RCA REVIEW

a technical journal

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

VOLUME XXII	SEPTEMBER 1961	NUMBER 3

CONTENTS

RCA REVIEW

BOARD OF EDITORS

Republication and Translation

SOME FACTORS AFFECTING APPLICABILITY OF OPTICAL-BAND RADIO (COHERENT LIGHT) TO COMMUNICATION*

BY

DAVID G. C. LUCK

RCA Advanced Military Systems
Princeton, N. J.

Summary—Numerous technical factors which determine the suitability of optical-band transmission, using coherent light, for future communication systems are discussed. Close similarity of coherent light to ordinary radio is stressed, while emphasizing the few differences that become important on going to a widely different frequency band. In particular, modifications necessary in the approach to problems of limitation of performance by noise are examined. Where performance in the older bands is limited by thermal-agitation noise in circuitry and surroundings, in the optical band it is limited by the inherent nature of the generation and detection processes for optical signals, as well as by ambient thermal radiation. Propagation effects are compared for the various radio bands, as are the attainable concentrations of signal energy in space and frequency. Characteristics of equipment for working with radio signals in the optical band are described briefly, with particular attention to coherent optical-power generators, power-supply conversion devices, modulators, and detectors. Some system-performance comparisons with the older bands are made which bring out the great advantage obtainable from ease of attaining high directivity in the optical band, in spite of the penalty for quantum-noise limitations. The importance of narrow-band carrier sources and sharp receiver selectivity in minimizing ambient noise is also shown, as is the heavy penalty for using noncoherent signal sources in the optical band.

INTRODUCTION

COMMUNICATION is one of the arts most likely to be affected strongly by the opening up of a major new area for application of electronic technology, which is now taking place. Extension of familiar radio techniques and concepts, in modified form, for use with electromagnetic radiation in the infrared and visible frequency bands may, indeed, prove to be one of the major breakthroughs in the history of electronics. Foreshadowed by a remarkably thorough technical prophecy published just at the end of 1958 by Schawlow and Townes,[1] the practice of the art of optical and infrared maser

---

[1] A. L. Schawlow and C. H. Townes, "Infrared and Optical Masers," *Phys. Rev.,* Vol. 112, p. 1940 (1958).

(molecular amplification by stimulated emission of radiation) was ushered in when the totally new phenomenon of strong stimulated emission of coherent-light radiation was observed in mid 1960 by Maiman[2] at Hughes Aircraft Company and by Collins and others[3] at Bell Telephone Laboratories. These light-maser devices are also known as "lasers." Another foundation pillar for the new art was foreshadowed by Forrester, Parkins and Gerjuoy[4] in 1947 when they suggested a test of the hypothesis that photoelectric emission is a true square-law phenomenon, able to generate output-current beats when illuminated by two separate light sources. Forrester's suggestion on detector action has been strikingly borne out by recent observations made by Javan[5] and others on coherent light. Thus two fresh approaches have provided the basis for a whole new technology using generators, amplifiers and frequency-changing mixers that work with coherent light. Work on the tremendous task of bringing this new technology to a useful stage is already commencing, and is likely to grow rapidly in scope and pace.

The term "coherent light" has introduced an aura of mystery which is already something of a thought deterrent, and indeed the detail of the underlying physical theory needs much improvement. This need not trouble the radio engineer excessively, however, since the physical properties of truly coherent light are, ideally, just those of the phase-coherent signals that he has been familiar with ever since the advent of continuous-wave radio. There is no need to be confused unduly by the troubles of the optical worker, who has grown up with the severe limitations imposed by the uncorrelated character of natural light, and who therefore finds the phenomenon of coherent light at significant power levels almost magical in its potentialities. What the radio engineer may now hope to gain is the ability to work freely in a new and very broad band of frequencies, higher by a factor

[2] T. H. Maiman, "Stimulated Optical Radiation in Ruby," *Nature*, Vol. 187, p. 493, Aug. 6 (1960).

[3] R. J. Collins. D. F. Nelson, A. L. Schawlow, W. Bond, C. G. B. Garrett, and W. Kaiser. "Coherence, Narrowing, Directionality, and Relaxation Oscillations in the Light Emission from Ruby," *Phys. Rev. Let.*, Vol. 5, p. 303, Oct. 1 (1960).

[4] A. T. Forrester, W. E. Parkins, and E. Gerjuoy, "On the Possibility of Observing Beat Frequencies between Lines in the Visible Spectrum," *Phys. Rev.* (L), Vol. 72, p. 728 (1947). E. Gerjuoy, A. T. Forrester, and W. E. Parkins. "Signal-to-Noise Ratio in Photoelectrically Observed Beats," *Phys. Rev.* (L) Vol. 73, p. 922 (1948).

[5] A. Javan, W. R. Bennett, Jr., and D. R. Herriott, "Population Inversion and Continuous Optical Maser Oscillation in a Gas Discharge Containing a He-Ne Mixture," *Phys. Rev. Let.*, Vol. 6, p. 106, Feb. 1 (1961).

of 30,000 than the present-day microwave frequencies. This straight-forward change is so great in magnitude, however, as to require rather drastic changes in system-application thinking, as well as a whole new set of techniques with its attendant problems. The birth pangs endured for microwaves are about to be repeated. As always in entering a new band, the noise phenomena that limit ultimate performance have new properties.
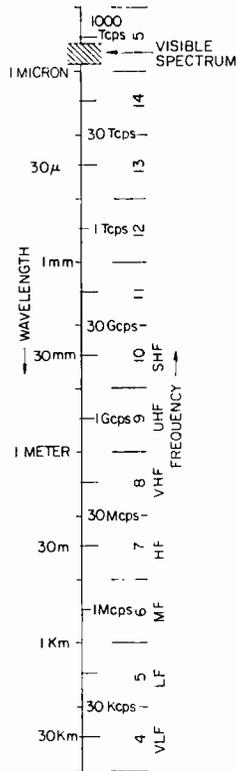
Fig. 1—Extended radio spectrum.

At the birth of a fast-moving new art with great growth potential, it is difficult to write anything that will still be valid and up-to-date when read, and detailed predictions of the pattern of development are almost sure to come to grief. Recognizing these pitfalls, this paper attempts to describe the basic character of some of the new machinery, and of some of the rules for deciding where it can be used to advantage, as well as to show samples of the pertinent play of numbers in

a rather unfamiliar range, in the hope that at least some of the material will still retain significance when it is read. Since many technical areas pertinent to useful application must be discussed to provide a rounded picture, none of these individually challenging areas can be treated very deeply here.

CHARACTERISTICS OF THE NEW WAVE BAND

Applying coherent light to communication needs is a new round of the old problem of occupying effectively a new frequency band, so it may be well to start by looking at the new band itself in relation to the radio spectrum of today. This emphasis is not meant to downgrade the very severe technique problems in getting new apparatus to operate well, which must be overcome before the new band can be used to full advantage. It is intended rather to follow the premise that those concerned with finding promising applications must press on from the assumption that many able people who are already beginning to attack technique problems will in due time find useful solutions. Figure 1 is a display of the familiar logarithmic scale of radio frequencies and wavelengths, showing the familiar named frequency bands, from VLF beginning at 3 kilocycles per second to SHF ending at 30 gigacycles (kilomegacycles) per second, together with the potential new bands out to 3000 teracycles (megamegacycles) per second,* noting also the CCIR band numbers.

Spectrum Space

The new art is beginning its development in the region of 300 teracycles (megamegacycles) per second, with a typical wavelength of 1 micron (one one-millionth meter), which lies in the near infrared. Unfortunately, it is not possible to reproduce in the 300-teracycle

---

*Because of the very wide range of values encountered. this paper makes free use of the extended set of multiplier prefixes recently standardized by the Institute of Radio Engineers. These are:

Prefix	Meaning	Prefix	Meaning
kilo-	$\times 10^3$	milli-	$\times 10^{-3}$
mega-	$\times 10^6$	micro-	$\times 10^{-6}$
giga-	$\times 10^9$	nano-	$\times 10^{-9}$
tera-	$\times 10^{12}$	pico-	$\times 10^{-12}$

Special case: 10^{-6} meter (1 micrometer) is called 1 micron.

This extension of the roster of available prefixes is a great convenience. It is clear, however, that there is great need for further extensions in both directions of the range of quantities that can be verbalized in this way.

frequency region of Figure 1 a line fine enough so that its thickness will represent to scale only the total bandwidth of the entire present radio spectrum through SHF! Even the entire visible spectrum, from about 400 to 800 teracycles per second, is rather narrow on the scale shown. Optical-maser action has already been observed between 100 and 450 teracycles per second, and extension over the range 10 to 1000 teracycles should come about fairly readily. This range is homogeneous enough in character to be called for convenience "the optical-frequency band," even though only one octave in the middle is directly visible. Ultimate limits for the new bands cannot yet be usefully stated, however, nor can any frequency region yet be regarded clearly as an optimum. Masers will probably prove operable all the way from the upper UHF through visible light and possibly well into the ultraviolet, though some practical upper limit seems sure to be reached far short of the million-million megacycles per second that is characteristic of x-rays. The term "iraser" already has had some use, and "uvaser" cannot be far behind. The actual experimental beginnings seem to lie in a good, practical middle ground. The extremely great spectrum space potentially available in the new bands is one of their most striking features.

Propagation

One thing that varies widely over the range shown in Figure 1 is the signal-impeding effect of the transmitting medium. Except for "whistler" modes in the lower VLF, the ionosphere is thoroughly opaque to signals throughout the lower-frequency radio spectrum up to the upper HF, especially at oblique incidence. Once the ionosphere becomes transparent, at VHF and above, however, it remains an open window all the way out insofar as the effects of electric charge are concerned. On the other hand, the thickest weather poses no obstacle to the lower radio frequencies, but extremely heavy rain or snow begins to give noticeable signal loss in the upper UHF, and by the middle of the SHF band this loss, due to strong multiple scattering, can become very severe. Once the absorption due to scattering in an atmosphere that is turbid with heavy water droplets or ice crystals closes in, however, the weather-blocked atmosphere never opens up again, but remains impenetrable for all higher frequencies, whether millimeter-wave, infrared, or visible light. It can be said about the new bands, then, that they have no all-weather capability, but can be used only *where* or *when* weather obscuration is not severe. This leaves them freely useful in open space, or in the high, clear

atmosphere above weather, or within weather-excluding structures such as building corridors or signal pipes, as well as outdoors at low altitudes on clear days.

Clear, non-ionized air is, of course, almost perfectly transparent for radio waves of any frequency below about 15 gigacycles per second (2-centimeter wave length). Above this, the various molecular and atomic resonances of the atmospheric gases provide strong selective absorption of electromagnetic signals of almost all frequencies. Water vapor and oxygen are major absorbers. Ionized molecules or atoms in the ionosphere simply show a somewhat altered pattern of absorbing resonances. There are a number of rather narrow, fortuitous "windows" in which no absorptions happen to pile up, but these are simply exceptions to the general rule that clear atmosphere is opaque to frequencies above about 18 gigacycles per second. The rather modest window at 35 gigacycles per second (8.6-millimeter wavelength) is well known as the K_A band of the radar art.

The most outstanding higher-frequency atmospheric window is that from about 400 to 900 teracycles per second (0.75 to 0.33 micron), the visible and near-ultraviolet region of the spectrum. One has only to look overhead on a clear, dark night, or to look from one mountain to another many miles away across a desert, to realize how startlingly transparent this atmospheric window can really be. At still higher frequencies, atmospheric absorption closes in finally, going over to multiple Rayleigh scattering from molecules themselves, and never again opens up completely. Another outstandingly good window is that from about 23 to 37 teracycles per second (13 to 8 microns) in the fairly far infrared. Other windows, of varying degrees of transparency, have been observed between 0.1 teracycle and 20 teracycles per second (3 millimeters and 15 microns, respectively), but this is a spectral region not yet thoroughly explored.

Atmospheric absorption of electromagnetic signals is affected strongly by many variables, and is far too complicated a subject to be treated extensively here. The important matter is that the new wave bands include two outstandingly good and relatively broad regions of clear-atmosphere transparency, with a number of narrower pass bands between them. Still another atmospheric effect, the slight bending of ray paths by nonuniformity of refractive index, also acts to limit attainable performance. It may further be noted in passing that blue-green light is one of a very few phenomena for which clear water, and even clear sea water, is reasonably transparent. Finally, propagation is along straight, unobstructed line-of-sight paths only.

Space Concentration

Another striking characteristic of coherent electromagnetic radiation of any frequency is that a plane wave front of finite area and uniform time phase will propagate as a parallel beam of radiation, spreading only through a diffraction angle which is fully determined by the wavelength of the radiation and the size and cross-sectional shape at the source of the original plane-wave front. This is so, of course, whether the wave front is produced by a closely spaced array of dipoles fed in phase, by a suitable reflector or lens fed by a single dipole, or by a suitably fed radiating horn. A corollary of the beam coherence is that a parallel, phase-coherent beam can always be focused, by suitable curved reflecting or refracting surfaces, so that its wave fronts are transformed to converging segments of concentric spheres, with the result that substantially all of its energy can be made to pass through a focal area of a few square wavelengths. All this is standard matter for the microwave-antenna engineer; the only things different in the new bands are the actual numbers, and these are very different. The new light sources offer a glimpse of heaven, however, for the optical worker, who has never been able to feed a dipole array in phase, nor to induce a single dipole to produce continuing radiation of significant power, nor in any way to produce a really sharp focal spot of high intensity.

A circular source of coherent plane waves of uniform strength produces at great distances, on a plane parallel to the source plane, a symmetrical distribution that has the form shown in Figure 2, where the small insert figure indicates the geometry involved. This is one of the few cases in which a closed analytical solution for the beam form is known, namely,

$$\frac{I}{I_0} = \left[2 \frac{J_1 \left(\dfrac{\pi\theta}{\theta_0} \right)}{\dfrac{\pi\theta}{\theta_0}} \right]^2 , \tag{1}$$

where I_0 is the maximum beam intensity at the center of the pattern, J_1 is the Bessel function of first kind and first order, θ is small enough to be indistinguishable from $\sin \theta$, and

$$\theta_0 = \frac{\lambda}{D} \text{ radians,} \tag{2}$$

with λ the wavelength and D the source diameter. The full width of the central beam at the first pattern nulls is $2.440\theta_0$, the full width at half intensity is $1.029\theta_0$, and the intensity at the first subsidiary maximum, or side-lobe peak, is 18.6 decibels down from the central or main-lobe maximum (the third side lobe is already down by 38 decibels). This case of a uniformly bright, circular, plane-wave source aperture is a convenient one to think about in comparing bands. Since any simple beam-width definition is arbitrary, and θ_0 is very close to
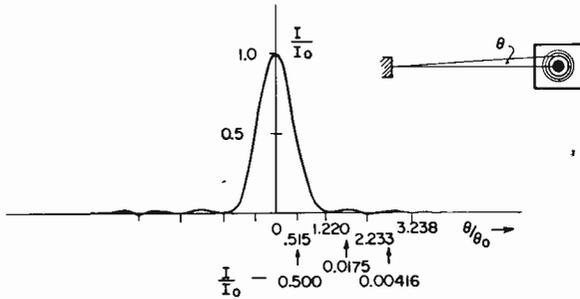
Fig. 2—Beam intensity pattern from distant circular source of coherent plane waves (uniform emittance).

the often-selected half-intensity width, θ_0 itself will be taken here as the beam width. The beam solid angle ω is then

$$\omega = \frac{\pi\theta_0^2}{4} = \frac{\pi\lambda^2}{4D^2} = \left(\frac{\pi}{4}\right)^2 \frac{\lambda^2}{A_s}, \tag{3}$$

where A_s is the source area. For the power within the beam, there is an average power gain of $4\pi/\omega$ over an isotropic radiator.

One of a few truly outstanding examples of the antenna-maker's art in radio technology to date is the 50-foot paraboloidal reflector at the Naval Research Laboratory at Bellevue, D. C., which is used for radio astronomy. This is shaped so accurately that it has been found capable of working substantially at the diffraction limit at 35 gigacycles per second, where its diameter of 1770 wavelengths gives by the above rule a beam width of 0.565 milliradians or 1.9 minutes of arc. This beam fills a solid angle of 2.5×10^{-7} steradian, giving a power gain of 77 decibels. Driving a dipole feeding this reflector at a level of 1 watt would, therefore, produce a radiated beam with an

intensity P/ω of 4.0 megawatts per steradian. Working against a like receiving reflector 10 megameters (10,000 kilometers, or one quadrant of the earth) away, with the two antenna-beam axes well aligned to each other, such a transmitter should ideally deliver 7.3 microwatts to a dipole at the receiving-reflector focus, in the absence of absorption along the 10-megameter path. Even allowing for considerable loss in imperfect practical equipment, this would still be a respectable signal to receive from a 1-watt transmitter 6200 statute miles away.

By far the most outstanding example of reflector making in existence today is 200 inches in diameter and is a paraboloid shaped accurately enough to work close to the diffraction limit at 600 teracycles per second, or more. This is located on Palomar Mountain, in southern California, and is used for photographic astronomy. Its diameter of 10 million wavelengths gives a beamwidth of 0.10 microradian, or 0.021 second of arc, thus providing a beam solid angle of 7.9×10^{-15} steradian with a directive power gain of 152 decibels. Driven coherently at 1 watt by a single radiating dipole at its focus, this reflector would produce a beam intensity of 127 terawatts (megamegawatts) per steradian. A like reflector for receiving, seen at a distance of 100 gigameters (2/3 of the mean distance from earth to sun, or 2/3 "astronomical unit" of distance), would subtend a solid angle of 2.0×10^{-21} steradian, so would be capable of intercepting 0.26 microwatt from this exceedingly sharp transmitted beam, and of focusing all of that power on a receiving dipole if transmitter and receiver beams were perfectly aligned.

The Palomar Mountain example concerns a beam-width capability not usable to full advantage in the lower atmosphere of the earth, because of local statistical fluctuations of atmospheric density, with consequent slight bendings of the beam. Arbitrariness again enters when one tries to state simply how narrow a beam can be used to advantage in the atmosphere, which varies considerably from time to time and place to place in degree of refractivity fluctuation. Astronomical experience would indicate that something of the order of 1 or 2 seconds of arc might usually be workable. At 300 teracycles per second (1 micron), a 2-second or 10-microradian coherent beam as defined above would require a plane-wave radiator diameter of 10 centimeters, or 4 inches. At 10 gigacycles per second, in the radar X band, a 10-microradian beam width would require a reflector diameter of 3.0 kilometers, or 9840 feet, with a surface accurate all over to about 0.1 inch. These numbers show well the gain in practicability of working with very sharp beams (or very high directional power gains) that should result from moving from microwaves to optical-

size waves; to be offset against this, of course, is a great loss in poor-weather operability. Even where there is no refracting atmosphere, the aiming problem involved in using effectively a beam width of 0.1 second or less is quite formidable.

One more example may be given to show that fairly sharp beams are, indeed, practical, even in very large-scale production and under field conditions of use. This is a lens-type receiving device with an aperture diameter of 10^4 wavelengths at 600 teracycles per second, giving a diffraction beam width of 0.10 milliradian, or 21 seconds of arc. The device, with excellent isolation from ambient vibrations or other disturbances, has a two-axis servo mount which is capable of very accurate tracking or flexible programmed scanning. Two of these remarkably effective devices are installed on each of us; they are called "eyes."

The other sort of space concentration, which involves the focusing of coherent plane-wave radiation by optics with large angular aperture, may also be important. 300-teracycle radiation that is fully space-coherent should be focusable in a spot of the order of 4 square microns, or 4×10^{-12} square meter, however big the source may be, so that a total power of 4 watts would permit a concentration of 1 terawatt per square meter. This is rather startling in its potential effect on anything material at the focal point. By way of comparison, the total density of emitted radiant power at the surface of the sun is only 63 megawatts per square meter. Smallness of the focal spot attainable with natural light is limited by size of the light source rather than by diffraction; therefore, focal concentration of energy in excess of the concentration at the source is unattainable with natural light.

There are certain situations in which extremely narrow beams do not appear useful in obtaining power economy, for the reason that an adequate signal must be made available throughout a large solid angle. Broadcast communication is one such situation, initial establishment of communication between stations in unknown relative positions is another, and radar search of a wide solid angle is a third (though radar tracking, once contact is made, does gain full advantage from narrow beams). Where it can be used, the extreme power concentration obtainable by working with coherent light is a very real and very important thing.

Frequency Concentration

When signaling through a background of broad-band noise, there is advantage in using a carrier with its unmodulated energy concentrated in an effective frequency band much narrower than that of the

modulating signal to be transmitted. This narrow-band character may be brought about by sophisticated spreading of the signal over a wide band through programmed processing, but such sophisticated spreading must start with a strongly time-coherent source, and when the useful signal is finally extracted by correlation against a copy of the spreading program, the effective carrier band will indeed be narrow.

Even a good crystal-controlled oscillator has some residual random fluctuation of both amplitude and frequency, mostly at very slow rates, and an appropriate question to ask is: what is the half-power bandwidth of the sharp, single spike that is the frequency spectrum of the crystal-oscillator output? This question, however, is not normally asked, so no clear answer is at hand for use as a point of reference in considering optical-band conditions. One may surmise that the half-power band width of the unmodulated output of a high-quality crystal oscillator running at 1 megacycle per second does not exceed 1 millicycle per second, or 1 part in one billion of the average frequency (consistent frequency drift is a separate matter, not considered in the statement just made).

There is already some indication that smaller fractional bandwidths than those of the best crystal oscillator can be had by the best maser generator techniques, either at microwave or optical frequencies. One part in 10^{12} or better seems to be indicated. Other light masers are not so outstanding and indeed seem to show fractional noise bands as high as 1 in 10^5. The best of these numbers is remarkable even by radio standards, but the worst is not very good on the radio scale. Especially pure, low-intensity sources of natural spectral lines can produce a fractional line width as low as 1 in 10^6, while more ordinary "monochromatic" natural-light sources of normally usable intensity, like sodium-vapor lamps, show line widths of the order of 1 in 10^5. Concentration of signal in frequency and prolonged phase coherence in time are, of course, merely two different ways of looking at the same thing.

In the present, early stage of the optical-radio art, limitations of devices prevent realization of ideally clean, spectrally pure signals. These limitations are in the nature of multiple oscillation modes of one sort and another, as well as of certain instances of strong fluctuation of amplitude, which obscure the possible purity of steady operation in a single, well-defined mode. Some of these multiple-mode difficulties may prove to be very fundamental in nature, and in that event may limit seriously the effectiveness of optical-band signals. Assuming, however, that such troubles will prove less fundamental and subject to correction by the application of sufficient ingenuity, it

appears that the optical band in time will not suffer by comparison with other radio bands on the score of spectral quality of carrier signals.

Noise

Each radio band is limited in its usefulness by the presence of meaningless extraneous signals, called noise. The most novel feature of the new bands appears to be the new (for the radio engineer, at least) set of rules that will describe the performance-limiting noise phenomena. In the time-honored radio bands, from a few kilocycles to at least two megacycles per second, the primary source of noise is the extremely high-powered natural phenomenon of lightning produced by thunderstorms. Man-made electrical disturbances take over the role of chief trouble maker at somewhat higher frequencies, but have pretty well dropped out at a hundred megacycles per second. From a few tens of megacycles per second to some hundreds, and even higher when the sun is strongly disturbed, "cosmic noise" from the sun and other astronomical objects becomes troublesome. In the gigacycle range, the atmosphere becomes able to radiate, as it begins to absorb. Noise representing thermal radiation from warm atmospheric constituents as well as from the warm earth then begins to become a significant limitation on ultimate radio performance. So much for extraneous limitations over the bands already occupied.

Beside completely extraneous noise from the environment, noise inherent in the working equipment can act to limit over-all performance. The most fundamental such source, for the ordinary radio bands, is the thermal motion of the conduction electrons in head-end electric circuits. The measure of thermal-noise power density, in watts per cycle per second of bandwidth, is simply kT, where k is Boltzmann's constant, a fundamental constant of thermal-atomic kinetics that has a value of 1.380×10^{-23} joule (watt-second) per degree Kelvin of absolute temperature T. The nature of the conductor that contains the electrons is of negligible importance. If an emission-limited space current of electrons takes part in the system operation, the fact that this current is composed of distinct electrons arriving at random, each with charge e of 1.602×10^{-19} coulomb, causes that space current to fluctuate statistically about its mean value, $\bar{\iota}$. The mean-square current fluctuation of this electronic "shot effect," per cycle per second of observing bandwidth, is just $2e\bar{\iota}$. There are additional, less fundamental sources of noise which represent less-than-ideal performance of active circuit elements. These are usually described by equipment "noise figures," and are always kept as low as is practicable, by the

continuing exercise of ingenuity to improve the state of the art. Truly emission-limited electron currents at low signal levels have customarily been avoided.

Over-all performance is limited by the combined effects of environmental and equipment noise. Where equipment noise tends to set the limit, as in the region of 3 gigacycles per second, thermal noise can be suppressed in magnitude either by strongly cooling the warm resistive elements that cause it, or by preceding such warm elements with a maser amplifier, to match a strong signal against the noise. Maser devices do not encounter serious thermal noise within themselves, but do begin to be affected by quantum noise, of which much more below.

a. Thermal Radiation

Noise conditions prevailing in the new bands relate to phenomena that have been negligible in more-familiar bands. These are certain phenomena of interaction between atomic particles of matter and electromagnetic radiation. Such interactions take place in discrete "quantum" events. The energy transformed in each quantum event is proportional to and depends only upon the frequency of the radiation involved; no smaller energy change can occur for that frequency. The constant of proportionality is a universal one, the Planck quantum constant, h, of 6.626×10^{-34} joule per cycle-per-second. The discrete energy-transforming events occur in essentially random fashion under natural conditions, with temperature affecting strongly the statistics of their occurrence.

