THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIII JULY 1954 NUMBER 4

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at $3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIII JULY 1954 NUMBER 4

Negative Resistance Arising from Transit Time in Semiconductor Diodes

By W. SHOCKLEY

(Manuscript received January 22, 1954)

The structural simplicity of two-terminal compared to three-terminal devices indicates the potential importance of two terminal devices employing semiconductors and having negative resistance at frequencies properly related to the transit time of carriers through them. Such negative resistances may be combined with unsymmetrically transmitting components, such as gyrators or Hall effect plates, to form dissected amplifiers *that may be made to simulate conventional three-terminal amplifiers and operate at high frequencies. The characteristics of several structures are analyzed on the basis of theory and it is found that negative resistances are possible for properly designed structures.*

1. NEGATIVE RESISTANCE AND DISSECTED AMPLIFIERS

Because the drift velocities of current carriers in semiconductors are smaller than the velocities attainable in vacuum tubes, transistor structures must be smaller to achieve comparable frequencies. In principle it is possible, of course, to make compositional structures (i.e., distributions of donors and acceptors) in semiconductor crystals on a scale much smaller than is possible for vacuum tubes. At present, however, the available techniques are limited and it may require many years before the ultimate potentialities are approached.

It is instructive, however, to speculate on some of these ultimate potentialities. For example a grain boundary formed of edge type dislocations is in a sense an analogue of a grid. Possibly it can be made into a grid by acting as a locus for an atmosphere of donors or acceptors. Evidently such a grid will approach the smallest spacing that can be achieved with any known form of matter. If the spacings perpendicular to the grid are made comparable to a mean free path of the carriers used, the device will operate like a vacuum tube with carrier velocities controlled by inertia rather than by mobility. It is not easy to conceive of a structure having the potentiality of operating at higher frequencies.

It is evident that the difficulty of making small structures increases with the number of electrodes. For example, it is now possible to make diodes which give usable rectification at frequencies above 10^{10} cps. In these the "working volume" is a very thin layer under the metal point. The thickness of this layer is controlled by surface treatments and the applied voltages. The diameter of the point, which is the minimum dimension mechanically controlled, is much larger than this thickness of the layer. In order to make a transistor of comparable frequency, it would be necessary to make structural elements having dimensions comparable to the thickness of the layer and this would be a much more exacting task than making the diode.

These considerations point out the importance of giving serious consideration to two-terminal structures as amplifying elements. It is possible, in principle at least, to have structures which are much smaller in one dimension than the other two and which exhibit negative resistance and thus give ac power at frequencies comparable to the reciprocal of the transit time across the small dimension.

The attractiveness of such negative resistance diodes for amplification is enhanced by the possibility of using them in *dissected amplifiers*[1,1] in combination with nonreciprocal elements such as gyrators or Hall effect plates. Combinations of negative resistance elements and nonreciprocal elements can lead to structures having gain and unsymmetrical transmission that simulate conventional amplifiers. The adjective *dissected* has been suggested for them since elements giving power gain are physically separated from those giving one-way transmission.

In this article we shall not consider the possible forms of dissected amplifiers, of which there are a wide variety. Instead we shall give an introductory treatment of some forms of negative resistance that may arise from transit time effects. In some cases the most instructive way of treating the structure is by way of the "impulsive impedance" and we devote most of the next section to considering this method.

2. THE IMPULSIVE IMPEDANCE AND NEGATIVE RESISTANCE

The impulsive impedance $D(t)$ for a two terminal device is defined in terms of its transient response to an impulse of current. Thus if the current through the device is

$$J(t) = J + j(t),$$

where J is the dc current and

$$j(t) = 0$$

except very near $t = 0$ and

$$\int j(t)\, dt = \delta Q,$$

then the voltage is

$$V(t) = V + v(t),$$

where V is the dc voltage and

$$v(t) = \delta Q\ D(t).$$

In other words, if in addition to the dc biasing current, a charge δQ is instantaneously forced through the circuit at time $t = 0$, the added voltage is $D(t)$ per unit charge. These equations also serve to introduce the notation used in this article:

Notation. In general, quantities that are functions of time or position will have the functional dependence explicitly indicated. In Sections 4 and 5, however, the symbol δ will be used to distinguish the transient parts δE and $\delta\rho$ from the dc parts of the electric field and charge density.

Capital $V(t)$ and $J(t)$ stand for total voltages and current. Without functional dependence upon (t) they are the dc parts. Similarly $v(t)$ and $j(t)$ are the ac or transient parts. A sinusoidal disturbance is represented by

$$v(t) = v \exp i\omega t,$$

$$j(t) = j \exp i\omega t.$$

where v and j are not functions of time. Where it is necessary to distinguish the displacement current at a particular location from the conduction current, as in the next section, we shall write

$$j(D,\ S_2,\ t),$$

meaning the displacement current across space charge region S_2 as a function of time.

In this section we shall treat J and j as circuit currents. In subsequent sections, we shall be concerned with current densities and shall use the same symbols.

The complex impedance of the device is evidently

$$Z(\omega) = v/j,$$

where v and j are the coefficients in the sinusoidal case.

In terms of the system of notation introduced above, $Z(\omega)$ may also be expressed in terms of $D(t)$ by expressing $j \exp i\omega t$ in terms of increments of charge

$$dQ = je^{i\omega t}\, dt,$$

and summing over all increments up to time t. This leads to

$$Z(\omega) = \int_0^\infty D(t) \exp(-i\omega t)\, dt.$$

A negative resistance will occur if

$$0 > \int_0^\infty D(t) \cos \omega t\, dt = (-1/\omega) \int_0^\infty D'(t) \sin \omega t\, dt,$$

the latter form coming from integration by parts for the case of $D(\infty) = 0$, the only situation treated in this article.

The use of $D(t)$ in analysing the potential merits of diode structures from the point of view of negative resistance is illustrated in Fig. 2.1. Here three cases of $D(t)$ together with certain cosine waves are shown. It is seen for case (a) that a negative real part of Z will be obtained. For case (b), the real part of Z is zero for the frequency shown; this represents a limit; for other frequencies, a positive real part will be ob-

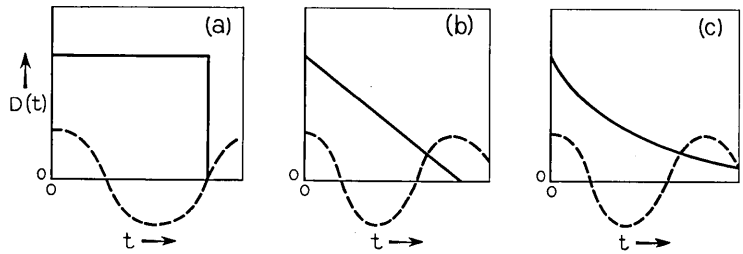


Fig. 2.1 — Some hypothetical $D(t)$ characteristics.

tained. Case (c) represents an exponential fall such as might occur for a capacitor and resistor in parallel. We shall discuss this example below.

The conclusions regarding (a) and (b) may be somewhat more easily seen from the corresponding $-D'(t)$ plots shown in Fig. 2.2. From part (a) it can be seen that the negative maximum in the sine wave at the end of the rectangular $D(t)$ plot is particularly favorable. From part (b) it is seen that no choice of ω will result in more negative area of sine wave than positive. For (c) it is evident that each positive half cycle of the sine wave gives a larger contribution than the following negative half cycle and hence that a positive resistance will be obtained.

For case (c), it is instructive to obtain the value of Z analytically by using

$$D(t) = C^{-1} \exp(-t/RC).$$

This leads correctly to

$$Z(\omega) = (R^{-1} + i\omega C)^{-1}.$$

For small values of ωRC, Z reduces to R; furthermore, for this case, the decay of $D(t)$ occurs while $\cos \omega t = 1$. Under these conditions

$$Z(\omega) = \int_0^\infty D(t)\, dt.$$

This result is useful for estimating the effect of quickly decaying contributions to $D(t)$. These evidently contribute a positive resistance to Z equal to the area under the $D(t)$ curve.

From these considerations it follows that an upward deviation from the linear fall in Fig. 2.1(b) towards Fig. 2.1(a) will result in negative resistance. In Sections 4 and 5 we shall see how particular structures may lead to such favorable, convex-upwards characteristics for $D(t)$.

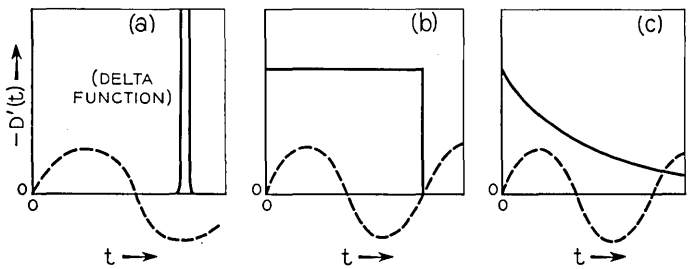


Fig. 2.2 — The $-D'(t)$ characteristic corresponding to fig. 2.1.

3. MINORITY CARRIER DELAY DIODE

As a first example we shall consider the behavior of the device shown in Fig. 3.1. We have chosen a p-n-p structure rather than an n-p-n so as to deal with positively charged carriers and thus avoid numerous minus signs in the equations. In this figure we have used capital letters P and N to designate specific regions, reserving the small letters to indicate carrier densities and conductivity types.

Several features that simplify the theoretical treatment should be noted:

(a) The P_1N junction is 100-fold more heavily doped on the P_1-side.

(b) The doping in the layer N varies exponentially with distance by a factor of 10 across the layer.

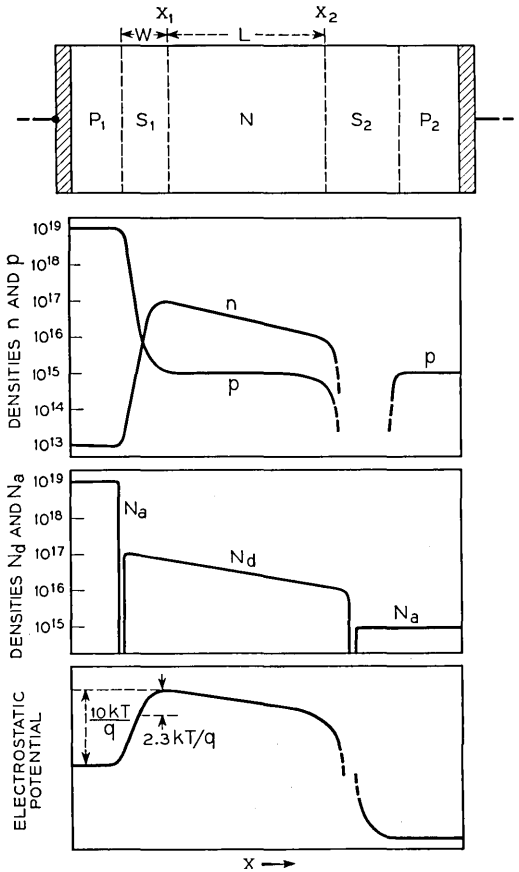


Fig. 3.1 — Constitution of minority carrier delay diode.

(c) Throughout N the concentration of holes is less by a factor of 10 than the electron concentration.

(d) The thickness of N is large compared to the depth of space charge penetration into it.

(e) The voltage drop across the space charge region S_2 is large compared to the other voltage drops.

The conditions lead at the operating frequency to the following consequences:

(1) The current across the first junction is carried preponderantly by holes.

(2) The hole drift in N is substantially unaffected by the ac field and thus represents the delayed diffusing and drifting current injected across the first junction.

(3) The ac voltage drop occurs chiefly across S_2.

We shall show below (Section 3.2) how (a) to (e) lead to (1) to (3), but we shall first bring out the importance of (1) to (3) by using them to give a simple treatment of the impedance of the diode.

3.1. *Calculation of Impedance.*

If the total current is

$$J(t) = J + je^{i\omega t}, \tag{3.1}$$

then the ac hole current across S_1 is also in the notation discussed in Section 2 with the addition of the symbol p to indicate holes

$$j(p, S_1, t) = je^{i\omega t}. \tag{3.2}$$

This current flows through the n-layer unaffected by the ac field and arrives at S_2 delayed and attenuated by a complex factor

$$\beta = |\beta| \exp(-i\theta). \tag{3.3}$$

Because of the high field in S_2, the transit time there is negligible so that the hole current arriving at P_2 is

$$j(p, S_2)e^{i\omega t} = \beta je^{i\omega t}. \tag{3.4}$$

In addition to this current, there is a dielectric displacement current in S_2 which is converted to hole conduction current in P_2. If the voltage drop across S_2 is

$$V(S_2, t) = V(S_2) + v(S_2)e^{i\omega t}, \tag{3.5}$$

then the ac displacement current is

$$j(D, S_2)e^{i\omega t} = i\omega C_2 v(S_2)e^{i\omega t}. \tag{3.6}$$

Now the total current is constant through the device, hence

$$j = j(p, S_2) + j(D, S_2) \tag{3.7}$$

which leads to

$$j = i\omega C_2 v(S_2)/(1 - \beta). \tag{3.8}$$

If $v(S_2)$ is substantially equal to the ac voltage across the unit, then the impedance is

$$\begin{aligned} Z &= v(S_2)/j = (1/i\omega C_2) + i\beta/\omega C_2 \\ &= (1/i\omega C_2) + (1/\omega C_2) \mid \beta \mid \exp i[(\pi/2) - \theta]. \end{aligned} \tag{3.9}$$

Evidently if $\theta > \pi$ and $\theta < 2\pi$, the second term will have a negative real part so that the diode will act as a power source.

If we neglect the ac electric field in N then β may be calculated in terms of the thickness $L = x_2 - x_1$ of the layer and the potential drop across the layer. This latter arises from the concentration ratio N_{d1}/N_{d2} between the two sides of N. Since the donor charge density is neutralized substantially entirely by electrons, and since almost no electron current flows, the electron concentration difference must result from a Boltzman factor (at $10^{17}/\text{cm}^3$ Fermi-Dirac statistics are not needed) and this leads to

$$\Delta V_n = (kT/q)ln(N_{d1}/N_{d2}) \tag{3.10}$$

for the potential drop across N. In N the electric field is thus

$$E = \Delta V_n/L. \tag{3.11}$$

The differential equation for hole concentration for a disturbance of frequency ω is

$$\dot{p} = i\omega p = -\mu p E \frac{\partial p}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2}. \tag{3.12}$$

This linear differential equation has two linearly independent solutions. These must satisfy at x_2, the left edge of the space charge layer S_2, the boundary condition that the hole density is practically zero.[3.1] The appropriate solution is

$$p = e^{i\omega t}\,[e^{k_1(x-x_2)} - e^{k_2(x-x_2)}], \tag{3.13}$$

where

$$Lk_1 \equiv (x_2 - x_1)k_1 = \alpha[1 + (1 + i\gamma)^{1/2}], \tag{3.14}$$

$$Lk_2 \equiv (x_2 - x_1)k_2 = \alpha[1 - (1 + i\gamma)^{1/2}] \tag{3.15}$$

where

$$\alpha = q\Delta V/2kT, \tag{3.16}$$

$$\gamma = 4\omega D_p/u^2, \tag{3.17}$$

$$u = \mu_p E = \mu_p \Delta V/L. \tag{3.18}$$

The current is

$$j(p, x, t) = q(up - D_p \partial p/\partial x) \tag{3.19}$$

and the ratio of currents at x_1 and x_2, which is β by definition, is

$$\beta = j(p, x_2, t)/(j(p, x_1, t),$$

$$= \frac{Lk_1 - Lk_2}{Kk_1 \exp(-Lk_2) - Kk_2 \exp(-Lk_1)} \tag{3.20}$$

$$= \frac{2(1 + i\gamma)^{1/2} e^{\alpha}}{[1 + (1 + i\gamma)^{1/2}] \exp \alpha(1 + i\gamma)^{1/2} - [1 - (1 + i\gamma)^{1/2}] \exp - \alpha(1 + i\gamma)^{1/2}}.$$

The phase lag in β must exceed 180° or π in order to give negative resistance. It can be seen that this phase factor must result from the first exponential in the denominator by the line of reasoning suggested below: The real part of the exponent is larger than the imaginary part. Hence the absolute ratio of the two exponentials is at least 2π. For this condition the second term in the denominator is negligible compared to the first. Hence the phase of (3.20) is determined largely by the first exponential. As a helpful approximation we may neglect the second term and write

$$\beta \doteq \frac{2(1 + i\gamma)^{1/2} \exp[\alpha - \alpha(1 + i\gamma)^{1/2}]}{1 + (1 + i\gamma)^{1/2}}. \tag{3.21}$$

Two limiting cases are worthy of special note:

(I) $\alpha \to 0$, uniform n-layer, $\gamma \to \infty$.

$$\beta \doteq 2 \exp -\alpha(i\gamma)^{1/2} = 2 \exp -(1 + i)(\omega/2D)^{1/2} L. \tag{3.22}$$

(II) $\alpha \to \infty$, $q\Delta V/kT \gg 1$, $\gamma \to 0$.

$$\begin{aligned} \beta &= \exp[-i\alpha\gamma/2 - \alpha\gamma^2/8], \\ &= \exp[-i(\omega L/u) - (DL/u)/(u/\omega)^2]. \end{aligned} \tag{3.23}$$

These expressions may be interpreted as follows: In case (I), flow is by diffusion and the propagation factors k_1 and k_2 take the form $\pm(\alpha/L)(i\gamma)^{1/2}$. For this case attenuation and delay terms in the exponential are equal, and the largest negative term occurs in Z when the phase angle is $5\pi/4$ (as may be verified by differentiation.) This leads to

$$\beta = 2(-1 + i)2^{-1/2} \exp(-5\pi/4) = 0.028(-1 + i), \quad (3.24)$$

which gives

$$Z = (1/\omega C_2)(-0.028 - i\,1.028), \quad (3.25)$$

the impedance of a condenser with a negative Q of 37. In order to make an oscillator by coupling this to an inductance, an inductance with a Q of more than 37 must be used. It is obviously advantageous to reduce the magnitude of the negative Q.

For case (II) in its ideal form, the ac current simply drifts through the n-layer without attenuation. This produces a phase lag of ω times the transit time L/u. If this were the only effect involved, a capacitor with a negative Q of less than unity could be produced. In addition, however, there is attenuation due to spreading by diffusion. This effect is dependent upon the ratio of the spread by diffusion $(DL/u)^{1/2}$ to the separation of planes of equal phase in the drifting hole current. This latter separation is $2\pi u/\omega$. The square of this ratio appears in the attenuation term in the second form of β.

We shall estimate the effect of the attenuation term by taking

$$\alpha\gamma/2 = 3\pi/2, \quad (3.26)$$

so that the desired phase shift is obtained. The attenuation term is $\gamma/4$ smaller than this so that if $\gamma/4$ is considerably less than one, the attenuation in β will be small while the phase shift is correct. If we take $3\pi/2$ for the value of $\alpha\gamma/2$, then the value of γ becomes

$$\gamma = 4\omega D/u^2 = 6\pi kT/q\Delta V. \quad (3.27)$$

Thus the approximation on which (II) is based fails unless $q\Delta V/kT > 18$, a value which implies an enormous range of concentration in the n-layer. We must, therefore, investigate the case of gradients in the n-layer by more complete algebraic procedures.

We shall denote by $-\theta_1$ the phase shift in β due to the exponential in equation (3.21). The total phase shift θ is somewhat less since the algebraic expressions give a small positive phase shift of at most about 15°, which vanishes for large and small values of γ. Similarly the attenuation of β arises chiefly from the real part of the exponential since the absolute

value of the algebraic expressions lies between 2 for $\gamma = 0$ and 1 for $\gamma = \infty$.

It is instructive to express the real part of the exponent in terms of α and θ_1. This is done as follows:

$$\alpha - \alpha(1 + i\gamma)^{1/2} = -\eta - i\theta_1 . \tag{3.28}$$

This can be solved for η and $(1 + i\gamma)^{1/2}$:

$$\eta = (\alpha^2 + \theta_1^2)^{1/2} - \alpha, \tag{3.29}$$

$$(1 + i\gamma)^{1/2} = [(\alpha^2 + \theta_1^2)^{1/2} + i\theta_1]/\alpha. \tag{3.30}$$

From there it is seen that for a fixed value of θ_1, the attenuation can be greatly reduced by increasing α. Unfortunately, this requires very large changes in concentration. For example with $\theta_1 = 3\pi/2$ and $\alpha = \theta_1$, the value of η is reduced to $\eta = 0.414\ \theta_1$. However, the value of potential difference is

$$q\Delta V/kT = 3\pi, \tag{3.31}$$

giving

$$N_{d1}/N_{d2} \doteq 10^4. \tag{3.32}$$

For the case shown in Fig. 3.1, $\alpha = 1.15$ and for $\theta_1 = 5\pi/4 = 3.9$ we obtain

$$\eta = 2.95 = ln_e\, 19.5 \tag{3.33}$$

This is an improvement of about 1 factor of e in the exponential compared to having $\Delta V = 0$. The value of β is

$$\beta = 0.082 \times \exp(-i\, 218°), \tag{3.34}$$

and this leads to

$$Z = (\omega C_2)^{-1}(-0.05 - i\ 1.065). \tag{3.35}$$

Thus at the operating frequency, the diode appears to be a capacitor with a negative Q of 21.

Increasing the concentration change to a factor of 100, so that $\alpha = 2.3$, gives

$$\eta = 2.25, \tag{3.36}$$

$$\beta = 0.18 < -220°, \tag{3.37}$$

$$Z = (\omega C_2)^{-1}(-0.116 - i\ 1.14), \tag{3.38}$$

$$Q = -10. \tag{3.39}$$

The calculations indicate that attenuation can be controlled to a considerable degree while maintaining the desired phase shift.

3.2. *Justification of Consequences (1), (2) and (3)*

In germanium at room temperature the product np is about 10^{27} under equilibrium conditions. At the first junction of Fig. 3.1 it is 10^{32}, implying a forward bias[3.2] of (kT/q) 2.3 $\times$ 5. In order to maintain this forward bias a flow of electrons must be furnished to N. There are several ways of accomplishing this. In the first place, the reverse bias across S_2 draws a reverse current of thermally generated electrons from P_2. This current can be controlled by controlling the lifetime and temperature in the P_2 region. Alternatively, electrons may be injected into P_2; some of these will diffuse to S_2 and arrive at N. Still another means of controlling the bias across S_1 is to make contact to N itself. Since only the dc bias need be controlled, the series resistance across N itself is unimportant; the source should be of high impedance.

The decrease in density of 10^4 across the junction in carrier concentration implies a potential difference of 9.2 (kT/q). Most of this potential difference occurs where the carrier concentration is negligible. Hence the space charge theory may be applied. Furthermore, the acceptor concentration is much higher than the donor concentration. Hence the space charge extends chiefly into the donor region and we may write[3.3]

$$\Delta V_1 = (2\pi q N_d/\kappa) W^2 \tag{3.40}$$

for the relationship between width W of the space charge region and voltage drop ΔV.

If this voltage drop has an ac component, then a charging current will be required to change W. This current is determined by the admittance

$$\omega C = \omega\kappa/4\pi W \tag{3.41}$$

of the space charge region.

At the same time injected hole and electron currents flow across the junction. The admittance associated with the hole current is approximately

$$A = (i\omega/D)^{1/2}\, \sigma_{p1}\,, \tag{3.42}$$

where σ_{p1} is the hole conductivity just inside the n-layer.[3.4] Actually, as discussed below, the admittance is somewhat higher.

The ratio of the admittances is

$$\frac{|A|}{\omega C} = \left[\frac{4\pi}{K}\,\frac{\omega\sigma_{p1}^2}{D}\cdot\frac{4\pi W^2}{\kappa\omega^2}\right]^{1/2}, \qquad = \left[\frac{4\pi\sigma_{p1}}{K\omega}\cdot\frac{q\Delta V}{kT}\cdot\frac{\sigma_{p1}}{\mu_P q N_D}\right]^{1/2}. \tag{3.43}$$

For our example this expression is much greater than 1 as may be seen as follows: The first fraction is the ratio of the dielectric decay constant to ω. This is 10^3 or more larger than ω need be. The next term is about 10 and the last term is the ratio of hole to electron density at x_1 and is about 10^{-2}. Hence the ratio of impedances is about 10:1.

We shall next consider why the expression for A for holes must be examined more closely. The admittance formula used above applies to the case of zero field to the right of the junction. The aiding field will increase the flow of holes into the n-layer and raise the admittance somewhat. Correcting for this will increase A in respect to ωC and will thus strengthen rather than weaken the argument.

Also in the expression for A, no account was taken of the transit time across the region W. If we assume a uniform field in this region for purposes of making estimates, then the solution of equation (3.20) may be applied. Since now u corresponds to drift velocity due to $9kT/q$ of voltage drop across W which is much less than L in length, it is evident that γ will be less by a large factor in this region compared to its value in N. This leads to the conclusion that phase lags will be unimportant in this region.

We have neglected the effect of electron injection into P_1. By the customary arguments for unsymmetrically doped junctions, it follows that this current is very small compared to the hole current.

This justifies consequence (1).

Consequence (2) may be justified as follows: At x_1 all the ac current $j \exp(i\omega t)$ is carried by holes. If a pure drift case occurred, the hole current might be reversed at some point in the n-layer and be $-j \exp(i\omega t)$. Under these conditions the electron current would have to be $2j \exp(i\omega t)$. Under no conditions, however, will the electron current be larger than this. This maximum possible electron current will require an electric field and this field will also affect the hole flow. Since the electron conductivity is at least 10 times larger than the hole conductivity, the hole current due to the ac field will only be about $\frac{1}{10}$ of j at most. Thus the hole current is only slightly affected by the ac field.

Consequence (3) follows from the fact that the reverse biased junction

S_2 has much higher impedance than S_1 . S_1 has higher impedance than that for hole injection into N. The impedance for hole injection into N corresponds to the hole conductivity in the n-layer over a distance comparable with the thickness of N. However, the impedance of N itself is that due to the much larger number of electrons in it and is thus much less than the impedance of S_1 . Thus it follows that the impedances across S_1 and across N are much less than across S_2 . This conclusion is not affected by the modification of impedance of S_2 due to hole flow across it.

3.3. *Modifications*

The treatment presented above has been based upon the conditions (a) to (e). Some of these are advantageous from the point of view of operation but others have been introduced to simplify the treatment. Among the latter is the condition that the current across S_1 is carried chiefly by holes. If the current were chiefly capacitative at this junction, then the voltage would lag 90° behind the current. This adds a desirable phase lag in the hole injection across S_1 and thus requires less phase shift in the n-layer. By suitably adjusting the ratio of capacitative and inductive admittances, a net improvement in Q may be obtained.

4. THE TRANSIENT RESPONSE IN A UNIPOLAR STRUCTURE

In the previous section the electric field produced by the injected holes had a negligible influence on the motion of the injected holes. In effect

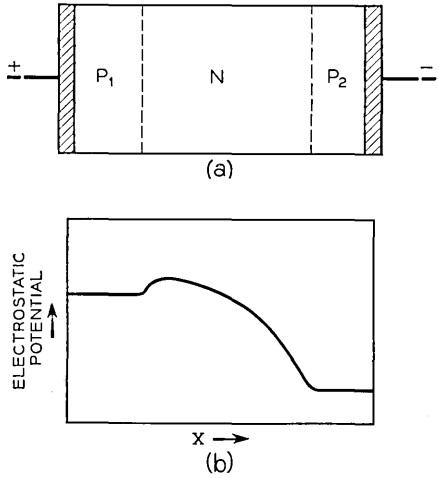


Fig. 4.1 — Space charge limited hole flow.

this was due to the bipolar nature of the mode of operation considered, the majority carriers in the region N acting to shield the minority carriers from their own space charge.

In this section we shall deal with unipolar diodes in which only one type of carrier is present in sufficient number to have a major effect. In these the influence of the space charge of the carriers upon their motion plays an important role.

Fig. 4.1 illustrates one example of the type of structure covered by the theory of this section. It is again a p-n-p structure like that considered in Section 3. However, in this case the dimensions, the donor density and the applied potential are such that the space charge "punches through" the device.[4.1] Under these circumstances a condition of space charge limited emission is set up so that holes are injected from the positive region P_1 to just such an extent that their flow is limited by their own space charge. This limitation is associated with the maximum of potential just inside N.

The potential maximum is evidently a "hook" for electrons generated thermally in P_2 and in N. Under some circumstances electrons may accumulate and form a layer in which there is no electron flow and hole flow is carried equally by diffusion and drift. Such *stagnant* regions will tend to be suppressed if P_1 is made of short lifetime material, so that electrons are siphoned out of N, or if p at the maximum is larger than p for intrinsic material and the lifetime is locally low.

We shall treat the transient response of this structure of Fig. 4.1 from the point of view of the *impulsive impedance* discussed in Section 2. Accordingly we suppose that a steady current J flows per unit area. At $t = 0$ an added pulse of current occurs carrying a total charge of δQ_i per unit area, the subscript "i" signifying initial condition. Our problem is to determine how this added charge is carried by a transient disturbance in the hole flow and what is the resultant dependence of voltage upon time; by definition the added voltage across the device is

$$v(t) = \delta Q_i D(t). \tag{4.1}$$

Since we are dealing with a planar model, we shall suppose that the initial condition at $t = 0$ corresponds to added charges δQ_i and $-\delta Q_i$ on the metal plates on the P-regions. These charges set up an added field

$$\delta E_i = \delta Q_i / K, \tag{4.2}$$

where

$$K = \kappa \epsilon_0 \tag{4.3}$$

in MKS units. The initial value $v(0)$ is then simply δE_i times the total width of the structure.

The first effect, which takes place in a negligible time in respect to the frequencies involved, is the dielectric relaxation of the field in P_1 and P_2. The added current due to δE_i leads to an exponential decay of δE in these regions with a transfer of δQ_i and $-\delta Q_i$ to the two boundaries of N. If P_1 and P_2 are thin compared to N, the resulting drop in $v(t)$ is small. In any event it can be shown by the reasoning at the end of Section 2 that this contribution to $D(t)$ adds simply the series resistance of P_1 and P_2 to the impedance.

The next effect is the transport of δQ_i on left side into N by hole flow over the potential maximum. It will be easier, however, to discuss this process after the treatment of the transient effects that occur in N itself. Consequently, we shall at this point assume that after a time, short compared with the important relaxation time in the structure, the disturbance of hole density is as shown in Fig. 4.2(a).

Fig. 4.2(a) shows added charges $+\delta Q_i$ and $-\delta Q_i$ produced by a disturbance denoted as δp in the hole density. The charge $-\delta Q_i$ on the right side is produced by an increased penetration of the space charge into P_2; it is similar to that produced by increasing reverse bias on a p-n junction.

Fig. 4.2(b) shows the corresponding disturbance in electric field. This disturbance is denoted by δE which is a function of x and t. Evidently

$$v(t) = \int_0^L \delta E(x, t)\, dx. \tag{4.4}$$

and this is the area under the δE curve.

The other parts of the figure indicate qualitatively a subsequent stage in the motion and decay of δp and δE. Our problem is to formulate mathematically this decay process. We shall treat the decay process in terms

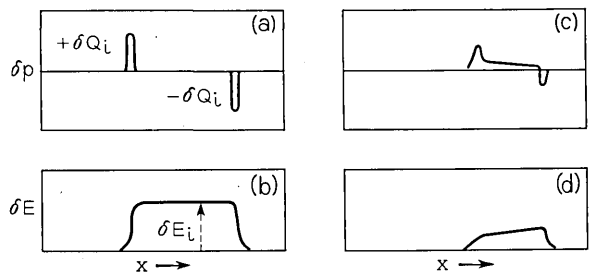


Fig. 4.2 — The initial stage and a subsequent stage of the transient.

of the effect of drift in the electric field and neglect the effects of diffusion. This procedure can be justified by the fact that as soon as a hole had reached a point where the potential has fallen by kT/q below the maximum, its flow is governed by drift rather than diffusion and the predominance of drift continues to increase towards the right.[4.2]

If drift in the field is the predominant cause of hole flow, then the equations governing the situation in N are

$$J + \delta J = (\rho + \delta\rho)(u + \delta u), \tag{4.5}$$

where the terms with δ represent the transient effects and those without represent the steady state solution, $\rho = qp$ is the charge density of the holes and u their drift velocity. The equation for the change of E with distance is

$$K(\partial/\partial x)(E + \delta E) = \rho_f + \rho + \delta\rho, \tag{4.6}$$

where ρ_f is the fixed charge density due to donors and acceptors. (We neglect any effect of traps.) The steady state equation for E is thus

$$K(dE/dx) = \rho_f + J/u. \tag{4.7}$$

In a region where ρ_f is independent of x, this equation may be reduced to quadratures by writing

$$K\, dE/(\rho_f + J/u) = dx; \tag{4.8}$$

the left side is then a known function of E through the dependence of u upon E.

It is convenient to introduce a time-like variable s which is the transit time for the dc solution. Evidently

$$ds = dx/u = K\, dE/(\rho_f u + J). \tag{4.9}$$

For the case of space charge limited current, s may be conveniently measured from the potential maximum. Even though the solution is invalid at that point, the integrals converge and the contribution from the region within kT/q of the maximum is small.

We shall assume that the equations for the steady state case have been solved and that the functional relationships are known between E, x, v and s.

The differential equation for δE may then be obtained as follows: To the left of the pulse in δp in Fig. 4.2(a), δE is zero. From equation (4.6) we have

$$K\partial E(x)/\partial x = \delta\rho. \tag{4.10}$$

Integrating this from the region where E is zero gives

$$K\delta E(x, t) = \int_0^x \delta\rho(x, t)\, dx. \tag{4.11}$$

Equation (4.11) states that the dielectric displacement at x is equal to the excess charge between the potential maximum and x. Evidently during the transient following Fig. 4.2(a), the rate of change of this extra charge is $-\delta J(x, t)$ since the dc current is flowing in at the left and an excess current δJ flows out at the right. Hence we have

$$\begin{aligned} K\partial\delta E/\partial t &= -\delta J, \\ &= -(\delta\rho u + \rho\delta u). \end{aligned} \tag{4.12}$$

For the change in drift velocity we may write

$$\delta u = (du/dE)\ \delta E = \mu^*\delta E. \tag{4.13}$$

For high electric fields u increases less rapidly than linearly with E and μ^* is less than the low-field mobility.[4.3] For very high fields μ^* is nearly zero and there are theoretical reasons for thinking that there may be a range in which μ^* is negative. We shall return to this point in the next section.

In Fig. 4.3 we show a diagrammatic representation of the transient so-

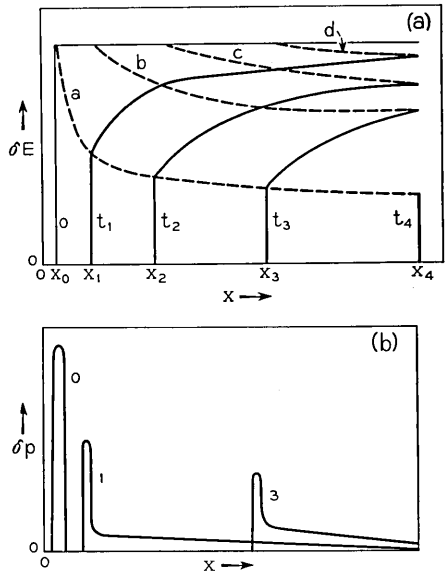


Fig. 4.3 — Graphical representation of the dependence of δE upon time.

lution. Each of the dashed lines represents the decay of δE as measured in a moving coordinate system: Thus we consider δE measured at a position $x(s_0 + t)$; this is a position that moves with the dc velocity u. This δE is evidently expressed in terms of $\delta E(x, t)$ by writing $x = x(s_0 + t)$:

$$\delta E \text{ in moving system} \equiv \delta E_m(s_0, t) = \delta E[x(s_0 + t), t]. \tag{4.14}$$

The differential equation for δE_m is

$$\begin{aligned} (\partial/\partial t)\, \delta E_m &= (\partial \delta E/\partial t)_x + (\partial \delta E/\partial x)_t \partial x/\partial t, \\ &= (\partial \delta E/\partial t)_x + (\partial \delta E/\partial x)_t u, \\ &= -(u\delta\rho + \rho\delta u)/K + (\delta\rho/K)u, \\ &= -(\rho\mu^*/K)\delta E = -\nu\delta E, \end{aligned} \tag{4.15}$$

where the quantity

$$\nu \equiv \rho\mu^*/m \tag{4.16}$$

is an effective dielectric relaxation constant being the *differential conductivity* $\rho\mu^*$ divided by the permittivity K.

Evidently ν is a function of position x only and may be expressed as $\nu(s)$ through the dependence of x upon s. Thus we may write

$$(\partial/\partial t)\delta E_m(s_0, t) = -\nu(s_0 + t)\delta E_m(s_0, t) \tag{4.17}$$

which has a solution

$$\delta E_m(s_0, t) = \delta E_m(s_0, 0) \exp\, [-g(s_0 + t) + g(s_0)], \tag{4.18}$$

where

$$g(s_0 + t) = \int_{s'}^{s_0+t} \nu(s)\, ds. \tag{4.19}$$

The lower limit s' is chosen for convenience so as to avoid singularities in $g(s)$. This integration shows that δE_m decays exponentially as the electrical field would decay in a material whose dielectric relaxation constant changed with time just as ν changes as observed on the moving plane.

Fig. 4.3 shows on the dashed lines the decay of δE_m on the moving planes. Since δE_m is zero to the left of the initial pulse in Fig. 4.2(a), it remains zero on all moving planes which follow the pulse of δQ_0. This justifies the statement made earlier. The solid curves labelled t_1, t_2 etc. show the spatial dependence of δE for times t_1, t_2, etc. after the charge δQ_1 is added.

The values of the transient voltage $v(t)$ at time t_1, for example, is the

integral under the curve t_1. This curve is zero for $x < x(t_1)$ and for $x > x(t_1)$ it is

$$\delta E(x, t_1) = (\delta Q_i/K) \exp [- g(s_0 + t_1) + g(s_0)], \tag{4.20}$$

where

$$x = x(s_0 + t). \tag{4.21}$$

If the total transit time across N is S so that

$$x(S) = L, \tag{4.22}$$

then

$$v(t_1) = \int_{x(t_1)}^{L} \delta E(x, t_1)\, dx. \tag{4.23}$$

From this expression we can derive the desired formula for $D(t)$. For this purpose the integral over dx is replaced by an integral over s. At time t the range of s is evidently from t to S and $dx = u(s)\, ds$. From this we obtain:

$$\begin{aligned} D(t) &= v(t)/\delta Q_i\,, \\ &= (1/K) \int_s^t \exp [- g(s) + g(s - t)] v(s)\ ds. \end{aligned} \tag{4.24}$$

From Fig. 4.3 we can see that there are competing tendencies in the decay of $D(t)$ some of which tend to produce the desired convex shape discussed in Section 2 and others the concave shape. The effect of the dielectric relaxation constant is adverse and tends to produce an exponential decay. On the other hand the advance of the pulse of holes from left to right in Fig. 4.2 proceeds in an accelerated fashion with the result that the range of x over which δE is not zero decreases at an accelerated rate. If the dielectric relaxation were zero, this would result in the desired convex upwards shape.

The resultant shape of the $D(t)$ curve is thus sensitive to the exact relationship of the transit time and dielectric relaxation. This can be illustrated by giving the results of analysis for a p-n-p structure, neglecting diffusion and considering μ to be constant. The solutions of the dc equations are readily obtained for this case and have been published.[4.4] For convenience we repeat them here:

$$E = (J/\mu\rho_f)(e^{\alpha s} - 1), \tag{4.25}$$

$$x(s) = \mu JK(\mu\rho_f)^{-2} (e^{\alpha t} - \alpha t - 1), \tag{4.26}$$

$$L = x(s) = (JK/\mu\rho_f^2)(e^{\beta} - \beta - 1), \tag{4.27}$$

$$\beta \equiv \alpha s.$$

From these it is found that

$$lng(s) = (1 - e^{-\alpha s}). \tag{4.29}$$

This leads to

$$\begin{aligned} D(t) &= (J/\mu\rho_f^2)\,[e^{\beta} + e^{\alpha t}\,(\alpha t - \beta - 1)] \\ &\equiv (J/\mu\rho_f^2)\,D(\beta, \alpha t). \end{aligned} \tag{4.30}$$

For $t = 0$ this reduces correctly to L/K.

Figure 4.4 shows the resulting shape of the D curves with β as a parameter. Large values of β correspond to cases in which the hole charge density is small compared to ρ_f and to relatively long relaxation constants. For them the desired convex upward shape results.

Figure 4.5(a) and 4.5(b) show the real and imaginary parts of the impedance expressed in terms of $Z(\beta, \theta)$:

$$\begin{aligned} Z(\omega) &= \int_0^T e^{-i\omega t} D(t)\, dt \\ &= (KJ/\mu^2\rho_f^3) Z(\beta, \theta), \end{aligned} \tag{4.31}$$

$$\theta = \omega T. \tag{4.32}$$

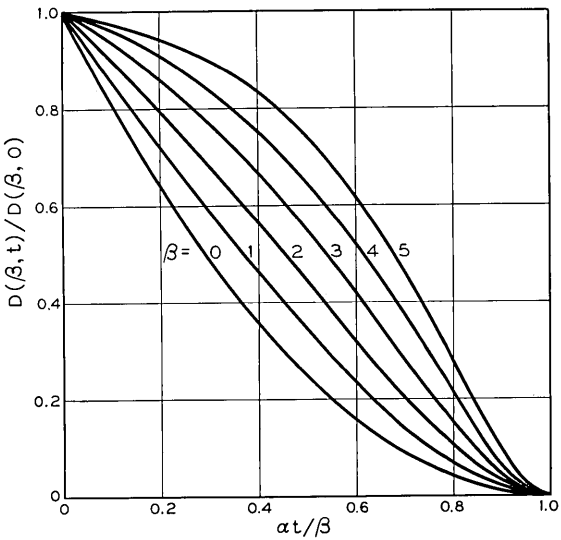


Fig. 4.4 — Impulsive impedance for various values of β in p-n-p structure.

It is seen that values of $-Q$ as small as about 10 can be obtained for $\beta \geqq 3$.

In the next section we shall consider modifications which may result from variations in μ^* and for changes in geometry.

We must return to the question of how the charge $+\delta Q_i$ passes the potential maximum. In order that the theory given above apply, it is necessary that the time required for δQ_i to enter the drift region be short compared to the transit time. At the potential maximum the charge density may be estimated by the methods previously dealt with in the theory of space charge limited emission. Initially $+\delta Q_i$ appears to the left of the maximum and the field at the maximum is δE_i. This field will

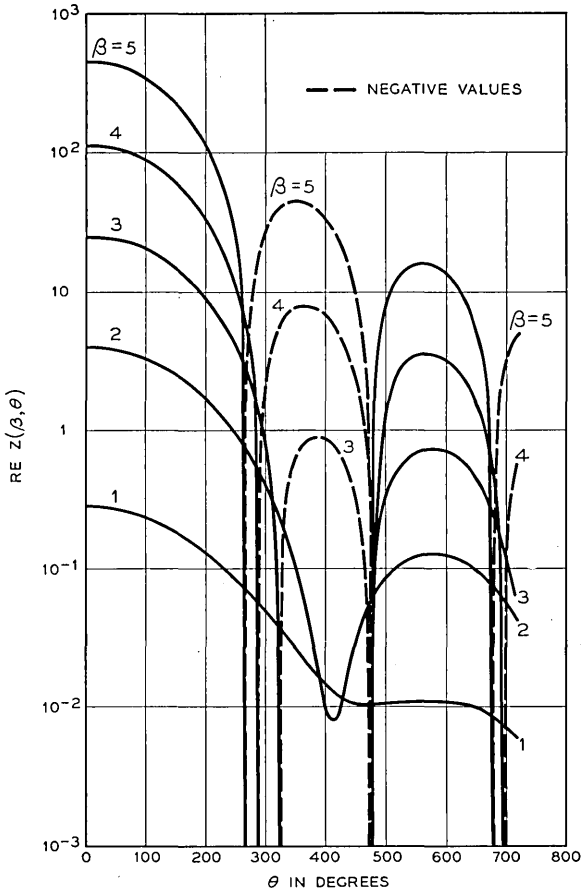


Fig. 4.5 — Impedance of a p-n-p structure. (c) Real part of impedance.

then relax with a relaxation constant of about $\mu\rho(\max)/K$ where $\rho(\max)$ is the hole charge density. Actually the relaxation may be somewhat quicker because the concentration gradient of the added holes also contributes to the flow over the maximum. Since the charge density at kT/q below the maximum is comparable to that at the maximum the entire relaxation process will proceed at about this rate. Thus a criterion for the applicability of the theory is that $K/\mu\rho(\max)$ be much less than S, the transit time or total decay time for $D(t)$.

5. MOBILITY AND GEOMETRY EFFECTS

5.1. *The Effect of a Region of Negative* μ^*

In very high electric fields holes may be expected on the basis of theory to exhibit a negative value of μ^*. This theory[5.1] is founded on the idea that a hole can lose energy to phonons at a certain maximum aver-

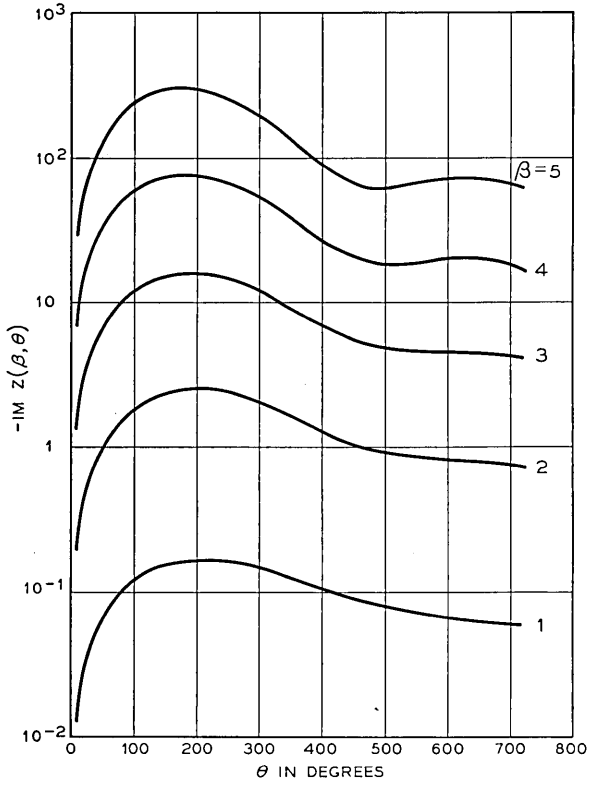


Fig. 4.5 — Impedance of a p-n-p structure. (b) Imaginary part of impedance.

age rate P_{max} . (P_{max} is the *staukonstante* of Krömer.) The holes probably achieve this rate when their energy is near the middle of the valence band. Under these conditions the power input from the electric field must be no greater than P_{max}:

$$qEu \leqq P_{max} \tag{5.1}$$

From this it follows that

$$u \leqq P_{max}/qE, \tag{5.2}$$

so that the drift velocity will decrease with increasing field at sufficiently high fields.

Furthermore, if the width of the valence band is less than the energy gap, then a hole cannot acquire enough energy to produce hole electron pairs. Thus in such a case, the negative resistance range should be reached before breakdown effects occur.

In Fig. 5.1 we illustrate the general trends of the u versus E curve, to be expected if the *staueffekt* occurs. As is indicated, the maximum drift velocity will be referred to as u_m . It occurs at a field E_m . Since we are here concerned with principles rather than details, no attempt has been made to indicate the square root range in which u is proportional to $E^{1/2}$. This range has been observed by E. J. Ryder[5.2] and shown by G. C. Dacey[5.3] to control hole flow in space charge limited hole currents in germanium and has been treated theoretically.[5.4] Dacey[5.5] has also investigated the effect of the square root law upon the $D(t)$ curves for the p-n-p structure of Section 4 and reports that the effects are so unfavorable that no negative resistance is to be expected.

The *staueffekt* opens the attractive possibility of making negative resistance devices in which the current decreases with increased dc voltage so that negative resistance will be exhibited over a wide frequency range. Unfortunately, when the boundary conditions are taken

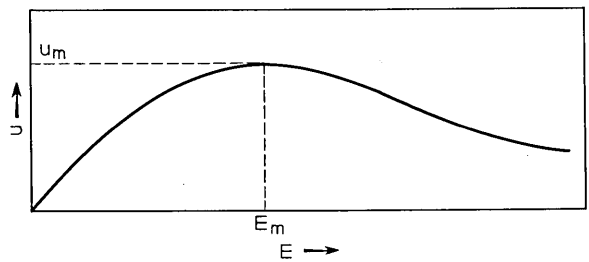


Fig. 5.1 — Qualitative representation of drift velocity versus field as affected by "staueffekt."

into account, it is found that a device in which most of the current flow occurs in a negative μ^* region does not necessarily show a dc negative resistance characteristic. On the other hand, such a structure may have a very favorable $D(t)$ characteristic.

We shall illustrate these conclusions by considering a $(p^+)p(p^+)$ structure having heavily doped ends, so that ohmic non-injecting contacts may be made to the ends. Fig. 5.2 shows the potential distribution and hole distribution for two cases of applied potential. The first case, represented in (a) and (b), corresponds to moderate fields such that the peak velocity u_m is not reached.

In the second case the voltage is so high that the average value of $E = V/L$ exceeds the critical field E_m. Under these conditions u is approximately equal to u_m over a large part of the P-region and a substantial portion of the voltage drop occurs near one end. An increase of the applied voltage occurs chiefly at this end with a small increase in current. Thus no negative dc resistance occurs.

The above conclusions are reached by considering the differential equation for the space charge again as in Section 4 neglecting diffusion and starting with $E = 0$ at the left edge of the P-region. This leads to

$$dx = K\, dE/[(J/u) - \rho_a], \tag{5.3}$$

where we have introduced

$$\rho_a = -\rho_f = -qN_a \tag{5.4}$$

for the charge density of the acceptors. Evidently if J exceeds J_m where

$$J_m = \rho_a u_m, \tag{5.5}$$

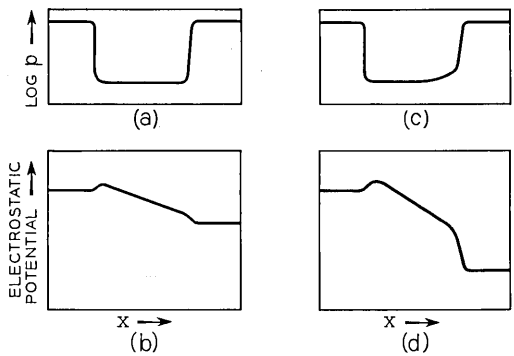


Fig. 5.2 — Hole density and potential distribution including influence of "staueffekt."

the denominator is everywhere positive and x increases monotonically with E. If J is only a little larger than J_m, there will be a large region of x in which E is nearly equal to E_m. Outside of this region, E increases much more rapidly with x. The relative scale of these distances may be estimated as follows: Suppose J is only a little larger than J_m, and consider the situation where E is about twice E_m so that u is about one half of u_m. Then

$$dx/dE \doteq K/\rho_a . \tag{5.6}$$

Under these conditions an increase of E by an additional E_m will require a distance

$$\Delta x \doteq KE_m/\rho_a . \tag{5.7}$$

If this value is much smaller than L, then the situation represented in Fig. 5.2(c) and (d) will occur. On the other hand if L is smaller than Δx, the region of space charge and high field will extend throughout most of the structure.

In any event equation (5.4) leads to positive resistance. This can be seen from the fact that increasing J always means a decrease in x for the same value of E and hence an increase in E at all values of x and thus an increase in voltage at any fixed value of x.

The above conclusion that a positive dc resistance will be exhibited by a structure like that discussed above may also be reached by considering the transient response. The theory of Section 4 may be once at be applied to this case by simply taking account of the fact that ν is negative for part of the structure and thus that δE_m increases with increasing s.

In Fig. 5.3 we illustrate a structure to which these considerations may be relatively simply applied, at least in a limiting case. It consists of four layers, the two outer being p^+ as before. Space-charge limited emission then enters the intrinsic layer which is of such a width that at its right hand boundary the electric field has a value E_3 that exceeds E_m. At this point the hole space charge is

$$\rho_3 = J/u_3 \tag{5.8}$$

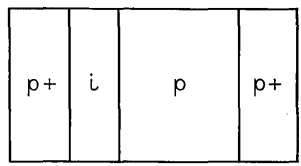


Fig. 5.3 — A structure having a region of uniform negative differential conductance.

where u_3 is $u(E_3)$. In the P-region this space charge is compensated by acceptors to produce a region of uniform field in which μ^* is negative.

If the P-region is wide compared to the I-region, then the transit time through it will also be relatively large. As a consequence δQ will be transferred quickly into the P-region. From that time on δE_m curves, like those of Fig. 4.3, will show an exponential increase with time and also with distance since for this case of constant u in the P-region, time and distance are linearly related. This will lead to a $D(t)$ of the form

$$D(t) = (u_3/K)(S - t) \exp | \mu^* \rho_3/K | t, \tag{5.9}$$

where the absolute value signs emphasize that for this case of negative μ^* there is a build-up in time. This form of D is always convex upwards and, in fact, if

$$S | \mu^* \rho_3/K | > 1, \tag{5.10}$$

it starts with a positive slope at $s = 0$ so that the transient voltage actually builds up initially with time.

Even an initially growing $D(t)$ does not give a negative resistance at low frequencies, however. As shown in Section 2, the dc resistance is simply the integral under the $D(t)$ curve and thus will still have a positive value.

5.2. *Convergent Geometry*

It is possible to obtain marked improvement of the $D(t)$ curves without the aid of the negative values of μ^*. This possibility is based upon convergent geometry. A possible case is illustrated in Fig. 5.4. In this case it is supposed that the field in the inner P-region is so large that a substantial reduction in μ^* has occurred. As a consequence, the decay of field in this region is relatively slow. Furthermore, since both the dc and

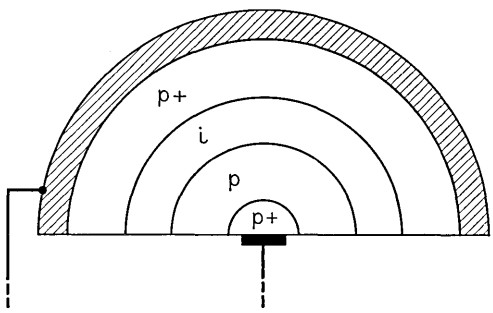


Fig. 5.4 — A convergent flow structure.

transient fields are high in this region, essentially because of the inverse square law, the principal contribution to $D(t)$ comes from this region. These two factors — relatively slow dielectric relaxation near the center and principal contribution to $D(t)$ from near the center — combine to give a $D(t)$ characteristic which holds up well until the pulse of injected holes reaches the inner region. This may result in a favorable convex upwards $D(t)$ characteristic.

ACKNOWLEDGEMENTS

The writer is indebted to a number of his colleagues for helpful discussions and to W. van Roosbroeck for the calculations for Fig. 4.4, and to R. C. Prim for Fig. 4.5.

REFERENCES

1.1. W. Shockley and W. P. Mason, J. of Appl. Phys., **25,** No. 5, p. 677, 1954.
3.1. W. Shockley, *Electrons and Holes in Semiconductors,* D. van Nostrand, N. Y., 1950, p. 312.
3.2. See Reference 3.1.
3.3. W. Shockley, B. S. T. J., **28,** p. 435, 1949, Section 2.4.
3.4. See References 3.1 or 3.3.
4.1. W. Shockley and R. C. Prim, Phys. Rev., **90,** pp. 753–758, 1953. See also G. C. Dacey, Phys. Rev., **90,** pp. 759–763, 1953, and W. Shockley, Proc. I.R.E., **40,** pp. 1289–1314, 1952.
4.2. W. Shockley and R. C. Prim, Reference 4.1.
4.3. E. J. Ryder, Phys. Rev., **90,** pp. 766–769, 1953, references.
4.4. See Shockley and Prim, Reference 4.1.
5.1. This theory of mobility in higher fields has been published by H. Krömer, Zeits f. Physik **134,** pp. 435–450, 1953. Krömer considers the theory in connection with the values of α in point contact transistors but does not explore it as a power source. The present writer derived the same result in a more primitive form in 1948 as a potential means of obtaining high frequency power. However, experimental results by E. J. Ryder did not show evidence of this effect. The effect may possibly have occurred without being recognized because of the absence of an adequate appreciation of the importance of the boundary conditions discussed in this section.
5.2. See Reference 4.3.
5.3. G. C. Dacey, Phys. Rev., **90,** pp. 759–763, 1953.
5.4. W. Shockley, B. S. T. J., **30,** pp. 990–1043, 1951.
5.5. Personal communication.

Transistors and Junction Diodes in Telephone Power Plants

By F. H. CHASE, B. H. HAMILTON and D. H. SMITH

(Manuscript received November 30, 1953)

This paper describes the use of junction diodes, reference voltage diodes, and junction transistors in regulated rectifiers for telephone power plants. It discusses the pertinent characteristics of these semiconductor devices, together with illustrative circuits in which they are used to control the flow of direct current power.

1. INTRODUCTION

Recent articles in the literature have treated the theory and properties of semiconductor devices. In particular, papers by Messrs. Shockley, Ryder, Wallace and others have emphasized the theoretical aspects of the new devices; their reliability, reproducibility and performance at high frequencies to name only a few.[1, 4, 5, 6, 7] In addition many papers have been published concerning their applications in the transmission and computer fields. There is also a field of application for these devices in the conversion and control of power, and this paper discusses some of these power applications.

1.1. *Scope*

The first three groups of sections in this discussion review the pertinent characteristics and practical engineering aspects of junction rectifier diodes (Section 2.1), reference voltage diodes (Section 2.2) and junction transistors (Section 2.3). The second three groups of sections concern respectively shunt transistor regulators (Section 3.1), series transistor regulators, (Section 3.2) and power regulating circuits employing magnetic amplifiers in combination with transistors and junction diodes (Section 3.3). The last two groups of sections treat specific applications (Sections 4.1 through 4.3).

2. DEVICE CHARACTERISTICS

2.1. *Junction Rectifiers*

A junction rectifier is made from a wafer cut from a single crystal of semiconductor material. The materials now being used for this purpose are germanium and silicon, but to date the use of germanium is more common than silicon. Pure germanium in its undisturbed or intrinsic state is a poor conductor; but its conductivity can be increased by disturbances such as cosmic rays, photons of light, external potentials, or by the addition of very small amounts of selected impurities. We are concerned here only with the addition of impurities. There are two classes of these impurities, called "donors" and "acceptors." The physical mechanism by which pure germanium becomes conductive depends on which of these two classes of impurities are present. Donor impurities result in a surplus of free electrons which can conduct current by negative charges passing through the germanium crystal. Thus the addition of donor impurities to pure germanium creates "n" type material. Presence of acceptor impurities results in a shortage of electrons creating "holes," which have positive charges. These holes are mobile and they can conduct current through the crystal.[7] Thus the addition of acceptor impurities to pure germanium creates "p" material.

When an abrupt change is made from p to n type material inside the crystal a rectifying junction exists at the boundary between the two materials. This p-n junction exhibits rectifier action in that it will conduct current every easily from p toward n; but, in its rectifier operating range, only minute currents can be made to flow from n toward p. We say that this junction has a low forward resistance and a high reverse resistance. All rectifiers have these characteristics to a greater or lesser degree and the p-n junction rectifier characteristics have been compared elsewhere to other rectifier devices.[2]

There are two methods of producing the junction inside the crystal. It can be obtained by growing part of the crystal from p type material and part from n type. This is called a "grown" junction. It can also be obtained by diffusing impurities into the crystal after it has been grown. This has been called an "alloy" process, a "fused junction" process, or a "diffused junction" process.

2.11. *Junction Rectifier Terminology*

Before discussing the characteristics of junction diodes, it may be helpful for the reader to consider the terminology employed. As in other

rectifying cells, there are two directions of current flow, forward and reverse. Each diode has a positive and a negative terminal, and we define the positive terminal as that terminal towards which forward current flows *within the diode.* Likewise, the negative terminal is that terminal towards which reverse current flows *within the diode.* In Fig. 1(a), terminal 1 is the negative terminal and terminal 2 the positive. The circuit convention for the diode is a shorthand method of indicating the polarity of the diode to the engineer. If a battery is connected to a diode as shown in Fig. 1(b), forward current will flow, and if connected per Fig. 1(c), reverse current will flow. If the battery is replaced by a source of alternating current, forward current will flow through the diode during the half cycle that terminal 1 is positive, and reverse current will flow during the half cycle that terminal 2 is positive. The rectifier is said to "conduct" during the first half cycle and to "block" during the second half cycle, for the resistance in the conducting direction is very much less than the resistance in the blocking direction.

The figure of merit of a diode is a measure of this ease of conduction and the effectiveness of the blocking action. The ease of conduction can readily be determined on a static basis by applying a dc voltage to the diode as shown in Fig. 1(b) and plotting forward current through the diode as a function of applied voltage. Likewise, the blocking characteristic can be determined if a circuit per Fig. 1(c) is employed.

2.12. *Typical Junction Rectifiers*

Fig. 2 is a photograph of several sizes of typical junction diodes. The diodes shown have a range of forward current from several milliamperes (Diode I) to hundreds of amperes (Diode IV). Diode I is made from a crystal of silicon and the balance are made from germanium. Most rectifying diodes have a particular field of use dictated mainly by their power handling capacity in the forward direction of current flow, although Diode I is of interest because of its unusual reverse or blocking characteristic, as will be pointed out later in this paper.

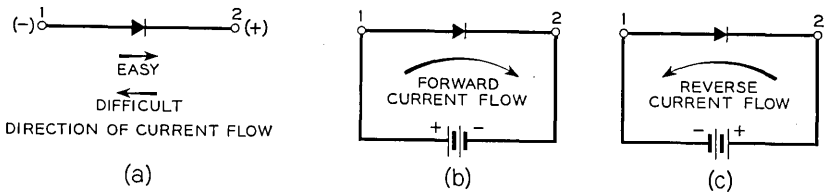


Fig. 1 — Rectifier terminology.

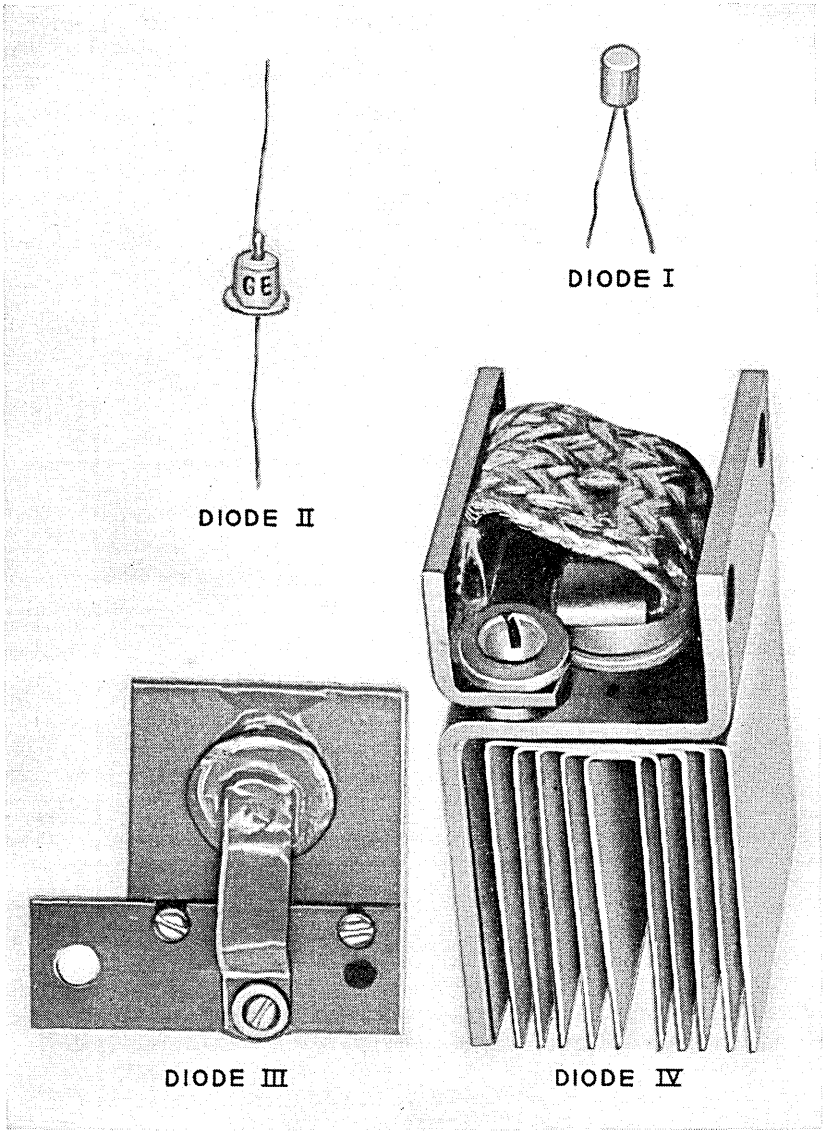


Fig. 2 — Typical junction rectifiers. Diodes II, III and IV, courtesy of the General Electric Company.

