Linear Pattern-Formation Near the Onset of Rayleigh-Benard Convection.** Phys Rev L 49(10):726–729, 1982.

Griffiths J. E., Sunder W. A., **Raman-Spectra and Phase-Transitions in $O_2PF_6$.** J Chem Phys 77(6):2753–2756, 1982.

Hall C. K., Helfand E., **Conformational State Relaxation in Polymers—Time-Correlation Functions.** J Chem Phys 77(6):3275–3282, 1982.

Hebard A. F., Fiory A. T., **Vortex Dynamics in Two-Dimensional Superconductors.** Physica B&C 110(1–3):1637–1648, 1982.

Hegarty J., Sturge M. D., Weisbuch C., Gossard A. C., Wiegmann W., **Resonant Rayleigh-Scattering From an Inhomogeneously Broadened Transition—A New Probe of the Homogeneous Linewidth.** Phys Rev L 49(13):930–932, 1982.

Heller A., **Electrochemical Solar-Cells.** Solar Energ 29(2):153–162, 1982.

Ho K. M., Fu C. L., Harmon B. N., Weber W., Hamann D. R., **Vibrational Frequencies and Structural-Properties of Transition-Metals Via Total-Energy Calculations.** Phys Rev L 49(9):673–676, 1982.

Kaplan A. E., Smith P. W., Tomlinson W. J., **Switching of Reflection of Light at Non-Linear Interfaces.** P Soc Photo 317:305–310, 1981.

Khanarian G., **Direct-Current Electric-Field Induced 2nd Harmonic-Generation in Flexible Molecules and Polymers.** J Chem Phys 77(5):2684–2687, 1982.

Lanzerotti L. J., et al., **$SO_2$ Ice and Applications to the Frosts of IO.** Astrophys J 259(2):920–929, 1982.

Leventhal M., et al., **Time-Variable Positron-Annihilation Radiation From the Galactic-Center Direction.** Astrophys J 260(1):L 1+,1982.

Levine B. F., Tsang W. T., Bethea C. G., Capasso F., **Electron-Drift Velocity-Measurement in Compositionally Graded $Al_xGa_{1-x}As$ by Time-Resolved Optical Picosecond Reflectivity.** Appl Phys L 41(5):470–472, 1982.

Lloyd J. R., Nakahara S., **Some Remarks on Void Growth in Thin Silver Films (Letter).** Thin Sol FI 92(1–2):L 61, 1982.

Loponen M. T., Dynes R. C., Narayanamurti V., Garno J. P., **The Time-Dependent Specific-Heat of Dielectric Glasses.** Physica B&C 110(1–3):1873–1879, 1982.

Marcus M. A., **Lack of 2nd-Order Phase-Transitions in Cubic Blue Phases in Landau Theory.** Molec Cryst 82(5):167–171, 1982.

McAfee K. B., Szmanda C. R., Hozack R. S., Johnson R. E., **Dynamics of Energy-Transfer During Charge-Exchange in Symmetric Diatomic Systems.** J Chem Phys 77(5):2399–2407, 1982.

McBrierty V. J., Douglass D. C., Wudl F., **A Nuclear Magnetic-Resonance Study of $(TMTSF)_2PF_6$.** Sol St Comm 43(9):679–682, 1982.

McBrierty V. J., Douglass D. C., Furukawa T., **Magnetic-Resonance and Relaxation in a Vinylidene Fluoride Trifluoroethylene Co-Polymer.** Macromolec 15(4):1063–1067, 1982.

McCrory J. C., **Electroanalytical Chemistry of Chlorinated Phenols at a Glassy-Carbon Electrode.** Analyt Chim 141(SEP):105–114, 1982.

Miller R. C., Gossard A. C., Tsang W. T., Munteau O., **Bound Excitons in P-Doped GaAs Quantum Wells.** Sol St Comm 43(7):519–522, 1982.

Miller R. C., Tsang W. T., Munteau O., **Extrinsic Layer at $Al_xGa_{1-x}As$-GaAs Interfaces.** Appl Phys L 41(4):374–376, 1982.

Miller T. A., Bondybey V. E., **Spectroscopy of Molecular-Ions (Review or Bibliog.)** Appl Sp Rev 18(1):105–169, 1982.

