RCA REVIEW

a technical journal

Published quarterly by
RCA LABORATORIES
in cooperation with all subsidiaries and divisions of
RADIO CORPORATION OF AMERICA

VOLUME XXV MARCH 1964 NUMBER 1

CONTENTS

RCA REVIEW is regularly abstracted and indexed by *Abstracts of Photographic Science and Engineering Literature, Applied Science and Technology Index, Bulletin Signalétique des Télécommunications, Chemical Abstracts, Electronic and Radio Engineer, Mathematical Reviews,* and *Science Abstracts* (I.E.E.-Brit.).

RCA REVIEW

ELECTROMAGNETIC PROPERTIES OF FINITE PLASMAS

By

M. P. Bachynski and K. A. Graf

RCA Victor Company, Ltd.,
Montreal, Canada

Summary—*The determination of plasma properties by the transmission or reflection of electromagnetic waves depends upon (1) the availability of a theory that adequately describes the physical situation and (2) experimental measurements that are amenable to theoretical interpretation. The major limitations are the finite size of the plasma, the effect of the boundaries of the plasma and the material boundaries that contain the plasma, and the nonuniformity of the plasma in space and time.*

In this paper, expressions are derived and typical numerical values presented for the effect on transmission, reflection, and absorption of electromagnetic waves of plasma and dielectric boundaries, refractive defocusing by plasmas of slab and cylindrical geometry for both plane and spherical incident waves, the effect of nonuniformity of the plasma both along the direction of propagation and normal to the direction of propagation, and for diffraction introduced by the finite size of a circular slab of plasma.

Introduction

THE ELECTRON DENSITIES found in many plasmas of interest correspond to plasma frequencies in the meter and centimeter wavelength range. Since the electrical properties of a plasma vary measurably in this range of frequencies, probing by low-strength radio signals has become a much used technique for determining the characteristics of the plasma. The accurate determination of plasma properties by this free-space method depends upon two factors, namely the availability of a theory that adequately describes the physical situation and experimental measurements that are amenable to theoretical interpretation.

In principal, the determination of the properties of a plasma from the phase change and attenuation that it introduces to an incident electromagnetic wave transmitted through or reflected from it is very simple. If the plasma is uniform and in the form of an infinite slab with sharp, well-defined boundaries, and if the incident field is a plane wave and the incident wave after interacting with the plasma can be accurately measured with a perfect system, then no discrepancy between theory and experiment should occur. This ideal situation is, however, impossible to realize.

In practice the plasma is finite in extent; it may be contained by material walls, the boundaries of the plasma may not be well-defined, and the plasma may be nonuniform in both space and time. The result is refraction, reflection, absorption, and diffraction phenomena that are not easy to define and interpret but the understanding of which is essential before accurate quantitative determination of plasma properties is possible.

Although a number of microwave free-space measurements of plasma have been reported[1-6] and some of the above limitations have been mentioned, there does not appear to have been a systematic attempt to assess the predictions of various simple theoretical models of the plasma or to develop theories to account for the influence of the aforementioned effects. In this paper, theoretical predictions are developed and typical numerical values presented for plasma effects such as plasma boundaries, refractive defocusing by the plasma, nonuniformity of the plasma, and diffraction introduced by the finite size of the plasma.

Effect of Boundaries on Transmission, Reflection and Absorption of Electromagnetic Waves by a Plasma

Many calculations on the effect of a plasma on an incident plane electromagnetic wave are based on a theoretical model in which the influence of the boundaries of the plasma is completely ignored. The electromagnetic wave is considered to traverse a region of plasma equal in extent to a given physical dimension, but the effect of reflection from the interface between the plasma and free-space and multiple reflections within the plasma are neglected (see Figure 1a). This "unbounded plasma" model thus predicts the attenuation and phase

---

[1] R. J. Jahn, "Microwave Probing of Ionized-Gas Flows," *Phys. Fluids*, Vol. 5, p. 678, June 1962.

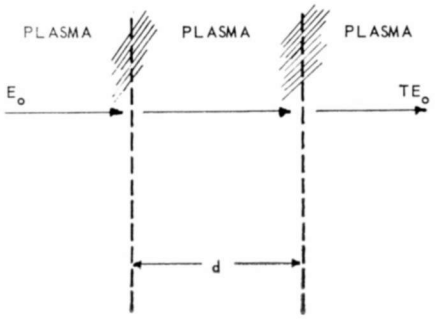
[2] P. W. Kuhns, "Microwave Measurements of Steady-State and Decaying Plasmas," *Trans. I.R.E. PGSET*, Vol. 8, p. 173, June 1962.

[3] R. Buser and W. Buser, "Determination of Plasma Properties by Free-Space Microwave Techniques," *Jour. Appl. Phys.*, Vol. 33, p. 2275, July 1962.

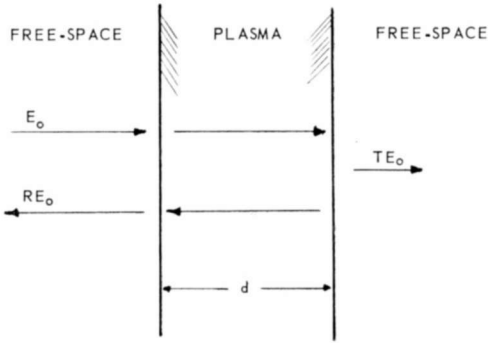
[4] R. Warder, M. Brodwin, and A. B. Cambel, "Sources of Error in the Microwave Diagnostics of Plasmas," *Jour. Appl. Phys.*, Vol. 33, p. 2868, Sept. 1962.

[5] G. R. Nicoll and J. Baser, "Comparison of Microwave and Langmuir Probe Measurements on a Gaseous Plasma," *Jour. Elect. Cont.*, Vol. XII, p. 23, June 1962.

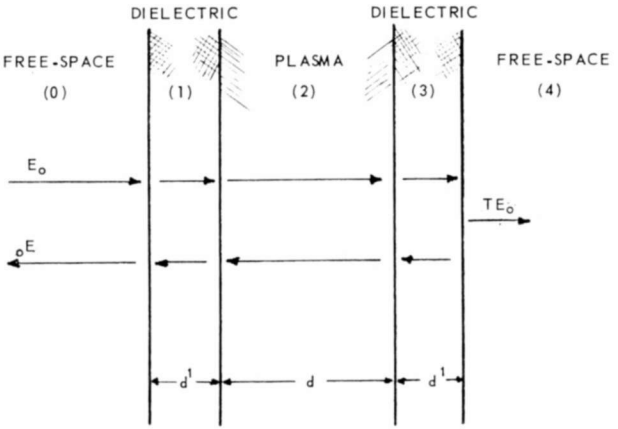
[6] L. Talbot, J. E. Katz, and C. L. Brundin, "Comparison Between Langmuir Probe and Microwave Electron Density Measurements in an Arc-Heated Low-Density Wind Tunnel," *Phys. Fluids*, Vol. 6, p. 559, April 1963.



(a)



(b)



(c)

Fig. 1—Theoretical models for determining the transmission and reflection of electromagnetic waves by a plasma: (a) unbounded plasma, (b) plasma slab bounded by free space, (c) plasma slab contained within dielectric plates in free space.

shift that the plasma would introduce to a plane homogeneous electromagnetic wave traversing a given distance in an infinite, uniform, isotropic plasma.

A more-realistic model (Figure 1b) considers a plane, homogeneous wave normally incident on a uniform, isotropic "plasma slab" bounded by free space. In this model both reflections from the interfaces and multiple reflections within the plasma are taken into account.

In a laboratory plasma, the plasma is very often contained by material walls. A theoretical model to take into account the effect of the material container is a slab of plasma bounded by two flat dielectric plates (as shown in Figure 1c).

Calculations of the attenuation and phase shifts introduced by a plasma on an incident electromagnetic wave based on these three models have been made and compared for various plasma parameters. The predictions of the different models give an indication of the range of validity of each model and the accuracy of measurement of plasma properties to be expected when free-space microwave techniques are used.

Unbounded Plasma Model

In uniform, neutral plasmas of electron density n and effective collision frequency ν, the dielectric constant can be written for a harmonic time varying field ($e^{j\omega t}$) as[7]:

$$K = K_r - jK_i = 1 - \left(\frac{\omega_p}{\omega}\right)^2 \left(\frac{1}{1 + (\nu/\omega)^2}\right) - j\left(\frac{\omega_p}{\omega}\right)^2 \left(\frac{\nu/\omega}{1 + (\nu/\omega)^2}\right)$$

$$= 1 - \frac{N}{1 + S^2} - j\frac{NS}{1 + S^2}, \tag{1}$$

where:

ω_p is the plasma frequency, $(ne^2/m\epsilon_0)^{1/2}$,

e,m are the electronic charge and mass, respectively,

ϵ_0 is the permittivity of free space,

ω is the radian radio frequency, and

N,S are normalized electron density and collision frequency parameters given by $N = (\omega_p/\omega)^2$ and $S = \nu/\omega$.

[7] M. P. Bachynski, T. W. Johnston, and I. P. Shkarofksy, "Electromagnetic Properties of High Temperature Air," *Proc. I.R.E.*, Vol. 48, p. 317, 1960.

For a plane homogeneous electromagnetic wave the propagation constant (γ), attenuation constant (α) and phase constant (β) can be written

$$\gamma = \alpha + j\beta, \tag{2a}$$

$$\alpha = k\left(\frac{|K| - K_r}{2}\right)^{1/2}, \tag{2b}$$

$$\beta = k\left(\frac{|K| + K_r}{2}\right)^{1/2}, \tag{2c}$$

where

$$|K| = (K_r^2 + K_i^2)^{1/2},$$

and $k = 2\pi/\lambda$ is the free-space wave number.

For a wave propagating a distance d in the plasma, the transmission coefficient, T, and the reflection coefficient, R, are given by

$$T = 1 - e^{-\alpha d}, \tag{3a}$$

$$R = 0. \tag{3b}$$

The attenuation of the wave after propagating a distance d in the plasma is given by

$$\alpha d = 2\pi\left(\frac{\alpha}{k}\right)\left(\frac{d}{\lambda}\right). \tag{4a}$$

The phase shift that occurs when β/k changes from unity (no plasma) to some value defined by the plasma is

$$\phi = 2\pi\left(\frac{d}{\lambda}\right)\left[1 - \left(\frac{\beta}{k}\right)\right]. \tag{4b}$$

The attenuation and phase shift for a path length, d, of 1.84 free-space wavelengths as function of electron density and collision frequency for an unbounded plasma are shown in Figure 2.

"Plasma Slab" Model

Application of the theory for an "unbounded plasma" is very simple. It is usually considered quite accurate if the refractive index

of the plasma is close to unity, in which case reflections from the plasma-free-space interfaces will be insignificant. To check the validity of such assumptions, calculations were performed to determine the attenuation and phase-shift of a plasma slab sharply bounded by free space. Writing the boundary conditions at each interface and

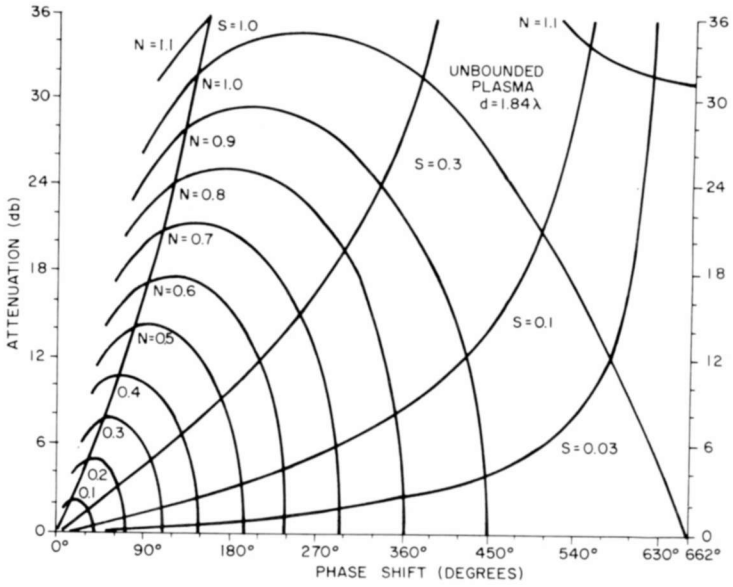


Fig. 2—Variation of attenuation and phase shift due to an unbounded plasma for various values of electron density and collision frequency.

solving the electromagnetic equations gives the transmission coefficient for normal incidence[8,9] as

$$T = \frac{E_T}{E_0} = [\cosh(\alpha d + j\beta d) + (Z_r - jZ_i)\sinh(\alpha d + j\beta d)]^{-1}. \qquad (5a)$$

The reflection coefficient (applied to the fields) is similarly obtained and found to be

---

[8] G. G. Cloutier, M. P. Bachynski, and K. Graf, "Antenna Properties in the Presence of Ionized Media," *AFCRL* Report No. 62-191.

[9] I. P. French, G. G. Cloutier, and M. P. Bachynski, "The Absorptivity Spectrum of a Uniform Anisotropic Plasma Slab," *Canadian Jour. Phys.*, Vol. 39, p. 1273, 1961.

$$R = T\left[Z_i \left(\frac{\beta/k}{\alpha/k} \right) + jZ_r \left(\frac{\alpha/k}{\beta/k} \right) \right] \sinh(\alpha + j\beta)d, \quad (5b)$$

where

$$Z_r = \frac{1}{2}\left(\frac{\beta}{k} \right)\left(\frac{|K|+1}{|K|} \right),$$

$$Z_i = \frac{1}{2}\left(\frac{\alpha}{k} \right)\left(\frac{|K|-1}{|K|} \right).$$

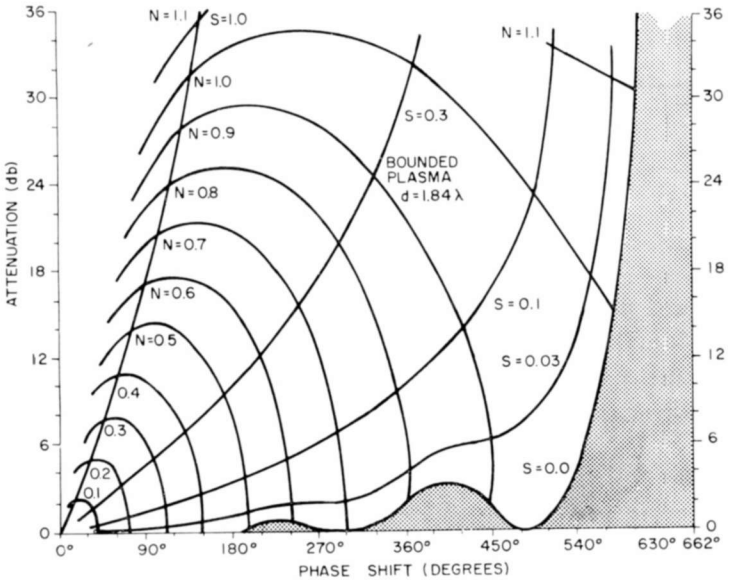


Fig. 3(a)—Variation of attenuation and phase shift due to a plasma slab bounded by free space for various values of electron density and collision frequency.

The attenuation and phase shift of the transmitted wave (with the phase referred to the second surface of the plasma slab) are plotted in Figure 3a.

Examination of Figure 3a shows that the effect of the slab boundaries is pronounced when the plasma is not very lossy and when the normalized electron density is greater than about 0.5. The effect of the boundaries is significant in modifying the amplitude and phase of the transmitted signal. If the normalized collision frequency (ν/ω) is greater than 0.1, the effect of the boundaries is small. The reflected

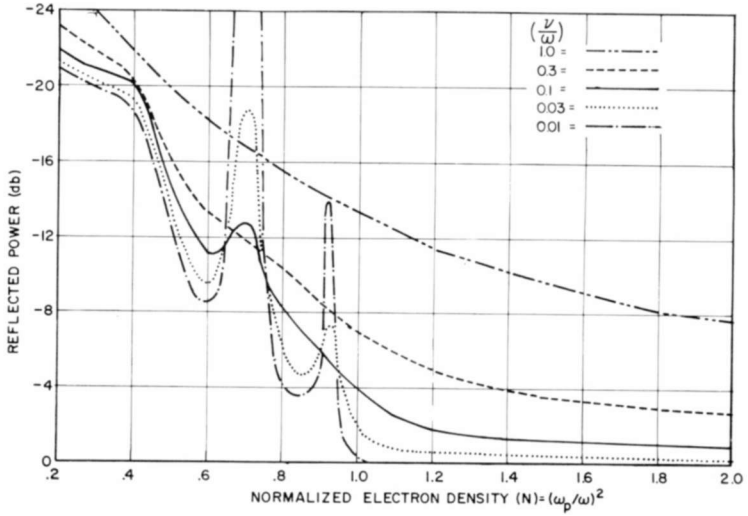


Fig. 3(b)—Variation of reflected power with normalized electron density due to a plasma slab bounded by free space for various values of collision frequency ($d = 1.84\lambda$).

energy and phase-shift of the reflected signal are shown in Figures 3b and 3c, respectively.

The incident energy that is not reflected or transmitted by the plasma slab is absorbed; the absorbed power, $A\omega$, is given by

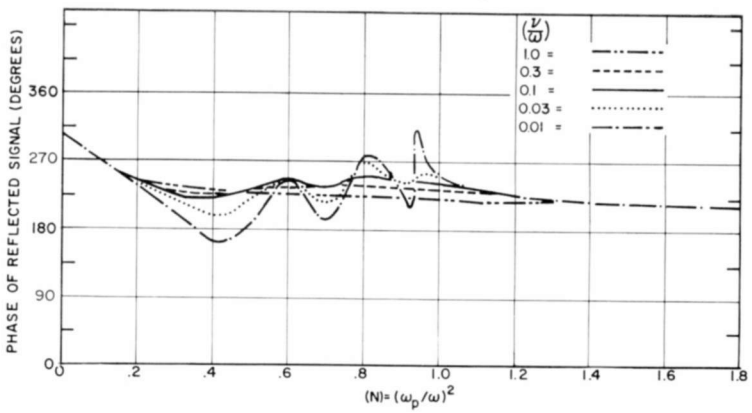


Fig. 3(c)—Variation of phase of reflected wave with normalized electron density due to a plasma slab bounded by free space for various values of collision frequency ($d = 1.84\lambda$).

$$A_\omega = 1 - RR^* - TT^* \tag{6}$$

where the R^*, T^* refer to the complex conjugates of R and T, respectively. The variation of A_ω for various plasma parameters is shown in Figure 3d.

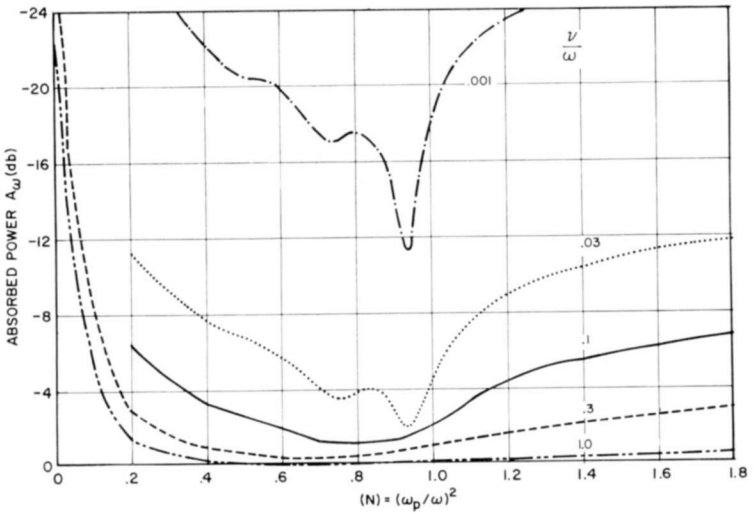


Fig. 3(d)—Variation of absorbed power with normalized electron density due to a plasma slab bounded by free space for various values of collision frequency ($d = 1.84\lambda$).

Plasma Slab Bounded by Dielectric Plates

Laboratory measurements on plasmas are usually significantly affected by the container in which the plasma is confined. Even when the index of refraction of the plasma is close to unity, the combined effect of reflection from the container and the plasma may be significant. To study the magnitude of this effect, calculations were made for a model consisting of a plasma slab bounded by two dielectric plates. The geometry was as shown in Figure 1c.

At normal incidence, the waves in the plasma and the dielectric will be plane waves. Nine "composite" waves representing all possible reflections are of interest. One can write all the boundary conditions for the continuity of the electric field and the magnetic field across the various interfaces. Solving the resulting equations for the transmission and reflection coefficient of the slab of plasma bounded by

dielectric plates in free-space gives:[10]

$$T = \left\{ \left[\cosh^2 \gamma' d' + \sinh^2 \gamma' d' + \left(\frac{Z_1}{Z_0} + \frac{Z_0}{Z_1} \right) \frac{\sinh 2\gamma' d'}{2} \right] \cosh \gamma d \right.$$
$$+ \frac{1}{2} \left\{ \left(\frac{Z_1}{Z_2} + \frac{Z_2}{Z_1} \right) \sinh 2\gamma' d' + \left(\frac{Z_2}{Z_0} + \frac{Z_0}{Z_2} \right) \cosh^2 \gamma' d' \right.$$
$$\left. \left. + \left(\frac{Z_1^2}{Z_0 Z_2} + \frac{Z_0 Z_2}{Z_1^2} \right) \sinh^2 \gamma' d' \right\} \sinh \gamma d \right\}^{-1} \tag{7a}$$

$$R = T \left\{ \left(\frac{Z_1}{Z_0} - \frac{Z_0}{Z_1} \right) \sinh \gamma' d' \cosh \gamma' d' \cosh \gamma d \right.$$
$$+ \frac{1}{2} \left[\left(\frac{Z_2}{Z_0} - \frac{Z_0}{Z_2} \right) \cosh^2 \gamma' d' + \right.$$
$$\left. \left. \left(\frac{Z_1^2}{Z_0 Z_2} - \frac{Z_0 Z_2}{Z_1^2} \right) \sinh^2 \gamma' d' \right] \sinh \gamma d \right\} \tag{7b}$$

where

γ, γ' are the propagation constants of the plasma and the dielectric plates, respectively.

d, d' are respectively the thickness of the plasma and a dielectric plate,

Z_0, Z_1, Z_2 are the impedances of free space, dielectric, and plasma, respectively.

If the dielectric plates are considered lossless, the propagation constant of the dielectric, γ', will be a pure imaginary.

$$\gamma' = j\beta' = jk \left(\frac{\beta'}{k} \right) = j\mu \frac{2\pi}{\lambda},$$

where μ is the index of refraction of the dielectric plates.

The impedances of the dielectric and plasma can be written in terms of the free-space impedance as

---

[10] M. P. Bachynski, G. G. Cloutier, and K. A. Graf, "Microwave Measurements of Finite Plasmas," *AFCRL* Report No. 63-161.

$$Z_1 = \frac{Z_0}{\sqrt{K_1}} = \frac{Z_0}{\mu},$$

$$Z_2 = \frac{Z_0}{\sqrt{K}} = Z_0\left(\frac{jk}{\alpha + j\beta}\right),$$

where K_1 is the dielectric coefficient of the dielectric.

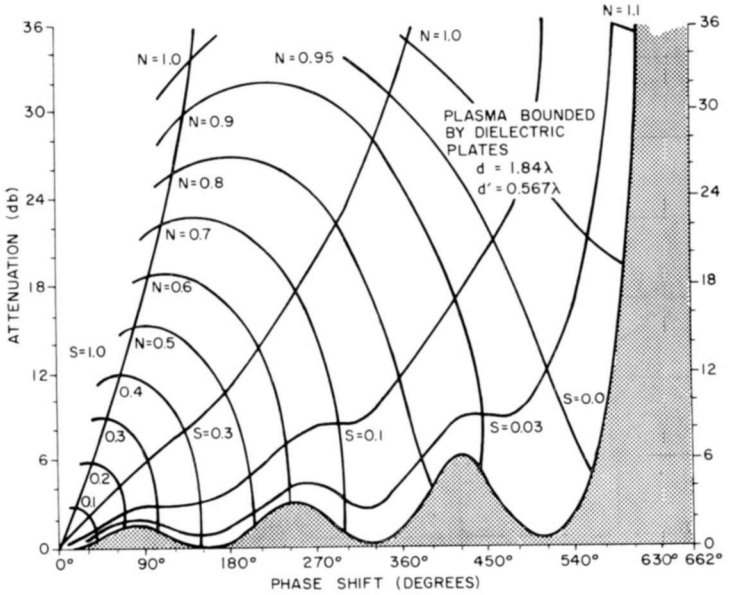


Fig. 4(a)—Variation of attenuation and phase shift due to a plasma slab contained within dielectric plates in free space for various values of electron density and collision frequency.

Calculations were performed for the same plasma slab dimensions and parameters as used in the plasma slab model. The refractive index of the dielectric plates was 1.58 (polystyrene) and the thickness 0.567 free-space wavelengths.

The optical path length of each dielectric plate was 0.9λ, which results in relatively small reflections from the plates for a normally incident plane wave. Had the plates been $(2n + 1)\lambda/4$ in optical path length, then very strong reflections would have occurred from the dielectric plates and would, therefore, have manifested themselves in the effect of the plasma-dielectric container on the transmitted and reflected electromagnetic fields (n is an integer).

The results of a computer calculation for the transmitted signal are shown in Figure 4a.

The phase-shift calculations have been "adjusted" on the plot so that there is no phase shift when the electron density is zero—a perfectly matched system. Similarly, the attenuation was taken as zero when the electron density was zero. For the calculations involving the dielectric plates, the signal transmitted through a composite slab (dielectric–plasma–dielectric) can be greater than the initial (zero electron density) value. This simply implies that the container (the two dielectric plates) reflects more signal when the electron density

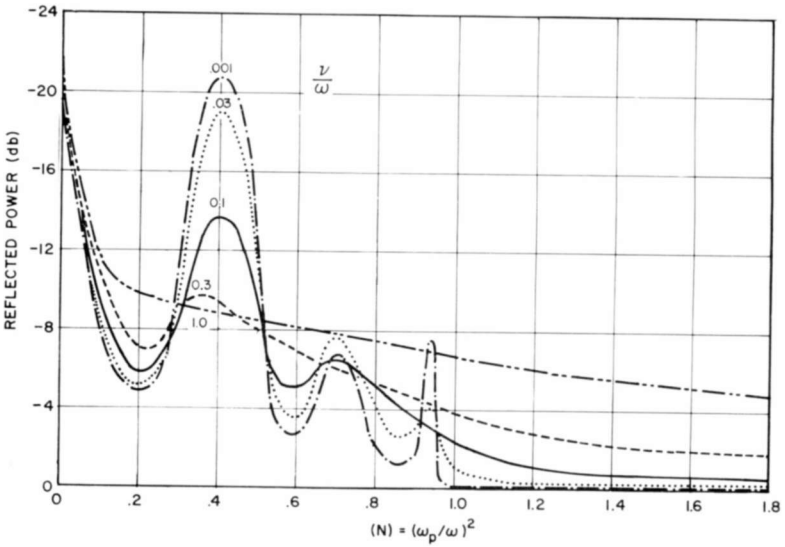


Fig. 4(b)—Variation of reflected power with normalized electron density due to a plasma slab contained within dielectric plates in free space for various values of collision frequency ($d = 1.84\lambda$, $d' = 0.567\lambda$).

is zero than for some other electron density. The varying refractive index of the plasma could "match" the two dielectric plates so that less signal is reflected. For the particular set of parameters μ, d', and d that were chosen for the computer calculations, this did not occur. Had it occurred, the attenuation would initially have appeared *negative*, i.e., the plasma would be "matching" the container to the incident fields.

It can be seen that the signal strength, even for small electron densities, is severely influenced by the dielectric plates. The phase of the transmitted signal is less severely affected. The magnitude and phase shift of the reflected signal are shown in Figures 4b and 4c.

The amount of power absorbed by the plasma is shown in Figure

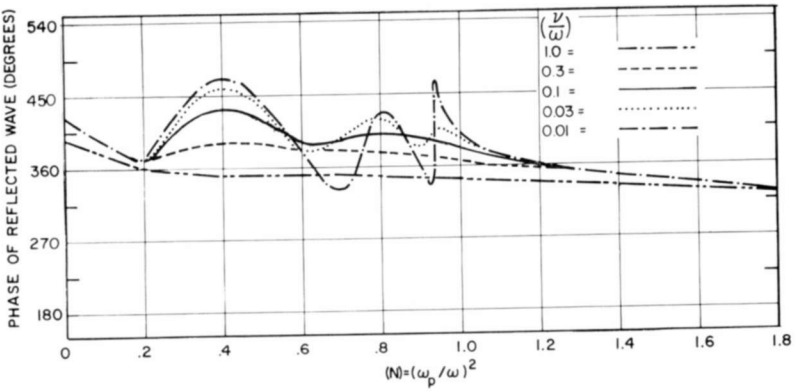


Fig. 4(c)—Variation of phase of reflected wave with normalized electron density due to a plasma slab contained within dielectric plates in free space for various values of collision frequency ($d = 1.84\lambda$, $d' = 0.567\lambda$).

4d. Since the dielectric plates were considered lossless, they did not absorb any power although they can influence the amount of power that the plasma absorbs.

Comparison of the Three Models

The plots of the attenuation and phase-shift dependence on electron density for the three models show that the phase-shift is perturbed

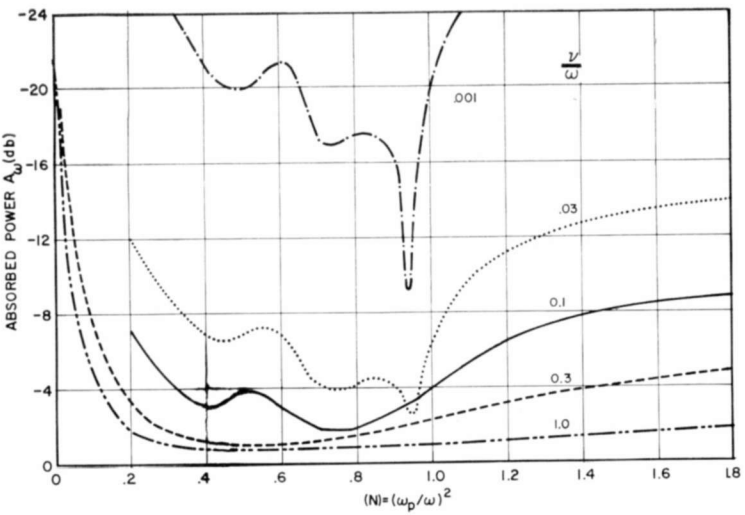


Fig. 4(d)—Variation of absorbed power with normalized electron density due to a plasma slab contained within dielectric plates in free space for various values of collision frequency ($d = 1.84\lambda$, $d' = 0.567\lambda$).

less by the interface conditions than the attenuation for low values of electron density. Since, very generally, phase shift is associated primarily with electron density and attenuation with the electron collision frequency, the effect of the interfaces makes the collision frequency more in doubt than the electron density.

A polar plot of the amplitude and phase of a transmitted signal, calculated for the three models, is shown for comparison in Figure 5.

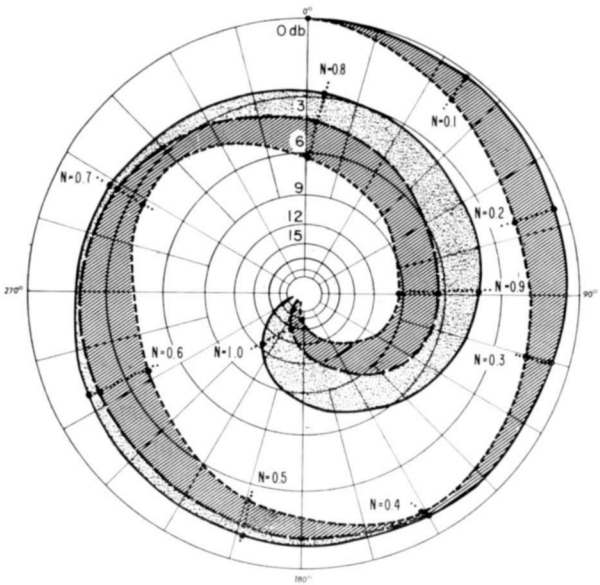


Fig. 5(a)—Comparison of attenuation and phase shift as predicted by the three theoretical models for the plasma for $\nu/\omega = 0.03$ ($\mu = 1.58$, $d = 1.84\lambda$, and $d' = 0.567\lambda$). The solid line is for the unbounded plasma, the long-dash line for the bounded plasma, and the short-dash line for the plasma bounded by dielectric plates.

The curve drawn in each case is for collision frequencies (ν/ω) of 0.03 and 0.1. Because of interface effects, and the thickness of plasma and dielectric plates used in the calculations, the deviation for the three curves is more pronounced at some electron densities than at others. As an example of the effect of the boundaries, note that for a phase shift of 360°, the electron density (N) measured by the three methods is 0.8.

The "unbounded" theory shows less than 3 db attenuation; when the dielectric plates and the interfaces are taken into account more than 6 db attenuation is obtained. This is a significant difference. Conversely, although it is not shown in Figure 5, it can be seen from

the earlier figures that an attenuation of 6 db would give a collision frequency of 0.03 by the "dielectric plate" calculations, and a collision frequency of about 0.06 by the "unbounded" theory. As the collision frequency increases, the differences between the various models become less and less significant.

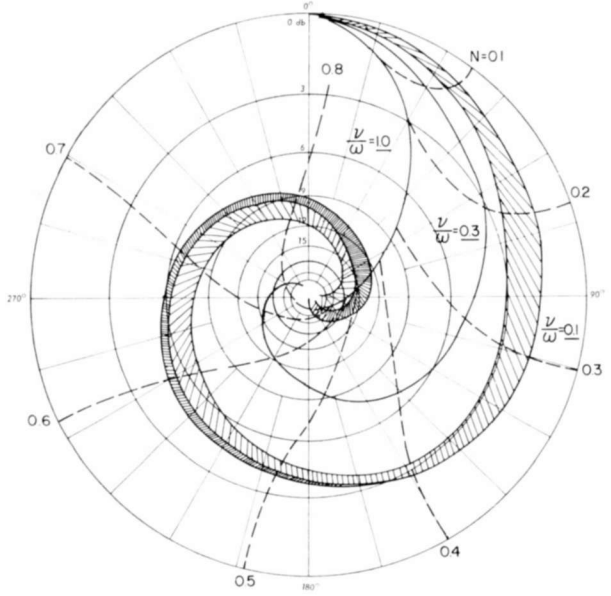


Fig. 5(b)—Comparison of attenuation and phase shift as predicted by the three theoretical models for the plasma for $\nu/\omega = 0.10$. Outer curve is for the "unbounded plasma model," innermost curve is for the "plasma bounded by dielectric plates." The distinction between the models becomes less significant with increasing collision frequency. For $\nu/\omega = 0.3$ and 1.0 all models predict about the same result ($\mu = 1.58$, $d' = 0.567\lambda$, and $d = 1.84\lambda$).

It is also interesting to note that the difference between the unbounded theory and the plasma slab model are *not* significant for low electron densities ($N < 0.4$); however, the effects of the dielectric plates on the plasma slab at these low electron densities *are* important due to the "matching" effects on the incident field.

Refractive Defocusing by Uniform Plasma Slabs and Cylinders

Microwave systems used for the free-space measurement of plasma properties can be broadly classified as to the type of incident wave front. The arrangements most often employed are of the type that

result in either an incident *plane* wave (by the use of auxiliary lenses), a *spherical* incident wave (unfocused point source) or a highly focused beam (using lenses or other focusing devices) to give a high degree of spatial resolution. These systems are illustrated in Figure 6.

Since the refractive index of the plasma ($\mu = K^{1/2}$) will, in general, not be equal to the refractive index of free space ($\mu = 1$), refraction

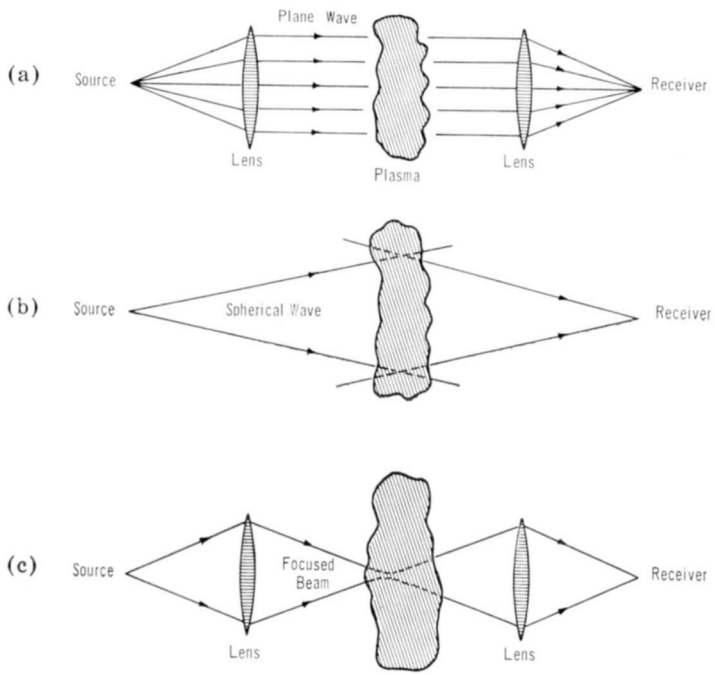


Fig. 6—Types of microwave systems used for free-space measurement of plasma properties: (a) plane wave system, (b) point source system, and (c) focussed system.

will occur at each boundary between plasma and free space. Using the concepts of geometric optics, the electromagnetic energy can be considered as traveling along ray paths or rays that are normal to the planes of constant phase of the wavefront. (We shall neglect in the sequel the situation which can arise[11] whereby the rays do not coincide with the direction of energy travel, i.e., the direction of the Poynting vector is not normal to the phase front, resulting in inhomogeneous

[11] K. A. Graf and M. P. Bachynski, "Transmission and Reflection of Electromagnetic Waves at a Plasma Boundary for Arbitrary Angles of Incidence," *Canadian Jour. Phys.*, Vol. 39, p. 1544, 1962.

plane waves.) The net result of the refraction is that the incident beam of energy is spread out or defocused by the plasma. (This is due to the fact that for the plasma $\mu < 1$; for a dielectric with $\mu > 1$, a focusing of the beam results.) The plasma can thus be considered as a lens of refractive index less than unity. The net result of this refractive defocusing is that the energy density of the radiation in the region where it can be measured by a microwave receiving system has been decreased not only by the amount of energy absorbed by the plasma, but also be the amount by which it has been spread out. Consequently, in order to obtain a measure of the energy absorbed by the plasma (and hence get a measure of collision frequency) some estimate of the refractive defocusing is essential.

Subject to the limitations of geometric optics (dimensions large compared to wavelengths, losses in plasma small, etc.) it is possible to derive expressions for the refractive defocusing by uniform plasma slabs and plasma cylinders. These are discussed subsequently.

Plane Wave Incident

For a plane wave incident normally on a slab of plasma, no refractive defocusing occurs, as shown in Figure 7a.

