THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 58 December 1979 Number 10

The Packing Problem for Twisted Pairs

By E. N. GILBERT

(Manuscript received June 26, 1979)

When wires are packed together in a bundle, as in a cable or on a shelf of a main distribution frame, the packing fraction f is the fraction of cross-sectional area of the bundle occupied by wire. With wires all the same radius, packing fractions as high as 0.90690 can be achieved. However, when the wires are pairs that have been twisted to avoid crosstalk, the packing fraction is much smaller. The largest obtainable packing fraction depends on other properties of the packing. For example, with pairs twisted by machine, all pairs twist at the same rate, and that influences the packing fraction. Several packing problems are considered, but most attention is given to a particularly regular kind of packing in which pairs twist about straight parallel axes located in a lattice arrangement. The densest lattice packing has packing fraction 0.56767. The densest lattice is a complicated one in which each wire touches 10 wires belonging to 6 other pairs. The numbers 10 and 6 cannot be increased even with nonlattice packings of pairs with straight parallel axes. These other packings are also conjectured to have packing fractions less than 0.56767, although only $f < 0.62240$ is proved.

I. INTRODUCTION

Pairs of telephone wires are often packed closely together in large numbers. These wires may belong to a cable or lie together on a shelf as jumper wires of a main distribution frame. To avoid inductive coupling, which produces crosstalk, the wire pairs are always twisted.

A twisted pair packing problem arose with a proposal for monitoring the accumulation of inoperative jumper pairs on a main distribution frame. Robert Graham of Western Electric has developed a technique for measuring the cross-sectional area of a bundle of jumper wires. Telephone company records determine the number of working jumper pairs in the bundle. The total number of pairs, working or inoperative, could be estimated from the measured area if the density of pairs in the bundle were known.

If each wire has radius r, a twisted pair has cross-sectional area $A_P = 2\pi r^2$. The number N of pairs in a bundle of area A_B is then

$$N = fA_B/A_P, \tag{1}$$

where f is the *packing fraction* (or *density*) of the bundle, the fraction of cross-sectional area filled by wire. Graham's measurements on spools of twisted pair wire suggest a value of f near 0.5. That is a much smaller packing fraction than could be achieved with single wires or untwisted pairs. To show that twisting the pairs reduces the packing fraction, this paper looks for packings that are as dense as possible. The problem takes several forms, depending on what regularities the packing may be assumed to have. For instance, do the pairs all twist around parallel, straight-line axes? If so, do they all twist at the same rate (in turns per foot)? The most regular packings are the "lattice" packings described in Section IV. The densest lattice packing will be found to have $f = 0.56767$.

The same mathematical problems arise in a different setting as follows. Suppose each dancer on a ballroom floor occupies a circular region. Dancing partners form pairs of circles in contact, and each pair rotates as the dance progresses. How densely can the floor be packed without causing couples to collide with one another?

II. PACKINGS

The pairs on the main frame are randomly packed, but there are several reasons for studying deterministic packings that maximize the fraction f. One reason is that the simplest mathematical models of random packing[1] produce low packing fractions. A more complicated random model will necessarily use some other special random process. But the random process that truly describes the main frame packing is not well understood; no special random model can be trusted completely. A packing that maximizes f, although special, has the virtue of giving a firm bound on the packing fraction actually achieved.

Another argument for maximizing f recognizes the tendency of gravity forces to pack the wires tightly. Indeed, the gravitational potential energy of a bundle of wires is minimized when the wires are packed as densely as possible. Of course, the bundle will usually have a different gravitationally stable configuration, but each time the bundle is disturbed, it tends to assume a new configuration of lower potential energy. This phenomenon may be illustrated by filling a jar with beans by pouring them in gently; then, shaking the jar will settle the beans and make room for more. Routine main frame maintenance includes "feathering" the wires, which probably helps to make the packing more dense.

Wires are assumed to be so nearly parallel to one another that they all appear as circles in any plane transverse section through the bundle. The two wires of one pair will always be represented as circles that touch. Figure 1 shows the well-known densest packing of circles in the plane.[2–4] The circles in this packing occupy a fraction $f = \pi/12^{1/2} = 0.90690$ of the plane. The circles within each horizontal row in Fig. 1 can be grouped into pairs of circles that touch. Thus, Fig. 1 can represent one cross section through a bundle of pairs of wires if the pairs are not twisted.

Arrangements like Fig. 1 often appear when circular disks are squeezed together on a flat tray. For twisted pairs, Fig. 1 would be very unlikely. The pairs must somehow twist so that they do not penetrate one another in moving from Fig. 1 to other cross sections farther along the wires.

Strictly speaking, it is possible to achieve $f = 0.90690$ in all cross sections, even with twisted pairs. Let Fig. 1 rotate bodily about some fixed center 0 to represent other cross sections. One full rotation of Fig. 1 then gives each pair one full twist. Of course, each twisted pair then forms a helix spiraling around an axis through 0, and so the pairs intertwine each other. Indeed, this intertwining cancels the crosstalk reduction that twisting the pairs tried to achieve.

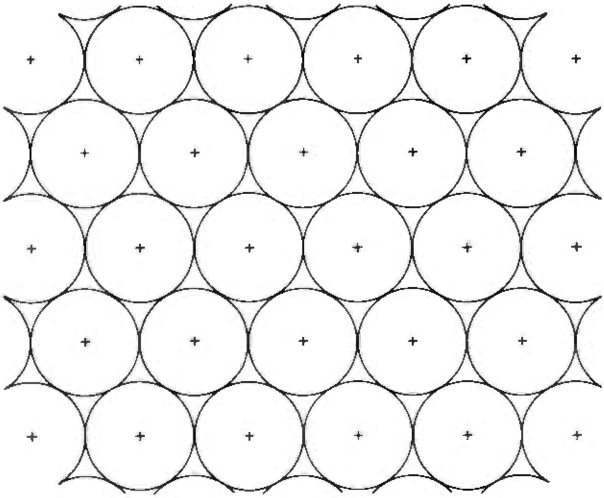


Fig. 1—The densest packing of nonoverlapping circles in the plane, $f = 0.90690$.

Very little intertwining occurs between different jumper pairs on the main frame. One way to prevent intertwining is to assume that the pairs twist about parallel straight-line axes. In a cross section, each axis appears as the point where the two circles of the pair touch. Requiring twisted pairs to have straight axes is a severe restriction. For example, it rules out Fig. 1 as a possible cross section; any rotations of pairs about the axes of Fig. 1 will cause some wires to intersect.

After assuming straight axes, one must make further assumptions about how pairs twist. One possibility is that all pairs twist at the same rate, in turns per foot. That assumption is reasonable if the pairs are cut from a reel of wire that has been twisted automatically by machine. Packings of wires that twist at the same rate will be considered in later sections. An opposite extreme is to assume that different pairs twist at rates that are not only unequal but incommensurable. Under that assumption, no two pairs can have axes lying within distance $4r$ of each other because two pairs with closer axes would overlap in some cross section. When axes are separated by at least $4r$, circles of radius $2r$ and centered at the axes do not intersect. The densest packing of such circles is again Fig. 1, now with circles of radius $2r$. In Fig. 2, these are the larger circles. Each contains two circles of radius r which

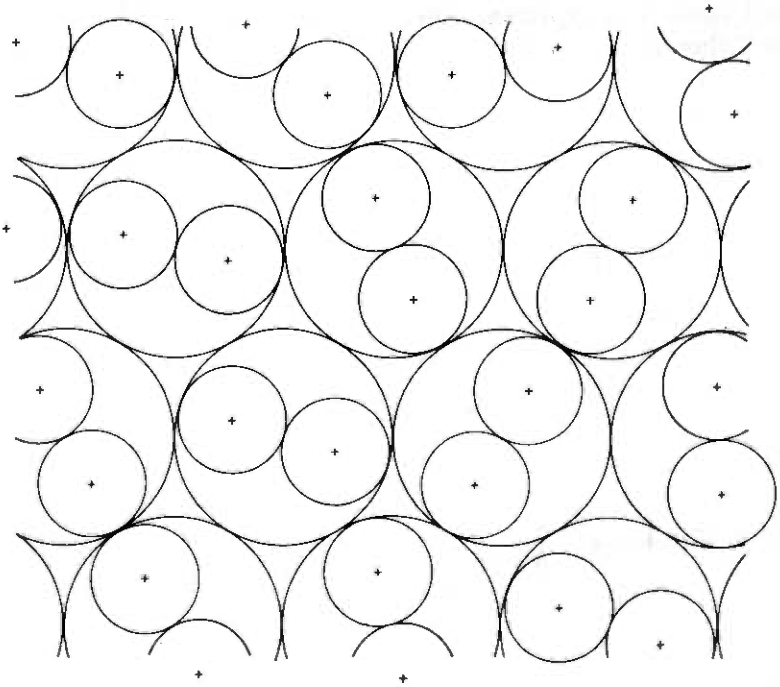


Fig. 2—The densest packing of twisted pairs with straight parallel axes and incommensurable twisting rates, $f = 0.45345$.

represent individual wires. The pair of circles of radius r fills exactly half the area of a circle of radius $2r$. Then Fig. 2 is exactly half as dense as Fig. 1; $f = \pi/(48)^{1/2} = 0.45345$ is the greatest packing fraction obtainable with incommensurable twist rates and straight parallel axes.

In this paper, pairs are always assumed to twist in the same sense, say, as a right-handed screw. Packings with pairs twisting in both senses can achieve other packing fractions. One can achieve $f = \pi/4 = 0.78540$ with half the pairs twisting in a right-handed sense and half twisting at the same rate in a left-handed sense.

III. EQUAL TWIST RATES

In a given cross section, each pair can be assigned a *phase angle* θ measured from the horizontal to a line between centers of the two wires. Figure 2 shows pairs of different phases. If all pairs twist at the same rate, the difference in phase between two pairs remains constant as one moves along the wires. It is no longer necessary to separate pair axes by distance $4r$. Theorem 1 below shows that the minimum allowed distance depends on the phase difference ϕ between the two pairs.

Figure 3 shows two pairs with axes at distance a apart. The constant phase difference for the two pairs is ϕ; one pair has a phase θ and the other has phase $\theta + \phi$.

Theorem 1: Suppose two pairs, with phase difference ϕ as shown in Fig. 3, twist about their straight parallel axes at the same rate. The smallest distance achieved between centers of wires in different pairs is $a - 2rM(\phi)$, where

$$M(\phi) = Max\{|\sin \tfrac{1}{2}\phi|, |\cos \tfrac{1}{2}\phi|\},$$

r is the radius of the wires, and a is the distance between the axes of the pairs.

Proof: The theorem is proved simply if Fig. 3 is regarded as the complex plane. Take the origin to be one pair axis. The other pair axis is at $a \exp(i\psi)$, where ψ is an angle from the horizontal to the line between axes. The two wires of the first pair have centers P_+ and P_- with

$$P_\pm = \pm r \exp(i\theta).$$

The centers Q_+ and Q_- of wires of the second pair are

$$Q_\pm = a \exp(i\psi) \pm r \exp\{i(\theta + \phi)\}.$$

One of the four distances to consider is $|Q_+ - P_+| = |a \exp(i\psi) + r \exp i\theta(\exp i\phi - 1)|$. Write

$$\xi = \exp\{i(\theta - \psi + \tfrac{1}{2}\phi + \tfrac{1}{2}\pi)\}$$

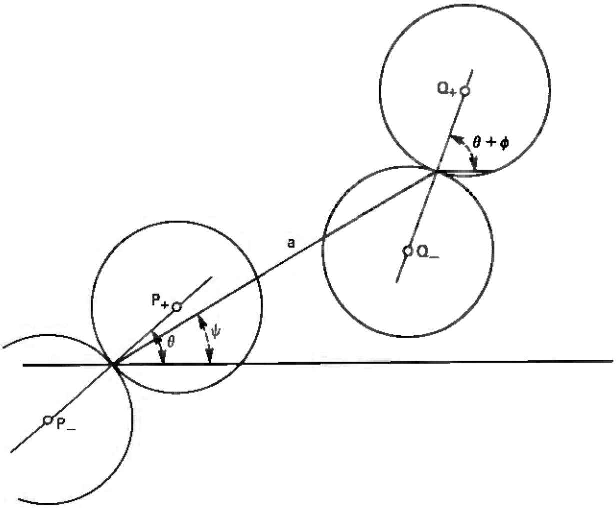


Fig. 3—Two twisted pairs with phase difference ϕ.

so that

$$|Q_+ - P_+| = |a + 2r\xi \sin \tfrac{1}{2}\phi|, \tag{2}$$

and likewise

$$|Q_- - P_+| = |a - 2r\xi \cos \tfrac{1}{2}\phi|, \tag{3}$$

$$|Q_+ - P_-| = |a + 2r\xi \cos \tfrac{1}{2}\phi|, \tag{4}$$

$$|Q_- - P_-| = |a - 2r\xi \sin \tfrac{1}{2}\phi|. \tag{5}$$

As θ varies, ξ moves on the unit circle $|\xi| = 1$. The four distances (2), $\cdots$, (5) have their maxima and minima at $\xi = \pm 1$. The smallest value attained by any of the four distances is either $a - 2r|\sin \tfrac{1}{2}\phi|$ or $a - 2r|\cos \tfrac{1}{2}\phi|$, as the theorem states.

Theorem 1 shows that the distance a between axes of two pairs with given phase difference ϕ can be only as small as

$$a = a(\phi) = 2r\{1 + M(\phi)\}. \tag{6}$$

This separation can be less than the $4r$ used in Fig. 2. The smallest allowed separation is obtained with $\phi = \pm 90°$;

$$a(\pm 90°) = (2 \pm 2^{1/2})r = 3.4142r.$$

For a given wire, say the one with center P_+, the closest approach to another wire center Q_+ or Q_- occurs when $\xi = \pm 1$, i.e., when

$$\theta - \psi + \tfrac{1}{2}\phi = 90° \text{ or } 270°.$$

Ordinarily $|Q_+ - P_+|$ and $|Q_- - P_+|$ have different minima and then the minimum distance $a - 2rM(\phi)$ is attained at only one of the two angles θ. If $a = a(\phi)$, each wire at 0 touches only one wire of the other pair. If $\phi = \pm 90°$, $|\sin \frac{1}{2}\phi| = |\cos \frac{1}{2}\phi|$ and $|Q_+ - P_+|$ has the same minimum value as $|Q_- - P_+|$. If $a = a(\phi) = a(90°)$, each wire at 0 touches both wires of the other pair.

Corollary: If pairs have straight parallel axes and twist at the same rate, the packing fraction cannot exceed

$$2\pi/\{27^{1/2} + 24^{1/2}\} = 0.62240.$$

Proof: Let $2R$ denote $(2 + 2^{1/2})r$. In any packing of twisted pairs, the distance between two pair axes is at least $2R$. Then circles of radius R, centered at the pair axes of a packing, do not overlap. The number ρ of circles per unit area, in any packing of nonoverlapping circles of radius R, satisfies

$$\rho \le \rho_0 = 1/(12^{1/2}R^2).$$

The maximum ρ_0 would be attained with Fig. 1 again, now with circles of radius R. Each circle of radius R represents one twisted pair of area $2\pi r.^2$ Then the packing fraction is $f = 2\pi r^2 \rho \le 2\pi r^2 \rho_0$. The bound simplifies to the number stated by the corollary.

One might try to achieve density 0.62240 by arranging pair axes in the same pattern as the centers in Fig. 1. Each pair would then be required to differ in phase by ±90° from each neighbor at distance $2R$. But that arrangement contains triples of pairs, each pair a neighbor of the other two. There is no way to assign phases to the pairs of such a triple. Packing fractions near 0.62240 are probably not obtainable. The more special packings in the next section have maximum packing fraction 0.56767.

IV. LATTICES

A *point lattice* is a discrete set of points forming a group under vector addition. Thus a point lattice must contain the origin 0 and the sum $P \pm Q$ of each pair of lattice points P, Q. Two-dimensional point lattices, the ones of interest here, can all be generated from pairs u, v of linearly independent vectors. Lattice points are then linear combinations

$$P_{ij} = iu \pm jv \tag{7}$$

of the generator vectors u, v, the coefficients i, j ranging over all integers. For example, circle centers in Fig. 1 form a point lattice with generators u, v both of length $2r$ and 60° apart.

A point lattice in a plane π can be used to construct a *lattice of twisted pairs*. Arrange pairs, all twisting at the same rate and having parallel straight axes normal to π, with a pair axis passing through each point of the point lattice. Let $\theta(P)$ denote the phase of the pair with axis at point P of π. The phases will be required to satisfy

$$\theta(P + Q) = \theta(P) + \theta(Q).$$

Then $\theta(0) = 0$ and all phases can be expressed in terms of twò parameters $\sigma = \theta(u)$, $\tau = \theta(v)$. At P_{ij} in (7), the phase $\theta_{ij} = \theta(P_{ij})$ is

$$\theta_{ij} = \theta(iu) + \theta(jv) = i\sigma + j\tau. \tag{8}$$

In a point lattice, each point P is like every other point Q. Adding $P - Q$ to every point merely translates the point lattice rigidly onto itself and carries Q to P. The lattice of twisted pairs has symmetries which are only slightly more complicated. The translation that carries Q to P need not leave the lattice of twisted pairs fixed because the phases $\theta(P)$ and $\theta(Q)$ may differ. However, (8) shows that the lattice of twisted pairs regains its original appearance if this translation is followed by turning every pair through a constant angle $\theta(P) - \theta(Q)$. With pairs that all twist at the same rate, rotating pairs through a fixed angle is equivalent to taking a different cross section through the wires. Thus the twisted pairs do have translation symmetries, although the translations have axial components.

A point lattice determines a tessellation of the plane into congruent parallelogram cells (for a detailed explanation, see Ref. 4). Each lattice point P determines a parallelogram cell with vertices P, $P + u$, $P + v$, $P + u + v$. A cell has area

$$A = |u|\,|v|\,|\sin\alpha|, \tag{9}$$

where α is the angle between u and v. The lattice points have density $\rho = 1/A$ points per unit area. Then a lattice of twisted pairs has packing fraction

$$f = 2\pi r^2/A. \tag{10}$$

To find a lattice of twisted pairs that maximizes f, one must find parameters $|u|$, $|v|$, α, σ, τ that minimize A. These parameters are allowed only values that keep wires from intersecting. For each pair P, P' of lattice points, with phases θ, θ', the separation $|P - P'|$ must be at least $a(\theta - \theta')$ as given by (6). That optimization problem has the following solution.

Theorem 2: The maximum packing fraction of lattices of twisted pairs is

$$f = \tfrac{1}{2}\pi/(2 + 32^{1/2})^{1/2} = 0.56766836 \cdots .$$

It is obtained with $\sigma = 0°$, $\tau = 90°$, $|u| = 4r$, $|v| = |v - u| = (2 + 2^{1/2})r$, *and* $\cos\alpha = 1/(1 + 2^{-1/2})$, *i.e.,* $\alpha = 54.14143°$.

The proof is long and will be deferred to the appendix. Figure 4 shows one cross section through the maximizing lattice of twisted pairs given by Theorem 2. Figure 4 also shows the parallelogram cells, determined by the generator vectors u and v. The vertices of parallelograms form the point lattice, representing the axes of the twisted pairs. The generators used in Fig. 4 are $u = (4r, 0)$, $v = (2, (2 + 32^{1/2})^{1/2})r = (2r, 2.7671r)$. The cell area is $A = 4(2 + 32^{1/2})^{1/2}r^2 = 11.06841r^2$.

In Fig. 4, certain wires touch. These contacts occur at midpoints of half the horizontal sides of cells. At two other places in the interior of each cell, wires almost touch. The very short gap between these wires is not apparent in a small drawing.

One of the parallelograms is shaded. Figure 5 shows what happens in this shaded cell as pairs rotate. The rotation angles, 9.141°, 80.859°, 90°, etc., were chosen to show contacts that occur between wires.

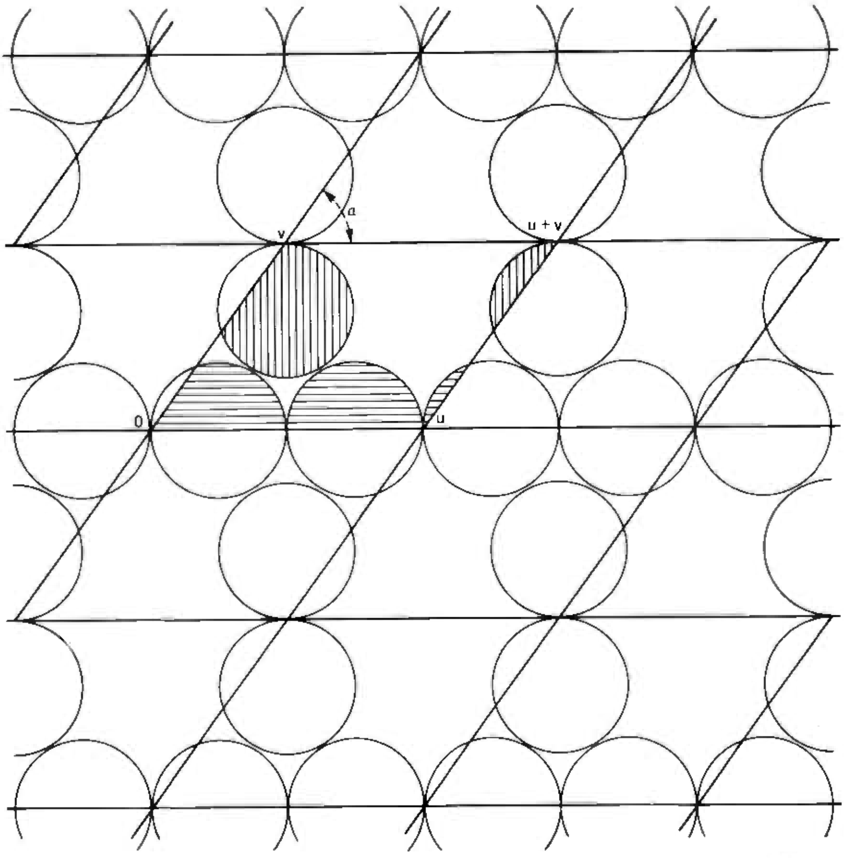


Fig. 4—One cross section through the densest lattice of twisted pairs, $\theta = 0°$.

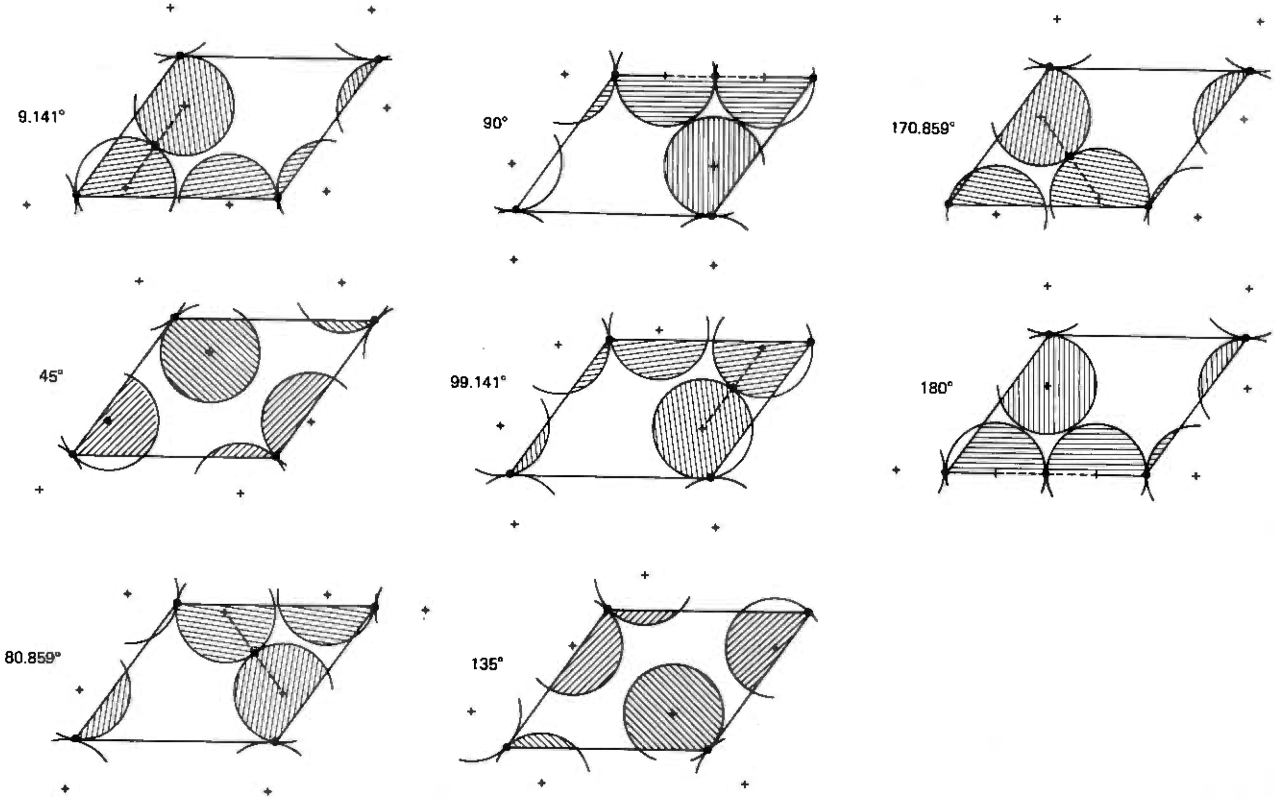


Fig. 5—History of a typical cell of the densest lattice of twisted pairs as θ increases.

Again there are some near misses (at 90°) and so Fig. 5 marks each point of true contact by a dot. A 180° rotation interchanges the wires in each pair and so Fig. 5 does not go beyond 180° rotation. Because of the lattice symmetry, each cell of the lattice goes through the same cycle as any other cell, but perhaps with different phase. In this special lattice, half the cells are in phase with the shaded cell and the rest are 90° out of phase.

Another view of the packing follows a single wire as it makes one full turn about its pair axis. Figure 6 shows the successive positions a, b, c, $\cdots$, j of the center of one wire as it comes into contact with other wires. Thus, a, b, c, $\cdots$ lie on a circle of radius r at the angles 9.141°, 80.859°, 99.141°, etc. The wire itself is not drawn but the wires it touches appear in their positions at contact. Each point a, b, c, $\cdots$ is connected by a line to the center of the contacted circle. The second wire of the chosen pair makes another contact whenever the first wire does but from a point 180° away. Thus the second wire goes through the cycle f, g, h, i, j, a, $\cdots$, e.

The contacts shown in Figs. 4, 5, and 6 occur because each pair has neighboring pairs at exactly the minimum allowed distance (6). In particular, $|u| = a(0°)$, $|v| = a(90°)$, $|v - u| = a(90°)$ and so the pair at P makes contact with the six pairs at $P \pm u$, $P \pm v$, and $P \pm (v - u)$. The pairs at $P \pm v$ and $P \pm (v - u)$ differ in phase from the one at P by 90°. Thus, as mentioned following eq. (6), a wire at P will touch

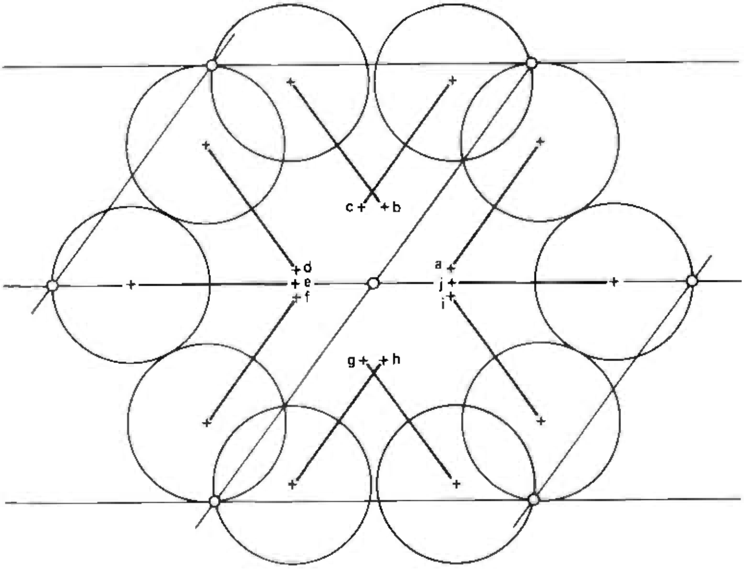


Fig. 6—The wires touched by a given wire of the densest lattice packing during one complete turn. Circular dots mark the pair axes.

both wires of pairs at $P \pm v$ and $P \pm (v - u)$. That accounts for the 8 contacts at positions a, c; b, d; f, h; g, i. At $P \pm u$, with phase difference 0°, only one wire is touched (at j and e). Thus the wire in Fig. 6 touches 10 wires belonging to 6 pairs.

The fact that each wire touches 10 wires belonging to 6 other pairs is an indication that the packing is very tight. The appendix proves the following theorem.

Theorem 3: Suppose all pairs have straight parallel axes and twist at the same rate. Then no wire can touch more than 6 other pairs nor more than 10 wires belonging to other pairs.

Note that the hypotheses of the theorem apply to packings more general than lattice packings.

Three twisted pairs will be said to form a *triplet* if each pair touches the other two. Pairs with axes P_1, P_2, P_3, and phases θ_1, θ_2, θ_3 form a triplet if $|P_i - P_j| = a(\theta_i - \theta_j)$ for the 3 choices of distinct subscripts i, j. One might expect many triplets in a dense packing, lattice or otherwise. Theorem 3 shows that no pair can belong to more than 6 triplets; the lattice packing of Theorem 2 achieves that number. In fact, each parallelogram cell in Fig. 4 is formed from two triangles, having vertices at axes of a triplet. Thus triplet triangles cover the entire plane of Fig. 4. Moreover, these triplet triangles have the least area possible.

Theorem 4: Suppose three twisted pairs with parallel straight axes and the same twisting rate form a triplet. The triangle with vertices at the axes of the three pairs has area at least

$$(2 + 32^{1/2})^{1/2} r^2 = 5.53420 r^2.$$

This minimum area is achieved if the three phase differences are 0°, ±90°, *and* ±90°.

This theorem is proved in the appendix.

V. CONCLUSION

Theorem 2 gives the packing fraction $f = 0.56767$ of the densest packing of twisted pairs having

(*i*) Pairs with parallel straight axes.

(*ii*) The same twisting rate for all pairs.

(*iii*) A lattice arrangement of pair axes and phases.

It may be that assumption (*iii*) can be dropped. Without assuming (*iii*), Theorems 3 and 4 show that pairs in the packing of Theorem 2 are as "close together" as possible in two senses that are not directly connected with f. But at present, without assuming (*iii*) one can only guarantee $f < 0.62240$ (Corollary to Theorem 1). To extend Theorem 2 without assuming (*iii*) may be difficult. Even for circles, the density maximizing property of Fig. 1 has only recently been proved without a lattice assumption.[3] For spheres in three dimensions, the astronomer J. Kepler conjectured that the face-centered cubic lattice packing is densest possible. Two centuries later, Gauss proved the conjecture for lattice packings, but there has been no proof without a lattice assumption.

Assumption (*i*) is too strong for most applications. It would be desirable to drop (*i*) and assume only that different pairs fail to intertwine. Relaxing (*i*) probably permits some increase in packing fraction. Figures 4 and 5 show that no pair ever touches more than two wires at a time in any cross section. Then, in any cross section, every pair is free to move slightly. If a bundle of pairs, packed as in Figs. 4 and 5, were surrounded by a cord and tied tightly, the axes would bend and the pairs would assume a denser packing in the plane of the cord. It should be possible to bend axes in Figs. 4 and 5 to obtain a packing fraction $f > 0.56767$ in all cross sections simultaneously.

In regard to assumption (*ii*), Section III mentioned a denser packing with pairs twisting in opposite senses, although still with the same absolute twist rates.

APPENDIX

Proofs of Theorems 2, 3, and 4

A1. Theorem 2, Part 1

Theorem 2 will be proved in two parts. The first part subjects a given lattice of twisted pairs to deformations that increase f and leave the lattice with generators u, v such that twisted pairs at 0, u, and v form a triplet. The second part of the proof is then also a proof of Theorem 4.

Many choices of u, v, σ, τ produce the same lattice of twisted pairs. For example, given one set of generators, another is obtained by changing v to $u + v$ and τ to $\sigma + \tau$. The pairs remain packed exactly as before although the point now called P_{ij} in (7) is the one formerly called $P_{i,i+j}$. This freedom to choose among different generators is used in the first part to obtain generators u, v with simplifying properties.

The deformations in the first part must be performed without causing wires to intersect. Because the pairs are symmetric to one another as explained in Section IV, it suffices to ensure that the pair at 0 never intersects a pair at any other point P. From (6) one obtains the requirement

$$|P| \geq 2r\{1 + M(\theta(P))\}. \tag{11}$$

Or, if $R(P)$ is defined to be the ratio

$$R(P) = |P|/\{2r[1 + M(\theta(P))]\},$$

the requirement is $R(P) \geq 1$ for all $P \neq 0$.

If a given lattice of twisted pairs has Min $R(P) = R_0 > 1$, then the lattice of P axes can be shrunk, moving each point P to P/R_0. Shrinking the lattice would increase f by a factor R_0^2. Hence a lattice of twisted pairs with maximum density must have points P such that $R(P) = 1$. One of these points will be taken for the generator vector u; $R(u) = 1$. This choice also determines $\sigma = \theta(u)$.

Figure 7 shows the point lattice of twisted pair axes. The points may be grouped in rows parallel to a central row of points $\cdots, -u, 0, u, 2u, \cdots$. These are horizontal rows in Fig. 4. Since $R(u) = 1$, $|u| \le 2r\{1 + M(0)\} = 4r$. From (11), any point P, except the origin 0, satisfies $|P| \ge 2r\{1 + M(90°)\} = (2 + 2^{1/2})r$. Thus

$$|P| > 0.85\,|u|.$$

Similarly, any point P except ku satisfies $|P - ku| > 0.85\,|u|$. Then it follows that the distance between the horizontal rows of points is at least $0.68\,|u|$. Now start to compress Fig. 7 linearly in a vertical direction. The compression only increases the packing factor. The compression must stop before the vertical separation between rows

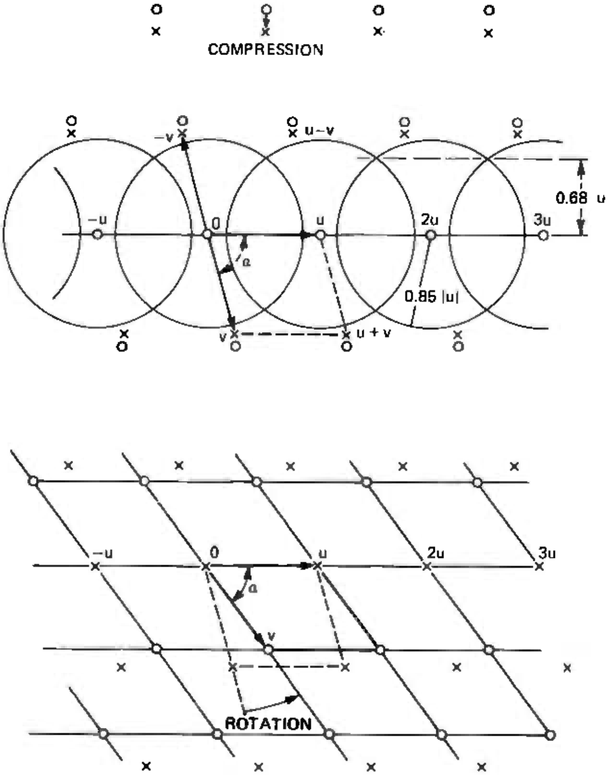


Fig. 7—Deformations of a lattice to increase packing fraction. Above: compression to produce a point v with $R(v) = 1$. Below: rotation of v to make $R(v - u) = 1$.

reaches $0.68\,|u|$ because, for some P, the ratio $R(P)$ will become equal to 1. A point P linearly independent of u and with $R(P) = 1$ can only lie in a row directly above or below the central row. For in a more remote row,

$$|P| > 2 \times 0.68\,|u| \geq 1.36(2 + 2^{1/2})r > 4r;$$

then $R(P) > 1$. A point P with $R(P) = 1$ and linearly independent of u will now be taken as the second generating vector v of the lattice. Because $R(-P) = R(P)$ there will be a choice between two vectors for v. Pick the vector that lies within 90° of u, as shown in Fig. 7; then $|\alpha| \leq 90°$ in (9). That determines $\tau = \theta(v)$.

Since $R(u) = R(v) = 1$, $|u|$ and $|v|$ are as small as (11) allows, for the given phases σ and τ. Another deformation of the lattice, changing α in (9) while holding $|u|$, $|v|$, σ, and τ fixed, may increase f. Because $|\alpha| \leq 90°$ the change must be in the direction of decreasing $|\alpha|$. The requirement $R(v - u) \geq 1$ sets a lower limit on the size of $|\alpha|$. Since $|u|$ and $|v|$ lie between $(2 + 2^{1/2})r$ and $4r$, $|v - u|$ would become smaller than $(2 - 2^{1/2})r$ and have $R(v - u) \leq (2 - 2^{1/2})/(2 + 2^{1/2}) < 1$ at $\alpha = 0$. Thus, with $R(u) = R(v) = 1$ and fixed σ and τ, A in (10) is always at least as large as the value given by (9) with $|\alpha|$ determined from the condition $R(v - u) = 1$. When $R(v - u) = R(u) = R(v) = 1$, twisted pairs at 0, u, v form a triplet and pairs at u, v, $u + v$ form another (recall the definition of triplet, given following Theorem 3). The two congruent triplet triangles $(0, u, v)$ and $(u, v, u + v)$ together constitute the parallelogram cell $(0, u, v, u + v)$. Given σ and τ, (10) shows that f is no greater than πr^2 divided by the area of a triplet triangle 0, u, v with $R(u) = R(v) = R(v - u) = 1$; i.e.,

$$|u| = 2r\{1 + M(\sigma)\} \tag{12}$$

$$|v| = 2r\{1 + M(\tau)\} \tag{13}$$

$$|v - u| = 2r\{1 + M(\tau - \sigma)\}. \tag{14}$$

The first part of the proof of Theorem 2 is now finished. It remains to adjust σ and τ to minimize the area of the triplet triangle 0, u, v. That will lead to $\sigma = 0°$, $\tau = 90°$, and prove Theorem 4. Another detail to verify is that the lattice determined by (12), (13), and (14) with $\sigma = 0°$, $\tau = 90°$ actually has $R(P) \geq 1$ for all $P \neq 0$. Since $R(P) \leq 1$ only if $|P| \leq 4r$, there are only a few lattice points to examine. A short calculation shows $R(P) \geq 1$, with equality holding only for $P = \pm u$, $\pm v$, and $\pm(v - u)$. These six vectors locate the axes of the six twisted pairs mentioned in Theorem 3.

A2. Theorem 2, Part 2 and Theorem 4

The proof of Theorem 4 will use a formula, of Heron of Alexandria, for the area of a triangle with sides of given lengths a, b, c.[5,6] Here

Table I—Packing fraction f of lattices of twisted pairs

τ \ σ =	0°	10°	20°	30°	40°	50°	60°
0°	0.4534						
10°	0.4546						
20°	0.4581	0.4569					
30°	0.4640	0.4616					
40°	0.4725	0.4689	0.4676				
50°	0.4839	0.4788	0.4763				
60°	0.4985	0.4918	0.4878	0.4865			
70°	0.5168	0.5082	0.5025	0.4997			
80°	0.5396	0.5286	0.5210	0.5165	0.5150		
90°	0.5677	0.5540	0.5439	0.5373	0.5341		
100°			0.5418	0.5332	0.5281	0.5265	
110°					0.5262	0.5227	
120°							0.5209

$a = |v - u|$, $b = |u|$, $c = |v|$, which depend on σ, τ as in (12), (13), (14). Heron's formula converts the cell area A in (9) (twice the area of a triplet triangle) to

$$A = \{s(s-a)(s-b)(s-c)\}^{1/2}, \tag{15}$$

where $s = (a + b + c)/2$. Table I shows how the packing fraction f, obtained from (15) and (10), depends on σ and τ.

Table I shows only values of σ and τ in the range

$$0 \le 2\sigma \le \tau \le 90^\circ + \tfrac{1}{2}\sigma. \tag{16}$$

Values of f for other angles can be obtained by exploiting symmetries in formulas (12), (13), (14), (15). Write $(\sigma', \tau') \approx (\sigma, \tau)$ if substituting σ', τ' for σ, τ leaves the three lengths a, b, c unchanged, except perhaps for a permutation. For example, $(\sigma + 180^\circ, \tau) \approx (\sigma, \tau) \approx (\sigma, \tau + 180^\circ)$ because $M(\theta)$ is a function with period 180°. Then σ and τ can be assumed nonnegative. Second, $(\tau - \sigma, \tau) \approx (\sigma, \tau)$, and so one can assume $\sigma \le \tau - \sigma$, i.e., $2\sigma \le \tau$. Finally $M(\theta)$ has a symmetry $M(180^\circ - \theta) = M(\theta)$, and so $(\sigma, 180^\circ - \tau + \sigma) \approx (\sigma, \tau)$ and $(180^\circ - \tau, 180^\circ - \tau + \sigma) \approx (\sigma, \tau)$. It suffices to require $\tau \le 180^\circ - \tau + \sigma$ or $\tau \le 90^\circ + \tfrac{1}{2}\sigma$.

Table I indicates a maximum of f at $\sigma = 0^\circ$, $\tau = 90^\circ$. However, for the sake of mathematical exactness, an analytical proof follows.

First, note that A is an increasing function of a, b, and c. To prove this, differentiate A^2 with respect to these variables. For example,

$$(8A/a)\frac{\partial A}{\partial a} = 2bc - a^2 \ge 2(2 + 2^{1/2})^2 - 4^2 > 0. \tag{17}$$

It now follows that A cannot have a maximum in the part of the set (16) where $\tau < 90^\circ$. For, in that part $M(\tau) \approx \cos \tau/2$ and, because $\tau - \sigma < 90^\circ$, too, $M(\tau - \sigma) = \cos\{(\tau - \sigma)/2\}$. For fixed σ, $|v|$ and $|v - u|$ are decreasing functions of τ while $|u|$ remains constant. Then (17) shows that A is decreasing and hence f can have no local maximum with $0 \le \tau < 90^\circ$.

The remaining portion of the range (16), with 90° ≤ τ, can be cut into three parts. These are

$$0 < \sigma \le 45^\circ, \qquad \tau = 90^\circ, \tag{18}$$

$$0 < \sigma \le 45^\circ, \qquad 90^\circ < \tau \le 90^\circ + \tfrac{1}{2}\sigma, \tag{19}$$

and

$$45^\circ \le \sigma, \qquad 2\sigma \le \tau \le 90^\circ + \tfrac{1}{2}\sigma \tag{20}$$

[note that the second inequality of (20) actually implies $45^\circ \le \sigma \le 60^\circ$]. In all three parts,

$$\begin{aligned} M(\sigma) &= \cos \sigma/2 \\ M(\tau) &= \sin \tau/2 \\ M(\tau - \sigma) &= \cos\{(\tau - \sigma)/2\}. \end{aligned} \tag{21}$$

Consider the range (20) first. Since $45^\circ \le \sigma \le 60^\circ$, one has $90^\circ \le \tau \le 120^\circ$ and $45^\circ \le \sigma \le \tau - \sigma \le 90^\circ - \tfrac{1}{2}\sigma \le 67.5^\circ$. Then

$$\begin{aligned} |u| &\ge (2 + 2^{1/2})r = 3.41421r \\ |v| &\ge (2 + 3^{1/2})r = 3.73205r \\ |v - u| &\ge (2 + 2 \cos 33.75^\circ)r = 3.66294r. \end{aligned}$$

When these minimum lengths are substituted for a, b, c in Heron's formula, one obtains a lower bound $A > 11.19573r^2$ and hence $f < 0.56121$ throughout (20). Thus these parameters σ, τ can never minimize A nor give a packing fraction as high as $f = 0.567668$, which is obtained with $\sigma = 0^\circ$, $\tau = 90^\circ$.

Next consider (19). Those inequalities imply $0 < \sigma \le 45^\circ$ and $90^\circ \le \tau \le 112.5^\circ$ so that

$$M(\tau) \le M(112.5^\circ) = M(67.5^\circ) < M(45^\circ) \le M(\sigma)$$

and

$$|v| < |u|. \tag{22}$$

To show that there is no local maximum of f with σ, τ satisfying (19), consider a small change from σ, τ to $\sigma + x$, $\tau + x$. Changing σ and τ by the same amount keeps $|v - u|$ constant but changes $|v|$ and $|u|$ in opposite directions. The effect on A is determined by differentiating. Equation (15) provides

$$\begin{aligned} 8A\frac{dA}{dx} &= |u|\,(2\,|v|\,|v - u| - |u|^2)\,\frac{d|u|}{dx} \\ &\quad + |v|\,(2|u||v - u| - |v|^2)\,\frac{d|v|}{dx}. \end{aligned} \tag{23}$$

The derivatives of $|u|$ and $|v|$ are obtainable from (12), (13), and (21):

$$\frac{d|u|}{dx} = -r \sin \sigma/2 < 0, \qquad \frac{d|v|}{dx} = r \cos \tau/2 > 0.$$

Then the inequality (22) can be used to simplify (23) to a bound

$$\begin{aligned}(8A/r)\frac{dA}{dx} &> |u|(2|v||v-u| - |v|^2)(-\sin \sigma/2)\\ &\quad + |v|(2|u||v-u| - |u||v|) \cos \tau/2\\ &= |u||v|(2|v-u| - |v|)(\cos \tau/2 - \sin \sigma/2).\end{aligned}$$

The inequalities (19) imply $\sin \sigma/2 < \sin 22.5°$ and $\cos \tau/2 > \cos 56.25° = \sin 33.75° > \sin \sigma/2$. Also, $2|v-u| - |v| \geq 2(2 + 2^{1/2})r - 4r > 0$. Then $dA/dx > 0$ in the range (19) and there can be no local maximum there.

The proof so far has shown that f is too small in range (20) to achieve a maximum there and that, elsewhere with $\tau \neq 90°$, f increases if (σ, τ) moves toward the line $\tau = 90°$. One must consider (18) and show that $dA/d\sigma > 0$ with τ fixed at 90°. Then $0 < \sigma \leq 45° < \tau - \sigma \leq 90° = \tau$, which implies $|v| < |v-u| < |u|$ because

$$\begin{aligned}|u| &= 2(1 + \cos \sigma/2)r\\ |v| &= (2 + 2^{1/2})r\\ |v-u| &= 2(1 + \cos\{(90° - \sigma)/2\})r.\end{aligned}$$

Now a formula like (23) may be written for $dA/d\sigma$. The proof that $dA/d\sigma > 0$ is similar to the one given for the range (19), here using the inequality $|v-u| < |u|$. That completes the proof of Theorems 2 and 4.

A3. Theorem 3

To prove Theorem 3, consider a wire twisting about an axis at the origin 0. Let $P_1, P_2, \cdots, P_K$ denote axes of neighboring pairs that this wire touches. The names $P_1, P_2, \cdots, P_K$ may be assigned in order of increasing polar angle about 0. Magnitudes $|P_k|$ must satisfy (6) with ϕ the phase difference $\phi_k = \theta(P_k) - \theta(0)$. Thus all $|P_k|$ lie between $(2 + 2^{1/2})r$ and $4r$. Also $|P_k - P_j| \geq (2 + 2^{1/2})r$. Let t_k denote the number of times the wire at 0 touches wires of the pair at P_k. Then $t_k = 2$ if $|P_k| = (2 + 2^{1/2})r$ (i.e., if $\phi_k = \pm 90°$), but otherwise $t_k = 1$. The total number of contacts is

$$T = t_1 + t_2 + \cdots + t_K$$

and the theorem states $K \leq 6$ and $T \leq 10$.

Let α_k denote the angle $P_k 0 P_{k+1}$. The K angles $\alpha_1, \cdots, \alpha_K$ fall into three types.

Type 1: If $t_k = t_{k+1} = 2$, then $|P_k| = |P_{k+1}| = (2 + 2^{1/2})r$ and $\phi_k = \pm\phi_{k+1} = \pm 90°$. Then $|P_{k+1} - P_k| \geq a(0°) = 4r$ and $\alpha_k \geq 71.7°$ follows from the cosine law.

Type 2: If $t_k = 2$ but $t_{k+1} = 1$, then $|P_k| = (2 + 2^{1/2})r$, $|P_{k+1}| \leq 4r$ and $|P_{k+1} - P_k| \geq (2 + 2^{1/2})r$. Then $\alpha_k \geq 54.1°$. The same bound holds if $t_k = 1$ and $t_{k+1} = 2$.

Type 3: If $t_k = t_{k+1} = 1$ then $|P_k|$ and $|P_{k+1}|$ may be as large as $4r$ and $|P_{k+1} - P_k| \geq (2 + 2^{1/2})r$. Then $\alpha_k \geq 50.5°$.

Let N_1, N_2, N_3 be the numbers of angles α_k of types 1, 2, 3. Then

$$71.7°\ N_1 + 54.1°\ N_2 + 50.5°\ N_3 \leq 360°. \tag{24}$$

Moreover,

$$T = \tfrac{1}{2}\{(t_1 + t_2) + (t_2 + t_3) + \cdots + (t_K + t_1)\}$$

and each term $(t_k + t_{k+1})$ has value 4, 3, or 2 according to the type 1, 2, or 3 of α_k. Thus

$$T = \tfrac{1}{2}\{4N_1 + 3N_2 + 2N_3\}. \tag{25}$$

Subject to the constraint (24), nonnegative integers N_1, N_2, N_3 give T a maximum value $T = 10$. That proves half of the theorem.

The other half is more delicate because the constraint (23) allows $N_3 = 7$, $N_1 = N_2 = 0$, $K = N_1 + N_2 + N_3 = 7$. An improved bound on α_k for type 3 is needed to prove $K \leq 6$. The angle $\alpha_k = 50.5°$ is not actually achievable because it would require both ϕ_k and ϕ_{k+1} to be 0° or 180° while $\phi_{k+1} - \phi_k = \pm 90°$. For given ϕ_k and ϕ_{k+1}, the smallest α_k is obtained with $|P_k| = a(\phi_k)$, $|P_{k+1}| = a(\phi_{k+1})$ and $|P_{k+1} - P_k| = a(\phi_{k+1} - \phi_k)$. Then the cosine law determines α_k as a function of ϕ_k and ϕ_{k+1}. The minimum α_k is found to occur at $\phi_{k+1} = 135°$, $\phi_k = 45°$. The details will be omitted. In this way, one finds $52.67° \leq \alpha_k$, $K \leq [360°/52.67°] = 6$, and the theorem is proved.

In (25) there are actually two ways to make $T = 10$. The solution $N_1 = 2$, $N_2 = 4$, $N_3 = 0$ corresponds to Figs. 4 and 5. Another solution $N_1 = 5$, $N_2 = N_3 = 0$ can represent an isolated arrangement of five twisted pairs with phase 90°, surrounding a central pair with phase 0°. That configuration cannot occur as part of a lattice. A lattice would also contain a pair at $P_2 - P_1$ with phase 0°, but that pair would conflict with the one at P_3.

REFERENCES

1. E. N. Gilbert, "Randomly Packed and Solidly Packed Spheres," Canadian J. Math., *16* (1964), pp. 286–298.

2. L. Fejes Tóth, *Lagerungen in der Ebene auf der Kugel und im Raum*, Berlin: Springer, 1953.
3. L. Fejes Tóth, *Regular Figures*, N.Y.: Macmillan (Pergamon), 1964.
4. D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, N.Y.: Chelsea, 1952.
5. Sir Thomas L. Heath, *A Manual of Greek Mathematics*, Oxford: Clarendon Press, 1931.
6. H. S. M. Coxeter, *Introduction to Geometry*, N.Y.: John Wiley, 1961.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

A General Characterization of Splice Loss for Multimode Optical Fibers

By S. C. METTLER

(Manuscript received June 20, 1979)

The Gaussian point transmission model for calculating optical fiber splice loss is extended to the general case of splice loss between fibers which differ in one or more intrinsic parameters—core radius, index of refraction profile shape, and maximum index of refraction difference between core and cladding. The model is first verified for splices with index-of-refraction profile mismatch. The average difference between calculated and measured splice loss due to profile parameter mismatch is 0.04 dB. Comparisons are also made between calculated and measured splice loss for ten different splices with mismatch in all three intrinsic parameters. The average difference between the calculated loss and the average of several measured losses for these ten cases was 0.06 dB. The additional losses introduced by transverse offset measured for one set of mismatched fiber splices agree with the calculated values within 0.1 dB. Loss due to misalignment of elliptical core fibers is calculated and measured with agreement within 0.06 dB for the maximum loss case. Both the model and the experimental data show that, for a given percentage mismatch, index of refraction profile parameter mismatch and core ellipticity contribute significantly less to splice loss than mismatch of core radius or numerical aperture. A family of curves for splice loss vs transverse offset is presented for various numerical aperture mismatches and core radius mismatches, since these parameters are typically the largest components of splice loss in practical fiber optic systems.

I. INTRODUCTION

One factor which must be considered in the development of fiber optical communication systems is the effect of fiber core parameter manufacturing variations on splice loss. These intrinsic fiber core parameters[1] are the maximum index-of-refraction difference between core and cladding, Δ, the index of refraction profile parameter, α, the radius, R, and core ellipticity, ϵ.

The development of a phenomenological Gaussian point loss model[2] allows an approximate analytic treatment of the loss induced by a butt-joint splice between fibers which differ in one or more intrinsic parameters. Previous models based on the uniform power distribution assumption[1] have exhibited only limited agreement between calculated and measured splice loss,[3, 4] whereas the Gaussian model gives good agreement with experimental data.

Previous work[2] developed the Gaussian point loss model and gave theoretical and experimental results for Δ mismatch, R mismatch and transverse offset. This model has been used to estimate system losses due to random splicing between fibers whose intrinsic parameter variation distributions are known.[5] An extension of this model to the case of index-of-refraction profile (α) mismatch or ellipticity (ϵ), and to combinations of intrinsic factors plus transverse offset, is presented in this paper. Although the extension of the model is straightforward, the resulting calculations require the use of approximate numerical quadrature techniques applied over irregular areas of integration. This paper presents a unified analysis of the various effects of intrinsic factors and transverse offset on splice loss utilizing the Gaussian model. Representative experimental data are also presented.

II. THEORETICAL DEVELOPMENT

An approximate analytical treatment of the loss in a butt-joint splice between two fibers with differing intrinsic parameters can be carried out using a Gaussian model.[2] This model assumes that a steady-state power distribution (after a long length of fiber) can be modeled as a Gaussian distribution of the power within the solid angle defined by the local numerical aperture, NA, at any point on the fiber core (Fig. 1). With this assumed power distribution at the input to a splice, the ratio of the power received, p_2, at any radial position, r, to the power transmitted, p_1, at that point is related to the ratio of the squares of the NA's at that point by the following equation:[2]

$$t(r) = \frac{p_2}{p_1} = \begin{cases} 1 + \dfrac{(\mathrm{NA}_2)^2}{(\mathrm{NA}_1)^2} p_0 - e^{(\mathrm{NA}_2)^2/(\mathrm{NA}_1)^2 \ln p_0} & \text{for} \quad \mathrm{NA}_2 < \mathrm{NA}_1 \\ 1 & \text{for} \quad \mathrm{NA}_2 > \mathrm{NA}_1, \end{cases} \tag{1}$$

where p_0 is the point defining the width of the Gaussian distribution corresponding to 0.1 of its maximum. The expression $t(r)$ is called the point transmission function.