Examination of the statistical mechanics of radiation in thermal equilibrium with solid matter (in the absence of strong, special, frequency-selective effects) yields an important expression, well known to workers in thermal radiation, for the dependence on both frequency and temperature of power radiated per unit area and per unit bandwidth (spectral radiant emittance) E_f, which is

$$E_f df = \frac{2\pi h}{c^2} \frac{f^3}{\epsilon^{hf/kT} - 1} df, \tag{4}$$

where c is the velocity of electromagnetic waves in vacuum, 2.998×10^8 meters per second, f is frequency in cycles per second, and ϵ is the base of natural logarithms. This so-called "black-body" or "complete radiation" function,[*] with some related concepts, is so important in

---

[*] Radiometrists prefer to work with radiant emittance per unit wavelength, E_λ, rather than per unit frequency, so a somewhat different function is found in their literature. In particular, the peak of E_λ occurs at a quite different frequency from that of E_f.

a version of Equation (4) in which a "quantum temperature" T_1 for the frequency under examination, given by

$$T_1 = \frac{h}{k} f = 4.801 \times 10^{-11} f_1 \text{ deg. Kelvin,} \tag{5b}$$

has been introduced for working convenience, and in which the emittance E_{f_1} for the frequency in question and the temperature T_1 has
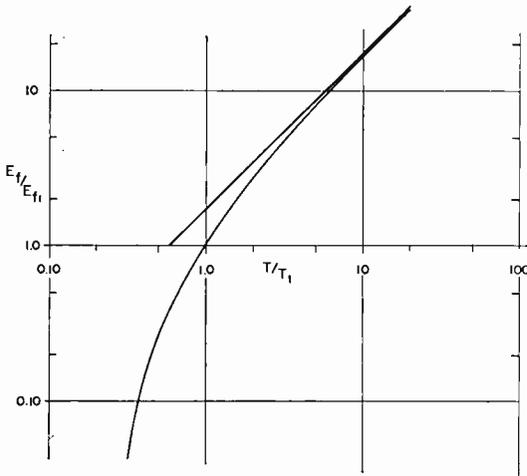
Fig. 4—Thermal radiation noise at constant frequency, power per unit bandwidth.

also been introduced for convenience; E_{f_1} is given as

$$E_{f_1} = \frac{2\pi}{\epsilon - 1} \frac{h}{c^2} f_1{}^3 = 2.696 \times 10^{-50} f_1{}^3 \quad \begin{array}{l} \text{joule per} \\ \text{square meter.} \end{array} \tag{6b}$$

The curve of Figure 4 applies for all frequencies, with no change in shape or orientation, simply by sliding along the temperature axis to accommodate any frequency. Equation (5b) represents the trade-off between temperature and frequency influences that is typical under quantum conditions.

It may be noted from the straight line on Figure 4 that, at temperatures high compared to the one for which kT equals hf at the frequency in use, the thermal-radiation emittance per unit bandwidth

becomes simply proportional to temperature, like thermal electron-agitation noise. In fact, the simple kT scale factor usually used is merely a very good low-frequency approximation in the case of thermal circuit noise also. A quantum formulation is required at sufficiently high frequencies, as Nyquist pointed out in his original publication on theory of thermal noise.[6] Table II displays quantum temperatures corresponding to a few typical frequencies, and shows plainly why noise considerations in the new bands may be expected to turn out quite different from those to which we have become accustomed.

Table II

Frequency	T_1 (°K)	E_{f_1} (w/m²/cps)
1 megacycle per second	4.8×10^{-5}	2.70×10^{-32}
3 gigacycles per second	0.132	7.28×10^{-22}
300 teracycles per second	13200	7.28×10^{-7}

Radio-wave generation probably takes place by discrete events among atomic particles, but even at the 35 gigacycles per second of K_A-band radar, the quantum temperature of these events is well below the boiling point of helium, so at ordinary temperatures the individual events go quite unnoticed in the thermal melee. In the new bands, even at a frequency as low as 100 teracycles per second, the quantum temperature is higher than any for which we can readily build furnaces, so electron agitation in room-temperature circuitry may be expected to look feeble by comparison with the discrete atomic-particle events of radiation generation and detection.

Thermal-radiation sources that follow accurately the precepts of Equation (4) are substantially nonexistent in nature, and are even hard to make in the laboratory. Familiarity with black-body radiation as a type of noise is important for two reasons, nevertheless. First, while natural radiators seldom follow Equation (4) exactly, all thermal-equilibrium radiators do act as if they were producing black-body radiation behind a curtain of selectively attenuating filters, or, in many cases that are called "gray bodies," simply behind one nonselective attenuating filter. In particular, no equilibrium thermal-radiation source can produce in any frequency interval more radiation than would be given out in the same frequency interval by a black body at

---

[6] H. Nyquist, "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.*, Vol. 32, p. 110 (1928).

the same temperature. Second, black-body radiation offers a convenient and well-understood way of becoming oriented to the impact of quantum effects on radio technology in the new bands.

As mentioned earlier, thermal radiation from earth and air just begins to become noticeable as a source of noise troubles, with very good receivers, in the region of a few gigacycles per second. This warm environment becomes a controlling source of noise in the 10 to 30-teracycle-per-second region of the new band, as Table I indicates, but has fallen off greatly by 100 teracycles. The frequency region from 100 to 1000 teracycles per second is dominated, not surprisingly, by the pervading presence of one wideband, continuously active noise source, which has a total output of 7.6×10^{17} watts per megacycle-per-second at its peak (340 teracycles per second, or 0.88 micron), with 3-decibel reduction at 140 and 640 teracycles. This ultimate in noise jammers, the sun, lays down all over the world and the space near it 2.70 microwatts per square meter per megacycle-per-second in the region of 340 teracycles per second. Its radiation must be effectively excluded from systems aiming at ultimate sensitivity.

b. Signal Quantization

The really significant limiting noise in the optical-frequency bands seems most likely to be the quantized origin and detection of the desired coherent-light signals themselves, a very novel state of affairs for the radio art. (This assumes that ingenuity in equipment design and system organization will keep the nuisance value of extraneous thermal radiation and equipment noise within bounds.) Detailed statistics of *coherent* generation and detection present an area in which the underlying theory is not in a satisfactory state, and in which suitable experimental data on which to base or assess more detailed theories has not yet been obtained. Some treatment of this subject of quantization noise is essential in assessing usefulness of the new bands for signalling, but no treatment can today be given with a high level of confidence.

A simple approach will be used here to give some feeling for the nature of the performance limitation by signal self-noise. Generation of optical-band coherent-signal power is assumed to take place by random occurrence of phase-locked quantum events, at an average rate corresponding, for the sharp radiation frequency involved, to the power produced. Detection is assumed to consist ultimately of liberation of photoelectrons which are collected in a strong field and, in effect, are counted as individual events. The average rate of occurrence of such detection events is assumed to be a constant fraction of the

average rate of occurrence of generation events, as a result of signal reduction by radiation losses between generator and detector, and by less than unity quantum efficiency of the detector. Modulation is assumed to be performed as an adjustment of the average generation rate. The "instantaneous rate" of occurrence of discrete detection events may be expected to exhibit random fluctuations about their modulated average rate. It is further assumed that fluctuations within one optical-frequency period are smoothed out by limited response speed of the detector, but that response is fast enough and modulation slow enough to avoid difficulty from loss of high modulation frequencies. Over a time in which a very large number n of events occurs, the root-mean-square value of the expected fluctuation is $\sqrt{n}$. The expectation is that a deviation of observed n from true n by more than $2\sqrt{n}$ will occur for less than one in 20 of many observations, of n events to each observation; this is the assumption of a "normal" statistical distribution.

The mean rate $\dot{n}_0$ of occurrence of generation events, each of energy hf, in the generation of power P_0, is

$$\dot{n}_0 = \frac{P_0}{hf}. \tag{7}$$

Modulation, as by a Faraday-cell light valve, results in a transmitted power P_t, where

$$P_t = [1 + m(t)]P_0, \tag{8}$$

$m(t)$ being a time-variable *intensity*-modulating function (light intensity varies linearly with input signal), with $|m(t)| < 1$, which impresses the desired signal on the transmission. Time average m of the modulation is assumed zero for simplicity. A fraction α of the transmitted signal reaches the receiving radiation collector as received power P_r. This is reduced by a loss fraction L in the receiver, to provide an effective rate of generation events, or signal quanta, applied to the detector of $\dot{n}_r$, where

$$\dot{n}_r = L\frac{P_r}{hf} = (1 + m)\,L\frac{\overline{P}_r}{hf} = (1 + m)\,\alpha L\frac{P_0}{hf}. \tag{9}$$

With a detector quantum efficiency q, the rate of electron-photo-emission detection events, $\dot{n}_p$, will be $q\dot{n}_r$, and the photocurrent i will

be $eq\dot{n}_r$. It is assumed that the phototube and operating conditions are chosen to keep the rate $\dot{n}_d$ of the thermionic or other dark-current electron emissions very much less than $\dot{n}_p$. In an observing time τ, there will be found both a short-time average $\bar{i}_\tau$ of the photocurrent, representing signal, and a mean-square departure $\overline{\Delta i^2}$ from the long-time average, representing noise. These are

$$\bar{i}_\tau = (1 + \overline{m}_\tau) \; eqL \, \frac{\bar{P}_r}{hf} \, , \tag{10}$$

and

$$\overline{\Delta i^2} = \frac{e\bar{i}}{\tau} = e^2 qL \, \frac{\overline{P}_r}{hf\tau} \, , \tag{11}$$

After a current gain G by electron multiplication and possible post amplification, signal and noise powers will appear in a load resistor R as i^2R products, giving

$$S = e^2 G^2 R \; (1 + \overline{m}_\tau)^2 \, L^2 q^2 \, \frac{\overline{P}_r{}^2}{h^2 f^2} \, , \tag{12}$$

and

$$N = e^2 G^2 R \, MLq \, \frac{\overline{P}_r}{hf\tau} \, , \tag{13}$$

where M is a factor somewhat above unity that represents excess noise in the process of amplification by secondary-emission multiplication of photoelectrons. Thus, at the output of a simple photodetector optical-radio receiver, the signal-to-noise ratio becomes

$$\frac{S}{N} = 4 \, \frac{Lq}{M} \, \frac{\overline{P}_r \tau}{hf} = 2 \, \frac{Lq}{M} \, \frac{\overline{P}_r}{hf\Delta f} \, , \tag{14}$$

where the factor $(1 + \overline{m}_\tau)^2$ has been taken as 4 to display maximum S/N capability.

In the case of an *amplitude*-modulated transmitter (light intensity varies as square of input signal) using, for example, a Kerr-cell light-valve modulator, and an optical-superheterodyne receiver with photoelectric mixer and linear final detector, the corresponding final expression differs from Equation (14) only in having $\overline{P}_r$ replaced by $2P_r$.

For comparison, an ordinary thermal-noise-limited radio receiver accepting an amplitude-modulated signal shows an input-signal power

$$S_0 = (1 + \overline{m_r})^2 \, L \, \overline{P_r}, \tag{15}$$

and an effective input-noise power

$$N_0 = \frac{kT_e}{2\tau}, \tag{16}$$

to give a signal-to-noise ratio of

$$\frac{S_0}{N_0} = 8L \frac{\overline{P_r}\tau}{kT_e} = 4L \frac{\overline{P_r}}{kT_e\Delta f}. \tag{17}$$

T_e in Equations (16) and (17) is simply an equivalent temperature that corresponds to all the noise present, making Equation (16) really only a truism. In modern high-band, low-noise receivers, however, this equivalent temperature is closely related to the low effective temperatures of ambient background seen by the antenna and of head-end coupling elements. For these, the well standardized noise-factor notation, with its nominal input temperature of 290 degrees Kelvin, is not as convenient as it was in the older art. Equation (17) holds, for example, for microwave-maser receivers, in which T_e can be of the order of 20 degrees Kelvin all told.

Other things being equal (as $\overline{P_r}$ may not be), the penalty paid for using a high carrier frequency in a system limited by detector-output-signal self-noise of quantum origin is evident on comparing Equations (14) and (17). Also evident is the penalty paid for quantum efficiency less than unity in the detector, which increases the effective coarseness of the signal grain. As an example of grain size, it may be noted that an energy change of 10^{-19} joule for one discrete event of signal or noise generation is so characteristic of the optical-frequency region that it is worth fixing in mind. At the standard exchange rates of 6.63×10^{-22} joule per teracycle-per-second, 1.38×10^{-23} joule per degree Kelvin, and 1.60×10^{-19} joule per volt. respectively, this 10^{-19} joule energy unit corresponds to radiation of 1 quantum at 151-teracycle frequency (1.99-micron wavelength), to thermal excitation of 1 atomic degree of freedom at 7250 degrees Kelvin, or to acceleration of 1 electron through a potential drop of 0.625 volt; it is also equal

to the energy stored in a 0.2-picofarad capacitance which is charged to 1 millivolt. In a 150-teracycle signal generator running at 1 milliwatt, 10^{16} generation events take place per second, or 33 per half cycle of the infrared-signal output. A photoelectric device with 10-per cent quantum efficiency would be releasing only $3\frac{1}{3}$ electrons per half cycle of coherent carrier if it received, unattenuated, the entire 1-milliwatt generated signal. Thus, at such frequencies, the signal graininess here discussed is no remote and abstract thing, but a very real source of potentially confusing statistical uncertainty.

If the observing time τ for one signal-sample decision, which cannot exceed one-half cycle of the highest signaling frequency that must be transmitted, is sufficient to ensure a statistically adequate sample of photoelectron emission at the detector, after all transmission losses, then the signal power observed on that sample (Equation (12)) should be in error by an amount greater than its r-m-s noise (Equation (13)) on barely more than 1 out of every 3 of a large number of observations. As an exercise in the behavior of small samples, not undertaken here, one may ask how few random detection events per information bit can provide a satisfactorily low error rate in binary pulse-code transmission. This would set a rather fundamental bit-rate capability limit for optical-radio communication.

The simple approach followed above in order to give some feeling for quantization noise is a tentative, middle-of-the-road one, used here by default, because of the nonexistence of adequate, firm knowledge of the subject, fully verified by experiment. It is an approach that seems fully justified in the case of natural light. Possibilities for the character of coupled processes of generation and detection of coherent-light radiation cover a wide range, however. At one extreme, photoelectric detector output might show, in addition to normal shot noise, a very large flicker at frequencies lower than the frequency difference between the two band edges of the very sharp coherent signal generated, corresponding to correlation of photoelectric emission over the duration of one signal-generation event. At the other extreme would be a rather smooth flow of photoelectric current, very crudely analogous to a thermionic current that has been smoothed by passage through a region of space charge. Since the first one of these limiting possibilities might seriously impede optical-maser applications to communication, it is encouraging, though not definitive, that the smooth beat notes observed between microwave-maser oscillators (ammonia-beam clocks) do not imply a violently fluctuating detected signal. No comparable observation made directly on optical-maser

signals was known to have been published when this was written. There is no established basis supporting strongly the more attractive limiting possibility of very smooth detector response, and there are considerations that can be applied to support the view that such smoothness is unlikely.

Consideration here has been limited to the effects of noise on amplitude or intensity-modulated signaling, for clarity of exposition. As more becomes known of the character of the self noise of the signal, careful analysis of the effect of such noise on frequency-shift types of modulation will be needed, as well as a more searching analysis of the amplitude-modulation situation, and an examination of more exotic signaling methods. Detailed character of the noise may place rather different limits on the usefulness of modulation of different sorts, and it will be important for optimum utilization that these differences be understood.

Power Capability

In the use of radio it is not merely the amount of r-f power generated that counts, but the amount radiated toward intended reception points. In broadcast service, the latter are usually distributed throughout a large solid angle as seen from the transmitter, so that extreme directional concentration of radiated power cannot be used. In fixed point-to-point communication, the maximum stably obtainable directional concentration is desirable.

Substantially any amount of power, either peak or average, that any one wishes to pay for can be *generated* at any frequency from the lowest VLF to the lower SHF region, perhaps to 6 gigacycles per second. The dollar cost may be very great, and a serious cost in developmental time lag may also exist, but within reasonable limits whatever is asked for can be had (perhaps it is just as well that nobody is about to offer to pay for a 1-gigawatt (1000-megawatt) continuous signal at 5 gigacycles). The art is there, waiting to be stretched. At the lowest frequencies, however, *radiation* is extremely inefficient because of size limitations of antenna structures, and very little directional gain is possible. From perhaps 2 megacycles per second up, very efficient radiation and increasingly effective directivity can be had. Quite significant generated power can be had up to about 36 gigacycles per second, with directional gain that is very good by past standards.

Between 40 and 300 gigacycles per second, 7.5 and 1-millimeter wavelength, respectively, a good deal of high-quality effort has gone toward seeking good generating devices for significant powers. Results

have been less than earth-shaking to date, because of the limitations imposed by the physical smallness of low-order resonant cavities. Expected need for additional spectrum space and high directivity has kept the field active. What power levels will be achieved in this region cannot now be predicted with confidence.

Above 300 gigacycles, very little work on strictly radio-style generators has been done. Work on high-power, broad-band noise-radiation generators, such as arc lamps and flash tubes, does not qualify for the radio category. Work in the region from 300 gigacycles to 30 teracycles per second is still at a low level, but verification of the possibility of optical-maser, or laser, generators has led to rapidly increasing interest in the region from about 30 to 900 teracycles per second (10 to ⅓ micron). It is far too early to predict what workable limits of power-generating capability will be reached in this region. Continuous-wave power of 10 milliwatts and pulse-peak power of 10 kilowatts already have been demonstrated in the 300-teracycle region, however, well within the first year since the first observation of the underlying phenomenon. Directional power-concentrating capabilities are striking, making this band interesting for fixed point-to-point application. On the other hand, for broadcasting tasks with requirements for sheer total power with broad beams, there may be stringent limits to the utility of the highest bands.

Doppler Effects

Change of signal frequency in transmission between points which have relative motion along the line joining them, the effect discovered over a century ago by C. J. Doppler, takes place for all radio signals. At low frequencies and low speeds, the Doppler effect may be used to measure relative speed, but does not affect practical communication. At 1 gigacycle per second, Doppler frequency shift is 1 cycle per second per foot per second of radial relative speed, giving a 2-kilocycle shift on transmissions from a Mach-2 aircraft. At 600 teracycles per second, one foot per second gives a 600-kilocycle Doppler shift. The natural habitat of optical-frequency radio is in space, where typical earth-satellite speeds are in the 20,000-foot-per second region and typical interplanetary speeds are in the region of 100,000 to 200,000 feet per second. Thus, for green-light communications in interplanetary space operations, Doppler shifts of frequency over the range up to ±120,000 megacycles per second will occur. This is sufficient to represent severe detuning in a sharply selective superheterodyne receiver, so making velocity-compensating facilities important in the design of optical-radio systems.

EQUIPMENT CHARACTERISTICS

Much of the technology of optical-band radio is so new that some description of it here may prove useful, particularly since it is rather different in detail from that familiar at lower frequencies, even in the microwave bands. This description will probably have only transitory validity in a time of whirlwind progress. Much other technology for the new art antedates all radio, and is very highly evolved; this classical optical art, the subject of a voluminous literature, lies in the field of manipulating the distribution in space and time of light radiation itself, and as little about it as possible will be repeated here. The intent here is to approach the new devices from the point of view of the potential user, describing their general nature and interface characteristics. This is in contrast to most papers on optical masers already published at the time of writing, which have been concerned primarily with the detail of the inner physical mechanisms of the devices themselves.

Organization

Organization of equipment into systems needed for optical-band radio working turns out, of course, to be entirely similar to the organization of more familiar radio systems. Figure 5 is a very general block diagram that includes, by removal of various blocks, a considerable assortment of possible one-way communication systems. Essential blocks are shown in solid outline, optional ones dotted. Radiation-beam interconnections, in the transmission band of frequencies, are shown dappled; electrical interconnections in other frequency ranges are shown solid, if essential, or dotted, if optional. New or unusual blocks are discussed in later sections.

The primary power source is considered conventional, and will not be mentioned further, but the power-supply energy-conversion process is unusual. The source of information for transmission, and the end-use device following reception, may be conventional, or may be novel, as invention dictates; one potential for novelty lies in far wider information bandwidths than have been customary in single-channel transmission systems. Modulation may be applied either in the process of coherent-light generation or to the light field when formed, as may seem best in individual cases; blocks (a) and (b) are alternatives, but one of them is essential. Directionality of radiation emission and of radiation collection are treated as essential, so that a means for aiming them becomes essential also. Fixed aiming may be done once and for all when building some systems.

At the receiving station, early interference-rejecting filtering is

virtually essential to combat broad-band interference. If the transmission-band amplifier ("r-f" amplifier) of block (c) is chosen for inclusion, it may swallow the head-end filtering function. As always, a nonlinear detector device is essential. Superheterodyne operation is optional at first, though it will probably become essential in time; it requires if chosen a local oscillator, with power-supply converter, and a second detector, the blocks (d_1), (d_2) and (d_3). Once blocks (d) are added, the function of additional intermediate-frequency amplification and filtering, block (e), is separately optional, but the choice to
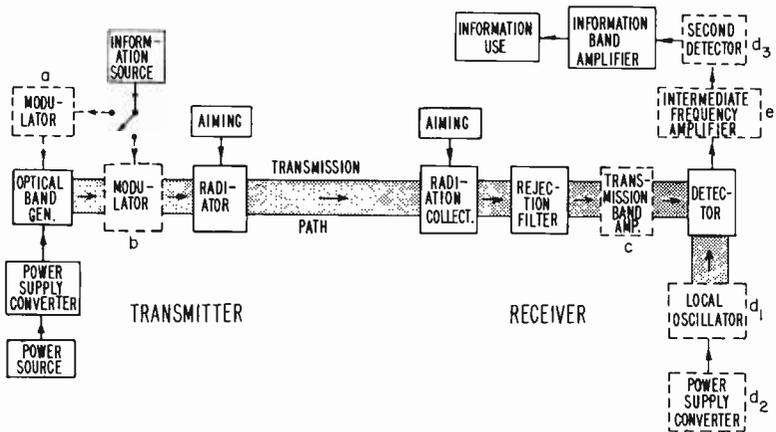
Fig. 5—Block diagram of system organization.

include (e) is likely to go with inclusion of (d) as time goes on. Amplification at the information band is essential, unless the separate need for it has been swallowed up by gain provided in block (e) or in the terminal information-using equipment. Means for tuning are essential, but are omitted in Figure 5 to keep the figure simple.

Optical-Band Power Generation

The starting point in making an optical-band power generator using maser (molecular amplification by stimulated emission of radiation) is to provide a large optical-frequency resonant cavity of extremely high selectivity; this can be done in many ways. One currently favored and long-known form is the Fabry–Perot interferometer etalon, which is a pair of parallel plane mirrors of high optical quality, both as to flatness and reflectivity, which are oriented accurately perpendicular to the line joining their centers and are separated by very

many wavelengths at the desired resonant frequency. The resonant mode is then a pattern of plane standing waves between the mirrors. If the mirrors are separated by fairly many times their diameter, and the side walls of the slender cavity, of which the mirrors form the ends, are highly absorbing (or absent) then all standing-wave modes except those due to waves traveling accurately perpendicular to the plane mirrors are strongly damped by loss to or through the sides. The cavity is filled with low-loss, dielectric material, of good optical quality, which is highly transparent at the desired resonant frequency; if this dielectric is a well-behaved solid, the mirrors are merely high-reflectance layers deposited on its optically flat and parallel ends.

The next step is to disperse in the transparent material filling the cavity a modest fraction of a material which, when fed by a suitable power supply, shows a negative absorption coefficient for the propagation of electromagnetic radiation of the desired frequency. This is simply a fluorescent material with very carefully selected atomic energy-level properties. To provide sufficient negative-absorption material to convert significant power, the resonant cavity must have a volume of very many cubic wavelengths for the very short optical waves; this means that the cavity will have a very large number of possible resonant modes.

Conceptually, and without quantum-mechanical niceties, the essential action of the specially selected material is quite simple. The desired action will be associated with transitions of the state of atomic excitation between two selected energy levels. The material will show a net resonant absorption for light of the desired frequency, with the strength of the absorption proportional to the excess Δn of the number of atoms per unit volume in the lower of the chosen states over atoms per unit volume in the upper state. Under thermal equilibrium such a lower-state excess with attendant absorption capability will always exist (unless, as one of the niceties, both states are then kept empty). The strength of the absorption is further determined by the strength of an electric-dipole moment μ associated with the chosen transition, and by the frequency width Δf of the absorption envelope. It has been shown[7] that

$$\alpha = \frac{(4\pi)^2 \mu^2 \Delta n}{\epsilon h \Delta f \lambda}, \tag{18}$$

---

[7] J. P. Wittke, "Molecular Amplification and Generation of Microwaves," *Proc. I.R.E.*, Vol. 45, p. 291 (1957).

where 4.34 α is the absorption in decibels per unit distance traveled, ϵ is the permittivity of space in units corresponding to those of the dipole moment, and λ is the wavelength of the absorbing resonance.

Supplying power to, or "pumping," the maser material is a fight against thermal equilibrium, so as to get more atoms per unit volume into the upper of the two working states than there are in the lower state, and thus obtain a negative Δn in Equation (18), with consequent negative absorption for light waves of the desired frequency. When power is applied, two things happen. One is that the absorption coefficient decreases as Δn decreases; the other is that ordinary light of the absorption-line frequency is emitted spontaneously through ordinary fluorescence. Both effects increase as more power is fed in.
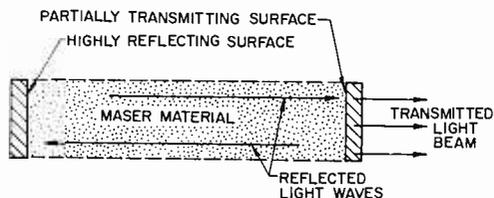
Fig. 6—Optical maser.