Fig. 3 is a plot of the static forward and reverse characteristics of the four diodes shown in Fig. 2. The characteristics were obtained using the circuits in Figs. 1(b) and 1(c), respectively, measurements being made in still air at room temperature. The curves in the first quadrant, $(+E + I)$ are the forward characteristics and the curves in the third quadrant $(-E\ -I)$ are the reverse characteristics. Notice that the scales are different in these quadrants. In general, at any other temperature the curves would shift their positions with respect to the reference axes. This must be taken into account by the circuit designer.

2.13. *Junction Temperature*

We will limit further discussion of general characteristics to those of Diode IV, for in many respects this is the most interesting rectifier for

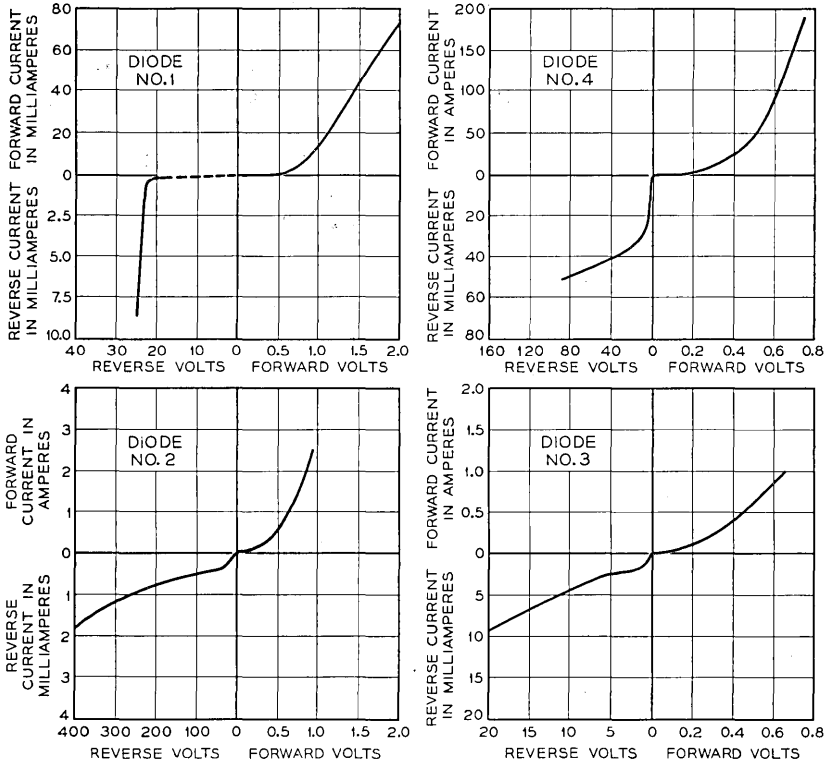


Fig. 3 — Junction rectifier static characteristics.

power applications. Laboratory experience indicates that it is not desirable to operate the junction of this diode above 65° to 70° centigrade. The value of this critical temperature is not accurately known on account of the difficulty in measuring the junction temperature inside of the crystal. However, below the critical temperature, those changes in characteristics which are associated with changes in junction temperature are reversible, that is, if the temperature is raised and then reduced, the characteristics will shift back to values previously experienced at the reduced temperature. Beyond the critical junction temperature any change in the reverse characteristics is permanent and has the effect of reducing the reverse resistance. In an operating circuit, this effect leads to progressively greater permanent damage to the diode. Lowered reverse resistance allows more reverse current to flow, increases the reverse power dissipation and elevates the temperature of the junction causing further reduction of the reverse resistance, and so on until the diode no longer blocks.

Thermal damage to the junction can be prevented by removing heat. This method is employed with the diode under discussion by forcing air through the cooling fins at a high velocity. The quantity of air needed depends on the amount of heat generated in the junction, the efficiency of the cooling fins and the temperature of the air employed for cooling. In most Bell System applications, the maximum temperature of the

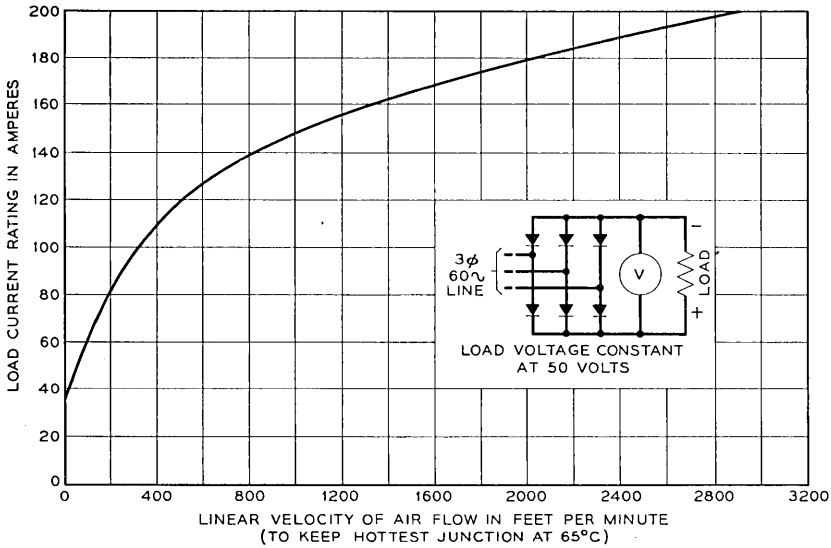


Fig. 4 — Junction rectifier forced cooling characteristics.

ambient air is 40°C, which permits the junction temperature to be 25 to 30°C above the air temperature before the critical value is reached.

A typical load-current versus air velocity curve is shown in Fig. 4. The curve is based on a 65°C junction temperature measured by thermocouples attached to the radiating structure near the junction and 40°C ambient air. Notice that the curve is taken with a working circuit composed of six diodes in a three-phase full wave bridge arrangement. In general, engineers developing rectifier circuits find that curves showing the properties of combinations of rectifying diodes are more useful than single diode characteristics, except where the properties of the diode are such as to make it useful as a valve, or as a reference standard, as is the case of Diode I in Fig. 2. This leads directly to a more detailed consideration of the blocking or reverse characteristics of junction rectifiers.

2.2. *Reference Voltage Diodes*

2.21. *General*

In the case of silicon junction diodes it has been possible to reduce the reverse current to a very low value for reverse voltages up to a value called the "saturation voltage." When the saturation voltage is reached the electrons and/or holes which comprise the leakage current are given sufficient energy to create other electron-hole pairs which add to the original reverse current. This process is cumulative and leads to large increases in current for small further increases in voltage. The effect is illustrated by the reverse voltage-current characteristic for Diode I in Fig. 3. This curve shows the reverse current to be quite low for voltages less than 22 volts. This portion of the characteristic is called the "high resistance region." As voltage is further increased the curve goes through a "transition region" to the "saturation voltage region" at 23 volts where voltage is nearly constant over a wide range of current. The voltage saturation characteristic makes the diode suitable for use as a source of reference potential in the control of power. Those readers who wish to study the basis of these properties will find the theory covered elsewhere in the literature.[3,8]

The rectifier selected for study in this Section is Diode I. This is a p-n junction rectifier made from silicon. It has been constructed to obtain a reasonably constant saturation voltage as shown in Fig. 5. In order to show the wide range of current values where this voltage is substantially constant, Fig. 5 is plotted to a logarithmic scale. In this connection it is interesting to note that the saturation voltage can be controlled in manufacture from a few volts to several hundred volts. This

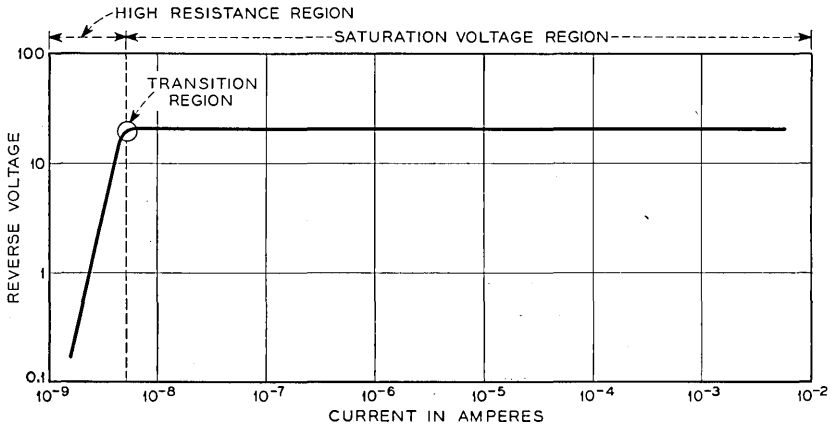


Fig. 5 — Reverse characteristics of a reference voltage diode.

range can be compared to the 60 to 150 volts range of cold cathode voltage regulator tubes which are also used as sources of reference potential.

2.22. *Saturation Voltage Utilized in Regulating Circuits*

In all check back (feedback) regulating circuits the potential to be regulated is compared to a reference potential. This comparison is a form of subtracting the two values so that the changes in the potential to be regulated produce a large percentage change in the difference or error voltage. The methods by which this is accomplished in direct current circuits are illustrated in Section 3. A stable source of reference potential is required for this type of regulation. When the saturation voltage of a silicon junction diode is used for this purpose, we have called the device a "reference voltage diode."

2.23. *Effect of Temperature on Saturation Voltage*

In order to evaluate the stability of Diode I in its saturation voltage region a small section of Fig. 5 has been redrawn in Fig. 6 using a linear scale. Additional curves are included in Fig. 6 to show the change of voltage with ambient temperature variations. The slope of the 30 degree curve in Fig. 6 is equivalent to a resistance of 200 ohms in series with a 23-volt battery with current flowing through this combination from an external source. The change of potential with ambient temperature is equivalent to a 0.07 per cent change per degree C. It should not be inferred that these are limiting values, for diodes have been tested which exhibit slopes of less than 10 ohms and temperature coefficients of less

than 0.01 per cent per degree C. The specific applications covered later in this discussion show methods to compensate for slope and temperature variation when necessary.

2.3. *Junction Transistor Action*

2.31. *Two-Rectifier Analysis*

In junction transistors there are two p-n junction rectifiers contained in the semiconductor material. Of the materials now in use germanium is the more prevalent. Remembering the results of adding donor and acceptor impurities to obtain n and p type materials covered in section 2.1 these two rectifiers are obtained by interposing a layer of p type material between two layers of n type making an *n-p-n transistor* or interposing a layer of n type material between two layers of p type making a *p-n-p transistor*. The electrical connections are designated as the collector terminal, the emitter terminal and the base terminal. Both types of transistors (n-p-n and p-n-p) have a rectifying junction between the collector and base terminals and another rectifying junction between the emitter and base terminals. The polarity of the collector and emitter rectifying junctions determines whether the transistor is n-p-n or p-n-p.

Figs. 7(a) and 7(b) are simplified diagrams illustrating respectively the internal circuits of n-p-n and p-n-p transistors. The figures show the characterizations of transistors by means of a two-rectifier analogy. Although a transistor may be somewhat over-simplified by this method of characterization, the analogy permits the power engineer to approximate the operation of transistors in familiar terms. Experience in the development of the circuits described later in this article has proven that the analogy is valid under circumstances where the operation of the transistor as a dc amplifier is of interest.

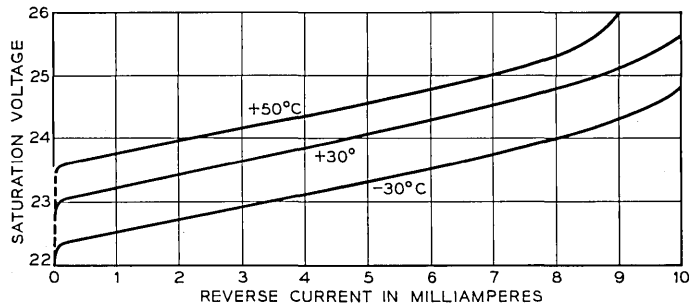
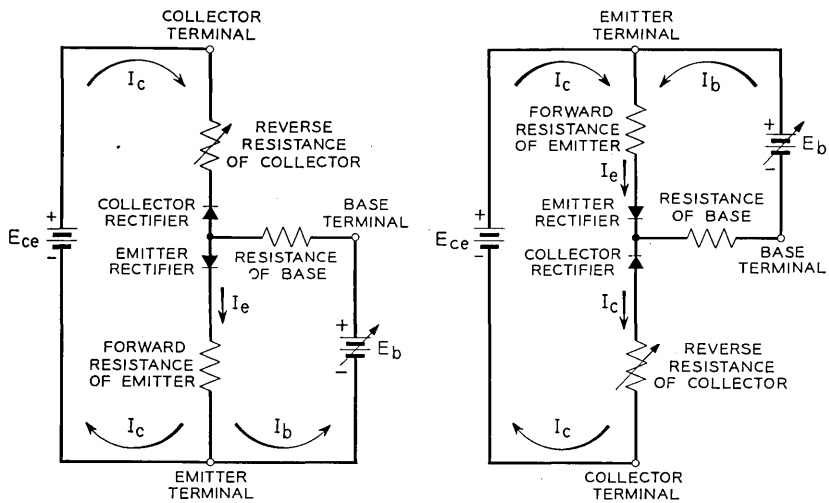


Fig. 6 — Saturation voltage characteristics of a reference voltage diode.

In an n-p-n transistor the collector and emitter terminals are the positive electrodes of the rectifiers, see Fig. 7(a), and in a p-n-p transistor the collector and emitter terminals are the negative electrodes of the rectifiers, see Fig. 7(b). The base terminal is the common point of the two rectifiers. In a given transistor each rectifier has a saturation voltage, usually stated in the characteristics, which must not be exceeded in normal operation. Thus, the saturation voltage of the collector rectifier determines the maximum instantaneous collector potential. The emitter rectifier also has a saturation voltage which determines the maximum potential which can be applied between the base and the emitter. The saturation voltage of the collector rectifier usually differs from the saturation voltage of the emitter rectifier.

2.32. *Transistor Action*

If a source of potential, E_{ce} in Figs. 7(a) and 7(b) is connected between the collector and emitter terminals, the resulting current will flow in series through the collector rectifier in its reverse direction and through the emitter rectifier in its forward direction. This is the direction of current flow for transistor action to take place. In Fig. 7 the reverse resistances of the collector rectifiers are shown and the forward resistances of the emitter rectifiers are also shown.



(a) n-p-n TRANSISTOR (b) p-n-p TRANSISTOR

Fig. 7 — Junction transistor analogy.

2.33. *Current Gain*

Now when a second relatively small potential is connected between the base and emitter rectifier (E_b in the sketches) additional current, I_e will flow through the emitter rectifier in the forward direction and I_c will also increase. This increase in I_c caused by the increase in I_e is transistor action. The increase in I_c is related to the increase in I_e by the factor alpha (α) as written below:

$$\Delta I_c = \alpha \Delta I_e . \tag{1}$$

The application of Kirchoff's current law to the sketches in Fig. 7 gives the change in I_b as follows

$$\Delta I_b = \Delta I_e - \Delta I_c . \tag{2}$$

By combining equations (1) and (2), ΔI_c can be written as a function of ΔI_b only

$$\Delta I_c = \frac{(\alpha)}{(1 - \alpha)} \Delta I_b . \tag{3}$$

The usual value of α for junction transistors is near but slightly less than unity. In a typical case α might be 0.98. This value when substituted in equation (3) shows the current gain of the transistor, $\Delta I_c/\Delta I_b$ to be 49. Most of the circuits discussed in this paper are based on equation (3).

It has been shown how a small change in base to emitter potential with a small change in base current effects a large increase in collector current at a higher voltage. This explains how large power gains, of the order of 60 db, can be obtained from the junction transistor.

The sketches in Fig. 7 do not show why this transistor action takes place. The reasons for it involve the use of such solid state physics terms as the migration of electrons and holes through a crystal lattice, and the interposition of junction barriers. The "why" for transistor action is very important in the manufacture of transistors, and it has been thoroughly covered in the literature.[1, 7] For present purposes it is only necessary to examine the static characteristics of an n-p-n transistor as shown in Fig. 8. This figure presents transistor characteristics in a manner which simplifies the explanation of the operation of the transistor control circuits covered later in this paper.

Referring to the curves in Fig. 8, it will be seen that in the straight portion of the 1.5-volt curve, a change of 50 microamperes in the base current will result in a change of about 2 milliamperes in the collector current. This illustrates the current amplification of transistors and the

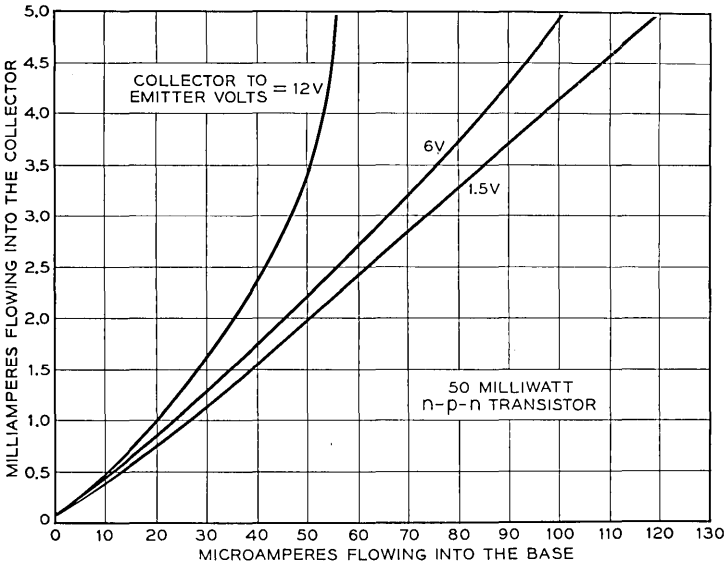


Fig. 8 — Junction transistor static characteristics.

current gain of this transistor is equal to 40. The measured α for this transistor was 0.976. Substituting in the current gain formula, equation (3) above, the calculated current gain is 40.6 which agrees with Fig. 8 within the accuracy of the measurements.

2.34. *Low Voltage Characteristics*

Again referring to the curves in Fig. 8, it will be seen that transistors operate at low collector to emitter potentials. The 1.5-volt curve is not the minimum potential at which this transistor will operate. Some transistors have good current amplification at potentials as low as two-tenths of a volt. When the base current is reversed, the characteristics in Fig. 8 can be extended to smaller collector current values. One might assume that the collector current can be reduced to zero by causing enough current to flow out of the base. This is not true. There is a minimum collector current, called the saturation current, and increasing current flow out of the base will not decrease the collector current below this value. This saturation current is assigned the symbol I_{c0}. This I_{c0} current is usually a few microamperes but it increases at the rate of 7 or 8 per cent per degree Centigrade increase in temperature of the collector junction. Transistors also have a critical junction temperature

which should not be exceeded under any operating conditions, and this must be kept in mind during the design of the regulating circuits.

2.35. *Equivalent Circuit of a Transistor*

Ryder and Kircher[4] have shown that it is possible to convert the sketches shown in Fig. 7 into a small signal equivalent circuit using alpha and the three characteristic resistances of the transistor. These resistances are the emitter resistance r_e, the base resistance r_b and the collector resistance r_c. Two forms of equivalent circuit are shown in Figs. 9(a) and 9(b). In the equivalent circuit in Fig. 9(a) the active portion of the transistor is characterized as a current generator. This equivalent circuit is more directly related to the physical processes occurring inside the transistor than the equivalent circuit in Fig. 9(b) which characterizes the active portion of the transistor as a voltage generator. Although both equivalent circuits are useful the one in Fig. 9(a) is preferred in power work because r_c is much larger than the load resistance in many cases and can be neglected. Typical values for the equivalent circuit parameters are given in the caption of Fig. 9. The use of the equivalent circuits are further discussed in some of the articles listed at the end of this paper. The article[5] by R. L. Wallace Jr. and W. J. Pietenpol is of particular interest in this connection.

2.36. *Typical Junction Transistors*

Fig. 10 is a photograph of two Bell System junction transistors made from germanium. The smaller one will dissipate 50 milliwatts, and the larger one is an exploratory model that will dissipate 2 watts when it is attached to a suitable heat sink. These transistors are hermetically sealed to protect them from the infiltration of moisture. The characteristics shown in Fig. 8 were measured using the smaller unit.

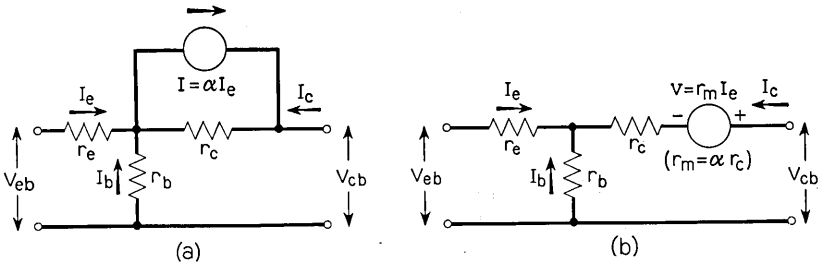


Fig. 9 — Junction transistor equivalent circuits. Typical values for a 50-milliwatt transistor: r_e, 25 ohms; r_b, 500 ohms; r_c, 5 megohms; α, 0.98; and r_m, 4.9 megohms.

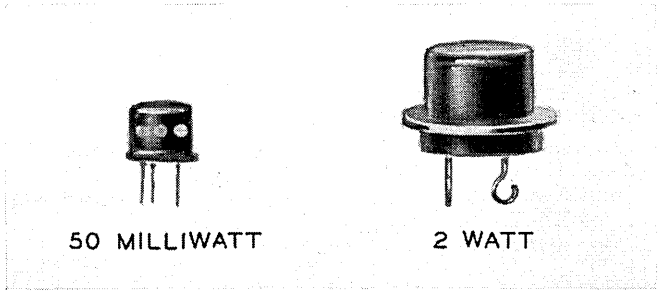


Fig. 10 — Typical junction transistors.

Thus, note that there are two kinds of transistors with respect to the polarity of the electrodes. The n-p-n transistor operates with positive collector potential and the p-n-p requires negative potential on the collector. Both will amplify current changes in the base circuit into much larger current changes in the collector circuit. The transistors have similar equivalent circuits and parameters but all of their operating potentials and currents are reversed. It is also significant that the normal direction of current flow is out of the base terminal of the p-n-p transistor and that the normal direction of current flow is into the collector terminal of the n-p-n transistor. Likewise, the normal direction of current flow is out of the collector terminal of the p-n-p transistor and into the base terminal of the n-p-n transistor. This relationship between direction of current flow in n-p-n and p-n-p transistors is called reversed or complementary symmetry, and enables the circuit designer to cascade direct coupled transistors, alternating n-p-n and p-n-p. This is not possible with vacuum tubes because there is no tube that will operate with negative plate potential. It will be shown how this complementary symmetry can be used to advantage in multistage direct current amplifier circuits.

3. TYPICAL REGULATING CIRCUITS

3.1. *Shunt Regulators*

3.11. *Simple Diode Regulator*

If a load is connected to a source of power, the current through the load and thus the voltage drop across the load will depend on the potential of the source of power, the internal impedance of the power supply and the load impedance. The voltage drop across the load can be made very nearly independent of these three parameters by employing a circuit known as a shunt regulator.

A shunt regulator is a variable current device, connected in parallel with the load. Both the load and the shunt regulator draw current from the source of power through a common impedance. The operating requirement for a good shunt regulator is that the voltage drop across it must remain constant over a certain range of current. Certain types of cold cathode tubes such as the VR-150-30(0D3) exhibit this effect. It has been determined that certain semiconductor junction diodes exhibit the same effect. Note that diode No. 1 that is shown in Fig. 3 has a reverse voltage drop of about 24 volts over a range of reverse current from less than 1 milliampere to almost 10 milliamperes. If such a device is connected in parallel with a load as in Fig. 11, shunt regulating action will take place. Consider the operation of the circuit in Fig. 11, first assuming that the load impedance is constant and that the potential of the power source increases. Additional current tends to flow from the source, but since the potential of the reference voltage diode designated "*s*" in Fig. 11 is fixed, this additional current develops an increased voltage drop across the regulating resistor, and the load voltage does not change. Similar reasoning can be applied to the case of a decrease in source voltage. Next assume constant source voltage and an increase or decrease in load resistance. This would normally tend to cause a change in the voltage drop across the load, but the shunt element draws respectively more or less current than normal and the load voltage again does not change.

The value of the regulated load potential in Fig. 11 is controlled by the saturation voltage of the diode and it cannot be adjusted to any other value. The accuracy of regulation is controlled by the slope of the reverse characteristics shown in Fig. 6. An additional limitation is that the usefulness of this type of regulator is controlled by the power handling capacity of the diode. The next section shows how these limitations can be circumvented by the addition of transistors.

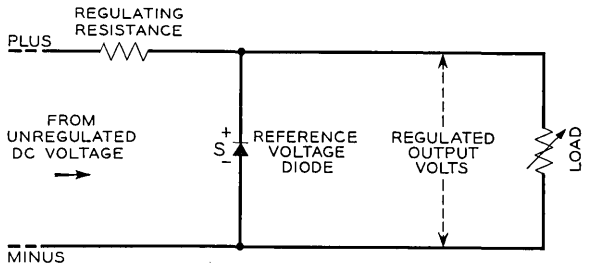


Fig. 11 — Simple shunt regulator.

3.12. *Transistor-Diode Regular*

To examine the operation of shunt transistor regulators consider the circuit shown in Fig. 12. In this, the transistor is shown by its standard convention where the upper slanting line represents the collector rectifier shown in Fig. 7 and the lower slanting line with the arrow on it represents the emitter rectifier. The direction of the arrow shows that, in this transistor, current flows out of the emitter, so it is an n-p-n transistor. The (*c*), (*b*) and (*e*) designations also help to locate the collector, base and emitter. Notice that the emitter current, shown by the arrow I_e, flows through the reference voltage diode in its reverse direction. The rectifier symbol with an adjacent "*s*" is a convention for this diode.

In Fig. 12 a portion of the load voltage is applied to the base of the transistor by means of the adjustable potentiometer. The potential of the emitter is held constant with respect to the negative output potential by the saturation voltage of the diode. The base-to-emitter voltage is thus equal to a proportion of the load voltage minus the saturation voltage. The potentiometer is adjusted so that the base potential is slightly positive with respect to the emitter when the desired value of voltage appears across the load. This value of load voltage is called the regulated voltage. Current I_b then flows into the base, current I_c flows through the regulating resistance, and $I_b + I_c$ combine to form I_e. Now assume that the regulated voltage (E) increases by an amount ΔE. The base voltage becomes more positive with respect to the negative terminal by the proportion of ΔE developed across points 1 and 2 of the potentiometer. Since the emitter potential is held constant by diode "*s*" and cannot change, the net effect is to increase the base to emitter potential. This change in base to emitter potential causes an increase in collector current,

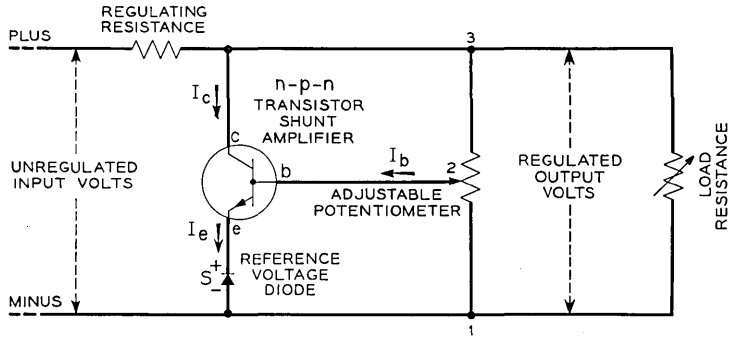


Fig. 12 — Transistor shunt regulator using one transistor.

a consequent increase in voltage drop in the regulating resistor, and a decrease in the load voltage. The correcting process continues until the load voltage returns to the regulated value, and takes only a small fraction of a second. The process is essentially the same for a decrease in load voltage, except that the base to emitter potential decreases, the collector current decreases, the voltage drop across the regulating resistor decreases and the load voltage rises to the regulated value.

The value of the regulated output voltage is determined by the adjustment of the potentiometer. Of course in a practical shunt regulator circuit, the adjustable range of the potentiometer would have to be limited to correspond with the operating range of the transistor. The maximum allowable positive potential between the base and the emitter is limited by the safe value of the maximum collector current. The maximum allowable negative potential between the base and the emitter is limited by the saturation voltage of the emitter rectifier.

The accuracy of this shunt regulator circuit is restricted by the slope of the characteristic curves for the reference voltage diode. All of the changes in base and collector currents required for regulation flow through this diode and cause changes in the saturation voltage. The addition of the transistor does not increase the accuracy of regulation but only allows adjustment of the regulated output potential to a value which is greater than the standard potential. However additional stages of transistor current amplification minimize the reference potential changes by restricting the range of current excursions through the diode. An example of a multistage shunt regulating circuit is given in Fig. 13.

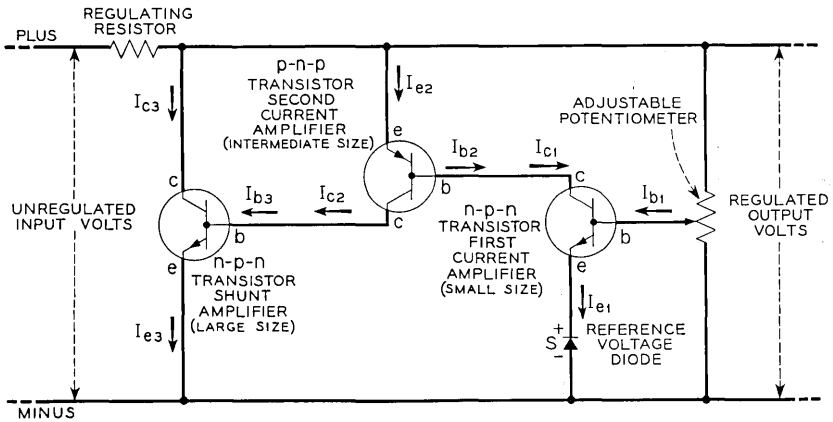


Fig. 13 — Transistor shunt regulator using three transistors.

3.13. *Multistage Transistor Shunt Regulator*

In Fig. 13 two additional transistors have been added to the simple shunt regulator of Fig. 12 in order to increase the accuracy of regulation. The first stage (subscript 1 is used for the transistor currents in this stage) compares the output potential to the reference voltage, drives the second stage (subscript 2) which in turn drives the third stage (subscript 3). The first stage transistor operates in a similar fashion to the transistor in Fig. 12, except that its collector current now is the base current of the second transistor. The collector current of the second transistor is the base current of the third transistor. The second and third stage transistors amplify the collector current of the first transistor. The shunt regulating current is the sum of the currents in all three transistors. An examination of Fig. 13 will reveal that the first transistor is an n-p-n, the second transistor is a p-n-p, and the third transistor is an n-p-n and that no coupling networks are used. This illustrates the advantages of complementary symmetry.

In Fig. 13 different sizes are specified for the three transistors. The transistor shown for the first current amplifier might be a 50-milliwatt transistor operating at a collector potential of about 10 volts. Then the maximum base current of the second stage p-n-p transistor should not exceed 5 milliamperes and, with an assumed current amplification of 20 times, the maximum collector current of the second stage could be 100 milliamperes Such a transistor has been developed. With 100 milliamperes flowing into the base of the large n-p-n transistor and an assumed current amplification of 20 times the maximum shunt regulator current would be about 2 amperes which would compensate for considerable load current variations. Large size transistors such as would be necessary in the third stage are now under exploratory development within the industry.[6]

The circuit in Fig. 12 can be modified to use a p-n-p transistor and several other modifications can be made. Similar modifications can be made in the circuit shown in Fig. 13. It is not within the scope of this article, however, to show all the permutations and combinations of transistor regulator circuits that are usable. Section 3.2 below covers some typical transistor series regulator circuits.

3.2. *Series Regulators*

Precise voltage control can be obtained with shunt regulators but series regulator circuits are usually more efficient. This comes about because the shunt regulator wastes the shunt current plus the voltage

drop across the regulating resistance whereas the series regulator wastes only the voltage drop across the series device. At light load the power dissipated in the shunt current is usually greater than the power dissipated in the series circuit. With a transistor used as the series regulator device this difference in efficiency is more pronounced because of the small collector voltage that can be used for the full load current. This collector voltage is the voltage drop across the series transistor as shown in Fig. 14.

3.21. *Simple Series Regulator*

Fig. 14 shows a simple transistor series regulator circuit. A p-n-p transistor is shown connected so that all the load current must pass through it. The comparison of the output voltage to the reference potential in the current amplifier of Fig. 14 is accomplished by holding the emitter at a constant potential with respect to the *positive* output terminal. Note the difference between this method and that covered in the previous section on the shunt regulator 3.12, where the emitter was held at a constant potential with respect to the *negative* output terminal. Now, when the output potential increases by an amount ΔE, the base voltage becomes more *negative* with respect to the positive terminal by the proportion of ΔE developed across points 2 and 3 of the potentiometer. Since the emitter cannot change with respect to the point of reference (the positive terminal), the net effect is to *decrease* the base to emitter potential and the collector current for an increase in output voltage. The collector current decrease is amplified by the current gain of the p-n-p series transistor to decrease the load current, reducing the output

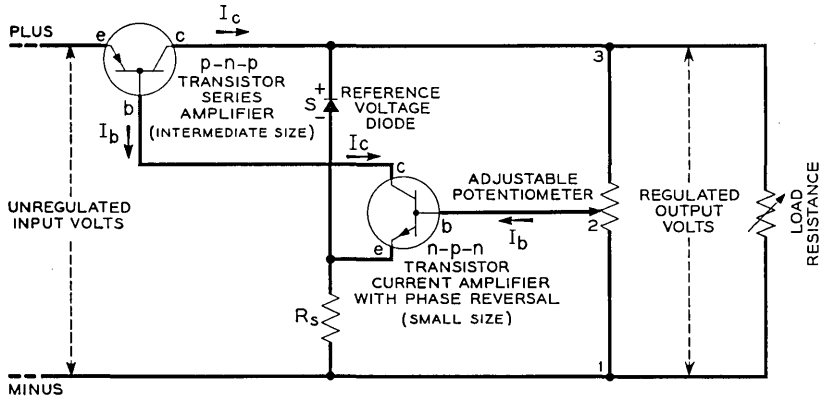


Fig. 14 — Transistor series regulator constant voltage regulation.

voltage, and thus regulating it. The value of the regulated output voltage is again determined by the adjustment of the potentiometer. The ohmic value of the R_s resistor in Fig. 14 is selected to keep the current flowing through the reference voltage diode in its saturation voltage region.

Fig. 14 is the simplest form of a transistor series regulator circuit. It requires two transistors whereas the most simple form of a transistor shunt regulator shown (Fig. 12) requires only one transistor. But the added current gain of the second transistor in Fig. 14 results in better regulation than can be obtained with Fig. 12. If desired the circuit in Fig. 14 can be modified to change the series transistor to the negative output lead by using the complementary p-n-p first current amplifier and an n-p-n series transistor. This illustrates another advantage of the complementary symmetry of the two types of transistors. Also, if more gain is required, additional transistor stages can be used employing the principles outlined above.

3.22. *Series Current Regulator*

The circuits covered so far regulate for constant output voltage. Similar transistor regulator circuits can be developed which will regulate for constant output current. One of these is shown in Fig. 15. In this circuit the load current produces a voltage drop across the regulating resistance and, in the n-p-n transistor, this voltage drop is compared to the reference voltage. The difference between these two potentials controls the n-p-n transistor base current and this base current is amplified by the current gain of both transistors to control the load current.

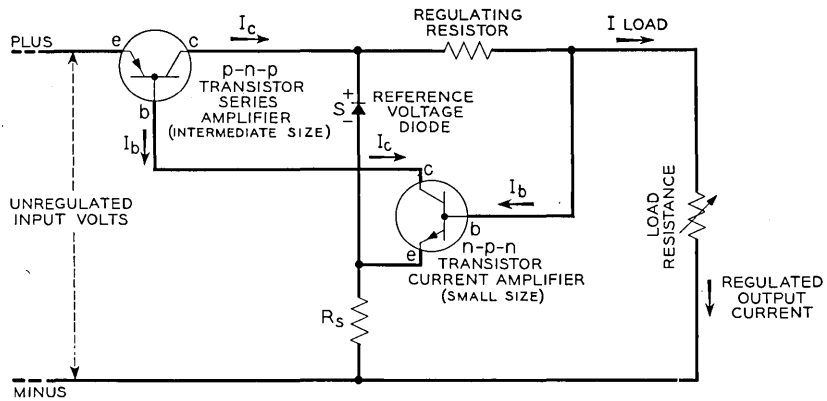


Fig. 15 — Transistor series regulator constant current regulation.

This circuit is phased so that the load current will be increased when it is too small and decreased when it is too large. The values of the regulating resistor and the reference voltage determine the value of the regulated load current. Additional current amplifier stages can be included or the circuit can be modified to change the series transistor to the minus lead as covered above.

3.3. *Transistors Combined With Magnetic Amplifiers*

3.31. *General*

Transistors can be used to control directly the flow of power to a load as pointed out in the sections on series and shunt regulators. However, their direct use is limited to moderate voltages (below 100 volts) or moderate currents (up to 1 ampere) with transistors now contemplated.

3.32. *Transistors as DC Preamplifiers*

In cases where regulation of higher power is required, it is expedient to combine transistor circuits with other devices having higher power-handling capacity. One type of combination is shown in Fig. 16, where a transistor is used to amplify weak dc error signals to a magnitude sufficient for driving a magnetic power amplifier.

In Fig. 16, emitter (*e*) of the n-p-n transistor is held at a fixed negative voltage with respect to the positive output of the power supply by the reference voltage diode ("*S*"). Another negative voltage derived from the output voltage of the power supply through potentiometer (*P*) is applied

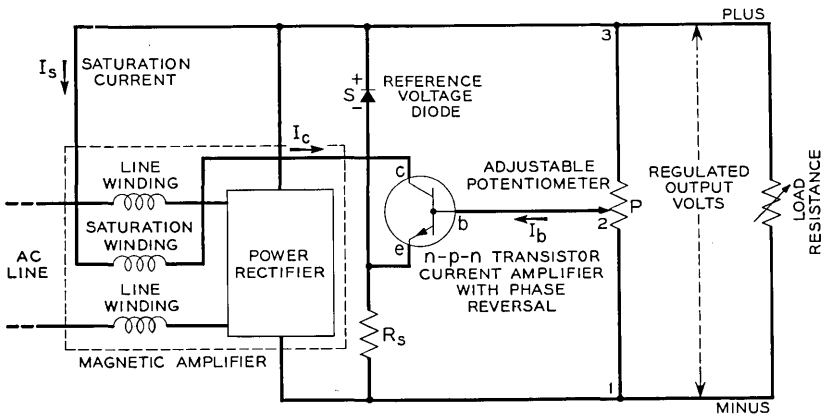


Fig. 16 — Transistor control circuit for a magnetic amplifier regulated rectifier with constant voltage regulation.

to base (b) of the transistor. This latter voltage is made a little smaller than the emitter voltage so that the base (b) is positive with respect to the emitter. Now assume that the load voltage increases for some reason such as an increase in the line voltage or a decrease in the load current. A portion of the increased load voltage appears across points 2 and 3 of potentiometer (P), and tends to make the base voltage more negative. Since the base is slightly positive with respect to the emitter, the net effect of making the base more negative is to decrease the base-to-emitter voltage. Through transistor action, the collector current, which is also the saturation current of magnetic amplifier, decreases and the ac impedance of the line windings rises. The line windings absorb more input voltage and the output voltage is brought back very nearly to the original value before the change.

The circuit of Fig. 16 is of interest because it can control larger amounts of power than can be handled by transistors alone and, in addition, it is capable of faster regulating action than an all-magnetic regulating circuit with the same loop gain. The use of the transistor in this circuit eliminates the need for one or more stages of milliwatt-size magnetic preamplifiers.

3.33. *Increased Gain in Voltage Regulators*

Additional amplification to improve the regulation can be added to Fig. 16 in two ways. Several stages of transistor current amplification can be added or more magnetic amplifier stages can be used. Of course

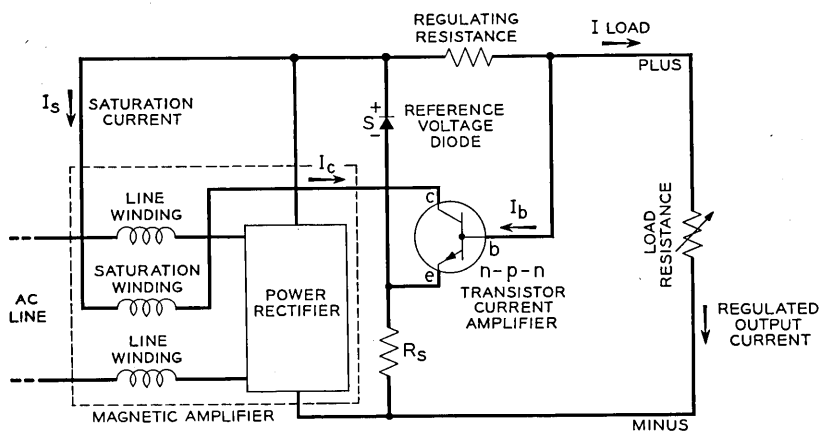


Fig. 17 — Transistor control circuit for a magnetic amplifier regulated rectifier with constant current regulation.

a combination of the two methods is also feasible. Additional magnetic amplifiers have the disadvantage of adding time delay. Transistor action likewise is not instantaneous because it takes a finite amount of time to move the charge over a finite distance in the crystal lattice. However transistor action is much faster than the time required to change the current in practical magnetic amplifiers.

3.34. *Current Regulators*

Fig. 17 shows a simple transistor control circuit to obtain constant current regulation with a magnetic amplifier regulated rectifier. The operation of this circuit is similar to Fig. 15 and its description will not be repeated.

3.35. *Temperature Effects*

One limitation of the foregoing transistor regulating circuits is the sensitivity of collector current to ambient temperature variations. The collector current increases with increasing temperature even if the base-to-emitter bias is held constant. This is the result of three factors. (1) I_{c0}, the uncontrolled portion of I_c increases greatly as covered in Section 2.34; (2) the emitter resistance (r_e) decreases causing I_b to increase, and (3) alpha changes. The effect of the temperature sensitivity of the collector can be greatly reduced by using a differential or push-pull circuit of the type illustrated in Fig. 18.

3.36. *"Push-Pull" DC Amplifier*

The push-pull circuit uses two emitter-coupled n-p-n transistors and is in many respects similar to a cathode-coupled vacuum tube amplifier.

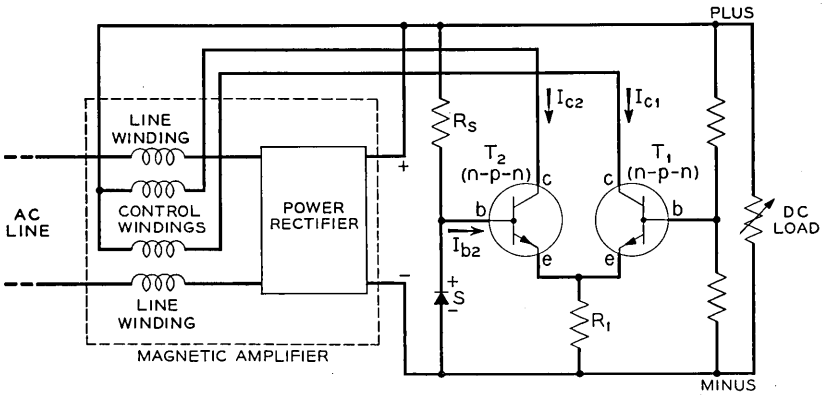


Fig. 18 — Push-pull transistor control circuit for a magnetic amplifier regulated rectifier with constant voltage regulation.

A voltage proportional to the regulated output is connected to the base of transistor (*T1*). Fixed resistors are shown in the base circuit of (*T1*) in Fig. 18, but a variable potentiometer could be used. The reference voltage diode "*s*" applies a constant reference voltage to the base of transistor (*T2*). If the output voltage tends to increase, more collector and emitter current flows in transistor (*T1*) due to the increase in its base-to-emitter voltage. This increase in emitter current of transistor (*T1*) flows through resistor (*R1*) and tends to raise the emitter voltage of transistor (*T2*). Since the base potential of transistor (*T2*) is fixed, the effect is to decrease the base-to-emitter voltage of (*T2*) and its collector and emitter currents decrease. The result is an increase in I_{c1} and an almost equal decrease in I_{c2}. If the two saturation windings on the magnetic amplifier are oppositely poled, the changes in I_{c1} and I_{c2} represent a net decrease in the control ampere turn input to the magnetic amplifier. As before, the magnetic amplifier responds by absorbing more voltage. If, however, I_{c1} and I_{c2} both increase equally due to an increase in ambient temperature, no net change is made in the control ampere turn input to the magnetic amplifier. Thus if the two transistors are perfectly matched, and the reference voltage diode has a low temperature coefficient, temperature changes will have little effect on the output regulated voltages.

A further advantage is the reduced variations in the current through the reference voltage diode. As in the case of the other circuits additional stages of transistor or magnetic amplification can be added to increase the loop gain and the precision of regulation.

4. APPLICATIONS

4.1. *General*

The last sections of this discussion cover some specific applications of the principles discussed above. Section 4.21 covers a one-stage transistor shunt regulated rectifier as a grid battery eliminator for phase controlled thyratron tube rectifiers. Section 4.22 covers a transistor voltage amplifier circuit as a grid battery eliminator for magnitude controlled thyratron tube rectifiers. A two-volt, three-ampere regulated rectifier covered in Section 4.31 illustrates how a low voltage, high current, regulated rectifier with a transistor and magnetic amplifier control circuit can be obtained. Section 4.32 covers a 65-volt, 200-ampere regulated rectifier for telephone central office battery charging. It uses the p-n junction rectifier devices covered in Section 2.1 and a modification

of the transistor control circuits for magnetic amplifier regulation covered in Section 3.36.

4.2. *Grid Battery Eliminators*

4.21. *Phase Controlled Thyratron Tube Rectifiers*

In thyratron tube regulated rectifiers the standard potential for the checkback regulator is often obtained from dry cells. The annual replacement of dry cell batteries is an appreciable maintenance expense, particularly in those cases where the rectifiers are installed in isolated or unattended locations. This section covers a transistor shunt regulated rectifier as a substitute for the dry batteries. Its circuit is illustrated in Fig. 19.

The circuit in Fig. 19 is the same as Fig. 12 with the addition of the compounding resistor and the thermistor. The compounding resistor is added to compensate for the slope of the reference voltage diode in its saturation voltage region (see Fig. 6). The thermistor is added to compensate for ambient temperature varitions of this diode and the transistor.

The compounding resistor adds ac line voltage compounding. The transistor base current regulating signal in Fig. 19 is increased by the compounding resistance whenever the ac voltage is increased. By selecting the proper ohmic value of the compounding resistor, the circuit in Fig. 19 can be arranged so it will deliver constant output voltage into a constant resistance load when the ac voltage is varied from 85 per cent

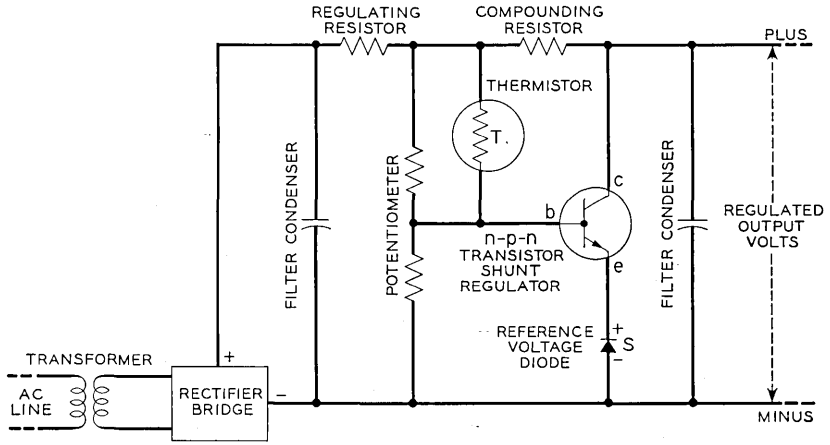


Fig. 19 — Grid-battery eliminator for phase-controlled thyratron rectifier.

to 115 per cent of its normal value. In the thyratron tube rectifiers, the circuit of Fig. 19 operates into a constant resistance load of several megohms. With such a high value of load resistance, the addition of the compounding resistor on the load side of the regulating resistor does not cause appreciable error. In fact, laboratory measurements on an experimental unit show that the compounding can be adjusted to obtain *improved* regulation of the thyratron tube rectifier when the grid battery eliminator is used in place of the normal grid battery. This is because the grid battery eliminator can be adjusted to over-correct for line voltage variations and thus compensate for the slight amount of residual line regulation error in the thyratron circuit.

The thermistor in Fig. 19 is a shunt element across one of the resitors in the potentiometer and a change of its resistance is equivalent to changing the potentiometer adjustment. The thermistor decreases its resistance with an increase of ambient temperature so it will change the output voltage when the temperature is changed. This output voltage change is opposed to the voltage changes resulting from the temperature effects in the reference voltage diode and the transistor. By selecting the proper thermistor and the proper ohmic values for the potentiometer resistors, these temperature variations will nearly cancel and the regulated output voltage will be temperature compensated.

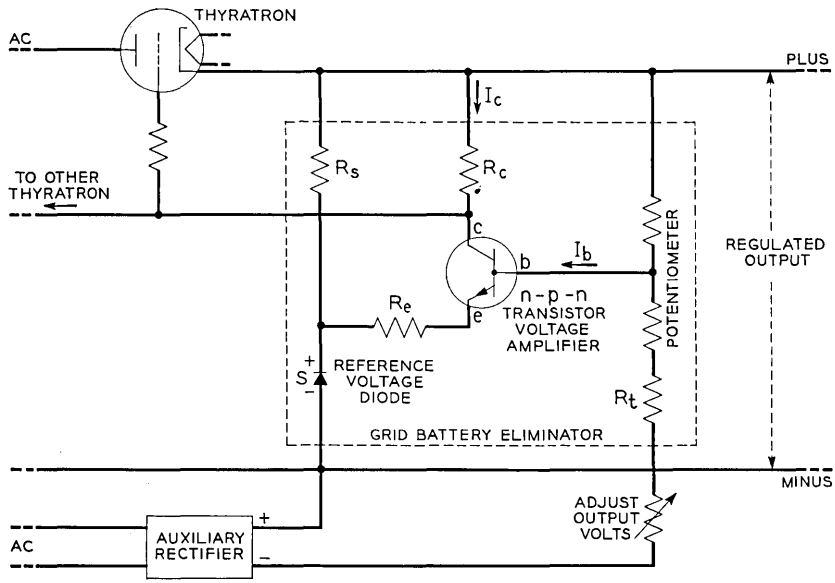


Fig. 20 — Grid-battery eliminator for magnitude controlled thyratron rectifier.

4.22. *Magnitude Controlled Thyratron Tube Rectifiers*

The grid battery eliminator covered in Section 4.21 is also usable in magnitude controlled thyratron tube regulated rectifiers but a simple, less expensive circuit can be used for this application. It is illustrated in Fig. 20. A simplified schematic of the thyratron rectifier is also shown in Fig. 20 and the grid battery eliminator is the portion of the circuit enclosed by the dotted line. It is actually a transistor voltage amplifier circuit. This type of circuit has not been covered previously in this discussion so its operation is described in some detail below.

Referring to Fig. 20 a portion of the output potential is compared to the reference potential by the base and emitter connections to the transistor. The difference between these two potentials causes the base current I_b to flow. This base current is amplified by the current gain of the transistor and it results in flow of collector current I_c, through the R_c collector resistance. The voltage drop across R_c is the negative grid potential applied to the thyratron tube. Now when the output potential is increased the base current is increased, the collector current is increased, the voltage drop across the R_c resistor is increased and the negative grid potential at the thyratron tube is increased. This will delay the firing of the thyratron and thus reduce the output potential.

If the ohmic value of the R_e resistor in Fig. 20 is zero the voltage amplification of this transistor circuit will be about 10, or a small change in the output potential will result in about 10 times this change in the thyratron grid potential. This is voltage amplification added to the circuit by the grid battery eliminator and a voltage gain of 10 is more than present circuits can use. The emitter resistance R_e reduces the voltage amplification of the grid battery eliminator to reasonable proportions.

The R_t resistance in the potentiometer circuit of Fig. 20 is wound with nickel resistance wire. Its positive temperature coefficient of resistance compensates the grid battery eliminator circuit for the temperature effects in the reference voltage diode and the transistor. This nickel wire resistance accomplishes the same result as the thermistor in Fig. 19. This is another method of compensating transistor regulator circuits for ambient temperature variations.

The "Adjust Output Volts" potentiometer and the auxiliary rectifier shown in Fig. 20 are part of the present magnitude controlled thyratron tube rectifiers. The auxiliary rectifier adds some ac line voltage compounding to the rectifier regulation. It is also used with a time delay relay circuit, not shown, to bias the grid potential of the thyratron tubes

so that they will not fire during the required rectifier starting time interval.

4.3. *Magnetic Amplifier Regulated Rectifiers*

4.31. *2-Volt, 3-Ampere Regulated Rectifier*

The control of direct current at low voltage levels has been complicated by the lack of inexpensive low voltage reference standards, and by the very poor efficiencies of most rectifiers at low voltage. The new semiconductor devices have made important contributions in this field. Germanium diodes with their relatively low forward resistance seem naturally suited for use at low voltages, and the high sensitivity of junction transistors likewise makes them an almost ideal amplifier of small dc potentials.

Fig. 21 illustrates the use of junction diodes, junction transistors and a magnetic amplifier combined to furnish a regulated 2-volt, 3-ampere dc power supply. It will be noticed that the circuit of Fig. 21 is very

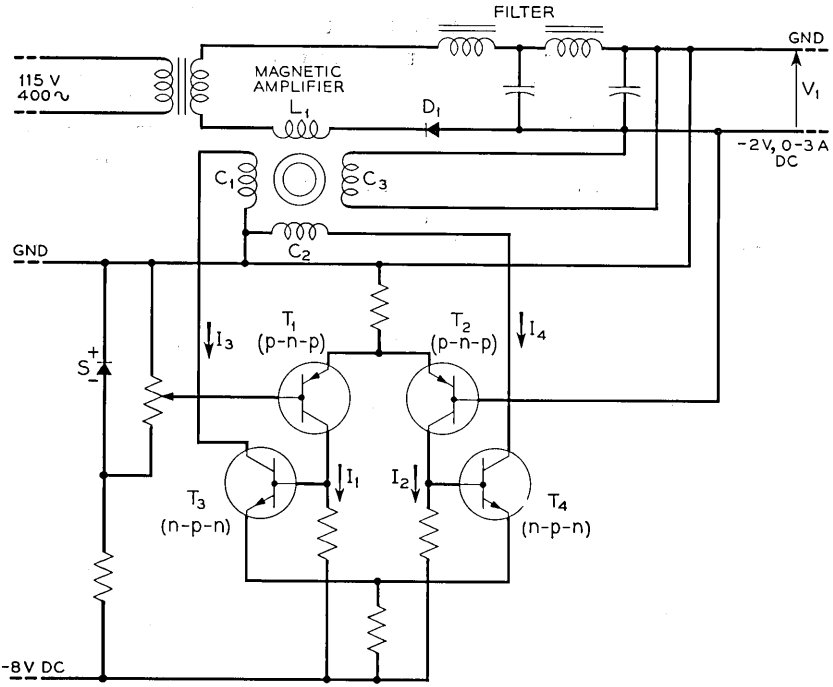


Fig. 21 — Two-volt three-ampere magnetic amplifier regulated rectifier.

similar to that in Fig. 18, except that an additional stage of current amplification has been added to the basic push-pull circuit.

Briefly, the regulating action is as follows. (1) The currents I_1 and I_2 respond in push-pull fashion to changes in output voltage *V1* as covered in Section 3.36, (2) currents I_1 and I_2 are amplified by the 2-watt n-p-n transistors (T_3) and (T_4), (3) the amplified currents (I_3) and (I_4) flow in control windings (C_1) and (C_2) of the magnetic amplifier to control the voltage absorbed by the power winding (*L*1), (4) this action regulates the average value of the voltage rectified by the germanium diode (*D*1), thus completing the feedback loop. Tests show that this circuit is capable of ± 1 per cent accuracy of the output voltage with a ± 15 per cent change in the line voltage and with load current variations of from 10 to 100 per cent of rated output current.

4.32. *65-Volt, 200-Ampere Germanium Rectifier*

Fig. 22 is a circuit sketch of a 65-volt, 200-ampere regulated rectifier suitable for charging and floating central office storage batteries. This rectifier employs six of the power rectifying cells with forced air cooling

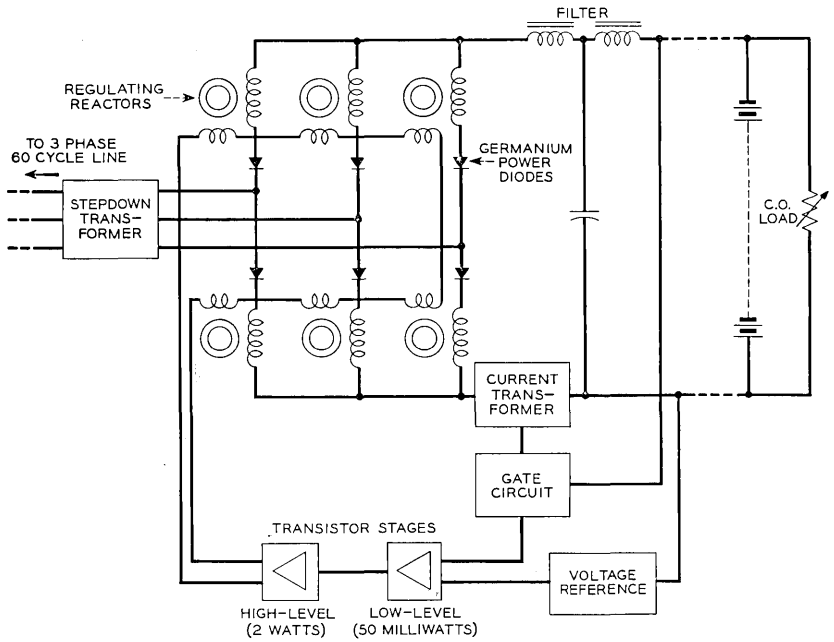


Fig. 22 — Sixty-five volt two-hundred ampere magnetic amplifier regulated rectifier.

described earlier (Diode IV, Fig. 2), a reference voltage diode (Diode I, Fig. 2), two 50-milliwatt, and two 2-watt junction transistors.

The dc output voltage of the rectifier is controlled by a high gain self saturating magnetic amplifier. High gain in the magnetic amplifier is achieved by using tapewound gapless nickel-iron cores having rectangular hysteresis loops. The control current for the magnetic amplifier is provided by 2, 2-watt n-p-n transistors acting in push-pull. The 2-watt transistors are driven by 2, 50-milliwatt p-n-p transistors also acting in push-pull. The circuit is similar to Fig. 21. Again, the reference potential is furnished by a reference voltage diode.

Where the rectifier is connected to storage batteries an additional feature known as "current droop" is needed to protect the rectifier. The output characteristic of the rectifier with current droop is shown in Fig. 23. This characteristic is obtained by coupling a signal proportional to load into the first stage transistor amplifier through a gating circuit. This signal is provided by a dc current transformer which is another form of magnetic amplifier. At currents below the "droop" value the current signal is blocked from the amplifier. At full load the gating circuit allows the current signal to take over and hold the output current constant over a wide range of output voltage. In Fig. 23, the performance

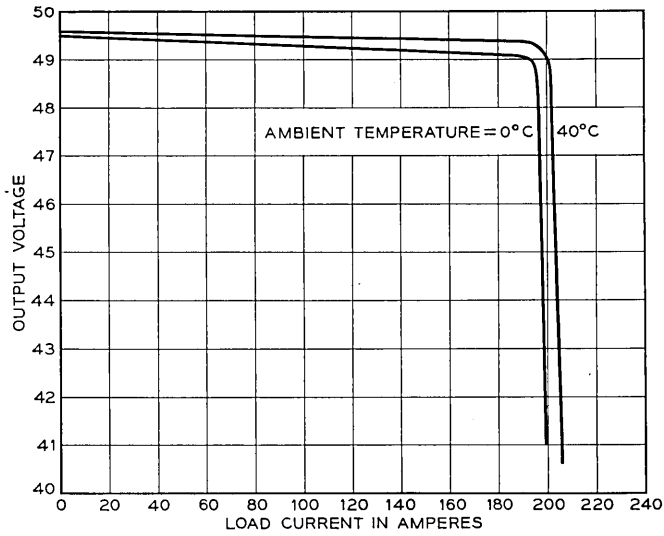


Fig. 23 — Output characteristics of experimental 65-volt 200-ampere germanium rectifier.

is shown over the range of ambient temperatures normally encountered in central offices.

5. CONCLUSIONS

It is seen from the above discussion that semiconductor junction diodes and transistors have a wide field of application in power conversion and control. Certain difficulties remain to be overcome, among which the variation of the device characteristics with ambient temperature appears to be the most troublesome at the present time. It has been shown that these variations with temperature can be minimized by two methods. First through the use of thermistors (negative temperature coefficient) or nickel-wire resistors (positive temperature coefficient) and second, through the employment of circuits in which the temperature variations of one element are balanced out by similar temperature variations in a complementary element. Thus, errors due to temperature changes can be minimized by further reduction of the sensitivity of the device characteristics to ambient temperature changes and by improved uniformity of the devices.

Another important aspect of the circuits covered in this paper is their freedom from dependence on auxiliary sources of dc potential. In most cases it is possible to power the regulating circuit directly from the regulated output, thereby eliminating the necessity for the transformers, rectifiers and filters usually needed to furnish plate potential for the regulating tubes and voltage standards.

The regulating circuits discussed in this paper are of the checkback type. In all of them, there must first be an error in the load voltage to start and maintain the regulating action. The load voltage will only return to precisely the original value if the regulating amplifier has infinite gain. These effects, however, are common to all closed-loop feedback regulating systems. Transistors and junction diodes, at their present stage of development seem well suited for use in checkback circuits having a high quality reference potential, for the feedback principle helps to minimize residual errors due to changes in the device characteristics with changes in ambient temperature.

Of course, the small size, long life and high efficiency of these semiconductor junction devices will also be very gratifying to the design engineers.

6. ACKNOWLEDGMENTS

The authors wish to acknowledge the assistance of C. W. Van Duyne of Bell Telephone Laboratories and the personnel in his group,

who were instrumental in obtaining the experimental data upon which this paper is based.

7. REFERENCES

1. W. Shockley, The Theory of p-n Junction in Semiconductors and p-n Junction Transistors, B.S.T.J., **28,** p. 435, 1949.
2. C. L. Rouault and G. N. Hall, A High-Voltage, Medium-Power Rectifier, I.R.E. Proc., **40,** p. 1519, Nov., 1952.
3. G. L. Pearson and B. Sawyer, Silicon p-n Junction Alloy Diodes, I.R.E. Proc., **40,** p. 1348, Nov., 1952.
4. R. M. Ryder and R. J. Kircher, Some Circuit Aspects of the Transistor, B.S.T.J. **28,** p. 367, 1949.
5. R. L. Wallace, Jr., and W. J. Pietenpol, Some Circuit Properties and Applications of n-p-n Transistors, B.S.T.J., **30,** p. 530, 1951.
6. R. N. Hall, Power Rectifiers and Transistors, Proceedings of the I.R.E. Proc., **40,** p. 1512, Nov., 1952.
7. W. Shockley, *Holes and Electrons in Semiconductors*, D. Van Nostrand, 1950.
8. K. G. McKay, Avalanche Breakdown in Silicon, Phys. Rev., **94,** p. 877, May 15, 1954.