Using the usual class of circularly symmetric profiles,[1]

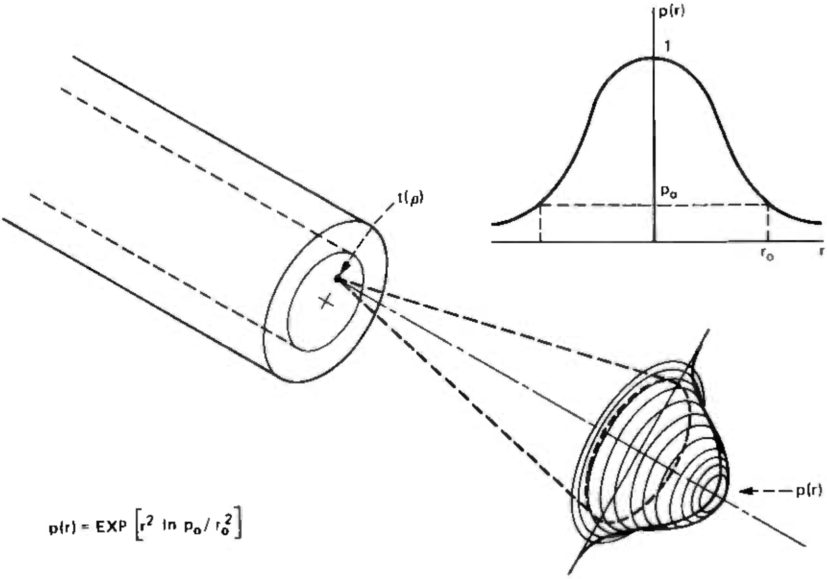


Fig. 1—Gaussian power distribution.

$$\mathrm{NA}(r) \simeq n_0 \sqrt{2\Delta} \left[1 - \left(\frac{r}{R} \right)^{\alpha} \right]^{1/2} \tag{2}$$

where $\Delta \simeq (n_0 - n_c)/n_0$ is small
n_0 = maximum core refractive index
n_c = refractive index of cladding
α = index profile parameter $[\alpha(r) = \alpha]$
R = fiber core radius.

Substituting eq. (2) into eq. (1) and defining

$$Q = \frac{(\mathrm{NA}_2)^2}{(\mathrm{NA}_1)^2} = \frac{\Delta_2(1 - r^{\alpha_2})}{\Delta_1(1 - k^{\alpha_1} r^{\alpha_1})} \tag{3}$$

gives the transmission coefficient at a radial distance r from the core axis,

$$t(r) = \begin{cases} 1 + Qp_0 - p_0^Q & \text{for} \quad Q < 1 \\ 1. & \text{for} \quad Q \geq 1, \end{cases} \tag{4}$$

where $k = R_2/R_1$ with R_2 normalized to 1. Subscripts 1 and 2 refer to the transmitting and receiving fibers respectively.

To find the total transmission through the splice, the point transmission function is multiplied by the input power of the transmitting fiber at each point and integrated over the area of core overlap. A transmission ratio is obtained by dividing by the total input power:

$$T = \frac{P_2}{P_1} = \frac{\displaystyle\int_0^{2\pi}\int_0^{x} t(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r\,dr\,d\theta}{\displaystyle\int_0^{2\pi}\int_0^{1/k} (1 - k^{\alpha_1} r^{\alpha_1})^2 r\,dr\,d\theta}, \tag{5}$$

where x = the lesser of 1 or $1/k$.

For circularly symmetric profiles with no transverse offset:

$$T = \frac{P_2}{P_1} = \frac{\displaystyle\int_0^{x} t(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r\,dr}{\alpha_1^2/2k^2(\alpha_1 + 1)(\alpha_1 + 2)}. \tag{6}$$

This formula represents the total intrinsic loss immediately after the splice. Substitution of eq. (4) for $t(r)$ into (6) results in an integral which must be solved numerically for $t(r) \neq 1$ except for the special case of $\alpha_1 = \alpha_2$ and $R_1 = R_2$.

The inclusion of transverse offset in the problem requires the evaluation of the double integral in the numerator of eq. (5) over the area of overlap of the two fiber cores as shown in Fig. 2:

$$\frac{P_2}{P_1} = \frac{\displaystyle\int\int t(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r\,dr\,d\theta}{2\pi\alpha_1^2/2k^2(\alpha_1 + 1)(\alpha_1 + 2)}. \tag{7}$$

Equation (7), when integrated over the area of core overlap, represents a general solution for any combination of intrinsic mismatch and transverse offset for short receiving fibers. Long receiving fibers require the use of the long length $t(r)$ as given in Ref. 2. The splice loss can be calculated by evaluating Q at each point and integrating eq. (7) numerically using the appropriate expression for $t(r)$. This numerical integration can sometimes be simplified by careful consideration of the particular splice parameters. The absence of transverse offset, for example, reduces the numerator to a single integral in r for circularly symmetric fibers. These results assume flat, clean, perpendicular fiber ends with index-of-refraction matching fluid and no angular or longitudinal offset. In addition, a steady-state distribution is assumed at the input to the splice, requiring either a long input fiber or a launched power distribution approximating the steady state.

III. COMPARISON OF EXPERIMENTAL AND THEORETICAL VALUES

Experimental verification of the Gaussian model has followed a step-by-step procedure. Splice losses were measured between selected sets

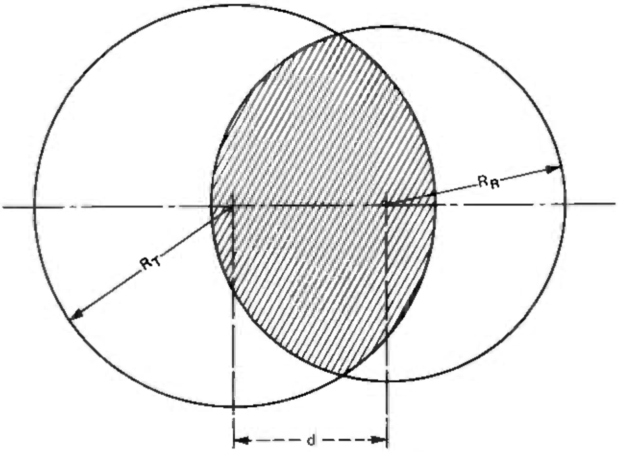


Fig. 2—Overlap regions for offset fiber cores.

of fibers which differed primarily in one intrinsic parameter. The results for Δ and radius mismatch were previously reported in Ref. 2, along with transverse offset results for both short and long receiving fibers. Results for α mismatch and combinations of intrinsic mismatch and transverse offset are given here.

All measurements were made using a pulsed, 0.82-μm, laser source with a pigtail which was loose-tube-spliced[6] to the transmitting fiber. The minimum length of the transmitting fiber was 550 m, except for one step index fiber (~250 m), to provide an approximate steady-state power distribution at the splice. The transmitting fiber was wound under tension to reproduce typical microbending losses found in some cables to simulate transmitting power distributions of interest. The receiving fiber was approximately 1 meter in length, though two tests were repeated using 10-m lengths to determine if any cladding modes were present. The results were essentially the same in either case.

IV. α MISMATCH

Splice loss due to α mismatch was measured for a few representative fiber pairs. The results are shown in Fig. 3 along with the general α mismatch curves generated by this model. The agreement between theory and measurement is good. For $\alpha_2 = 1.5$ and $\alpha_1 = 2.0$, a 25-percent mismatch in α, the splice loss is less than 0.2 dB. Sensitivity of splice loss to α mismatch is therefore substantially less than that for Δ or core radius mismatch.[2]

Fibers with combinations of α and Δ or α and radius mismatch were also spliced to observe the effects of combinations of factors. The results are given in Table I, where the first three columns indicate the percentage difference of each intrinsic parameter of the transmitting

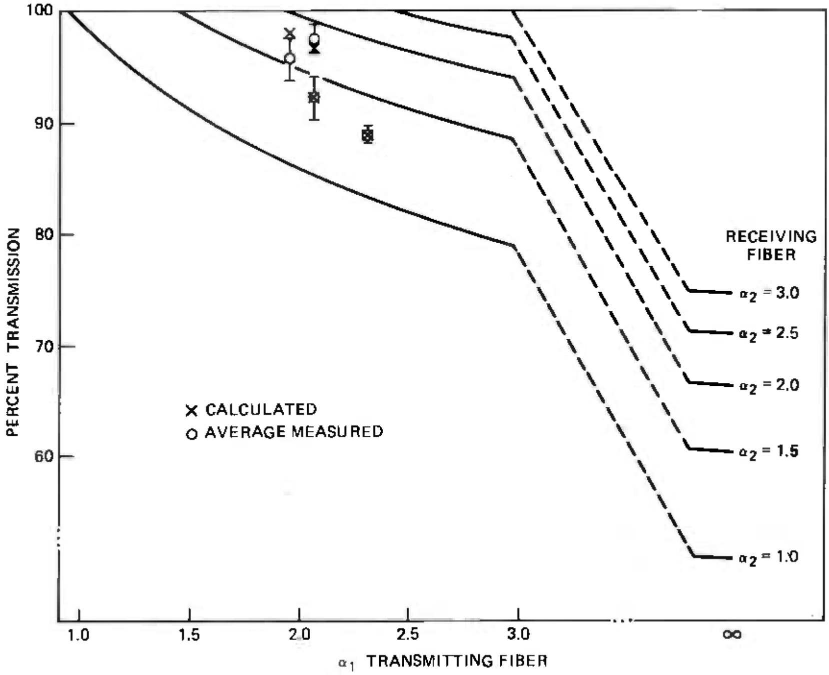


Fig. 3—Transmission vs α mismatch.

Table I—Calculated and measured splice losses between selected fiber pairs for α mismatch verification

Percentage Difference*			Predicted Loss (dB)	Measured Loss (dB)			Number of Readings
α	Δ	R		Low	Avg.	High	
+26.1	0	+6.3	0.374	0.258	0.357	0.458	6
+20.9	−3.5	0	0.105	0.053	0.110	0.156	7
−32.5	+37.6	0	0.558	0.564	0.614	0.667	5
+13.1	0	−1.1	0.074	0.101	0.189	0.276	7
+∞	−3.5	−14.8	0.864	0.442	0.585	0.674	8

* +26.1 indicates that the α of the transmitting fiber is 26.1 percent higher than the α of the receiving fiber.

fiber from those of the receiving fiber with the sign indicating the direction of the change (+ indicates a larger transmitting fiber parameter). The average measured loss is given, as well as the range and number of measurements. Each measurement involves disassembly of the splice, fracture of new ends on both fibers, and reassembly of the splice.

The range of loss measurements for some combinations of parameter mismatch is relatively large. Several factors contributed to this variation. (*i*) The detector used for early measurements degraded in its repeatability properties and was replaced by a different detector for later measurements. These repeatability problems were primarily due to lack of precision in positioning the fiber relative to the detector. (*ii*) Failure of the lock-in amplifier to track frequency drift in the source was found to cause measurement variations of as much as ±0.1 dB. (*iii*) End quality and contamination, along with small variations in the positions of the fibers in the splices, also contribute to measurement variations. (*iv*) Fiber profile asymmetries may also contribute slightly to measurement variations.

An attempt was made to test the limits of applicability of the model by applying it to a splice between step-index ($\alpha = \infty$) and approximately parabolic-index fibers. The results, as shown in Table I, indicate that the model gives a fair estimate of the loss even in this extreme case when the Δ and radius mismatches are included in the calculation. The shortness of the step-index transmitting fiber in this case (250 m) may have caused an incomplete filling of the transmitting mode structure.

The results for general intrinsic parameter mismatch measurements are given in Table II. It can be seen that some measurements agree very well with the calculated values, while others vary on both the high and low sides. The average relative difference between theoretical and experimental values, $(1/n) \sum$ (calculated loss − measured loss), was 0.03 dB for all graded-index fiber data in Tables I and II. The average absolute difference, $(1/n) \sum$ |calculated loss − measured loss|, was 0.11 dB. It should be noted that inaccuracies in measuring α, Δ, and core radius as well as variations of α over the core probably account for some of the difference between theoretical and experimental values given here. The experimental data were accumulated over a three-month period using several different detectors. All the data are reported here. No systematic errors are apparent.

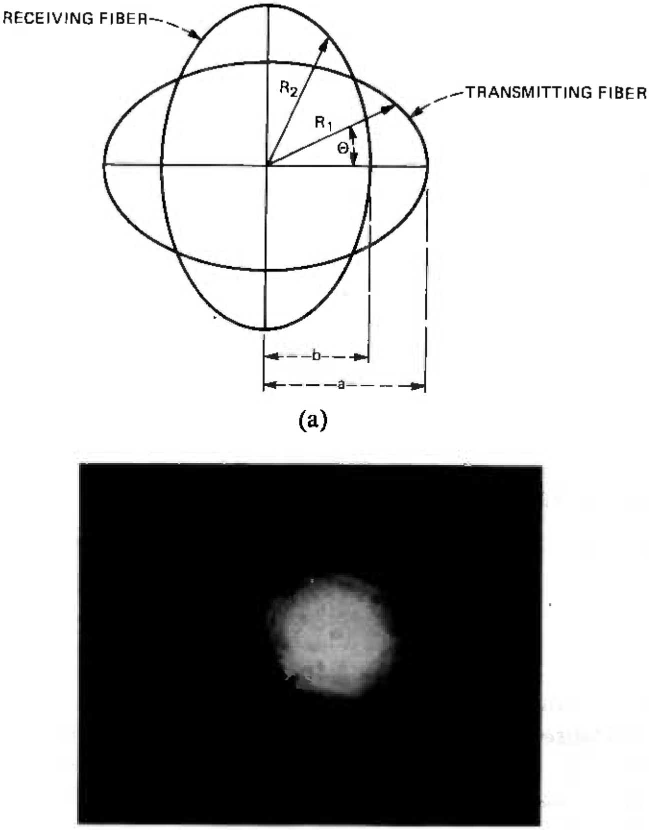
V. CORE ELLIPTICITY

Measurements of splice loss due to core ellipticity were made as a function of the angle between the major axes of the transmitting and receiving fibers (Fig. 4). This produced an approximately sinusoidal curve with zero loss at the 0° and 180° points and maximum loss at the 90° and 270° points. This maximum can be compared to the value calculated from the model.

The measurements were limited by a lack of suitable fibers with significant core ellipticity as well as by the low magnitude of the losses involved. Several fibers were selected which had core ellipticities of 4, 10, and 16.5 percent. The fibers with 4- and 10-percent ellipticity were too short to permit measurements with long fiber lengths after the splice. The maximum loss measured was less than 0.05 dB for the 4-percent fiber and less than 0.1 dB for the 10-percent fiber. The

Table II—Calculated and measured splice loss for general parameter mismatch verification

Percentage Difference			Predicted Loss	Measured Loss (dB)			Number of
α	Δ	R	(dB)	Low	Avg.	High	Readings
−5.1	−0.22	+0.4	0.001	0.059	0.090	0.156	6
+5.1	+0..22	−0.4	0.032	0.102	0.168	0.216	5
+33.3	−1.2	+7.8	0.495	0.468	0.497	0.537	5
+4.5	+7.6	+2.7	0.195	0.075	0.155	0.321	8
+24.6	+19.9	+0.8	0.579	0.275	0.296	0.310	5
−10.5	+50.0	+0.8	1.27	0.817	1.03	1.42	24
+21.0	+13.3	−1.9	0.340	0.171	0.216	0.260	5
+31.3	+13.3	−3.0	0.437	0.237	0.264	0.321	5
−14.6	+45.9	−1.9	0.996	0.703	0.876	1.06	12
−30.4	+36.8	+0.8	0.536	0.610	0.684	0.774	7



(a)

(b)

Fig. 4—Elliptical core fibers. (a) 90-degree core misalignment. (b) 16.5 percent ellipticity fiber core.

repeatability of the measurements is a few hundredths of a decibel. This prevented an accurate determination of the effect of the ellipticity. These losses were approximately the same as those predicted by the Gaussian model (0.025 dB for 4 percent and 0.08 dB for 10 percent) (Fig. 5) and significantly less than those predicted by the uniform power model (0.12 dB for 4 percent and 0.325 dB for 10 percent). Figure 4b shows the core cross section of the 16.5-percent fiber. Although not a perfect ellipse, this was the only fiber available with sufficient ellipticity to give a splice loss large enough to be accurately measured.

The 16.5-percent ellipticity fiber was an extremely high loss fiber, ~50 dB/km for long sections of fiber. The near-field patterns for a 100-m length and an 800-m length were essentially the same, and it appeared that the high-loss characteristics of the fiber produced full mode coupling in the 100-m length. For this reason, it was felt that the

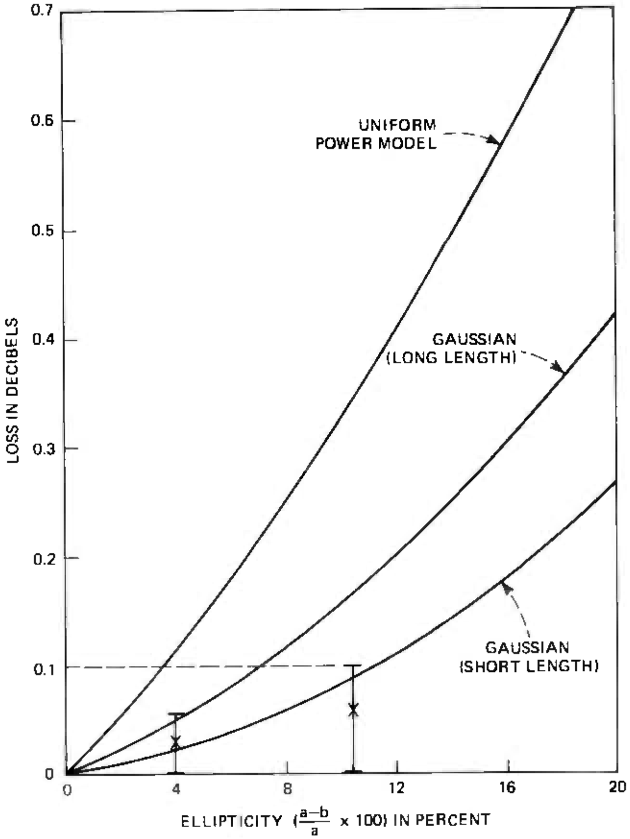


Fig. 5—Maximum splice loss due to core ellipticity.

100-m length was sufficient to satisfy the assumption (in both Gaussian and uniform power models) of a steady-state power distribution entering the splice.

Figure 6 shows the splice loss for this fiber with a 1-m length after the splice as a function of the angle between the semi-major axes of the ellipses. The general shape of a curve through the data points and the average maximum loss of 0.13 dB are in good agreement with the expected shape and the value calculated from the model of 0.19 dB. The discrepancies may be partly attributable to the deviation from pure ellipticity of the core cross section as shown in Fig. 4b and to the shortness of the transmitting fiber. The difference between the average maximum measured loss and the loss calculated with the Gaussian model was only 0.06 dB. This was approximately the same as the noise and repeatability properties of the measurement system.

Figure 7 shows the results when a 100-m length of receiving fiber is used. Again the general shape of the curve agrees with that expected, although more noise is apparent in the data. The average maximum splice loss was 0.18 dB which is in fair agreement with the calculated value of 0.31 dB. Measurements with longer lengths of receiving fiber were attempted, but the results were inconclusive because of the high attenuation and reduced signal-to-noise ratio.

VI. INTRINSIC MISMATCH PLUS TRANSVERSE OFFSET

Experimental and theoretical values for splice loss vs transverse offset with intrinsic parameter mismatch are shown in Fig. 8. The fibers used in this part correspond to line 9 in Table II. The experimental points represent the average of four readings, two left of center and two right of center, as the receiving fiber was traversed across the transmitting fiber. The agreement is very good.

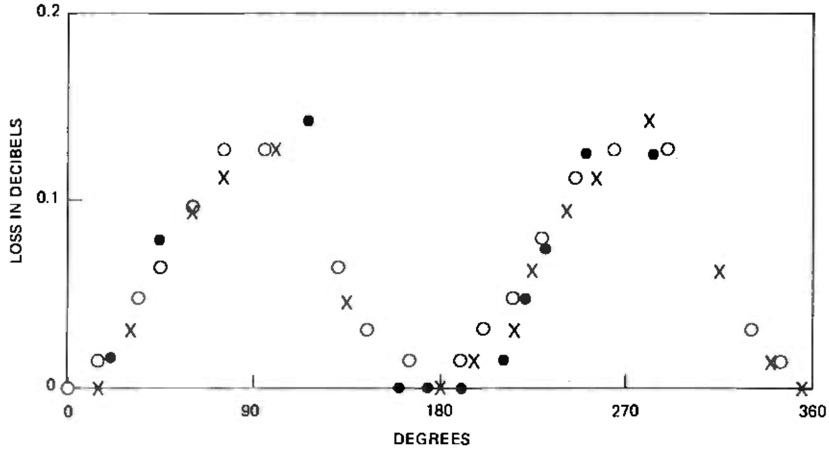


Fig. 6—Core ellipticity splice loss, short length.

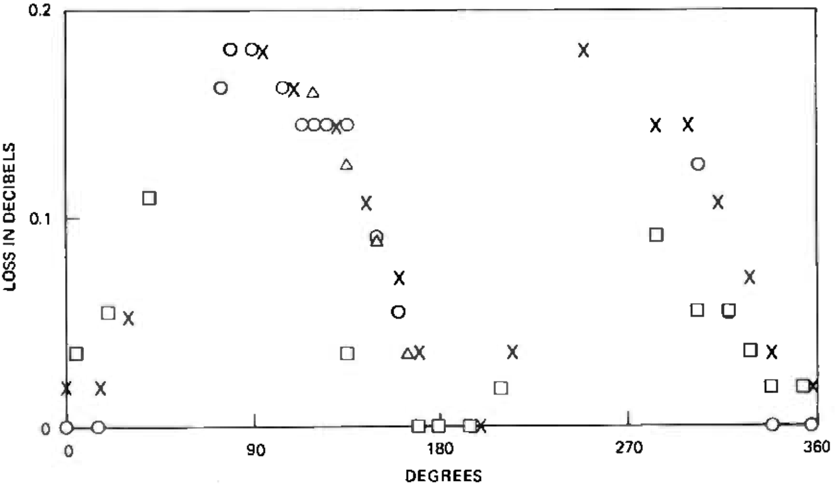


Fig. 7—Core ellipticity splice loss, long length.

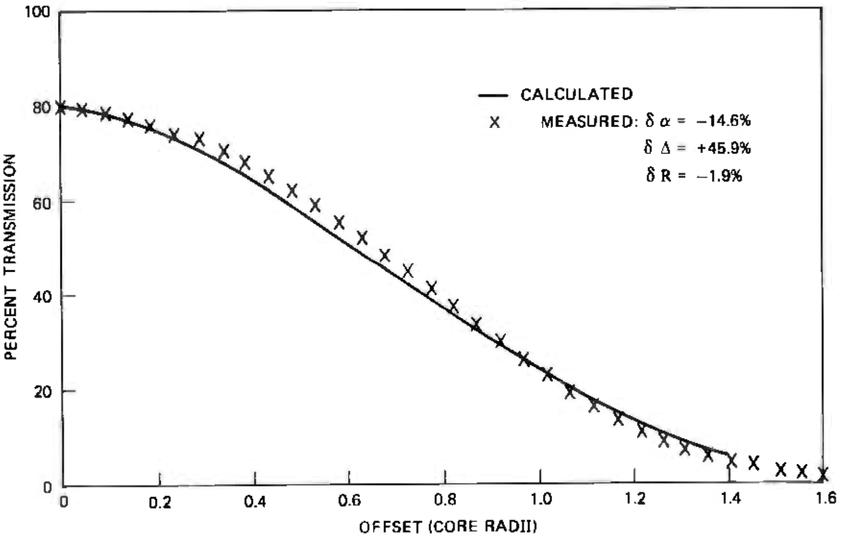


Fig. 8—Intrinsic parameter mismatch and transverse offset.

The model can also be used to generate parametric families of curves such as Fig. 9 and Fig. 10, showing the importance of the different intrinsic parameters when combined with an extrinsic parameter. Δ mismatch and transverse offset are shown in Fig. 9 and radius mismatch and transverse offset in Fig. 10, since they are the chief intrinsic and extrinsic factors contributing to splice loss at this time. In particular, Fig. 9 emphasizes that, if the receiving fiber NA is greater than the transmitting fiber NA, the sensitivity to transverse offset is significantly reduced, especially for small offsets. Sensitivity to transverse

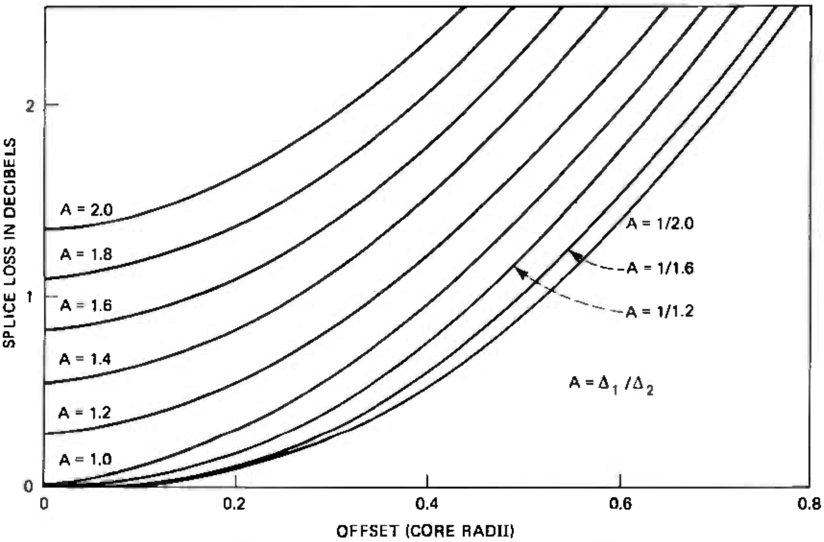


Fig. 9—Splice loss due to Δ mismatch and transverse offset.

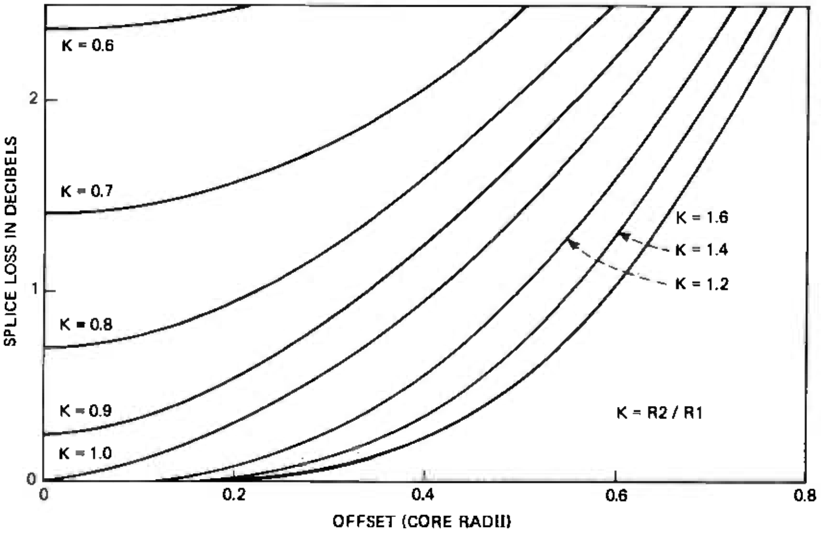


Fig. 10—Splice loss due to radius mismatch and transverse offset.

offset is slightly reduced for a receiving fiber NA less than the transmitting fiber NA, in fact, the maximum sensitivity to transverse offset occurs for identical fibers.

VII. CONCLUSION

It has been demonstrated that the Gaussian point transmission model gives good estimates of splice loss due to intrinsic parameter mismatch and transverse offset. The effects of combinations of these factors on splice loss has been characterized and shown to be significant. Effects due to α and ϵ mismatch are smaller for a given percentage mismatch than for Δ or core radius mismatches.

VIII. ACKNOWLEDGMENT

The author is grateful to C. M. Miller for his advice and for calculating the long length and the uniform power model losses and to D. N. Ridgway for his assistance with the experimental measurements. Thanks are also due to M. J. Saunders for providing the intrinsic parameters of the fibers used in this experiment, to F. T. Stone for providing both the 16.5-percent elliptical core fiber and the photograph of its core, and to M. I. Schwartz for help in organizing this paper.

APPENDIX

This section provides more detail on the application of this theory to various general classes of fiber splices. The integrals in eq. (7) cannot be performed exactly, except in a few special cases. Numerical evaluation is usually required, but considerable simplification can sometimes result from careful consideration of the particular splice parameters. The case of intrinsic parameter mismatch with no ellipticity and no transverse offset is examined first.

A1. Splices without transverse offset

Zero offset splices can be broken down into three classes for the particular family of radially symmetric profiles assumed here.

Class I contains splices with all transmitting parameters greater than or less than all corresponding receiving parameters. Class Ia is illustrated in Fig. 11a.

$$\text{Class Ia:} \begin{cases} \alpha_1 \geq \alpha_2 \\ \Delta_1 \geq \Delta_2 \\ R_1 \geq R_2 \end{cases} \qquad \text{Class Ib:} \begin{cases} \alpha_1 \leq \alpha_2 \\ \Delta_1 \leq \Delta_2 \\ R_1 \leq R_2 \end{cases}$$

For Ia

$$P_2 = \int t(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r dr. \tag{8}$$

For Ib the transmission equals unity.

Class II splices have one intersection oι the core profile. Class IIa is shown in Fig. 11b.

$$\text{Class IIa:} \begin{cases} \Delta_1 > \Delta_2 \\ R_1 < R_2 \end{cases} \qquad \text{Class IIb:} \begin{cases} \Delta_1 < \Delta_2 \\ R_1 > R_2 \end{cases}$$

In both these cases, the transmission is unity where NA$_2$ exceeds NA$_1$ and a function of r over the remainder of the profile. The crossover point, r', is determined by equating the NAs and solving the resulting nonlinear equation numerically:

$$\Delta_1(1 - k^{\alpha_1}r^{\alpha_1}) = \Delta_2(1 - r^{\alpha_2}). \tag{9}$$

Then

$$P_2 = \begin{cases} \displaystyle\int_0^{r'} t(r)(1 - k^{\alpha_1}r^{\alpha_1})^2 r dr + \int_{r'}^{1} (1 - k^{\alpha_1}r^{\alpha_1})^2 r dr & \text{IIa} \\ \displaystyle\int_0^{r'} (1 - k^{\alpha_1}r^{\alpha_1})^2 r dr + \int_{r'}^{1/k} t(r)(1 - k^{\alpha_1}r^{\alpha_1})^2 r dr & \text{IIb} \end{cases} \tag{10}$$

Class III contains all remaining possibilities. Class IIIa is shown in Fig. 11c.

$$\text{Class IIIa:} \begin{cases} \Delta_1 \geq \Delta_2 \\ \alpha_1 < \alpha_2 \\ R_1 \geq R_2 \end{cases} \qquad \text{Class IIIb:} \begin{cases} \Delta_1 \leq \Delta_2 \\ \alpha_1 > \alpha_2 . \\ R_1 \leq R_2 \end{cases}$$

Class III mismatches may or may not have the crossovers shown in Fig. 11c. Those that do not cross over correspond to Fig. 11a; however, it is necessary to solve eq. (9) to determine whether a crossover occurs. If no r' exists, the loss can be calculated using the appropriate formula for Class I. If an r' exists, the second crossover point r'' must also be found to determine the appropriate areas of integration. Again, the transmission is unity where NA$_2$ is greater than NA$_1$ and is a function of r over the rest of the profile. Then

$$P_2 = \begin{cases} \displaystyle\int_0^{r'} t(r)(1 - k^{\alpha_1}r^{\alpha_1})^2 r dr + \int_{r'}^{r''} (1 - k^{\alpha_1}r^{\alpha_1})^2 r dr \\ \displaystyle\qquad + \int_{r'}^{1} t(r)(1 - k^{\alpha_1}r^{\alpha_1})^2 r dr & \text{IIIa} \\ \displaystyle\int_0^{r'} (1 - k^{\alpha_1}r^{\alpha_1})^2 r dr + \int_{r'}^{r''} t(r)(1 - k^{\alpha_1}r^{\alpha_1})^2 r dr \\ \displaystyle\qquad + \int_{r''}^{1/k} (1 - k^{\alpha_1}r^{\alpha_1})^2 r dr & \text{IIIb} \end{cases} \tag{11}$$

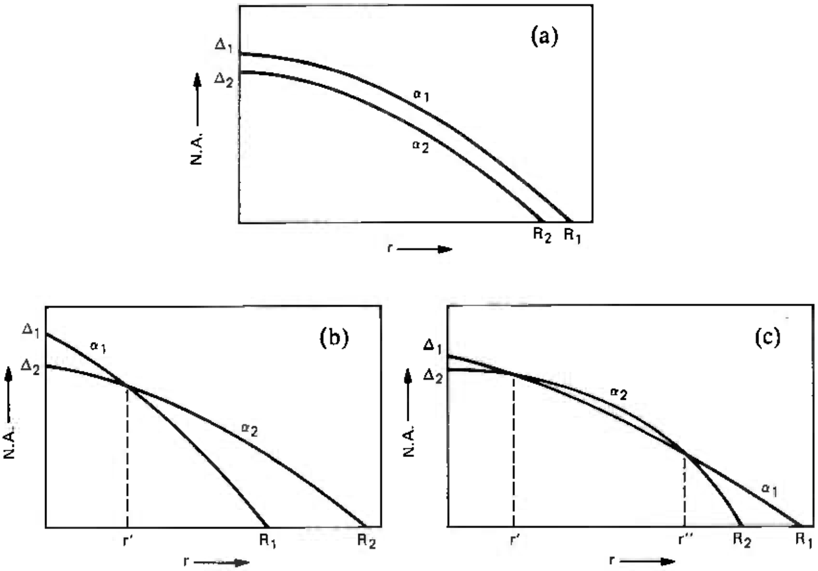


Fig. 11—Intrinsic mismatch profiles. (a) Class Ia. (b) Class IIa. (c) Class IIIa.

Two special cases also occur when $\Delta_1 = \Delta_2$ or $\alpha_1 = \alpha_2$.

If $\Delta_1 = \Delta_2$ (Class III):

$$r' = k^{\left(\frac{\alpha_1}{\alpha_2 - \alpha_1}\right)}. \tag{12}$$

In this special case, r' exists only if $R_1 > R_2$ and $\alpha_1 < \alpha_2$ or $R_1 < R_2$ and $\alpha_1 > \alpha_2$. Otherwise, the only crossover is at $r = 0$.

If $\alpha_1 = \alpha_2 = \alpha$ (Class II):

$$r' = \left(\frac{\Delta_2 - \Delta_1}{\Delta_2 - k^{\alpha}\Delta_1}\right)^{1/\alpha}. \tag{13}$$

If $\alpha_1 = \alpha_2$ in Class I or III, there is no crossover.

A2. Splices with transverse offset

The addition of transverse offset greatly complicates the problem. The four cases of Ref. 4 must be included and the integrals evaluated over the areas of overlap. The power transmitted at any point is the product of the point transmission function, $t(r)$, and the power available to be transmitted at that point.

$$p(r) = t(r)(1 - k^{\alpha_1} r^{\alpha_1})^2. \tag{14}$$

However, since the cores are not centered, one of the NA functions must be transformed to the center of the other core. The integral can then be performed numerically setting $t(r)$ equal to unity at any point at which NA_2 is greater than NA_1. In the cases in which the two core boundaries do not intersect, Cases I and III of Ref. 4, one obtains a double integral referenced to the center of the smaller core. This is illustrated in Fig. 12a. The definition of d is the normalized separation of the fiber core centers.

Case I: $R_1 < R_2$, $d < 1 - 1/k$.

$$d = \frac{\text{Transverse Offset}}{R_2} \;(R_2 = \text{receiving fiber core radius}) \tag{15}$$

$$P_2 = 2\int_0^{\pi}\int_0^{1/k} t_1(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r\,dr\,d\theta \tag{16}$$

$$t_1(r) = \begin{cases} 1 + Q_1 p_0 - p_0^{Q_1} & Q_1 < 1 \\ 1 & Q_1 \geq 1 \end{cases}. \tag{17}$$

$$Q_1 = \frac{\Delta_2[1 - (r^2 + d^2 - 2rd\cos\theta)^{\alpha_2/2}]}{\Delta_1(1 - k^{\alpha_1} r^{\alpha_1})}. \tag{18}$$

Case III: $R_1 > R_2$, $d < 1/k - 1$.

$$P_2 = 2\int_0^{\pi}\int_0^{1} t_2(r)[1 - k^{\alpha_1}(r^2 + d^2 - 2rd\cos\theta)^{\alpha_1/2}]^2 r\,dr\,d\theta \tag{19}$$

$$t_2(r) = \begin{cases} 1 + Q_2 p_0 - p_0^{Q_2} & Q_2 < 1 \\ 1 & Q_2 \geq 1 \end{cases}. \tag{20}$$

$$Q_2 = \frac{\Delta_2(1 - r^{\alpha_2})}{\Delta_1[1 - k^{\alpha_1}(r^2 + d^2 - 2rd\cos\theta)^{\alpha_1/2}]}. \tag{21}$$

Cases II and IV of Ref. 4 present two different possible conditions of overlap requiring different approaches to the numerical integration. These are illustrated in Figs. 12b and 12c. The area of overlap is divided into regions which permit the use of the functions $t_1(r)$ and $t_2(r)$ defined above. The two separate conditions are distinguished by the following criteria:

Case II: $R_2 > R_1$ and $d > 1 - 1/k$ (interchange R_2 and R_1 in Fig. 12).

$$\text{Condition 1:}\quad 1 - \frac{1}{k^2} - d^2 \leq 0:\quad \text{Fig. 12b}$$

$$\text{Condition 2:}\quad 1 - \frac{1}{k^2} - d^2 > 0:\quad \text{Fig. 12c}$$

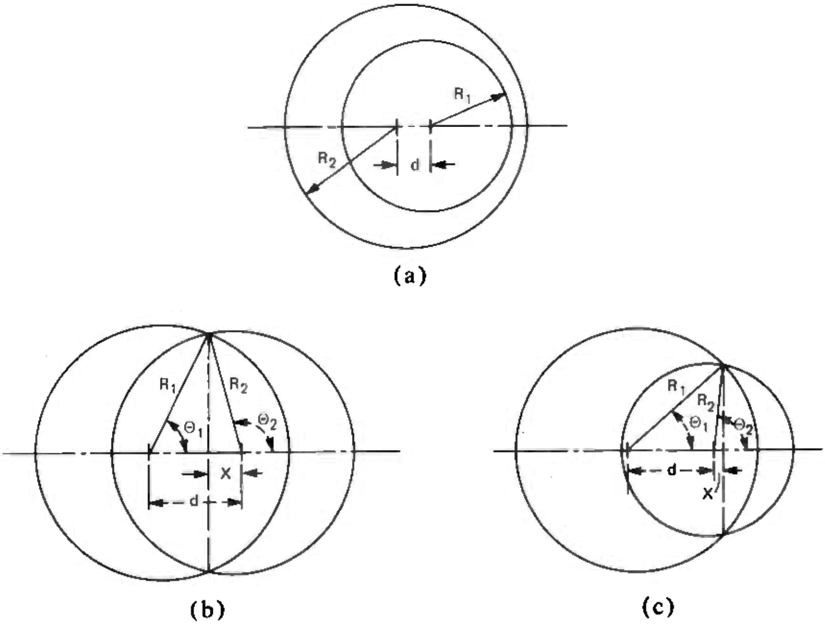


Fig. 12—Parameter mismatch and transverse offset. (a) Cases I and III offset. (b) Condition I $\theta_2 > \pi/2$. (c) Condition II $\theta_2 < \pi/2$.

Case IV: $R_2 < R_1$ and $d > 1/k - 1$.

$$\text{Condition 1:} \qquad \frac{1}{k^2} - 1 - d^2 \leq 0: \quad \text{Fig. 12b}$$

$$\text{Condition 2:} \qquad \frac{1}{k^2} - 1 - d^2 > 0: \quad \text{Fig. 12c.}$$

In practice, the two different conditions in Cases II and IV are distinguished by the sign of $\cos \theta_2$.

Case II:

$$\cos \theta_1 = \frac{1 + d^2 - 1/k^2}{2d} \tag{22}$$

$$\cos \theta_2 = \frac{1 - d^2 - 1/k^2}{2d}. \tag{23}$$

Condition I: $\cos \theta_2 < 0$ (Fig. 12b).

$$P_2 = \int_0^{\theta_1} \int_{\frac{d-x}{\cos\theta}}^{1} t_2(r)[1 - k^{\alpha_1}(r^2 + d^2 - 2rd \cos \theta)^{\alpha_1/2}]^2 r dr d\theta$$

$$+ \int_{\pi-\theta_2}^{\pi} \int_{\frac{x}{\cos\theta}}^{1/k} t_1(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r dr d\theta, \tag{24}$$

where

$$x = \frac{d^2 + 1/k^2 - 1}{2d}. \tag{25}$$

Condition II: $\cos \theta_2 > 0$ (Fig. 12c).

$$P_2 = \int_0^{\theta_1} \int_{\frac{d+x}{\cos\theta}}^{1} t_2(r)[1 - k^{\alpha_1}(r^2 + d^2 - 2rd \cos \theta)^{\alpha_1/2}]^2 r dr d\theta$$

$$+ \int_0^{\theta_2} \int_0^{\frac{x}{\cos\theta}} t_1(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r dr d\theta,$$

$$+ \int_{\theta_2}^{\pi} \int_0^{1/k} t_1(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r dr d\theta \tag{26}$$

where

$$x = \frac{1 - d^2 - 1/k^2}{2d}. \tag{27}$$

Case IV:

$$\cos \theta_1 = \frac{1/k^2 + d^2 - 1}{2d} \tag{28}$$

$$\cos \theta_2 = \frac{1/k^2 - d^2 - 1}{2d}. \tag{29}$$

Condition I: $\cos \theta_2 < 0$ (Fig. 12b).

$$P_2 = \int_0^{\theta_1} \int_{\frac{d-x}{\cos\theta}}^{1/k} t_1(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r dr d\theta$$

$$+ \int_{\pi-\theta_2}^{\pi} \int_{\frac{x}{\cos\theta}}^{1} t(r)[1 - k^{\alpha_1}(r^2 + d^2 - 2rd \cos \theta)^{\alpha_1/2}]^2 r dr d\theta, \tag{30}$$

where

$$x = \frac{1 + d^2 - 1/k^2}{2d}. \tag{31}$$

Condition II: $\cos \theta_2 > 0$ (Fig. 12c).

$$P_2 = \int_0^{\theta_1} \int_{\frac{d+x}{\cos \theta}}^{1/k} t_1(r)(1 - k^{\alpha_1} r^{\alpha_1})^2 r dr d\theta$$

$$+ \int_0^{\theta_2} \int_0^{\frac{x}{\cos \theta}} t_2(r)[1 - k^{\alpha_1}(r^2 + d^2 - 2rd \cos \theta)^{\alpha_1/2}]^2 r dr d\theta$$

$$+ \int_{\theta_2}^{\pi} \int_0^1 t_2(r)[1 - k^{\alpha_1}(r^2 + d^2 - 2rd \cos \theta)^{\alpha_1/2}]^2 r dr d\theta, \qquad (32)$$

where

$$x = \frac{1/k^2 - d^2 - 1}{2d}. \qquad (33)$$

Application of the Gaussian model to the case of elliptical mismatch is restricted to identical fibers with no transverse offset. The splice loss in this case will vary from zero, when the axes of the elliptical cores are perfectly aligned, to some maximum, when they are at right angles to each other. For identical elliptical fibers, the Gaussian model was used to calculate the maximum loss.

Figure 4a illustrates the geometry of the maximum loss case. For identical fibers, $\alpha_1 = \alpha_2 = \alpha$, $\Delta_1 = \Delta_2$ and $a_1 = a_2$ and $b_1 = b_2$ where a and b are the semimajor and semiminor axes of the elliptical core and the subscripts 1 and 2 refer to the transmitting and receiving fibers, respectively. The symmetry of the system shown was used to reduce the calculation of the fraction of the power transmitted through the splice to one quadrant.

At any angle θ,

$$k(\theta) \equiv \frac{R_2}{R_1} = \sqrt{\frac{a^2\sin^2\theta + b^2\cos^2\theta}{a^2\cos^2\theta + b^2\sin^2\theta}}. \qquad (34)$$

Then Q, the ratio of the numerical apertures squared at any point in the splice, is

$$Q(r, \theta) = \frac{1 - r^{\alpha}}{1 - k^{\alpha} r^{\alpha}} \qquad (35)$$

for $0 \leq \theta \leq \Pi/4$, $r \leq R_2$ and for $\Pi/4 \leq \theta \leq \Pi/2$, $r \leq R_1$. The total transmission ratio with R_2 normalized to unity is

$$T = \frac{P_2}{P_1}$$

$$= \frac{\int_0^{\pi/4}\int_0^1 t(r,\theta)(1-k^\alpha r^\alpha)^2 r dr d\theta + \int_{\pi/4}^{\pi/2}\int_0^{1/k}(1-k^\alpha r^\alpha)^2 r dr d\theta}{\int_0^{\pi/2}\int_0^{1/k}(1-k^\alpha r^\alpha)^2 r dr d\theta}. \tag{36}$$

The point transmission from 45° to 90° in the first quadrant is unity since the receiving numerical aperture is larger than the transmitting numerical aperture in this region. Solution of the integrals in eq. (36) by numerical methods gives the results shown in Fig. 5 for the maximum loss due to core ellipticity. The long length loss was found by correcting for the additional loss in a long fiber due to mode redistribution.[2]

The uniform power model loss was found using eq. (36) with $t(r, \theta)$ set equal to Q, the ratio of the squares of the numerical apertures, using $1 - k^\alpha r^\alpha$ as the near-field power function. The difference between the Gaussian model and the uniform power model is appreciably greater for elliptical mismatch than for the other intrinsic mismatches due to the greater difference between the two models near the core-cladding boundary.

REFERENCES

1. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52*, No. 9 (November 1973), pp. 1563–1578.
2. C. M. Miller and S. C. Mettler, "A Loss Model for Parabolic-Profile Fiber Splices," B.S.T.J., *57*, No. 9 (November 1978), pp. 3167–3180.
3. Haruhiko Tsuchiya et al., "Double Eccentric Connectors for Optical Fibers," Appl. Opt., *16*, No. 5 (May 1977), pp. 1323–1331.
4. C. M. Miller, "Transmission vs. Transverse Offset for Parabolic-Profile Fiber Splices with Unequal Core Diameters," B.S.T.J., *55*, No. 7 (September 1976), pp. 917–927.
5. C. M. Miller, "Effects of Fiber Manufacturing Variations on Graded Index Fiber Splices," Proc. 4th European Conference on Optical Communication, Genoa, Italy, September 12–15, 1978.
6. C. M. Miller, "Loose Tube Splices for Optical Fibers," B.S.T.J., *54*, No. 7 (September 1975), pp. 1215–1225.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

Three-Stage Multiconnection Networks Which Are Nonblocking in the Wide Sense

By F. K. HWANG

(Manuscript received June 20, 1979)

A multiconnection network deals with the connections of pairs $\{(X, Y)\}$ *where* X *is a subset of the input terminals and* Y *is a subset of the output terminals. We study the conditions under which a three-stage Clos network is nonblocking for such connections. We show that the number of middle switches needed for nonblocking depends on the routing strategy. Therefore the networks satisfying the conditions are networks nonblocking in the wide sense. We also derive formulas for computing the minimum numbers of crosspoints required by such networks.*

I. INTRODUCTION

A *three-stage Clos network*, denoted by $\nu(m, n_1, r_1, n_2, r_2)$, consists of r_1 rectangular $(n_1 \times m)$ input switches, m rectangular $(r_1 \times r_2)$ middle switches and r_2 rectangular $(m \times n_2)$ output switches. There is exactly one link connecting each input switch to each middle switch and one link connecting each middle switch to each output switch. The n_1r_1 inlets of the input switches are called *input terminals* and the n_2r_2 outlets of the output switches are called *output terminals*.

Let I denote the set of input terminals and O the set of output terminals. A connecting pair in the classical sense is a pair (x, y): $x \in I$, $y \in O$ requesting to be connected. Masson and Jordan[1] generalized the definition of a connecting pair to be a pair (x, Y): $x \in I$, $Y \subseteq O$ such that x is to be connected to every output terminal in Y. This definition was further generalized in Ref. 2 so that a connecting pair is a pair (X, Y): $X \subseteq I$, $Y \subseteq O$ such that each terminal in X is to be connected to every terminal in Y. A network dealing with this type of connecting pairs is called a multiconnection network.[2] In practice, we often need only consider X and Y with limited cardinalities. Let $|S|$ denote the cardinality of a set S. Then in a (q_1, q_2) multiconnection network, we are only concerned with connecting pairs $(X, Y), |X| \leq q_1, |Y| \leq q_2$.

A multiconnection network is *nonblocking* if, regardless of what state the network is currently in, a connecting pair involving only idle terminals can always be connected by a subgraph of the network which is link-disjoint to all subgraphs connecting previous pairs. A multiconnection network is *nonblocking in the wide sense*, following Beneš' definition[3] for the classical single-connection case, if it is nonblocking when a particular routing (connection) strategy is followed. Practical networks which are nonblocking in the wide sense for classical assignments rarely exist. In this paper, we show that such networks exist for the multiconnection case.

II. FOUR ROUTING STRATEGIES

Suppose (X, Y) is the current pair to be connected. It is commonly assumed[1,2] that the rectangular switches have the fan-in, fan-out property, i.e., any subset of inlets can be connected simultaneously to any subset of outlets. Therefore, it suffices to consider the pair (X, Y) consisting of at most one terminal from each input switch and at most one terminal from each output switch. For, if we can connect one input (output) terminal to $Y(X)$, then all terminals in the same input (output) switch can be connected to $Y(X)$. Therefore we may assume $r_1 \geq q_1$ and $r_2 \geq q_2$ without loss of generality. We now discuss four possible routing strategies.

Strategy 1: Find $|X||Y|$ middle switches each connecting a distinct pair $(x, y), x \in X, y \in Y$.

Strategy 2: Find $|X|$ middle switches each connecting a distinct pair $(x, Y), x \in X$.

Strategy 3: Find $|Y|$ middle switches each connecting a distinct pair $(X, y), y \in Y$.

Strategy 4: Find one middle switch connecting the pair (X, Y).

We now compute the number of middle switches needed under each strategy so that the pair (X, Y) can always be connected. To avoid discussions of uninteresting modifications, we assume $r_1 \geq q_1q_2n_1$ and $r_2 \geq q_1q_2n_2$.

Theorem 1: *$\nu(m, n_1, r_1, n_2, r_2)$ is nonblocking as a (q_1, q_2) multiconnection network under Strategy 1 if and only if $m \geq q_2n_1 + q_1n_2 - 1$.*

Proof: Consider the connection of the pair $(x, y), x \in X, y \in Y$. The input switch which contains x can be already connected to at most $n_1q_2 - 1$ distinct middle switches under Strategy 1. This is because there are only n_1 inlets in the switch and each inlet has at most q_2 connections except that x can have at most $q_2 - 1$ connections. Similarly, the output switch which contains y can be already connected to at most $n_2q_1 - 1$ distinct middle switches under Strategy 1. In the worst case, the $q_2n_1 - 1$ middle switches and the $q_1n_2 - 1$ middle switches are disjoint. However, if we have $(q_2n_1 - 1) + (q_1n_2 - 1) + 1$ middle switches, then one middle switch must be available to connect the pair (x, y). Since it is also clear that the worst case can happen, the "only if" part of Theorem 1 is also proved.

Theorem 2: *$\nu(m, n_1, r_1, n_2, r_2)$ is nonblocking as a (q_1, q_2) multiconnection network under Strategy 2 if and only if $m \geq n_1 + q_1q_2n_2 - q_2$.*

Proof: Consider the connection of the pair (x, Y). The input switch which contains x can be already connected to at most $n_1 - 1$ distinct middle switches under Strategy 2. Each output switch in Y can be already connected to at most $(n_2q_1 - 1)$ distinct middle switches under Strategy 2. Since $|Y| \leq q_2$, Theorem 2 follows from a worst-case argument similar to the one given in Theorem 1.

Theorem 3: *$\nu(m, n_1, r_1, n_2, r_2)$ is nonblocking as a (q_1, q_2) multiconnection network under Strategy 3 if and only if $m \geq q_1q_2n_1 + n_2 - q_1$.*

Proof: Analogous to the proof of Theorem 2.

Theorem 4: *$\nu(m, n_1, r_1, n_2, r_2)$ is nonblocking as a (q_1, q_2) multiconnection network under Strategy 4 if and only if $m \geq q_1n_1 + q_2n_2 - q_1 - q_2 + 1$.*

Proof: Consider the connection of the pair (X, Y). The input switches in X can be already connected to at most $|X|(n_1 - 1) \leq q_1(n_1 - 1)$ distinct middle switches under Strategy 4. Similarly, the output switches in Y can be already connected to at most $|Y|(n_2 - 1) \leq q_2(n_2 - 1)$ distinct middle switches. Theorem 4 follows immediately from a worst-case argument.

We note that, for $q_1 = q_2 = 1$, Theorems 1, 2, 3, and 4 are all reduced to the famous Clos Nonblocking Theorem.[4]

The existence of networks nonblocking in the wide sense can now be easily shown. For example, assume $q_2n_1 + q_1n_2 - 1 > m \geq q_1n_1 + q_2n_2 - q_1 - q_2 + 1$. Then the network is nonblocking under Strategy 4 but not necessarily nonblocking under any other strategy, for instance, Strategy 1.

III. COMPUTING THE NUMBERS OF CROSSPOINTS

For given numbers of input terminals and output terminals $N_1 = n_1r_1$, $N_2 = n_2r_2$, we would like to determine n_1, n_2 and m such that $\nu(m, n_1, r_1, n_2, r_2)$ is nonblocking in the wide sense for (q_1, q_2) multiconnection networks and has a minimum number of crosspoints. The optimal solutions for n_1, n_2 and m, of course, depend on which routing strategy we adopt. However, we will give a mathematical formulation general enough for all four cases.