When Δn is reduced to zero, so is the net absorption. At still higher inputs, the maser substance will provide gain as a traveling-wave light amplifier. When this gain becomes sufficient (Δn having been made sufficiently negative) to override the losses of a round trip, twice the length of the resonant cavity with a reflection at each end, the loop gain of the entire device as an amplifier with feedback exceeds unity, and oscillation sets in, generating coherent light. The strength of oscillation, of course, is limited by whatever nonlinear effect first comes into play to cut back the loop gain.

Figure 6 shows diagrammatically the entire, deceptively simple, optical-band coherent generator, less power supply but complete with directional radiator. It should be noted from the figure that output can be drawn by allowing some leakage of light by transmission through a reflecting cavity end. The cavity serves a threefold purpose: it concentrates the traveling-wave field to facilitate prolonged interaction with the excited maser atoms; it folds the field back on itself to provide strong feedback for oscillation; also, it provides a selective feedback loop to give a very sharp resonance or pure frequency when oscillating. The output band is sharpened, with respect to the absorp-

tion resonance width Δf of the working substance, by a very large effective dynamic Q, which preliminary analysis by Schawlow and Townes[1] has indicated to be of the order of

$$\frac{1}{Q} = \frac{\delta f}{\Delta f} = \frac{4\pi h f \Delta f}{P} , \tag{19}$$

where δf is the width of the very sharp oscillator output and P is the output power. The cavity as well as the choice of line affects strongly the frequency finally produced, though this must be quite well centered on the spectrum line.

At the time of writing, four materials have been made to give maser operation in the optical band: triply ionized chromium[3] dispersed in crystalline aluminum oxide (ruby); doubly ionized samarium[8] dispersed in crystalline calcium fluoride; neutral gaseous neon[5] dispersed in helium gas; and triply ionized uranium[9] dispersed in crystalline calcium fluoride. Several attempts to use alkali-metal vapors have so far fallen short of success, but look likely to succeed in time. The chromium and samarium masers work at frequencies close to 430 teracycles per second (0.70-micron wavelength), the neon one at several frequencies near 260 teracycles (1.15 micron), and uranium near 120 teracycles (2.5 microns). That is, success to date has been confined to generation of deep-red and infrared light. It is generally anticipated that very many usable substances will be found, so that there will in time be usefully solid coverage of a very extensive band.

Observed results to date, though very striking, remain far short of the limits set by the idealized simple theory, both as to beam sharpness and frequency sharpness. Ruby crystals, in particular, very obviously prefer not to work as a single, coherent whole, but seem to behave rather as bundles of fairly independent fibers. They also appear to like to act as relaxation oscillators, showing a spiky emission envelope with time. Uranium also shows relaxation oscillations, but samarium and neon appear willing to emit smoothly with time. Neon has given continuous-wave output at a few milliwatts, but the more powerful and compact crystalline devices have so far permitted only pulsed operation, at a relaxation-pulse-peak power as high as 10 kilo-

---

[8] P. P. Sorokin and M. J. Stevenson, "Solid-State Optical Maser Using Divalent Samarium in Calcium Fluoride," *IBM Jour. of Res. and Dev.*, Vol. 5, p. 56, Jan. (1961).

[9] P. P. Sorokin and M. J. Stevenson, "Stimulated Infrared Emission from Trivalent Uranium," *Phys. Rev. Let.*, Vol. 5, p. 557, Dec. 15 (1960).

watts in the case of ruby. Continuous operation of some solid optical masers is anticipated in the near future. As always, average power attainable will evidently be limited by the disposal of waste heat due to losses, at least in the case of the solid devices. Fortunately, efficiency of the desired effect is expected to be fairly high.

Even in the case of the relatively well-behaved gaseous-neon laser, there is evidence from intra-signal beats found when its output is observed with a square-law detector that operation is taking place in several resonant modes. The anticipated penalty for using cavities that operate in resonant modes of extremely high order is evidently being paid. Multiple-mode operation, and even multiple-region operation, as in ruby, seem likely to restrict severely the utility of optical-maser devices for communication. Intensive effort to reduce such defects is needed. Results with the neon device indicate how outstandingly successful the effort to break out of the limitations of thermal equilibrium has been. It is estimated that about 5 milliwatts has been obtained in a single mode over an area about 2 centimeters in diameter and with a line width of less than 300 cycles per second at 260 tera-cycles per second (quantum temperature about 12500 degrees Kelvin), for a spectral emittance of 60 milliwatts per square meter per cycle per second. Reference to Equations (4b), (5b) and (6b) and to Figure 4 indicates that it would take a source at about 100 million degrees Kelvin to produce thermal-equilibrium radiation of such extreme spectral intensity.

Power generation in local oscillators for optical-superheterodyne receivers will be basically similar to power generation for transmitters. Detailed design, of course, will be different. In the case of the receiver oscillator, more attention may be paid to frequency stability and purity, and much less to maximizing power output. No separate discussion of the local oscillator is presented in this paper, but the great importance of having such devices should be kept in mind. Much of what has been said here also relates closely to maser *amplifiers* working directly in the optical band.

Maser devices for the optical band have characteristics that differ in several important respects from those of microwave masers. The need to use a high-order resonator to couple in enough working material to provide significant power has already been mentioned, as has the attendant difficulty of securing single-mode operation. The drastic difference in quantum temperature has also been mentioned (see Table II); this results in optical-band devices needing refrigeration only to keep them from burning up and for such niceties as keeping the lower working level empty,[8] since ordinary temperatures of the working

substances are far below the quantum temperatures of the primary transitions being used. Strong cryogenic measures are necessary, however, to make microwave masers work at all, because of the very low quantum temperatures of the low-energy transitions primarily involved in that case. Again, optical-band masers, with their relatively energetic transitions, concern mostly electric-dipole moments of atomic states, and electric coupling to the radiation field; microwave masers, with their far weaker working transitions, are concerned mainly with magnetic-dipole atomic moments and magnetic-field output coupling.

Finally, in comparing optical with microwave masers, *spontaneous* emission of narrow-band noise (ordinary fluorescence) competes with *stimulated* emission of coherent *optical* radiation, though spontaneous emission is practically negligible in the *microwave* case. Relative to the same desired output, spontaneous emission may be about 10^{13} times stronger in the optical than in the microwave band. The ratio of stimulated-emission power output to spontaneous-emission output, both per unit volume, can be found from expressions given by Schawlow and Townes[1] to be

$$\frac{P_{st}}{P_{sp}} = \pi\rho\left(\frac{\lambda}{2\pi}\right)^3 \frac{\Delta n}{h\Delta f n_2} \tag{20}$$

where Δf is the width in frequency of the atomic resonance involved, ρ is the density of resonant radiation energy that is stimulating the emission, n_2 is the number per unit volume of atoms in the upper state of the laser transition, and Δn is the excess of atoms per unit volume in the upper state over those in the lower state of the transition. Assuming an empty lower state, the radiant-energy density in the resonant cavity required to make stimulated emissions occur at precisely the spontaneous rate, for 430-teracycle radiation with 1-gigacycle atomic-resonance width, is 0.15 millijoule per cubic meter, which corresponds to a radiant flux density of 11 kilowatts per square meter in the resonant wave in the cavity.

Spontaneous emission represents both a loss, which loads the power-supply and cooling devices, and a source of interfering noise. Fortunately, the long slender cavities used to reject unwanted modes in the optical band, as well as the narrow beam in the wanted mode, also permit very strong directional rejection of spontaneous-emission noise (by a factor of perhaps 10^7 or more); selective enhancement of the wanted signal by the cavity band narrowing of Equation (19) also may help by a large factor. Spontaneous emission, however, seems likely to play a major part in setting an upper limit on operating frequencies that are attainable in practice with maser effects.
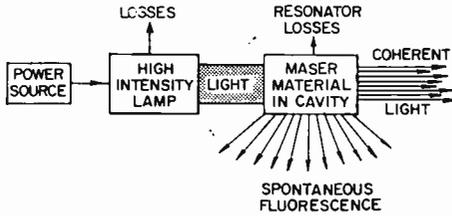
Fig. 7—Optical pumping of light maser.

Power-Supply Conversion

As with other masers, the procedure for driving maser generators of coherent light is to immerse them in something energetic that can deliver with some selectivity just the desired excitation to the power-converting atoms of the maser substance. No dissertation on pumping of masers is intended here, but a brief description of the basic processes will be included. Actually, power-supply difficulties and niceties have done a good deal to slow the development of masers in all bands. Because these supplies must accomplish extreme inversions of thermal-equilibrium conditions, there have been phases of considerable doubt as to practical feasibility of doing the job at all; that road block now seems to be behind us.

One preferred basic method of pumping optical-band masers is to create an extremely intense bath of ordinary light and immerse the maser material in it, transparent-dielectric dispersal medium and all. This can work with solid, liquid or gaseous masers. The other currently effective method, workable only with gases, is to excite the atoms of a carrier gas by collisions with the energetic electrons of an electric discharge, then let these excitation carriers transfer energy in the desired way to the more dilute maser-gas atoms by a second set of collisions. Figures 7 and 8 illustrate the mechanization of these
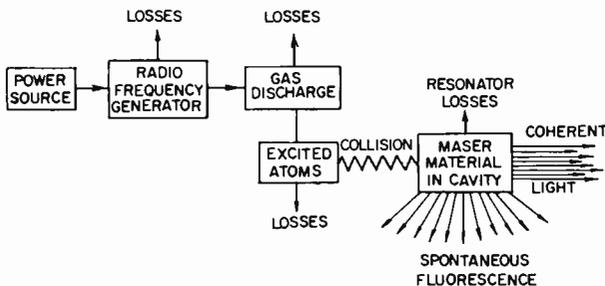
Fig. 8—Electrical-discharge pumping of optical maser.

power-supply methods, and may serve to indicate why optical-band maser generators have not yet become models of over-all efficiency.

Other methods of excitation, at present less important than the two above, are receiving active attention. One would employ electron-beam impacts, another the mechanism of electroluminescence directly in the maser material or the dispersal material carrying it. Excited-atom sorting in hot-gas beams, as used in ammonia-beam or atomic-hydrogen maser-oscillator clocks is being considered as well. The extreme care and sophistication of the work of selecting and optimiz-ing substances and interactions that has had to be done to bring the art of maser excitation to its present state remains hidden in this brief mention. Reference to the original publications is necessary to understand the rules guiding what has been done and the vast amount of delicate work that still remains to be done.

Modulation and Tuning

Methods of modulating optical-band coherent generators are in a rather primitive state. Intensity modulation by control of pumping power fed to the maser may be feasible within limits, above the oscillation threshold, but has yet to be proved satisfactory. Some degree of frequency modulation is expected to be possible through Stark or Zeeman effect from electric or magnetic fields applied directly to the maser substance, possibly leading to amplitude modulation by detuning of the generating substance relative to the resonant cavity.

Modulation of the coherent light can be done subsequent to its generation, by light valves of various sorts. Kerr or Faraday-cell light valves, modifying the polarizing properties of optical materials by action of electric and magnetic fields, may be used. Also, the shift of fringe patterns from a mechanically driven interferometer may pro-vide the basis for a light-valve modulator. Reflection from a moving mirror can give Doppler modulation of frequency, as well. All of these and probably numerous other expedients remain to be tried.

One thing that should become clear from what little can now be said is that the process of modulating coherent light may well prove a serious limitation on utilization of the wide-band potential of the optical bands. Losses in driving Kerr or Faraday cells in broad-band fashion, with attendant problems of cell cooling, may well restrict their utility to modulation-frequency bands of only a few megacycles. Mechanical motions of a good fraction of a light wavelength, needed for effective interferometric or Doppler modulation, are hard to obtain at even a few megacycles. Strong Stark or Zeemann modulation over

a very wide band may further aggravate the cooling problem of the power maser itself.

Narrow-band modulation does not appear to pose excessive problems. Neither does it exploit the capability of the new band. Table III compares this capability with past practice, to indicate what should become available.

Tuning is a function needed increasingly as a new band develops and becomes occupied. It has not yet been needed in the optical band. Zeeman tuning of microwave masers, by adjusting a fairly strong ambient magnetic field, has proved feasible. Stark tuning of optical

Table III—Channel Widths in Various Frequency Bands

Band	Frequency	Channel Width	Per Cent Width
Standard Broadcast	1 megacycle per second	10 kilocycles per second	1
VHF Television	60 megacycles per second	6 megacycles per second	10
Military UHF	300 megacycles per second	100 kilocycles per second	0.03
Optical	300 million mega-cycles per second	30 thousand megacycles per second	0.01

masers, within limits, is a significant possibility. Control of temperature, also, may prove useful for tuning. Because of the very highly selective resonant cavities used with optical masers, tracking between cavity and atomic tuning will be needed, at least over the range between successive cavity modes. Channel selection, except for the tightest packing, will be a matter of choice of working substance, or of a particular atomic energy transition within a working substance.

Control of Radiation

Permitting escape of plane-wave radiation from a resonant cavity by partial transmission through a plane reflecting surface on one end, as in Figure 6, is a direct analog of feeding a flat "bedspring" array of dipole antennas all in phase. As with the array, a sharp beam broadside to the radiating surface is the direct result. No further beam-forming elements are necessary, because the form and feeding of the array have done the entire job. There should be no surprise that even a 1-millimeter transmission spot on the flat end of a plane-wave optical resonator emits a beam which has a very small angular spread; even this small spot is the analog of a phased array of *millions* of antenna elements.

As the optical-radio art develops, fine control of main-beam shape and side-lobe levels by controlled tapering of amplitude variation across the plane-wave front should become practical. Other shapes of wave front at the radiating surface, in particular segments of spherical fronts, should also become feasible by use of modified cavity-resonance modes.

When further beam tailoring after initial formation is desired, the whole art of reflective and refractive image-forming optics is available for use. For example, the beam leaving a plane-wave resonator through any hole big enough to pass a really strong signal will be very sharp, as in Figure 9a. If it is desired to spread the radiated
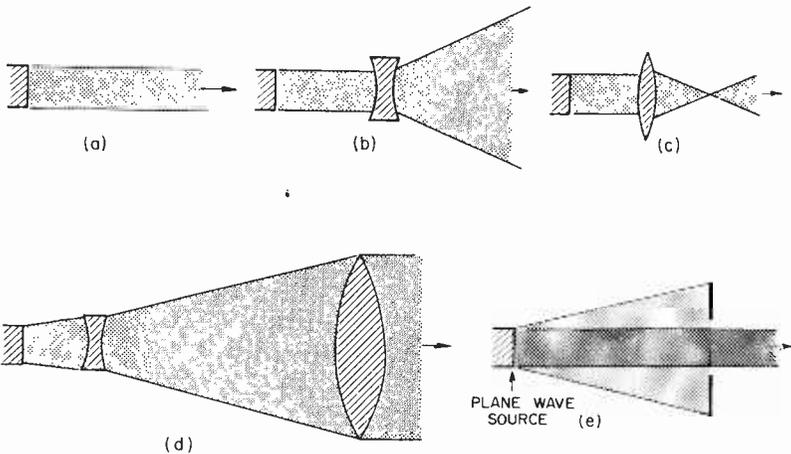
Fig. 9—Control of radiation.

beam over a considerable angle, to cover a spread of receivers, some modification is necessary. The spreading can be done by passing the sharp beam axially through a simple diverging lens, as shown in Figure 9b. If, on the other hand, it is desired to concentrate the beam energy on a very small area, this can be done with a simple converging lens, as in Figure 9c. Energy concentrations of an order not producible by other known means can be had in this way. If an exceedingly parallel beam is required, an inverted telescope can be used, as in Figure 9d, to reduce the angular spread of the beam by the ratio of the focal length of the small lens to that of the large lens. Plane mirrors can always be used to take the beam around corners.

Collection and concentration of incoming light energy at a receiving terminal will normally employ the simple arrangement of Figure

9c, with a detecting device in the focal region. Some cases will arise where the telescope of Figure 9d, traversed by light in the normal magnifying direction, can with advantage be placed in front of the concentrator of Figure 9c, the two being connected by the narrow beam of nearly parallel light leaving the telescope. One very simple but very useful optical device, the apertured shadow mask of Figure 9e, should not be overlooked. This is especially useful for keeping thermal radiation from entering a receiver obliquely. The optical lever arm between mask aperture and lens aperture that is necessary to make this work effectively has not usually been available in the longer-wavelength situation of even microwave radio, but can often be arranged in optical-band working.

Operation with extremely directive beams calls for extremely accurate beam aiming. It is just as well that aiming to a fraction of a second of arc has been a routine practical operation in astronomy for many decades. If it were not for this accumulated evidence, the radio engineer surely would never believe that what the astronomer does habitually could be made technically feasible at all.

For use later on, some relations between received and transmitted power will be set down here for the two cases of transmission into a specified solid angle and of maximum beam sharpness. A requirement that signals arriving within a specified solid angle be accepted may degrade receiver sensitivity, but need not do so if proper steps are taken, or it may lead to excess acceptance of thermal-radiation noise, but measures to avoid this are also possible. Separate cases for different receiver directivities, therefore, are not set forth. In the case of uniform distribution of transmitted power P_0 over a specified solid angle Ω_t, a receiving collector having area A_r and looking at the transmitter over a distance R will receive power P_r given by

$$P_r = \frac{P_0 A_r}{\Omega_t R^2}, \qquad (21)$$

independently of transmission frequency. In the case of maximum beam sharpness, if diffraction is limiting parallelism of the beam, the solid angle Ω_t will be $K^2\lambda^2/A_t$, where A_t is the transmitting-radiator area and K is a numeric representing the distribution of amplitude and phase over the coherent wave front in the aperture A_t (K may include a small near-field correction as well). Thus, for this case,

$$P_r = \frac{P_0 A_t A_r}{K^2\lambda^2 R^2}. \qquad (22)$$

For a wave front of circular section with constant amplitude and phase, K becomes $\pi/4$ between the half-power points of the main pencil-beam lobe for $R \gg A_t/\lambda$.

Transmission-Band Filtering and Amplification

Since thermal radiation from unwanted sources (the sun in particular) is very broad band, its interfering effect against relatively narrow-band coherent-light signals can be minimized by making the first element of an optical-band receiver a sharp filter that transmits only the wanted-signal band. As the optical band comes into extensive use, rejection of other narrow-band signals may also be needed. For superheterodyne operation, rejection of image-band signals and noise is a further function for which head-end filtering is needed. There is a well-developed art of making moderately sharp optical filters that work by light interference in thin films. Transmission of these in their pass band is severely penalized, however, in trying to make them extremely sharp. Filters with a pass-band width of 0.1 per cent of their center frequency are usually considered very sharp, though at 400 teracycles per second this still represents a 400-gigacycle pass band. Much sharper over-all filtering can be achieved by applying the output of the 0.1-per cent pre-filter to an optical interferometer and making provision to mask out unwanted frequencies after the interferometer, where radiation at such frequencies is traveling in slightly different directions from radiation at the wanted frequencies. This is the most effective method known for passive optical-frequency sorting.

When interferometer filtering is done, the possibility of placing maser material in the interferometer cavity and gaining a head-end "r-f" amplifier arises. Use of optical-band maser amplification raises again the question of the quantum statistics of interaction between coherent radiation and excited atoms of maser substance, in this case possibly with cavity resonance weaker than that of the transmitter. This question, so important to estimating the possibility of signal-to-noise improvement through use of such an amplifier, has yet to be answered fully. At any rate, it is evident that usefulness of laser head-end amplifiers will depend upon application of means to separate their spontaneous-emission output from their wanted-signal output. If amplification can be done without introducing significant additional noise, which might be found possible, such enrichment of signal flow before converting radiant signal to discrete electron-liberation events would be decidedly advantageous.

Detection

In order to extract information from received signals, it is neces-
sary to convert them, in some manner, out of the optical transmission
band. This requires some action other than linear transmission,
whether the conversion is directly to the frequency band of the modu-
lation carried or to some intermediate frequency. There is some
thought that detection may be attainable in optical-maser devices
themselves, but this remains to be clarified and verified. In the mean-
time, it appears that all detection must be done by known types of
photoelectric devices, whether photoemissive or photoconductive, and
that these are in fact well suited to the task by virtue of being
basically square-law devices. Such detectors have three faults: they
show imperfect efficiency, providing less than one charge-carrier liber-
ation upon receiving an amount of energy that corresponds to one
radiation-generating event; they show some liberation of dark-current
charges even without arrival of any stimulating radiation; and they
show limited speed of response when examined complete with means
for output coupling.

Most suitable of the detectors now available are electron-emitting
light-sensitive surfaces in vacuum, with subsequent multiplication of
the photoelectrons by secondary-electron emission. The best of these
devices show good efficiency (up to 30 per cent) at favorable optical
frequencies, rather low dark current even at room temperature, and
a spread of response delays that should permit working at output
frequencies up to 10 gigacycles per second (there is a homogeneous
over-all time lag in response which does not prevent the output current
from following fast variations of optical input). These devices can
act, effectively, to permit counting of individual electron emissions.
Noise enhancement in the secondary-emission amplification process is
real but slight (1 or 2 decibels), and the output level is high enough to
avoid contamination by noise in post-amplifiers.

These seemingly ideal devices have one serious defect: their high
quantum efficiency is limited to optical frequencies above 650 teracycles
per second, in the blue, violet, and ultraviolet regions of the spectrum.
Present optical masers work at frequencies below 450 teracycles per
second, in the deep red and infrared. Some future lasers will no doubt
work in the blue and ultraviolet, but a good detector for the ones in
the infrared will still be needed. Breakthroughs to provide electron-
multiplier photocells with high efficiency in the infrared are not clearly
predictable today, and might be associated with a dark-current problem
if we had them.

While the speed of response of electron-multiplier photocells seems very good (its upper limit has never been firmly established, but is beyond 1 gigacycle per second), and will probably be adequate to handle any modulation band that we will know how to create for some time to come, this response may prove no more than adequate to permit working at the intermediate frequencies desirable in truly broad-band superheterodyne receivers. Photomultipliers seem fairly sure to make good frequency-converting mixers when illuminated by light from a local oscillator of high purity, like that from a neon laser working in a single mode, as well as by the incoming signal. The question of their capability to work to really high intermediate frequencies is therefore an important one.

Present sensitive devices that show high quantum efficiency for red and infrared light are photoconductors, with incident radiation acting to alter their electrical conductivity. Little is understood about the inner workings of the most sensitive of the near-infrared photoconductive substances, lead sulphide, but its response appears rather slow, and its dark current considerable. Specialized semiconductor substances that respond quite remarkably at the lower infrared frequencies are not strikingly good at the high-frequency end of the infrared, and excessive dark current can be avoided only by strong refrigeration. The intrinsic photoconductive response of very pure silicon offers high quantum efficiency in just the frequency region of the light from chromium (ruby), samarium and neon lasers, the basic mechanism of charge liberation is thought to be very fast, and excessive dark current seems avoidable even at room temperature. Lack of clear needs has to date restricted the development of silicon-junction photoconductive cells, however. To realize the potential of these cells for fast response, particularly when coupled to a following broad-band amplifier, might call for very considerable development effort. Bloembergen has pointed out another way of attacking the problem of fast, efficient infrared detection.[10]

Intermediate-Frequency Amplification

Superheterodyne selectivity will be necessary to use the color purity of coherent-light signals to best advantage for avoiding disturbance by thermal radiation, as well as to occupy efficiently the spectrum space in the optical band. To gain most from these advantages, it is necessary to have a good, quiet intermediate-frequency amplifier, with

---

[10] N. Bloembergen, "Solid State Infrared Quantum Counters," *Phys. Rev. Let.*, Vol. 2, p. 84 (1959).

appropriate bandwidth and the maximum selectivity compatible with that bandwidth.

This is another area requiring development if it is not to limit exploitation of the broad-band potential of the great new region of spectrum space. Even a 5-gigacycle low-noise amplifier with 1-giga-cycle bandwidth is today a stiff order, though possible. When it is remembered that the frequency of a 1-millimeter wave is only 0.1 per cent of that of a neon laser, the appropriateness of using millimeter-wave intermediate-frequency amplifiers in optical-band radio, with perhaps 50-gigacycle bandwidth, becomes evident. It is today a little hard to conceive such technology, or the frequency-changing mixers needed to drive such amplifiers. It is also hard today to conceive the tasks that will one day make these things necessary. Nevertheless, it should be borne in mind that one important use of millimeter-wave technology may turn out to be intermediate-frequency amplification and selectivity for use with optical-band radio.

Final Detection and Information-Band Amplification

So long as the intermediate frequencies chosen do not exceed a few tens of gigacycles and the actual information bands are not broader than a few tens of megacycles per second, problems of final detection and amplification for optical-band radio will be routinely solvable by use of available components and equipment. Should a need for unitary information bands of tens of gigacycles per second arise, however, and should development of the art make possible modulation and detection of light over such bandwidths, then new final-detector coupling methods, and new "video" amplifiers of quite extraordinary capability will also have to be developed.

In all necessary areas, it may be seen from what has been said earlier that at least crudely appropriate equipment for optical-band radio systems is already available. This will permit a quick start on such communication applications of coherent light as offer early promise, subject to some initial restriction of performance. It should also be evident, however, that real exploitation of the optical band, particularly in the directions of using very wide-band channels and of getting maximum effectiveness at all carrier frequencies, will require major development effort in substantially all relevant equipment areas.