Wire Straightening and Molding for Wire Spring Relays

By A. J. BRUNNER, H. E. COSSON and R. W. STRICKLAND

(Manuscript received January 19, 1954)

The basic design of the wire spring relay departs from conventional relay design in many ways. Translation of some of these design departures into commercial relay manufacture has necessitated the development of new machines and new methods because those available were incapable of producing to the new design requirements. Two developments in this category involved the straightening of large quantities of small diameter wire and the molding of a multiplicity of straightened wire inserts into phenolic resin blocks. The manner in which these developments were reduced from problems to practice is the subject of this paper.

Part I — Automatic Wire Straightening

Ordinarily wire is received from suppliers on spools or reels. In the spooling operation a spiral bend is placed in the wire which persists when it is unspooled. For use as a wire spring in the wire spring relay this spooling bend must be removed if the wire is to be positioned with the precision required for the desired functioning of the relay. It is necessary, also, to have the wire free of bends if automatic manufacturing methods are to be employed. For these reasons, it is important that the nickel silver and silicon copper wire used in the wire spring relay be straightened as the initial operation in the manufacture of wire block assemblies or "combs" for these relays.

Wire straightening can be accomplished by cold working the wire under controlled conditions until sufficient stress has been built up, particularly at the surface, to make the wire resist bending efforts. The degree of straightness required is governed, of course, by the desired performance of the comb in the operation of the wire spring relay. For the 0.0226-inch nickel-silver wire used in the twin wire comb this has been established, for example, as a deviation not exceeding 0.010-inch from absolute straightness measured at the contact end of the comb,

i.e. 2⅝-inches from its anchorage point in the phenol resin block. This degree of straightness is satisfactory also for the automatic manufacture of relay combs in which a multiplicity of straightened wires are guided into a molding die and positioned so accurately that they can be permanently imbedded in phenolic resin to the close dimensional limits necessary for ultimate assembly into relays.

EXPERIMENTAL WORK

Wire Straightening

The original experimental work on wire straightening was done at the Bell Telephone Laboratories to aid in establishing the feasibility of a wire spring relay design. After eliminating other approaches it was decided to straighten the wire in a motor-driven machine by pushing the wire through carefully oriented dies in a rotating head. The wire produced in this manner was known to have a twist but was adequate for making model parts. Subsequently, Western Electric development engineers made a survey of available commercial wire straightening machines. A machine was purchased which, while not intended for straightening wire of the small diameters used in wire spring relays, was capable of modification. Among the important things learned from the operation of this machine were first, it is preferable to push instead of pull wire through the rotating die head because of interference at the puller due to twist in the straightened wire; second, much of the twist can be removed from the straightened wire by spinning the spool of raw wire counter to the direction of the driven die head; and third, it appeared that a simpler approach than spinning the spool of raw wire would be to pass the wire through a second straightening head rotated in the opposite direction from the first. On the basis of these observations, a Hawthorne-designed experimental straightening machine was constructed. This machine featured two die heads independent of each other and counter rotating in operation. Five individual sets of die blocks, with provision for spacing adjustment as found on the rotary die holder of the commercial straightener, were retained in each head. Subsequently, this experimental machine was used for an extended series of tests to determine such things as the optimum spacing between individual dies, the proper offset from the center line of the head for each die, the best ratio of opposing head speeds, and the maximum rate of wire feed with respect to the rotational speed of the die holder head required to produce straight wire in the diameters employed in the wire spring relay.

Since it had become evident by the time this study was well advanced

that a multiple head machine would be required, a second experimental machine was built. This machine, Fig. 1, designed with the driving mechanism and spacing allowances considered necessary for an automatic multiple head straightener, had only one double head capable of straightening a single wire. A major change, to be discussed later, was replacement of the five adjustable die blocks in each head by a pair of opposing die blades contoured to provide a wire passage space between them identical to the predetermined path previously forced upon the wire by the five die sets. These die blades were retained by a spindle keyed to the drive mechanism. To accommodate the double head feature, a rotating unit consisting of two spindles coupled together was employed. Much of the remaining experimental work, such as optimum rotational speed of the spindles, the effect of different configurations of the die blade wire path surfaces, rate of wire feed, etc. was performed with this machine. Except for minor changes it became the prototype for the automatic multiple head machines constructed later.

Straightened Wire Storage

In contrast to the manual operations needed to assemble the springs and phenol fibre insulators of U- and Y-type relay spring pile-ups, it was planned from the beginning to mold straightened wire into phenolic resin by automatic means so that unit assemblies would be obtained for

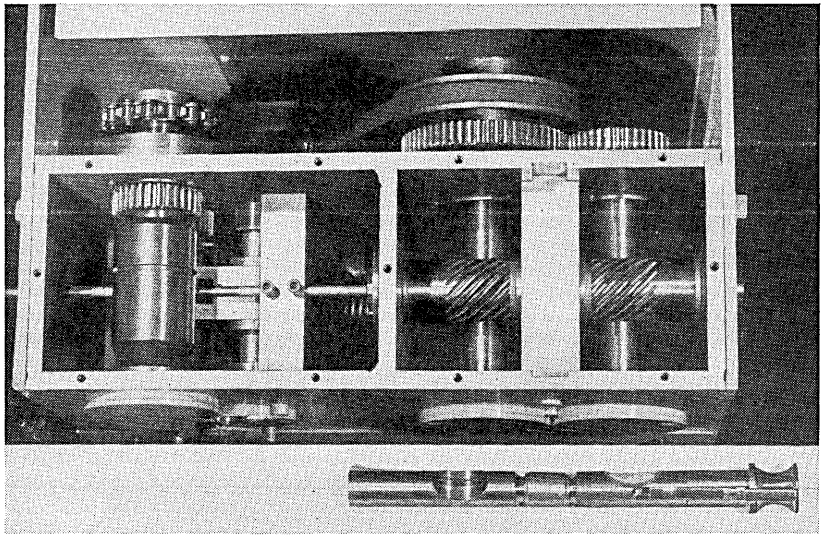


Fig. 1 — Experimental machine with one double head, prototype of 24 and 30 double head wire straightening machines.

the wire spring relay. The labor economy of the latter procedure is obvious. To implement this plan it was necessary that straightened wire be available at the molding press in the quantities required to prevent loss of molding time. To be successful it was important that interruptions to the regular recurrence of molding cycles, such as rethreading the multiplicity of wires into molding dies, be kept to a minimum.

The original planning envisioned a battery of single strand wire straighteners operating continuously to make relatively long lengths of straightened wire. How to store this wire between the wire straightener and the molding press presented the real problem. An early attempt toward a solution involved winding straightened wire on 36-inch diameter reels until sufficient length had been accumulated for eight hours' molding time. These reels would be mounted ahead of the molding press as shown in Fig. 2, and changed at the end of each eight hour shift.

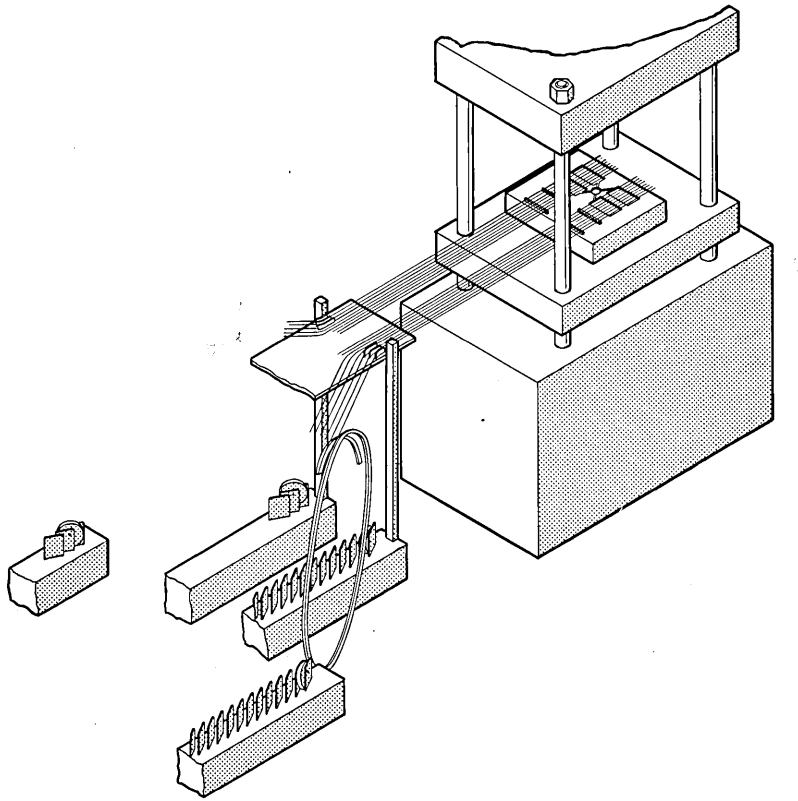


Fig. 2 — Sketch showing handling of straightened wire on storage reels.

Initial efforts indicated that this procedure was practicable. However, a new shipment of nickel silver wire revealed that, while not detectably different from previous shipments, the new wire took a permanent set on the 36-inch reels thereby making it unusable at the molding press. Principally because of the incipient possibility of straightened wire acquiring a set when not stored on flat surfaces, this reel approach was abandoned. Another effort consisted of providing a multiplicity of straight storage tubes, of either metallic or plastic material, into one end of each of which an eight hour supply of a single strand of wire was pushed by the wire straightening machine and from the other end of which a molding press would withdraw its requirement of wire (Fig. 3). This was found unworkable because often the wire straighteners were unable to push the required length of wire into the tubes due to the lead end becoming snarled from twist in the wire. It was decided, finally, to discard the idea of continuously straightening and storing wire in favor of placing multiple head machines adjacent to the molding presses and operating them only as required. This meant increasing the straightening machine investment because intermittent operation of the straighteners necessitated more wire straightening facilities. A compensating factor was the elimination of investment in storage facilities. It was found that interrupting the continuous operation of the straightening heads had no detectable effect on the characteristics of the straightened wire. Accordingly, multiple head automatic wire straighteners are now placed adjacent to the molding press and operated at a speed slightly greater than the wire consumption of the molding dies. Automatic control of the length of a partial loop of wire extending from the wire straightener assures an adequate supply of wire at the molding press. The ultimate length of continuous straightened wire available to the molding press by this arrangement is governed only by the length of raw wire on the spool and is sufficient for many operating shifts.

AUTOMATIC MULTIPLE HEAD WIRE STRAIGHTENER

Both 24- and 30-double head automatic wire straighteners have been built by the Western Electric Company. The 24-double head straighteners are used in making combs for the AF, AG and AJ type general purpose wire spring relays and the 30-double head machines for the 286, 287 and 288 type multi-contact relays. In practice the phenol resin molding operation is accomplished in four-cavity dies with the cavities arranged symmetrically about the center of the die. Thus two forward cavities face the wire straightener with the remaining two cavities in tandem. When making the twelve wire single comb of general purpose

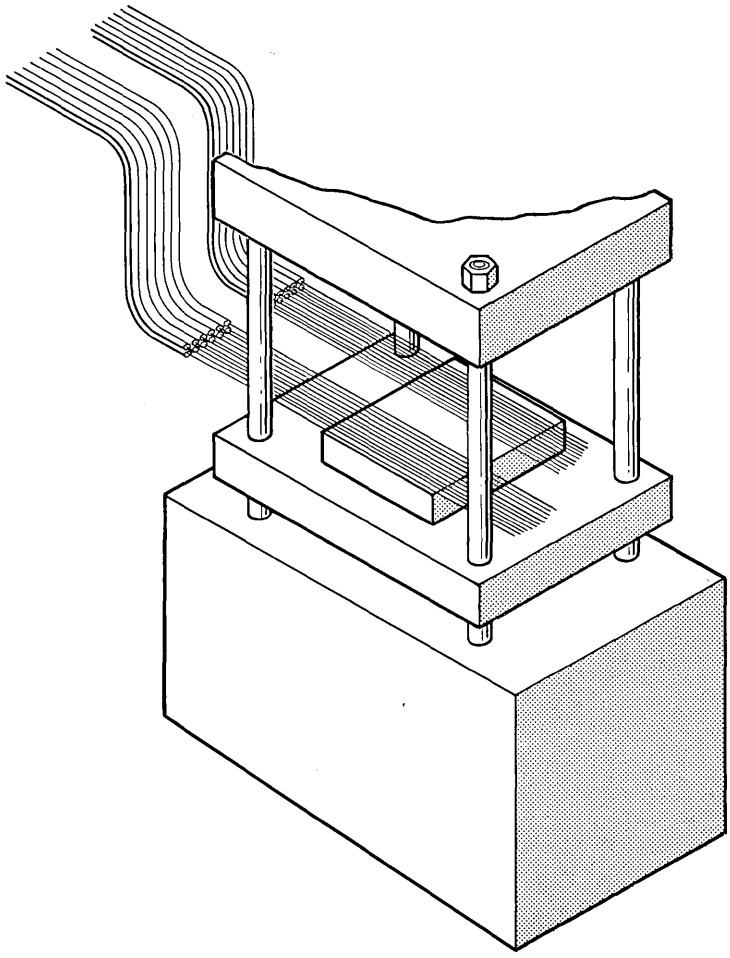


Fig. 3 — Sketch showing method of handling straightened wire from storage tubes.

relays, half of the wires from the 24-double head straightener are guided into each forward cavity. When the twenty-four wire twin wire comb is being made, on the other hand, the entire production of a 24-double head straightener is guided into each forward cavity and two wire straighteners are necessary for each molding press.

The 30-double head straighteners are arranged similarly when the fifteen wire single wire comb and the thirty wire twin wire comb of the multi-contact relay are being molded.

WIRE SUPPLY

One spool of raw wire is cradled in the wire straightener, Fig. 4, for each wire required in the molded comb. There are three sizes of wire straightened, 0.0200 and 0.0226-inch diameter wire for the twin wire comb of multi-contact and general purpose relays, respectively, and 0.0400-inch diameter wire for the single wire comb of both relays. The smaller wires are nickel silver while the 0.0400-inch wire is a silicon-copper alloy. All three wires are in the hard temper range.

Originally the wire was pulled from the spools by the drive (pusher) roll of the wire straightener. However, the pulling force required varied widely from spool to spool. The result was an unequal rate of wire feed through the straightening heads. To avoid this, a capstan with an individual pulley adjustment for each wire was added to the machine. This capstan, in addition to pulling the wire from the supply spools, meters the amount of wire fed into the straightener. An occasional adjustment of individual capstan pulleys is all that is necessary now to assure production of straightened wire at a uniform rate from every head.

WIRE STRAIGHTENING MECHANISM

Fig. 5 shows the wire straightening mechanism. Some of the spools of raw wire are visible to the right below the 24 wires, in this instance, being pulled from the capstan pulleys by the grooved shaft mounted just inside the machine housing. This shaft has 24 grooves, one for each wire, which mate with twelve spring tensioned twin grooved wobble rolls underneath to provide the means for pushing the wires through the straightener heads. Both the grooved shaft and the twelve mating rolls are power driven. The wires are pushed through the tubes to the left of the grooved shaft which guide them to the spindles in the straightening heads. These heads, arranged in two vertical rows, make it possible for a common spiral geared drive shaft to rotate all 24 heads at identical speed. This arrangement, however, causes twelve of the heads to rotate clockwise and twelve to rotate counterclockwise. The opposite twists produced in the upper and lower wires under this circumstance are corrected for by the double head arrangement in which the second set of heads rotate counter to the first set.

DIE BLADES AND SPINDLES

Inside each head is a removable spindle for retaining a pair of contoured wire straightening die blades. The spindles are suitably keyed

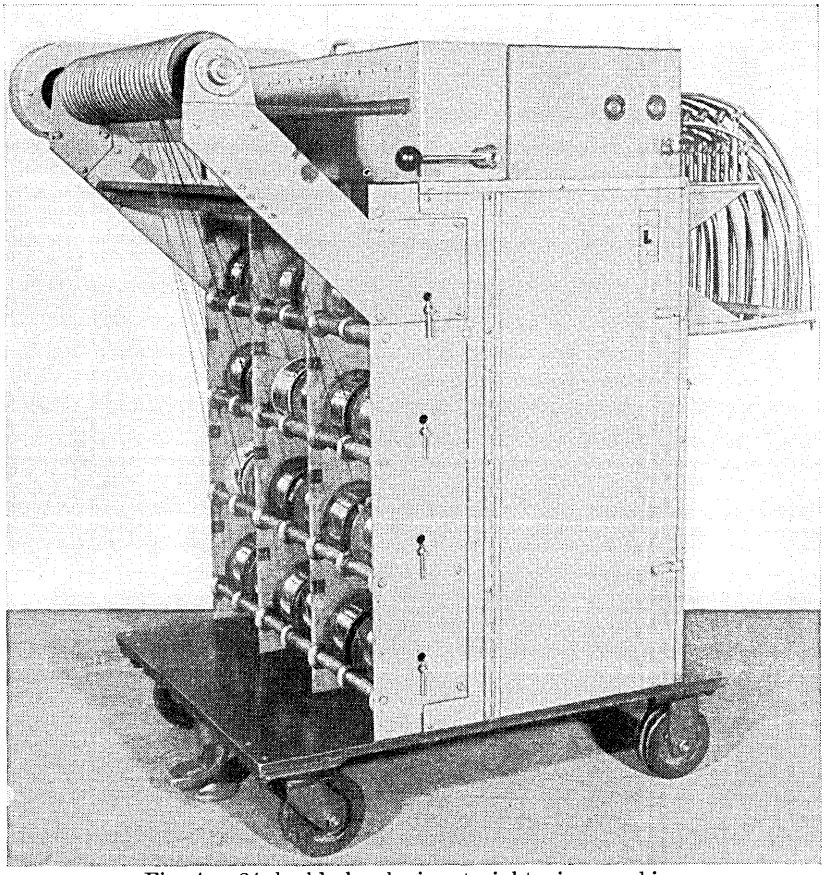


Fig. 4 — 24 double head wire straightening machine.

to the heads to assure rotation. To conform to the double head design two spindles, joined by a loose coupling, are used. This is illustrated in Fig. 6 which also pictures two sets of die blades removed from the spindle slot. The space between each pair of contoured die blades as positioned for the photograph shows clearly the path of the wire during its transit of the rotating heads.

The die blades maintain the same spacing, offset and length that constituted the desired wire path through the five individually adjusted sets of die blocks of the early straightening machines. The continuity of a die blade is accomplished simply by bridging what had been air spaces between the individual die elements and removing enough metal to prevent wire contact in the bridging sections.

The die blades are used not only to conserve space but also to minimize die costs. The latter is accomplished by making them from inexpensive sheet metal on a punch press and discarding them as soon as wear has destroyed their usefulness for wire straightening. Unlike the individual die blocks used in the previous rotary head straighteners, it is not necessary to groove these die blades to direct the flow of wire through the head. There is a slot milled into each spindle to hold the die blades as shown in Fig. 6. The walls of these slots guide the wire and limit its sideways movement in much the same manner as the grooves in the individual die blocks. The actual thickness of the die blade was established as slightly more than that of the diameter of the largest wire to be straightened for wire spring relay combs. Thus, one slot of uniform width is milled into each spindle allowing interchangeability of spindles regardless of the diameter of the wire to be straightened.

Wire in its transit through the die blades is flexed and burnished to the extent required to produce the desired degree of straightness. It is not rotated during the straightening operation but may acquire twist and even a spiral threadlike burnished appearance from rotation of the die blade surfaces.

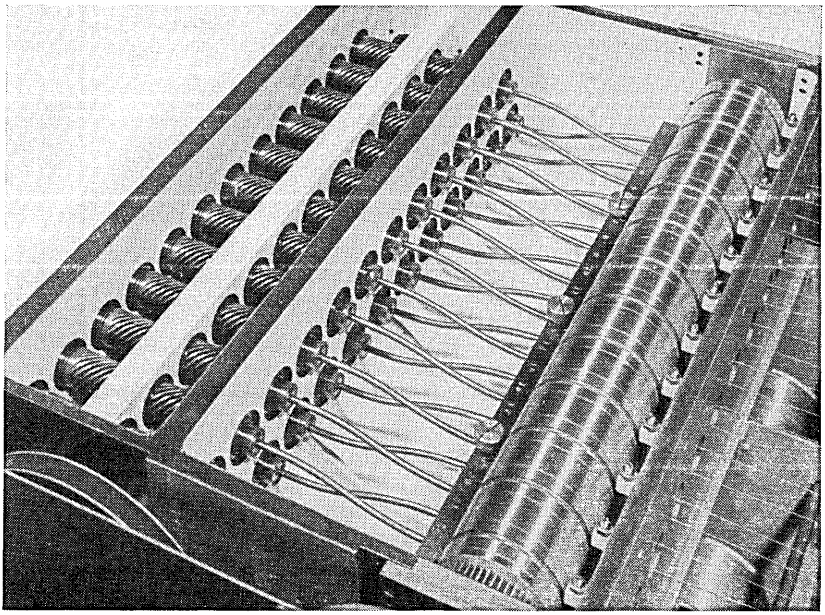


Fig. 5 — Straightening mechanism of 24 double head wire straightening machine.

OBSERVATIONS ON WIRE STRAIGHTENING

The straightening operation affects some physical properties of the wire. Tensile strength is reduced about 10 per cent while elongation is increased around 50 per cent. The diameter of the straightened wire is usually from 0.5 to 1.0 per cent greater than that of the raw wire with commensurate loss in wire length. Both straightness and twist appear to be dependent in large part upon the contours of the die blades. Thus far these contours have been determined by trial and error on the adjustable die block straightener, although general relationships, especially with respect to wire size, are becoming evident. It is expected that further study and experience will establish bases on which contours can be calculated with accuracy.

Twist imparted to the wire by the rotating action of the spindle has been found difficult to measure. What is referred to as twist is actually

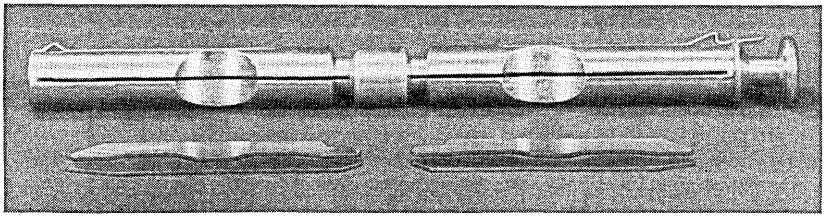


Fig. 6 — Double spindle showing slots and complement of two sets of die blades.

radial distortion of the wire about its longitudinal axis resulting from partial release of internal stresses remaining in the wire after straightening. Further release of internal stresses may occur when the wire ends of the twin wire comb are formed before welding, in which event mislocation of contacts will result, Fig. 7. This is objectionable from the standpoint both of subsequent manufacturing operations and of relay performance. The internal stresses are caused by the crank action applied to the wire surface while it is passing between the die blades in the rotating spindles. Internal stress which is not apparent until after its release, as by forming, has been designated as "residual twist".

A rough approximation of the amount of residual twist in wire can be obtained by measuring what has been termed "apparent twist". Apparent twist is the amount of visible rotation at the end of a wire after leaving the straightener. It can be measured in degrees of rotation per foot of wire straightened. When the apparent twist is low, usually the residual twist also is low. A working range for permissible apparent

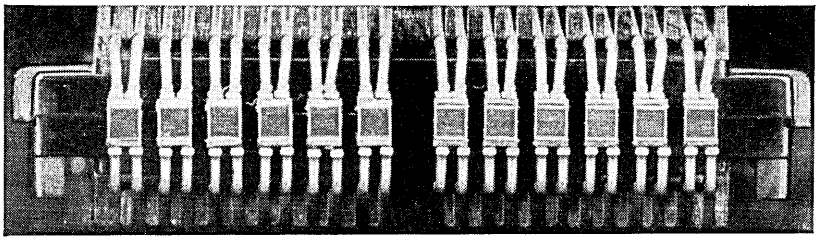


Fig. 7 — Photograph showing the affect of residual twist in the upper set of wires as compared to freedom from residual twist in the lower wires.

twist has been established which has been successful generally in maintaining acceptable limits on residual twist.

There are three variables which largely control the quality of straightened wire. The first is the physical properties of the wire itself. Although a shipment of wire may, on the basis of the sampling method employed, meet specification requirements limiting physical and chemical characteristics, an occasional spool or part spool of wire can be expected which will be enough outside limits to cause unsatisfactory straightness and unmanageable residual twist.

The second variable affecting straightness and twist in wire processed on multiple head machines lies in small differences between the rotating head assemblies. While all critical dimensions of the die blades, spindles, and spindle housings are held to close tolerances, it is possible to obtain an accumulation of dimensional deviations in some spindle assemblies of such magnitude as to cause appreciable difference in wire twist and sometimes in wire straightness. It has been necessary, therefore, to provide means for balancing such dimensional variations.

The third variable is wear on working surfaces of the die blades. Continued sliding of wire over die blade surfaces eventually produces grooves which decrease offset and increase clearance in the wire path. It has been found in general that, as the die blades wear, twist decreases until it eventually reverses direction. Simultaneously, straightness may improve to a critical point from which it rapidly deteriorates. Accordingly, replacement of die blades must be made before wear has rendered them ineffective.

CONCLUSIONS

Satisfactory performance of the multi-head wire straightening machines described has been demonstrated during the pilot plant period of wire spring relay manufacture. Further refinements in the means

for controlling known variables must be made, however, to assure the reliability demanded of heavy duty mass production machines.

Part II — Automatic Molding of Wire Spring Relay Block Assemblies

Parallel with the effort directed toward development of wire straightening facilities, an investigation was undertaken by Western to develop automatic facilities for molding an array of straightened wires into small plastic blocks spaced at specified intervals. These blocks were designed not only to hold the wires securely and to locate them accurately but also to insulate them from each other electrically. The design engineers at Bell Telephone Laboratories had decided, after evaluation of the physical properties of available plastic molding materials, that a thermosetting phenolic type resin would best provide the characteristics needed for wire spring relay block assemblies. Proceeding on this information, Western Electric development engineers reviewed the merits of molding methods adaptable to embedding a multiplicity of inserts, wires in this instance, into phenolic resin. Such economic factors as molding time and material cost were balanced against molding problems like shrinkage and flow characteristics. It appeared from this review that transfer molding offered the most favorable possibilities. It appeared also that the shortest practicable molding cycle might be achieved by preforming the phenolic resin material, preheating these preforms electronically and automatically feeding them into the molding die. Further study of the molding problems indicated that molding presses for this purpose would have to be specially designed, particularly if multicavity dies were to be used. The special design features, such as wider spacing between the tie rods, provision of an electronic preform heater, and microtiming devices, are discussed later as they become pertinent to the description of the machines finally adopted.

AUTOMATIC MOLDING MACHINE

Hydraulic molding presses appeared to offer the most advantages for this job. Essentially, these consist of two opposed hydraulic rams mounted vertically; the lower and more powerful ram providing the force required to close and hold closed the split die employed and the upper ram providing the force needed to transfer the phenolic resin in a softened or plastic state into the die cavities.

Unusually wide spacing between the tie rods of the press had to be provided to accommodate the complex progressive die required to make

the molding operation automatic. This die had to be designed not only to mold resin in multiple cavities but also to remove the plastic cull, index the molded blocks, and shear the completed assemblies. Provision had to be made, in addition, for mounting an electronic preform heater and for space to install an appropriate conveyor from the heater to the plastic transfer point of the die. These features were incorporated into an experimental prototype machine, shown in Fig. 8, operated under laboratory conditions. Adjustable microtiming devices were engineered into this machine to control the sequence of operations precisely and automatically.

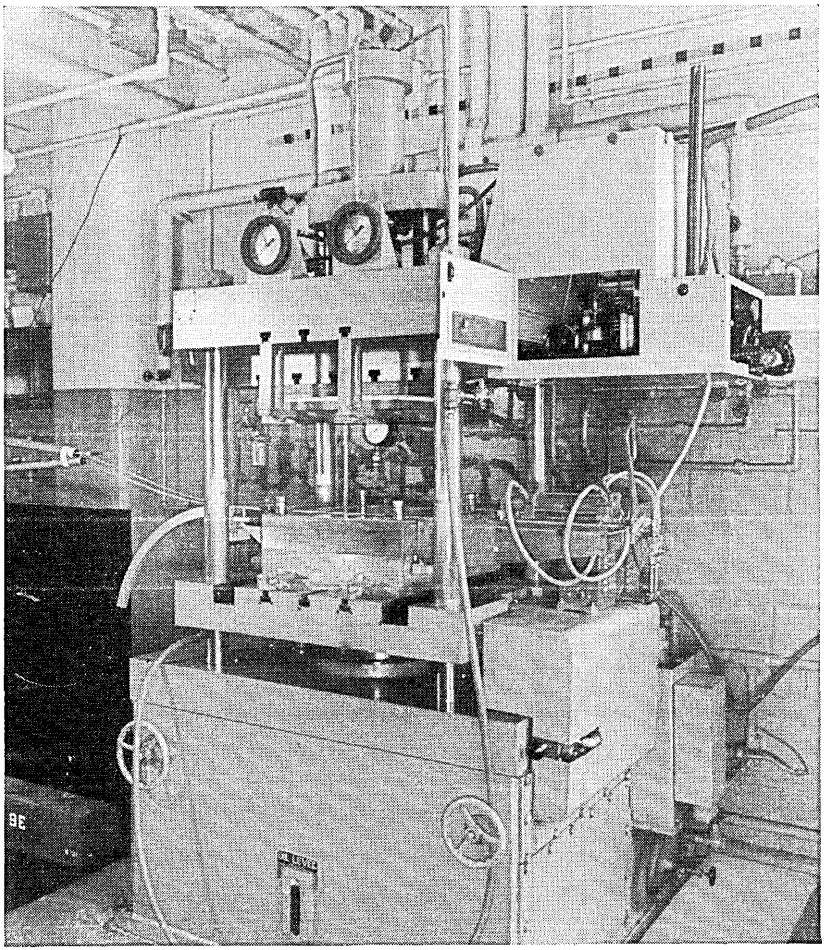


Fig. 8 — Experimental prototype automatic transfer molding machine.

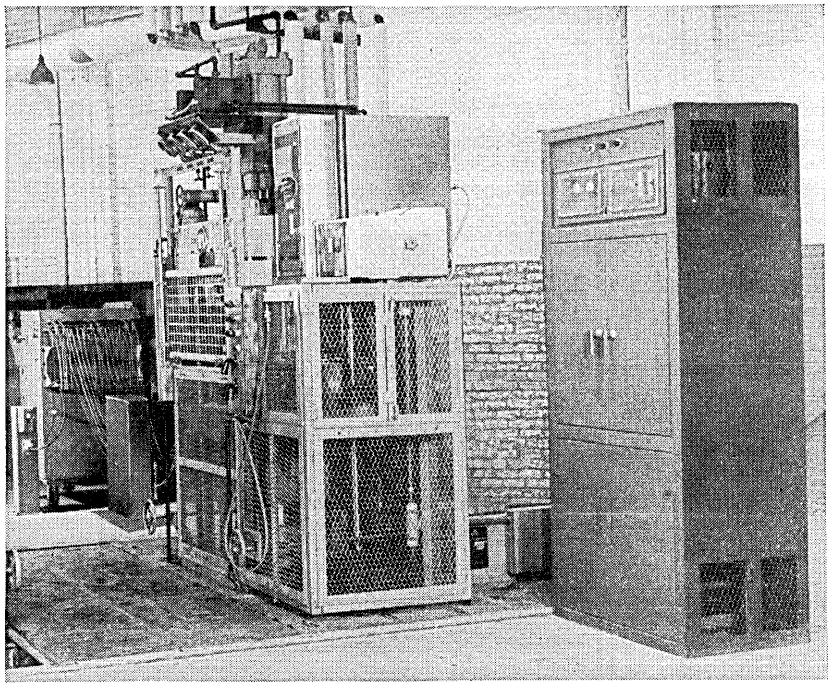


Fig. 9 — Typical installation of wire straightening and molding machines.

The final embodiment of the molding machine is shown in Fig. 9. The wire straightener included in the photograph is positioned a short distance from the molding machine to allow straightened wire to form a partial loop between the two. This permits the wires to leave the straightener at a constant speed from a fixed position and to enter the die at intervals controlled by the molding cycle and move vertically with the opening and closing motions of the lower half of the die. The straightening machine is started and stopped by an electrical control which assures the desired size partial wire loop at all times.

PHENOLIC RESIN PREFORMS

To operate on an automatic basis, it is important to have the molding compound in such form that it can be handled easily and that the charges are of uniform weight and size. This is done by compressing the bulky granular compound as received from suppliers into small carefully dimensioned cylinders or "preforms". Mechanical presses, capable of turning out preforms in multiple at each stroke, are used at Hawthorne

for this purpose. To obtain uniformly low water content in these preforms, it is necessary to store them in an air conditioned chamber for a three week period to assure attaining equilibrium conditions.

PREFORM HEATING

An electronic preform heater is a means of increasing molding machine production by shortening the time the phenolic resin must be retained in the molding die during each cycle. This is done by adding to the preform, just before it enters the die, much of the heat required to plasticize the resin. Thus, as one charge of compound is being molded in the die cavity, another is being preheated as part of the molding cycle. The amount of preheat that can be permitted is limited by the extent to which heat induced chemical reaction can be tolerated outside the molding cavity and varies with the size and contour of the die cavity. The rate at which the preform is heated influences its consistency at a given temperature.

A feed mechanism has been provided to convey the preform from a magazine to the electronic heater and thence to the die. This mechanism consists of a horizontal guide plate extending between the two grids in the upper part of the press. The preform is pushed along this guide plate by a metal bar mounted on endless roller chains at each side of the guide plate. In operation, a preform from the magazine is pushed to a position between the electrodes of the dielectric heater. The push bar is then backed away a small distance so that it will not interfere in the inductive field. Upon completion of the preheating operation, the push bar shoves the heated phenolic preform to and beyond a drop off position above the open end of the transfer cylinder of the molding die. The conveyor continues to operate until the push bar has removed a new preform from the magazine and loaded the heater in preparation for the next cycle.

PRESS CONTROLS

The electrical controls or "brain" of the production unit are housed in a cabinet adjacent to the molding machine. The operation of the press, the electronic heater, the feed mechanism and the pneumatic device on the die are all coordinated into precise sequences by micro-adjustment of these controls, Fig. 10. Any operational sequence or length of cycle desired can be established for repetitive manufacture. On the other hand, the press and the electronic heater can be placed on manual control at any time.

THE MOLDING DIE

Experimental Work

No attempt will be made to discuss the many ideas on mold design which were conceived, evaluated and either discarded or improved in arriving at the designs now employed. The investigation was undertaken

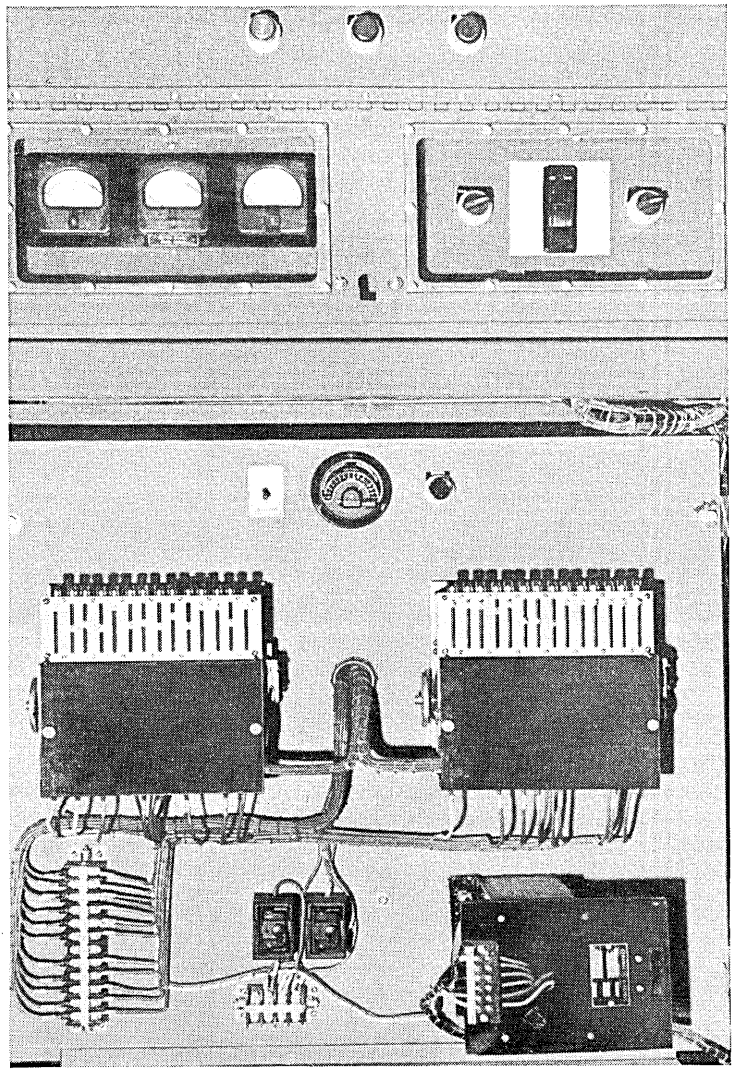


Fig. 10 — Cycle timing and electronic heater controls for molding machine.

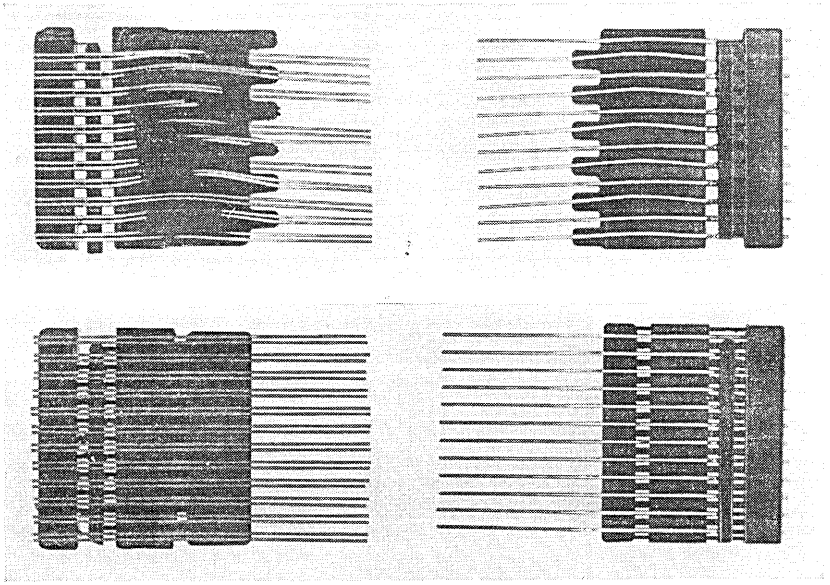


Fig. 11 — Cutaway sections of resin blocks showing inadequate versus adequate wire supports in molding die.

with no significant experience in molding closely spaced arrays of small diameter wires into phenolic resin. The initial effort demonstrated convincingly, Fig. 11, that small diameter wires cannot be embedded in resin by high pressure molding techniques without the liberal use of wire supports. These are needed to prevent individual wires from being deformed by the pressure of the plastic as it is forced into the cavity. One fundamental observed in die design subsequently was to keep all wire spans inside die cavities as short as possible.

As expected, early studies demonstrated that lack of cross sectional symmetry caused combs to have a marked tendency to warp. While warping could be reduced by increasing the time the resin was retained in the molding cavity, this partial solution was unsatisfactory from the standpoints both of warpage and product cost. Accordingly, every effort was made, consistent with relay design requirements, to depart as little as possible from symmetrical die cavity design and where symmetry could not be achieved, to attempt to distribute the resin mass uniformly on each side of the center line of the wires.

The importance of symmetry, together with the desire to keep molding flash to a minimum, influenced Western development engineers to design the earlier experimental die cavities with the wire inserts centered at the

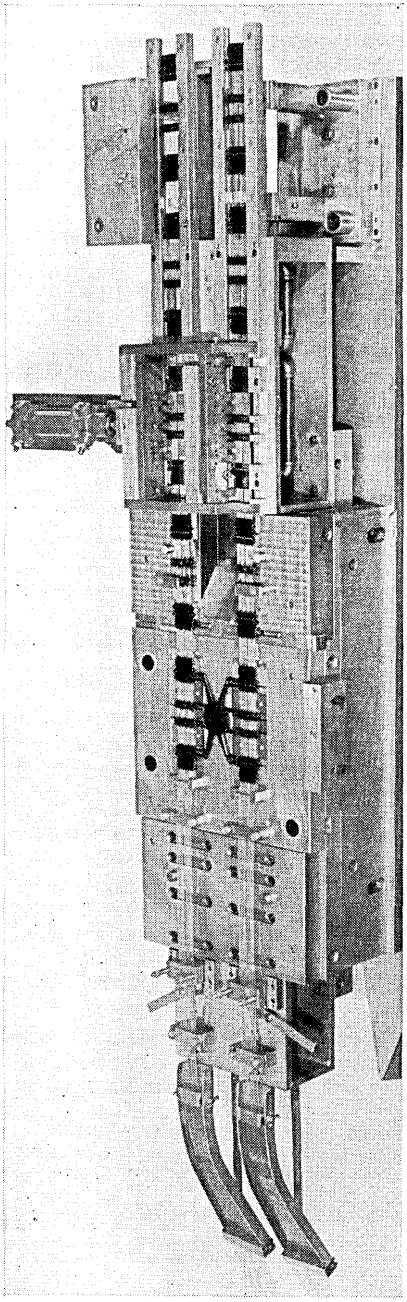


Fig. 12 — Lower half of typical automatically operated four cavity die.

die parting line. In the course of die development work it became apparent that reduction in both die cost and die construction time could be effected in some instances by confining the grooves for locating and supporting the wire array to one half of the die cavity. This type of design had the added advantage of eliminating annoying problems relating to precise registration of the upper and lower die halves during the molding operation.

The design of the die cavities now used for high production molding represents development work based on such considerations as those outlined above. The die cavities are similar but not the same for both twin and single wire combs. They differ in the location of the parting line which is flush with the top of the wire array in the twin wire comb die and principally at the center line in the single wire comb. The latter arrangement was dictated by two considerations, (1) Use of a U shaped groove in one half the die cavity for a wire as large as 0.040-inch diameter, resulted in excessive molding flash and (2) The chance of obtaining completely filled fins on both sides of the forward block was enhanced by centering the wires.

THE DIE

Fig. 12 shows the lower half of the complex automatically operated four cavity die which was evolved from simple single cavity hand loaded units. That this evolution may not be completed is indicated by the fact that the section of the die operated by the air cylinder shown at the top of the photograph and intended to remove molding flash from the wires is no longer used. Much better flash removal is effected by blasting the combs with ground walnut shells after the molding operation has been completed. The operation of the automatic die, as related to the manufacture of single wire combs containing twelve 0.040-inch silicon copper wires, will be described under five sub-heads in the same sequence as the progression of the wire array through the five sections of the die. Similar progressive operations are performed in making twin wire combs. The die consists of an upper half attached immovably to the head of the press and a lower half mounted on a hydraulic ram which raises and lowers to close and open the die.

WIRE GUIDING

The single wire comb die uses one 24-double head wire straightener. Fig. 13 shows how twelve continuous strands of wire from the straightener are guided to the required spacing in each comb array. Spring ten-

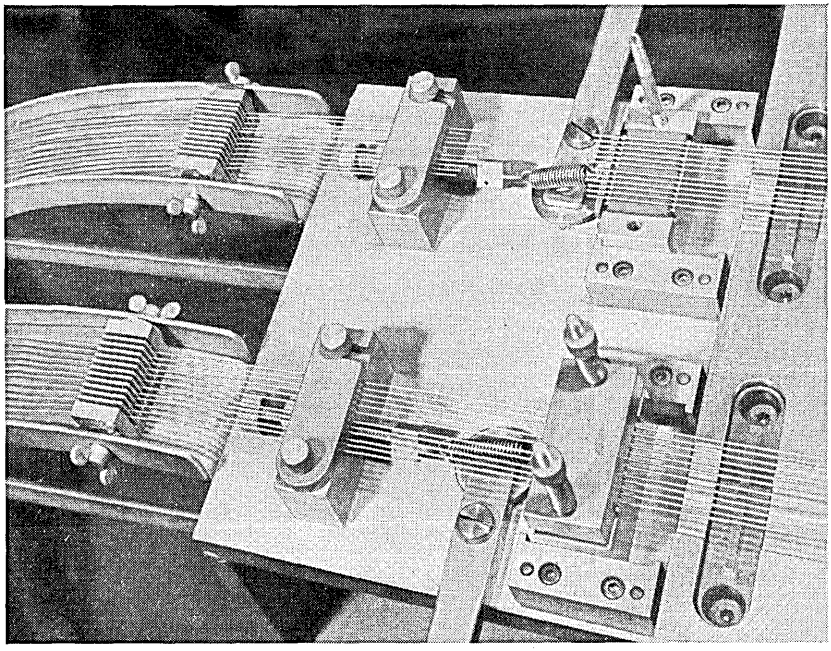


Fig. 13 — Wire guide section of automatic die.

sioned sliding blocks with spring loaded felt cleaning pads hold the wires taut during the operations that follow.

ANCHORING

Because phenolic resin does not wet and, therefore, does not bond with most metals, mechanical means must be used to prevent the wires from turning in the resin block of the finished combs. This precaution is necessary because wire wrapping tools are employed in wiring relays into equipment. The wrapping tool puts a torque on the wire which may, if the wire is not securely anchored, turn it in the resin block. Any movement, obviously, will mislocate the contact welded onto the wire and cause maladjustment of the relay. To mechanically prevent turning, a section of each half of the die is provided with accurately spaced mating blade inserts, Fig. 14, which press flat lands or "anchors" on the wire each time the die is closed. When the wires are indexed subsequently these anchors are located on that portion of the wire embedded in the resin.

MOLDING

Dies with four cavities are used for all comb molding operations. The cavities are located around the plastic transfer area as illustrated in Fig. 15. Above the transfer area a cylindrical opening extends vertically through the upper die half to permit passage of the transfer ram. By closing the die, the transfer area is enclosed except for small orifices leading to each of the four cavities. The heated preform is dropped into the center of the transfer area ahead of the transfer ram. This ram, operated by hydraulic pressure and heated by contact with the hot die, forces or transfers much of the heat softened resin through the orifices and into the die cavities. The resin residues which are left in the passages between the ram and the cavity gates upon completion of the molding cycle are called runners. The heat of the die, approximately 360°F., further softens the plastic resin enabling the pressure of the ram to force it into intimate contact with all die cavity surfaces. Continued application of heat and pressure hardens or cures the resin by the process of polymerization, after which it is permanently infusible.

To successfully operate on an automatic basis, the cluster of wire, plastic blocks, ram slug, and runners must be removed from the molding cavity. This is complicated in the single wire comb because the plastic block at the contact end of the wires has very thin fin sections

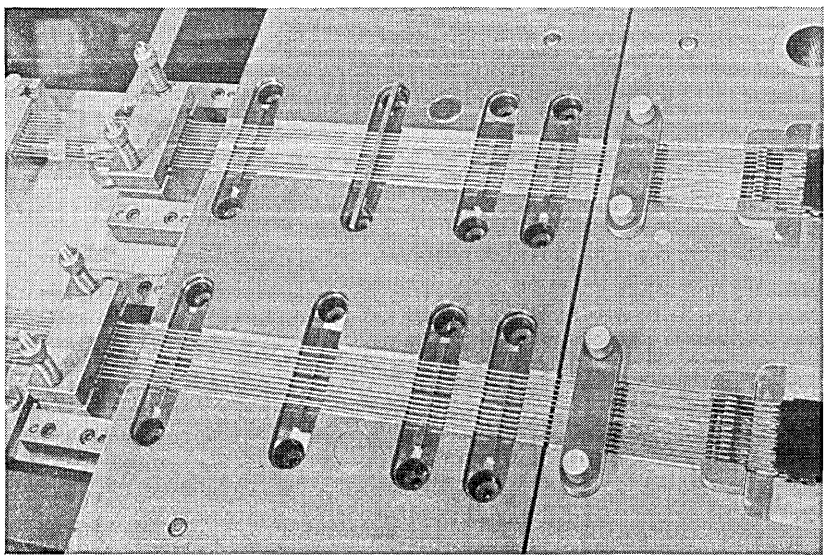


Fig. 14 — Anchoring section of automatic die.

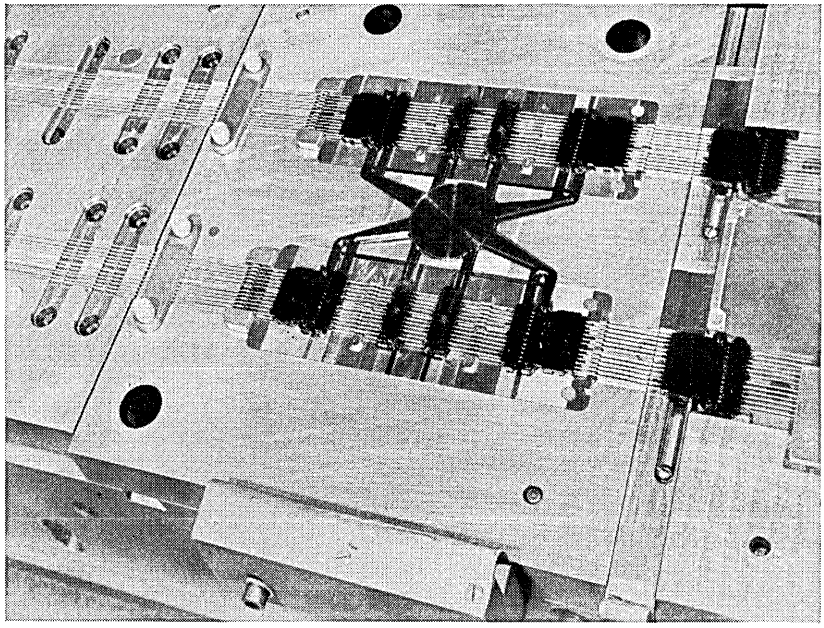


Fig. 15 — Molding section of automatic die.

projecting into both halves of the die cavity. These fins, used to guide the mating twin contact wires in the finished relay, must be held to close limits dimensionally and are thin to the point of fragility. To insure satisfactory removal of the resin blocks from the upper half of the die while simultaneously retaining them in place in the lower half when the die is opened, spring loaded ejector pins in the upper die half push down against the cured blocks. The ejector pins operate until the wire-resin cluster has been ejected from the upper die. Subsequently they are retracted by contact with reset pins which butt against the lower die half when the die is closed. The openings in the die half which accommodate these ejector pins serve as air vents when the resin is entering the cavity. To prevent the plastic cull, i.e., the resin slug and associated runners, from breaking off and damaging the die on the next molding cycle, the pressure of the transfer ram is maintained on the slug until the upper die surface has been cleared.

The hydraulic ram bearing the lower die half continues to withdraw until the lower ejector pins can function. These ejector pins are more numerous and more complex in design than those in the upper die half because, in addition to ejecting the resin blocks from the die cavities,

they aid in guiding the progression of wire inserts through the die and in locating the incoming wires in the die cavities. The operation of this ejection or "knock-out" consists of freeing the wires and resin blocks from the die cavities and then moving the newly molded cluster horizontally from the lower die cavity. When the die surface has been cleared by this indexing operation, compressed air is blasted across the die to remove loose molding flash which might be present. With the completion of the indexing operation, the ejector pins are restored to their original position by spring pressure. As a precaution, reset pins are provided in case this spring pressure should be inadequate.

COOLING AND CULL REMOVAL

The next operations are those of cooling the plastic to minimize warpage and removing the cull. Because the larger plastic block of the single wire comb has surface irregularities on one side and is smooth on the

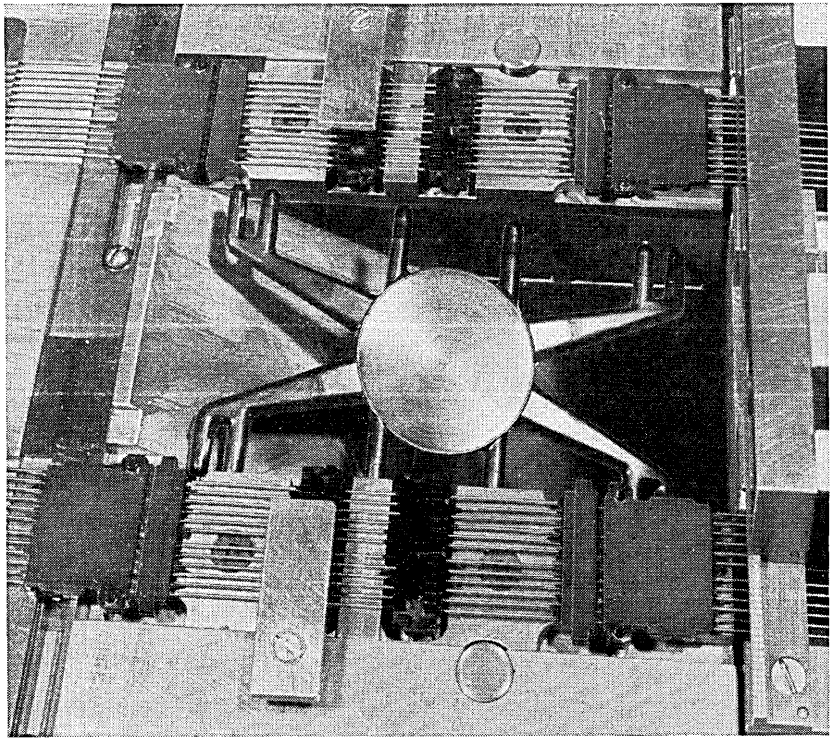


Fig. 16 — Cooling and culling section of automatic die.

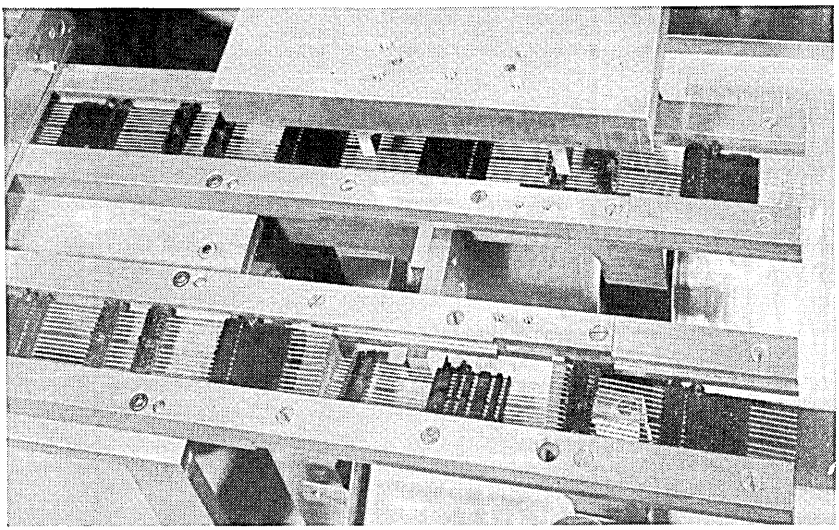


Fig. 17 — Indexing and shearing section of automatic die.

other, there is some tendency for it to warp upon cooling. To prevent this, when the die is closed the plastic embedded wire assemblage is clamped against spring loaded steel pads operating between water cooled plates and retained in this position during the next molding cycle. The closing of the die also causes the cull to be sheared off, Fig. 16. This waste material slides down a chute for removal from the press. A locating stop is built into this section of the die to establish the proper spacing for inserting a new progression of wires through the die should such action become necessary as when wire from new spools must be placed in the dies. This locating stop can be used also as a check at any time to determine whether the parts are being indexed the proper distance.

INDEXING AND SHEARING

The mechanism for indexing the continuous strands of wire through the die is located beneath and parallel to tracks built into the last die section to accommodate and guide them. It is driven by a timer controlled pneumatic cylinder to which feed heads are attached by a yoke. The feed heads reciprocate on rails beneath the guide track and transmit their motion to the ladderlike train of assemblies by a spring loaded dog which engages the resin blocks on the forward stroke. The return stroke is carried out after the press is closed.

The last operation performed by the die is that of shearing the wire in the molded assembly to form individual combs. One set of opposing shearing details for each line of molded assemblies is adjustably mounted on the base plate of the lower die half, Fig. 17. When the die is closed, the upper details butt against the upper die plate thereby forcing the cutting shears upon the wire. To reduce the force required, the cutting blades are so tapered that they cut each wire in succession.

Upon termination of the cutting operation, the parts fall free of the guide rails into chutes leading to the front of the press, thus completing the molding operation.

CONCLUSION

The original objective of embedding a multiplicity of straightened small diameter wires in phenolic resin blocks (Fig. 18) on a commercial basis has been accomplished. These wire spring relay parts are being produced at low cost to the required dimensional accuracy in automatic molding machines.

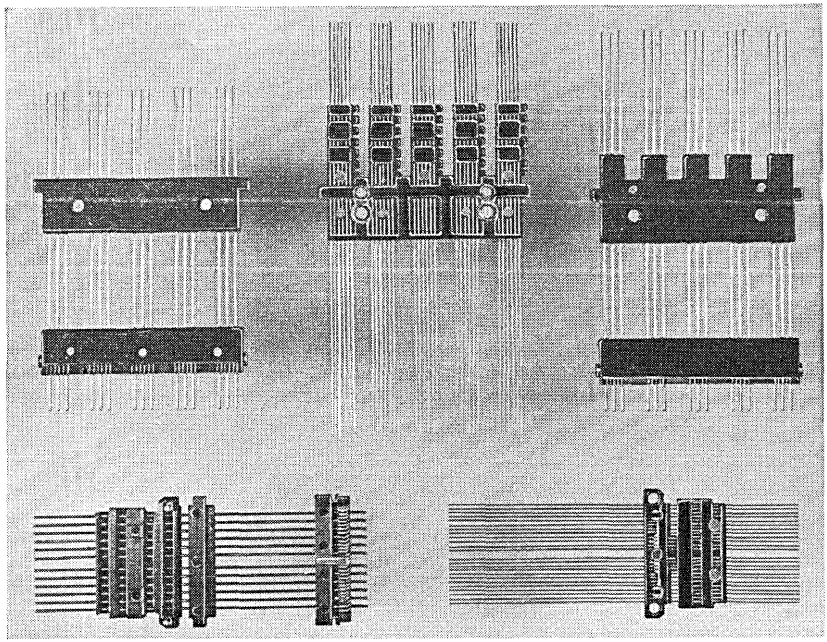


Fig. 18 — Wire block assemblies as manufactured for wire spring relays.

ACKNOWLEDGMENT

The authors wish to acknowledge the many contributions of fellow engineers to the development work which they have reported. These engineers include; in the wire straightening study, C. Paulson, A. E. Swickard, L. L. Mazza and the late N. K. Engst; and in the molding investigation, F. A. Schultz.

Some Fundamental Problems in Percussive Welding

By ERIC EDEN SUMNER

(Manuscript received February 5, 1954)

The basic processes of percussive welding are presented. Large variations in arc duration result from the spread in initiation separation, magnetic bridging effects, and the amplifying effect of evaporation. Higher voltages are shown to decrease the relative spread of initiation separation. An analysis of bridging suggests minimizing the ratio of current to separation. A welding circuit offering independence from arc-duration variations is developed. The use of a capacitative transmission line, or approximations thereto, has resulted in greatly improved process control.

INTRODUCTION

Early work in percussive welding goes back to late in the nineteenth century. Both applications and accounts in literature are relatively rare. However, this type of welding should have considerable applicability in view of some rather outstanding advantages:

1. The fact that the arc potential is approximately 15 volts permits the addition of considerable energy within a very short time and, relative to resistance welding, small currents for shorter times are possible. This allows the welding electrodes to be placed well away from the weld zone without overheating of adjacent areas. Effects of deflection due to the high electrode clamping forces can be minimized.
2. The compatibility problem between the materials and geometries of the parts to be welded are eased relative to the slower butt welding method.
3. The welds produced in a controlled process are quite strong and can well approach the intrinsic strength of the parts to be welded.
4. The percussive welding process is very fast. Use in high speed automatic production is advantageous.

The problem treated in this paper arose during a very short study program at Bell Telephone Laboratories, Inc. in connection with a new

relay development.* The paper, therefore, does not purport to be a thorough study of all phenomena of interest. The purpose is rather to enumerate the major fundamental problems and to present first order solutions. Perhaps other groups having further interest in the process will carry on basic research along the lines indicated.

BASIC PROCESS

Basically the process consists of an electrical circuit which stores energy and maintains a voltage across the two parts to be welded. This is illustrated in Fig. 1 where a wire is to be welded to a small rectangular block. A mechanical appendage, the "gun," holds one of these parts and moves towards the other, stationary part. At a separation x_0, the arc initiation separation, the airgap breaks down and the arc is initiated. While the arc heats the opposing surfaces, forming a thin layer of molten metal on both parts, they are being brought closer together and finally come into contact, extinguishing the arc. The joint now cools and the weld is made.

A properly controlled process poses two design problems. First, it is necessary to select materials that are reasonably compatible and geometries which allow each part to reach the desired temperatures. The second design problem is the choice of the proper electric circuit. This paper is primarily concerned with the latter problem.

Material selection may be dictated by other considerations, but it is necessary to choose materials which are capable of producing a sound joint. Irregularities, such as gas pockets, possibly due to a low boiling point component are to be avoided. Geometries must be chosen such that in the presence of heat conduction away from the surfaces to be welded, the average temperatures of both surfaces exceed their melting points but stay below their boiling points.

A proper electric circuit for percussive welding has to supply sufficient energy to produce thin molten layers on both parts. It is undesirable to

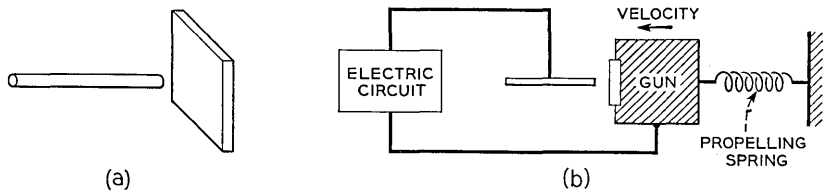


Fig. 1 (a) and (b) — Percussive welding. (a) Parts to be welded. (b) Process diagram.

* A. C. Keller, A New General Purpose Relay for Telephone Switching Systems. B.S.T.J., pp. 1023–1067. November, 1952.

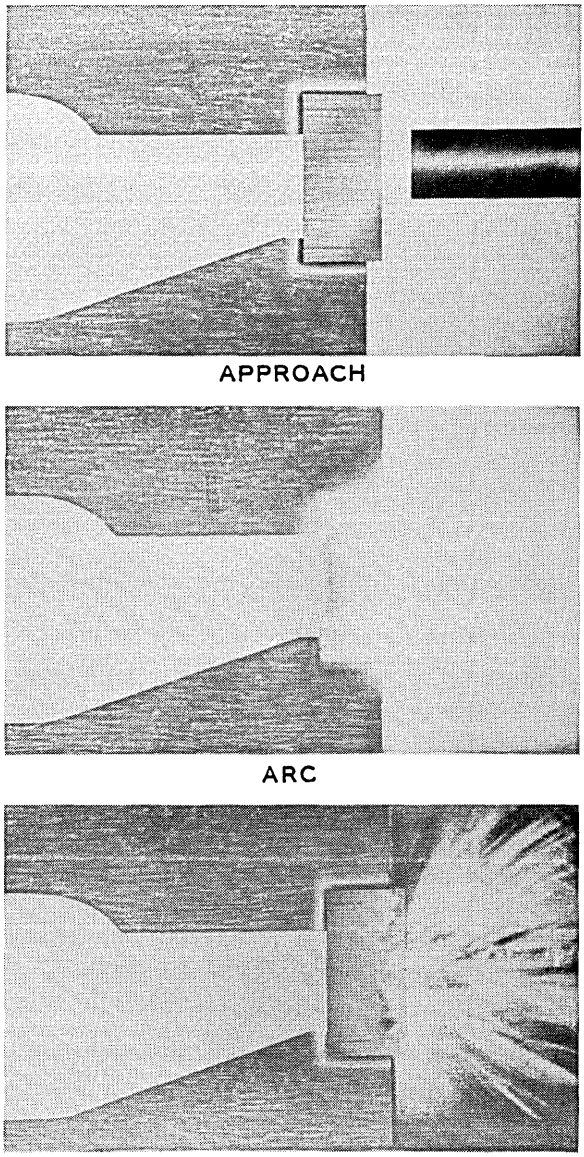


Fig. 1 (c) — High-speed photographs of welding operation.

supply excessive energy because of the large amounts of material that are then burned off, and objectionable weld flash is produced. Probably the major problem is to control the energy supplied by the circuit or, indirectly, the control of arc duration.