Let Q be the number of crosspoints for (m, n_1, r_1, n_2, r_2). Then

$$Q = r_1 n_1 m + m r_1 r_2 + r_2 m n_2$$
$$= m\left(N_1 + \frac{N_1}{n_1}\frac{N_2}{n_2} + N_2\right).$$

Assume

$$m = u n_1 + v n_2 - w,$$

where u, v and w are nonnegative constants. Setting the first partial derivatives of Q with respect to n_1 and n_2 to zero, we obtain

$$\frac{\partial Q}{\partial n_1} = u\left(N_1 + \frac{N_1}{n_1}\frac{N_2}{n_2} + N_2\right) - (u n_1 + v n_2 - w)\frac{N_1 N_2}{n_1^2 n_2} = 0$$
$$\frac{\partial Q}{\partial n_2} = v\left(N_1 + \frac{N_1}{n_1}\frac{N_2}{n_2} + N_2\right) - (u n_1 + v n_2 - w)\frac{N_1 N_2}{n_1 n_2^2} = 0.$$

Solving for n_1 and n_2, we obtain

$$n_1 = n_2 v/u$$

and n_2 is the unique real root (easily verified by standard methods) of the cubic equation

$$v^2(N_1 + N_2)n_2^3 - uvN_1N_2n_2 + uwN_1N_2 = 0.$$

Let Q_i, $i = 1, 2, 3, 4$, denote the minimum Q under Strategy i, namely, m is replaced by $q_2 n_1 + q_1 n_2 - 1$, $n_1 + q_1 q_2 n_2 - q_1$, $q_1 q_2 n_1 + n_2 - q_2$ and $q_1 n_1 + q_2 n_2 - (q_1 + q_2 - 1)$, respectively. Then we will select Strategy j such that

$$Q = Q_j = \min_{i=1,2,3,4} Q_i.$$

For example, let $Q_1 = \min_{i=1,2,3,4} Q_i$. Approximating m by $q_2 n_1 + q_1 n_2$, we obtain the solution

$$n_1 = \sqrt{\frac{q_2 N_1 N_2}{q_1(N_1 + N_2)}}, \qquad n_2 = \sqrt{\frac{q_1 N_1 N_2}{q_2(N_1 + N_2)}}.$$

Substituting back in Q, we obtain

$$Q \cong \left(q_2\sqrt{\frac{q_1 N_1 N_2}{q_2(N_1 + N_2)}} + q_1\sqrt{\frac{q_2 N_1 N_2}{q_1(N_1 + N_2)}}\right)$$
$$N_1 + \frac{N_1 N_2}{\sqrt{\frac{q_1 N_1 N_2}{q_2(N_1 + N_2)}}\sqrt{\frac{q_2 N_1 N_2}{q_1(N_1 + N_2)}}} + N_2 \Bigg)$$
$$= 4\sqrt{q_1 q_2 N_1 N_2 (N_1 + N_2)}.$$

IV. SOME CONCLUDING REMARKS

Strategy 1 has been adopted in previous works on multiconnection networks[1,2] and Theorem 1 was proved in Ref. 2 with the special case $q_1 = 1$ and $q_2 = r_2$ first proved in Ref. 1. A $(1, r_2)$ multiconnection network under Strategy 1 was called an expansion network by Masson.[5] In Ref. 6, Masson considered $(1, r_2)$ multiconnection networks under a routing strategy which is a weakened version of Strategy 2, namely to use Strategy 2 whenever possible. Under this strategy, he stated the result that $(n_1, n_1, r_1, n_2, r_2)$ is nonblocking if $r_2 \leq (2n_1/n_2)$ where $[x]$ is the smallest integer not exceeding x. However, the following example shows that this result is incorrect. Consider a network $\nu(3, 3, 2, 2, 3)$. Then

$$r_2 = \left[\frac{2n_2}{n_1}\right] = 3$$

satisfying the condition of Masson's result. However, $\nu(3, 3, 2, 2, 3)$ is not nonblocking even as a classical single connection network, since it does not satisfy the necessary and sufficient condition $m \geq n_1 + n_2 - 1$ of the Clos Nonblocking Theorem.

V. ACKNOWLEDGMENT

The author thanks V. E. Beneš for many helpful comments.

REFERENCES

1. G. Masson and B. Jordan, "Generalized Multistage Connection Networks," Networks, *2* (1972), pp. 191–209.
2. F. K. Hwang, "Rearrangeability of Multiconnection Three-Stage Networks," Networks, *2* (1972), pp. 301–306.
3. V. E. Beneš, "Algebraic and Topological Properties of Connecting Networks," B.S.T.J., *41*, 1962, pp. 1249–1274.
4. C. Clos, "A Study of Nonblocking Switching Networks," B.S.T.J., *32*, 1953, pp. 406–424.
5. G. Masson, "On Rearrangeable and Nonblocking Switching Networks," Conf. Records of 1976 IEEE Inter. Conf. Commun., *1* (1976), pp. 7-1 to 7-7.
6. G. Masson, "Upperbounds on Fanout in Connection Networks," IEEE Trans. Circuit Theory, *20* (1973), pp. 222–230.

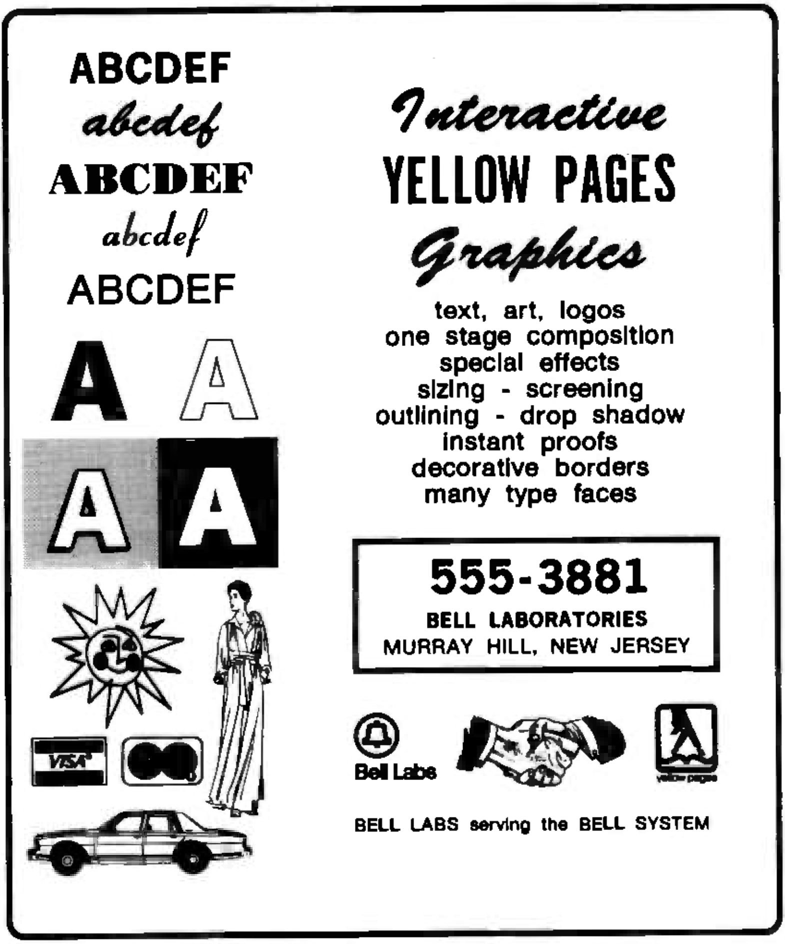
Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

An Interactive Terminal for the Design of Advertisements

By B. E. CASPERS and P. B. DENES

(Manuscript received July 9, 1979)

The advertisement below was created on the experimental graphics terminal described in this paper and demonstrates its features.



ABCDEF
abcdef
ABCDEF
abcdef
ABCDEF

A A
A A

Interactive
YELLOW PAGES
Graphics

text, art, logos
one stage composition
special effects
sizing - screening
outlining - drop shadow
instant proofs
decorative borders
many type faces

555-3881
BELL LABORATORIES
MURRAY HILL, NEW JERSEY

BELL LABS serving the BELL SYSTEM

The terminal is intended for the interactive design of advertisements. The work is part of our research on replacing photographic by digital methods of picture handling in typesetting; the results may help reduce the cost of producing Yellow Pages directories. Controls of the terminal were planned with human engineering in mind and thus are easy to learn and use.

The experimental terminal consists of a color TV display, a keyboard, a lightpen, a facsimile-type hard copier, and a fast picture scanner/digitizer which are all connected to a minicomputer equipped with disk and tape. It enables the user to interactively position, size, crop, and edit pictures as well as text; pictures can be scanned and digitized in a few seconds, corrected, and stored for future use; easy-to-use commands are provided for enhancing the appearance of text or pictures by automatically "outlining," "shadowing," or "screening" interactively selected areas. Finished advertisements are either stored on disk or outputted to tape; the data on the output tape are in a form ready for use on a CRT phototypesetter. The advertisement can also be outputted more quickly but with reduced quality using the hard copier.

Ways of outputting graphics on the typesetter with high quality and speed have also been studied; an average output speed of 1.5 seconds per square inch was achieved.

Field trials of the terminal are now in progress. Their goal is to study the type of personnel best suited to operate the terminal, to establish easy-to-use operating procedures, to make time and motion comparisons with existing production techniques, and to identify modifications for improved operation. Results already available are discussed in this paper.

I. TYPESETTING TECHNIQUES—OLD AND NEW

Traditional typesetters used lead type to obtain an image of the page to be printed, while more modern machines use computer-controlled CRTs. Indeed, modern printing technology uses computers not only for typesetting but also for preparing, editing, storing, and paginating the material to be printed. So far, however, computers have been used for processing only the text to be printed, while pictures are handled photographically. Photographic preparation of the picture material includes sizing and cropping, as well as such enhancements as "screening" (simulated gray scale), outlining, and shadowing; finally, the photographs are pasted manually on the page, into spaces appropriately left blank by the text typesetting process. Such photographic handling is used extensively, it is labor-intensive, and therefore it is expensive.

II. OBJECTIVES OF OUR TYPESETTING RESEARCH

Our aim is to replace these *photographic* methods of picture handling by *digital* methods. This would result in strongly reduced involvement of manual labor in typesetting tasks such as Yellow Pages directory production: It would eliminate pasting down pictorial material, much of the photographic laboratory work, and the manual retrieval/storage of artwork and advertisements. Digital methods also open the way for interactive design of the material to be typeset: Pictures (and text) can be moved, sized, cropped, or enhanced in other ways, and the changes can be viewed immediately on a display screen.

The principal steps needed for digital handling of pictures are:

(*i*) Scanning/digitizing of the original art.

(*ii*) Coding of pictures for efficient storage.

(*iii*) A design terminal for laying out advertisements and for adding special effects such as screening.

(*iv*) Appropriate programming of the typesetter to produce the pictures on its output medium.

It should be noted that digital typesetting of pictures requires that the picture data be converted into one code for efficient storage, into another code for aesthetic processing such as sizing and outlining, and into yet another code for fast typesetting. Means for converting quickly from one code to another is therefore a key element of the entire process; it occupies much of our attention, and work on it continues.

Bell Laboratories research has already made significant progress in finding methods of coding digitized pictures for efficient storage.[1-3]

This paper describes our research on an interactive terminal and on certain aspects of finding the best code for controlling typesetters. We also describe how the design of the man-machine interface was changed to simplify the use of the terminal by staff not familiar with computer operations. We have conducted a field trial to test the usability of this type of terminal. A more detailed trial is now starting which includes time and motion study comparisons with conventional production methods, attempts to arrive at a better definition of personnel requirements, and studies of various operational enhancements.

III. THE INTERACTIVE TERMINAL

The first version of our advertisement design terminal was implemented on a Honeywell DDP 224 laboratory computer.[4, 5] An improved version with many new features was transferred to a stand-alone minicomputer so that it could be installed for field trials at directory production departments of telephone companies. This is the experimental terminal described in this paper. The terminal performs the following functions:

(*i*) It allows text and picture material to be viewed and adjusted relative to each other; pictures and text can be moved, sized, and cropped, and text can be edited; the appearance of text or pictures can be enhanced by outlining, shadowing, and "screening" (shading).

(*ii*) The entire advertisement is coded suitably and then outputted to a digital electrostatic printer (200 lines/inch resolution).

(*iii*) The advertisement is also coded for output to a CRT typesetter (720 lines/inch resolution).

IV. THE HARDWARE

The principal hardware components used for the experimental terminal are shown in Fig. 1 and detailed below (of particular interest are the TV scan display and the scanner; reasons for their selection are given):

(*i*) A 16-bit word-length minicomputer (Data General Nova 830) with 48k words of memory, a tape drive, and three 2.4 Mbyte disks. The main purpose of the tape output is to transport advertisement definitions computed in typesetter code by the terminal to the typesetter for final outputting. The disks provide about 1.5 Mbyte buffer space for program execution and about 6 Mbytes for storing picture and advertisement specifications. A 96 Mbyte disk was later added to accommodate the larger number of logos and fonts needed for the extended trial, which is now starting.

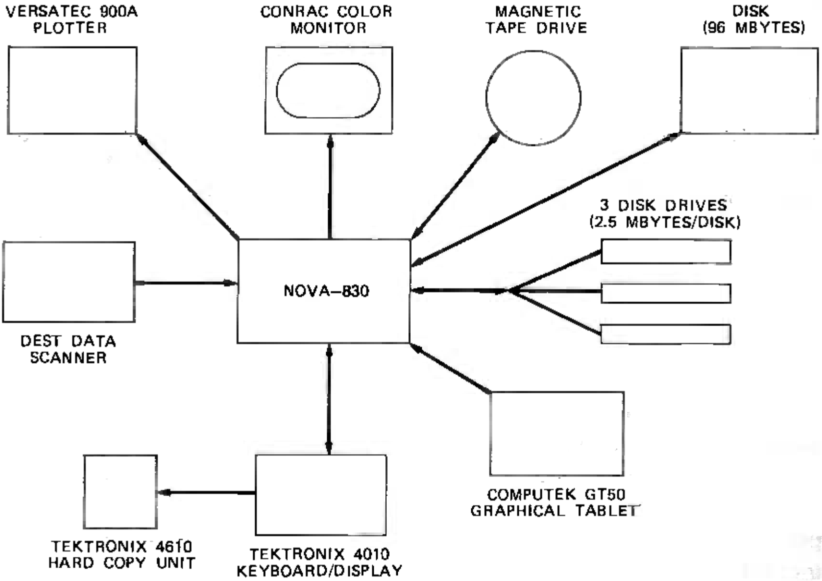

VERSATEC 900A PLOTTER
CONRAC COLOR MONITOR
MAGNETIC TAPE DRIVE
DISK (96 MBYTES)
3 DISK DRIVES (2.5 MBYTES/DISK)
NOVA–830
DEST DATA SCANNER
COMPUTEK GT50 GRAPHICAL TABLET
TEKTRONIX 4610 HARD COPY UNIT
TEKTRONIX 4010 KEYBOARD/DISPLAY

Fig. 1—Block diagram of terminal.

(*ii*) A keyboard/display (Tektronix 4010). Its main purpose is to provide the "function keys" by which the layout design process is controlled. The keyboard/display is also used for typing in text and for outputting messages to the operator.

(*iii*) A tablet/stylus (Computek) for easy communication with the displayed material by "pointing."

(*iv*) A color television monitor (Conrac) with buffer memory provides the main display on which the operator can inspect entire advertisements and individual scanned pictures. The entire display is defined in terms of 512 × 508 = 260096 picture elements; each picture element (pixel) is specified by two bits which allow selection of one of four colors (or four levels of gray scale). The entire picture is stored in a 520192-bit 600-ns buffer memory; it is read out at the rate of 10 pixels per microsecond to produce the video signal for the TV monitor.

Display Screen Considerations. The advantages of the TV scan display used for this work include the ability to display several 100000 pixels without flicker, to change any one pixel within a few microseconds without affecting the rest of the display, and to control the color of individual pixels for attracting attention to or otherwise label individual parts of the display. An alternative method for rapidly displaying the considerable volume of pixels needed for this application is the "storage screen" used for the Tektronix 4014 computer terminal. No display buffer is needed, because the storage screen provides its own picture memory. Therefore, even more picture data can be displayed than on the TV screen, because refresh rate and memory size limitations do not apply. However, the storage screen has no gray scale or color, and the entire screen has to be erased in order to erase even a single pixel.

(*v*) A 240 lines/inch scanner (Dest Data), which scans a full 8 × 10-inch page in 5 seconds (it was slowed to 12 seconds because our original disk drives were not fast enough to record the scan data). A fast scanner is an essential part of the interactive terminal; it allows convenient revision of artwork by using pen on paper (the method artists are familiar with), followed by rapid rescanning. The required scan resolution is about 700 lines/inch. Scanners of this resolution cost well over $100,000, however, which was our reason for selecting a less expensive 240-lines/inch device. The necessary resolution can still be achieved by scanning pictures which are three times larger than the size to be printed. An additional reason for selecting this scanner was that it uses a line of 2048 rigidly assembled photosensitive elements to digitize individual scan lines; it thus eliminates the need for rapidly moving parts, particularly the sensitive scanning mirrors. It has indeed proved itself to be a very reliable instrument.

(*vi*) An electrostatic printer (Versatec) which has facsimile quality of 200 lines/inch. This printer provides immediate copies of designed advertisements, although of lower quality than the typesetter; it could be located remotely and connected by telephone line to the terminal (or to any central store of digitally stored advertisements or of "speculative art"). The picture in Fig. 3 was outputted on the electrostatic printer: its quality can be compared with that of the typesetter output shown in the abstract.

V. NOVA SOFTWARE

The terminal runs under MRDOS, Data General's Mapped Real-Time Disk Operating System, but user familiarity with MRDOS is not required. Most of the software was programmed in Data General Fortran V. Some bit and pixel operations were coded in assembler to increase execution speed. Also in assembler are the I/O routines for such nonstandard MRDOS devices as the tablet, the scanner, the TV display, and the Versatec. All I/O to disk and magnetic tape is managed by calls to the appropriate MRDOS subroutines. A number of load-on-call overlays are employed to accommodate the terminal programs in the available core space.

VI. USER ENVIRONMENT

The principal way the user communicates with the terminal is by operating the keyboard and the tablet/stylus. The layout is shown on the TV screen and immediately indicates any changes made by the user. The display colors are white, red, and blue on black background.

Man-Machine Communication Considerations. Experience has shown that considerable attention must be paid to the design of a command language to make control of the terminal easy for noncomputer people. The first terminal[4, 5] used picture control commands that were close relatives of the UNIX*[6] text editor commands. They consisted of abbreviated 1- and 2-letter command names typed on the keyboard. Text editing was also available and was done by context search; use of the stylus was limited to positioning operations.

Discussions with advertisement make-up experts indicated the need for an easier-to-operate command structure. Light buttons were considered, but would have further reduced the already small display area. Therefore, a command structure utilizing a function keyboard was chosen, similar to one we have already used successfully[7] in a speech signal manipulating program.

Each command consists of two parts: function and argument. Each part is specified by one key on the keyboard. A template laid over the keyboard relabels the keys for that purpose (Fig. 2). The template offers the following advantages: Labels on the template allow more meaningful names than 1- or 2-letter abbreviations, functions and

* Trademark of Bell Laboratories.

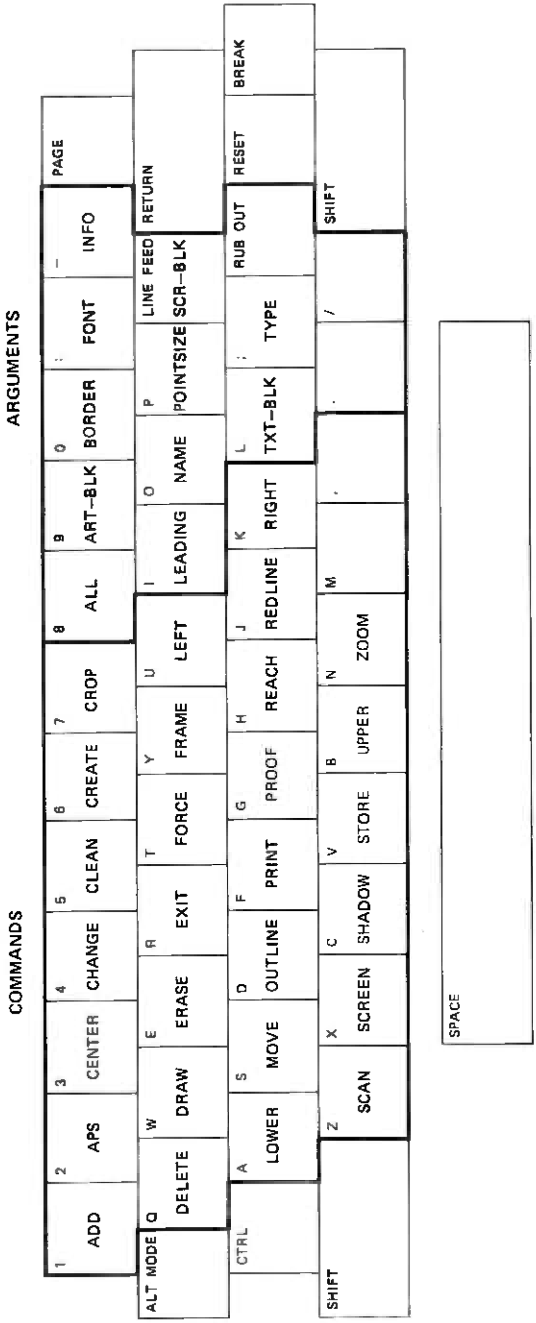


Fig. 2—Arrangement of commands and of their arguments on the keyboard.

arguments can be grouped separately, irrespective of their names, and the keyboard can still serve as a regular text input device.

Most commands are completed with the stylus by pointing to the additional information required. For example, text editing is performed by pointing directly to the character string to be changed followed by the replacement text from the keyboard. The terminal always mirrors the stylus position on the tablet as a tracking cross (cursor) on the TV screen. The cross consists of a thin vertical and horizontal line extending across the full screen. The data for the cursor lines are exclusively ORed with the display buffer. Thus, the lines remain always visible, though their color changes from the normal white-over-black background to black over white, red over blue, etc; this change of color is very helpful in the alignment of blocks. The desired cursor position can be signaled to the terminal by pressing on the stylus and thereby activating a pushbutton on its tip.

To get the best possible display resolution, the program scales each advertisement so that its longer boundary fits the full TV screen. In addition, a ZOOM command is available to enlarge any smaller area within the advertisement to the full size of the screen. All commands remain available to the zoomed display.

Color is used in various ways to guide the user:

(*i*) Colored template labels help to identify function and argument groups.

(*ii*) Red text on the screen warns that the particular text line exceeded its allotted space.

(*iii*) Solid blue areas indicate screened regions.

(*iv*) A blue rectangle shows the initial, a red rectangle the new location or size under stylus control for blocks being moved or resized.

(*v*) All the blocks of the applicable type are outlined in white, whenever a block needs to be identified.

More complex commands utilize a menu approach, where all possible options are listed for the user, who then selects one from the keyboard. All but the "quit this command" option terminate back to the menu.

The bell on the keyboard alerts the user to events such as illegal command terminations, completion of various stages in the execution of a lengthy command (e.g., drop-shadowing), or request of a user response to a command's option menu.

An abort pushbutton situated on the right-hand side of the keyboard lets the user abandon execution of the current command.

VII. GRAPHICS SOFTWARE

7.1 Blocks

The terminal uses layout blocks similar to those used in current production methods for the composition and paste-up of pieces of text and artwork. Blocks are the key feature of the terminal, since all commands operate on the blocks themselves (e.g., MOVE) or on the information contained within them (e.g., CHANGE TYPE).

Blocks are rectangular areas which can contain either text or artwork. Their size and position are shown by red outlines during the layout process. Blocks can be created, deleted, modified, and moved. It is permissible for them to overlap. In addition, the artwork contained in a block may be replaced, recropped, and/or edited. Type within a block may be edited, centered, left/right justified and its font, point size, or leading modified. Also, blocks may be framed (e.g., a box placed around a telephone number) with the thickness of the frame under user control.

Each block is allocated 34 bytes (maximum of 30 blocks per ad). The information stored for each block includes its size, position within the ad space, whether it is to be framed, and the thickness of the frame. For art blocks, it also includes the artwork's disk file name and its cropping parameters, which determine what part of the artwork will actually appear within the block. The information for screen blocks is identical to that of art blocks, except for the additional screen density parameter. The text block information includes a justification parameter and pointers to the block's first and last text line in the text buffer.

7.2 Fonts

A font is generally defined as the complete assortment of characters of one size and style needed for ordinary composition. Each font character is digitally represented as a series of points within a rectangular area, called the character matrix. Different character sizes are plotted by enlarging or reducing the matrix accordingly.

To generate a character on the phototypesetter, the typeface, point size, and character identity must be specified. This information is used at composition time to retrieve a coded version of the character matrix from the typesetter's font library and to output the font shape using the typesetter's own special decoding circuitry. The terminal's software also includes dot matrices for the characters of several fonts; they are used to produce the character shapes on the Versatec and the TV screen. The characters are plotted point by point, and they are sized by horizontal and vertical scaling of the dot matrix.

The experimental terminal's font library currently contains 16 frequently used Yellow Pages fonts. A separate program prepares and installs each font into the library of the terminal. Additional fonts can be added at any time. The resolution of the installed typefaces is equivalent to a 48-point master on the typesetter. This is also the largest permissible character size in a display ad. The accuracy was needed tor the production of high-quality outlined and drop-shadowed type, which is generated by the terminal and plotted as art on the typesetter.

The fonts are stored in one disk file with a record format of 512 bytes/record. The first record is an index to the font data in the file. Each index entry consists of three 2-byte parameters: (*i*) the font number, (*ii*) a pointer to the font's first data record within the file, and (*iii*) the font's master sizes available on the typesetter. Each font can have a maximum of 128 characters.

For each font, the character data are preceded by four records, each containing two 256-byte tables. The tables contain such entries as the width and height of each character matrix, the left and right side bearing values, the amount of clearance above each character, the starting records of the character matrices, and the number of data records per character. The data are run-length encoded. Each character starts at a record boundary.

7.3 Text

There are two ways to enter text into the layout: Lines are typed in from the keyboard or text is read in from a separately prepared text file. Advance text preparation is more suited for the production of new ads, while keyboard input seems more advantageous for updating existing ads.

Three parameters control the appearance of type in one text line. (*i*) The font number specifies the typeface, (*ii*) the point size defines the size of the characters, and (*iii*) the leading parameter determines the extra spacing between the characters of consecutive text lines. Leading is executed before the characters are plotted and can possess negative values.

All text information is stored in a separate array (currently, 2000 bytes). Text lines, together with their font, point size, and leading parameters, are stored in the array upward. Pointers to the text lines run from the top of the array down. Each pointer consists of two entries, an array subscript to the start of each line and the line's character count.

Text lines are always linked to a text block. Deletion of a block results in the removal of its type as well. Type cannot be moved across text blocks.

Text commands perform substitution, insertion, deletion, and upper/lowercase conversion. Font, point size, and leading parameters may be altered separately or combined into a single command. The above commands can be applied to any string of consecutive characters within a text line, or to a sequence of entire text lines, or to a whole text block. Text in a block can be centered, left-justified or right-justified. Special effects such as outlining or drop-shadowing may be applied to individual text lines. The outlined or shadowed text can be screened.

7.4 Artwork

The terminal handles two-tone (black-and-white) pictures such as ink drawings, logos, trademarks, and other illustrations which are scanned and stored digitally on disk files. Logos and artwork which might be needed for more than one advertisement are kept under a separate file directory and are accessible to all advertisements. Special artwork for individual ads is stored under the same directory as the layout specifications.

All artwork is digitized on the terminal. The difference between typesetter and scanner resolutions (3:1) can be compensated for by enlarging the input picture to three times the intended output size. Each scan results in data from the entire 8 × 10-inch scanning surface. Thus, one or more pictures may be digitized together, as long as they fit the scanning surface. The inputted data are immediately displayed on the TV screen, where individual pictures can be isolated and saved on disk. The cropping boundaries are defined by using the stylus/tablet to position vertical and horizontal lines.

Art data are currently stored without compression because the 96 Mbyte disk should have sufficient capacity to store picture data for the time-and-motion trial now in progress. However, if needed, average compression of about 5:1 would not be difficult to implement by applying the run-length coding programs we already use for storing font data. The pel data are stored on disk, scan line after scan line. A dummy scan line at the beginning of the file is used for storing scan parameters. The number of 16-bit words in each scan line is stored in the first 16-bit field of the dummy line; the total number of picture scan lines in the file is stored in the second 16-bit field.

Artwork is first scaled to fill the art block, often resulting in different horizontal and vertical scale factors. The FORCE command adjusts either the width or the height of the art block, so that the ratio is the same as that of the original art. The side to be changed is under user control.

Artwork may be screened, outlined, or drop-shadowed. A pencil/eraser function is also available for minor touchups. More extensive corrections should be made on the original, which should then be rescanned.

7.5 Special functions

In Yellow Pages production, the most frequently used artistic effects that involve time-consuming photographic techniques are screening, outlining, and drop-shadowing, shown in Fig. 3. Algorithms to automate these processes were developed by Franklin[8] as part of an interactive picture manipulating system. The algorithms were then adapted and programmed into suitable commands for the terminal.

Outlining and drop-shadowing are normally applied to text only, to create custom lettering from conventional fonts. The commands, however, are completely general and can be used for interesting effects on artwork also. Representative specifications of thickness of outline and size and direction of shadow were selected in advance and are not

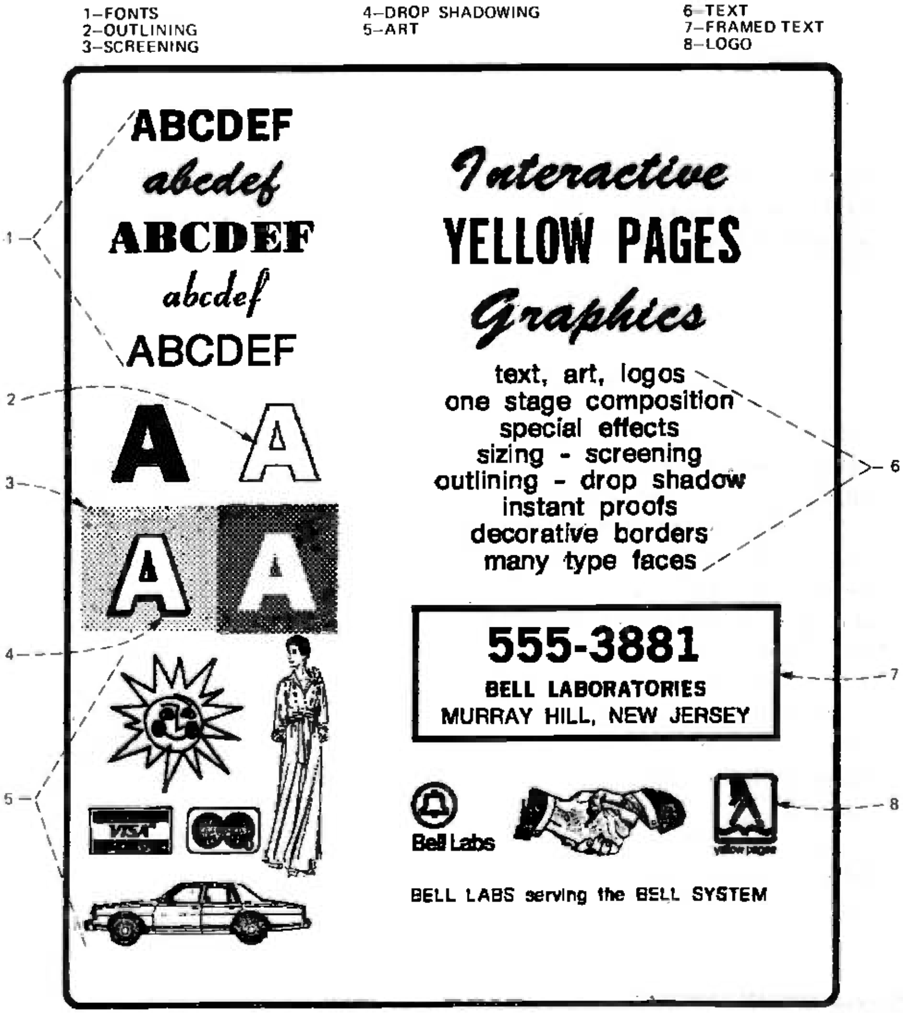


Fig. 3—Sample advertisement showing most of the terminal's capabilities. This picture was outputted on the electrostatic printer. Its quality can be compared with that of the typesetter output shown on page 2189.

under user control. The specificatioñs are:

Width of outline:	10 decipoints (= ⅒th point)
Width of shadow:	30 decipoints
Position of sun:	Northeast.

The specifications are for 48-point objects; the thicknesses vary proportionately for other sizes.

7.5.1 Outlining

This operation converts a black-and-white (input) picture into a new (output) picture in which only the black-and-white transitions of the input are shown.

For this purpose a 3 × 3 window is shifted bit by bit across the entire input picture. The center element of every matrix position is modified according to the status of its surrounding elements and transferred to the corresponding position of the output picture. If all surrounding elements are zero, the center element is transferred as zero (no threshold). If any surrounding element is nonzero, the center element is transferred complemented (possible threshold).

The generated outline in this case is always just one pixel wide. Larger windows would generate thicker outlines but would require a greatly increased number of operations. The terminal, therefore, always uses a 3 × 3 window and generates the required thickness on the output picture by plotting a series of points for every threshold value found.

In the process of outlining text, the terminal first converts a text block into an art block and a text line into a digital picture file. The font data used for conversion come from the terminal's local font library.

7.5.2 Drop-shadowing

This operation generates a new (output) picture consisting of the outlined (input) picture with a shadow added. Initially, we again intended to use an algorithm by Franklin. However, this was changed soon to a different method,[9] which required less core space and avoided the processing of individual bits.

Franklin's algorithm computes the shadow by shifting a matrix bit by bit across the entire input picture. The size of the matrix determines the length of the shadow. The corner element of the matrix selected to compute the corresponding output element determines the direction of the shadow. An output element is zero if the input corner element is nonzero or if the corner element and all the other elements along its diagonal are zero. An output element is nonzero if the corner element is zero and any of the other elements along the diagonal is nonzero. To achieve the drop-shadowed effect, the outlined picture and the shadow are ORed together.

The new algorithm combines an input scan line, the corresponding output scan line and as many subsequent output lines as are needed for the length of the shadow; initially, all output scan lines are zeroed. The shadow is calculated by first ORing the scan lines of the initial and the outlined picture together. These data are ORed to the corresponding output scan line. The data are then shifted left by one bit and ORed to the following output (shadow) scan line. This operation is repeated, always shifting the input data one more bit until all the shadow scan lines are processed. Finally, to mask out any hidden shadows, the current output scan line is exclusively ORed with the initial picture scan line and written on the output file. Before processing the next input line, the first shadow scan line becomes the current output scan line, the second shadow scan line becomes the first one, and so on; the last line is zeroed.

7.5.3 Screening

In this operation, the computer program traces the boundaries of any area enclosed by a continuous line and then fills this area with screening dots of selected density. The area to be screened is identified on the TV screen by pointing to any part of it with the stylus. The screened area is recorded on a separate screen file.

On the final output, screened areas are shown as regions filled with evenly spaced spots of uniform size. The chosen spot size determines the perceived gray level (screen density). Areas of different density are recorded in separate files. If no screen file exists at the start of the screening operation, the file is created and zeroed.

The tracing algorithm employs a stack to manage the starting points for the boundary search. The first stack entry is the initial starting point. When the stack becomes empty the search is complete. For each starting point, the scan line to the left and right of the initial position is searched for a zero to nonzero boundary. The elements between the left and right boundaries are then set to nonzero on the corresponding screen line. Finally the immediate scan line above and below is searched between the two boundaries from left to right for nonzero to zero transitions. (A nonzero element is assumed beyond the left boundary.) The transition point (zero) coordinates are pushed on the stack. This procedure is repeated for all points on the stack. The scan lines being searched are the picture scan lines ORed with their corresponding screen lines. To improve efficiency and minimize the required I/O, the stack is first searched for another point on the current scan line. If found, this point is popped and it becomes the next starting point; if the search fails, the last point entered will be popped from the stack.

7.5.4 Pencil/eraser

This function uses the tablet/stylus to simulate a pencil/eraser. It affects the artwork on the TV screen and, simultaneously, the data on disk. The function is best suited to correct blemishes and to straighten out lines. It was not intended for freehand drawing or curves. Within its scope, it has proved to be convenient and easy to use.

At the start, the user selects a particular section of the picture as the work area. For this purpose, the program first shows the complete picture on the screen. The chosen work area is then redrawn on the full screen and the pencil/eraser "tip" appears on the screen.

The tip is shown as a white (pencil) or red (eraser) rectangular area and follows the movement of the stylus. Actual drawing/erasing occurs only while the pen switch is depressed, with the mode depending on the status of the draw/erase sense switch. In erase mode, the pels covered by the tip are set to zero, in draw mode, to nonzero. The size of the tip is adjustable. Adjustment is initiated by the TIP pushbutton at the right-hand side of the keyboard, which cancels the current tip and starts the cursor for creating a new one. The task to be accomplished should influence the shape of the new tip. For example, a small square tip is useful for fine detail and isolated blemishes, a tip in the shape of a narrow horizontal or vertical bar is more suited to smooth out ragged edges along horizontal or vertical lines.

The user can specify a scale factor, which controls how large each picture element will appear on the screen. Larger scale factors are more suited for work with intricate detail, because the tip is less sensitive.

User requests, other than a tip size change, are initiated by the ABORT pushbutton, which terminates the draw/erase loop and causes return to the options selection loop.

The pencil/eraser function is available both as a command and as a menu item under the SCAN command.

VIII. COMMAND SUMMARY

Command names are listed below in alphabetical order; a list of their possible arguments is included in square brackets following each command name. A more detailed description of the commands can be found in the appendix.

ADD	[border, name, type]
APS	[all]

CENTER	[all, type, text-block]
CHANGE	[border, name, type, font, point size, leading, text-block, art-block, screen-block]
CLEAN	[all, art-block, screen-block]
CREATE	[art-block, text-block]
CROP	[art-block, screen-block]
DELETE	[all, name, type, text-block, art-block, screen-block]
DRAW	[all, text-block, art-block, screen-block]
ERASE	[all, text-block, art-block, screen-block]
EXIT	[all]
FORCE	[art-block, screen-block, point size, leading, font]
FRAME	[text-block, art-block, screen-block]
LEFT	[all, type, text-block]
LOWER	[all, type, text-block]
MOVE	[text-block, art-block, screen-block]
OUTLINE	[text-block, art-block]
PRINT	[info, all, border, type, text-block, art-block, screen-block]
PROOF	[all]
REACH	[text-block, art-block, screen-block]
REDLINE	[all]
RIGHT	[all, type, text-block]
SCAN	[all]
SCREEN	[art-block]
SHADOW	[text-block, art-block]
STORE	[all]
UPPER	[all, type, text-block]
ZOOM	[all, art-block]

Argument names are self-explanatory, except for ALL. In some cases, ALL implies "everything" as in ERASE ALL (blocks); in some cases ALL implies "everything of a certain kind" as in LOWER ALL (type); in some cases it is simply a dummy argument to conform to the standard command sequence as in SCAN ALL.

9. OUTPUTTING TO THE TYPESETTER

9.1 Overview

The final task of the terminal is to output the entire advertisement (text and pictures) on a typesetter; this typesetter output is then used for making the printing plate. Our terminal computes commands for the Autologic APS-4 typesetter which are then recorded on magnetic tape. The tape is later mounted on the typesetter and the advertisement is outputted without further attention.

The APS-4 was used for our experiments because it was the typesetter available where the terminal is being field-tested. Later the more modern APS-5 was made available to us and our research profited by being allowed to use it. The knowledge gained with the APS-4 and APS-5 should be easily transferable to any other CRT (or laser) typesetter.

Considerable study, testing, and typesetter adjustments were needed before the typesetter would produce pictures with the quality and speed needed for directory production. This was so even though the APS-4 (and other CRT typesetters now in use at telephone companies) already has basic picture-drawing capabilities which are used for drawing font shapes when outputting text. However, these font drawing methods could not conveniently be adapted to drawing more complex images. Another way available for drawing pictures was to adapt the APS's "rule" commands, even though they were originally intended to draw single straight lines only. We found that considerable testing (by ourselves), tuning, and modification (by the manufacturer) was needed to make this method usable; even then, its quality was not entirely satisfactory and its speed slower than desirable. Finally, a new method was introduced for graphics output on the APS-5 (called the "graphics" mode), and we were its first users.

As a result of these experiments, we have demonstrated a reliable method for typesetter graphics which produces pictures whose quality, for directory purposes, compared well with those produced by photography. Also, speed trials using about a half-dozen four-column pages showed that entire pages filled with representative graphics material and text were produced in about 90 seconds, which was considered a very satisfactory performance.

9.2 Graphics on the APS-4 and APS-5

9.2.1 The nibble code

The instruction set of the APS-4 and APS-5 includes the usual beam positioning commands and specifications for font, point size, and leading. Text characters are painted by closely spaced vertical strokes of the CRT beam. The font shapes are specified in terms of the end points of the vertical strokes. The stroking data are coded into a form of run-length-differential code. Individual codes consist of 2-bit units, which is the reason for calling them "nibble" code (by contrast to the 8-bit byte). The APS-4 and APS-5 use a hardwired nibble decoder whose output controls the CRT beam. This decoder can only deal with a maximum of four vertical strokes in one "column" and is therefore not convenient or efficient for more complex shapes.

9.2.2 The rule command

The APS machines also have "rule" commands which generate horizontal lines of selectable width and height. The lines are again produced by a series of closely spaced vertical strokes. This is the first of the two different methods we used for typesetting graphics. Its principal drawback is that fractional data about rule lengths are not handled adequately (rounding troubles), so that the end points of a number of chained "rules" along a single horizontal line quickly become inaccurate; this results in a ragged appearance of right-hand outlines of complex pictures. Most of these troubles were remedied by Autologic after we demonstrated shortcomings to them. However, in pictures that consist of large numbers of short "rules" (such as a screen pattern), the trouble has not completely disappeared. The method is also relatively slow.

Pictures to be typeset using rule commands are drawn by plotting individual scan lines. For each line, the picture elements are analyzed for zero and nonzero run-lengths. Nonzero sequences are used to set rule widths, and zero sequences are translated into beam positioning commands. Any required scaling is performed by direct sampling of the input data. The height of rules is normally set to one decipoint. The height of the rule is increased only when the picture has to be enlarged. In this way, the rule has to be drawn only once, instead of plotting the data several times as required by the ratio of enlargement.

Screen data are typeset similarly, except that nonzero sequences are translated into dot sequences instead of solid black rules. Traditionally, screen measures are expressed in diagonal lines per square inch, with screen dots arranged along parallel lines angled at 45 and 135 degrees to a horizontal line. On the typesetter, however, we plot *horizontal* screen lines. Each dot is drawn as a single square rule; its size is determined by the screen density. The dot spacing (in decipoints) along horizontal lines for a given typesetter can be computed to

$$s = sqrt(2)*r/d$$

where r is the resolution of the typesetter (in decipoints/inch) and d is the diagonal screen measure (in screen lines/inch). Vertically, screen rows are repeated in intervals of half the horizontal spacing, with alternate rows being staggered by the same amount. For example, an 85-line screen on the APS-4 with 720 decipoints per inch resolution has a horizontal spacing of 12 decipoints. Consecutive screen rows are 6 decipoints apart, with every other row indented by 6 decipoints.

Before plotting a screened data file, the program initializes the width and height of a screen rule. Also, a basic command loop is set up in one of the APS buffers: Draw one rule, move beam horizontally to start of next rule, decrement loop count, quit loop if count is zero, otherwise repeat loop. Subsequently, screen sequences scan can be plotted by just two APS commands: Set loop count to the number of screen dots required and execute the command string.

9.2.3 The graphics mode

Recently, Autologic developed an alternative graphics feature on the APS-5 for faster and more flexible typesetting. Again, vertical stroking under the control of the hardwired nibble decoder is utilized. The method is based on a special font with 256 characters identified by an 8-bit code. Each character consists of up to four vertical strokes in a single column. The pattern of strokes in each special character is the same as the pattern of ones in the 8-bit code which specifies it. Pictures are typeset in a swath of eight horizontal scan lines at a time; 8-bit character codes for specifying a picture are derived by slicing the eight scan lines vertically at every pixel position. The "graphics" mode provides a repeat (run-length) code to avoid repeating identical codes. Thus fewer bits are needed to specify a picture, resulting in less storage and higher output speeds.

For screen data the scan lines are logically ANDed with the screen pattern before remapping; otherwise, the process is the same as for artwork.

One item that has not yet been resolved concerns lens magnification. For text, the typesetter program compensates for the different lenses, and the resulting output always has the same size. For the "graphics" mode, no lens compensation is performed. Artwork with a scale factor of 1.0 will be properly sized only on a machine with a 2.22X lens. To obtain the same result on a machine with a 2.66X lens, the user must adjust the artwork's scaling parameters to 0.8326.

9.3 Results

Of major interest in typesetting graphics are output speed and quality.

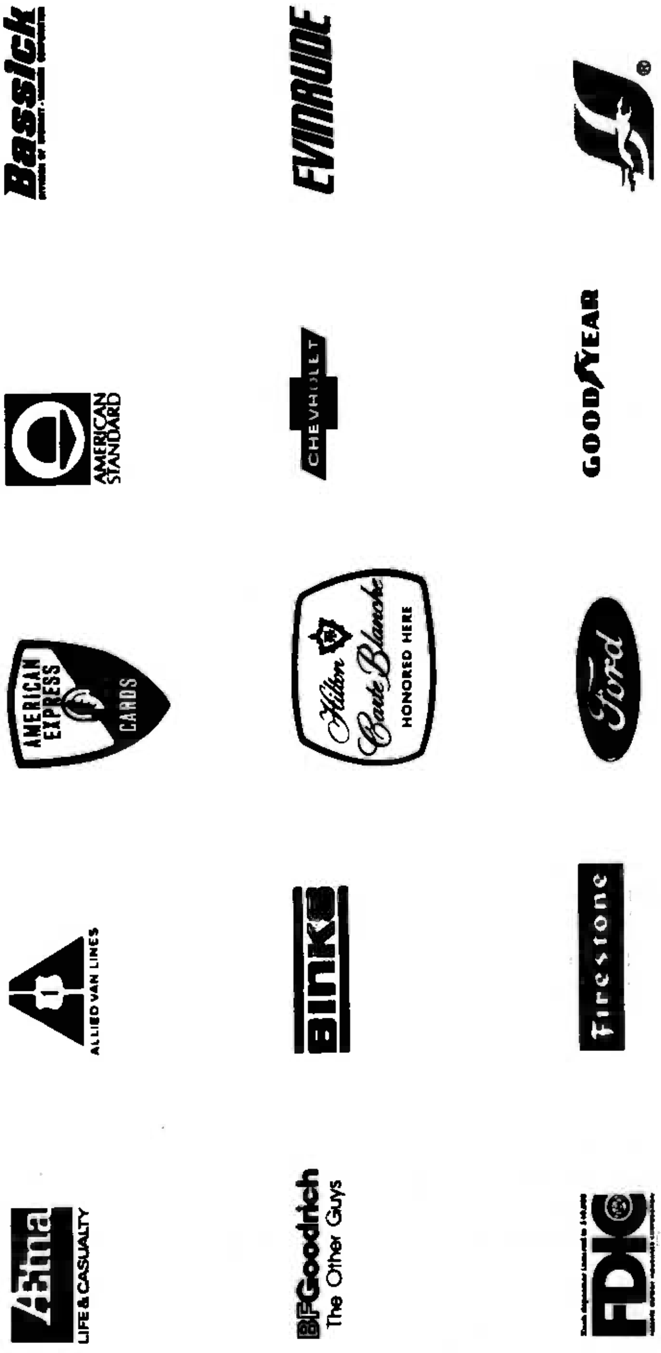
To obtain a meaningful timing comparison, we produced the same page of 30 logos (Fig. 4) by the "rule" mode on the APS-4 and on the APS-5 and by the new "graphics" mode on the APS-5. The logos had been scanned previously and were outputted in six rows on a single page using the rule and the graphic modes.

The typesetter timing tests for the entire page of 30 logos gave the following results:

(*i*) Using rule commands the APS-4 was over four times faster than the APS-5.

(*ii*) Using the graphics mode (on the APS-5), was seventeen times faster than the rule commands on the APS-5 and four times faster than the rule commands on the APS-4.

The output quality of artwork (except for screens) produced by either the rule or the graphics methods was very high. The graphics mode also produced excellent, uniform-looking screen patterns. Occasionally, a series of dots in a column appeared fainter but were still uniformly spaced. This effect can be explained by the grain of the paper and varies with screen measure and density. Screen patterns produced by rule commands on the APS-5 showed regularly spaced


Aetna
LIFE & CASUALTY
ALLIED VAN LINES
AMERICAN EXPRESS
CARDS
AMERICAN STANDARD
Bassick
BFGoodrich
The Other Guys
BINKS
Hilton
Carte Blanche
HONORED HERE
CHEVROLET
EVINRUDE
FDIC
Firestone
Ford
GOODYEAR

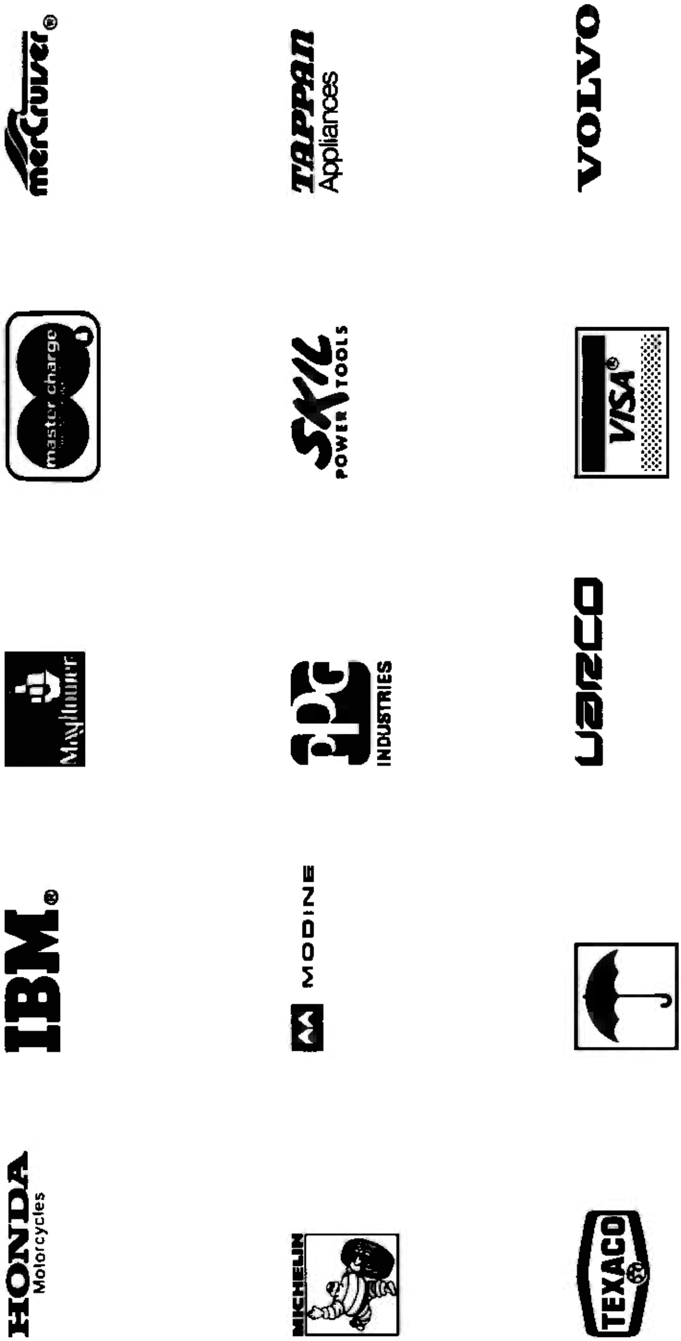


Fig. 4—Thirty logos used for timing.

thin vertical stripes. They seem to be introduced by the APS-5 when mapping absolute decipoint values to actual beam distances, at which point the lens magnification also has to be considered. Various attempts to remedy this problem did not entirely succeed. Results on the APS-4 varied with the particular machine used.

X. FIELD TRIALS

Two different field trials of the experimental terminal have been or are being carried out. The goal of the first trial—which has already been completed—was to establish what type of personnel is best suited to operate the terminal, to gain experience for establishing operating routines, and to identify modifications for better operation, particularly for making the operator's task simpler. The goal of the second trial—now getting under way—is to make cost and time comparisons between the new terminal and the conventional methods of producing advertisements.

In the first trial, two persons were trained in the use of the terminal. One had years of artistic layout experience, the other trained clerical personnel for various tasks. After four weeks, a series of program changes were installed to improve the reliability of the terminal, to increase feedback to users, and to adopt certain terminologies more meaningful to the users. The following changes were made:

(*i*) The command key sequence was standardized to two keys followed by a carriage return.

(*ii*) The capability of scanning partial pages was removed; it was more efficient always to scan the whole page and crop the material on the TV screen.

(*iii*) Some new messages, such as "scan in progress," were added to inform and reassure the user on the status of commands which required more than 1 or 2 seconds for execution.

(*iv*) Some terminology was changed to names more meaningful to users: pictures became "art," text lines and text strings became "type," output from the electrostatic printer was renamed "proof."

(*v*) Automatic saving on disk of the layout specifications after every command was added to facilitate painless continuation after an unexpected interruption or program abort. After these changes, evaluation of the terminal proceeded smoothly.

The lessons learned from the first trial were as follows:

(*i*) The terminal could indeed be a valuable tool.

(*ii*) Clerical personnel can operate the terminal, especially if the advertisement worksheet—scanned and displayed on the TV screen—could be used as a guide for positioning and sizing the layout blocks.

(*iii*) The terminal should be used mainly for the input and layout of artwork; text should be prepared separately in advance and only modified online.

(*iv*) A pencil/eraser option must exist to edit the scanned data, such as to delete erroneous spots and possibly touch up logos and artwork.

(*v*) There should be a second trial to make a more detailed comparison (including a time and motion study) between our terminal and conventional production methods; a large number of ads (100 or more) would have to be prepared by both methods.

(*vi*) A clerk must be trained and become proficient in the use of the terminal before the study can begin.