A plane wave incident on a uniform, cylindrical plasma will be refracted. With reference to Figure 7a let the extreme ray of an incident beam of radiation of radius a be intercepted by a plane of half-width A located a distance R from the center of the cylinder of plasma of radius r. The angles of incidence and refraction are θ_i and θ_n, respectively, while the other parameters are defined in the diagram. Using Snell's Law and geometric considerations, it is easy to show that in the small-angle limit

$$\frac{a}{A} = \frac{r}{r + 2\left(\frac{1}{\mu} - 1\right) R}.$$

When the refractive index of the plasma is unity ($\mu = 1$), then $a = a_0$ or

$$\frac{a_0}{A} = 1.$$

For an extreme ray of radius A, the effective radius of the beam of incident radiation which is intercepted is a. The effect of the plasma

is to reduce the radius of the incident beam that is intercepted from a_0 to a. A measure of the refractive defocusing effect in one dimension is then

$$\eta = \frac{a}{a_0} = \frac{r}{r + 2\left(\frac{1}{\mu} - 1\right) R} = \frac{1}{1 + 2\left(\frac{1}{\mu} - 1\right) \frac{R}{r}}. \quad (8)$$

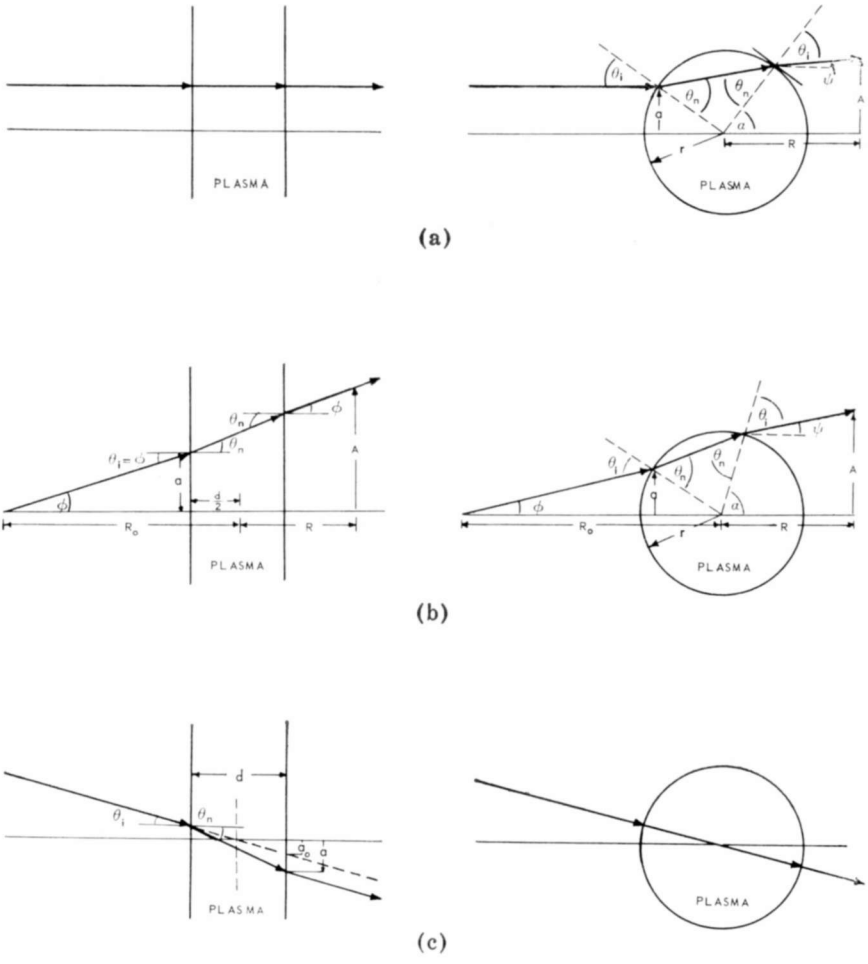


Fig. 7—Refractive defocusing introduced by a uniform plasma for (a) plane incident wave, (b) spherical incident wave, and (c) focused incident beam.

We shall call η the "refractive defocusing coefficient" or in most cases the "refractive defocusing." Note that if $\mu > 1$, then $\eta > 1$, i.e., focusing occurs.

For a plane incident wave, the defocusing coefficient η is a measure of the beam of energy intercepted by a receiver of aperture dimension A, located at R in the presence of the cylinder of plasma relative to that intercepted when there is no plasma cylinder. Note that for a plane incident wave the spreading or defocusing of energy occurs only in the plane normal to the cylinder axis and no defocusing effect is present along the axis of the cylinder.

Spherical Incident Wave (Point Source)

A spherical wave incident upon a uniform slab or cylinder of plasma will result in refractive defocusing of the incident beam as shown in Figure 7b.

To determine the refractive defocusing for a spherical wave incident upon a uniform slab of plasma we can proceed as before. The result is

$$\frac{a}{A} = \frac{R_0 - d/2}{R_0 + R - d\left\{1 - \dfrac{\cos\phi}{\sqrt{\mu^2 - \sin^2\phi}}\right\}}.$$

In the absence of plasma, $a = a_0$, $\mu = 1$, so that

$$\frac{a_0}{A} = \frac{R_0 - d/2}{R_0 + R}.$$

Hence

$$\eta = \frac{a}{a_0} = \frac{1}{1 - \dfrac{d}{R + R_0}\left(1 - \dfrac{\cos\phi}{\sqrt{\mu^2 - \sin^2\phi}}\right)}. \tag{9a}$$

In the small-angle approximation, $\cos\phi \to 1$, $\sin\phi \to 0$, and

$$\eta = \frac{1}{1 + \left(\dfrac{1}{\mu} - 1\right)\dfrac{d}{R + R_0}}. \tag{9b}$$

For a spherical wave emanating from a point source, the total energy received depends on the cross-sectional area of the beam normal to the direction of propagation, i.e., it is proportional to a^2. Hence the reduction in received power due to refractive defocusing is given by η^2. A typical variation of η^2 with electron density is shown in Figure 8.

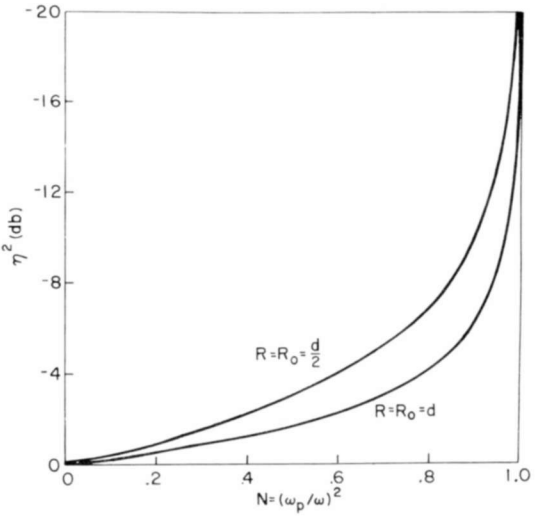


Fig. 8—Reduction in received power due to refractive defocusing by a uniform slab of plasma when a spherical wave is incident on the plasma.

For a spherical wave incident on a uniform cylinder of plasma, following the procedure as before, (see Figure 7b) the relevant equations become

$$\frac{a}{A} = \frac{r(R_0 - r)}{r(R + R_0) + 2\left(\frac{1}{\mu} - 1\right)RR_0}.$$

In the absence of the cylinder of plasma,

$$\frac{a_0}{A} = \frac{R_0 - r}{R + R_0};$$

the refractive coefficient is

$$\eta = \frac{a}{a_0} = \frac{1}{1 + 2\left(\frac{1}{\mu} - 1\right)\frac{RR_0}{r(R_0 + R)}}. \tag{10}$$

A result similar to Equation (10) was obtained previously by Heald[12,13]. Notice that this is the refractive defocusing (power) for a line source parallel to the axis of the cylinder of plasma, i.e., a cylindrical incident wave. For a point source, the reduction in power due to refractive defocusing is given by the product of Equations (9) and (10), since the defocusing will be two dimensional.

Focused Beam

A focused beam incident on a slab of plasma and focused at the center of the slab will be defocused as shown in Figure 7c. From Snell's Law and geometric considerations we arrive at

$$\eta = \frac{a_0}{a} = \left(\frac{2}{\mu}\cos\theta_i - 1\right)^{-1} \sim \frac{\mu}{2 - \mu}. \tag{11}$$

The reduction in *power* for a focused beam incident on a slab of plasma will be proportional to η^2, since η is the refractive defocusing along a radius of the incident beam and the total incident power is proportional to the area or (radius)^2 of the incident beam.

An incident beam focused at the center of a uniform cylinder of plasma will not suffer refractive defocusing in the plane normal to the axis of the cylinder (see Figure 7c). There will, however, be refractive defocusing in the direction along the axis of the cylinder since in the axial direction the cylinder will present a plane rather than cylindrical surface. This refractive defocusing is given by Equation (11).

Effects of Nonuniformity of Plasma

Consideration was given earlier to the phase change and attenuation introduced by a slab of uniform plasma to the transmitted and reflected fields of an incident microwave signal. In this section we

---

[12] M. A. Heald, "The Application of Microwave Techniques to Stellarator Research," Princeton Univ. Project Matterhorn Report MATT-17, August 1959.

[13] C. B. Wharton, "International Summer Course in Plasma Physics," *Danish Atomic Energy Comm. Report* No. 18, p. 579, 1960.

shall consider the effect that a variation in electron density with position in a plasma slab has on the phase change, attentuation, and refractive defocusing. Variation of the plasma properties both in the direction of propagation and normal to the direction of propagation are considered.

Plasma Properties Varying in Direction of Propagation

No Boundaries

Consider a plane wave incident normally on a slab of plasma of thickness d, as shown in Figure 9a. Let the electron density be a

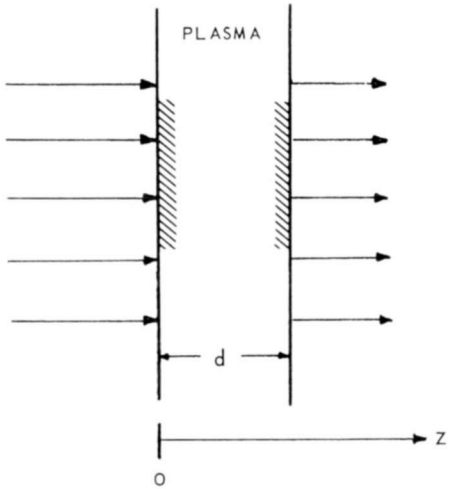


Fig. 9(a)—Plane wave incident on plasma slab whose properties vary in direction of propagation (z-direction).

function of position in the slab in the z-direction only (the direction of propagation). Since the plane wave is incident normally on the slab of plasma no refractive defocusing effects will occur even if the plasma properties vary in the direction of propagation.

Initially, neglect the effect of the plasma boundaries so that the reflected wave and multiple internal reflections within the plasma can be ignored. (We shall return to these later.) This is a reasonable assumption for a dilute plasma or a very lossy plasma.

Considering only the wave transmitted through the plasma then, the phase change $\Delta\phi$ and attenuation A_α introduced by the plasma are

$$\Delta\phi = k \int_0^d \left(\frac{\beta(z)}{k} - 1 \right) dz, \tag{12a}$$

$$A_\alpha = k \int_0^d \frac{\alpha(z)}{k}\, dz, \tag{12b}$$

where $\beta(z)$, $\alpha(z)$ are the phase and attenuation coefficients, respectively, and $k = 2\pi/\lambda$ is the wave number in free space.

It is advantageous to normalize the phase change and attenuation with respect to the thickness of the slab. Setting $s = z/d$ yields

$$\Delta\Phi = \frac{\Delta\phi}{kd} = \int_0^1 \left(\frac{\beta(s)}{k} - 1 \right) ds, \tag{13a}$$

$$\Lambda = \frac{A_\alpha}{kd} = \int_0^1 \frac{\alpha(s)}{k}\, ds. \tag{13b}$$

The effect on a plane wave introduced by the plasma slab is then

$$\exp\{-kd(\Lambda - j\,\Delta\Phi)\}.$$

When the losses in the plasma are small, $K_r \gg K_i$ (this is the only type of plasma for which present free-space microwave techniques are applicable), we can write

$$\frac{\beta(s)}{k} = K_r^{\frac{1}{2}} = \sqrt{1 - N(s)}, \tag{14}$$

where

$$N(s) = \frac{\omega_p^2(s)}{\omega^2} = \frac{e^2\, n(s)}{m\, \epsilon_0 \omega^2}.$$

For a dilute lossless plasma $N(s) \ll 1$, so that

$$\frac{\beta(s)}{k} - 1 \cong -\frac{N(s)}{2},$$

and

$$\Delta\Phi = -\frac{1}{2}\int_0^1 N(s)\, ds. \tag{15}$$

Thus for a dilute plasma the change in phase depends only on the total electron density along the path and not on the electron density distribution. In subsequent calculations we shall *not* make the dilute plasma approximation in considering the effect of the form of the electron distribution on the phase of the transmitted electromagnetic wave but will retain the restriction that $K_r \gg K_i$.

The effect of the electron density profile on the phase of an electromagnetic wave transmitted through a plasma has been considered by Wharton[14] and by Motley and Heald.[15] We shall adopt the slightly more general results due to Johnston.[16]

Consider the electron density profile in the slab of plasma to be given by a "barn roof" type of distribution (as shown in Figure 9b) of the form

$$\begin{aligned} N &= \frac{A}{1-A} N_m s, && 0 < s < (1-A), \\ N &= N_m \left[A + \frac{(1-A)}{A} (s - \{1-A\}) \right] && \\ & && (1-A) < s < 1, \end{aligned} \tag{16}$$

where

N_m is the maximum normalized electron density,

$A (\leqslant 1)$ is the height of the "shoulder" and is also the ratio of the average electron density to the maximum electron density, i.e., the average electron density in the slab is AN_m.

The normalized phase change introduced by a slab of plasma of this form of electron distribution is then

---

[14] C. B. Wharton and D. M. Sleger, "Microwave Determination of Plasma Density Profiles," *Jour. Appl. Phys.*, Vol. 31, p. 428, 1960.

[15] R. Motley and M. A. Heald, "Use of Multiple Polarizations for Electron Density Profile Measurements in High Temperature Plasmas," *Proc. Symp. on Millimetre Waves*, p. 141, Polytechnic Press, New York, 1960.

[16] M. P. Bachynski, I. P. Shkarofsky, and T. W. Johnston, *Plasmas and the Electromagnetic Field*, Chapter 13, Addison Wesley Publishing Company, Inc., Reading, Mass. (in press).

$$\Delta\Phi = \int_0^{1-A} \left[1 - N_m \frac{A}{1-A} s \right]^{1/2} ds$$

$$+ \int_{1-A}^{1} \left[1 - N_m \left\{ A + \frac{1-A}{A} (s - \{1-A\}) \right\} \right]^{1/2} ds - 1 \tag{17}$$

$$= \frac{2}{3N_m} \left\{ \frac{1-A}{A} + \left(\frac{A}{1-A} - \frac{1-A}{A} \right) (1 - AN_m)^{3/2} - \frac{A}{1-A} (1 - N_m)^{3/2} \right\} - 1.$$

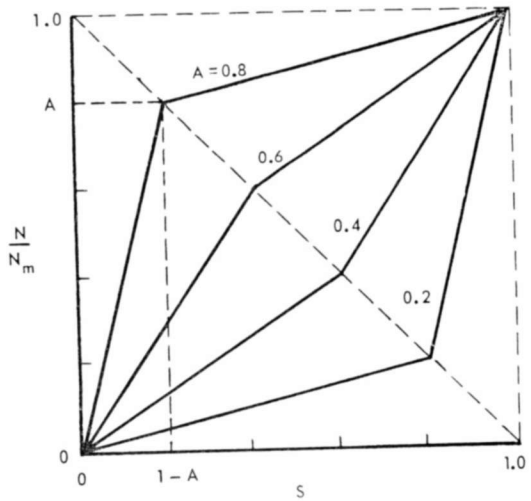


Fig. 9(b)—Electron density distribution in the plasma slab as function of position. N_m is the maximum or peak normalized electron density. A is the height of the shoulder of the "barn roof" type of distribution.

For a uniform slab, $A = 1$, $N = N_m$ and we go back to the original integral (Equation (13a)) to obtain

$$(\Delta\Phi)_{A=1} = (1 - N_m)^{1/2} - 1. \tag{18}$$

A plot of $\Delta\Phi$ versus N_m for different density profiles (different values of A) is shown in Figure 10a. If we now normalize the results to

correspond to slabs of equal total electron content (equal values of AN_m) the result is shown in Figure 10(b). The striking feature to note is that the phase is quite *insensitive* to the electron density profile, even for densities very near to the critical density, but depends almost exclusively on the total electron content. It is thus *impossible* with phase measurements performed at a *fixed* frequency to ascertain

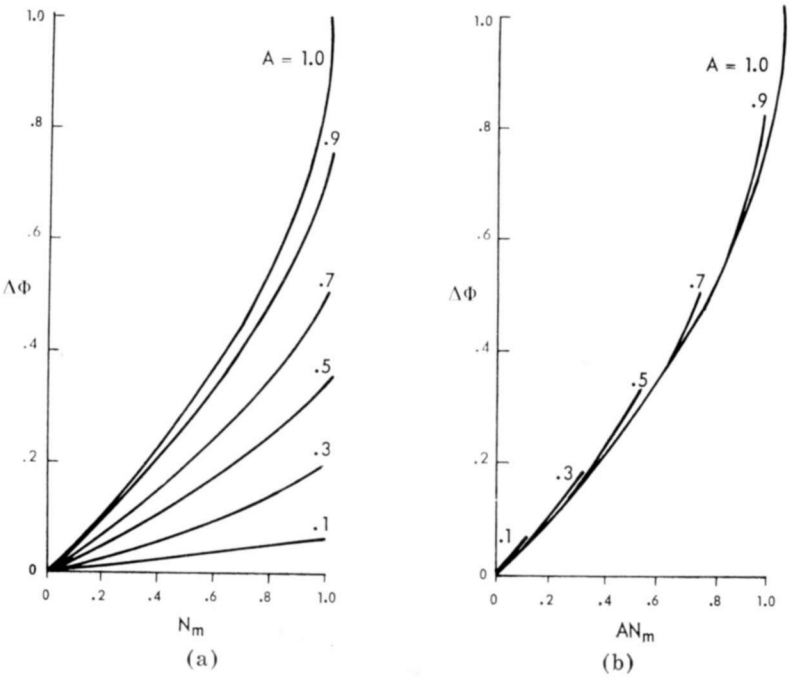


Fig. 10—Variation of phase change introduced by a plasma (a) with normalized electron density for different spatial distributions of electron density and (b) with average normalized electron density for different spatial distributions of electron density (after Johnston[16]).

with any degree of accuracy the electron density profile. Only measurements performed at a number of different radio frequencies on the same plasma can hope to give an indication of the electron density distribution. This corresponds to keeping A fixed (same plasma conditions) and varying N_m (by changing the radio frequency).

An indication of the dependence of the attenuation on the electron density profile is of value for analyzing experimental data. To a first approximation, the effective collision frequency is independent of the electron density and is considered as a constant throughout the slab in the ensuing discussion. In a plasma where $K_r \gg K_i$, the attenuation coefficient becomes

$$\frac{\alpha(s)}{k} \cong \frac{K_i}{2K_r^{1/2}} = \frac{\nu}{\omega} \frac{N(s)}{\sqrt{1-N(s)}}. \tag{19}$$

For the "barn roof" distribution of electron densities given by Equation (16) the normalized attenuation coefficient is

$$\Lambda = \frac{\nu}{\omega} \int_0^1 \frac{N(s)\, ds}{(1-N(s))^{1/2}}$$

$$= \frac{\nu}{\omega} \left[\int_0^{1-A} \frac{\dfrac{A}{1-A} N_m s \, ds}{\left[1 - \dfrac{A}{1-A} N_m s\right]^{1/2}} \right.$$

$$\left. + \int_{1-A}^{1} \frac{N_m \left[A + \dfrac{1-A}{A} (s - \{1-A\}) \right]}{\left[1 - N_m \left\{ A + \dfrac{1-A}{A} (s - \{1-A\}) \right\} \right]^{1/2}} ds \right]. \tag{20}$$

These are standard integrals which yield

$$\Lambda = \frac{\nu}{\omega} \frac{4}{3N_m} \left[\frac{1-A}{A} - \left(\frac{1-A}{A} - \frac{A}{1-A} \right) \left(1 + \frac{AN_m}{2} \right) \left(1 - AN_m \right)^{1/2} \right.$$

$$\left. - \frac{A}{1-A} \left(1 + \frac{N_m}{2} \right) \left(1 - N_m \right)^{1/2} \right]. \tag{21}$$

For a uniform slab ($A = 1$, $N = N_m$) we go back to the initial expression (Equation (13b)) to obtain

$$(\Lambda)_{A=1} = \frac{\nu}{\omega} \frac{N_m}{(1-N_m)^{1/2}}.$$

A plot of $\Lambda/(\nu/\omega)$ versus N_m for different density profiles is shown in Figure 11a. Again normalizing the results to total electron content of the slab, (Figure 11b) reveals that the effect of the density distri-

bution is not apparent until $N_m > 0.8$, i.e., the electron density must be at least 80 percent of the cutoff density before the attenuation becomes sensitive to the electron density profile. It is thus apparent that single-frequency measurements of either or both phase and attenuation will give little information on the electron density distribu-

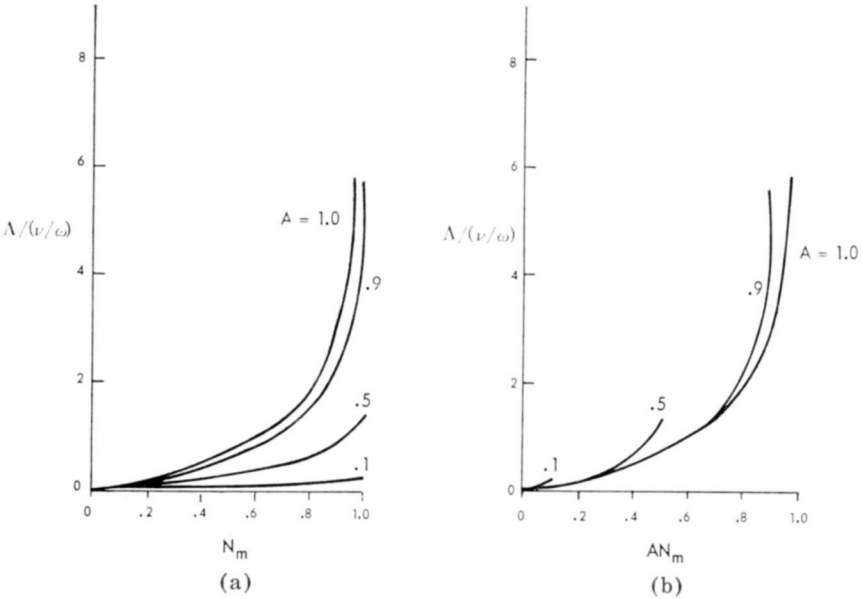


Fig. 11—Variation of normalized attenuation introduced by a plasma (a) with normalized electron density for different spatial distributions of electron density and (b) with average normalized electron density for different spatial distribution of electron density.

tion throughout the plasma. Simultaneous probing at multiple frequencies offers some hope in this direction.

Effect of Boundaries

When the effect of boundaries is taken into account for a plane wave incident normally on a slab of plasma, a part of the incident field is reflected and a part is transmitted. There are then four measurable parameters—the phase and amplitude of the transmitted wave and the phase and amplitude of the reflected signal.

The effect of the electron density varying in the direction of propagation (z-direction) on the fields reflected and transmitted by a plasma slab have been considered by a number of people (see, for example,

Budden[17,18]), with probably the best set of numerical results being recently obtained by Albini and Jahn.[19] Albini and Jahn solve the nonlinear wave equation

$$\nabla^2 E + k^2 K(z)\, E = 0$$

by machine computation for various distributions of the electron density. Of particular interest to this work is their numerical results

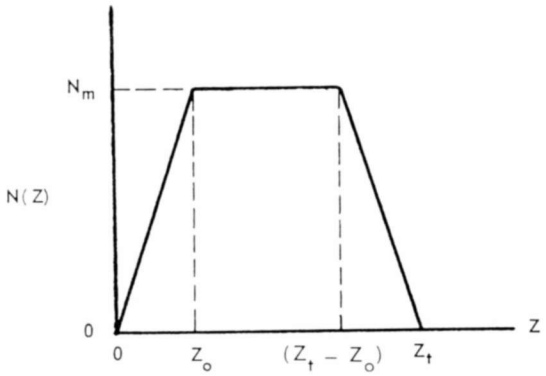


Fig. 12—Trapezoidal distribution of electron density in a plasma slab as used by Albini and Jahn[19,20] in computing effect of spatial electron distribution on transmission and reflection of electromagnetic waves.

for a slab with "trapezoidal" electron distribution, i.e., a uniform slab of plasma bounded by symmetric linear ramps of electron density. Such a trapezoidal electron distribution (in the notation of Albini and Jahn) is shown in Figure 12. Note that changing Z_0/λ is equivalent to changing the electron density profile, whereas changing Z_t/λ simply changes the thickness of the slab. The *total* electron density over a cross section of the slab is $N_m(Z_t - Z_0)$, while the *average*

[17] K. G. Budden, *Radio Waves in the Ionosphere*, Cambridge Univ. Press, 1961.

[18] G. R. Nicoll and J. Basu, "Reflection and Transmission of an Electromagnetic Wave by a Gaseous Plasma," *IEE* (London) *Monograph No. 498E* January 1962.

[19] F. A. Albini and R. G. Jahn, "Reflection and Transmission of Electromagnetic Waves at Electron Density Gradients," *Jour. Appl. Phys.*, Vol. 32, p. 75, Jan. 1961.

[20] F. A. Albini and R. G. Jahn, "Reflection and Transmission of Electromagnetic Waves at Electron Density Gradients," Tech. Note No. 3, Guggenheim Jet Propulsion Centre, Calif. Inst. of Technology, Pasadena, Oct. 1960.

electron density is $N_m(1 - Z_0/Z_t)$. We use the numerical results of Albini and Jahn, but present them in a slightly different form.

The value of the total electron content of a plasma as a universal normalizing parameter was illustrated earlier for the case of an un-

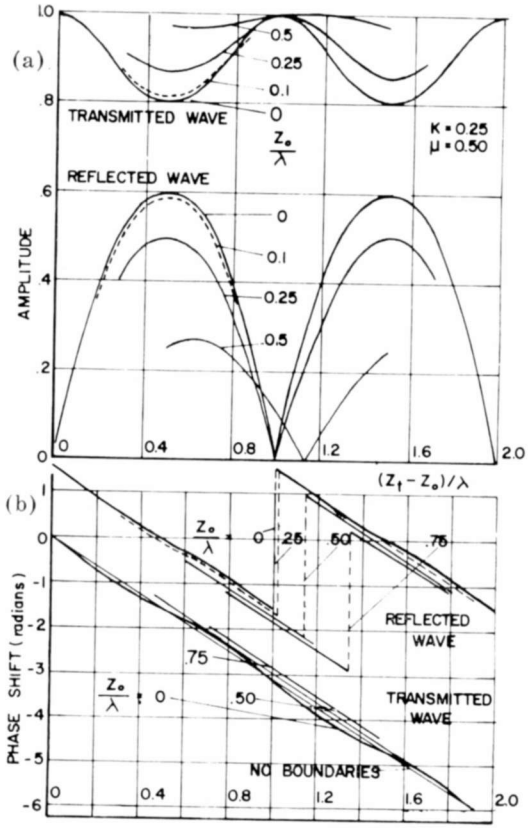


Fig. 13—Normalization to average electron density of the computations due to Albini and Jahn[19,20] showing the effect of the spatial distribution of electron density on (a) amplitude of transmitted wave for $K = 0.25$ and (b) phase shift of reflected wave and transmitted wave for $K = 0.25$.

bounded plasma. Taking the numerical results of Albini and Jahn and plotting them against $(Z_t - Z_0)/\lambda$ for different values of the "ramp" distance Z_0/λ results in the curves shown in Figure 13. Note that $(Z_t - Z_0)/\lambda$ is the normalized width of a uniform plasma slab of electron density equal to the maximum density of the trapezoidal distribution and containing the same number of electrons as the slab of

width Z_t and having a trapezoidal distribution of electrons along its width.

Figure 13a shows the variation in amplitude of the reflected and transmitted signals with slab thickness for different "ramp" distances (Z_0/λ) for a lossless plasma of dielectric coefficient $K = 0.25$. As

(c)

(d)

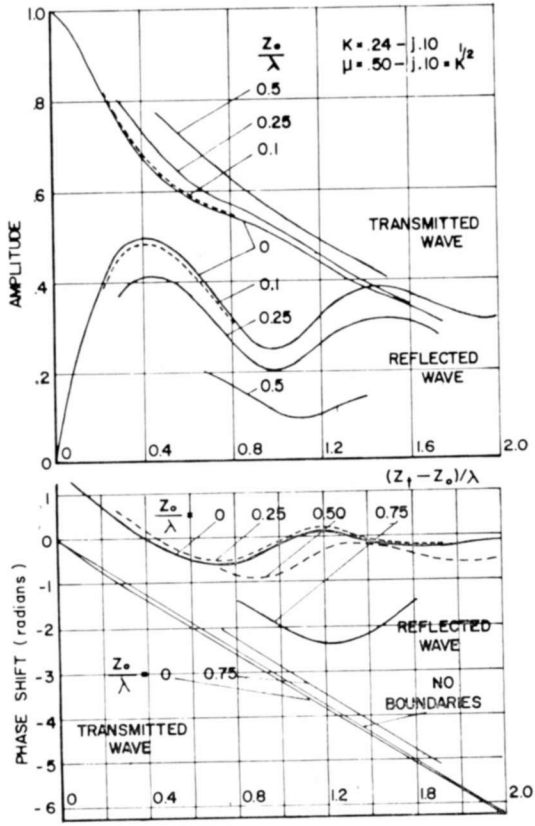


Fig. 13 (cont.)—Effect of spatial distributions of electron density on (c) amplitude of transmitted wave for $K = 0.24 - j0.10$ and (d) phase shift of reflected wave and transmitted wave for $K = 0.24 - j0.10$.

expected, the more gradual the ramp, the better is the "match" of the plasma; hence more of the signal is transmitted through the slab and less is reflected from the plasma. The important point to notice is that the positions of the maxima and minima of both the reflected and transmitted signals depend on the *total* electron density or the physical length of a uniform plasma slab, and not on the actual physical dimen-

sions of the slab. This is the case until the ramp dimensions become quite significant ($Z_0/\lambda \geq 0.50$).

The effect of losses on the behavior of the amplitude of the transmitted and reflected waves is shown in Figure 13c for a plasma of dielectric coefficient $K = 0.24 - j0.10$. In the presence of losses, the signals become less sensitive to the shape of the boundaries. In particular, the transmitted signal does not depend significantly on the shape of the electron density profile. As before, the minima and maxima, which have become drastically damped, occur at the same position for ramp distances up to 0.5λ when the slab dimensions are normalized to those of a uniform slab.

The phase of the transmitted and reflected waves[20] can also be put in a normalized form that shows their dependence on the total electron content and relative insensitivity to the shape of the electron density profile. Albini and Jahn plot the total phase shift of the transmitted wave Φ_T upon passing through a plasma slab of thickness Z_t and include the free-space path as well. To put this result into the form of the phase change introduced by the plasma $\Delta\Phi$ it is necessary to subtract from Φ_T the phase change in a path length Z_t in free-space. Thus,

$$\Delta\phi_t = 2\pi \frac{Z_t}{\lambda} - |\Phi_T|.$$

Figures 13b and 13d show plots of the phase shift of the transmitted wave introduced by the plasma normalized to total electron density. The phase shift is very nearly the same as calculated for the unbounded plasma. The effect of boundaries is to make the phase undulate slightly about the no-boundary value. The influence of losses does not introduce any significant modifications. The density profile changes the phase very slightly—by an amount which, because of the present-day precision of plasma microwave measurements, cannot be used to give any reliable information on the density profile of the plasma.

For the phase of the reflected signal, consider the reflections to occur from the slab as if the boundary were located at the midpoint between where the plasma starts and where the maximum electron density has been reached. This is again replacing the slab by an equivalent (in total electrons) uniform slab of the maximum density. We, therefore, take the phase of the reflected wave Φ_R as calculated by Albini and Jahn and add $(2\pi/\lambda)Z_0$ to their result (since the effec-

tive boundary of the slab is considered to be at $Z_0/2$ and the wave has to travel this distance twice). Thus

$$\Delta\phi_r = \Phi_R + \frac{2\pi}{\lambda} Z_0.$$

Plots of $\Delta\phi_r$ versus $(Z_t - Z_0)/\lambda$ (i.e., slab width) are shown in Figures 13b and 13d. Only at values of $Z_0/\lambda > 0.5$ does the character of the reflected phase depart notably from that of a uniform slab whose electron density is the same as the maximum of the trapezoidal electron distribution and which contains the same total number of electrons as the trapezoidal slab.

Plasma Properties Varying Normal to Direction of Propagation

Consider a plane wave normally incident upon a plasma slab as shown in Figure 14a. The properties of the slab are constant throughout the thickness of the slab, but depend on the distance from the center of the plasma slab. That is, the electron density $n(r)$ varies in the direction normal to the direction of propagation. At the incident boundary of the slab $(z = 0)$, the phase front of the incident plane wave coincides with the front face of the slab. At the second boundary of the slab $(z = d)$ the phase of the wave emanating at a height r above the center line of the slab is

$$\Delta\phi = kd\,[1 - \sqrt{1 - N(r)}\,].$$

The phase difference between the wave coming through the slab at height r and the wave coming out at the center of the slab $(r = 0)$ is

$$\Delta\phi(r) - \Delta\phi(0) = kd\,[-\sqrt{1 - N(r)} + \sqrt{1 - N_0}], \qquad (22)$$

where $$N_0 = \frac{n_0 e^2}{m\epsilon_0\omega^2},$$

n_0 is the electron density along the width of the slab at position $r = 0$,

$$N(r) = n(r)\,\frac{e^2}{m\epsilon_0\omega^2}.$$

Equation (22) is thus the equation of the phase front (surface of constant phase) of the wave emanating from the plasma; the front

has the form shown in Figure 14b (provided $N_0 > N(r)$). When $N(r) > N_0$, the curvature is in the opposite direction.

In physical space the important parameter is the optical path

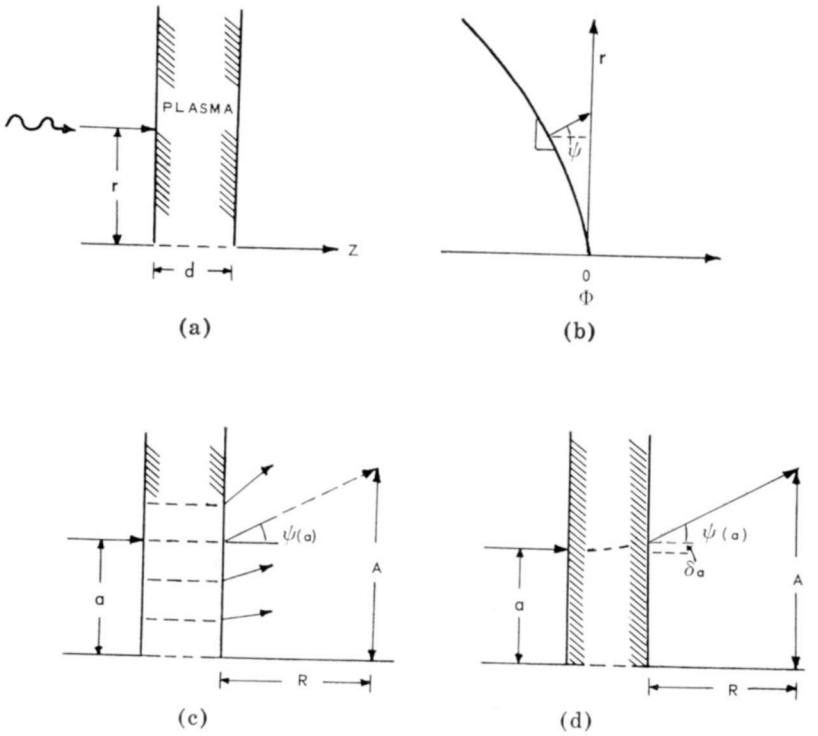


Fig. 14—Notation for discussing energy propagation in a plasma whose properties vary in direction normal to direction of propagation: (a) plane wave incident at distance r from center of plasma whose properties depend on distance from center, (b) surface of constant phase of wave emerging from plasma, (c) approximate path of ray in plasma and emerging from plasma, and (d) accurate path of ray in plasma and emerging from plasma.

length. The surface of constant path length is given by

$$\Phi(r) = \frac{\Delta\phi(r) - \Delta\phi(0)}{k} = d[-\sqrt{1-N(r)} + \sqrt{1-N_0}]. \qquad (23)$$

This is the shape in free space of the planes of constant phase given in terms of physical dimensions.

Considering that the phase fronts are orthogonal to the direction of energy travel, the ray incident upon the slab of plasma at height r is refracted so that it emerges at some angle $\psi(r)$ as shown in Figure 14c. The angle $\psi(r)$ is given by

$$\tan \psi(r) = -\frac{d\Phi(r)}{dr} = -\frac{d}{2}\frac{dN(r)/dr}{\sqrt{1-N(r)}}. \tag{24}$$

If the beam of incident radiation of radius a is intercepted by a plane of half-width A located at distance R from the second surface of the slab, then the refractive defocusing introduced by the nonuniform electron density variation of the slab, η, is given by

$$\eta = \frac{a}{A}.$$

(The true situation is shown in Figure 14d. The normally incident ray at height a undergoes continuous refraction as it traverses the slab and emerges at a height $(a + \delta a)$ traveling in the direction $\psi(a)$. Subsequently we shall assume $\delta a \ll a$ and consider the incident ray to travel at the same height in the slab, but to emerge at angle $\psi(a)$.) In practice this is probably a good assumption since the presence of the dielectric plates of the plasma container will tend to counterbalance the refractive defocusing and in effect make δa small.

The effect of the nonuniform electron distribution in the plasma slab can be thought of as a plasma lens of constant electron density but shaped so as to give the same phase change to an incident plane wave as does the plasma slab.

We can thus write

$$\tan \psi(a) = \frac{A-a}{R} = -\frac{d}{2}\frac{(dN(r)/dr)_{r=a}}{\sqrt{1-N(a)}}, \tag{25}$$

or

$$\eta = \frac{a}{A} = 1 + \frac{Rd}{2A}\frac{dN(r)/dr}{\sqrt{1-N(a)}}. \tag{26}$$

A number of variations of the electron density with direction normal to the direction of propagation, and the corresponding refractive defocusing coefficient, are listed in Table I. A convenient distribution

for laboratory experimental purposes is the parabolic distribution, particularly since Γ represents the amount by which the electron density has decreased at the edge of the experimental plasma container relative to the electron density at the center. Despite the fact that the resulting equation for $\eta(a)$ is transcendental it can be very readily solved for fixed values of A.

Table I

Electron Distribution in Plasma Slab $N(r)$	$\frac{dN(r)}{dr}$	$\eta(a)$
$N(r) = N_0$ (constant)	0	1
$N(r) = N_0(1 - \Gamma(r/r_0))$ (linear)	$-\Gamma N_0/r_0$	$1 - \frac{d}{2r_0}\frac{R}{A}\frac{\Gamma N_0}{\sqrt{1-N(a)}}$
$N(r) = N_0(1 - \Gamma(r/r_0)^2)$ (parabolic)	$-2\Gamma N_0(r/r_0^2)$	$\frac{1}{1 + d\frac{R}{r_0^2}\frac{\Gamma N_0}{\sqrt{1-N(a)}}}$
$N(r) = N_0 \exp\{-\Gamma(r/r_0)^2\}$ (Gaussian)	$-2\Gamma N(r)(r/r_0^2)$	$\frac{1}{1 + d\frac{R}{r_0^2}\frac{\Gamma N(a)}{\sqrt{1-N(a)}}}$
$N(r) = N_0 J_0\left(\frac{\Gamma r}{r_0}\right)$	$-N_0\frac{\Gamma}{r_0}J_1(\Gamma r/r_0)$	$1 - \frac{d}{2r_0}\frac{R}{A}\frac{\Gamma N_0 J_1(\Gamma r/r_0)}{\sqrt{1-N(a)}}$

r_0 = radius of actual plasma (finite in experiment)
r_0 = radius of laboratory plasma bottle.