DURATION OF ARC

In this section the variables affecting arc duration will be discussed. During the arcing period the gun moves essentially at constant speed. The arc time may then be said to be equal to the distance traveled after arc initiation divided by the gun velocity.

A. Initial Separation

The voltage at which the arc is initiated is primarily a function of the separation between the two electrodes. A series of static voltage breakdown tests was made in order to define the distribution of initiation separation under conditions to be expected in production welding of a block to a wire. The block material was 70-30 per cent cupro-nickel. The wire was 0.040-inch diameter silicon copper. Tests were taken with the wire end flat or terminated in a 60° conical point while the block surface was maintained flat. Industrial contamination as may very well be present in a productiln machine was simulated by the addition of a thin oil film on each of the opposing surfaces.

The results are summarized in graph form in Fig. 2. Plotted are the three σ limits* for the conditions indicated. Better arc initiation separation control is obtained with flat as compared with pointed wire ends, and clean as compared with oil contaminated wire ends. The ratio of maximum to minimum arc initiation separations to be expected is considerably lower for high voltages than low voltages. This fact alone makes operation at voltages in excess of 1,000 volts desirable.

In addition there exists an initiatory time lag† between the time that the separation reaches the static breakdown value and the moment of actual initiation. Arc duration variations due to this phenomenon are reduced by an increase in applied voltage.

B. Evaporation of Material

The arc does not cover the whole surface but is concentrated on a small area which is being heated, therefore, at a rate considerably in

* All but three out of 1,000 welds are expected to fall within these limits. It is to be noted that in view of the high reliability often required of this type of weld, conditions even further removed than three σ limits may have to be considered.

† Field Emission of Electrons in Discharges by Llewellyn, Jones and E. T. de la Perrelle, Proc. Roy. Soc. A, **216**, p. 267, 1953.

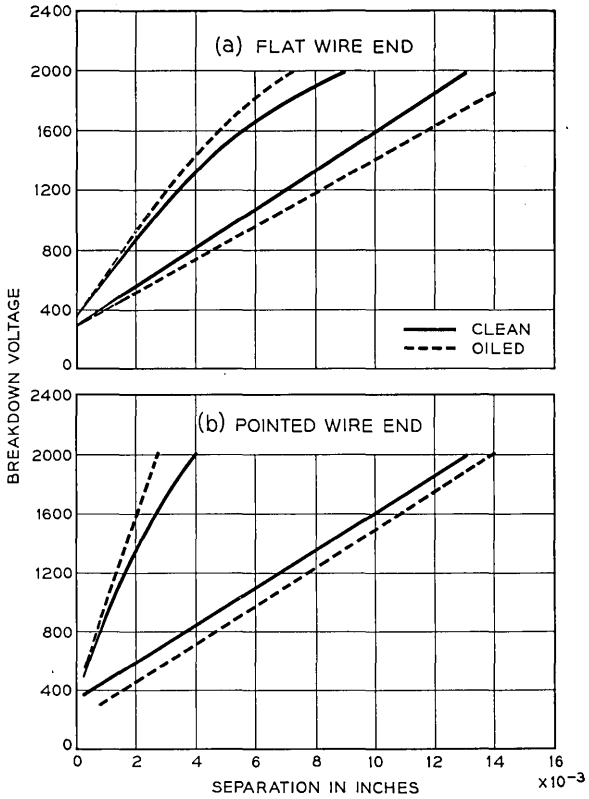


Fig. 2 — Three sigma limits of breakdown voltages between a plane surface and (a) a flat end and (b) pointed end of a round cross-section wire.

excess of that which one would compute as average heating. As a result very high temperatures occur in the arc region and material is evaporated. This somewhat increases the arc length and the arc moves on to a new spot. As a result of this mechanism some material is burned off in the arcing period and the arc duration is therefore longer than the initiation separation divided by the approach velocity. It is quite difficult to predict the amount of material evaporated. In lieu of more extensive experiments it can be stipulated that the evaporated material should be proportional to the energy input minus the energy directly radiated to the surroundings.

It is of interest, however, to note that the phenomenon of evaporation tends to increase the spread of arc duration as computed from the initiation separation alone. This is simply due to the fact that a large initiation

separation means a longer arc duration, providing a larger energy input and hence increased evaporation. This means that the moving part has to travel further before the arc is extinguished thus further increasing the arc duration.

C. Bridging

Early experiments indicated a variation in arc duration greater than that which could be explained on the basis of variation of initiation separation and burn-off alone. It was realized that there was a third phenomenon involved. This phenomenon in which metal filaments form and extinguish the arc prematurely will be called bridging. As a demonstration of bridging the two surfaces to be welded were spaced 0.002 inches apart and voltage applied between them. The resultant arc produced a molten filament between them and a weld was formed.* With the materials used the welds produced were somewhat porous and not too strong but tests were much too fragmentary to properly evaluate this process.

Electrostatic forces are too small to account for the bridging phenomenon. A speculative explanation on the basis of magnetic forces may be attempted. Let us first examine a uniform liquid filament carrying a current I. Application of the electromagnetic stress tensor demonstrates the presence of radially compressive pressures equal to:

$$P = \frac{\mu_0 I^2 r^2}{8\pi^2 a^4}, \tag{1}$$

where

I equals total current carried by filament.
r equals radial distance from center of filament.
μ_0 equals permeability of filament material.
a equals radius of filament.

In the presence of these compressive stresses the filament will tend to elongate. Consider now the two surfaces of the parts to be welded covered with a thin film of molten material. If due to the turbulence caused by the arc a small filament forms on one surface it may tend to elongate in the presence of magnetic forces and bridge the gap between the surfaces. A detailed dynamical analysis of the formation and stability of these metal bridges is quite difficult. The following treatment is a very rough

* This may actually be an alternate method of welding with the advantage of offering excellent dimensional control.

model which will show that consideration of the magnetic forces does yield an explanation verifying the order of magnitude of the bridging observed in experiments.

The model simulates the bridging phenomenon by the translational motion of a small cylindrical filament across the gap between the two surfaces. It is argued that the magnetic energy originally stored in the arc should be comparable to the kinetic energy of the moving filament.

The magnetic energy residing within a small cylindrical filament of diameter d, length ℓ, carrying a current I is:

$$\mathcal{E}_m = \frac{\mu_0 \ell I^2}{16\pi} . \tag{2}$$

If such a filament moves at constant velocity through a distance ℓ in time t its kinetic energy will be

$$\mathcal{E}_K = \tfrac{1}{2} M v^2 = \frac{\pi d^2 \ell^3 \rho}{8t^2} , \tag{3}$$

where ρ is the density of the filament material. If the two energies are comparable then the bridging time is roughly:

$$t \sim \frac{\sqrt{2}\pi d\ell}{I} \cdot \sqrt{\frac{\rho}{\mu_0}} \tag{4}$$

If we assume the following as reasonable numbers:

$$\begin{aligned} I &= 1{,}000 \text{ amperes}, \\ \ell &= 3 \text{ mils}, \\ d &= 20 \text{ mils}, \\ \rho &= 10 \text{ gm/cm}^3, \end{aligned}$$

then equation (4) gives a transition time of 15×10^{-6} seconds. This figure is of the same order of magnitude as bridging times observed during experiments.

The rough model used here is only one stage more refined than purely dimensional analysis but seems to give reasonable agreement with experiments. Of importance is the design guide offered by equation (4). By means of proper choice of the welding circuit it is possible to select an arbitrary current versus time relationship. In order to avoid bridging effects, which are, of course, very erratic, the bridging time t should be maximized. By equation (4) the ratio of current to separation should, therefore, be minimized. Stated in words this means that if large currents are necessary for the process (this will be shown to be desirable

in a later section) they should be confined to a period when the separation between the surfaces to be welded is quite large. The current should be sharply decreased as the surfaces approach each other.

DESIGN OF IDEAL WELDING CIRCUIT

In the previous section it has been shown that arc duration will vary over a wide range. This suggests that the system be designed in such a way as to be independent of arc duration.

The procedure will be to start with the desired temperature versus time relationship of the two opposing surfaces. From this the corresponding current versus time relationship can be found and finally the circuit giving such a current distribution selected. Obviously, the "safest" temperature-time relationship is one where the temperature is kept constant at the desired level T. The corresponding current distribution will be derived on the basis of one-dimensional heat flow. Let

u = temperature
T = desired temperature at surface
a^2 = diffusivity of material
x = distance in direction of heat flow
A = cross-sectional area over which heat flow occurs
K = heat conductivity of material
V_m = voltage across arc
i = transient current

Start with the differential equation for one-dimensional heat flow:

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}. \tag{5}$$

For the boundary conditions:

$$\begin{aligned} u &= 0 && t < 0 \\ u\,|_{x=0} &= T && t > 0 \end{aligned} \tag{6}$$

The solution is:

$$u = T\left(1 - \frac{2}{\sqrt{\pi}} \int_0^{x/2a\sqrt{t}} e^{-\beta^2}\, d\beta\right). \tag{7}$$

In order to determine the heat input required we must find the gradient at the surface. Expanding equation (7):

$$\frac{\partial u}{\partial x} = \frac{2T}{\sqrt{\pi}}\left[\frac{1}{2a\sqrt{t}} - \frac{x^2}{(2a\sqrt{t})^3} + \frac{x^4}{2!(2a\sqrt{t})^5} \cdots\right], \tag{8}$$

$$\left.\frac{\partial u}{\partial x}\right|_{x=0} = \frac{T}{a\sqrt{\pi t}}. \tag{9}$$

The required heat input must now be set equal to that supplied by the welding circuit:*

$$\tfrac{1}{2} i V_m = \frac{KAT}{a\sqrt{\pi t}}, \tag{10}$$

or

$$i = \frac{2KA}{aV_m\sqrt{\pi}} \frac{T}{\sqrt{t}}. \tag{11}$$

The current time relationship represented by equation (11) is that due to a capacitative transmission line working into a short circuit. Since the arc voltage is considerably lower than the voltage to which the line is charged, it is substantially a short circuit.

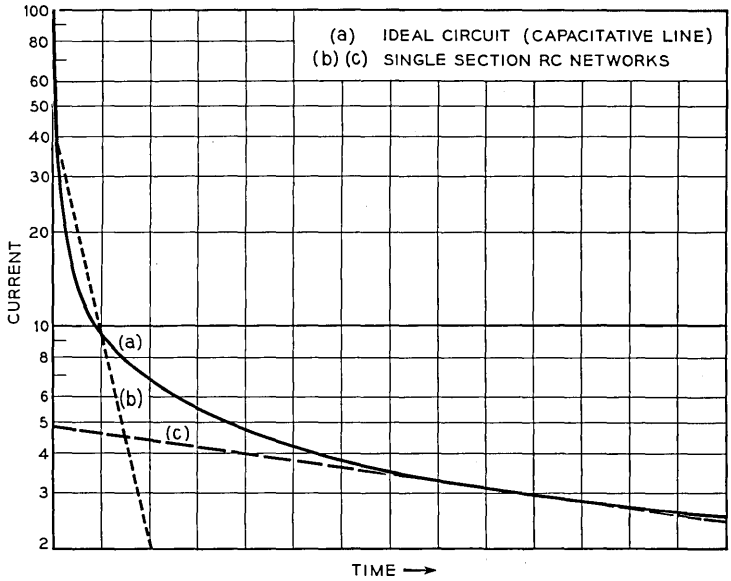


Fig. 3 — Current time characteristic. (a) Ideal circuit (capacitative line). (b) and (c) single section RC networks.

* It will be assumed that the energy of the arc is divided equally by the two opposing surfaces.

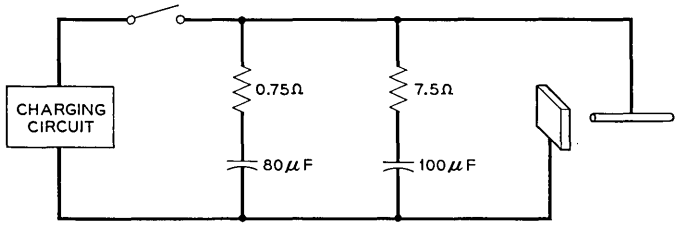


Fig. 4 — Practical welding circuit.

SELECTION OF A PRACTICAL CIRCUIT

It is probably not practical to use a distributed constant capacitative line having a current time relationship as plotted in Fig. 3. The line can, however, be approximated to the desired degree of accuracy by means of a series of r-c sections. The current discharge curve for a single r-c section plotted on the semilog graph of Fig. 2 will be a straight line. Clearly this is not a very good approximation of the ideal curve. If the constants are adjusted such that the desired initial high currents are met, the current will decay too fast allowing the surface to cool prematurely for long arc durations. If the constants are adjusted to match the desired curve for long arc times, the initial heating will be insufficient and short arc durations will produce poor welds.

A fairly good approximation of Fig. 2 can be obtained by as little as two r-c sections in parallel (Fig. 4).* Use of this circuit has resulted in considerable improvement not only in the uniformity of the welds obtained but curiously enough in the control of arc duration. The reason for the latter phenomenon is that with the multiple section circuit the desired bridging characteristics can be met much more closely.

THE MECHANICAL STRUCTURE

Relatively little is known about the effect of the mechanical design of the welding apparatus on the process. Basically, the mechanical constants of interest are the mass and velocity of the gun when the arc is being extinguished and the forces propelling the gun. The gun contains kinetic energy part of which is absorbed during the impact of the two parts to be welded. The remaining part will tend to produce rebounding of the gun. Clearly the weld must have cooled sufficiently when the gun draws back such that it can withstand the forces tending to pull it apart. The time allowed for cooling is then roughly one-half the period deter-

* The circuit configuration shown is equivalent to two L sections of a lumped constant line as usually depicted.

mined by the mass of the gun and the stiffness of the stationary part to be welded.

The effects on weld quality of that hammer blow produced by the gun are not well understood but there is some evidence that this blow may be advantageous in producing intimate mixing.

SUMMARY

The basic processes of percussive welding have been discussed. Large variations in arc duration are caused by the spread in the initiation separation, bridging phenomena, and the amplifying effect of evaporation.

The relative spread of initiation separation is minimized by working at high voltages, in excess of 1,000 volts. Bridging, which causes premature extinguishing of the arc, is minimized by maintaining the ratio of current to separation at a minimum.

A welding circuit offering independence from arc duration variations has been developed on the basis of one-dimensional heat flow. The analysis presented suggests a capacitative transmission line, which can however be approximated by two or more r-c sections. Greatly improved process control has been effected with this circuit.

ACKNOWLEDGMENT

The author gratefully acknowledges the assistance of J. J. Madden in all phases of experimentation connected with this project. Miss L. Mitchell performed the study of breakdown voltage. S. P. Morgan suggested the model of bridging time.

Automatic Contact Welding in Wire Spring Relay Manufacture

By A. L. QUINLAN

(Manuscript received January 19, 1954)

Welding of precious metal contacts to the new wire spring relay has presented some unusual manufacturing problems. As the name implies, the springs of these relays are wires. The contacts through which electrical circuits are established consist of small blocks of palladium accurately and securely welded to one end of these wires. The wires, arranged in a parallel array, are imbedded for part of their length in molded phenol plastic to form parts which will be designated as "combs" in this paper. There are two kinds of combs, those with the wires arranged in pairs called twin wire combs, and single wire combs. Different welding techniques are required, each of which will be described separately.

INTRODUCTION

The wire spring relay, Fig. 1, was designed with such advantages over U and Y type relays as higher operating speed, longer life, lower power consumption and lower cost, as described in a recent article in this Journal.* The lower cost will be achieved largely by reduction of assembly labor time, by reduction of adjustment effort after assembly due to greater precision in the manufacture of component parts and by extensive use of automatic manufacturing processes. To attain these goals the closest cooperation has been necessary, particularly during the design stage, between Bell Telephone Laboratories relay engineers and Western Electric development engineers. Small lots of wire spring relays of several of the early designs were manufactured by the Western to furnish the Laboratories with relays for testing. These operations provided Western development engineers with valuable manufacturing experience.

The present design of relay has been in production on a pilot plant basis. The contact welders have operated as individual units during

* A. C. Keller, A New General Purpose Relay for Telephone Switching Systems, B.S.T.J., Nov., 1952.

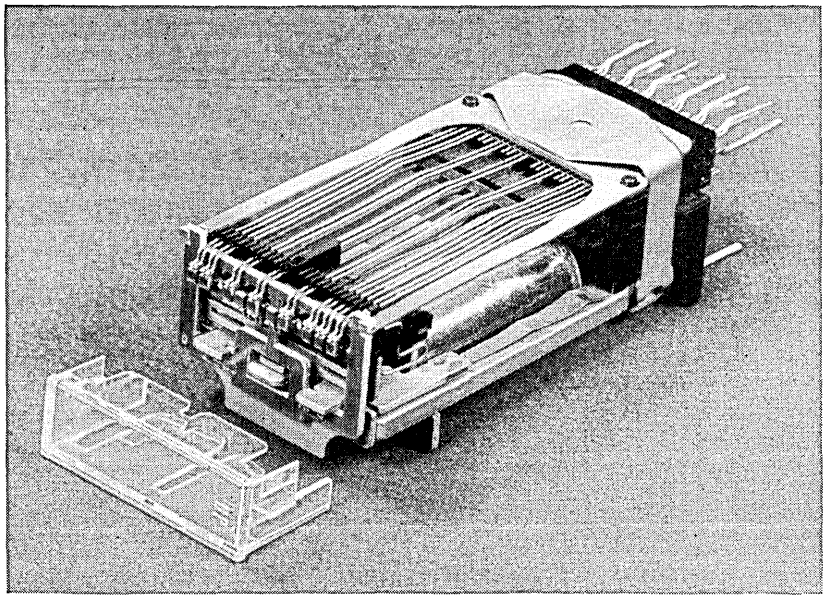


Fig. 1 — Wire spring relay.

that period as contrasted to their inclusion as components in automatic welding and wire forming lines now being placed in operation. The performance reported herein is based on individual operation.

Part I — Automatic Multiple Resistance Welding of Twin Wire Combs

The problem to be solved by Western Electric development engineers was that of welding small blocks of palladium, a precious metal, to flattened ends of 0.0226-inch diameter nickel silver wires molded in twin wire combs, Fig. 2. In addition to a secure weld, close limits had to be

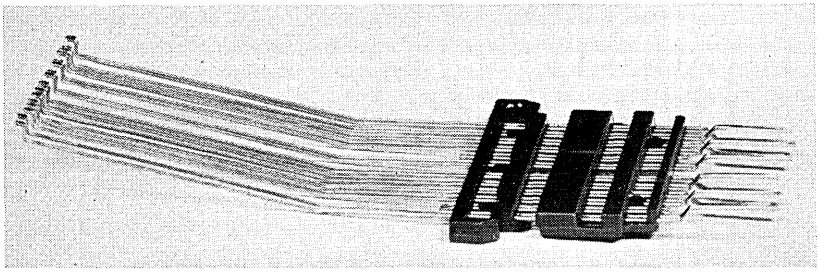


Fig. 2 — Twin wire comb.

met on the location of the precious metal. For example, Fig. 5, the upper contact surfaces had to be located with a precision of 0.004 inch. The contact itself had to be centered laterally with respect to the straight portion of the wire within 0.004 inch. Variables in the wire materials, such as elasticity, make such limits difficult to meet. The wires in the twin combs are oriented in pairs to provide bifurcated contacts, i.e., the contacts on both wires of any pair mate with one stationary contact of the single wire comb to furnish two current paths. Any number of pairs up to a maximum of twelve may be required by a code of relays. To conserve precious metal, contacts are welded only to those wires needed for a particular comb. The wires not required are clipped from the comb by a hydraulic press operated die in the automatic welding and wire forming line.

The 24-circuit capacitor discharge resistance welders, one of which is shown in Fig. 3, were designed and constructed for welding twin wire contacts. As stated before, this welder is one of the units in a welding and wire forming line. Combs from the molding operation are delivered to the beginning of the line in magazines. By fully automatic means they are removed from the magazine, carried by a reciprocating conveyor through each machine unit in the line and, when fully processed, placed in another magazine. The line of machine units, Fig. 4, forms the contact end of the wires, degreases the wires, welds the contacts to the wires, coins the contacts to the specified dimensions, forms the terminals, clips the ends to length, tension bends the wires, removes unnecessary wires and tins the terminal ends.

REASONS FOR SELECTION OF PROJECTION TYPE RESISTANCE WELDING

A capacitor discharge resistance projection welder was chosen for this job for the following reasons:

1. It is capable of welding automatically in a line of other machines.
2. It can provide the fast rate of temperature build-up required to prevent excessive heating of the small nickel silver wire ends.
3. A capacitor, charged to a fixed voltage, offers a good means of controlling weld energy within narrow limits.
4. Resistance projection welding can concentrate heat at the point required. The use of a projection at the welding interface lowers the electrical current needed to a value which can be handled by electrodes without excessive heating. By placing a projection on the metal with the lower electrical resistance, in this instance the palladium, the temperature of the palladium can be increased in the weld zone as compared to that of the nickel silver wire, thus securing a better heat balance at the joint

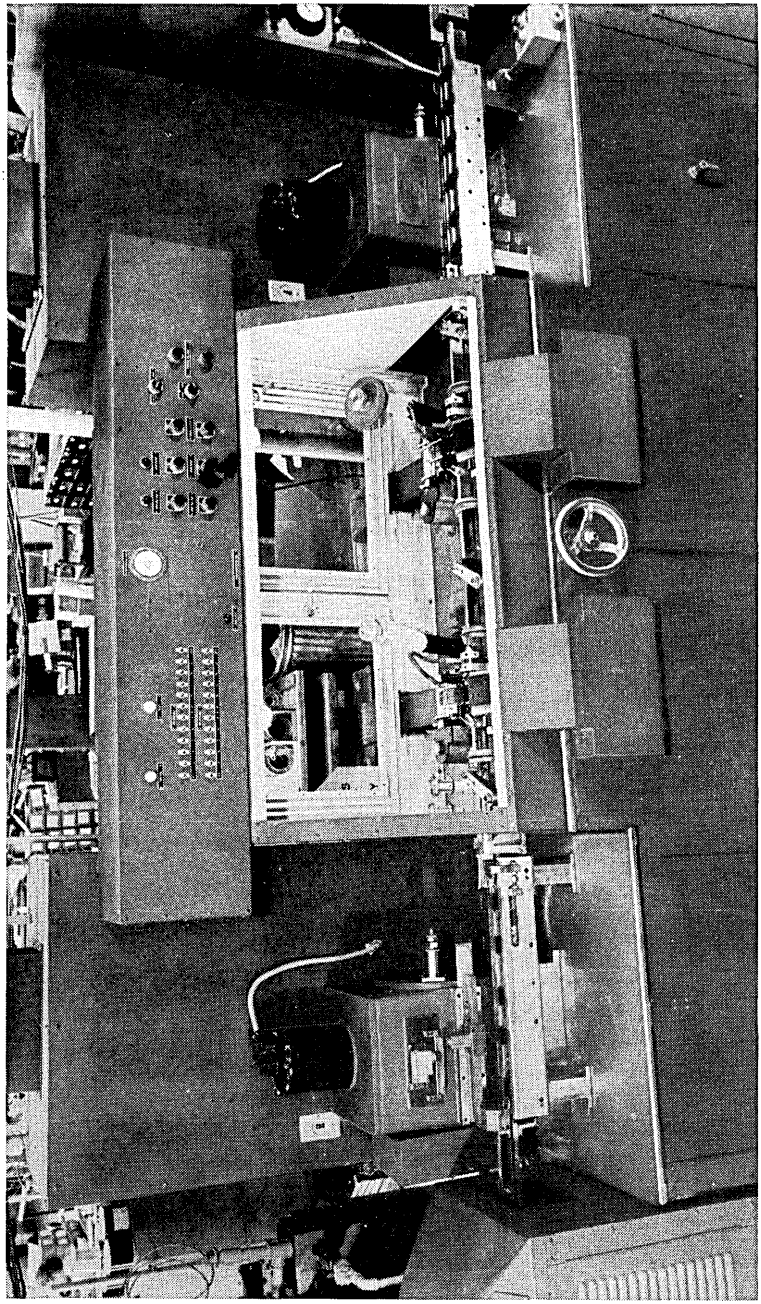


Fig. 3 — General purpose apparatus twin comb welders.

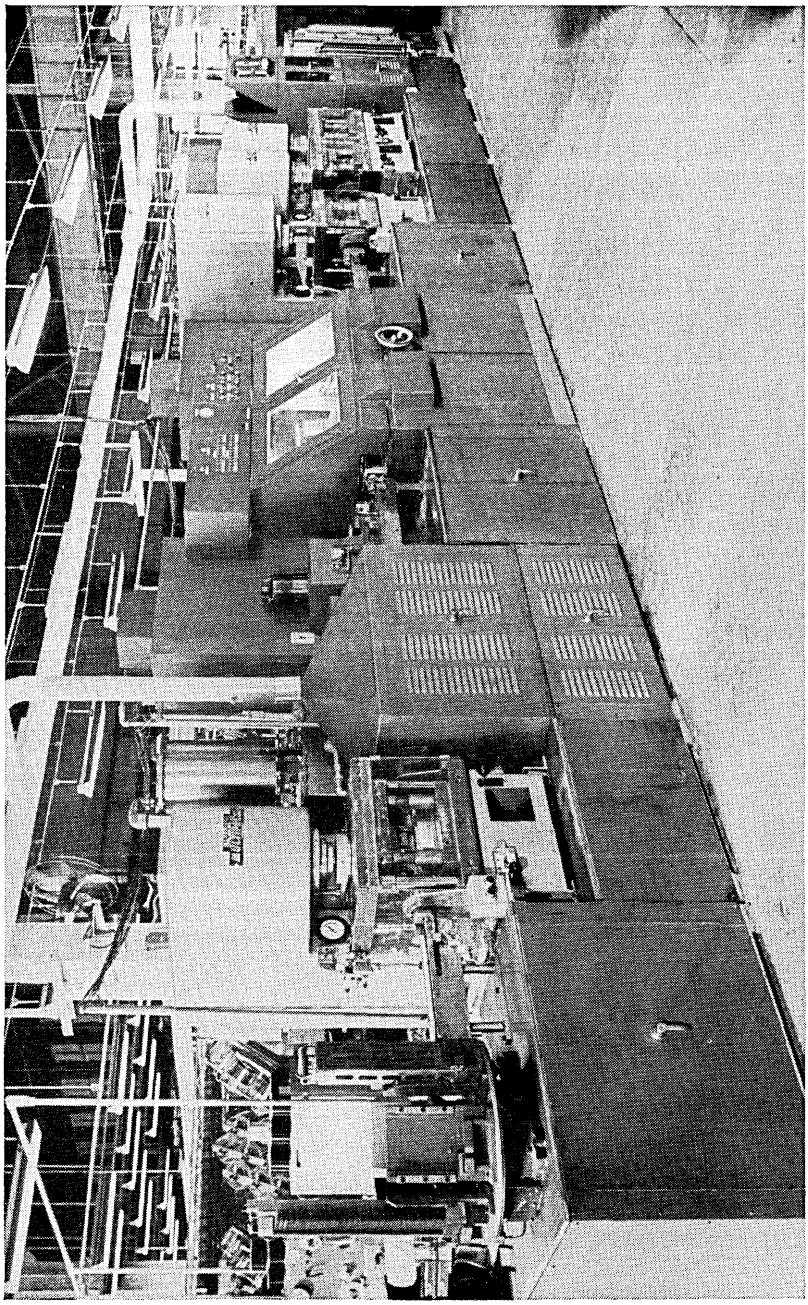


Fig. 4 — Line of machine units.

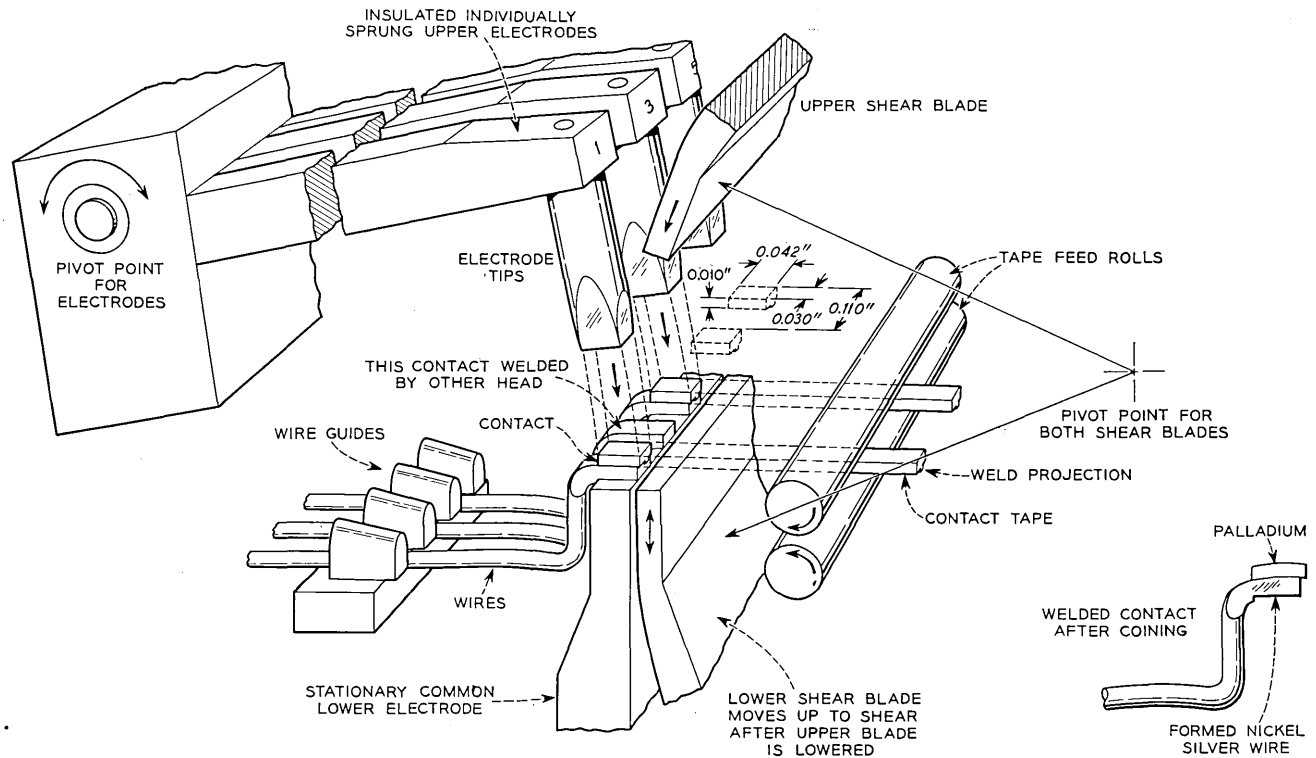


Fig. 5 — Sketch of a welding head.

The projection also confines the location of the weld to a central area of the wire end. When the weld nugget is confined to a central area, cold surrounding metal helps prevent flashing out of hot metal and results in stronger welds. To further centralize the weld nugget and to limit the projection area in contact with the nickel silver wire, the wire surface is preformed to a large radius at right angles to the weld bead.

WELDER HEAD OPERATION

Welding is done in two stages. Head No. 1 welds the odd numbered wires of each pair, after which the comb is advanced and head No. 2 welds the even number wires. This is necessary because of the small distance between the wire centers. The electrodes in each head are spaced to match the interval between odd or even numbered wire centers. Fig. 5 shows an isometric sketch of a welding head. The sequence of operations for either head is as follows:

1. The twin wire comb is advanced to a locating nest and lowered into position with the reference holes in the plastic engaged on pilot pins.
2. As the comb is lowered, the contact wires enter guide slots of a rake at a point adjacent to the plastic.
3. This rake is moved toward the ends of the wires, thereby spacing and positioning them as shown in Fig. 5. A plastic spacing member is incorporated in the relay for aligning the contacts in the relay assembly.
4. The palladium contact metal, in tape form, is advanced the proper distance to provide a contact of the specified length. These tapes, parallel with the wires, extend from guide slots for a distance slightly greater than a contact length and project over the wire ends. The tape guide slots, incorporated in a split holder, consist of steel inserts embedded in phenol fibre so designed that when the upper section is shifted laterally with respect to the lower section, the steel inserts clamp or release the tapes. Since the rake is located with reference to the tape guide, accurate spacing of contacts in the relay assembly is secured regardless of minor deviations of position of the wire ends at the welding machine. The phenol fibre mounting electrically insulates the tapes from one another to prevent part of the weld current from passing through adjacent tapes.
5. The upper electrodes, on the end of cantilever arms, are brought down on the palladium-nickel silver wire assembly.
6. The capacitors are discharged and the welds are made.
7. The upper electrodes are pivoted upward out of the way.
8. The wire guide rake is returned to a position near the plastic.
9. The upper shear blade is lowered onto the contact ends clamping them against the lower electrode.

10. The lower shear blade is moved upward, shearing off the tapes.
11. The upper shear blade is raised and the welded comb is transferred either to the second similar welder head or to the next operation.

The upper electrode tips are pins of special electrode material fixed in the ends of the supporting cantilever spring members by tapered joints, Fig. 5. Each electrode member is insulated electrically from the others and from ground by a coating of Teflon and a slotted phenol fibre guide block (not shown in Fig. 5) at the end adjacent to the electrode tips. Teflon furnishes the necessary insulation in a minimum of space and is slippery enough to allow free movement of any member. The cantilever construction makes a very light electrode assembly with the quick follow through necessary to maintain pressure on the weld area as the projection on the palladium tape is melted during the short weld cycle. A deflection of one-sixteenth inch at the tips will produce a force of about ten pounds, which is ample for this welding job. The working surfaces of the electrode tips are dressed approximately every 20,000 parts by an abrasive wheel mounted on an arbor and rotated back and forth by hand with the electrode tips pressed against the flat side of the wheel. In this manner all electrodes are dressed to a uniform length and angle simultaneously. After repeated dressings have reduced the tip length by approximately one-eighth inch, new tips are inserted.

MECHANICS OF THE WELDER

The welder is driven by an electric motor connected to the main cam shaft through a suitable reduction gear. This cam shaft is located just under the table top and extends the length of the machine. A mechanical overload clutch is provided on the output shaft of the reduction gear. All cams for the head movements are located on the main shaft. A solenoid operated clutch stops the main cam shaft when necessary. When no comb is in the weld position the weld control circuit is held open and solenoids shift either of the tape feed cams axially out of line with their followers to prevent the feeding of tape. This prevents burning the electrodes, saves precious metal and avoids loose contact pieces which might interfere with succeeding welds. A reciprocating conveyor, extending through the machine about four inches above the table top, carries the combs through the two weld positions. Excessive load, as when a part jams, will trip a microswitch and stop the machine.

TAPE SUPPLY

The contact tapes are advanced the amount required to provide the exact contact lengths by means of opposing rubber-faced feed rolls in

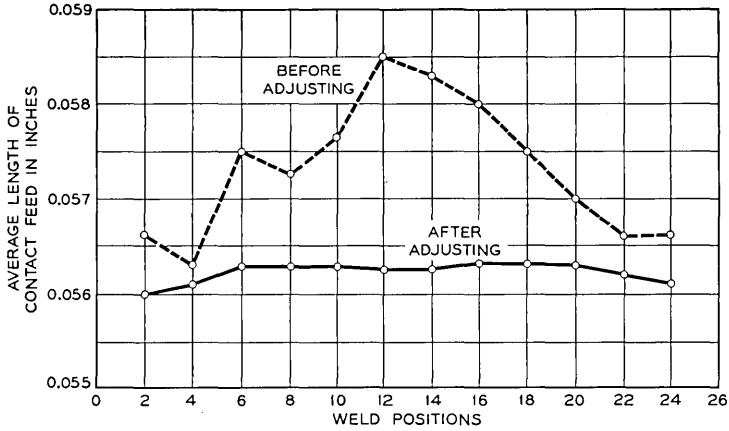


Fig. 6 — Contact length before and after adjusting hardness of rubber feed rolls and omitting gears between them.

each head. Fig. 6 shows the results of an early test on feed accuracy. Improved accuracy was obtained by providing free running tape reels. All portions of the tapes and reels are insulated from each other to prevent loss of weld energy.

ELECTRICAL FUNCTIONS

Electrically the welder has 24 separate weld circuits. Each circuit, shown schematically in Fig. 7, includes a capacitor which is charged through its own thyratron to a predetermined high voltage maintained by a voltage regulating transformer. The capacitor discharges through a thyratron and a transformer to produce the customary low voltage-high current weld energy. The transformers are located under the table near the welding heads.

The electronic equipment is mounted in cabinets to the right and left of the welder proper, as shown in Fig. 3. The capacitors are located in low cabinets, on either side, together with their solenoid controlled safety short-circuiting system. The cabinet above the welding heads contains control switches and relays. Toggle switches are provided for cutting off the weld energy for each of the 24 circuits.

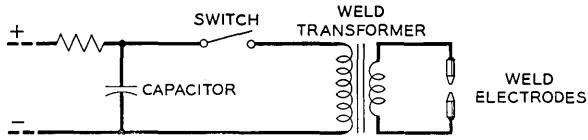


Fig. 7 — Schematic of weld circuit.

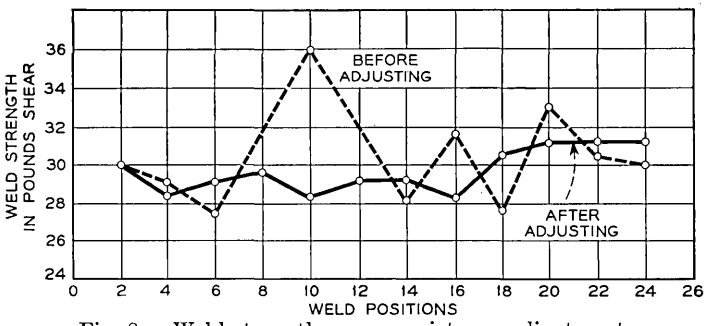


Fig. 8 — Weld strength versus resistance adjustment.

The welds occur in rapid succession in each head thereby preventing electrical interference between the many circuits involved. When combs which require less than a full complement of contacts are to be welded, the tapes in the not wanted positions are removed from the feed rolls and the toggle switches on the control cabinet are set to direct the weld energy in the sequence desired. To aid in obtaining uniform weld strength from the 24 circuits, all secondary leads from the transformers to the upper welding electrodes are of the same length. To further aid in effecting uniform weld strength, rheostats are provided in the high voltage side of the discharge circuit. Fig. 8 shows the result of adjusting these rheostats to balance the weld strengths produced by all circuits. A longer pulse would give the heat produced more time to be conducted away from the weld zone and would tend to heat more of the wire end, to the point, perhaps, of melting it completely. The three millisecond welding time has proven to be satisfactory for this job. However, the shorter the duration of the weld the higher will be the current peak required to obtain the necessary heat; the higher the current peak, the greater will be the likelihood of burning the electrodes and shortening electrode life. Adjustment of the pressure on the electrodes can be employed as a compensating factor. As the pressure is increased the tendency to burn the contacting surfaces is decreased, but the weld projection on the palladium is flattened correspondingly and the temperature of the weld, for a given current, is lowered accordingly. These variables can be adjusted to maintain optimum balance between uniformity of weld strength and electrode life.

CONCLUSIONS

More than ten million welds from this automatic machine have offered convincing proof of its capabilities. Fig. 9 indicates that good weld

strengths are being obtained. These values are shear strengths obtained on a dial indicator type of gage, reading directly in pounds, when the contact is sheared from the wire along the wire axis. A minimum test requirement of ten pounds has been established.

Part II — Automatic Percussion Welding of Single Wire Combs

Percussion welding is not new. Early work goes back beyond the beginning of the century, but little application has been made of it and only a meager amount of literature is available. However, this method has a real field of usefulness as the application described in this paper will show. The original Vang process, wherein a capacitor charged to a high potential, often several thousand volts, is discharged across the gap between parts as they approach each other under a propelling force, is a good general description of the method used. The arc so produced heats the abutting surfaces before they collide so that a very thin layer of metal is brought to welding temperature. The propelling force, continuing to act, brings the parts together percussively and the weld is made. Little metal is heated and little heat penetrates the adjoining metal; therefore, the heat balance problem is greatly minimized and different metals weld together with little trouble. There is, however, the problem of protecting personnel from high voltage. Also, the two surfaces being welded must be insulated electrically from each other. This excludes the use of this process for joining the ends of the same piece of metal as in making a ring.

The project undertaken by Western Electric development engineers in the case of the single wire combs was the development of a machine for

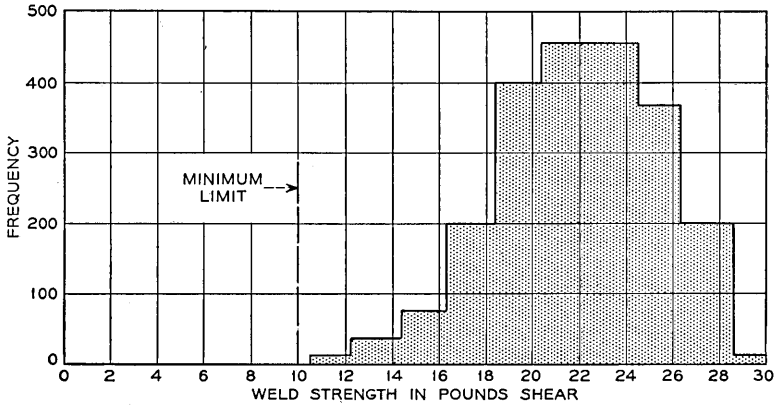


Fig. 9 — Weld strength distribution.

the automatic multiple percussion welding of contact blocks to the ends of an array of small wires extending less than a quarter of an inch from molded phenolic plastic. This array, fixed in plastic, forms a comb which is illustrated in Fig. 10. When completed this comb becomes the stationary contact member of the wire spring relay. The small blocks of metal on the ends of the wires are cut from a composite tape of which a small portion near the top and/or the bottom surface is palladium. There is a family of combs to be welded, depending on the number and type of contacts required by the code of relays into which each comb is assembled. Unlike the twin wire combs, wires which do not require contacts are left in the single wire combs, primarily to facilitate reading terminal locations during wiring into equipment. All top palladium contact surfaces must be located in the same plane across the 12 wire positions of the comb within a tolerance of ± 0.002 inch to meet design requirements. In addition, other dimensions for locating the precious metal must be held to close limits for reasons of precious metal economy.

The contact blocks for the wire comb are welded to the wire ends by the automatic percussion welder, Fig. 11, which is a unit in an automatic welding and forming line, Fig. 12, similar to but not identical with the

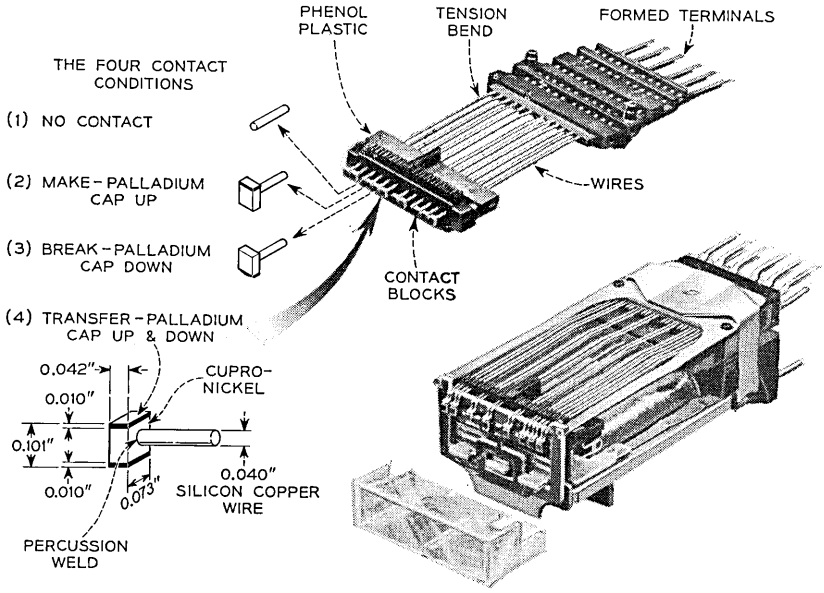


Fig. 10 — Single-wire comb with percussion welded contacts. Also view of wire spring relay.

line used for twin wire comb manufacture. These lines, by being fully automatic in operation, are interesting examples of automation.

REASONS FOR SELECTION OF PERCUSSION WELDING

Percussion welding was selected for this process for the following reasons:

1. The electrodes may be placed well away from the weld zone. The lesser current required for arc welding as compared to resistance welding makes it possible to conduct this current through the wires without heating them appreciably. The electrodes must be placed away from the wire ends so the clamping force will not deflect the wires from their normal position, thus causing a misalignment of contact surfaces.

2. A suitable heat balance in the weld zone can be obtained readily. If the slower butt welding method were used this would be more difficult because of the unequal size and differing electrical and heat conductivities of the abutting parts.

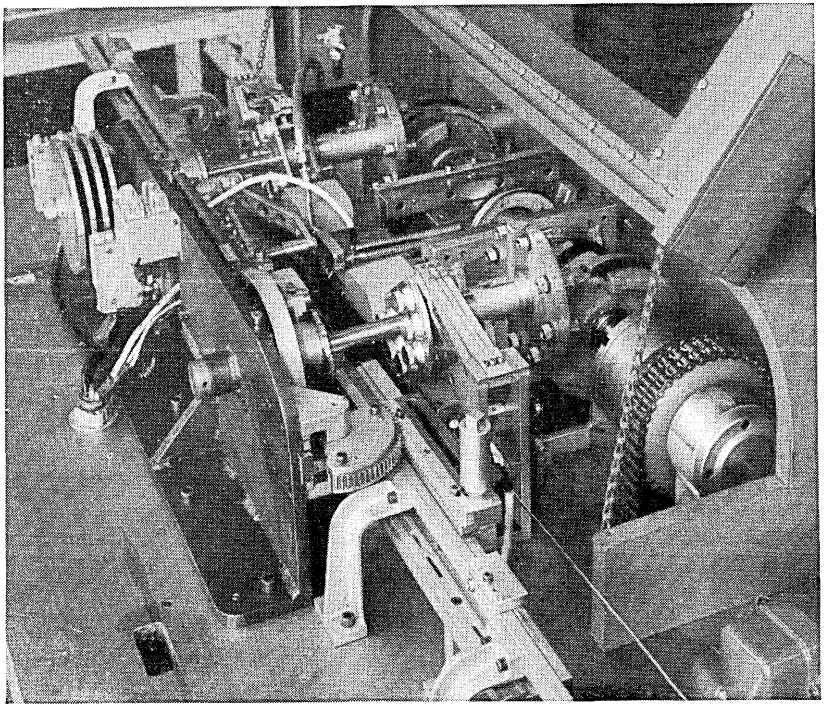


Fig. 11 — Percussion welder for contacts on single-wire combs.

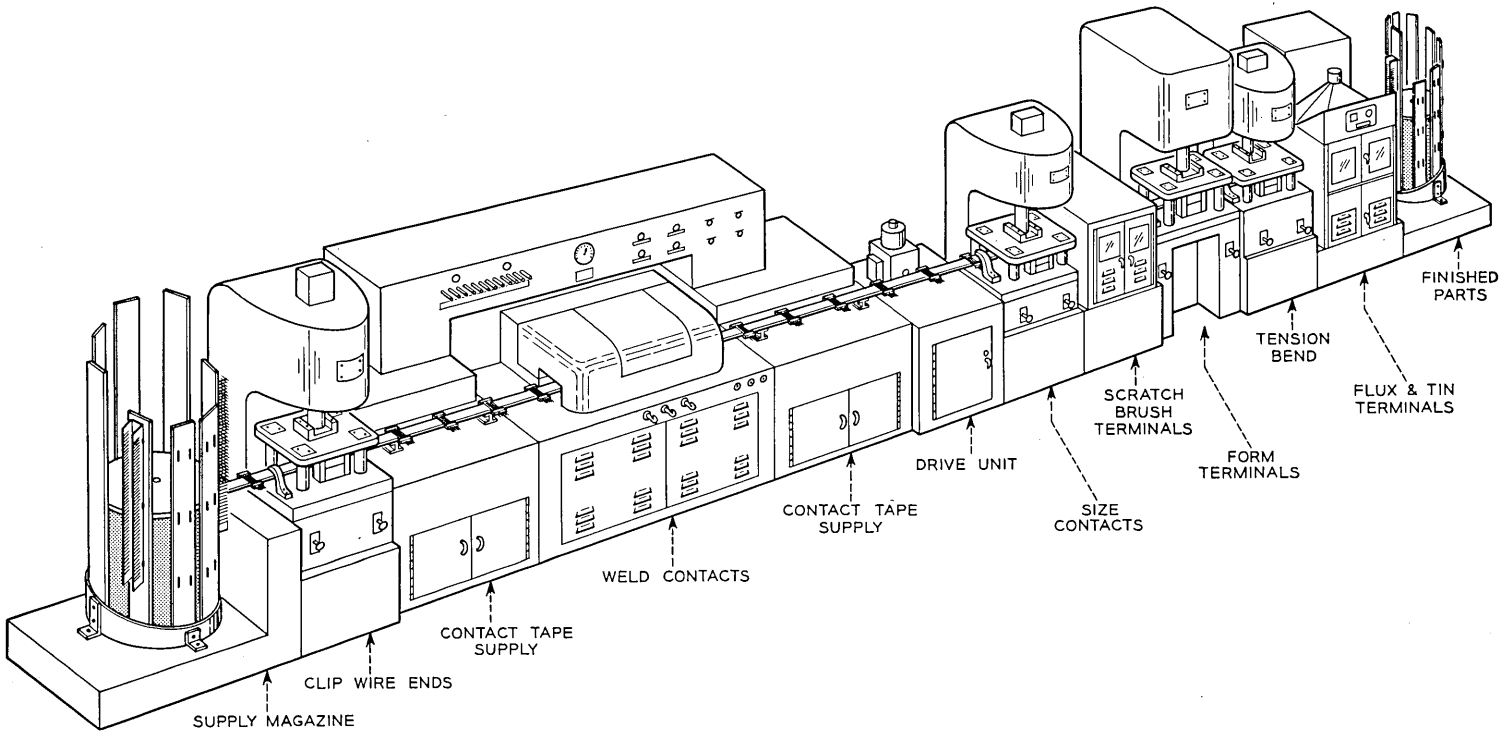


Fig. 12 — Single-wire comb line.

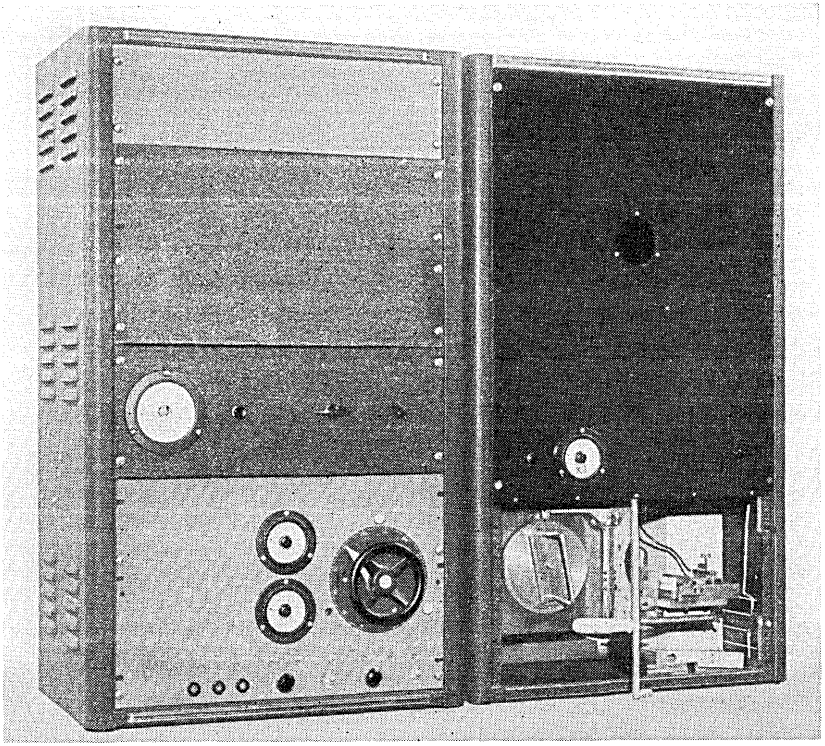
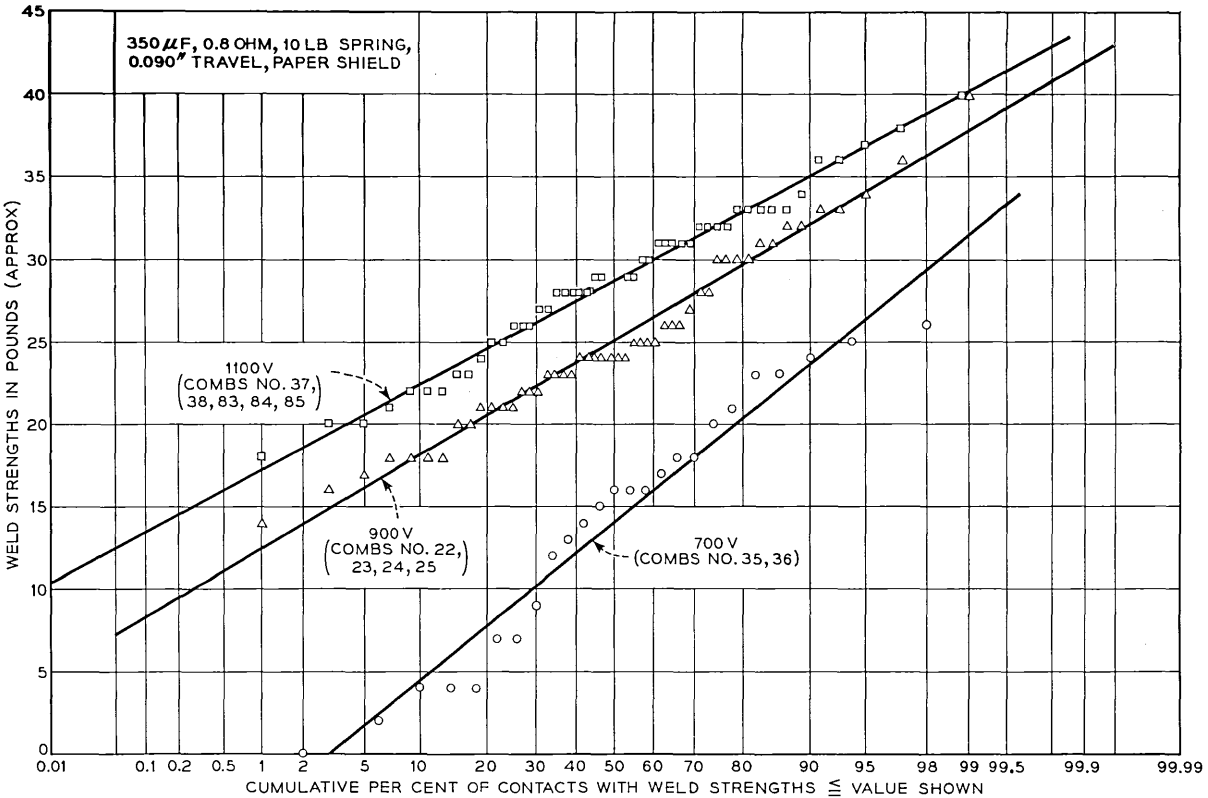


Fig. 13 — Experimental percussive welder for welding contacts to a molded comb on a "one at a time" basis.

3. The fast welding time recommends its use in high speed automatic welding machines.

EXPERIMENTAL WORK

The earliest experimental work was performed on a simple welding fixture with a spring loaded sliding jaw for retaining a contact sized piece of metal and a clamp for holding the wire. Some welds were produced but they varied widely in strength. The next fixture built, Fig. 13, was designed to weld contacts to a molded comb on a one at a time basis. A lightweight spring jawed slider held the contact and guided its travel along a fixed path. This slider was propelled by a lever actuated by a spring and controlled by a cam. The cam was powered by a variable speed drive. An extended study failed to show good weld results except at high speeds when the lever was unable to follow the cam surface and moved



Figs. 14 — Effect of voltage variation on weld strength. (Typical establishing data.)

freely, controlled only by its mass and the applied spring force. From this study the presently used spring actuated welding gun has evolved.

To determine the operating conditions that would be required, an investigation was made in which a group of 48 contacts was welded under carefully controlled conditions following which one condition was varied and the tests repeated. Many curves were established in this way for such variable factors as voltage, capacitance, resistance, spring force, and distance traveled; all related to weld strength. Weld strength was measured by a hook-pull gage developed at Bell Telephone Laboratories. The "burn-off" i.e., reduction in length of the wire and contact as a result of the arc heat, was determined in many of these tests because it usually provided a measure of weld strength uniformity. Many other conditions were investigated such as the weldability of different metals; the effect of cleanliness and of contamination on the joining metals; the influence of the physical form of the joining metals, such as pointed, rounded, or flat surfaces on wire ends; the effect of atmospheric conditions, of the presence of various gases, or a stream of compressed air, and of shields in or near the weld zone. The nature of the weld flash deposit was studied. Neither streamers of base metal which might extend over the edges of the palladium caps nor loosely adhering and easily dislodged metallic particles could be tolerated. Charted data from some of these tests are shown in Figs. 14, 15, and 16.

Another phase of the investigation was concerned with obtaining such

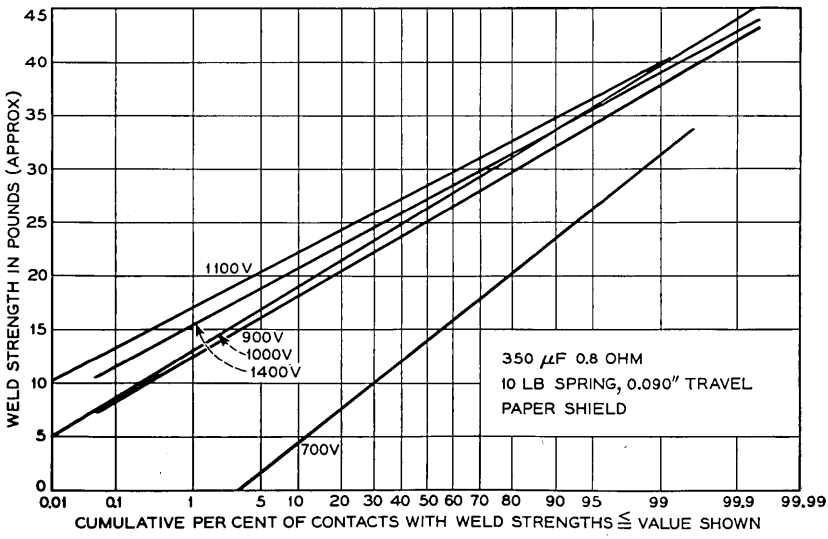


Fig. 15 — Effect of voltage variation on weld strength.

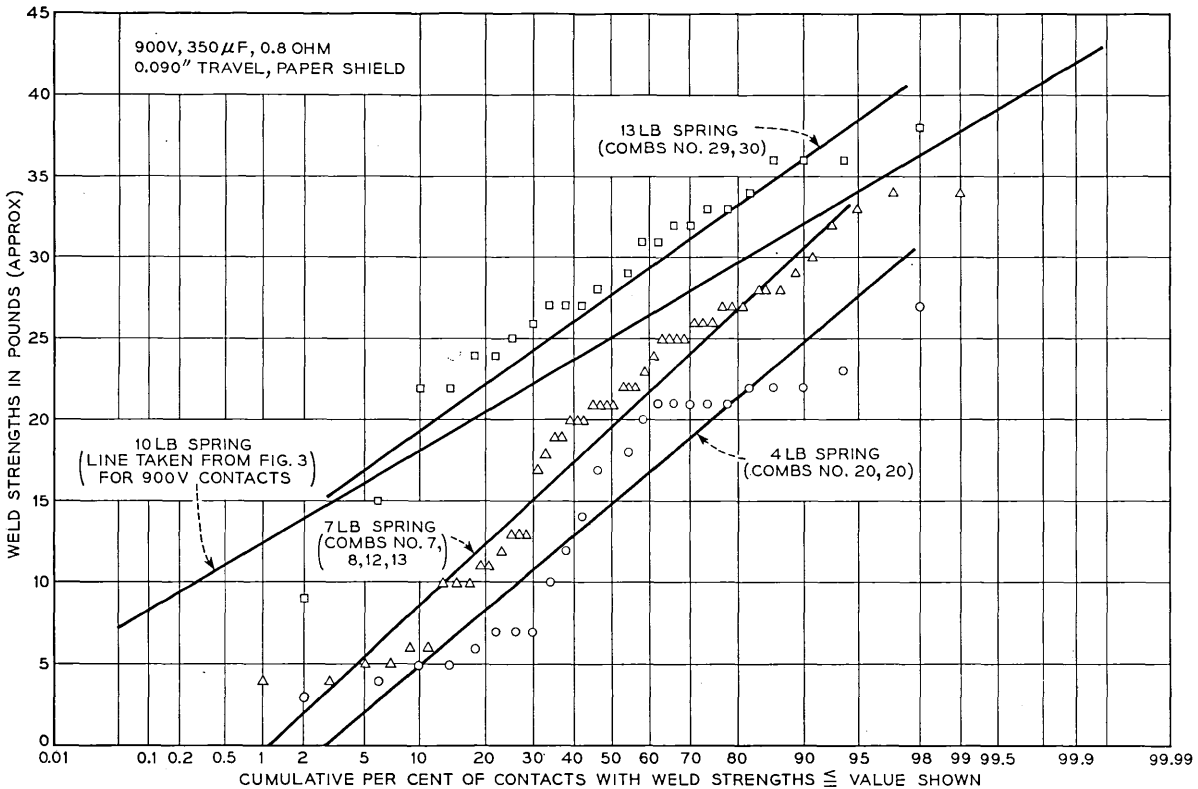


Fig. 16 — Effect of spring pressure on weld strength.

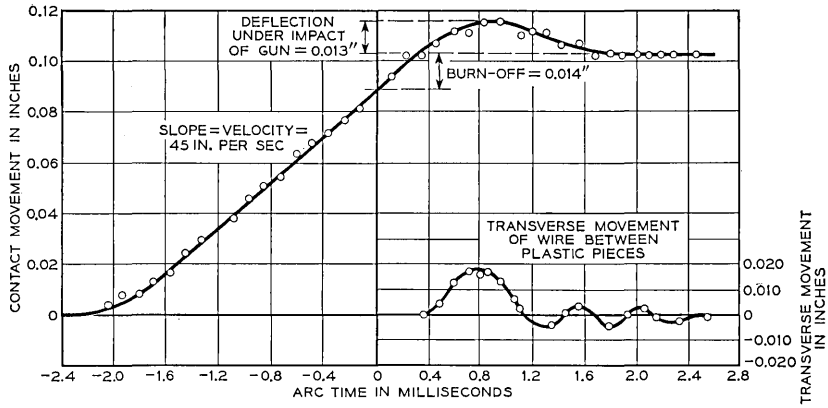


Fig. 17 — Motion of contact during percussion welding plotted from measurements on film.

evidence of what occurred during the welding operation as could be revealed by high speed motion pictures. Fig. 17 shows the kind of information obtained from these films; notably the speed of approach, the amount of burn-off, and the deflection of the comb wires back of the front plastic block upon impact by the welding gun.

In connection with a study directed toward a better understanding of the mechanism of percussion welding undertaken by E. E. Sumner of Bell Telephone Laboratories tests were made with an electro-capacitive transducer, oscilloscope, and polaroid camera arrangement for the purpose of recording variations in the velocity of the welding gun. The characteristics of the current discharge during welding was recorded by another oscilloscope-camera setup. Burn-off was measured and broken weld surfaces on the contacts were examined and photographed. These tests pointed to the value of a parallel capacitor discharge circuit as a solution to the arc duration variability problem. A commercial trial of parallel capacitor circuits at Western Electric demonstrated the merit of this type of circuit, which will be described in more detail later. Calculations from the transducer traces indicated that the striking force of the contact and contact holder upon impact with the comb wire is less than seventy-five pounds. The welding gun was operated on a reed mounting during Mr. Sumner's study.