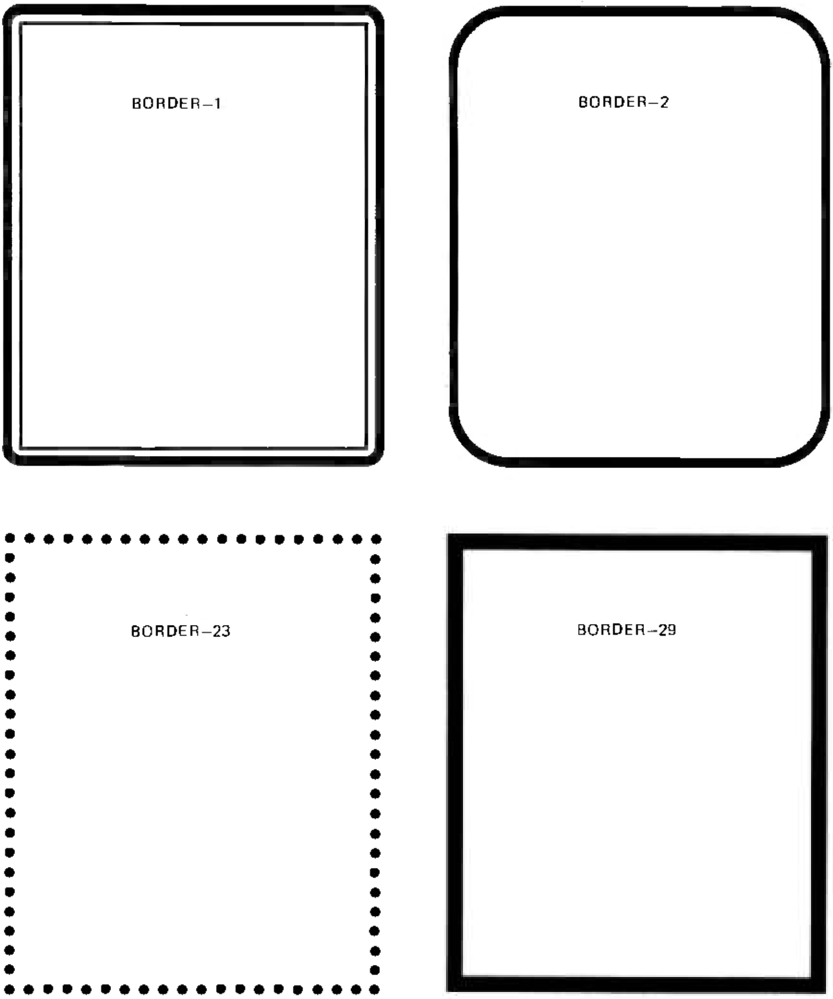


Fig. 5—Four of the ten decorative borders available on the terminal.

(*vii*) To produce real ads, the terminal must have a reasonable number of different fonts, decorative borders (Fig. 5), and adequate storage facilities for approximately 100 illustrations, logos, and stock art.

The second field trial has just started. Its objective is to collect data for a cost and time comparison between the current advertisement production method and the experimental terminal method. To establish costs, the layout process is subdivided into individual steps (such as scanning of worksheet, scanning of artwork, adding a screen) and each step is to be timed separately. A set of 50 advertisements was selected which covered the variety of material considered typical of those appearing in Yellow Pages directories. Statistics on the type of revision work normally encountered were also collected. In the first part of the trial, the 50 advertisements will be created afresh and timed on the experimental terminal, just as if they were new advertisements. Later, the same time measurements will be made for "revising" existing advertisements, using the 50 advertisements prepared in the first part of the trial. Separate estimates will be made of the economic and other advantages that result from greatly shortening the time span between closing a directory and its printing. Further, the results from this field trial should help in the assessment of future terminal needs and define areas for further study.

The terminal had to be expanded considerably to accommodate the much greater variety of work and the simulated production environment. Preparations for the trial included:

(*i*) The installation of a 96 Mbyte disk for storing the additional fonts, illustrations, and logos.

(*ii*) The installation of 15 additional typefaces and the introduction of run-length coding for storing them in less disk space.

(*iii*) Programs to generate 10 stock borders by rule.

(*iv*) Pencil/eraser capabilities.

(*v*) A program for separate text input in advance.

(*vi*) The capability of framing text, such as placing a box around a telephone number.

(*vii*) Selection and training of a clerk in the use of the terminal.

All preparations have now been completed and the trial is under way.

XI. CONCLUSIONS

The work described in this paper demonstrates that modern directory production methods are able to process pictures with the same ease with which they currently handle text. In our experimental terminal, photographic methods were replaced by digital methods to position, size, crop, screen, outline, shadow, and store pictures conveniently and efficiently; methods for typesetting artwork with high speed and quality have also been established. As a result, directory pages can now be typeset in a single pass, without paste-ups, and corresponding production cost reductions and sales advantages can be expected.

The new techniques demonstrated on our experimental terminal are to be included in the full production prototype for the preparation, maintenance, and conversion of all Yellow Pages art material which is now being developed by New York Telephone Company with AT&T support.

XII. ACKNOWLEDGMENTS

We would like to express our thanks for the imaginative, forward-looking, and substantial help and encouragement we received from Bell System companies. In particular, we are grateful to P. J. Desmond and B. R. Jansson (New York Telephone), M. Jensen (Northwestern Bell), J. C. O'Neel (Pacific Telephone), and E. R. Buckstine (AT&T). B. R. Moore, Marlene D. Bierig, and J. A. Zanette (all at Pacific Telephone) contributed expert advice and hard work in helping us improve the all-important man-machine communication aspects of the terminal and in carrying out the field trials. We thank R. J. Cain and H. H. Finney (both at New York Telephone) for their help in making the APS-4 and APS-5 produce acceptable graphics. We also wish to acknowledge gratefully the contribution of O. C. Jensen of Bell Laboratories, who designed, constructed, and maintained the terminal's graphics hardware.

APPENDIX

Command Description

The functions of the various commands are briefly explained here. Multipurpose commands appear under each appropriate subheading. Detailed operating instructions can be found in the user's manual.[10]

A.1 General block commands

CHANGE is used to adjust the dimensions of a specified block within the work space.

CREATE generates a new block of the specified type.

DELETE deletes the indicated block and its associated information from the layout.

DRAW, ERASE operate on the TV display only. DRAW displays the selected block on the screen, ERASE erases it. The block will still appear on the final output.

FRAME lets the user specify an outline or "frame" around a block for the final output. The thickness of the outline is specified in ⅒ point units.

MOVE repositions the specified block within the layout space.

PRINT prints the coordinates (relative to the top left corner of the advertisement) of the designated block and, for art or screen blocks, the associated data file name and, for text blocks, the text lines together with their font, leading, and point size parameters.

REACH allows the user to address a block completely contained within another block of the same type.

A.2 Text handling commands

ADD allows the user to add type to an empty block or to insert or append type to existing text lines in a block, or it can get the data for all text blocks from a previously prepared text file.

LEFT, CENTER, RIGHT are used to justify or to center text lines in the specified text block.

CHANGE is used to edit character strings or text lines, or modify font, leading, and/or point-size parameters.

DELETE is used to delete character strings or text lines.

UPPER, LOWER convert character strings, text lines, or text blocks to upper- or lowercase. LOWER is the more frequently used command, since all input from the Tektronix keyboard is initially in uppercase.

FORCE sets the text parameters of the specified block back to their default values. This command is useful after a change of the text parameters causes type to exceed the allocated block space.

PRINT is used to get a listing of the indicated text lines with their current font, leading, and point-size values.

A.3 Art processing commands

ADD, CHANGE are used to associate a picture data file with a specified art block.

CLEAN invokes the pencil/eraser function to touch up the artwork in the current art block.

CROP is used to frame the original artwork. Only the data inside the frame will appear in its associated art block. This feature allows the same artwork to be used for various ads, with each ad showing different parts of the stored artwork.

DELETE deletes irretrievably the specified picture data file from disk.

FORCE adjusts the parameters of art or screen blocks in such a way that the associated picture has the same scale factor in both the horizontal and vertical dimension, and thus appears undistorted. This is important for trademarks and logos. For other artwork, the layout artist may actually prefer the effect of having the artwork appear condensed or expanded along one dimension. FORCE allows the user to specify which of the four sides of the art block should remain unchanged in size as well as position within the workspace.

PRINT lists the file name associated with the data of an art or screen block.

SCAN is the command to convert artwork from a paper drawing into digital form for handling by the terminal. The command employs an options menu to perform various operations connected with scanning: crop the scanned data, store the data under a user assigned name, copy data from one file or device to another for backup or restoration purposes, print some file characteristics, and use the pencil/eraser.

A.4 Commands for producing special effects

OUTLINE produces a new picture data file that contains only the outlines of the initial artwork or the initial text characters.

SHADOW produces another picture file that contains the outlined picture with a shadow added.

SCREEN permits the user to specify areas for screening.

A.5 Commands for producing output on the typesetter or on the electrostatic printer

APS computes and records on magnetic tape the commands to drive the Autologic APS-4 or APS-5 phototypesetters to produce the final version of the ad.

PROOF calculates and then plots an image of the ad on the electrostatic printer.

A.6 Miscellaneous other commands

CHANGE is also used to assign or change the decorative border around the ad.

EXIT is the command to wrap up the layout program properly.

PRINT can also be used to get the overall statistics of the current ad. In this case, the size of the ad, the number of blocks of each type, the number of text lines, and the decorative border type are listed.

REDLINE reverses the current status of either showing or not showing the red outlines around all blocks during the layout process.

STORE is used to save the layout specifications under a user specified name. The information includes everything except the picture and screen data files.

ZOOM allows the user to enlarge a specific part of the ad across the full screen. All commands remain available to the zoomed ad.

REFERENCES

1. A. J. Frank. "High Fidelity Encoding of Two-Level, High Resolution Images," IEEE Inter. Conf. of Commun., June 1973, pp. 26-5—26-10.
2. A. J. Frank and R. H. Groff, "On Statistical Coding of Two-Tone Image Ensembles," Proc. SID, *17/2,* Second Quarter 1976, pp. 102-110.
3. R. G. Todd, "A Hardware Decoder for Two-Dimensionally Compressed Pictures," unpublished work.

4. I. G. Gershkoff, "An Interactive System for Page Layout Design," unpublished work.
5. P. B. Denes and I. G. Gershkoff, "An Interactive System for Page Layout Design," Proc. ACM Conf., Nov. 1974, pp. 212–221.
6. *UNIX*™ Programmer's Manual, Bell Laboratories, March 1977.
7. L. H. Nakatani, Computer-Aided Signal Handling for Speech Research," J. Acoust. Soc. Amer., *61*, (Apr 1977), pp 1056–1062.
8. D. L. Franklin, "An Interactive Picture Manipulation System," unpublished work.
9. K. C. Knowlton, Bell Laboratories, private communication.
10. B. E. Caspers, "User's Guide for the Yellow Pages Design Terminal," unpublished work.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

Application of Clustering Techniques to Speaker-Trained Isolated Word Recognition

By L. R. RABINER and J. G. WILPON

(Manuscript received July 25, 1979)

Speaker-trained, isolated word recognizers have achieved notable success in a wide variety of applications. The training for such systems generally involves a single (or sometimes two) replication(s) of each word of the vocabulary by the designated talker. Word reference templates are then formed directly from these replications. In recent work on speaker-independent word recognition, it has been shown that statistical clustering procedures provided an effective way for determining the structure in multiple replications of a word by different talkers. Such techniques were then used to provide a set of reference templates based on the clustering results. In this paper we discuss the application of clustering techniques to speaker-trained word recognizers. It is shown that significant improvements in recognition accuracy are obtained when using templates obtained from a clustering analysis of multiple replications of a word by the designated talker. It is also shown that recognition accuracy did not change with time (over a 6-month period) for any of the subjects tested, thereby indicating that the reference templates were reasonably stable.

I. INTRODUCTION

Although a great deal has been learned about isolated word speech recognition systems,[1-14] several key issues are not as well understood as others. One such issue is the manner in which the word reference templates for such a system are obtained. To date, there have been at least three distinct ways of obtaining templates, including:

(*i*) Casual training in which the designated talker (for a speaker-trained system) speaks each word of the vocabulary (one or more times) and a reference template is created for each spoken word.[3,4] Thus, for casual training, there is a direct correspondence between a spoken token of the word and the reference template.

(*ii*) Averaging methods in which the designated talker (for a speaker-trained system) or a set of talkers (for a speaker-independent system) speaks the word a number of times and a weighted, time-normalized average of the feature sets for that word is used as the reference template.[1,7,15]

(*iii*) Statistical clustering methods in which a set of talkers speak the word and a statistical pattern recognition algorithm is used to group the feature sets of the tokens into a set of clusters.[14,16] The similarity of tokens within a cluster is high (small intratoken distances), whereas the similarity of tokens in different clusters is low (large intertoken distances). Reference templates are obtained by representing each cluster by a single template (either using a minimax approach,[14] or via averaging techniques[17]). Thus, a word is generally represented by a *set* of templates rather than one or two templates.

The third method above, the statistical approach, has been successfully applied to a speaker-independent word recognizer for a variety of vocabularies.[14,17,18] It is the purpose of this paper to show how this technique can be applied in a speaker-trained system to further increase their accuracy and robustness over systems in which the reference templates are obtained by casual training.

The organization of this paper is as follows. In Section II we review the operation of the basic word recognizer and the clustering procedures. In Section III we present the experimental procedures used to obtain the data for training and testing the system. The statistics of the clustering for each of three talkers are presented in Section IV, and the recognition accuracy as a function of key system parameters is given in Section V. Finally, Section VI discusses the results and their implications for practical implementations of word recognition systems.

II. REVIEW OF THE WORD RECOGNITION SYSTEM

The word recognizer, shown in Fig. 1, is similar to the one originally proposed by Itakura,[3] and has been used in a variety of applications.[4,13,14,16-18] Telephone line input signals (100- to 3200-Hz bandwidth) are digitized at a 6.67-kHz rate, and a p = 8th-order autocorrelation analysis is performed on overlapping frames of N = 300 samples (45 ms), with an overlap of 200 samples between frames. Prior to the autocorrelation analysis, each frame of data is preemphasized with a first-order digital network with transfer function $(1 - 0.96\ z^{-1})$ and windowed by a 300-sample Hamming window. If we denote the lth preemphasized, windowed frame of speech as $\hat{x}_l(n)$, $0 \le n \le N - 1$, then

$$\hat{x}_l(n) = \hat{x}(l \cdot S + n) \cdot w(n) \qquad 0 \le n \le N - 1, \quad 0 \le l \le L - 1, \quad (1)$$

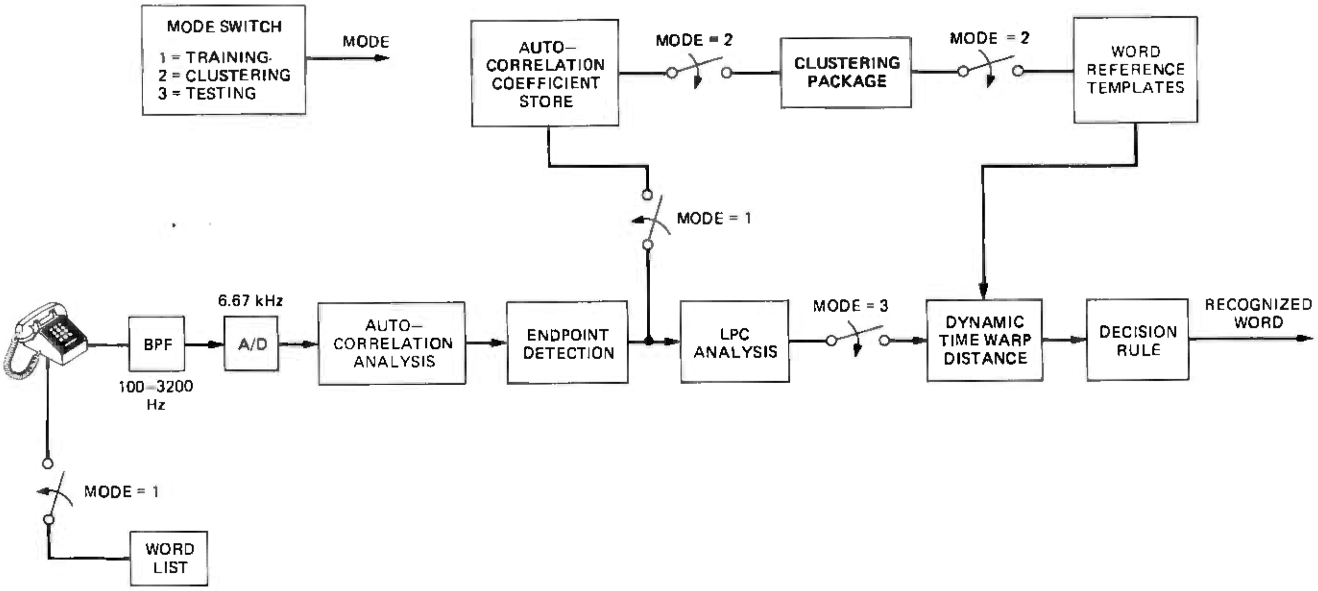


Fig. 1—Block diagram of the word recognizer.

where $\hat{x}(n)$ is the preemphasized speech, $w(n)$ is a Hamming window, $S = 100$ samples (15 ms) is the shift in samples between adjacent frames, and L is the number of frames in the recording interval. The autocorrelation coefficients of the lth frame, $R_l(m)$ are given by

$$R_l(m) = \sum_{n=0}^{N-l} \hat{x}_l(n)\hat{x}_l(n+m) \quad 0 \le m \le p \tag{2}$$

$$= \sum_{n=0}^{N-1} \hat{x}(lS+n)\hat{x}(lS+n+m). \tag{3}$$

The zeroth autocorrelation coefficient of each frame ($R_l(0)$) is the energy in the frame. The time pattern of $R_l(0)$ (i.e., $R_l(0)$ vs l) is used to determine the end-point boundaries of the spoken, isolated word in a simple manner based on the measured energy of the background noise in the recording environment.[14]

As noted in Fig. 1, there are three modes of operation of the word recognizer, namely, training, clustering, and testing (normal usage). As discussed earlier, for many speaker-trained systems, the training mode is simply a recording of each word of the vocabulary and the clustering (i.e., formation of word reference templates) is a direct conversion from stored autocorrelation coefficients to the format required for the word reference template. (For this recognizer, the word reference templates are stored as frames of autocorrelated linear predictive coding (LPC) coefficients. This is explained later in this section.) For a statistical clustering approach, however, training consists of storing sets of autocorrelation coefficients for a number of replications of each word of the vocabulary, and clustering consists of grouping the replications of each word into clusters and creating a single word reference template for each cluster.

For the third mode of the system, namely the testing or normal usage mode, following end-point detection, an LPC analysis of each frame is performed (the autocorrelation method), and each autocorrelation frame is converted to a normalized form as follows. If we denote the frames of autocorrelations in the test word as $R_i(m)$, $i = 1, 2, \cdots, NT$, and the LPC prediction residual of each frame as E_i, then the test frame parameters are given as

$$V_i(m) = \frac{R_i(m)}{E_i} \qquad 0 \le m \le p, \quad 1 \le i \le NT. \tag{4}$$

For the word reference templates if we denote the jth frame of LPC coefficients (derived from the autocorrelation coefficients) as $a_j(k)$, $0 \le k \le p$, then the jth reference frame parameter set is given as

$$P_j(m) = 2 \sum_{k=0}^{p} a_j(k)a_j(k+m) \qquad 1 \le m \le p \tag{5a}$$

$$= \sum_{k=0}^{p} [a_j(k)]^2 \qquad m = 0. \tag{5b}$$

A distance can now be defined between the ith test frame and the jth reference frame as [3,19]

$$d(i, j) = d(V_i, P_j) = \log \left[\sum_{m=0}^{p} V_i(m) P_j(m) \right]. \tag{6}$$

The distance measure of eq. (6) has been shown to be an effective measure for comparing sets of LPC coefficients in a variety of applications,[3,19-22] and it can be computed with $(p + 1)$ multiplications and additions and one logarithm.

The next step in the recognizer is to compare the test word against each stored word reference template. A dynamic time-warping algorithm is used to optimally align in time the test and reference patterns and to give the average distance associated with the optimal warping path. The average distance for the qth template of the rth reference word is

$$\bar{D}_{r,q} = \frac{1}{NT} \left[\min_{w_{r,q}(i)} \sum_{i=1}^{NT} d(i, w_{r,q}(i)) \right], \tag{7}$$

where $w_{r,q}(i)$ is the optimally determined warping path. The final step in the process is to choose the recognized word based on the set of average distances $\bar{D}_{r,q}$. The most common decision rule is the minimum distance rule which chooses the word r^* such that

$$\bar{D}_{r^*,q^*} \leq \bar{D}_{r,q} \quad \text{all} \quad r, q \tag{8}$$

for some value q^*. An alternative and more powerful decision rule (for the case of multiple reference templates) is the K-nearest neighbor rule (KNN), which says that for each word r, the distances $\bar{D}_{r,q}$ are reordered according to average distance so that

$$\bar{D}_{r,[1]} \leq \bar{D}_{r,[2]} \leq \cdots \leq \bar{D}_{r,[Q]}, \tag{9}$$

where Q is the number of templates for the rth word, and the KNN rule says to choose the word r^* such that

$$\sum_{k=1}^{K} \bar{D}_{r^*,[k]} \leq \sum_{k=1}^{K} \bar{D}_{r,[k]} \quad \text{all} \quad r. \tag{10}$$

For $K = 1$ the KNN rule is the minimum distance rule. Unless otherwise noted, a value of $K = 2$ was used in the recognition tests in this paper.

2.1 The clustering procedure

The clustering analysis is based on the fully automatic technique (unsupervised with averaging—UWA) described in Ref. 17. It was assumed that we begin with M replications of each word in the vocabulary and, based on the pairwise dynamic time-warped average distance between words, the M tokens are grouped into P disjoint clusters, ω_i, such that

$$\Omega = [t_1, t_2, \cdots, t_M] = \bigcup_{i=1}^{P} \omega_i, \tag{11}$$

where $t_1, t_2, \cdots, t_M$ are the M tokens in the set. The total number of clusters, P, is determined automatically by the clustering procedure. Each cluster, ω_i, is represented by a prototype $\hat{x}_i$. Based on the work of Rabiner and Wilpon,[17] the tokens within cluster ω_i are averaged (using dynamic time warping for time alignment) to give the prototype $\hat{x}_i$. Word reference templates are determined as the prototypes $\hat{x}_i$ corresponding to the $\hat{P}$ largest clusters, i.e., for a single template we choose the prototype of the cluster with the most tokens; for a two-template representation, we choose the prototypes of the two clusters with the largest number of tokens, etc.

The grouping of the M tokens into P clusters is based on splitting of the set Ω by iteratively determining cluster centers (based on a minimax criterion) and cluster points based on a given distance threshold. Ultimately, all M tokens are assigned to one of the clusters. A cluster may consist of a single outlier token whose distance to all other tokens in Ω is greater than the distance threshold of the procedure. The final set of P clusters is ordered based on size of the clusters, and the averaged centers of the $\hat{P}$ largest clusters are retained as the $\hat{P}$ word reference templates.

III. EXPERIMENTAL PROCEDURES

To test the effectiveness of the clustering analysis for a speaker-trained system, three talkers trained the recognizer of Fig. 1. One of the three talkers was the first author of this paper. The other two talkers were experienced workers in the area of speech processing. All three talkers were instructed to speak the words naturally, but in an isolated format. No specific motivation for good performance was employed, as the talkers' interest in the area was considered sufficient. The vocabulary for these tests consisted of the letters A to Z, the digits 0 (zero) to 9, and the command words STOP, ERROR, and REPEAT for a total of 39 words. This vocabulary is an extremely difficult one (especially when recorded over telephone lines, as was done here), but one which has utility in a wide range of practical applications.[23]

Each talker spoke the 39-word vocabulary (in a random order) three times per session over a one-month period for a total of 50 replications for each word in the vocabulary. A total of 17 sessions was used, with only two recordings in the last session.

A clustering analysis was performed for each talker, and a set of word reference templates was obtained. For speaker-independent systems, a total of $\hat{Q} = 12$ reference templates per word was used. For comparison purposes, the same number of templates was obtained for the speaker-trained vocabulary. However, results are also given for a variable number of templates per word.

To test the system, each of the three talkers spoke the 39-word vocabulary five times per session for a total of 10 sessions. Each session was at least two weeks after the preceding one; thus, a total of at least 20 weeks was used to obtain the 10 test sets.

Additional analyses were performed to show the effects of reduced training on the recognition accuracy. To do this, we simply used fewer training runs in the clustering analysis. As such, results are presented for cluster sets based on 24, 12, and 6 replications of the word vocabulary during the training phase.

IV. CLUSTER STATISTICS

Based on the clustering analysis, a set of objective statistics on the clusters can be given which indicates how the tokens cluster. In accordance with past experience with these clustering algorithms, the following statistics appear to be most meaningful:

(*i*) Number of clusters per word. A cluster is defined as a set with at least two tokens.

(*ii*) Number of outliers per word. An outlier is a token that does not fall into one of the clusters above, i.e., its distance to all other tokens in the training set exceeded a threshold.

(*iii*) Quality ratio, σ, defined as the ratio of the average intercluster distance (as defined between cluster prototypes) to the average intracluster distance (as defined between cluster tokens).

(*iv*) Size of largest cluster—i.e., the number of tokens in the largest set.

This set of cluster statistics gives an excellent picture of how the M tokens are distributed in the feature space of the problem being studied.

Table I gives the statistics of the word clusters for the three talkers used in this investigation. Included in the table are averaged, minimum, and maximum values of the cluster statistics for each of the three talkers. The statistics in Table I were obtained from clustering the 50 replications of each word for each talker. It is seen that the average values of all statistics are about the same for all three talkers. Typically, about 6 clusters per word were sufficient to include all nonoutlier tokens. Included in the six clusters were, on the average, 38 of the 50 tokens, with about 20 of the 50 tokens in the biggest cluster. The

Table I—Statistics of the word clusters for the three talkers

	Subjects								
	LRR			AER			SWC		
	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
Number of clusters per word	6	1	10	6	3	10	6	3	10
Number of outliers per word	13	3	17	12	5	17	11	5	18
Quality ratio (σ)	2.90	1.99	4.06	2.92	2.52	3.86	2.68	2.12	3.41
Size of largest cluster	20	9	46	20	7	35	21	9	31

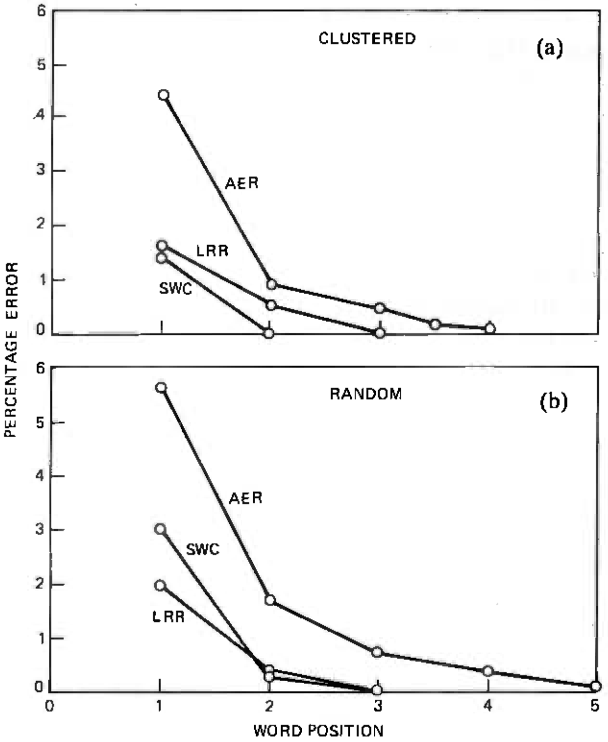


Fig. 2—Recognition error as a function of word position for the three talkers for (a) clustered templates and (b) randomly chosen templates.

quality ratios of between 2.6 and 2.9 indicate good cluster separation for each of the talkers.[16]

V. RECOGNITION RESULTS

Recognition results on the total of 1950 words (50 replications of the 39-word vocabulary) for each of the three talkers (LRR, AER were male, SWC was female) are presented in Figs. 2 and 3. Figure 2a shows a series of plots of the percentage errors as a function of word position for the three talkers for reference templates obtained from the clustering analysis. Word error rate for the *k*th word position is the percentage of words which were not within the top *k* candidates. A total of 12 templates per word was used in these tests. Overall error rates of 1.4

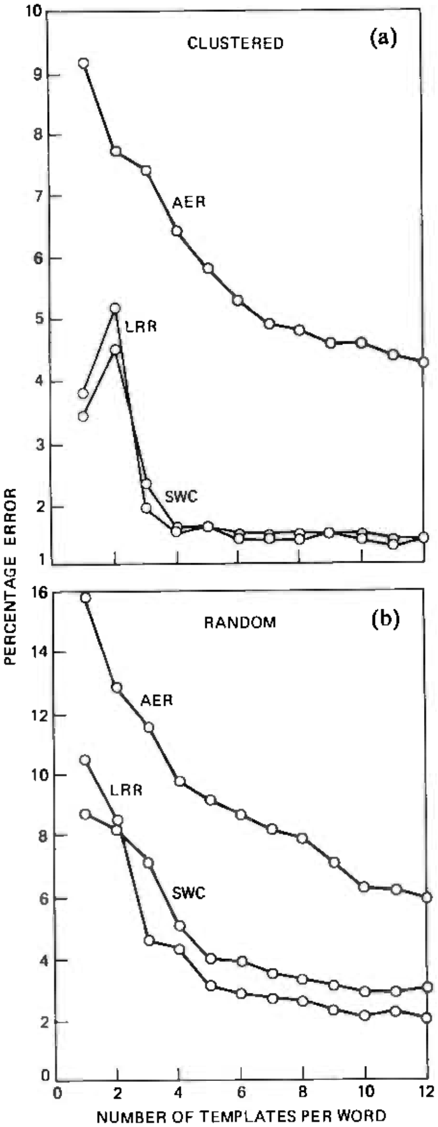


Fig. 3—Recognition error as a function of the number of templates per word (top choice candidate) for the three talkers for (a) clustered templates and (b) randomly chosen templates.

percent (SWC), 1.6 percent (LRR), and 4.4 percent (AER) were obtained for the three-talkers for the top recognition candidate (i.e., the first-choice candidate). The error rate was below 1 percent for the top two candidates (word position 2) for all three talkers.

For comparison purposes, a set of word templates was created by randomly choosing tokens from the 50 replications of each word and creating one reference template directly from each token. Again, a total of 12 templates per word was used. Figure 2b shows the error scores for the random set of word templates for the three talkers. Overall error rates of 3 percent (SWC), 2 percent (LRR), and 5.6 percent (AER) were obtained for the top recognition candidate. Although these error rates were somewhat higher than the scores obtained from the clustered template set, the differences are reasonably small and indicate that the clustering analysis is unnecessary if we are using 12 templates per word. In such a case, a random selection of word templates is essentially equivalent.

It is shown in Fig. 3, however, that the results given in Fig. 2 are not a complete picture of the effectiveness of the clustering analysis. Figure 3a shows plots of percentage error for the top recognition candidate as a function of the number of templates per word for the clustered template set, and Fig. 3b shows similar plots for randomly chosen templates. For talkers LRR and SWC, it is readily seen that the error rate does not change for more than four templates per word for the clustered data. For talker AER, the error rate decreases by about 0.6 percent as the number of templates per word increases from 6 to 12. Thus, Fig. 3a indicates that from 4 to 6 templates per word obtained via clustering give comparable recognition accuracies to 12 templates per word obtained in the same manner.

A totally different picture emerges from Fig. 3b for the case of randomly chosen templates. (Note that the vertical scale of Fig. 3b is different from the vertical scale of Fig. 3a.) It is seen that the percentage error decreases steadily as the number of (random) templates per word increases until about 10 templates per word. Thus for randomly chosen templates a substantially larger number of templates per word are required than for templates obtained from a clustering analysis. An alternative way of stating this is that recognition accuracies from 3 to 4 templates per word (obtained from the clustering analysis) are comparable to those obtained from 10 to 12 randomly chosen templates per word.

5.1 Confusions among the words

The spoken letters of the alphabet form one of the most difficult of word recognition vocabularies because of the high confusability among sets of the letters.[14,23] A major advantage of the clustering analysis is that confusions among many of the subsets are entirely eliminated. The major confusions, for all three talkers, were in the equivalence class of letters containing B, C, D, E, G, P, T, V, and Z. The confusion matrices for this class for the three talkers (for 12 templates per word) are shown in Table II. For talker SWC, 26 of the 27 errors were within this equivalence set; for talker LRR all 31 errors occurred within the equivalence set; for talker AER, 72 of the 85 recognition errors occurred within the equivalence class—however, one confusion was with a word outside the set. (Nine of the remaining errors were A, K confusions.)

Table II shows that each talker had one or more letters in the major equivalence class which were hard to reliably recognize; however, for all three talkers most letters in the hard equivalence class were correctly recognized. This result again demonstrates the power of the clustering analysis in determining the structure of each word in the vocabulary.

5.2 Recognition accuracy vs time

An important aspect of a speaker-trained word recognizer is the stability of the reference templates as a function of time. For casually

Table II—Confusion matrices of the equivalence class with B, C, D, E, G, P, T, V, Z for the three talkers

		Recognized Word										
		B	C	D	E	G	P	T	V	Z	Other	
Spoken Word	B	46		1			1	1			1	LRR
	C		50									
	D	4		35			4	4	3			
	E				50							
	G					50						
	P						44	6				
	T						5	45				
	V	1							49			
	Z									50		
Spoken Word	B	36		3			6	4	1			AER
	C		48							2		
	D			31				17	2			
	E	3			45		2					
	G					50						
	P						39	10			1	
	T			3				46			1	
	V	1	1				1		47			
	Z		12						1	36	1	
Spoken Word	B	47		2	1							SWC
	C		50									
	D	1		48	1							
	E			1	49							
	G					48		2				
	P						49	1				
	T					3		47				
	V								50			
	Z		4						10	36		

trained, nonadaptive systems, the reference templates often degrade with time and the system must be retrained.[1] Since training is so simple for these systems, this generally does not pose a problem. However, some mechanism must be provided for detecting the degradation of the reference templates and retraining the system.

For a clustering analysis method of obtaining reference templates, it is imperative that the templates be robust in time, i.e., that no degradation in recognition accuracy occurs, since the training procedure is a long and involved one. To demonstrate that the reference templates from this system are indeed robust, Fig. 4 shows plots of the error rate vs time for each of the three talkers. It is seen that over the 20-week period of testing, only small changes occur in the recognition accuracy.

5.3 The effects of reduced training

Since the amount of training used to obtain the accuracies reported here was quite extensive (50 repetitions of each word), the effects of reduced training on the recognition scores are important to understand. Thus, the clustering analysis was redone using subsets of the 50 replication training data. The subsets included the first 24, 12, and 6 replications. Before discussing the results, two points should be noted. Each recording session consisted of three consecutive replications of the word list. Thus the three subsets constitute eight, four, and two recording sessions. This is important since it was found that a high degree of correlation existed between tokens within a given recording session.

The second point of note is that, for the 12 replications training set, the maximum number of clusters was limited to six (including outliers), and for the six replication training set the maximum number of clusters was limited to four. The reason for this limitation is that more than six (or four) meaningful clusters cannot be obtained from the reduced

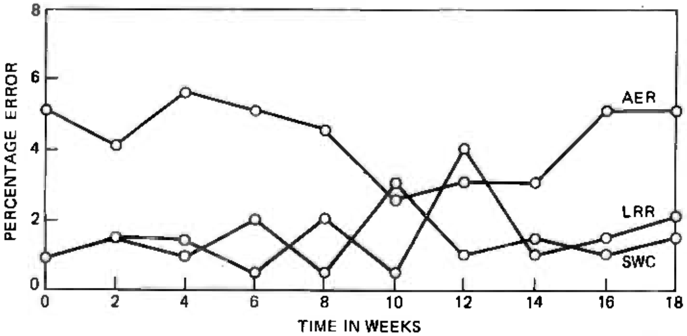


Fig. 4—Recognition error as a function of time for the three talkers.

number of tokens. For recognition purposes, the KNN = 1 rule was used for the 12 and 6 replication template sets.

The recognition results for the reduced training sets are shown in Figs. 5 and 6. Figure 5 shows plots of percentage error as a function of word position for each training set and for each talker. Figure 6 shows plots of percentage error vs the number of templates per word for

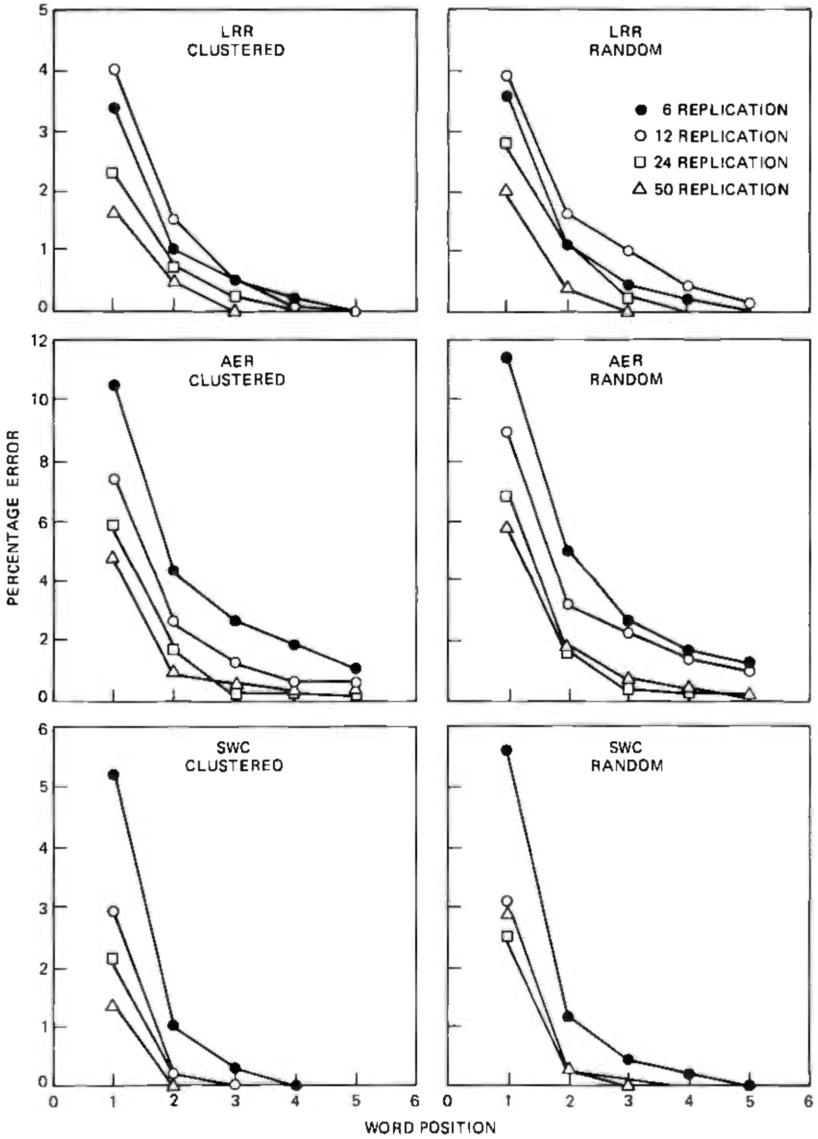


Fig. 5—Recognition error as a function of the word position for different numbers of training sets for the three talkers for both clustered and random templates.

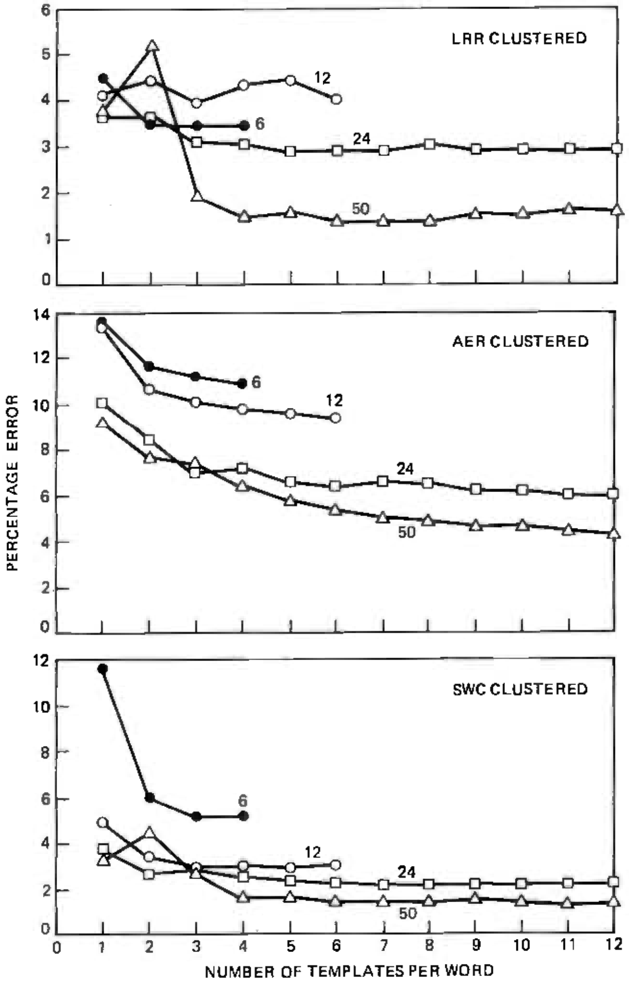


Fig. 6—Recognition error as a function of the number of templates per word and the number of training sets for the three talkers for the clustered templates.

these cases. It is seen that, in all cases, the reduced training set leads to increased error in recognition. In reducing the size of the original training set (from 50 to 24 replications), the error increased about 1.5 percent, on the average, for the three talkers. In going from 50 to 12 replications for training, the error increased by about 2.5 percent for the three talkers, and in going from 50 to 6 replications, the error increased by about 3.3 percent.

The results given above indicate that increased training always gave better templates from the clustering analysis and reduced the recognition error rate.

5.4 Comparisons to casually trained systems

The recognition system of Fig. 1 was casually trained to each of the three talkers by having them speak the vocabulary twice and creating reference templates directly from the spoken words. The recognition tests were then rerun using the casually obtained word templates. The average error rates (over the 50 replications) for the three talkers were 6.5 percent for talker LRR, 12.9 percent for talker AER, and 12.5 percent for talker SWC. Since the overall error rates for the clustered data were 1.6, 4.3, and 1.4 percent for these talkers, respectively, reductions in error rate of 3.9, 8.6, and 11.1 percent were obtained. These error reductions represent a substantial improvement in the recognition.

VI. DISCUSSION

The main result of this paper is the demonstration that statistical clustering techniques can be applied equally well to speaker-trained word recognizers as they have been to speaker-independent ones. It was shown that, with sufficient training and through the use of well-developed clustering algorithms, extremely high recognition scores can be obtained, even with vocabularies as difficult as the letters of the alphabet. This result indicates that, if a user is sufficiently motivated to spend the time necessary to train a word recognizer, he can reliably use the recognizer with a range of vocabularies in a wide variety of applications.[1,23]

An important consideration in a practical implementation of a system like the one described in this paper is to keep the number of reference templates as small as possible. It was shown that about 4 to 6 templates per word were sufficient for the given vocabulary. It is anticipated that for alternative, less complex vocabularies even fewer templates per word would be required. The templates themselves appear to be stable with time as the recognition scores did not change appreciably through the 20 weeks of testing.

One point in question about this work is that only three (experienced) talkers were used. We can only speculate on what the results would be for a larger set of talkers. It is believed that the clustering approach would be highly effective for any talker. (It should be especially good for an inexperienced one who has a lot of replication-to-replication variability in the way he says the words.) As such, the conjecture is that even larger improvements in recognition accuracy over casual training would be obtained by using this system for a wide range of talkers.

Finally, it was shown that the clustering analysis could be bypassed with sufficient training, if a large number of randomly chosen templates (10 to 12) were used to represent each word in the vocabulary. If computational complexity was not an issue, this result could be useful for some applications.

VII. SUMMARY

We have shown that statistical clustering techniques can be applied to a speaker-trained, isolated word recognition system to provide significant improvements in recognition accuracy over casually trained systems. The amount of training required for such a system is fairly extensive. Thus, this method would probably be limited to applications requiring extremely difficult vocabularies (e.g., the letters of the alphabet), or those in which very high recognition accuracies are required.

VIII. ACKNOWLEDGMENTS

The authors acknowledge the diligence and time of Susan Christianson (SWC) and Aaron Rosenberg (AER) of Bell Laboratories for participating as subjects in an experiment that lasted the better part of six months.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *64*, No. 4 (April 1976), pp. 487–501.
2. P. Vicens, "Aspects of Speech Recognition by Computer," Ph.D. Thesis, Stanford Univ., April 1969.
3. F. Itakura, "Minimum Prediction Residual Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-23*, No. 1 (February 1975), pp. 67–72.
4. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System Over Dialed-Up Telephone Lines," J. Acoust. Soc. Am., *60*, Supplement No. 1 (November 1976), p. S12 (Abstract).
5. S. R. Hyde, "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature," in *Human Communication: A Unified View*, E. E. David, Jr., and P. B. Denes, eds., New York: McGraw-Hill, 1972, pp. 399–438.
6. G. M. White and R. B. Neely, "Speech Recognition Experiments With Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-24*, No. 2 (April 1976), pp. 183–188.
7. M. B. Herscher and R. B. Cox, "An Adaptive Isolated-Word Speech Recognition System," Conf. Rec. 1972 Conf. Speech Comm. and Proc., *AD-742236*, (April 1972), pp. 89–92.
8. B. Gold, "Word Recognition Computer Program," MIT Res. Lab of Electronics, Tech. Report 452, June 1966.
9. J. M. Shearme and P. F. Leach, "Some Experiments With a Simple Word Recognition System," IEEE Trans. Audio and Electroacoustics, *AU-16*, No. 2 (June 1968), pp. 256–261.
10. V. M. Velichiko and N. G. Zagoruiko, "Automatic Recognition of 200 Words," Int. J. Man-Machine Studies, *2* (1970), p. 23.
11. C. F. Teacher, H. G. Kellet, and L. R. Rocht, "Experimental, Limited Vocabulary Speech Recognizer," IEEE Trans. Audio and Electroacoustics, *AU-15*, No. 3 (September 1967), pp. 127–130.
12. A. Ichikawa, Y. Nakamo, and K. Nakata, "Evaluation of Various Parameter Sets in Spoken Digits Recognition," IEEE Trans. Audio and Electroacoustics, *AU-21*, No. 3 (June 1973), pp. 202–209.
13. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-26*, No. 1 (February 1978), pp. 34–42.
14. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," IEEE

Trans. Acoustics, Speech, and Signal Proc., *ASSP-27*, No. 4 (August 1979), pp. 336–349.

15. M. R. Sambur and L. R. Rabiner, "A Statistical Decision Approach to the Recognition of Connected Digits," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-24*, No. 6 (December 1976), pp. 550–558.
16. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-27*, No. 2 (April 1979), pp. 134–141.
17. L. R. Rabiner and J. G. Wilpon, "Considerations Applying Clustering Techniques to Speaker Independent Word Recognition," J. Acoust. Soc. Amer., *66* (September 1979), pp. 663–673.
18. L. R. Rabiner and J. G. Wilpon, "Speaker Independent, Isolated Word Recognition for a Moderate Size (54 Word) Vocabulary," IEEE Trans. Acoustics, Speech, and Signal Proc., 1979.
19. J. M. Tribolet, L. R. Rabiner, and M. M. Sondhi, "Statistical Properties of an LPC Distance Measure," IEEE Trans. Acoustics, Speech, and Signal Proc., 1979.
20. J. Makhoul, R. Viswanathan, L. Cosell, and W. Russell, "Natural Communication with Computers," BBN Rep. No. 2976, December 1974.
21. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-24*, No. 6 (December 1976), pp. 488–494.
22. D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner, and J. Goodman, "Objective and Subjective Performance of Tandem Connections of Waveform Coders with a LPC Vocoder," B.S.T.J., *58*, No. 5, (March 1979), pp. 601–629.
23. A. E. Rosenberg and C. E. Schmidt, "Recognition of Spoken Spelled Names Applied to Directory Assistance," J. Acoust. Soc. Amer., *62*, Supplement No. 1 (December 1977), p. 563 (Abstract).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

A Family of Active Switched Capacitor Biquad Building Blocks

By P. E. FLEISCHER and K. R. LAKER

(Manuscript received May 10, 1979)

Two closely related, low-sensitivity, active switched capacitor filter topologies are presented. Each of these circuits comprises two operational amplifiers and at most nine capacitors. The topologies have been carefully constructed so that they are immune to the various parasitic capacitances normally present in switched capacitor networks. One filter topology is capable of realizing any stable biquadratic z-domain transfer function, while the second one is only slightly less than fully general. Most commonly used transfer functions can be realized with either topology and will require only seven capacitors. The choice between the two topologies will generally be made on the basis of total capacitance required, dynamic range behavior, and sensitivity. A complete set of synthesis equations is given for both circuits which cover both the general and all the important special cases of the biquadratic transfer function. Finally, several examples are given which illustrate the synthesis procedures and the versatility of the filter topologies.

I. INTRODUCTION

Active RC building blocks which realize a biquadratic transfer function[1,2] have played an increasingly important role in supplying filters for the Bell System. Thus, at the present time, STAR building blocks enjoy a very large volume of production and there is a constant effort directed toward the reduction of their costs.[3,4]

Perhaps the most promising recent development in filtering is the emergence of active switched capacitor filters.[5–8] These filters are fully MOS realizable and therefore enjoy the cost advantages of LSI integrated circuit realizations. Further, they can be expected to share in the future cost reductions due to VLSI.

Since biquadratic building blocks have played such a dominant role in the active RC field, they are also expected to be of great importance for active switched capacitor (SC) filter realizations. Many of these basic biquadratic blocks can then be realized on a single chip to implement higher order filter functions.

The purpose of this paper is to describe a pair of similar active SC biquadratic filter topologies[9] capable of realizing any stable biquadratic z-domain transfer function. In arriving at these networks, we have been particularly careful to eliminate the effects of parasitic capacitances. Such capacitances arise mainly from two sources.

First is a sizable (approximately 10 percent) parasitic capacitance from the bottom plates of the capacitors to the epitaxial layer (ac ground). By using two operational amplifiers, it is possible to arrange matters so that the parasitic capacitors are connected either to a voltage source (operational amplifier output) or ground, thereby eliminating any effects due to them. Second are parasitic capacitances from the MOS switches to ground via the power supplies. By using only certain switching arrangements,[10] the parasitics again can be forced to appear at harmless locations.

After a description of the general biquad circuits, a full set of synthesis relations will be developed for them. The paper concludes with several examples that demonstrate the flexibility and general usefulness of these circuits.

II. GENERAL CIRCUIT TOPOLOGY

The transfer function of an active-RC biquad is biquadratic in the Laplace transform variable s:

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{cs^2 + es + d}{s^2 + as + b}. \tag{1}$$

In contrast, the active-SC biquad voltage transfer function is biquadratic in the z-transform[11] variable z, where $z = e^{sT_s}$ and T_s is the sampling interval, viz.,

$$\frac{V_{\text{out}}}{V_{\text{in}}}(z) = \frac{N(z)}{D(z)} = \frac{\gamma + \epsilon z^{-1} + \delta z^{-2}}{1 + \alpha z^{-1} + \beta z^{-2}}. \tag{2}$$

This is the general biquadratic transfer function we seek to realize. Of course, by suitable choice of the numerator coefficients, the different generic transfer functions, such as LP, BP, and HP, can be obtained.

All the filter topologies to be considered in this paper are special cases of the general active-SC biquad shown in Fig. 1a. This circuit bears a close resemblance to the three amplifier biquad;[2] however, because of the inversion[7] inherent in capacitor A, the inverter of the active-RC biquad is not needed. In effect, the circuit consists of two integrators, the first stage being inverting while the second stage is noninverting. Damping is provided by the capacitor E and the switched

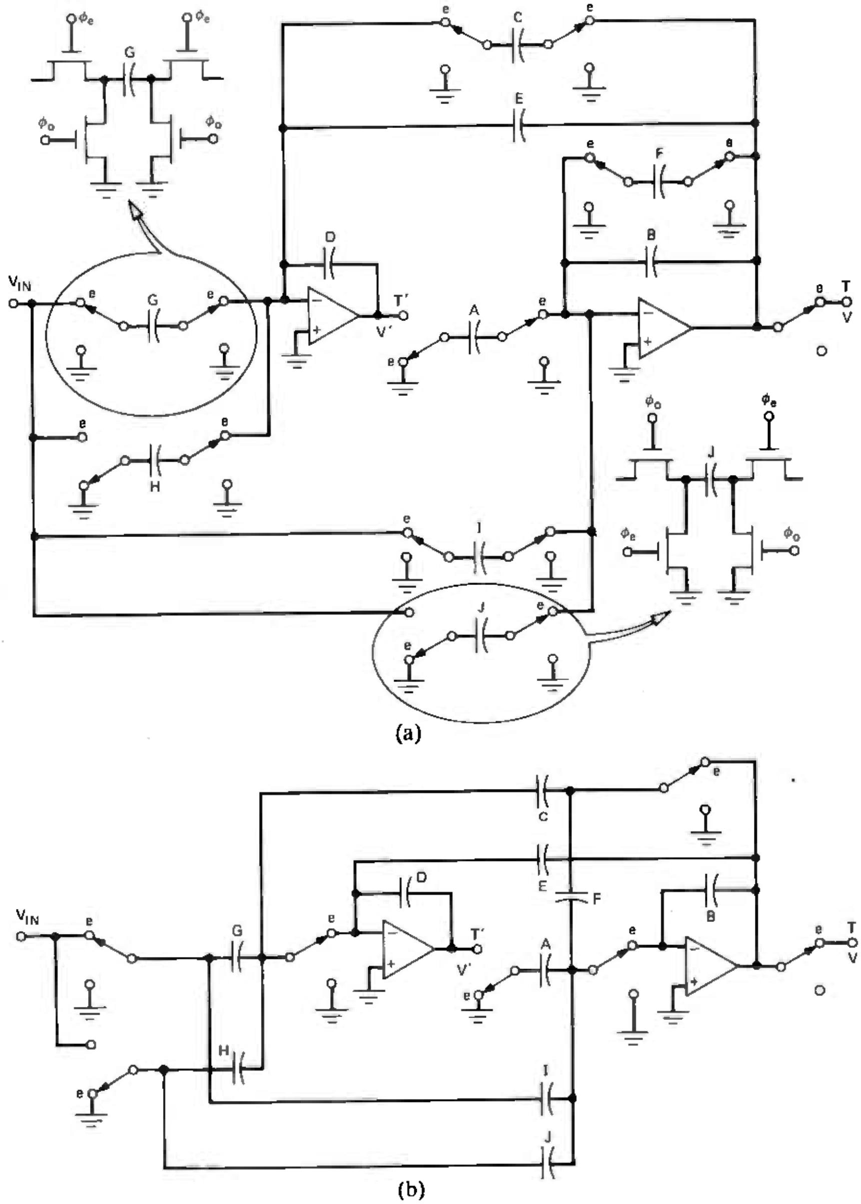
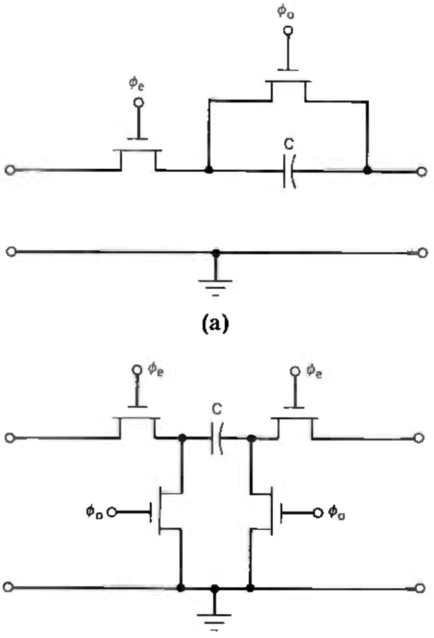


Fig. 1—(a) General active sc topology. (b) General active sc topology (minimum switch configuration).

capacitor *F*. In any particular application, only one of these will be present, leaving a total of nine capacitors, but for analysis purposes it is convenient to handle the two cases together.