Numerical results for a parabolic distribution of electron densities as determined from Equation (26) are shown in Figure 15. As can be seen, if the plasma is nonuniform then very strong attenuation effects can be obtained due to these refractive effects.

The geometrical-optics type of refractive defocusing that has been considered cannot take into account the phase difference between the various rays (from different radii) as they reach the receiving antenna. In order to do this, resort must be made to diffraction theory as shown in the next section.

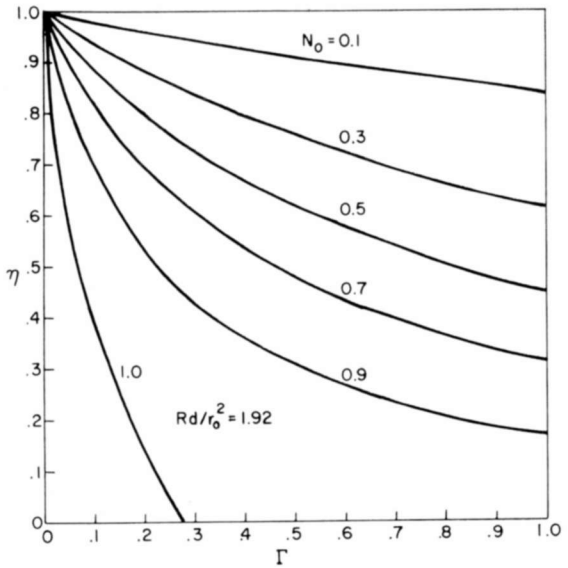


Fig. 15(a)—Refractive defocusing effect due to plasma properties varying normal to direction of propagation as function of plasma nonuniformity for different electron densities at center of plasma. (Parabolic distribution of electron density in direction normal to propagation and $A = r_0$ is assumed.)

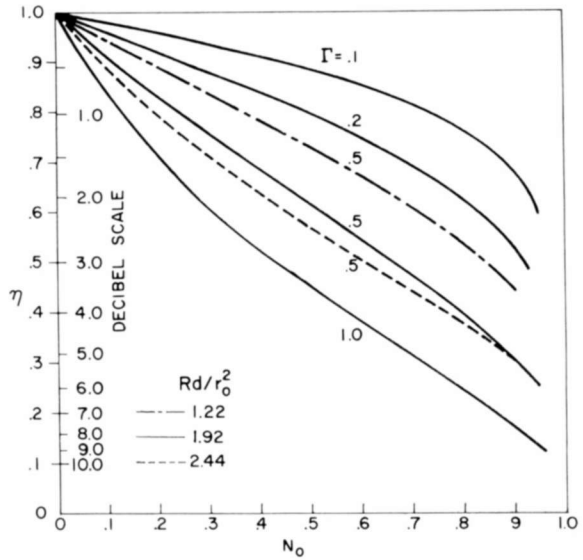


Fig. 15(b)—Refractive defocusing effect due to plasma properties varying normal to direction of propagation as function of electron density at center of plasma for different geometrical arrangements and degrees of nonuniformity. ($A = r_0$ is assumed.)

Electromagnetic Wave Propagation through Laboratory Plasmas as a Diffraction Problem

Most laboratory-scale plasmas are only a few wavelengths in extent, and hence when the properties of the plasma are to be measured using electromagnetic waves, diffraction will play a major role in deter-

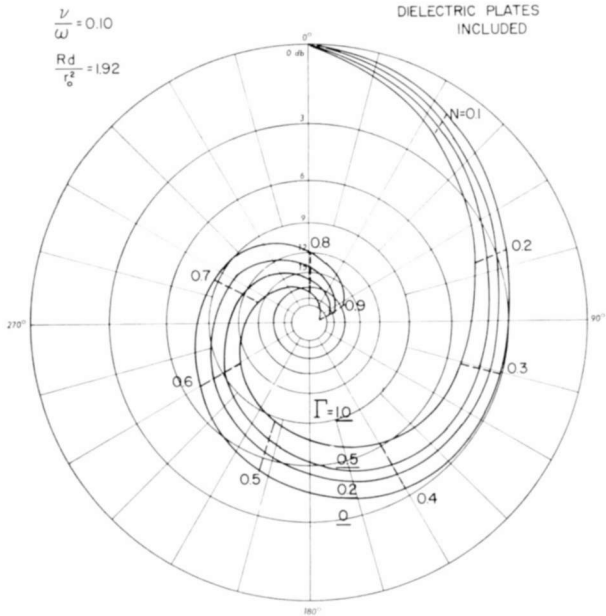


Fig. 15(c)—Phase and attenuation of a plane incident wave transmitted through a slab of plasma including the effect of dielectric plates and of lateral defocusing in one dimension for a parabolic distribution of electron density in direction normal to direction of propagation. ($A = r_0$)

mining the electromagnetic energy that emanates from the plasma, and its distribution in space.

When considering diffraction phenomena, the Kirchhoff scalar diffraction formula, although not rigorous in its formulation, has enjoyed considerable success when applied to actual physical problems. Using the scalar diffraction formula, the field u at point p may be written

$$u(p) = -\frac{1}{4\pi} \iint_S \left\{ \frac{e^{-jks}}{s} \frac{\partial \psi}{\partial n} - \psi \frac{\partial}{\partial n} \left(\frac{e^{-jks}}{s} \right) \right\} dS, \qquad (27)$$

where s is the distance of the field point p to the surface of integration,

ψ is the value of incident field (amplitude and phase) at the element of integration,

n is the normal derivative in the plane of integration,

S is the surface of integration.

Application to Point Source Illuminating Finite Plasma Slab

To consider the diffraction phenomena introduced by a finite plasma, let a source of electromagnetic energy be situated at point S (see Figure 16a) a distance $(R - d)$ from a uniform slab of plasma of thickness d. The exit pupil of the system is an aperture of radius r located at the exit position of the plasma slab. (In practice,[8] it is found that the exit pupil of a finite plasma container determines the major diffraction effects, so that the above model is a good approximation to a cylindrical slab of plasma of radius r and thickness d.) This exit pupil is taken as the surface of integration. The problem is then to determine the incident field over the exit pupil and perform the integration according to Equation (27) in order to evaluate the field at the point p.

The field incident from the source must pass through the plasma slab before it reaches the point of integration $p'(r,\phi)$ where r,ϕ are the polar co-ordinates in the exit pupil. With reference to Figure 16b, let L_1 be the path of the incident radiation in free space and L_2 the path of the incident radiation in the plasma which reaches the point p'. Let the refractive index of the plasma be $K^{1/2} = (\beta/k) - j\alpha/k$. The phase of the incident field at p' is then

$$\phi_i = k\left(L_1 + \frac{\beta}{k} L_2\right).$$

and the amplitude of the incident field is

$$U = U_0(\phi)\,\frac{\exp\{-k(\alpha/k)L_2\}}{R},$$

where U_0 is the free-space radiation pattern of the source representing both the strength and directivity of the radiation, and

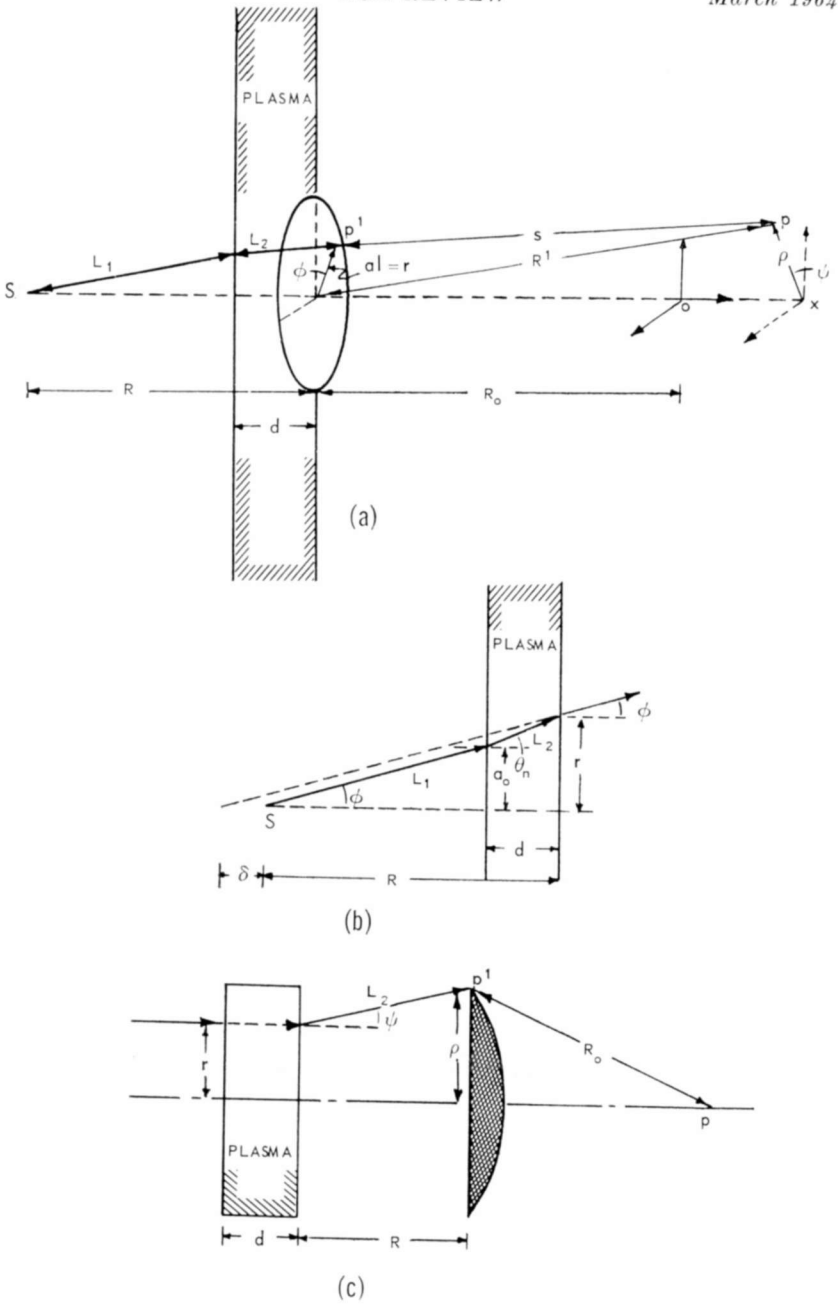


Fig. 16—(a) Geometry for derivation of diffraction due to a plasma slab located in front of a circular aperture in a metal screen; (b) optical distance traveled by radiation from source to exit pupil of diffracting system; and (c) optical distance traveled by radiation through plasma and from exit pupil to field point p.

$\exp\{-k(\alpha/k)L_2\}$ represents the attenuation of the incident field in its passage through the plasma.

(The effects due to the boundaries of the plasma have been neglected.) Again with reference to Figure 16b we can write

$$\sin\phi = \frac{\beta}{k}\sin\theta_n,$$

$$\tan\phi = \frac{a}{R-d} = \frac{r}{R+\delta},$$

$$\tan\theta_n = \frac{r-a}{d}.$$

For most practical purposes we can set $\sin\phi \sim \tan\phi$, $\sin\theta_n \sim \tan\theta_n$, which leads to

$$\frac{a}{r} = \frac{1}{1+\dfrac{d}{(R-d)(\beta/k)}},$$

$$\delta = d\left(\frac{1}{(\beta/k)} - 1\right).$$

We can write

$$L_1 = [(R-d)^2 + a^2]^{1/2} \sim R - d + \frac{a^2}{2(R-d)} \quad \text{for a} \ll (R-d),$$

$$= (R-d) + \frac{r^2}{2(R-d)} \frac{1}{\left(1+\dfrac{d}{(R-d)(\beta/k)}\right)^2} = (R-d) + gr^2,$$

where

$$g = \frac{1}{2(R-d)\left(1+\dfrac{d}{(R-d)(\beta/k)}\right)^2}$$

and

$$L_2 = [d^2 + (r-a)^2]^{1/2} \sim d + \frac{(r-a)^2}{2d} \quad \text{for } (r-a) \ll d,$$

$$= d + \frac{r^2}{2d} \frac{1}{\left[1 + \left(\frac{R}{d} - 1\right)\frac{\beta}{k}\right]^2} = d + fr^2,$$

where

$$f = \frac{1}{2d\left[1 + \frac{\beta}{k}\left(\frac{R}{d} - 1\right)\right]^2}.$$

The incident field over the aperture is thus

$$\psi = \frac{U_0(r)}{R} \exp\{-\alpha d - jk(R + ((\beta/k) - 1)d) - \alpha f r^2 - jk(g + (\beta/k)f)r^2\}. \quad (28)$$

The distance s from the field point p to the point of integration p' can be written

$$s = [(R_0 + x)^2 + \rho^2 + r^2 - 2\rho r \cos(\psi - \phi)]^{1/2}.$$

Letting $R' = [(R_0 + x)^2 + \rho^2]^{1/2}$,

$$s = R'\left[1 + \frac{r^2 - 2\rho r \cos(\psi - \phi)}{(R')^2}\right]^{1/2} \sim R' + \frac{r^2}{2R'} - \frac{2\rho r}{2R'}\cos(\psi - \phi). \quad (29)$$

For the far field, the diffraction integral can be written

$$u(p) = \frac{jk}{2\pi} \iint_S \psi \frac{e^{-jks}}{s} dS. \quad (30)$$

We now make the normalization $r = al$, where a is the radius of the exit pupil and $0 \leq l \leq 1$. Using Equations (29) and (30), the field at the point p can be written

$$u(p) = \frac{jka^2}{2\pi RR'} \exp\left\{-jk\left[R + R' + d\left(\frac{\beta}{k} - 1\right)\right] - \alpha d\right\}$$
$$\int_0^1 \int_0^{2\pi} U_0(l) \exp\left\{-\alpha f a^2 l^2 - jk\left[g + f\left(\frac{\beta}{k}\right) + \frac{1}{2R'}\right] a^2 l^2 + jk\left[\frac{\rho a}{R'}\right] l \cos(\psi - \phi)\right\} l dl d\phi. \quad (31)$$

Equation (31) includes the usual far-field approximation. If a computer is available or if greater accuracy is desired, then the exact values of for ψ and s can be used.

Let

$$P = k\left(g + \frac{\beta}{k} f + \frac{1}{2R'}\right) a^2,$$

$$Q = \frac{ka}{R'} \rho,$$

$$\theta = k\left(R + R' + \left(\frac{\beta}{k} - 1\right) d\right).$$

Then

$$u(p) = j \frac{a^2}{\lambda RR'} \exp\{-j\theta - \alpha d\} \int_0^1 \int_0^{2\pi} U_0(l) \exp\{-\alpha f a^2 l^2 - jPl^2 + jQl \cos(\psi - \phi)\} l dl d\phi. \quad (32)$$

The integration with respect to ϕ is readily executed to give

$$u(p) = j \frac{(2\pi a^2)}{\lambda RR'} \exp\{-j\theta - \alpha d\} \int_0^1 U_0(l) \exp\{-\alpha f a^2 l^2\} J_0(Ql) l dl, \quad (33)$$

where J_0 is the zero-order Bessel function of the first kind. For the

field along the principal axis of the system $Q = 0$ so that

$$u(p) = \frac{j(2\pi a^2)}{\lambda RR'} \exp\{-j\theta - \alpha d\} \int_0^1 U_0(l) \exp\{-\alpha f a^2 l^2 - jPl^2\}\, ldl. \quad (34)$$

Consider then the field along the principal axis for a lossless plasma ($\alpha = 0$) that is uniformly illuminated from a point source located at S (i.e., $U_0(l) = u_0 = \text{constant}$). In this case

$$\frac{u(p)}{u_0} = j\frac{(2\pi a^2)}{\lambda RR'} \exp\{-j\theta\} \int_0^1 \exp\{-jPl^2\}\, ldl$$

$$= j\frac{(2\pi a^2)}{\lambda RR'} \frac{1}{2} \exp\left\{-j\theta - \frac{jP}{2}\right\}\left[\frac{\sin(P/2)}{P/2}\right]. \quad (35)$$

The intensity along the principal axis is thus

$$I(p) = \frac{k^2 a^4}{4(RR')^2}\left[\frac{\sin(P/2)}{P/2}\right]^2. \quad (36)$$

This is just the field along the principal axis of a circular aperture. To study the influence of the plasma we must consider the parameter P. After some algebra we can write

$$\left(\frac{P}{2}\right) = \frac{ka^2}{4}\left(\frac{1}{R'} + \frac{1}{R + d\left(\frac{1}{\beta/k} - 1\right)}\right)$$

$$\sim \frac{ka^2}{4}\left(\frac{1}{R'} + \frac{1}{R} - \frac{d}{R^2}\left(\frac{1}{\beta/k} - 1\right)\right).$$

The effect of the plasma is thus to decrease the value of $(P/2)$ since the term $(\beta/k)/[d + (R-d)\beta/k]$ decreases as β/k decreases. The effect of the plasma is to shift the axial radiation pattern of the system.

When the plasma is lossy ($\alpha \neq 0$), and if the directivity of the

incident radiation can be approximated in the form

$$U_0(l) = u_0 \exp\{-\beta_0 l^2\}$$

then Equation (34) can be written

$$\frac{u(p)}{u_0} = \frac{j(2\pi a^2)}{\lambda RR'} \frac{1}{2} \exp\left\{-j\theta - \frac{j}{2}(P - j[\alpha f a^2 + \beta_0])\right\} \left\{ \frac{\sin\left[\dfrac{P - j(\alpha f a^2 + \beta_0)}{2}\right]}{\left[\dfrac{P - j(\alpha f a^2 + \beta_0)}{2}\right]} \right\} \quad (37)$$

The intensity along the principal axis is then

$$I(p) = \frac{k^2 a^4}{4(RR')^2} \frac{\exp\{-\beta_0 - \alpha f a^2\}}{\left(\dfrac{P^2 + (\beta_0 + \alpha f a^2)^2}{4}\right)} \left[\frac{1}{2}\{\cosh(\beta_0 + \alpha f a^2) - \cos P\}\right] \quad (38)$$

The effect of the losses in the plasma on the intensity is the same as the effect of using a directive antenna.

Application to Plane Wave Illumination of Plasma Whose Properties Change in Radial Direction

Consider a plane wave incident on a slab of plasma as shown in Figure 16c. The plasma is considered to be nonuniform with the electron density $N(r)$ depending upon distance r from the center of the plasma. A ray incident upon the plasma at r emerges from the plasma at angle ψ. Assume a perfect lens is located a distance R from the second surface of the slab. The energy at the focus of the lens is then readily determined if we know the field distribution incident upon the lens. We can thus write

$$u(p) = \frac{jk}{2\pi} \int_S \psi \frac{\exp\{-jkR_0\}}{R_0} dS = \frac{jk}{2\pi} \frac{\exp\{-jkR_0\}}{R_0} \int_S \psi \, dS, \quad (39)$$

where ψ is the field incident on the lens. The incident field at the point p' is then

$$\psi(p') = \psi_0 \exp\{-jd\sqrt{K} - jkL_2\}, \tag{40}$$

where ψ_0 is the amplitude of the field incident on the plasma. Now

$$L_2 = [R^2 + (\rho - r)^2]^{1/2} = R\,[1 + \tan^2\psi]^{1/2} \sim R + \frac{R}{2}\tan^2\psi.$$

From Equation (25)

$$\tan\psi(r) = -\frac{d}{2}\frac{dN(r)/dr}{\sqrt{1-N(r)}}.$$

If we restrict ourselves to a parabolic electron variation (any of the other variations listed in Table I could be used as well) then

$$\tan\psi(r) = \frac{d\,N_0\,\Gamma}{\sqrt{1-N(r)}}\frac{r}{r_0^2}.$$

Hence we can write

$$\tan^2\psi = \frac{\left(\frac{d}{r_0}\Gamma N_0\right)^2\left(\frac{r}{r_0}\right)^2}{1-N_0\left(1-\Gamma\left(\frac{r}{r_0}\right)^2\right)} \sim \left(\frac{d}{r_0}\Gamma N_0\right)^2(1+N_0+N_0^2)$$

$$\left(\frac{r}{r_0}\right)^2 = 2S\left(\frac{r}{r_0}\right)^2 \tag{41}$$

where

$$S = \frac{1}{2}\left(\frac{d}{r_0}\Gamma N_0\right)^2(1+N_0+N_0^2).$$

Thus

$$L_2 = R + RS\left(\frac{r}{r_0}\right)^2,$$

and

$$\sqrt{K} = \frac{\beta}{k} - j\frac{\alpha}{k} = \left[1 - \frac{N_0(1-\Gamma(r/r_0)^2)}{1+\nu^2/\omega^2}(1+j\nu/\omega)\right]^{1/2}$$

$$\cong \left[1 - \frac{N_0'}{2}\left(1+j\frac{\nu}{\omega}\right)\left[1+\frac{N_0'}{4}\left(1+j\frac{\nu}{\omega}\right)\right]\right] + \frac{N_0'}{2}\Gamma\left(1+j\frac{\nu}{\omega}\right)\left[1+\frac{N_0'}{2}\left(1+j\frac{\nu}{\omega}\right)\right]\left(\frac{r}{r_0}\right)^2$$

$$= P + Q\left(\frac{r}{r_0}\right)^2,$$

where $N_0' = N_0/(1+\nu^2/\omega^2)$,

P, Q are complex.

Thus we have

$$\psi(p') = \psi_0 \exp\left\{-jk(R+Pd) - jk(RS+Qd)\left(\frac{r}{r_0}\right)^2\right\}. \tag{42}$$

The field at the field point p can thus be written

$$u(p) = \frac{jk}{2\pi}\frac{1}{R_0}\exp\{-jk(R_0+R) - jkPd\}\int_0^{\rho_0}\int_0^{2\pi} \exp\left\{-jk(RS+Qd)\left(\frac{r}{r_0}\right)^2\right\}\rho d\rho d\phi, \tag{43}$$

where ρ_0 is the radius of the lens. It has been established that

$$\frac{r}{\rho} = \eta,$$

where η is the coefficient of refractive defocusing. Hence

$$\rho d\rho \cong \frac{1}{\eta^2} r dr.$$

Further make the substitution

$$r = r_0 l,$$

where r_0 is the radius of the plasma, and

$$0 \leq l \leq 1.$$

Then

$$u(p) = \frac{jk}{\eta^2} \frac{r_0^2}{R_0} \exp\{-jk(R_0 + R) - jkPd\} \int_0^{\eta(\rho_0/r_0)} \exp\{-jk\,(RS + Qd)\,l^2\}\, l\,dl$$

$$= \frac{jk}{2} \rho_0^2 \frac{1}{R_0} \exp\{-jk(R_0 + R) - j(Pkd + W)\} \left[\frac{\sin W}{W}\right] \quad (44)$$

where

$$W = \frac{k}{2}(RS + Qd)\eta^2(\rho_0^2/r_0^2).$$

Note that P and W are complex so that the expression is not as simple as it seems. When the plasma and the lens are the same size, $\rho_0 = r_0$, and the values calculated for η in the preceding section are applicable. (Otherwise η is calculated from Equation (26) in which A is set equal to ρ_0). In the limiting case of

(a) uniform plasma, $\Gamma = 0$, so that $S = 0$, $Q = 0$ and $W = 0$;

$$u(p) = \frac{jk}{2} \rho_0^2 \frac{1}{R_0} \exp\left\{-jk\,[R_0 + R] - jkd\left[1 - \left(\frac{N_0'}{2}\right)\left(1 + \frac{N_0'}{4}\right)\right] - \frac{\nu}{\omega} \frac{N_0'}{2}\left(1 + \frac{N_0'}{2}\right) kd\right\} \quad (45)$$

(b) lens against plasma, $R = 0$, $\eta = 1$ and $W = \frac{k}{2} Qd$;

$$u(p) = \frac{jk}{2} \rho_0^2 \frac{1}{R_0} \exp\left\{-jkR_0 - jk\left[P + \frac{Q}{2}\right] d\right\} \left[\frac{\sin\left\{\frac{kd}{2} Q\right\}}{\left\{\frac{kd}{2} Q\right\}}\right]. \quad (46)$$

Numerical results for a parabolic distribution of electrons are shown in Figure 17, which illustrates the effect of the nonuniformity in electron density, the effect of the distance of the lens from the plasma, and the effect of collision frequency. One can, for example, perform measurements at different distances from the plasma in order to determine the degree of nonuniformity of the plasma. Note that

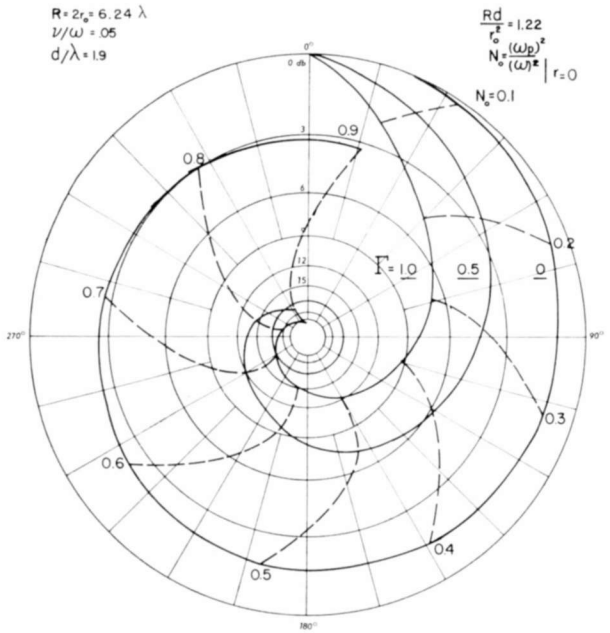
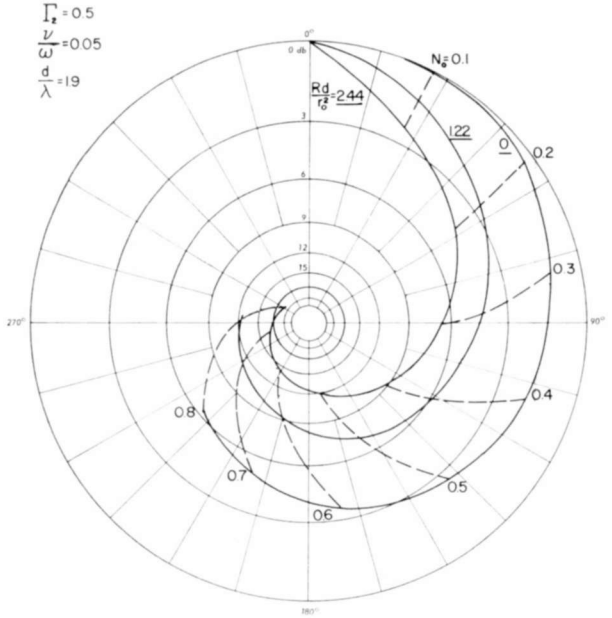


Fig. 17(a)—Phase and intensity of an incident plane wave transmitted through a slab of plasma and diffracted by a circular metal screen forming the exit pupil of the microwave optics system showing the effect of nonuniformity in electron density in the direction normal to the direction of propagation. (A parabolic distribution of electron density in the lateral direction and $A = r_0 = \rho_0$ is assumed.)

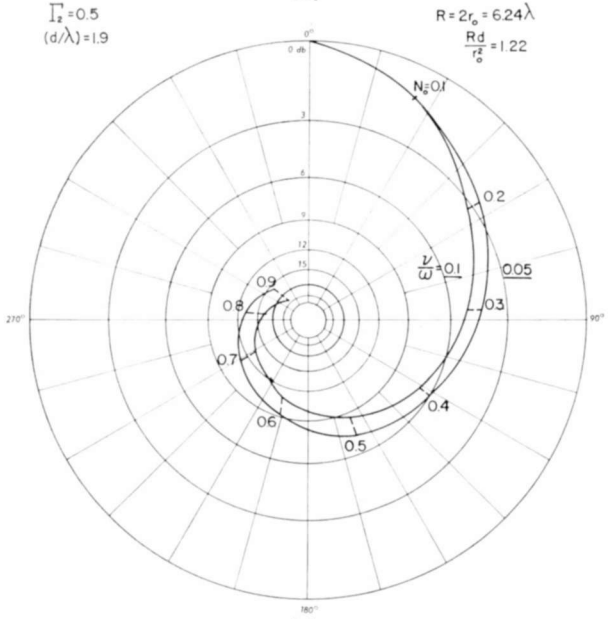
the nonuniformity in the electron density of the plasma and the geometry of the arrangement (Rd/r_0^2) have a far greater influence on the electromagnetic wave passing through the plasma than does the collision frequency.

Conclusion

Analytic expressions and typical numerical results have been presented for the effect on transmission, reflection, and absorption of electromagnetic waves of plasma and dielectric boundaries, refractive



(b)



(c)

Fig. 17 (cont.)—Phase and intensity of an incident plane wave transmitted through a slab of plasma and diffracted by a circular metal screen forming the exit pupil of the microwave optics system showing (b) the effect of the distance of the receiving lens from the exit pupil and (c) the effect of various values of collision frequency. ($A = r_0 = \rho_0$)

defocusing by slabs and cylinders of plasma for plane wave and spherical wave incidence, the effect of nonuniformity of the plasma both along the direction of propagation and normal to the direction of propagation, and for diffraction effects due to the finite size of a plasma.

The influence of these effects is significant for any laboratory measurements of plasmas using free-space microwave techniques. In many instances they predominate over the effect of the parameters of the plasma that are being determined and limit the amount (accuracy and detail) of information regarding the plasma that can be obtained. It is, therefore, essential either to minimize these effects or to take them into account in any quantitative interpretation of experimental measurements.

Acknowledgments

This work was carried out under Air Force Cambridge Research Laboratories Contract AF 19(604)-7334. The authors are indebted to Dr. G. G. Cloutier for numerous stimulating discussions during the course of the investigations and for the numerical computation of some of the results.

MICROWAVE TUNNEL-DIODE AMPLIFIERS WITH LARGE DYNAMIC RANGE

By

R. Steinhoff and F. Sterzer

RCA Electronic Components and Devices,
Princeton, N. J.

Summary—*The large-signal behavior of tunnel-diode amplifiers is analyzed, and curves for calculating the gain-saturation characteristics of gallium arsenide, germanium, and gallium antimonide tunnel-diode amplifiers are presented. Agreement between theory and experiment is good. The design of microwave tunnel-diode amplifiers with dynamic ranges of about 90 decibels (for 1 mc noise bandwidth) is then discussed, and experimental results obtained with such amplifiers are given.*

Introduction

Conventional microwave tunnel-diode amplifiers use germanium or gallium antimonide tunnel diodes with peak currents of only a few milliamperes. Such amplifiers generally have noise figures in the 3- to 5-decibel range, power outputs of several microwatts, and dynamic ranges of less than 70 decibels* for power gains of about 15 decibels. This paper describes microwave amplifiers that use gallium arsenide tunnel diodes with peak currents of more than 20 milliamperes. Although these GaAs amplifiers have a higher noise figure (NF $\approx$ 6 decibels) than amplifiers using Ge or GaSb diodes, they can deliver hundreds of microwatts of output power and have dynamic ranges that exceed 90 decibels. Cascaded amplifiers consisting of low-noise Ge or GaSb first stages followed by a GaAs second stage can combine 3- to 5-decibel noise figures with high power outputs and large dynamic ranges.

In the first section of this paper, the large-signal behavior of tunnel-diode amplifiers is analyzed, and curves for calculating the gain-saturation characteristics of GaAs, Ge, and GaSb tunnel-diode amplifiers are presented. The next section discusses microwave amplifiers that use high-current GaAs diodes, and gives design procedures and experimental results. The final section discusses the use of cascaded amplifiers to obtain high power output with low noise figures.

* In this paper dynamic range is defined on the basis of 1-mc noise bandwidth and 3-db gain compression.

Large-Signal Analysis of Tunnel-Diode Amplifiers

The power gain, G, of a circulator-coupled tunnel-diode amplifier of the type shown in Figure 1 is given by[1,2]

$$G = \frac{(R_0 + R)^2 + X^2}{(R_0 - R)^2 + X^2}, \tag{1}$$

where R_0 is the characteristic impedance of the transmission line con-

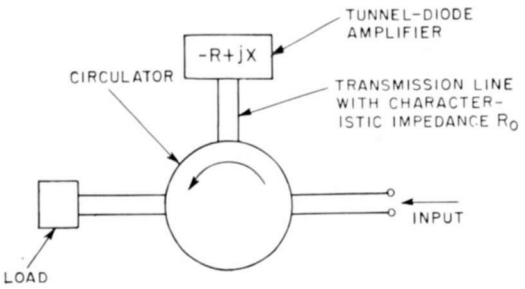


Fig. 1—Circulator-coupled tunnel-diode amplifier.

necting the circulator to the tunnel-diode amplifier and

$$Z = -R + jX$$

is the impedance of the tunnel-diode amplifier. It is assumed that the circulator is ideal and is matched to the transmission lines connected to its ports.

At the resonance frequency of the amplifier, Equation (1) simplifies to

$$G = \frac{(R_0 + R_R)^2}{(R_0 - R_R)^2} = \frac{\left(1 + \dfrac{R_R}{R_0}\right)^2}{\left(1 - \dfrac{R_R}{R_0}\right)^2}, \tag{2}$$

---

[1] K. K. N. Chang, "Low-Noise Tunnel-Diode Amplifier," *Proc. I.R.E.*, Vol. 47, p. 1268, July 1959.

[2] M. E. Hines and W. W. Anderson, "Noise Performance Theory of Esaki (Tunnel) Diode Amplifiers," *Proc. I.R.E.*, Vol. 48, p. 789, April 1960.

where R_R is the value of Z at resonance. Equation (2) is plotted in Figure 2.

An a-c equivalent circuit of the amplifier that is general enough for most practical applications is shown in Figure 3. Here the tunnel diode is shunted by a parallel circuit admittance $G_c + jB_c$. The equiva-

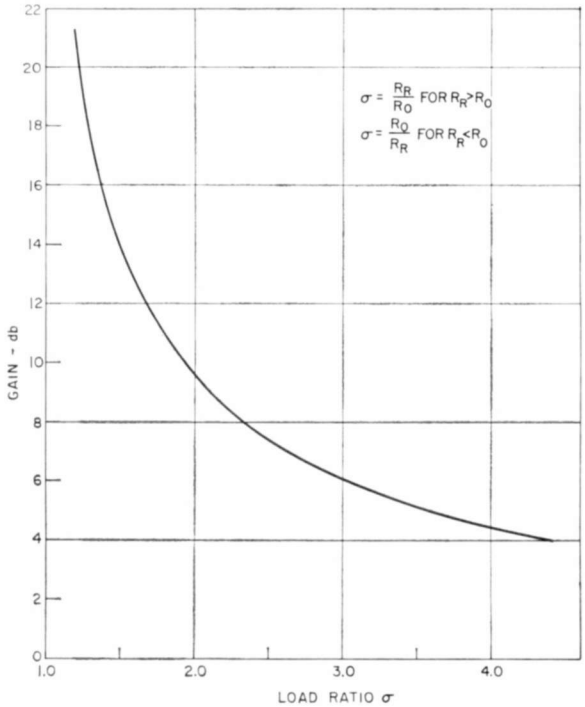


Fig. 2—Gain versus load ratio at resonant frequency for a circulator-coupled tunnel-diode amplifier.

lent circuit of the diode itself consists of three elements connected in series—an inductance L_d, a resistance r_d, and a voltage-dependent a-c junction resistance R_d shunted by a voltage-dependent junction capacitance C_d*. In the limit of vanishingly small r-f signals, R_d is given by

$$(R_d)_s = \frac{dV_d}{dI_d}, \tag{3}$$

* The effects of the variation of C_d with voltage are usually small, and are neglected in this analysis.

where V_d is the voltage across the diode junction and I_d is the current through R_d. For finite r-f signals, an effective negative resistance $(R_d)_e$, which is defined as the ratio of the fundamental components of the junction r-f voltage and current, is used.**

The power output of a tunnel-diode amplifier can be written

$$P_{out} = P_{in} + P_d - P_s - P_c, \tag{4}$$

where P_{in} is the power input to the amplifier, P_d is the power generated by the negative resistance of the diode junction, P_s is the power lost

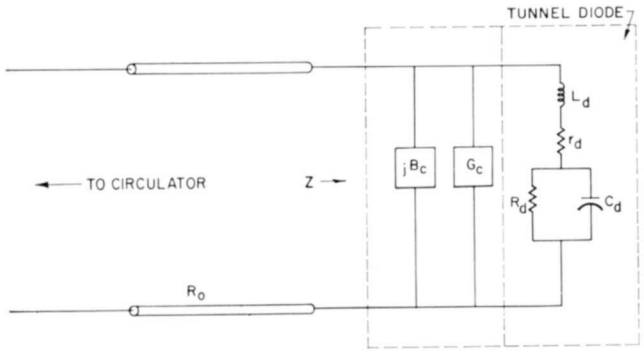


Fig. 3—Equivalent circuit of a tunnel-diode amplifier.

in the series resistance of the diode, and P_c is the power lost in the amplifier circuit. Now

$$P_d = \frac{V_0^2}{2(R_d)_e}, \tag{5}$$

where V_0 is the peak r-f voltage across the diode junction. Also

$$P_s = \frac{V_0^2}{2(R_d)_e^2} r_d \left[1 + \omega^2 C_d^2 (R_d)_e^2\right] \tag{6}$$

** Because R_d is nonlinear, harmonics of the input frequency are generated for finite r-f signals and the output of the amplifier is in general nonsinusoidal. Throughout this paper, the power gain and the impedance of the amplifier are, therefore, defined in terms of only the fundamental components of power, voltage, and current.

$$= \frac{V_0^2}{2(R_d)_e^2}\left[r_d + (R_{\min} - r_d)\left(\frac{f}{f_c}\frac{(R_d)_e}{R_{\min}}\right)^2 \right], \tag{7}$$

where $R_{\min}$ is the minimum value of $(R_d)_s$ and f_c is the resistive cutoff frequency of the diode and is given by

$$f_c = \frac{\sqrt{\dfrac{R_{\min}}{r_d} - 1}}{2\pi R_{\min} C_d}. \tag{8}$$

Finally,

$$P_c = \left| \frac{2Z}{Z + R_0} \right|^2 P_{in} R_0 G_c. \tag{9}$$

Thus

$$P_{out} = \frac{V_0^2 \left\{ 1 - \dfrac{1}{(R_d)_e}\left[r_d + (R_{\min} - r_d)\left(\dfrac{f}{f_c}\dfrac{(R_d)_e}{R_{\min}}\right)^2 \right] \right\}}{2(R_d)_e \left[1 - \dfrac{1}{G} + \left| \dfrac{2Z}{Z + R_0} \right|^2 \dfrac{R_0 G_c}{G} \right]}. \tag{10}$$

To calculate the power output of a tunnel-diode amplifier from Equation (10), the dependence of $(R_d)_e$ on V_0 must be known. This dependence was calculated for a typical GaAs tunnel diode by the use of the following tenth-degree power series approximation of its I–V characteristics:

$$I_d = \sum_{n=0}^{10} a_n V_d^n. \tag{11}$$

Equation (11) is plotted in Figure 4 together with the measured I–V characteristic of a GaAs tunnel diode (normalized to the diode peak current, I_p). The figure shows that the power-series approximation is excellent.