SYSTEM CONSIDERATIONS

Preparatory to discussion of system aspects, it is convenient to combine the transmission attenuation (Equations (21) and (22)) with

the received signal-to-noise (Equations (14) and (17)) to provide over-all expressions indicating how operating conditions and equipment parameters interact to determine system signal-to-noise ratio in various situations. This is done in Table IV and Figure 10, which show both analytically and graphically the signal-to-noise ratio expressions, at maximum intensity modulation, for four different working conditions. Noise reduction through signal integration is indicated in terms of duration of observing time τ allotted to reach a single readout decision (transmissible information bandwidth Δf is ideally the reciprocal of 2τ). Similarity between detector-output signal-fluctuation-limited (self-noise) cases and thermal-noise-limited cases is stressed by introducing for the latter a "thermal equivalent wavelength" λ_c, which is $hc/(kT_c)$, where T_c is simply that effective input-end temperature needed to account for the noise of the thermally limited system. No credit is allowed for the twofold improvement to be expected from superheterodyne operation in the self-noise-limited cases.

Expressions are given in a form normalized to hold as many factors as possible the same for all four cases in order to facilitate comparison of performance. The difference between broadcast and point-to-point cases is recognized as just one of usable transmitting directional power gain. No special case is shown for reception from a solid angle exceeding that of the sharpest receiver beam; it is assumed that this can be had without loss of performance by image-forming optics used with a mosaic of detectors, to provide many differently aimed receiving channels in parallel. One thing that should be noted is that, unless future research reveals surprisingly favorable statistics of coherent-light generation and detection, signal-quantum (self-noise) limitation of performance will be inevitable at any wavelengths much below $hc/(kT_c)$, however much one might wish to remain limited only by low-temperature thermal noise. Conversely, at wavelengths much above $hc/(kT_c)$, limitation by thermal noise (or worse) cannot be avoided, even if self-noise limitation were greatly desired. This analysis has been too simple to provide results for the transition region near $hc/(kT_e)$.

Discussion of all possible trade-offs among the four conditions displayed is hardly feasible here, but a few comparisons will now be made. Observation period τ and collector diameter D_r, as well as receiver loss factor L, will be considered not to change between cases. Average transmitted power will be considered scaled to working range so as to keep P_0/R^2 unchanged, while hc/D_r turns out to be a convenient unit of detector-input energy with which to compare transmitted

Table IV—Signal-to-Noise Ratios in Various Situations

Type of System → Type of Service ↓	S/N	Self-Noise Limited	Thermal-Noise Limited
(general)	$\bar{P}_r$	(14) $L \cdot \dfrac{\bar{P}_{r\tau}}{hf} \cdot \dfrac{4q}{M} = L \cdot \dfrac{\bar{P}_r}{hf\Delta f} \cdot \dfrac{2q}{M}$ $= L \cdot \dfrac{\bar{P}_{r\tau}}{hc} \cdot \lambda \cdot \dfrac{4q}{M}$	(17) $L \cdot \dfrac{\bar{P}_{r\tau}}{kT_e} \cdot 8 = L \cdot \dfrac{\bar{P}_r}{kT_e\Delta f} \cdot 4 =$ $L \cdot \dfrac{\bar{P}_{r\tau}}{hc} \cdot \lambda_e \cdot 8$
Broadcast (Specified Solid Angle)	(21) $\dfrac{P_0 A_r}{\Omega_t R^2} =$ $\dfrac{P_0 D_r^2}{R^2} \cdot \dfrac{G_t}{16}$	$L \cdot \dfrac{P_0\tau A_r}{hf\Omega_t R^2} \cdot \dfrac{4q}{M} =$ $L \cdot \dfrac{P_0\tau D_r^3}{hcR^2} \cdot G_t \cdot \dfrac{\lambda}{D_r} \cdot \dfrac{q}{4M}$	$L \cdot \dfrac{P_0\tau A_r}{kT_e\Omega_t R^2} \cdot 8 =$ $L \cdot \dfrac{P_0\tau D_r^3}{hcR^2} \cdot G_t \cdot \dfrac{\lambda_e}{D_r} \cdot \dfrac{1}{2}$
Point-to-Point (Sharpest Beam)	(22) $\dfrac{P_0 A_t A_r}{K^2\lambda^2 R^2} =$ $\dfrac{P_0 D_r^2}{R^2} \cdot \dfrac{D_t^2}{\lambda^2}\left(\dfrac{\pi}{4K}\right)^2$	$L \cdot \dfrac{P_0\tau A_t A_r}{hcK^2\lambda R^2} \cdot \dfrac{4q}{M} =$ $L \cdot \dfrac{P_0\tau D_r^3}{hcR^2} \cdot \dfrac{D_t^2}{D_r^2} \cdot \dfrac{D_r}{\lambda}\left(\dfrac{\pi}{4K}\right)^2\dfrac{4q}{M}$	$L \cdot \dfrac{P_0\tau A_t A_r}{kT_e K^2\lambda^2 R^2} \cdot 8 =$ $L \cdot \dfrac{P_0\tau D_r^3}{hcR^2} \cdot \dfrac{D_t^2}{D_r^2} \cdot \dfrac{D_r\lambda_e}{\lambda^2}\left(\dfrac{\pi}{4K}\right)^2 8$

P_0 = average transmitted power; P_r = average power into receiving collector; f = transmitted frequency; Δf = transmitted bandwidth; λ = transmitted wavelength; λ_e = thermal equivalent wavelength; $hc/(kT_e)$; L = receiver input loss factor; M = photomultiplier excess-noise factor; q = detector quantum efficiency; K = transmitting radiator illumination factor; A_t, D_t = transmitting radiator area, diameter; A_r, D_r = receiving collector area, diameter; Ω_t = transmitted-beam solid angle; G_t = transmitting directional power gain; τ = observation time for one decision; R = range of transmission; S = utilization-band maximum signal power; N = utilization-band r-m-s noise power.

energy per decision, $P_0\tau$. Transmitter power gain G_t is treated as a specified system parameter for both cases of broadcast operation. In both point-to-point or sharpest-beam cases, the circular transmitting aperture is considered to be uniformly illuminated, both as to phase and intensity, so that the beam-shape factor $\pi/(4K)$ is unity, and to have the same diameter as the circular collecting aperture, so that D_t/D_r is unity also. The greater the signal-to-noise ratio, the finer grained and more certain is the decision that can be made within one observation time τ as to strength of transmitted signal.
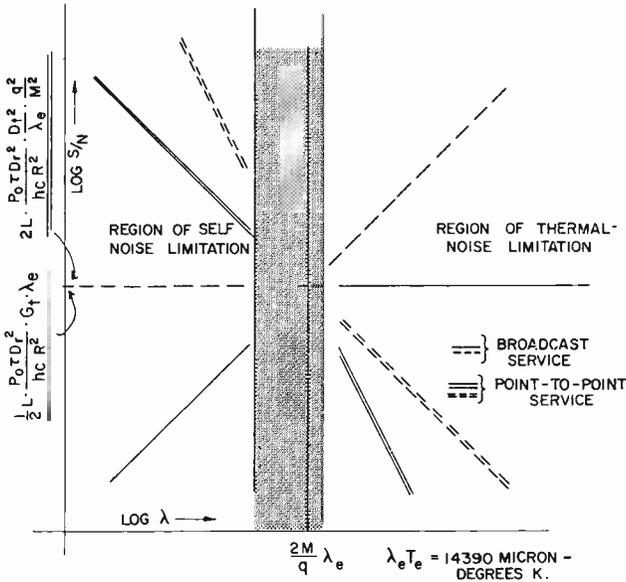
Fig. 10—System comparison.

Figure 10 shows the four situations of Table IV, with the performance in broadcast service plotted as single lines and that in maximum-directivity point-to-point service as double lines. The separation of the total range of possible conditions into mutually exclusive regions of thermal-noise limitation at long wavelengths and of self-noise limitation at short wavelengths is evident, as is the "gray area" around the thermal-equivalent wavelength, λ_c. Performance of each type of system in its own working range is shown by solid lines; extrapolated performance that would result if the same behavior could have been extended to the other region is shown dotted, but must be regarded

as inherently unattainable. It is amusing to note that in every case the unattainable situation is the attractive one.

It is clearly evident that maximum-directivity service is best at the shortest wavelengths, at which only self-noise-limited systems can be had, and that if a thermal-noise-limited system is necessary, it too should be worked at the shortest wavelength that is feasible for such systems. Likewise, it is evident, as was to be expected, that self-noise-limited operation has nothing to offer in broadcast service of specified directivity, and is worse the shorter the wavelength. Thermal-noise-limited broadcast-service operation can be seen not to be critical as to wavelength so long as effective temperature attainable does not depend on wavelength. Of course, all this neglects the possibility of a selectively attenuating propagation medium.

It can also be seen that if operation of the two types of system could be extended without deterioration into the region around λ_e, their signal-to-noise performance would be equal at a wavelength $2M\lambda_e/q$, for either type of service; such a long wavelength is quite outside the region reachable under clear self-noise-limited conditions. The signal-to-noise expression at equality, for each type of service, is indicated on the figure. Turning from the figure to the table, it is evident from the four input expressions that, while directivity improves, if permitted to do so, as the inverse square of the wavelength, signal-to-noise in a self-noise-limited receiver improves directly with the wavelength, leaving only a net advantage of the inverse first power by going to the very short wavelengths now attainable under self-noise limitation. Even an inverse first power, however, provides a major advantage when a wavelength change of 10,000 to 1 is available.

One of the important new areas about to be penetrated by communication technology is that of space operations. This will involve a quite different set of working ranges. Figure 11 is an attempt to provide graphically a view of the relations among distances of some useful mental calibration points, by marking them on a logarithmic scale of distance. The natural emptiness of the 100-to-1 range of distances between Earth-to-Moon and Earth-to-Venus at average close approach should be noted, as should the vast 5000-to-1 empty region between the greatest interplanetary distance, the diameter of the orbit of Pluto, and the smallest (for us) interstellar distance. Also to be noted are the signal-travel-time calibration points marked. Typical time delays are $1\frac{1}{4}$ seconds (one way) to the moon, $2\frac{1}{4}$ minutes to Venus at close approach, and $1\frac{1}{4}$ hours to Saturn at close approach.

A few numbers may make the information shown in Table IV and Figure 10 more concrete. Assume a receiver-input light-loss factor

L of ½, a multiplier noise factor *M* of 2, and a quantum efficiency *q* of 0.1 (this will require either better long-wave detector cells or somewhat shorter-wave optical masers than we have now), working with circular transmitting and receiving radiation aperture diameters, D_t and D_r, of 0.3 meter (1 foot). Assume further that the communication task is point-to-point transmission of binary pulse-code signals
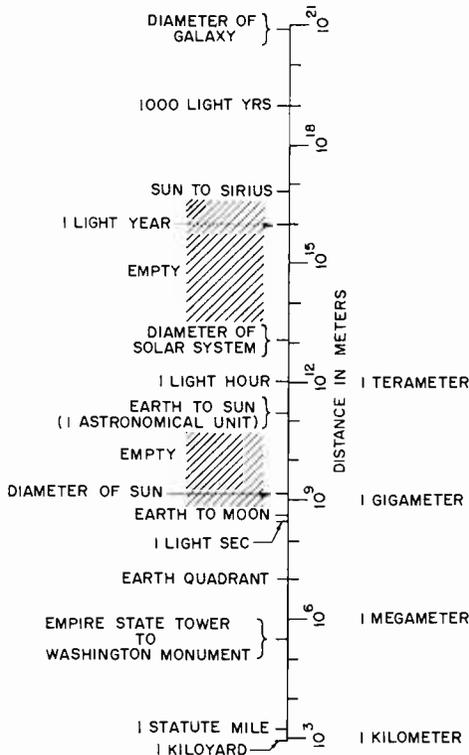
Fig. 11—Relative distances.

at 1 megabit per second, for an observing time τ of 1 microsecond. Let the working wavelength λ of the quantum-limited system be 0.55 micron (green light at 550 teracycles per second), and the wavelength of the thermally limited system be 7.5 millimeters (40 gigacycles per second), with thermal limitation corresponding to an equivalent temperature T_e of 40 degrees Kelvin for ambient background, antenna, and input-coupling circuits (for an equivalent wavelength λ_e of 360 microns).

Finally, consider the transmitter power to be adjusted to hold P_0/R^2 constant at 10^{-18} watt per square meter as range changes,

with 1 microwatt transmitted for 1 megameter range (622 statute miles, or 1000 kilometers). This calls for 1 watt to work 1 gigameter (2.6 times the distance to the moon), or 1 megawatt for 1 terameter (1.3 times as far as Earth to Jupiter at close approach), with a one-way travel time of 1 hour for the signal in the last case. The received-energy reference hc/D_r is 6.63×10^{-25} joule, and the normalized single-bit signal energy $P_0 \tau D_r^2/R^2$ is 0.9×10^{-25} joule, giving the basic bit-energy factor $P_0 \tau D_r^3/(hcR^2)$ the value of 0.136, independently of the working range chosen. For the thermally limited system, the factor $8LD_r \lambda_e/\lambda^2$ is 7.7, giving over-all a signal-to-noise ratio of unity. For the self-noise-limited system, the factor $4qLD_r/(M\lambda)$ is 5.45×10^4, and the resulting signal-to-noise ratio is 7.4×10^3. If only binary, on-off decisions are required, a S/N of 64 will limit the expected rate of wrong decisions to about 1 in 16000.

Evidently, the optical system in the example just given has considerable power to spare, while the millimeter-wave version is inadequate. If 10-foot dishes had been assumed rather than 1-foot ones, the megawatt of power to reach 1 terameter could have been scaled down to 100 watts, with the optical system retaining the same generous signal-to-noise ratio of 7400, and the millimeter-wave one still inadequate. With 10 kilowatts and 3-meter (10-foot) dishes, the entire solar system could be covered generously for point-to-point, one-way communication; extreme transmission lags would frustrate anything resembling normal two-way conversations. This assumes, of course, very stable radiator mounting and pointing, and no atmosphere in the way to scintillate. With 1 megawatt average and 30-meter (100-foot) dishes, a 1-megabit-per-second signal with a self noise of slightly under ¼ per cent could be laid down in the vicinity of our nearest-neighbor star, with a propagation-time lag of 4.3 years. That is not yet an urgent task, and we can not yet predict in what year it will become sensible to ask for 1 megawatt of average coherent-light power. At the other end of the scale, with 30-millimeter lens diameters, 1 microwatt transmitted becomes appropriate for working 10 kilometers at night, with beams broad enough not to be disturbed by normal atmospheric scintillation, and with a signal-to-noise ratio of 740 even under 1 decibel per kilometer of atmospheric absorption.

An example of the conditions encountered in trying to work with natural light may now be useful. Consider again the task of binary communication at 1 megabit per second over a range of 1 terameter (Earth to Jupiter at fairly close approach). It has been indicated above that 100 watts in a beam of width 0.183 microradian (diffraction width from a 3-meter radiating aperture at 550 teracycles per second)

can provide a signal-to-noise ratio of over 7000 at that range. A circular area of 4-millimeter diameter on the incandescent carbon of a high-intensity arc (total radiance about 22 megawatts per steradian per square meter) can produce a radiant intensity of 100 watts per steradian within the spectral range of about 300 to 750 teracycles, after throwing away the power-rich but photoelectrically weak long-wave portion of the arc spectrum. A 4-millimeter-diameter non-coherent source at the focal point of a reflector having a focal length of 22 kilometers will produce a beam parallel to within 0.183 micro-radian. To subtend at the source 1 steradian, in order to gather the full 100 watts required, an optical aperture of f:0.8 is needed. That is, a reflector of 22-kilometer focal length must have a diameter of 28 kilometers (92,000 feet). This reflector also must have its whole surface held to the intended shape with a tolerance of a fraction of a wavelength of visible light.

The above example shows what sort of apparatus natural-light optics would need to approach the performance of medium-powered coherent-light radio using a 3-meter diameter reflector (well within the state of the art today). The fact that the change from coherent to noncoherent radiation requires the entire optical system to be scaled up linearly in the ratio of the needed noncoherent-source diameter to the coherent wavelength, 7300 to 1, is hardly surprising. While 100-watt optical-maser output was not known to have been obtained continuously at the time of writing, a continuous emittance of 1 watt per square centimeter appears a realistic goal to anticipate. This would require fully coherent operation of an 11-centimeter diameter laser of circular cross section to get 100 watts; this could readily be matched to a 3-meter diameter beam by the simple inverted telescope of Figure 9d, using a demagnification of 27 times. The coherent approach has the further advantage of high monochromaticity, needed to combat broad-band background noise.

By way of contrast, at the other end of the scale, 1-microwatt transmission over 10 kilometers can be done rather compactly even with natural light, and that from a fairly monochromatic source. This means simply that the small power required can be provided by a natural-source spot only a few wavelengths in diameter, so that optical elements used in handling the radiation need to be scaled up only slightly from the coherent case.

It is true that optical masers produce spectral radiances, or powers emitted per unit area, per unit solid angle, and per unit frequency that are greater by factors on the order of 10^{12} than the radiance of the surface of the sun. However, the sun has a projected area of

1.5×10^{18} square meters and can fill a tremendous frequency band with noise interference. If much sunlight leaks or is scattered into a communication receiver, the system will no longer reach the fine performance limit set by the fluctuations in its own signal, but will rather be limited by statistical fluctuation of the stray light. This is a highly probable situation, since darkness is in very short supply throughout the solar system. The immediate surroundings of any transmitter at which a receiver may be looking are likely to be in full sunlight, as is the transmitting radiator itself. Such circumstances necessitate extraneous-noise rejection by maximum optical selectivity, which can probably only be achieved by the complication of a superheterodyne receiver.

Even a bright planet or star in the narrow sensitive cone of the receiver might upset a system working at a $P_0\tau/R^2$ of 10^{-28} joule per square meter (100 watts for 1 microsecond at 1 terameter range). A star of bolometric magnitude 1.0 lays down a radiant flux density of 10^{-8} watt per square meter at the top of the earth's atmosphere. No fully detailed noise accounting will be given here, but this corresponds to about 80,000 photons per second per square meter arriving in a frequency interval of 1 gigacycle per second near a peak-radiance frequency of 340 teracycles per second, or on the average ½ photon per microsecond on the 7 square meters of a 3-meter diameter receiving collector. There should then be an r-m-s fluctuation of about one photon, or r-m-s noise energy of 2.4×10^{-19} joule, for 1-gigacycle bandwidth into such a collector in 1 microsecond, the duration of a single signal bit. It is evident that by using directive power gain exceeding 10^{11}, the distant transmitter working at 10^{-28} joule per square meter per bit can remain substantially undisturbed by a first-magnitude star in the receiver field of view. It should be remembered that an increase of 1.0 in apparent magnitude of an astronomical source represents a decrease of 4.0 decibels in radiant flux laid down on the earth, and that there are altogether about 20 objects of bolometric magnitude* brighter than 1.0 in the entire sky, in addition to the sun and moon.

---

* Astronomers use several varieties of magnitude, depending on the method of measurement used. Examples are visual, photovisual, photographic, photoelectric, radiometric, and bolometric apparent magnitudes (the last is a measure of total watts per unit area delivered on top of the atmosphere). Bolometric and photovisual scales are so defined as to show nearly equal magnitude values for objects of spectral class G2, such as our sun, which is a fairly typical star. Relations among the various scales of magnitude are complex, and color corrections between different scales can amount to several decibels for objects of certain spectral classes.

Unless the relative positions of points between which communication is desired are accurately known in advance, in terms of a readily available reference system, establishment of communication using extremely sharp beams may become exceedingly difficult. Use of mosaic detectors can permit a sharp-beam receiver to monitor an appreciable solid angle continuously for signals, but there is no comparable help for the transmitter. It is necessary for the transmitter beam to be kept very narrow to maintain effective power, and therefore to be aimed very exactly at the receiver in order for it to be detected at all. This situation is greatly aggravated when distances are such that a prolonged round-trip time lag is involved in informing the transmitter that it has, in fact, illuminated a receiver. *Technique* areas requiring major development have been pointed out earlier; acquisition of initial contact when using very sharp beams is probably the area of *system* problems that is most in need of development effort.

Sudden and rather unexpected availability of vast new frequency bands, with good likelihood that adequate technology will develop rapidly, will influence considerably the course of future development of related areas. One such area is the millimeter-wave band. Nowadays, whenever a new problem appears for which millimeter waves are suggested as a solution, it is going to be necessary to make a careful comparison with coherent-light methods, to determine which really offers the more attractive solution. There may well be many tasks for which the optical band is clearly the better. It seems likely that a period of fast, striking, and relatively easily achieved technical advances lies ahead for this band.

Summary of Pertinent Factors

Some of the broad ideas useful in guiding communication-system synthesis that have been described are:

—Optical-band coherent radio is in most respects like other radio. Main points of difference are:

  Ease of obtaining extreme directivity.

  Tremendous available spectrum.

  Limitation of capability by fluctuation noise inherent in generation and detection of the wanted signal itself.

  Difficulty in eliminating unwanted modes to produce a pure signal.

—Turbidity of the propagation medium blocks optical-band radio completely, as it does millimeter-wave radio. Either can be used only *where* or *when* the weather is clear (space, indoors).

—Bands of clear atmospheric transmission are rare above 15 giga-cycles per second. Particularly good ones lie in the infrared, visible and near ultraviolet regions between 15 and 1500 tera-cycles per second.

—Atmospheric fluctuations limit the sharpness of directivity that can be used.

—The main source of extraneous noise in the optical bands is very wide-band thermal radiation from hot bodies. To give serious noise in the visible or ultraviolet, interfering bodies must be very hot.

—Where special applications in which light is received only by multiple scattering from turbidity are considered, sharpness of original beam offers no evident advantage to offset the limited power capability of coherent light. Great frequency sharpness, however, does offer advantage in combating broad-band noise from scattered natural light.

—Extreme receiver selectivity, and sharpness and stability of wanted signal, together with high directivity, help to hold dis-turbance by thermal-radiation noise in check.

—The basic performance limitation, when strong extraneous noise is absent, is by fluctuation of the wanted signal itself, due to the inherently discrete nature of the signal-generating and signal-detecting processes at these frequencies.

—Exact knowledge of the important self-noise properties awaits further research, and the range of possible results is great, with consequent temporary uncertainty as to attractiveness of the optical bands for critical signaling-system applications.

—Relative merit of different types of modulation may depend strongly upon the detailed character of the limiting noise.

—When critically narrow-band coherent-optical radio is used on very high-speed vehicles, Doppler frequency shifts become large and must be provided for.

—Optical-band systems will be organized on the same scheme as radio systems in other bands.

—Optical-band coherent-power generators, power-supply convert-ers, light modulators, and photoelectronic detectors are now available in early form.

—A monumental task of development remains to be done, particularly in clearing up spurious signals produced by the generators, extending generator frequency coverage, providing more efficient power-supply conversion, and speeding up the response of modulators and detectors, while also making them more efficient.

—Radiation-handling optical technology is highly developed; formation and aiming of exceedingly sharp beams are commonplace.

—When extreme directivity can be used, optical-band radio becomes very attractive. Working to interplanetary distances seems certainly feasible. Interstellar working, not yet needed, will require greatly advanced technology.

—When a requirement for broadcast communication prevents use of extremely narrow transmitting beams, the optical bands are not attractive.

—Whenever millimeter-wave systems are considered, a comparison with corresponding optical-band systems should be made, in order to choose the more suitable of the two approaches.

ACKNOWLEDGMENT

SYNCHRONIZATION OF PULSE TRAINS*

By

Jacques Dutka† and Alan A. Meyerhoff

RCA Surface Communications Division,
New York, N. Y.

Summary—Consideration is given to the problem of synchronizing a receiver which is presented with a train of binary pulses divided into long frames in which the individual marker pulses are indistinguishable from the signal pulses. A method is proposed for obtaining synchronization. An expression is derived for the probability that at least N frames are required for synchronization. This expression is evaluated and the expected number of frames required for synchronization is derived.

INTRODUCTION

THE transmission of information in digital form is usually accomplished through the use of pulse trains composed of signal elements evenly spaced in time. Often, these signal elements are grouped into equal-length frames for identification purposes, and the beginning of each frame is marked by a pulse somehow different from those in all the other frame positions. When the marker pulse is designed to have a greater duration than the ordinary pulses, or some other readily distinguished characteristic, the receiver can find it easily; but such characteristics are often undesirable because the transmission system, in accommodating them, exhibits reduced performance in handling the information-bearing pulses.

To overcome this difficulty, the marker pulse can be made identical to the signal pulses in size and shape. For example, in the transmission of binary information, the first pulse in each frame can be made a mark always. Of course, all transmissions are finite in duration, and there is a possibility that the information on some pulse position other than the marker is represented by a mark throughout the message. The longer the message, the less likely this possibility becomes, but this type of transmission is restricted in that the messages must be rather long continuous streams of information with an allowable time at the beginning during which no information is re-

---
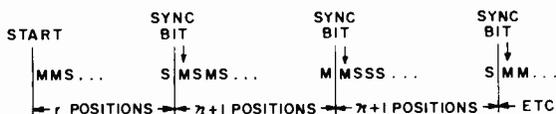
Fig. 1

ceived. An example of such a message source is a pulse code modulation (PCM) multiplex which puts out millions of pulses in a few seconds, which will run for hours or days at a time, and in which the loss of a second in establishing frame synchronization is unimportant.

In the case of a rather simple device for extracting the marker position, expressions are found here for the time required for the process. Since this quantity is statistical, the results are expressed in terms of probabilities.

DESCRIPTION OF THE PROBLEM

A receiving device is presented with a long train of regularly spaced binary pulses which is divided into frames each of which contains $n + 1$ pulses. It is assumed that marks and spaces are equiprobable and that the pulses are independent of each other except that every $n + 1$ th pulse is definitely a mark (the synchronization bit) for just one of the $n + 1$ pulse positions in the frame (see Figure 1).