The transmission zeros are realized via the multiple feed-forward paths consisting of switched capacitors G, H, I, and J. It is seen later that, typically, no more than two of these capacitors are needed to realize the useful biquadratic transfer functions. Thus, most often it will be found that only seven capacitors are needed. Although the circuit schematic shown in Fig. la facilitates understanding the circuit, a more efficient implementation can be obtained by allowing similarly switched capacitors to share a common switch. Rearranging the circuit schematic in this way results in the minimum switch configuration shown in Fig. 1b. One can readily verify that the electrical behaviors of these circuits are identical.

To minimize[8,10] the deleterious effects of switch parasitics, we have avoided the use of toggle-switched capacitors[7] in which one end of the capacitor is permanently connected to ground. Furthermore, it is noted that the sc elements shown in Figs. 2a and 2b are equivalent in function. However, by discharging the capacitor C through ground, as per Fig. 2b (and the insert in Fig. 1), the parasitic switch capacitances are also discharged. For ideal operational amplifiers, the parasitic switch capacitances have absolutely no effect on the operation of the biquad. It is expected that the effect will only be negligibly enhanced[12] by the nonideal character of practical operational amplifiers.



(a)

(b)

Fig. 2—Discharge type switched capacitors. (a) Typical realization. (b) Parasitic free realization.

The close analogy of SC filters to active-RC biquads has already been mentioned. In particular, note that the circuit with $E = 0$ is very similar in spirit to the three-operational amplifier multiple input biquad[13] except for the absence of the inverter, which is not needed for the SC filter. However, active-SC biquads offer even further versatility which to this point has not been exploited. Because of the inability to trim capacitors and the relative cost factors, practical active-RC biquads are constructed to be canonic in capacitors, namely, two. This constraint is unnecessarily placed on active-SC topologies when they are derived from an active-RC topology via a resistor-to-switched-capacitor replacement.[5-7,14] As we show, one can achieve interesting and beneficial results when this constraint is removed.

2.1 The transfer functions

Before deriving the transfer functions which characterize the general active-SC network of Fig. 1, it is necessary to consider the timing of the switches. Note that the schematics for some typical SC elements are shown as inserts in Fig. 1.

For simplicity in analysis, we assume the clocks ϕ_e and ϕ_o to be nonoverlapping with 50-percent duty cycle, as in Fig. 3. It is noted that switches clocked by ϕ_e close instantaneously at the $2k\tau$ (even) time instants and those clocked by ϕ_o close instantaneously at the $(2k + 1)\tau$ (odd) time instants. We also assume that the input signal is sampled and held over a full clock period, 2τ, as shown in Fig. 3. In fact, the switch phasings of the SC biquad have been chosen to operate with this kind of input. Under these conditions, we have

$$V^o_{\text{in}}(z) = z^{-1/2}\, V^e_{\text{in}}(z). \tag{3}$$

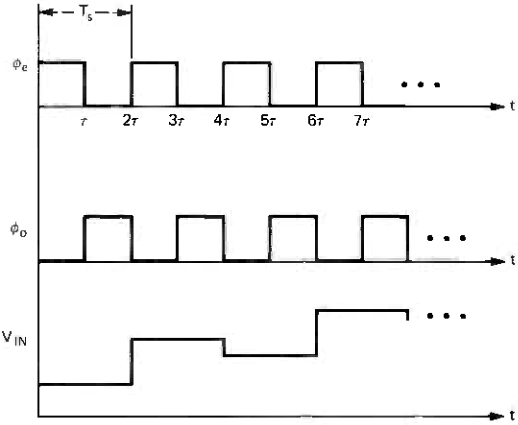


Fig. 3—Clock and input waveforms.

It follows then[16] that the output voltages appearing at the operational amplifier outputs are also held for the full clock period and change only at the even sampling instants:

$$V^{o}_{\text{out}}(z) = z^{-1/2} V^{e}_{\text{out}}(z). \tag{4}$$

This fact can also be surmised by inspection of Fig. 1a. Appendix A shows that the full-cycle S/H assumption can be relaxed. For the present, however, since both the input and the outputs are fully held, the even transfer functions provide all the information we need. For simplicity, therefore, the "even" superscript will be omitted from here on.

Let us now derive the voltage transfer functions T' and T for the networks of Fig. 1. As noted previously, these transfer functions are most conveniently expressed in the z-domain where $z = e^{sT_s}$ and $T_s = 2\tau$ represents a full clock period. Any SC network containing biphased switches can be conveniently transformed into a z-domain equivalent circuit.[15,16] With this equivalent circuit, one can then apply all the network tools available to continuous, linear, time-invariant networks. When the input signal is held over the full clock period, one can readily obtain[16] the equivalent circuit given in Fig. 4.

The desired transfer functions are then derived using straightforward nodal analysis:

$$T \triangleq \frac{V}{V_{\text{in}}}$$

$$= \frac{-Az^{-1}(G - Hz^{-1}) - D(1 - z^{-1})(I - Jz^{-1})}{Az^{-1}(C + E - Ez^{-1}) + D(1 - z^{-1})(F + B - Bz^{-1})} \tag{5a}$$

$$= -\frac{DI + (AG - DI - DJ)z^{-1} + (DJ - AH)z^{-2}}{D(F + B) + (AC + AE - DF - 2DB)z^{-1} + (DB - AE)z^{-2}}, \tag{5b}$$

and

$$T' \triangleq \frac{V'}{V_{\text{in}}}$$

$$= \frac{(I - Jz^{-1})(C + E - Ez^{-1}) - (G - Hz^{-1})(F + B - Bz^{-1})}{Az^{-1}(C + E - Ez^{-1}) + D(1 - z^{-1})(F + B - Bz^{-1})} \tag{6a}$$

$$= \frac{\begin{gathered}(IC + IE - GF - GB)\\ + (FH + BH + BG - JC - JE - IE)z^{-1} + (EJ - BH)z^{-2}\end{gathered}}{D(F + B) + (AC + AE - DF - 2DB)z^{-1} + (DB - AE)z^{-2}}. \tag{6b}$$

Before undertaking further analysis, some extraneous degrees of freedom will be eliminated. First, we arbitrarily set $A = B$. It may be shown that the net effect of this choice is to remove our ability to

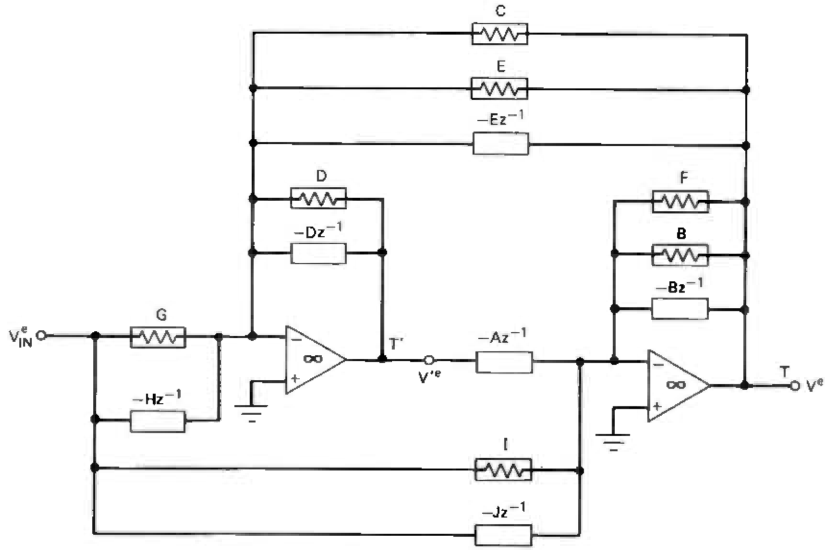


Fig. 4—z-domain equivalent circuit for the biquad in Fig. 1.

control the gain constants associated with T and T' simultaneously. Later we will show how, through scaling techniques, this degree of freedom can be restored to the circuit. Second, it is clear by examining the circuit of Fig. 1 that there is an arbitrary impedance scaling associated with each of the two stages. Thus, the two groups of capacitors (C, D, E, G, H) and (A, B, F, I, J) may each be arbitrarily and independently scaled without changing the transfer functions. Accordingly, we arbitrarily choose $B = 1$ and $D = 1$. This choice will also be ultimately overridden by minimum capacitor realization considerations.

In view of the above, we have

$$A = B = D = 1. \tag{7}$$

Substituting these into (5) and (6) yields the following simplified transfer functions:

$$T = -\frac{I + (G - I - J)z^{-1} + (J - H)z^{-2}}{(F + 1) + (C + E - F - 2)z^{-1} + (1 - E)z^{-2}} \tag{8}$$

and

$$T' = \frac{\begin{array}{c}(IC + IE - FG - G)\\ + (FH + H + G - JC - JE - IE)z^{-1} + (EJ - H)z^{-2}\end{array}}{(F + 1) + (C + E - F - 2)z^{-1} + (1 - E)z^{-2}}. \tag{9}$$

Let us first examine the salient features of the transfer function T. Note that its poles are determined by C, E, and F, and its zeros by G,

H, I, and J. The fact that the poles and zeros are independently adjustable may prove to be useful in some filter or equalizer applications. It is also clear that the three numerator coefficients are independently adjustable thus permitting arbitrary zeros to be realized. The fact that any stable poles are also realizable will be demonstrated later.

Regarding the transfer function T', we first observe the obvious fact that its poles are identical to those of T. We note, however, that its zeros are formed in a more complicated fashion and they do not have the aforementioned independence property. Nevertheless, there are cases where T' provides a more economical realization of a given transfer function than T.

The zero-forming pairs (I, J) and (G, H) have some interesting alternate realizations. Using the techniques of Ref. 16, it may be shown that the pair I and J, for example, may be realized as shown in Fig. 5. Thus, one capacitor is eliminated not only when $I = 0$ or $J = 0$ but also when $I = J$. The equivalence of a pair of switched capacitors to one unswitched capacitor is quite fascinating in view of the switched-capacitor—resistance equivalence so commonly assumed in dealing with SC networks.

When $I = J$, not only is one capacitor eliminated, but sensitivity is usually improved since I and J now "track" perfectly. Even when $I \neq J$, the transformations may be useful to reduce sensitivity or to lower the total capacitance needed for a given case. For example, if $I = 13$ pf and $J = 12$ pf, transformation (a) can be applied, yielding new element values $I - J = 1$ pf and $J = 12$ pf. The transformation is obviously reversible; thus, the converse transformation can be applied to element pairs C,E and B,F. It should be noted that in arriving at these transformations we have assumed, as in Fig. 1, that terminals 1 and 2 are connected to a full clock period S/H voltage source and to a virtual ground, respectively. In Appendix A, we show that the full cycle S/H assumption can be conditionally relaxed.

Note that the SC elements in Fig. 5 all possess a z-domain admittance of the form $I - Jz^{-1}$. If the inverting switched capacitor J were replaced with a noninverting switched capacitor,[16] a z-domain admittance of the form $I + Jz^{-1}$ would be obtained. At times, it might be convenient from a synthesis point of view to have $I + Jz^{-1}$ admittances; however, as mentioned earlier, noninverting toggle-switched capacitors do introduce parasitics;[8,10] therefore, we avoid using this element. In spite of this omission, the circuit is capable of realizing all stable biquadratic transfer functions.

2.2 The E- and F-circuits

One final simplification we can make to the general biquad in Fig. 1 involves the elements E and F. As already mentioned, E and F are

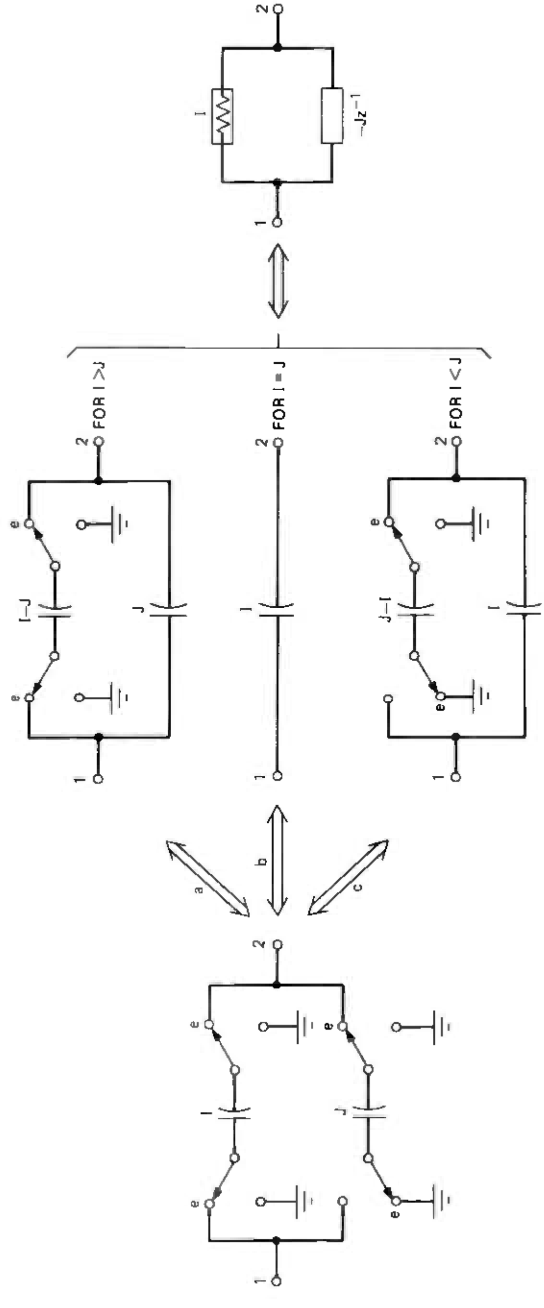


Fig. 5—sc element transformations (ports 1 and 2 are assumed to be buffered).

redundant elements in the sense that they both provide damping. Consequently, they need not both be present in the same circuit. It is, therefore, convenient to define an "E-circuit" in which $E \neq 0$ and $F = 0$ and an "F-circuit" in which $F \neq 0$ and $E = 0$.

The transfer functions for these two circuits are

$$T'_E = \frac{I + (G - I - J)z^{-1} + (J - H)z^{-2}}{1 + (C + E - 2)z^{-1} + (1 - E)z^{-2}} \tag{10a}$$

$$T'_E = \frac{(IC + IE - G) + (H + G - JC - JE - IE)z^{-1} + (EJ - H)z^{-2}}{1 + (C + E - 2)z^{-1} + (1 - E)z^{-2}} \tag{10b}$$

and

$$T_F = -\frac{\hat{I} + (\hat{G} - \hat{I} - \hat{J})z^{-1} + (\hat{J} - \hat{H})z^{-2}}{(\hat{F} + 1) + (\hat{C} - \hat{F} - 2)z^{-1} + z^{-2}} \tag{11a}$$

$$T'_F = -\frac{(\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C}) + (\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G})z^{-1} + \hat{H}z^{-2}}{(\hat{F} + 1) + (\hat{C} - \hat{F} - 2)z^{-1} + z^{-2}}. \tag{11b}$$

The "hats" are placed on the F-circuit elements in order to distinguish them from the E-circuit elements.

Let us briefly examine these transfer functions. Note that the numerators of T_E and T_F are identical, while the numerators of T'_E and T'_F are quite different. Thus for a given design in which the desired output is V, the T' as well as the unscaled dynamic range of V' may be quite different for the two networks. The analogous situation is obtained if the desired output is V'. These differences will ultimately affect the final scaled capacitor values and the total capacitance required to realize the circuit. Significant sensitivity differences between the four possible realizations of a given transfer function may also exist. These points will be illustrated via examples in Section IV.

2.4 Sensitivities

The sensitivities for the E- and F-circuits are at least comparable to any active-RC biquad. One can arrive at this conclusion by examining the Q and ω_o relations and associated sensitivities for a moderate-to-high-Q resonant response.

For any pair of complex conjugate poles in the z-domain, one can write the denominator as:

$$D(z) = 1 + \alpha z^{-1} + \beta z^{-2} \tag{12a}$$

$$= (1 - re^{j\theta}z^{-1})(1 - re^{-j\theta}z^{-1})$$

$$= 1 - 2r\cos\theta z^{-1} + r^2z^{-2}. \tag{12b}$$

By analogy to the continuous case, the following equations involving the resonant frequency ω_o and Q can be written:

$$\theta \approx 2\pi \frac{\omega_o}{\omega_s} = \omega_o T_s \tag{13}$$

and (14)

$$\frac{1}{2Q} \approx \frac{1-r}{\theta} = \frac{1-r}{\omega_o T_s},$$

which implies

$$r \approx 1 - \frac{\omega_o T_s}{2Q}. \tag{15}$$

Therefore, it follows from (12) through (15) that

$$\alpha \approx -2\left(1 - \frac{\omega_o T_s}{2Q}\right) \cos \omega_o T_s \tag{16a}$$

$$\beta \approx \left(1 - \frac{\omega_o T_s}{2Q}\right)^2. \tag{16b}$$

Whenever $\omega_o T_s \ll 1$, i.e., the sampling rate is high and $Q \gg 1$, the above expressions may be further approximated:

$$\begin{aligned}\alpha &\approx -2\left(1 - \frac{\omega_o T_s}{2Q}\right)\left(1 - \frac{\omega_o^2 T_s^2}{2}\right) \\ &\approx -2 + \frac{\omega_o T_s}{Q} + \omega_o^2 T_s^2 \end{aligned} \tag{17a}$$

and

$$\beta \approx 1 - \frac{\omega_o T_s}{Q}. \tag{17b}$$

Consider first the E-circuit. After suitable manipulations, the denominator of (5) becomes (with $F = 0$)

$$D_E(z) = 1 + \left(-2 + \frac{AC}{DB} + \frac{AE}{DB}\right)z^{-1} + \left(1 - \frac{AE}{DB}\right)z^{-2}. \tag{18}$$

Comparing this with (12a) and (17) immediately yields

$$\frac{\omega_o T_s}{Q} \approx \frac{AE}{DB} \tag{19a}$$

and

$$\omega_o^2 T_s^2 \approx \frac{AC}{DB}. \tag{19b}$$

Therefore,

$$\omega_o T_s \approx \left(\frac{AC}{DB}\right)^{1/2} \tag{20a}$$

and

$$Q \approx \frac{1}{E}\left(\frac{DBC}{A}\right)^{1/2}. \tag{20b}$$

Similarly, it may be shown that, for the F-circuit,

$$\omega_o T_s \approx \left(\frac{\hat{A}\hat{C}}{\hat{D}\hat{B} + \hat{D}\hat{F}}\right)^{1/2} \tag{21a}$$

and

$$Q \approx \left[\frac{\hat{A}\hat{C}}{\hat{D}\hat{F}}\left(1 + \frac{\hat{B}}{\hat{F}}\right)\right]^{1/2}. \tag{21b}$$

From (20) and (21), it is seen that ω_o and Q are controlled by the ratios of four or five capacitors. Furthermore, it is clear that

$$|S_x^{\omega_o}| \leq ½ \quad \text{and} \quad |S_x^Q| \leq 1, \tag{22}$$

where x denotes any capacitor in the E- or F-circuits. This situation compares favorably to the active-RC case,[1,2,17] where a minimum of four passive elements, namely, two RC products, determines ω_o. Since, in practice, ratios of capacitors can be more tightly controlled than individual resistors and capacitors, the active-SC realization can be expected to be superior to the active-RC case with respect to initial (untuned) response as well as temperature and aging variations.*

Even though, in practice, the ω_o variation is usually the most significant contributor to the overall variation in the response,[17] we note that the Q sensitivity of the active-SC circuit is also low.

The overall circuit sensitivity is also affected, of course, by the contribution of the numerator. This aspect of the problem is not amenable to a general analysis and has to be handled on a case-by-case basis. It has been our experience that T designs provide lower sensitivity realizations than T' designs. It can be seen from (5) and (6) that the numerator of T is simpler than the numerator of T'. Although one might expect the pole-zero tracking afforded by the T' designs to yield lower sensitivities for some applications, we find that, because of the cancellations which occur in the coefficients, larger sensitivities are often incurred.

* Since T_s is normally derived from a very stable clock, it is assumed to be invariant.

III. SYNTHESIS

The synthesis of the biquad begins with the identification of the desired transfer function. This determination ultimately depends upon the frequency domain specifications which can then be transformed in some manner to a z-domain transfer function. The individual biquad transfer functions are then obtained by factoring this higher order z-domain transfer function.

Once the desired biquadratic transfer function is identified, the unscaled capacitor values are determined from (10) or (11) for the E- or F-circuit, respectively. Once this basic design is obtained, the final step consists of scaling the capacitors to adjust the dynamic range at the output of the other operational amplifier. It is then convenient to rescale the admittances in each of the two stages to obtain a minimum capacitance value of 1 unit in each stage. The actual minimum value of capacitance which can be realized depends on the technology, the desired precision of the transfer function, and the estimated effects of parasitics. A minimum capacitance of 1 pf will be used in this paper.

3.1 z-domain biquadratic transfer functions

Like digital filters, switched-capacitor filters are most conveniently synthesized from a z-domain transfer function. Several methods are available[11] for obtaining a z-domain transfer function from frequency domain specifications. Perhaps the most useful of these is based on the bilinear transformation[11] which can be shown to preserve "maximally flat" or "equal ripple" properties. This method starts with a suitably prewarped analog transfer function in the s-domain. This analog transfer function is then converted to a z-domain transfer function using the bilinear transformation,

$$s = \frac{2}{T_s}\frac{1 - z^{-1}}{1 + z^{-1}}, \tag{23}$$

where, it is recalled, T_s denotes the full clock period. The application of the bilinear transform to an analog biquadratic transfer function yields a z-domain transfer function which is also biquadratic. Of consequence to the ultimate form of low-pass and bandpass transfer functions is the fact that analog zeros at $s = \infty$ map into finite z-domain zeros at $z = -1$. Other transformations,[8] such as

$$s = \frac{1}{T_s}\frac{1 - z^{-1}}{z^{-1}}, \tag{24}$$

do not have this mapping property. Using (24), the mapping of the poles is also somewhat different than that obtained via the bilinear transformation.

3.2 Pole placement

At this point, it is appropriate to consider the stability and realizability of the proposed circuits. It is, of course, desirable to be able to realize all stable pole positions. Stability for a biquad can be conveniently expressed[18] in the α, β parameter space as the area within the shaded triangle shown in Fig. 6. The upper parabolic area of the triangle represents the α, β values for stable, complex poles. The remainder of the upper triangular area, where $\beta > 0$, corresponds to real pole pairs which lie pairwise to the left or right of $z = 0$, while the lower triangular area, where $\beta < 0$, corresponds to real poles which lie on alternate sides of $z = 0$. Clearly, the upper portion of the triangle, i.e., $\beta > 0$, represents most of the useful pole locations for frequency selective filters.

Consider first the E-circuit realizability properties. Comparing (2) and (10) yields

$$\alpha = E + C - 2 \tag{25a}$$

$$\beta = 1 - E \tag{25b}$$

and, therefore,

$$\alpha + \beta = C - 1. \tag{25c}$$

Since $E \geq 0$ and $C \geq 0$, we immediately have from (25a) and (25c):

$$\alpha \geq -2 \tag{26a}$$

$$\beta \leq 1 \tag{26b}$$

$$\alpha + \beta \geq -1. \tag{26c}$$

Thus, the α, β values realizable with the E-circuit are confined within the wedge-like area shown in Fig. 7a. This area includes all the stable

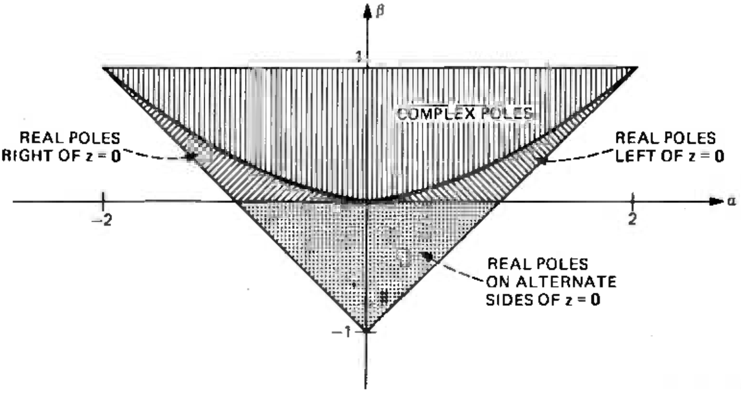

β
1
COMPLEX POLES
REAL POLES RIGHT OF z = 0
REAL POLES LEFT OF z = 0
−2
2
α
REAL POLES ON ALTERNATE SIDES OF z = 0
−1

Fig. 6—Triangle of stable pole positions for $D(z) = 1 + \alpha z^{-1} + \beta z^{-2}$.

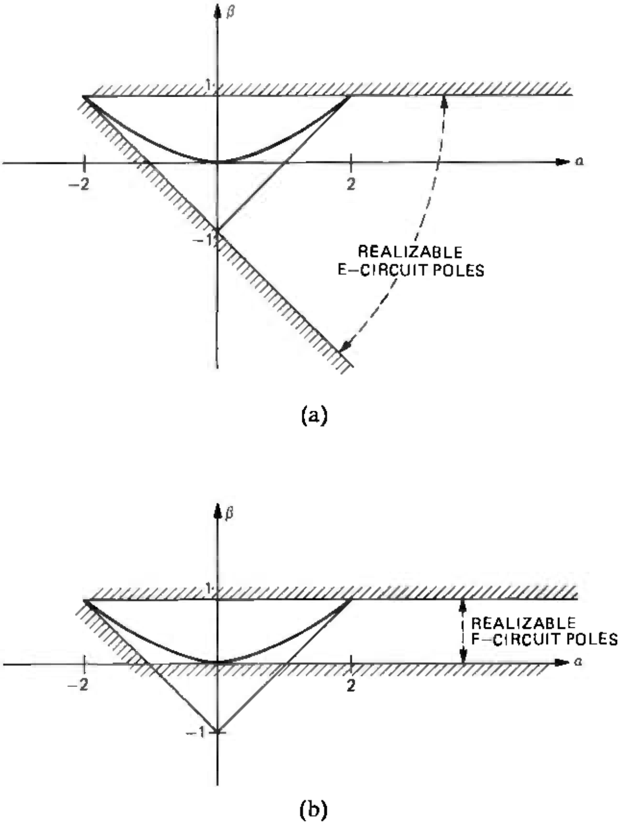


Fig. 7—Pole placement realizability conditions for (a) *E*-circuit and (b) *F*-circuit.

region as well as a portion of the remaining unstable area. *E*-circuits which are unstable must possess real poles.

The *F*-circuit realizability conditions can be derived similarly from (11):

$$0 \leq \beta \leq 1 \tag{27a}$$

and

$$\alpha + \beta \geq -1. \tag{27b}$$

These equations define the semiwedge-like area shown in Fig. 7b. Although this area is more restricted than that representing realizable *E*-circuit poles, the *F*-circuit is seen to be able to realize all the useful stable pole positions. The only stable case not realizable with the *F*-circuit has two real poles of opposite signs. Also, unstable *F*-circuits must have real poles of the same sign.

Now that we have established the realizability conditions for the *E*-

and F-circuits, let us state the pole placement synthesis equations in terms of the z-domain transfer function coefficients α and β. For the E-circuit, the synthesis equations can be stated as

$$E = 1 - \beta \tag{28a}$$

and

$$C = 1 + \beta + \alpha. \tag{28b}$$

Similarly, for the F-circuit we have

$$\hat{F} = \frac{1 - \beta}{\beta} \tag{29a}$$

and

$$\hat{C} = \frac{1 + \alpha + \beta}{\beta}. \tag{29b}$$

Equations (28) and (29) yield nonnegative values for E, C and $\hat{F}$, $\hat{C}$ within the realizability areas sketched in Figs. 7a and 7b, respectively. Also note that $\hat{F}$ and $\hat{C}$ for the F-circuit can be obtained from values calculated for E and C for the E-circuit using the simple relations

$$\hat{F} = \frac{E}{1 - E} \tag{30a}$$

and

$$\hat{C} = \frac{C}{1 - E}. \tag{30b}$$

As noted previously, it is often convenient to derive the z-domain transfer function from the prewarped analog requirements via the bilinear transformation given by (23). Once an appropriate prewarped s-domain transfer function has been determined, one can then derive simple synthesis relations for the pole-determining capacitors in terms of its coefficients. Let the denominator of the prewarped s-domain transfer function be

$$D(s) = s^2 + as + b. \tag{31}$$

Then, substituting for s the bilinear transformation of (23) and equating the coefficients of the resulting z-domain quadratic polynomial with those of the denominator of T_E (or T'_E) yields

$$E = \frac{aT_s}{1 + \frac{aT_s}{2} + b\frac{T_s^2}{4}} \tag{32a}$$

and

$$C = \frac{bT_s^2}{1 + \frac{aT_s}{2} + b\frac{T_s^2}{4}} \cdot \tag{32b}$$

Similar expressions for the F-circuit may be written

$$\hat{F} = \frac{aT_s}{1 - \frac{aT_s}{2} + \frac{bT_s^2}{4}} \tag{33a}$$

and

$$\hat{C} = \frac{bT_s^2}{1 - \frac{aT_s}{2} + \frac{bT_s^2}{4}} \cdot \tag{33b}$$

In summary, when the desired z-domain biquadratic transfer function is known, capacitor values E, C or $\hat{F}$, $\hat{C}$ are readily evaluated according to (28) or (29), respectively. Alternatively, when an appropriately prewarped transfer function is known, the capacitor values E, C or $\hat{F}$, $\hat{C}$ can be evaluated in terms of the coefficients of the prewarped analog function and the sampling period T_s according to (32) or (33), respectively.

3.3 Zero placement

Before deriving the synthesis relations for the zeros, it is instructive to list the z-domain transfer functions for the well-known generic forms; namely, low-pass (LP), high-pass (HP), bandpass (BP), low-pass notch (LPN), high-pass notch (HPN), and all-pass (AP).[19] The numerators, with reference to (2), for these generic forms are listed in Table I. The LP and BP functions are particularly interesting in that there are several different forms which can be used. These forms are referred to in Table I as LP_{ij} and BP_{ij}, where i denotes the number of $1 + z^{-1}$ factors and j the number of z^{-1} factors. As already noted, the zeros at $z = -1$ arise only when the bilinear transform is used. These transfer functions, of course, exhibit steeper cutoff in the vicinity of half the sampling rate, but they may not afford the most economical realization. As a rule of thumb, the additional cutoff will become less and less important as the sampling frequency increases with respect to the pole-zero locations, i.e., as $\omega_o T_s \rightarrow$ small.

The number of z^{-1}, i.e., delay, terms will usually be immaterial. In these cases, economy of realization or perhaps sensitivity considerations might indicate the best choice. Note, however, that if there is additional feedback around any biquad block, the delay term becomes critical.

Table I—Generic biquadratic transfer functions

Generic Form	Numerator $N(z)$
LP 20 (bilinear transform)	$K(1 + z^{-1})^2$
LP 11	$Kz^{-1}(1 + z^{-1})$
LP 10	$K(1 + z^{-1})$
LP 02	Kz^{-2}
LP 01	Kz^{-1}
LP 00	K
BP 10 (bilinear transform)	$K(1 - z^{-1})(1 + z^{-1})$
BP 01	$Kz^{-1}(1 - z^{-1})$
BP 00	$K(1 - z^{-1})$
HP	$K(1 - z^{-1})^2$
LPN	$K(1 + \epsilon z^{-1} + z^{-2}),\ \epsilon > \alpha/\sqrt{\beta},\ \beta > 0$
HPN	$K(1 + \epsilon z^{-1} + z^{-2}),\ \epsilon < \alpha/\sqrt{\beta},\ \beta > 0$
AP	$K(\beta + \alpha z^{-1} + z^{-2})$
General	$\gamma + \epsilon z^{-1} + \delta z^{-2}$

As already noted in connection with (10) and (11), T_E and T_F have identical numerators except for the $1/(1 + \hat{F})$ gain constant term. For convenience, the numerator of T_E is repeated below:

$$N(z) = -I + (G - I - J)z^{-1} + (J - H)z^{-2}. \tag{34}$$

It is evident that there are enough degrees of freedom here to choose the three coefficients independently and thus realize arbitrary zero locations; the fact that the leading coefficient is nonpositive is a trivial constraint. In Table II a complete set of design equations is given for the special generic transfer functions of Table I as well as for the general case. For each of the cases, a "simple" solution is also listed. These simple solutions, which are not unique, lead to fewer number of capacitors by setting as many of the capacitors G, H, I, and J as possible to zero, or by having $G = H$ or $I = J$ which, according to Fig. 5, also eliminates a capacitor as well as some switches.

It should be noted that the F-circuit capacitors $\hat{G}$, $\hat{H}$, $\hat{I}$, and $\hat{J}$ are related to G, H, I, and J by

$$\hat{x} = (1 + \hat{F})x \quad \text{where} \quad x = G, H, I, J. \tag{35}$$

For this reason a separate table need not be given for the synthesis of the zeros of T_F.

Similar zero-placement synthesis conditions are listed for transfer functions T'_E and T'_F in Tables III and IV, respectively. It is noted that these synthesis equations require prior knowledge of either E, C or $\hat{F}$, $\hat{C}$, in contrast to the T_E or T_F case.

This completes the material on zero-placement. It will be recognized that, for any given transfer function, four alternative realizations exist, i.e., T_E, T'_E, T_F, and T'_F. In the case of LP or BP designs, additional degrees of freedom exist, as there may be a choice of transfer functions (see Table I). At this point, we cannot state any general rule for

Table II—Zero placement formulas for T_E and T_F

Filter Type	Design Equations	Simple Solution
LP 20	$I = \|K\|$ $G - I - J = 2\|K\|$ $J - H = \|K\|$	$I = J = \|K\|$ $G = 4\|K\|, \quad H = 0$
LP 11	$I = 0$ $G - I - J = \pm\|K\|$ $J - H = \pm\|K\|$	$I = 0, \quad J = \|K\|$ $G = 2\|K\|, \quad H = 0$
LP 10	$I = \|K\|$ $G - I - J = \|K\|$ $J - H = 0$	$I = \|K\|, \quad J = 0$ $G = 2\|K\|, \quad H = 0$
LP 02	$I = 0$ $G - I - J = 0$ $J - H = \pm\|K\|$	$I = J = 0$ $G = 0, \quad H = \|K\|$
LP 01	$I = 0$ $G - I - J = \pm\|K\|$ $J - H = 0$	$I = J = 0$ $G = \|K\|, \quad H = 0$
LP 00	$I = \|K\|$ $G - I - J = 0$ $J - H = 0$	$I = \|K\|, \quad J = 0$ $G = \|K\|, \quad H = 0$
BP 10	$I = \|K\|$ $G - I - J = 0$ $J - H = -\|K\|$	$I = \|K\|, \quad J = 0$ $G = H = \|K\|$
BP 01	$I = 0$ $G - I - J = \pm\|K\|$ $J - H = \mp\|K\|$	$I = 0, \quad J = \|K\|$ $G = H = 0$
BP 00	$I = \|K\|$ $G - I - J = -\|K\|$ $J - H = 0$	$I = \|K\|, \quad J = 0$ $G = H = 0$
HP	$I = \|K\|$ $G - I - J = -2\|K\|$ $J - H = \|K\|$	$I = J = \|K\|$ $G = H = 0$
HPN and LPN	$I = \|K\|$ $G - I - J = \|K\|\epsilon$ $J - H = \|K\|$	$I = J = \|K\|$ $G = \|K\|(2 + \epsilon), \quad H = 0$
AP ($\beta > 0$)	$I = \|K\|\beta$ $G - I - J = \|K\|\alpha$ $J - H = \|K\|$	$I = \|K\|\beta, \quad J = \|K\|$ $G = \|K\|(1 + \beta + \alpha) = \|K\|C$ $H = 0$
General ($\gamma > 0$)	$I = \gamma$ $G - I - J = \epsilon$ $J - H = \delta$	$I = \gamma$ $J = \delta + x$ $G = \gamma + \delta + \epsilon + x$ $H = x \geq 0$

Note: $\hat{G} = G(1 + \hat{F})$, $\hat{H} = H(1 + \hat{F})$, $\hat{I} = I(1 + \hat{F})$, and $\hat{J} = J(1 + \hat{F})$.

Table III—Zero placement formulas for T_E'

Filter Type	Design Equations	Simple Solution
LP 20	$IC + IE - G = \pm\|K\|$ $H + G - JC - JE - IE = \pm 2\|K\|$ $EJ - H = \pm\|K\|$	$I = \frac{\|K\|(4E + C)}{EC}, \quad J = \frac{\|K\|}{E}$ $G = \frac{\|K\|(2E + C)^2}{EC}, \quad H = 0$
LP 11	$IC + IE - G = 0$ $H + G - JC - JE - IE = \pm\|K\|$ $EJ - H = \pm\|K\|$	$I = \frac{\|K\|(2E + C)}{EC}, \quad J = \frac{\|K\|}{E}$ $G = \frac{\|K\|(E + C)(2E + C)}{EC}, \quad H = 0$
LP 10	$IC + IE - G = \pm\|K\|$ $H + G - JC - JE - IE = \pm\|K\|$ $EJ - H = 0$	$I = \frac{2\|K\|}{C}, \quad J = 0$ $G = \frac{\|K\|(E + C)^2}{EC}, \quad H = 0$
LP 02	$IC + IE - G = 0$ $H + G - JC - JE - IE = 0$ $EJ - H = \pm\|K\|$	$I = \frac{\|K\|(E + C)}{EC}, \quad J = \frac{\|K\|}{E}$ $G = \frac{\|K\|(E + C)^2}{EC}, \quad H = 0$
LP 01	$IC + IE - G = 0$ $H + G - JC - JE - IE = \pm\|K\|$ $EJ - H = 0$	$I = \frac{\|K\|}{C}, \quad J = 0$ $G = \frac{\|K\|(E + C)}{C}, \quad H = 0$
LP 00	$IC + IE - G = \pm\|K\|$ $H + G - JC - JE - IE = 0$ $EJ - H = 0$	$I = \frac{\|K\|}{C}, \quad J = 0$ $G = \frac{\|K\|E}{C}, \quad H = 0$
BP 10	$IC + IE - G = \pm\|K\|$ $H + G - JC - JE - IE = 0$ $EJ - H = \mp\|K\|$	$I = J = \frac{\|K\|}{E}$ $G = \frac{\|K\|(2E + C)}{E}, \quad H = 0$
BP 01	$IC + IE - G = 0$ $H + G - JC - JE - IE = \pm\|K\|$ $EJ - H = \mp\|K\|$	$I = J = 0$ $G = 0, \quad H = \|K\|$
BP 00	$IC + IE - G = \pm\|K\|$ $H + G - JC - JE - IE = \mp\|K\|$ $EJ - H = 0$	$I = 0, \quad J = 0$ $G = \|K\|, \quad H = 0$
HP	$IC + IE - G = \pm\|K\|$ $H + G - JC - JE - IE = \mp 2\|K\|$ $EJ - H = \pm\|K\|$	$I = J = 0$ $G = H = \|K\|$
HPN and LPN	$IC + IE - G = \pm\|K\|$ $H + G - JC - JE - IE = \pm\|K\|\epsilon$ $EJ - H = \pm\|K\|$	See general solution below
AP	$IC + IE - G = \pm\|K\|\beta$ $H + G - JC - JE - IE = \pm\|K\|\alpha$ $EJ - H = \pm\|K\|$	See general solution below

Table III—*Continued*

Filter Type	Design Equations	Simple Solution
General $\delta > 0$	$IC + IE - G = \gamma$ $H + G - JC - JE - IE = \epsilon$ $EJ - H = \delta$	$I = \frac{\gamma + \delta + \epsilon}{C} + \frac{\delta}{E}, \quad J = \frac{\delta}{E}$ $G = I(C + E) - \gamma \quad H = 0$

selecting the optimum transfer function from these four functions. We can show for specific designs that considerable differences in total capacitance and sensitivity can be obtained for equivalent T_E, T_F, T'_E, and T'_F designs.

Hopefully, as we gain more experience in active-sc design, some insights will be acquired to shorten the design procedure. In the meanwhile, enough alternatives have to be tried until a satisfactory solution is obtained.

3.4 Capacitor value scaling

The synthesis equations given in the previous subsections result in unscaled capacitor values. To complete the synthesis, some scaling is required. The first order of business is to adjust the voltage level at the "secondary" output. If this voltage is too high, overloads will result, while if it is too low, unnecessary noise penalties may be taken.

Although the voltage levels may be obtained using analysis techniques,[16] the simplest procedure is to simulate the unscaled circuit[20] on a program such as CAPECOD.[21] This also serves as a confirmation of the correctness of the design.

To adjust the voltage level V', i.e., the flat gain of T', without affecting T, only the capacitors A and D need be scaled. More precisely, if it is desired to modify the gain constant associated with V' according to

$$T' \rightarrow \mu T', \tag{36}$$

then it is only necessary to scale A and D as

$$(A, D) \rightarrow \left(\frac{1}{\mu} A, \frac{1}{\mu} D\right). \tag{37}$$

The gain constant associated with T remains invariant under this scaling. The correctness of this procedure follows directly from simple signal flow graph concepts. Note that A and D are the only two capacitors connected to the operational amplifier output node.

In a similar fashion, it can be shown that, if the flat gain associated with V is to be modified, i.e.,

Table IV—Zero placement formulas for T_F'

Filter Type	Design Equations	Simple Solution
LP 20	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \lvert K\rvert(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = 2\lvert K\rvert(1 + \hat{F})$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	$\hat{I} = 0, \quad \hat{J} = \dfrac{\lvert K\rvert(2 + \hat{F})^2}{\hat{C}}$ $\hat{G} = \lvert K\rvert, \quad \hat{H} = \lvert K\rvert(1 + \hat{F})$
LP 11	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = 0$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = \lvert K\rvert(1 + \hat{F})$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	$\hat{I} = 0, \quad \hat{J} = \dfrac{\lvert K\rvert(1 + \hat{F})(2 + \hat{F})}{\hat{C}}$ $\hat{G} = 0, \quad \hat{H} = \lvert K\rvert(1 + \hat{F})$
LP 10	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \pm\lvert K\rvert(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = \pm\lvert K\rvert(1 + \hat{F})$ $\hat{H} = 0$	$\hat{I} = 0, \quad \hat{J} = \dfrac{\lvert K\rvert(2 + \hat{F})}{\hat{C}}$ $\hat{G} = \lvert K\rvert, \quad \hat{H} = 0$
LP 02	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = 0$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = 0$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	$\hat{I} = 0, \quad \hat{J} = \dfrac{\lvert K\rvert(1 + \hat{F})^2}{\hat{C}}$ $\hat{G} = 0, \quad \hat{H} = \lvert K\rvert(1 + \hat{F})$
LP 01	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = 0$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = \pm K(1 + \hat{F})$ $\hat{H} = 0$	$\hat{I} = 0, \quad \hat{J} = \dfrac{\lvert K\rvert(1 + \hat{F})}{\hat{C}}$ $\hat{G} = 0, \quad \hat{H} = 0$
LP 00	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \pm\lvert K\rvert(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = 0$ $\hat{H} = 0$	$\hat{I} = \dfrac{\lvert K\rvert(1 + \hat{F})}{\hat{C}}, \quad \hat{J} = 0$ $\hat{G} = 0, \quad \hat{H} = 0$
BP 10	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = -\lvert K\rvert(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = 0$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	$\hat{I} = \dfrac{\lvert K\rvert(1 + \hat{F}),}{\hat{C}}, J = \dfrac{\lvert K\rvert(1 + \hat{F})^2}{\hat{C}}$ $\hat{G} = 0, \quad \hat{H} = \lvert K\rvert(1 + \hat{F})$
BP 01	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = 0$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = -\lvert K\rvert(1 + \hat{F})$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	$\hat{I} = 0, \quad \hat{J} = \dfrac{\lvert K\rvert\hat{F}(1 + \hat{F})}{\hat{C}}$ $\hat{G} = 0, \quad \hat{H} = \lvert K\rvert(1 + \hat{F})$
BP 00	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \pm\lvert K\rvert(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = \mp\lvert K\rvert(1 + \hat{F})$ $\hat{H} = 0$	$\hat{I} = \hat{J} = \dfrac{\lvert K\rvert(1 + \hat{F})}{\hat{C}}$ $\hat{G} = \hat{H} = 0$
HP	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \lvert K\rvert(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = -2\lvert K\rvert(1 + \hat{F})$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	$\hat{I} = 0, \quad \hat{J} = \dfrac{\lvert K\rvert\hat{F}^2}{\hat{C}}$ $\hat{G} = \lvert K\rvert, \quad \hat{H} = \lvert K\rvert(1 + \hat{F})$
HPN and LPN	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \lvert K\rvert(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = -\lvert K\rvert\epsilon(1 + \hat{F})$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	See general solution below
AP	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \lvert K\rvert\beta(1 + \hat{F})$ $\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = \lvert K\rvert\alpha(1 + \hat{F})$ $\hat{H} = \lvert K\rvert(1 + \hat{F})$	See general solution below
General	$\hat{G}\hat{F} + \hat{G} - \hat{I}\hat{C} = \gamma(1 + \hat{F})$	$I = x \geq 0$

Table IV—*Continued*

Filter Type	Design Equations	Simple Solution
$\delta > 0$	$\hat{J}\hat{C} - \hat{F}\hat{H} - \hat{H} - \hat{G} = \epsilon(1 + \hat{F})$	$\hat{J} = \dfrac{\delta(1 + \hat{F})^2 + \epsilon(1 + \hat{F}) + \gamma}{\hat{C}} + \dfrac{x}{1 + \hat{F}}$
	$\hat{H} = \delta(1 + \hat{F})$	$\hat{G} = \gamma + \dfrac{C}{1 + \hat{F}} x$ $\hat{H} = \delta(1 + \hat{F})$

$$T \rightarrow \nu T, \tag{38}$$

the following capacitors must be scaled:

$$(B, C, E, F) \rightarrow \left(\frac{1}{\nu} B, \frac{1}{\nu} C, \frac{1}{\nu} E, \frac{1}{\nu} F\right) \tag{39}$$

Once satisfactory gain levels have been obtained at both outputs, it is convenient to scale the admittances associated with each stage so that the minimum capacitance value in the circuit becomes unity. This makes it easier to observe the maximum capacitance ratios required to realize a given circuit and also serves to "standardize" different designs so that the total capacitance required can be readily observed. The two groups of capacitors which are to be scaled together are listed below.

Group 1: (C, D, E, G, H).

Group 2: (A, B, F, I, J).

Note that capacitors in each group are distinguished by the fact that they are all incident on the same input node of one of the operational amplifiers.

This completes the design process for synthesizing practical SC-biquad networks. In the next section, a detailed example is given to demonstrate each step of the design.

IV. DESIGN EXAMPLES

In this section, some illustrative examples will be given. The first example is a low-pass notch network whose design is followed through, step by step, to illustrate the design procedure. The second example is a bandpass. For this case, eight different designs are displayed to demonstrate the versatility of the active-SC topologies and to provide some insight into the relative merits of different realizations.

4.1 Low-pass notch circuit

The transfer function to be realized will be based on the s-domain transfer function shown below:

$$T(s) = \frac{0.891975s^2 + (1.140926 \times 10^8)}{s^2 + 356.047s + (1.140926 \times 10^8)}. \tag{40}$$

This transfer function provides a notch frequency of $f_z = 1800$ Hz, a peak corresponding to a quality factor $Q_p = 30$ at $f_p = 1700$ Hz and 0 dB dc gain. The assigned sampling frequency is 128 kHz, i.e., $T_s = 7.8125\ \mu s$.

The z-domain transfer function is conveniently obtained via the bilinear transformation shown in (23). Because the band-edge frequency of 1700 Hz is much less than the sampling rate, it is not necessary to prewarp the $T(s)$ given in (40). Applying the bilinear transformation to (40) yields, after some algebraic manipulations:

$$T(z) = 0.89093\,\frac{1 - 1.99220z^{-1} + z^{-2}}{1 - 1.99029z^{-1} + 0.99723z^{-2}}. \tag{41}$$

Note that in obtaining this transfer function a high degree of numerical precision is required. However, this does not result in high sensitivities, since the capacitor ratios define only the departures from -2 and $+1$ in the above terms.

Only the T_E and T_F realizations of the above circuit will be given here, as these circuits are more economical in the number of capacitors required for realization. The synthesis itself is straightforward. The capacitors C, E or $\hat{C}$, $\hat{F}$ are determined from (28) or (29), respectively, while the capacitors G, H, I, J, or $\hat{G}$, $\hat{H}$, $\hat{I}$, $\hat{J}$ are obtained from the "simple solution" entry in Table II. Finally, of course, A, B, D or $\hat{A}$, $\hat{B}$, $\hat{D}$ are set equal to unity according to (7). The resulting unscaled capacitor values are given in the appropriate columns of Table V. Note that, in this table and all succeeding ones, the hats are omitted from the F-circuit capacitors for notational convenience. Also note that since $I = J$ these two switched capacitors are replaced by the unswitched capacitor K, ($K = I = J$), in accordance with Fig. 5.

At this point, the unscaled E- and F-circuits were simulated via CAPECOD.[20,21] These results confirmed that the T_E and T_F were both correct. In particular, the maximum gain in both these realizations was approximately 10.56 dB. However, the maximum gains for T'_E and T'_F were very low. It was decided to increase these gains also to a maximum of 10.56 dB. In this way, the first stage is no more susceptible to overloads than the second stage. Specifically, it was found that

$$T'_{E\,\text{MAX}} \approx -11.05 \text{ dB}, \qquad T'_{F\,\text{MAX}} \approx -10.96 \text{ dB}. \tag{42}$$

Therefore, in accordance with (36),

$$\mu = 12.0365, \qquad \hat{\mu} = 11.9124. \tag{43}$$

Using these factors to rescale A, D and $\hat{A}$, $\hat{D}$, respectively, as given in (37), yields the "dynamic range adjusted" capacitor values shown in Table V. Finally, the capacitors associated with each operational amplifier stage are separately rescaled so that the minimum capacitance value becomes 1 pF. These "final" values are also shown in Table V.

In comparing the "final" realizations, we note that the F-circuit requires roughly 12 times the total capacitance of the E-circuit, in spite of the fact that the initial values were almost identical. Thus, alternative designs must be carried to completion before they can be meaningfully compared. It should be noted that other practical examples exist where the F-circuit designs are dramatically more efficient than the corresponding E-circuit designs.

As a final step, Monte Carlo simulations on the two circuits were carried out, assuming each capacitor to have a flat, independent ±1 percent tolerance. It should be pointed out that independent 1-percent capacitor tolerances represent a pessimistic estimate in view of today's technology. Since capacitor deviations, whether they are due to manufacturing tolerances, temperature variations, or aging, are highly correlated, the capacitor ratios are recognized to be achievable[6,7] with considerably better precision. These results, given in Table V, show both circuits to be excellent, with the E-circuit being slightly superior. Note that σ_1 is the absolute standard deviation at 1 Hz, while σ_{1700} is the standard deviation of the relative gain at 1700 Hz with respect to 1 Hz.

Table V—Low-pass notch realization

	E-Circuit			*F*-Circuit		
Capacitor (pf)	Initial	Dynamic Range Adjusted	Final	Initial	Dynamic Range Adjusted	Final
A	1.0000	0.08308	1.0000	1.0000	0.08395	30.1895
B	1.0000	1.0000	12.0365	1.0000	1.0000	359.629
C	0.00694	0.00694	2.5035	0.00696	0.00696	1.0000
D	1.0000	0.08308	29.9613	1.0000	0.08395	12.0591
E	0.00277	0.00277	1.0000	—	—	—
F	—	—	—	0.00278	0.00278	1.0000
G	0.00694	0.00694	2.5035	0.00696	0.00696	1.0000
H	—	—	—	—	—	—
I	—	—	—	—	—	—
J	—	—	—	—	—	—
$K(I = J)$	0.89093	0.89093	10.7238	0.89340	0.89340	321.293
ΣC (pF)	—	—	59.7	—	—	726.1
σ_1 (dB)	—	—	0.068	—	—	0.068
σ_{1700} (dB)	—	—	1.233	—	—	1.271

4.2 High Q bandpass circuits

This example demonstrates the versatility of the topology and examines the various tradeoffs this versatility provides. As noted in the previous sections and highlighted in Tables I through IV, we have the following degrees of freedom in realizing a bandpass active-SC biquad.

(*i*) The transfer function; namely, BP 00, BP 01, or BP 10, as shown in Table I.

(*ii*) The circuit realization; namely, the *E*- or *F*-circuits.

(*iii*) The output port; namely, T_E or T'_E for the *E*-circuit and T_F or T'_F for the *F*-circuit.

Using the "simple" solutions given in Tables II through IV, these freedoms yield 12 different design possibilities for a bandpass biquad realization. In selecting a design from these 12 possibilities, we adopt the following criterion: The circuit must meet all frequency domain requirements within acceptable tolerances. The circuit that satisfies this criterion, while requiring an estimated minimum silicon area for its realization, is in our view the best design. The primary factor that determines the required silicon area is total capacitance. Of secondary importance are the number of capacitors and the number of switches, i.e., an unswitched capacitor consumes less area than a switched capacitor of the same capacitance value. Let us now consider the various design possibilities in view of these considerations.

The transfer function to be realized will be based on the following *s*-domain transfer function:

$$T(s) = \frac{2027.9s}{s^2 + 641.28s + (1.0528 \times 10^8)}, \tag{44}$$

which possesses a center frequency $f_o = 1633$ Hz, a quality factor $Q = 16$, and a peak gain of 10 dB at f_o.

For the active-SC design, the assumed sampling frequency is 8 kHz, i.e., $T_s = 125\ \mu s$. One method for obtaining the *z*-domain transfer function is the application of the bilinear transform given by (23) to the prewarped *s*-domain transfer function. The prewarped *s*-domain function is obtained by adjusting the desired 3-dB frequencies f_l and f_h according to the relation[11]

$$\tilde{f}_{l,h} = \frac{1}{\pi T_s} \tan\,(\pi f_{l,h} T_s), \tag{45}$$

where $\tilde{f}_{l,h}$ denotes the prewarped 3-dB frequencies. Upon calculating $\tilde{f}_l$ and $\tilde{f}_h$, the following prewarped transfer function is determined:

$$\tilde{T}(\tilde{s}) = \frac{3159.2\tilde{s}}{\tilde{s}^2 + 999.0289\tilde{s} + (1.4285 \times 10^8)}. \tag{46}$$

Substituting for the prewarped complex frequency variable $\bar{s}$ in (46), the bilinear transform (23) yields the following BP 10 z-domain transfer function:

$$T(z) = \frac{0.1219(1 - z^{-1})(1 + z^{-1})}{1 - 0.5455z^{-1} + 0.9229z^{-2}}. \tag{47}$$

In accordance with the synthesis procedures given in Section III and in Tables II through IV, the four possible "simple" BP 10 designs were evaluated. The total capacitance required for each of these realizations is listed in Table VI. The capacitance values were scaled in the same manner as described in the previous example. As noted in Section II, switched capacitors, G and H, when equal, can be replaced by a single unswitched capacitor. All four designs were simulated on CAPECOD[20,21] to verify the response and to determine their statistical behavior. The Monte Carlo simulations were carried out at the 1633-Hz peak frequency, assuming each capacitor to have a flat, independent ±1 percent tolerance. Note that σ_{1633} is the standard deviation of the absolute gain at 1633 Hz. The results of these simulations are listed in Table VI.