If the r-f component in V_d is assumed to be purely sinusoidal,* then

$$V_d = V_B + V_0 \cos \omega t, \tag{12}$$

* In general, V_d contains harmonics of the input frequency. The effect of these harmonics is usually small and is neglected here.

and

$$I_d = \sum_{n=0}^{10} a_n (V_B + V_0 \cos \omega t)^n \tag{13}$$

$$= I_0 + I_1 \cos \omega t + I_2 \cos 2\omega t + \cdots,$$

so that

$$(R_d)_c = \frac{V_0}{I_1}. \tag{14}$$

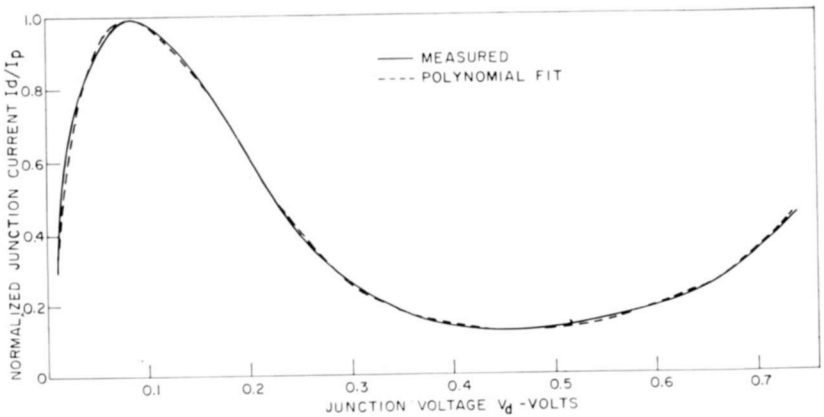


Fig. 4—Normalized I–V characteristic of GaAs tunnel diode.

Figure 5 is a plot of calculated values of $(R_d)_c/R_{\min}$** as a function of V_0 for a GaAs diode that has the I–V characteristics shown in Figure 4. Also shown in the figure are experimental points obtained from the measured gain-saturation curve of a microwave amplifier. Agreement between calculated and measured values is good.

The I–V characteristics of typical Ge and GaSb tunnel diodes are shown in Figure 6;

$$\text{for the Ge diode,} \quad R_{\min} \approx \frac{0.120}{I_p},$$
$$\text{for the GaSb diode,} \quad R_{\min} \approx \frac{0.060}{I_p}. \tag{15}$$

** Most practical tunnel-diode amplifiers are d-c biased at the minimum negative resistance point. For the GaAs diode of Figure 4, $R_{\min} \sim 0.22/I_p$. (MKS units are used throughout this paper.)

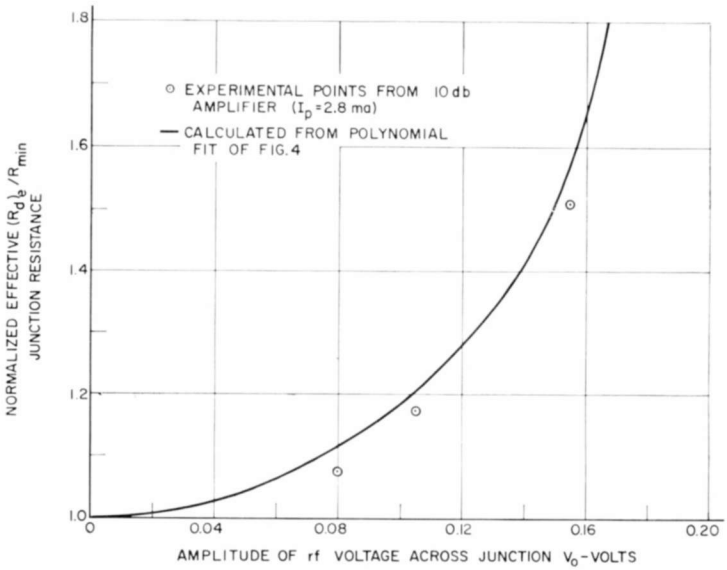


Fig. 5—Normalized effective junction resistance versus amplitude of r-f voltage across the junction for a GaAs tunnel diode having the characteristics shown in Figure 4.

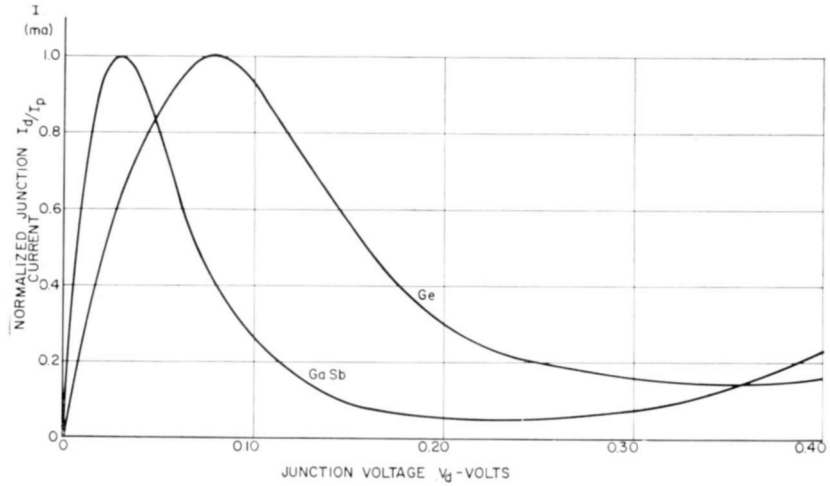


Fig. 6—Normalized *I–V* characteristics of GaSb and Ge tunnel diodes.

Plots of $(R_d)_e/R_{min}$ versus V_0 are given for these diodes in Figure 7. These curves, like the experimental points of Figure 5, were calculated from measured gain-saturation characteristics.

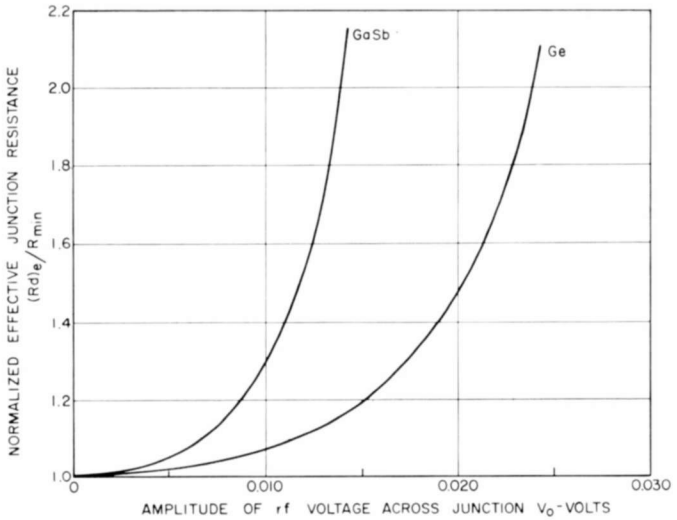


Fig. 7—Normalized effective junction resistance versus amplitude of r-f voltage across the junction for GaSb and Ge tunnel diodes.

Design of High-Power Tunnel-Diode Amplifiers

Choice of Diode Material

In general, a tunnel diode cannot be prevented from oscillating and therefore cannot be used in a stable amplifier unless[3]

$$L_d < 3(R_d)_s{}^2 C_d. \tag{16}$$

In practice, stabilization of a tunnel diode is extremely difficult if $L_d > R_{min}{}^2 C_d$, and hence the maximum allowable value of L_d is usually

$$(L_d)_{max} = R_{min}{}^2 C_d. \tag{17}$$

The voltage-gain–bandwidth product (for $\mathcal{G} \gg 1$) of a single-tuned

[3] L. I. Smilen and D. C. Youla, "Stability Criteria for Tunnel Diodes," *Proc. I.R.E.*, Vol. 49, p. 1206, July 1961.

circulator-coupled tunnel-diode amplifier is given by

$$G_v B \sim \frac{1}{\pi R_{\min} C_d}. \tag{18}$$

Substituting this relation into Equation (17) gives

$$(I_p)_{\max} \sim \frac{\alpha}{\pi G_v B L_d} \tag{19}$$

where $\alpha = I_p R_{\min}$. The relative values of α for GaAs, Ge, and GaSb diodes are

$$(\alpha)_{GaAs} : (\alpha)_{Ge} : (\alpha)_{GaSb} = 1 : 0.5 : 0.27. \tag{20}$$

Equations (19) and (20) show that for the same values of voltage-gain–bandwidth product and inductance, the maximum usable peak current is considerably greater for GaAs diodes than for either Ge or GaSb diodes.

The maximum power generated by the negative resistance of the diode is directly proportional to the maximum value of the diode peak current and the maximum value of the peak voltage across the diode junction; i.e.,

$$(P_d)_{\max} \propto (I_p)_{\max} (V_0)_{\max} \tag{21}$$

where $(V_0)_{\max}$ is determined by the maximum allowable large-signal gain depression of the amplifier, i.e., the maximum allowable value of $(R_d)_e/R_{\min}$. For example, from Figures 5 and 7, for $((R_d)_e/R_{\min})_{\max} = 1.5$,

$$(V_0)_{\max GaAs} = 0.1495, \quad (V_0)_{\max Ge} = 0.0206, \quad (V_0)_{\max GaSb} = 0.0115, \tag{22}$$

and from Equations (20), (21), and (22),

$$(P_d)_{\max GaAs} : (Pd)_{\max Ge} : (P_d)_{\max GaSb} = 1 : 0.069 : 0.021. \tag{23}$$

Thus, the power output of amplifiers using GaAs tunnel diodes can be many times larger than the power output of amplifiers using Ge or GaSb diodes.

Diode Housing

High-power tunnel-diode amplifiers must use diodes with high peak currents. Because $(I_p)_{max}$ is proportional to $1/L_d$ (see Equation (19)), it is important that the value of L_d be held to a minimum. For diodes housed in conventional ceramic "pill" packages, L_d ranges from about 100 to 600 picohenries. In our experiments the diodes used were mounted in a recently developed stripline package (see Figure 8). The

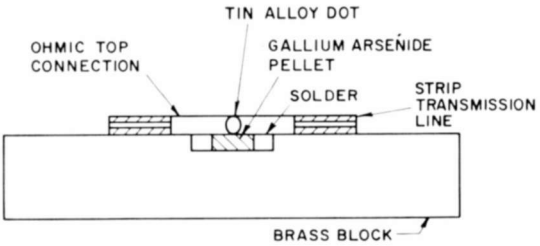


Fig. 8—Low-inductance tunnel-diode package.

inductance of these packages is estimated to be of the order of 50 picohenries, and they can be mounted into strip transmission-line circuits with a minimum of discontinuity between package and circuit.

Experimental Amplifiers

We have built experimental L-band tunnel-diode amplifiers using GaAs diodes having peak currents of the order of 20 milliamperes and voltage-gain–bandwidth products of the order of 3.0 gigacycles ($R_{min}^2 C_d \approx 800$ picohenries). The diodes were mounted in re-entrant strip transmission-line resonators of the type discussed in detail in Reference (4). Figure 9 shows typical curves of gain and power output of a GaAs tunnel-diode amplifier as a function of power input. The dynamic range of this amplifier is about 90 decibels. This range is more than two orders of magnitude greater than the range of conventional 1-milliampere Ge diode amplifiers having the same gain.

Power outputs of the magnitude illustrated in Figure 9 by no means represent the maximum values that can be achieved at micro-

---

* This development was carried out by RCA under U.S. Army Signal Corps sponsorship.

[4] F. Sterzer and D. E. Nelson, "Tunnel Diode Microwave Oscillators," *Proc. I.R.E.*, Vol. 49, p. 744, April 1961.

wave frequencies. Methods to increase the power output include paralleling of amplifier circuits, use of more than two diodes in a single amplifier circuit, use of high-current tunnel diodes with distributed junctions mounted in very low inductance packages, and cascaded amplifiers.

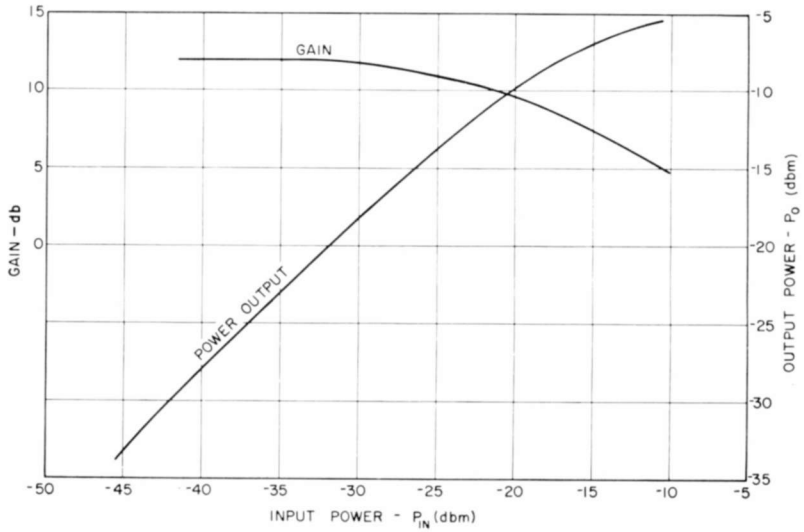


Fig. 9—Gain and power output versus power input for a 22-milliampere GaAs tunnel-diode amplifier.

Cascading of Amplifiers

For tunnel-diode amplifiers using similar diodes, it is generally true that the lower the gain of the amplifier, the higher the power output at which the amplifier starts saturating (see Equation (1) and Figures 5 and 7). Thus, if two similar amplifiers are cascaded, they will saturate at a higher power level than a single amplifier having the same gain as the cascaded amplifier. This fact is illustrated in Figure 10 where the power output of lossless single-stage and two-stage cascaded amplifiers are compared. The figure shows that the cascaded amplifiers have significantly greater saturated power output and dynamic range.

Cascading can also be used to improve the power-handling capabilities of low-noise tunnel-diode amplifiers. The minimum noise figures of tunnel-diode amplifiers (if negligible losses and high gain are assumed) are approximately as follows:

Diode Material	*Minimum Noise Figure (decibels)*
GaAs	4.9
Ge	3.8
GaSb	2.8

Thus, while GaAs diodes have the highest power-handling capability, they also have the highest noise figure. To combine the low-noise

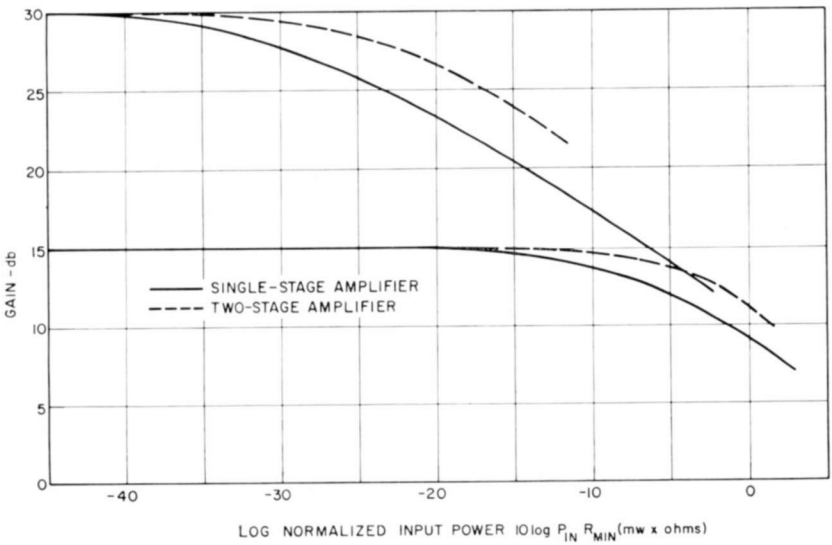


Fig. 10—Gain-saturation characteristics of lossless single-stage and two-stage GaAs amplifiers. Each stage in the cascaded amplifiers has the same gain.

properties of Ge and GaSb diodes with the high power capability of GaAs diodes, a low-noise Ge or GaSb amplifier can be cascaded with a GaAs power amplifier. The calculated saturation characteristics of single-stage Ge and GaAs and cascaded (first stage Ge, second stage GaAs) amplifiers are compared in Figure 11. The calculations assume that the Ge and GaAs diodes used in the cascaded amplifiers have the same minimum negative resistance (i.e., the peak current of the GaAs diodes is twice the peak current of the Ge diodes). The last assumption was made (and the power input of Figure 11 normalized with respect to R_{min}) to make possible meaningful comparisons of the various amplifiers, since the minimum value of R_{min} is independent of

diode material and is only a function of the gain–bandwidth product and the series inductance of the diode (see Equations (15) and (16)).

Figure 11 shows that while the noise figure of the cascaded Ge–GaAs amplifiers is only slightly higher than the noise figure of a single-stage Ge amplifier, the cascaded amplifiers have significantly

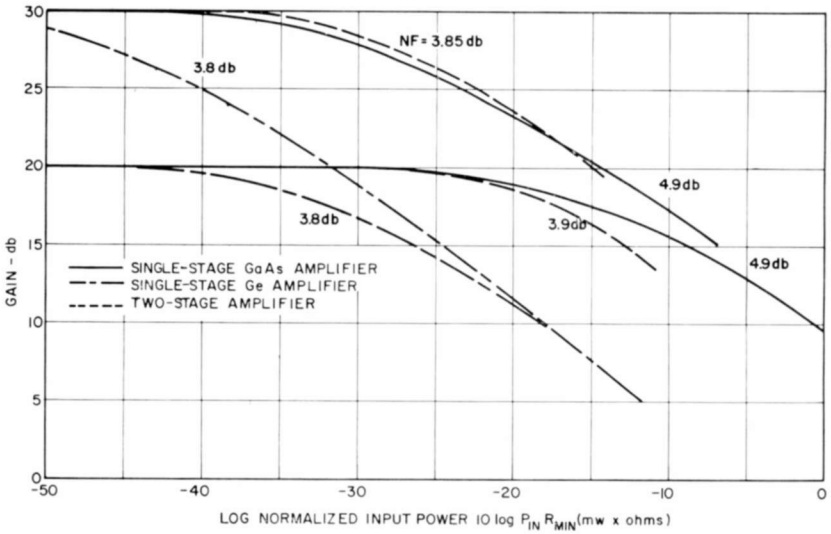


Fig. 11—Gain-saturation characteristics of lossless single-stage Ge and GaAs amplifiers, and two-stage Ge–GaAs amplifiers. Each stage in the cascaded amplifiers has the same gain.

better saturation characteristics and a much greater dynamic range. For example, the dynamic range of the 20-decibel cascaded amplifier is 15 decibels greater than that of the single-stage 20-decibel Ge amplifier.

Comparison between cascaded Ge–GaAs and single-stage GaAs amplifiers shows that the cascaded amplifiers have significantly lower noise figures. The cascaded 30-decibel amplifier has a higher saturated power output than the single-stage amplifier, while for the 20-decibel amplifier the situation is reversed, i.e., the saturated power output of the single-stage GaAs amplifier is greater than that of the cascaded amplifier.

Acknowledgment

The authors wish to thank E. T. Casterline and R. D. Gold for supplying the GaAs tunnel diodes, H. C. Johnson and T. E. Walsh for performing some of the calculations, and D. E. Nelson and A. Presser for valuable discussions.

TECHNIQUES FOR DIGITAL COMMUNICATION VIA SATELLITES

By

F. Assadourian and E. M. Bradburd

RCA Communications Systems Division
New York, N. Y.

Summary—*Synchronization techniques are treated for digital transmission over subsynchronous and synchronous satellite links which form parts of general communication networks interconnecting several nodes that perform multiplexing functions. The primary emphasis is on bit synchronization and identification on a link basis. A bit-transport equation is developed to relate numbers of transmitted and received pulses, taking into account variable path delays. When differences between clocks at the ends of a link are included, the result is useful in determining storage requirements for time buffering that must be inserted to maintain bit synchronization. For subsynchronous satellites, handover techniques are discussed for the preservation of bit integrity during switching from one satellite to the next.*

Introduction

TECHNIQUES are discussed for handling some of the synchronization problems that arise in digital communication via subsynchronous and synchronous satellite repeaters. The complexity of these problems depends upon the kind of communication network in which the satellite links are employed. Synchronization considerations are usually simplest for networks of a single link and most difficult for networks of several links connecting a number of nodes.

The general case to be treated is shown in Figure 1 with the solid lines radiating from the two nodes representing long-haul transmission paths and the dotted lines representing connections to local subscribers. Each switching node is assumed to demultiplex incoming digital pulse streams from various links and switch some of the outgoing data to the satellite link in new multiplexed arrangements. Synchronization disturbances over the satellite link arise from two basic sources—first, instabilities and relative inaccuracies of terminal clocks used for data timing; and second, path-delay variations due to satellite motion. The need for handover (switching from satellite to satellite) imposes additional constraints in subsynchronous satellite data links.

The satisfactory operation of a data link in the above general

network involves both (1) bit synchronization (and identification) and (2) frame synchronization. The first refers to synchronism between pulses at the sending and receiving ends of a link, and the second refers to the detection of the transmitted framing pattern at the receiving end. Both are needed for proper demultiplex/multiplex functioning at a node that differs in local clock rates from other nodes

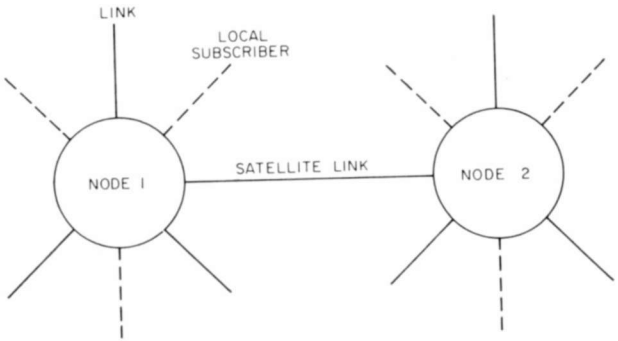


Fig. 1—Portion of communication network.

and that has to remultiplex portions of incoming pulse streams arriving at varying rates.

Bit synchronization can be achieved, for example, by slaving bit timing associated with an input pulse stream to a local clock through buffering techniques. Methods of obtaining bit timing may be found in References (1) and (2). An analysis of a particular buffering technique is discussed in the present paper.

The recognition of the transmitted framing pattern at the receiver can be accomplished through a parallel search, a serial search, a combination of the two, or a suitable examination of the spectrum of the incoming pulse stream. A framing pattern may, in particular cases, consist of periodically spaced marks (or alternate marks and spaces) for purposes of counting off multiplexed channel positions. Analyses useful in forming quantitative estimates of the times required to

[1] E. M. Bradburd and F. Assadourian, "Digital Transmission in Media of Variable Time Delay," *7th MIL-E-CON Conference Proceedings*, Sept. 1963.

[2] O. E. DeLange, "The Timing of High-Speed Regenerative Repeaters," *Bell Syst. Tech. Jour.*, p. 1455, Vol. 37, Nov. 1958.

identify framing patterns by one or more of the above methods can be found in References (1), (3), and (4).

Handover problems and techniques are discussed in some detail, particularly the instantaneous type needed to preserve bit integrity during handover. Brief tables of useful numerical parameters for synchronous and subsynchronous satellite systems are provided at the end of this paper and applied to illustrative examples.

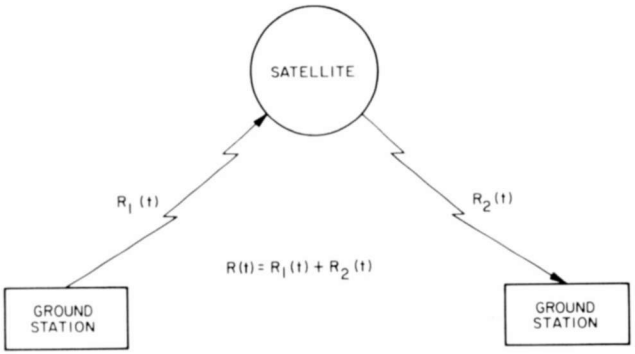


Fig. 2—Satellite range geometry.

Bit-Transport Equation for Satellite Link

A bit-transport equation is developed for a satellite repeater link to serve as the basis of synchronization considerations with the aid of Figures 2 and 3. First let the sending end of the link transmit $N_T(t_0, t_0 + t)$ pulses of constant width Δ during a time t at the rate of $n_T = \dot{N} = 1/\Delta$ pulses per second (pps), where the dot denotes time derivative. For a total path delay (up and down) of $\tau = R/c$, where R is the total path length and c is the velocity of light, the pulses will be received during the interval from $t_0 + \tau(t_0)$ to $t_0 + t + \tau(t + t_0)$. If N_R represents the number of received pulses, then

[3] J. Dutka and A. A. Meyerhoff, "Synchronization of Pulse Trains," *RCA Review*, p. 410, Vol. 22, Sept. 1961.

[4] M. Masonson, "Power Spectra in Digital Transmissions," Appendix 5E, Progress Report, VII and VIII Quarters, Vol. II, Jan. 1 to June 30, 1961, UNICOM, BTL.

$$N_R\,[t_0 + \tau(t_0),\ t_0 + t + \tau(t_0 + t)\,] = N_T(t_0, t_0 + t). \qquad (1)$$

Since the moving satellite repeater produces time-varying path delays, the received pulses appear to be compressed or expanded into the interval of length $t + \tau(t_0 + t) - \tau(t_0)$. If this length is changed to t, then it can be shown with the aid of Figure 3 that the bit-transport equation becomes approximately

$$N_R\,[t_0 + \tau(t_0),\ t_0 + \tau(t_0) + t] \approx N_T\,[t_0, t_0 + t] + N_T\,[t_0 + t,\ t_0 + t + \tau(t_0) - \tau(t_0 + t)]. \qquad (2)$$

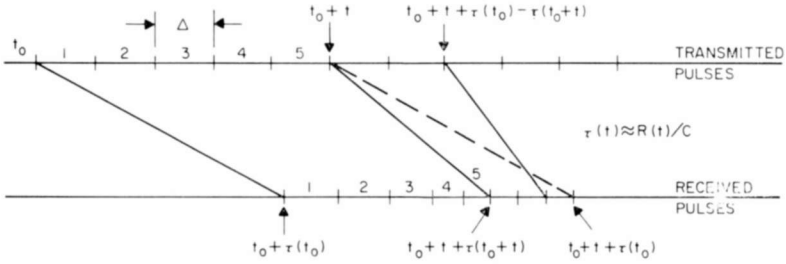


Fig. 3—Bit transport representation.

It is readily seen that

$$N_R(t) \approx \frac{t - \tau_D}{\Delta}, \quad \tau_D = \tau(t_0 + t) - \tau(t_0), \qquad (3)$$

and

$$n_R(t) = \dot{N}_R(t) \approx \frac{1}{\Delta}(1 - \dot{\tau}) = \frac{1}{\Delta}(1 + u), \qquad (4)$$

where the fractional doppler shift, u, arises from

$$\dot{\tau} = \dot{R}(t)/c \approx -(\Delta f)/f = -u.$$

Since Equation (4) assumes that the same frequency is used for both directions, the equation must be modified if such is not the case. According to Equation (3), the number of received pulses depends upon the difference in delay for the first and last pulses. Also, a stable transmitter clock and a slowly varying doppler shift have been assumed.

Transmitter clock instability changes n_T to

$$n_T(t) \approx \frac{1}{\Delta}[1 + \epsilon_T(t)], \quad (\epsilon_T)_{\text{rms}} = \epsilon, \tag{5}$$

where the average $\bar{\epsilon}_T$ over the period of one day, for example, may be zero. The received pulse rate is now

$$n_R(t) \approx \frac{1}{\Delta}[1 + \epsilon_T(t) + u]. \tag{6}$$

No local-oscillator instabilities of the satellite repeater appear in Equation (6) because the repeater shifts the entire input r-f spectrum as a unit without producing differential effects among spectral components.

For a high percentage of time, n_R is bounded by

$$\frac{1}{\Delta}[1 - 3\epsilon - U] < n_R(t) < \frac{1}{\Delta}[1 + 3\epsilon + U], \; U = |u|_{\max}. \tag{7}$$

In a sample case, if ϵ is 10^{-7} and U is 10^{-5}, then the maximum change in n_R is around 2×10^{-5}, which is insignificant in synchronization schemes that are required only to derive timing information from the received pulse stream, as in the case of a network of one link or a few cascaded links.

Bit Synchronization

In applications that demand a tight bit synchronism between sending and receiving ends of a link (such as in Figure 1 if the links are encrypted), a buffer can be inserted at the receiving end to absorb the effects of path-delay variations and differences between timing clocks at the link terminals. In this manner, pulses can be sent at one rate and read out correctly after reception at the slightly different rate provided by a local clock. The result is useful in multiplex operations in that all incoming pulse streams are read out at a node in the right sequence with a single local clock. Also, in the case of link encryption, the sequence of pulses generated at encryption can be decrypted after reception by maintaining the correct phasing between key generators.

The time buffers, which are called stores here, may be shift registers, delay lines, magnetic tape loops, etc. Received pulses arriving

at rate n_R are clocked into the store with timing information derived from these pulses. When the store of capacity C bits is, say, half-full, it is read out with timing obtained from a local clock. Read-in and read-out have an initial separation of $C/2$ bits which may change in time. If the separation either becomes zero or exceeds C bits, then timing errors are produced. For subsynchronous satellite links, magnetic-tape loops are particularly useful in conjunction with feedback loops to slave read-in to read-out rates.

The buffer approach is now analyzed for a satellite link. First, for the incoming pulses, Equation (2) is rewritten

$$N_R\,[t_0+\tau(t_0),\ t_0+\tau(t_0)+t] \approx \int_{t_1}^{t_1+t} \frac{1+\epsilon_T(t)}{\Delta}\,dt - \frac{\tau_D}{\Delta}. \quad (8)$$

Since store read-out starts some time after read-in, the number of pulses read out (N_0) during a time t at the rate of the local clock is

$$N_0(t_1, t_1+t) \approx \int_{t_1}^{t_1+t} \frac{1+\delta+\epsilon_R(t)}{\Delta}\,dt, \quad (9)$$

where δ is the normalized relative inaccuracy between the sending and local clocks, and ϵ_R is the instability of the local clock.

Assume now that ϵ_R has the same average (over a day) and r-m-s values as ϵ_T. Then, for a large storage capacity C, the accumulated difference over the interval t between the numbers of pulses read in and out of the store is bounded, for a high percentage of the time, by

$$S = \frac{C}{2} = |N_0 - N_R|_{\max} = \frac{t}{\Delta}\,(|\delta| + 6\epsilon) + \frac{|\tau_D|}{\Delta}. \quad (10)$$

Here S is the maximum slippage in bits between read-in and read-out.

A few examples will illustrate the implications of Equation (10).

With no buffering, C/2 is replaced by 1/2 for maximum slippage of 1/2 bit, and $\tau_D = 0$. Loss of bit synchronization can occur after a time $t = \Delta/[2(|\delta| + 6\epsilon)]$ due to relative clock differences. For example, if $|\delta| = \epsilon$ and t is to be 24 hours, then

$$|\delta| = \epsilon \approx 0.8 \times 10^{-6}\Delta,$$

where Δ is in seconds. For 1000 pps data, both clock stability and relative clock accuracy must be better than 10^{-9}; for 100,000 pps data, the value becomes 10^{-11}.

With a buffer store of 200 pulses, the clock requirements are reduced to 10^{-7} for 1000 pps data and 10^{-9} for 100,000 pps data.

The maximum delay variation for a satellite at 6,000 miles altitude may be approximately 30 milliseconds in one hour. With no buffering and only time-delay variation, the maximum allowable data rate is only 16 pps for a maximum slippage of 1/2 bit.

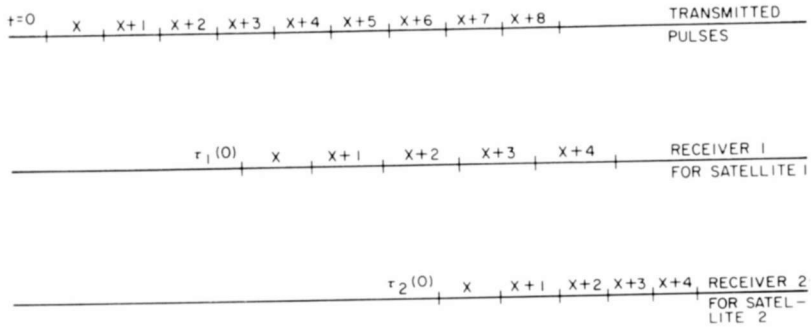


Fig. 4—Pulse representations for satellite handover.

For clock differences less than about 10^{-8}, these differences can be disregarded, and the previous delay variations dominate. To avoid store runout or overflow, C must be about $6 \times 10^{-2}/\Delta$. For 1000 pps data, C is 60 pulses, and for 100,000 pps data, C increases to 6000. It is evident that shift-register stores are impractical for these figures. However, magnetic-tape loops can be used.

Satellite Digital Handover Techniques

Digital transmission via subsynchronous satellites involves handover of the communications link from satellite to satellite at various times. Available techniques depend upon the number of antennas per ground site, the tolerable complexity of handover circuitry, and bit-integrity requirements.

With two antennas (each complete with transmitter and receiver) per ground site, either fast or instantaneous handover becomes possible. While one antenna at each end of a link is tracking one satellite during communications, the remaining pair of antennas can acquire and track the next satellite. Then, for a period of time there will be 2 received data streams, as shown in Figure 4. At handover, there

can be an abrupt discontinuity which, if uncorrected, takes the form of a sudden apparent gain or loss of pulses. For example, in a system of satellites at 6000 miles altitude, the maximum discontinuity between any 2 satellites in usable orbital positions is about 30 milliseconds.

Instantaneous Handover

With special circuitry, instantaneous handover can maintain continuity of information flow (i.e., bit integrity). Two techniques are

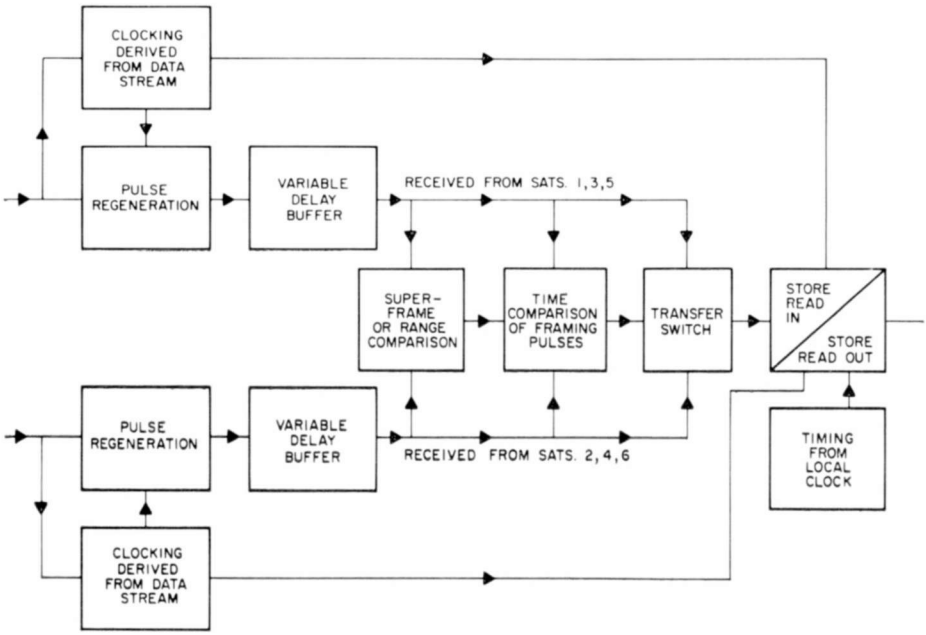


Fig. 5—Instantaneous handover circuit with stores.

illustrated in Figure 5. Both can be applied in handover either at unequal or at equal satellite path lengths. Figure 6 illustrates these situations. The end store in Figure 5 is used when tight synchronism is needed between the sending and receiving ends of the satellite link. Although Figure 5 will be used illustratively to develop concepts, it may be replaced in actual practice by Figure 7, as will be explained later.

In both techniques illustrated in Figure 5 it is assumed that the data stream received from each satellite is separately examined to locate the positions of framing pulses before the two data streams are aligned for handover from one to the next. Consequently, the time

required for this step sets a minimum time after the beginning of the overlap in time of the two path delay curves and after which handover becomes possible. It is assumed that handover is performed between framing pulses. In typical cases, it can be shown that the time required to locate the framing pulse position in a data stream is at most a few seconds (see Reference (1)).

Equal Path Lengths

For handover at equal satellite path lengths, curves 1 and 2 of Figure 6 are applicable, and the variable-delay buffer sections of

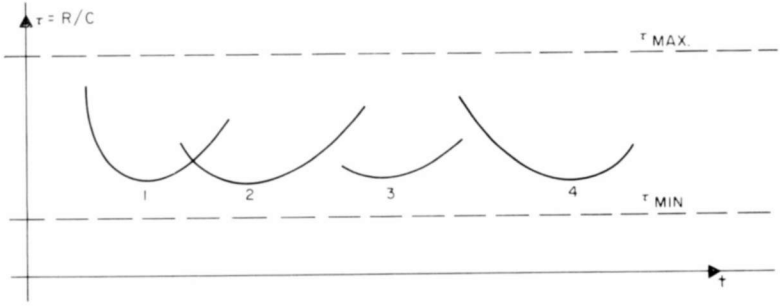


Fig. 6—Satellite path delay curves.

Figure 5 are not needed. The implication in Figure 4 is that the received data streams tend to slide past each other and achieve time coincidence of corresponding pulses during some interval. After this interval is recognized, instantaneous handover can be accomplished anywhere within it.

Both techniques in Figure 5 use coarse and fine alignments of data streams. One approach achieves coarse alignment with the aid of periodic super-framing pulses inserted in a data stream as markers. These must be distinguishable from regular framing or other periodic pulses either in amplitude or coding and can occur anywhere. They should be locatable within a given data stream in a few seconds.

Super-framing pulses should be spaced by more than twice the maximum delay difference in the two satellite paths. When the first time-comparison circuit in Figure 5 measures a spacing of less than the maximum delay difference between any 2 super-framing pulses, one taken from each data stream, then corresponding super-frames have been identified. Next, when these super-framing pulses are within a frame length of each other, corresponding framing pulses

within the super-frames are identifiable, and the second time-comparison circuit of Figure 5 or other correlation type circuit can be actuated for fine alignment. Finally, when corresponding framing pulses overlap in this circuit, instantaneous handover may be performed between framing pulses.

An estimate is now made of the length of the time interval during which instantaneous-handover switching must be executed. The maximum rate at which the two received data streams slide past each other is given by

$$m_{MAX} \approx \frac{2U}{\Delta} \text{ pps}, \quad U = \left| \frac{\Delta f}{f} \right|_{MAX}. \tag{11}$$

If T_{MIN} is the minimum time of overlap of the two data streams to within the fraction $\pm k$ of a pulse width, then

$$T_{MIN} = \frac{k\Delta}{U} = \frac{2k}{m_{MAX}}. \tag{12}$$

For example, if the maximum doppler shift is 10^{-5}, the transmitted pulse width Δ is 25 microseconds and $k = 0.2$, then T_{MIN} is 1/2 second. If the frame length is 65 pulses, then this interval contains around 300 framing pulses in each data stream. Any coincident pair chosen from these can be used for switching between the data streams.

In some applications the use of super-framing pulses for coarse alignment may be undesirable. Another approach, as shown in Figure 5, is to make a range comparison of the two satellite paths to determine when corresponding frames in the two data streams become spaced by a half-frame or less. After this point, the procedure can follow the rest of the previous super-framing approach.

An analysis of the range comparison approach is given. If there are F pulses per frame, then the frame rate is $1/F\Delta$, and a half-frame occupies $F\Delta/2$ seconds. Corresponding framing pulses in the two data streams become spaced by less than a half frame when the path-length difference for the two satellites reduces to less than R_D, with

$$R_D = 9.3 \times 10^4 F\Delta \text{ miles}. \tag{13}$$

The path lengths are measured with known accuracy, and fine alignment should not be initiated until their measured difference reaches r_D as given by

$$r_D + 2\epsilon_R < R_D, \quad r_D - 2\epsilon_R > 0. \tag{14}$$

Here $\pm\epsilon_R$ is the accuracy of either path-length measurement. The setting of r_D is needed to guarantee that the actual path difference is less than R_D.

Path-length difference information need not be supplied until the two data streams are, say, one frame apart. Since the maximum possible closing rate of the two digital streams is approximately $2U$ seconds per second, the minimum time for closure by a half-frame is $F\Delta/(4U)$ seconds. Path-length difference data supplied several times during this closure would seem to be adequate for determining when fine alignment should start.