Mollenauer L. F., Vieira N. D., Szeto L., **Mode-Locking by Synchronous Pumping Using a Gain Medium with Microsecond Decay Times.** Optics Lett 7(9):414–416, 1982.

Murray C. A., Allara D. L., Hebard A. F., Padden F. J., **Determination of Sample Morphology of Multilayered Structures Used in Surface Enhanced Raman-Scattering Experiments.** Surf Sci 119(2–3):449–478, 1982.

Nelson E. D., Kaufman S., **Release Rate Calorimetry of PVC Compounds Containing Antimony Oxide and Iron-Oxide.** J Fire Flam 13(2):79–103, 1982.

Olego D., Chang T. Y., Silberg E., Caridi E. A., Pinczuk A., **Compositional Dependence of Band-Gap Energy and Conduction-Band Effective Mass of $In_{1-x-y}Ga_xAl_yAs$ Lattice Matched to InP.** Appl Phys L 41(5):476–478, 1982.

Penzias A. A., **Laser Patents (Letter).** Science 217(4565):1082, 1982.

Raghavachari K., Haddon R. C., Starnes W. H., **Primary Event in the Thermal Dehydrochlorination of Pristine Polyvinyl-Chloride—Intermediacy of A Cyclic Chloronium Ion.** J Am Chem S 104(19):5054–5056, 1982.

Ramakrishnan T. V., Sur K., **Theory of a Mixed-Valent Impurity.** Phys Rev B 26(4):1798–1811, 1982.

Riley J. E., **Ultrahigh Sensitivity Uranium Analyses Using Fission-Track Counting—Further Analyses of Semiconductor Packaging Materials.** J Rad Chem 72(1–2):89–99, 1982.

Schluter M., **Theoretical—Models of Schottky Barriers.** Thin Sol Fi 93(1–2):3–19, 1982.

Schmidt P. F., Glascock M. D., **Applications and Problems of Parametric Counting.** J Rad Chem 72(1–2):231–244, 1982.

Stiles K. R., Lee J. W., **Automated Conductivity Profiler for Multilayer GaAs-(AlGa)As Structures.** Rev Sci Ins 53(9):1449–1451, 1982.

Stillinger F. H., Weber T. A., **Asymmetry Between Protons and Proton Holes in Gas Phase Neutralization Reactions.** Molec Phys 46(6):1325–1333, 1982.

Stone F. T., **Separation of Total-Loss Data Into Its Absorption and Scattering Components—A More Accurate Model for Fiber Loss.** Appl Optics 21(15):2721–2726, 1982.

Sturge M. D., **Citation Classic—Optical-Absorption of Gallium—Arsenide Between 0.6 and 2.75 eV.** CC/Phy Chem1982(38):20, 1982.

Swaminathan V., Wagner W. R., Schumaker N. E., Miller R. C., **Anomalous Bands in the Photo-Luminescent Spectra From GaAs-(Al,Ga)As Double Heterostructures.** Thin Sol Fi 93(1-2):195-205, 1982.

Thomas G. A., et al., **Temperature-Dependent Conductivity of Metallic Doped Semiconductors.** Phys Rev B 26(4):2113-2119, 1982.

Tolk N. H., et al., **Optical Radiation From Photon-Stimulated Desorption of Excited Atoms.** Phys Rev L 49(11):812-815, 1982.

Tomita A., **Interferometric-Technique for Measuring Small Gaps in Single-Mode and Multimode Fiber Connectors (Letter).** Appl Optics 21(15):2655-2656, 1982.

Tsang W. T., **Novel Optoelectronic Devices Prepared by Molecular-Beam Expitaxy.** P Soc Photo 317:66-73, 1981.

Tsang W. T., Miller R. C., Capasso F., Bonner W. A., **High-Quality InP Grown by Molecular-Beam Epitaxy.** Appl Phys L 41(5):467-469, 1982.

Tsang W. T., et al., **Molecular-Beam Epitaxially Grown 1.3 $\mu$m GaInAsP/InP Double-Heterostructure Lasers.** Electr Lett 18(18):785-786, 1982.

Tung R. T., Poate J. M., Bean J. C., Gibson J. M., Jacobson D. C., **Epitaxial Silicides.** Thin Sol Fi 93(1-2):77-90, 1982.