Early experimental work on percussion welding showed that small gas pockets in the weld zone were causing weak welds. Nickel silver was used at that time. Its component of easily volatilized zinc was suspected of causing the trouble. Then various metals and alloys were tested and silicon-copper was chosen for the wires in the comb. This alloy,

however, was not suited for use as the base metal of the contact block because it was difficult to weld in the roll welding process used for fabrication of the contact tape. A 70-30 per cent cupro-nickel was selected for this base metal because it approximated the resistance of the palladium, a necessary condition for roll welding, and it did not have an easily volatile component to weaken the percussion welds.

Another subject investigated was the speed of approach of the parts during the percussion welding operation. The comb and its array of

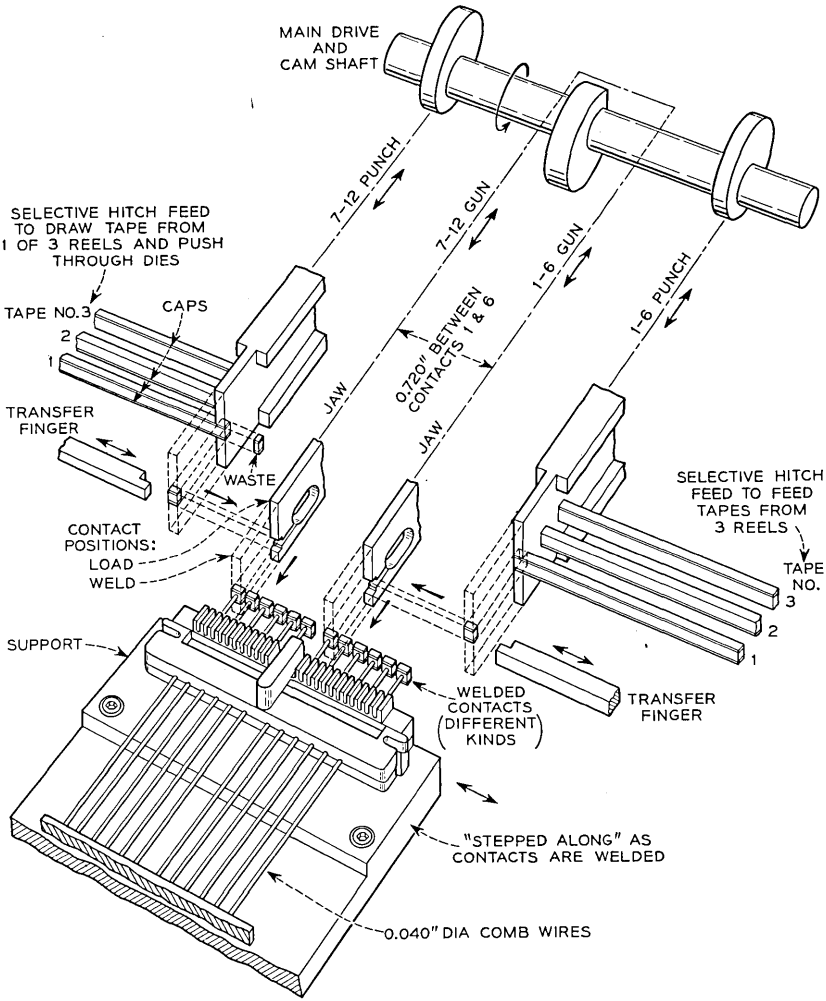


Fig. 18 — Schematic of the percussion welder.

wires is held stationary and the contact block is moved to the wires to make the weld. The speed of approach is an important factor in controlling the arc duration and the impact force. Various speeds of the contact block carriage from approximately 1 to 80 inches per second were tried. The best welds were obtained by a spring actuated gun which attained a velocity at impact of approximately 40 inches per second.

AUTOMATIC WELDER OPERATION

The automatic percussion welder contains duplicate welder heads which are mirror images of each other. Two contacts are welded at a time, one on each half of the comb, Fig. 18. After each welding operation or one cycle, the comb is indexed to the next pair of wire centers. Six cycles complete the welding of twelve contacts, or a lesser number if required. The guns do not weld at exactly the same time. There is an interval of one degree of revolution of the main shaft between the firing points to prevent electrical or mechanical interference. The welder was designed to select the type of contact, cut it from the tape and weld it in one cycle, to avoid the handling problems associated with precutting and magazining contact blocks.

TAPE SUPPLY AND SELECTOR

One of four contact conditions will apply for each wire; (1) no contact, (2) contact with palladium cap up, (3) palladium cap down, or (4) palladium cap both up and down. This requires three reels of tape on the right for one head and three on the left for the other. Adjustable knobs on both right and left tape feed cams are set for one of the four tape feed conditions for each wire position of the comb. Thus, any combination of contact conditions can be set up to make parts for any code of relay.

CONTACT SHEAR AND TRANSFER

The three tapes enter the shearing die through individual openings. However, only that tape selected by the tape selector is fed into the die and subsequently sheared by one punch stroke, Fig. 18. The tape is punched from such a direction that the base metal is not dragged over the boundary line into the palladium zone. This avoids contamination of the palladium. As the contact is blanked out, the walls of a notch in the punch confine it on the precious metal sides to prevent distortion. The punch delivers the contact to a transfer position at the end of the shearing stroke. There a transfer finger pushes the contact out of the punch notch, through a guide channel and into the waiting gun jaws.

WELDING GUNS

The welding gun is a light reciprocating member that carries two opposing steel fingers or jaws to receive the contact block from the transfer finger mentioned above. The jaw opening is a few thousandths of an inch less than the nominal height of the contact, however, the edges are beveled so that when a contact is pressed against them they spring open, the contact enters and is held securely in place. After welding the jaws are pulled off the contact. At the extreme return travel of the gun any contact which might remain in the jaws because it was not properly welded is removed by an ejector blade. When in the loading position, a portion of the blade stops the travel of the contact through the jaw opening so it is held in a uniform position and will be located on the wire with precision.

GUN MASS CONSIDERATION

During welding the comb must be supported accurately to meet the close contact location requirements. It must be supported securely to withstand the impact of the guns, or weak welds may result. The mass of the gun is important. Evidence indicates that more uniform and higher weld strengths are obtained with lightweight guns. A magnesium gun weighing about 60 grams produced better welds than did the original steel gun weighing about 130 grams. A newly installed steel gun weighing about 30 grams appears to be even more satisfactory. Other guns are being designed and tested to further check various features. A striking force of approximately 75 pounds tends to loosen the wires in the plastic and to produce weak welds. The 60 and 30 gram guns propelled at about 40 inches per second at impact produce less than half this force. The velocity during the arcing period is important to control the amount of heating. One-half cycle of vibration of jaw and wire after impact is the time available for the weld to freeze before a tension strain is placed on it. This time has been measured in the laboratory by the use of a transducer. Weak welds resulted when the time was less than 1 millisecond.

ELECTRICAL FUNCTIONS

Fundamentally, each weld circuit includes a capacitor which is charged during a small portion of each cycle and subsequently discharged through a resistance in series with the weld. During the charging period, which is controlled by a cam and microswitch, the contact and the wire end of the comb are separated electrically at the weld point. A mul-

tiple leaf brush under considerable pressure connects the one side of the circuit to the terminal end of the individual wire being welded. The gun, and through it the contact block, is connected to the other side of the circuit. After a cam frees the gun, the spring propells it toward the wire end and an electrical arc is established by the high potential (900 to

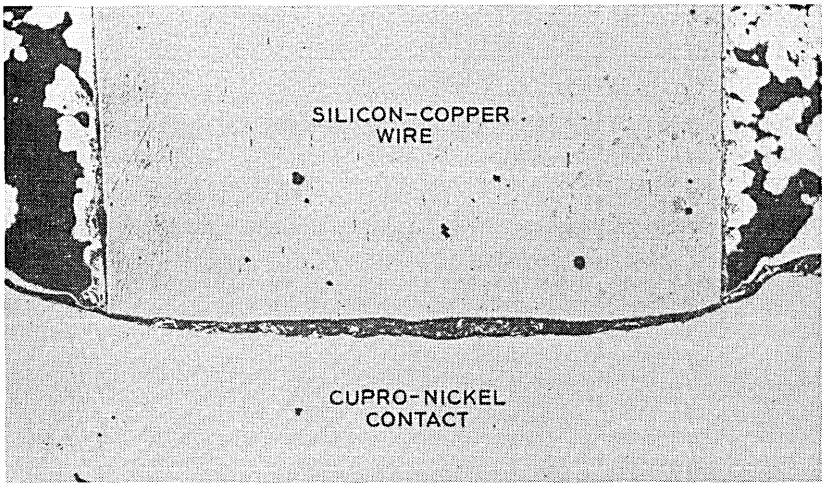


Fig. 19 — Section of a percussion weld. Original amplification 100 ×.

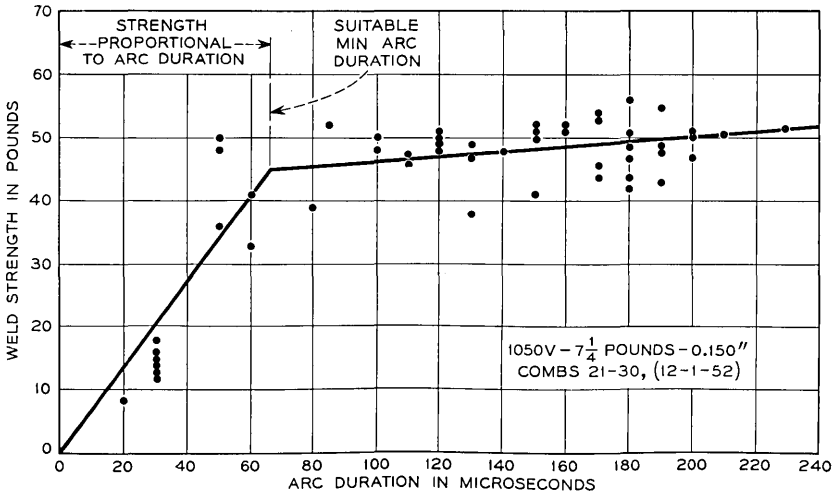


Fig. 20 — Weld strength versus arc duration.

1800 volts) just before the parts touch. The arc initiates itself when the gap has been reduced to a few thousandths of an inch. A portion of the abutting surfaces of both the wire and the contact base metal (normally 0.005 inch to 0.010 inch) are melted and expelled in liquid and gaseous states before the molten surfaces are forced together. The arc is extinguished when it can no longer melt and expel metal to maintain a gap. Under good operating conditions it persists from 0.1 to 0.4 milliseconds. Nearly all of the heated metal is expelled from the joint during the welding operation as illustrated by Fig. 19. This micrograph of a typical sectioned percussion weld shows only a 0.001 inch to 0.002 inch thick layer which was melted or heated sufficiently to change the structure.

A small stream of compressed air is directed into the weld area to remove gaseous, possibly ionized, arc products so that they will not

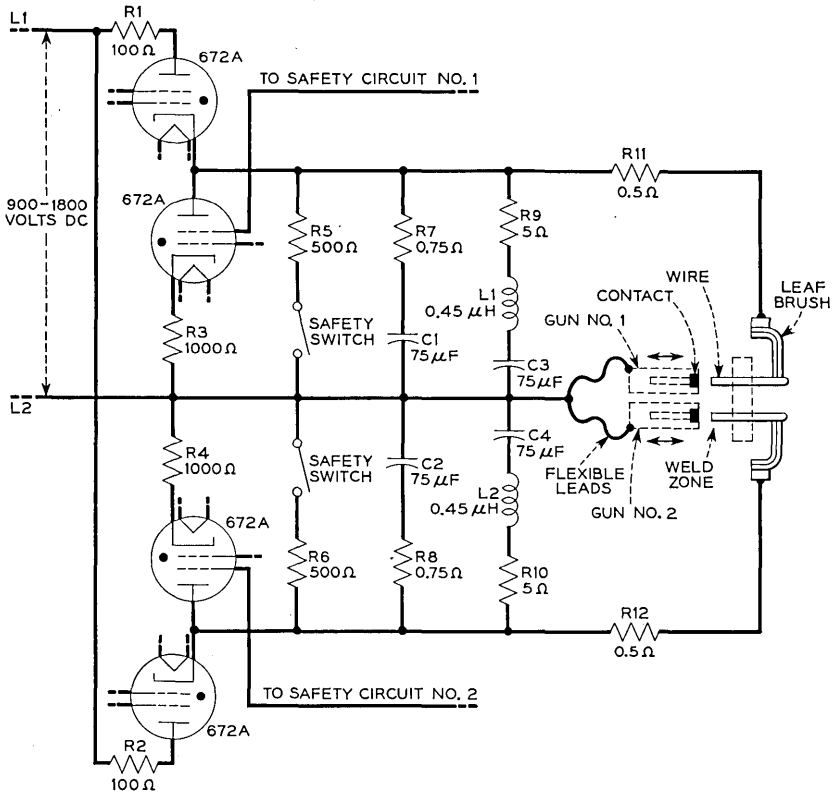


Fig. 21 — Schematic of percussion welder circuit.

interfere with the initiation of the arc during the next weld cycle. Limited tests made with nitrogen and helium atmospheres gave no indication of improvement in weld quality.

In the course of the many studies employing different kinds of instrumentation, an electronic counter was used to measure successive arc durations. Variations from 20 to 230 microseconds were observed for the prevailing conditions during early tests. The contacts welded were measured for strength of weld and a correlation found between short arc durations (up to 65 microseconds) and weak welds, Fig. 20.

Circuit impedance was found to be important in producing uniform weld strength. The assistance provided by the Laboratories in the early work led to better control of arc duration. Hawthorne continued this study and arrived at the circuit, Fig. 21 which greatly improved the uniformity of welds. The improved circuit resulted in less arcing in the jaws, thus prolonging their life and precision. Uniformity of burn-off and of weld strength both showed marked improvement. Fig. 22 shows a typical weld strength distribution for this circuit based on standard

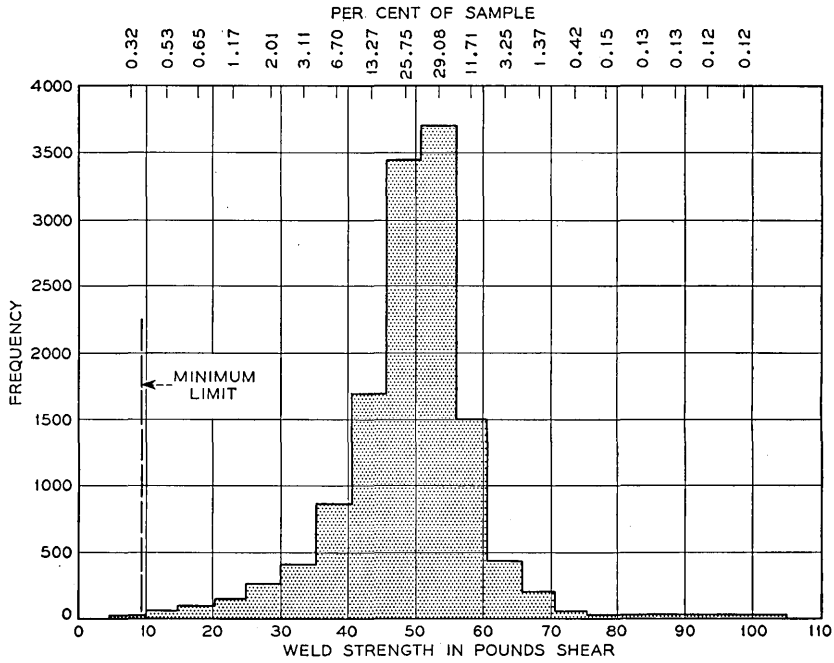


Fig. 22 — Weld strength distribution for 13,200 samples tested from 600,000 welds made in production.

shear test data. In a limited number of samples which were tension tested failure occurred more frequently in the wire than in the weld.

JAW LIFE

Jaw life is dependent upon many factors, but one of importance is the prevention of accidental conditions which establish an arc directly to the jaws. The jaws are adjusted so they will not touch the wire end and thereby discharge the capacitors if no contact is in the jaws. However, if only a very short length of contact metal is cut off and put in the jaws it may start the arc but there may not be sufficient material or the material may not be held securely enough to keep the arc from burning the jaw surfaces. Another possible trouble condition occurs when the end of the wire is misplaced so it touches a jaw. A safety circuit is provided with microswitches which trigger thyratrons to discharge the capacitors when either of these conditions occur. Signal lamps are provided to indicate at which microswitch the trouble is occurring. A reset microswitch is used in the test for shorted contacts so the indication can be held to a later time in the cycle. At the end of the stroke this switch is reset. After the part is nested approximately a quarter of a second is available to test for safe welding conditions. A cam-actuated microswitch controls the time of test and, if conditions are at fault, the weld energy is discharged through a thyratron before the parts are brought together for welding.

SAFETY

Safety for personnel from high voltage is provided by door switches, solenoid released shorting bars and bleeder resistors on the capacitors. Safety from mechanical jams is provided by a slip clutch on the main drive gear and by a pull out clutch and automatic stop switch located in the comb transfer drive mechanism.

CONCLUSIONS

After making millions of welds by automatic percussive welding methods, it is found to be a method well suited to this job. Accuracy of location and good weld strength are obtained. This welding method is especially useful where speed and precision are desired, and where joints must be made between dissimilar metals. Metals of high heat conductivity and high electrical conductivity join readily by this method.

ACKNOWLEDGMENTS

It is not possible to give due credit to all engineers who contributed to the development of these welding machines and to the preparation of this paper. Particular mention, however, will be made of H. Beedy, D. R. Brown, R. L. Kessel, G. A. Mitchell, J. M. Roach, J. P. Seider, R. Spillar, and the late Dr. B. J. Babbitt who, as Western Electric development engineers, engaged in various phases of this investigation. In addition to acknowledging the part played by numerous Bell Laboratories engineers, I desire especially to express my appreciation to D. C. Koehler and J. J. Madden for their able assistance while loaned to Western Electric during the early stages of the percussion welding study.

Electronic Relay Tester

By T. E. DAVIS and A. L. BLAHA

(Manuscript received September 24, 1953)

An electronic relay tester has been developed to gauge the contacts of a relay while it is pulsing. The device provides a visual presentation of the gauging position of up to 16 contacts at one time. The equipment has been developed primarily as a means for rapid inspection and concurrent adjustment (when needed) in relay manufacture. It may also be used as a laboratory instrument for studies of timing, chatter, and other performance characteristics of relays.

INTRODUCTION

Electronic equipment has been developed to gauge up to 16 relay contacts simultaneously as the relay operates or releases in pulsing. The system provides a visual presentation on an oscilloscope of the gauge points where the relay armature opens or closes the contact. The position of the armature is shown by the horizontal position of the scope beam. Each particular contact is represented on the scope by one of 16 horizontal lines generated by the motion of the armature. A vertical step on a line indicates the gauge point at which the contact operates.

The device consists of three main parts: (1) An electrostatic gauge to indicate the position of the relay armature, (2) An electronic scanner that continually switches across all the contact pairs, and (3) A brightness control circuit to intensify the beam of the scope when the armature is moving. Fig. 1 shows a schematic diagram of this equipment.

As shown on Fig. 2 the scope, test relay jack and electrostatic gauge are mounted on a bench. The scanner, power supplies and control circuits are mounted underneath the bench.

ELECTROSTATIC GAUGE

As the relay armature moves, its position is indicated at all times by the horizontal position of the scope beam in response to the output voltage of the electrostatic gauge, which is applied to the horizontal plates.

This output voltage varies with the position of the armature which serves as one plate of an air capacitor controlling the frequency of an oscillator. The voltage across a coupling impedance in the plate circuit of the oscillator is rectified to provide the output voltage of the gauge.

The gauge which is shown schematically on Fig. 3 may be divided functionally into three main parts: (1) An oscillator which operates in the region of 100 megacycles; (2) A detector sensitive to the frequency of the oscillator and consequently also sensitive to the position of the armature; and (3) A "zero set" cathode follower which connects the detector output to the input of a dc amplifier. The dc amplifier has suffi-

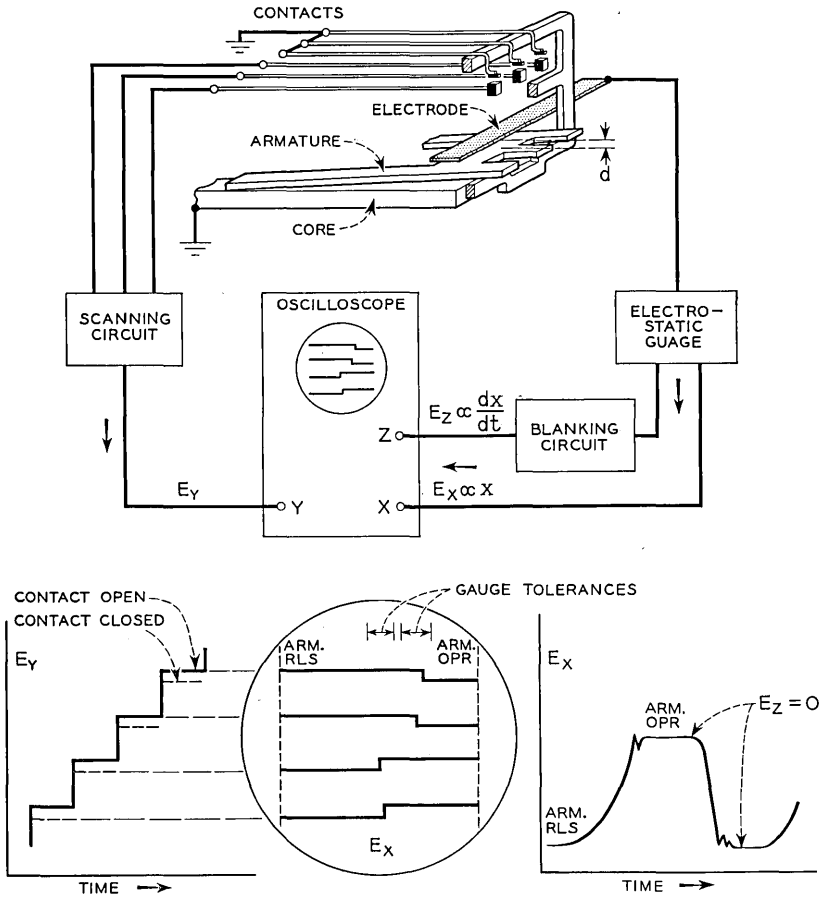


Fig. 1 — Schematic diagram of electronic relay tester.

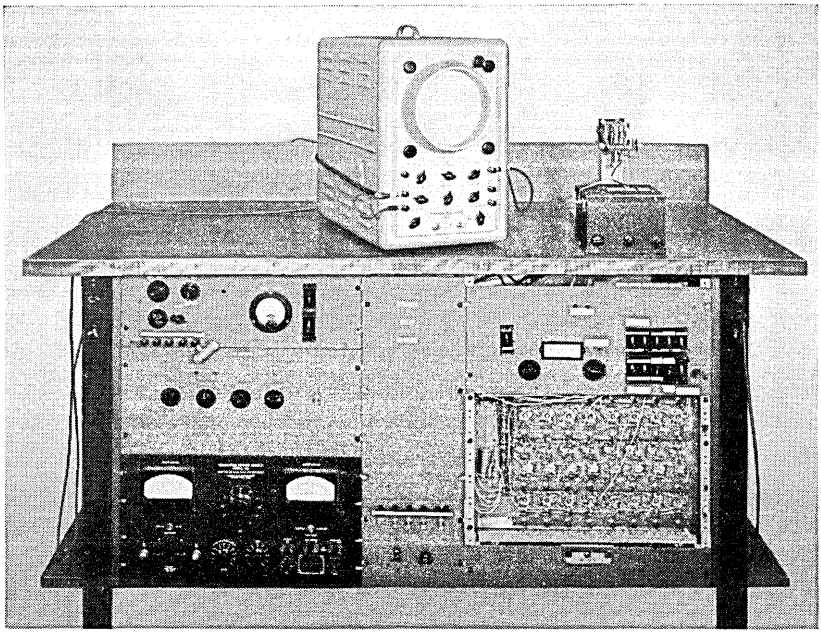


Fig. 2 — Photograph showing arrangement of equipment.

cient gain to provide adequate deflection voltage to the horizontal plates of the scope.

The oscillator comprises a 6AU6 pentode tube and an associated tank circuit. The tube is connected as an electron coupled oscillator with the energy for the oscillations being supplied by the screen circuit. The plate output is a modulated current that passes through the screen and suppressor grids. Considerable isolation of the coupling circuit from the oscillator is obtained by this arrangement.

An antiresonance circuit similar to the tank circuit is used to couple the plate of the oscillator tube to the rectifier. Since the electron coupled oscillator acts as a high impedance generator or constant current source the voltage across the coupling impedance is proportional to the value of the impedance. Hence as the oscillator frequency is shifted by a change in the position of the armature a corresponding shift occurs in the ac voltage across the coupling impedance and rectifier tube (6AL5). The output voltage of the rectifier also shifts with the input to the rectifier since the changes are slow compared with the time constant of the output circuit of the rectifier.

In normal operation the output of the rectifier may vary from 6 volts

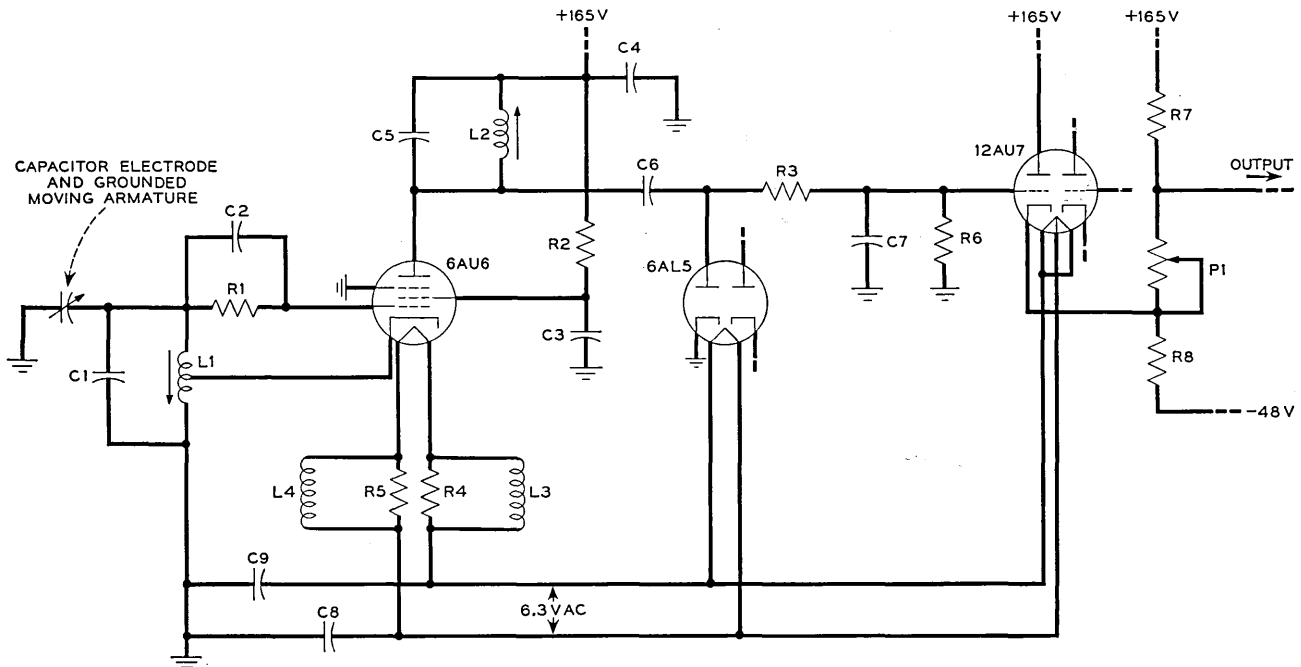


Fig. 3 — Schematic circuit of electrostatic gauge.

to 10 volts with 10 volts appearing when the armature is operated. A typical voltage-displacement curve is shown on Fig. 4. This output is applied to the grid of the "zero set" cathode follower and its bias is set so the output will be from approximately -2 volts to $+2$ volts. This voltage is applied to the horizontal dc amplifier of the scope.

Theory of Operation of Gauge

Since the capacitance probe is in the tank circuit the frequency of the oscillator changes with variations in the separation of the probe electrodes. Hence a displacement of the moving electrode causes a change in the frequency but no change in the amplitude of the alternating current through the pentode plate. The radio frequency voltage that appears across the coupling impedance is the product of the plate current and the coupling impedance.

Since the current is constant in amplitude:

$$E = I_0 Z,$$

where $E = rf$ voltage applied to rectifier, $I_0 = rf$ plate current from oscillator, and $Z =$ coupling impedance.

Since a change in E is due to a change in the oscillator frequency and the resultant change in Z;

Let $$f = \frac{\omega}{2\pi} = \text{oscillator frequency},$$

$$f_2 = \frac{\omega_2}{2\pi} = \text{resonance frequency of coupling circuit},$$

$$Q = \frac{\omega_2 L_2}{R},$$

and

$$Y = 1 - \frac{f^2}{f_2^2}.$$

where $L_2 =$ inductance of coupling circuit, and $R =$ resistance of coupling circuit.

The coupling impedance can be written:

$$Z = \frac{\omega L_2 Q}{\sqrt{Q^2 Y^2 + 1}} \tag{1}$$

At resonance $Z_{\max.} = \omega_2 L_2 Q$.

Since the maximum value of E is obtained when Z is at a maximum:

$$\frac{E}{E_{\max}} = \frac{Z}{Z_{\max}} = \frac{1}{\sqrt{Q^2Y^2 + 1}}, \tag{2}$$

provided ω/ω_2 is near unity.

Probe and Grid Circuit of Oscillator

The oscillator frequency is determined by the resonant frequency of the tank circuit. Fig. 3 shows that the tank comprises the capacitance of the probe plus a fixed capacitance and an inductance. The probe which is assumed to be a parallel plate condenser has a capacitance which is inversely proportional to the plate separation.

Let: d be plate separation of probe, d_1 be separation at resonance (where $f = f_2$),

$$x = \frac{d}{d_1}.$$

Then the variable capacitance of probe is given by:

$$C_p = C_1 \frac{d_1}{d} = \frac{C_1}{x}$$

and the total capacitance of the tank by:

$$C_0 + C_p = C_0 + \frac{C_1}{x},$$

where C_0 = fixed capacitance of tank including fixed part of probe capacitance.

Let:

$$\frac{C_1}{C_0} = K,$$

then

$$\omega^2 = \frac{1}{L_1 C_0 \left(1 + \frac{K}{x}\right)} \quad \text{and} \quad \omega_2{}^2 = \frac{1}{L_1 C_0 (1 + K)},$$

where ω and ω_2 are radian frequencies of the oscillator at probe separations d and d_1 respectively and L_1 is the effective inductance of the tank circuit. It follows that:

$$\frac{\omega^2}{\omega_2^2} = \frac{1 + K}{1 + \frac{K}{x}},$$

and therefore:

$$\begin{aligned} Y = 1 - \frac{\omega^2}{\omega_2^2} &= 1 - \frac{x + Kx}{K + x}, \\ &= \frac{1 - x}{1 - \frac{x}{K}}. \end{aligned} \tag{3}$$

This expression for Y may be substituted in (2) giving:

$$\frac{E}{E_{\max}} = \frac{1}{\sqrt{Q^2Y^2 + 1}} = \frac{1}{\sqrt{1 + Q^2\left(\frac{1 - x}{1 + \frac{x}{X}}\right)^2}}. \tag{4}$$

For $x/K \gg 1$ (4) may be written:

$$\frac{E}{E_{\max}} \frac{1}{\sqrt{1 + K^2Q^2\left(\frac{1 - x}{x}\right)^2}}. \tag{5}$$

Equation (5) gives the voltage across the rectifier as a function of probe separation and the product KQ. It can be used to determine the optimum values of: (1) K the ratio of variable to fixed oscillator capacitance, (2) the Q of the coupling impedance, and (3) the range of probe

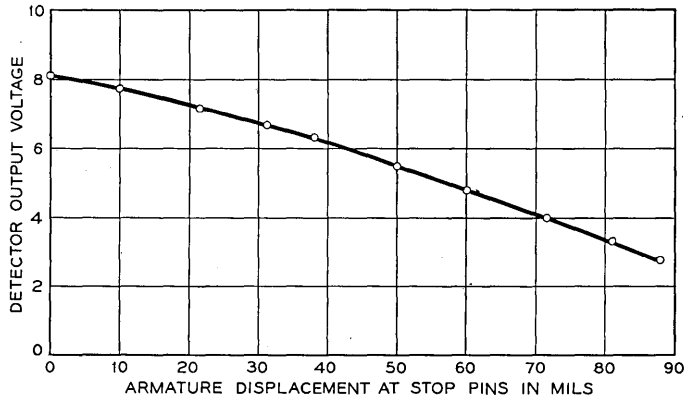


Fig. 4 — Typical output voltage versus displacement curve of the electrostatic gauge.

separations over which a prescribed degree of linearity can be obtained. It can be shown that the maximum range of linearity is obtained with $(KQ)^2 = 2$ and that the center of this range is at $x = \frac{1}{2}$. Fig. 5 is a graph of (5) against x with $(KQ)^2$ set at 2.

SCANNING CIRCUIT

The scanning circuit is a high speed electronic switching and detecting device which rapidly selects successive contacts, checks whether the contact is opened or closed, and displays this information on the vertical plates of the scope. Each contact is assigned a particular vertical reference level. When gauging, each level appears as a horizontal line with a vertical step on the line indicating the armature position for the contact operation.

Electron tubes designated "C" on Fig. 6 serve as switches to connect one contact at a time to the vertical plates of the scope. Each contact is connected to the grid circuit of a C tube. In order to test a contact its

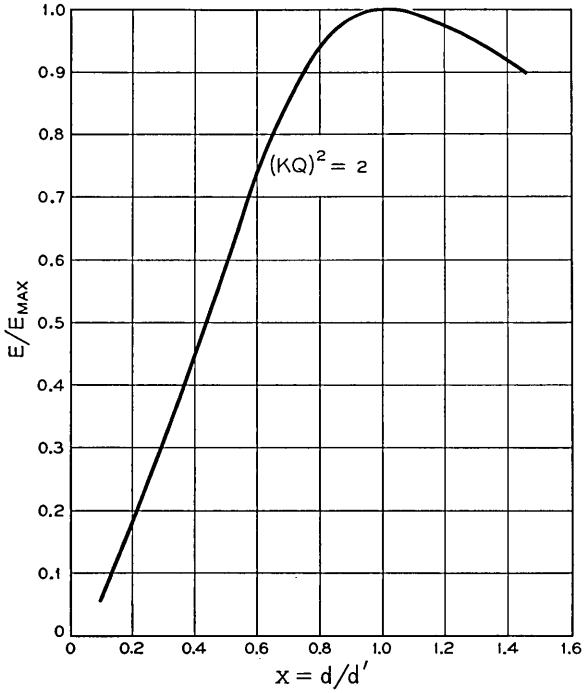


Fig. 5 — Curve showing E/E_{max} versus x with $(KQ)^2 = 2$ computed from equation (5).

associated C tube is made to conduct by shifting the grid voltage. All of the C tube cathodes are connected together to a common cathode resistor which couples the voltages from the contacts to the vertical amplifier of the scope. The plate current that flows when a C tube is fired causes a voltage drop through the common cathode resistor which deflects the scope beam to the proper vertical level. The level is set to the desired value by adjusting the plate circuit resistance. Since the contact under test is connected to the grid circuit, the grid voltage and the cathode voltage are shifted a small amount by opening or closing the contact. That is, when a C tube is fired the vertical plate voltage jumps to one value when the contact is open and it jumps to another slightly different value when the contact is closed. The 16 C tubes corresponding to the 16 contacts under test are fired one at a time in succession by 16 associated multivibrator stages. The multivibrators are connected in a ring so that each stage is fired by the preceding stage. When a stage is fired it holds its C tube in a conducting condition for two microseconds.

Fig. 6 is a detailed schematic of the single stage multivibrator (tubes A and B) with an associated modulator (C) tube. The multivibrators are normally in a stable waiting state and go into a temporary unstable state only when a transient is applied. Normally the A tube is cut off and the B tube is conducting. When a pulse is applied to the A tube grid the A tube conducts and the B tube is cut off. After two microseconds the A tube reverts back to its waiting state and sends a pulse to the next stage. When the B tube is cut off it provides a flat two microsecond pulse through the coupling circuit to the associated C tube.

Each multivibrator stage consisting of the A and B tubes with other circuit elements is mounted on a plug-in turret which may easily be changed in case of trouble.

BRIGHTNESS CONTROL CIRCUIT

When pulsing a relay the armature remains on the operated and released positions for a relatively long time interval. If the intensity of the scope beam is allowed to remain constant the spots at the ends of the traces are bright enough to fog the entire scope face. This makes it difficult to see the relatively weak lines that are caused by the motion of the armature. Therefore a control circuit is provided to brighten the scope trace only when the relay armature is in motion. The circuit shown on Fig. 8 provides the voltage required to intensify the cathode beam of the scope. This voltage is obtained by differentiating the output voltage of the electrostatic gauge. As the latter is proportional to the

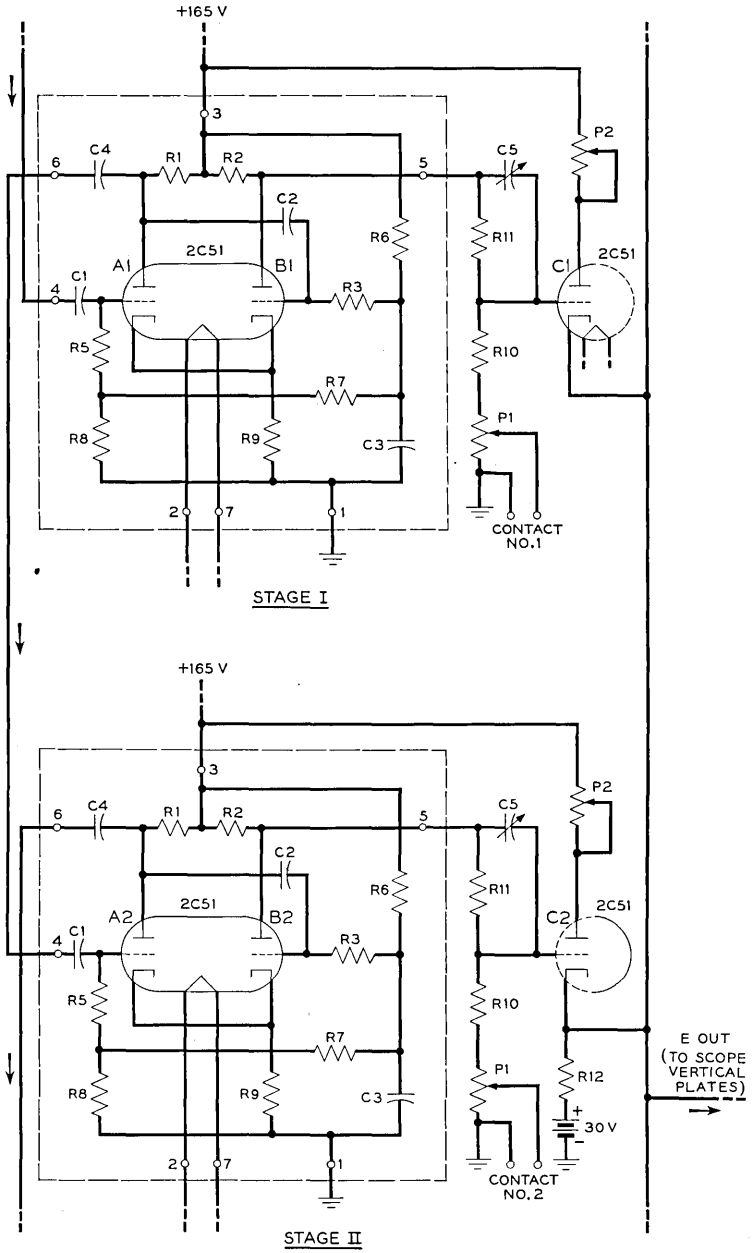


Fig. 6 — Schematic circuit of two switching stages showing modulator tubes (C) and their associated multivibrator tubes (A and B).

armature position, the differentiated voltage is proportional to the armature velocity. After amplification the voltage from the gauge is applied to a tube with a transformer in the plate circuit. The voltage across the secondary coil of the transformer is the differentiated voltage, proportional to the armature velocity. It is applied to a germanium diode bridge, which may be connected as a half-wave or full-wave rectifier by the selector switch. The rectified voltage is amplified and clipped and then applied to the Z axis terminals of the scope. The scope trace is brightened during operate, during release, or during both in accordance with the setting of the bridge selector switch.

GAUGING A RELAY

The relay is plugged into a holding fixture with a jack for the coil and contact terminals. The jack connects the relay coil to the circuit which provides power for operating the relay. It also connects the relay contact terminals to the scanning circuit.

The electrostatic transducer electrode is mounted on a bracket which is attached to the front end of the fixture by means of a hinged bracket.

After the relay is plugged into the jack, it is clamped and the gauge electrode is rotated into position near the armature. Then the "zero set" potentiometer on the dc amplifier is used to align the operated armature position with the zero marking on the calibrated horizontal scale of the scope.

The contact selector switch is used to select the combination of contacts to be scanned such as: all contacts, all breaks, or all makes. For convenience in checking relays with 8 contacts or less a switch on the scanning circuit is used to connect 8 of the multivibrators in a ring instead of 16 so that only 8 lines appear on the scope. These 8 lines may be shifted to obtain greater spacing for easier reading. A typical gauging pattern on an oscilloscope screen is shown on Fig. 7.

This also improves the gauging accuracy which depends upon the amount of armature motion between successive scanning dots. For 16 horizontal lines on the screen the time interval between successive dots is 32 microseconds. For an armature velocity of 30 inches per second the corresponding distance between successive dots on one line is about 1 mil-inch. If 8 lines are used this distance is about 0.5 mil-inches. The gauging error resulting from this dot definition is actually less than indicated because successive operations of the armature give a small random variation in the dot locations.

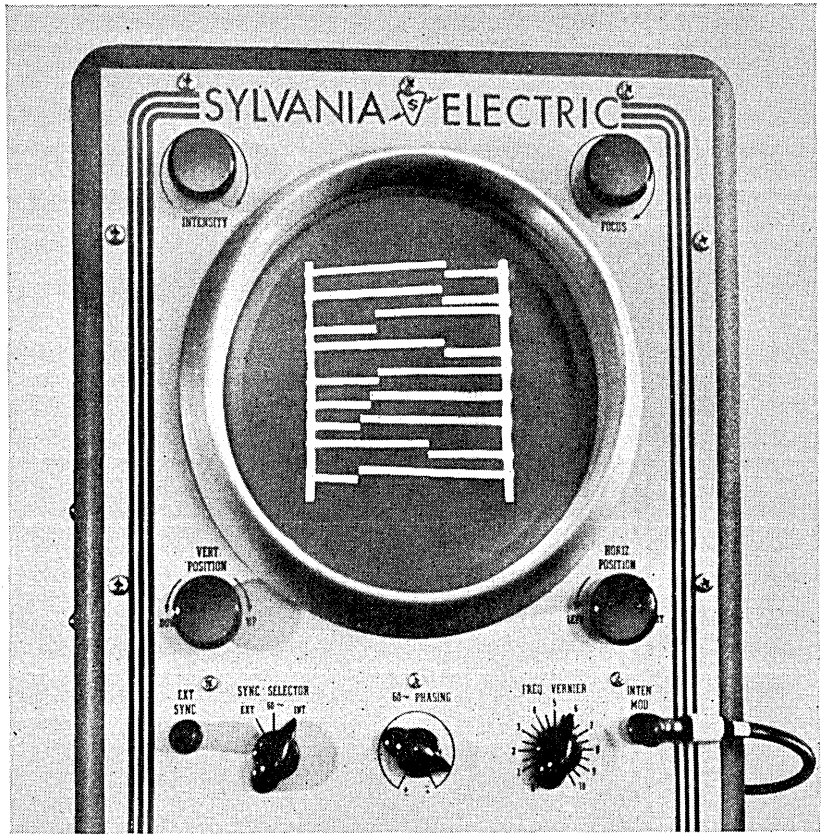


Fig. 7 — Photograph of the 7-inch oscilloscope with a typical gauging pattern on the screen.

The brightness control switch is used to compare gauging when the relay is operating with the corresponding gauging when it is releasing. An apparent difference of 2 to 3 mil-inches is noticed between the operating gauge point of a contact and the releasing gauge point. This measurement change is ascribed to the tip flexure or follow of the contact wire. Contact closure occurs when the contacts first touch, in advance of the short follow travel during which the tension of the contact wire is transferred from the card to the contact. In dynamic operation, the contact opens when it is first struck by the card, before the tension would be transferred statically from the contact to the card. Thus the dynamic contact closure points agree with the static gauging, and differ from the dynamic opening point by the amount of contact follow. Good

correlation is obtained between the static and dynamic gauge measurements if make contacts are gauged on operate and break contacts on release.

OTHER RELAY CHARACTERISTICS

Switches are provided to permit other relay charcateristics to be observed on the scope such as armature motion, contact chatter, and the contact operate and release times. The contact gauging is obtained with the horizontal plates connected to the electrostatic gauge. With this setting, the time for a dot to move across the screen may be less than 3 milliseconds (depending on the armature transit time) and contact chatter may be observed during this time.

With a time base sweep on the horizontal plates, the armature motion

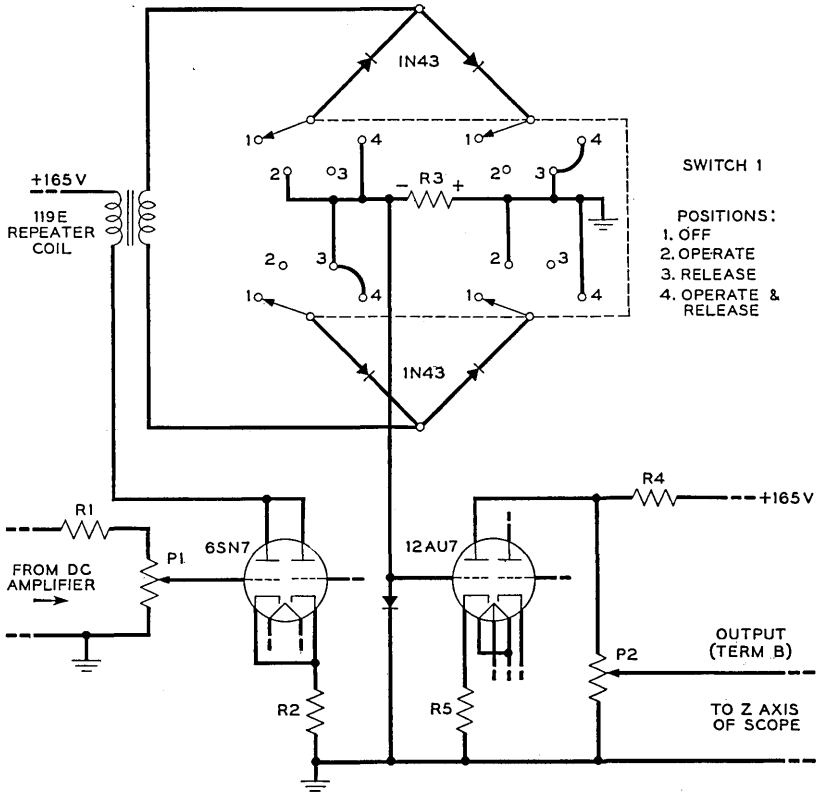


Fig. 8 — Schematic of oscilloscope brightening circuit.

may be put on the vertical plates. The chatter of 16 contacts may also be observed at one time to obtain a quick check of the contact performance. Operate and release times of the contacts may be read directly from the scope if a pip is put on the vertical axis when the test relay coil is energized. The front end of the relay is clamped for all measurements.

Switches are provided to select different combinations of contacts to be checked with either a time base horizontal sweep or with the armature displacement as the horizontal sweep.

DISCUSSION

The electronic relay tester was designed primarily for rapid inspection of wire spring relays. It provides a means for observing the position of the armature for the operation of all contacts simultaneously while the relay is pulsing. This method has several advantages over thickness gauges which are frequently used: (1) The contacts are gauged under conditions of use, that is, the armature moves as in normal operation without touching the gauging device; (2) Inspection is more rapid since the go-no-go positions are displayed on the scope face. Relay spring combinations requiring a sequence of contact operation are readily observed while the relay operates in a normal way; and (3) When relay adjustments are made the results of the adjustment are shown immediately so the contact operate points are centered within the desired range giving more margin for changes that may occur with use. This device is readily changed to measure other relay characteristics that may be of interest such as operate or release time, contact chatter and armature rebound.

ACKNOWLEDGMENTS

The authors are indebted to J. H. McConnell who designed a similar electrostatic transducer and to R. L. Peek, Jr., who derived the equations for it. Mr. Peek also suggested the method used in the scanning circuit.

Topics in Guided Wave Propagation Through Gyromagnetic Media

Part II—Transverse Magnetization and the Non-Reciprocal Helix

By H. SUHL and L. R. WALKER

(Manuscript received March 30, 1954)

Propagation through a gyromagnetic medium in a direction normal to a uniform magnetizing field is considered. Geometrical arrangements which make this propagation non-reciprocal are described. A few illustrative examples are discussed briefly. The non-reciprocal helix, of importance in traveling wave tube work, is treated at length.

1. INTRODUCTION

1.1. *General Remarks about Non-Reciprocal Propagation*

Part I of this paper began with a brief discussion of some of the microwave properties of two gyromagnetic media; the gas discharge plasma and the ferrite. The remainder of Part I was devoted to the analysis of the mode spectrum in a cylindrical waveguide filled with one of the media and placed in an axial magnetic field. It was demonstrated that the natural modes in such a guide are right- and left-circularly polarized waves which travel with different phase velocities. Accordingly a plane polarized mode, which to some approximation can be regarded as the sum of right and left circular modes, will, in traversing a section of the guide, undergo Faraday rotation, just like a plane wave in the unbounded medium. It is true that the presence of the guide wall has a drastic effect on the course of the rotation with magnetizing field, changing it, sometimes beyond recognition, from that prevailing in the unbounded medium. Nevertheless the principle remains the same; confinement of the wave to a guide merely modifies quantitatively the Faraday effect for plane waves. In optics, where practically plane waves are almost always employed in this connection, the non-reciprocal nature of this effect is so familiar that it hardly requires restatement here.

By contrast, Part II of this paper deals with devices whose non-reciprocal operation depends *in principle* as well as in numerical detail on the disposition of the boundary, or, more generally, on geometrical configuration. These devices employ magnetizing fields transverse to the propagation direction. Some electromagnetic field configurations are unaffected by such a dc field, but, whenever the rf magnetic field in the case of a ferrite (or the rf electric field for a plasma) has a component normal to the dc magnetic field, this is no longer the case. For, now, the magnetization (or the charges) will be caused to precess about the dc field, giving extra terms in Maxwell's equations and a resultant change in the propagation. This change may be simply an alteration in phase velocity, the propagation remaining reciprocal. This is the case, for example, for the propagation of plane waves in an infinitely extended medium [Cotton-Mouton effect]. Here, since every direction of propagation normal to the dc field is physically equivalent to any other and, in particular, to the opposite direction, no non-reciprocity can arise.

For reciprocity to be preserved in the presence of the dc magnetic field is, however, exceptional and requires a certain amount of geometrical symmetry in the system. That non-reciprocity may be expected in asymmetrical systems may be foreseen if we consider a system, typical of those to be treated in this paper, in which all the rf fields are independent of the coordinate along which the dc magnetic field is pointing. The relevant conducting boundaries and any interfaces between ferrite (plasma) and air are all surfaces parallel to the direction of propagation and lying in the dc magnetic field direction. Suppose the system to be divided into two parts by another surface of a similar kind and examine the surface impedance of one of the parts (which should contain some gyromagnetic material). If the propagation direction be reversed it is necessary to reverse the magnetic field to retrieve a situation in the part considered *geometrically* equivalent to the original. But, since the precession of the magnetization (or charges) about the dc magnetic field has a definite sense, the magnetic or electric current associated with this precession will be reversed when the dc field is reversed. Thus, the properties of the medium are altered and the surface impedance will be different for the two directions of propagation. In general, the surface impedance of the other part of the system will not compensate for this distinction between the two directions and we shall find different propagation constants for opposing directions. An exception will occur if the system contains a surface about which it has geometrical symmetry, for, then, compensation clearly takes place about this surface and the system is reciprocal.

An example of a simple non-reciprocal system is indicated in Fig. 1(a). Here a slab of ferrite is inserted into a rectangular waveguide parallel to the narrow walls and closer to one of them. Several workers have demonstrated that this arrangement and a similar arrangement in a circular waveguide are non-reciprocal for what is essentially the dominant mode.[1, 3, *] When the slab is centered in the guide we have a plane of symmetry and the non-reciprocity vanishes.

Another configuration of the transverse field type is represented by the system shown in Fig. 1(b). Here a hollow ferrite cylinder is magnetized circumferentially and propagates a TE_{0n}-mode. It is clear that any arrangement of this sort, which might, in principle, include conducting sheaths, internally or externally, or might have the ferrite extending to an indefinitely large or small radius, cannot have any symmetry

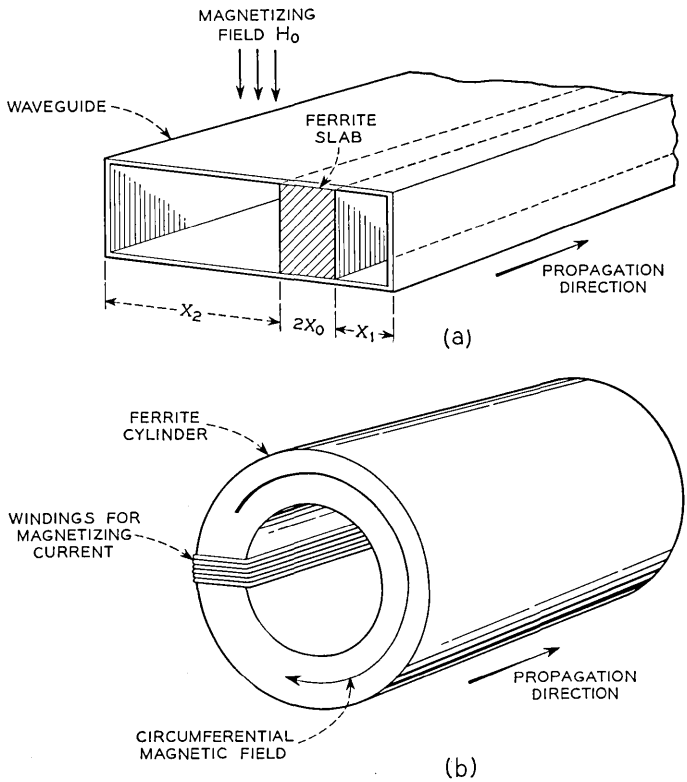


Fig. 1 — (a) Rectangular waveguide and ferrite slab. (b) Circumferentially magnetized ferrite cylinder.

* It is expected that an article on this subject by S. E. Miller, A. G. Fox and M. T. Weiss will appear in a forthcoming issue of the JOURNAL.

about a cylinder coaxial with the ferrite. Thus, the internal and external impedances of such a system at any coaxial cylinder can only compensate accidentally (perhaps at a single frequency) and non-reciprocity is the rule.

It is frequently asserted without qualification that for non-reciprocity a further condition upon the relevant rf field is that its projection upon a plane normal to the magnetizing field be elliptically or circularly polarized in the limit of vanishing magnetization. The argument is based on the consideration, in itself correct, that the effective material constants are different for right- and left-circularly polarized field vectors. Suppose that the magnetization direction is y. Then the tensor relating B to H is (see Part I, Section 2.1):

$$\begin{Vmatrix} \mu & 0 & -j\kappa \\ 0 & \mu_0 & 0 \\ j\kappa & 0 & \mu \end{Vmatrix}.$$

For right- and left-circular fields with $H_z = \pm jH_x$, therefore, the medium is isotropic in the plane transverse to the dc field with permeabilities $\mu + \kappa$, $\mu - \kappa$ respectively. Since opposite circular polarizations accompany opposite propagation directions, (see for example, Fig. 2) the permeabilities, and hence the propagation constants, are different for opposite propagation directions. It is then argued that the field must already be circularly, or at least elliptically polarized to start with, if non-reciprocal effects are to result from application of the magnetization. However, the argument is true only for effects of first order in the magnetization. For general values of magnetization, the rf field, even if linearly polarized to begin with, will become elliptically polarized, and

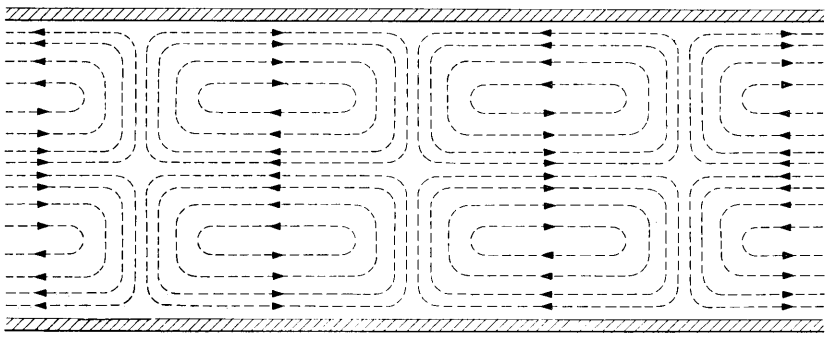


Fig. 2 — Magnetic lines of force parallel to the broad side of a rectangular waveguide.

non-reciprocity will occur. It is understandable that this double function of the magnetization (conversion to elliptic polarization plus creation of differences in permeability) leads to higher order effects. However, inasmuch as for the ferrite and plasma the magnetization can produce large changes, the requirement of elliptic polarization in zero magnetic field cannot be regarded as essential in practice. These considerations are demonstrated by a simple example in Section 2.2.

The devices considered in this paper actually are such that the electric or the magnetic field vector in the plane normal to the field is elliptically polarized even in the absence of the gyromagnetic medium. For example, the magnetic lines of force of the TE_{10} mode in a rectangular waveguide form two sets of closed loops in a plane parallel to the wide sides (see Fig. 2) and repeating every wavelength. This pattern moves bodily down the guide with the phase velocity of the mode, so that an observer stationary at any point not at the center or at the narrow walls of the guide sees a magnetic field rotating at the signal frequency and tracing out a generally elliptic path. The sense of the rotation is opposite on opposite sides of the center plane, and depends on the propagation direction. The conditions outlined in the previous paragraph are therefore satisfied; introduction of a ferrite slab magnetized as in Fig. 1(a) will yield first order non-reciprocal effects.

The problems considered here are such that the electromagnetic fields do not vary in the direction of magnetization. Under these conditions the field can be split into a TE and TM field satisfying different wave equations. In general, the two fields are coupled through the boundary conditions. Most of the paper is devoted to the analysis of the non-reciprocal helix, a problem that has recently gained importance in connection with high power traveling-wave-amplifiers.[4] The conventional amplifier suffers from a limitation on its maximum useful gain; waves reflected from the output end will make the tube "sing" above a certain critical gain. Ferrites offer the possibility of preferentially suppressing these backward waves and so of increasing the permissible gain by a large amount.* In section 2.3 the "flat" helix (one of infinite radius) is considered. For the slow waves employed in practice a rather complete treatment is possible in this case of planar geometry. In Section 3 the cylindrical helix is treated. Inasmuch as the solutions involve functions for which no extensive tables exist, the treatment has to be more sketchy.

* More specifically, in high-power traveling-wave tubes the large beam current employed may be above the critical value required for backward wave oscillations due to spatial harmonics of the helix structure. In such cases the larger attenuation of backward waves will permit a higher beam current and therefore stable amplification to higher power levels.

Thus the loss is neglected and only the non-reciprocal phase characteristic is considered. The losses have then to be determined approximately by differentiation of the phase characteristics.

A few further problems were considered as illustrations of the general principles. One case, that of the plasma filling the space above an impedance sheet can actually be solved analytically and provides a particularly clear demonstration of the non-reciprocity. The case of the rectangular waveguide with a ferrite slab has already been considered extensively elsewhere, and only the results for a thin slab are given here. A problem with cylindrical symmetry is taken up in Section 3.3: a cylindrical waveguide fitted with a circumferentially magnetized cylinder

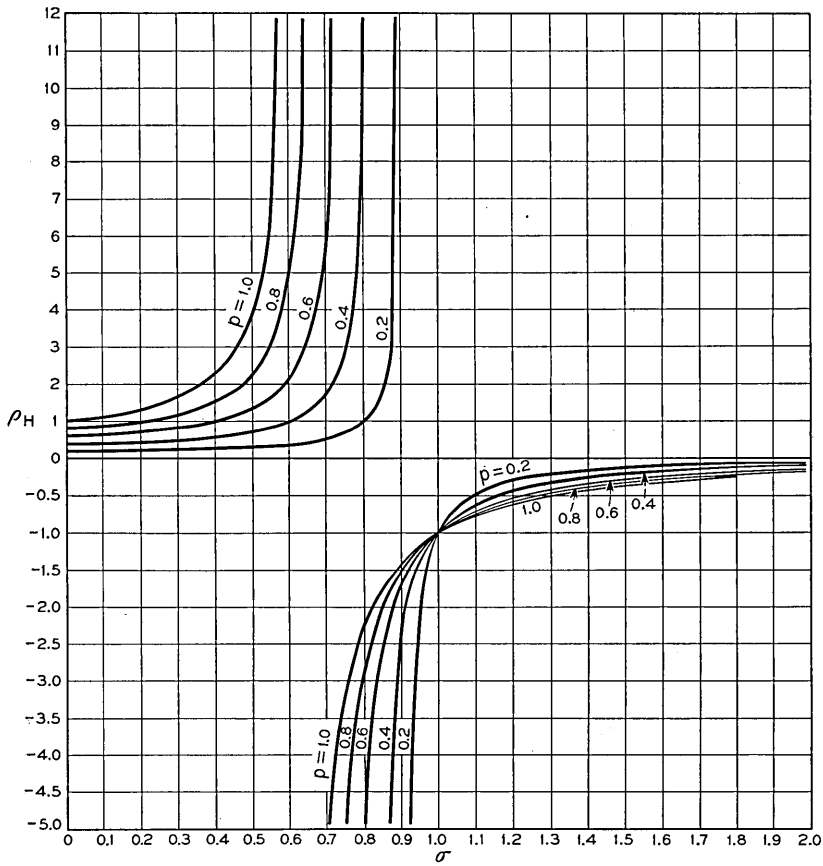


Fig. 3 — ρ_H versus σ for various p.

of ferrite. Again the discussion had to be sketchy in view of the scarcity of information on the functions that solve the problem.

2. PLANAR GEOMETRY

2.1. *Fields and Impedances*

In this section we consider planar transverse field problems which are characterized by the following conditions. The dc magnetic field is of uniform strength H_0 within the gyromagnetic medium and points along the y-axis. All rf field components are independent of the y coordinate. We discuss the ferrite case first, then indicate how the results are to be translated for the plasma.

For the orientation of the dc magnetic field which has been chosen the permeability matrix is of the form:

$$\begin{Vmatrix} \mu & 0 & -j\kappa \\ 0 & \mu_0 & 0 \\ j\kappa & 0 & \mu \end{Vmatrix},$$

μ and κ are, in general, even and odd functions of H_0; the permeability of unmagnetized ferrite is taken to be μ_0 as in free space. Following the procedure of Part I we shall assume specifically for μ and κ the formulae given by Polder's treatment of the dynamics of the medium. Thus, we have the expressions (for the case of no loss):

$$\begin{aligned} \frac{\mu}{\mu_0} &= \frac{1 - p\sigma - \sigma^2}{1 - \sigma^2}, \\ \frac{\kappa}{\mu_0} &= \frac{p}{1 - \sigma^2}, \text{ and} \\ \rho_H &= \frac{\kappa}{\mu} = \frac{p}{1 - p\sigma - \sigma^2}, \end{aligned} \tag{1}$$

where σ is the ratio of the precession frequency, $\frac{|\gamma|}{2\pi} H_0$, to the signal frequency and p is the ratio of a frequency, $\frac{|\gamma|}{2\pi} M_0/\mu_0$, associated with the saturation magnetization, M_0, to the signal frequency. It should be noted that p and σ have always the same sign. The behavior of μ and κ as functions of σ was shown in Fig. 1(a) and (b) of Part I. ρ_H is shown as a function of σ in Fig. 3. The dielectric constant of the ferrite is taken to be ϵ. For reasons given in Part I, $|\ p\ |$ is assumed less than unity.