To illustrate the range-comparison approach, let $\Delta = 25$ microseconds, $F = 65$ and $U = 10^{-5}$. According to Equation (13), the two data streams are separated by a half frame of 0.812 millisecond when $R_D = 151$ miles. If the range accuracy is 30 miles, then a range setting r_D of 91 miles implies, by Equation (14), a true range difference lying between 31 and 151 miles. The lower bound represents a minimum spacing between corresponding frames in the two data streams of about 166 microseconds, leaving adequate time for fine alignment. Furthermore, if range data is desired during closure of the data streams from one-frame to half-frame separation, which takes around 40 seconds, then a range-difference reading every few seconds during this interval should suffice.

In summary, the range-difference technique for coarse alignment does not require the insertion of special marker pulses as in the superframing approach, but has the disadvantages that it requires information about complete satellite path lengths and imposes limits on data rates.

In some applications it may be inconvenient to perform instantaneous handover at equal ranges because of satellite assignment difficulties. In other cases, as illustrated in Figure 6 between curves 2 and 3 and curves 3 and 4, the condition of equal ranges may never be reached (for example, in a system of satellites with randomly inclined orbits). The dashed horizontal lines represent path-delay (hence range) bounds fixed by zenith and horizon conditions. The curves terminate at points beyond which particular satellites are out of view of either end of the satellite link.

Unequal Path Lengths

Handover at unequal ranges can be accomplished by setting the variable delays in Figure 5 at appropriate positions. They are time

buffers and, as described previously, can be shift registers, delay lines, magnetic tape loops, etc. For example, if a shift register store is used, a desired delay can be inserted by reading bits into the store until it contains the right number of bits before starting read out. If a magnetic-tape loop arrangement is used to combine the variable delay and the end store and to slave the read-in rate to the read-out rate provided by a local clock with a feedback loop, Figure 5 may be

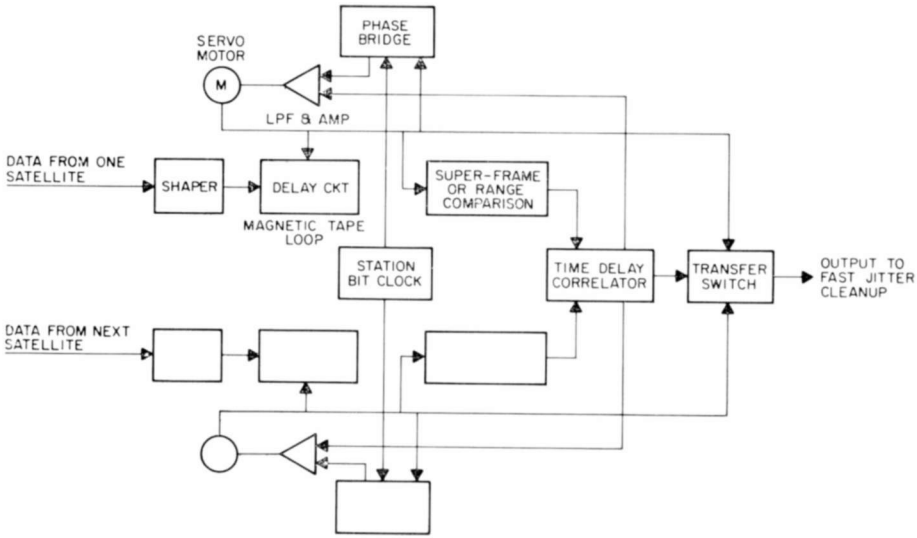


Fig. 7—Instantaneous handover with magnetic tape loops.

replaced by Figure 7. The fast-jitter cleanup at the end of the latter is essentially a small-capacity store to remove the effects of flutter, etc.

The purpose of either variable delay in Figure 5 is to introduce a delay in the shorter satellite path to enable handover at a time when the two path lengths are unequal. This process can be applied to cases represented by any consecutive pair of curves in Figure 6, but is considered here only for the last 3 curves, which do not intersect. They are grouped into the extreme cases of Figures 8 and 9.

Figure 8 shows a case in which the satellite in use before handover (the setting satellite) yields a shorter path than the rising satellite for a number of intervals of required service. To start the operation (see Figure 5) the undelayed data stream is read into the end store. Then variable delay 1 is set at $\Delta_1\tau$, withholding pulses from entry into the end store. This quantity must be calculated from the desired

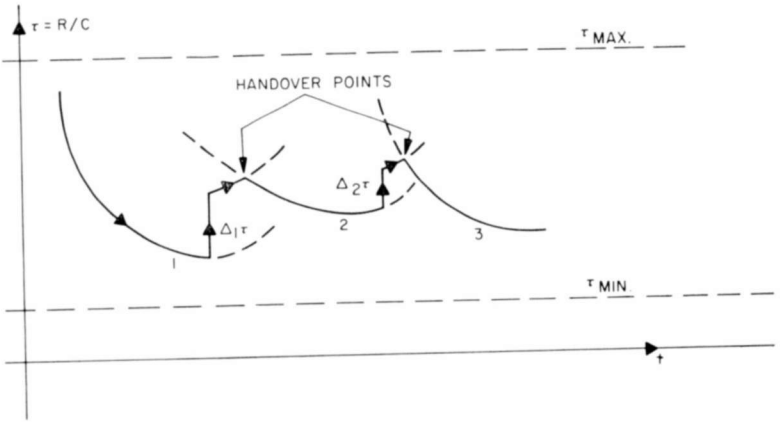


Fig. 8—Satellite path delay curves.

location and accuracy of the handover point for synchronization of the two data streams. At handover, read-in for the end store is derived from the pulse in data stream 2 which follows the last one used in stream 1, where both pulses are preferably framing pulses. The pulses in variable delay 1 are then dumped and the bias is removed. While the end store receives from stream 2, the gap between read in and read out resulting from the previous delay introduced in stream 1 tends to be reduced until a new delay is added for stream 2.

The capacity of the end store should be adequate to handle both the maximum difference, $\tau_{MAX}-\tau_{MIN}$, and differences in setting and

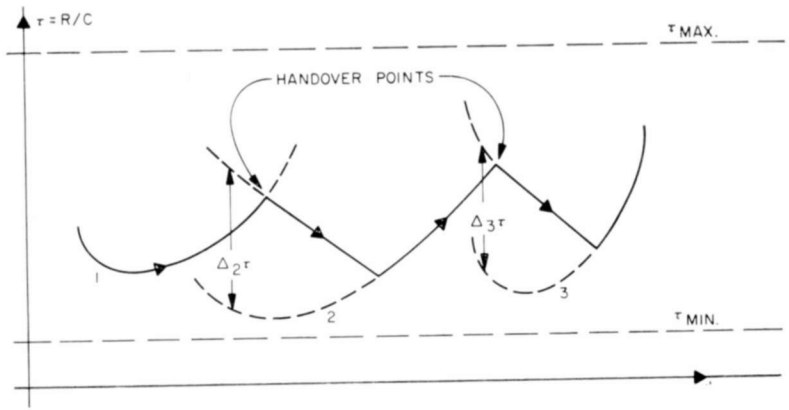


Fig. 9—Satellite path delay curves.

stability of the local clocks at the ends of the satellite link. The reason for considering the indicated delay difference is that, in accordance with the discussion following Equation (10), excursions of the composite arrowed curve in Figure 8 always remain within the prescribed bounds. The maximum required setting for either variable delay in Figure 5 can be selected as the above maximum delay difference, but will generally be considerably less.

Figure 9 shows another extreme case. The operation in Figure 5 again starts with the read-in of data stream 1 into the end store. Then, at some particular time, a bias delay of $\Delta_2\tau$ is injected in the path for stream 2. After handover, the end store receives stream 2, which tends to slow down and create a gap in the end store if no action is taken. After handover from stream 2 to a delayed stream 3, the bias delay and pulses stored in the variable delay can be removed for stream 2. However, the gap it produced in the end store remains. Furthermore, with enough consecutive satellite passes of the present type, the gap in the end store tends to increase until there is store runout unless it has a larger capacity than needed in the previous case of Figure 8.

Since a large number of Figure 9 curves are not likely to occur consecutively in practice, the present problem has perhaps been exaggerated. Also, with enough satellites, it can be avoided by proper satellite assignments. If the problem proves to be sufficiently severe, however, there are at least two available courses of action. First, as indicated by the composite arrowed curve of Figure 9, the read-in of data stream 2 (and, later, stream 3) into the end store can be speeded up. The capacity of the end store and the maximum variable delay settings can then be given the same values as for Figure 8.

Second, the problem can be eliminated by using the magnetic-tape-loop approach of Figure 7. The present system can then function as shown in Figure 10. Super-framing (or other marker) and framing pulses are used to measure actual arrival times of marker points in each data stream. The maximum and minimum delay of possible paths is known from geometrical considerations of satellite heights and distances between terminals. Hence the system buffer can be designed so that a variable delay is added to the path delay to yield an approximately constant total system delay. This total system delay can be measured by comparing the arriving data markers after buffering with those generated in a highly accurate local clock. The buffer thus compensates as well for clock run-out between the two ground stations.

As shown, the total *system* delay is designed to exceed the maximum path delay by some margin large enough to insure at least one

full day of operation without readjustment of the buffer stores due to clock run-out. The two tape buffers shown in Figure 7 thus equalize the system delay for each satellite path so that handover can be made at any convenient time in the overlap of the periods of mutual visibility of the rising and setting satellites.

As shown in Figure 8, this overlap interval must be of sufficient duration so that the rising satellite buffer circuits can lock on to bit, framing, and super-framing timing in the received bit stream. There must also be sufficient time to achieve proper servo tracking of the receiving buffer.

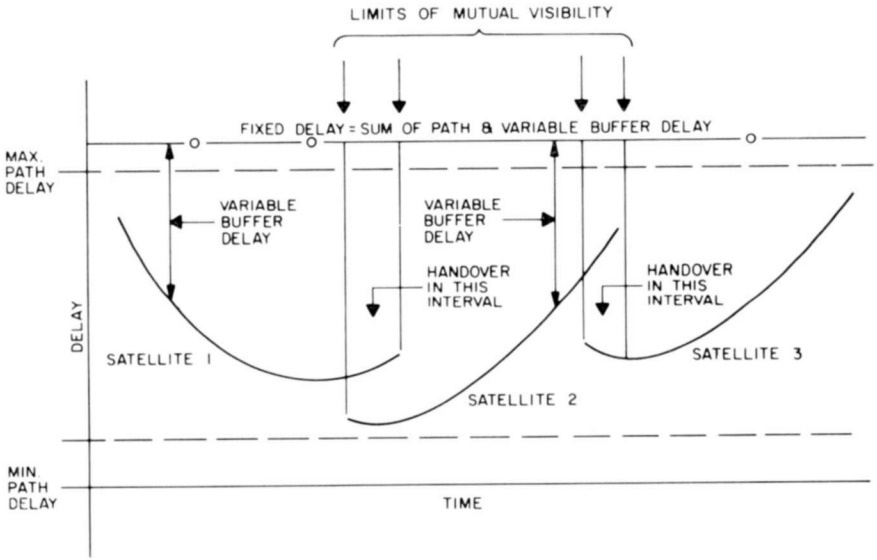


Fig. 10—Variable delay compensation and handover principle.

The following estimates are based on Reference (1). Bit timing can be derived from the arriving stream in approximately 25 bit intervals. Framing can be derived in around 3 seconds for a megacycle bit rate and 64 bits per frame. The servo tracking lock up should be possible in a comparable time. Super-framing can be derived as soon as it arrives if it is unique with respect to the data stream. The system will thus require at least six seconds of mutual visibility of rising and setting satellites before handover can be made.

Notice that in this method of use of buffer stores, there is no restriction for the equality of path delay at time of transfer, or for the rising or setting paths to be consistently of shorter or longer relative delay.

Numerical Illustrations

Background Tables

In Table I, the time interval, t, during which the maximum delay variation, $|\tau_D|$, occurs for a satellite path between two stations in close proximity has been estimated from the approximation

$$t \approx \frac{\alpha}{\omega_S}, \quad \cos \alpha = \frac{R_e}{R_e + H}, \quad \omega_S = \frac{2\pi}{T} \tag{15}$$

Table I—Subsynchronous Satellites

H (statute miles)	T (hours)	$\lvert\tau_D\rvert$ (milliseconds)	t (hours)	U (parts in 10^{-5})
2000	2.6	24	0.35	2.8
3000	3.3	30	0.52	2.2
6000	5.6	31	1.04	1.3
9000	8.5	34	1.69	0.89
12000	11.3	35	2.41	0.31

where R_e = earth radius,

α = angle subtended at the earth center by the satellite positions at zenith and at the horizon with respect to the two ground stations,

H = altitude of circular satellite orbit, and

T = orbital period.

The above approximation is best for inclined (close to polar) satellites and poorest for equatorial satellites, which require the insertion of the earth angular velocity. The formula must obviously be refined as needed in any detailed applications.

The values for U for equal up and down frequencies and for the two ground stations with satellite tracking down to the horizon (0°) have been obtained from the approximation

$$U \approx \frac{2R_e\omega_S}{c}, \tag{16}$$

which is also most accurate for orbital inclinations close to 90°.

In Table II the two ground sites are assumed to be at the equator

Table II—Synchronous Satellites

Orbit Inclination, i, (degrees)	$\lvert\tau_D\rvert$ (milliseconds)	t (hours)
10	1.1	6
20	3.2	6
30	7.5	6

under the crossover of the satellite figure eight pattern to obtain extreme values.

In Table III a "stationary" satellite is assumed (0° inclination) with a daily periodic variation in altitude between $H - \Delta H$ and $H + \Delta H$. Extreme values are shown for two ground sites in close proximity under the satellite.

Bit Synchronization for Synchronous Satellites

If a maximum slippage between transmission and reception of 1/2 pulse is permitted in 24 hours, and if a receiver store with a capacity of C pulses is employed, then Equation (10) shows that the data rate is limited by

$$n_T < \frac{C}{2} \frac{1}{8.64 \times 10^4(\lvert\delta\rvert + 6\epsilon) + \lvert\tau_D\rvert} \quad \text{pps.} \quad (17)$$

Two conclusions may be drawn. First, there is no point in making the first term of the denominator less than, say, a tenth or so of the second term, i.e., $\lvert\delta\rvert + 6\epsilon < 10^{-6}\,\lvert\tau_D\rvert$. In this case, Equation (17) reduces to $C > 2n_T\lvert\tau_D\rvert$. Second, the choice of C is now made to depend entirely upon n_T and τ_D.

Table III—Synchronous Satellites (With Periodic Variation in Altitude)

ΔH (statute miles)	$\lvert\tau_D\rvert = \frac{4\Delta H}{c}$ (milliseconds)	t (hours)
20	0.43	6
50	1.07	6
100	2.1	6
200	4.3	6
500	10.7	6

For example, if $|\tau_D|$ is selected from Tables II or III, it lies between 0.5 and 10 milliseconds. The first inequality becomes $|\delta| + 6\epsilon < 0.5 \times 10^{-9}$ to 10^{-8}, which will be satisfied if both stability and relative clock inaccuracy figures are better than around 0.7×10^{-10} to 1.4×10^{-9}. The second inequality yields $C > 10^{-3} n_T$ to $2 \times 10^{-2} n_T$.

If the store capacity is chosen in observance of the above conditions, then, since $|\delta|$ and ϵ have been made very small, there should be no loss of bit integrity due to store runout during periods of several days. On the other hand, if C is not sufficient to cope with τ_D, then it is possible for bit synchronization to be lost during a six-hour interval. Although these results for synchronous satellites may seem surprising at first, it must be remembered that they depend upon the assumption of delay variations of 1/2 millisecond or more in a 6-hour interval, as compared to fractions of a microsecond for microwave or tropospheric scatter transmission.

Bit Synchronization for Subsynchronous Satellites

With the same assumptions for bit synchronization as before, Equation (17) is applicable. Since Table I shows that τ_D does not vary much over the selected range of satellite altitudes, a figure of 30 milliseconds is assumed for illustrative purposes. One now obtains $|\delta| + 6\epsilon < 3 \times 10^{-8}$, which will be satisfied if both $|\delta|$ and ϵ are less than around 5×10^{-9}. The required store capacity becomes $C > 6 \times 10^{-2} n_T$, which implies that C should vary from 60 to 3000 to accommodate data rates going from 1000 pps to 50,000 pps. If C is not large enough to handle τ_D, then bit synchronization may be lost during a time interval of duration determined by the satellite altitude (½ to 2½ hours for altitudes between 2000 and 12,000 miles).

ELLIPSOMETRY—A VALUABLE TOOL IN SURFACE RESEARCH

By

K. H. Zaininger and A. G. Revesz

RCA Laboratories
Princeton, N. J.

Summary—*Ellipsometry is a technique that allows the determination of the optical properties of a surface, or the optical properties and thickness of a thin film, by measuring the effect of reflection on the state of polarization of polarized light. In this paper, the fundamental equation governing ellipsometry is developed starting from the problem of reflection and refraction of light at a boundary between two homogeneous, isotropic media, and reflection from a film-covered surface. A pictorial representation and the classical mathematical specification of polarized light is given. Various solutions of the ellipsometry equation are discussed, the actual ellipsometer is described, and experimental techniques are outlined. Areas of applications are summarized and the value of ellipsometry is examined in terms of possible errors and obtainable accuracy. Finally, some of the deviations in the optical properties of thin films from those of the bulk are briefly outlined.*

Introduction

Interest in the physical properties of thin films has rapidly increased within the last two decades. Much of this increased interest was brought about by advances in high-vacuum techniques. Initially, investigations were concentrated on evaporated thin films used mainly for various optical purposes. In the last few years, however, interest has also developed in insulating, semiconducting, and metallic thin films that are used in both active and passive electronic devices, especially in connection with integrated electronics.

Various methods are available for the determination of the properties of thin films as for example weighing, electrical measurements, and electron microscopy. In addition, there are a large variety of optical methods that in many cases are preferable to nonoptical methods. Optical methods have the advantage that they permit the investigation of surfaces.

For the study of surfaces and films on substrates, three basically different optical methods of investigation are available.

(1) In photometric measurements the amplitudes of incident (generally normal incidence) and reflected or transmitted rays are measured. The main areas of application are determination of optical

constants and absorption peaks to obtain information concerning composition and structure of materials.

(2) In interference measurements, the phases of two rays reflected by surfaces differing in height are measured and, depending on the experimental arrangement, film thickness and/or surface structure are determined. Outstanding examples are single- and multiple-beam interferometry, and phase-contrast and interference microscopy.

(3) In polarization measurements, the ellipticity of the reflected light is determined; thus the technique utilizing this principle is generally called ellipsometry. The main applications of ellipsometry are the determination of optical constants of reflecting surfaces and measurement of index of refraction and thickness of films on substrates.

The relative advantages among these optical methods depend on a number of factors including the objective of the measurements, possible restrictions on sample size and preparation, and the sensitivity and accuracy desired. It will subsequently be shown that ellipsometry is superior to the other methods for many applications. For example, sensitivity to the presence of very thin films is very high and the real part of the index of refraction of absorbing media can be determined with great accuracy. Furthermore, no elaborate sample preparation is required and the method is nondestructive. For these reasons ellipsometry can be a valuable tool in thin-film and surface research.

In this paper we are exclusively concerned with ellipsometry; the purpose is to outline to the nonspecialist the theory, experimental technique, and applications of ellipsometry, and to save him the trouble of reading the numerous articles that are scattered throughout the literature and that are, in many cases, so specific that they tend to discourage rather than encourage the general use of ellipsometry.

Classical Theory of Film Optics

Reflection and Refraction of Light at a Boundary between Two Isotropic Media—Fresnel Formulas

In order to aid in understanding the optics of thin films,[1-9] the problem of determining the light reflected and transmitted at a boundary separating two media will first be reviewed.

For a homogeneous isotropic material characterized by time-independent dielectric permittivity ϵ, magnetic permeability μ, and electrical conductivity σ, containing no space charge, so that $\nabla \cdot \mathbf{E} = 0$, Maxwell's equations can be combined to result in the well-known vector wave equation[10]

$$\nabla^2 \mathbf{A} - \mu\epsilon \frac{\partial^2 \mathbf{A}}{\partial t^2} - \mu\sigma \frac{\partial \mathbf{A}}{\partial t} = 0, \tag{1}$$

where $\mathbf{A} = \mathbf{E}$ or $\mathbf{A} = \mathbf{H}$. Equation (1) together with Maxwell's equations, determines the propagation of electromagnetic waves in this medium.

The problem of reflection and refraction of light at a planar boundary between two isotropic, homogeneous media is usually solved by applying the boundary conditions to the sinusoidal, electromagnetic plane wave solutions of Equation (1). In such an analysis it is expedient to treat the case of plane waves with electric vectors vibrating parallel (p) to the plane of incidence separately from those with vectors vibrating normal (s) to the plane of incidence. Because of the linearity of the wave equation, superposition is allowed and other cases can then be conveniently analyzed by decomposition into p and s components, followed by superposition of the resulting solutions.

Let us define a coordinate system in the conventional manner so that the z-axis is the direction of propagation of the light wave, and the y-axis lies in the plane of discontinuity. We define the plane of incidence as that plane which contains both the z-axis and the normal to the plane of discontinuity. The angle of incidence, ϕ_0, is the angle between the z-axis and the normal to the plane of discontinuity.

Let us also specify the amplitude of the electric vector of a wave traveling in the positive direction in the n^{th} medium and polarized with the electric vector parallel to the plane of incidence by $E_{np}{}^{+}$. We use $E_{ns}{}^{+}$ for the component of the electric vector perpendicular to the plane of incidence. A minus-sign superscript denotes a wave traveling in the negative direction.

The analysis of this problem leads to the following results (see Figure 1):

(a) Law of Reflection:

$$\phi_0 = \phi_0', \tag{2}$$

i.e., the angle of incidence equals the angle of reflection.

(b) Snell's Law of Refraction:

$$n_0 \sin \phi_0 = n_1 \sin \phi_1 = n_2 \sin \phi_2 = \cdots \tag{3}$$

(c) Fresnel Reflection and Transmission Coefficients:

$$\frac{E_{0(p)}{}^{-}}{E_{0(p)}{}^{+}} = \frac{n_0 \cos \phi_1 - n_1 \cos \phi_0}{n_0 \cos \phi_1 + n_1 \cos \phi_0} \equiv r_{01(p)}, \tag{4}$$

$$\frac{E_{1(p)}{}^{+}}{E_{0(p)}{}^{+}} = \frac{2n_0 \cos \phi_0}{n_0 \cos \phi_1 + n_1 \cos \phi_0} \equiv t_{01(p)}, \tag{5}$$

$$\frac{E_{0(s)}^{-}}{E_{0(s)}^{+}} = \frac{n_0 \cos \phi_0 - n_1 \cos \phi_1}{n_0 \cos \phi_0 + n_1 \cos \phi_1} \equiv r_{01(s)}, \tag{6}$$

$$\frac{E_{1(s)}^{+}}{E_{0(s)}^{+}} = \frac{2n_0 \cos \phi_0}{n_0 \cos \phi_0 + n_1 \cos \phi_1} \equiv t_{01(s)}. \tag{7}$$

Here n_0 and n_1 are the optical constants of the two media and ϕ_0 and ϕ_1 are the angles of propagation in the two media.

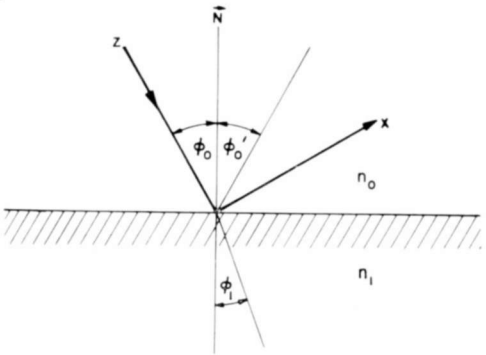


Fig. 1—Reflection and refraction of light at a plane boundary between two media.

When both media are transparent the optical constants are their respective refractive indices. In this case, they are real numbers and are given by

$$n = \sqrt{\frac{\mu\epsilon}{\mu_0\epsilon_0}}, \tag{8}$$

where the zero subscript indicates free space. If we define

$$\epsilon \equiv \epsilon_r \epsilon_0 \tag{9}$$

and

$$\mu \equiv \mu_r \mu_0 \tag{10}$$

then, for the case of $\mu_r = 1$, Equation (8) reduces to

$$n = \sqrt{\epsilon_r}. \tag{11}$$

In this case of two dielectrics, all terms in Snell's Law as well as in the Fresnel equations are real.

Equations (2) to (7) are also valid for the case of absorptive media provided that we use a "complex index of refraction" to characterize these materials. In that case all the terms appearing in these equations may be complex, depending on the specific situation. The meaning of complex trigonometric functions is intimately related to the inhomogeneous nature of the waves in an absorbing medium where planes of constant amplitude do not necessarily coincide with planes of constant phase. A discussion of this problem is, however, beyond the scope of this paper. Complex Fresnel coefficients indicate, of course, that the reflected and refracted rays suffer a phase shift (at the interface) which is neither zero nor 180°.

The complex index of refraction is defined by

$$\tilde{n} \equiv n - ik, \tag{12}$$

where n and k fulfill the following relationships:

$$n^2 - k^2 = \frac{\mu\epsilon}{\mu_0\epsilon_0} = \mu_r\epsilon_r, \tag{13}$$

and

$$nk = \frac{\mu\sigma}{2\omega\mu_0\epsilon_0} = \frac{\mu_r\sigma}{2\omega\epsilon_0}. \tag{14}$$

For most cases of practical interest $\mu_r = 1.0$ and hence

$$n^2 - k^2 = \epsilon_r \tag{15}$$

and

$$nk = \frac{\sigma}{2\omega\epsilon_0} \tag{16}$$

giving

$$n^2 = \frac{\epsilon_r}{2}\left[1 + \left(1 + \frac{\sigma^2}{\omega^2\epsilon^2}\right)^{1/2}\right], \tag{17}$$

$$k^2 = \frac{\epsilon_r}{2}\left[1 - \left(1 + \frac{\sigma^2}{\omega^2\epsilon^2}\right)^{1/2}\right]. \tag{18}$$

It should be kept in mind that σ is the conductivity and ϵ_r the permittivity at the optical frequency concerned, and they are not generally equal to their respective d-c or low-frequency values.

Reflection from a Film-Covered Surface

The results of the previous section may conveniently be utilized for determining the reflection and transmission of plane waves at the boundary of two optically different media of semi-infinite extent separated by a uniform film of a third medium. Let us restrict ourselves to those cases in which the first medium (the immersion medium) is transparent, isotropic and homogeneous, and the other two (substrate and film) may, in general, be absorbing, but are also isotropic and homogeneous.

As seen in the previous section, a beam that strikes an interface between two optically different media is broken up into reflected and

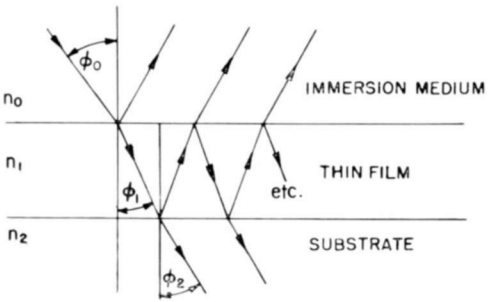


Fig. 2—Reflection and refraction of light at a planar thin film on a substrate.

transmitted components. In the case considered in this section, this division occurs every time the beam reaches the immersion-medium–film and film–substrate interfaces (Figure 2). The total reflected and transmitted beams are then obtained by summing up these multiply reflected and multiply transmitted rays.*

This summation is easily carried out for the case of a single film and the result is conveniently expressed in the form of generalized, complex Fresnel coefficients for the reflection and transmission of plane waves with electric vectors vibrating parallel and normal to the plane of incidence. The reflection coefficient is given by

$$\rho_{(p)} = \frac{r_{01(p)} + r_{12(p)} e^{-2i\delta}}{1 + r_{01(p)} r_{12(p)} e^{-2i\delta}} \tag{19}$$

* A more elegant way of solving not only this case but also multilayer problems is to employ the concept of optical or wave impedance[3] and use the transmission-line analogy, or to use matrix methods.[3,11,12]

and the transmission coefficient by

$$\tau_{(p)} = \frac{t_{01(p)} t_{12(p)} e^{-i\delta}}{1 + r_{01(p)} r_{12(p)} e^{-2i\delta}} \tag{20}$$

with $p = \mathrm{p}$ or $p = \mathrm{s}$ for parallel and normal electric vectors, respectively, and the change in phase of the beam on traversing the film given by

$$\delta = 2\pi n_1 \left(\frac{d}{\lambda_0} \right) \cos \phi_1 = \frac{2\pi d}{\lambda_0} \left(n_1^2 - \sin^2 \phi_0 \right)^{1/2}. \tag{21}$$

Here d is the film thickness and λ_0 the vacuum wavelength of the radiation. (It must be realized that if the film is absorbing, $\cos \phi_1$ is complex, making δ also complex. The meaning of this is beyond the scope of this paper.)

Since $r_{01(s)} \neq r_{01(p)}$ and $r_{12(s)} \neq r_{12(p)}$, it can be seen from Equation (19) that the components of the incident light that are perpendicular and parallel to the plane of incidence are unequally attenuated and unequally shifted in phase upon reflection.

For a more general discussion of the optical behavior of a single film, including anisotropic films, as well as a treatment of multilayer problems the reader is referred to more specialized literature (e.g., References (1)-(9)).

Polarized Light

Since ellipsometry requires the measurement of elliptically polarized light it will be useful to review polarization in general as well as to examine carefully how such elliptically polarized light is characterized.

Plane Waves in a Nonconductive Medium

For a proper understanding of polarization the meaning of plane electromagnetic waves must be elucidated, and this can best be done for a nonconducting medium.

In a Cartesian coordinate system the vector wave equation is simply a set of three scalar equations, one for each of the rectangular components of the vector. (In other coordinate systems it is considerably more complicated or even impossible to write the fields in terms of scalar functions.)

For $\sigma = 0$ the scalar wave equation is then given by

$$\nabla^2 U - \mu\epsilon \frac{\partial^2 U}{\partial t^2} = 0, \tag{22}$$

where U may be any Cartesian component of **E** or **H**. If we assume a sinusoidal time dependence of the form

$$U(x,y,z,t) = u(x,y,z)\,T(t) = u(x,y,z)\,e^{i\omega t}, \tag{23}$$

then the scalar wave equation reduces to the form

$$\nabla^2 u + \gamma^2 u = 0, \tag{24}$$

where the wave number or propagation constant is

$$\gamma = \omega\sqrt{\mu\epsilon} = \frac{\omega}{v}, \tag{25}$$

and the phase velocity is given by

$$v = \frac{1}{\sqrt{\mu\epsilon}}. \tag{26}$$

In rectangular coordinates we may set

$$u(x,y,z) = X(x)\,Y(y)\,Z(z). \tag{27}$$

Separation then results in

$$\begin{aligned} \frac{\partial^2 X}{\partial x^2} + \gamma_x^2 &= 0, \\ \frac{\partial^2 Y}{\partial y^2} + \gamma_y^2 &= 0, \\ \frac{\partial^2 Z}{\partial z^2} + \gamma_z^2 &= 0, \end{aligned} \tag{28}$$

with

$$\gamma_x^2 + \gamma_y^2 + \gamma_z^2 = \gamma^2 = \frac{\omega^2}{v^2}. \tag{29}$$

Equation (22) has the well-known plane-wave solutions

$$U = c \exp \{i(\omega t \pm \gamma \cdot \mathbf{r})\} \tag{30}$$

which represent waves propagating along the (positive or negative) direction of γ. In the general case c is complex because it includes a constant phase shift. One must also remember that it is only the real part of U that represents the actual field.

If we restrict ourselves to propagation along the z-direction, then

$$\gamma_x = \gamma_y = 0, \tag{31a}$$

$$\gamma_z = \gamma \tag{31b}$$

and we have $$U(z, t) = c \exp \{i(\omega t - \gamma z)\}. \tag{30a}$$

This indicates that such waves do not vary as a function of either x or y. However, we must still consider the vector nature of the electromagnetic fields (i.e., U can be any Cartesian component of **E** or **H**) and the requirement that they satisfy Maxwell's equations. By substituting Equation (31a) into Maxwell's equations it can easily be shown that E_z and H_z must both be zero. This means that electromagnetic plane waves in nonabsorbing dielectrics must be transverse, i.e., electric and magnetic field vectors lie in planes normal to the direction of propagation. In addition, Maxwell's equations also show that associated with each transverse component of **E** there is a magnetic field that is in (time) phase with it and at right angles to it.

Pictorial Representation of Polarized Light

If the direction of the transverse **E** vector is constant in time the wave is said to be linearly or plane polarized. Such a plane-polarized wave is the simplest component into which light can be decomposed.

Because the wave equation is a linear equation, any complicated electromagnetic wave with given frequency, propagating in a certain direction, can advantageously be built up by a superposition of individual plane waves of the same frequency with different amplitudes, directions, and phases, but all propagating in the same direction.

(1) Plane or linearly polarized waves are waves for which the electric field vector always lies in a given direction. Such waves are obtained when all the superposed waves have the electric field in the same direction (with arbitrary phase) or if they are in different directions but are exactly in phase. Linear polarization is often characterized by the expression "plane of polarization." This term is quite

ambiguous since in radio engineering one usually describes polarization by the plane of the electric vector, whereas in optics one uses the magnetic field to specify the polarization. In order to avoid ambiguity it is best to give a complete specification as, for example, "polarized with the electric field in the horizontal plane."

(2) Elliptically polarized light is the result of a combination of two uniform plane waves of the same frequency but of different phases, magnitudes, and orientations of the field vectors, and receives its name from the fact that the terminus of the electric field vector traces an elliptic path in the plane normal to the direction of propagation. Elliptic polarization is the most general type of polarization and includes the other two types (linear and circular) as special cases.

Each type of polarization is, in turn, characterized by polarization forms. Linear polarization includes an infinite number of polarization forms, differing as to azimuth, i.e., angle between plane of polarization and reference plane. Circular polarization includes two forms, differing as to direction in which **E** rotates when viewed by looking in the direction of propagation (i.e., right-handed or left-handed polarization). Elliptic polarization includes an infinite number of forms, differing as to azimuth, ellipticity (ratio of minor to major axis of ellipse), and direction of rotation.

Mathematical Specification of Polarized Waves

Several methods are available for describing polarized light.[13] The more sophisticated of these are the Poincaré sphere, the Stokes vector, and the Jones vector. These methods provide direct insight into certain difficult problems, and permit great simplification in many calculations involving the influence of polarizers and retarders upon a wave. The methods are quite useful but are too specialized to warrant description here; only the classical specification will be considered.

In the coordinate system adopted, the plane of incidence is the x-z plane. The angle between the electric field vector at the interface between the two media (i.e., at $z = 0$) and the plane of incidence is the azimuth of the electric field and is denoted by α. The components of the vector **E** in the direction of the coordinate axes are then either parallel or perpendicular to the plane of incidence and are given by

$$E_x = E_{(p)} = E \cos \alpha, \tag{32a}$$

$$E_y = E_{(s)} = E \sin \alpha. \tag{32b}$$

Two arbitrary vibrations with the same frequency, and amplitudes

A and B, parallel to the two coordinate axes, can be represented as

$$E_x = A \cos(\omega t + \theta_x), \tag{33a}$$

$$E_y = B \cos(\omega t + \theta_y). \tag{33b}$$

If the two vibrations are in phase ($\theta_x - \theta_y = 0$) or in opposite phase ($\theta_x - \theta_y = \pm\pi$) the ratio of the above equations is

$$\frac{E_y}{E_x} = \pm\frac{B}{A}, \tag{34}$$

which is simply the equation of a straight line in the x-y plane, and the light is said to be linearly polarized.

In general, however, the components in a light wave have arbitrary phase. In that case the ratio of Equation (33a) and (33b) results, after proper manipulation, in the equation of an ellipse, namely,

$$\left(\frac{E_x}{A}\right)^2 + \left(\frac{E_y}{B}\right)^2 - 2\left(\frac{E_x}{A}\right)\left(\frac{E_y}{B}\right)\cos\Delta = \sin^2\Delta, \tag{35}$$

where

$$\Delta \equiv \theta_x - \theta_y. \tag{36}$$

The terminus of the light vector traces out an ellipse that is inscribed in a rectangle $2A$, $2B$ (see Figure 3). From Equation (35) one can clearly see that the semiaxes of the ellipse do not coincide with the coordinate axes.

We can, however, choose a coordinate system in which the semiaxes of the ellipse will be parallel to the coordinate axes, so that the term E_xE_y/AB in Equation (35) will vanish (Figure 3). Let us denote the new coordinate system (ξ,η). From the theory of linear transformations we know that the rotation of a coordinate system is given by

$$E_\xi = E_x \cos\chi + E_y \sin\chi \tag{37a}$$

$$E_\eta = -E_x \sin\chi + E_y \cos\chi \tag{37b}$$

where χ is the angle between the ξ-axis (i.e., the major semiaxis of the ellipse) and the x-axis. In the new coordinate system the ellipse is given by

$$E_{\xi} = a \cos (\omega t + \theta_0) \tag{38a}$$

$$E_{\eta} = \pm b \sin (\omega t + \theta_0) \tag{38b}$$

where a and b are the semiaxes of the ellipse. We introduce the double sign in the equation for E_{η} so that the two possible ellipticities (positive or negative) can be taken into consideration. Substituting Equation (33) into Equation (37), setting Equations (37) and (38) equal,

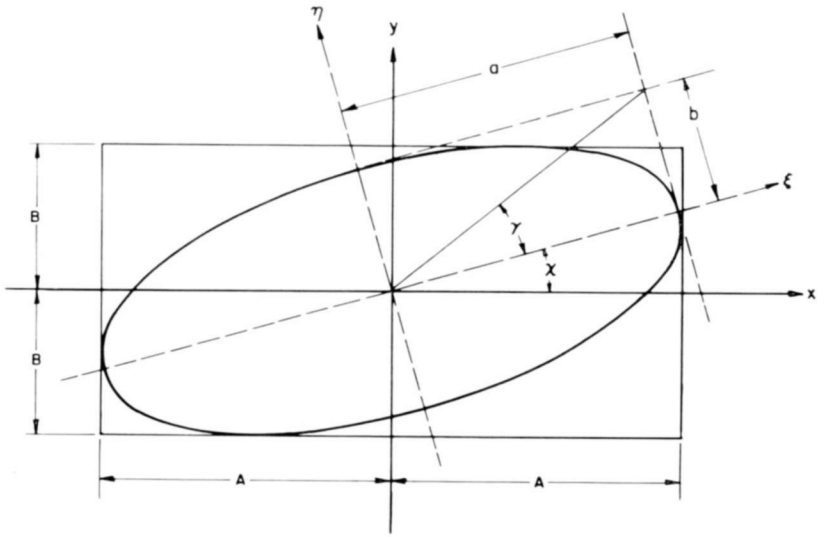


Fig. 3—Characterization of a general inclined ellipse.

and expanding all the trigonometric functions we get the following relations:

$$\tan 2\chi = \tan 2\psi \cos\Delta \tag{39}$$

$$\sin 2\gamma = \pm \sin 2\psi \sin\Delta \tag{40}$$

$$\tan\Delta = \frac{\pm \tan 2\gamma}{\sin 2\chi} \tag{41}$$

where

$$\tan\psi \equiv \frac{B}{A} = \frac{|E_{(p)}|}{|E_{(s)}|}, \tag{42}$$

and

$$\tan\gamma \equiv \frac{b}{a}. \tag{43}$$

We see now that the ellipse, which we first characterized from physical reasoning by ψ and Δ (i.e., amplitude ratio and phase difference of the components of light parallel and perpendicular to the plane of incidence), can, from geometrical considerations, also be characterized by the inclination of its major axis with respect to the x-z plane of the original coordinate system (i.e., plane of incidence), χ, and by the ellipticity, given by the ratio of minor to major semiaxis, $\tan \gamma = b/a$. Since these two sets of two quantities each describe the same ellipse,

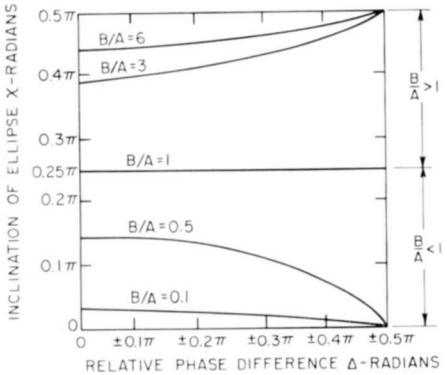


Fig. 4—Inclination of ellipse versus relative phase difference with amplitude ratio as a parameter.

there must be a relation between them. These relations are given by Equations (39), (40), and (41); they are very important in the theory of ellipsometry, since it is χ and γ that are determined experimentally, whereas the thickness and the optical properties of the film are given as functions of ψ and Δ.