Wagner R. E., Tomlinson W. J., **Coupling Efficiency of Optics in Single-Mode Fiber Components.** Appl Optics 21(15):2671-2688, 1982.

Walsh W. M., Wudl F., Aharon-Shalom H. E., Rupp L. W., Vandenberg J. M., Andres K., Torrance J. B., **Itinerant-Electron Antiferromagnetism Precursor to Superconductivity in an Organic Conductor.** Phys Rev L 49(12):885-888, 1982.

Wemple S. H., Flahive P. G., Allyn C. L., Schlosser W. O., Iglesias D. E., **Design Principles for Source Via Power GaAs-FETs.** Inst Phys C1982(63):437-442, 1982.

Wertheim G. K., Remeika J. P., Buchanan D. N., **Electronic-Structure of $BaPb_{1-x}Bi_xO_3$.** Phys Rev B 26(4):2120-2123, 1982.

Whangbo M. H., Walsh W. M., Haddon R. C., Wudl F., **Band-Structure of $(TMISF)_2X$.** Sol St Comm 43(8):637-639, 1982.

Wood D. L., Nassau K., **Refractive-Index of Cubic Zirconia Stabilized With Yttria.** Appl Optics 21(16):2978-2981, 1982.

Zweier J. L., Peisach J., Mims W. B., **Electron-Spin Echo Studies of the Copper-Complexes of Conalbumin.** J Biol Chem 257(17):314-316, 1982.

## SOCIAL AND LIFE SCIENCES

Bauer D. W., Miller J., **Stimulus-Response Compatibility and the Motor System.** Q J Exp P-A 34(AUG):367-380, 1982.

Blumberg W. E., **A Statistical-Analysis of the Publications and Collaborations of Fanelli, Alessandro, Rossi (Review or Bibliog.).** Adv Exp Med 148:7-19, 1982.

Chance B., Powers L., Ching Y., **Structure and Function of the Redox Site of Cytochrome-Oxidase (Review or Bibliog.).** Adv Exp Med 148:95-109, 1982.

Daniels S. R., Bates S., Lukin R. R., Benton C., Third J., Glueck C. J., **Cerebrovascular Arteriopathy (Arteriosclerosis) and Ischemic Childhood Stroke.** Stroke 13(3):360-365, 1982.

Egan D. E., Grimes-Farrow D. D., **Differences in Mental Representations Spontaneously Adopted for Reasoning.** Mem Cognit 10(4):297-307, 1982.

Kitagawa T., et al., **Evidence for Hydrogen-Bonding of Bound Dioxygen to the Distal Histidine of Oxycobalt Myoglobin and Hemoglobin.** Nature 298(5877):869-871, 1982.

Lund A. M., **Nebraska Symposium on Motivation—Cognitive Processes (1980) (Book Review).** Am J Psycho 95(2):333-335, 1982.

Meyer D. E., Smith J. E. K., Wright C. E., **Models for the Speed and Accuracy of Aimed Movements.** Psychol Rev 89(5):449-482, 1982.

Ogawa S., Lee T. M., **Proton Stoichiometry of Adenosine 5'-Triphosphate Synthesis in Rat-Liver Mitochondria Studied by P-31 Nuclear Magnetic-Resonance.** Biochem 21(18):4467-4473, 1982.

Teich M. C., Prucnal P. R., Vannucci G., Breton M. E., McGill W. J., **Multiplication Noise in the Human Visual-System at Threshold. 3. The Role of Non-Poisson Quantum Fluctuations.** Biol Cybern 44(3):157-165, 1982.

## SPEECH AND ACOUSTICS

Fay D., **Substitutions and Splices—A Study of Sentence Blends.** Linguistics 19(7-8):717-749, 1981.

Pierrehumbert J., Liberman M., **Fundamental—Frequency in Sentence Production—Cooper W. E., Sorensen J. M (Book Review).** Cont Psycho 27(9):690-692, 1982.

Xydeas C. S., Evci C. C., Steele R., **Sequential Adaptive Predictors for ADPCM Speech Encoders.** IEEE Commun 30(8):1942-1954, 1982.

# CONTENTS, FEBRUARY 1983

**THE BELL SYSTEM TECHNICAL JOURNAL** is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/ Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts* (*Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts*), *Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development*, and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System