The receiver must determine the positions of the synchronization bits in the pulse train. A relatively simple method for accomplishing this, shown in Figure 2, begins by examining the first pulse arriving at the receiver, where the probability that this is the rth pulse before the sync pulse in its frame is $1/(n + 1)$, and $1 \leq r \leq n + 1$. If this
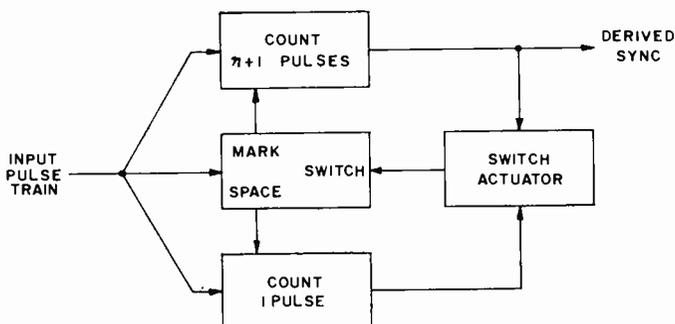
Fig. 2

first pulse examined is a mark, the receiver counts forward $n + 1$ pulses, and examines the last pulse of this count to determine whether this is a mark or a space. If the pulse which is initially examined is a space, then the receiver only counts forward one pulse before the next examination is made. This process is continued as illustrated. Thus whenever a space occurs, the receiver advances one pulse toward the correct sync position. After enough spaces have been so examined, the sync position is established, and an unending stream of sync marks is examined thenceforth. The occurrence of an error in an early sync position can extend the time required.

The process described above should be terminated after a fixed interval of time so that the receiver will not be thrown out of sync by a single error in the marker position.

ANALYTICAL FORMULATION

To evaluate the effectiveness of the proposed method, the probability is found that at least a specified time will be required for the true sync position to be determined. As a first step, the probability is derived that exactly s frames plus the fraction of a frame taken up by r pulses are required for synchronization. It will be recalled that whenever a mark is examined, a frame time is used up, while whenever a space is encountered, the time used is that allotted to a single pulse. Accordingly, the conditions for exactly s frames plus r pulses to be required are:

1. The examination process by the receiver must have begun exactly r pulses before the sync position, the probability of this being $1/(n + 1)$;

2. Exactly $s + r$ pulses had to be examined before the sync position was reached;

3. The last of these pulses had to be a space, the probability of this being $1/2$;

4. Of the remaining $r + s - 1$ pulses examined, s had to be marks, the probability of this event being equal to that of obtaining exactly s heads when $r + s - 1$ random tosses of a coin are made.

Thus the probability that exactly s frames plus r pulses are required for synchronization is

$$\frac{1}{n+1}\frac{1}{2}\left(\frac{1}{2}\right)^{r+s-1}\binom{r+s-1}{s}.$$

The probability that at least N frames are required for synchronization[*] is found by first summing the above probability for each of the pulses in a frame so that r ranges from 1 to n. (The value $n + 1$ cannot be assumed by r, for this implies immediate synchronization, which indeed occurs with probability $1/(n + 1)$.) Next, the result so obtained is summed over the number of frames, so that s ranges over the values $N, N + 1, N + 2, \ldots$. Thus the required probability is given by

$$p_n(N) = \frac{1}{n + 1} \sum_{s=N}^{\infty} \sum_{r=1}^{n} \frac{1}{2^{r+s}} \binom{r + s - 1}{s}.$$

Once the expression for $p_n(N)$ is evaluated, specific values for N and n can be chosen, corresponding to the time allowed for synchronization and the size of the frames respectively.

EVALUATION OF THE REQUIRED PROBABILITY

It is shown in the Appendix that $p_n(N)$ can be expressed as

$$p_n(N) = \frac{1}{n + 1} \left[nT(N, N + n - 1) - NT(N + 1, N + n - 1) \right]$$

where $T(r, n)$ represents the probability of obtaining at least r tails when n random tosses are made. That is

$$T(r, n) = \sum_{j=r}^{n} \binom{n}{j} \frac{1}{2^n}.$$

Extensive tables of this cumulative binomial probability are given in References (1) and (2). For values outside the range of these tables, the tables of binomial coefficients in Reference (3) and the table of

---

[*] While, in general, the specified maximum time for synchronization will correspond to an integral number N of frames plus some r additional pulses for convenience, the time taken is that which corresponds to N frames. The method of analysis in the text could be readily extended to cover the more precise case if desired.

[1] *Annals of Computation Laboratory of Harvard University*, Vol. 35, "Tables of the Cumulative Binomial Distribution," Cambridge, 1955.

[2] *Ordnance Corps Pamphlet ORDP 20-1*, Tables of the Cumulative Binomial Probabilities, Aberdeen, Maryland, September 1952.

[3] *Royal Society Mathematical Tables*, Vol. 3, "Tables of Binomial Coefficients," London, 1954.

factorials in Reference (1) can be used. Approximations to any desired degree of accuracy can be obtained by the methods described in Reference (4).† In particular, it is shown there that for $2r \geqq n$

$$A \frac{x^2}{1 + x^2} \leqq T(r, n) \leqq A,$$

where

$$A = \left(\begin{array}{c} n \\ r \end{array}\right) \left(\frac{1}{2}\right)^n \frac{r + 1}{2r - n + 1}$$

and $x = (2r - n)/\sqrt{n}$. These bounds are useful when $T(r, n)$ is small. In a particular case where the specified time allowed for synchronization is ½ second while the pulse rate is 19,200 per second with a sync pulse every 72 pulses, the values for N and n become 133 and 71 respectively. For this case,

$$1.87 \times 10^{-7} < p_{71}(133) < 3.07 \times 10^{-7},$$

a result indicating practical certainty of synchronization within the time allotted.

It is also shown in the Appendix that the expected number of frames required for synchronization is $n/2$ if the case of immediate synchronization is excluded.

APPENDIX

The evaluation of $p_n(N)$ is somewhat lengthy. To simplify the presentation, it will be helpful to derive some preliminary results.

Let p denote the probability that an event occurs and let n independent trials be made. *The probability that the event will occur at least r times, $P(r, n)$, is*

$$\sum_{s=r}^{n} \left(\begin{array}{c} n \\ s \end{array}\right) p^s q^{n-s} = p^r \sum_{s=0}^{n-r} \left(\begin{array}{c} r + s - 1 \\ s \end{array}\right) q^s, \qquad (1)$$

where $q = 1 - p$.

――――――
⁴ R. R. Bahadur, "Some Approximations to the Binomial Distribution Function," *Annals of Mathematical Statistics*, Vol. 31, pp. 43-54, (1960).

† The normal probability integral approximation which is discussed, e.g., in J. V. Uspensky's, *Introduction to Mathematical Probability*, New York, 1937, p. 135, is not applicable in the case of interest since the relative magnitude of the error term in the approximation turns out to be too large.

Proof:

Consider the incomplete beta function ratio

$$I = \frac{\displaystyle\int_0^p x^{r-1} (1-x)^{n-r} \, dx}{\displaystyle\int_0^1 x^{r-1} (1-x)^{n-r} \, dx}.$$

The denominator is the beta function

$$B(r, n-r+1) = \frac{1}{r \dbinom{n}{r}}.$$

In the numerator let $x = p(1-t)$, so that

$$I = r \binom{n}{r} p^r \int_0^1 (1-t)^{r-1} (q+pt)^{n-r} \, dt.$$

Expanding the factor $(q+pt)^{n-r}$ in the integrand and evaluating the resulting beta functions gives

$$I = r \binom{n}{r} p^r \sum_{s=0}^{n-r} p^s q^{n-r-s} \binom{n-r}{s} \frac{(r-1)!s!}{(r+s)!}$$

$$= \sum_{s=0}^{n-r} p^{r+s} q^{n-r-s} \binom{n}{r+s} = \sum_{s=r}^{n} \binom{n}{s} p^s q^{n-s}.$$

If, on the other hand, we substitute $x = pt$ in the numerator of I, then

$$I = r \binom{n}{r} p^r \int_0^1 t^{r-1} [(1-t) + tq]^{n-r} \, dt.$$

Expanding the term in brackets on the right and evaluating the beta functions obtained gives

$$I = p^r \sum_{s=0}^{n-r} q^s r \binom{n}{r} \binom{n-r}{s} \frac{(r+s-1)!(n-r-s)!}{n!}$$

$$= p^r \sum_{s=0}^{n-r} \binom{r+s-1}{s} q^s.$$

Another useful result is the formula for summation by parts from the calculus of finite differences in the form

$$\sum_{r=1}^{n} u_r v_r = u_n \sum_{r=1}^{n} v_r + \sum_{r=1}^{n-1} (u_r - u_{r+1}) \sum_{s=1}^{r} v_r . \qquad (2)$$

Proof:

Let $V_r = \sum_{s=1}^{r} v_s$ and $V_1 = v_1$. Then

$$\sum_{r=1}^{n} u_r v_r = u_n V_n + \sum_{r=1}^{n-1} (u_r - u_{r+1}) V_r$$

$$= u_n V_n + u_1 V_1 + u_2 (V_2 - V_1) + u_3 (V_3 - V_2) +$$

$$\cdots + u_{n-1} (V_{n-1} - V_{n-2}) - u_n V_{n-1}$$

$$= u_1 v_1 + \sum_{r=2}^{n} u_r (V_r - V_{r-1}) = \sum_{r=1}^{n} u_r v_r .$$

The last preliminary result which will be used is that $P(r, n)$, the probability that an event of constant probability p occurs at least r times in n independent trials, satisfies the partial difference equation

$$P(r, n) = pP(r-1, n-1) + qP(r, n-1) ; \quad 1 \leq r \leq n. \qquad (3)$$

Proof:

For the event to occur at least r times in n trials, one of the following mutually exclusive cases must occur: the event occurs at least $r-1$ times in the first $n-1$ trials and then occurs again on the nth trial; or the event occurs at least r times in the first $n-1$ trials and then fails to occur on the nth trial. Equation (3) now follows by the Addition Theorem.

It will now be proved that $p_n(N)$, the probability that at least N frames will be required for synchronization, is given by

$$p_n(N) = \frac{1}{n+1} \ [nT(N, N+n-1) - NT(N+1, N+n-1)] \ \ (4a)$$

or, alternatively,

$$p_n(N) = \frac{1}{n+1}\left[(n-N)\, T\,(N+1, N+n-1) \right.$$
$$\left. + n\binom{N+n-1}{N}\left(\frac{1}{2}\right)^{N+n-1} \right] \tag{4b}$$

where $T(r, n)$ denotes the probability of obtaining at least r tails when a coin is tossed at random n times.

Proof:

It is shown in the text that

$$p_n(N) = \frac{1}{n+1}\sum_{s=N}^{\infty}\sum_{r=1}^{n}\frac{1}{2^{r+s}}\binom{r+s-1}{s}$$

or,

$$p_n(N) = \frac{1}{n+1}\sum_{r=1}^{n}\left[\sum_{s=N}^{\infty}\frac{1}{2^{r+s}}\binom{r+s-1}{s}\right].$$

Now the sum in brackets on the right can be rewritten in the form $\sum_{s=N}^{\infty} = \sum_{s=0}^{\infty} - \sum_{s=0}^{N-1}$. Moreover

$$\sum_{s=0}^{\infty}\frac{1}{2^{r+s}}\binom{r+s-1}{s} = \frac{1}{2^r}\left(1-\frac{1}{2}\right)^{-r} = 1.$$

Thus

$$\sum_{s=N}^{\infty}\frac{1}{2^{r+s}}\binom{r+s-1}{s} = 1 - \frac{1}{2^r}\sum_{s=0}^{N-1}\binom{r+s-1}{s}\frac{1}{2^s}.$$

From Equation (1), the second sum on the right represents the probability of obtaining at least r heads when $N+r-1$ random tosses of a coin are made. The complement of this number represents the probability $T(N, N+r-1)$ of obtaining at least N tails when $N+r-1$ random tosses of a coin are made. Hence

$$p_n(N) = \frac{1}{n+1}\sum_{r=1}^{n} T(N, N+r-1).$$

To evaluate the sum on the right, the summation by parts formula, Equation (2), is applied to yield

$$\sum_{r=1}^{n} T(N, N + r - 1) \cdot 1 = T(N, N + n - 1) \cdot \sum_{r=1}^{n} 1 +$$

$$+ \sum_{r=1}^{n-1} (T(N, N + r - 1) - T(N, N + r)) \cdot \sum_{s=1}^{r} 1.$$

Now by Equation (3),

$$T(N, N + r) = \frac{1}{2} T(N - 1, N + r - 1) + \frac{1}{2} T(N, N + r - 1).$$

Substituting this in the second sum on the right,

$$\sum_{r=1}^{n} T(N, N + r - 1) = nT(N, N + n - 1) - \sum_{r=1}^{n-1} \left(\frac{r}{2}\right)\binom{N + r - 1}{N - 1}\left(\frac{1}{2}\right)^{N+r-},$$

since it follows from the definition of $T(r, n)$ that

$$T(N, N + r - 1) - T(N - 1, N + r - 1) = -\binom{N + r - 1}{N - 1}\left(\frac{1}{2}\right)^{N+r-1}.$$

Thus

$$p_n(N) = \frac{1}{n + 1}\left[nT(N, N + n - 1) - \frac{1}{2}\sum_{r=1}^{n-1} r\binom{N + r - 1}{N - 1}\left(\frac{1}{2}\right)^{N+r-1} \right]$$

$$= \frac{1}{n + 1}\left[nT(N, N + n - 1) - \frac{N}{2^{N+1}}\sum_{s=0}^{n-2}\binom{N + s}{s}\left(\frac{1}{2}\right)^{s} \right].$$

From Equation (1) and the definition of $T(r, n)$, it follows that

$$p_n(N) = \frac{1}{n + 1}\left[nT(N, N + n - 1) - NT(N + 1, N + n - 1) \right].$$

Alternatively, since

$$T(N, N + n - 1) - T(N + 1, N + n - 1) = \binom{N + n - 1}{N}\left(\frac{1}{2}\right)^{N+n-1},$$

another form for $p_n(N)$ which is simpler for computational purposes is

$$p_n(N) = \frac{1}{n+1}\left[(n-N)\ T(N+1, N+n-1) + n\binom{N+n-1}{N}\left(\frac{1}{2}\right)^{N+n-1}\right].$$

(It may be shown directly that the sum on the right is positive for all positive integers n and N.)

To obtain the expected number of frames required for synchronization, the sum

$$\sum_{N=0}^{\infty}\ N[p_n(N) - p_n(N+1)]$$

must be evaluated. From the definition of $p_n(N)$,

$$p_n(N) = \frac{1}{n+1}\ \sum_{s=N}^{\infty}\sum_{r=1}^{n}\frac{1}{2^{r+s}}\binom{r+s-1}{s},$$

we find that

$$\begin{aligned}
\sum_{N=0}^{\infty}\ N[p_n(N) - p_n(N+1)] &= \frac{1}{n+1}\ \sum_{N=0}^{\infty}\sum_{r=1}^{n}\frac{N}{2^{r+N}}\binom{r+N-1}{N}\\
&= \frac{1}{n+1}\ \sum_{r=1}^{n}\frac{1}{2^r}\sum_{N=1}^{\infty}\frac{1}{2^N}\frac{(r+N-1)!}{(N-1)!(r-1)!}\\
&= \frac{1}{n+1}\ \sum_{r=1}^{n}\frac{r}{2^{r+1}}\sum_{M=0}^{\infty}\frac{1}{2^M}\binom{r+M}{M}\\
&= \frac{1}{n+1}\ \sum_{r=1}^{n}\frac{r}{2^{r+1}}\left(1-\frac{1}{2}\right)^{-(r+1)}\\
&= \frac{1}{n+1}\ \sum_{r=1}^{n}r = \frac{n}{2}.
\end{aligned}$$

GENERATION OF LINEAR BINARY SEQUENCES*

By

H. CHARNEY AND C. MENGANI

RCA Surface Communications Division.
New York, N. Y.

Summary—This paper presents criteria and techniques which are useful in the design of digital equipment to perform the following and related functions: linear pseudo-random word generation, linear digital-pattern generation. Analytical techniques are utilized that result in criterion formulas which are of aid to the designer. One sufficient condition completely specifies the configuration of a linear word generator of desired repetition rate. A technique is outlined which facilitates the design of linear equipment to synthesize particular digital patterns.

INTRODUCTION

THIS paper treats binary sequences and the linear configurations which potentially yield them, in terms of certain polynomials. Figure 1 contains the general configuration which is considered. It is composed simply of a shift register, and the associated circuitry which combines various bits of a binary sequence in a linear manner and introduces the result to the input of the register, thus perpetuating the sequence.
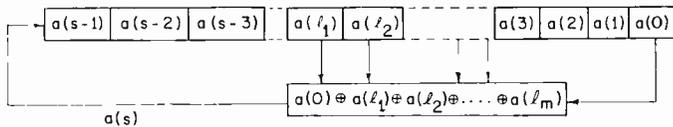
$$A \oplus B = AB' + A'B$$

$$A \oplus B \oplus C = AB'C' + A'BC' + A'B'C + ABC$$

Fig. 1

Polynomials are derived by means of the Z transform operating upon the governing circuit equation as well as the general repetition equation. Information relating sequences to configurations is derived from these polynomials. A basic system equation is derived which yields general information pertaining to output sequences (when

---

treating the analysis aspect) and general information pertaining to possible configurations (when considering the synthesis aspect). Techniques for obtaining precise sequences and configurations are described.

ANALYSIS

Let $n =$ interval of time, $a(\alpha) =$ state of the α stage at $t = 0$, and $a(\alpha + n) =$ state of the α stage at $t = n$. From Figure 1,

$$a(s) = a(0) \oplus a(l_1) \oplus a(l_2) \oplus \cdots \oplus a(l_m),$$

and, in general,

$$a(s + n) = a(n) \oplus a(l_1 + n) \oplus a(l_2 + n) \oplus \cdots \oplus a(l_m + n). \quad (1)$$

Since Equation (1) is a linear modulo-2 recursion equation,

$$\overline{Z}[a(s + n)] = \overline{Z}[a(n) \oplus a(l_1 + n) \oplus \cdots \oplus a(l_m + n)], \quad (2)$$

where $\overline{Z}[f(n)] = F(z)$ is the Z transform of $f(n)$.

The Z transform is defined as follows:

If $f(n) = f(0), f(1), f(2), \cdots f(n),$

$$\overline{Z}[f(n)] = F(z) = f(0) + z^{-1} f(1) + z^{-2} f(2) + \cdots + z^{-n} f(n),$$

$$\overline{Z}[f(n + 1)] \quad = f(1) + z^{-1} f(2) + z^{-2} f(3) + \cdots + z^{-n+1} f(n)$$

$$= z[f(0) + z^{-1} f(1) + z^{-2} f(2) + \cdots + z^{-n} f(n)] - z f(0),$$

so that $\overline{Z}[f(n + 1)] = z F(z) - z f(0).$

But

$$\overline{Z}[f(n)_{\bmod 2}] = \overline{Z}[f(n)]_{\bmod 2},$$

and therefore Equation (2) may be rewritten

$$\overline{Z}[a(s + n)] = \overline{Z}[a(n)] \oplus \overline{Z}[a(l_1 + n)] \oplus \cdots \oplus \overline{Z}[a(l_m + n)]. \quad (3)$$

Applying the Z transform, term by term,

$$z^s a(z) \oplus z^s a(0) \oplus z^{s-1} a(1) \oplus \cdots \oplus z a(s - 1) = a(z) \oplus$$

$$z^{l_1} a(z) \oplus z^{l_1} a(0) \oplus z^{l_1-1} a(1) \oplus \cdots \oplus z a(l_1 - 1) \oplus$$

$$z^{l_2} a(z) \oplus z^{l_2} a(0) \oplus z^{l_2-1} a(1) \oplus \cdots \oplus z a(l_2 - 1) \oplus$$

$$\cdots \cdots \oplus z^{l_m} a(z) \oplus z^{l_m} a(0) \oplus z^{l_m-1} a(1) \oplus \cdots \cdots \oplus z a(l_m - 1).$$

$$(4)$$

Grouping all $a(z)$ terms on the left-hand side of the equation and all other terms on the right hand side, the result is

$$(z^s \oplus z^{l_1} \oplus z^{l_2} \oplus \cdots z^{l_m} \oplus 1) \, a(z) = P_0(z),$$

where $P_0(z)$ is a polynomial of z whose terms depend solely upon initial conditions.

Solving for $a(z)$,

$$a(z) = \frac{P_0(z)}{S(z)}, \tag{5}$$

where

$$S(z) = z^s \oplus z^{l_1} \oplus z^{l_2} \oplus \cdots \oplus z^{l_m} \oplus 1$$

and is determined by the circuit configuration employed.

For a repeating sequence with a repetition interval of k,

$$a(n + k) = a(n),$$

and

$$\overline{Z}[a(n + k)] = \overline{Z}[a(n)],$$

or

$$z^k \, a(z) \oplus z^k \, a(0) \oplus z^{k-1} \, a(1) \oplus \cdots \oplus z \, a(k-1) = a(z). \tag{6}$$

Solving for $a(z)$

$$a(z) = \frac{P(z)}{z^k \oplus 1}, \tag{7}$$

where $P(z)$ is a polynomial in z which describes the generated sequence.

Combining Equations (5) and (7),

$$\frac{P_0(z)}{S(z)} = \frac{P(z)}{z^k \oplus 1},$$

or

$$P(z) = \frac{(z^k \oplus 1) \, P_0(z)}{S(z)}. \tag{8}$$

This equation relates the transform $P(z)$ of the generated sequence having a repetition period k, to the transform $P_0(z)$ of the initial conditions by means of the system function $1/S(z)$.

In order that $P(z)$ have positive exponents only, the numerator in the right hand side of the equation must be factorable by $S(z)$. $P_0(z)$ by itself cannot be factored by $S(z)$ because $S(z)$ contains a maximum exponent of s and a unity term, but $P_0(z)$ does not contain a unity term nor an exponent greater than s. Therefore, fundamental values of k must be found to permit the numerator to be factorable by $S(z)$. All such fundamental values of k are permissible repetition periods.

If $S(z)$ is a prime polynomial, $P_0(z)/S(z)$ is an irreducible fraction. Consequently, the value of k is a function of $S(z)$ only, and is independent of $P_0(z)$. In other words, k is independent of initial conditions. Therefore, only one fundamental value of k exists.

In the generation of the output sequence, a particular combination in the register will uniquely and always yield only one other combination, i.e.,

$$A \rightarrow B \rightarrow C \rightarrow \cdots \rightarrow A.$$

For a register of size s, there are 2^s such combinations. Therefore the relationship

$$k_1 p_1 + k_2 p_2 + \cdots + k_v p_v = 2^s \tag{9}$$

holds, where p_i is the number of groups associated with k_i, and $k_i p_i$ is the number of combinations associated with k_i.

The combination all zeros is a degenerate case and results in a kp value of unity, thereby modifying Equation (9):

$$k_1 p_1 + k_2 p_2 + \cdots + k_v p_v = 2^s - 1. \tag{10}$$

When only one fundamental value of k exists, as in the case of a prime $S(z)$, Equation (10) reduces to

$$kp = 2^s - 1. \tag{10a}$$

If $2^s - 1$ is a prime number, unity is the only permissible value of p, and

$$k = 2^s - 1, \tag{11}$$

which is the maximum possible repetition period. To recapitulate, the following two conditions are sufficient to achieve the maximum repetition period:

(1) $S(z)$ is a prime polynomial,

(2) $2^s - 1$ is a prime number.

If $2^s - 1$ is not a prime number, k is bounded by Equation (10a), but can be determined exactly by the following procedure. The term $z^k + 1$ must be divisible by $S(z)$ resulting in polynomial quotient whose lowest-order term is unity and with a remainder of zero. By imposing these criteria, values of k are found in terms of the exponents of $S(z)$, $(s, l_1, l_2,$ etc). For example, in the case where $S(z) = z^s \oplus z^l \oplus 1$, the division may be performed in the following manner:

$$z^{k-s} \oplus z^{k-2s+l}$$

$$z^s \oplus z^l \oplus 1 \quad \overline{\smash{\big)}\ z^k \oplus 1}$$

$$z^k \oplus z^{k-s+l} \oplus z^{k-s}$$

(r_1) $\qquad z^{k-s+l} \oplus z^{k-s} \oplus 1$

$$z^{k-s+l} \oplus z^{k-2s+2l} \oplus z^{k-2s+l}$$

(r_2) $\qquad z^{k-s} \oplus z^{k-2s+2l} \oplus z^{k-2s+l} \oplus 1.$

The first two partial remainders are designated r_1 and r_2. Successive partial remainders will be

(r_3) $\qquad z^{k-2s+2l} \oplus z^{k-2s} \oplus 1$

(r_4) $\qquad z^{k-2s} \oplus z^{k-3s+3l} \oplus z^{k-3s+2l} \oplus 1$

(r_5) $\qquad z^{k-3s+3l} \oplus z^{k-3s+2l} \oplus z^{k-3s+l} \oplus z^{k-3s} \oplus 1$

(r_6) $\qquad z^{k-3s+2l} \oplus z^{k-3s+l} \oplus z^{k-3s} \oplus z^{k-4s+4l} \oplus z^{k-4s+3l} \oplus 1$

(r_7) $\qquad z^{k-3s+l} \oplus z^{k-3s} \oplus z^{k-4s+4l} \oplus z^{k-4s+2l} \oplus 1$

(r_8) $\qquad z^{k-3s} \oplus z^{k-4s+4l} \oplus z^{k-4s+l} \oplus 1$

(r_9) $\qquad z^{k-4s-4l} \oplus z^{k-4s} \oplus 1$

(r_{10}) $\qquad z^{k-4s} \oplus z^{k-5s+5l} \oplus z^{k-5s+4l} \oplus 1.$

This yields the following values of k:

(r_2) $\qquad\qquad k = 2s - l, \quad s = 2l$

(r_4) $k = 3s - 2l, \quad s = 3l$

(r_8) $k = 4s - l, \quad\;\; s = 4l$

(r_{10}) $k = 5s - 4l, \quad s = 5l$

etc.