It can be seen from Equation (39) that the inclination of the ellipse, χ, is a function of both the amplitude ratio B/A and the phase difference Δ. Only in the case where $A = B$ is the inclination at a constant angle ($\pm\pi/4$), independent of Δ.

$$\tan 2\chi = \left(\tan 2 \arctan \frac{1}{1} \right) \cos\Delta = \tan \left[2 \times \frac{\pi}{4} \right] \cos\Delta \to \infty, \quad \chi = \frac{\pi}{4}. \tag{44}$$

For $B/A > 1$ the ellipse rotates clockwise for increasing Δ, starting at $\chi = \psi = \arctan (B/A)$ for $\Delta = 0$ to $\chi = 0$ for $\Delta = \pi/2$. For $B/A < 1$ the ellipse rotates counterclockwise from $\chi = \psi = \arctan (B/A)$ for $\Delta = 0$ to $\chi = \pi/2$ for $\Delta = \pi/2$, as shown in Figure 4.

Ellipsometry

The Fundamental Equation of Ellipsometry

The absolute changes in amplitude and phase, given by the Fresnel coefficients for parallel and normal electric vectors [Equation (19)], can be investigated experimentally by intensity and interference methods, respectively. Relative changes of amplitude and phase can be conveniently studied by reflecting a polarized light beam from the surface under study, and examining the changes in the polarization of the beam. These relative changes can be expressed by the ratio of the generalized Fresnel reflection coefficients for the p-wave to that for the s-wave. This results in the fundamental equation of ellipsometry:

$$\frac{\left(\frac{E_{(p)}}{E_{(s)}}\right)_{refl}}{\left(\frac{E_{(p)}}{E_{(s)}}\right)_{inc}} = \frac{\left(\frac{E_{refl}}{E_{inc}}\right)_{(p)}}{\left(\frac{E_{refl}}{E_{inc}}\right)_{(s)}}$$

$$= \frac{\rho_{(p)}}{\rho_{(s)}} \equiv \frac{\tan \psi_{refl}}{\tan \psi_{inc}} \exp\{i(\Delta_{refl} - \Delta_{inc})\} \equiv e^{i\Delta} \tan \psi, \tag{45}$$

where $\tan \psi_{refl} \equiv \left(\frac{|E_{(p)}|}{|E_{(s)}|}\right)_{refl}$ and $\Delta_{refl} \equiv (\theta_{(p)} - \theta_{(s)})_{refl}$.

Analogous definitions hold for the incident wave.

In general, the incident light is plane polarized, with the plane of vibration of the electric vector inclined at $\pm\pi/4$ with respect to the plane of incidence. In that case,

$$\tan \psi_{inc} = \left(\frac{|E_{(p)}|}{|E_{(s)}|}\right)_{inc} = 1; \quad \Delta_{inc} = 0, \tag{46}$$

so that

$$e^{i\Delta} \tan \psi = \left(\frac{E_{(p)}}{E_{(s)}}\right)_{refl}. \tag{47}$$

Since, in general, $|E_{(p)}| \neq |E_{(s)}|$ and $\theta_{(p)} \neq \theta_{(s)}$, we see that the reflected wave is elliptically polarized. The ellipsometer measures experimental quantities that allow the determination of ψ and Δ (as will be shown later).

Because of Snell's Law, the Fresnel coefficients in Equation (45) can all be expressed in terms of the optical constants of the media bounding the reflective interfaces and the angle of incidence in the ambient medium. If Equations (4), (6), and (21) are substituted into Equation (45), and the resulting equation separated into its real and imaginary parts, there results one equation for ψ and one for Δ, each as functions of the angle of incidence in the immersion medium, the vacuum wavelength of the light, the index of refraction of the substrate, and the thickness and refractive index of the film. All of these quantities can be independently determined, or are fixed constants, except for the properties of the film.

Solutions of the Ellipsometry Equation

Since the above-mentioned equations for ψ and Δ are transcendental for film-covered surfaces, they can only be solved by either making appropriate simplifying assumptions, or by graphical or numerical methods. For film-free surfaces explicit solutions can be obtained.

(a) *For the film-free case*, i.e., $d = 0$, Equation (45) (in combination with Equation (46)) can be solved for the optical constants of the substrate, resulting in

$$n^2 - k^2 = \sin^2\phi_0 \left[1 + \frac{\tan^2\phi_0(\cos^2 2\bar{\psi} - \sin^2 2\bar{\psi}\sin^2\bar{\Delta})}{(1 + \sin 2\bar{\psi}\cos\bar{\Delta})^2} \right], \tag{48a}$$

$$2nk = \frac{\sin^2\phi_0 \tan^2\phi_0 \sin 4\bar{\psi}\sin\bar{\Delta}}{(1 + \sin 2\bar{\psi}\cos\bar{\Delta})^2}, \tag{48b}$$

where the bars over ψ and Δ indicate a film-free surface.

(b) *For very thin films* ($d \ll \lambda$) Drude expanded the exponential terms in Equation (45) in a power series of (d/λ), discarding terms of higher order than the first to get

$$\Delta = \bar{\Delta} - \alpha d, \tag{49a}$$

$$\psi = \bar{\psi} + \beta d, \tag{49b}$$

where

$$\alpha = \left(\frac{4\pi}{\lambda_0}\right) \left[\frac{\cos\phi_0 \sin^2\phi_0 (\cos^2\phi_0 - a)\left(\frac{1}{n_1^2} - 1\right)}{(\cos^2\phi_0 - a)^2 + a_1^2} \right], \tag{50a}$$

$$\beta = \left(\frac{2\pi}{\lambda_0}\right)\left[\frac{\cos\phi_0 \sin 2\bar{\psi} \sin^2\phi_0 a_1 \,(1 - n_1^2 \cos^2\phi_0)\left(\frac{1}{n_1^2} - 1\right)}{(\cos^2\phi_0 - a)^2 + a_1^2}\right], \tag{50b}$$

$$a = \frac{n_2^2 - k_2^2}{(n_2^2 + k_2^2)^2}, \tag{51a}$$

$$a_1 = \frac{2n_2 k_2}{(n_2^2 + k_2^2)^2}. \tag{51b}$$

$\bar{\Delta}$ and $\bar{\psi}$ as well as n and k for the film must be determined by independent experiments. Drude's equation and variations of it are first-order approximations and are valid for film thicknesses small compared to one wavelength, i.e., $d \ll \lambda_0$ or $d \cong$ 50-100 Å. By a binomial expansion one gets a second-order approximation to exact theory that is valid for absorbing films as thick as 1000 Å.

Approximate theories are advantageous in the sense that it is relatively easy to solve the equations to give film properties from experimental measurements.

(c) *For thin films with a thickness not small compared with* λ, exact theory must be used. Winterbottom[4] presents graphical solutions of the exact equations, showing the functional relationships between film properties and ellipsometer measurements. Vašiček[14] gives tables to evaluate measurements of transparent films on glass. Archer[15] and McCrackin et al[16], among others, applied computer techniques to the solution of the transcendental equations. The result can be either a graphical representation of the dependence of ψ and Δ on the properties of the film or specific numerical answers.

The Ellipsometer

The ellipsometer is an instrument that allows the determination of the optical constants and thickness of thin films by analyzing the elliptically polarized light reflected from a thin film on a reflecting substrate.

As has been shown, the optical constants and the thickness of the thin film are determined by the amplitude ratio $\psi = \arctan(B/A)$ and the phase difference Δ of the components of the reflected light parallel and perpendicular to the plane of incidence. For transparent films measurement at one angle of incidence is sufficient, but for

absorbing films two measurements at different angles of incidence are required. The quantities ψ and Δ, in turn, are related to the inclination χ and ellipticity $\tan \gamma = b/a$ by Equations (39) — (41). The ellipsometer, finally, enables the experimental determination of χ and γ.

From the discussions of the mathematical specification of polarized waves, it can be seen that there are only two coordinate systems in which the rotating **E** vector can be resolved into two perpendicular components out of phase by $\pi/2$, namely those in which the coordinate axes are parallel to the major and minor semiaxes of the ellipse.

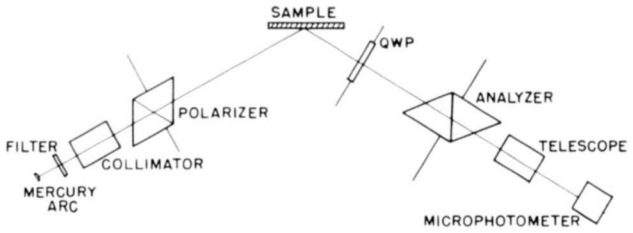


Fig. 5—Schematic representation of ellipsometer.

Therefore, the only way elliptically polarized light can be transformed into plane-polarized light with the help of a quarter wave plate (QWP) (which introduces a $\pi/2$ phase shift between vibrations parallel to its fast and slow axes, see Appendix), is by aligning the fast and slow axes of the QWP with the major and minor axes of the ellipse. The result is plane-polarized light at an angle of $\beta = \chi + \gamma$ with respect to the plane of incidence. The angle that the QWP makes with respect to the plane of incidence is χ. The angle β can be determined with an analyzing Nicol prism. At extinction the analyzer azimuth is $\pi/2 + \chi + \gamma$.

Once χ and γ are experimentally determined, Equations (40) and (41) are used to determine ψ and Δ; from these the thickness and optical constants of the film can be determined.

An ellipsometer, schematically represented in Figure 5, is a polarizing spectrometer with collimator and telescope arms swinging in a plane with provisions for reading the angles of incidence and reflection on a large fixed circle. The polarizer is a Nicol prism (Glan-Thompson prism) mounted in a divided circle on the collimator, and the analyzer is a similar prism mounted on the telescope. The compensator is also mounted in a divided circle; it is approximately a quarter-wave plate

for the light used and is attached either to the telescope or the collimator depending on the experimental technique to be used (see below). The light source, mounted on the collimator can, for example, be a mercury arc with a filter for isolating the 5461 Å line. Extinction settings are determined by using a photomultiplier microphotometer, mounted on the telescope, as a detector.

Experimental Technique

Alignment of the polarizer and analyzer prisms is the first step in setting up the ellipsometer for operation. Alignment of a prism simply means the determination of the scale reading for which the plane of vibration of the light transmitted by the prism is parallel with the plane of incidence. Correct alignment is critical for the accurate determination of optical constants of surfaces; it is not quite as critical for the measurement of the thickness and index of refraction of thin films.[16]

Alignment can be achieved by utilizing the fact that if randomly polarized light is incident upon a dielectric at the Brewster angle, the reflected ray is linearly polarized. In principle this is quite simple, but in practice great care must be taken and the method outlined by McCracken et al[16] is recommended.

With the collimator and telescope aligned and analyzer and polarizer in crossed position, the QWP can be mounted in its divided circle and turned until an extinction setting is obtained. Two such settings can be observed; in one the fast axis is parallel to the polarizer and in the other perpendicular to it. This determines the direction of the two axes. A simple reflection experiment must be performed in order to differentiate between the two axes. The relative retardation of the plate is then determined by use of an auxiliary plate in a known azimuth, a method devised by A. B. Winterbottom.[4] The QWP is usually a thin doubly refractive crystal such as mica. Because of the thinness of the mica crystals required (0.035 mm) it is difficult to cleave them to produce retarders of exactly $\pi/2$. Even though the phase retardation of a particular plate is not exactly $\pi/2$ one can still use it to convert elliptically polarized light into plane-polarized light, but in the analysis of the results corrections must be applied to Equations (39) — (41). These corrections take different forms, depending on the measurement method applied.

It is clear that $e^{i\Delta} \tan \psi$ can be determined either by finding the parameters of elliptically polarized light obtained by reflecting plane polarized light which was incident with the plane of vibration inclined

at $\pi/4$ with respect to the plane of incidence, or by finding the parameters of elliptically polarized light that gives plane polarized light upon reflection.

Method #1: In this method, plane-polarized light with the plane of vibration of the electric field vector inclined $\pi/4$ with respect to the plane of incidence is reflected from the film. The reflected beam is elliptically polarized and the ellipticity and inclination of the reflected light are determined by the azimuths of the QWP (mounted on the telescope) and the analyzer. The QWP azimuth is χ, whereas the analyzer azimuth is $\pi/2 + \chi + \gamma$.

Method #2: Here one attempts to find the parameters of elliptically polarized light that gives plane polarized light upon reflection. This method is, from an experimental point of view, very convenient since it allows one to mount the QWP on the stationary collimator instead of the movable telescope arm, and with it one can compensate any phase difference with a QWP fixed in $\pm\pi/4$ azimuth. For any given surface there are several combinations of polarizer, analyzer, and compensator scale settings that result in extinction. An excellent discussion of this problem is given in Reference (16). In the following, all azimuthal angles are considered to be positive in the counterclockwise direction from the plane of incidence when looking into the light beam, the QWP azimuth is $+\pi/4$ and the analyzer azimuth is always in the fourth quadrant. The relative phase retardation, Δ, and the amplitude ratio, $\tan\psi$, are then given by the following relations:[15]

$$\tan\Delta = \sin\beta \tan\left(\frac{\pi}{2} - 2P_0\right), \tag{52}$$

$$\cos 2L \equiv -\cos\beta \cos 2P_0, \tag{53}$$

$$\tan\psi = \cot L \tan(-A_0), \tag{54}$$

where β is the actual relative retardation of the QWP, and P and A are the polarizer and analyzer azimuth angles, respectively. The zero subscript indicates extinction settings.

The extinctions settings can be obtained by several methods that differ in accuracy and sensitivity. Approximate extinction settings can be obtained by alternately adjusting analyzer and polarizer until minimum light transmission occurs. If more accurate extinction settings are required, graphical plots of the intensity of the transmitted beam versus polarizer and analyzer settings may be used.

Another, more accurate method, is as follows. First, approximate extinction settings are made by adjusting P and A for minimum light transmission. Then the exact extinction setting for the polarizer P_0 is determined by measuring P at equal intensities on each side of the minimum, and averaging these two values. The polarizer must then be set at P_0 and the same method is reapplied to find the correct extinction setting for the analyzer, A_0. This method is, in principle, based on the fact that the light intensity at the detector, I, is symmetric about P_0 for any setting of A, and symmetric about A_0 if $P = P_0$; thus[15]

$$I \propto \sin^2(A - A_0) + \sin 2A_0 \sin 2A \sin^2(P - P_0). \tag{55}$$

Since it is not assured that the initial extinction setting for A_0 is correct, the above procedure must be successively repeated until the results are within the required tolerances or within the sensitivity of the instrument.

For ultrasensitivity, a Faraday cell and appropriate electronics as well as a phase-sensitive detector can be added to the standard detection equipment.[17]

Applications

The physics of thin films and surfaces is very extensive and offers numerous areas for the successful application of ellipsometry. The following discussion relates to areas that might be of interest to workers in solid-state physics, inorganic and physical chemistry, and physical electronics.

Optical Constants of Film-Free Materials

Optical constants of dielectrics, semiconductors[15,18,19], and conductors[20,21] can be determined by ellipsometry. In order to achieve maximum sensitivity for these measurements a judicious choice of a proper angle of incidence must be made. This problem has been analyzed for absorbing media by Ditchburn.[22] He finds that for metals the angle of incidence should be approximately the principal angle for the particular material under consideration, whereas for semiconductors there are certain limits to the angle of incidence within which it is preferable to measure Δ or ψ for two angles of incidence. He also shows which method is most sensitive when the constants are in certain ranges. Even though this method is very sensitive, it is prone to errors due to unintentional thin surface films. It is clearly necessary in many cases to take into account the effect of such surface films in

evaluating the optical constants from ellipsometry data by applying to the experimental measurements the corrections given by the Drude equations, namely Equation (49), or to produce and measure the surface in ultra high vacuum.[23] The presence of such a film is easily established because for light reflected at the principal angle, $\Delta = \pi/2$ and the value of χ can be obtained from Figure 4.

It should be realized, in connection with absorbing materials, that the optical constants determined are characteristic only of the material within a few skin depths from the surface. This is not necessarily a disadvantage because it allows the study of surface effects.

Optical Constants and Thickness of Thin Films

The application of ellipsometry for the determination of the properties of thin nonabsorbing films is straightforward. Depending upon the thickness of the film, one of the solutions of the ellipsometry equations outlined previously must be used. The angle of incidence is usually chosen to give maximum sensitivity for the determination of the film thickness. No general rule is available to make this choice, and the angle of incidence will depend upon the particular substrate, film, and ambient medium. Smith and Hacskaylo[24] present curves and equations to illustrate the dependence of the sensitivity on the experimental parameters. Mertens et al[21] find that for organic films on metals the maximum sensitivity for thickness determination is achieved if the angle of incidence is close to the principal angle of the substrate. Studies have been performed on a number of materials, including silicon[15,25,26,27], titanium[28], copper[4], aluminum[4], iron[4], stereate films on metals[21,29], and dielectric films on glass.[30,31]

For absorbing films this method is only meaningful when the attenuation is sufficiently low that the light emerging from the film–ambient interface after reflection from the film–substrate interface can still be detected. The analysis of elliptically polarized light that results from the reflection of linearly polarized light does not furnish more than two parameters. For this reason one must perform two measurements at different angles of incidence; this will result in four equations for the three unknown parameters (n, k, d) that characterize an absorbing film.

Study of Physical and Chemical Processes on Surfaces

Ellipsometry allows the investigation of a large variety of different physical and chemical processes,[32] and its main advantage in such applications is the fact that measurements can be made *in situ* in both

gaseous and liquid ambients, are nondestructive, and are, in most cases, more sensitive than those of other methods.

The processes amenable to ellipsometric study can be classified in various ways. From a physico-chemical point of view it seems most natural to make a distinction as to whether or not a new phase has been formed on the surface.

Surface Layer with Optical Properties Different from Those of the Bulk

It is well known that mechanical treatments change the structure of surface layers to varying depths depending on the material and the

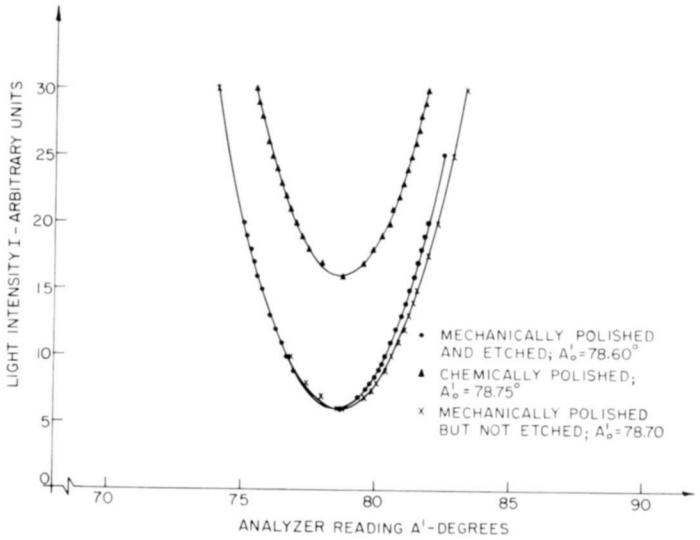


Fig. 6—Light intensity at the detector versus analyzer setting for (111)-oriented silicon wafers treated in various ways. For the particular experimental arrangement used here ψ is obtained from A_0' by the relation $\psi = 91.43^\circ - A_0'$.

treatment. Such changes are revealed by deviations of the optical properties (among others) from those of the bulk. Archer,[15] for example, finds by very precise ellipsometer determinations of the optical constants of germanium that etched surfaces give results that correspond to the bulk properties of very pure germanium, while mechanically polished surfaces give results similar to those of highly doped germanium. The present authors find that the curve of light intensity at the detector versus analyzer angle for mechanically polished silicon is shifted with respect to the curve for chemically polished silicon (see Figure 6). The angle for the minimum of the curve representing

a mechanically polished sample is different from that for the chemically polished one, indicating that the two specimens have different optical properties. There is also a difference in the minimum light intensities and slopes of the curves, most probably due to variation in surface roughness. This problem has as yet not been analyzed in the literature.

Diffusion of impurities into surface layers might, in certain cases, manifest itself also in a change in the optical properties of the surface layers. Such changes have been observed for boron diffusion into silicon by the authors and for silicon immersed into a molten mixture of $LiNO_3$ and KNO_3.[26]

Space-charge layers at the surface of semiconductors are present because of the existence of surface states; they can, under certain conditions, also be present in insulating layers, especially in oxide layers grown on metals or semiconductors. Due to the contact potential at the interface between two materials there might also be a space-charge layer. Even though a space-charge layer does not constitute a different phase of the material, it still could have different optical properties. Mertens et al,[21] for example, stipulate that the space-charge layer in an insulator on a metal substrate is the cause for a position-dependent absorption in this dielectric.

Adsorption and Desorption Studies

Because of the many advantages outlined above, adsorption and desorption studies from liquids or gases *in situ* are conveniently carried out by the use of ellipsometry. Thus, adsorption isotherms for water and various organic liquids on single-crystal silicon[33] and the thickness and index of refraction of adsorbed polystyrene films on chromium[34] have been determined, and optical measurements on thin films of condensed gases at low temperatures have been made.[35]

Study of Deposition, Growth, and Dissolution of Films

This category represents perhaps the largest area for application of ellipsometry and for this reason a major portion of the available literature in ellipsometry is concerned with problems falling into this classification. Deposition of films from the liquid or gas phase (evaporation) can be investigated for various materials in order to establish the pertinent process parameters, and their influence on the deposited or grown material.

In growth studies the main interest thus far has been in the investigation of oxidation kinetics, and typical examples are the oxidation of copper[4], aluminum[4,36], and iron[4], room-temperature oxidation of

germanium and silicon[25], oxidation of silicon at elevated temperatures in various ambients[26,27], and oxidation of titanium[28]. Another interesting experiment in this area is the *in situ* electrochemical study of film formation and growth on mirror electrodes immersed in an aqueous electrolyte[37].

Little has been reported on the study of dissolution processes by the use of ellipsometry. The authors have studied the dissolution of oxide films obtained by thermal oxidation of GaAs and found that the constitution of these films changed when the film–substrate interface was approached, and that there was an accumulation of As and/or As_2O_3 at the interface.[19]

Although the theory of ellipsometry is very old, it has not been applied to the extent that its usefulness would warrant. Only a small portion of the possible areas of application have been investigated; a great variety of new experiments can be envisioned.

Evaluation

Errors and Accuracy

Random and systematic (instrumental) errors occur, of course, in the use of the ellipsometer, and they are eventually carried over to the derived optical constants of substrates and film parameters.

Multiple reflections in the optical system produce a number of beams of decreasing intensity and of different states of polarization that influence the settings of both polarizer and analyzer, and can, under certain circumstances, introduce errors of the order of one degree of azimuth. These errors cannot be eliminated by averaging several independent extinction settings, and it is impossible to compute corrections with certainty. The only possibility for reducing the influence of these effects is through the use of optical components having nonreflecting coatings.[4]

There is also the possibility of errors due to birefringence in the optical components. Little is known concerning their magnitude, and in order to reduce their influence care must be taken to avoid any strain in the optical system.

The sensitivity of the photometer is another important parameter that might limit the accuracy of the instrument, especially in applications where the cross section of the incident beam is reduced so that the specimen can be scanned. In extreme cases it might be necessary to cool the photomultiplier tube in order to reduce noise problems.

Systematic errors may arise from errors in the initial alignment of polarizer, analyzer, and QWP, and from an error in the determination of the relative retardation of the QWP.

In order to fully appreciate the importance of these errors and the ones introduced during the measurement process, the procedure used to obtain the final results from the experimental data will be briefly outlined. The extinction settings for polarizer, analyzer, and QWP which, in addition to the above-mentioned alignment errors also contain random errors, are used in Equations (52)-(54) (in connection with the relative retardation of the QWP, which also contains an alignment error) to obtain ψ and Δ. From these, in turn, the values of n, k, and d are determined. Because of this complicated procedure, the propagation of errors is obscure. It is possible that by taking two or more of the 32 independent sets of the extinction settings[16] and performing suitable averaging, the errors due to misalignment could be reduced to such an extent that they could be neglected in comparison to others. The lack of a rigorous error analysis for both systematic and random errors is in part responsible for the large discrepancies in the accuracy of results reported by various workers.

Limitations Introduced by the Properties of the Specimen

Lateral inhomogeneities of the film or substrate and/or variations in film thickness within the area covered by the light beam cannot be determined by ellipsometry. They might simply cause a decrease in the accuracy, and the results obtained would represent the average or effective parameters for the area examined[38].

Optical properties of films having an inhomogeneity along the normal to the boundary have been treated by Abelès.[39] The problem of anisotropy was discussed by Winterbottom,[4] and experiments in this field were performed recently.[40-41]

Differences between Optical Properties of Thin Films and Bulk Materials

Whenever the thickness of a film is much smaller than the wavelength of the light used, the measured properties generally begin to differ from those characteristic of the bulk material. The meaning of the thickness of such a film becomes ambiguous, first because, as is well known from the thermodynamics of interfaces, no sharp boundaries can exist between two phases, and second, because of the granular structure of many of these films (especially of those obtained by evaporation).

The optical properties of metal films having such granular structure have been explained by two different theories. David[42] interprets the variation of the optical constants with thickness of the film on the basis of ellipsoids of revolution having the same optical constants as the bulk material but being separated by voids. Fragstein and

Römer[13] assume that the optical constants of thin films are different from those of the bulk material.

Dielectric thin films may be granular, but are not necessarily so. This is especially true for films obtained by oxidation, and even more so if the resulting film is amorphous. In such a case, however, the optical properties might still be different from those of the bulk material. This could be explained as follows. The phase velocity of the light wave in the film is not necessarily the same as in the bulk because the internal field in the film could obviously be quite different from that in the bulk as a result of dimensional effects alone, even if other typical surface phenomena were not considered. Thus, for instance, anisotropy can be expected even if the bulk material is isotropic. Macroscopic concepts, such as the relationship between the index of refraction, dielectric constant and polarizability, as given by the Clausius–Mosotti and Lorentz–Lorenz equations, are not valid a priori for thin films because the very assumptions involved in their derivation are not necessarily fulfilled.

Conclusion

Ellipsometry is a useful method for the study of surface phenomena because it allows the measurement of the properties of very thin films in a large variety of combinations of substrates, film, and immersion medium. These measurements are nondestructive, can be performed *in situ* (even in liquids), and one measurement generally allows the determination of two parameters.

Even though the relations between the optical properties and the measured quantities are complicated, the desired information can readily be extracted through the use of computers.

The ellipsometer, like all precision optical instruments, is a delicate piece of equipment and requires extreme care in its construction, maintenance, and operation. Because multiple reflections and birefringence in the optical components introduce large instrumental errors, a proper choice for the parts employed must be made so as to minimize these effects.

The sensitivity of the method is quite high and, by proper statistical analysis of experimental results, a very high accuracy can be obtained—in many cases as high as five significant figures—if the instrumental errors are properly taken into account.

Appendix—Description of the Quarter-Wave Plate

The function of a quarter-wave plate (QWP) is to produce a phase

shift of $\pi/2$ between two electric vectors perpendicular to each other. This can be achieved by the proper use of a birefringent crystal.

In general, the optical behavior of birefringent crystals is quite complicated. It is the main subject of the optics of crystals and, quite obviously, cannot be treated here. For the understanding of retardation plates it will suffice to say that there are two characteristic directions, perpendicular to each other, and both lying in the face of the plate. Plane-polarized light that is normally incident upon the face of this plate, has different phase velocities, depending on which of the two characteristic directions the electric vector is parallel to. One usually distinguishes between the "fast" and "slow" directions of vibrations of a plate. The fast axis is, of course, in the direction of vibrations that have the greatest phase velocity and the lowest index of refraction (n_f); the slow axis is in the direction of vibrations with the lowest velocity and the highest index of refraction (n_s). For studying the effect of a retardation plate on normally incident light (generally elliptically polarized), it is expedient to decompose it into two plane-polarized beams with electric vectors parallel to fast and slow axes. These two beams have the same geometrical path length inside the plate, but their optical path lengths are dependent on the indices of refraction, so that the optical path difference is

$$l = t(n_s - n_f), \tag{56}$$

and the phase difference between the two beams upon emerging from the plate is

$$\beta = 2\pi \frac{t}{\lambda_0} (n_s - n_f), \tag{57}$$

where t is the thickness of the plate.

A retarding plate with a thickness that produces a phase difference of $\pi/2$ is called a quarter-wave plate.

We see then that if two perpendicular **E** vectors, which are out of phase by $\pm\pi/2$, are normally incident upon a QWP in such a way that they are parallel to the fast and slow axes of the QWP, they will emerge with a phase difference of 0 or π, and plane-polarized light will result.

ACKNOWLEDGMENT

The authors would like to thank R. J. Evans for assistance in the experimental work.

References

1. A. B. Winterbottom, "Optical Methods for the Determination of Films on Metals Based on Polarization and Interference Phenomena," p. 802, Chapter of *Metallic Corrosion, Passivity and Protection*, by U. R. Evans, E. Arnold and Co., London, 1946.

2. H. Mayer, *Physik Dünner Schichten*, Teil I, Wissenschaftliche Verlagsgesellschaft, Stuttgart, 1950.

3. O. S. Heavens, *Optical Properties of Thin Solid Films*, Butterworths, London, 1955.

4. A. B. Winterbottom, *Optical Studies of Metal Surfaces*, I. Kommisjon Hos. F. Bruns Bokhandel, Trondheim, 1955.

5. H. Wolter, "Optik Dünner Schichten," in *Handbuch der Physik*, edited by S. Flügge, Vol. 24, p. 461, Springer Verlag, Berlin, 1956.

6. P. Bousquet and P. Rouard, "Constantes Optiques et Structure des Couches Minces," *Jour. Phys. Radium*, Vol. 21, p. 873, 1960.

7. O. S. Heavens, "Optical Properties of Thin Films," in *Reports on Progress in Physics*, ed. by A. C. Stickland, Vol. 23, p. 1, The Physical Society, London, 1960.

8. A. Vašiček, *Optics of Thin Films*, North-Holland Publishing Co., Amsterdam, 1960.

9. F. Abelès, "Methods for Determining Optical Parameters of Thin Films," in *Progress in Optics*, Vol. 2, p. 250, ed. by E. Wolf, North-Holland Publishing Co., Amsterdam, 1963.

References 1 to 9 are standard treatments of the optics of thin films and contain extensive bibliographies of earlier work. For this reason the subsequent references are limited to more recent publications.

10. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, New York, 1959.

11. F. Abelès, "The Propagation of Electromagnetic Waves in Stratified Media," *Ann. Phys.* (Paris), Vol. 3, p. 504, 1948.

12. F. Partovi, "Theoretical Treatment of Ellipsometry," *Jour. Opt. Soc. Amer.*, Vol. 52, p. 918, Aug. 1962.

13. W. A. Shurcliff, *Polarized Light*, Harvard Univ. Press, Cambridge, 1962.

14. A. Vašiček, "Tables for the Determination of the Refractive Index and of the Thickness of the Thin Film by the Polarimetric Method," *Jour. Opt. Soc. Amer.*, Vol. 37, p. 979, Dec. 1947.

15. R. J. Archer, "Determination of the Properties of Films on Silicon by the Method of Ellipsometry," *Jour. Opt. Soc. Amer.*, Vol. 52, p. 970, Sept. 1962.

16. F. L. McCrackin, E. Passaglia, R. Stromberg, and H. L. Steinberg, "Measurement of the Thickness and Refractive Index of Very Thin Films and the Optical Properties of Surfaces by Ellipsometry," *Jour. Res. Natl. Bur. Std.*, Vol. 67A, p. 363, 1963.

17. M. Weingart and A. R. Johnston, "Electronics Polarimeter Techniques," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

18. R. J. Archer, "Optical Constants of Germanium," *Phys. Rev.*, Vol. 110, p. 354, 1958.

19. K. H. Zaininger and A. G. Revesz, "Ellipsometric Investigations of Oxide Films on GaAs," Colloquium on the Optics of Solid Thin Layers, Marseille, France, Sept. 1963.

20. S. Roberts, "On Determining Optical Constants of Metals in the Infrared," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

21. F. P. Mertens, P. Theroux, and R. C. Plumb, "Some Observations on the Use of Elliptically Polarized Light to Study Metal Surfaces," *Jour. Opt. Soc. Amer.*, Vol. 53, p. 788, 1963.

22. R. W. Ditchburn, "Some New Formulas for Determining the Optical Constants from Measurements on Reflected Light," *Jour. Opt. Soc. Amer.*, Vol. 45, p. 743, Sept. 1955.

23. J. F. Dettorre, T. G. Knorr, and D. A. Vaughan, "Application of Ellipsometry to the Study of Phenomena on Surfaces Prepared in Ultrahigh Vacuum," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

24. R. C. Smith and M. Hacskaylo, "Determination of the Optimum Experimental Sensitivity for Thickness Measurements by the Drude Technique," Symposium on the Ellipsometer and

Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

25. R. J. Archer, "Optical Measurement of Film Growth on Silicon and Germanium Surfaces in Room Air," *Jour. Electrochem. Soc.*, Vol. 104, p. 619, Oct. 1957.
26. A. G. Revesz and K. H. Zaininger, "Some Properties of SiO_2 Films on Silicon," Colloquium on the Optics of Solid Thin Layers, Marseille, France, Sept. 1963.
27. B. H. Claussen and M. Flower, "An Investigation of the Optical Properties and the Growth of Oxide Films on Silicon," *Jour. Electrochem. Soc.*, Vol. 110, p. 983, Sept. 1963.
28. R. C. Menard, "Optical Measurement of Oxide Thickness on Titanium," *Jour. Opt. Soc. Amer.*, Vol. 52, p. 427, April 1962.
29. A. Rothen, "Improved Method to Measure the Thickness of Thin Films with a Photoelectric Ellipsometer," *Rev. Sci. Instr.*, Vol. 28, p. 283, April 1957.
30. A. Vašiček, "Polarimetric Methods for the Determination of the Refractive Index and the Thickness of Thin Films on Glass," *Jour. Opt. Soc. Amer.*, Vol. 37, p. 145, March 1947.
31. J. Rassow, "Über ein graphisches Verfahren zur polarisationsoptischen Bestimmung von Dicke und Brechungsindex beliebig dicker, nichtabsorbierender, homogener Schichten auf Glas," *Zeit. für Physik*, Vol. 168, p. 353, 1962.
32. J. Kruger, "Use of Ellipsometry for *In-Situ* Studies of the Oxidation of Metal Surfaces Immersed in Aqueous Solutions," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.
33. R. J. Archer, "The Measurement of the Adsorption of Vapors and Gases on Silicon by the Method of Ellipsometry," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.
34. R. R. Stromberg, W. Passaglia, and D. J. Tutas, "The Study of Adsorption from Solution by the Method of Ellipsometry," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.
35. J. Kruger and W. J. Ambs, "Optical Measurements on Thin Films of Condensed Gases at Low Temperatures," *Jour. Opt. Soc. Amer.*, Vol. 49, p. 1195, Dec. 1959.

36. M. A. Barrett, "Optical Study of the Formation and Stability of Anodic Films on Aluminum," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

37. A. K. N. Reddy, M. A. V. Devanathan, and J. O'M. Bockris, "Ellipsometry in Electrochemical Studies," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

38. F. L. McCrackin and J. Colson, "Computational Techniques for the Use of the Exact Drude Equations in Reflection Problems," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

39. F. Abelès, "Optical Properties of Inhomogeneous Films, Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films," Washington, D. C., Sept. 1963.

40. I. V. Cathcart and G. F. Petersen, "Studies on Thin Oxide Films on Copper Crystals Using a Polarizing Spectrometer," Symposium on the Ellipsometer and Its Use in the Measurement of Surfaces and Thin Films, Washington, D. C., Sept. 1963.

41. W. T. Pimbley, "Optical Properties of Thin Adsorbed Films on Metal Surfaces Undergoing Stress," *Jour. Opt. Soc. Amer.*, Vol. 52, p. 1410, Dec. 1962.

42. E. David, "Deutung der Anomalien der optischen Konstanten dünner Metallschichten," *Zeit. für Physik*, Vol. 114, p. 389, 1939.

43. C. v. Fragstein and H. Römer, "Über die Anomalie der optischen Konstanten," *Zeit. für Physik*, Vol. 151, p. 54, 1958.

INTRODUCTORY STATISTICS AND SAMPLING CONCEPTS APPLIED TO RADAR EVALUATION

By

Remo J. D'Ortenzio

RCA Missile and Surface Radar Division
Moorestown, N. J.

Summary—*The analysis of any type of digital data requires a working knowledge of statistics and sampling theory. This paper presents some fundamentals that are particularly useful in analyzing radar data. Included are definitions, least-mean-squares curve-fitting equations, sampling and bandwidth considerations, data smoothing, and basic power spectrum concepts. Portions of the contents are also applicable to many nonradar situations.*

Introduction

THIS paper is intended to serve as a guide for specifying the contents of radar test data packages and the data analysis methods to be used in evaluating them. The material presented here is not intended to provide a rigorous theoretical background in statistics. Rather, it is meant to give a practical understanding of some of the items to be considered when collecting and analyzing digital data. Although the discussions and examples are primarily associated with radar, the ideas presented can be applied to a large variety of situations possessing similar or analogous properties.

In most radars digital data is recorded at a fixed sampling frequency. Some typical questions that arise concerning the processing and analysis of this data are:

(a) How much data should be called for? (length of test run)
(b) Should every data point be processed or should some be ignored to minimize data analysis and computer time?
(c) How valid will the results be?
(d) How are curve fits utilized to determine radar precision and accuracy?
(e) How does one specify the polynomial order of the curve fit?
(f) What are the statistical effects of averaging or smoothing data points that are not independent?
(g) What does digital sampling do to the spectrum of the parameter being measured?

An attempt has been made in this paper to answer these questions and others without introducing excessive mathematical complexities.

It is hoped that the contents will serve as a foundation on which the reader can build by consulting more-elegant treatments of this subject in the literature.

DEFINITIONS

Consider a parameter measured n times; the values obtained are $r_1, r_2, r_3 \cdots r_n$. This set of data points $\{r_i\}$ defines a random variable R whose statistics are expressed as follows:

Mean

The mean, average value, or expected value of R is denoted by $\overline{R}$ or $E(R)$ and is given by

$$\overline{R} = E(R) = \frac{1}{n} \sum_{i=1}^{n} r_i. \tag{1}$$

Standard Deviation

The standard deviation of R is a measure of the "scatter" of R about its mean, $\overline{R}$, and is denoted by σ_R, where

$$\sigma_R = \sqrt{\frac{1}{n} \sum_{1}^{n} (r_i - \overline{R})^2}. \tag{2}$$

Variance

The variance of R, called $\mathrm{Var}(R)$, is simply the square of the standard deviation σ_R

$$\mathrm{Var}(R) = \sigma_R^2 = \frac{1}{n} \sum_{1}^{n} (r_i - \overline{R})^2. \tag{3}$$

$\mathrm{Var}(R)$ is actually the average value of the square of the deviation of each data point from its mean, and is sometimes represented as shown in Equation (4).

$$\mathrm{Var}(R) = E(R - \overline{R})^2 = \overline{(R - \overline{R})^2}. \tag{4}$$

By expanding $(R - \overline{R})^2$ (see Appendix I), $\mathrm{Var}(R)$ can also be expressed as

$$\mathrm{Var}(R) = E(R^2) - [E(R)]^2 = \overline{R^2} - (\overline{R})^2. \tag{5}$$

Equation (5) states that $\mathrm{Var}(R)$ can also be considered as the average value of the square of the data points minus the square of their average value.