Comparing the four designs, it is apparent that the F-circuit, using either T_F or T'_F, requires less capacitance and is less sensitive than either of the E-circuit designs. The 0.25-dB standard deviation for the F designs indicates the good stability of these circuits. Although the T'_F design consumes slightly less capacitance (~3 pF) than the T_F design, it requires one more capacitor and four additional switches. Depending on the layout, the additional capacitor, switches, and connections can more than offset the total capacitance advantage. With this reasoning, we are inclined to recommend the T_F design.

As noted earlier, one of our degrees of freedom is the choice of the BP transfer function. Exercising this freedom, as we shall soon demonstrate, can significantly impact the character of the designed circuit. Alternative realizations can be obtained by altering the z-domain transfer function in (47) to achieve a BP 00 function. An appropriate BP 00 function is obtained from (47) by removing the zero at one-half the sampling rate and adjusting gain K from 0.1219 to 0.1953 to preserve the desired 10-dB peak gain. The desired BP 00 transfer function is then

Table VI—Comparison of BP 10 and BP 00 biquad realizations

Case		E-Circuit (T_E)	F-Circuit (T_F)	E-Circuit (T'_E)	F-Circuit T'_F)
(1) BP 10, 10-dB	ΣC (pF)	55.0	51.2	75.1	48.3
gain at 1633 Hz	σ_{1633} (dB)	0.2738	0.2524	0.9932	0.2569
(2) BP 00, 10-dB	ΣC (pF)	46.7	32.7	39.6	32.8
gain at 1633 Hz	σ_{1633} (dB)	0.2852	0.2554	0.2833	0.2570

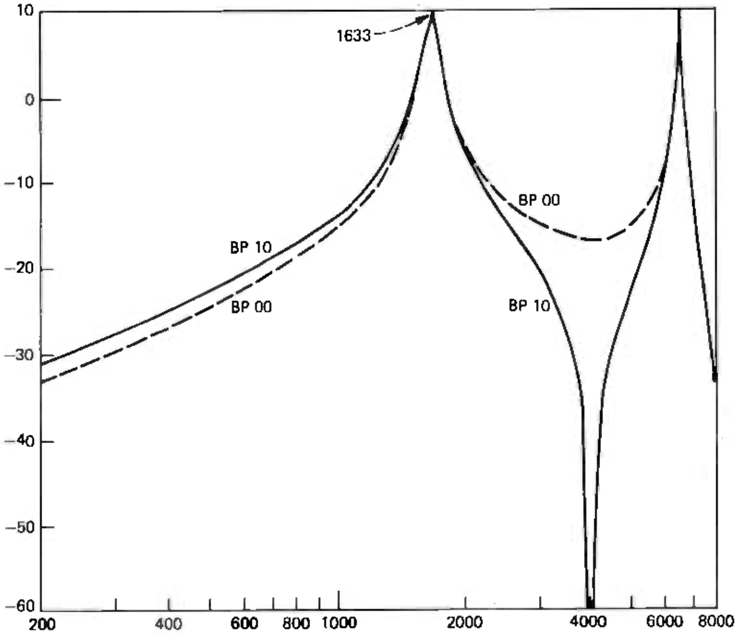


Fig. 8.—Frequency responses for the BP 10 and BP 00 designs with $Q = 16$ and $f_0 = 1633$ Hz.

$$T(z) = \frac{0.1953(1 - z^{-1})}{1 - 0.5455z^{-1} + 0.9229z^{-2}}. \tag{48}$$

Before carrying out the circuit designs, let us examine the frequency domain behavior of (48). The frequency responses for both the BP 10 and BP 00 transfer functions are plotted in Fig. 8. The BP 00 response is seen to be equivalent to the BP 10 response except for frequencies near 4000 Hz, where (47) possesses a zero. In any event, the BP 00 response was considered adequate. The appropriately scaled E- and F-circuit designs are listed in Table VI. In contrast to the first example, the F-circuit designs are seen to require less total capacitance and to be less sensitive than the E-circuit designs. More important, however, is the comparison between the BP 10 and the BP 00 designs. The BP 00 F-circuit designs are seen to require about 20 pF less total capacitance while sacrificing nothing. As far as the F-circuit BP 00 designs are concerned, our inclination is to recommend the T'_F design which requires four fewer switches than the T_F design.

V. CONCLUSIONS

Two closely related two-amplifier, active-SC filter topologies have been presented. These circuits have been constructed so that they are immune to parasitic capacitances normally present in SC networks.

The first topology, the E-circuit, has been shown to permit the realization of arbitrary stable z-domain biquadratic transfer functions at either of its two outputs. The second topology, the F-circuit, has been shown to be only slightly less general in that only certain unimportant pole pairs (real poles of opposite signs) are not realizable. A complete set of synthesis equations is presented for both of these circuits. Since every desired biquadratic transfer function has at least four alternative realizations, the designer can choose among these to best satisfy his economic and sensitivity requirements. If the "simple solutions" given in the tables do not satisfy his requirements, many other possible realizations also exist, especially if an LP or BP is being designed.

Several examples have been given to demonstrate the design process and to highlight the many degrees of freedom these biquad topologies provide.

APPENDIX

In the text, we have assumed that the input signal is sampled and held for the full clock period. While this assumption simplifies the analysis, it is by no means necessary. Thus, consider the more general case where the clock period is still T_s but the desired input signal is sampled and held only for the interval τ_e, ($\tau_e < T_s$). The subscript "e" here is meant to imply the even phase of the clock period. The odd phase of the clock period is referred to as τ_o ($\tau_o = T_s - \tau_e$). The input during this phase is assumed to be "undesirable." These concepts are also shown in Fig. 9.

In certain special cases, the circuits of Fig. 1 will continue to perform correctly even with this less restricted class of inputs. This happens whenever $H = 0$ and $J = 0$. This is readily confirmed by observing that the input voltage during τ_o is coupled into the circuit only via the two capacitors H and J. When these are both absent, the input during the odd clock phase is simply not "seen" by the circuit.

In general, however, the circuits of Fig. 1 must be modified by reversing the switch phasings of the switched capacitors A, H, and I. The resulting active-SC circuits are shown in Fig. 10. Note that the topology is so arranged that only the input during the even phase is coupled into the circuit. Thus, the input during the odd phase is again immaterial. One slight constraint on the operation of this circuit is that now the "correct" output is also only obtained during the even phase. Thus, if a fully held output signal is desired, the circuits of Fig. 10 will have to be followed by a suitable sample-and-hold circuit.

The proof of the above statement is most conveniently obtained by using the equivalent circuit techniques given in Ref. 16. For convenience in analysis, the duty cycle is assumed to be 50 percent, i.e., $\tau_e =$

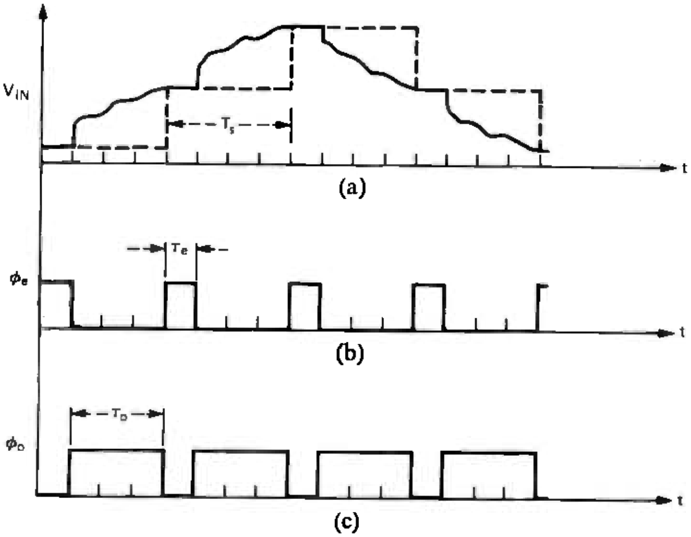


Fig. 9—Waveforms when input is not fully held.

$\tau_o = \frac{1}{2}T_s$; however, this does not detract from the generality of the results.

The equivalent circuit given in Fig. 11 can be readily constructed by substituting the equivalent circuits[16] for each SC element and operational amplifier in the circuit schematic given in Fig. 10. Due to the new switch phasings, $V_{\rm in}^o$ does not enter the filter. Writing nodal charge equations at the four virtual ground nodes of this circuit yields the following system of equations:

$$GV_{\rm in}^e + DV'^e - Dz^{-1/2}V'^o + (C+E)V^e - Ez^{-1/2}V^o = 0, \qquad (49)$$

$$-Hz^{-1/2}V_{\rm in}^e + DV'^o - Dz^{-1/2}V'^e + EV^o - Ez^{-1/2}V^e = 0, \qquad (50)$$

$$IV_{\rm in}^e + (F+B)V^e - Bz^{-1/2}V^o = 0, \qquad (51)$$

and

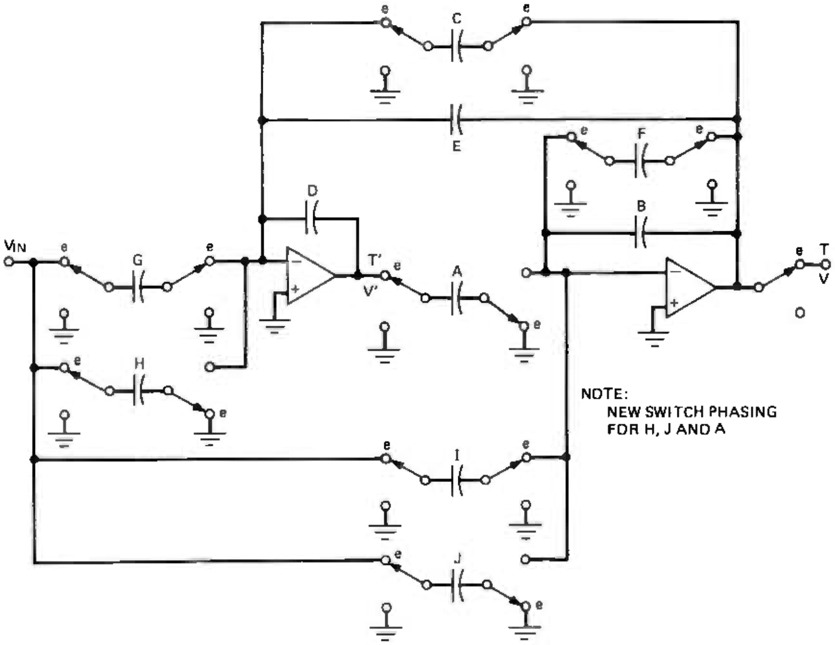
$$Jz^{-1/2}V_{\rm in}^e - Az^{-1/2}V'^e + BV^o - Bz^{-1/2}V^e = 0. \qquad (52)$$

Algebraically eliminating V^o and V'^o from these equations results in the following pair of equations:

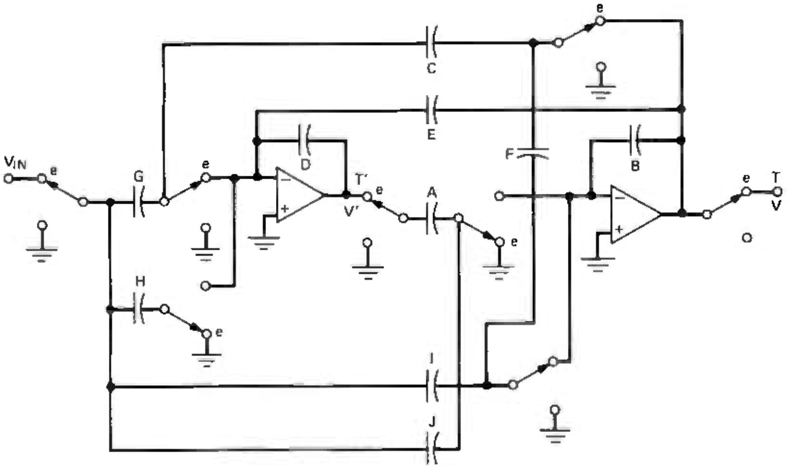
$$(I - Jz^{-1})V_{\rm in}^e + (F + B - Bz^{-1})V^e - Az^{-1}V'^e = 0 \qquad (53)$$

and

$$(G - Hz^{-1})V_{\rm in}^e + D(1 - z^{-1})V'^e + (C + E - Ez^{-1})V^e = 0. \qquad (54)$$



(a)



(b)

Fig. 10—(a) General biquad topology for input signals which are not held constant over the full clock period. (b) General biquad topology for input signals which are not held constant over the full clock period (minimum switch configuration).

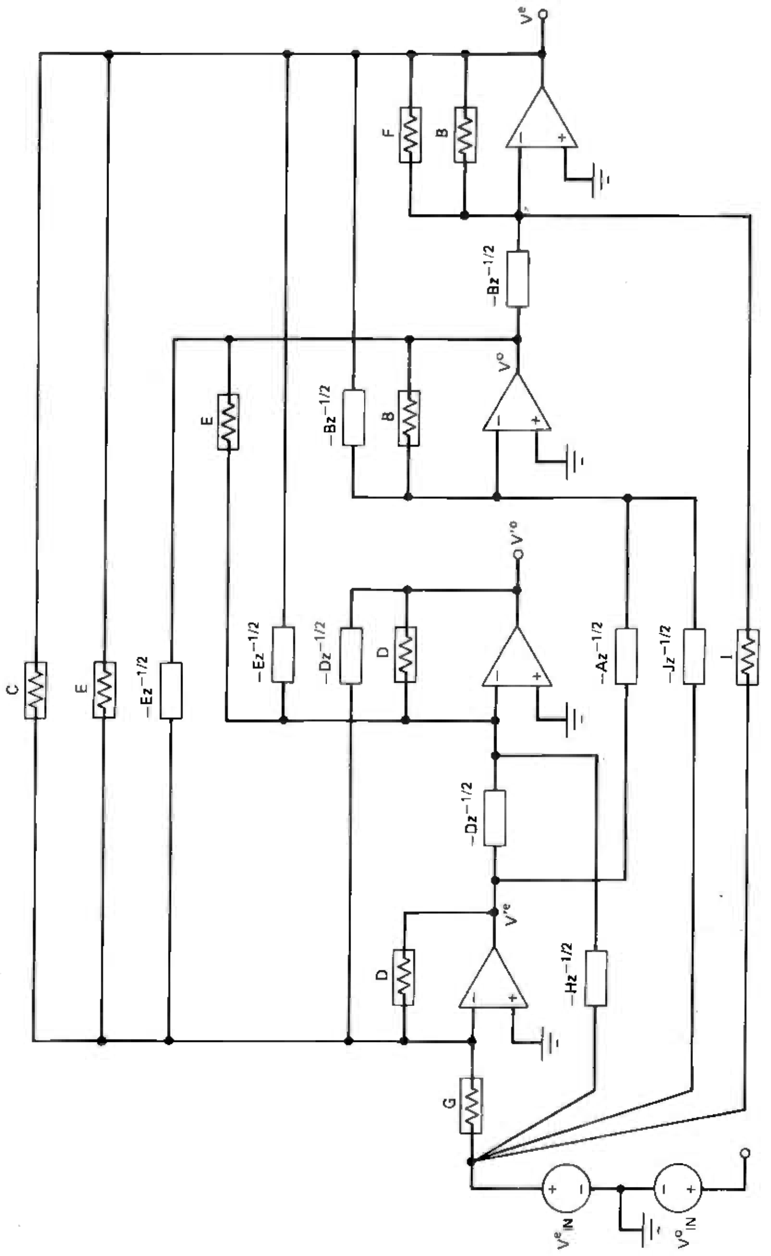


Fig. 11—z-domain equivalent circuit for the active SC biquad in Fig. 10.

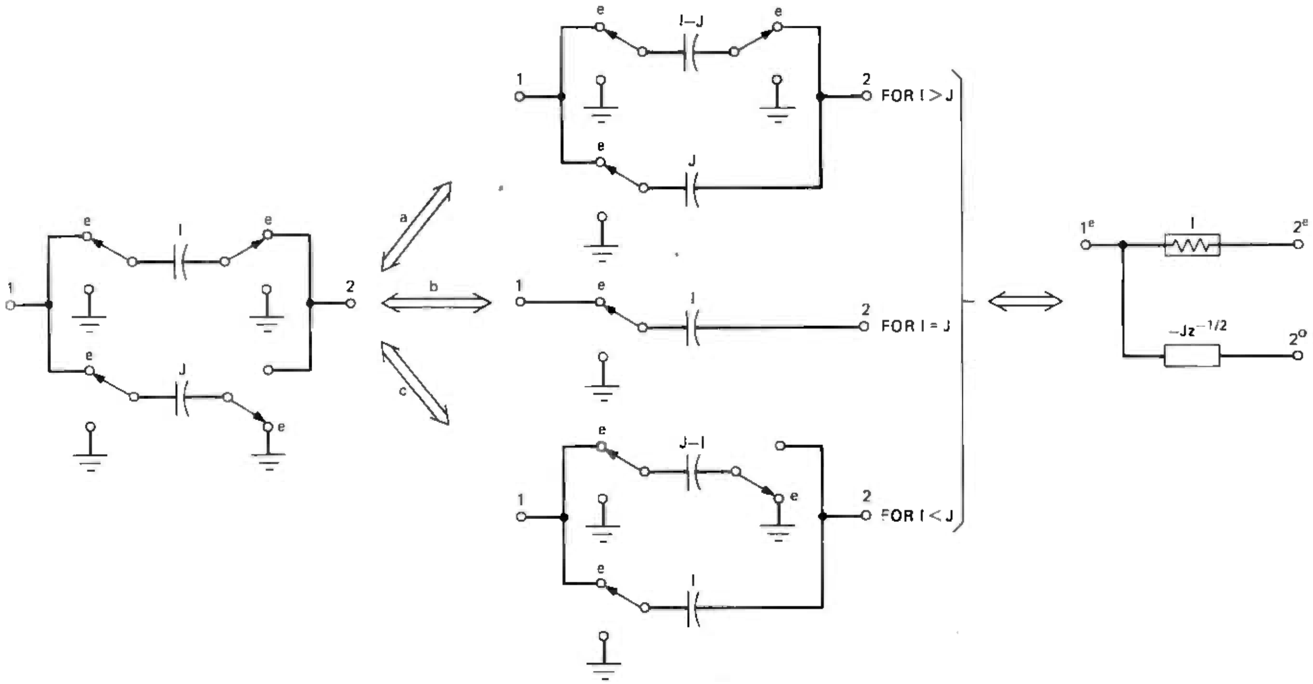


Fig. 12—sc element transformations for the biquad in Fig. 10 (ports 1 and 2 are assumed to be buffered).

Equations (53) and (54) can be readily verified to be the nodal equations which characterize the equivalent circuit given in Fig. 4. Thus, during the even phase the transfer functions will again be those given in (5) and (6), thus proving our contention. Note carefully, however, that, during the odd phase, the transfer functions which relate V_{in}^e to V^o and V'^o are quite different from those which characterize the even phase operation. In fact, we can express the two odd outputs (V^o and V'^o) as functions of the even outputs (V^e and V'^e) and the even input (V_{in}^e); i.e.,

$$V^o = z^{1/2}\left[\left(1+\frac{F}{B}\right)V^e + \frac{I}{B}V_{\text{in}}^e\right] \tag{55}$$

and

$$V'^o = z^{1/2}\left[z^{-1}V'^e - \frac{E}{D}\left(1+\frac{F}{B}-z^{-1}\right)V^e + \left(\frac{EI}{DB}-\frac{H}{D}z^{-1}\right)V_{\text{in}}^e\right]. \tag{56}$$

It is noted that, for $F = I = 0$, $V^e = z^{-1/2}V^o$; thus, V is held for the full clock period. On the other hand, when $E = H = 0$, V' is held for the full clock period. Thus, in some special cases, at least, a fully held output can be obtained. Finally, SC-element equivalences similar to those given in Fig. 5 can be used to reduce sensitivity and/or total capacitance. These element equivalences and their common z-domain equivalent circuit are given in Fig. 12. As in Fig. 5, these equivalences are based on the assumption that port 1 is voltage-driven and port 2 is connected to an operational-amplifier virtual ground.

REFERENCES

1. J. J. Friend, C. A. Harris, and D. Hilberman, "STAR: An Active Biquadratic Filter Section," IEEE Trans. Circuits and Systems, *CAS-22* (February 1975), pp. 115–121.
2. L. C. Thomas, "The Biquad: Part I—Some Practical Considerations," and "The Biquad: Part II—A Multi-Purpose Active Filtering System," IEEE Trans. Circuit Theory, *CT-18* (May 1971), pp. 350–357 and pp. 358–361.
3. W. Worobey and J. Rutkiewicz, "Tantalum Thin-Film RC Circuit Technology for a Universal Active Filter," IEEE Trans. Parts, Hybrids, and Packaging, *PHP-12* (December 1976), pp. 276–282.
4. P. E. Fleischer and J. J. Friend, "Active Filters at Bell Laboratories," presented at the 1977 IEEE International Symposium on Circuits and Systems, April 1977.
5. J. T. Caves et al., "Sampled Analog Filtering Using Switched Capacitors as Resistor Equivalents," IEEE J. Solid State Circuits, *SC-12*, No. 6 (December 1977), pp. 592–599.
6. B. J. Hosticka, R. W. Brodersen, and P. R. Gray, "MOS Sampled Data Recursive Filters Using Switched Capacitor Integrators," IEEE J. Solid State Circuits, *SC-12*, No. 6 (December 1977), pp. 600–608.
7. G. M. Jacobs, "Practical Design Considerations for MOS Switched Capacitor Ladder Filters," unpublished work.

8. B. J. White, G. M. Jacobs, and G. F. Landsburg, "A Monolithic Dual-Tone Multifrequency Receiver," Digest of the 1979 ISSCC (February 1979), pp. 36–37.
9. P. E. Fleischer and K. R. Laker, internal memorandum, November 8, 1978.
10. P. E. Fleischer, Laboratory Notebook, May 26, 1978.
11. L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1975, Ch. 4.
12. A. J. Vera, unpublished work.
13. P. E. Fleischer and J. Tow, "Design Formulas for Biquad Active Filters Using Three Operational Amplifiers," Proc. IEEE, *61*, No. 5 (May 1973), pp. 662–663.
14. G. C. Temes, "The Derivation of Switched Capacitor Filters from Active-RC Prototypes," Electron. Lett., *14*, No. 12 (June 8, 1978), pp. 361–362.
15. C. F. Kurth and G. S. Moschytz, "Nodal Analysis of Switched Capacitor Networks," IEEE Trans. Circuits and Systems, *CAS-26* (February 1979), pp. 93–104, and "Two-Port Analysis of Switched Capacitor Networks Using Four-Port Equivalent Circuits," IEEE Trans. Circuits and Systems, *CAS-26* (March 1979).
16. K. R. Laker, "Equivalent Circuits for the Analysis and Synthesis of Switched Capacitor Networks," B.S.T.J., *58* (March 1979), pp. 727–767.
17. P. E. Fleischer, "Sensitivity Minimization in a Single Amplifier Biquad Circuit," IEEE Trans. Circuits and Systems, *CAS-23*, No. 1 (January 1976), pp. 45–55.
18. H. W. Schüssler, *Digitale Systeme Zur Signalverarbeitung*, Berlin: Springer-Verlag, 1973, p. 38.
19. S. A. Tretter, *Introduction to Discrete-Time Signal Processing*, New York: John Wiley, 1976, pp. 219–222.
20. P. E. Fleischer, "Computer Analysis of Switched Capacitor Networks in the Frequency Domain," unpublished work.
21. R. M. M. Chen et al., "L5 System: Role of Computing and Precision Measurements," B.S.T.J., *53*, No. 10 (December 1974), pp. 2249–2267.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

The *BELLPAC** Modular Electronic Packaging System

By W. L. HARROD and A. G. LUBOWE

(Manuscript received May 1, 1979)

The BELLPAC* *system is a family of electronic packaging modules being used in the physical design of more than 40 new Bell Laboratories-developed systems. The* BELLPAC *system consists of a set of circuit packs, connectors (both circuit pack and backplane), and shelf hardware. A range of circuit pack sizes and interconnection densities is provided to match system packaging needs. Present elements include circuit pack connectors with pin-outs ranging from 50 to 300, circuit pack sizes ranging from 30 to 100 square inches, and circuit pack technologies ranging from simple, low-density, epoxy glass (or epoxy-coated metal) circuits up to fine-line multilayer boards. In this paper, we review the physical design of the* BELLPAC *system. We also describe the large body of design and manufacturing support information available to system development organizations using* BELLPAC *hardware.*

I. HISTORY AND DEVELOPMENT GOALS

The *BELLPAC** system, formerly known as CDCP (Common Design Circuit Packaging), grew out of a working committee established in 1975 and led by J. G. Brinsfield of the Interconnection Technology Laboratory at Bell Laboratories in Whippany, New Jersey. Committee representatives from each system development area, from Western Electric, and from the electronic components area established the following goal for their work:

> By the use of standard physical design, to reduce the costs and time intervals required to develop and manufacture new systems.

The committee also established requirements that were felt neces-

* Trademark of Western Electric.

sary for the wide acceptance of a standardized set of hardware building blocks. Among these were the following objectives:

(*i*) Design packaging modules to provide the correct trade-offs between flexibility and cost.

(*ii*) Serve a large enough customer base to provide economies of scale.

(*iii*) Provide continual interaction between packaging developers and users on requirements, design status, schedules, and costs.

(*iv*) Provide off-the-shelf prototype hardware.

(*v*) Provide timely documentation of components, assemblies, specifications, and application guidelines.

(*vi*) Insure that proper computer aids are available for the design of printed wiring boards.

(*vii*) Demonstrate the feasibility of production early in the development cycle.

(*viii*) Plan ahead for manufacturing buildup.

(*ix*) Provide extensions to the hardware family to meet new requirements while maintaining compatibility with existing designs.

The development of *BELLPAC* hardware was mainly the responsibility of the Interconnection Technology Laboratory at Bell Laboratories in Whippany, although invaluable contributions were made by several of the system development organizations. During the development of the hardware building blocks, the objectives listed above were aggressively pursued, and progress was monitored by the committee.

In our view, the development objectives have been successfully met. The number of projects which have chosen to use *BELLPAC* hardware is now large enough to guarantee high-volume manufacturing benefits to even very low-running projects. A broad range of physical design options has been employed in the projects using the *BELLPAC* system, indicating that the trade-offs between flexibility and costs were correct for a majority of the users. Recently, some specialized parts have been added for very high-volume applications. These parts will be available to low-volume users as well, but will have less flexibility in application. The committee which originally served as the steering group for the development of the *BELLPAC* system has now become a part of the *BELLPAC* System Users Forum. Regular meetings are held to review design and manufacturing status.

II. PHYSICAL DESIGN CONCEPTS

The *BELLPAC* hardware family covers a broad range of options in board sizes, pin-out densities, and shelf configurations; these options, however, all stem from a small set of parts and a common design concept. The way in which these parts are designed and assembled is summarized here.

The exploded view of Fig. 1 illustrates the physical design concept. Circuit card connectors contain receptacle contacts. These contacts mate with 25-mil square pins which are press-fit into an epoxy glass backplane. The circuit packs, connectors, and backplanes are described in more detail in later sections. Proper alignment of the circuit card to the pin field is assured by the parts labeled *ramp* and *spacer-aligner* in Fig. 1. The spacer-aligner contains precision-molded apertures that fit over the pins in the backplane and thus align the protrusions on the spacer-aligner. These protrusions serve to align the ramps to the pin field. The ramps, in turn, guide the circuit cards into position.

The support structure for the *BELLPAC* backplane is the mounting plate indicated in Fig. 1. The backplane assembly is self-aligning via precision holes and target pins, so that no special jigging or fixturing is required.

The apparatus housings fasten to the mounting plates and provide support for the circuit cards. Card guides are plastic tracks that snap into the apparatus housing where required. This approach provides minimal blockage of air flow while allowing a modularity of 0.25 in. in circuit card spacing.

As illustrated, a lever for insertion and withdrawal of the circuit cards is incorporated, and a hinged designation strip for identification of circuit card position is provided.

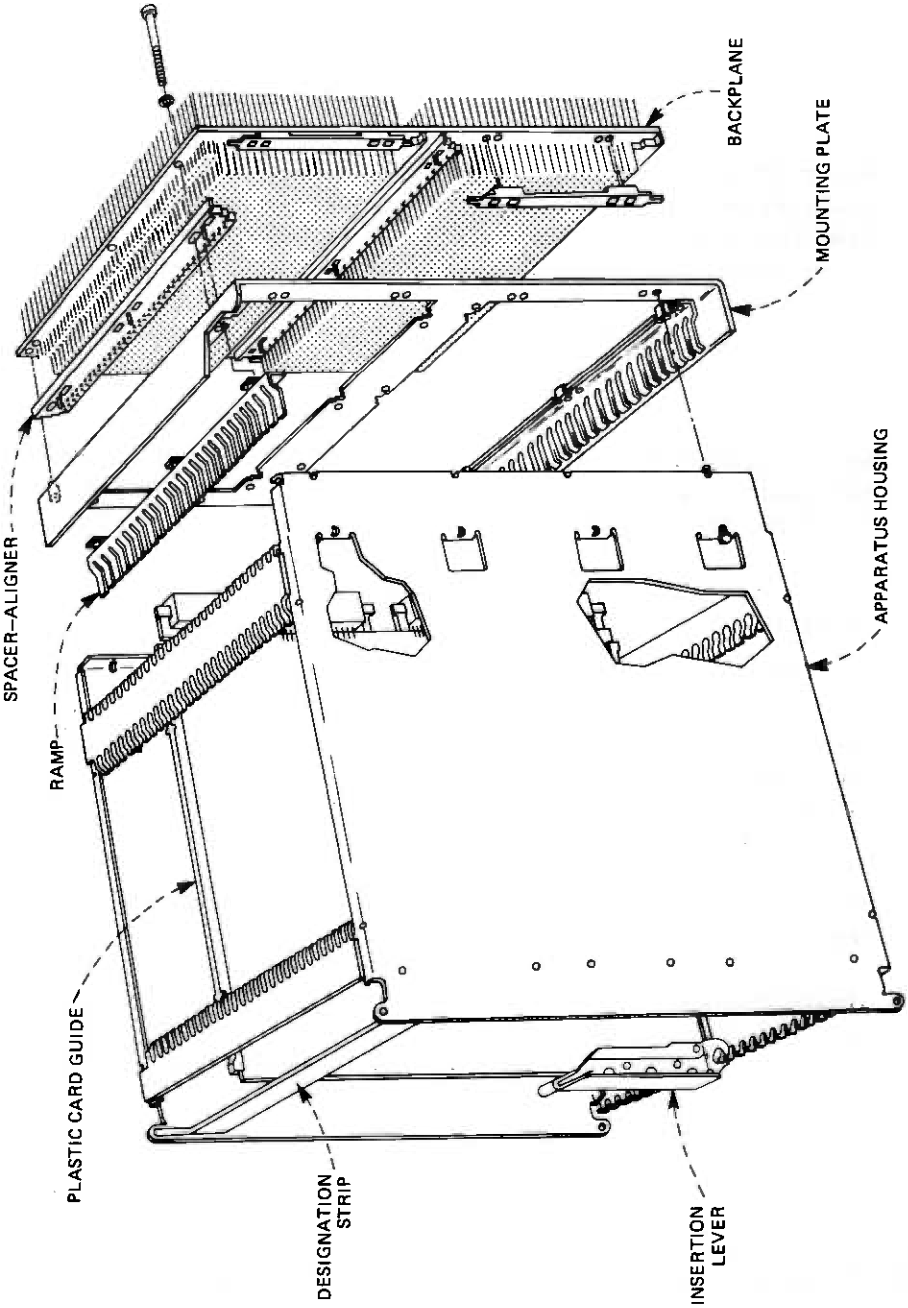
2.1 Compliant pin and circuit pack connectors

A key element of the packaging system is a compliant pin which is press-fit into printed wiring backplanes. The reliability of the compliant pin-to-backplane interface has been established by studies at Bell Laboratories over the past five years. Portions of this work are covered in Ref. 1.

The major experience to date has been accumulated with pins manufactured to Bell Laboratories specification by Winchester Electronics in Oakville, Connecticut. The compliant region, which is of Winchester's design, is in the center section of the pin shown in Fig. 2. (Reference 2 gives further details of the compliant region design.) The large square shoulder section of the pin appears on the circuit pack side of the backplane and is provided as an aid to insertion tool design. Also shown in Fig. 2 is the contact which is the basis for all *BELLPAC* connectors.

A cross-sectional view of the compliant section before insertion is shown in Fig. 3 and after insertion in Fig. 4. The compression of the pin and the intimate contact between the pin and the surrounding plated through hole are clearly shown.

The circuit card connectors that mate with the compliant pins all utilize contacts of the type shown in Fig. 2. The contacts are assembled into plastic housings to provide a family of circuit pack connectors.


SPACER–ALIGNER
RAMP
PLASTIC CARD GUIDE
DESIGNATION STRIP
INSERTION LEVER
APPARATUS HOUSING
MOUNTING PLATE
BACKPLANE

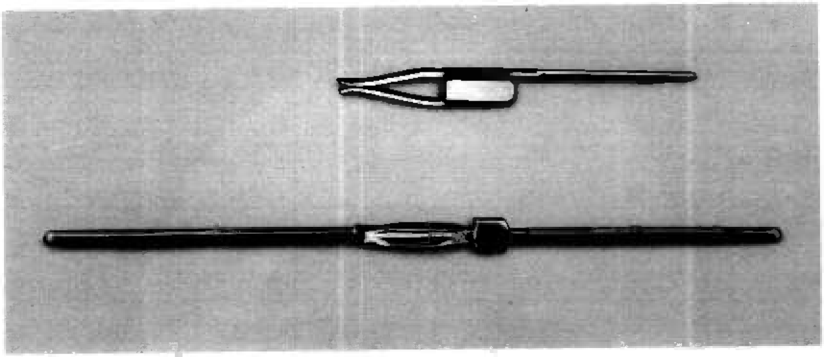


Fig. 2—Compliant pin and connector contact.

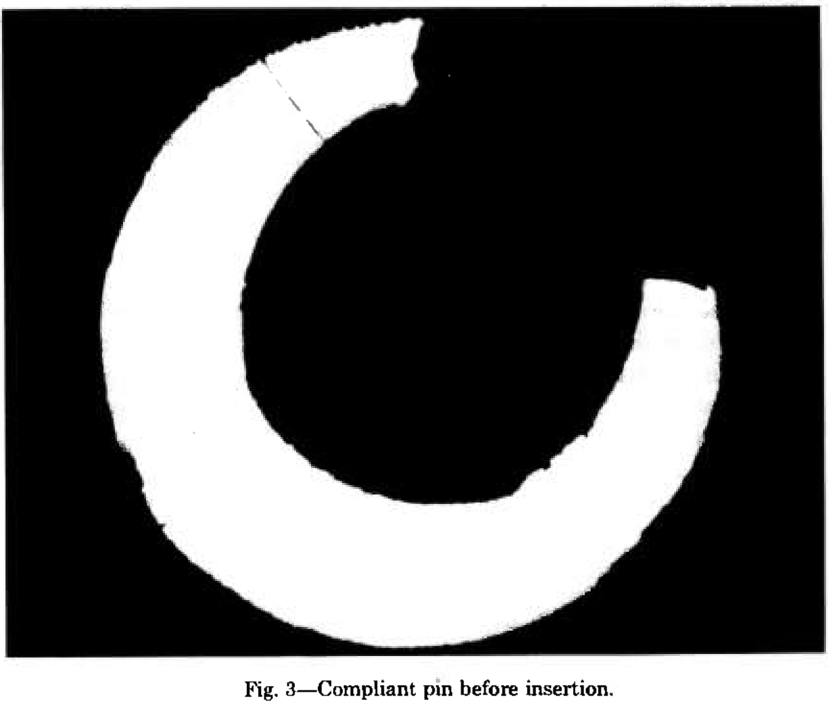


Fig. 3—Compliant pin before insertion.

The connector family is modular in two dimensions. Nominal heights (which correspond to circuit card heights) of 4, 6, and 8 inches are available. The connectors provide variable pinout densities by accessing varying numbers of columns of backplane pins. Connectors are available to mate with 2, 3, 4, and 6 columns of pins. Figure 5 illustrates the *BELLPAC* connector family and indicates the number of contacts available for each code.

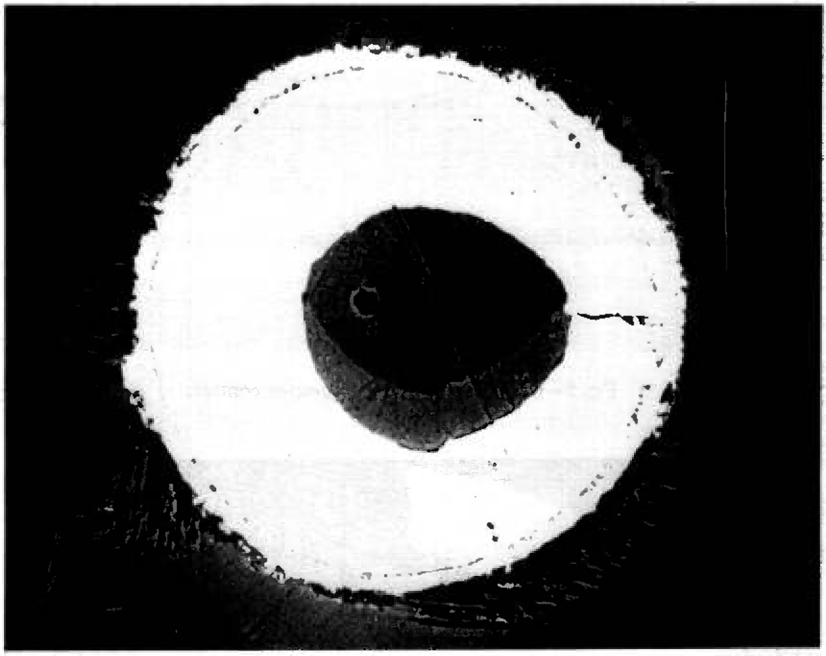


Fig. 4—Compliant pin after insertion.

The lower density connectors with two or three columns of contacts provide 16 or 24 pin-outs per inch of circuit pack height. These connectors are attached to circuit packs by heat-staking plastic posts on the connectors so that they plastically deform and completely fill corresponding holes in the circuit cards. Electrical connections are made by soldering tails of the connector contacts into plated through holes in the circuit cards at the same time as other components are soldered to the circuit packs.

The higher-density connectors with four or six columns of contacts provide 32 or 48 pin-outs per inch of circuit pack height. These connectors are provided with ears so that they may be riveted to the circuit cards. Electrical connections are made in a separate mass soldering operation which reflows the solder on the tails of the connector contacts and their corresponding leads on the circuit packs.

2.2 Compliant pin backplanes

A large degree of design flexibility is inherent in the backplane system. All pins are placed on 0.125-in.-grid positions; however, only those columns of pins required for mating with circuit card connectors or other connectors need be installed. Interconnections among pins in the backplane can be provided by any combination of printed wiring (double-sided or multilayer), manual wiring, automatic wire-wrap, and backplane cables (switchboard or tape). All pins are designed to allow three wire-wrap levels or two wire-wrap levels and one backplane connector engagement on the wiring side of the backplane. On the circuit pack side, the pins may extend either of two heights above the backplane to allow early make/late break capability.

A backplane arranged to accommodate twenty-seven 8-in. high circuit cards is illustrated in Fig. 6. This backplane is approximately 24 in. wide and contains about 3000 compliant pins. A backplane of the same size with a full complement of pins contains 10,800 pins.

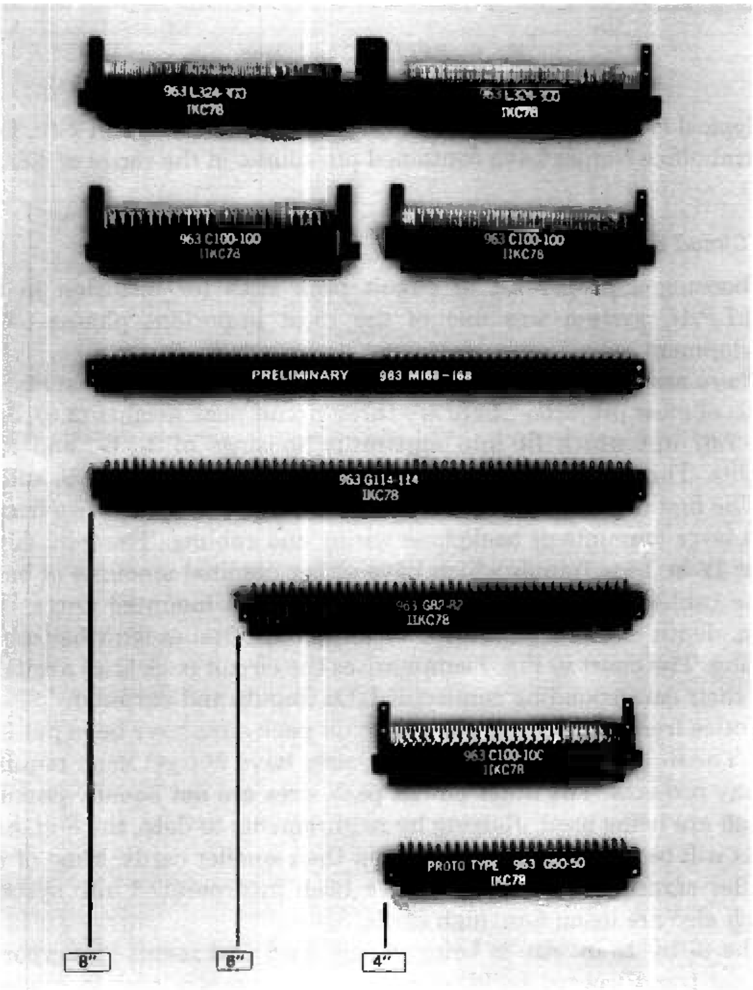


Fig. 5—*BELLPAC* connector family.

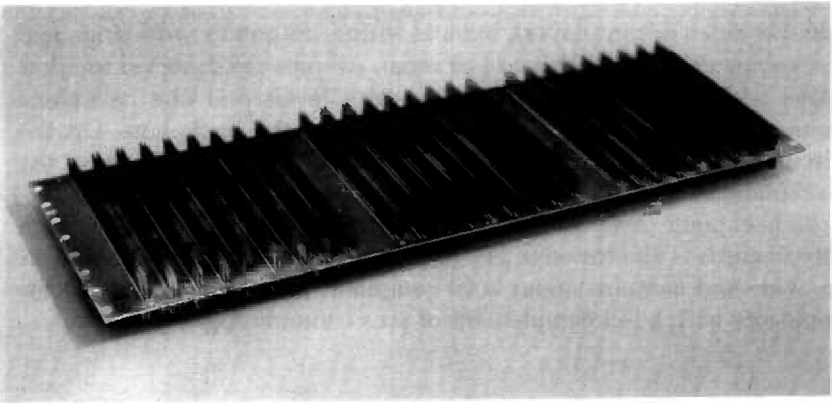


Fig. 6—*BELLPAC* backplane.

Typical 8-in. high backplanes designed to date for use in 2-ft., 2-in. central office frames have contained pin counts in the range of 4000 to 8000.

2.3 Circuit packs

Choosing a proper set of circuit pack sizes for inclusion in the *BELLPAC* system was one of the most important phases of its development.

There are eight circuit pack sizes, a sufficient number to satisfy the needs of most projects. There are three circuit pack heights (3.67, 5.67, and 7.67 in.) which fit into apparatus housings of 4-, 6-, and 8-in. heights. Three nominal circuit pack depths are available: 7, 9, and 13 in. The first depth is tailored for use in 12-in. base central office frames with large amounts of backplane wiring and cabling. The 9-in. depth is for 12-in. base frames which have either nominal amounts of backplane cables or cabling accommodated in front-mounted ducts. The 13-in. depth is provided for use in 18-in. base frames with backplane cabling. The chart in Fig. 7 summarizes the circuit pack sizes available and their corresponding connector I/Os (inputs and outputs).

Notice from Fig. 7 that only six circuit pack sizes have been put into use. The 6- by 7-in. and 6- by 13-in. sizes have not yet been required by any projects. The other circuit pack sizes are not equally popular, but all are being used. Judging by requirements to date, the 8-in. high cards will be much more widely used than smaller cards. Most of the smaller cards being used now have been incorporated into systems which also are using 8-in. high cards.

The 8- by 13-in. size is being widely used and seems to support a general trend toward building much more complex plug-in modules than in past systems.

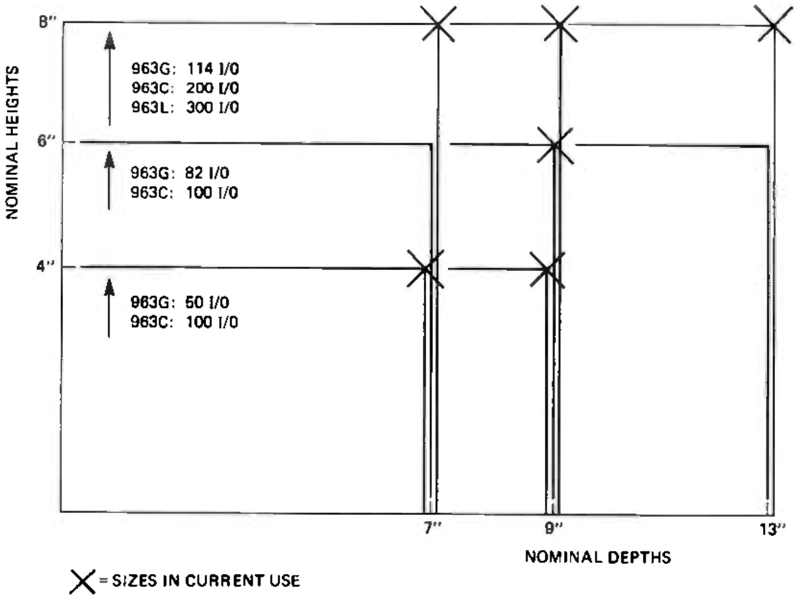


Fig. 7—Circuit pack sizes.

Circuit packs may be fabricated in a variety of technologies. Choice of circuit pack technology is dictated by many factors, including cost, electrical performance, thermal performance, and interconnection density. *BELLPAC* system designs are currently supported in the following circuit pack styles:

(*i*) Double-sided epoxy glass (both conventional and fineline with bus bars).

(*ii*) Double-sided epoxy-coated metal.

(*iii*) 4-layer multilayer board.

(*iv*) 6-layer multilayer board.

(*v*) Wire-wrap.

(*vi*) Quick connect.

The last two board styles are designed for rapid system breadboarding and are available as off-the-shelf parts to be wired by the user.

By way of example, an 8-in. high by 9-in. deep circuit pack is shown in Fig. 8. This double-sided rigid card has a 114-pin connector and low wiring density. By contrast, other cards designed in *BELLPAC* system technology have packaged as many as 150 DIPS (dual in-line packages) on 6-layer, fineline multilayer boards.

2.4 Common features

An extensive set of drawings has been generated to define common features for *BELLPAC* circuit cards. The term "common features"

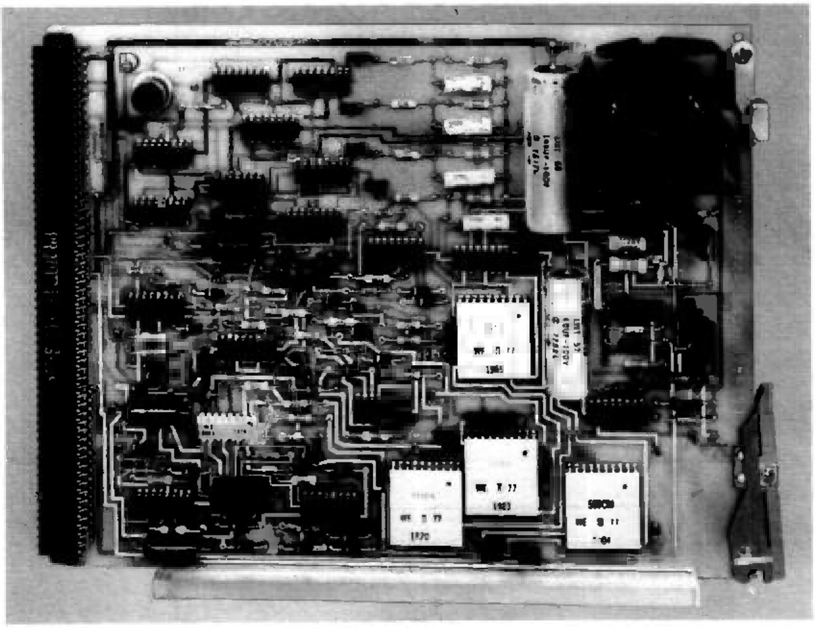


Fig. 8—*BELLPAC* circuit pack example.

refers to the fact that all those circuit pack characteristics common to a particular class or style of *BELLPAC* circuit cards have been identified and defined. There are two types of common features: physical common features and artwork common features. The physical common features define exact board dimensions, tooling holes, void areas for components, connector mounting holes, etc. The artwork common features define connector attachment lands, pattern feature sizes, board layups for multilayer boards, and a very large number of other printed pattern features.

The development and dissemination of common feature information have provided the following advantages:

(*i*) Facilitates standardization and characterization.

(*ii*) Improves routability and manufacturability.

(*iii*) Allows efficient board topology libraries to be built and used.

(*iv*) Reduces design cycles: Less manual input data required. Fewer input errors.

(*v*) Allows control and dissemination of design and manufacturing changes.

III. ELECTRICAL CHARACTERIZATION

The goal of this effort is to provide a complete electrical characterization of the *BELLPAC* components so that the electrical system designer may concentrate on those detailed questions specific to the particular system.

Studies have been made on the electrical properties of *BELLPAC* connectors, cables, and circuit packs. The transmission properties of the connector family were studied and found to be adequate for the circuit pack needs of almost all present Bell System projects; that is, those with signal rise times of 2 ns or greater or with bandwidths of 175 MHz or less. Another study showed the stability of the electrical connection from the circuit pack to the backplane through the connector. Changes of less than 0.80 milliohm were observed over the lifetime of the connector (200 insertions and withdrawals) with worst-case temporal changes (within 30 seconds of insertion) of less than 0.08 milliohm.

The transmission properties of the connector were found to be dependent upon the grounding pattern used. Similarly, proper attention to grounding patterns is important for the proper use of flat cable. In particular, a study showed that stacks of Western Electric-manufactured PVC flat cable, when properly grounded, have sufficiently low pulse crosstalk to allow the replacement of more expensive coax, shielded wires, or Teflon* flat cable.

Studies of pulse transmission properties (characteristic impedance, propagation delay, rise time, and bandwidth) were made earlier in rather general terms for various circuit pack styles. The development of the *BELLPAC* system, with its specified circuit pack styles and common features, enabled this work to be expanded upon and applied directly to the *BELLPAC* system styles.[3] Detailed evaluation of crosstalk properties, which are strongly geometry-dependent, became possible. Table I is adapted from Ref. 3. (Some material is presented in the table on styles not currently supported in the *BELLPAC* system, namely, the bonded board and the 8-layer multilayer board, or MLB.) The reference presents theoretical results and theoretical scaling laws which extend the application of the crosstalk results to arbitrary pulse signals, periodic signals, and random signals. The material has been used for choice of an appropriate circuit pack style, for crosstalk estimation (either manually or for post-routing analysis, using computer-aided design, or CAD), for estimation of conductor capacitance and inductance, and to estimate the effects of proposed new dielectrics or geometries.

Similarly, an earlier study on current-carrying capacity[4] is being applied and extended to encompass all the *BELLPAC* system styles of printed wiring. Once again, the standardization associated with the *BELLPAC* system makes this detailed analysis feasible.

* Trademark of E. I. Du Pont de Nemours Company.

Table I—Summary of the pulse transmission properties of various circuit pack styles

Circuit Pack Style	Characteristic Impedance (ohms)	Propagation Delay per Ft. (ns/ft)	Tr (ns)	Band-width (MHz)	Maximum Interlayer Pulse Crosstalk (Near-End) (percent)	Maximum Intralayer Pulse Crosstalk (Near-End) (percent)	For more details, see the following figures in the appendix to Ref. 3
Wire wrap	125 ± 50 160 ± 35	1.4 1.3	2.0 1.8	250 278	—	40 35	4
Extender board	70 ± 5	1.8	1.3	385	0.3	1.6	5
Double-sided (epoxy)	150 ± 20 150 ± 20	1.5 1.5	2.6 2.6	190 190	21 24	39 34	6
Double-sided (metal)	98 ± 8 83 ± 6	1.5 1.5	2.6 2.6	190 190	1.2 3.2	15 13	7
Bonded board	95 ± 10 85 ± 10	1.5 1.5	2.6 2.6	190 190	38 44	21 19	8
4L MLB (EXT P/G)	95 ± 35 85 ± 30	1.8 1.8	2.5 2.5	200 200	20 21	30 16	9
6L MLB (EXT P/G)	75 ± 30 70 ± 35	1.8 1.8	2.5 2.5	200 200	40 46	32 16	10
6L MLB (INT P/G)	68 ± 3 61 ± 3	1.9 1.9	1.8 1.8	278 278	0.5 0.5	20 15	11
6L MLB (INT P/G, surface routing)	85 ± 25 75 ± 15	1.5 (surface), 1.8 1.5 (surface), 1.8	1.8 1.8	278 278	22 26	16 14	12
8L MLB (INT P/G)	85 ± 25 75 ± 15	1.9 1.9	1.8 1.8	278 278	20 24	18 12	13

IV. COSTING AND PARTITIONING COMPUTER AIDS

The computer-aided design specialists at Bell Laboratories are responsible for ensuring that existing and future Bell Laboratories computer aids to design are readily applicable to the *BELLPAC* system. In addition, the specialists are developing one specific, new, standalone program, a *BELLPAC* system costing and partitioning analysis program.

The goal is to develop an interactive system that will help the physical designer answer some questions which arise during the design process. A large number of parameters must be considered by the designer, including circuit pack parameters (size, type, spacing, and technology) connector mix, hardware costs, design intervals, power dissipation, electrical bandwidth, and many others. An output from one problem solution may well be the input to another. Some problems are quite straightforward, such as obtaining the cost (prototype or production) and the parts list for a specific shelf assembly. Others are more subtle, such as determining an appropriate division of available frame space into various heights or apparatus housings and circuit packs, with appropriate pin-outs per circuit pack, under various physical, thermal, or electrical constraints. It became apparent that one interactive system could efficiently handle many of these questions. In cases where much is known (such as where a parts list is required), the user will enter the known data. In other cases, theoretical relationships will be necessary to produce some of the needed data.

Much effort has been expended to determine the proper environment for this program. The decision has been made to program in C for a *UNIX** system environment with compatibility to other environments being maintained.

V. ASSEMBLY

Traditionally, the physical designer (e.g., Bell Laboratories) has not specified manufacturing or assembly methods, except as they may be implied by the end-product requirements. The production methods are then left up to the manufacturer (e.g., Western Electric). With design of the *BELLPAC* system, the designer has accepted the responsibility of ensuring the availability of workable, efficient, and cost-effective methods of assembly.