It is noted that this particular division does not result in solutions for all possible S and l proportionalities. The reason is traced to assumptions made while carrying out the long division procedure. For example, the second partial remainder (r_2) contains four terms. The arrangement of these terms presupposes the condition $k - s > k - 2s + 2l$, or $S > 2l$; hence all solutions obtained at or beyond this point conform to this condition. Successive assumptions at each partial remainder specify the above family of solutions.

When $S(z)$ is not a prime polynomial, it may be factored into its prime factors. Since $P_0(z)$ may take the form of any polynomial up to degree s, depending upon the initial condition, groups of initial conditions exist which will permit $P_0(z)$ to be factored by all possible combinations of the prime factors of $S(z)$. Consider the initial conditions which permit the factoring of $P_0(z)$ by one combination of prime factors of $S(z)$, with $z^k + 1$ being divisible by the remaining prime factors, as before. This yields a k value associated with the group of initial conditions. Thus there is a value of k associated with each combination of prime factors of $S(z)$. For u prime factors, the total number of combinations, and therefore possible k values, are

$$C = 2^u - 1. \tag{12}$$

The k value for each individual prime factor, when all other prime factors are cancelled by $P_0(z)$, can be found by the method previously described. In this case, the highest exponent of each prime factor will be less than s, so that the k values will be small and perhaps less than s. Following is an example of how the k values are found for the prime factors of $S(z)$.

Let $S(z) = z^5 \oplus z \oplus 1 = (z^2 \oplus z \oplus 1)\,(z^3 \oplus z^2 \oplus 1).$

Then,

$$P(z) = \frac{(z^k \oplus 1)\,P_0(z)}{(z^2 \oplus z \oplus 1)\,(z^3 \oplus z^2 \oplus 1)}$$

for initial conditions of $P_0(z)$ which are factorable by the cubic prime factor,

$$P(z) = \frac{(z^k \oplus 1)\, P_0{}'(z)}{z^2 \oplus z \oplus 1},$$

and, from Equation (11),

$$k_1 = 2^2 - 1 = 3,$$

since the denominator and $2^2 - 1$ are prime. Similarly,

$$k_2 = 2^3 - 1 = 7.$$

Since $2^u - 1 = 3$, a third value, k_3, exists. This value of k is determined from the preceding equations relating k, s and l, as follows:

$$s = 5l, \qquad\qquad \text{from } S(z)$$

then, since $l = 1$,

$$k_3 = 5s - 4l = 5 \times 5 - 4 = 21.$$

This result may be verified by Equation (10)

$$3 + 7 + 21 = 2^5 - 1 = 31.$$

In the example, three binary sequences with repetition rates 3, 7, and 21, are generated. Each sequence depends upon the initial conditions in the register. There is a group of 3 initial conditions which result in $k_1 = 3$; 7 initial conditions which result in $k_2 = 7$, and the remaining 21 possible initial conditions give rise to a k_3 value of 21.

The initial conditions associated with k_1 and k_2 may be established by equating $P_0(z)/z$ to each of the prime factors of $S(z)$ in turn and solving for the coefficients of $P_0(z)$. This will yield only one initial condition for k_1 and k_2; the other initial conditions may be obtained from the binary sequences generated by the known initial conditions.

Thus, to find the initial conditions for $k_2 = 7$ in the preceding example, set

$$P_0(z) = z(z^2 \oplus z \oplus 1),$$

which results in

$$a\,(0) = 0$$
$$a\,(1) = 0$$
$$a\,(2) = 1$$
$$a\,(3) = 1$$
$$a\,(4) = 1$$

from the expression for $P_0(z)$ of Equation (4). This initial condition yields the output sequence

$$0\ 0\ 1\ 1\ 1\ 0\ 1$$

The remaining six initial conditions which yield this sequence can be obtained from the sequence itself; they are

$a\,(0)$	$a\,(1)$	$a\,(2)$	$a\,(3)$	$a\,(4)$
0	1	1	1	0
1	1	1	0	1
1	1	0	1	0
1	0	1	0	0
0	1	0	0	1
1	0	0	1	1

Those initial conditions that yield the 21-bit sequence may be found by setting $P_0(z) = z$. This is a valid initial-condition polynomial that can not be factored by the square or cubic portion of $S(z)$, and hence must place the register in its $k = 21$ mode.

Consider two nonprime system functions, $S_1(z)$ and $S_2(z)$, such that

$$S_1(z) = A(z)B(z),$$

$$S_2(z) = A(z)C(z),$$

and

$$P_1(z) = \frac{(z^{k_1} \oplus 1)\,P_{01}(z)}{A(z)B(z)}, \tag{13}$$

$$P_2(z) = \frac{(z^{k_2} \oplus 1)\,P_{02}(z)}{A(z)C(z)}. \tag{14}$$

An initial condition of $P_{01}(z)$ which will factor $B(z)$ is

$$P_{01}{}'(z) = ZB(z).$$

Similarly,

$$P_{02}{}'(z) = ZC(z).$$

Rewriting Equations (13) and (14),

$$P_1(z) = \frac{(z^{k_1} \oplus 1)z}{A(z)},$$

$$P_2(z) = \frac{(z^{k_2} \oplus 1)z}{A(z)},$$

but, since only one k value exists for $A(z)$,

$$k_1 = k_2.$$

Therefore,

$$P_1(z) = P_2(z),$$

i.e., the same output sequence is generated by a given factor of $S(z)$ regardless of the other factors of $S(z)$.

SYNTHESIS

Given a binary sequence of repetition period k, it is desired to synthesize a system of the type shown in Figure 1, which will repetitively generate the given sequence. Rearranging Equation (8),

$$P_0(z) = \frac{S(z)P(z)}{Z^k \oplus 1}.$$

It is observed that $P(z)$ cannot be factored by $z^k \oplus 1$. Therefore, to satisfy the conditions for $P_0(z)$, $z^k \oplus 1$ must be factorable by $S(z)$ or must contain a factor common to $S(z)$. Consequently, $z^k \oplus 1$ may not be a prime polynomial, and its prime factors may be found. Any combination of these prime factors may wholly describe $S(z)$ or may be contained in $S(z)$. If $S(z)$ contains factors not common to $z^k \oplus 1$, the uncommon factors determine the initial conditions required to obtain the output sequence but do not affect the repetition period, k.

Having found the prime factors of $z^k \oplus 1$, all combinations of these factors which satisfy the conditions,

$$k \geq s,$$

$$2^s - 1 \geq k$$

are considered to be potential $S(z)$ polynomials or factors of potential $S(z)$ polynomials. To obtain the smallest possible register, the lowest

possible values of s are considered first. Each potential $S(z)$ is tested by attempting to generate the given sequence $P(z)$ with the initial conditions specified in $P_0(z)$. The initial conditions are specified by the first s bits in the output sequence. If the test fails, another potential $S(z)$ is tried until $s = k$. The value of $s = k$ results in a shift register initially loaded with the entire output sequence which is recirculated in the register by direct feedback from output to input. Values of s can be initially tested by checking that no group of s successive bits repeat in the output sequence.

Let
$$z^k \oplus 1 = S(z)A(z).$$

Then
$$P_0(z) = \frac{S(z)P(z)}{S(z)\ A(z)} = \frac{P(z)}{A(z)}.$$

$A(z)$ must be a factor of $P(z)$ to obtain a realizable $P_0(z)$. The maximum possible exponent of $P(z)$ is k, whereas the highest exponent of $A(z)$ is $k - s$. Therefore, the division of $A(z)$ into $P(z)$ results in a maximum possible exponent of s, which is a criterion for a realizable $P_0(z)$. Hence if $A(z)$ is a factor of $P(z)$, a realizable $P_0(z)$ must result. Indeed, an initial condition is given by the result of the division of $P(z)$ by $A(z)$. Thus $z^k \oplus 1$ and $P(z)$ must be examined for common factors. The common factor with the largest exponent results in the smallest possible value of s and therefore the smallest possible register required to generate the specified sequence. Example: Let it be required to generate the following sequence:

$$00011010001101 - - - - - - - - - -.$$

For this sequence $k = 7$, and

$$P(z) = z^4 \oplus z^3 \oplus z = z(z^3 \oplus z^2 \oplus 1)$$

$$z^7 \oplus 1 = (z \oplus 1)(z^3 \oplus z \oplus 1)(z^3 \oplus z^2 \oplus 1).$$

The common factor is $z^3 \oplus z^2 \oplus 1$. Therefore,

$$S(z) = z^4 \oplus z^3 \oplus z^2 \oplus 1$$

$$P_0(z) = z,$$

$$s = 4.$$

CONCLUSIONS

One necessary and one sufficient condition have been established for the generation of linear binary sequence of most efficient (largest possible under circuit restrictions) period.

Necessary:

$$S(z) \text{ is prime.}$$

Sufficient:

$$S(z) \text{ is prime,}$$
$$2^s - 1 \text{ is prime.}$$

Permissible periods were determined by means of factors of the system function polynomial. Actual periods were obtained under division. Actual repeating patterns were obtained by a substitution of a derived valid initial condition into the configuration equation. A method for the synthesis of a specific period as well as a particular pattern was outlined. Efficiency is guaranteed by establishing the common factor of $z^k \oplus 1$ and $P(z)$ with the greatest exponent.

REFERENCES

A. N. Radchenko and V. I. Filippov, "Shift Registers With Logical Feedback and Their Use as Counting and Coding Devices," *Automation and Remote Control*, Vol. 20, No. 11, p. 1467, Nov. (1959).

H. R. Charney, "An Algebraic Approach to Sequential Switching Circuit Problems," *Polytechnic Institute of Brooklyn, MRI 811-60*, June 1960.

N. Zierler, "Linear Recurring Sequences," *Group Report 34-63* (Revision) *Lincoln Labs, MIT*, Lexington, Mass., Dec. 2, 1957.

STATISTICAL PROPERTIES OF M-ARY FREQUENCY-SHIFT-KEYED AND PHASE-SHIFT-KEYED MODULATED CARRIERS*

By

H. Zabronsky

RCA Surface Communications Division,
New York, N. Y.

Summary—This paper presents methods for determining the statistical properties of frequency-shift-keyed and phase-shift-keyed modulated carriers through a limiter plus other devices, in the presence of white Gaussian noise. Closed form expressions are obtained for the transition probabilities in mistaking one signal for another.

A nonlinear element such as a limiter offers unusual analytic difficulties. It is hoped that the methods presented in this paper will be useful in the solution of other problems involving signals plus noise through nonlinear devices.

INTRODUCTION

IN Part I of this paper, a study is made of the statistical properties of an M-ary frequency-shift-keyed (FSK) modulated carri.r with additive noise, transmitted through a nonlinear system comprised of an i-f amplifier followed by a limiter and a discriminator. The modulating signal consists of a random sequence of equally spaced pulses, where each pulse is multiplied by some modulating frequency ω_i $(i = 1,...M)$. The pulse shape is arbitrary but fixed, while the noise is taken as white Gaussian. Part I may be regarded as an extension of a paper by S. O. Rice.[1]

The joint probability density function of the input modulating signal and the instantaneous frequency deviation at the output is determined in closed form for each instant of time. Using this result, the transition probabilities, at each time, of mistaking a frequency ω_i for a frequency ω_j are obtained in terms of the signal-to-noise ratio and the bandwidth of the i-f amplifier.

---

* Manuscript received 1 August 1961.

[1] S. O. Rice, "Statistical Properties of a Sine Wave Plus Random Noise," *Bell Sys. Tech. Jour.*, Vol. 27, p. 109, January, 1948.

In Part II, results are obtained for the statistical properties of M-ary differentially coherent and coherent phase-shift-keyed (PSK) modulated carriers with additive noise transmitted through an i-f amplifier followed by a limiter. The modulating signal consists of a random sequence of equally spaced phase angles b_i $(i = 1,...M)$. The noise is again taken to be white Gaussian. In the differentially coherent case, the joint probability density function of the modulating signal and the phase difference, determined by sampling the phase in two adjacent pulse intervals, is obtained. This result is then used to determine the transition probabilities of mistaking a phase b_i for a phase b_j. In the case of coherent detection, the joint probability density function of the modulating signal and the phase is obtained. This result is used to determine the transition probabilities of mistaking a phase b_i for a phase b_j.

Furthermore, the results in Part II are used to obtain results for the case of an M-ary FSK-modulated carrier with additive noise transmitted through an i-f amplifier followed by a limiter, discriminator, and integrator.

In all four cases, equations are derived for the error rate as a function of the signal-to-noise ratio.

In general, the presence of a limiter increases the error rate for a given signal-to-noise ratio in comparison with linear devices which lack the limiter. In certain cases there is little or no difference but in no case is there an improvement. The essential steps in the derivations are given in the hope that the methods presented may find application in future problems dealing with nonlinear devices.

Part I—FSK Through a Narrow-Bandpass Filter Followed by a Limiter and Discriminator

The current in which we are interested may be written as

$$I(t) = P \cos\left(\omega t + \int_0^t e(t)\,dt\right) + n(t). \tag{1}$$

where P and ω are constants, t is time, $n(t)$ is a Gaussian random process, and $e(t)$ is a random signal, i.e.,

$$e \equiv e(t) = \omega_i F(t), \qquad n\alpha \leqq t < (n+1)\alpha \tag{2}$$

$$(i = 1, \cdots M; \; n = 0, 1, 2, \cdots)$$

where $F(t)$ is an arbitrary function, and a is a fixed interval of time. Let

$$A \equiv A(t) = \int_0^t e(t)\, dt. \tag{3}$$

It is assumed that $\alpha \gg 1/\omega$, so that cos A and sin A are slowly varying functions of time compared to sin ωt and cos ωt. Furthermore, the power spectrum of the noise is assumed to lie within a narrow band about ω. Then, $n(t)$ may be written[2]

$$n(t) = n_c \cos \omega t - n_s \sin \omega t, \tag{4}$$

where

$$n_c = \sum_{n=1}^{N} C_n \cos (\omega_n t - \omega t - \phi_n)$$

$$n_s = \sum_{n=1}^{N} C_n \sin (\omega_n t - \omega t - \phi_n) \tag{5}$$

$\phi_1, \cdots \phi_N$ are angles distributed uniformly over the range $(0, 2\pi)$ and,

$$C_n = [2\omega (f_n) \Delta f]^{\frac{1}{2}},$$
$$\omega_n = 2\pi f_n,$$
$$f_n = n\Delta f_n,$$
$$\Delta f = \frac{1}{T}. \tag{6}$$

$\omega(f)$, the power spectrum of the noise, is taken to be symmetric about $\omega/2\pi$. The finite Fourier series expansion for $n(t)$ is valid for $0 \leq t \leq T$. If $\omega(f)$ vanishes for $f > F$, and N and T are chosen so that $N/T \equiv N\Delta f = F$, then all of the conditions for the approach of the series to a normal law with the given power spectrum are valid as T approaches infinity. Thus,

$$I(t) = (n_c + P \cos A) \cos \omega t - (n_s + P \sin A) \sin \omega t$$
$$= R \cos (\omega t + \theta), \tag{7}$$

[2] S. O. Rice, "Mathematical Analysis of Random Noise," Parts I and II, *Bell Sys. Tech. Jour.*, Vol. 23, July, 1944; Part III, Vol. 24, p. 46, January, 1945.

where

$$n_s + P \sin A = R \sin \theta,$$
$$n_c + P \cos A = R \cos \theta. \tag{8}$$

Thus,

$$R^2 = (n_c + P \cos A)^2 + (n_s + P \sin A)^2,$$
$$\tan \theta = \frac{n_s + P \sin A}{n_c + P \cos A}. \tag{9}$$

The joint probability density function $p(R, \dot{R}, \theta, \dot{\theta}, e)$, is determined in closed form for each instant of time. This is used to obtain the joint density function $p(\dot{\theta}, e)$ of the instantaneous frequency deviation $\dot{\theta}$ and the modulating signal $e(t)$, by successively eliminating the variables R, θ, $\dot{R}$. Let

$$C \equiv n_s + P \sin A = R \sin \theta,$$
$$D \equiv n_c + P \cos A = R \cos \theta \tag{10}$$

Then,

$$\dot{C} = \dot{n}_s + P\dot{A} \cos A = \dot{R} \sin \theta + R \dot{\theta} \cos \theta$$
$$\dot{D} = \dot{n}_c - P\dot{A} \sin A = \dot{R} \cos \theta - R \dot{\theta} \sin \theta$$

In order to find $p(R, \dot{R}, \theta, \dot{\theta}, e)$, the joint probability density function $p(C, D, \dot{C}, \dot{D}, e)$ is found first. The desired density function is then found by a transformation of coordinates.

The density function is obtained by calculating the inverse Fourier transform of the characteristic function (Appendix 1). The following result is obtained

$$f(C, D, \dot{C}, \dot{D}, e) = \frac{1}{(2\pi)^2 \, b_0 b_2} \sum_{j=1}^{M} \delta(e - \omega_j F) \sum_m \frac{m!}{M^{m+1} \, (r_1! \ldots r_M!)}$$
$$\exp\left\{ -\frac{(H_j - C)^2 + (J_j - D)^2}{2b_0} - \frac{(\omega_j J_j F - \dot{C})^2 + (\omega_j H_j F + \dot{D})^2}{2b_2} \right\}, \tag{11}$$

where $r_1 + \cdots + r_M = m$, $\delta(x)$ is the delta function, and

$$H_j = P \sin \left[(r_1 \omega_1 + \ldots r_M \omega_M) \; G_a + \omega_j G_\tau \right],$$

$$J_j = P \cos \left[(r_1 \omega_1 + \ldots + r_M \omega_M) \; G_a + \omega_j G_\tau \right]$$

$$G_t = \int_0^t F(t) \; dt,$$

$$b_n = (2\pi)^n \int_0^\infty \omega(f) \, (f - f_q)^n df$$

$$f_q = \frac{\omega}{2\pi}. \tag{12}$$

H_j and J_j are actually functions of $r_1, \cdots r_M$ appearing in the inner sum on the right side of Equation (11). With this understanding, the notation becomes less cumbersome. The same applies for the symbol h_j where

$$h_j = (r_1 \omega_1 + \cdots + r_M \omega_M) + \omega_j G_\tau,$$

so that $$H_j = P \sin h_j, \quad J_j = P \cos h_j.$$

In accordance with Equations (9) and (10), a transformation of coordinates is made, carrying $C, D, \dot{C}, \dot{D}, e$ into $R, \dot{R}, \dot{\theta}, \theta, e$. It is easily verified that the Jacobian of this transformation is $R^2 \, dR \, d\dot{R} \, d\theta \, d\dot{\theta} \, de$. Furthermore, the variable θ may be integrated out of Equation (11) in view of the identity[3]

$$\frac{1}{2\pi} \int_0^{2\pi} \exp \left[a \cos \theta + b \sin \theta \right] d\theta \equiv I_0 \{ \sqrt{a^2 + b^2} \}. \tag{13}$$

Thus $p(R, \dot{R}, \dot{\theta}, e)$, the density function of the random variables $R, \dot{R}, \dot{\theta}, e$, may be written

---

[3] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Second Edition, London, 1958.

$$p(R, \dot{R}, \dot{\theta}, e) =$$

$$\frac{R^2}{2\pi \, b_0 b_2} \sum_{j=1}^{M} \delta\,(e - \omega_j F)\, I_0 \left\{ P \sqrt{\left(\frac{R}{b_0} + \frac{F\,\omega_j\,R\,\dot{\theta}}{b_2} \right)^2 + \left(\frac{F\,\dot{R}\,\omega_j}{b_2} \right)^2} \right\} \times$$

$$\exp \left[-\frac{P^2 + R^2}{2b_0} - \frac{(PF\omega_j)^2 + \dot{R}^2 + R^2\dot{\theta}^2}{2b_2} \right] \tag{14}$$

The variable θ has been eliminated first. The variable $\dot{R}$ will be integrated out next. $p(R, \dot{\theta}, e)$, the joint probability density function of R, $\dot{\theta}$ and e, is obtained by integrating both sides of Equation (14) from minus infinity to plus infinity. After exchanging the orders of summation and integration and using an integral identity involving Bessel functions, the resulting infinite series is summed by means of Neumann's addition theorem. These details are presented in Appendix II. The following result is obtained:

$$p(R, \dot{\theta}, e) = \int_{-\infty}^{\infty} p(R, \dot{R}, \dot{\theta}, e)\; d\dot{R} =$$

$$\frac{R^2}{Mb_0 (2\pi \, b_2)^{1/2}} \exp \left\{ -\frac{P^2 + R^2}{2b_0} - \frac{R^2\dot{\theta}^2}{2b_2} \right\} \times$$

$$\sum_{j=1}^{M} \delta\,(e - \omega_j F) \exp \left\{ -\frac{(PF\omega_j)^2}{4b_2} \right\} \times$$

$$\sum_{n=0}^{\infty} \epsilon_n\,(-1)^n I_{2n} \left\{ PR \left(\frac{1}{b_0} + \frac{F\omega_j\dot{\theta}}{b_2} \right) \right\} I_n \left\{ \frac{(PF\omega_j)^2}{4b_2} \right\}$$

$$\tag{15}$$

where ϵ_n is Neumann's factor, i.e., $\epsilon_0 = 1$, $\epsilon_n = 2$ for $n > 0$.

The variable R is integrated out next. This may be accomplished by integrating both sides of Equation (15) from zero to infinity with respect to R and employing an identity relating an infinite integral of Bessel functions to the confluent hypergeometric function $_1F_1$ which is then re-expressed in terms of Bessel functions. The resulting infinite series is summed, as before, by means of Neumann's addition

theorem (Appendix III). Thus, finally, the joint density function $p(\dot{\theta}, e)$, of the instantaneous frequency deviation $\dot{\theta}$ and the modulating signal $e(t)$ may be written

$$
p(\dot{\theta}, e) =
$$
$$
\frac{\psi^{1/2} e^{-\rho}}{2M (1 + \psi \dot{\theta}^2)^{3/2}} \sum_{j=1}^{M} \delta(e - \omega_j F) e^{\zeta_j} [I_0(\zeta_j)
$$
$$
+ 2z_j (I_0(\zeta_j) + I_1(\zeta_j))] \tag{16}
$$

where

$$
\psi = b_0/b_2,
$$
$$
\rho = P^2/(2b_0) = \text{signal-to-noise ratio},
$$

$$
\zeta_i = \frac{\rho}{2} \left[\frac{1 + \psi\{\dot{\theta}^2 - (\dot{\theta} - \omega_j F)^2\}}{1 + \psi\dot{\theta}^2} \right]
$$

$$
z_j = \zeta_j - \frac{\rho \psi (F \omega_j)^2}{2}.
$$

Furthermore, the conditional probability $p(\dot{\theta}|s = \omega_j F)$ may be written

$$
p(\dot{\theta}|s = \omega_j F) = p(\dot{\theta}, \omega_j F)/p(s = \omega_j F) =
$$
$$
\frac{\psi^{1/2} e^{-\rho}}{2(1 + \psi \dot{\theta}^2)^{3/2}} e^{\zeta_j} [I_0(\zeta_j) + 2z_j (I_0(\zeta_j) + I_1(\zeta_j))], \tag{17}
$$

since

$$
p(s = \omega_i F) = \frac{1}{M} \qquad (i = 1, \cdots M).
$$

Transition Probabilities

The transition probabilities p_{ij}, at each time, in mistaking a frequency ω_i for a frequency ω_j, may now be determined using Equation (17).

Let p_{ij} denote the probability, at each time, that a transmitted frequency $\omega + \omega_i$ is interpreted as $\omega + \omega_j$ by the receiver. Let the ω_i $(i = 1, \cdots M)$ be so numbered as to constitute a finite monotone increasing set. Define $\omega_0 = -\infty$, $\omega_M = +\infty$. The entire real axis is divided into $M + 1$ intervals (finite or semi-infinite) by the points

ω_i $(i = 1, \cdots M)$. Then the following is an adequate decision procedure. If the sampled frequency $\dot{\theta}$ lies in the interval $(\omega_i + \omega_{i-1})/2$, $(\omega_i + \omega_{i+1})/2$, the frequency will be interpreted as ω_i. Thus,

$$p_{ij} = \int_{(\omega_j + \omega_{j-1})/2}^{(\omega_j + \omega_{j+1})/2} p(\dot{\theta}|\omega_i F) \ d\dot{\theta}$$

$$= \frac{e^{-\rho}}{2} \int_{(\phi_j + \phi_{j-1})/2}^{(\phi_j + \phi_{j+1})/2} e^{\zeta_i} \left[I_0\left(\zeta_i\right) + 2z_i\left\{I_0\left(\zeta_i\right) + I_1\left(\zeta_i\right)\right\} \right] \frac{du}{(1+u^2)^{3/2}}, \qquad (18)$$

where

$$\phi_i = \omega_i \, \psi^{1/2} \, F,$$

$$\zeta_i = \frac{\rho}{2} \left[1 - \frac{(u - \phi_i)^2}{1 + u^2} \right]$$

$$z_i = \zeta_i + \frac{\rho \phi_i^2}{2}.$$

Binary FSK

In the case $M = 2$, $\omega_2 = -\omega_1$, the transition probabilities may be written

$$p_{11} = p_{22} = \frac{e^{-\rho}}{2} \int_0^\infty e^{\zeta} \left[I_0\left(\zeta\right) + 2z\left\{I_0\left(\zeta\right) + I_1\left(\zeta\right)\right\} \right] \frac{du}{(1+u^2)^{3/2}}. \quad (18a)$$

For an ideal bandpass filter centered on f_q,

$$\omega(f) = \omega_0 \ \text{for} \ f_a < f < f_b,$$

$$= 0 \ \text{elsewhere}.$$

Thus,

$$b_0 = \omega_0 \ (f_b - f_a),$$

$$b_2 = \frac{\pi^2}{3} \omega_0 \ (f_b - f_a)^3.$$

PART II — COHERENT AND DIFFERENTIALLY COHERENT PSK

In this part, statistical properties of M-ary differentially coherent and coherent PSK modulated carriers with additive noise transmitted through an i-f amplifier followed by a limiter are considered.