Root-Mean-Square Value

The root-mean-square (r-m-s) *value* ρ of a set of data is strictly defined as

$$\rho = \sqrt{\frac{1}{n}\sum_{1}^{n} r_i^2} = \sqrt{\overline{R^2}}. \tag{6}$$

This particular parameter is very seldom directly used to characterize a set of data. It gives a measure of "scatter" of the data points from a *zero reference*. The more widely used term is the root-mean-square *error* described below.

Root-Mean-Square Error

The root-mean-square (r-m-s) *error* α_P is the measure of the "scatter" of a set of data about some reference point P. For the completely general case, *the r-m-s error of the data points about* P is given by

$$\alpha_P = \sqrt{\frac{1}{n}\sum_{1}^{n} (r_i - P)^2}. \tag{7}$$

Note that if P is the mean value of the data, then α_P is identical to σ_R. Stated another way, *the r-m-s error of a set of data about its mean is equal to the standard deviation of the data.*

If, on the other hand, P is set equal to zero, then α_P becomes identical to the data r-m-s *value* ρ.

The more general case occurs when P is neither $\overline{R}$ nor zero. Such a case may arise for example when a set of radar data is taken to determine the ability of the radar to measure the range of a boresight tower whose exact range is known. If the true or surveyed range is R_T and the set $\{r_i\}$ represents the radar measurements, the total r-m-s error α_P of the radar is given by Equation (7) with P set equal to R_T. A typical set of range data obtained in the presence of noise is shown in Figure 1.

Note that the average of the data points, $\overline{R}$, generally differs from the true range, R_T. This gives rise to a bias error B defined by

$$B = \overline{R} - R_T. \tag{8}$$

The *r-m-s error* of the measurements about R_T is

$$\alpha_{RT} = \sqrt{\frac{1}{n}\sum_{1}^{n} (r_i - R_T)^2}, \tag{9}$$

and the *r-m-s error* about $\overline{R}$ (or the standard deviation) is

$$\alpha_{\overline{R}} = \sigma_R = \sqrt{\frac{1}{n}\sum_{1}^{n} (r_i - \overline{R})^2}. \tag{10}$$

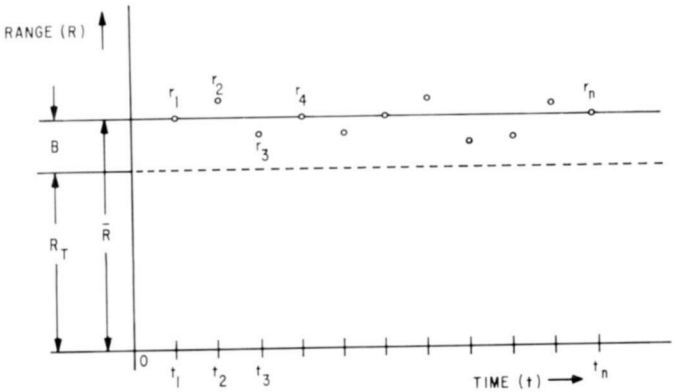


Fig. 1—Typical range measurements of boresight tower.

Combining Equations (1), (8), (9), and (10) shows that

$$\alpha_{RT}^2 = \sigma_R^2 + B^2. \tag{11}$$

The complete derivation of Equation (11) is shown in Appendix II.

The interesting point to observe in Equation (11) is that the total r-m-s error α_{RT} can be broken down into two components—the standard deviation σ_R and the bias B. More is said about these in relation to the definitions of precision and accuracy that follow.

Precision and Accuracy

The *International Dictionary of Physics and Electronics* defines the *precision* of a measuring device as the degree of reproducibility among a group of independent measurements of the same true value made under specified conditions. The *accuracy* is the quality of correctness or freedom from error.

Barton[1] applies these definitions such that *the r-m-s radar range accuracy is the total r-m-s error of a set of data points* $\{r_i\}$ *with respect to the true range* R_T. This includes both bias errors and noise errors and is mathematically equivalent to α_{RT} as defined in Equation (9).

Precision is used as a measure of noise error only, since noise interferes with the ability of the radar to reproduce a range measurement. Thus *r-m-s precision is defined as the r-m-s error of the readings about their mean,* $\overline{R}$. This is equivalent to the value σ_R defined in Equation (10).

Equation (11) shows that the r-m-s precision and accuracy differ because of the bias error B. This equation demonstrates two interesting facts. First, the square of the r-m-s accuracy is the sum of the squares of the r-m-s precision and bias errors and secondly, under certain "low noise" conditions, a set of radar data can be very precise (low σ_R) but highly inaccurate (large bias).

Least-Mean-Squares Curve Fitting

Basic Concepts

The true value of a parameter frequently varies with time. For the case of radar range measurements, this situation would arise when making range measurements on a target moving with respect to the radar. On other occasions the true value might remain constant, but the readings of the measuring device may slowly drift with time. This drift might be caused by a bad component in the measuring equipment or it could conceivably be part of a low-frequency oscillation that looks like drift when examined over a relatively short period of time.

For cases such as these, calculation of the precision and accuracy of the measuring device as described in the preceding section would lead to erroneous results. Suppose, for example, that a radar was making range measurements on a balloon moving slowly away from the radar. A typical set of such data is shown in Figure 2.

The mean ($\overline{R}$) of this data would not be too meaningful by itself. At best, it would be an estimate of the average range of the balloon between times t_1 and t_n. In order to handle this type of data, some sort of curve fit is required.

If the value of the true range is known to vary approximately linearly with time, a straight-line curve fit should be specified. The most common type of curve fit used is that from the method of least mean squares.

The equation of the line representing the least-mean-squares curve

fit to Figure 2 can be expressed as

$$R_c(t) = A_0 + A_1 t, \tag{12}$$

where A_0 and A_1 are chosen such that the residual variance δ^2 of the data points with respect to the derived curve is a minimum. The residual variance δ^2 is defined as

$$\delta^2 = \frac{1}{n} \sum_{i=1}^{n} [r_i - R_c(t_i)]^2, \tag{13}$$

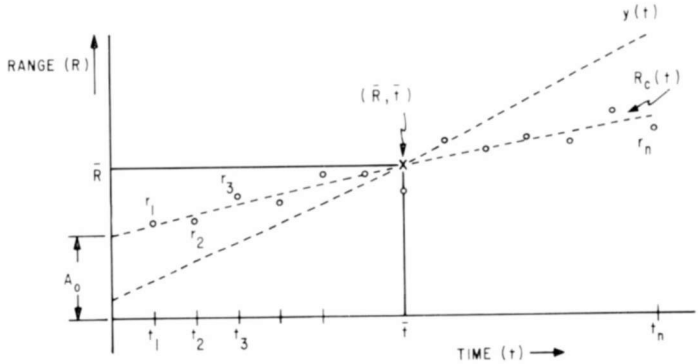


Fig. 2—Range measurements during a balloon track.

where the quantity $r_i - R_c(t_i)$ is called a *data point residual* and is denoted by d_i.

$$d_i = r_i - R_c(t_i). \tag{14}$$

Stated another way, the straight line is derived so that the variance (or mean square) of the residuals is a minimum. The quantity δ is the residual r-m-s error, but is usually referred to simply as the r-m-s error.

When A_0 and A_1 are properly calculated the average residual is zero. Thus, for a least-mean-squares curve fit (of any order polynomial)

$$\frac{1}{n} \sum_{1}^{n} d_i = \frac{1}{n} \sum_{1}^{n} [r_i - R_c(t_i)] = 0. \tag{15}$$

The more general expression for a curve fit is given by the k^{th} order polynomial;

$$R_c(t) = A_0 + A_1 t + A_2 t^2 + A_3 t^3 \cdots + A_k t^k. \tag{16}$$

The order of the polynomial specified for curve fits must be based on some a priori knowledge of the data. For example, if a missile re-entry motion is known to contain radial acceleration and jerk components, at least a third-order polynomial would be required for range-data curve fitting. Specifying too high an order polynomial is undesirable because in the limit, the fitted curve will theoretically be connecting each data point, thereby reducing the residual variance toward zero and obscuring the radar range tracking errors. On the other hand, specifying a linear curve fit to range data obtained on a radially accelerating target would tend to exaggerate the range tracking errors. However, it would be quite acceptable to fit a linear curve over a small segment of data extracted from a large data record whose overall range versus time varied—for example in a quadratic or cubic fashion—provided that the *segment* being analyzed was approximately linear. Thus, the length of the data interval is another factor to consider when selecting the order of polynomial.

In actual practice, when computer facilities are used, computational difficulties will set in long before the order of the polynomial approaches the number of data points. As the order of the polynomial is increased, one may actually find the residual variance reaches some minimum and thereafter begins to increase. This will occur when the computational errors become significantly larger than the errors in the data being analyzed. Most data reduction centers have pre-programmed routines for least-mean-square curve fits. They also have programs available for determining the appropriate order of the polynomial to be used when there is no a priori knowledge of how the data behaves. Generally, it is not practical to fit polynomials above orders of seven or eight without special procedures.

Calculation of the r-m-s error of a set of range measurements made on a boresight tower has been described. In this calculation a linear curve fit would provide useful information about the nature of the errors even though the target is known to be fixed in range. For example, if the calculated r-m-s error is large, it could be caused by excessive noise or by a drift that occurred during the test run. Furthermore, as previously mentioned, the drift could actually be part of a low-frequency oscillation. In any case, a linear curve fit to the data points would provide much in the way of diagnostic information. If

the slope of the line is zero, then one can be sure there was no drift. On the other hand, if the slope is not zero, there is an indication that some drift did occur. One method commonly used to determine whether any drifts are cyclic is to take data on a known target over a relatively long time interval and run a power-spectrum analysis on the data points. This is discussed later.

It should be noted that curve fitting is by no means restricted to the principal of least mean squares. The reasons for the widespread use of this method are its direct association with the concept of variance as a measure of data dispersion and its mathematical simplicity as compared to other methods. One could, for example, specify a curve fit that would minimize the maximum of the absolute values of the residuals, or the average of the absolute values of the residuals, or the average of the residuals raised to an even-integer power. Each of these methods, however, presents excessive mathematical complexities and they are generally avoided except for special situations.

Figure 2 illustrates why one would not attempt to fit a curve to minimize the average of the residuals raised to the first power (or any odd-integer power). Consider the line $y(t)$. By simple inspection one observes that it represents a poor fit to the data. However, it is very possible that $y(t)$ possesses the properties of fitting the data with an average residual of zero. It is possible, in fact, to have an infinite number of lines of different slopes, each of which has a zero average residual.

Linear Least-Mean-Squares Curve Fit Equations

Given a set of data points $\{r_i\}$, the following equation is to be derived:

$$R_c(t) = A_0 + A_1 t, \tag{12}$$

such that the residual variance,

$$\delta^2 = \frac{1}{n} \sum_{i=1}^{n} [r_i - (A_0 + A_1 t_i)]^2, \tag{17}$$

is minimized. The problem is thus reduced to finding the values of A_0 and A_1 that minimize the residual variance δ^2.

The necessary condition for δ^2 to be a minimum is that the partial derivatives of δ^2 with respect to A_0 and A_1 be zero.

$$\frac{\partial(\delta^2)}{\partial A_0} = \frac{\partial(\delta^2)}{\partial A_1} = 0. \tag{18}$$

Operating on Equation (17) as indicated in Equation (18) results in two linear equations with two unknowns. It appears at first glance that solving for A_0 and A_1 could theoretically result in either a maximum or minimum for δ^2. However, each term in Equation (17) is non-negative; therefore δ^2 must have a non-negative minimum. Solving the two linear equations for A_0 and A_1 gives only one solution, and hence this must be the solution for minimizing δ^2. Therefore the criteria of Equation (18) represents both necessary and sufficient conditions for minimizing δ^2. The development of the equations for determining the values of A_0 and A_1 follows:

$$\frac{\partial(\delta^2)}{\partial A_0} = -\frac{2}{n} \sum [r_i - (A_0 + A_1 t_i)], \tag{19}$$

Equating the right-hand of Equation (19) to zero, dividing through by $-(2/n)$, expanding, and rearranging yields

$$\sum r_i = nA_0 + A_1 \sum t_i. \tag{20}$$

Also,

$$\begin{aligned} \frac{\partial(\delta)^2}{\partial A_1} &= -\frac{2}{n} \sum [r_i - (A_0 + A_1 t_i)] [-t_i] \\ &= \frac{2}{n} \sum [r_i t_i - A_0 t_i - A_1 t_i^2]. \end{aligned} \tag{21}$$

Equating the right-hand side of Equation (21) to zero, dividing through by $2/n$, expanding, and rearranging yields

$$\sum r_i t_i = A_0 \sum t_i + A_1 \sum t_i^2, \tag{22}$$

where all summations extend from $i = 1$ to $i = n$. Solving Equations (20) and (22) simultaneously yields the desired values of A_0 and A_1. Note that if the data is made symmetrical in time about $t = 0$, so that $\sum t_i = 0$, then the solution is given simply by

$$A_0 = \frac{1}{n} \sum r_i, \tag{23}$$

$$A_1 = \frac{\sum r_i t_i}{\sum t_i^2}. \tag{24}$$

Second-Order Polynomial Least-Mean-Squares Curve-Fit Equations

Given a set of data points $\{r_i\}$, the following equation is to be derived:

$$R_c(t) = A_0 + A_1 t + A_2 t^2. \tag{25}$$

The function to be minimized is given by

$$\delta^2 = \frac{1}{n} \sum_{i=1}^{n} [r_i - (A_0 + A_1 t + A_2 t^2)]^2. \tag{26}$$

The function δ^2 is minimized if

$$\frac{\partial(\delta)^2}{\partial A_0} = \frac{\partial(\delta^2)}{\partial A_1} = \frac{\partial(\delta^2)}{\partial A_2} = 0. \tag{27}$$

Performing the partial differentiations indicated by Equation (27) and simplifying yields the three simultaneous equations which must be solved for A_0, A_1 and A_2;

$$\sum r_i t_i^2 = A_0 \sum t_i^4 + A_1 \sum t_i^3 + A_2 \sum t_i^2, \tag{28}$$

$$\sum r_i t_i = A_0 \sum t_i^3 + A_1 \sum t_i^2 + A_2 \sum t_i, \tag{29}$$

$$\sum r_i = A_0 \sum t_i^2 + A_1 \sum t_i + A_2 n, \tag{30}$$

where all summations extend from $i = 1$ to $i = n$.

Note again that considerable simplification in solving the above can be achieved if the data is made symmetrical in t_i so that $\sum t_i = \sum t_i^3 = 0$.

Sampling Considerations

Suppose one wishes to estimate the mean, $\overline{X}$, and standard deviation, σ, of an infinite population by means of sampling. A sample of size n is drawn and its mean, U, and standard deviation, S, are calculated. These calculated sample values approach the actual values only

as the sample size n approaches infinity (provided of course that no bias exists in the measuring equipment). However, by choosing a finite sample size n, certain statistical inferences can be made about the relationship of the sample characteristics to those of the population.

When n is very small (of the order of 30 or less) the relationships between the sample values U and S, and the actual values $\overline{X}$ and σ, are quite subtle and will not be considered here. Fortunately most radar data to be analyzed consists of numbers of independent data points much larger than 30. Furthermore the data is usually distributed in a normal or near normal form (especially when taken under conditions of large signal-to-noise ratios). These two factors afford a considerable simplification in relating the estimates to the actual and in determining the validity or the "confidence level" of the estimates. (The concept of a normal distribution is described in Appendix III.)

Sample Size—Confidence Levels

If it is assumed that the population is normally (or near normally) distributed and the sample size, n, is larger than 30, then the following results can be utilized with negligible error.

For a sample of size n there is a certain probability, or confidence level, associated with how well the estimates U and S compare with the "true" values $\overline{X}$ and σ. In a radar situation the number of independent data points required to estimate a particular parameter is usually a compromise between the confidence level desired and the practicality of processing a large set of data.

In Figure 3 the per cent error, E, in estimating σ versus the sample size n, is plotted with confidence level as a parameter. Table I relates the error in estimating $\overline{X}$ with the sample size n for various confidence levels. The remainder of this section is directed toward the interpretation and use of the results shown in Figure 3 and Table I rather than the arguments used in obtaining them. The development of Figure 3 and Table I is described in Appendix IV and References (2) and (6).

Figure 3 can best be interpreted by stating that if n samples are taken and their standard deviation, S, is used to estimate σ, there is a certain probability (or confidence level) that S will be within $\pm E\%$ of σ. For example, with $n = 100$ there is a 90.0% probability that the calculated value of S will be within $\pm 11.5\%$ of σ. In equation form

$$\text{Probability} \left\{ \sigma(1 - 0.115) \leq S \leq \sigma(1 + 0.115) \right\} = 0.900, \tag{32}$$

or

$$\text{Probability} \left\{ \frac{S}{1+0.115} \leq \sigma \leq \frac{S}{1-0.115} \right\} = 0.900. \qquad (32a)$$

Thus there is a 90.0% confidence that the true value of σ lies between $S/1.115$ and $S/0.885$.

Table I—Estimation of $\overline{X}$

Confidence Level (%)	*Maximum Difference* between* $\overline{X}$ *and* U
99.7	$\pm 3.00 \dfrac{\sigma}{\sqrt{n}}$
95.5	$\pm 2.00 \dfrac{\sigma}{\sqrt{n}}$
90.0	$\pm 1.65 \dfrac{\sigma}{\sqrt{n}}$
80.0	$\pm 1.28 \dfrac{\sigma}{\sqrt{n}}$
68.3	$\pm 1.00 \dfrac{\sigma}{\sqrt{n}}$

* The expressions shown apply when the population standard deviation σ is known. When σ is not known, the sample standard deviation S can be used in place of σ with negligible error provided the sample size n is large (greater than 30).

Table I is used to determine how good a measure of $\overline{X}$ is obtained by using the sample mean, U, as an estimate. If the true σ is known, the expressions in Table I are used to determine the confidence levels of the estimate. However, if the value of σ is not known (as is usually the case) the sample standard deviation, S, must be calculated and used in its place. When the sample size n is greater than 30, the substitution of S for σ in the expressions of Table I results in a negligible error. Suppose that $n = 100$ and σ is not known; Table I shows that if the estimate of $\overline{X}$ is U and that of σ is S, then

$$\text{Probability} \left\{ \overline{X} - \frac{3S}{\sqrt{100}} \leq U \leq \overline{X} + \frac{3S}{\sqrt{100}} \right\} \approx 0.997, \qquad (33)$$

or

$$\text{Probability} \left\{ U - \frac{3S}{\sqrt{100}} \leq \overline{X} \leq U + \frac{3S}{\sqrt{100}} \right\} \approx 0.997. \tag{33a}$$

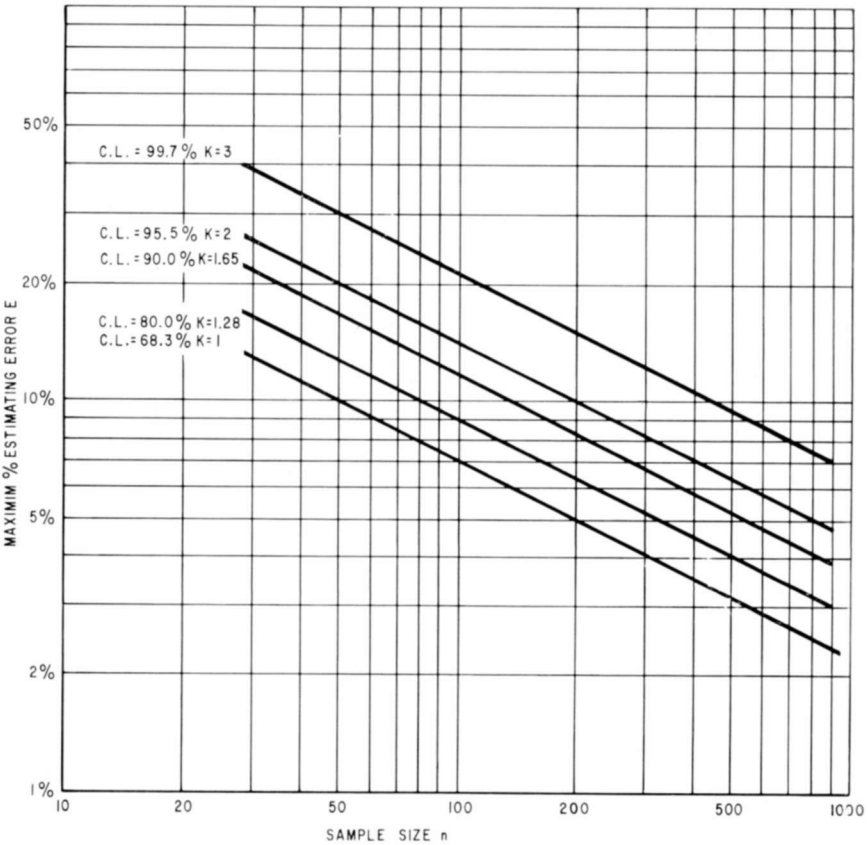


Fig. 3—Maximum per cent error in estimating the σ of a normal population versus sample size for various confidence levels (C.L.). (Equations of curves shown: $E = 100\,K/\sqrt{2n}$; values of K shown.)

Effects of Bandwidth and Sampling on R-M-S Error Calculations

Marcum[3] states that if white Gaussian noise is passed through a narrow-band i-f filter whose overall 3-decibel bandwidth is Δf, it is probably a good approximation to assume that samples of the noise envelope taken $1/\Delta f$ seconds apart are statistically independent. This of course serves only as an approximate quantity which changes for

different filter shapes. If the filtering is done at the video level instead of i-f and the low-pass 3-decibel video bandwidth is B, statistical independence is assumed every $1/2B$ seconds.

Suppose a set of radar measurements is bandwidth limited by a 10 cps i-f filter and the radar computer has a fixed sampling frequency of 100 pps. For this special case, data points spaced 0.1 second apart (every tenth data point) are assumed statistically independent.

Using these assumptions, one can determine the length of data record required to obtain a certain confidence level with a certain per cent error. For example, if a 99.7 per cent probability is desired that a measured r-m-s error does not differ from its true value by more than 10 per cent, then Figure 3 shows that about 450 *independent* data points are required. Therefore, for a 10 cps bandwidth, 45 seconds of data would be required (10 independent samples per second). In order to minimize computer time in determining r-m-s error one should select every tenth point in the data. If all the data points are used instead of every tenth, there would be a predictable improvement in the confidence level of the measurement, but this improvement would generally be small. One significant factor to emphasize, however, is that using all the data points can only improve the estimate; it will never make the results less valid. The only disadvantage is that the additional processing time required may not be warranted.

Data Smoothing

Oftentimes it is desirable to average groups of adjacent data points to smooth out some of the effects of noise. Suppose a set of data points r_1, r_2, $r_3 \cdots r_n$ are known to be distributed in a Gaussian or near-Gaussian form. Next assume that smoothing is accomplished by dividing the data points into adjacent groups of ten, the first group being comprised of r_1 through r_{10}, the second of r_{11} through r_{20}, etc. If the averages of each group of ten are calculated to be $\overline{X}_1$, $\overline{X}_2 \cdots \overline{X}_{n/10}$ these can be looked at as a new set of data points that are "smooth" with respect to the original points.

If the points $\{r_i\}$ are independent, then the standard deviation of the set $\{\overline{X}_j\}$ will be approximately $1/\sqrt{n}$ times the standard deviation of the original data. For the particular case chosen, n is ten and thus the standard deviation of the smoothed data is $1/\sqrt{10}$ or about 0.3 times that of the original data.

If the points $\{r_i\}$ are not independent, the analysis required is elaborate. Appendix V shows how correlation techniques can be used to calculate (under specific conditions of filter shape, bandwidth, and sampling frequency) the standard deviation of a set of smoothed data

points that are dependent. Averaging groups of ten points from 100 pps data obtained from a 5 cps bandwidth, single-tuned low-pass filter reduces the standard deviation to 0.67 of the value calculated for the "raw" data.

The resulting reductions in standard deviations can be interpreted in terms of a filtering process. If the raw data is obtained from a low-pass filter of bandwidth B_1 and has a standard deviation σ_1, then the *effective* value B_2 of the filter bandwidth which would have given rise to the same standard deviation σ_2 obtained by smoothing is

$$B_2 = \frac{\sigma_2^2}{\sigma_1^2} B_1. \tag{34}$$

Averaging ten independent data points, for example, would give a σ_2 of $1/\sqrt{10}$ times σ_1. The effective bandwidth B_2 of this smoothing process is thus given by

$$B_2 = \left(\frac{\sigma_1}{\sqrt{10}} \right)^2 \frac{1}{\sigma_1^2} B_1 = \frac{B_1}{10}. \tag{35}$$

It must be noted, however, that Equation (34) is valid only if the degree of smoothing is considerable. This situation simulates a white-noise input to a narrow-band filter, from which Equation (34) is derived.

All the discussions in this section are also directly applicable to the case where raw data points r_1 through r_{10} are averaged first, then points r_2 through r_{11}, then points r_3 through r_{12}, etc. The resulting smoothed data will have the same standard deviation as did the averages of discrete groups of ten. In fact, for a given number of raw data points the confidence level of estimating r-m-s error or standard deviation with this type of smoothing would be greater than for the discrete case. The only disadvantage to this technique is the additional processing time required which may not be justified in terms of the increase in confidence level.

Power Spectrum

Definitions and Basic Concepts

The power content of any voltage waveform is defined as the power that the voltage would develop across a resistance of one ohm. If white noise having W watts per cycle per second or W watts per unit bandwidth is applied to a filter of narrow bandwidth whose transfer func-

tion is $G(f)$, the *power spectrum or spectral density* $P(f)$ of the waveform out of the filter is given by

$$P(f) = W|G(f)|^2. \tag{36}$$

$P(f)$ is basically a measure of the power per unit bandwidth. A plot of $P(f)$ versus f would show the relative contributions of different frequency bands to the total power. The area under the $P(f)$ versus f curve represents the total average power that the waveform would develop across a one-ohm resistance.

$$P_{av} = \int_0^\infty P(f)\, df = \int_0^\infty W|G(f)|^2\, df. \tag{37}$$

Note that P_{av} is also the square of the r-m-s value of the waveform. Thus the area under the data power spectrum curve should be numerically equal to the square of the data r-m-s value. For example, if radar range measurements are taken during a balloon track, it may be desirable to curve fit the data and calculate residuals between the curve fit and data points. The r-m-s tracking error is then defined as the r-m-s value of the residuals. Suppose that this gives rise to some value σ_1. A power spectrum of these same residuals would give a plot of their relative frequency distribution. The area under this power spectrum curve should theoretically be equal to σ_1^2. However, power spectrum of sampled data cannot generally be calculated as accurately as r-m-s errors; thus there would be some discrepancies between the results in a practical case. One method for calculating the power spectrum of sampled data is described in Appendix VI.

Equation (36) suggests that if a set of radar measurements is band limited by some filter $G(f)$, the power spectrum of the data will in general tend to reproduce the absolute square of the filter response. The resulting plot of $P(f)$ versus f can be analyzed to observe any deviations or sharp peaks that may indicate undesired oscillations.

Spectrum of Sampled Data

If some continuous data has a certain power spectrum characteristic that is to be determined after sampling, the question arises of how the power spectrum of the sampled data is related to the continuous data. (The discussion is restricted to sampled data that is equispaced in time.) A general answer to the above question is that sampling introduces sideband frequencies into the spectrum which

cause the original power spectrum of the continuous data to be reproduced in shape, at frequencies centered about multiples of the sampling frequency.

To understand this phenomenon, it is desirable to examine a situation where the continuous data is sampled for a finite time τ at intervals of T seconds by what is called the unit sampling function $u(\tau,t)$ shown in Figure 4.

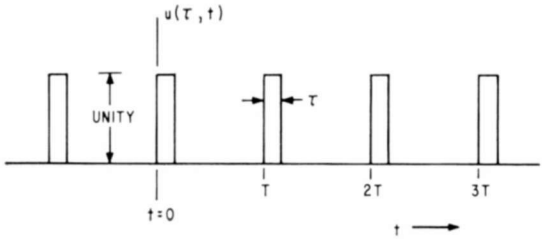


Fig. 4—The unit sampling function $u(\tau,t)$.

The Fourier series representation of this function is given by

$$u(\tau,t) = \frac{\tau}{T} + \frac{2\tau}{T}\sum_{n=1}^{\infty}\left\{\frac{\sin\dfrac{n\pi\tau}{T}}{\dfrac{n\pi\tau}{T}}\cos n\left(\frac{2\pi}{T}t - \frac{\pi\tau}{T}\right)\right\} \tag{38}$$

where the quantity $n\pi\tau/T$ is the phase angle $n\theta_s$ of the n^{th} harmonic and $(2\tau/T)\,(\sin n\pi\tau/T)/(n\pi\tau/T)$ is the amplitude of the n^{th} harmonic.

If some continuous waveform $x(t)$ is sampled by the unit sampling function $u(\tau,t)$ as in Figure 5, the output $y(t)$ is given by the product of $u(\tau,t)$ and $x(t)$ as shown in Equation (39).

$$y(t) = u(\tau,t)\,x(t) = \frac{\tau}{T}x(t) + \frac{2\tau}{T}\sum_{n=1}^{\infty}\left\{\left[\frac{\sin\dfrac{n\pi\tau}{T}}{\dfrac{n\pi\tau}{T}}\cos n\left(\frac{2\pi}{T}t - \frac{\pi\tau}{T}\right)\right]x(t)\right\} \tag{39}$$

In order to analyze the effects of sampling on the frequency spectrum, it is convenient to consider some $x(t)$ waveform that contains only one frequency component f_0.

Let
$$x(t) = V \cos(2\pi f_0 t + \theta_0), \tag{40}$$

$$f_s = 1/T, \tag{41}$$

and
$$\theta_s = \frac{\pi\tau}{T}. \tag{42}$$

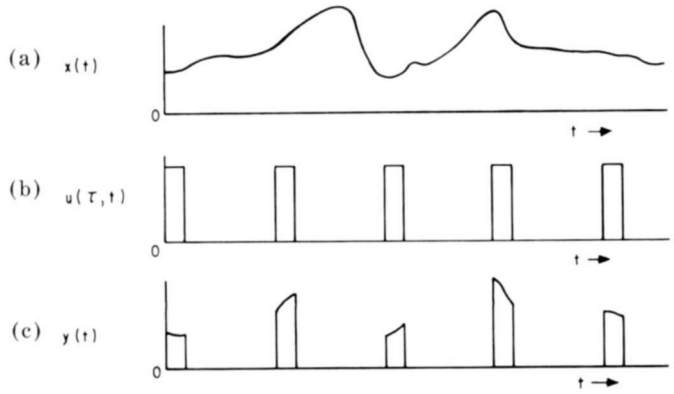


Fig. 5—(a) A continuous waveform $x(t)$, (b) the unit sampling function $u(\tau,t)$, and (c) the sampled waveform $y(t)$.

Then Equation (39) becomes

$$y(t) = \frac{\tau}{T} V \cos(2\pi f_0 t + \theta_0) + \frac{2\tau}{T} V \sum_{n=1}^{\infty} \left\{ \frac{\sin n\pi f_s \tau}{n\pi f_s \tau} \cos n(2\pi f_s t - \theta_s) \cos(2\pi f_0 t + \theta_0) \right\}. \tag{43}$$

Using the identity

$$\cos A \cos B \equiv 1/2\,[\cos(A + B) + \cos(A - B)], \tag{44}$$

Equation (43) becomes

$$y(t) = \frac{\tau}{T} V \cos(2\pi f_0 t + \theta_0) + \frac{\tau}{T} V \sum_{n=1}^{\infty} \frac{\sin n\pi f_s \tau}{n\pi f_s \tau} \Big\{ \cos[2\pi (nf_s + f_0)t - n\theta_s + \theta_0] + \cos[2\pi (nf_s - f_0)t - n\theta_s - \theta_0] \Big\}. \quad (45)$$

Equation (45) shows that the sampling process reproduces the original frequency f_0 (attenuated by a factor τ/T) and also introduces sideband frequencies $nf_s \pm f_0$ whose magnitudes have a $(\sin x)/x$ gross spectrum (see Figure 6).

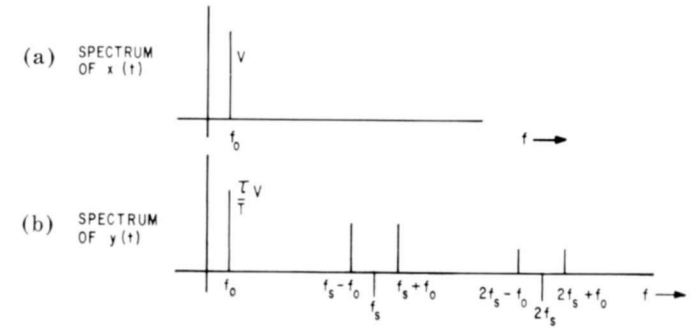


Fig. 6—Spectrum (a) of $x(t) = V\cos(2\pi f_0 t + \theta_0)$ and (b) of $y(t)$.

When the original signal $x(t)$ is a composite periodic wave consisting of j frequency components ranging from zero to f_0 cps, the frequency spectrum of $x(t)$ would contain j vertical lines. The height of these vertical lines would be equal to the magnitude of the corresponding frequency component. Thus for this case the frequency spectrum of the sampled signal would contain the lines of the original spectrum plus sideband frequencies at $nf_s \pm f$ where f is the frequency of any single line in the original spectrum.

Now, if the composite signal $x(t)$ contains an infinite number of frequency components within its frequency band, then in the limit, $x(t)$ becomes nonperiodic and its frequency spectrum approaches a continuous curve. The spectrum $|g_x(f)|$ of such a signal and the corresponding spectrum $|g_y(f)|$ of the sampled signal are shown in Figure 7.

Figure 7 clearly illustrates that *the sampling frequency f_s must be at least twice the highest significant frequency component in the original spectrum.* If f_s were less than $2f_0$ in Figure 7, the sideband

lobes of the sampled spectrum would distort the main lobe. One "rule of thumb" commonly used in practice is to choose f_s equal to or greater than $3f_0$.

Sometimes, when a power density spectrum is calculated from a set of residuals derived from a least-mean-squares curve fit, excessive spectral components may appear in the region of zero frequency. One possible cause of this phenomenon is the use of a polynomial curve fit

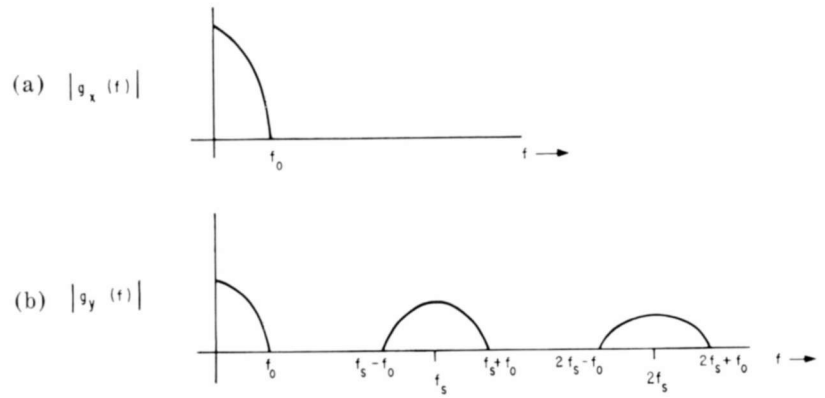


Fig. 7—Frequency spectrum (a) of a nonperiodic signal of a low-pass frequency band and (b) of the sampled signal.

whose order is too low. For example, if a parameter R actually varied quadratically with time, and a linear curve fit was specified, the residuals would tend to exhibit a cyclic variation with a fundamental frequency of about $1/T$ cps, where T is the total time duration of the data record (in seconds). In general, T is relatively large, and thus the effects show up in the vicinity of zero frequency.

Appendix I—Derivation of Variance Expression (Equation (5))

Given a random variable X with mean value $\overline{X}$, the variance of X is defined by

$$\text{Var}\,(X) = \overline{(X - X)^2}. \tag{46}$$

If the right-hand member is expanded, Equation (46) becomes

$$\text{Var}\,(X) = \overline{X^2 - 2\overline{X}X + (\overline{X})^2}. \tag{47}$$

The average value of a sum is equal to the sum of the averages; therefore

$$\text{Var } (X) = \overline{X^2} - \overline{2\bar{X}X} + \overline{(\bar{X})^2}. \tag{48}$$

The average value of a constant is the same constant, and the average value of a constant times a random variable is equal to the constant times the average value of the random variable. Therefore

$$\text{Var } (X) = \overline{X^2} - 2\bar{X}\bar{X} + (\bar{X})^2 = \overline{X^2} - 2(\bar{X})^2 + (\bar{X})^2 = \overline{X^2} - (\bar{X})^2. \tag{49}$$

Appendix II—Derivation of Total Mean-Square-Error Expression (Equation (11))

Refer to Equations (1), (8), (9) and (10). Note that

$$B^2 = (\bar{R} - R_T)^2 \tag{50}$$

$$\begin{aligned}
\alpha_{RT}^2 &= \frac{1}{n} \sum (r_i - R_T)^2 \\
&= \frac{1}{n} \sum (r_i^2 - 2r_i R_T + R_T^2) \\
&= \frac{1}{n} \sum r_i^2 - \frac{2R_T}{n} \sum r_i + \frac{1}{n} \sum R_T^2 \\
&= \frac{1}{n} \sum r_i^2 - 2R_T \bar{R} + R_T^2
\end{aligned} \tag{51}$$

$$\begin{aligned}
\sigma_R^2 &= \frac{1}{n} \sum (r_i - \bar{R})^2 \\
&= \frac{1}{n} \sum (r_i^2 - 2r_i \bar{R} + (\bar{R})^2) \\
&= \frac{1}{n} \sum r_i^2 - \frac{2\bar{R}}{n} \sum r_i + \frac{1}{n} \sum (\bar{R})^2 \\
&= \frac{1}{n} \sum r_i^2 - 2(\bar{R})^2 + (\bar{R})^2 \\
&= \frac{1}{n} \sum r_i^2 - (\bar{R})^2.
\end{aligned} \tag{52}$$

Subtracting Equation (52) from (51) yields

$$\alpha_{RT}^2 - \sigma_R^2 = (\overline{R})^2 - 2\overline{R}\,R_T + R_T^2$$
$$= (\overline{R} - R_T)^2. \tag{53}$$

Comparing Equations (50) and (53) shows that

$$B^2 = \alpha_{RT}^2 - \sigma_R^2 \tag{54}$$

and Equation (11) is proved.

Appendix III—The Normal Distribution

This section is intended to provide a simplified version of the concept of a normal (or Gaussian) distribution. The reader who is interested in a comprehensive treatment of the normal distribution, or probability distributions in general, can consult References (2), (6), and (7).

When a population or set of data is said to be normally distributed, the probabilistic relationships describing the statistics of the data are defined by the equations and curves shown in Figure 8.