When the condition, $\partial/(\partial y) \equiv 0$, is put into Maxwell's equations the latter are found to be separable into two sets:

$$-\frac{\partial H_y}{\partial z} = j\omega\epsilon E_x, \tag{2a}$$

$$\frac{\partial H_y}{\partial x} = j\omega\epsilon E_z, \tag{2b}$$

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = -j\omega\mu_0 H_y, \tag{2c}$$

and

$$-\frac{\partial E_y}{\partial z} = -j\omega[\mu H_x - j\kappa H_z], \tag{3a}$$

$$\frac{\partial E_y}{\partial x} = -j\omega[j\kappa H_x + \mu H_z], \tag{3b}$$

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = j\omega\epsilon E_y. \tag{3c}$$

It is to be stressed that such a separability is possible only when the rf fields do not vary along the dc magnetic field. The sets of equations (2) and (3) correspond to the separate equations for H_z and E_z which arise from (13) of Part I when β is there set equal to zero. The first set describes a TM field of the familiar type, whose propagation through the medium is unaffected by the magnetic field. The second set describes a TE field whose components, because of the presence of κ, are connected by different relations from those which exist in an unmagnetized medium. The separability of the two fields is equivalent to saying that they are not coupled by the medium itself, but they may, of course, be coupled at the boundaries.

We may write (3a) and (3b) in the form

$$-j\omega(\mu^2 - \kappa^2)H_x = j\kappa\frac{\partial E_y}{\partial x} - \mu\frac{\partial E_y}{\partial z}, \tag{4a}$$

$$-j\omega(\mu^2 - \kappa^2)H_z = \mu\frac{\partial E_y}{\partial x} + j\kappa\frac{\partial E_y}{\partial z}, \tag{4b}$$

and upon eliminating H_x and H_z, the wave equation for E_y is found to be:

$$\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial z^2} + \omega^2\epsilon\frac{\mu^2 - \kappa^2}{\mu}E_y = 0, \tag{5}$$

where E_y (and also the H's) are evidently propagated in the ferrite as

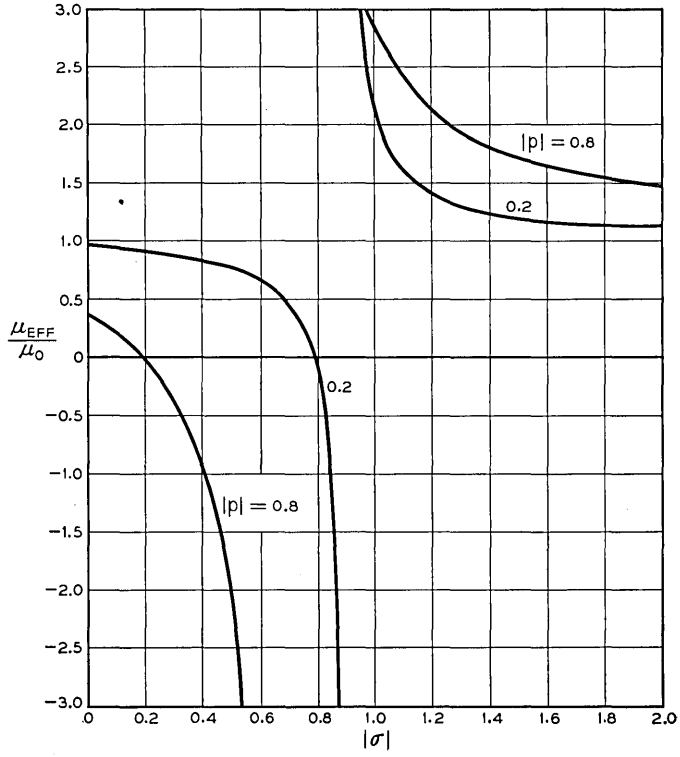


Fig. 4 — $\frac{\mu_{\text{eff}}}{\mu_0}$ versus $|\sigma|$.

though the latter had an effective permeability, $(\mu^2 - \kappa^2)/\mu = \mu_{\text{eff}}$. This permeability may assume any value from $+\infty$ to $-\infty$ as may be seen by writing it in terms of σ and p. From the Polder formulae, in fact,

$$\mu_{\text{eff}} = (\mu^2 - \kappa^2)/\mu = \mu_0 \frac{1 - (p + \sigma)^2}{1 - p\sigma - \sigma^2}. \tag{6}$$

As it should be, this is an even function of magnetic field. μ_{eff}/μ_0 decreases from $1 - p^2$ to 0 as $|\sigma|$ rises from 0 to $1 - |p|$; it decreases from 0 to $-\infty$ as $|\sigma|$ runs from $1 - |p|$ to $\sqrt{1 + p^2/4} - |p|/2$ and finally decreases from ∞ to 1 as $|\sigma|$ increases indefinitely above $\sqrt{1 + p^2/4} - |p|/2$. Its behavior is indicated in Fig. 4. It should be recalled from Part I that "$\sigma = 0$" is an abbreviation for the very small magnetic field necessary to saturate the ferrite.

A brief examination of the propagation of plane waves shows the more

significant features of transmission in this medium. Since no direction in the x–z plane can be a preferred one, the plane wave may be assumed to travel in the z-direction with variation, $e^{-j\beta z}$. Then, from equations (4) and (5)

$$\beta^2 = \omega^2 \epsilon \mu_{\text{eff}} ,$$

and

$$H_x = \frac{-\beta\mu}{\omega(\mu^2 - \kappa^2)} E_y ,$$

$$H_y = \frac{j\beta\kappa}{\omega(\mu^2 - \kappa^2)} E_y ,$$

Since μ_{eff} is negative between $| \sigma | = 1 - | p |$ and $| \sigma | = \sqrt{1 + p^2/4} - | p |/2$, the medium is cut off for plane waves in this range of magnetic field (at a fixed signal frequency). H is elliptically polarized when the medium is magnetized and $| H_z |/| H_x | = | \kappa |/| \mu | = | \rho_H |$. We may also put

$$H_x + jH_z = \frac{-\beta E_y}{\omega(\mu - \kappa)} ,$$

$$H_x - jH_z = \frac{-\beta E_y}{\omega(\mu + \kappa)} ,$$

But $| H_x + jH_z |$ and $| H_x - jH_z |$ are proportional to the amplitudes of the left-handed and right-handed circularly polarized components of the magnetic field. The medium may thus be considered to exhibit the permeability,

$$\mu - \kappa = \mu_0 \left(1 - \frac{p}{1 - \sigma} \right) ,$$

for left-handed components and the permeability,

$$\mu + \kappa = \mu_0 \left(1 + \frac{p}{1 + \sigma} \right)$$

for right-handed components. The effective permeability is essentially a parallel combination of these two permeabilities and the medium may propagate ($\mu_{\text{eff}} > 0$) even when $\mu - \kappa$ is negative, as will happen for $\sqrt{1 + p^2/4} - | p |/2 < \sigma < 1$.

Since the medium itself has no non-reciprocal properties it is clear that if the latter are to arise they must do so as a result of interaction between the medium and its surroundings. The boundaries of the ferrite

at which matching will be necessary are surfaces parallel to the y-axis and here we need an expression for H_{tang}/E_y where H_{tang} is a tangential magnetic field at the surface. For the moment we will consider the admittance looking into the ferrite and take tangential components in the counterclockwise sense. From equation (4), then,

$$\frac{H_{\text{tang}}}{E_y} = \frac{j\mu \dfrac{1}{E_y}\dfrac{\partial E_y}{\partial \nu}}{\omega(\mu^2 - \kappa^2)} - \frac{\kappa \dfrac{1}{E_y}\dfrac{\partial E_y}{\partial \sigma}}{\omega(\mu^2 - \kappa^2)}, \tag{7}$$

where $\partial/(\partial\nu)$ is a normal derivative (outward) and $\partial/(\partial\sigma)$, a tangential derivative. It is possible, although by no means essential, to interpret the terms of (7) in the following fashion: the first term is just the admittance of a normal TE mode propagating in the interior of the ferrite (which is to have the permeability, μ_{eff}); the second term is to be ascribed to an independent surface current,

$$\frac{-\kappa \dfrac{\partial E_y}{\partial \sigma}}{\omega(\mu^2 - \kappa^2)}.$$

Using this picture one may see how non-reciprocity arises in a simple case. If the ferrite be bounded by the planes $x = x_1$ and $x = x_2$, and the z-variation is of the form $e^{-j\beta z}$, the admittances due to the surface currents are $+j\kappa\beta/\omega(\mu^2 - \kappa^2)$. If β reverses its sign the surface currents are interchanged. If now the external admittances on the two sides of the slab are unequal (and, of course, themselves reciprocal) for given values of ω and $|\beta|$, there is no reason why β and $-\beta$ should simultaneously solve the matching problem.

Almost all the above considerations may be taken over to the case of the plasma. Here the TE fields will be undisturbed by the magnetic field (but the dielectric constant is altered by the presence of the charge from its free space value). Equations (4) and (5) are now replaced by

$$j\omega(\epsilon^2 - \eta^2)E_x = -\epsilon \frac{\partial H_y}{\partial z} + j\eta \frac{\partial H_y}{\partial x}, \tag{8a}$$

$$j\omega(\epsilon^2 - \eta^2)E_z = j\eta \frac{\partial H_y}{\partial z} + \epsilon \frac{\partial H_y}{\partial x}, \tag{8b}$$

$$\frac{\partial^2 H_y}{\partial x^2} + \frac{\partial^2 H_y}{\partial z^2} + \omega^2\mu_0 \frac{\epsilon^2 - \eta^2}{\epsilon} H_y = 0 \tag{9}$$

for the TM fields. Here ϵ and η are the diagonal and off-diagonal terms of the dielectric matrix which is of the same form

$$\begin{Vmatrix} \epsilon & 0 & -j\eta \\ 0 & \epsilon_1 & 0 \\ j\eta & 0 & \epsilon \end{Vmatrix}$$

as the permeability matrix of the ferrite. The equations of motion for the plasma* lead to the expressions

$$\begin{aligned} \epsilon_1 &= \epsilon_0 (1 - q^2), \\ \epsilon &= \epsilon_0 \left(1 + \frac{q^2}{\sigma^2 - 1}\right), \\ \eta &= \epsilon_0 \frac{q^2\sigma}{\sigma^2 - 1}, \\ \epsilon_{\text{eff}} &= \frac{\epsilon^2 - \eta^2}{\epsilon} = \epsilon_0 \frac{(1 - q^2)^2 - \sigma^2}{(1 - q^2) - \sigma^2}, \end{aligned} \tag{10}$$

where σ is now the ratio of the cyclotron resonance frequency;

$$\frac{1}{2\pi} \frac{|e|\mu_0}{m} H_0,$$

in a field H_0, to the applied frequency and q is the ratio of the plasma frequency to applied frequency; ϵ_{eff} behaves with magnetic field in much the same way as μ_{eff}. It is a constantly decreasing function of σ and is negative between $|\sigma| = 1 - q^2$ and $|\sigma| = \sqrt{1 - q^2}$, going to infinity at the latter value. The left and right handed dielectric constants, $\epsilon - \eta$ and $\epsilon + \eta$ are given by $\epsilon_0[1 - q^2/(1 - \sigma)]$ and $\epsilon_0[1 - q^2/(1 + \sigma)]$.

2.2. *Examples of Non-reciprocal Systems*

We now discuss briefly three examples of non-reciprocal systems as illustrations of particular points. As an example of a system which can be analyzed very easily and completely we consider a plasma occupying the region, $x > 0$ and bounded at $x = 0$ by a sheet of constant impedance. This impedance is to depend upon frequency but not on the propagation constant. It will be written as $j\sqrt{\mu_0/\epsilon_0}Z(\omega)$ where μ_0 and ϵ_0 are free space values. A practical realization of such a sheet might consist of a very large number of similar fins of negligible thickness and separation, parallel to the y-axis and attached normally to a conducting plane $x =$ constant. The fields between separate fins are uncoupled and E_z is uniform between fins. For such an arrangement, $Z(\omega) = \tan \omega\sqrt{\epsilon_0\mu'_0}x_0$,

* See Section 2.2 of Part I.

where x_0 is the depth of the fins. If the z-variation of the fields is $e^{-j\beta z}$ and the waves are guided, the x-dependence in the plasma must be as exp $-\sqrt{\beta^2 - \omega^2\mu_0\epsilon_{\text{eff}}}\, x$. From equation (8b)

$$j\omega(\epsilon^2 - \eta^2)E_z = [\eta\beta - \epsilon\sqrt{\beta^2 - \omega^2\mu_0\epsilon_{\text{eff}}}\,]H_y .$$

Matching at $x = 0$ gives

$$\frac{\eta\beta - \epsilon\sqrt{\beta^2 - \omega^2\mu_0\epsilon_{\text{eff}}}}{j\omega(\epsilon^2 - \eta^2)} = j\sqrt{\frac{\mu_0}{\epsilon_0}}\, Z(\omega).$$

This yields

$$\left[\beta - \eta\omega\sqrt{\frac{\mu_0}{\epsilon_0}}\, Z(\omega)\right]^2 = \omega^2\mu_0\epsilon + \omega^2\frac{\mu_0}{\epsilon_0}\epsilon^2 Z^2(\omega)$$

or

$$\frac{\beta}{\omega\sqrt{\mu_0\epsilon_0}} = \frac{\eta}{\epsilon_0}\, Z(\omega) \pm \sqrt{\frac{\epsilon}{\epsilon_0} + \frac{\epsilon^2}{\epsilon_0^2}\, Z^2(\omega)}\,. \tag{11}$$

The non-reciprocity is clearly exhibited, since the two values of β are not equal and opposite. The solution (11) is valid only if $\beta^2 > \omega^2\mu_0\epsilon_{\text{eff}}$, corresponding to guided waves.

In the second example we assume that the region between conducting planes at $x = 0$ and $x = x_0$ is filled by a plasma. When no magnetic field is present E_z is supposed to vanish and E_x is uniform across the gap. The unperturbed field is then plane polarized (TEM). The magnetic field is now applied parallel to the y axis to that part of the gap between $x = 0$ and $x = x_1$. E_z in the magnetized plasma is now given by

$$E_z = E_0 \sin\sqrt{\omega^2\mu_0\epsilon_{\text{eff}} - \beta^2}\, x$$

since it must vanish at $x = 0$. The z variation is again exp $-j\beta z$. H_y may be found from equation (8b) and is

$$H_y = \frac{j\omega E_0}{\omega^2\mu_0\epsilon - \beta^2}\,[\eta\beta \sin\sqrt{\omega^2\mu_0\epsilon_{\text{eff}} - \beta^2}\, x - \epsilon\sqrt{\omega^2\mu_0\epsilon_{\text{eff}} - \beta^2}\cos\sqrt{\omega^2\mu_0\epsilon_{\text{eff}} - \beta^2}\, x].$$

The admittance of the magnetized section at $x = x_1$, is thus,

$$\frac{H_y}{E_z} = \frac{j\omega}{\omega^2\mu_0\epsilon - \beta^2}\,[\eta\beta - \epsilon\sqrt{\omega^2\mu_0\epsilon_{\text{eff}} - \beta^2}\cot\sqrt{\omega^2\mu_0\epsilon_{\text{eff}} - \beta^2}\, x_1],$$

and, analogously, that of the unmagnetized part is

$$\frac{H_y}{E_z} = \frac{j\omega}{\omega^2\mu_0\epsilon_1 - \beta^2}\,[-\epsilon_1\sqrt{\omega^2\mu_0\epsilon_1 - \beta^2}\cot\sqrt{\omega^2\mu_0\epsilon_1 - \beta^2}\,(x_0 - x_1)],$$

where $\epsilon_1 = \epsilon_0(1 - q^2)$. Suppose now that the applied field is weak, so that $\omega^2\mu_0\epsilon \sim \omega^2\mu_0\epsilon_1 \sim \beta^2$. Then, the cotangents may be expanded and by equating admittances one obtains

$$\frac{j\omega}{\omega^2\mu_0\epsilon - \beta^2}\left[\eta\beta - \frac{\epsilon}{x_1}\right] = \frac{j\omega\epsilon_1}{(\omega^2\mu_0\epsilon_1 - \beta^2)(x_0 - x_1)},$$

or

$$\frac{\beta^2}{\omega^2\mu_0\epsilon_1} = 1 + \frac{\epsilon - \epsilon_1}{-\epsilon_1 + \left(\eta\beta - \frac{\epsilon}{x_1}\right)(x_0 - x_1)}. \tag{12}$$

Since $\epsilon - \epsilon_1$ is of the second order in σ this equation may be solved by substituting the unperturbed value of β, $\pm \omega\sqrt{\mu_0\epsilon_1}$, in the right hand side. It is clear that although the system is non-reciprocal, it will be so only to third order in σ. This system, therefore, illustrates the fact pointed out in Section 1.1 that even when the fields are plane polarized in the absence of a dc magnetic field, non-reciprocity may arise, although it may be very small in weak fields.

The third example to be considered is one which has been referred to in Section 1.1, namely that in which a strip of ferrite is placed across the short dimension of rectangular wave guide, see Fig. 1(a). In view of the fact that this problem has been discussed with great thoroughness by Lax, Button and Roth, we shall, after deriving the characteristic equation, consider only the case of a very thin strip. Let the thickness of the strip be $2x_0$ and the distance from its two faces to the nearest guide wall be x_1 and x_2 respectively. The admittances at the two faces are then

$$\frac{j}{\omega\mu_0}\cot\sqrt{\beta_0^2 - \beta^2}\,x_1$$

and

$$\frac{-j}{\omega\mu_0}\cot\sqrt{\beta_0^2 - \beta^2}\,x_2$$

respectively, where $\beta_0^2 = \omega^2\mu_0\epsilon_0$. Inside the ferrite

$$\frac{\partial^2}{\partial x^2} \equiv -[\omega^2\epsilon\mu_{\text{eff}} - \beta^2] = -(\beta_f^2 - \beta^2),$$

and

$$-j\omega(\mu^2 - \kappa^2)H_z = \mu\frac{\partial E_y}{\partial x} + \kappa\beta E_y.$$

Thus, immediately within the ferrite,

$$\frac{1}{E_y}\frac{\partial E_y}{\partial x} = -j\omega\mu_{\text{eff}}\left(\frac{H_z}{E_y}\right)_{\text{external}} - \rho_H\beta.$$

If the two faces of the ferrite are $x = -x_0$ and $x = x_0$ we then have

$$\left(\frac{1}{E_y}\frac{\partial E_y}{\partial x}\right)_{x=-x_0} = \frac{\mu_{\text{eff}}}{\mu_0}\cot\sqrt{\beta_0^2 - \beta^2}\,x_1 - \rho_H\beta = A,$$

and

$$\left(\frac{1}{E_y}\frac{\partial E_y}{\partial x}\right)_{x=x_0} = -\frac{\mu_{\text{eff}}}{\mu_0}\cot\sqrt{\beta_0^2 - \beta^2}\,x_2 - \rho_H\beta = B.$$

If we write

$$\frac{1}{E_y}\frac{\partial E_y}{\partial x} = \sqrt{\beta_f^2 - \beta^2}\tan\left(\sqrt{\beta_f^2 - \beta^2}\,x\right),$$

and make use of the boundary conditions we obtain

$$\tan 2\sqrt{\beta_f^2 - \beta^2}\,x_0 = \frac{\sqrt{\beta_f^2 - \beta^2}\,(A - B)}{\beta_f^2 - \beta^2 - AB}. \tag{13}$$

The non-reciprocity is clearly contained in the odd power of β in A and B.

For small thickness we replace the tangent by its argument, and, substituting for A and B on one side of the equation, obtain

$$\frac{\mu_{\text{eff}}}{\mu_0}\left[\cot\sqrt{\beta_0^2 - \beta^2}\,x_1 + \cot\sqrt{\beta_0^2 - \beta^2}\,x_2\right] = 2x_0[\beta_f^2 - \beta^2 - AB],$$

or

$$\sin\sqrt{\beta_0^2 - \beta^2}\,(x_1 + x_2) = 2x_0\frac{\mu_0}{\mu_{\text{eff}}}\sin\sqrt{\beta_0^2 - \beta^2}\,x_1\sin\sqrt{\beta_0^2 - \beta^2}\,x_2[\beta_f^2 - \beta^2 - AB]. \tag{14}$$

Since the guide is almost empty, we may write

$$\sqrt{\beta_0^2 - \beta^2}(x_1 + x_2) = \pi - \delta,$$

where δ is small. Or,

$$\beta = \beta_1 + \frac{2\pi\delta}{\beta_1(x_1 + x_2)^2},$$

where

$$\beta_1^2 = \beta_0^2 - \pi^2/(x_1 + x_2)^2.$$

Writing $\beta = \beta_1$ in the right hand side of equation (14) and noting that if

$$\sqrt{\beta_0^2 - \beta_1^2}\, x_1 = \frac{\pi x_1}{x_1 + x_2} = \theta$$

then $\sqrt{\beta_0^2 - \beta_1^2}\, x_2 = \pi - \theta$, we have

$$\begin{aligned}\delta &= 2x_0 \frac{\mu_0}{\mu_{\text{eff}}} \sin^2\theta \left[\beta_f^2 - \beta_1^2 - \left(\frac{\mu_{\text{eff}}}{\mu_0}\cot\theta - \rho_H\beta_1\right)\left(\frac{\mu_{\text{eff}}}{\mu_0}\cot\theta - \rho_H\beta_1\right)\right], \\ &= 2x_0 \frac{\mu_0}{\mu_{\text{eff}}} \left[\left(\beta_f^2 - (1 + \rho_H^2)\beta_1^2\right)\sin^2\theta - \left(\frac{\mu_{\text{eff}}}{\mu_0}\right)^2 \cos^2\theta \right. \\ &\qquad \left. + 2\frac{\mu_{\text{eff}}}{\mu_0}\rho_H\beta_1 \cos\theta \sin\theta\right]. \end{aligned} \tag{15}$$

The non-reciprocal part of β is thus $4\pi x_0(x_1 + x_2)^{-2}\rho_H \sin 2\theta$. This has a maximum value for $\theta = \pi/4$ or $3\pi/4$ and, hence, $x_1 = (x_1 + x_2)/4$ or $3(x_1 + x_2)/4$. This result may be understood qualitatively by considering the fields in the guide before the ferrite is inserted. We then have $E_y = E_0 \sin(\pi x/a)$ where $a = x_1 + x_2$ and consequently,

$$H_x = -\frac{\beta}{\omega\mu}\sin\frac{\pi x}{a} \qquad \text{and} \qquad H_z = \frac{j\frac{\pi}{a}}{\omega\mu}\cos\frac{\pi x}{a}.$$

The amplitudes of the left- and right-handed components of circular polarization at a point are then proportional to

$$-\beta\sin\frac{\pi x}{a} - \frac{\pi}{a}\cos\frac{\pi x}{a} \qquad \text{and} \qquad -\beta\sin\frac{\pi x}{a} + \frac{\pi}{a}\cos\frac{\pi x}{a}.$$

The difference in the squares of these amplitudes is $2\beta(\pi/a)\sin(\pi x/a)\cos(\pi x/a)$ and this is proportional to the difference between the energy stored at x in the left-handed wave and in the right-handed wave. It is plausible that this should be a measure of the non-reciprocal effect produced by a thin piece of ferrite at x.

2.3. *The Plane Helix*

In dealing with transverse field problems with cylindrical geometry we shall consider non-reciprocal propagation along a helix which is surrounded by circumferentially magnetized ferrite. The analysis of this problem is rather cumbersome and it is advantageous to study first an analogous plane problem. The simplicity thus gained allows us to examine somewhat more complicated problems. The "plane helix,"

to be treated here, is a sheet of negligible thickness, lying in the plane $x = 0$, which conducts only in a single direction making an angle ψ with the y-axis. In this direction it will be supposed lossless. In addition we assume that the regions, $x < 0$ and $0 < x < x_0$ are empty, while the space $x' > x_0$ is filled with ferrite. As usual the magnetic field is along the y-axis and the fields are independent of y. The problem is clearly the limit for very large radius of that of an empty, helically-conducting cylinder, surrounded by an infinitely thick shell of ferrite, circumferentially magnetized, the whole system carrying fields with no angular variation.

We first consider the boundary conditions for the plane helix, after noting that it is evident that both TE and TM fields will be required. The tangential electric field on either side of the sheet must necessarily be at right angles to the direction of conduction since the conductivity is infinite. Further, the tangential electric field must be continuous through the sheet. Hence, if the field normal to the direction of conduction is E_0 (omitting here and elsewhere the factor $e^{-j\beta z}$), we have

$$E_z^+ = E_z^- = E_0 \cos \psi,$$
$$E_y^+ = E_y^- = -E_0 \sin \psi,$$

where the symbols $+$ and $-$ refer to $x > 0$ and $x < 0$ respectively. Again, since current cannot flow normal to the direction of conduction, the tangential magnetic field along the latter must be continuous through the sheet or

$$(H_z^+ - H_z^-) \sin \psi + (H_y^+ - H_y^-) \cos \psi = 0.$$

The boundary conditions may be combined into a single equation, by introducing admittances, in the form

$$\left(\frac{H_z^+}{E_y^+} - \frac{H_z^-}{E_y^-}\right) \sin^2 \psi = \left(\frac{H_y^+}{E_z^+} - \frac{H_y^-}{E_z^-}\right) \cos^2 \psi. \tag{16}$$

The left-hand side refers to the TE fields and the right to TM fields. In the empty regions surrounding the sheet

$$\left[\frac{\partial^2}{\partial x^2} - (\beta^2 - \omega^2 \epsilon_0 \mu_0)\right] H_y = \left[\frac{\partial^2}{\partial x^2} - (\beta^2 - \beta_0^2)\right] H_y = 0,$$

$$E_z = \frac{1}{j\omega\epsilon_0} \frac{\partial H_y}{\partial x},$$

and

$$\left[\frac{\partial^2}{\partial x^2} - (\beta^2 - \beta_0^2)\right] E_y = 0, \qquad H_z = \frac{-1}{j\omega\mu_0}\frac{\partial E_y}{\partial x},$$

with $\beta_0^2 = \omega^2\epsilon_0\mu_0$. If the waves are to be guided, $\beta^2 > \beta_0^2$, and, then, for $x < 0$, we have $\partial/\partial x \equiv \sqrt{\beta^2 - \beta_0^2}$. Thus

$$\frac{H_y^-}{E_z^-} = \frac{j\omega\epsilon_0}{\sqrt{\beta^2 - \beta_0^2}},$$

and

$$\frac{H_z^-}{E_y^-} = -\frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\mu_0}.$$

If the admittances at the surface of the ferrite are H_y^f/E_z^f and H_z^f/E_y^f, then H_y^+/E_z^+ and H_z^+/E_y^+ are given by the impedance transformation:

$$\frac{H_y^+}{E_z^+} = \frac{j\omega\epsilon_0}{\sqrt{\beta^2 - \beta_0^2}} \cdot \frac{\frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\epsilon_0}\frac{H_y^f}{E_z^f} - \tanh\sqrt{\beta^2 - \beta_0^2}\, x_0}{1 - \frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\epsilon_0}\frac{H_y^f}{E_z^f}\cdot\tanh\sqrt{\beta^2 - \beta_0^2}\, x_0},$$

and

$$\frac{H_z^+}{E_y^+} = -\frac{\sqrt{\beta^2 - \beta_0^2}}{j\omega\epsilon_0} \cdot \frac{-\frac{j\omega\mu_0}{\sqrt{\beta^2 - \beta_0^2}}\frac{H_z^f}{E_y^f} - \tanh\sqrt{\beta^2 - \beta_0^2}\, x_0}{1 - \left[\frac{-j\omega\mu_0}{\sqrt{\beta^2 - \beta_0^2}}\frac{H_z^f}{E_y^f}\right]\cdot\tanh\sqrt{\beta^2 - \beta_0^2}\, x_0}.$$

Within the ferrite the TM-fields fall off as $\exp -\sqrt{\beta^2 - \omega^2\epsilon\mu_0}x$ and the TE-fields as $\exp -\sqrt{\beta^2 - \omega^2\epsilon\mu_{\text{eff}}}x$. We then have

$$\frac{H_y^f}{E_z^f} = \frac{-j\omega\epsilon}{\sqrt{\beta^2 - \omega^2\epsilon\mu_0}},$$

and

$$\frac{H_z^f}{E_y^f} = \frac{j\omega\mu_{\text{eff}}}{\sqrt{\beta^2 - \omega^2\epsilon\mu_{\text{eff}}} - \beta\kappa/\mu}.$$

These results may now be collected and substituted in equations (16). The equation of condition so obtained is

$$(\beta^2 - \beta_0^2)\frac{A + 1}{A + \tanh\sqrt{\beta^2 - \beta_0^2}x_0} = \beta_0^2\cot^2\psi\,\frac{1 - B}{1 - B\tanh\sqrt{\beta^2 - \beta_0^2}x_0}, \qquad (17)$$

where

$$A = \frac{\mu_{\text{eff}}}{\mu_0} \frac{\sqrt{\beta^2 - \beta_0^2}}{\sqrt{\beta^2 - \omega^2\epsilon\mu_{\text{eff}}} - \beta\rho_H},$$

and

$$B = -\frac{\epsilon}{\epsilon_0} \frac{\sqrt{\beta^2 - \beta_0^2}}{\sqrt{\beta^2 - \omega^2\epsilon\mu_0}}.$$

We shall assume that we are dealing with slow waves (β large). This is the case of greatest practical interest and is usually ensured by making $\cot \psi$ large. Assuming that the waves are slow we simplify the equation (12) and find certain values for β. We may then ascertain in what ranges of σ and p the simplifying assumptions and the results are consistent.

In equation (17), then, we replace all square roots by $| \beta |$ and $\beta^2 - \beta_0^2$ by β^2, obtaining

$$\beta^2 = \beta_0^2 \cot^2 \psi \frac{A + \tanh | \beta | x_0}{A + 1} \cdot \frac{1 - B}{1 - B \tanh | \beta | x_0},$$

with

$$A = \frac{\mu_{\text{eff}}/\mu_0}{1 - \rho_H \operatorname{sgn} \beta} = \frac{\mu + \kappa \operatorname{sgn} \beta}{\mu_0},$$

and

$$B = -\epsilon/\epsilon_0$$

where $\operatorname{sgn} \beta = 1$ for $\beta > 0$ and $\operatorname{sgn} \beta = -1$ for $\beta < 0$. Taking first the simplest case of no separation ($x_0 = 0$), this becomes

$$\beta^2 = \frac{\beta_0^2(1 + \epsilon/\epsilon_0) \cot^2 \psi}{1 + \dfrac{1 - \rho_H \operatorname{sgn} \beta}{\mu_{\text{eff}}/\mu_0}} = \frac{\beta_0^2(1 + \epsilon/\epsilon_0) \cot^2 \psi}{1 + \dfrac{\mu_0}{\mu + \kappa \operatorname{sgn} \beta}}.$$

Substituting the expressions (6) and (1) for μ_{eff}/μ_0 and ρ_H we are led to*

$$\beta_+ = \beta_0 \cot \psi \cdot \sqrt{\frac{1}{2}(1 + \epsilon/\epsilon_0)} \sqrt{\frac{\sigma + 1 + p}{\sigma + 1 + p/2}}, \tag{18a}$$

$$\beta_- = -\beta_0 \cot \psi \cdot \sqrt{\frac{1}{2}(1 + \epsilon/\epsilon_0)} \sqrt{\frac{\sigma - 1 + p}{\sigma - 1 + p/2}} \tag{18b}$$

* Since reversal of the magnetization is equivalent to interchange of the propagation directions, we are at liberty to consider σ and p always positive, and to deal with the two cases $\beta > 0$ and $\beta < 0$ separately.

The β_+ mode propagates for all σ, $\frac{\sigma + 1 + p}{\sigma + 1 + p/2}$ declining steadily from $\frac{1 + p}{1 + p/2}$ to unity. The β_- mode on the other hand is cut off, with $\beta^2 = 0$ at $\sigma = 1 - p$ and then reappears with $\beta^2 = \infty$ at $\sigma = 1 - p/2$. The behavior of the two modes is shown in Fig. 5. Self-consistency requires that $\beta^2 \gg \omega^2\epsilon\mu_0$ or $\omega^2\epsilon \mid \mu_{\text{eff}} \mid$, whichever is the greater. The condition $\beta^2 \gg \omega^2\epsilon \mid \mu_{\text{eff}} \mid$ fails to be met near

$$\sigma = \sigma_0 = \sqrt{\frac{p^2}{4} + 1} - \frac{p}{2},$$

when

$$\frac{\mid \sigma - \sigma_0 \mid}{\frac{p}{2}\sqrt{\frac{p^2}{4} + 1} \pm \left(\frac{p}{2} - 1\right)} < \frac{2}{(1 + \epsilon/\epsilon_0)\cot\psi}$$

The condition $\beta^2 \gg \omega^2\epsilon\mu_0$ will fail for β_- near $\sigma = 1 - p$. The range for this to occur is given by

$$1 - p - \sigma < \frac{p/2}{\frac{1}{2}(1 + \epsilon/\epsilon_0)\cot\psi - 1}.$$

The extension of the Polder formulae to the case of a lossy ferrite was given in Part I, (Section 2.1). From the results given there one may write

$$\mu + \kappa \operatorname{sgn}\beta = 1 + p\,\frac{\sigma(1 + \alpha^2) - \operatorname{sgn}\beta + j\alpha}{\sigma^2(1 + \alpha^2) - 1 + 2j\alpha\sigma}$$

where α is a damping parameter. Substitution of these expressions in the slow wave formula for β^2 will give the effect of loss on the propagation constant. In Fig. 6 the complex value of β_- is shown for $\alpha = 0.1$ and several values of p. The imaginary part of β_+ is small and varies only slightly with σ. Fig. 7 shows the initial loss ($\sigma = 0$) for three values of α and a range of p values. From a knowledge of β as a function of $\sigma = \omega_H/\omega$ and $p = \omega_M/\omega$ it is possible to calculate the loss, Im $\beta_-/(\beta_0\sqrt{1 + \epsilon/\epsilon_0}$ cot ψ), as a function of frequency when the magnetic field and saturation magnetization of the ferrite are held constant. Fig. 8 shows the results of such calculations. It should be noted that the horizontal scale is linear in σ or $1/\omega$ and that the vertical scale implicitly contains the frequency in the form of $1/\beta_0$. Both of these distortions of scale tend to

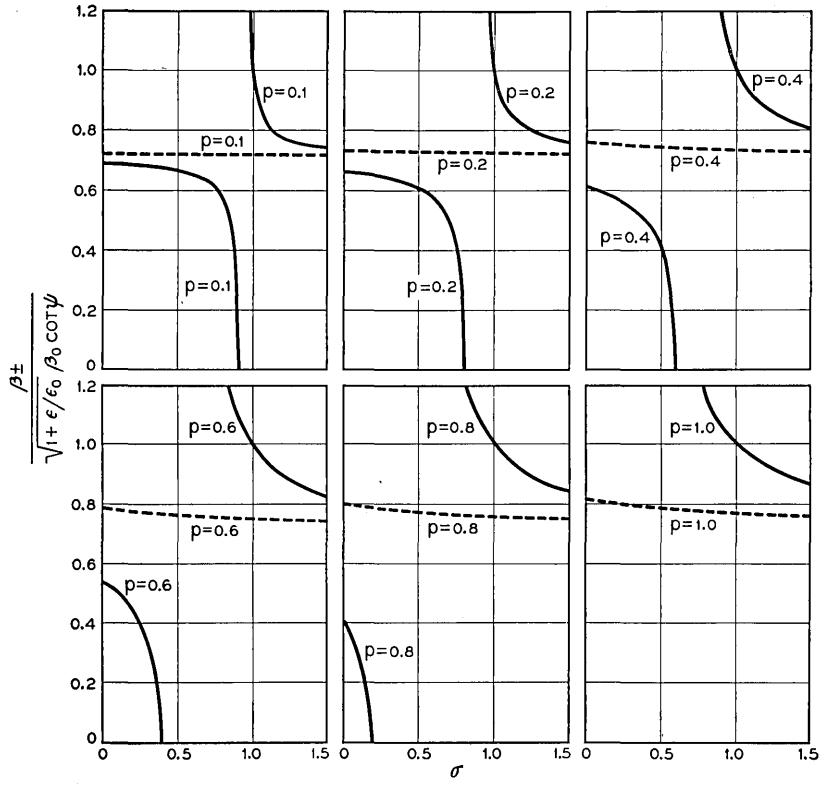


Fig. 5 — The non-reciprocal propagation constants of the flat helix. The dotted curves represent β_+, the solid curves β_-. (Loss free case).

produce an appearance of sharpness in the variation of the loss at higher frequencies.

If the slow wave assumption be made again it is possible to obtain solutions for the case in which the ferrite does not touch the helix and the latter is lossless. With the slow wave approximation, (17) becomes

$$\frac{\beta_\pm{}^2}{\beta_0{}^2 \cot^2 \Psi} = \frac{1 + \dfrac{\epsilon}{\epsilon_0}}{1 + \dfrac{\epsilon}{\epsilon_0} \tanh \mid \beta \mid x_0} \cdot \frac{\dfrac{\mu + \kappa \operatorname{sgn} \beta}{\mu_0} + \tanh \mid \beta \mid x_0}{\dfrac{\mu + \kappa \operatorname{sgn} \beta}{\mu_0} + 1}$$

If we write $\mid \beta \mid x_0 = u$, then the above equation expresses $\beta_\pm$ in terms of u. At the same time $x_0 = u/\mid \beta_\pm(u) \mid$ and we evidently have a parametric representation of $\beta_\pm$ and x_0. The results of such computations

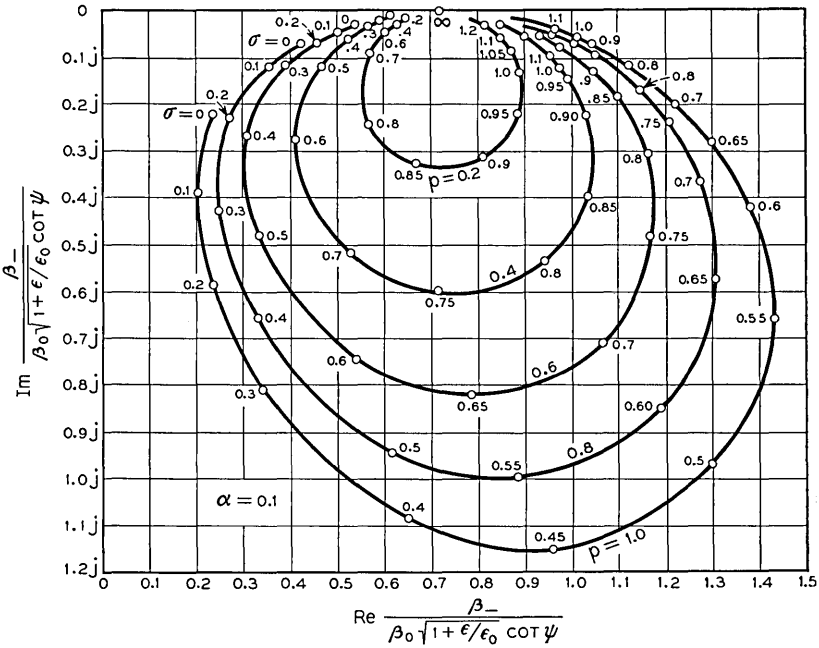


Fig. 6 — Real and imaginary parts of the reverse propagation constant β_- of a flat helix versus σ for various p and for $\alpha = 0.1$ (the parameter σ is marked along the curves). The forward propagation constant has a very small imaginary part which hardly varies with σ.

are shown in Figs. 9(a), (b) and (c), where β is plotted against x_0 for various fixed σ for two values of p. It is to be noted that the introduction of any characteristic length or scale into the problem, such as is provided here by the distance x_0 immediately produces a great complication in the mode spectrum. The plane helix with the ferrite in contact may be thought of as a highly degenerate problem.

To carry out loss calculations using the appropriate expressions for $\mu + \kappa$ sgn β would be very tedious in the separated case. However, it was pointed out in Part 1 that to order α the expressions for μ and κ are given correctly if we put $\sigma + j\alpha$ in place of σ in the lossless formulae and that, in consequence, for small α, the imaginary part of the propagation constant is approximately given by

$$\alpha \frac{\partial \beta}{\partial \sigma}$$

In Figs. 10(a) and 10(b) the loss calculated in this way for the cases considered in Figs. 9 is shown.

3. CYLINDRICAL GEOMETRY

3.1. *Impedances*

In this section we consider systems with cylindrical symmetry about the propagation direction. All boundaries, those of the circuit as those of the medium, are coaxial circular cylinders, and the medium is assumed to be magnetized circumferentially. The practical means for bringing about such a magnetization — for example, thin wires threaded through a cylinder of ferrite and carrying a dc current, Fig. 1(b) — are assumed to effect the electromagnetic field to a negligible extent. As in the case of planar geometry, we restrict ourselves to fields which have no variation along the magnetizing field; that is, in the azimuthal direction. Only the ferrite is considered here; the results for a plasma are obvious corollaries. The magnetizing field and the dc magnetization

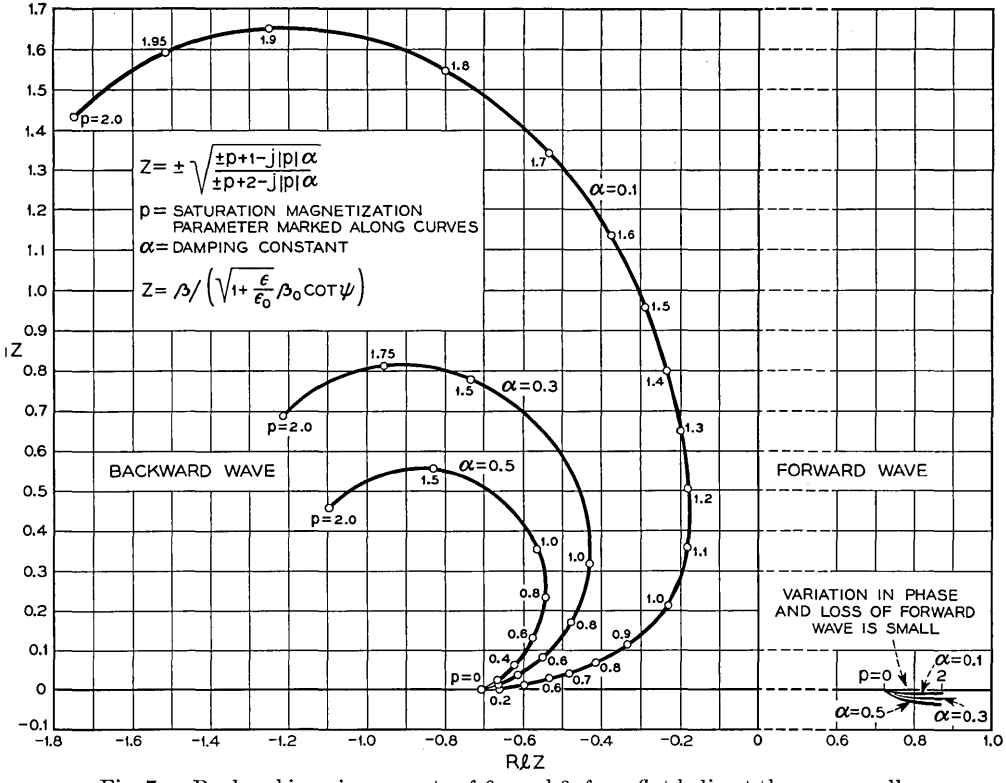


Fig. 7 — Real and imaginary parts of β_+ and β_- for a flat helix at the very small magnetizing field required to saturate the ferrite.

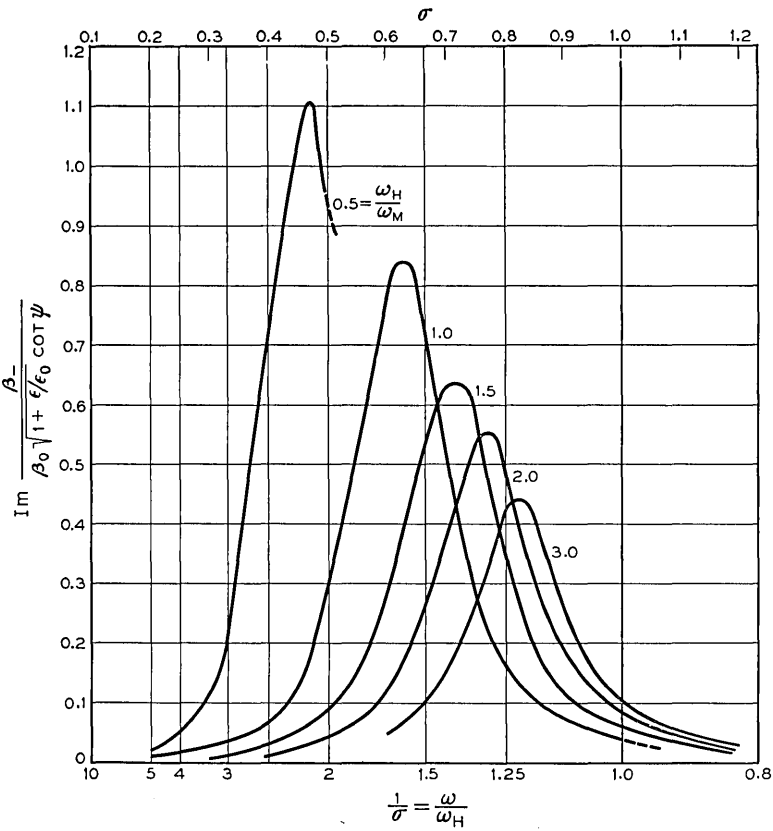


Fig. 8 — Attenuation of reverse wave versus frequency at a fixed magnetic field and saturation magnetization for various values of $\omega_H/\omega_M = (\mu_0 H)/M$. The curves for 0.5 and 1 are discontinued when p reaches unity.

are supposed to be independent of radial distance from the cylinder axis. For the geometries employed in practice, this will be a reasonably good assumption. Thus it is possible to relate the components of B and H in cylindrical coordinates (r, φ, z) through the tensor

$$\begin{Vmatrix} \mu & 0 & -j\kappa \\ 0 & \mu_0 & 0 \\ j\kappa & 0 & \mu \end{Vmatrix},$$

where μ and κ are given by the Polder relations (1) and are independent of position.

Written in cylindrical coordinates, Maxwell's equations in the ferrite are therefore

$$j\beta H_\varphi = j\omega\epsilon_1 E_r\,, \tag{20a}$$

$$-j\beta H_r - \frac{\partial H_z}{\partial r} = j\omega\epsilon_1 E_\varphi\,, \tag{20b}$$

$$\frac{1}{r}\frac{\partial}{\partial r} rH_\varphi = j\omega\epsilon_1 E_z\,, \tag{20c}$$

$$j\beta E_\varphi = -j\omega\mu(H_r - j\rho_H H_z), \tag{20d}$$

$$-j\beta E_r - \frac{\partial E_z}{\partial r} = -j\omega\mu_0 H_\varphi\,, \tag{20e}$$

$$\frac{1}{r}\frac{\partial}{\partial r} rE_\varphi = -j\omega\mu(j\rho_H H_r + H_z)\,. \tag{20f}$$

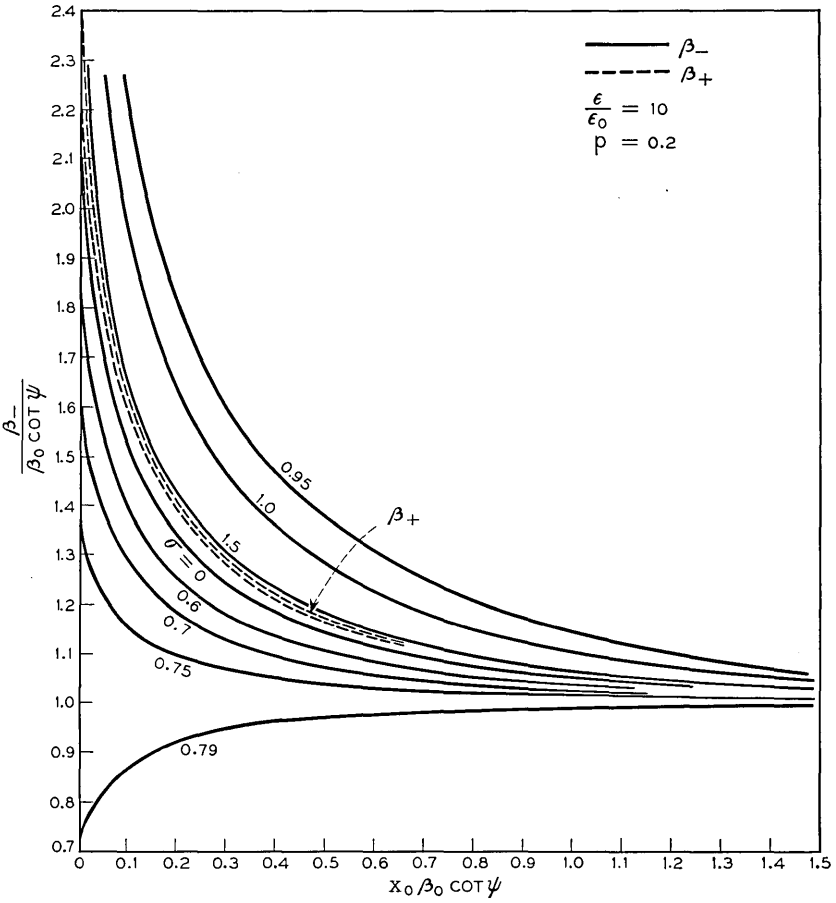


Fig. 9 — Propagation constants for a flat helix separated from the ferrite by a distance x_0 for various values of σ. (Loss-free case) (a) β_- and β_+ for $p = 0.2$, (b) β_- for $p = 0.8$, (c) β_+ for $p = 0.8$. In (a), above, the dotted lines bound all σ values.

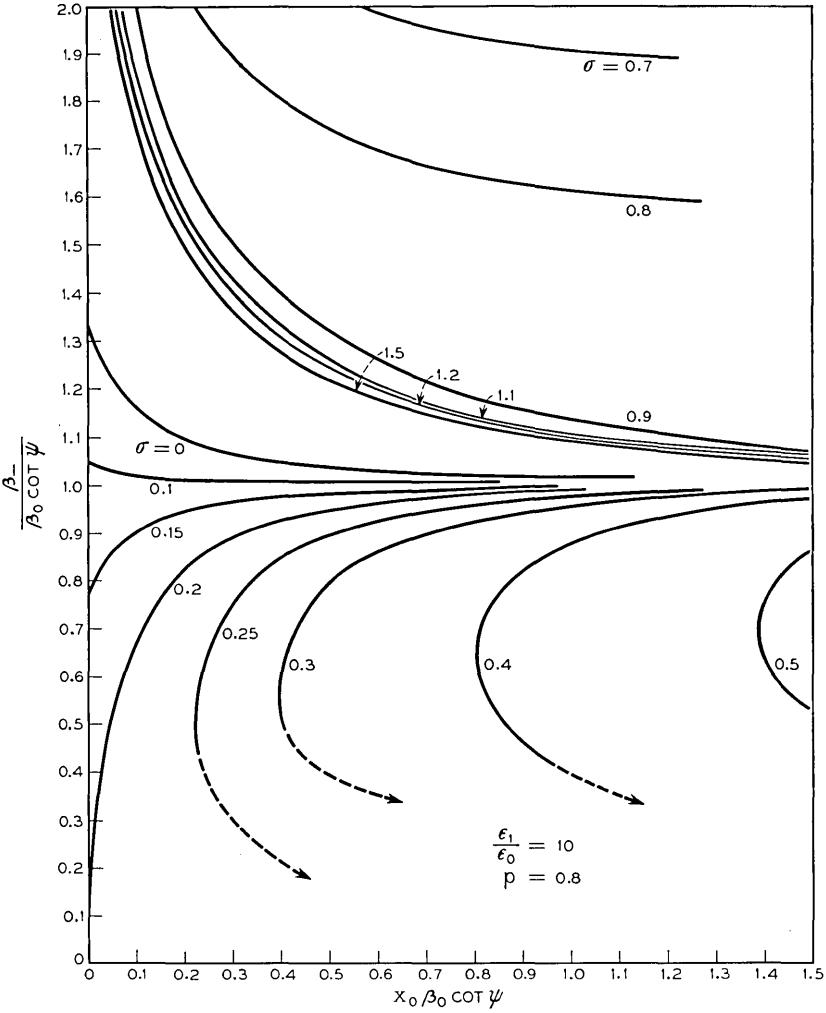


Fig. 9(b) — See Fig. 9.

As in the case of planar geometry, the field-components can be grouped into TE and TM parts; only the TE part will depend explicitly on the gyromagnetic properties of the medium. Equations (20a, 20c) and (20e) determine the TM field. E_r and H_φ can be eliminated from them, yielding the familiar wave equation

$$\frac{1}{r}\frac{\partial}{\partial r}\, r\, \frac{\partial E_z}{\partial r} + (\beta_1^{\,2} - \beta^2)E_z = 0; \qquad \beta_1^{\,2} = \omega^2\mu_0\epsilon_1\,, \tag{21}$$

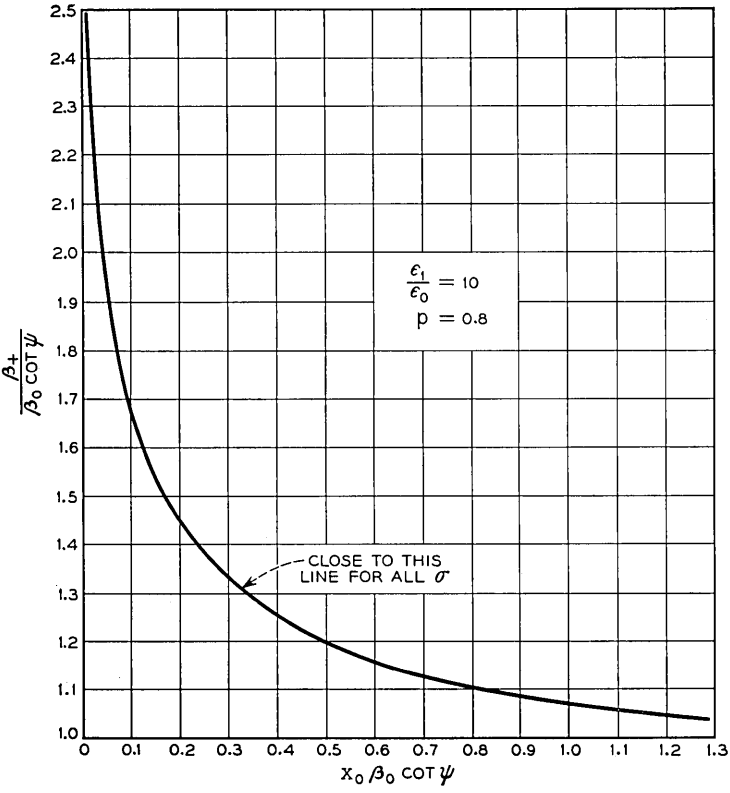


Fig. 9(c) — See Fig. 9.

whose solutions are zero order Bessel functions (or linear combinations thereof) of a kind depending on the region under consideration. Thus if $\beta > \beta_1$, and the region occupied by the medium includes the cylinder axis, the modified Bessel function $I_0\,(r\sqrt{\beta^2 - \beta_1^2})$ has to be chosen if the field is to be finite at $r = 0$; if the medium extends from a finite r to infinity, the function $K_0\,(r\sqrt{\beta^2 - \beta_1^2})$, regular at infinity, is selected. Correspondingly, if $\beta < \beta_1$, the Bessel and Hankel function $J_0\,(r\sqrt{\beta_1^2 - \beta^2})$ and $H_0^{1,\,2}\,(r\sqrt{\beta_1^2 - \beta^2})$ replace I_0 and K_0 respectively.

In terms of the appropriate solution of (21), the remaining field components are

$$E_r = -j\,\frac{\beta}{\beta_1^2 - \beta^2}\,\frac{\partial E_z}{\partial r}\,; \qquad H_\varphi = -\frac{j\omega\epsilon_1}{\beta_1^2 - \beta^2}\,\frac{\partial E_z}{\partial r}\,.$$

The tangential admittance for the TM field is thus

$$Y_{\mathrm{TM}} = \frac{H_\varphi}{E_z} = -\frac{j\omega\epsilon_1}{\beta_1^2 - \beta^2}\,\frac{\partial}{\partial r}\log E_z\,. \tag{22}$$

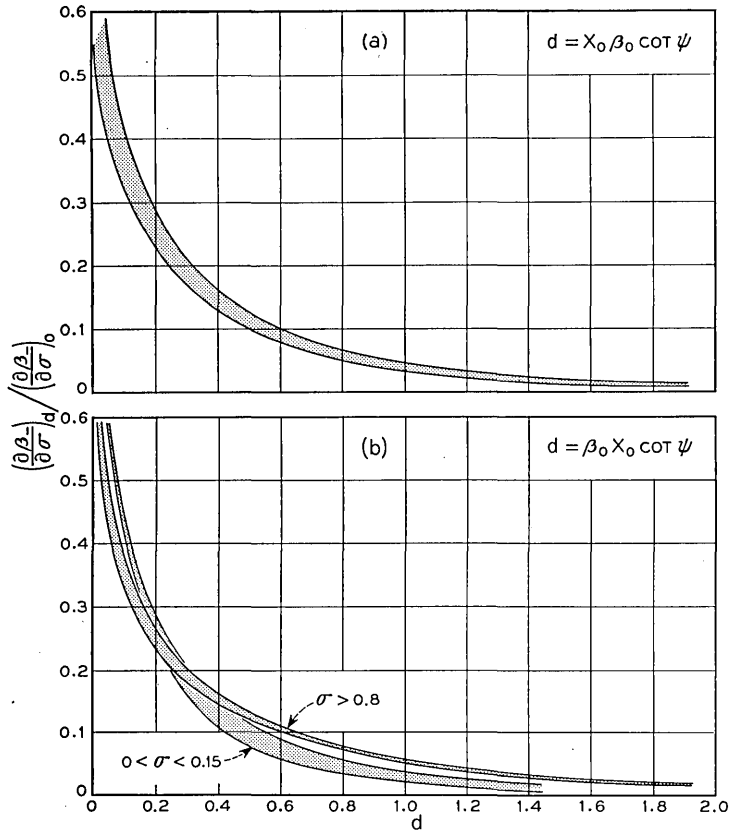


Fig. 10 — Ratios of reverse attenuation for a flat helix separated from the ferrite by a distance x_0 to the reverse attenuation at $x_0 = 0$. Computations made from the approximate slope-formula for the loss, where applicable. (a) $p = 0.2$. (b) $p = 0.8$. In (a) all applicable σ values lie in the shaded region.

For example, if $\beta > \beta_1$, and the medium occupies all space from a finite r to infinity

$$Y_{TM} = \frac{j\omega\epsilon_1}{\alpha_1}\frac{K'_0(\alpha_1 r)}{K_0(\alpha_1 r)} = \frac{-j\omega\epsilon_1 K_1(\alpha_1 r)}{\alpha_1 K_0(\alpha_1 r)}, \tag{23}$$

where $\alpha_1 = \sqrt{\beta^2 - \beta_1^2}$.

The field components of the TE field are determined by equations (20b), (20d) and (20f). Elimination of E_φ and H_r from these gives

$$\frac{1}{r}\frac{\partial}{\partial r} r \frac{\partial H_z}{\partial r} + \left(\beta_f^2 - \beta^2 - \frac{\rho_H \beta}{r}\right) H_z = 0, \tag{24}$$

where $\beta_f^2 = \omega^2\mu\epsilon_1 (1 - \rho_H^2)$. The term $\beta_f^2 - \beta^2$ in the bracket is already

familiar from the planar case; it depends on the magnetic field and on the propagation direction in a purely reciprocal way. The term $\rho_H\beta/r$, however, reverses sign when either magnetizing field or propagation constant changes sign. The solutions of equation (24) are therefore different for opposite propagation directions (or opposite magnetizations). Thus, in contrast to the planar case, where non-reciprocity arose only through the boundary conditions, the cylindrical case is inherently non-reciprocal.

In the absence of the last term in the bracket, equation (24) would be solved by zero order Bessel functions, just like equation (21). In the presence of this term, the solutions become confluent hypergeometric functions. Different changes of variable bring these functions into forms known by different names and notations. One such change leads to Laguerre functions, another to Whittaker functions. We shall choose the latter representation, since it is closely related to Bessel functions, and numerical tables seem about equally scarce for all representations. In equation (24), let $\beta^2 - \beta_f^2 = \alpha_2^2$, and let $\alpha_2 r = y$. Then it becomes

$$\frac{1}{y}\frac{d}{dy}\, y \frac{dH_z}{dy} - \left(1 - \frac{2\chi}{y}\right) H_z = 0, \tag{25}$$

where $\chi = -\beta\rho_H/2\alpha_2$. Further, let $y = x/2$, and $H_z(y) = h(x)/\sqrt{x}$; then equation (25) becomes

$$\frac{d^2h}{dx^2} + \left(\frac{1}{4x^2} + \frac{\chi}{x} - \frac{1}{4}\right) h = 0, \tag{26}$$

which is the standard form of the equation for zero order Whittaker functions. It is a special case of the equation for μ^{th} order Whittaker functions:

$$\frac{d^2h}{dx^2} + \left(\frac{\frac{1}{4} - \mu^2}{x^2} + \frac{\chi}{x} - \frac{1}{4}\right) h = 0, \tag{27}$$

The solution of equation (27) which is regular near zero is denoted in the literature by $M_{\chi,\mu}(x)$; it is proportional, in the limit $\chi = 0$, to the Bessel function $I_\mu(x/2)$. The solution of equation (27) regular at infinity is denoted by $W_{\chi,\mu}(x)$, and in the limit $\chi = 0$, is proportional to $K_\mu(x/2)$. The factors of proportionality are found in Appendix I.

In this notation, the solutions for H_z are thus

$$\frac{M_{\chi,0}(2\alpha_2 r)}{\sqrt{2\alpha_2 r}}, \qquad \frac{W_{\chi,0}(2\alpha_2 r)}{\sqrt{2\alpha_2 r}} *.$$

* If $\beta < \beta_2$, both argument and suffix χ are imaginary. These functions are then related to J_0 and H_0 respectively.