The output, $v(t)$, of a PSK-modulated sine wave through an i-f amplifier may be written[4]

$$v(t) = e^{j\omega_0 t} [1 + (e^{jb} - 1) h(t)], \tag{19}$$

where ω_0 is the angular frequency of the carrier, b the phase shift, and $h(t)$ the response of the low frequency equivalent of the i-f amplifier to a unit step. Since the transfer function of the associated low-pass i-f amplifier may be written

$$A(\omega) = \exp \left\{ -\frac{\omega^2}{8\pi^2 \sigma_f^2} \right\} \quad , \tag{20}$$

then

$$h(t) = \sigma_f \sqrt{2\pi} \int_{-\infty}^{t} \exp \left\{ -\frac{x^2}{2\sigma_f^2} \right\} dx, \tag{21}$$

where

$$\sigma_f \sigma_t = \frac{1}{2\pi}.$$

σ_f may be expressed in terms of the bandwidth, B, of the i-f as follows:

$$B = 2\pi \sigma_f \sqrt{\pi}. \tag{22}$$

Consequently, the output of the i-f, corresponding to an input of the form $P \cos \omega_0 t$, is

$$s(t) = Ph \cos(\omega_0 t + b) + P(1 - h) \cos \omega_0 t \tag{23}$$

$$= \{Ph \cos b_1 + P(1 - h)\} \cos \omega_0 t - Ph \sin b \sin \omega_0 t,$$

where $h = h(t)$.

---

[4] I. Gumowski, "Transient Response in FM," *Proc. I.R.E.*, Vol. 42, p. 819, May, 1954.

Combining the output of the i-f given above with the noise current flowing in the output of the i-f,

$$s(t) + n(t) = R \cos (\omega_0 t + \phi), \tag{24}$$

where

$$R \cos \phi = Ph \cos b + P(1 - h) + n_c,$$

$$R \sin \phi = Ph \sin b + n_s.$$

$R = R(t)$ is a slowly varying function of time compared with $\cos \omega_0 t$, and $\phi = \phi(t)$ satisfies the inequality $-\pi < \phi \leqq \pi$.

The output of an ideal limiter for the wave represented by Equation (24) as $R \cos (\omega_0 t + \phi)$, is $\phi \equiv \phi(t)$.

Suppose that samples of the phase ϕ are made in two successive pulse intervals. After a certain time the transient in the first interval had died out, so that the wave output at the beginning of the second interval may be written $P \cos (\omega_0 t + b_1)$ where b_1 is the phase shift of the PSK in the first of the intervals. If the phase shift in the second interval is b_2, the output of the i-f may be written

$$s_2(t) = Ph \cos (\omega_0 t + b_1 + b_2) + P(1 - h) \cos (\omega_0 t + b_1). \tag{25}$$

Writing the signal plus noise in the phase envelope form, the following equations are obtained:

$$s_2(t) + n_2(t) = R_2 \cos (\omega_0 t + \phi_2),$$

$$C_2 = R_2 \cos \phi_2 = Ph \cos (b_1 + b_2) + P(1 - h) \cos b_1 + n_{c2}, \tag{26}$$

$$D_2 = R_2 \sin \phi_2 = Ph \sin (b_1 + b_2) + P(1 - h) \sin b_1 + n_{s2}.$$

Rewriting Equation (24),

$$s_1(t) + n_1(t) = R_1 \cos (\omega_0 t + \phi_1),$$

$$C_1 = R_1 \cos \phi_1 = Ph \cos b_1 + P(1 - h) + n_{c1}, \tag{27}$$

$$D_1 = R_1 \sin \phi_1 = Ph \sin b_1 + n_{s1}.$$

The angles ϕ_1 and ϕ_2 satisfy the inequalities $-\pi < \phi_1 \leqq \pi, -\pi < \phi_2 \leqq \pi$. Moreover, since the noise is white Gaussian, $n_{c1}, n_{s1}, n_{c2}, n_{s2}$

are mutually independent, normally distributed random variables with zero mean and variance

$$\psi_0 = \int_0^\infty \omega(f)\, df \tag{28}$$

The conditional probability $p(\psi)$ of the phase difference $\psi = \phi_2 - \phi_1 \bmod 2\pi$, $-\pi < \psi \leqq \pi$, subject to the condition that $b = b_1$ in the first interval and $b = b_2$ in the second interval, in which samples of the phase ϕ are taken, will now be found.

Since n_{c1} and n_{s1} are independent, normally distributed random variables, with zero mean and variance ψ_0, the joint probability density function for the random variables C_1 and D_1 defined in Equations (27) may be written

$$(C_1, D_1) = \frac{1}{2\pi\psi_0} \exp\left[\frac{-\{C_1 - Ph\cos b_1 - P(1-h)\}^2 - \{D_1 - Ph\sin b_1\}^2}{2\psi_0} \right]. \tag{29}$$

$f_1(C_1, D_1)$ may be written as a density function $f(R_1, \phi_1)$ of the random variables R_1, ϕ_1, after transforming the coordinates C_1, D_1 to R_1, ϕ_1 by means of the relations $C_1 = R_1\cos\phi_1, D_1 = R_1\sin\phi_1$. The density function $p_1(\phi_1)$ of the random variable ϕ_1 may be obtained by integrating the joint density function $f(R_1, \phi_1)$ from $R_1 = 0$ to $R_1 = \infty$. The density function $p_2(\phi_2)$ of the random variable ϕ_2 may be found in exactly the same manner. The following results are obtained:

$$P_1(\phi_1) = \frac{1}{2\pi} \exp\left[\frac{-P^2 r_1{}^2}{2\psi_0} \right] +$$

$$\frac{Pr_1\cos(\phi_1 - \theta_1)}{\pi(2\psi_0)^{\frac{1}{2}}} \left[\frac{\sqrt{\pi}}{2} + \frac{\sqrt{\pi}}{2}\operatorname{erf}\left\{ \frac{Pr_1\cos(\phi_1 - \theta_1)}{(2\psi_0)^{\frac{1}{2}}} \right\} \right] \times$$

$$\exp\left[\frac{-P^2 r_1{}^2 \sin^2(\phi_1 - \theta_1)}{2\psi_0} \right],$$

$$P_2(\phi_2) = \frac{1}{2\pi} \exp\left[\frac{-P^2 r_2{}^2}{2\psi_0} \right] +$$

$$\frac{Pr_2 \cos (\phi_2 - \theta_2 - b_1)}{\pi (2\psi_0)^{\frac{1}{2}}} \left[\frac{\sqrt{\pi}}{2} + \frac{\sqrt{\pi}}{2} \operatorname{erf} \left\{ \frac{Pr_2 \cos (\phi_2 - \theta_2 - b_1)}{(2\psi_0)^{\frac{1}{2}}} \right\} \right] \times$$

$$\exp \left[\frac{-P^2 r_2^2 \sin^2 (\phi_2 - \theta_2 - b_1)}{2\psi_0} \right], \tag{30}$$

where the following new constants are defined:

$$r_1 \cos \theta_1 = h \cos b_1 + 1 - h,$$

$$r_1 \sin \theta_1 = h \sin b_1,$$

$$r_2 \cos \theta_2 = h \cos b_2 + 1 - h,$$

$$r_2 \sin \theta_2 = h \sin b_2. \tag{31}$$

ϕ_1 and ϕ_2 are independent random variables defined in the intervals $-\pi < \phi_1 \leqq \pi, -\pi < \phi_2 \leqq \pi$ with probability density functions $p_1(\phi_1)$, $p_2(\phi_2)$ given by Equations (30). The probability density function of the random variable $\psi = (\phi_2 - \phi_1) \bmod 2\pi, -\pi < \psi \leqq \pi$ will now be determined in terms of p_1 and p_2.

The density function $f(x)$, for the random variable $x = \phi_2 - \phi_1$ is given by the convolution integral

$$f(x) = \int_{-\infty}^{+\infty} p_1(y) \, p_2(y + x) \, dy = \int_{-2\pi}^{2\pi} p_1(y) \, p_2(y + x) \, dy, \tag{32}$$

for $-2\pi < x \leqq 2\pi$. Thus,

$$f(x) = \int_{-\pi}^{\pi - x} p_1(y) \, p_2(y + x) \, dy; \ x \geqq 0$$

$$f(x) = \int_{-\pi - x}^{\pi} p_1(y) \, p_2(y + x) \, dy; \ x \leqq 0. \tag{33}$$

Thus, using the symbol P for probability,

$$P \{\phi_2 - \phi_1, (\bmod 2\pi) \leqq x\} = P \{-\pi \leqq \phi_2 - \phi_1 \leqq x\} +$$

$$+ P \{\pi \leqq \phi_2 - \phi_1 \leqq x + 2\pi\}; \qquad x \leqq 0, \tag{34}$$

$$P\left\{\phi_2 - \phi_1 \pmod{2\pi} \leqq x\right\} = P\left\{-\pi \leqq \phi_2 - \phi_1 \leqq x\right\} +$$

$$P\left\{-2\pi \leqq \phi_2 - \phi_1 \leqq x - 2\pi\right\} + P\left\{\pi \leqq \phi_2 - \phi_1 \leqq 2\pi\right\}; \quad x \geqq 0.$$

$$(35)$$

Denoting the density function of the random variable ψ by $p(x)$, we have, on differentiating Equations (34) and (35) with respect to x,

$$p(x) = f(x) + f(x - 2\pi), \text{ for } x > 0$$

$$p(x) = f(x) + f(x + 2\pi), \text{ for } x < 0. \tag{36}$$

Substituting in Equations (33),

$$p(x) = \int_{-\pi}^{\pi - x} p_1(y)\, p_2(y + x)\, dy + \int_{\pi - x}^{\pi} p_1(y)\, p_2(x + y - 2\pi)\, dy, \ (x > 0)$$

$$p(x) = \int_{-\pi - x}^{\pi} p_1(y)\, p_2(y + x)\, dy + \int_{-\pi}^{x - \pi} p_1(y)\, p_2(x + y + 2\pi)\, dy, \ (x < 0).$$

$$(37)$$

Since, in the case at hand, p_1 and p_2 are periodic functions with period 2π,

$$p(x) = \int_{-\pi}^{\pi} p_1(y)\, p_2(x + y)\, dy \qquad \text{for } -\pi < x \leqq \pi. \tag{38}$$

$p_2(x + y)$ is now the function obtained by substituting $x + y$ for ϕ_2 in the second of Equations (30).

For purposes of symmetry, it will be more convenient to consider the density function of the random variable $\xi = \psi - b_2$. Denoting this density function by $\bar{p}(x)$,

$$\bar{p}(x) = p(x + b_2) = \int_{-\pi}^{\pi} p_1(y)\, p_2(x + y + b_2)\, dy$$

$$= \int_{-\pi}^{\pi} p_1(y + b_1)\, p_2(x + y + b_1 + b_2)\, dy$$

$$(39)$$

$$\text{for, } -\pi + b_2 < x \leqq \pi + b_2.$$

It will be convenient to employ the following substitutions:

$$\beta_1 = b_1 - \theta_1, \beta_2 = b_2 - \theta_2$$

$$x = \frac{Pr_1}{(2\psi_0)^{\frac{1}{2}}} , y = \frac{Pr_2}{(2\psi_0)^{\frac{1}{2}}} \qquad (40)$$

Substituting in Equations (30), $p_1(\phi_1 + b_1)$ and $p_2(\phi_2 + b_1 + b_2)$ may be written

$$p_1(\phi_1 + b_1) = \frac{e^{-x^2}}{2\pi} +$$

$$\frac{x \cos (\phi_1 + \beta_1)}{\pi} \left[\frac{\sqrt{\pi}}{2} + \frac{\sqrt{\pi}}{2} \operatorname{erf} \{x \cos(\phi_1 + \beta_1)\} \right] e^{- x^2 \sin^2 (\phi_1 + \beta_1)},$$

$$p_2(\phi_2 + b_1 + b_2) = \frac{e^{-y^2}}{2\pi} +$$

$$\frac{y \cos (\phi_2 + \beta_2)}{\pi} \left[\frac{\sqrt{\pi}}{2} + \frac{\sqrt{\pi}}{2} \operatorname{erf} \{y \cos (\phi_2 + \beta_2)\} \right] e^{- y^2 \sin^2 (\phi_2 + \beta_2)}.$$

$$(41)$$

$p_1(\phi_1 + b_1)$ and $p_2(\phi_2 + b_1 + b_2)$ are expanded in Fourier series in $\phi_1 + \beta_1$ and $\phi_2 + \beta_2$, respectively. This is explained in detail in Appendix IV. The following results are obtained:

$$p_1(\phi_1 + b_1) = \sum_{n=o}^{\infty} A_n \cos n(\phi_1 + \beta_1)$$

$$A_o = \frac{1}{2}\pi$$

$$A_{2n} = \frac{x^{2n} e^{- x^2} n!}{\pi (2n)!} \{{}_1F_1(n + 1; 2n + 1; x^2) - 1\}$$

$$A_{2n+1} = \frac{xe^{- x^2/2}}{2\sqrt{\pi}} \left[I_n\left(\frac{x^2}{2} \right) + I_{n+1}\left(\frac{x^2}{2} \right) \right]$$

$$(42)$$

$$p_2(\phi_2 + b_1 + b_2) = \sum_{o=u}^{\infty} B_n \cos n(\phi_2 + \beta_2),$$

$$B_o = \frac{1}{2\pi} .$$

$$B_{2n} = \frac{y^{2n} e^{-y^2} n!}{\pi (2n)!} \{ {}_1F_1(n+1; 2n+1; y^2) - 1 \}$$

$$B_{2n+1} = \frac{y e^{-y^2/2}}{2\sqrt{\pi}} \left[I_n\left(\frac{y^2}{2}\right) + I_{n+1}\left(\frac{y^2}{2}\right) \right] \qquad (43)$$

Equations (42 and 43) may be written

$$p_1(y + b_1) = \frac{1}{2\pi} + \sum_{m=1}^{\infty} A_m \cos my \cos m\beta_1 - A_m \sin my \sin m\beta_1,$$

$$(x + y + b_1 + b_2) = \frac{1}{2\pi} + \sum_{m=1}^{\infty} B_m \cos my \cos m(x + \beta_2) - B_m \sin my \sin m(x + \beta_2).$$

Applying Parseval's theorem,

$$\bar{p}(x) = \frac{1}{2\pi} + \pi \sum_{m=1}^{\infty} A_m B_m \cos m(\beta_2 - \beta_1 + x) \qquad (44)$$

Error Rate

The probability density function $\bar{p}(x)$, obtained in Equation (44) will now be used to obtain the transition probabilities in mistaking one signal for another, in the most important cases.

For $b_2 = -b_1 = \pi/2$, the transition probabilities $p_{ij}(i, j = 1, 2)$ may be written,

$$p_{ii} = \int_{-\pi/2}^{\pi/2} \bar{p}(x) dx, i = 1, 2$$

$$p_{ij} = 1 - p_{ii}, i \neq j \qquad (45)$$

In this case, moreover, $\theta_2 = -\theta_1$, and $\beta_2 - \beta_1 = (b_2 - b_1) - (\theta_2 - \theta_1) = \pi - 2\theta_2$. Thus,

$$\cos m(\pi + x - 2\theta_2) = (-1)^m \cos m(x - 2\theta_2),$$

$$\int_{-\pi/2}^{\pi/2} \cos m(\pi + x - 2\theta_2) dx = \frac{2(-1)^{n+1}}{2n+1} \cos 2(2n+1)\theta_2, \ (m = 2n+1)$$

$$= 0, \ (m = 2n).$$

Thus,

$$\int_{-\pi/2}^{\pi/2} \bar{p}(x) dx = \frac{1}{2} - \frac{re^{-r}}{2} \sum_{n=o}^{\infty} \left[I_n\left(\frac{r}{2}\right) + I_{n+1}\left(\frac{r}{2}\right) \right]^2 \frac{(-1)^n}{2n+1} \cos 2(2n+1)\theta_2,$$

where,

$$r = \frac{p^2 r_i^2}{2\psi_0} = \rho[h^2 + (1-h)^2]. \tag{46}$$

From Equation (46), the following equation for p_{ii} is obtained (Appendix V):

$$p_{ii} = \frac{1}{2} - \frac{re^{-r}}{4} \int_{2\theta_2 - \frac{\pi}{2}}^{2\theta_2 + \frac{\pi}{2}} [\cos x \, I_0(r \cos x) + I_1(r \cos x)] dx. \tag{47}$$

The same results are obtained for $b_2 = -b_1 = -\pi/2$. For $b_2 = b_1 = \pm \pi/2$, $\beta_2 - \beta_1 = 0$. Thus, the value of p_{ii} may be obtained by substituting $2\theta_2 = \pi$ in Equation (47). The following result is obtained:

$$p_{ii} = \frac{1}{2} - \frac{re^{-r}}{2} \int_0^{\pi/2} [\cos x \, I_0(r \cos x) + I_1(r \cos x)] dx. \tag{48}$$

Equation (48) may be simplified (Appendix VI.) to obtain the final result:

$$p_{ii} = 1 - \frac{e^{-r}}{2} ; \quad i = 1, 2$$

$$p_{ij} = \frac{1}{2} e^{-r} ; \quad i \neq j. \tag{49}$$

Thus, the transition probabilities have been determined for any sampling time, and expressed in a fairly simple form. If samples of the phase are made near the ends of the pulse intervals, we may assume

that $h(t) = 1$. This implies that $r_1 = r_2 = 1$, $\theta_1 = b_1$, $\theta_2 = b_2$, $\beta_1 - \beta_2 = 0$, $x^2 = y^2 = P^2$. Consequently, in all cases,

$$p_{ii} = 1 \frac{e^{-\rho}}{2} \; ; \quad i = 1, 2$$

$$p_{ij} = \frac{1}{2} e^{-\rho} \; ; \quad i \neq j. \tag{50}$$

The more general case of 2^s equally spaced PSK signals, where s is any positive integer, is treated next. Assuming that samples of the phase are made near the ends of the pulse intervals, we may take $h = 1$ and $\beta_2 - \beta_1 = 0$. The density function $\bar{p}(x)$ may be written,

$$\bar{p}(x) = \frac{1}{2\pi} + \pi \sum_{m=1}^{\infty} B_m{}^2 \cos mx. \tag{51}$$

It is assumed that the 2^s signals have the following symmetrical spacing;

$$b_k = \frac{\pi}{2^s} + \frac{k\pi}{2^{s-1}} \; ; \quad k = 0, 1, \cdots 2^s - 1. \tag{52}$$

Due to symmetry, it is seen that the transition probability p_k in mistaking b_r for b_{r+k} is given by,

$$p_k = \int_{b_{k-1}}^{b_k} \bar{p}(x) \, dx \, . \tag{53}$$

In particular, the probability p_o of interpreting a signal correctly may be written

$$p_o = \int_{-b_0}^{b_0} \bar{p}(x) \, dx \, . \tag{54}$$

Carrying out the indicated integration and summing the resulting series, the following result is obtained:

$$
p_k = \frac{1}{2^s} + \frac{\rho e^{-\rho}}{4} \int_{b_{k-1}}^{b_k} [\cos x \, I_0 \, (\rho \cos x) + I_1 \, (\rho \cos x)] \, dx +
$$

$$
\frac{e^{-2\rho}}{\pi} \sum_{n=1}^{\infty} \left\{ \frac{n!}{(2n)!} \right\}^2 \frac{\rho^{2n}}{n} \{ {}_1F_1 \, (n+1; 2n+1; \rho) - 1 \}^2 \sin \frac{n\pi}{2^{s-1}} \cos \frac{nk\pi}{2^{s-2}}
$$

$$(55)$$

Results similar to those derived in Part II of this paper are contained in Reference (5), which deals with the performance of digital phase modulation systems in Gaussian noise in the absence of a limiter. In particular, the probability of error for binary transmission with phase comparison detection given in Equation (10) of Reference (5) coincides with Equation (50) of this paper. Furthermore, the more general Equation (49) is identical in form to Cahn's Equation (10).

Coherent PSK

This case is considerably simpler than the differentially coherent PSK treated in the previous section. Since the methods of the last section apply here, in even simpler form, it will be sufficient to state the main results.

Let M phase angles corresponding to the M signals be written

$$
\frac{\pi}{M}, \frac{\pi}{M} + \left(\frac{2\pi}{M} \right), \ldots, \frac{\pi}{M} + (M-1) \frac{2\pi}{M} . \tag{56}
$$

Then the transition probabilities may be written

$$
p_{ij} = \frac{1}{M} + 2 \sum_{m=1}^{\infty} \frac{A_m}{m} \sin \frac{m\pi}{M} \cos m \left(\frac{2\pi \, |j-i|}{M} + \beta \right) . \tag{57}
$$

---

[5] C. R. Cahn, "Performance of Digital Phase-Modulation Communication Systems," *Trans. I.R.E. PGCS*, Vol. CS-7, No. 1, p. 3, May, 1959.

In particular,

$$p_{ii} = \frac{1}{M} + 2 \sum_{m=1}^{\infty} \frac{A_m}{m} \sin \frac{m\pi}{M} \cos m\beta . \tag{58}$$

where,

$$\beta = b_1 - \theta$$

$$r \cos \theta = h \cos b_1 + 1 - h$$

$$r \sin \theta = h \sin b_1$$

$$x = Pr / (2\psi_0)^{\frac{1}{2}}$$

in accord with the notation of the last section.

For $M = 2$ (binary PSK) the following result is obtained:

$$p_{ii} = \frac{1}{2} + \frac{1}{2} \operatorname{erf} (x \cos \beta). \tag{59}$$

In particular, for a sample made near the end point of a pulse interval we have $\beta = 0$ and $x = \rho^{\frac{1}{2}}$. Hence

$$p_{ii} = \frac{1}{2} + \frac{1}{2} \operatorname{erf} \rho^{\frac{1}{2}} . \tag{60}$$

For quaternary coherent PSK and sampling near the endpoint of a pulse interval,

$$p_{ii} = \frac{\sqrt{\pi}}{2M} \operatorname{erf} \left(\frac{\rho}{2} \right)^{\frac{1}{2}} + \frac{1}{2} \sum_{k=0}^{\infty} \frac{(-1)^k A_{2(2k+1)}}{2k+1} . \tag{61}$$

PART III — FSK THROUGH A NARROW-BANDPASS FILTER FOLLOWED BY A LIMITER, DISCRIMINATOR AND INTEGRATOR

The output of the discriminator, at a given time t, is $\dot{\theta}(t)$. Thus the output of the integrator, operating from t_1 to t_2 is

$$\int_{t_1}^{t_2} \dot{\theta}(t) \, dt = \theta(t_2) - \theta(t_1) . \tag{62}$$

In the absence of noise, the output of the integrator would be

$$\int\limits_{t_1}^{t_2} \dot{\theta}\,(t)\,dt = \int\limits_{t_1}^{t_2} \omega_i\,dt = \omega_i\,(t_2 - t_1)\ . \tag{63}$$

Thus we seek the density function of the random variable, ψ, where

$$\psi = \theta\,(t_2) - \theta\,(t_1) - \omega_i\,(t_2 - t_1)\ \mathrm{mod}\ 2\pi\ . \tag{64}$$

$$(-\pi \leqq \psi \leqq \pi)\ .$$

This is exactly the problem that was solved in Part II. Therefore, the density function of the random variable ψ may be written

$$f(u) = \frac{1}{2\pi} + \frac{\rho e^{-\rho}}{4} \sum_{m=0}^{\infty} \left[\ I_m\left(\frac{\rho}{2}\right) + I_{m+1}\left(\frac{\rho}{2}\right)\right]^2 \cos\,(2m+1)\ u$$

$$+ \frac{e^{-2\rho}}{\pi} \sum_{m=1}^{\infty} \frac{\rho^{2n}}{\left[\dfrac{(2n)\ !}{n\ !}\right]^2} \{{}_1F_1\ (n+1;2n+1;\rho) - 1\}^2 \cos 2mu, \tag{65}$$

where $\rho = x^2 = P^2/(2\psi_o) = $ signal-to-noise ratio.

Writing $\Delta = t_2 - t_1$ for the integration time, and assuming that the modulating frequencies are

$$f_i = -\frac{b}{2} + \frac{(i-1)}{n-1};\ \omega_i = 2\pi f_i \tag{66}$$

where b is the total bandwidth in cycles per second, we may define a criterion for correct reception of a signal as follows. If $\theta_2 - \theta_1$ lies in the interval $[\frac{1}{2}\,(\omega_i + \omega_{i-1}),\ \frac{1}{2}\,(\omega_i + \omega_{i+1})]$, the signal is interpreted as ω_i. Thus the error rate may be written

$$\int\limits_{-\pi b\Delta/(n-1)}^{\pi b\Delta/(n-1)} f(u)\,du = 2 \int\limits_{0}^{\pi b\Delta/(n-1)} f(u)\,du, \tag{67}$$

since $f(u)$ is an even function.