Figure 8a, represented by $p(X)$, is the non-cumulative form of the normal distribution usually referred to as the normal probability *density* function, while Figure 8b is the cumulative form represented by $P(X)$ and called the normal probability *distribution* function. If one states that a population is normal with mean $\overline{X}$ and standard deviation σ and a single random sample X is drawn, the probability that the random variable X will be less than or equal to some value X_1, is given by the shaded area of Figure 8a. Expressed mathematically:

$$\text{Probability}\{-\infty \leq X \leq X_1\} = \int_{-\infty}^{X_1} p(X)\,dX$$
$$= \int_{-\infty}^{X_1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X-\overline{X})^2}{2\sigma^2}\right\} dX. \tag{55}$$

Unfortunately, the integral of Equation (55) cannot be expressed in

closed form. A series expansion and term by term integration of $p(X)$ is necessary to evaluate the integral. The definite integral from $-\infty$ to any value X_1 can then be computed to any desired accuracy by choosing an appropriate number of terms. The results so obtained constitute the cumulative form of the normal distribution $P(X)$, where

$$P(X) = \int_{-\infty}^{X} p(X)\,dX. \tag{56}$$

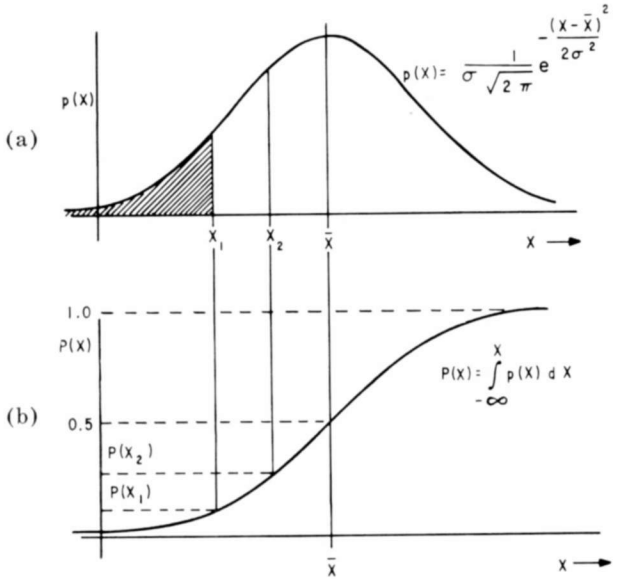


Fig. 8—Forms of the normal distribution; (a) noncumulative (probability density function), and (b) cumulative (probability distribution function).

Extensive tables, such as those appearing in Burington and May[2] have been computed for evaluating $P(X)$. The results are shown graphically in Figure 8b. Here the probability $P(X_1)$ that a random selection will be less than or equal to X_1 is read directly from the curve. It will be stated here, without proof, that $P(\infty)$, given by the total area under $p(X)$ is equal to unity. This is a necessary condition for $p(X)$ to be called a density function and for $P(X)$ to be termed a distribution function.

Given the same normal population, suppose it is desired to find the probability that a point selected at random lies between X_1 and X_2.

This is given graphically by the area under the curve of Figure 8a contained between X_1 and X_2. Expressed mathematically

$$\text{Probability}\{X_1 \leq X \leq X_2\} = \int_{X_1}^{X_2} p(X)\,dx \tag{57}$$

$$= \int_{-\infty}^{X_2} p(X)\,dx - \int_{-\infty}^{X_1} p(X)\,dx. \tag{58}$$

Table II—Probability that a Random Sample X Drawn from a Normal Population Will Lie within $\pm K\sigma$ of the Mean $\overline{X}$

K	Probability $\{\overline{X} - K\sigma \leq X \leq \overline{X} + K\sigma\}$
1.00	0.683
1.28	0.800
1.65	0.900
2.00	0.955
3.00	0.997

Thus the desired probability can be determined by reading from the curve at Figure 8b the values $P(X_2)$ and $P(X_1)$ and computing the difference $P(X_2) - P(X_1)$. Thus, Equation (58) can be re-written as

$$\text{Probability}\{X_1 \leq X \leq X_2\} = P(X_2) - P(X_1). \tag{59}$$

Oftentimes, it is desirable to know what the probability is that a point selected at random will lie within a certain number of standard deviation units from the mean $\overline{X}$. This probability is

$$\text{Probability}\{\overline{X} - K\sigma \leq X \leq \overline{X} + K\sigma\} = \int_{\overline{X}-K\sigma}^{\overline{X}+K\sigma} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X-\overline{X})^2}{2\sigma^2}\right\} dx. \tag{60}$$

The results for various values of K can be found in the tables of Burington and May.[2] A few representative values are given in Table II.

For example there is a 0.683 probability that a value selected at random from a normally distributed population with mean $\overline{X}$ and standard deviation σ will lie within $\pm 1.00\sigma$ from $\overline{X}$.

The importance of the normal distribution stems from the fact that a very large class of statistical data tends to be normal or near normal. Such is the case with radar data, where the presence of noise causes data to be scattered about its true value—usually in a normally distributed manner.

Appendix IV—Development of Confidence Levels

Suppose that a sample of size n is drawn from an infinite, normally distributed population whose mean is $\overline{X}$ and standard deviation is σ. Next, the sample mean u_1 and standard deviation s_1 are calculated. Taking a second sample, also of size n, would give rise to a sample mean and standard deviation u_2 and s_2. A third attempt would give u_3 and s_3 etc. If this procedure were repeated m times (m approaching infinity) the values $u_1, u_2, u_3 \cdots u_m$ and $s_1, s_2, s_3, \cdots s_m$ would represent two sets of random variables U and S characterized by their own probability distributions. If the sample size n is large (>30) certain useful approximations can be made regarding the distributions of U and S and their relationship to the population mean and standard deviation $\overline{X}$ and σ. Under the specified assumptions, the random variables U and S are each normally distributed as shown in Figures 9 and 10.

Confidence Levels in Estimating $\overline{X}$

Figure 9 shows that the mean of all the sample means $u_1, u_2, u_3 \cdots u_m$ is equal to the mean of the population $\overline{X}$ and the standard deviation of the sample means is given by $\sigma/\sqrt{n}$ (see References (6) or (7)). Thus, if one were interested in how well $\overline{X}$ could be estimated by drawing a sample of size n and calculating its mean U, the following reasoning would apply.

The probability that U lies between $\overline{X} - q$ and $\overline{X} + q$ is given by the shaded area of Figure 9. This area for any given q could be determined by using tables of the normal distribution found in Reference (2). For example, the probability (or area) that results when $q = 1.65$ standard deviations, is found to be 0.900. Thus there is a 90.0 per cent probability, or confidence level, that

$$\overline{X} - \frac{1.65\sigma}{\sqrt{n}} \leq U \leq \overline{X} + \frac{1.65\sigma}{\sqrt{n}}.$$

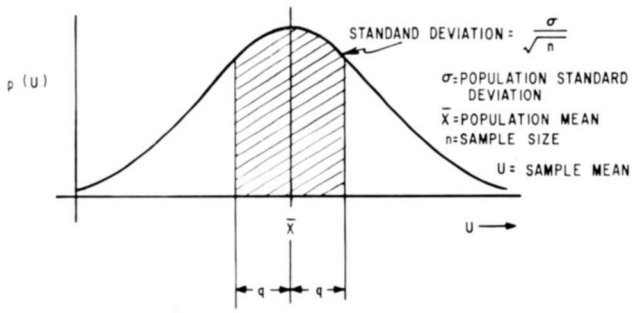


Fig. 9—The distribution of sample means u_1, u_2, u_3, $\cdots$ u_m (noncumulative form).

If q is set equal to K standard deviations, a corresponding probability (or area) can be similarly determined for any specified K. Some representative values of confidence levels for various K are shown in Table II.

In the previous paragraph, it was assumed that the σ of the population was known. If σ is not known (which is usually the case) the sample standard deviation S can be used in its place, *provided that n is large* (>30), with negligible error. When σ is not known and n is small, the determination of confidence levels for estimating $\overline{X}$ is more complex and will not be considered here.

Confidence Levels in Estimating σ

Figure 10 shows that the mean of all the sample standard deviations s_1, s_2, s_3, $\cdots$ s_m is approximately equal to the standard deviation of the population, σ, and that the standard deviation of the sample standard deviations is approximately $\sigma/\sqrt{2n}$. (When n is small, these approximations do not hold. For this case, the relationships are quite subtle and will not be considered here.)

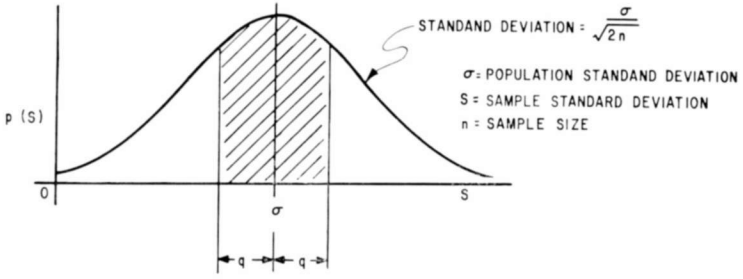


Fig. 10—The distribution of sample standard deviations s_1, s_2, s_3, $\cdots$ s_m (noncumulative form).

If S is used to estimate σ, the probability that $\sigma - q \leq S \leq \sigma + q$ would be given by the shaded area under the curve of Figure 10. For example, the probability that results for $q = 1.28$ standard deviations (found from the tables in Reference (2)) is 0.800. Thus

$$\text{Probability} \left\{ \sigma - \frac{1.28\sigma}{\sqrt{2n}} \leq S \leq \sigma + \frac{1.28\sigma}{\sqrt{2n}} \right\} = 0.800, \tag{61}$$

or

$$\text{Probability} \left\{ -\frac{1.28}{\sqrt{2n}} \leq \frac{S - \sigma}{\sigma} \leq + \frac{1.28}{\sqrt{2n}} \right\} = 0.800. \tag{62}$$

Equation (62) shows that there is an 80.0 per cent probability, or confidence level, that the per cent error in estimating σ is less than $128/\sqrt{2n}$ per cent. Figure 3 illustrates some relationships between the per cent error and sample size n for various confidence levels. The general relationship is given by

$$E(\%) = \frac{100K}{\sqrt{2n}}, \tag{63}$$

where the confidence level is related to K as shown in Figure 3.

Appendix V—Derivation of the Standard Deviation of a Set of Smoothed Data Using Correlation Methods

The specific case to be considered is the calculation of the standard deviation of the average of 10 successive dependent data points (spaced 0.01 second apart) at the output of a single-tuned low-pass filter whose 3-decibel bandwidth is 5 cps.

Suppose some radar parameter is sampled at 100 pps and the resultant samples are curve fitted with some least-mean-squares polynomial. The resulting residuals have a zero average value and a residual r-m-s error σ, whose magnitude is a measure of the "noise content" of the data.

One question that sometimes arises is, "If σ is the r-m-s value of the residuals, what would be the r-m-s value σ_1 of the averages of ten successive data points; that is to say, if the data points are divided into discrete groups of ten each and the average of each group of ten was calculated, what would be the predicted value of their standard deviation σ_1 in terms of the σ of the raw data?" The relationship between σ_1 and σ can be analytically determined using correlation

techniques, but only if the filter characteristic that gives rise to the output data is accurately known.

If the filter in question can be represented by an equivalent low-pass single-tuned filter whose 3-decibel bandwidth is B, the autocorrelation function $\phi(\tau)$ of the output residuals for a wide band noise input is given by

$$\phi(\tau) = \sigma^2 \exp\{-2\pi B\tau\} \tag{64}$$

where σ is the standard deviation or r-m-s value of the output residuals about the polynomial curve fit. (By a previous definition σ^2 is called the residual variance.)

This equation is derived by the autocorrelation relationship

$$\phi(\tau) = \int_{-\infty}^{\infty} S(f) \exp\{j2\pi f\tau\}\, df = \int_{-\infty}^{\infty} N_0 \,|G(f)|^2 \exp\{j2\pi f\tau\}\, df \tag{65}$$

where

$S(f)$ is the data's spectral density $= N_0 \,|G(f)|^2$,

$G(f)$ is the filter transfer function $= \dfrac{1}{1 + j\dfrac{f - f_0}{B/2}}$,

$N_0 =$ noise power per unit bandwidth at the filter input.

For the case in question, B is 5 cps and Equation (64) becomes

$$\phi(\tau) = \sigma^2 \exp\{-31.4\tau\}. \tag{66}$$

The physical interpretation of an autocorrelation function is that the *average product* of all points spaced τ seconds apart is given by $\phi(\tau)$.

In order to compute σ_1^2 by brute force, residual data points one through ten would be averaged and their square calculated. This would be repeated say r times for r successive groups of 10 data points. The average of the squares of the group averages would give the value of σ_1^2. Expressed mathematically,

$$\sigma_1^2 = \frac{1}{r}\left\{\left[\sum_{1}^{10} \frac{d_i}{10}\right]^2 + \left[\sum_{11}^{20} \frac{d_i}{10}\right]^2 \cdots + \left[\sum_{n-9}^{n} \frac{d_i}{10}\right]^2\right\}. \tag{67}$$

If the first bracketed term is expanded, the result is

$$\left[\sum_{1}^{10}\frac{d_i}{10}\right]^2=\frac{1}{100}\left\{\sum_{1}^{10}d_i^2+2\sum_{1}^{9}d_id_{i+1}+2\sum_{1}^{8}d_id_{i+2}\cdots+2\sum_{1}^{1}d_id_{i+9}\right\} \tag{68}$$

Repeating this for the other bracketed terms in Equation (67) gives similar results. If this is repeated r times the resulting expressions show that there are a total of

$10r\,d_i^2$ terms	(10 per group times r groups)
$9r\,d_id_{i+1}$ terms	(9 per group times r groups)
$8r\,d_id_{i+2}$ terms	(8 per group times r groups)
$7r\,d_id_{i+3}$ terms	(7 per group times r groups)
.	
.	
.	
$r\,d_id_{i+9}$ terms	(1 per group times r groups)

If r is large, meaning a large number of residual data points, then the sum of any one class of terms could be replaced by their average value times the number of terms. Expressed mathematically,

$$\begin{aligned}\sum d_i^2 &= 10r\,\overline{d_i^2}\\ 2\sum d_id_{i+1} &= 2\times 9r\,\overline{d_id_{i+1}}\\ 2\sum d_id_{i+2} &= 2\times 8r\,\overline{d_id_{i+2}}\\ &\vdots\\ 2\sum d_id_{i+9} &= 2\times 1r\,\overline{d_id_{i+9}}\end{aligned} \tag{69}$$

where the bar denotes the average value. Note that $\overline{d_i^2}$ is the average product of all points spaced 0 seconds apart, $\overline{d_id_{i+1}}$ is the average

product of all points spaced 0.01 second apart, etc. These terms therefore become $\phi(0)$, $\phi(0.01)$, $\phi(0.02)$, etc. Equation (67) then becomes

$$\sigma_1^2 = \frac{1}{100r}\left\{\begin{array}{l} 10r\phi(0) \\ +\,18r\phi(0.01) \\ +\,16r\phi(0.02) \\ +\,14r\phi(0.03) \\ \cdot \\ \cdot \\ \cdot \\ +\;\;2r\phi(0.09) \end{array}\right\} \tag{70}$$

Cancelling out the r's and substituting the appropriate values for $\phi(0)$, $\phi(0.01)$, etc., (as computed from Equation (66)) gives

$$\sigma_1^2 = \frac{1}{100}\left\{\begin{array}{l} 10\sigma^2 \\ +\,18\ (0.731\sigma^2) \\ +\,16\ (0.553\sigma^2) \\ +\,14\ (0.391\sigma^2) \\ +\,12\ (0.274\sigma^2) \\ +\,10\ (0.208\sigma^2) \\ +\;\;8\ (0.153\sigma^2) \\ +\;\;6\ (0.111\sigma^2) \\ +\;\;4\ (0.081\sigma^2) \\ +\;\;2\ (0.059\sigma^2) \end{array}\right\} = 0.45\sigma^2. \tag{71}$$

Therefore

$$\sigma_1 \approx 0.67\sigma. \tag{72}$$

Appendix VI—A Technique for Calculating Power Spectrum of Equi-Spaced Digital Data

This appendix describes a method developed by Real and Cannady[8] for calculating the power density spectrum for data given at equal space intervals. The computation equations are based on the use of autocorrelation techniques as described by Blackman and Tukey.[5] Only the mechanics of performing the calculations are presented here.

Restrictions

(a) The input data must be derived from a stationary process; that is, one whose statistics are time invariant.

(b) The maximum frequency computed for points spaced Δt seconds apart is

$$f_{\max} = \frac{1}{2\Delta t} \text{ cps.} \tag{73}$$

Note that $1/\Delta t$ is equal to the sampling frequency f_s. It has been pointed out that f_s must be at least twice the most significant frequency component in the spectrum in order that no "folding over" of the sampled data spectrum takes place.

(c) For N points spaced Δt seconds apart, a minimum frequency increment $\Delta f_{\min}$ is defined by

$$\Delta f_{\min} = \frac{2}{N\Delta t} \text{ cps.} \tag{74}$$

Computation Equations

Suppose a set of data $\{r_i\}$ has been curve fitted by a curve $R_c(t)$ and the residuals d_i have been calculated;

$$d_i = [r_i - R_c(t_i)]. \tag{75}$$

Note that if $R_c(t)$ is a least-mean-squares curve fit, the average residual is zero.

(a) Calculate for all m in the set $0 \leq m \leq N-1$ the autocovariance functions R_m given by

$$R_m = \frac{1}{N-m} \sum_{i=1}^{N-m} d_i d_{i+m}. \tag{76}$$

(b) Define a quantity E_m such that

$$\begin{aligned} E_m &= 1, & 0 < m < N-1, \\ E_m &= 1/2, & m = 0,\ N-1. \end{aligned} \tag{77}$$

(c) Calculate the values of L_k — the *apparent* line power at frequency $k\Delta f/4$ for all k in the set $0 \leq k \leq N-1$;

$$L_k = 4\Delta t \sum_{m=0}^{N-1} R_m E_m \cos \frac{km\pi}{N-1}. \tag{78}$$

(d) Calculate the smoothed power density Q_k at frequency $k\Delta f/4$ for all k in the set $0 \leq k \leq N-1$;

$$\begin{aligned} Q_0 &= 1/2\,[L_0 + L_1], \\ Q_k &= 1/4\,[L_{k-1} + 2L_k + L_{k+1}], \\ Q_{N-1} &= 1/2\,[L_{N-2} + L_{N-1}]. \end{aligned} \tag{79}$$

References

[1] D. K. Barton, "Final Report, Instrumentation Radar AN/FPS-16 (XN-2)," Contract Noas 55-869C, AD250 500; 1959.

[2] Burington and May, *Handbook of Probability and Statistics*, Handbook Publishers, Inc., Sandusky, Ohio, 1953.

[3] J. L. Marcum, "A Statistical Theory of Target Detection by Pulsed Radar," and Mathematical appendix, *Trans. IRE PGIT*, Vol. IT-6, April 1960.

[4] Julius T. Tou, "Digital and Sampled-Data-Control Systems," McGraw-Hill, New York, 1959.

[5] R. B. Blackman and J. W. Tukey, "The Measurement of Power Spectra," Dover Publications, New York, N.Y., 1957.

[6] Acheson J. Duncan, *Quality Control and Industrial Statistics*, Richard D. Irwin, Inc., Homewood, Ill., 1955.

[7] Paul G. Hoel, *Introduction to Mathematical Statistics*, John Wiley & Sons, Inc., New York, N. Y., 1947.

[8] Philip Real and Cynthia Cannady, "AAPDSI—Power Density Spectrum Subroutine," IBM SHARE Program.

RCA Technical Papers†

Fourth Quarter, 1963

Any request for copies of papers listed herein should be addressed to the publication to which credited.

"Analysis of Niobium Stannide," K. L. Cheng, *Chemist-Analyst* (October) 1963

"BC-7 Stereo/Dual Channel Audio Consolette," A. J. May, *Broadcast News* (October) 1963

"Bounds on Threshold Gate Realizability," R. O. Winder, *Trans. IEEE PTGEC* (Correspondence) (October) 1963

"Design of Satellite Tape Recorders After Tiros I," A. D. Burt, S. P. Clurman, and T. T. Wu, *Jour. S.M.P.T.E.* (October) .. 1963

"The Display Storage Tube as an Information Display Device," R. P. Stone, *Proc. East Coast Symposium* (October) 1963

"The Distribution of Cathode Sublimation Deposits in a Receiving Tube, as Determined by X-Ray Spectrometric Scanning," V. Raag, E. P. Bertin, and R. J. Longobucco, *Electron Tube Techniques* (October) 1963

"Effect of CdS on the Electroluminescence of ZnS:Cu Halide Phosphors," A. Dreeben, *Jour. Electrochem. Soc.* (October) 1963

"General Synthesis of Tributary Switching Networks," J. Sklansky, *Trans. IEEE PTGEC* (October) 1963

"High-Band 25 KW TV Transmitter," H. E. Small, *Broadcast News* (October) 1963

"Improved High Resolution Electron Gun for Television Cameras," S. Gray, P. C. Murray, and O. J. Ziemelis," *Jour. S.M.P.T.E.* (October) 1963

"Introduction to Color TV Transmission," J. W. Wentworth, *Broadcast News* (October) 1963

"Network Color Transmission," H. C. Gronberg, *Broadcast News* (October) 1963

"New 16-MM TV Film Projector," J. C. Adison, *Broadcast News* (October) 1963

"New Type of Two-Stream Plasma Instability," S. Tosima and R. Hirota, *Jour. Appl. Phys.* (October) 1963

"Proposals for Ordered Sequential Detection of Simultaneous Multiple Responses," H. Weinstein, *Trans. IEEE PTGEC* (Correspondence) (October) 1963

"Radioisotopes in Semiconductor Science and Technology, Part 1," W. Kern, *Semiconductor Products* (October) 1963

"Response of Photoconducting Imaging Devices with Floating Electrodes," H. S. Sommers, Jr., *Jour. Appl. Phys.* (October) ... 1963

"Superconductors at Work," E. R. Schrader, *The New Scientist* (October) 1963

"Tracing Distortion—Its Cause and Correction in Stereodisk Recording Systems." E. C. Fox and J. G. Woodward, *Jour. Aud. Eng. Soc.* (October) 1963

"Optical Quenching of Metastable Hydrogen," W. Zernik, *Phys. Rev.* (October 1) 1963

"Field and Angular Dependence of Critical Currents in Nb_3Sn," G. W. Cullen, G. D. Cody, and J. P. McEvoy, Jr., *Phys. Rev.* (October 15) 1963

† Report all corrections to *RCA Review*, RCA Laboratories, Princeton, N. J.

"Fluorescence of Naphthacene Vapor," R. Williams and G. J. Goldsmith, *Jour. Chem. Phys.* (October 15) 1963

"Thermal Conductivity of III-V Compounds at High Temperatures," E. F. Steigmeier and I. Kudman, *Phys. Rev.* (October 15) .. 1963

"Analysis of Antimony Telluride and Bismuth Telluride Mixture Containing Selenium and Iodine," K. L. Cheng and B. L. Goydish, *Anal. Chem.* (November) 1963

"Crossmodulation in Transistorized AM Auto Radio Receivers," J. A. Kuklis, *Trans. IEEE PTGBTR* (November) 1963

"Crossmodulation in Transistorized TV Tuners," H. Thanos, *Trans. IEEE PTGBTR* (November) 1963

"Description of Research on Automatic Theory Formation," S. Amarel, *Current Research and Development in Scientific Documentation*, No. 13 (November) 1963

"Fixed, Associative Memory Using Evaporated Organic Diode Arrays," M. H. Lewin, H. R. Beelitz, and J. A. Rajchman, *AFIPS Conf. Proc.*, Vol. 24, p. 101 (November) 1963

"Internal Power Dissipation in GaAs Solar Cells," M. F. Lamorte, *Advanced Energy Conversion* (November) 1963

"Laminated Ferrite Memory," R. Shahbender, C. Wentworth, K. Li, S. Hotchkiss, and J. Rajchman, *AFIPS Conf. Proc.*, Vol. 24, p. 77 (November) 1963

"Low-Noise Transistor UHF Amplifier," P. E. Kolk, T. J. Robe, and W. A. Pond, *RCA Ham Tips* (November) 1963

"Ordering and Disordering Processes in Cu_3Au-III," L. R. Weisberg and S. L. Quimby, *Jour. Phys. and Chem. of Solids*, p. 1251 (November) 1963

"Push-Pull Optical Modulators and Demodulators," F. Sterzer, *Appl. Optics* (November) 1963

"Q-Switched $CaWO_4:Nd^{3+}$ Laser," D. Karlsons and T. Falvey, *Jour. Appl. Phys.* (Communications) (November) 1963

"Radiation Studies on GaAs and Si Devices," J. J. Wysocki, *Trans. IEEE PTGNS* (November) 1963

"Radioisotopes in Semiconductor Science and Technology, Part 2," W. Kern, *Semiconductor Products* (November) 1963

"Resolution of Electrostatic Storage Targets," I. M. Krittman, *Trans. IEEE PTGED* (November) 1963

"Tunnel Diodes as Millimeter Wave Detectors and Mixers," P. E. Chase and K. K. N. Chang, *Trans. IEEE PTGMTT* (Correspondence) (November) 1963

"The Tunnel Resistor," J. T. Wallmark, L. Varettoni, and H. Ur, *Trans. IEEE PTGED* (November) 1963

"The Use of Close Spacing in Chemical Transport Systems for Growing Epitaxial Layers of Semiconductors," F. H. Nicoll, *Jour. Electrochem. Soc.*, p. 1165 (November) 1963

"Use of High-Emissivity Coating for Picture-Tube Heaters," R. K. Schneider, *Trans. IEEE PTGBTR* (November) 1963

"New Millimeter Wave Device—Beam-Plasma Amplifier," G. A. Swartz, *Electronics* (November 8) 1963

"A Discussion of Planar Sense Arrangements for Superconductive Continuous Film Memories," R. W. Ahrons, *Doctoral Dissertation*, Polytechnic Institute of Brooklyn (November 14) 1963

"Magnetic Moment of a Solid-State Plasma," A. R. Moore and J. O. Kessler, *Phys. Rev.* (November 15) 1963

"Paramagnetic Resonance of Trivalent Manganese in Rutile (TiO_2)," H. J. Gerritsen and E. S. Sabisky, *Phys. Rev.* (November 15) 1963

"Using Microcircuits in High-Resolution Range Counters," L. C. Drew, *Electronics* (November 22) 1963

"An Analysis of the Gain–Bandwidth Limitations of Solid-State Triodes," A. Rose, *RCA Review* (December) 1963
"Chemical Polishing of Silicon with Anhydrous Hydrogen Chloride," G. A. Lang and T. Stavish, *RCA Review* (December) 1963
"Comments on 'Heat-Sinking Techniques for Power Transistors in a Space Environment,'" D. F. Metz and R. A. Smith (Correspondence), *Trans. IEEE PTGSET* (December) 1963
"Coplanar-Electrode Insulated-Gate Thin-Film Transistors," P. K. Weimer, F. V. Shallcross, and H. Borkan, *RCA Review* (December) 1963
"The CTC 15: RCA's Newest Color Chassis," A. Hilderbrand, *Radio-Electronics* (December) 1963
"Electric-Field-Enhanced Precipitation of Li in Ge," J. Blanc and M. S. Abrahams, *Jour. Appl. Phys.* (Communications) (December) 1963
"Epitaxial Deposition of Silicon and Germanium Layers by Chloride Reduction," E. F. Cave and B. R. Czorny, *RCA Review* (December) 1963
"Epitaxial Deposition of Silicon by Thermal Decomposition of Silane," S. R. Bhola and A. Mayer, *RCA Review* (December) 1963
"Epitaxial Growth from the Liquid State and Its Application to the Fabrication of Tunnel and Laser Diodes," H. Nelson, *RCA Review* (December) 1963
"Epitaxial Growth of GaAs Using Water Vapor," G. E. Gottlieb and J. F. Corboy, *RCA Review* (December) 1963
"The Etching of Germanium Substrates in Gaseous Hydrogen Chloride," J. A. Amick, E. A. Roth, and H. Gossenberger, *RCA Review* (December) 1963
"Evaluation of Cadmium Selenide Films for Use in Thin-Film Transistors," F. V. Shallcross, *RCA Review* (December) 1963
"The Field-Effect Transistor—A Review," J. T. Wallmark, *RCA Review* (December) 1963
"Gas Phase Equilibria in the System GaAs-I_2," D. Richman, *RCA Review* (December) 1963
"The Growth of Germanium Epitaxial Layers by the Pyrolysis of Germane," E. A. Roth, H. Gossenberger, and J. A. Amick, *RCA Review* (December) 1963
"The Growth of Single-Crystal Gallium Arsenide Layers on Germanium and Metallic Substrates," J. A. Amick, *RCA Review* (December) 1963
"High-Power Epitaxial Silicon Varactor Diodes," H. Kressel and M. A. Klein, *RCA Review* (December) 1963
"Laminated Ferrite Memory," R. Shahbender, K. Li, C. Wentworth, S. Hotchkiss, and J. A. Rajchman, *RCA Review* (December) 1963
"Observation of Plasma Density Waves in n-InSb in Transverse Magnetic Fields," M. Toda, *Japanese Jour. Appl. Phys.* (December) 1963
"Pulse Distribution Amplifier With New Pulse Re-Forming Technique," A. J. Banks, *Jour. S.M.P.T.E.* (December) 1963
"An S-Band Traveling-Wave Maser with a 30 Per Cent Tunable Bandwidth," D. J. Miller and H. B. Yin, *Proc. IEEE* (Correspondence) (December) 1963
"Surface Waves Along a Plasma-Air Boundary," L. W. Zelby, *Proc. IEEE* (Correspondence) (December) 1963
"Technical Papers Interest Survey," C. W. Fields, *Trans. IEEE PTGEWS* (Correspondence) (December) 1963
"Transfer Characteristics of Field-Effect Transistors," W. A. Bösenberg, *RCA Review* (December) 1963
"Transients in Markov Chains," J. Sklansky and K. R. Kaplan, *Trans. IEEE PTGEC* (Correspondence) (December) 1963
"Transport of Gallium Arsenide by a Close-Spaced Technique," P. H. Robinson, *RCA Review* (December) 1963

"Vapor-Phase Synthesis and Epitaxial Growth of Gallium Arsenide," N. Goldsmith and W. Oshinsky, *RCA Review* (December) ... 1963

"Charge Transport in Copper Phthalocyanine Single Crystals," G. H. Heilmeier and S. E. Harrison, *Phys. Rev.* (December 1) 1963

"Optical Spectra of Exchange Coupled Mn++ Ion Pairs in ZnS: MnS," D. S. McClure, *Jour. Chem. Phys.* (December 1) 1963

"Quantum Oscillations in the Absorption of Helicon Waves in Solids," J. J. Quinn, *Physics Letters*, p. 235 (December 1) .. 1963

"Survey of the Spectra of the Divalent Rare-Earth Ions in Cubic Crystals," Z. Kiss and Coauthor, *Jour. Chem. Phys.* (December 15) 1963

"Carbon Blacks for Organic Depolarized Batteries," G. S. Lozier and J. B. Eisen, *Proc. Power-Sources Conf.* 1963

"Computer Memories: Remarks on Possible Future Developments," J. A. Rajchman, *Proc. of Symposium on Switching Theory in Space Technology*, p. 207 1963

"Design and Application of High-Current Tunnel Diodes to DC-AC Inverters," F. M. Carlson and P. Gardner, *Proc. Power-Sources Conf.* 1963

"Electron Spin Resonance of Copper Phthalocyanine," S. E. Harrison and J. M. Assour, *Low Symposium on Paramagnetic Resonance*, Academic Press, Inc., p. 855 1963

"Extraction of High-Density Electron Beams from Synthesized Plasmas," A. L. Eichenbaum, *Rept. of the 23rd Ann. MIT Conf. on Phy. Electronics* 1963

"Properties of Non-Metallic Materials: Survey of 1962 Literature," P. Wojtowicz, *Magnetic Materials Digest*, M. W. Lads Publishing Co., Philadelphia, Pa. 1963

"Reflectivity Measurements on InSb-In_2Te_3 and InAs-In_2Te_3 Alloys and on Pure InSb, InAs and In_2Te_3," D. L. Greenaway and M. Cardona, *Report of the Int. Conf. on the Physics of Semiconductors* (July) 1963

"The Relay Communication Satellite," W. A. Schreiner, *IEEE Wescon Show and Convention, Part 6, Communications; Space Electronics* 1963

"Si-Ge Thermocouple Development," V. Raag, *Proc. Power-Sources Conf.* 1963

"Ultimate Limits of Microelectronics," J. T. Wallmark, *Conf. Record Conf. on the Impact of Microelectronics*, p. 89 1963

"Vacuum Techniques for Fabricating Integrated Cryoelectric Computer Devices," G. W. Leck, *Trans. of the 10th Nat. Vacuum Symposium*, p. 397 1963

"Vapor Deposition of Nb_3Sn," J. J. Hanak, *Metallurgy of Advanced Electronic Materials* 1963

AUTHORS

F. ASSADOURIAN received the B.S., M.S. and Ph.D. degrees in Mathematics from New York University in 1935, 1936, and 1940, respectively. From September 1937 to September 1942 he was an instructor of Mathematics at New York University, and from 1942 to 1944 he served as Associate Professor of Mathematics at Texas Technological College, Lubbock, Texas. In June of 1944 he joined the Westinghouse Research Laboratories in East Pittsburgh, Pa. From 1946 to 1956, he was a member of the staff of the Federal Telecommunication Laboratory of International Telephone and Telegraph, first as Development Engineer and finally as a Senior Project Engineer working on microwave circuits and devices, electromagnetic propagation, and other related projects. He joined the RCA Surface Communications Division in New York in 1956, where he has worked on communication system analysis and design, including SSB technique, intermodulation distortion in FDM-FM and multicarrier repeaters and troposcatter. He is presently a Senior Staff Scientist, heading an advanced transmission group active in analysis and design of satellite communications systems.

MORREL P. BACHYNSKI received the B.E. degree in Engineering Physics in 1952 and the M.Sc. degree in 1953, both from the University of Saskatchewan, and the Ph.D. degree from McGill University in 1955. Until October 1955, as a member of the Eaton Electronics Research Laboratory, McGill University, he was engaged in investigations on microwave optics. Since that time, he has been with the RCA Victor Research Laboratories, Montreal, Canada, concerned primarily with electromagnetic wave propagation, microwaves and plasma physics. He is presently Director of the Microwave and Plasma Physics Laboratories. Dr. Bachynski is a senior member of the Institute of Electrical and Electronics Engineers, a member of the American Physical Society and of the Canadian Association of Physicists. He served on the Defence Research Board of Canada Advisory Committee on Gas Dynamics (1960-1963), is Chairman of Commission VI of the Canadian National Committee of the International Scientific Union (URSI) and is associated with the Graduate School of McGill University.

ERVIN M. BRADBURD obtained the BSEE and MSEE from Columbia University in 1941 and 1943, respectively. From 1944 to 1947 he attended the Polytechnic Institute of Brooklyn, and is presently attending Columbia University to complete the requirements for a Ph.D. in Electrical Engineering. From 1943 to 1954, Mr. Bradburd was with Federal Telecommunications Laboratories. While there, he directed development and design of navigational aids, television transmitters, and PTM radio-relay communication systems. In 1954 he joined Olympic Radio & Television, where he was in charge of research and development of military and commercial electronic equipment. Mr. Bradburd joined RCA Surface Communications Division in 1956 as a Senior Engineer. He was promoted to his present position as Manager of the Advanced Communications Techniques Group in 1957. From the fall of 1959 through the spring of 1961 he was responsible for transmission system design for the UNICOM Project under a joint RCA-BTL-ITT program. Mr. Bradburd is a member of Tau Beta Pi, and was awarded the Illig Medal for scholarship upon his graduation from Columbia University.

REMO J. D'ORTENZIO received the degrees of B.E.E. and M.S. (Mathematics) from Rensselaer Polytechnic Institute in 1956 and 1958, respectively. From 1956 to 1958, he was employed by Rensselaer Polytechnic Institute as an Instructor of Electrical Engineering. In 1958, he joined RCA as a design engineer in the Range Tracking group of the RCA Missile and Surface Radar Division at Moorestown, N. J., working on the analysis, design and testing of various radar range trackers, signal processors and target detection systems. He is currently assigned as a systems engineer associated with the definition, analysis and implementation of radar signal processes for future weapon systems. He is a member of Tau Beta Pi and Eta Kappa Nu.

KURT GRAF graduated in 1956 with a degree in Engineering Physics from the University of Saskatchewan. He returned to the University of Saskatchewan in 1958 and obtained his M.Sc. in 1960. After graduating as an engineer, Mr. Graf worked for the Defence Research Board at C.A.R.D.E. and D.R.T.E. At D.R.T.E. and while working on the Master's degree, he was engaged in research on radio wave propagation and upper-atmosphere physics. After obtaining the Master's degree, he joined the RCA Victor Research Laboratories working in plasma physics. At present Mr. Graf is studying for a Ph.D. at the University of Toronto, Institute of Aerospace Sciences.

AKOS G. REVESZ received the Dipl. Ing. degree from the Technical University of Budapest, Hungary, in 1950. He worked as assistant in Physical Chemistry Department of the University and later as staff member in the Iron and Metal Research Institute. From 1951 until 1956 he was associated with the Tungsran Company in Budapest, working on the development of various semiconductor devices. In 1957 he joined the Philips Co., Eindhoven, Holland, where he worked in research and development of electrolytic and solid-state capacitors. Mr. Revesz joined RCA Laboratories in 1959. He is engaged in research related to thin films and, in particular, to problems related to growth.

REYNOLD STEINHOFF received the B.S. degree in Electrical Engineering in 1942 from Newark College of Engineering where he is presently doing graduate study. During the period from 1942-1946, he worked as a power-transmitting-tube development and production engineer at Westinghouse and Federal Telephone and Radio Corporation. After a period of service in the U. S. Army, he returned to vacuum-tube engineering at Federal Telephone and Radio Corporation. A period of one year (1953) was spent as assistant research physicist at the Columbia University Radiation Laboratory performing investigation of low-field operation of millimeter and K-band magnetrons. He returned to Federal Telephone and Radio Corporation as Engineering leader of the magnetron development group. In 1955, he joined the RCA magnetron development group in Harrison, New Jersey. During 1959, Mr. Steinhoff was assigned to the RCA Laboratories, Princeton, where he participated in the early development work on tunnel-diode oscillators and amplifiers. He returned to the magnetron engineering development group in 1960. In March of 1962 Mr. Steinhoff was assigned to the Microwave Applied Research group at Princeton where he is continuing work on tunnel diode oscillators and amplifiers.

FRED STERZER received the B.S. degree in physics from the College of the City of New York in 1951, and the M.S. and Ph.D. degrees in physics from New York University in 1952 and 1955, respectively. From 1952 to 1953 he was employed by the Allied Control Corporation, New York, New York. During 1953 and 1954 he was an instructor in physics at the Newark College of Engineering, Newark, New Jersey, and a research assistant at New York University. He joined the RCA Electron Tube Division in Harrison, New Jersey, in October, 1954, and transferred to the Princeton, New Jersey branch in 1956, where he is now Manager of the Microwave Applied Research Group. His work has been in the field of microwave spectroscopy, microwave tubes, parametric amplifiers, tunnel-diode microwave amplifiers, frequency converters and oscillators, microwave computing circuits, and light modulators and demodulators.

Dr. Sterzer is a member of Phi Beta Kappa, Sigma Xi, the American Physical Society, and the Institute of Electrical and Electronics Engineers.

KARL H. ZAININGER received the B.E.E. degree (magna cum laude) from City College of New York in 1959, the M.S.E. and M.A. degrees from Princeton University in 1961 and 1962, respectively, and expects to receive the Ph.D. in Engineering Physics from Princeton University in 1964. In 1959 he joined the staff of RCA Laboratories where he has been engaged in research on microwave phase-locked parametric amplifiers, semiconductor devices such as the oscillistor and MOS diode, oxidation and optical properties of GaAs and Si, and ellipsometry. Mr. Zaininger is a member of Tau Beta Pi, Eta Kappa Nu, and the Institute of Electrical and Electronics Engineers.