This does not infringe on the traditional prerogatives of the manufacturer, since much of the development of assembly equipment and methods is still performed by Western Electric, specifically at the Engineering Research Center. The designer's function is to disseminate

* Trademark of Bell Laboratories.

this information, to help avoid duplication, and to solve particular problems where the designer has particular expertise. These goals are met through leadership or membership on standing committees, through the maintenance of a prototype assembly shop at Bell Laboratories in Whippany and through special studies in critical areas. We discuss each of these topics briefly.

5.1 Committees

There are two ongoing groups of interest here: As part of the *BELLPAC* Forum, mentioned earlier, approximately every three months a group of Bell Laboratories and Western Electric engineers meets to disseminate and discuss the latest manufacturing and assembly developments, problems, and successes. The second group is the Western *BELLPAC* Manufacturing Task Force. This group has Western Electric Department Chief representation from Interconnection Engineering, from Corporate Engineering, and from locations involved in component manufacture, assembly, and assembly tooling development. In addition, there is a Bell Laboratories representative. The group coordinates the initial and on-going manufacturing process utilizing a corporate perspective. It also oversees development activities to avoid duplications or omissions.

5.2 BELLPAC prototype assembly

A prototype assembly shop and laboratory is in operation at Bell Laboratories in Whippany, and is available for alignment, heat-staking, and soldering of connectors to *BELLPAC* circuit packs. This shop is also capable of assembling apparatus housings and backplanes and inserting compliant pins into backplanes. Measurement facilities for end product inspection are available.

Some equipment available is shown in Fig. 9, including (counterclockwise from upper right) alignment of connectors to the circuit pack, heat-staking of the connector, and the end-product inspection for connector to board squareness and for connector tail protrusion below the board. Equipment is either identical to that used in production or similar enough that experience gained can be transferred. For example, at present, one post is heat-staked at a time, while the production equipment stakes all 14 posts simultaneously. However, the staking cycles (time, temperature, and pressure) are sufficiently similar to enable experience gained in one location to be used in another.

5.3 Specific studies

At first glance, the *BELLPAC* system warp requirements for printed wiring products appear to be more stringent than usual. Actually, since

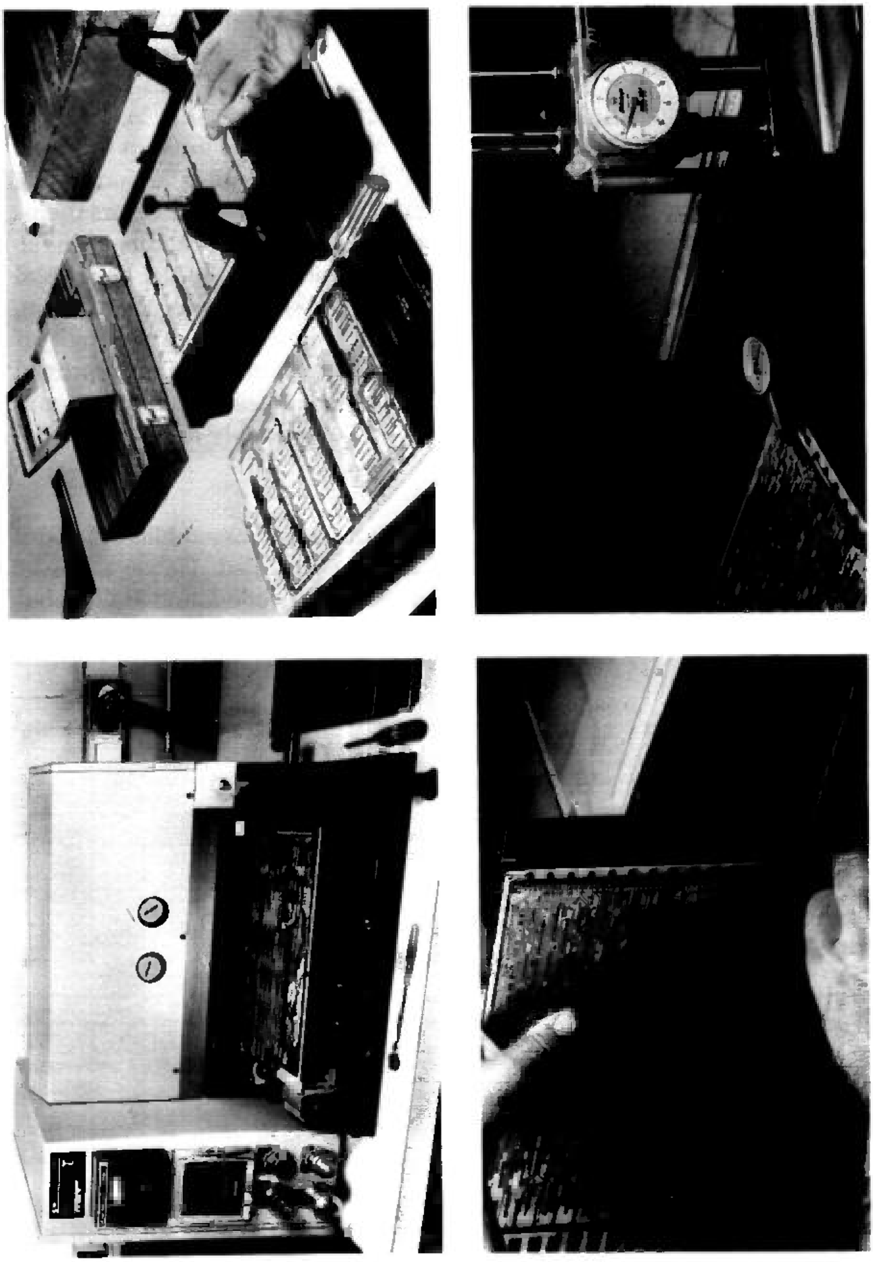


Fig. [illegible]—BELLPAC [illegible]

the requirements are to be measured in a use-related, rather than in a conventional, manner, it is expected they will prove to be a relaxation from the usual (except for printed wiring boards, or PWBs, on 0.5-in. centers). The concept of use-related measurement of warp is illustrated in Fig. 10.

The designers capitalize upon the fact that the *BELLPAC* system physical design is "tight" enough that a considerable amount of board warpage can be tolerated because of the straightening action of the card guides and apparatus housing. Furthermore, the connectors themselves maintain the PWBs straight enough (in the vertical direction) so that, if the board mates properly with the ramp (i.e., in the horizontal direction), it will mate properly with the backplane pins. In addition, detailed studies of proper soldering techniques (e.g., fixturing), of improvements in standard wave-soldering machines and of improved soldering machines have been made. These studies will continue, and others will be initiated as needs arise in the assembly area.

VI. REPAIR

Repair of *BELLPAC* printed wiring boards and assemblies, like any other PWBs or PWB assemblies, is already specified. However, in at least two new areas repair developments are needed, and Bell Laboratories is committed to supplying these needs. These areas are the

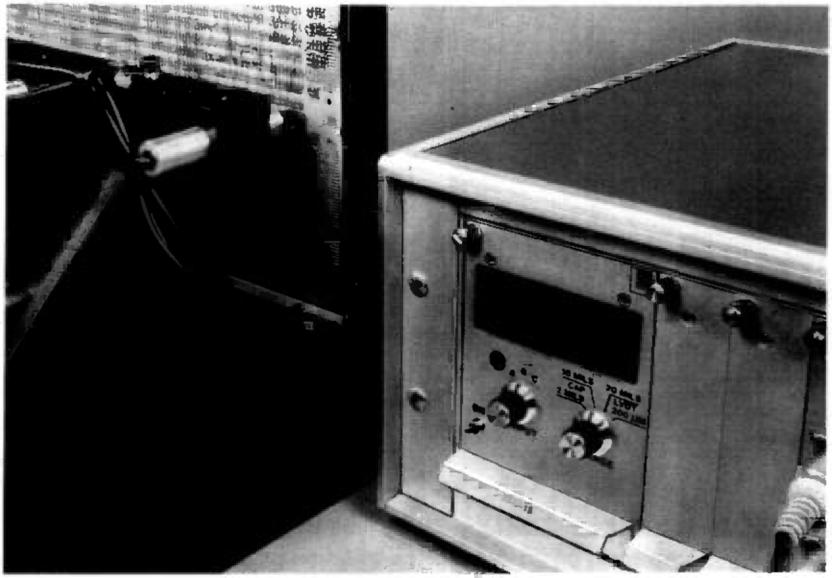


Fig. 10—*BELLPAC* system warp measurements.

repair of connectors and the repairs associated with, or required by, the use of compliant pins press-fit into backplanes.

Repair techniques for connectors are available for individual contacts and for individual contacts of connector assemblies. Satisfactory techniques also exist for the removal of connectors from PWB assemblies.

The use of compliant pin backplane assemblies has eliminated the need for soldering and thus further increased printed backplane reliability. (The reliability of the pin-PWB connection has been established in detail.) This has made possible the removal of defective pins and the insertion of replacement pins by simple procedures, employing hand-held tools.

A new repair need has occurred because of the use of compliant pins. When repairing or modifying multilayer printed wiring boards, one may wish to remove the electrical connection of the backplane pin to the backplane plated through hole (PTH). The present techniques assume that the pin is mechanically secured by a connector housing. Since this is not true for the compliant pin used in the *BELLPAC* system, mechanical retention as well as electrical isolation must be provided by the repair method. Two repair techniques have been developed to meet this need. In the first, the PTH is drilled out and replaced by an isolation bushing, and a standard pin is inserted. In the second, an insulated pin is inserted into the PTH in place of the original pin; this version is ideal where the connections in the MLB do not need to be broken, since the potentially dangerous hole-drilling operation is not needed. Isolation bushings and insulated pins are expected to be used for both repairs and modifications.

Welding procedures are also being investigated to avoid soldering when adding modification or repair wiring to compliant pins. These and other repair techniques will continue to be developed as needed.

VII. PROJECT APPLICATIONS

Currently, over 40 projects use *BELLPAC* hardware. Many more are studying its applicability to their needs. *BELLPAC* hardware was first shipped to an operating company in May, 1978. The first system, called AMARC, uses 8- by 13-in. circuit packs—generally as low-density double-sided boards carrying large numbers of relays and other large components. Even though the component density was low, applications required the use of 200 pin connectors on some cards. The AMARC shelf assemblies mount within relay rack housings for compatibility with minicomputer equipment. We mention these details because, in several ways, this application typifies systems which capitalize on the benefits provided by *BELLPAC* hardware. The system was partitioned optimally by choosing appropriate building block sizes from the range available. Sophisticated interconnection products were readily available, allowing development intervals to be kept short. And low costs were achieved, even though manufacturing volumes were low.

Not all *BELLPAC* system projects are low volume, of course. Two electronic switching projects now in development are being completely packaged with *BELLPAC* hardware and will generate very high manufacturing volumes. These high volume programs and others like them will keep hardware and design costs low for all users.

VIII. SUMMARY

The *BELLPAC* system has become a successful technology for packaging Bell System electronic hardware. It consists of a proven set of components and associated documentation. It has been shown to provide both cost and time savings to system area projects. Its flexibility provides the capability for packaging a variety of different systems in a compatible manner. And, most important, the *BELLPAC* system has the development support from both the systems areas and the component areas to assure that the technology will continue to evolve to meet the packaging needs of the next generation of hardware.

IX. ACKNOWLEDGMENTS

Many people have contributed to the design and implementation of the *BELLPAC* system. The authors particularly wish to acknowledge the efforts of J. G. Brinsfield in formulation of the concept, of H. J. Scagnelli in the successful physical design of the hardware system, and of C. L. Winings in the design and implementation of the connector system.

REFERENCES

1. P. J. Tamburro, "Press-Fit Pins in Printed Circuit Boards—Third, Fourth, Fifth and Sixth Test Series," Tenth Annual Connector Symposium Proceedings, October 19–20, 1977.
2. G. W. Schwindt, "Round Hole/Round Pegs—The Second Generation," Proceedings of the Technical Program, National Electronics Packaging and Production Conference, February 28, March 1, 2, 1978.
3. A. J. Rainal, "Transmission Properties of Various Styles of Printed Wiring Boards," B.S.T.J., *58*, No. 5 (May–June, 1979), pp. 995–1025.
4. A. J. Rainal, "Temperature Rise at a Constriction in a Current Carrying Printed Conductor," B.S.T.J., *55*, No. 2 (February 1976), pp. 233–269.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

Numerical Integration of Stochastic Differential Equations

By E. HELFAND

(Manuscript received June 28, 1979)

A procedure for numerical integration of a stochastic differential equation, by extension of the Runge-Kutta method, is presented. The technique produces results which are statistically correct to a given order in the time step. Second- and third-order approximations are explicitly displayed.

I. INTRODUCTION

Systematic work on numerical solution of stochastic differential equations (SDEs) seems not to have kept pace with the considerable analytical developments. This parallels the lag which existed between the analytical and numerical study of ordinary differential equations near the turn of the century (which is perhaps understandable in view of the difficulty of implementing even straightforward algorithms at the time). In the last few years, there has been a burst of activity in performing Brownian dynamics computer simulations[1] to gain insight into motions in complex physical systems. Little attention seems to have been paid, though, to the systematic development of the numerical techniques in most of these works.

In the present paper, the Runge-Kutta (RK) approximation for deterministic differential equations (DDEs) is extended to SDEs. Although we have not as yet explicitly considered other popular numerical schemes, we feel that the techniques utilized here should have wider applicability. For the sake of simplicity, several further restrictions are placed on the discussions in this paper. These, we believe, can ultimately be removed by fairly simple means.

(*i*) We shall work only with a single equation rather than a set of n equations. It has been explicitly verified that the second-order approximation carries over in a straightforward manner to sets (and in our studies of polymers[2] we used it for 600 simultaneous equations).

However, bear in mind Butcher's demonstration[3] that extra conditions arise with the RK method when one generalizes fifth-order (5_O) and higher schemes from single equations to sets.

(*ii*) We present explicit results only for low-order algorithms, second (2_O) and third (3_O) order, although the principles of higher order extensions will be written down.

(*iii*) Finally, we restrict attention to a simple SDE, which is of the general form occurring in Brownian motion theory. This is

$$dx/dt = f(x) + A(t), \tag{1}$$

where the $A(t)$ are Gaussianly distributed random variables with mean zero and covariance

$$\langle A(t)A(t')\rangle = \xi\delta(t - t') \tag{2}$$

(white noise). The extension to $f(x,t)$ appears to involve little new, but makes the presentation more cumbersome.

In Section II, we review the RK technique for DDEs. After defining more clearly what is meant by numerical solution of an SDE in Section III, we explicitly extend the 2_O RK method to SDEs and outline the generalization to any order. A 3_O RK scheme is presented in the appendix. Section IV is a brief discussion of the question of accuracy. The concluding remarks indicate areas for future studies.

Abbreviations used throughout the paper are listed in Table I.

II. SUMMARY OF THE RK APPROXIMATION FOR DDEs

To set the stage, it will be useful to review[4] briefly the application of the RK technique to the DDE

$$dx/dt = f(x). \tag{3}$$

Of course, this equation can be solved by quadrature, but not when x and f are vectors, or when f is a function of x and t (the RK procedure for the latter case is presented in most standard texts[4] and does not differ greatly from the case we are considering).

Begin by writing down the solution of eq. (3) as a series in the time step s:

$$x(s) = x_o + sf_o + ½\, s^2 f_o f_o' + (⅙) s^3 (f_o f_o'^2 + f_o^2 f_o'') + \cdots, \tag{4}$$

Table I—Summary of abbreviations

SDE	Stochastic differential equation
DDE	Deterministic differential equation
RK	Runge-Kutta
k_O	kth order
l_S	l stages
m_G	m Gaussian random variables per step

where $f_o^{(n)}$ denotes the nth derivative of f evaluated at x_o. The aim of many numerical procedures is to present an algorithm which, when expanded in s, matches the series (4) to a given order, k, in s. Of course, merely evaluating eq. (4) will do that, but a further aim is to avoid the determination of derivatives of f. Thus, in the RK theory one goes from an initial condition x_o to $x(s)$ in l stages by the general procedure

$$g_1 = f(x_o),$$

$$g_2 = f(x_o + \beta_{21} s g_1),$$

$$g_3 = f(x_o + \beta_{31} s g_1 + \beta_{32} s g_2),$$

$$\cdots$$

$$g_l = f(x_o + \beta_{l1} s g_1 + \cdots + \beta_{l,l-1} s g_{l-1}), \tag{5}$$

$$x_s = x_o + s(A_1 g_1 + A_2 g_2 + \cdots + A_l g_l). \tag{6}$$

The ½ $l(l+1)$ parameters $A_1, \cdots, A_l, \beta_{21}, \cdots, \beta_{l,l-1}$ are to be selected so that an expansion of eq. (6) in powers of s matches Eq. (4) through order k. Only for $k \le 4$ can a kth order (k_O) RK calculation be done in k stages (k_S). For $k \ge 5$, a larger number of stages than the order is necessary to provide enough parameters to match the true series.

In the $2_O 2_S$ RK, the parameters must satisfy

$$A_1 + A_2 = 1, \tag{7}$$

$$A_2 \beta_{21} = ½. \tag{8}$$

This illustrates the common occurrence of less equations than parameters. The user then has the freedom to select some parameters (one in the present case) for convenience, or to achieve the smallest error estimates.[5]

III. GENERALIZATION OF RK METHOD TO SDEs

An SDE does not have a definite solution. When we say that we are numerically integrating an SDE, we mean that we are generating a statistically representative trajectory. Furthermore, as in numerical integration of a DDE, we are not going to generate the full trajectory, but only values of x at discrete times: $x(0)$, $x(s_1)$, $x(s_1 + s_2)$, $\cdots$. Let us be more specific. The stochastic process x (or set of processes) specified by eq. (1) is Markovian. Thus, the process is completely specified by the conditional probability density function $p(x, s \mid x_o)$, which gives the probability density of observing x at time s, given the value x_o of the variable at time zero. What we seek is a method of selecting a value x_s with statistics correct to kth order in s. By this, we mean that the moments $\langle x_s^q \rangle$ are all correctly given to $O(s^k)$; i.e., there exists a sequence C_q such that for sufficiently small s

$$|\langle x_s^q \rangle_a - \langle x(s)^q \rangle_e| \leq C_q s^k, \tag{9}$$

for all positive integers q. The average $\langle\ \rangle_a$ is over the ensemble generated by the approximate process, while $\langle\ \rangle_e$ is over the exact process.

An approximation algorithm will involve generation of some random numbers. Naturally, if $p(x,s \mid x_o)$ is known, all that need be done is to generate a single uniformly distributed random number, u, for each step and solve the equation $p(x,s \mid x_o) = u$ for x. We shall see that a 1_O approximation is equivalent to linearizing f (since f'' does not enter until $O(s^2)$). For a linear SDE, $p(x,s \mid x_o)$ is a well-known Gaussian.[6] Use of this Gaussian as an approximate process has been suggested.[7,8] This is practical for a single variable x, but for a large set, the amount of matrix manipulation is overwhelming.

In the RK extension to be discussed, for each step of time s, m independent Gaussianly distributed variables, Z_i (or m sets, $\mathbf{Z}_i$), will be needed. These have

$$\langle Z_i \rangle = 0, \tag{10}$$

$$\langle Z_i Z_j \rangle = \delta_{ij}. \tag{11}$$

An approximation which requires m Z's will be said to be m-fold Gaussian, abbreviated m_G.

Now we shall present a parallel to the RK procedure for SDEs. Again begin by developing a "power series" expansion for the solution of the SDE (1):

$$dx/dt = f(x) + A(t). \tag{1}$$

This may be done by iteration and Taylor series expansion:

$$x(s) = x_o + \int_0^s ds_1 f\left\{x_o + \int_0^{s_1} ds_2 f[x_o + \cdots] + w_o(s_1)\right\} + w_o(s) \tag{12}$$

$$= x_o + sf_o + \frac{1}{2} s^2 f_o f_o' + \frac{1}{6} s^3 (f_o f_o'^2 + f_o^2 f_o'') + \cdots + S, \tag{13}$$

$$S = \{w_o(s)\} + \{f_o' w_1(s)\} + \left\{\frac{1}{2} f_o'' \int_0^s ds_1 w_o^2(s_1)\right\}$$
$$+ \left\{f_o'^2 w_2(s) + f_o f_o''[s w_1(s) - w_2(s)]\right.$$

$$+ \frac{1}{6} f_o''' \int_0^s ds_1 w_o^3(s_1) \Big\}$$

$$+ \Big\{ \frac{1}{2} f_o' f_o'' \int_0^s ds_1 (s - s_1) w_o^2(s_1) + \frac{1}{2} f_o f_o''' \int_0^s ds_1 s_1 w_o^2(s_1)$$

$$+ \frac{1}{24} f_o^{(iv)} \int_0^s ds_1 w_o^4(s_1) \Big\} + \cdots; \tag{14}$$

$$w_n(s) = \int_0^s ds_1 w_{n-1}(s_1), \quad n > 0, \tag{15}$$

$$= \int_0^s ds_1 \frac{(s - s_1)^n}{n!} A(s_1), \quad n \geq 0 \tag{16}$$

($w_o(s)$ is the Wiener process). The term S is a stochastic process. Its various parts, set off in braces, have orders in probability $s^{1/2}$, $s^{3/2}$, s^2, $s^{5/2}$, s^3, $\cdots$ (N.B.: there is no s^1 term). A stochastic variable v will be said to have an order s^k in probability if

$$| (v^q) | \leq K_q s^{qk}, \tag{17}$$

for all positive integers q, a set of constants K_q, and sufficiently small s. The w_n are correlated Gaussian random variables with mean zero and covariances

$$(w_n(s) w_m(s)) = \xi s^{n+m+1} / n! m! (n + m + 1), \tag{18}$$

$$\langle w_o(s) w_o(t) \rangle = \xi \min(s, t).$$

$$\langle w_o(s) w_1(t) \rangle = \begin{cases} \tfrac{1}{2} \xi t^2, & t \leq s, \\ \tfrac{1}{2} \xi s (2t - s), & t \geq s, \end{cases} \tag{19}$$

$$\cdots$$

The statistics of the stochastic part of the trajectory are embodied in the moments of S which, from eqs. (18) to (20), are

$$\langle S \rangle = \tfrac{1}{4} s^2 \xi f_o'' + s^3 \xi (\tfrac{1}{12} f_o' f_o'' + \tfrac{1}{6} f_o f_o''' + \tfrac{1}{24} \xi f_o^{(iv)}) + \cdots, \tag{20}$$

$$\langle S^2 \rangle = s\xi + s^2 \xi f_o' + s^3 \xi (\tfrac{2}{3} f_o'^2 + \tfrac{2}{3} f_o f_o'' + \tfrac{1}{3} \xi f_o''') + \cdots, \tag{21}$$

$$\langle S^3 \rangle = \tfrac{7}{4} s^3 \xi^2 f_o'' + \cdots. \tag{22}$$

For the expansion through $O(s^3)$, the terms of S nonlinear in the w's do not contribute to the moments $\langle S^4 \rangle$ and higher. Thus, these moments are related to the second moment by the usual Gaussian formulas: $\langle S^4 \rangle = 3 \langle S^2 \rangle$, etc.; i.e., the cumulants vanish to $O(s^3)$. The point is that, if $\langle S^2 \rangle$ is properly given, so will $\langle S^k \rangle$, $k \geq 4$, be. The aim of a k_O numerical scheme will be to match not only the nonstochastic terms of the series solution for x, eq. (13), but also to match all the moments of the stochastic term.

We delay the presentation of the general extension of the RK approximation and first explicitly display a $2_O 2_S 1_G$ scheme. Consider the algorithm:

$$g_1 = f(x_o + s^{1/2}\xi^{1/2}\,\lambda_1 Z), \tag{23}$$

$$g_2 = f(x_o + s\beta g_1 + s^{1/2}\xi^{1/2}\,\lambda_2 Z), \tag{24}$$

$$x = x_o + s(A_1 g_1 + A_2 g_2) + s^{1/2}\xi^{1/2}\,\lambda_o Z. \tag{25}$$

Z is a single Gaussian random variable with mean zero and variance unity, generated for each time step s. Using these equations, x can be developed in power of $s^{1/2}$ to $O(s^2)$:

$$x = x_o + (A_1 + A_2)sf_o + A_2\beta s^2 f_o f_o'' + \cdots + \tilde{S}, \tag{26}$$

$$\tilde{S} = \lambda_o Z s^{1/2}\xi^{1/2} + (A_1\lambda_1 + A_2\lambda_2)Z s^{3/2}\xi^{1/2} f_o'$$

$$+ \tfrac{1}{2}(A_1\lambda_1^2 + A_2\lambda_2^2)Z^2 s^2 \xi f_o'' + \cdots . \tag{27}$$

The moments of $\tilde{S}$ through $O(s^2)$ are

$$\langle \tilde{S} \rangle = \tfrac{1}{2}(A_1\lambda_1^2 + A_2\lambda_2^2)s^2\xi f_o'' + \cdots \tag{28}$$

$$\langle \tilde{S}^2 \rangle = \lambda_o^2 s\xi + 2(A_1\lambda_1 + A_2\lambda_2)\lambda_o s^2 \xi f_o' \cdots . \tag{29}$$

To $O(s^2)$, the moments $\langle \tilde{S}^3 \rangle$ and higher involve only the linear terms of $\tilde{S}$, so they are Gaussianly related to $\langle \tilde{S}^2 \rangle$. Matching the deterministic part of eq. (26) to (13), eq. (28) to (20), and eq. (29) to (21), we find as equations for the parameters:

$$A_1 + A_2 = 1, \tag{30}$$

$$A_2\beta = \tfrac{1}{2}, \tag{31}$$

$$\lambda_o^2 = 1, \tag{32}$$

$$(A_1\lambda_1 + A_2\lambda_2)\lambda_o = \tfrac{1}{2}, \tag{33}$$

$$A_1\lambda_1^2 + A_2\lambda_2^2 = \tfrac{1}{2}. \tag{34}$$

The sign of λ_o is immaterial since it multiplies a symmetric random variable. There are five equations and six parameters. A convenient solution set is

$$A_1 = A_2 = \tfrac{1}{2}, \tag{35}$$

$$\beta = 1, \tag{36}$$

$$\lambda_o = 1, \tag{37}$$

and either

$$(\lambda_1 = 0, \lambda_2 = 1), \tag{38a}$$

or

$$(\lambda_1 = 1, \lambda_2 = 0). \tag{38b}$$

With this as background, the general procedure for constructing a $k_O l_S m_G$ approximation should be clear. Consider the m Gaussian random variables as a vector $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_m)$. Also define $l + 1$ vectors of parameters, each of dimension m:

$$\begin{aligned}
\lambda_o &= (\lambda_{o1}, 0, 0, \cdots) \\
\lambda_1 &= (\lambda_{11}, \lambda_{12}, 0, \cdots) \\
&\cdots \\
\lambda_l &= (\lambda_{l1}, \lambda_{l2}, \cdots, \lambda_{lm}).
\end{aligned} \tag{39}$$

The number of scalar λ parameters is $m(l - \tfrac{1}{2} m + \tfrac{3}{2})$. The generalization of the RK algorithm is

$$\begin{aligned}
g_1 &= f(x_o + s^{1/2}\xi^{1/2}\, \lambda_1 \cdot \mathbf{Z}), \\
g_2 &= f(x_o + s\beta_{21} g_1 + s^{1/2}\xi^{1/2}\, \lambda_2 \cdot \mathbf{Z}), \\
&\cdots \\
g_l &= f(x_o + s\beta_{l1} g_1 + \cdots + s\beta_{l,l-1} g_{l-1} + s^{1/2}\xi^{1/2}\, \lambda_l \cdot \mathbf{Z}),
\end{aligned} \tag{40}$$

$$x = x_o + s(A_1 g_1 + \cdots + A_l g_l) + s^{1/2}\xi^{1/2}\, \lambda_o \cdot \mathbf{Z}. \tag{41}$$

The A's and β's are subject to the usual RK equations since, for $\xi = 0$, the DDE is recovered. The equations for the λ's are obtained by expanding eq. (41), in powers of $s^{1/2}$ to order s^k and separating off a stochastic term $\tilde{S}$. Each term of the moments of $\tilde{S}$ has the form of a product of a numerical coefficient, an integral power of s and of ξ, a product of powers of f_o and its derivatives, and a product of the A, β and λ parameters (the λ parameters enter only as dot products of the λ vectors). This term is equated to the term of the exact moments of S with the same powers of s, ξ, f_o, and derivatives of f_o [see eqs. (20) to (22)]. The result is a set of equations for the λ's, and the number of Gaussians must be chosen so that there are a sufficient number of parameters to satisfy these equations. One Gaussian will do for $2_O 2_S$, and two Gaussians for $3_O 3_S$ (see the appendix).

IV. ACCURACY

The accuracy of a numerical scheme for integrating a DDE can be judged on the basis of its ability to determine trajectories for analytically soluble equations. The schemes for SDEs can only be judged on a statistical basis. For example, the probability density, $p(x, t)$, for the random process x defined by eq. (1) satisfies the Fokker-Planck equation

$$\frac{dp}{dt} = -\frac{d}{dx}\left[f(x)p - \frac{1}{2}\xi\frac{d}{dx}p\right]. \tag{42}$$

This has a stationary solution

$$p_o(x) = N(\xi)\exp\left[2F(x)/\xi\right], \tag{43}$$

$$F(x) = \int^{x} f(x')dx', \tag{44}$$

$$1/N(\xi) = \int_{-\infty}^{\infty} \exp\left[2F(x')/\xi\right]dx' \tag{45}$$

(assuming that the density is normalizable). For a stable approximation scheme, the distribution of x will also approach a stationary probability density. One could attempt to test the overall "goodness of fit" of the observed to the theoretical density function.[9] An easier procedure is to assume that eq. (43) holds and to obtain an estimate of ξ, for instance by maximum likelihood estimation.[9] The estimated ξ is then compared with the exact ξ. We have used this technique and have clearly observed how the estimate improves with decreasing step size s. However, no systematic studies have been carried out yet to determine whether the error decreases as s^{k+1}.

In general, one is interested in the complete comparison of the transition probability $p(x, s \mid x_o)$ for the SDE and the numerical scheme. This is embodied in the spectral resolution, for the exact process and the approximation, of $p(x,s \mid x_o)$ regarded as an integral kernel. Here studies performed on exactly soluble systems would be of value.

A question related to accuracy is: How long a trajectory need one run to reduce statistical error in some property to acceptable levels? The answer depends on the time, τ, for decay of correlation of that property. New statistical information is only generated in a time of $O(\tau)$.[10] Therefore, a simulation of total time t will lead to a decrease of error like $(\tau/t)^{1/2}$. Some systems cannot be described in such a clear-cut fashion since they have a spectrum of relaxation times, some of which may be very long. In such cases, there may be an advantage in reinitializing the run to break correlations.

V. DIRECTIONS FOR FURTHER RESEARCH

The specific procedures displayed in this paper are illustrative of the manner in which standard numerical techniques can be extended to stochastic differential equations. There are several general directions in which further research may be aimed.

5.1 More general SDEs

The numerical schemes should be directed toward more general SDEs. The extension to sets of equations has been mentioned. More general forms of SDEs than eq. (1) are

$$dx = f(x,t)dt + \phi(x, t)dw_u(t) \tag{46}$$

or

$$\frac{dx}{dt} = f(x, t, A(t)). \tag{47}$$

Another generalization is that A may be other than Gaussianly distributed. Also, in the physical literature there is increased attention being directed to stochastic integrodifferential equations, representing processes with memory, such as[11,12]

$$\frac{dx}{dt} = f(x) + \int_0^t d\tau K(\tau)x(t-\tau) + A(t), \tag{48}$$

$$(A(t)A(t+\tau)) \propto K(\tau), \tag{49}$$

or more generally,[13]

$$\frac{dx}{dt} = f(x) + \int_0^t d\tau G[\tau, x(t-\tau)] + A, \tag{50}$$

with A and G related by a generalized fluctuation-dissipation theorem.

5.2 Other numerical schemes

It would be interesting to develop stochastic versions of other numerical schemes used for DDEs. One may raise the objection to any multistep procedure that it does violence to the Markovian nature of the process. One would have to reuse the random variables, Z_i, for several steps to eliminate the spurious memory to the desired order.

5.3 General principles

There are many matters, which are the standard fare of the deterministic numerical analyst, that should be placed in a stochastic context. The question of accuracy has been raised. Another is stability. A third question is that of step-by-step error estimation. An interesting problem arises in developing the analog of step-size adjustment and the criteria for when it is necessary. Imagine that such criteria exist and a particularly large Z triggers the call for step-size adjustment. The new Z's that are generated should not be independent of the old Z's.

Finally, as a general problem, the matter of computational speed should be considered. To gather statistical data, long trajectories must be run, sometimes on systems of many degrees of freedom. It is urgent that there be an analysis of various procedures with respect to their relative speeds, for a given accuracy.

APPENDIX

$3_O 3_S 2_G$ Procedure

To carry out a 3_O procedure requires three stages and two Gaussian random variables. The explicit algorithm is eqs. (40) and (41) with $l = 3$. The parameters must satisfy the equations

$$A_1 + A_2 + A_3 = 1, \tag{51}$$

$$A_2\beta_{21} + A_3(\beta_{31} + \beta_{32}) = 1/2, \tag{52}$$

$$A_2\beta_{21}^2 + A_3(\beta_{31} + \beta_{32})^2 = 1/3, \tag{53}$$

$$A_3\beta_{32}\beta_{21} = 1/6, \tag{54}$$

$$\lambda_{o1} = 1, \tag{55}$$

$$A_1\lambda_{11} + A_2\lambda_{21} + A_3\lambda_{31} = 1/2, \tag{56}$$

$$A_1|\lambda_1|^2 + A_2|\lambda_2|^2 + A_3|\lambda_3|^2 = 1/2, \tag{57}$$

$$A_1\lambda_{11}^2 + A_2\lambda_{21}^2 + A_3\lambda_{31}^2 = 1/3, \tag{58}$$

$$A_1|\lambda_1|^2\lambda_{11} + A_2|\lambda_2|^2\lambda_{21} + A_3|\lambda_3|^2\lambda_{31} = 1/4, \tag{59}$$

$$A_2\beta_{21}\lambda_{21} + A_3(\beta_{31} + B_{32})\lambda_{31} = 1/4 \tag{60}$$

$$\begin{aligned} &|A_1\lambda_1 + A_2\lambda_2 + A_3\lambda_3|^2 + 2(A_2\beta_{21}\lambda_{11} \\ &\quad + A_3\beta_{31}\lambda_{11} + A_3\beta_{32}\lambda_{21}) = 2/3. \end{aligned} \tag{61}$$

The first four equations are the usual ones for a 3_O RK approximation. They leave two degrees of freedom. A widely used solution is

$$A_1 = 2/9, \quad A_2 = 3/9, \quad A_3 = 4/9; \tag{62}$$

$$\beta_{21} = 1/2, \quad \beta_{31} = 0, \quad \beta_{32} = 3/4. \tag{63}$$

With this set, the remaining seven equations can be solved for the λ parameters. The solution is

$$\lambda_{o1} = 1, \quad \lambda_{11} = 0, \quad \lambda_{21} = \tfrac{1}{2}, \quad \lambda_{31} = \tfrac{3}{4}. \tag{64}$$

There are four solutions for the $\lambda_{l2's}$, two of which are complex. The real solutions are either

$$\begin{aligned} \lambda_{12} &= 0.245538, \\ \lambda_{22} &= -0.023225, \\ \lambda_{32} &= 0.544169, \end{aligned} \tag{65}$$

or

$$\begin{aligned} \lambda_{12} &= -1.34583, \\ \lambda_{22} &= 1.24987, \\ \lambda_{32} &= 0.385032. \end{aligned} \tag{66}$$

Solution (65) is probably superior because it uses less of the Z_2 process. (All the λ_{l1} and/or all the λ_{l2} may be reversed in sign as an acceptable solution, as is evident since they multiply symmetrically distributed random numbers.)

REFERENCES

1. Workshop on Stochastic Problems in Molecular Dynamics and Macromolecular Dynamics, Centre Européen de Calcul Atomique et Moleculaire, Orsay, 1978; Workshop on Stochastic Molecular Dynamics, National Resource for Computation in Chemistry, Woods Hole, 1979.
2. E. Helfand, Z. R. Wasserman, and T. A. Weber, J. Chem. Phys., *70* (1979), p. 2016.
3. J. C. Butcher, J. Australian Math. Soc., *3* (1963), p. 202.
4. A. H. Stroud, *Numerical Quadrature and Solution of Ordinary Differential Equations*, New York: Springer-Verlag, 1974.
5. T. E. Hull, W. H. Enright, B. M. Fellen, and A. E. Sedgwick, SIAM J. Numer. Anal., *9* (1972), p. 603.
6. M. C. Wang and G. E. Uhlenbeck, Rev. Modern Phys., *17* (1945), p. 323.
7. J. D. Doll and D. R. Dion, Chem Phys. Lett., *74* (1975), p. 386; cross correlations between B_x and B_u appear to have been neglected in this paper.
8. E. Helfand, J. Chem. Phys., *69* (1978), p. 1010.
9. N. R. Mann, R. E. Schafer, and N. D. Singpurwalla, *Methods for Statistical Analysis of Reliability and Life Data*, New York: John Wiley, 1974.
10. T. R. Koehler and P. A. Lee, J. Comp. Phys., *22* (1976), p. 319.
11. R. Zwanzig, in *Statistical Mechanics*, ed. by S. A. Rice, K. F. Freed, and J. C. Light, Chicago: U. of Chicago Press, 1972.
12. M. Shugard, J. C. Tully, and A. Nitzan, J. Chem. Phys., *66* (1977), p. 2534; A. Nitzan, M. Shugard, and J. C. Tully, *ibid*, *69* (1978), p. 2525.
13. H. Mori, H. Fujisaka, and T. Shigematsu, Prog. Theor. Phys. (Kyoto), *51* (1974), p. 109.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

Coefficient Inaccuracy in Transversal Filtering

By A. GERSHO, B. GOPINATH, and A. M. ODLYZKO

(Manuscript received August 20, 1978)

Coefficient inaccuracy in transversal filters is known to degrade the frequency response, particularly in stopband regions. The magnitude of the stopband degradation increases with the number of stages n, the length of the impulse response. A widely used formula for the error in frequency response is proportional to $\sqrt{n}$. By extending recent results on random trigonometric polynomials, we show that for random additive coefficient errors with variance σ^2, the error $\Delta H(\omega)$ in frequency response for large n is such that

$$\max_{\omega} |\Delta H(\omega)| \simeq \sigma\sqrt{n \log n}$$

where log denotes the natural logarithm. This result leads to an absolute bound on attainable stopband rejection for any band-select transversal filter with given coefficient inaccuracy. In particular, the result places a definite limitation on the quality of band-select filtering that can be achieved using a CCD split-electrode filter. It also implies bounds for the peak sidelobes of random radar arrays.

I. INTRODUCTION

In recent years, the transversal filter has emerged as an essential signal-processing structure for a large variety of applications in communication systems. A few of these applications are matched filtering in radar or spread-spectrum systems, equalization in data receivers, echo cancellation for satellite communications, and band-select digital filters. The term "transversal filter" originally referred to the continuous-time tapped delay line structure where an output is formed from a weighted sum of the tap voltages. The same basic function has also been achieved using lumped networks to approximate the delay sections. More recently, transversal filters have been realized with digital circuitry using shift registers and digital multipliers, operating on a sampled and quantized input signal. The most recent development is the emergence of two new technologies, charge-coupled devices (CCDs) and surface acoustic wave (SAW) devices which allow the realization of discrete-time transversal filters without the need for analog-to-digital conversion.

The new technological advances now offer the possibility of realizing transversal filters with hundreds and perhaps even thousands of tap-weight stages on a single integrated-circuit chip. These developments suggest that extremely sophisticated signal-processing functions can readily be obtained. Specifically, with a sufficient number of taps, a transversal filter can be designed to approximate virtually any specified frequency response as closely as desired. However, the inevitable inaccuracies in implementing the desired weighting coefficients result in a departure of the actual frequency response from the predesigned frequency response which increases with the number of tap-weight stages. In digital filtering, coefficient values can be made as accurate as needed, but at the price of increasing hardware costs. With the CCD or SAW technologies, there are fundamental limits on attainable accuracy. Also, in adaptive filtering, the weight-adjustment algorithm results in a steady-state coefficient inaccuracy. It is therefore necessary to have a quantitative knowledge of the degradation in performance of the transversal filter as a function of the coefficient inaccuracy and the number of stages.

For most applications, the appropriate performance measure for the realized transversal filter is the maximum deviation in frequency response magnitude from the desired values over the particular frequency band of interest. In this paper, we focus on this performance measure by examining the *error-frequency response* due to coefficient inaccuracy and show that under very general conditions the maximum magnitude is given asymptotically by $\sigma\sqrt{n \log n}$, where n is the number of stages, σ is the rms coefficient inaccuracy, and log denotes the natural logarithm. Several other closely related results and implications are also presented.

Since the attainable quality of a designed filter increases with n, the number of stages, and for a given coefficient inaccuracy the degradation increases with n, the question arises: Is it possible to realize a filter with arbitrarily high quality in spite of a given coefficient inaccuracy if n is made sufficiently large? We make this question more precise later and show that the answer is negative for low-pass filtering with a transversal filter structure when "quality" is measured by the amount of stopband rejection. In other words, a limit on filter accuracy implies a limit on attainable filter quality regardless of the number of stages used. The results of this paper provide a tool for determining the ultimate limitation on transversal filter performance associated with a particular technology or a particular adaptive algorithm for weight adjustment.

In CCD transversal filters, the split-electrode method requires that the tap weights be scaled so that the maximum magnitude of the coefficient values is unity. The pattern generator used in making the photomasks for CCD fabrication introduces a quantization error whose peak size is a fixed fraction of the maximum coefficient magnitude. Now, for most applications, increasing the number of stages to be realized corresponds to including additional coefficient values representing the tail of the desired impulse response. Consequently, increasing n does not alter the scaling of the coefficient values for CCD implementation. As a result, a coefficient error can indeed be modeled as an additive random variable whose variance does not depend on the desired coefficient value.

A problem that is very similar to that considered above occurs in the theory of random arrays.[1] These are arrays consisting of fewer elements than conventional phased arrays, with the locations of the elements in the array picked randomly. Such arrays are less costly than conventional phased arrays, but this advantage is gained at the cost of increasing the peak sidelobes. Our main result shows how big those sidelobes can be expected to become.

II. PROBLEM FORMULATION

Regardless of the particular application, the transversal filter may be described by its frequency response, $H(\omega)$, which has the general form

$$H(\omega) = \sum_{k=0}^{L-1} \alpha_k e^{jk\omega}, \tag{1}$$

where ω is the normalized frequency variable, L is the number of stages, $j = \sqrt{-1}$, and the coefficients α_k are real-valued numbers specified by the designer. Since $H(\omega)$ is periodic, only the frequency interval $0 \le \omega \le 2\pi$ need be considered.

We note, in passing, that (1) also describes the discrete Fourier transform, so that the results of this paper are also applicable to studying the effect of approximate representations of given data values on the Fourier transform of the data.

A special case of transversal filters, of particular interest in band-select filter design, arises when the coefficients are chosen to have the symmetry property:

$$\alpha_k = \alpha_{L-k-1} \qquad \text{for} \qquad 0 \le k \le (L-1). \tag{2}$$

When L is odd, this condition results in a linear phase transfer function having the form

$$H(\omega) = e^{+j\omega n} \sum_{k=0}^{n} b_k \cos k\omega, \tag{3}$$

with $n = (L - 1)/2$, and

$$b_k = 2\,\alpha_{n-k} \qquad \text{for} \qquad k \neq 0, \quad b_0 = \alpha_n.$$

Implementation of the coefficients α_k for the general form (1) or b_k for the linear phase form (2) inevitably results in the introduction of errors or inaccuracies. We denote the actual (inaccurate) value realized as α'_k, or as b'_k for the linear phase case. Then the kth coefficient error is the difference $\epsilon_k = \alpha'_k - \alpha_k$, or $\epsilon_k = b'_k - b_k$ in the linear phase case. The realized transfer function then differs from the desired transfer function by the *error transfer function* defined as

$$f_L(\omega) = \sum_{k=0}^{L-1} \epsilon_k e^{jk\omega} \tag{4}$$

in the general case or, in the linear phase case:

$$g_n(\omega) = e^{+j\omega n} \sum_{k=0}^{n} \epsilon_k \cos k\omega. \tag{5}$$

Clearly, the error transfer function, if known, provides a full description of the degradation in performance of the realized filter from the desired performance in the absence of inaccuracies.

Since the errors, ϵ_k, are generally not known prior to fabrication of the filter, they are modeled most effectively as random variables whose distribution depends on the particular mechanism involved in fabricating the tap weights. In digital filtering, the errors are due to coefficient quantization and are usually modeled as uniformly distributed random variables. The error terms for different coefficients, being independently produced, can reasonably be assumed to be independent random variables.

Additive error components were used by Knowles and Olcayto[2] for modeling coefficient quantization in recursive filters. Chan and Rabiner[3] applied this approach for transversal filters and evaluated the rms values of $f_L(\omega)$ and $g_n(\omega)$ at a particular frequency. They assumed mutually independent and uniformly distributed errors ϵ_k resulting in rms values for the error transfer function proportional to $\sqrt{L}$, or $\sqrt{n}$ in the linear phase case. By taking the maximum over all frequencies of the rms deviation, a frequency-independent upper bound on the error transfer function is obtained which is valid at any particular frequency with high probability.

More recently, Heute[4,5] noted that the bounds of Chan and Rabiner underestimate the degradation due to the maximum of $|g_n(\omega)|$ over the frequency band. It is this latter measure of degradation that is meaningful in most applications. Chan and Rabiner's bound is not a high probability upper bound for the maximum ripple magnitude taken on by $g_n(\omega)$. Heute proposed a heuristic upper bound for the maximum of $|g_n(\omega)|$ which has the form $Q[a + bn + (cn + d)^{1/2}]$, where a, b, c, and d are constants and Q is the peak amplitude of the uniformly distributed error terms ϵ_k. His bound gave an improved fit to simulated data for values of n up to 128. We shall see later that Heute's bound, which for large n grows linearly with number of stages n, grossly overestimates the degradation as n becomes much larger than 100.

Andrisano and Calandrino[6] assumed that the error transfer function is a Gaussian process and found an (implicit) bound on stopband rejection as the solution of a transcendental equation.

In this paper, we take as the measure of degradation due to coefficient inaccuracy,

$$D_L = \max_{\omega \epsilon \Omega} |f_L(\omega)| \tag{6}$$

for the general transversal filter and

$$M_n = \max_{\omega \epsilon \Omega} |g_n(\omega)| \tag{7}$$

for the linear phase transversal filter, where Ω is a particular frequency band of interest. We assume the errors ϵ_k are mutually independent random variables with a common distribution satisfying certain regularity conditions that include the uniform and normal distributions as special cases.

We establish here for the first time that the maximum frequency response errors D_n and M_n are asymptotically (for large n) given by $\sigma\sqrt{n \log n}$ where σ is the rms coefficient error. Although the result is asymptotic, Lawrence and Salazar[7] found that it was moderately accurate in one study of a low-pass filter with only 33 taps. Application of the result to low-pass filter performance is examined briefly in this paper and more extensively in Ref. 8 and 9. Until the report of our result,[8] the correct behavior of the error frequency response magnitude had apparently not been recognized in the digital filtering literature.

The existing mathematical results most closely related to our work are due to Halasz[10] who considered random trigonometric sums with coefficients that take on the values ±1 with equal probability. While too restrictive to apply to transversal filters, his methodology was useful in deriving our more general upper bound on the maximum error frequency response.

Our main result is also applicable to the analysis of random arrays, and in particular to that of statistical arrays.[1] These are arrays consisting of k isotropic radiators placed among n positions $(n > k)$ that are spaced $\lambda/2$ apart $(\lambda = \text{wavelength})$, with the k positions to be occupied by the k elements determined at random. The array factor of such an array is defined as

$$f(u) = \sum_{r=0}^{n-1} g_r \, e^{\pi jru}, \tag{8}$$

where $g_r = 1$ if the rth position is occupied by a radiator, and $g_r = 0$ otherwise. This can be rewritten as

$$f(u) = \frac{k}{n} \sum_{r=0}^{n-1} e^{\pi jru} + \sum_{r=0}^{n-1} \epsilon_r \, e^{\pi jru}, \tag{9}$$

where $\epsilon_r = 1 - k/n$ for the k values of r for which $g_r = 1$, and $\epsilon_r = -k/n$ otherwise. The first sum above represents (except for the k/n multiplier) the array factor of a conventional phased array. The random choice of the positions for the radiators corresponds to letting the ϵ_r be independent random variables, assuming the value $1 - k/n$ with probability k/n, and the value $-k/n$ with probability $1 - k/n$. If we assume that $k \sim \alpha n$ as $n \to \infty$, then our theorem shows that, with the probability approaching 1 as $n \to \infty$, the second sum in (9) will never be significantly larger than $\sqrt{1-\alpha} \sqrt{n \log n}$ and that, conversely, it will get that large on any subinterval. This result, which had been derived only heuristically before,[1] explains why random arrays are usually not very satisfactory.

III. STATEMENT OF MATHEMATICAL RESULTS

As we saw in Section II, the errors in the realized transfer functions are given by $\sum_{k=0}^{L-1} \epsilon_k \, e^{jk\omega}$, or $\sum_{k=0}^{n} \epsilon_k \cos k\omega$, or $\sum_{k=0}^{n} \epsilon_k \sin k\omega$. Hence the distribution of the random variables ϵ_k will depend on the model for the sources of inaccuracies. For digital implementation, the usual model assumes ϵ_k to be independently distributed uniformly between $-\Delta$, $+\Delta$, where Δ is the maximum error due to truncation of the coefficients of the filter. In other situations, a Gaussian distribution may be more appropriate. But, as we shall see, the asymptotic behavior of the maximum magnitude of the error is not dependent on the exact nature of the distribution. It depends only on a few functionals of the distribution.

The results presented here rely on an important assumption about the distribution of the ϵ_k. We assume throughout that the ϵ_k have mean zero and finite sixth moment, so that the characteristic function $E(e^{jx\epsilon_k})$ of ϵ_k is such that

$$E(e^{jx\epsilon_k}) = \exp\left[-\sum_{r=2}^{5} a_r x^r + O\,(x^6) \right] \tag{10}$$

for x in some nontrivial interval $[-d, d]$. (Note that $a_2 > 0$ if the ϵ_k are not identically zero.) Condition (10) is satisfied for most probability density functions of practical interest.

Now we are ready to state the main result:

Theorem: *Let* ϵ_k, $k = 1, 2, \cdots$ *be a sequence of independent identically distributed random variables satisfying* (10).

Then there exist constants C_1 *and* C_2*, not depending on* n*, such that*

$$\max_{0\le\theta\le 2\pi} \left| \sum_{k=1}^{n} \epsilon_k \, e^{kj\theta} \right| \le \sqrt{2a_2}\,\sqrt{n \log n} + C_1 \sqrt{\frac{n}{\log n}} \log \log n$$

holds with probability $\ge 1 - C_2 (\log n)^{-4}$. *Furthermore, if* Ω *is any subinterval of* $[0, 2\pi]$ *of length* $\ge (\log n)^{-1}$ *and* α *is any real number, then*

$$\max_{\theta\epsilon\Omega} Re \left\{ e^{j\alpha} \sum_{k=1}^{n} \epsilon_k \, e^{kj\theta} \right\} \ge \sqrt{2a_2}\,\sqrt{n \log n} - C_1 \sqrt{\frac{n}{\log n}} \log \log n$$

holds with probability $\ge 1 - C_2 (\log n)^{-4}$.

Thus, with high probability, max $|f(\theta)|$ is about $\sqrt{2a_2}\,\sqrt{n \log n}$. The proof is outlined in Section V.

Remark 1. By choosing α appropriately, we can conclude that each of $\sum \epsilon_k \cos(k\theta)$, $\sum \epsilon_k \sin(k\theta)$ becomes large on any long θ interval with high probability.

Remark 2: The estimates presented here are not the best possible ones. For example, the interval Ω in the lower bound proof can be of size $n (\log n)^{-\lambda}$ for any $\lambda > 0$.

IV. APPLICATION TO LOW-PASS FILTERS

The usual specifications for FIR low-pass filters are shown in Fig. 1.[11] A design problem is to find the smallest n such that

$$\left| \sum_{k=0}^{n-1} \alpha_k \cos k\theta \right|$$

lies between $1 - \delta_1$ and $1 + \delta_1$ in the passband, i.e., for $\theta \epsilon [0, F_p]$ and between 0 and δ_2 in the stopband, i.e., for $\theta \epsilon [F_s, \pi]$. Estimates for n given δ_1, δ_2, F_p, F_s are given in Ref. 12. However, the validity of the estimates in Ref. 12 for regions of practical interest is not proven. An empirical relationship is given in Ref. 11.

As n increases, smaller δ_2, δ_1 and $F_s - F_p$ are possible. Hence, a question that is usually raised is: Given that the α_k's cannot be realized exactly, what can be said about the minimum δ_2 possible if the distribution of ϵ_k, the error in α_k, is known. If α_k's could be realized exactly, arbitrarily small values of δ_2 can be obtained by making n large. However, ϵ_k's introduce errors that grow with n as seen from the theorem. So there is a trade-off between errors introduced by inaccur-

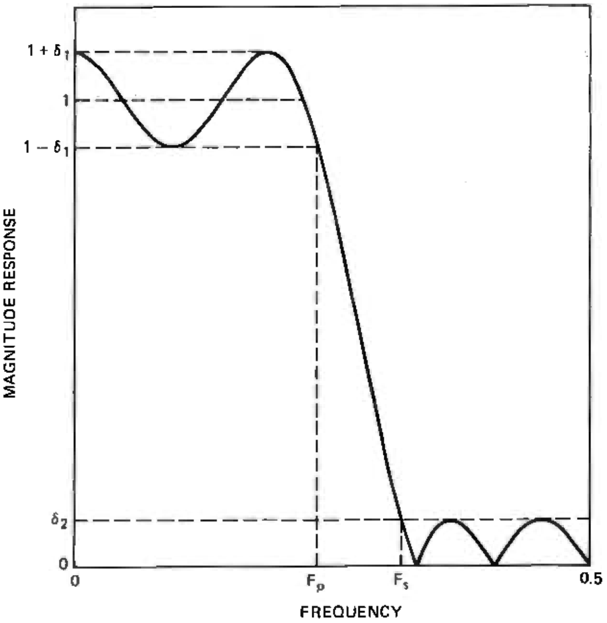


Fig. 1—Frequency as fraction of 2 π.

acies in α_k's and the improvement in performance with increasing n. For an example, we consider the stopband rejection, $20 \log_{10} \delta_2$, as a figure of merit with given values of δ_1 and F_p, F_s. The empirical formula gives the following relation for $\hat{n}$, the minimum n required to achieve a stopband rejection of $20 \log_{10} \delta_2$.

$$\hat{n} = c_1 \log \delta_2 + c_2, \tag{11}$$

where c_1 and c_2 are constants depending on δ_1 and $F_s - F_p$.[11] For fixed point digital implementations, if the coefficients of the filter, the α_k's, are truncated to d bits, then the "error" in α_k is generally modeled as a uniform random variable ϵ_k having values between -2^{-d} and $2^{-d} = \Delta$. For this model, a_2 of the theorem is $1/6\ \Delta^2$. Hence, the maximum error e_n due to these inaccuracies,

$$e_n = \max_{0 \le \theta \le 2\pi} \left| \sum_{k=0}^{n-1} \epsilon_k \cos k\theta \right|, \tag{12}$$

is such that

$$\frac{e_n}{\sqrt{n \log n}} \rightarrow \sqrt{2a_2} = \sqrt{\frac{\Delta^2}{3}} = \frac{\Delta}{\sqrt{3}} \quad \text{as} \quad n \rightarrow \infty \tag{13}$$

and

$$\left| e_n - \frac{\Delta}{\sqrt{3}} \sqrt{n \log n} \right| < c \sqrt{\frac{n}{\log n}} \log \log n \tag{14}$$

with probability $\geq 1 - O((\log n)^{-4})$.