Once the appropriate H_z is determined, H_r and E_φ are given by

$$\begin{aligned} H_r &= \frac{j}{\beta^2 - \omega^2\mu\epsilon_1}\left(\beta\frac{\partial H_z}{\partial r} - \omega^2\mu\epsilon_1\rho_H H_z\right), \\ E_\varphi &= -\frac{j\omega\mu}{\beta^2 - \omega^2\mu\epsilon_1}\left(\frac{\partial H_z}{\partial r} - \rho_H\beta H_z\right). \end{aligned} \tag{28}$$

The tangential impedance for TE-fields is thus

$$Z_{\text{TE}} = \frac{E_\varphi}{H_z} = \frac{\omega\mu}{j(\beta^2 - \omega^2\mu\epsilon_1)}\left(\frac{\partial}{\partial r}\log H_z - \rho_H\beta\right). \tag{29}$$

The reader will recall that for isotropic media ($\rho_H = 0$) the numerator of the right hand side of equation (29) can always be expressed as the ratio of first order to zero order Bessel functions by virtue of relations like $K_0'(\chi) = -K_1(\chi)$, $I_0'(\chi) = I_1(\chi)$ and so on. Analogous results hold true in the present case. Suppose, for example, that we are dealing with a geometry such that the correct H_z is

$$H_z = \frac{W_{\chi,0}(2\alpha_2 r)}{\sqrt{2\alpha_2 r}} = R_{\chi 0}(2\alpha_2 r), \qquad \text{say}$$

Then

$$\frac{1}{H_z}\frac{\partial H_z}{\partial r} = 2\alpha_2 R_{\chi 0}'/R_{\chi 0},$$

and

$$Z_{\text{TE}} = \frac{2\alpha_2\omega\mu}{j(\beta^2 - \omega^2\mu\epsilon_1)}\frac{R_{\chi 0}' + \chi R_{\chi 0}}{R_{\chi 0}}.$$

It is shown in Appendix I that

$$R_{\chi 0}' + \chi R_{\chi 0} = (\chi - \tfrac{1}{2})\,R_{\chi 1}.$$

Therefore

$$\begin{aligned} Z_{\text{TE}} &= \frac{\omega\mu\alpha_2(2\chi - 1)}{j(\beta^2 - \omega^2\mu\epsilon_1)}\frac{R_{\chi 1}(2\alpha_2 r)}{R_{\chi 0}(2\alpha_2 r)}, \\ &= \frac{\omega\mu\alpha_2(2\chi - 1)}{j(\beta^2 - \omega^2\mu\epsilon_1)}\frac{W_{\chi,1}(2\alpha_2 r)}{W_{\chi,0}(2\alpha_2 r)}. \end{aligned} \tag{30}$$

A similar difference relation shows that if the region is such that

$$H_z = M_{\chi,0}(2\alpha_2 r)/\sqrt{2\alpha_2 r},$$

then

$$Z_{\mathrm{TE}} = \frac{\omega\mu\alpha_2(\tfrac{1}{4} - \chi^2)M_{\chi,1}(2\alpha_2 r)}{j(\beta^2 - \omega^2\mu\epsilon_1)M_{\chi,0}(2\alpha_2 r)} . \tag{31}*$$

In the unmagnetized case equations (30) and (31) reduce to

$$-\frac{\omega\mu_0\alpha_1}{j(\beta^2 - \omega^2\mu_0\epsilon_1)}\frac{K_1(\alpha_1 r)}{K_0(\alpha_1 r)} = -\frac{\omega\mu_0 K_1(\alpha_1 r)}{j\alpha_1 K_0(\alpha_1 r)} , \tag{32}$$

and to

$$\frac{\omega\mu_0}{j\alpha_1}\frac{I_1(\alpha_1 r)}{I_0(\alpha_1 r)} . \tag{33}$$

One might be led to believe that the search for solutions of Maxwell's equations with angular dependence $e^{jn\varphi}$ will lead to n^{th} order Whittaker functions (just as in the isotropic case this dependence leads to n^{th} order Bessel functions). Such is not the case, however. Unless $n = 0$, one is led to two simultaneous second order equations for E_z and H_z, and the character of the problem is changed completely.

3.2. *The Cylindrical Helix*

We are now in a position to derive the characteristic equation for a close wound cylindrical helix and approximated by a helically conducting sheet surrounded by ferrite. We confine the discussion to the case in which the ferrite is in actual contact with the helix, Fig. 11; the case of finite separation discussed for the planar helix (Section 2.3) would be too cumbersome here. Losses are neglected. If they are not excessive, they can be deduced from the curves for the propagation constant in the loss free sample by differentiation, as outlined in Sections 2.1 and 4.15, Part I.

The boundary conditions are just the same as in the planar case. In Section 2.3 they were stated in terms of admittances, and it is only necessary to substitute for these the admittances just derived for cylindrical geometry. Thus for H_y/E_z we substitute Y_{TM}, and for H_z/E_φ we write $Y_{\mathrm{TE}} = 1/Z_{\mathrm{TE}}$.

If superfixes i and e refer to the inside and outside of the helix (in Section 2.3 on the plane helix i and e were denoted by "$-$", and "$+$"), the characteristic equation is

$$\{Y_{\mathrm{TM}}^{(i)} - Y_{\mathrm{TM}}^{(e)}\}_{r=r_0}\cot^2\psi = (Y_{\mathrm{TE}}^{(i)} - Y_{\mathrm{TE}}^{(e)})_{r=r_0} , \tag{34}$$

where r_0 is the radius of the helix.

* The appearance of different factors $(2\chi - 1)$ and $(\tfrac{1}{4} - \chi^2)$ is simply due to the way the functions W, M are normalized in the literature.

For waves bound to the helix, $Y_{TE}^{(e)}$ is to be derived from that solution of (24) which tends to zero ar $r \to \infty$. This solution is $W_{\chi,0}(2\alpha_2 r)/\sqrt{2\alpha_2 r}$, and so $Y_{TE}^{(e)}$ is given by equation (30)

$$Y_{TE}^{(e)} = \frac{1}{Z_{TE}} = \frac{j}{\omega\mu} \frac{\beta^2 - \omega^2\mu\epsilon_1}{\alpha_2(2\chi - 1)} \frac{W_{\chi,0}(2\alpha_2 r)}{W_{\chi,1}(2\alpha_2 r)}.$$

Similarly, from equation (23), we have, for bound waves, with $E_z \sim K_0$,

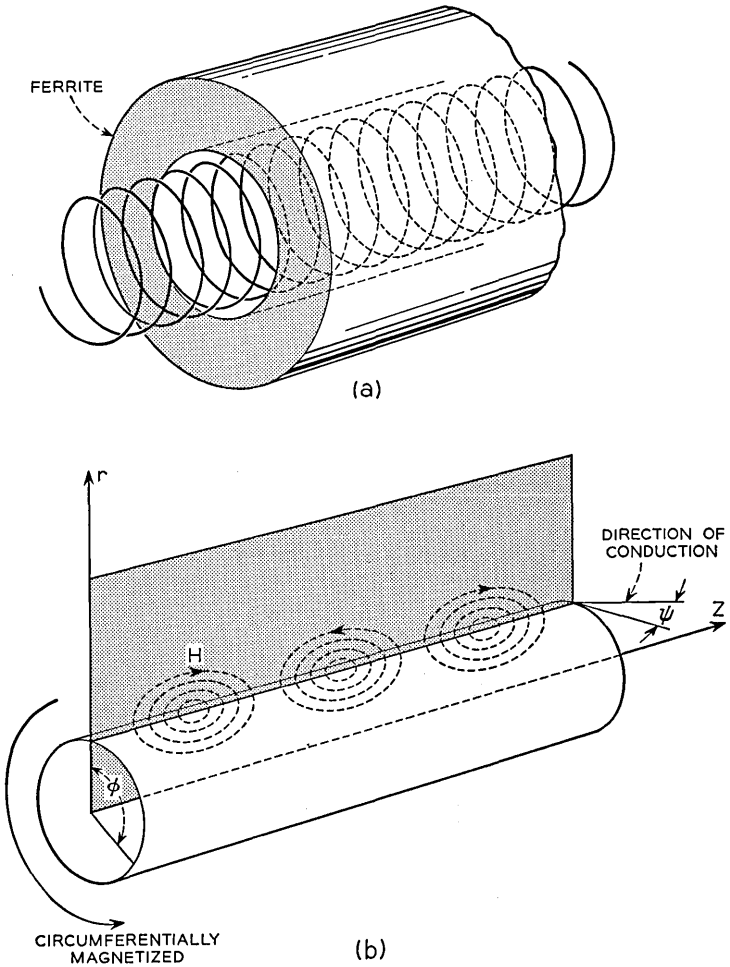


Fig. 11 — (a) Cylindrical helix surrounded by ferrite, (b) Magnetic field lines projected onto a plane containing the axis of the helix.

$$Y_{\mathrm{TM}}^{(e)} = -\frac{j\omega\epsilon_1}{\alpha_1}\frac{K_1(\alpha_1 r)}{K_0(\alpha_1 r)}.$$

Inside the helix, we require the solution for an isotropic region which remains regular as $r \to 0$. Accordingly

$$Y_{\mathrm{TE}}^{(i)} = \frac{j\alpha_0}{\omega\mu_0}\frac{I_0(\alpha_0 r)}{I_1(\alpha_0 r)}; \qquad \alpha_0 = \sqrt{\beta^2 - \beta_0^2},$$

$$\beta_0^2 = \omega^2\mu_0\epsilon_0$$

and

$$Y_{\mathrm{TM}}^{(i)} = j\frac{\omega\epsilon_0}{\alpha_0}\frac{I_1(\alpha_0 r_0)}{I_0(\alpha_0 r_0)},$$

where ϵ_0, μ_0 are the dielectric constant, and permeability of vacuum. Combining these expressions in (34), we obtain after slight rearrangement

$$\frac{I_1(\alpha_0 r_0)}{I_0(\alpha_0 r_0)} + \frac{\alpha_0}{\alpha_1}\frac{\epsilon_1}{\epsilon_0}\frac{K_1(\alpha_1 r_0)}{K_0(\alpha_1 r_0)} = \left[\alpha_0^2\frac{I_0(\alpha_0 r_0)}{I_1(\alpha_0 r_0)} - \frac{\mu_0\alpha_0}{\mu\alpha_2}\frac{\beta^2 - \omega^2\mu\epsilon_1}{(2\chi - 1)}\frac{W_{\chi,0}(2\alpha_2 r)}{W_{\chi,1}(2\alpha_2 r)}\right]\tan^2\psi \tag{35}$$

which determines β.

A complete solution of equation (35) is out of the question. However, as in the planar case, for the slow waves used in traveling wave tube work, the equation may be simplified so that solutions may be computed rather easily. For electron velocities usually employed the resultant β must be about $10\beta_0$. Therefore in equation (35) it will be permissible to neglect all the quantities β_0^2, β_1^2, β_2^2, $\omega^2\mu\epsilon_1$, in comparison with β^2, except in the narrow ranges of magnetic field such that μ or $\mu(1 - \rho_H^2)$ becomes very large. This will occur near $\pm\sigma_0$ where $\sigma_0 = -p/2 + \sqrt{p^2/4 + 1}$, and near $\sigma = 1$. A solution obtained by assuming a large β must be self-consistent; that is, it can be credited only in regions where it does, in fact, predict large β. However, in Section 2.3 it was shown for the plane helix that in any practical case the ranges of magnetic fields so excluded are very narrow, even in the loss-free case, and one may suppose that this is true also in the cylindrical case.

For slow waves, each of the α's reduces to $|\beta|$; the absolute value sign derives from the fact that the positive square root is implied in the definition of the α's. Therefore

$$\chi \to \frac{-\rho_H\beta}{2|\beta|} = -\frac{\rho_H}{2}\operatorname{sgn}\beta. \tag{36}$$

Now the suffix χ of the Whittaker functions no longer depends on the

magnitude of β, and it is chiefly for this reason that further progress is possible. For large β, equation (35) can be written

$$\beta^2 = \beta_0^2 \cot^2 \psi \frac{\dfrac{I_1(|\beta| r_0)}{I_0(|\beta| r_0)} + \dfrac{\epsilon_1}{\epsilon_0} \dfrac{K_1(|\beta| r_0)}{K_0(|\beta| r_0)}}{\dfrac{I_0(|\beta| r_0)}{I_1(|\beta| r_0)} - \dfrac{1}{\dfrac{\mu}{\mu_0}(2\chi - 1)} \dfrac{W_{\chi,0}(2|\beta| r_0)}{W_{\chi,0}(2|\beta| r_0)}}. \tag{37}$$

where χ is now given by equation (36). Equation (37) is now solved by

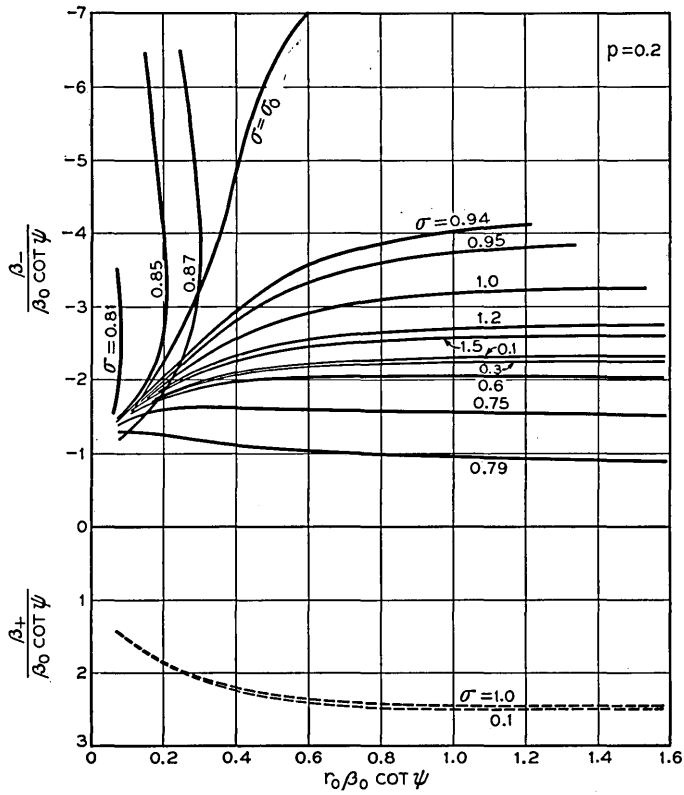


Fig. 12 — Reduced forward and reverse propagation constants versus reduced radius of a cylindrical helix (loss-free case) for various σ and p. The range $\sigma_1 < \sigma < \sigma_0$ where

$$\sigma_1 = \sqrt{1 + \frac{p^2}{4}} - \frac{p}{3} - \frac{p}{2} \quad \text{and} \quad \sigma_0 = \sqrt{1 + \frac{p^2}{4}} - \frac{p}{2}$$

contains an infinity of shape resonances and is not shown here. (a), above, $p = 0.2$. (b) $p = 0.6$. (c) $p = 1.0$.

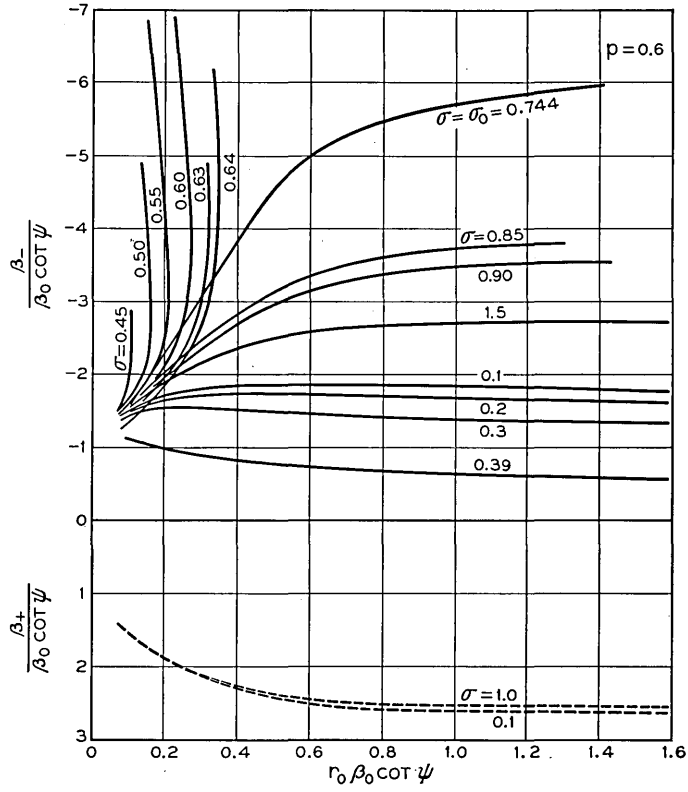


Fig. 12(b) — See Fig. 12.

the following procedure: Introduce the parameter

$$\frac{u}{2} = |\beta| r_0 . \tag{38}$$

For a given ρ_H and u (or σ and p), and a given sign of β, each value of u determines β through equation (37), and then r_0 through equation (38). Thus β can be plotted versus r_0. The procedure is repeated for the opposite sign of β (and therefore the opposite sign of χ). A different curve of β versus r_0 is then obtained. Thus for a given value of r_0, the "forward" and "backward" propagation constants are different in magnitude. The results (computed for a typical ratio $\epsilon_1/\epsilon_0 = 10$) are conveniently stated in terms of $\bar{\beta} = \beta/(\beta_0 \cot \psi)$ and $\bar{r}_0 = r_0\beta_0 \cot \psi$ and are shown in Fig. 12(a) to (c), and again, for fixed $\bar{r}_0$, in Fig. 13(a) to (e). We note that for $\bar{r}_0$ in excess of about 1.5, the results are almost the same as those

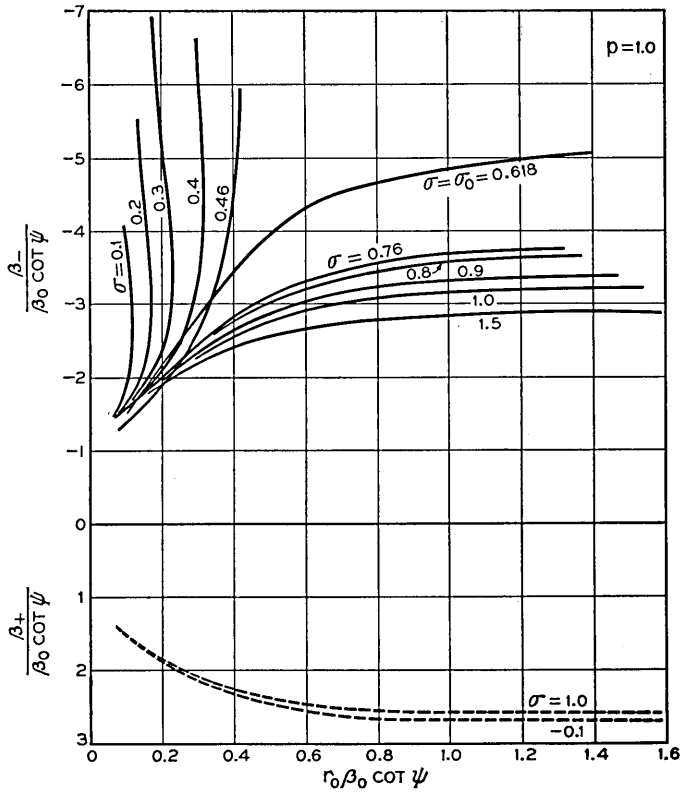


Fig. 12(c) — See Fig. 12.

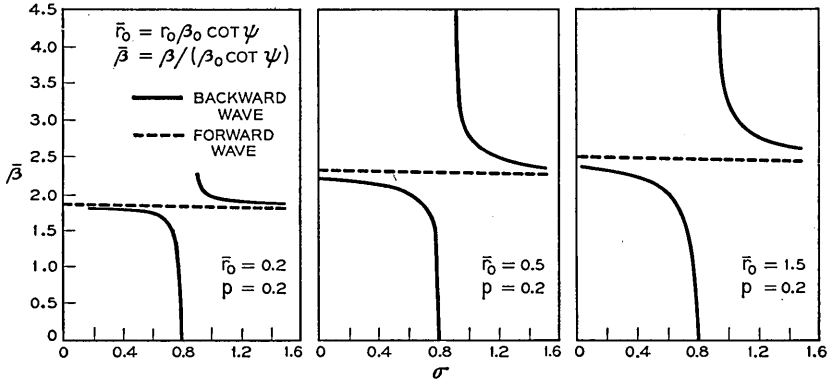


Fig. 13 — Reduced forward and reverse propagation constants versus σ for various reduced radii of a cylindrical helix. (Loss free case). The region $1 - p < \sigma < \sigma_0$ is omitted. (a), above, $p = 0.2$. (b) $p = 0.4$. (c) $p = 0.6$. (d) $p = 0.8$. (e) $p = 1.0$.

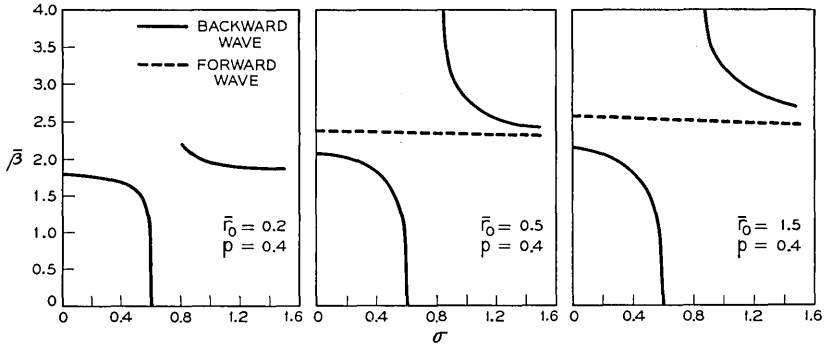


Fig. 13(b) — See Fig. 13.

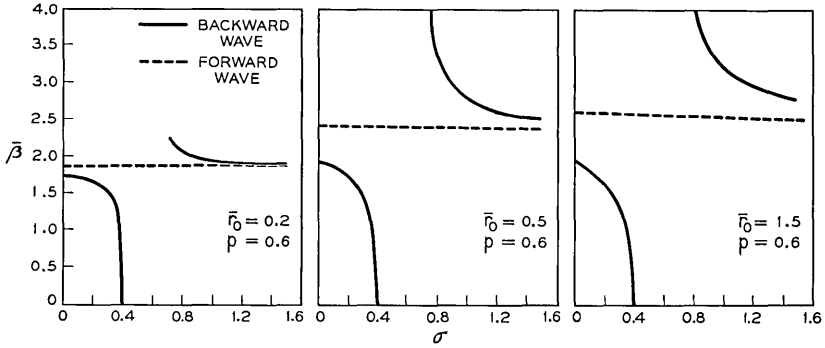


Fig. 13(c) — See Fig. 13.

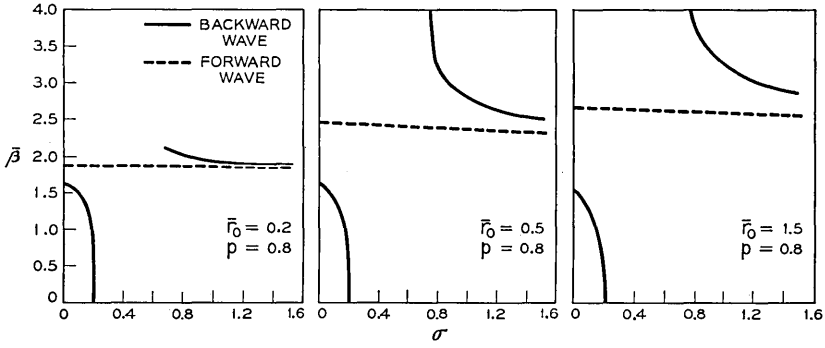


Fig. 13(d) — See Fig. 13.

for a flat helix. This is due to the fact that the large dielectric constant reduces the circuit wavelength so much that the radius appears infinite by comparison once it exceeds 1.5. In traveling-wave tube practice, however, $\bar{r}_0$ is generally below 1.5.

The behavior of the $\bar{\beta} - \bar{r}_0$ curves can be understood from the behavior of the ratio

$$\frac{W_{\chi,0}(u)}{W_{\chi,1}(u)} = Z_\chi(u)$$

and of the coefficient $1/[\mu/\mu_0\ (2\chi - 1)]$. Suppose first that χ is positive. When χ exceeds zero only slightly, $Z_\chi(u)$ behaves essentially like $K_0(u/2)/K_1(u/2)$. This function is always positive; it begins at 0 when $u = 0$ with a vertical tangent and steadily increases to unity as $u \to \infty$. $Z_\chi(u)$ varies in the same way, see Figs. 14 and 15, in the range $0 < \chi < \frac{1}{2}$. For $\frac{3}{2} > \chi \geqq \frac{1}{2}$, $W_{\chi,0}$, and therefore Z_χ, has a zero which increases from $u = 0$ at $\chi = \frac{1}{2}$ to $u = 1$ at $\chi = \frac{3}{2}$. Accordingly $Z_\chi(u)$ in $\frac{3}{2} >$

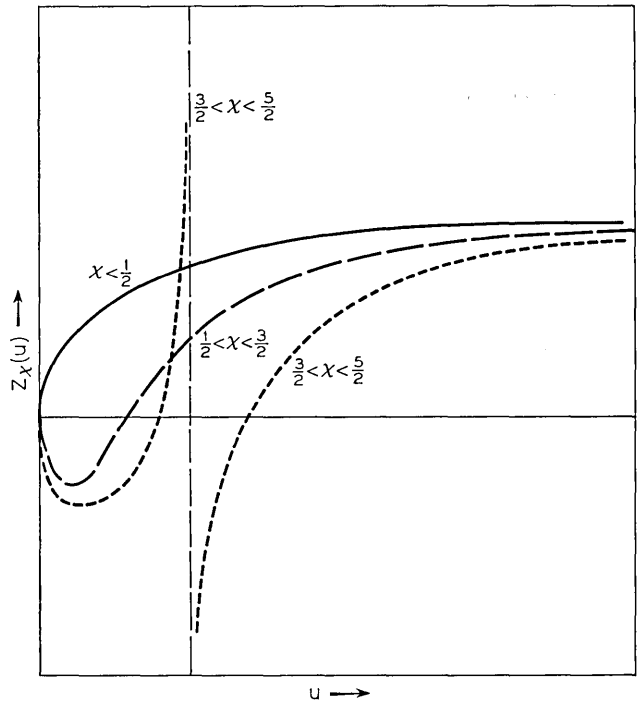


Fig. 14 — Schematic behavior of the function $Z_\chi(u) = W_{\chi,0}(u)/W_{\chi,1}(u)$ in various ranges of χ.

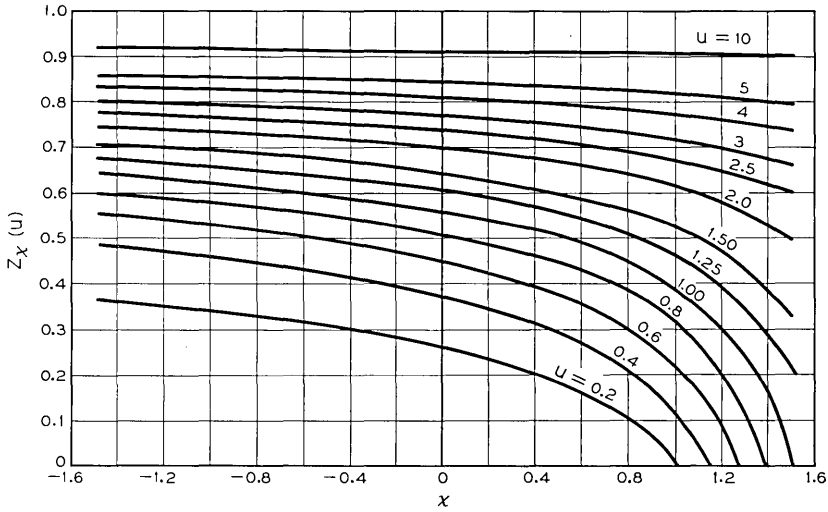


Fig. 15 — The function $Z_\chi(u)$ versus χ for various u, in the range $-\frac{3}{2} < \chi < \frac{3}{2}$.

$\chi > \frac{1}{2}$ starts from 0 at $u = 0$ with a downward-directed vertical tangent, achieves a negative minimum, then increases, through its zero, to the asymptotic value unity as $u \to \infty$. The minimum becomes deeper as χ approaches $\frac{3}{2}$. At $\chi = \frac{3}{2}$, $W_{\chi,1}(u)$ developes a zero at $u = 0$, which steadily moves to larger u as χ increases further towards $\frac{5}{2}$. At the same time the zero of $W_{\chi,0}$ already discussed moves from 1 to $2 + \sqrt{2}$, and a new zero arises at $u = 0$, $\chi = \frac{3}{2}$, which increases to $2 - \sqrt{2}$ as χ approaches $\frac{5}{2}$, but which lags behind the zero of $W_{\chi,1}(u)$. The function $Z_\chi(u)$ now has a pole and two zeros, and behaves as shown in Fig. 14. This process continues; each time χ passes $(2n + 1)/2$, a new zero and a new pole appear. (For a detailed list of poles and zeros the reader is referred to Appendix I). To apply these results, we first resort to the Polder relations.

In terms of σ, p, we have, for β negative

$$\chi = \tfrac{1}{2} \frac{p}{1 - p\sigma - \sigma^2}$$

and

$$\frac{1}{\frac{\mu}{\mu_0}(2\chi - 1)} = \frac{1 - \sigma}{p - 1 + \sigma} = A, \qquad \text{say.}$$

The characteristic equation is

$$\bar{\beta}^2 = \frac{\dfrac{I_1(u/2)}{I_0(u/2)} + \dfrac{\epsilon_1}{\epsilon_0}\dfrac{K_1(u/2)}{K_0(u/2)}}{\dfrac{I_0(u/2)}{I_1(u/2)} - AZ_\chi(u)}; \qquad |\bar{\beta}|\bar{r}_0 = u/2 \tag{37}$$

and can now be discussed in terms of σ at a fixed p. Suppose that $p < 1$. Then A is negative for $\sigma < 1 - p$. In the same range, $\chi < \frac{1}{2}$, so that Z behaves essentially as K_0/K_1. Therefore both numerator and denominator are positive for all u, and the ratio tends to the planar result

$$\bar{\beta}^2 = \frac{1 + \dfrac{\epsilon_1}{\epsilon_0}}{1 - A}$$

as $u \to \infty$. As $u \to 0$, $\bar{\beta}^2$ tends to zero along the vertical, as can be shown by an examination of the various functions near $u = 0$. For $\sigma < 1 - p$, the course of the $\bar{\beta}^2$ versus u-curves is as shown schematically in Fig. 16(a), and it is easily seen that the $\bar{\beta}$ versus $\bar{r}_0$ curves run in essentially the same way, Fig. 12(a) to (c). However, as σ approaches $1 - p$, the $\bar{\beta}^2$ versus u curves steadily fall, until at $\sigma = 1 - p$, $\bar{\beta}^2 = 0$ for all finite u, since $A = -\infty$.

As σ passes $1 - p$, A changes sign and at the same time χ passes $\frac{1}{2}$ so that Z_χ acquires a zero. As σ varies from $1 - p$ to $1 - p/2$, A decreases from $+\infty$ to unity. Therefore, while $u < u_1$, the zero of Z_χ, $1 - AZ_\chi$ is positive; however as u increases beyond u_1, $\{I_0(u/2)/I_1(u/2)\} - AZ_\chi(u)$ eventually passes zero, since $Z_\chi(u)$ and $I_0(u/2)/I_1(u/2)$ both approach unity. On the other hand the numerator of equation (37) is

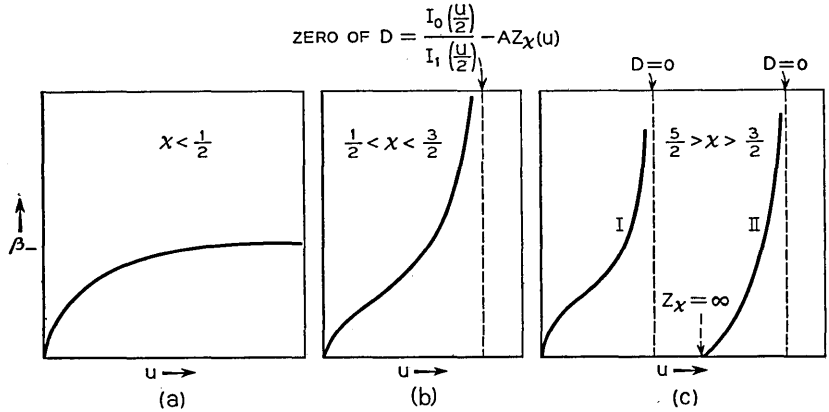


Fig. 16 — Schematic variation of β_- with u. a) $\chi < \frac{1}{2}$; b) $\frac{1}{2} < \chi < \frac{3}{2}$; c) $\frac{3}{2} < \chi < \frac{5}{2}$.

always positive; therefore $\bar{\beta}^2$ approaches infinity at $I_0(u/2)/I_1(u/2) - AZ_\chi(u) = 0$, and no real values of $\bar{\beta}$ exist thereafter (see Fig. 16b). Since this "cut-off" occurs at a finite value of u, the corresponding value of r_0 is zero. This explains the bulging of the corresponding $\bar{\beta} - \bar{r}_0$ curves in Fig. 12(a) to 12(c).

The next major change in the curves occurs when χ exceeds $\frac{3}{2}$, (that is, σ exceeds

$$\sigma_1 = -\frac{p}{2} + \sqrt{1 + \frac{p^2}{4} - \frac{p}{3}}.)$$

For $p < 2$, σ_1 is still less than $1 - (p/2)$, so that, initially at any rate, A is still greater than unity. In addition to the infinity of β^2 just discussed, a further infinity arises between $u = 0$, and the pole of $Z_\chi(u)$, as is seen from Fig. 16(c). β^2 increases from zero at $u = 0$ to this infinity, thereafter it is negative, until the pole of $Z_\chi(u)$ is reached. Thereupon it resumes at $\beta^2 = 0$ and approaches infinity at the zero of the denominator $[I_0(u/2)/I_1(u/2)] - AZ_\chi(u)$ already discussed. Thus there are now two branches of the $\bar{\beta}^2 - u$ curve; their corresponding traces in the $\bar{\beta} - \bar{r}_0$ plane are shown schematically in Fig. 17. (The computations on which Fig. 12(a) to 12(c) were based were broken off at $\sigma = \sigma_1$.)

A further branch is added each time χ increases beyond a number of the form $(2n + 1)/2$ (σ increases beyond

$$\sigma_n = -\frac{p}{2} + \sqrt{1 + \frac{p^2}{4} - \frac{p}{2n + 1}}).$$

These all resemble the two branches just discussed, until $n >$

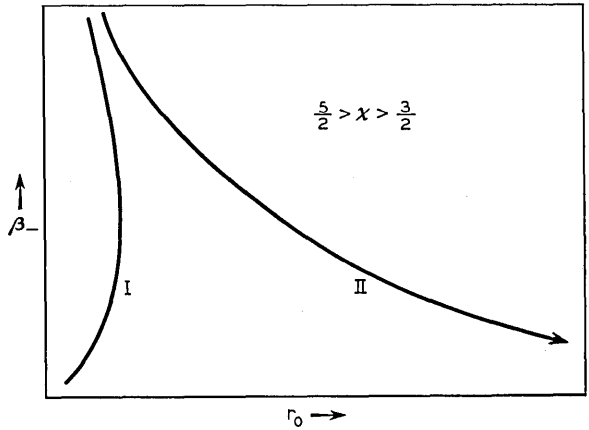


Fig. 17 — Schematic variation of β_- with r_0 for $\frac{3}{2} < \chi < \frac{5}{2}$.

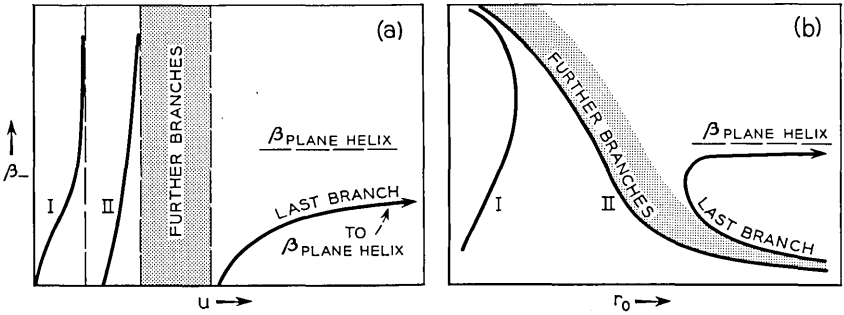


Fig. 18 — (a) β_- versus u when $1 - (p/2) < \sigma < \sigma_0$. (b) β_- versus r_0 under the same circumstances.

$\frac{1}{2}\left(\frac{4}{p} - 1\right)$. When this occurs, $\sigma_n > 1 - (p/2)$, so that there will be a value σ between σ_{n-1} and σ_n beyond which the denominator no longer decreases through zero as $u \to \infty$, but approaches a finite positive value, Fig. 18(a). Accordingly β^2 approaches a finite positive value, and cut-off of the extreme right hand branch, Fig. 16(c), no longer occurs. The corresponding $\bar{\beta}$ versus $\bar{r}_0$ branch is as in Fig. 18(b).

As $n \to \infty$ ($\sigma_n \to \sigma_0 = \sqrt{1 + (p^2/4)} - p/2$) the number of branches increases to infinity. This situation resembles that in the completely filled waveguide (Part I), where we found an infinity of modes ("Shape-resonances") in the range $\sigma_0 < \sigma < 1$. In the present case, however, they are to be found in the range $1 - p < \sigma < \sigma_0$.

When $\sigma = \sigma_0 + 0$, χ is infinite and negative. The function $Z_\chi(u)$ is then constant and equal to unity. A is less than unity, and the denominator of equation (27) has no zeros. The $\bar{\beta}$ versus $\bar{r}_0$ curve is now "normal" again, see Fig. 12(a). As σ increases further, the curve falls (since A decreases steadily to -1 as $\sigma \to \infty$), and no more qualitative changes occur.

3.3 *Cylindrical Waveguides*

As pointed out before, the fact that the propagation problem in the cylindrical case can always be integrated in terms of Whittaker functions when the fields show no angular variation is an accident, and in view of the lack of numerical tables, not a particularly fortunate one. Only in special cases (like that of the slow-wave helix) is the text-book information on these functions of any great utility. In general, it will be more convenient to solve the differential equations numerically. However, for completeness, we shall state some of the formal results for a cylindrical waveguide containing a cylinder of circumferentially magnetized ferrite, and propagating a TE_0 mode.

First we consider a waveguide, radius r_0, into which is fitted a hollow cylinder of ferrite, outer radius r_0, inner radius r_1. In that cylinder, the magnetic field H_z may be taken to be a superposition $AS_{\chi 0}(2\alpha_2 r) + BR_{\chi 0}(2\alpha_2 r)$ where, as before,

$$\alpha_2^2 = \beta^2 - \omega^2\epsilon_1\mu(1 - \rho_H^2); \qquad S(x) = \frac{M(x)}{\sqrt{x}}; \qquad R(x) = \frac{W(x)}{\sqrt{x}}.$$

$$\chi = -\frac{\beta\rho_H}{2\alpha_2}.$$

α_2 may be either real or positive imaginary. [In choosing this combination we depart from the usual practice of taking a superposition of J_0 and N_0 in the isotropic case. Were we to follow this practice, it would be necessary to define a new function $R_{\chi\mu}(2jx)e^{j(\mu\pi/2)} + R_{\chi\mu}(-2jx)e^{-j\mu\pi}$ to correspond to $N_\mu(x)$. Our choice corresponds to taking a combination of $J_0(x)$ and one of the Hankel functions $H_0(x)$ in the isotropic case. Since the functions H, J, N are linearly dependent, this will not affect the results.]

In view of the difference relations, equation (39) in Appendix I, and of equation (29) we obtain for the impedance in the ferrite

$$\frac{E_\varphi}{H_z} = -\frac{j\omega\mu\alpha_2}{(\beta^2 - \omega^2\mu\epsilon_1)} \frac{[A(\tfrac{1}{4} - \chi^2)S_{\chi 1}(2\alpha_2 r) + B(2\chi - 1)R_{\chi 1}(2\alpha_2 r)]}{[AS_{\chi 0}(2a_2 r) + BR_{\chi 0}(2a_2 r)]}.$$

A and B must be adjusted so that this quantity vanishes at r_0, the guide wall. This gives

$$\frac{E_\varphi}{H_z} = -(2\chi - 1)(\tfrac{1}{4} - \chi^2)\frac{j\omega\mu\alpha_2}{\beta^2 - \omega^2\mu\epsilon_1} \frac{W_{\chi,1}(2\alpha_2 r_0)M_{\chi,1}(2\alpha_2 r) - M_{\chi,1}(2\alpha_2 r_0)W_{\chi,1}(2\alpha_2 r)}{(2\chi-1)W_{\chi,1}(2\alpha_2 r_0)M_{\chi,0}(2\alpha_2 r) - (\tfrac{1}{4} - \chi^2)M_{\chi,1}(2\alpha_\chi r_0)W_{\chi,0}(2\alpha_2 r)}.$$

In the vacuum, between $r = 0$ and $r = r_1$, H_z is $I_0(\alpha_0 r)$ and the impedance is $E_\varphi/H_z = -j(\omega\mu_0/\alpha_0)I_1(\alpha_0 r)/I_0(\alpha_0 r)$ where $\alpha_0^2 = \beta^2 - \omega^2\mu\epsilon_0$. At $r = r_1$, the ferrite-vacuum interface, the two impedances must be equal. Thus we obtain the characteristic equation

$$\frac{\mu_0}{\alpha_0}\frac{I_1(\alpha_0 r_1)}{I_0(\alpha_0 r_1)} = (2\chi - 1)(\tfrac{1}{4} - \chi^2)\frac{\alpha_2\mu}{\beta^2 - \omega^2\mu\epsilon_1} \frac{W_{\chi,1}(2\alpha_2 r_0)M_{\chi,1}(2\alpha_2 r_1) - M_{\chi,1}(2\alpha_2 r_0)W_{\chi,1}(2\alpha_2 r_1)}{(2\chi - 1)W_{\chi,1}(2\alpha_2 r_0)M_{\chi,0}(2\alpha_2 r_1) - (\tfrac{1}{4} - \chi^2)M_{\chi,1}(2\alpha_2 r_0)W_{\chi,0}(2\alpha_2 r_1)}.$$

(It is understood that for "normal" waveguide propagation α_0 will be imaginary, and the I will be replaced by J). As a simple illustration we

consider the case in which the ferrite cylinder fills the waveguide completely. E_φ is then proportional to $M_{\chi 1}(2\alpha_2 r)$, and so β is determined from the condition

$$M_{\chi 1}(2\alpha_2 r_0) = 0.$$

When "normal" propagation prevails (β less than the natural propagation constant of the medium $\omega\sqrt{\epsilon_1\mu(1 - \rho_H^2)}$) both α_2 and χ are imaginary, equal to $j\alpha_2'$ and $j\chi'$, say. Under these circumstances little is known about the zeros of M. However, it is possible to say something about the solution for large radial mode numbers. It follows from Erdélyi et al.[2] **1,** p. 278, formula (2), that for large argument

$$M_{j\chi',1}(2j\alpha_2' r_0) = \text{const}\cdot\sin[\alpha_2' r_0 + \chi' \log \alpha_2' r_0 + \chi' \log 2 + \Phi(\chi') - \pi/4].$$

where $\Phi(\chi') = \arg \Gamma(\tfrac{3}{2} + j\chi')$. The zeros of this expression are at

$$\alpha_2' r_0 + \chi' \log \alpha_2' r_0 = \frac{5n\pi}{4} - \chi' \log 2 - \Phi(\chi'). \qquad n = \text{a large integer}$$

This equation may be solved graphically by setting $\alpha_2' r_0 = u$, assigning values to β, ρ_H, u (and hence to χ', α_2'). From a solution u one then finds

$$r_0 = u/\alpha_2'.$$

M also has zeros for real α_2, χ, if χ is large enough. Thus the waveguide will support waves with a β^2 greater than $\omega^2\mu\epsilon_1(1 - \rho_H^2)$. It is shown in Reference 2 (**1,** p. 289) that when χ is between $\tfrac{3}{2}$ and $\tfrac{5}{2}$, M has one zero, when χ is between $\tfrac{5}{2}$ and $\tfrac{7}{2}$, M has two zeros and so on. Suppose that ρ_H is negative, $= -|\rho_H|$. Then

$$\chi = \frac{\beta|\rho_H|}{\sqrt{\beta^2 - \beta_2^2}}.$$

For real positive β, this equation has a solution for β if $|\rho_H| < \chi$:

$$\beta = \frac{\chi\beta_2}{\sqrt{\chi^2 - \rho_H^2}}.$$

If $\tfrac{3}{2} < \chi < \tfrac{5}{2}$, M will have a zero $u(\chi)$ depending on the value of χ. Thus the equations

$$\beta = \frac{\chi\beta_2}{\sqrt{\chi^2 - \rho_H^2}}; \qquad r_0 = \frac{u(\chi)}{\sqrt{\beta^2 - \beta_2^2}} = \frac{\sqrt{\chi^2 - \rho_H^2}}{|\rho_H|\beta_2} u(\chi)$$

solve the propagation problem parametrically. Similarly when χ is between $\tfrac{5}{2}$ and $\tfrac{7}{2}$, there are two zeros of M given by two functions

$u_1(\chi)$ and $u_2(\chi)$. There are now two possible modes, with the same restrictions on ρ_{II}. An additional mode arises each time χ is allowed to pass a number of the form $(2n + 1)/2$. It is to be noted that these modes are not confined to the resonance range. For β positive, they can exist in the range $\infty > \sigma > \sigma_0$ and in the range $-\sigma_0 < \sigma < 0$.

Appendix I. Some Properties of Whittaker Functions Used in This Paper

I. Relation to Bessel functions

$$\lim_{x\to 0} \frac{M_{\chi,\mu}(2jx)}{\sqrt{2jx}} = 2^{2\mu}\Gamma(\mu + 1)j^{\mu}J_{\mu}(x),$$

$$\lim_{x\to 0} \frac{M_{\chi,\mu}(2x)}{\sqrt{2x}} = 2^{2\mu}\Gamma(\mu + 1)I_{\mu}(x),$$

$$\lim_{x\to 0} \frac{W_{\chi,\mu}(2x)}{\sqrt{2x}} = \frac{K_{\mu}(x)}{\sqrt{\pi}},$$

$$\lim_{x\to 0} \frac{W_{\chi,\mu}(-2jx)}{\sqrt{-2jx}} = \frac{\sqrt{\pi}}{2} j^{\mu+1}H_{\mu}^{(1)}(x), \qquad \text{and}$$

$$\lim_{x\to 0}\left[\frac{W_{\chi,\mu}(2jx)}{\sqrt{2jx}} e^{j(\mu\pi/2)} + \frac{W_{\chi,\mu}(-2jx)}{\sqrt{-2jx}} e^{-j(\mu\pi/2)}\right] = -\sqrt{\pi}N_{\mu}(x)$$

II. Difference relations

The following results can be obtained either by reference to Erdélyi,[2], 1 pp. 258, 254, by differentiation and subsequent integration by parts of integrals such as

$$W_{\chi,\mu}(x) = \frac{x^{\mu+1/2}\, e^{-x/2}}{\Gamma(\mu + \tfrac{1}{2} - \chi)} \int_0^{\infty} e^{-x\tau}\tau^{\mu-\chi-1/2}(H(1 + \tau)^{\mu+\chi-1/2}\, d\tau$$

or by observing that combinations of the form

$$\sqrt{x}\,\frac{d}{dx}\frac{W_{\chi,0}(x)}{\sqrt{x}} + \chi W_{\chi,0}(x)$$

satisfy Whittaker's equation with $\mu = 1$. In the last mentioned method, the required constant multiplying the first order Whittaker function can be obtained by reference to the limiting behavior for small x. If $R_{\chi\mu} = W_{\chi,\mu}(x)/\sqrt{x}$ and $S_{\chi\mu} = M_{\chi,\mu}(x)/\sqrt{x}$ the results are

$$\begin{aligned} R_{\chi 0}' + \chi R_{\chi 0} &= (\chi - \tfrac{1}{2})R_{\chi 1}, \\ S_{\chi 0}' + \chi S_{\chi 0} &= \tfrac{1}{2}\,(\tfrac{1}{4} - \chi^2)\, S_{\chi 1}. \end{aligned} \tag{39}$$

III. ZEROS

When $\chi = (2n + 1)/2$, $n = 1, 2 \ldots$, $W_{\chi,\mu}/x^{\mu-1/2}$ reduces to a polynomial times a function of χ. This may be inferred from the asymptotic expansion,

$$W_{\chi,\mu} = e^{-(x/2)}x^{\chi}$$
$$\left(1 + \sum_{n=1}^{\infty} \frac{[\mu^2 - (\chi - \tfrac{1}{2})^2][\mu^2 - (\chi - \tfrac{3}{2})^2] \cdots [\mu^2 - (\chi - n + \tfrac{1}{2})^2]}{n!x^n}\right),$$

which terminates if $\mu = 1$ and $\chi = n + \tfrac{1}{2}$ ($n = 1, 2 \ldots$), or from the fact that for these values of the suffixes, W reduces to the generalized Laguerre polynomial

$$L^{(2)}_{\chi-(3/2)}(x).$$

Similarly when $\chi = (2n + 1)/2$, $n = 0, 1, 2 \ldots$, $W_{\chi,0}$ reduces to the Laguerre polynomial

$$L_{\chi-(1/2)}(x).$$

The zeros at the critical values of χ are given in the following table

χ	Zeros of $W_{\chi,0}$	Zeros of $W_{\chi,1}$
$\frac{1}{2}$	0	None
$\frac{3}{2}$	0; 1	0
$\frac{5}{2}$	0; 0.586; 3.414	0; 3
$\frac{7}{2}$	0; 0.416; 2.294; 6.290	0; 6; 2
$\frac{9}{2}$	0; 0.323; 1.746; 4.537; 9.395	0; 1.517; 4.312; 9.171
$1\frac{1}{2}$	0; 0.26356; 1.413; 3.596; 7.086; 12.641	0; 1.227; 3.413; 6.903; 12.458

Between $n + \frac{1}{2}$ and $n + \frac{3}{2}$, $W_{\chi,0}$ has $n + 1$ zeros ($n = 0, 1, 2 \ldots$) and $W_{\chi,1}$ has n zeros ($n = 1, 2 \ldots$).

The zeros of $M_{\chi,1}$ coincide with those of $W_{\chi,1}$ when $\chi = n + \frac{1}{2}$, and at those values of χ only. The functions $M_{\chi,1}$ and $W_{\chi,1}$ then are proportional to each other.*

IV. THE RICCATI-EQUATION FOR THE IMPEDANCE FUNCTION $W_{\chi,0}/W_{\chi,1}$.

The computations concerning the cylindrical helix required a study of the function

$$Z(u) = \frac{W_{\chi,0}(u)}{W_{\chi,1}(u)}.$$

* At these critical values of χ, the solution to the problem of the hollow cylinder of ferrite in the waveguide breaks down, since M and W are then not independent. A further independent solution must then be constructed.

We will show that $Z_\chi(u)$ satisfies a non-linear first order differential equation of Riccati-type. From the difference relations, equation (36), we have

$$(\chi - \tfrac{1}{2})\frac{W_{\chi,1}(u)}{W_{\chi,0}(u)} = \frac{W_{\chi,0}'(u)}{W_{\chi,0}(u)} - \frac{1}{2u} + \chi$$

$$= \frac{1}{g}, \qquad \text{say}$$

and from Whittaker's equation

$$\frac{d}{du}\left(\frac{W_{\chi,0}'}{W_{\chi,0}}\right) + \left(\frac{W_{\chi,0}'}{W_{\chi,0}}\right)^2 + \left(\frac{1}{4u^2} + \frac{\chi}{u} - \frac{1}{4}\right) = 0.$$

Therefore

$$\frac{d}{du}\frac{1}{g} + \frac{1}{g^2} + \frac{1}{g}\left(\frac{1}{u} - 2\chi\right) + (\chi^2 - \tfrac{1}{4}) = 0,$$

or

$$\frac{dg}{du} = 1 + \left(\frac{1}{u} - 2\chi\right) g + (\chi^2 - \tfrac{1}{4})\, g^2.$$

Finally, let

$$Z = (\chi - \tfrac{1}{2})\, g(u) = \frac{W_{\chi,0}(u)}{W_{\chi,1}(u)}.$$

Then

$$\frac{dZ}{du} = (\chi - \tfrac{1}{2}) + \left(\frac{1}{u} - 2\chi\right)Z + (\chi + \tfrac{1}{2})\, Z^2.$$

Since this equation is satisfied by $M_{\chi,0}(u)/M_{\chi,1}(u))$ as well as by $W_{\chi,0}(u)/W_{\chi,1}(u)$, a selection has to be made from all the possible solutions of this equation. We require the one which for large u approaches unity. But for large u the equation is

$$\frac{dZ}{du} = (1 - Z)\,[\chi - \tfrac{1}{2} - (\chi + \tfrac{1}{2})\, Z],$$

whose integral is

$$Z = \frac{(\chi - \tfrac{1}{2})Ae^u - 1}{(\chi + \tfrac{1}{2})Ae^u - 1}$$

For large u, therefore, the solution is either unity, when $A = 0$, or else $(\chi - \tfrac{1}{2})/(\chi + \tfrac{1}{2})$, $A \neq 0$. The solution with $A \neq 0$ corresponds to the M functions; that with $A = 0$ to the W-functions.

The case $A = 0$ was integrated on an analogue-computer, and the results are shown in Fig. 14(b). The computation was restricted to the range $|\chi| < 3/2$. Beyond these values, the helix-problem was discussed only qualitatively.

REFERENCES

1. M. L. Kales, H. N. Chait, and N. G. Sakiotis, Letter to the Editor, J. Appl. Phys., **24,** No. 6.
2. Erdélyi, Magnus, Oberhettinger and Tricomi, *Higher Transcendental Functions*, **I,** McGraw-Hill, 1953.
3. E. H. Turner, I. R. E. Proc., **41,** p. 937, July, 1953.
4. J. S. Cook, R. Kompfner, and H. Suhl, Non-Reciprocal Loss in Traveling Wave Ferrite Attenuators, Letter to Editor, I. R. E. Proc., to be published.

Theoretical Fundamentals of Pulse Transmission — II

By E. D. SUNDE

(Manuscript received September 23, 1953)

PART II.

Part I of this paper dealt with various idealized transmission characteristics and with methods of evaluating pulse distortion resulting from various system imperfections. In Part II the resultant transmission impairments or limitations on pulse transmission rates are discussed for systems with low-pass, symmetrical band-pass and asymmetrical band-pass characteristics, and a comparison made of the transmission performance of double and vestigial sideband systems. The limitation on channel capacity imposed by random imperfections in the transmission-frequency characteristic, as compared to random noise, is also discussed.

12. IMPULSE CHARACTERISTICS AND PULSE TRAIN ENVELOPES

In pulse modulation systems pulses are transmitted in various combinations to form pulse trains, and at the receiving end the envelope of the pulse train is sampled at regular intervals to determine the amplitudes of the transmitted pulses. As a result of pulse overlaps there may be appreciable distortion of the pulse train envelope, which may cause errors in reception or noise, depending on the type of system. To evaluate transmission impairments, or limitations imposed on transmission capacity to avoid excessive transmission impairments from pulse distortion, it is necessary to establish basic relations between the impulse characteristic of the system and the envelope of the received pulse train.

In Fig. 42 are shown three transmitted pulses of different peak amplitudes, A_{-1}, A_0 and A_1, transmitted at intervals τ with the first and third

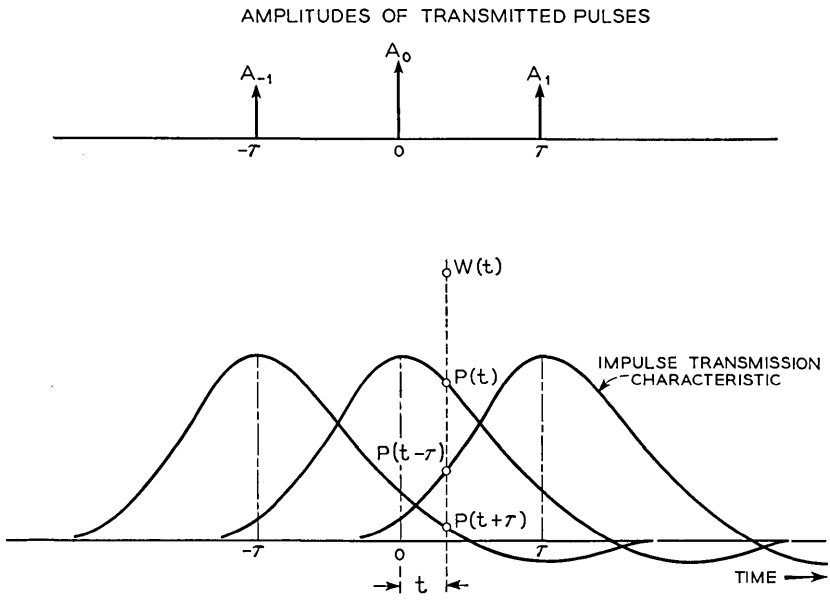
pulse overlapping into the middle pulse. The instantaneous amplitude of the received train at a time t_0 referred to the peak amplitude of the middle pulse is

$$\begin{aligned} W(t_0) &= A_{-1}P(t_0 - \tau) + A_0P(t_0) + A_1P(t_0 + \tau), \\ &= \sum_{n=-1}^{1} A_nP(t_0 + n\tau). \end{aligned} \tag{12.01}$$

When the sequence of pulses transmitted at uniform intervals τ extends between $n = -\infty$ and ∞, the instantaneous amplitude of the pulse train at time t_0 is

$$W(t_0) = \sum_{n=-\infty}^{\infty} A_nP(t_0 + n\tau). \tag{12.02}$$

The above equation gives the instantaneous value $W(t_0)$ for any selected combination of transmitted pulses. The transmitted pulses may have any value within certain limits, as when they represent signal samples in a pulse amplitude modulation system, or may assume two or



$$W(t) = A_{-1}P(t+\tau) + A_0P(t) + A_1P(t-\tau)$$

Fig. 42 — Formulation of expression for pulse train envelope in terms of impulse characteristic and amplitudes of transmitted pulses.

more discrete values as in pulse code modulation systems. In pulse position modulation, $A_n = 0$ except at the instants pulses of a given amplitude are transmitted, and n may not necessarily be an integer. In pulse duration modulation, $A_n = 1$ over the intervals $n\tau$ of varying duration in which pulses are transmitted, and zero otherwise. Equation (12.02) is thus a general formulation of the wave shape of a received pulse train, applicable to various pulse modulation methods.

Inserting (2.09) in (12.02) with $R_- + R_+ = R$ and $Q_- - Q_+ = Q$ and taking $\psi_r = 0$ without loss of generality

$$\begin{aligned} W(t_0) &= \sum_{n=-\infty}^{\infty} A_n[\cos \omega_r(t_0 + n\tau)R(t_0 + n\tau) + \sin \omega_r(t_0 + n\tau)Q(t_0 + n\tau)], \\ &= \cos \omega_r t_0 \sum_{-\infty}^{\infty} A_n[\cos \omega_r n\tau R(t_0 + n\tau) + \sin \omega_r n\tau Q(t_0 + n\tau)] \\ &+ \sin \omega_r t_0 \sum_{-\infty}^{\infty} A_n[\cos \omega_r n\tau Q(t_0 + n\tau) - \sin \omega_r n\tau R(t_0 + n\tau)]. \end{aligned} \tag{12.03}$$

The envelope of the wave at the sampling instant $t_0 = 0$ is

$$\overline{W}(0) = (\bar{R}^2 + \bar{Q}^2)^{1/2}, \tag{12.04}$$

$$\begin{aligned} \bar{R} &= \sum_{-\infty}^{\infty} A_n[\cos \omega_r n\tau R(n\tau) + \sin \omega_r n\tau Q(n\tau)], \\ \bar{Q} &= \sum_{-\infty}^{\infty} A_n[\cos \omega_r n\tau Q(n\tau) - \sin \omega_r n\tau R(n\tau)]. \end{aligned} \tag{12.05}$$

For the particular case of a low-pass system

$$Q = 0, \text{ and } \omega_r = 0,$$

so that

$$W(0) = \sum_{n=-\infty}^{\infty} A_n P(n\tau). \tag{12.06}$$

A band-pass characteristic can be obtained with the aid of band-pass filters at the transmitting or receiving ends, or at both ends of a system, and the equations developed previously for the impulse characteristic tacitly assumed such an arrangement. Equivalent performance can, however, also be secured by methods which are usually employed in practice, and to which the equations also apply. Impulses can thus be applied to a low-pass pulse shaping network or filter, and the output used to modulate a carrier. There will then be a symmetrical distribution of

sidebands with respect to the carrier, equivalent to a band-pass characteristic, with the spectrum of the sideband frequencies determined by the characteristic of the low-pass filter. The equivalent of an asymmetrical band-pass characteristic can be obtained by suppressing part of the upper or lower sideband with the aid of filters.

Although the mathematical formulation with both methods is essentially the same when ω_r is identified with the carrier frequency ω_c, with impulse excitation the phase of ω_r is fixed in relation to the envelope but is independent of it with carrier modulation. By proper choice of the pulse interval τ in (12.03), such that $\cos \omega_r (t_0 + n\tau) = \cos \omega_r t_0$ or $\omega_r \tau = 2\pi m$, $m = 0, 1, 2 \cdots$ it is possible with impulse excitation to obtain the same relation between the reference or carrier frequency as when the output of a low-pass filter is used to modulate a carrier. In the above case the pulses are transmitted at intervals $\tau = m/f_r = m/f_c$, corresponding to multiples of the duration of a carrier cycle. Since the duration of a carrier cycle is ordinarily small in relation to the pulse interval, there is essentially no important difference in the rate at which pulses can be transmitted with the above two methods. However, with band-pass filters the exact relationship of pulse intervals to the carrier frequency may be difficult to maintain with simple instrumentation, while this is no problem with carrier modulation. For this reason, and since the performance is otherwise equivalent, only the basic relationships with carrier modulation will be discussed further.

Assuming that $\cos \omega_r(t_0 + n\tau) = \cos \omega_r t_0$, as discussed above, equation (12.03) becomes

$$W(t_0) = \cos \omega_r t_0 \sum_{-\infty}^{\infty} A_n R(t_0 + n\tau) + \sin \omega_r t_0 \sum_{-\infty}^{\infty} A_n Q(t_0 + n\tau). \quad (12.07)$$

The envelope at the sampling point is accordingly

$$\overline{W}(0) = \left(\left[\sum_{-\infty}^{\infty} A_n R(n\tau) \right]^2 + \left[\sum_{-\infty}^{\infty} A_n Q(n\tau) \right]^2 \right)^{1/2}. \quad (12.08)$$

In ideal transmission systems there would be no pulse overlaps or intersymbol interference, and the amplitude of the pulse train at the sampling instant would be

$$\overline{W}(0) = A_0[R^2(0) + Q^2(0)]^{1/2}. \quad (12.09)$$

This condition could be realized with sufficient pulse spacing. However, the objective in the design of efficient pulse systems is to determine the minimum pulse spacing consistent with tolerable intersymbol inter-

ference and thus the maximum transmission capacity or optimum performance in other respects for a given bandwidth. In the following sections this problem is discussed further.

13. TRANSMISSION LIMITATIONS IN SYMMETRICAL SYSTEMS

In a symmetrical system the amplitude characteristic has even symmetry and the phase characteristic odd symmetry with respect to a properly chosen frequency. A low-pass transmission system is thus symmetrical with respect to zero frequency, when the negative frequency range is included. A double sideband system is symmetrical if the amplitude characteristic has even and the phase characteristic odd symmetry with respect to the mid-band frequency.

Equation (12.06) applying to a low-pass system or baseband transmission may be written

$$W(0) = A_0P(0) + \sum_{n=1}^{\infty} [A_nP(n\tau) + A_{-n}P(-n\tau)]. \qquad (13.01)$$

Let it be assumed that pulses of varying but discrete amplitudes are transmitted, with a maximum peak amplitude equal to $A_{\max}$ and a minimum peak amplitude $A_{\min}$. If q pulse amplitudes are employed, the difference between peak amplitudes is then $(A_{\max} - A_{\min})/(q - 1)$. Let P^+ designate positive values of $P(n\tau)$ and P^- the absolute value of negative amplitudes.

The maximum value of $W(0)$ when a pulse of amplitude A_0 is transmitted at the sampling point $n = 0$ is then

$$W_{\max} = A_0P(0) + \sum_{n=1}^{\infty} A_{\max}[P^+(n\tau) + P^+(-n\tau)] - \sum_{n=1}^{\infty} A_{\min}[P^-(n\tau) + P^-(-n\tau)] \qquad (13.02)$$

The minimum amplitude of $W(0)$ when a pulse of the next higher amplitude $A_0 + (A_{\max} - A_{\min})/(q - 1)$ is transmitted becomes

$$W_{\min} = \left(A_0 + \frac{A_{\max} - A_{\min}}{q - 1}\right) P(0) - \sum_{n=1}^{\infty} A_{\max}\,[P^-(n\tau) + P^-(-n\tau)] + \sum_{n=1}^{\infty} A_{\min}\,[P^+(n\tau) + P^+(-n\tau)]. \qquad (13.03)$$

To permit distinction between the two pulse peaks it is necessary that

$W_{\min}$ be greater than $W_{\max}$. The difference $M = W_{\min} - W_{\max}$, which represents the margin for distinction between pulse amplitudes, becomes

$$M = \frac{A_{\max} - A_{\min}}{q - 1} P(0) - (A_{\max} - A_{\min}) \sum_{n=1}^{\infty} [P^{+}(n\tau) + P^{-}(n\tau) + P^{+}(-n\tau) + P^{-}(-n\tau)], \tag{13.04}$$

or:

$$M = (A_{\max} - A_{\min}) \left[\frac{P(0)}{q - 1} - \sum_{n=1}^{\infty} | P(n\tau) | + | P(-n\tau) | \right], \tag{13.05}$$

where $| P(\pm n\tau) |$ designates the absolute values of the impulse characteristic.