Using the limit (13) to indicate expected deterioration in performance, we can arrive at a design rule. If coefficients are truncated to d bits, then the minimum achievable δ_2 before the random errors become comparable to δ_2 itself is given by:

$$\frac{2^{-d}}{\sqrt{3}} \sqrt{(c_1 \log \delta_2 + c_2) \log (c_1 \log \delta_2 + c_2)} = \delta_2. \tag{15}$$

Putting $\delta_2 = 2^{-s}$,

$$\frac{2^{-d}}{\sqrt{3}} = \frac{2^{-s}}{\sqrt{(-c_1' s + c_2) \log (-c_1' s + c_2)}}, \tag{16}$$

where $c_1' = c_1 \log 2$.

From the above formula, we can estimate the required precision for the coefficients for a given value of $\delta_2 = 2^{-s}$.

In design of CCD filters, a similar formula can be used. In situations where the tap-weight errors can be modeled by a Gaussian random variable with a standard deviation Δ, then a_2 for our theorem is $\Delta^2/2$. The minimum achievable δ_2 satisfies

$$\sqrt{(c_1 \log \delta_2 + c_2) \log (c_1 \log \delta_2 + c_2)} = \frac{\delta_2}{\Delta}. \tag{17}$$

Solving for δ_2, we can estimate the optimum value of n.

As an illustration of the effect of coefficient inaccuracy on limiting the stopband rejection of a low-pass filter, Fig. 2 shows how the best achievable rejection depends on the number of stages, n, for various values of the transition width $\Delta F = F_S - F_P$. These curves were calculated by solving the empirical formula of Ref. 11 for δ_2 and adding to it the maximum error $\sigma\sqrt{n} \log n$. This gives an expression for the best attainable stopband rejection in the presence of coefficient errors, as a function of n, δ_1, and ΔF. For additional curves obtained in this way, see Ref. 8. Computation also shows that varying the allowed passband ripple δ_1 has a negligible effect on the maximum attainable stopband rejection. It is evident that coefficient inaccuracy places an ultimate limitation on the attainable quality of a low-pass filter implemented with the transversal structure.

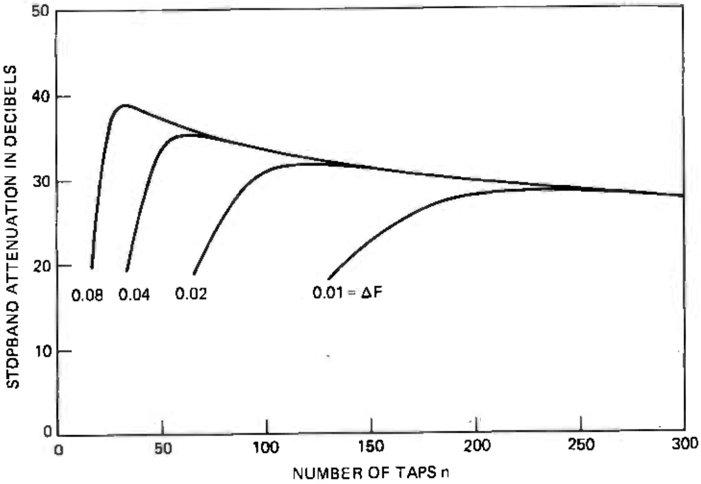


Fig. 2—Best obtainable stopband attenuation for a low-pass transversal filter in the presence of coefficient inaccuracies. Root-mean-square coefficient error = 0.001, passband ripple = 0.0122. Curves are shown for four values of the transition width. Note that, for each curve, an optimum number of taps exist. Reducing the passband ripple allowable has the effect of shifting these curves to the right while reducing the peak value of each curve.

V. ACKNOWLEDGMENTS

We wish to thank Prof. W. Schuessler who brought to our attention the work of Heute and the inadequacy of the $\sqrt{n}$ bound. Prof. P. Erdös kindly told us about the work of Halasz.[10] We also benefited from discussions with L. A. Shepp.

APPENDIX

Here we outline the main steps in the proof, which follows that of Halasz,[10] in which he assumed ϵ_k to be ±1. (An earlier proof of a slightly weaker result had been outlined by Whittle.[13]) Results that are incidental to the main line of reasoning are collected at the end of this outline. Let

$$f(\alpha, \theta) = \sum_{k=1}^{n} \epsilon_k \cos(k\theta + \alpha) = \mathrm{Re}\left(e^{j\alpha} \sum_{k=1}^{n} \epsilon_k e^{jk\theta}\right).$$

(i) We construct a nonnegative function $u(x) \le 1$ which can be used to indicate in an approximate sense the set of values of x that exceed given values. Let M_1, M_2, $D > 0$ be given numbers. Then $u(x)$ is zero for $-M_2 \le x \le M_1$, and $u(x) = 1$ for $x \ge M_1 + D$ or $x \le -M_2 - D$. In the interval $[-M_2 - D, -M_2]$ and $[M_1, M_1 + D]$, $u(x)$ is 40 times differentiable and $u^{(r)}(x) = O(D^{-r})$ as $D \uparrow \infty$, for $0 \le r \le 40$.

For deriving the upper bound, we proceed as follows:

(*ii*) Put $M_1 = M_2 = M = \sqrt{2a_2}\sqrt{n \log n} + gD \log \log n$ where $D = \sqrt{n/\log n}$ and $g = 20\sqrt{a_2/2}$.

(*iii*) Let $v_1(t) = 1/2\pi \int_{-\infty}^{\infty} (1 - u(x))e^{-jtx}\, dx$. Then $|v_1(t)| = O(M)$ and $|t^r v_1(t)| = O(D^{-r+1})$, $1 \leq r \leq 40$.

(*iv*) Let $G = \int_0^{2\pi} d\alpha \int_0^{2\pi} d\theta\, u(f(\alpha, \theta))$ and $v(t) = \delta(t) - v_1(t)$, where $\delta(t)$ is the Dirac delta function. Then

$$G = \int_0^{2\pi} d\alpha \int_0^{2\pi} d\theta \int_{-\infty}^{\infty} \exp\left[\mathrm{j}t \sum_1^n \epsilon_k \cos(k\theta + \alpha)\right] v(t)dt.$$

Using the properties of v_1, we can then show (for details, see the section at the end of this proof) that

$$E(G) = O(n^{-1}(\log n)^{-10}).$$

(*v*) Let $T = \max_{\alpha,\theta} |f(\alpha, \theta)|$. Then using the inequalities

$$\left|\frac{\partial}{\partial\theta} f(\alpha, \theta)\right| \leq Tn, \qquad \left|\frac{\partial}{\partial\alpha} f(\alpha, \theta)\right| \leq T$$

we can show that $G \leq 1/(n \log^2 n) \Rightarrow T \leq M + 2D$ for large enough n.

(*vi*) The result from step (*iv*) implies

$$Pr\left\{G \geq \frac{1}{n \log^2 n}\right\} = O\left(\frac{1}{(\log n)^8}\right).$$

Hence, using step (*v*)

$$\max_{\alpha,\theta} |f(\alpha, \theta)| \leq \sqrt{2a_2}\sqrt{n \log n} + (g + 2)\sqrt{\frac{n}{\log n}} \log \log n$$

$$\text{with probability } \geq 1 - O\left(\frac{1}{(\log n)^8}\right).$$

The derivation of the lower bound is more difficult, but similar. We will only outline the proof. We examine the values of $f(\alpha, \theta)$ at the points $\theta_m = 2\pi \cdot (2m - 1)/2n$, for $1 \leq m \leq n$.

(*vii*) Let $M_1 = M = \sqrt{2a_2}\sqrt{n \log n} - gD \log \log n$, and $M_2 = 2M$. Let S be a subset of the integers from 1 to n with cardinality greater than $n(\log n)^{-1}$ and put $F = \sum_{m \in S} u(f(\alpha, \theta_m))$.

As in the derivation of the upper bound, we can find the asymptotic behavior of the first two moments of F using the properties of u.

(*viii*) We can show $E(F) \geq c_3 |S| n^{-1}(\log n)^{21/2}$ for some constant $c_3 > 0$, and $E(F^2) - E^2(F) = O(|S| n^{-1}(\log n)^{21/2}) + O(|S|^2 n^{-2}(\log n)^7)$.

(*ix*) Now

$$\Pr\{\max_{m \in S} f(\alpha, \theta_m) \geq M \text{ or } \min_{m \in S} f(\alpha, \theta_m) \leq -2M\} \geq 1 - \Pr\{F = 0\}$$

by the definition of n. But $\Pr\{\min f(\alpha, \theta_m) \leq -2M\} = O((\log n)^{-8})$ from the upper bound, so that

$$\Pr\{\max_{m \epsilon S} f(\alpha, \theta_m) \geq M\} \geq 1 - \Pr(F = 0) - O((\log n)^{-8}).$$

Further, $\Pr\{F = 0\} \leq \Pr\{(F - E(F))^2 \geq E^2(F)\}$, so by Chebyshev's inequality

$$\Pr\{F = 0\} \leq \frac{E((F - E(F))^2)}{E^2(F)}.$$

(x) Using the bounds from step ($viii$), $\Pr\{F = 0\} = O((\log n)^{-9/2})$. Therefore, using the definition of M_1, M_2, and D, we have:

$$\begin{aligned}\Pr\left\{\max_{m \epsilon S} f(\alpha, \theta_m) \geq \sqrt{2a_2}\,\sqrt{n \log n} - g\sqrt{\frac{n}{\log n}} \log \log n\right\} \\ \geq 1 - \Pr\{F = 0\} - O((\log n)^{-8}) \\ \geq 1 - O((\log n)^{-9/2}).\end{aligned}$$

Details of Step (iv)

From the definition of $v(t)$ [see step (iv)], we can show that

$$\int_{-\infty}^{\infty} |t|^r |v(t)| dt = O(D^{-r}) \qquad 1 \leq r \leq 18 \tag{18}$$

$$\int_{|t|>d/2} |t|^r |v(t)| dt = O(D^{-19}). \tag{19}$$

Since the Fourier transform of $t^r v(t)$ is $j^{-r} u^{(r)}(x)$,

$$\begin{aligned}\left|\int_{-\infty}^{\infty} e^{-\beta t^2} t^r v(t) dt\right| &= \frac{1}{2\sqrt{\pi\beta}} \left|\int_{-\infty}^{\infty} u^{(r)}(x) e^{-x^2/4\beta} dx\right| \\ &= O\left(\beta^{-1/2} D^{-r} \int_{|x| \geq M} e^{-x^2/4\beta} dx\right) \\ &= O\left(\frac{\sqrt{\beta}}{MD^r} e^{-M^2/4\beta}\right)\end{aligned} \tag{20}$$

uniformly in $\beta > 0$, $0 \leq r \leq 18$.

Similarly, for $\beta > 0$

$$\int_{-\infty}^{\infty} \exp\left(-\beta t^2 \sum_{k=1}^{n} \cos^2(k\theta + \alpha)\right) t^r v(t) dt = O\left(\frac{\sqrt{\beta u}}{MD^2} \exp(-M^2/Q)\right), \tag{21}$$

where

$$Q = 4\beta \sum_{k=1}^{n} \cos^2 (k\theta + \alpha).$$

Further,

$$\sum_{k=1}^{n} \cos^2 (k\theta + \alpha) = \frac{n}{2} + \frac{1}{2} \sum_{k=1}^{n} \cos 2(k\theta + \alpha)$$

and

$$\sum_{k=1}^{n} \cos 2(k\theta + \alpha) \leq \frac{1}{|\sin \theta|}.$$

So

$$\sum_{k=1}^{n} \cos^2 (k\theta + \alpha) \leq \begin{cases} n & \forall \theta, \alpha \\ \dfrac{n}{2} + \dfrac{n}{2 \log n} & \text{for } \dfrac{\pi}{2}\dfrac{\log n}{n} \leq |\theta|, \end{cases}$$

since

$$|\sin \theta| \geq \frac{n}{\log n} \quad \text{for} \quad \frac{\pi}{2}\frac{\log n}{n} \leq |\theta| \leq \pi - \frac{\pi}{2}\frac{\log n}{n}.$$

Therefore

$$\int_0^{2\pi} \exp \frac{-M^2}{4\beta \sum_{k=1}^{n} \cos^2 (k\theta + \alpha)} \, d\theta = O(e^{-M^2/2\beta n}), \tag{22}$$

whence

$$\int_0^{2\pi} d\alpha \int_0^{2\pi} d\theta \int_{-\infty}^{\infty} \exp\left[-\beta t^2 \sum_{k=1}^{n} \cos^2 (k\theta + \alpha) \right] t^r v(t) dt$$

$$= O\left(\frac{\sqrt{\beta n}}{MD^r} e^{-M^2/2\beta n} \right). \tag{23}$$

Using (10) and (23) above, we can derive the result of step (*iv*) as follows:

Since ϵ_k are independent,

$$E\left(\exp\left[jt \sum_{k=1}^{n} \epsilon_k c_k \right] \right) = \prod_{k=1}^{n} E e^{jt\epsilon_k c_k}$$

$$= \exp\left[-\sum_{l=2}^{5} a_l t^l \sum_{k=1}^{n} c_k^l + O(nt^6) \right] \quad \text{for} \quad |t| \leq d$$

from assumption (10), where c_k denotes $\cos(k\theta + \alpha)$. Further,

$$
\begin{aligned}
E\left(\exp\left[jt \sum_{k=1}^{n} \epsilon_k c_k\right]\right) = {} & \exp\left[-a_2 t^2 \sum_{k=1}^{n} c_k^2\right] \\
& + \sum_{j=3}^{5} a_j t^j \sum_{k=1}^{n} c_k^j \exp\left[-a_2 t^2 \sum_{k=1}^{n} c_k^2\right] \\
& + \frac{1}{2}\left\{\sum_{j=3}^{5} a_j t^j \sum_{1}^{n} e_k^j\right\}^2 \exp\left[-a_2 t^2 \sum_{k=1}^{n} c_k^2\right] \\
& + O(nt^6) + O(n^3 |t|^9) \quad \text{for} \quad |t| \leq d, \qquad (24)
\end{aligned}
$$

since

$$e^{-a} = e^{-b} + (b - a)e^{-b} + ½\,(b - a)^2 e^{-b} + O(|b - a|^3)$$

uniformly for $b \in \mathscr{R}$, $b \geq 0$, $a \in \mathscr{C}$, $\mathrm{Re}(a) \geq 0$. Now we consider

$$E\left(\int_{-\infty}^{\infty} \exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right] v(t)dt\right) = \int_{-\infty}^{\infty} E\left(\exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right]\right) v(t)dt;$$

the expression on the right-hand side of (24) can be substituted for the integrand in the interval $|t| \leq d$. Outside this interval, we can use the simple bound $|\exp[jt \sum_1^n \epsilon_k c_k]| \leq 1$, and arrive at:

$$
\begin{aligned}
E \int_{-\infty}^{\infty} \exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right] v(t)dt &= \int_{-d}^{d} E\left(\exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right]\right) v(t)dt \\
& \quad + O\left(\int_{|t| \geq d} |v(t)|dt\right) \\
&= \int_{-\infty}^{\infty} E\left(\exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right]\right) v(t)dt \\
& \quad + O\left(\int_{|t| \geq d} |v(t)|dt\right) \\
& \quad + O\left(n \int_{|t| \geq d} |t|^5 |v(t)|dt\right) \\
& \quad + O\left(n^2 \int_{|t| \geq d} t^{10} |v(t)|dt\right)
\end{aligned}
$$

$$+ O\left(n \int_{-\infty}^{\infty} t^6 \,|\, v(t) \,|\, dt\right)$$

$$+ O\left(n^3 \int_{-\infty}^{\infty} |\, t \,|^9 \,|\, v(t) \,|\, dt\right). \qquad (25)$$

From (18) and (19), we see that the right-hand side is

$$= \int_{-\infty}^{\infty} E\left(\exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right]\right) v(t)dt + O(nD^{-6}) + O(n^2 D^{-19}) + o(n^3 D^{-9})$$

$$= \int_{-\infty}^{\infty} E\left(\exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right]\right) v(t)dt + O\left(\frac{(\log n)^{9/2}}{n^{3/2}}\right). \qquad (26)$$

To find the asymptotic behavior of $E(G)$, we use (21). After integrating with respect to α,θ, we have, for each of the terms in (24), with expressions in square brackets corresponding to those in (24),

$$\int_0^{2\pi} d\alpha \int_0^{2\pi} d\theta \int_{-\infty}^{\infty} \exp\left[-a_2 t^2 \sum_{1}^{n} c_k^2\right] v(t)dt = O\left(\frac{\sqrt{n}}{M} e^{-M^2/2a_2 n}\right)$$

$$\int_0^{2\pi} d\alpha \int_0^{2\pi} d\theta \sum_{j=3}^{5} a_j \sum_{1}^{n} c_k^j \int_{-\infty}^{\infty} t^j \exp[\ ] v(t)dt = O\left(n \frac{\sqrt{n}}{MD^3} e^{-M^2/2a_2 n}\right)$$

$$\int_0^{2\pi} d\alpha \int_0^{2\pi} d\theta \int_{-\infty}^{\infty} \left\{\sum_{j=3}^{5} a_j t^j \sum c_k^j\right\}^2 \exp[\ ] v(t)dt$$

$$= O\left(n^2 \frac{\sqrt{n}}{MD^6} e^{-M^2/2a_2 n}\right).$$

Therefore, collecting the previous results, we have

$$E(G) = \int_0^{2\pi} d\alpha \int_0^{2\pi} d\theta \int_{-\infty}^{\infty} E\left(\exp\left[jt \sum_{1}^{n} \epsilon_k c_k\right]\right) v(t)dt$$

$$= O\left(\frac{\sqrt{n}}{M} e^{-M^2/2a_2 n} + \frac{(\log n)^{9/2}}{n^{3/2}}\right) = O(n^{-1}(\log n)^{-10}).$$

REFERENCES

1. B. D. Steinberg, *Principles of Aperture and Array System Design*, New York: John Wiley & Sons, 1976.

2. J. B. Knowles and E. M. Olcayto, "Coefficient Accuracy and Digital Filter Response," IEEE Trans. on Circuit Theory, *CT-15* (March 1968), pp. 31–41.
3. D. S. K. Chan and L. R. Rabiner, "Analysis of Quantization Errors in the Direct Form for Finite Impulse Response Digital Filters," IEEE Trans. Audio and Electroacoustics, *AU-21* (August 1973), pp. 354–356.
4. U. Heute, "Koefficienten—Empfindlichkeit nicht—rekursiver Digitalfilter in direkter Struktur," Dissertation, Institut für Nachrichtentechnik Universität Erlangen, 1974.
5. U. Heute, "Necessary and Efficient Expenditure for Non-Recursive Digital Filters in Direct Structure," European Conf. on Circuit Theory and Design, IEEE Conf. Pub. No. 116 (July 1974), pp. 13–19.
6. O. Andrisano and L. Calandrino, "Tap Weight Tolerance Effects in CCD Transversal Filtering," Alta Frequenza, *45*, 1976, pp. 739–745.
7. V. B. Lawrence and A. C. Salazar, "Effects of Finite Coefficient Precision on FIR Filter Spectra," Proc. IEEE International Conf. Acoustics, Speech, Signal Processing, April 1979, pp. 378–379.
8. A. Gersho, "Charge Transfer Filtering," Proc. IEEE, *67*, No. 2 (February 1979), pp. 196–218.
9. A. Gersho, B. Gopinath, and A. Odlyzko, "Coefficient Inaccuracy in FIR Filters," Proc. Int'l. Symp. Acoustics Speech & Signal Processing, 1979, pp. 375–377.
10. G. Halasz, "On a Result of Salem and Zygmund Concerning Random Polynomials," Studia Scient. Math Hung., *8* (1973), pp. 369–377.
11. L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1975.
12. W. Fuchs, "A Problem on Approximation by Polynomials," unpublished work, 1975.
13. P. Whittle, "Sur la Distribution du Maximum d'un Polynome Trigonometrique à Coefficients Aléatories," Le Calcul des Probabilities et ses Applications, Centre National de la Recherche Scientifique, Paris, 1959 (Colloques Internationaux du C.N.R.S. 1958).

Contributors to This Issue

Barbara E. Caspers, 1961, Electrical Engineering Associate's Diploma, Siemens & Halske, Germany; Bell Laboratories, 1964—. Mrs. Caspers initially worked on networks for transmission systems. In 1966 she joined the Speech and Communication Research Department and was involved in research on operating systems for real-time speech and graphics applications.

Peter B. Denes B.Sc., M.Sc. (Electrical Engineering), Manchester University, England; Ph.D. (Engineering), University of London, England; Bell Laboratories, 1961—. From 1946 to 1961, Mr. Denes was a lecturer at University College, London, where he was in charge of the laboratory of the Phonetics Department. At Bell Laboratories, his research interests have been primarily in speech, graphics and improved man-computer communication. He became head of the Speech and Communication Research Department in 1967. Fellow, Acoustical Society of America. He has served as Associate Editor for Speech Communication, Journal of the Acoustical Society of America, and as vice president, Summit (N.J.) Speech School.

Paul E. Fleischer, B.E.E., 1955, M.E.E., 1956, Dr. Eng. Sc. (E.E.), 1961, New York University; Instructor, New York University, 1958–1961; Bell Laboratories, 1961—. Mr. Fleischer's earlier work was concerned with the application of analog, digital, and hybrid computation to the design and optimization of electric networks. More recently, his work has been in the study and application of active networks for the realization of filters and equalizers. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Allen Gersho, B.S. (E.E.), 1960, Massachusetts Institute of Technology; M.S., 1961, and Ph.D., 1963, Cornell University; Bell Laboratories, 1963—. A member of the Mathematics and Statistics Research Center at Murray Hill, Mr. Gersho's research activities have been in the analysis and modeling of engineering problems in communications, circuits, and systems, including, in particular, adaptive equalization, adaptive quantization, and filtering with charge-coupled devices. Senior member, IEEE; Editor, *IEEE Communications Magazine.*

Edgar N. Gilbert, B.S. (Physics), 1943, Queens College; Ph.D. (Mathematics), 1948, Massachusetts Institute of Technology; M.I.T. Radiation Laboratory, 1944–46; Bell Laboratories, 1948—. Mr. Gilbert is a member of the Mathematics and Statistics Research Center at Bell Laboratories, specializing in communication theory and other applications of probability and combinatorial mathematics.

B. Gopinath, M.Sc. (Math.), 1964, University of Bombay; Ph.D. (E.E.), 1968, Stanford University; research associate, Stanford University, 1967–1968; Alexander von Humboldt research fellow, University of Göttingen, 1971–1972; Bell Laboratories, 1968—. Mr. Gopinath is engaged in applied mathematics research in the Mathematics and Statistics Research Center.

W. L. Harrod, B.S.M.E., 1962, Texas Tech. University; M.S.E.M., 1964, Rutgers University; M.S.M.S., 1971, and Ph.D., 1975, Northwestern University; Bell Laboratories, 1962—. Mr. Harrod has worked on the physical design of hardware for government and military switching systems and on the development of technology for applying thin-film and silicon integrated circuits in electronic switching systems. From 1968 through 1977, he supervised a group responsible for a variety of projects related to the design, development, and evaluation of new hardware for packaging and interconnecting integrated circuits and other components in electronic switching systems. From 1977 until the present, he has been supervisor of a group having responsibility for the physical design and characterization of the *BELLPAC* system of electronic packaging hardware. Member, IEPS, Tau Beta Pi.

Eugene Helfand, B.S., 1955, Polytechnic Institute of Brooklyn; M.S., 1957, Ph.D. 1958 (Chem.), Yale University; Bell Laboratories, 1958—. Mr. Helfand has worked on various theoretical aspects of chemistry and materials science including transport theory, liquids, superconductivity, and critical phenomena. His current interests are in polymer science and dynamical simulation. Guggenheim Fellow, 1968. Member, American Physical Society, American Chemical Society, Sigma Xi, Phi Lambda Upsilon.

Frank K. Hwang, B.A., 1960, National Taiwan University; M.B.A., 1964, City University of New York; Ph.D. (Statistics), 1968, North Carolina State University; Bell Laboratories, 1967—. Mr. Hwang has been engaged in research in statistics, discrete mathematics, computing algorithms, and switching networks in the Mathematics and Statistics Research Center.

Kenneth R. Laker, B.S.E.E., 1968, Manhattan College, M.S.E.E., 1970, Ph.D. (E.E.), 1973, New York University; Captain, USAF, Air Force Cambridge Research Laboratories, 1973–1977; Bell Laboratories, 1977—. Mr. Laker has worked in the areas of surface acoustic wave devices and active networks. Since joining Bell Laboratories he has been engaged in various development and exploratory activities associated with active-RC and active-switched capacitor filters. Member, Eta Kappa Nu, Sigma Xi; Senior Member, IEEE; Administrative Committee, IEEE CAS Society; Co-chairman, Optical, Microwave, and Acoustical Circuits Technical Committee, IEEE CAS Society; IEEE SU Group.

Anthony G. Lubowe, A. B., 1957; B.S. (M.E.), 1958; M.S., 1959; Eng. Sc.D., Engineering Mechanics, 1961, Columbia University; Bell Laboratories, 1961—. Mr. Lubowe first worked on methods of orbit determination and prediction used for the Telstar communications satellite experiment, for the Apollo project, and for NASA and Department of Defense studies. Later work included experimental study of acoustic wave propagation in soil and exploratory development for a digital echo canceller. Since 1971, he has been in the Interconnection Technology Laboratory. He is presently supervisor of the *BELLPAC* Prototype and Assembly Group. Professional Engineer, N.J., N.Y.; member, ASME, Tau Beta Pi, Phi Beta Kappa.

Stephen C. Mettler, B.S., 1962, U.S.A.F. Academy; M.S. (Physics), 1972, Ph.D. (M.E.), 1976, Purdue University; Bell Laboratories, 1976—. Mr. Mettler is presently engaged in optical fiber splicing studies. Member, Optical Society of America.

Andrew M. Odlyzko, Ph.D., 1975, Massachusetts Institute of Technology; Bell Laboratories, 1975—. Mr. Odlyzko works in various areas of mathematics, including error-correcting codes, combinatorics, analysis, probability theory, and analysis of algorithms.

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

Jay G. Wilpon, B.S. (mathematics), A.B. (economics), (cum laude), 1977, Lafayette College; Bell Laboratories, 1977—. At Bell Laboratories, Mr. Wilpon has been engaged in speech communications research and is presently concentrating on problems of speech recognition.

Papers by Bell Laboratories Authors

PHYSICAL SCIENCES

AC Electroluminescence in Thin Film ZnSe:Mn. J. Shah and A. E. DiGiovanni, Appl. Phys. Lett., *33* (Dec 15, 1978), pp. 995–7.

Allowed Character of the 1900AA Band Borazine. M. B. Robin and N. A. Kuebler, J. Mol. Spectrosc., *70*, No. 3 (Jun 1, 1978), pp. 472–5.

Aluminum Gallium Arsenide (DH) Pump Laser for Photoluminescence Lifetime Measurements. R. J. Nelson, Rev. Sc., Instrum., *49*, No. 12 (Dec 1978), pp. 6103–8.

Analysis of Tin Nickel Electroplate by Secondary Ion Mass Spectrometry, Ion Scattering, Spectrometry, and Rutherford Backscattering. R. Schubert, J. Electrochem. Soc., *125*, No. 8 (Aug 1978), pp. 1215–8.

Analytic Approximations for the Fermi Energy in (Aluminum, Gallium) Arsenide. W. B. Joyce, Appl. Phys. Lett, *32*, No. 10 (May 15, 1978), pp. 680–1.

Angle-Resolved Photoemission from Surfaces and Adsorbates. N. V. Smith, J. Phys. (Paris), *39*, No. 4 (1978), p. 161.

Cadmium Sulfide/Indium Phosphide and Cadmium Sulfide/Gallium Arsenide Heterojunctions by Chemical Vapor Deposition of Cadmium Sulfide. M. Bettini, K. J. Bachmann, and J. L. Shay, J. Appl. Phys., *49*, No. 2 (Feb 1978), pp. 865–70.

Chemical Kinetics of the Reactions of $SiCl_4SiBr_4$, $GeCl_4$, $POCl_3$, and BCl_3 with Oxygen. W. G. French, L. J. Pace, and V. A. Foertmeyer, J. Phys. Chem., *82*, No. 20 (Oct 1978), pp. 2191–4.

Comparative Study of Annealed Neon, Argon, Krypton, Ion Implanted Damage in Silicon. A. G. Cullis, T. E. Seidel, and R. L. Meek, J. Appl. Phys., *49*, No. 10 (Oct 1978), pp. 5188–98.

Compton Profile of Lithium Hydride. W. A. Reed, Phys. Rev. B, *18* (1978), p. 1986.

The Compton Profile of Urea. W. A. Reed, L. C. Snyder, H. J. Guggenheim, T. A. Weber, and Z. R. Wasserman, J. Chem. Phys., *69*, No. 1 (Jul 1, 1978), pp. 288–96.

Copper Chloride: More Facts Generate More Thoughts on High Temperature Superconductivity. J. A. Wilson, Phil. Mag. B, *38* (1978), pp. 427–44.

Core Hole Screening in Lanthanide Metals. G. Crecelius, G. K. Wertheim, D. N. E. Buchanan, Phys. Rev. B, *18*, No. 12 (Dec 15, 1978), pp. 6519–24.

A Coulometric Analysis of Iron (II) in Ferrites Using Chlorine. P. K. Gallagher, Am. Cer. Soc. Bull., *57*, No. 6 (1978), pp. 576–8.

Current Distribution Leveling Resulting from Auxiliary Bipolar Electrodes. W. Engelmaier, T. Kessler, and R. Aikire, J. Electrochem. Soc., *125*, No. 2 (Feb 1978), pp. 209–16.

Depolarized Rayleigh Spectroscopy in the N-Alkanes. G. D. Patterson, C. P. Lindsey, and G. R. Alms, J. Chem. Phys., *69* (1978), p. 3250.

Determination of the E(K) Relation for a Surface State on Gold (111). Z. Hussain and N. V. Smith, Phys. Lett. A, *66* (Jun 26, 1978), pp. 492–4.

Deuteron Quadrupole Coupling in Molecules. L. C. Snyder, J. Chem. Phys., *68*, No. 1 (1978), pp. 291–4.

The Deuteron Quadrupole Coupling Constant in Carbon Deuterium (3) Fluoride. L. C. Snyder, J. Chem. Phys., *68*, No. 1 (1978), pp. 340–1.

Diffraction of Photoelectrons Emitted from Core Levels of Tellurium and Sodium Atoms Adsorbed on Nickel (001). D. P. Woodruff, D. Norman, B. W.

Holland, N. V. Smith, H. H. Farrell, and M. M. Traum. Phys. Rev. Lett., *41*, No. 16 (Oct 16, 1978), pp. 1130–3.

Directional Photoemission from Two-Dimensional Systems. N. V. Smith and P. K. Larsen, in *Photoemission and Electronic Structure of Surfaces*, B. Fenesbacher, B. Fitton, and R. F. Willis, eds., New York: John Wiley & Sons, 1978, p. 407.

Distant Intramolecular Interaction Between Identical Chromophores: The N-PI* Excited States of P-Benzoquinone. J. Goodman and L. E. Brus, J. Chem. Phys, *69* (1978), p. 1604.

Double Phase Matching Function. D. F. Nelson, J. Opt. Soc. Amer., *68*, (Dec 1978), pp. 1780–1.

The Effect of Melt Flow Phenomena on the Perfection of Czochralski Grown Gadolinium Gallium Garnet. D. C. Miller, A. J. Valentino, and L. K. Shick, J. Cryst. Growth, *44* (1978), pp. 121–34.

On the Effect of a Partial Sink in Binary Diffusion in Thin Films. H. G. Tompkins, J. Appl. Phys., *49*, No. 1 (Jan 1978), pp. 223–8.

The Effect of Thermal Fluctuations on the Harmonics and Lock-in Transition of Charge Density Waves. S. A. Jackson and P. A. Lee, Phys Rev B, *18*, No. 6 (Sep 15, 1978), pp. 2500–5.

Effects of Atomic Order in Alpha- and Beta-Phase AgCd Alloys Studied by X-Ray Photoelectron Spectroscopy. G. Crecelius and G. K. Wertheim, Phys. Rev. B, *18*, No. 12 (Dec 15, 1978), pp. 6525–30.

Electric Deflection and Molecular Structure. W. E. Falconer, Israel J. Chem., *17* (1978), pp. 31–6.

Electrical Characterization of Heterostructure Lasers. W. B. Joyce and E. W. Dixon, J. Appl. Phys., *49*, No. 7 (Jul 1978), pp. 3719–28.

Electrochromism in Anodic Iridium Oxide Films. S. Gottesfeld, Joe McIntyre, G. Beni, and J. L. Shay, Appl. Phys. Lett., *33*, No. 2 (Jul 15, 1978), pp. 208–10.

Electronic Structure and Spectra of Small Rings. VI. Multiphoton Ionization Spectra of the Saturated Three-Membered Rings. M. B. Robin and N. A. Keubler, J. Chem. Phys., *89*, No. 2 (Jul 15, 1978), pp. 806–10.

An Evolved Gas Analysis System. P. K. Gallagher, Thermochim. Acta, *26* (1978), pp. 175–83.

Exact Effect of Surface Roughness on the Reverberation Time of a Uniformly Absorbing Spherical Enclosure. W. B. Joyce, J. Acoust. Soc. Amer., *64* (Nov 1978), pp. 429–36.

Extraction of Information from Glassy Spectra. G. E. Peterson, A. Carnevale, and C. R. Kurkjian, Phys. Chem. Glass., *21* (1978), p. 34.

The Four-Point Bend Test for Measuring the Ductility of Brittle Coatings. C. C. Lo, J. Electrochem. Soc., *125*, No. 3 (Mar 1978), pp. 400–3.

Gettering of Surface and Bulk Impurities in Czochralski Silicon Wafers. G. A. Rozgony and C. W. Pearce, Appl. Phys. Lett., *32*, No. 11 (Jun 1, 1978), pp. 747–9.

A High Efficiency Tunable Picosecond Dye Laser. D. Huppert and P. M. Rentzepis, J. Appl. Phys., *49*, No. 2 (Feb 1978), pp. 543–8.

Hypersonic Attenuation in the N-Alkanes. G. D. Patterson and C. P. Lindsey, J. Appl. Phys., *49* (1978), pp. 5039–41.

Hypersonic Relaxation and the Glass-Rubber Relaxations in Polypropylene Glycol. G. D. Patterson, D. C. Douglass, and J. P. Latham, Macromolecules, *11* (1978), pp. 263–5.

Interaction of Nitrogen Oxide and Copper at Various Relative Humidities. R. Schubert, J. Electrochem. Soc., *125*, No. 7 (Jul 1978), pp. 1114–6.

Interfacial Recombination in Gallium Aluminium Arsenide–Gallium Arsenide Heterostructures. R. J. Nelson, J. Vacuum Sci. Technol, *15*, No. 4 (Jul-Aug 1978), pp. 1475–7.

Interlaminar Thermoelastic Stresses in Layered Beams. P. B. Grimado, J. Therm. Anal., *1* (Jul 1978), pp. 75–86.

Investigation of Agitation Effects on Electroplated Copper in Multilayer Board Plated-Through-Holes in a Forced Flow Plating Cell. W. Engelmaier and T. Kessler, J. Electrochem. Soc., *125* (Jan 1978), pp. 36–43.

Laser Induced Nuclear Orientation: Intersection of Laser and Nuclear Spectroscopy. M. G. Feld, M. Burns, P. Pappas, and D. E. Murnick, Hyperfine Interact., *4* (1978), pp. 56–60.

Lattice Dynamics of Lanthanum Antimonide and Praseodymium Antimonide. D. B. McWhan, C. Vettier, L. D. Longinotti, and G. Shirane, Phys. Rev. B, *18* (1978), pp. 4540–1.

Low-Loss, Single-Mode Fibers with Different Boron Dioxide-Silicon Dioxide Compositions. G. W. Tasker, W. G. French, J. R. Simpson, P. Kaiser, and H. M. Presby, Appl. Opt., *17,* No. 11 (Jun 1, 1978), pp. 1836–42.

The Magnetic Behavior of an S = ½ Amorphous Antiferromagnet. R. B. Kummer, R. E. Walstedt, S. Geschwind,V. Narayanamurti, and G. E. Devlin, Phys. Rev. Lett., *40,* No. 16 (Apr 17, 1978), pp. 1098–1100.

Magnetic Susceptibility of an Amorphous Spin Glass: Au-Si-Mn. J. J. Hauser and J. V. Waszczak, Phys. Rev., *18* (Dec 1, 1978), pp. 6206–12.

Metastable Alloy Layers Produced by Implantation of Silver (+) and Tantalum (+) Ions into Copper. A. G. Cullis, J. A. Broders, J. K. Hirvonen, and J. M. Poate, Phil. Mag., *37* (1978), pp. 615–30.

Microstructure and Magnetism in Amorphous Rare Earth Transition Metal Alloys: I. Microstructure. H. J. Leamy and A. G. Dirks, J. Appl. Phys., *49* (Jun 1978), pp. 3430–6.

Minority-Carrier Lifetimes and Internal Quantum Efficiency of "Surface-Free" Gallium Arsenide. R. J. Nelson and R. G. Sobers, J. Appl. Phys., *49* (Dec 1978), p. 6103.

Molecular Orbital Theory of the Compton Profile of Fluorine (2). L. C. Snyder and T. A. Weber, J. Chem. Phys., *68,* No. 6 (Mar 15, 1978), pp. 2974–9.

Molecular Recognition and Self-Organization in Fluorinated Hydrocarbons. F. H. Stillinger and Z. Wasserman, J. Phys. Chem., *32,* No. 8 (Apr 20, 1978), pp. 929–40.

Molecular SCF Calculations of Model (111) Surface States and Relaxation of Diamond. L. C. Snyder and Z. Wasserman. Surf. Sci., *31,* No. 2 (Feb 1978), pp. 407–13.

A Narrow Bandwidth Picosecond Laser. D. Huppert and P. M. Rentzepis, Appl. Phys. Lett., *32,* No. 4 (Feb 15, 1978), pp. 241–4.

Neutron Scattering at 4.5 GPA and 20 K. D. B. McWhan in *High-Pressure Science and Technology,* K. D. Timmerhaus and M. S. Barber, eds., New York: Plenum Press, 1979, pp. 292–6.

Nonlinear NMP in Helium-3B with Arbitrary Orientation. M. Liu and W. F. Brinkman, J. Low Temp. Phys., *30,* No. 4 (Feb 1978), pp. 551–7.

A Note on the Elusive 1E2G State in the Two-Photon Spectrum of Benzene. V. Vaida, M. B. Robin, and N. A. Kuebler, Chem. Phys. Lett., *58,* No. 4 (1978), pp. 557–60.

Observation and Analysis of the Depolarized Rayleigh Doublet in Isotropic MBBA and the N-Alkanes. G. P. Alms and G. D. Patterson, J. Colloid, Interfac. Sci., *63* (1978), pp. 184–92.

Optically Modulated Electron Beam Studies of Gallium Phosphide. S. C. Dahlberg, Surf. Sci., *75* (1978), p. 256.

Optically Modulated Electron Beam Studies of Silicon (III). S. C. Dahlberg, Phys. Rev. B., *17* (1978), pp. 4757–64.

Optical Surface Waves in Periodic Layered Media. P. Yeh, A. Yariv, and A. Y. Cho, Appl. Phys. Lett., *32* (Jan 15, 1978), pp. 104–5.

Periodic Regrowth Phenomena Produced by Laser Annealing of Ion Implanted Silicon. H. J. Leamy, G. A. Rozgonyi, T. T. Sheng, and G. K. Celler, Appl. Phys. Lett., *32,* No. 9 (May 1, 1978), pp. 535–7.

Piscosecond Absorption Studies of Excess Electrons in Ammonia, Amines, and Ethers. D. Huppert, P. H. Avouris, and P. M. Rentzepis, J. Phys. Chem., *82* (1978), pp. 2282–6.

Picosecond Dynamics of Primary Electron-Transfer Processes in Bacterial Photosynthesis. P. Avouris, K. S. Peters, and P. M. Rentzepis, Biophys. J., *23,* No. 2 (Aug 1978), pp. 207–17.

Picosecond Self-Induced Transparency and Photon Echoes in Sodium Vapor. H. M. Gibbs and P. Hu, in *Piscosecond Phenomena,* C. V. Shank, E. P. Ippen, and S. L. Shapiro, eds., Berlin: Springer-Verlag, 1978, pp. 336–7.

Plasma Assisted Etching Techniques for Pattern Delineation. C. M. Melliar-Smith and C. J. Mogab in *Thin Film Processes,* J. L. Vossen and W. Kern, eds., New York: Academic Press, 1978, pp. 497–556.

Polarization Model Representation of Hydrogen Fluoride for Use in Gas and Condensed Phase Studies. R. H. Stillinger, Int. J. Quantum Chem., *14* (1978), pp. 649–57.

Positron-Annihilation Momentum Profiles in Aluminum: Core Contribution and the Independent Particle Model. K. G. Lynn, J. R. MacDonald, R. A. Boie, L. C. Feldman, J. D. Gabbe, M. F. Robbins, E. Bonderup, and J. Golduchinko, Phys. Rev. Lett., *38* (Jan 31, 1977), pp. 241–4.

Preparation and Characterization of Tetra(2,4-Pentanedionato)-Hexa(Benzotriazolato)-Penta-Copper (II). J. H. Marshall, Inorg. Chem., *17* (Dec 1978), pp. 3711–3.

Preparation and Electropolishing of Thin Gold Disc Specimens for Transmission-Electron-Microscope Examinations. W. F. Beck and S. Nakahara, Metallography, *11* (1978), pp. 347–54.

Pressure-Broadened Linewidths of the R(9.5)3/2 Nitric Oxide Transition. R. E. Richton, Appl. Opt., *17*, No. 10 (May 15, 1978), pp. 1606–9.

Pressure Dependence of Magnetic Excitations on Samarium Sulfide. D. B. McWhan, S. M. Shapiro, J. Eckert, H. A. Mook, and R. J. Birgeneau, Phys. Rev. B, *18* (1978), p. 3632.

Primary Intermediates in the Photochemical Cycle of Bacteriorhodopsin. M. L. Applebury, K. S. Peters, and R. M. Rentzepis, Biophys. J., *23* (1978), pp. 375–82.

Proton Transfer and Tautomerism in an Excited State of Methyl Salicylate. J. Goodman and L. E. Brus, J. Am. Chem. Soc, *100* (1978), pp. 7472–4.

P-State Pairing and the Ferromagnetism of Zirconium Zinc (2). B. T. Matthias and C. P. Enz, Science, *201* (Sep 1, 1978), pp. 828–9.

Raman Study of the Oxygen Fluoride Plus Vanadium Fluoride (5) Reaction: Isolation and Identification of an Unstable Reaction Intermediate. J. G. Griffiths, A. J. Edwards, W. A. Sunder, and W. E. Falconer, J. Fluorine Chem., *11* (1978), pp. 119–42.

Singlet-Triplet Anticrossings Between the Doubly Excited 3(1)K State and the G(3D)3 Sigma (G+) State of Hydrogen (2). R. S. Freund, T. A. Miller, R. Jost, and M. Lombardi, J. Chem. Phys. *68*, No. 4 (Feb 15, 1978), pp. 1683–8.

Sintering of High Density Ferrites. M. F. Yan and D. W. Johnson, Jr., in *Processing of Crystalline Ceramics*, H. Palmour, R. F. Davies, and T. M. Hare, eds., New York: Plenum Publishing, 1978, pp. 393–402.

Slow Positron Emission from Metal Surfaces. A. P. Mills, Jr., P. M. Platzman, and B. L. Brown, Phys. Rev. Lett., *41* (1978), p. 1076.

Sol-Gel Behavior and Image Formation on Poly(glyoidyl Methacrylate) and its Copolymers with Ethyl Acrylate. E. D. Feit, M. E. Wurtz, and G. W. Kammlott, J. Vacuum Sci. Technol, *15*, No. 3 (May-Jun 1978), pp. 944–7.

Spin and Orbital Dynamics of Superfluid Helium-3. W. F. Brinkman and M. C. Cross in *Progress in Low Temperature Physics, VIIA*, D. F. Brewer, ed., New York: Elsevier (North Holland), 1978, pp. 105–90.

Stochastic Classical Trajectory Approach to Relaxation Phenomena: I. Vibrational Relaxation of Impurity Molecules in Solid Matrices. M. Shugard, J. C. Tully, and A. Nitzan, J. Chem. Phys., *69*, No. 1 (Jul 1, 1978), pp. 336–45.

Study of Melting and Freezing in the Gaussian Core Model by Molecular Dynamic Simulation. F. H. Stillinger and T. A. Weber, J. Chem. Phys., *68*, No. 8 (Apr 15, 1978), pp. 3637–44.

Subpicosecond Vibrational Relaxation in Calcium (2) in Rare Gas Solids. V. E. Bondybey and C. Albiston, J. Chem. Phys., *63*, No. 7 (Apr 1, 1978), pp. 3172–6.

Tantalum Ion Transport Number During the Anodic Oxidation of Beta-Tantalum Films. N. Schwartz and W. M. Augustyniak, J. Electrochem. Soc., *125* (1978), p. 1812.

Tellurium-Induced Defects in LPE Aluminum (.36) Gallium (.84) Arsenide. W. R. Wagner, J. Appl. Phys., *49*, No. 1 (Jan 1978), pp. 173–80.

MATHEMATICS

Application of Fast Kalman Estimation to Adaptive Equalization. D. D. Falconer and L. Ljung, Trans. Commun., *26*, No. 10 (Oct 1978), pp. 1439–46

Fast Calculation of Gain Matrices for Recursive Estimation Schemes. L. Ljung, M. Morf, and D. D. Falconer, Int. J. Contr., *27*, No. 1 (Jan 1978), pp. 1–9.

Fixed Point Error Analysis of the Winograd Fourier Transform Algorithm. R. W. Patterson and J. H. McClellan, IEEE Trans. Acoust. Speech Single Process., *5* (Oct 1978), pp. 447–55.

A Generalization of Line Connectivity and Optimally Invulnerable Graphs. F. Boesch and S. Chen, SIAM J. Appl. Math, *34*, No. 4 (Apr 1978), pp. 657–65.

An Implicit Enumeration Algorithm for Sequencing Policies Applied to Telephone Switching Facilities. L. J. Ackerman, H. Luss, and R. S. Berkowitz, IEEE Trans. Syst. Man. Cybern., *8*, No. 4 (Apr 1978), pp. 296–300.

A Linear Programming Approach to Geometric Programs. J. J. Dinkel, W. H. Elliott, and G. A. Kochenberger. Nav. Res. Logis. Quart., *25*, No. 1 (Mar 1978), pp. 39–53.

Unicyclic Realizability of Degree Lists. F. Boesch and F. Haray, Networks, *8*, No. 2 (Aug 1978), pp. 93–6.

COMPUTING

ASIS-77 (Meeting Report). D. T. Hawkins, Online, *2*, No. 1 (Jan 1978), pp. 72–3.

Bibliometrics of the Online Information Retrieval Literature. D. T. Hawkins, Online Rev., *2*, No. 4 (1978), pp. 345–52.

Data Base Subject Index. H. H. Teitelbaum and D. T Hawkins, Online, *2*, No. 2 (Apr 1978), pp. 16–21.

Fortran 77. T. A. Gibson and J. C. Noll, Commun. ACM, *21* (Oct 1978), pp. 806–20.

The Literature of Noble Gas Compounds. D. T. Hawkins, J. Chem. Inform. Comput. Sci., *18*, No. 4 (1978), pp. 190–9.

Mathematical Programming Approaches to System Partitioning. J. L. Uhrig, IEEE Trans. Syst. Man. Cybern., *8*, No. 7 (Jul 1978), pp. 540–8.

Multiple Data Base Searching: Techniques and Pitfalls. D. T. Hawkins, Online, *2*, No. 2 (Apr 1978), pp. 9–15.

On-Line Information Retrieval Bibliography. D. T. Hawkins, Online Rev., *2*, No. 1 (Mar 1978), pp. 63–106.

On-Line Searching Technique: Retrieving Every Metallic Element Using Registry Numbers. B. A. Stevens, Online, *2*, No. 3 (Jul 1978), p. 67.

On the Rapid Solution of Differential Equations by Microprocessors. C. A. Cooper and D. I. Moldovan. Modeling Simulation, *9*, No. 4 (1978), pp. 1465–71.

Toward an Understanding of (Actual) Data Structures. W. L. Honig and C. R. Carlson, Comput. J., *21*, No. 2 (May 1978), pp. 98–104.

Unconventional Uses of On-Line Retrieval Systems. D. T. Hawkins, J. Amer. Soc. Inform. Sci, *29*, No. 4 (Jul 1978), p. 209.

Workshop Report: The New and the Not-So-New. M. F. Slana and G. G. Dumas, Computer, *11*, No. 3 (Mar 1978), pp. 47–51.

ENGINEERING

Bending of a Nonlinear Rectangular Beam in Large Deflection. D. D. Lo and S. Das Gupta, J. Appl. Mech., *45*, (Mar 1978), pp. 213–5.

A Conductor Crossing Problem in Magnetic Bubble Memories. W. Strauss, J. Appl. Phys., *49*, No. 3 (Pt. 2) (Mar 1978), pp. 1897–9.

An Experimental Display Structure Based on Reversible Electrodeposition. I. Camlibel, S. Singh, H. J. Stocker, L. G. Van Uitert, and G. J. Zydzik, Appl. Phys. Lett, *33*, No. 9 (Nov 1978), pp. 793–4.

Experimental Speakerphone System for Teleconferencing. J. E. Wert, D. J. McLean, J. R. Nelson, and J. L. Flanagan, J. Acoust. Soc. Amer., *64* (Dec 1978), pp. 1561–5.

Fiberguide to Metal Hermetic Seal. W. W. Benson, D. R. MacKenzie, T. C. Rich, and I. Camlibel, Appl. Opt., *17*, No. 15 (Aug 1, 1978), pp. 2271–2.

Lightwave Fiber Tap. M. A. Karr, R. C. Rich, and M. DiDomenico, Appl. Opt., *17*, No. 14 (Jul 15, 1978), pp. 2215–8.

MOS Control of Switches in Single Mode GaAsAlGaAs Optical Rib Waveguides. J. C. Shelton, F. K. Reinhart, and R. A. Logan, IEEE Trans. Electron. Dev., *24*, No. 9 (1977), p. 1198.

Optical Picosecond Spectroscopy. D. Huppert, P. M. Pentzepis, and G. E. Busch, Opt. Eng., *17* (Jan–Feb 1978), pp. 82–9.

Request Queuing for Magnetic Bubble Memories. P. I. Bonyhard and F. B. Hagedorn, IEEE Trans. Mag., *14*, No. 2 (Mar 1978), pp. 37–40.

Rib Waveguide Switches with MOS Electrooptic Control for Monolithic Integrated Optics in Gallium Arsenide-Aluminum (X) Gallium (1-X) Arsenic. J. C. Shelton, F. K. Reinhart, and R. A. Logan, Appl. Opt., *17,* No. 16 (1978), pp. 2548–55.

A Simple, Reliable, Optical Fiber-to-Metal Hermetic Seal. T. C. Rich and I. Camlibel, Amer. Ceram. Soc. Bull., *57,* No. 2 (Feb 1978), pp. 234–5.

Single-Mode Gallium Arsenide-Aluminum (X) Gallium (1-X) Arsenic Rib Waveguide Switches. J. C. Shelton, F. K. Reinhart, and R. A. Logan, Appl Opt., *17,* No. 3 (1978), pp. 890–1.

Stress in Glass Fibers Induced by the Draw Force. L. Rongved, Appl. Mech., *45,* No. 4 (Dec 1978), pp. 765–72.

Theoretical Derivatives of the Electrical Characteristic of a Junction Laser Operated in the Vicinity of Threshold. T. L. Paoli, J. Quantum Electron., *14,* No. 1 (Jan 1978), pp. 62–8.

SOCIAL AND LIFE SCIENCES

Syllables as Concatenative Phonetic Units. O. Fujimura and J. B. Lovins in *Syllables and Segments,* A. Bell and J. B. Hooper, eds., Amsterdam: North Holland, 1978.

Computer Model to Characterize the Air Volume Displaced by the Vibrating Vocal Cords. J. L. Flanagan and K. Ishizaka, J. Acoust. Soc. Amer., *63,* No. 5 (May 1978), pp. 1559–65.

Prose Retention: Recognition Test Effects and Style Memory. R. E. Christiaansen, D. J. Dooling, and T. F. Keenan, Bull. Psychonom. Soc., *11* (1978), pp. 383–6.

A Self-Oscillating Model of the Glottal Sound Sources. K. Ishizaka and J. L. Flanagan, J. Acoust. Soc. Japan, *34,* No. 3 (Mar 1978), pp. 122–31.

Subjective Detection of Differences in Variance from Small Samples. R. L. Pike and W. R. Ferrell, Organ. Behav. Hum. Per., *22* (1978), pp. 262–78.

Techniques for Expanding the Capabilities of Practical Speech Recognizers. J. L. Flanagan, S. E. Levinson, L. R. Rabiner, and A. E. Rosenberg in *Trends in Speech Recognition,* W. Lea, ed., Englewood Cliffs: Prentice-Hall, 1978.

MANAGEMENT AND ECONOMICS

The Basic Principles of Teleconferencing. C. Stockbridge and R. J. Miller, Commun. News (Dec 1978), pp. 82–3.

Capacity Constrained Peak Load Pricing, Quart. J. Econ., R. E. Dansby, *42,* No. 3 (Aug 1978), pp. 387–99.

A Communication Network for Distributed Data Acquisition and Control in Industrial Plant. S. Chandra, IEEE Trans. Ind. Electron. Contr. Instrum., *25,* No. 3 (Aug 1978), pp. 206–12.

Human Performance Considerations in Complex Systems. H. C. Holt and F. L. Stevenson, J. Syst. Manag. (Oct 1978), pp. 14–20.

Selection of Microform Readers and Reader Printers. C. H. Robertshaw, J. Microgr., *12,* No. 2 (Nov-Dec 1978), pp. 70–2.

Two Conferences Address Local Digital Switching and Transmission in Zurich and Atlanta. R. W. Wyndrum, Commun. Soc. Mag., *16* (Jul 1978), pp. 23–4.