Equation (13.05) shows that for a given value of q the margin depends on the maximum pulse excursion $A_{\max} - A_{\min}$ and is thus the same with $A_{\max} = 1$ and $A_{\min} = 0$ as with $A_{\max} = 0.5$ and $A_{\min} = -0.5$. As an example, equation (13.05) shows that with two pulse amplitudes, $q = 2$, it is possible to distinguish between pulses and spaces, or between positive and negative pulses, if the sum of the absolute values of the impulse characteristic at all the sampling points, excluding 0, is less than the amplitude P(0) of the impulse at sampling point 0.

The maximum margin against errors is obtained without pulse overlaps, i.e. when the summation term in (13.05) is zero, and is

$$M_{\max} = (A_{\max} - A_{\min}) \frac{P(0)}{q - 1}. \tag{13.06}$$

The ratio of the margin M as given by (13.05) to the maximum margin becomes:

$$M/M_{\max} = 1 - \frac{q - 1}{P(0)} \sum_{n=1}^{\infty} | P(n\tau) | + |P(-n\tau) |. \tag{13.07}$$

This equation may be employed to determine the maximum possible pulsing rate for a given impulse characteristic and number of pulse amplitudes, obtained when $M/M_{\max} = 0$, or to determine the margin for a given pulse transmission rate. An example of the latter application is illustrated in Fig. 43, which shows the margin $M/M_{\max}$ in per cent, obtained when (13.07) with $q = 2$ is applied to the curves shown in Fig. 23 for various degrees of delay distortion. The pulse interval is taken as $\tau = 1/2f_1 = 1/f_{\max}$, where f_1 is the frequency at the 6 db down

point on the amplitude characteristic and $f_{max} = 2f_1$. Under this condition there is no intersymbol interference in the absence of phase distortion.

The above equations apply to peak intersymbol interference, obtained by taking the maximum positive and negative values of the summation term in (13.01). As discussed in previous sections, certain types of transmission system imperfections give rise to pulse distortion extending over long time intervals, such as fine structure deviations over the transmission band, a low-frequency cut-off and pronounced band-edge phase deviations. Evaluation of peak intersymbol interference is then rather difficult, and a more convenient approximate method is to evaluate rms intersymbol interference, which can be related to rms deviation in the transmission frequency characteristic by methods discussed previously. Peak intersymbol interference may then be estimated by applying a peak factor between 3 and 4, depending on the type of transmission distortion.

If $P_0(n\tau)$ designates an ideal impulse characteristic, which is zero for $n = \pm 1, \pm 2$ etc., the deviation from the ideal envelope of a pulse

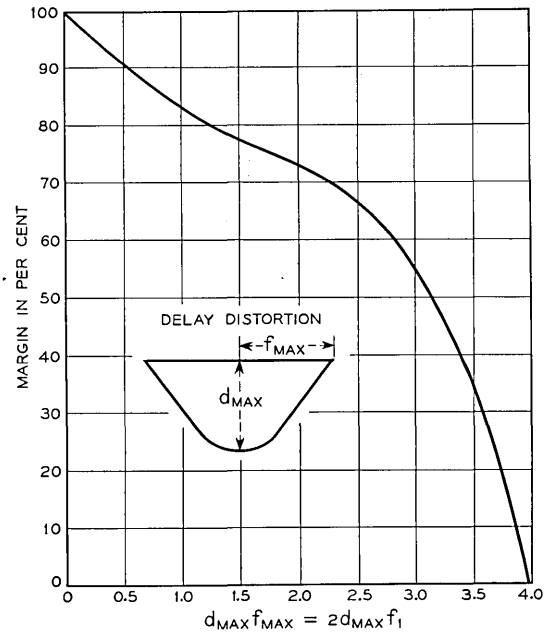


Fig. 43 — Margin against excessive peak interference in systems employing two pulse amplitudes with intervals between pulses $\tau = \tau_1 = 1/2f_1 = 1/f_{max}$. for impulse transmission characteristic as shown in Fig. 23.

train may be written

$$\Delta W(0) = W(0) - W_0(0) = \sum_{n=-\infty}^{\infty} A_n[P(n\tau) - P_0(n\tau)]. \quad (13.08)$$

The rms deviation becomes, with $\Delta P(n\tau) = P(n\tau) - P_0(n\tau)$

$$\underline{\Delta} W(0) = \underline{A} \left(\sum_{n=-\infty}^{\infty} [\Delta P(n\tau)]^2 \right)^{1/2}, \quad (13.09)$$

$$\cong \underline{A} \left(\frac{1}{\tau} \int_{-\infty}^{\infty} [\Delta P(t)]^2 \, dt \right)^{1/2}, \quad (13.10)$$

$$= \underline{A} P(0) \underline{U}. \quad (13.11)$$

$\underline{A}$ is the rms amplitude of the transmitted pulses and $\underline{U}$ the rms intersymbol interference referred to unit amplitude of the received pulses. Expressions for $\underline{U}$ applying to fine structure imperfections in the transmission frequency characteristic were given in Section 8, for a low-frequency cut-off in Section 9 and for band-edge phase deviations in Section 10.

For balanced pulse systems employing positive and negative pulses, rms intersymbol interference in the positive and negative directions will be equal. For such systems the maximum value of the summation in (13.02) becomes $k\underline{W}(0)$ and in (13.03) $-k\underline{W}(0)$, where k is the peak factor. Equation (13.04) is then replaced by

$$\begin{aligned} M &= \frac{A_{\max} - A_{\min}}{q - 1} P(0) - 2k\underline{A} P(0) \underline{U}, \\ &= 2A_{\max} P(0) \left[\frac{1}{q - 1} - k\underline{U}(\underline{A}/A_{\max}) \right], \end{aligned} \quad (13.12)$$

when $A_{\min} = -A_{\max}$.

In a balanced pulse system employing q pulse amplitudes, i.e., $q/2$ positive and $q/2$ negative amplitudes, with equal steps $2A_{\max}/(q - 1)$ between pulse amplitudes, the following relation applies if all amplitudes have equal probability.

$$\underline{A}/A_{\max} = \left[\frac{q + 1}{3(q - 1)} \right]^{1/2}. \quad (13.13)$$

Hence,

$$M = \frac{2A_{\max} P(0)}{q - 1} \left[1 - k \left(\frac{q^2 - 1}{3} \right)^{1/2} \underline{U} \right]. \quad (13.14)$$

As mentioned before, the factor k may be as high as 4, in which case the

maximum tolerable rms intersymbol interference $\underline{U}$ referred to unit peak amplitude of the received pulses becomes for $M = 0$:

$q =$	$\underline{2}$	$\underline{4}$	$\underline{8}$
$\underline{U} =$	0.25	0.112	0.054

In (13.14) and in the above table, $\underline{U}$ is the maximum tolerable rms intersymbol interference from all sources, such as fine structure imperfections over the transmission band, band-edge phase distortion and a low-frequency cut-off. Interference from these various sources may be combined on a root-sum-square basis.

In the above evaluation of rms intersymbol interference a balanced pulse system was assumed. An unbalanced system can be obtained by superposing on a balanced system an infinite sequence of pulses of equal amplitude and polarity at uniform intervals as indicated in Fig. 44. This superposed system will give rise to a fixed intersymbol interference or displacement of the received pulse train, which does not alter the margin for distinction between pulse amplitudes and which can be corrected by a fixed bias at the receiving end if necessary. For this reason, in the case

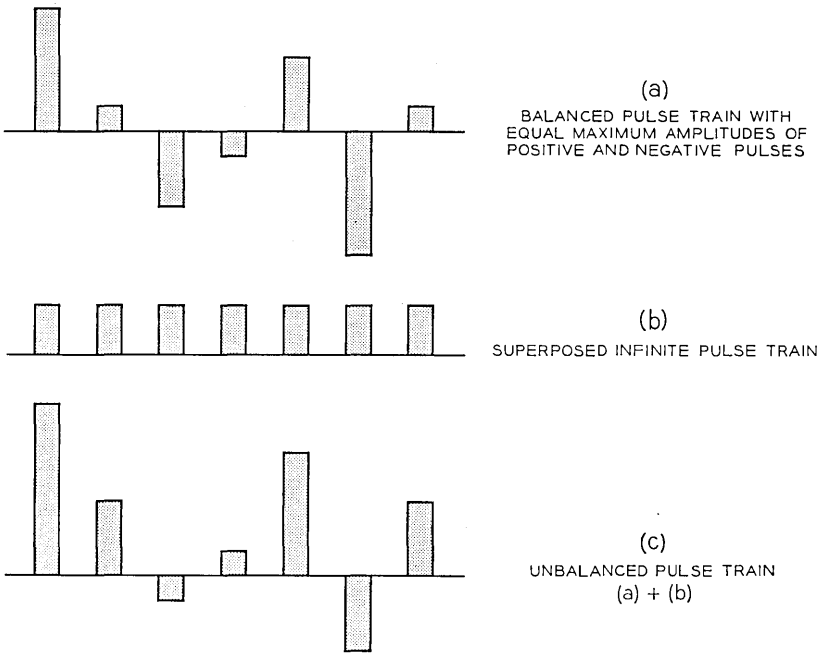


Fig. 44 Derivation of an unbalanced from a balanced pulse train by superposition of an infinite train of pulses of equal amplitude.

of an unbalanced system, only the balanced component need to be considered in evaluating rms intersymbol interference, which will thus be the same whether or not the system is balanced. As shown previously, peak intersymbol interference, or the margin for distinction between pulse amplitudes, depends only on the peak to peak pulse excursion and is thus the same for unbalanced as for balanced systems. It may be noted here that for a balanced system the transmitted power is a minimum for a given margin in pulse reception, as is the interference in other systems that may be caused by the transmitted pulses.

For a symmetrical band-pass system, rather than a low-pass system as discussed above, $Q(n\tau) = 0$ in (12.08). The envelope of the pulse train then becomes

$$\overline{W}(0) = \sum_{n=-\infty}^{\infty} A_n R(n\tau), \tag{13.15}$$

where $R(n\tau) = R_-(n\tau) + R_+(n\tau) = 2R_+(n\tau)$, with R_- and R_+ given by (2.10).

Since (13.15) is of the same form as (13.01), the relationships established above for low-pass systems also apply to symmetrical band-pass systems, with $R(n\tau)$ replacing $P(n\tau)$. $R(n\tau)$ will have the same shape as $P(n\tau)$, but will be greater by a factor 2, which will appear as a multiplier in the various expressions and hence not alter the requirements on tolerable pulse distortion or intersymbol interference.

14. TRANSMISSION LIMITATIONS IN ASYMMETRICAL SIDEBAND SYSTEMS

The formulation of transmission limitations imposed by pulse distortion in asymmetrical sideband systems is complicated by the presence of the quadrature component in the impulse transmission characteristic. Of particular interest are the transmission limitations with vestigial sideband as compared with double sideband transmission, assuming the same bandpass characteristic in both cases, a question which has been dealt with in literature for systems with a linear phase characteristic[4,11] Relationships (2.18) and (2.19) facilitate a comparison also for systems with phase distortion, as shown in the following.

If the envelope of the impulse characteristic with double sideband transmission is $\overline{P}(t)$, the in-phase and quadrature components with vestigial sideband transmission are given by (2.19), with $\omega_y = \omega_s$ or

$$\begin{aligned} R &= R_- + R_+ = \cos(\omega_s t - \psi_s)\,\overline{P}(t), \\ Q &= Q_- - Q_+ = \sin(\omega_s t - \psi_s)\,\overline{P}(t). \end{aligned} \tag{14.01}$$

If t is so chosen that $\omega_s t - \psi_s = 0$, and the time with respect to this value of t is designated t_0, then

$$\begin{aligned} R(t_0) &= \cos \omega_s t_0 \bar{P}(t_0), \\ Q(t_0) &= \sin \omega_s t_0 \bar{P}(t_0). \end{aligned} \tag{14.02}$$

An application of this method to the impulse characteristic shown in Fig. 23 for $b = 15$ radians is illustrated in Fig. 45.

In order to compare vestigial with double sideband transmission, it suffices to evaluate the in-phase and quadrature components at the sampling instants. With $\tau = \pi/2\omega_s = 1/4f_s$, the in-phase and quadrature components at times $m\tau$, for $m = 0 \pm 1, \pm 2$, etc., will be as illustrated in Fig. 46.

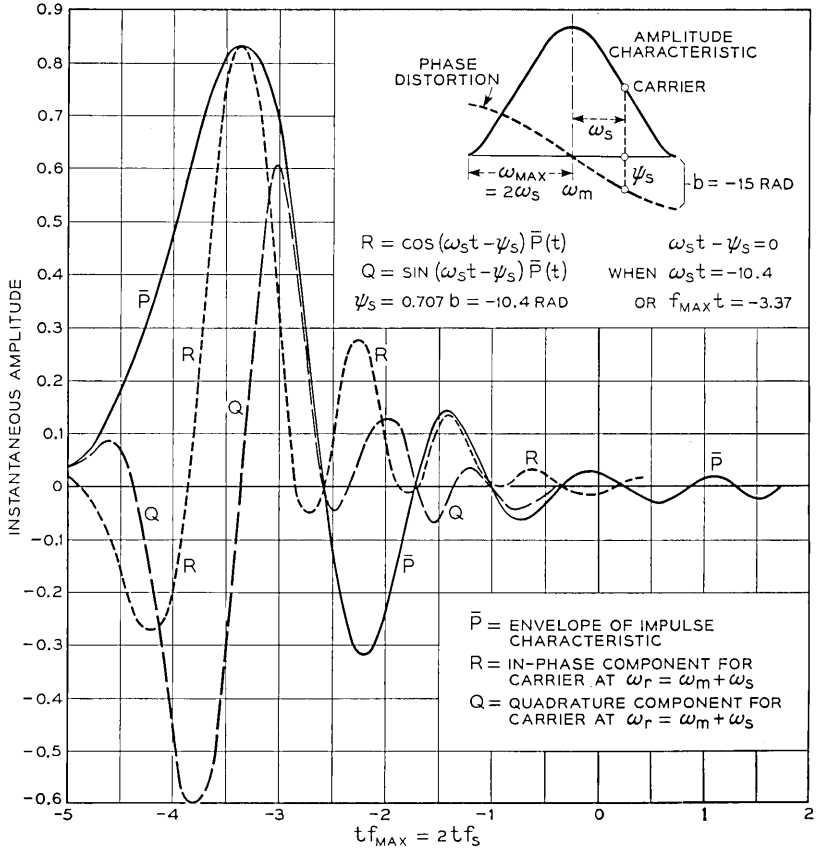
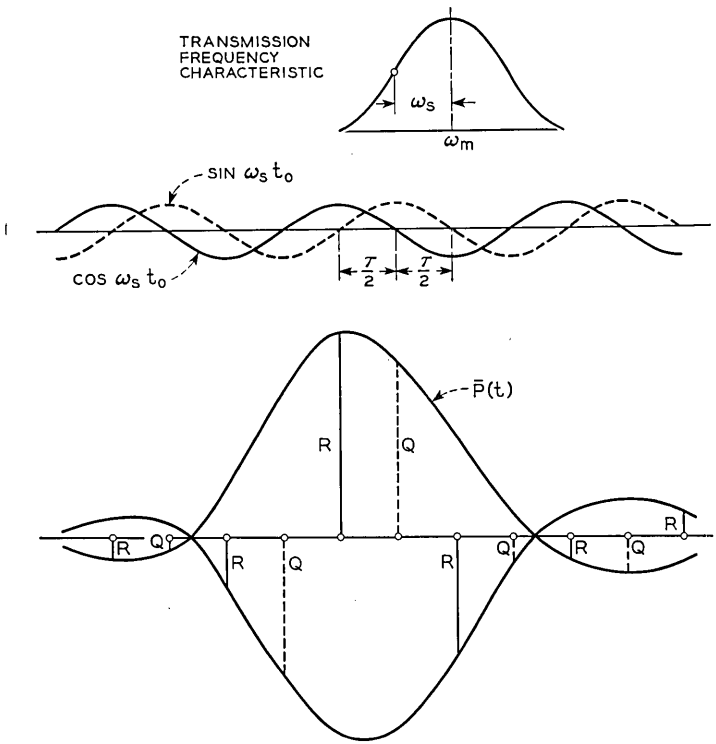


Fig. 45 — Determination of in-phase and quadrature components of impulse characteristic for vestigial side-band transmission.

With double sideband transmission, pulses would be transmitted at the points $m = 0, \pm 2, \pm 4$, etc. At these points the quadrature components vanish, as indicated in the above figure, and the in-phase components are the same in amplitude as with double sideband transmission. Thus, if pulses were transmitted at the same rate as with double sideband transmission, the sum of the absolute values of the in-phase components at the sampling points would be identical with the sum of the absolute values of the envelope with double sideband transmission. It follows from the criteria established in Section 13 that for this particular pulse transmission rate the effect of pulse distortion would be the same with both transmission methods. With an ideal transmission frequency



$\bar{P}(t)$ = ENVELOPE OF IMPULSE CHARACTERISTIC
R = IN-PHASE COMPONENTS AT SAMPLING INSTANTS
Q = QUADRATURE COMPONENTS AT SAMPLING INSTANTS
$\tau = 1/\omega_s$ = PULSE INTERVALS WITH DOUBLE SIDE BAND TRANSMISSION

Fig. 46 — In-phase and quadrature components of impulse characteristic with vestigial side-band transmission.

characteristic having a linear phase shift, there would be no intersymbol interference with either method for the above rate of pulse transmission.

Assume next that the pulse transmission rate is doubled and that the quadrature component is eliminated. This is possible if the carrier frequency is transmitted and is derived at the receiving end with the aid of filters and applied in proper phase to a product demodulator, a method known as homodyne detection. At the points $m = 1, 3, 5$, etc., there would then be no quadrature components and no in-phase components. The sum of the absolute values of the in-phase components at the other sampling points, $m = 2, 4$, etc., would be the same as with double sideband transmission. It follows that the transmission capacity (pulsing rate) can be doubled by vestigial sideband transmission if the quadrature component is eliminated by homodyne detection, for the same margin against excessive intersymbol interference as with double sideband transmission.

An increase in transmission capacity can be realized with vestigial sideband transmission without elimination of the quadrature component by homodyne detection, although a two-fold increase is then possible only if the phase characteristic is linear, as discussed below. Vestigial sideband transmission can be employed without transmission of the carrier, or with a fixed level of carrier in the absence of pulses and a higher level in the presence of pulses. The latter method is equivalent to the transmission of two or more pulse amplitudes, with the minimum amplitude greater than zero. With this method the effect of the quadrature component on the envelope of a pulse train can be reduced, and even eliminated provided the phase characteristic is linear. In the following, vestigial sideband transmission with two pulse amplitudes at twice the double sideband pulsing rate is discussed, for the case in which the minimum pulse amplitude is finite rather than zero.

With pulses transmitted at twice the double sideband rate, i.e., with the interval between pulses equal to $\tau = \pi/2\omega_s$, equation (12.08) for the envelope becomes in view of (14.02)

$$\overline{W}(0) = \left(\left[\sum_{-\infty}^{\infty} A_n \cos \omega_s n\tau \overline{P}(n\tau)\right]^2 + \left[\sum_{-\infty}^{\infty} A_n \sin \omega_s n\tau \overline{P}(n\tau)\right]^2\right)^{1/2}. \qquad (14.03)$$

At the even sampling points, i.e., $n = 0, 2, 4 \cdots$, $\cos \omega_s n\tau = \pm 1$ and the in-phase components may be written

$$R(\pm 2m\tau) = \pm \overline{P}(\pm 2m\tau), \qquad m = 0, 1, 2 \cdots$$

At the odd sampling points, i.e., $n = 1, 3, 5 \cdots$, $\sin \omega_s n\tau = \pm 1$ and

the quadrature components may be written

$$Q[\pm(2m-1)\tau] = \pm\overline{P}[\pm(2m-1)\tau], \qquad m = 1, 2, 3$$

Let

$$\begin{aligned}
\sum R^+ &= \sum_{m=1}^{\infty} [R^+(2m\tau) + R^+(-2m\tau)], \\
\sum R^- &= \sum_{m=1}^{\infty} [R^-(2m\tau) + R^-(-2m\tau)], \\
\sum Q^+ &= \sum_{m=1}^{\infty} Q^+[(2m-1)\tau] + Q^+[-(2m-1)\tau], \\
\sum Q^- &= \sum_{m=1}^{\infty} Q^-[(2m-1)\tau] + Q^-[-(2m-1)\tau],
\end{aligned} \tag{14.04}$$

where R^+, Q^+ designate positive values and R^-, Q^- the absolute values of negative amplitudes of the in-phase and quadrature components.

Let it be assumed that two pulse amplitudes are employed, $A_{\min}$ and $A_{\max}$. When the minimum amplitude is transmitted, the maximum value of the envelope is obtained by considering the maximum positive overlaps of the in-phase components in conjunction with the maximum value of the quadrature component. The value thus obtained is

$$\begin{aligned}
\overline{W}_{\max} = [(A_{\min} R(0) + A_{\max} \textstyle\sum R^+)^2 \\
+ (A_{\max} \textstyle\sum Q^- - A_{\min} \textstyle\sum Q^+)^2]^{1/2}
\end{aligned} \tag{14.05}$$

It is assumed that $\sum Q^- > \sum Q^+$, otherwise Q^- and Q^+ would be interchanged in the last term.

When the maximum amplitude is transmitted, the minimum value of the envelope is obtained by considering the maximum negative overlaps of the in-phase components, in conjunction with the minimum value of the quadrature component, which gives

$$\begin{aligned}
\overline{W}_{\min} = [(A_{\max} R(0) - A_{\max} \textstyle\sum R^-)^2 + A^2_{\min} (\textstyle\sum Q^- \\
- \textstyle\sum Q^+)^2]^{1/2}.
\end{aligned} \tag{14.06}$$

The margin for distinction between $A_{\min}$ and $A_{\max}$ is $M = W_{\min} - W_{\max}$ and becomes

$$\begin{aligned}
M = A_{\max} [(R(0) - \textstyle\sum R^-)^2 + \mu^2(\textstyle\sum Q^+ - \textstyle\sum Q^-)^2]^{1/2} \\
- A_{\max} [(\mu R(0) + \textstyle\sum R^+)^2 + (\textstyle\sum Q^- - \mu \textstyle\sum Q^+)^2]^{1/2},
\end{aligned} \tag{14.07}$$

where

$$\mu = A_{\min}/A_{\max}.$$

The margin for a unit difference $A_{max} - A_{min}$, i.e. $M_1 = M/(A_{max} - A_{min})$ becomes:

$$M_1 = \frac{1}{1-\mu}\left([(R(0) - \sum R^-)^2 + \mu^2(\sum Q^- - \sum Q^+)^2]^{1/2} - [(\mu R(0) + \sum R^+)^2 + (\sum Q^- - \mu \sum Q^+)^2]^{1/2}\right). \quad (14.08)$$

The special case of an ideal transmission characteristic as shown in Fig. 47 will be considered first. In this case

$$R(0) = 1 \qquad R(2\tau) = 0 \qquad R(-2\tau) = 0$$
$$R(4\tau) = 0 \qquad R(-4\tau) = 0$$
$$Q(\tau) = 0.5 \qquad Q(-\tau) = -0.5$$
$$Q(3\tau) = 0 \qquad Q(-3\tau) = 0$$

so that:

$$\sum R^+ = 0 \qquad \sum R^- = 0$$
$$\sum Q^+ = 0.5 \qquad \sum Q^- = 0.5$$

Equation (14.08) in this case simplifies to

$$M_1 = \frac{1}{1-\mu}\left(1 - \left[\mu^2 + \frac{1}{4}(1-\mu)^2\right]^{1/2}\right). \quad (14.09)$$

For various values of $\mu = A_{min}/A_{max}$ the margin for unit difference in

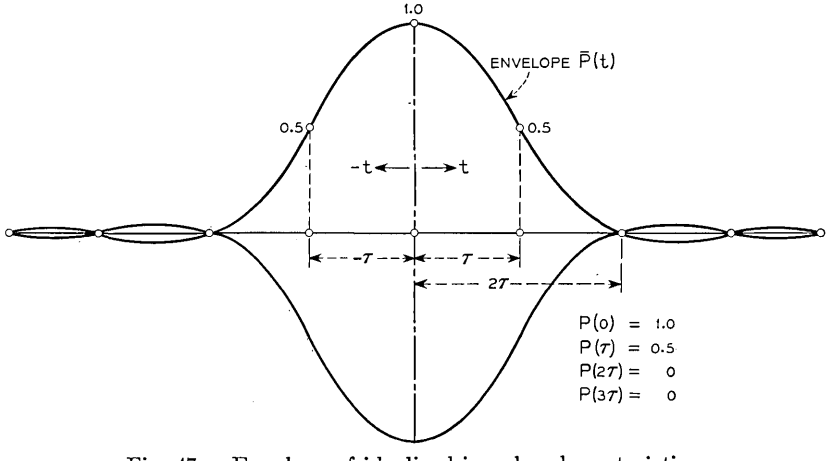


Fig. 47 — Envelope of idealized impulse characteristic.

pulse amplitudes becomes:

μ	0	0.2	0.3	0.5	0.8	0.9	1.0
M_1	0.50	0.69	0.77	0.87	0.96	0.998	1.0

Thus, for an ideal impulse characteristic as assumed above, the quadrature component gives rise to 50 per cent maximum intersymbol with $\mu = 0$, and to negligible intersymbol interference when $\mu \geqq 0.8$ or greater. By way of comparison, the margin would be zero with double sideband transmission at the rate assumed above, i.e., twice the normal double sideband rate. This follows from (13.07) when it is considered that $P(\pm\tau) = \frac{1}{2} P(0)$, $P(\pm 2\tau) = 0$, so that the sum of the absolute values of the impulse characteristic at the sampling points is equal to $P(0)$ and thus $M/M_{\max} = 0$.

Elimination of the effect of the quadrature component by the above method is contingent on a symmetrical impulse characteristic, i.e., $\overline{P}(n\tau) = \overline{P}(-n\tau)$, a condition which can be realized only with a linear phase shift. Furthermore, the in-phase components must vanish at the sampling points, which entails an ideal amplitude characteristic. In the presence of phase distortion the effect of the quadrature component cannot be eliminated but may be reduced by proper choice of the ratio μ, as discussed below for a transmission characteristic with moderate phase distortion.

As an example consider an impulse characteristic as shown in Fig. 23 for $b = 5$ radians. The in-phase and quadrature components at the various sampling points are in this case

$$R(0) = 0.97 \qquad R(-2\tau) = -0.09 \qquad R(2\tau) = 0.13$$
$$R(-4\tau) \geqq 0 \qquad R(4\tau) \geqq 0$$
$$Q(-\tau) = -0.54 \qquad Q(\tau) = 0.44$$
$$Q(-3\tau) \geqq 0 \qquad Q(3\tau) = -0.03$$
$$Q(-5\tau) \geqq 0 \qquad Q(5\tau) \geqq 0$$

Hence

$$\sum R^+ = 0.13 \qquad \sum R^- = 0.09 \qquad \sum Q^+ = 0.44 \qquad \sum Q^- = 0.57$$

Equation (14.08) in this case becomes

$$M_1 = \frac{1}{1-\mu}\left((0.88^2 + 0.13^2\mu^2)^{1/2} - [(0.97\mu + 0.13)^2 + (0.57 - 0.44\mu)^2]^{1/2}\right).$$

For various values of $\mu = A_{\min}/A_{\max}$ the margin for unit difference in pulse amplitudes becomes

μ	0	0.2	0.3	0.4	0.5	0.6	0.7	0.75
M_1	0.30	0.375	0.40	0.375	0.34	0.25	0.13	0

The optimum condition is thus in the above particular case obtained with $\mu \leqq 0.3$, with a comparatively small variation in transmission performance for any value of μ between 0 and 0.5.

In the above discussion of vestigial sideband transmission, modulation of a carrier was assumed, with elimination of one sideband except for the wanted vestige. The equivalent performance can be secured by application of impulses to a band-pass transmission characteristic with the proper interval between pulses in relation to the midband frequency, as discussed below:

When equation (12.03) is written with respect to the midband frequency, $\omega_r = \omega_m$, and a symmetrical amplitude characteristic is assumed so that $Q = 0$, the following relation obtained.

$$\begin{aligned} W(t_0) &= \cos \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \cos \omega_m n\tau R(t_0 + n\tau) \\ &\quad - \sin \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \sin \omega_m n\tau R(t_0 + n\tau), \end{aligned} \tag{14.10}$$

in which R may be replaced by $\overline{P}$, the envelope of the impulse characteristic.

Let it be assumed that τ is so chosen that $\cos \omega_m n\tau = \cos n\pi/2$ in which case $\sin \omega_n n\tau = \sin n\pi/2$. The above equation then becomes

$$\begin{aligned} W(t_0) &= \cos \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \overline{P}(t_0 + n\tau) \cos n\pi/2 \\ &\quad - \sin \omega_m t_0 \sum_{n=-\infty}^{\infty} A_n \overline{P}(t_0 + n\tau) \sin n\pi/2. \end{aligned} \tag{14.11}$$

The in-phase and quadrature components of the envelope at the sampling instant $t_0 = 0$ are accordingly

$$\begin{aligned} \bar{R}(0) &= \sum_{n=-\infty}^{\infty} A_n \overline{P}(n\tau) \cos n\pi/2, \\ \bar{Q}(0) &= \sum_{n=-\infty}^{\infty} A_n \overline{P}(n\tau) \sin n\pi/2. \end{aligned} \tag{14.12}$$

Pulses in even positions, i.e., A_0, A_2, A_4, etc., will thus contribute an in-phase but no quadrature component while pulses in odd positions

A_1, A_3, A_5, etc., will contribute a quadrature but no in-phase component. It will be recognized from Fig. 42 that this is the same condition as encountered in vestigial sideband transmission with pulses in the latter case transmitted at intervals $\tau = \pi/2\omega_s = 1/4\, f_s$.

To realize the above condition with pulses applied to a band-pass filter, it is necessary that in (14.10)

$$\omega_m n\tau = n\pi(\tfrac{1}{2} + N), \tag{14.13}$$

where N is an integer, or that

$$\tau = \frac{\pi(1 + 2N)}{2\omega_m} = \frac{1 + 2N}{4f_m}. \tag{14.14}$$

The interval between pulses must thus be an integral number of half-cycles plus one quarter cycle of the midband frequency f_m, as illustrated for a particular case in Fig. 48. When f_m is large in relation to the sideband frequency this condition can be achieved with substantially the same pulse spacing as with vestigial sideband transmission. To secure exactly the same rate of pulse transmission it is necessary that

$$\tau = 1/4\, f_s,$$

which, in conjunction with (14.14), gives

$$N = \tfrac{1}{2}\,(f_m/f_s - 1). \tag{14.15}$$

Thus, if $f_m = 5f_s$, $N = 2$ and the interval τ between pulses as obtained from (14.14) is 1.25 cycles of f_m. If $f_m = 10f_s$, $N = 4.5$ and it is not possible to have exactly the same pulsing rate as with vestigial sideband transmission, since N must be an integer. It is then necessary to take $N = 4$ or 5. With $N = 4$ equation (14.14) gives $\tau = 9/40f_s$ and with $N = 5$, $\tau = 11/40f_s$. This compares with $\tau = 1/4f_s = 10/40f_s$ with vestigial sideband transmission, so that there is a minor difference in pulse intervals with the two methods.

15. DOUBLE VERSUS VESTIGIAL SIDEBAND SYSTEMS

From the preceding discussion it follows that, for the same bandwidth and margin against interference from characteristic distortion, a two-fold increase in transmission capacity can be approached with vestigial over double sideband transmission. This assumes that the carrier is transmitted at the proper level and that the phase characteristic is linear, or that otherwise homodyne detection is used to cancel the effect of the quadrature component.

For the same bandwidth, the same transmission capacity can be realized with a double sideband system employing four pulse amplitudes as with a vestigial sideband system with two pulse amplitudes. However, the latter type of system will have a greater tolerance to interference from characteristic distortion than the former. This follows when it is considered that in a quaternary system the maximum tolerable interference is $\frac{1}{6}$ the maximum pulse amplitude, as compared to $\frac{1}{2}$ the maximum pulse amplitude in a binary system. With $\mu = A_{min}/A_{max} = 0$, the quadrature component reduces the margin by a factor of 0.5, so that the maximum tolerable interference in relation to the maximum pulse

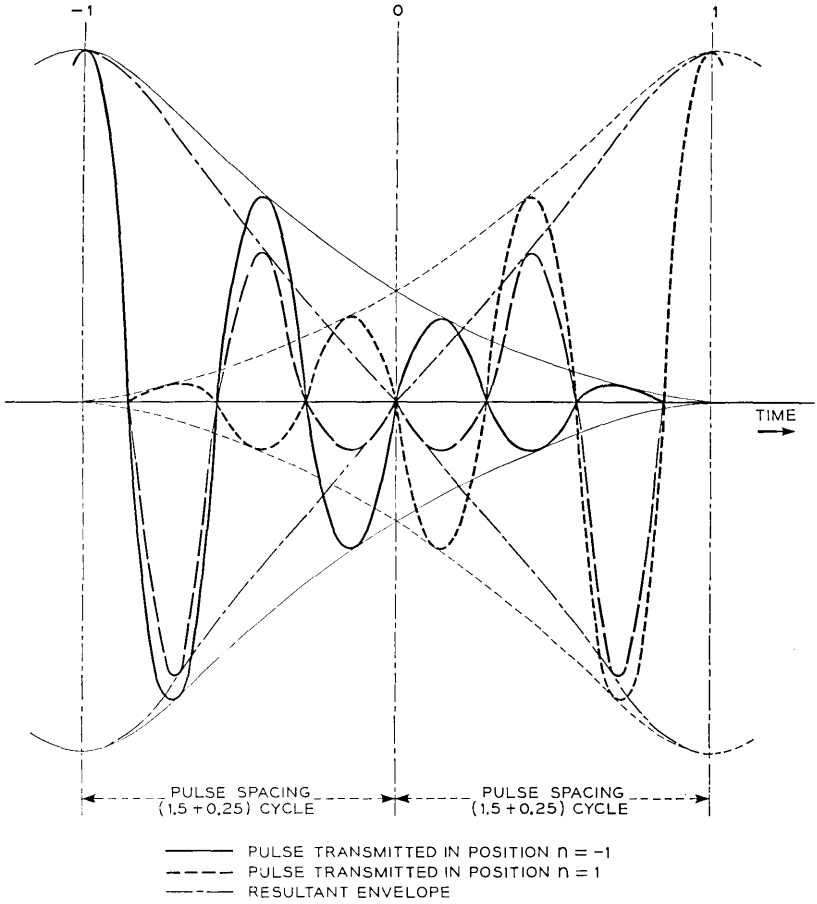


Fig. 48 — Impulse exitation of band-pass system with pulse spacing selected to provide equivalent of vestigial side-band transmission.

amplitude is ¼ as compared to ⅙ for a quaternary system. If the phase characteristic is linear and the carrier is transmitted at the optimum level, or if homodyne detection is used, the effect of the quadrature component is cancelled. The maximum tolerable interference is then ½ as compared with ⅙ for a quarternary double sideband system.

In the presence of phase distortion, a substantial advantage can also be realized with a binary vestigial system, which can be illustrated by considering the example in Section 14. For the optimum condition $\mu = 0.4$, the margin is reduced by a factor 0.4 and is thus 0.2. For a quaternary double sideband system the factor by which the margin is reduced is given by (13.07), with $q = 4$ and with

$$\frac{1}{P(0)} \sum_{n=1}^{\infty} | P(n\tau) | + | P(-n\tau) | = \frac{1}{R(0)} [\sum R^{+} + \sum R^{-}],$$

where $R(0) = 0.97$, $\sum R^{+} = 0.13$ and $\sum R^{-} = 0.09$, as in the example in Section 14. The reduction in margin thus obtained is $M/M_{\max} = 0.32$. Hence the maximum tolerable interference for a quarternary double sideband system is $0.32/6 = 0.053$ as compared with 0.20 for a binary vestigial sideband system under the optimum condition $\mu = 0.4$.

For the same transmission capacity and same number of pulse amplitudes, a substantial transmission advantage may be realized with vestigial over double sideband transmission in circuits with pronounced phase distortion, owing to the circumstance that a two-fold reduction in bandwidth with vestigial sideband transmission may afford a sub-

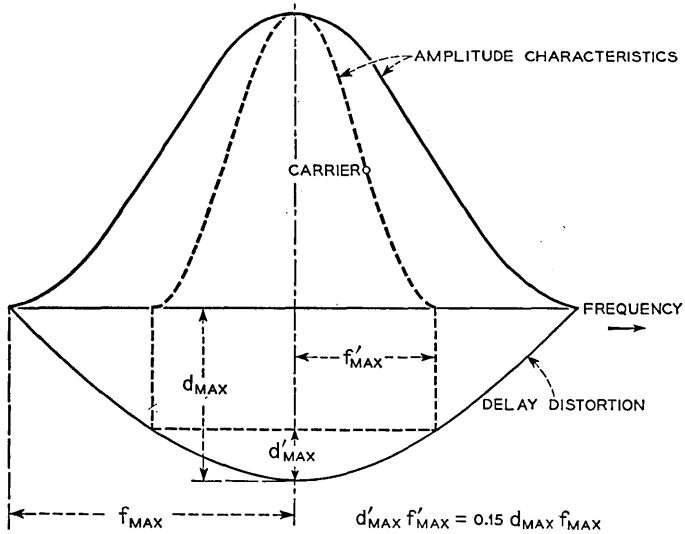


Fig. 49 — Comparison of double and vestigial side-band transmission in the presence of delay distortion.

stantial reduction in delay distortion over the transmission band. This is illustrated in Fig. 49, where a cosine variation in transmission delay is assumed. With a two-fold reduction in bandwidth, the product $d'_{max}f'_{max}$ for vestigial sideband transmission is about 15 per cent of the product $d_{max}f_{max}$ for double sideband transmission. Thus, with $d_{max}f_{max} = 8.3$, $d'_{max}f'_{max} = 1.25$, corresponding to $b = 5$ radians, as assumed in the example in Section 14. Vestigial sideband transmission is in this case feasible with an adequate margin, about 40 per cent of the maximum margin in the absence of phase distortion. Double sideband transmission would not be possible, as is evident from Fig. 43, since it would be necessary to have $d_{max}f_{max}$ less than 4, as compared with 8.3 in the above case.

The above discussion of vestigial vs double sideband transmission pertains to the effects of characteristic distortion rather than noise, and the relative complexity of terminal equipment was disregarded. Because of the simpler terminal equipment with double sideband transmission, this method is ordinarily used where bandwidth is not a primary consideration, as for example in providing a number of telegraph channels over a voice frequency circuit.

16. LIMITATION ON CHANNEL CAPACITY BY CHARACTERISTIC DISTORTION

For an idealized channel of bandwidth f_1 with a transmission-frequency characteristic as shown in Fig. 7, the transmission capacity in bits per second for a signal of average power P in the presence of random noise of average power N can with sufficiently complicated encoding methods approach the limiting value given by Shannon:[12]

$$C = f_1 \log_2 (1 + P/N). \tag{16.01}$$

The above expression also applies to certain other idealized channels with a linear phase characteristic, when f_1 is defined as in Fig. 10. In all of these cases the integral of the area under the amplitude characteristic, or the equivalent bandwidth, is f_1.

By way of comparison, for pulse code modulation systems the channel capacity is of the same basic form as (16.01), namely[13]:

$$C = f_1 \log_2 \left(1 + \frac{P}{KN}\right), \tag{16.02}$$

where $K \cong 8$. Thus about an 8-fold increase in signal power is required to attain the same channel capacity as with the idealized but impracticable encoding system underlying (16.01).

The above expressions give the limitation on channel capacity imposed by random noise. From the discussion in Sections 13 and 14 it follows that a limitation is placed on channel capacity by characteristic distor-

tion, in the absence of noise. In idealized communication theory, characteristic distortion has been disregarded in determining channel capacity on the premise that unlike random noise it is predictable and can therefore be corrected, at least in principle. In actual systems, however, complete elimination though possible in principle cannot be accomplished in practice. The resultant limitation on transmission capacity may be as important as that imposed by the maximum signal power that can actually be provided to override noise.

In the following it will be assumed that correction of amplitude and phase deviations is made by equalization, so that the amplitude and phase characteristics are as assumed for an ideal channel, except for small fine structure residual deviations as illustrated in Fig. 30. These small fine structure deviations may be regarded as of random nature in the sense that they differ among channels and cannot be predicted, although for a given system they would remain fixed in the absence of temperature variations or changes in amplifiers with age.

From equation (13.12) it follows that the maximum number of pulse amplitudes or quantizing levels as limited by characteristic distortion is obtained from the relation

$$\frac{1}{q-1} = k\underline{U}\underline{A}/A_{\max}, \tag{16.03}$$

or

$$q = 1 + \frac{1}{k\underline{U}} A_{\max}/\underline{A}. \tag{16.04}$$

In the absence of characteristic distortion, the maximum number of pulse amplitudes as limited by an rms noise amplitude $\underline{A}_n$ or a peak noise amplitude $k\underline{A}_n$ is obtained from the following relation for a balanced pulse system.

$$\frac{A_{\max}}{q-1} = k\underline{A}_n, \tag{16.05}$$

or

$$q = 1 + \frac{\underline{A}}{k\underline{A}_n} A_{\max}/\underline{A}. \tag{16.06}$$

Comparison of (16.04) and (16.06) shows the following equivalence between intersymbol interference and noise from the standpoint of limitation on the permissible number of pulse amplitudes

$$\underline{U} = \underline{A}_n/\underline{A}, \tag{16.07}$$

or

$$\underline{U}^2 = D = N/P. \tag{16.08}$$

This means that random characteristic distortion has the same effect as a random noise power $N = DP$, where D is a distortion factor.

In view of the above equivalence, the channel capacity of a PCM system in the presence of random characteristic distortion, but without noise, as obtained by substitution of (16.08) in (16.02) becomes

$$C = f_1 \log_2 \left(1 + \frac{1}{KD}\right). \tag{16.09}$$

With random interference from both characteristic distortion and noise, the interfering powers add directly, so that for a PCM system

$$C = f_1 \log_2 \left(1 + \frac{1}{K(D + N/P)}\right). \tag{16.10}$$

The equivalence (16.08) was established above on the basis of discrete pulse amplitudes, but it is independent of q and would thus apply also for continuous signals. On this basis it would apply for any method of modulation or of encoding signals and the maximum channel capacity as given by (16.01) would be modified to

$$C = f_1 \log_2 \left(1 + \frac{1}{D + N/P}\right), \tag{16.11}$$

It follows from the above that for any modulation method the tolerable distortion factor is directly related to the average signal-to-noise ratio. Thus two modulation methods which are equivalent from the standpoint of signal-to-noise ratio are also equivalent from the standpoint of tolerable rms distortion, provided faithful reproduction of the transmitted signal is required, as assumed here.

From (8.14) the following relation is obtained between the distortion factor $D = \underline{U}^2$ and small rms deviations $\underline{a}$ (nepers) and $\underline{b}$ (radians) in the amplitude and phase characteristics

$$D = \underline{a}^2 + \underline{b}^2. \tag{16.12}$$

In order that characteristic distortion may be disregarded in comparison with noise, it is necessary that $D \ll N/P$ or

$$\underline{a}^2 + \underline{b}^2 \ll N/P. \tag{16.13}$$

For example, in communication systems employing the same bandwidth as the original signal, such as a pulse amplitude modulation system, a representative signal-to-noise ratio would be about 40 db, or $N/P = 10^{-4}$. In order that characteristic distortion may be disregarded in this case, it would be necessary for both $\underline{a}$ and $\underline{b}$ to be substantially

less than 10^{-2} nepers and radians respectively. This would correspond to an rms gain deviation over the transmission band well below 0.08 db and an rms deviation from a linear phase characteristic well below 0.6 degrees. Since these tolerances are difficult to realize in actual systems, at least for wire circuits, characteristic distortion rather than noise may impose a limitation on channel capacity of systems employing about the same bandwidth as the original signal.

In accordance with (16.01), the bandwidth can in principle be halved without change in channel capacity if the signal-to-noise ratio is squared, i.e., if $N/P = 10^{-8}$ rather than 10^{-4} in the previous example. The tolerable rms amplitude and phase deviations would then be

$$\underline{a}^2 + \underline{b}^2 \ll 10^{-8}.$$

Thus both a and b would have to be substantially smaller than about 10^{-4}, which would preclude a substantial bandwidth saving in practical systems from the standpoint of characteristic distortion, even if it were feasible from the standpoint of signal power required to override noise.

The above considerations apply when faithful reproduction of the transmitted signal is required, as for example in data transmission. In speech transmission considerable distortion can be tolerated, a circumstance which permits appreciable phase distortion in the usual frequency division system without noticeable impairment of intelligibility, but which cannot be taken advantage of in time division pulse systems because of the resultant intersymbol interference. The characteristics of speech sounds also permit a reduction in the bandwidth of the original transmitted signal, by such devices as vocoders or frequency compandors, without excessive impairment of intelligibility.

ACKNOWLEDGMENTS

This paper is based on studies in connection with the application of pulse systems to wire circuits, suggested by M. L. Almquist and J. T. Dixon and carried out under their direction. Some of these studies were made on behalf of the Signal Corps, under contract W-36-039-SC-38115. The Signal Corps Engineering Laboratories have consented to the publication of results obtained in these studies. The writer had the benefit in these studies of discussions with C. B. Feldman and W. R. Bennett, and of some of their memoranda on pulse modulation systems. He is also thankful for comments and advice from F. B. Llewellyn and others in connection with preparation of the paper.

REFERENCES

See Part I.

Bell System Technical Papers Not Published in this Journal.

ALLIS, W. P., see ROSE, D. J.

ANDERSON, P. W.,[1] MERRITT, F. R.,[1] REMEIKA, J. P.,[1] and YAGER, W. A.[1]

Magnetic Resonance in Fe_2O_3, Phys. Rev., **93**, pp. 717–718, Feb. 15, 1954.

BICKELHAUPT, C. O.[2]

Fifty Years Ago, Telephony, **146,** pp. 20–22, Feb. 20, 1954.

CLOGSTON, A. M.[1] and HEFFNER, H.[1]

Focussing of an Electron Beam by Periodic Fields, J. Appl. Phys., **25,** pp. 436–447, April, 1954.

CONWELL, E. M., see DEBYE, P. P.

DARROW, K. K.[1]

Solid State Electronics, Research, **7,** pp. 209, Jan., 1954; pp. 46–53, Feb., 1954; pp. 94–100, March, 1954.

DEBYE, P. P.,[1] and CONWELL, E. M.[4]

Electrical Properties of N-Type Germanium, Phys. Rev., **93,** pp. 693–706.

DUNN, H. K.[1]

Remarks on a Paper entitled "Multiple Helmholtz Resonators", Letter to the Editor, Acous. Soc. Am., J., **26,** p. 103, Jan., 1954.

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.
[4] Sylvania Electric Products, Bayside, New York.

ELLIS, W. C.[1] and FAGEANT, J.[1]

Orientation Relationships in Cast Germanium, J. Metals, **6,** Pt. 2, pp. 291–294, Feb., 1954.

ELLIS, W. C.[1] and GREINER, E. S.[1]

Production of Acceptor Centers in Germanium and Silicon by Plastic Deformation, Letter to the Editor, Phys. Rev., **92,** pp. 1061–1062, Nov. 15, 1953.

FAGEANT, J., see ELLIS, W. C.

FINE, M. E.,[1] VAN DUYNE, H.,[1] and KENNEY, NANCY T.[1]

Low-Temperature Internal Friction and Elasticity Effects in Vitreous Silica, J. Appl. Phys., **25,** pp. 402–405, March, 1954.

FULLER, C. S., see PEARSON, G. L.

FULLER, C. S., see SEVERIENS, J. C.

GALT, J. K.,[1] YAGER, W. A.,[1] and MERRITT, F. R.[1]

Temperature Dependence of Ferromagnetic Resonance Fine Width in a Nickel Iron Ferrite — A New Loss Mechanism, Phys. Rev., **93,** pp. 1119–1120, March, 1954.

GERMER, L. H.[1]

Arcing at Electrical Contacts on Closure. Part IV — Activation of Contacts by Organic Vapor, J. Appl. Phys. **25,** pp. 332–335, March, 1954.

GOHN, G. R.,[1] GUERARD, J. P.,[1] and HERBERT, G. J.[1]

The Mechanical Properties of Some Nickel Silver Alloy Strips, Proc. A.S.T.M., **54,** Jan., 1954.

GREINER, E. S., see ELLIS, W. C.

GUERARD, J. P., see GOHN, G. R.

[1] Bell Telephone Laboratories.

HAGSTRUM, H. D.[1]

Reflection of Ions as Ions or as Metastable Atoms at a Metal Surface, Phys. Rev., **93,** p. 652, Feb., 1954. Abstract of paper presented at Gaseous Electronics Conference Oct. 22–24, 1953.

HEFFNER, H., see CLOGSTON, A. M.

HERBERT, G. J., see GOHN, G. R.

JOHNSON, J. B.,[1] and MCKAY, K. G.[1]

Secondary Electron Emission from Germanium, Phys. Rev., **93,** pp. 668–672, Feb. 15, 1954.

KAHN, A. H., see TESSMAN, J. R.

KENNEY, NANCY, see FINE, M. E.

KOCK, W. E.[1]

Use of the Sound Spectrograph for Appraising the Relative Quality of Musical Instruments, Letter to the Editor, Acous. Soc. Am., J., **26,** pp. 105–106, Jan., 1954.

KRETZMER, E. R.[1]

An Amplitude Stabilized Transistor Oscillator, Proc. I.R.E., **42,** pp. 391–401, Feb., 1954.

LEWIS, H. W.[1]

Search for the Hall Effect in a Superconductor — Experiment, Phys. Rev., **92,** pp. 1149–1151, Dec. 1, 1953.

LINVILL, J. G.[1]

RC Active Filters, Proc. I.R.E., **42,** pp. 555–564, March, 1954.

MACCOLL, L. A.[1]

Geometrical Properties of Two-Dimensional Wave Motion. Am. Math. Monthly, **61,** pp. 96–103, Feb., 1954.

[1] Bell Telephone Laboratories.

MACHLUP, S.[1]

Noise in Semiconductors: Spectrum of a Two-Parameter Random Signal, J. Appl. Phys., **25,** March, 1954.

MATTHIAS, B. T.[1]

Transition Temperatures of Superconductors, Phys. Rev., **92,** pp. 874–876, Nov. 15, 1953.

McKAY, K. G., see JOHNSON, J. B.

MERRITT, F. R., see ANDERSON, P. W.

MERRITT, F. R., see GALT, J. K.

MORIN, F. J.[1]

Lattice-Scattering Mobility in Germanium, Phys. Rev., **93,** pp. 62–63, Jan. 1, 1954.

MORIN, F. J., see PEARSON, G. L.

PEARSON, G. L.,[1] and FULLER, C. S.[1]

Silicon p-n Junction Power Rectifiers and Lightning Protectors, Proc. I.R.E., **42,** pp. 760, April, 1954.

PEARSON, G. L.,[1] READ, W. T., JR.,[1] and MORIN, F. J.[1]

Dislocations in Plastically Deformed Germanium, Phys. Rev., **93,** pp. 666–667, Feb. 15, 1954.

PFANN, W. J.[1]

Comment on Paper by Tiller, Jackson, Rutter and Chalmers, Letter to the Editor, Acta Metallurgica, **1:** pp. 763–764, Nov., 1953.

PFANN, W. G.[1]

Redistribution of Solutes by Formation and Solidification of a Molten Zone, J. Metals, **6,** Pt. 2, pp. 294–297, Feb., 1954.

[1] Bell Telephone Laboratories.

PIERCE, J. R.[1]

Coupling of Modes of Propagation, J. Appl. Phys., **25,** pp. 179–183, Feb., 1954.

READ, W. T., JR.[1]

Dislocations or What Makes Metals So Weak? Metal Progress, **65,** pp. 101–106, 168, 170, 172, Feb., 1954.

READ, W. T., JR., see PEARSON, G. L.

REMEIKA, J. P., see ANDERSON, P. W.

ROSE, D. J.,[1] and ALLIS, W. P.[1]

Transition from Free to Ambipolar Diffusion, Phys. Rev., **93,** pp. 84–93, Jan. 1, 1954.

RYDER, R. M.,[1] and SITTNER, W. R.[1]

Transistor Reliability Studies, Proc. I.R.E., **42,** pp. 414–419, Feb., 1954.

SCHLAACK, N. F.[1]

Development of the LD Radio System, I.R.E., Trans., P.G.C.S., **2,** pp. 29–38, Jan., 1954.

SEVERIENS, J. C.,[1] and FULLER, C. S.[1]

Mobility of Impurity Ions in Germanium and Silicon, Letter to the Editor, Phys. Rev., **92,** pp. 1322–1323, Dec. 1, 1953.

SHIVE, J. N., see SLOCUM, A.

SHOCKLEY, W.,[1] see TESSMAN, J. R.

SITTNER, W. R., see RYDER, R. M.

SLOCUM, A.,[1] and SHIVE, J. N.[1]

Shot Dependence of p-n Junction Phototransistor Noise, Letter to the Editor, J. Appl. Phys., **25,** p. 406, March, 1954.

[1] Bell Telephone Labaratories.

SMITH, C. S.[1]

Piezoresistance Effect in Germanium and Silicon, Phys. Rev., **94,** pp. 42–49, April 1, 1954.

STILES, K. P.[2]

Overseas Radiotelephone Services of A. T. & T. Co., I.R.E., Trans., P.G.C.S., **2,** pp. 39–44, Jan., 1954.

STUBBS, R. R.[5]

Telephone Service is a Big Bargain, Telephony, **146,** pp. 20–21, 43, March 13, 1954.

TESSMAN, J. R.,[6] KAHN, A. H.,[6] and SHOCKLEY, W.[1]

Electronic Polarizabilities of Ions in Crystals, Phys. Rev., **92,** pp. 890–895, Nov. 15, 1953.

THOMAS, D. E.[1]

Single Transistor FM Transmitter, Electronics, **27,** pp. 130–133, Feb., 1954.

VAN DUYNE, H., see FINE, M. E.

VALDES, L. B.[1]

Resistivity Measurements on Germanium for Transistors, Proc. I.R.E., **42,** pp. 420–427, Feb., 1954.

WILFONG, J. C., JR.[7]

The Telephone, an Instrument of Culture, Telephony, **146,** pp. 22–23, 45, March 20, 1954.

WOJCIECHOWSKI, B. M.[3]

Capacitance Gage Checks Cable Sheath Thickness, Electronics, **27,** pp. 134–137, April, 1954.

YAGER, W. A., see GALT, J. K.

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.
[3] Western Electric Company, Inc.
[5] Southern Bell Telephone Company.
[6] Department of Physics, University of California, Berkeley, Calif.
[7] Chesapeake and Potomac Telephone Company.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

BIELING, C. A., see EDELSON, D.

BOWN, R.

Vitality of a Research Institution and How to Maintain It, Monograph 2207.

CAMPBELL, W. H.

Current Status of Fretting Corrosion, Monograph 2160.

CONWELL, E. M.

High Field Mobility in Germanium with Impurity Scattering Dominant, Monograph 2158.

DACEY, G. C.

Space-Charge Limited Hole Current in Germanium, Monograph 2157.

EDELSON, D., BIELING, C. A., and KOHMAN, G. T.

Electrical Decomposition of Sulfur Hexafluoride, Monograph 2175.

GRAY, MARION C.

Legendre Functions of Fractional Order, Monograph 2164.

GRISDALE, R. O.

The Formation of Black Carbon, Monograph 2161.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

GROTH, W. D., AND SLADE, F. D.

Principles of Tape-to-Card Conversion in the AMA System and Mechanized Billing of AMA Toll Messages, Monograph 2190.

KOHMAN, G. T., see EDELSON, D.

PIERCE, J. R., and WALKER, L. R.

"Brillouin Flow" With Thermal Velocities, Monograph 2191.

PRIM, R. C., see SHOCKLEY, W.

READ, W. T., JR.

Dislocations or What Makes Metals so Weak? Monograph 2211.

ROSE, D. J.

The Transition from Free to Ambipolar Diffusion, Monograph 2205.

SHOCKLEY, W., and PRIM, R. C.

Space-Charge Limited Emission in Semiconductors, Monograph 2156.

SLADE, F. D., see GROTH, W. B.

THOMAS, D. E.

Low-Drain Transistor Audio Oscillator, Monograph 2204.

THURMOND, C. D.

Equilibrium Thermochemistry of Solid and Liquid Alloys of Germanium and of Silicon, Monograph 2210.

WALKER, L. R., see PIERCE, J. R.

WILKINSON, R. I.

Random Picture Spacing with Multiple Camera Installations, Monograph 2188.

Contributors to this Issue

ALBERT L. BLAHA, B.S. in E.E., Polytechnic Institute of Brooklyn, 1950; Bell Telephone Laboratories, 1936–. In 1937 and 1938, Mr. Blaha was in the quartz crystal development shop. Since then he has been primarily concerned with the testing and development of relays. During World War II he worked on magnetostriction type sonar devices.

A. J. BRUNNER, B.S. in M.E., Lewis Institute, 1934; Western Electric Company, 1920–. During Mr. Brunner's early association with Western Electric he was included in an engineering group developing special machines for the manufacture of telephone products. Later he worked on the development of die casting processes. For the past decade his assignments have been in the field of plastic molding. He did notable work in connection with molding 500-type handset parts and is presently engaged in engineering the facilities required to manufacture molded parts for the wire spring relay. He is the holder of numerous patents.

F. HAROLD CHASE, University of Illinois, 1921; Western Electric Company, 1917–1918; Bell Telephone Laboratories, 1921–. Until joining the Power Engineering Group in 1943 he was concerned with the design of carrier system equipment and maintenance practices on toll equipment. Since 1951, he has been developing new uses for transistors in the control of power equipment.

H. E. COSSON, B.S. in M.E., Michigan College of Mining and Technology, 1949; Allis Chalmers Manufacturing Company, 1949–51; A. O. Smith Corp., 1951; Western Electric Company, 1951–. Mr. Cosson served thirty-one months during World War II, fifteen of which were on a naval aircraft carrier. Since joining the development engineering group at Western Electric Company, he has worked on problems associated with straightening wire for the wire spring relay. Junior member A.S.M.E.

THOMAS E. DAVIS, B.S. in E.E., University of Arizona, 1928; Bell Telephone Laboratories, 1928–. He has been concerned with apparatus development projects, including those related to microphones, handsets

and echo suppressors for long telephone lines. During World War II he worked on underwater sound systems for the Navy and was awarded the Naval Ordnance Development Award by the U. S. Navy. Since then he has been working with the wire-spring relay and line concentrator for the No. 5 crossbar system. Member American Institute of Electrical Engineers and Tau Beta Pi.

B. H. HAMILTON, B.S. in E.E., University of Kansas, 1949; Bell Telephone Laboratories 1950–. With the Laboratories he has worked on development of equipment to power the L3 carrier system and has been concerned with fundamental studies of new types of regulated rectifiers. Member American Institute of Electrical Engineers, Tau Beta Pi, Sigma Tau, Sigma Xi, Kappa Eta Kappa.

A. L. QUINLAN, B.S. in E.E., University of Kansas, 1921; Western Electric Company, 1921–. Prior to World War II, Mr. Quinlan worked extensively on the development of manufacturing methods and machines for loading coils. He was granted patents on loading coil case designs and on toroidal coil winding machines and was co-author of an article, *Recent Improvements in Loading Apparatus for Telephone Cables*, published in the A.I.E.E. Journal, Dec. 1947. During the war he had engineering assignments on gun director, precision coil manufacture and vacuum tube projects. Since then he has developed manufacturing facilities and methods for welding precious metal contacts to telephone switching apparatus. Notable among these are the roll welding of contact tape to crossbar switch multiples and the resistance and percussion welding of contacts to wire spring relays. Member A.I.E.E.

WILLIAM SHOCKLEY, B.Sc., California Institute of Technology, 1932; PH.D., Massachusetts Institute of Technology, 1936; Teaching Fellow, M.I.T., 1932–1936; Bell Telephone Laboratories 1936–1942; Director of Research, Antisubmarine Warfare Operations Research Group, Division of War Research, Columbia University, 1942–1944; Expert Consultant, Office of the Secretary of War, 1944–1945; Bell Telephone Laboratories 1945–. Appointed Director of Transistor Physics Research December 1, 1953, he had directed the group which invented the point-contact transistor. During the past six years he has made many contributions to solid state physics particularly in connection with the transistor. In addition to solid state physics and semiconductors, his work has also included vacuum tube and electron multiplier design, studies of various physical phenomena in alloys, radar development and magnetism. Medal for Merit, U. S.

War Department, 1946; Air Force Association Citation of Honor, 1951; Morris Liebmann Memorial Prize, I.R.E., 1952; Oliver E. Buckley Solid State Physics Prize, American Physical Society, 1953; Certificate of Appreciation, Department of Army, 1953; Comstock Prize, National Academy of Sciences, 1954. Fellow of American Physical Society; Senior Member Institute of Radio Engineers; Member of National Academy of Sciences, Tau Beta Pi, Sigma Xi. For the past few months he has been on leave to California Institute of Technology for teaching and study in the field of solid state physics.

DONALD H. SMITH, B.S. in E.E., University of Minnesota, 1944; Bell Telephone Laboratories, 1947–. After working with the Systems Department of the Laboratories on trial installations, Mr. Smith was concerned with rectifiers and regulating systems in power development. He is currently in charge of the group doing long-range engineering on power development. Member of A.I.E.E. and the Amateur Astronomers Association.

R. W. STRICKLAND, B.M.E., University of Florida, 1951; Western Electric Company, 1951–. Mr. Strickland served two and a half years in the U. S. Armed Forces prior to receiving his degree. Since coming to the Western Electric Company, he has been active in the development of equipment and processes for molding of plastic components of the wire spring relay. Junior member A.S.M.E.

HARRY SUHL, B.Sc., University of Wales, 1943; PH.D., Oriel College, University of Oxford, 1948. Admiralty Signal Establishment, 1943–46; Bell Telephone Laboratories, 1948–. Dr. Suhl conducted research on the properties of germanium until 1950 when he became concerned with electron dynamics and solid state physics research. His current work is in the applied physics of solids. Member of the American Institute of Physics and Fellow of the American Physical Society.

ERIC E. SUMNER, B.M.E., Cooper Union, 1948; M.A. Degree in Physics, Columbia University, 1953; Instructor of Physics, Cooper Union, 1947–48; Non-resident instructor of Massachusetts Institute of Technology on *Probability and Statistics — Applications to Sampling and Quality Control*, summer, 1950; Bell Telephone Laboratories, 1948–. Mr. Sumner was given rotational assignments in apparatus, switching, and Television transmission development and switching research, and has worked on a number of projects, including the card translator, the mag-

netic drum, video transmission evaluator, vibrating reed selector, development of wire-spring relay, trouble recording apparatus development, and transistor circuitry for a subscriber line concentrator. He is currently engaged in developing small functional circuits for an electronic switching system. Member of Tau Beta Pi and Pi Tau Sigma.

ERLING D. SUNDE, E.E., Technische Hochschule, Darmstadt, Germany, 1926. Brooklyn Edison Company, 1927; American Telephone and Telegraph Company, 1927–1934; Bell Telephone Laboratories, 1934–. Mr. Sunde's work has been centered on theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. Author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series Book. Member of the A.I.E.E., the American Mathematical Society, and the American Association for the Advancement of Science.

LAURENCE R. WALKER, B.Sc. and PH.D., McGill University, 1935 and 1939; University of California, 1939–41. Radiation Laboratory, Massachusetts Institute of Technology, 1941–1945; Bell Telephone Laboratories, 1945–. Dr. Walker has been primarily engaged in research on microwave oscillators and amplifiers. At present he is a member of the physical research group concerned with the applied physics of solids. Fellow of the American Physical Society.