Choosing a value for Δ equal to half the reciprocal of the bandwidth, we may write,

$$p_{ii} = 2 \int_0^{\pi/2(n-1)} f(u)\,du. \tag{68}$$

Substituting Equation (65) into Equation (68) and carrying out the integration, the following equation for the error rate is obtained:

$$p_{ii} = \frac{1}{2(n-1)} + \frac{\rho e^{-\rho}}{2} \sum_{m=0}^{\infty} \frac{\left[I_m\left(\dfrac{\rho}{2}\right) + I_{m+1}\left(\dfrac{\rho}{2}\right) \right]^2}{2m+1} \sin \frac{(2m+1)\pi}{2(n-1)}$$
$$+ \frac{e^{-2\rho}}{\pi} \sum_{m=1}^{\infty} \frac{\rho^{2m}}{m\left[\dfrac{(2m)!}{m!}\right]^2} \left\{ {}_1F_1\left(m+1; 2m+1; \rho\right) - 1\right\}^2 \sin\left(\frac{m\pi}{n-1}\right). \tag{69}$$

For binary FSK $(n=2)$

$$p_{ii} = \tfrac{1}{2} + \frac{\rho e^{-\rho}}{2} \sum_{m=0}^{\infty} \left[I_m\left(\frac{\rho}{2}\right) + I_{m+1}\left(\frac{\rho}{2}\right) \right]^2 \frac{(-1)^m}{2m+1}. \tag{70}$$

For quaternary FSK $(n=4)$

$$p_{ii} = \frac{1}{6} + \frac{\rho e^{-\rho}}{2} \sum_{m=0}^{\infty} \frac{\left[I_m\left(\dfrac{\rho}{2}\right) + I_{m+1}\left(\dfrac{\rho}{2}\right) \right]^2}{2m+1} \sin \frac{(2m+1)\pi}{6}$$
$$+ \frac{e^{-2\rho}}{\pi} \sum_{m=1}^{\infty} \frac{\rho^{2m}}{m\left[\dfrac{(2m)!}{m!}\right]^2} \left\{ {}_1F_1\left(m+1; 2m+1; \rho\right) - 1\right\}^2 \sin \frac{m\pi}{3}. \tag{71}$$

ACKNOWLEDGMENT

APPENDIX I

The joint density function $f(C, D, \dot{C}, \dot{D}, e)$ is obtained by calculating the inverse Fourier transform of the characteristic function in the following manner:

$$\text{ave. exp } i(sC + rD + u\dot{C} + v\dot{D} + we) =$$

$$\text{ave. exp } i(sn_s + rn_c + u\dot{n}_s + v\dot{n}_c) \times$$

$$\text{ave. exp } (iP\ [s\sin A + r\cos A + ue\cos A - ve\sin A] + iew)$$

Now from Reference (1), p. 16,

$$\text{ave. exp } i(sn_s + rn_c + u\dot{n}_s + v\dot{n}_c)$$

$$= \exp\ [-\tfrac{1}{2}\ (b_0\{r^2 + s^2\} + b_2\{u^2 + v^2\})\ ],$$

where

$$b_n = (2\pi)^n \int_0^\infty \omega(f)\ (f - f_q)^n\ df,$$

$$f_q = \omega/2\pi.$$

Assuming that $P(e = \omega_i F) = 1/M$ $(i = 1, \dots M)$, where P denotes probability, and writing $t = ma + \tau$; $o \leqq \tau < a$, the joint density $f(A, e)$ may be written

$$f(A, \omega_i F) = \frac{m!}{M^{m+1}\ (r_1!\ r_2!\ \dots\ r_M!)},$$

where

$$A = (r_1\omega_1 + \dots + r_M\omega_M)\ G_a + \omega_i G\tau$$

for each i, $i, = 1, \dots M$ and $r_1 + \dots r_M = + m$.   Thus,

$$\text{ave. exp } \{iP[s\sin A + r\cos A + ue\cos A - ve\sin A] + i_\omega e\} = \sum_{j=1}^{M} \sum_{m}$$

$$\frac{m!}{M^{m+1}(r_1!\dots r_M!)}\ \exp[iP\{(s - \omega_j Fv)\sin[(r_1\omega_1 + \dots + r_M\omega_M)\ G_a + \omega_j G_\tau] +$$

$$+ (r + \omega_j Fu)\ \cos\ [(r_1\omega_1 + \dots + r_M\omega_M)\ G_a + \omega_j G_\tau]\} + i_\omega\ \omega_j F].$$

Writing the density function $f(C, D, \dot{C}, \dot{D}, e)$ as the inverse Fourier transform of the characteristic function,

$$f(C, D, \dot{C}, \dot{D}, e) = \frac{1}{(2\pi)^5} \int\int\int\int\int \exp\ [-isC - irD - iu\dot{C} - iv\dot{D} - i_\omega e]$$

$\exp\left[-\frac{1}{2}\left(b_0\{r^2+s^2\}+b_2\{u^2+v^2\}\right)\right]$ ave. $\exp\{iP[s\sin A + r\cos A + ue\cos A - ve\sin A] + i\omega e\}\ drdsdudvd\omega,$

and employing the identities

$$\int_0^\infty e^{-a^2x^2}\cos bx dx = \frac{\sqrt{\pi}}{2a}\, e^{-b^2/(4a^2)}\ (a>0)$$

$$\frac{1}{2\pi}\int_{-\infty}^\infty e^{i\omega(-e+\omega_j F)}\,d\omega = \delta\,(e-\omega_j F)$$

the expression for $f(C, D, \dot{C}, \dot{D}, e)$ reduces to Equation (11) of the text.

APPENDIX II

In order to integrate the variable $\dot{R}$ out of Equation (14), the following identity is employed:

$$\int_{-\infty}^\infty I_o(\sqrt{x^2+y^2})e^{-a^2x^2}\,dx =$$

$$\frac{\sqrt{\pi}\ e^{1/(8a^2)}}{\alpha}\,[I_o(y)I_o(1/(8\alpha^2)) + 2\sum_{n=1}^\infty (-1)^n\,I_{2n}(y)I_n(1/(8\alpha^2))\,].\ (72)$$

The above identity may be seen in the following way. According to Neumann's expansion theorem[3],

$$I_o(\sqrt{x^2+y^2}) = I_o(x)I_o(y) + 2\sum_{n=1}^\infty (-1)^n\,I_{2n}(x)I_{2n}(y).$$

Furthermore (Ref. (6), Vol. 2, p. 92),   (73)

$$\int_{-\infty}^\infty e^{-a^2x^2}\,I_{2n}(x)\,dx = \frac{\sqrt{\pi}}{\alpha}\,e^{1/(8a^2)}\,I_n(1/(8\alpha^2)).$$ (74)

---

[6] Bateman Manuscript Project, *Higher Transcendental Functions*, Vols. 1 and 2, McGraw-Hill Book Co., Inc., New York, N. Y., 1953.

Thus, substituting Equation (73) into the left hand side of Equation (14) and integrating term by term, Equation (15) is obtained.

Appendix III

In order to integrate the variable R out of Equation (15) the following identity is employed (Ref. (7), p. 198):

$$\int_0^\infty e^{-a^2x^2} x^{\mu-1} J_\nu(bx)\, dx = \frac{b^\nu\, \Gamma\left(\dfrac{\mu+\nu}{2}\right)\, {}_1F_1\left(\dfrac{\mu+\nu}{2}; \nu+1; \dfrac{-b^2}{4a^2}\right)}{2^{\nu+1}\, \alpha^{\mu+\nu}\, \Gamma(\nu+1)}\; .$$

(75)

For $\mu = 3$, $\nu = 2n$, $b = i$,

$$\int_0^\infty x^2 \exp\{-\alpha^2 x^2\}\, I_{2n}(x)\, dx =$$

(76)

$$\frac{\Gamma\left(\dfrac{2n+3}{2}\right)}{2^{2n+1}\,\Gamma(2n+1)\,\alpha^{2n+3}}\, {}_1F_1\left(\dfrac{2n+3}{2}; 2n+1; \dfrac{1}{4\alpha^2}\right),$$

since $J_{2n}(ix) = (-1)^n I_{2n}(x)$.

Furthermore (Ref. (6), Vol. 1, pp. 254 and 265),

$${}_1F_1\left(\dfrac{2n+1}{2}; 2n+1; z\right) = 2^n\, \Gamma(n+1)\, e^{z/2} \left(\dfrac{2}{z}\right)^n I_n\left(\dfrac{z}{2}\right),$$

(77)

$$\frac{d}{dz}\, {}_1F_1\left(\dfrac{2n+1}{2}; 2n+1; z\right) =$$

(78)

$$\frac{(2n+1)}{2z}\left[{}_1F_1\left(\dfrac{2n+3}{2}; 2n+1; z\right) - {}_1F_1\left(\dfrac{2n+1}{2}; 2n+1; z\right)\right],$$

and[3]

---

[7] W. Gröbner and N. Hofreiter, "Bestimmte Integrale," *Zweiter Teil*, Springer-Verlag, Austria, 1950.

$$\frac{d}{dz}\left[\left(\frac{2}{z}\right)^n I_n\left(\frac{z}{2}\right)\right] = \frac{1}{2}\left[\left(\frac{2}{z}\right)^n I_{n+1}\left(\frac{z}{2}\right)\right].$$

(79)

Thus

$${}_1F_1\left(\frac{2n+3}{2}; 2n+1; z\right) =$$

$$2^n\,\Gamma(n+1)\,e^{z/2}\left(\frac{2}{z}\right)^n\left[I_n\left(\frac{z}{2}\right) + \frac{z}{2n+1}\left\{I_n\left(\frac{z}{2}\right) + I_{n+1}\left(\frac{z}{2}\right)\right\}\right].$$

(80)

Substituting Equation (80) into Equation (76) and reducing the gamma functions by means of Legendre's duplication formula,

$$2^{2z-1}\,\Gamma(z)\,\Gamma(z + \tfrac{1}{2}) = \sqrt{\pi}\,\Gamma(2z),$$

(81)

the following equation is obtained:

$$\int_0^\infty x^2 \exp\{-\alpha^2 x^2\}\, I_{2n}(x)\, dx =$$

(82)

$$\frac{(2n+1)\sqrt{\pi}}{4\alpha^3}\, e^{1/(8\alpha^2)}\left[I_n(1/8\alpha^2) + \frac{I_n(1/(8\alpha^2)) + I_{n+1}(1/(8\alpha^2))}{4\alpha^2(2n+1)}\right]$$

Integrating both sides of Equation (15) from 0 to infinity with respect to R, and substituting Equation (82) for each term of the right side, Equation (16) is obtained after an application of Neumann's expansion theorem and several standard recursion identities among Bessel functions.

APPENDIX IV

$p_1(\phi_1 + b_1)$ may be expanded in a Fourier series in $\phi_1 + \beta_1$ in the following way:

$$\frac{\sqrt{\pi}}{2}\, \mathrm{erf}\,\{x\cos(\phi_1 + \beta_1)\} = \int_0^{x\cos(\phi_1+\beta_1)} e^{-v^2}dy$$

(83)

$$= x \cos(\phi_1 + \beta_1) \int_0^1 \exp\{-x^2 \cos^2(\phi_1 + \beta_1)z^2\}\, dz.$$

Thus,

$$x \cos(\phi_1 + \beta_1)\frac{\sqrt{\pi}}{2}\, \mathrm{erf}\,\{x \cos(\phi_1 + \beta_1)\}\, \exp\{x^2 \cos^2(\phi_1 + \beta_1)\} =$$

(84)

$$\frac{x^2\{1 + \cos 2(\phi_1 + \beta_1)\}}{2} \int_0^1 \exp\left\{\frac{x^2(1 - z^2)\lfloor 1 + \cos 2(\phi_1 + \beta_1)\rfloor}{2}\right\}\, dz.$$

Since

$$e^{y\cos\theta} = I_o(y) + 2\sum_{n=1}^{\infty} I_n(y)\cos n\,\theta\,,$$

(85)

$$(\cos\theta)\, e^{y\cos\theta} = I_o'(y) + 2\sum_{n=1}^{\infty} I_n'(y)\cos n\theta,$$

we may obtain the coefficients A_m in the expansion

$$p_1(\phi_1 + b_1) = \sum_{m=0}^{\infty} A_m \cos m(\phi_1 + \beta_1)$$

(86)

by substituting Equation (85) into Equation (84) with appropriate values of y and θ. The following results are obtained:

$$A_o = \frac{e^{-x^2}}{2\pi}\left\{1 + x^2 \int_0^1 \exp\left\{\frac{x^2(1 - z^2)}{2}\right\}\left[I_o\left\{\frac{x^2(1 - z^2)}{2}\right\}\right.\right.$$

$$\left.\left. + I_1\left\{\frac{x^2(1 - z^2)}{2}\right\}\right]\, dz\right\},$$

$$A_{2n+1} = \frac{xe^{-x^2/2}}{2\sqrt{\pi}}\left[I_n\left(\frac{x^2}{2}\right) + I_{n+1}\left(\frac{x^2}{2}\right)\right],$$

$$A_{2n} = \frac{x^2\, e^{-x^2}}{\pi} \int_0^1 \exp\left\{\frac{x^2(1 - z^2)}{2}\right\}\left[I_n\left\{\frac{x^2(1 - z^2)}{2}\right\}\right.$$

$$+ I_n' \left\{ \frac{x^2 (1-z^2)}{2} \right\} \Bigg] dz . \tag{87}$$

The integrals in Equation (87) will now be evaluated. Since[3]

$$I_n(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{x\cos\phi} \cos n\phi \, d\phi, \tag{88}$$

then

$$e^x [I_n(x) + I_n'(x)] = \frac{1}{2\pi} \int_0^{2\pi} e^{x(1+\cos\phi)} (1 + \cos \phi) \cos n\phi \, d\phi$$

$$= \frac{1}{\pi} \int_0^{2\pi} e^{2x\cos^2\frac{\phi}{2}} \cos^2 \frac{\phi}{2} \cos n\phi \, d\phi. \tag{89}$$

Therefore,

$$\int_0^1 \exp \left\{ \frac{x^2(1-z^2)}{2} \right\} \left[I_n \left\{ \frac{x^2(1-z^2)}{2} \right\} + I_n' \left\{ \frac{x^2(1-z^2)}{2} \right\} \right] dz$$

$$= \frac{1}{\pi} \int_0^1 dz \int_0^{2\pi} e^{x^2(1-z^2)\cos^2\theta} \cos^2 \theta \cos 2n\theta \, d\theta. \tag{90}$$

Since

$$\int_0^{2\pi} \cos^{2m} \phi \cos 2n\phi \, d\phi = 2^{-2m} (2\pi) \begin{pmatrix} 2m \\ m+n \end{pmatrix}, \tag{91}$$

and

$$e^{x^2(1-z^2)\cos^2\theta} \cos^2\theta = \sum_{m=o}^{\infty} \frac{\{x^2(1-z^2)\}^m}{m!} \cos^{2m+2} \theta, \tag{92}$$

then

$$\int_0^{2\pi} e^{x^2(1-z^2)\cos^2\theta} \cos^2\theta \cos 2n\theta \, d\theta =$$

$$2\pi \sum_{n=m}^{\infty} \frac{\{x^2(1-z^2)\}^m}{m!} 2^{-2m-2} \begin{pmatrix} 2m+2 \\ m+1+n \end{pmatrix}. \tag{93}$$

Consequently Equation (90) may be written

$$2 \int_0^1 dz \sum_{m=n}^{\infty} 2^{-2(m+1)} \binom{2m+2}{m+1+n} \frac{x^{2m}(1-z^2)^m}{m!} \tag{94}$$

Since

$$\int_0^1 (1-z^2)^m \, dz = \int_0^{\pi/2} \cos^{2m+1}\phi \, d\phi = \frac{2^{2m}(m!)^2}{(2m+1)!}, \tag{95}$$

Equation (94) may be written

$$\sum_{m=n}^{\infty} \frac{x^{2m}(m+1)!}{(m+1+n)!(m+1-n)!} = \frac{n! \, x^{2n-2}}{(2n)!} \, ({}_1F_1(n+1; 2n+1; x^2)-1), \tag{96}$$

where ${}_1F_1$ is a confluent hypergeometric function. Therefore,

$$A_o = \frac{1}{2\pi},$$

$$A_{2n} = \frac{x^{2n} e^{-x^2} n!}{\pi (2n)!} \{{}_1F_1(n+1; 2n+1; x^2) - 1\}. \tag{97}$$

The coefficients B_m of the Fourier series expansion of $p_2(\phi_2 + b_1 + b_2)$ in terms of $\phi_2 + \beta_2$ may be obtained in exactly the same manner.

Appendix V

Equation (47) for p_{ii} is obtained in the following manner:

Since

$$I_n(x) + I_{n+1}(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} e^{x\cos\theta} \cos\frac{\theta}{2} \cos(n+\tfrac{1}{2})\theta \, d\theta, \tag{98}$$

$$[I_n(x) + I_{n+1}(x)]^2 =$$

$$\frac{1}{\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{x(\cos\theta + \cos\phi)} \cos\frac{\theta}{2} \cos\frac{\phi}{2} \cos(n+\tfrac{1}{2})\theta \cos(n+\tfrac{1}{2})\phi \, d\theta \, d\phi \tag{99}$$

Let $\theta - \phi = 2\zeta$, $\theta + \phi = 2\psi$.

Then

$$J = \frac{\partial(\theta, \phi)}{\partial(\zeta, \psi)} = 2, \qquad\qquad d\theta\, d\phi = 2d\zeta\, d\psi$$

$$\cos\theta + \cos\phi = 2\cos\frac{\theta + \phi}{2}\cos\frac{\theta - \phi}{2} = 2\cos\zeta\cos\psi,$$

$$2\cos\frac{\theta}{2}\cos\frac{\phi}{2} = \cos\frac{\theta + \phi}{2} + \cos\frac{\theta - \phi}{2} = \cos\zeta + \cos\psi,$$

$$2\cos(n + \tfrac{1}{2})\theta\cos(n + \tfrac{1}{2})\phi = \cos(2n + 1)\zeta + \cos(2n + 1)\psi.$$

Since the range of integration may be taken to be the rectangle for which $0 \leqq \zeta \leqq \pi, -\pi \leqq \psi \leqq \pi$, then

$$\{I_n(x) + I_{n+1}(x)\}^2 =$$

$$\frac{1}{2\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{2x\cos\zeta\,\cos\psi}\,[\cos\zeta + \cos\psi]\cos(2n + 1)\zeta\, d\zeta\, d\psi.$$

$$(100)$$

Thus,

$$p_{ii} = \frac{1}{2} - \frac{re^{-r}}{4\pi^2}\sum_{n=0}^{\infty}\frac{(-1)^n}{2n + 1}\times$$

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{r\cos\zeta\cos\psi}\,[\cos\zeta + \cos\psi]\,[\cos\,(2n + 1)(\zeta - 2\theta_2)]\, d\zeta d\psi.$$

$$(101)$$

If the order of summation and integration in Equation (4) is inverted, noticing that

$$\sum_{n=0}^{\infty}\frac{(-1)^n\cos(2n + 1)x}{2n + 1} = +\frac{\pi}{4}, \quad -\frac{\pi}{2} < x < \frac{\pi}{2},$$

$$= -\frac{\pi}{4}, \quad -\pi < x < -\frac{\pi}{2},$$

$$= -\frac{\pi}{4}, \qquad \frac{\pi}{2} < x < \pi,$$

then Equation (47) for p_{ii} results.

APPENDIX VI

Equation (48) may be simplified in the following manner (Ref. (3) pp. 328, 374):

$$H_{1/2}(z) = \left(\frac{2z}{\pi}\right)^{1/2} \int_0^{\pi/2} J_1(z \sin\theta)\, d\theta = \left(\frac{2z}{\pi}\right)^{1/2} \int_0^1 \sin zt\, dt.$$

$$(102)$$

Thus,

$$\int_0^{\pi/2} I_1(r \cos\zeta)\, d\zeta = \frac{\cosh r - 1}{r} \tag{103}$$

Furthermore, (Ref. (3) p. 373),

$$\left(\frac{2z}{\pi}\right)^{1/2} \int_0^{\pi/2} J_0(z \sin\theta)\, \sin\theta\, d\theta = J_{1/2}(z) = \left(\frac{2}{\pi z}\right)^{1/2} \sin z. \tag{104}$$

Thus,

$$\int_0^{\pi/2} I_0(r \cos\zeta)\, \cos\zeta\, d\zeta = \frac{\sinh r}{r}. \tag{105}$$

Adding Equations (103) and (105),

$$\int_0^{\pi/2} \{I_0(r \cos\zeta)\, \cos\zeta + I_1(r \cos\zeta)\}\, d\zeta = \frac{e^r - 1}{r}. \tag{106}$$

QUANTIZATION IN COHERENT AND QUADRATURE RECEPTION OF ORTHOGONAL SIGNALS*

By

GILBERT LIEBERMAN

RCA Aerospace Communications and Controls Division,
Camden, N. J.

Summary—A review is given of the calculation of error probability for M-ary digital communication in white Gaussian noise using equal-energy equally probable orthogonal signals. The analysis is given for both coherent and quadrature detection with ideal analog signal processing and for digital processing using time and amplitude quantization. The effects of amplitude quantization are calculated for an arbitrary number of equally spaced intervals (except the end intervals). These effects are shown to be essentially the same for coherent M-ary and quadrature M-ary reception of binary orthogonal signals, under the assumption of small input signal-to-noise ratio, and large signal-bandwidth–duration product. The quantization boundaries are expressed in terms of a single variable K. An expression for the quantization intervals is given as a function of K. From this function, optimum-quantization boundaries giving minimum-quantization loss have been calculated as a function of the number of intervals.

INTRODUCTION

FOR certain digital communication applications, it is desirable to be able to use broad-band signals in conjunction with coherent reception.[1,2] Such a system can be designed to operate reliably for low input signal-to-noise ratios. By employing long-duration signals, the signal energy can be made large enough to meet a given specification on error probability. When a large signal-bandwidth–duration product, WT, is required, where W is small and T is large, situations exist in which it is advantageous to employ digital correlation processing. In order to use digital processing, the receiver inputs must be quantized both in time and in amplitude. Since the transmitter can choose band-limited signals, time-quantization losses can easily be made negligible. In the case of amplitude quantization, it is necessary to specify the manner in which the quantization should be

---

[1] D. Middleton, *Statistical Communication Theory*, McGraw-Hill Book Co., Inc., New York, N. Y., 1960. Contains many references to earlier work.
[2] Matched Filter issue, *Trans. I.R.E. PGIT*, June 1960.

applied, and to determine how this quantization affects the receiver error-probability performance.

An explicit general expression has been derived for the quantization loss in the case of uniform quantization intervals. Given a prescribed number of intervals, quantization boundaries can be determined which minimize the quantization loss. Graphs are presented which show the loss as a function of the choice of boundaries.

The simplicity of the results is due to the assumption of large signal-bandwidth–duration product, WT, and small input signal-to-noise ratio. This permits the use of the "M-position experiment formula"[3-5] for error probability calculations in coherent reception with quantization. For quadrature reception with quantization, the formula given by Reiger[6,7] for M-ary noncoherent frequency-shift keying (FSK) can be used.

GENERAL

Consider an M-ary symmetric communication channel with equal-energy orthogonal signals. That is, there is a distinct signal, $S_j(t)$, which corresponds to each of M equally probable messages. The orthogonality implies that

$$\int_0^T S_i(t) S_j(t)\, dt = T\delta_{ij}. \tag{1}$$

The Kronecker delta, δ_{ij}, is unity for $i = j$, and zero otherwise. When i equals j, Equation (1) expresses an assumption that each of the M signals has equal energy, or equal unit average power in the interval $(0, T)$. When i does not equal j, Equation (1) implies zero interaction energy between pairs of signals.

The receiver r-f input signal plus noise is defined to be

[3] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals*, M.I.T. Rad. Lab. Series No. 24, p. 171, McGraw-Hill Book Co., Inc., New York, N. Y., 1950.

[4] R. H. Urbano, "Analysis and Tabulation of the M Positions Experiment Integral and Related Error Function Integrals," *AFCRC, Electronic Research Directorate TR-55-100*, April, 1955. AD 79394.

[5] V. A. Kotelnikov, *The Theory of Optimum Noise Immunity*, McGraw-Hill Book Co., New York, N. Y., 1959. Translation of a 1947 doctoral thesis by R. A. Silverman.

[6] S. Reiger, "Error Rates in Data Transmission," *Proc. I.R.E.*, Vol. 46, p. 919, May, 1958.

[7] G. L. Turin, "The Asymptotic Behavior of Ideal M-ary Systems," *Proc. I.R.E.*, Vol. 47, p. 93, January, Correspondence, 1959.

$$x_1(t) = AS_i(t) \cos (\omega_0 t) + n_1(t)$$

$$= [AS_i(t) + I_c(t)] \cos (\omega_0 t) - I_s(t) \sin (\omega_0 t). \qquad (2)$$

where $\omega_0 = 2\pi f_0$, and A is the input signal amplitude. This expression states that the i-th transmitted signal multiplies a carrier of frequency f_0, and is combined with additive noise $n_1(t)$. It is further assumed that $n_1(t)$ is stationary, band-limited, zero mean, white Gaussian noise with bandwidth $2W$, spectral density N_0, and mean frequency f_0. It will also be assumed that the signals $S_i(t)$ are band limited in the interval $(0, W)$. The functions $I_c(t)$ and $I_s(t)$ in Equation (2) are the noise inphase and quadrature components, respectively.[8] These components are independent, identically distributed Gaussian processes with the same variance, $2N_0W$, as $n_1(t)$. In an ideal synchronous detector, the signal plus noise, $x_1(t)$, can be translated down to zero frequency before correlation, by low-pass filtering the product of $x_1(t)$ and $\cos (\omega_0 t)$. The resulting signal plus noise, aside from a negligible factor of two, can be written

$$x(t) = AS_i(t) + I_c(t). \qquad (3)$$

The translated signal has mean-square value A^2. The translated noise has mean-square-value $2N_0W$. The noise function $I_c(t)$ is band-limited and white, with spectral density $2N_0$ in the frequency range $(0, W)$. It is interesting to observe that the r-f signal-to-noise average power ratio equals $A^2/(4N_0W)$, whereas the translated signal-to-noise average power ratio equals $A^2/2N_0W$. The translated signal-to-noise ratio is larger by a factor of two due to the elimination of the noise term $I_s(t)$ in the translation process.

Equations (2) and (3) are consistent with each other, however, since the ratio of r-f signal energy, $A^2T/2$, to r-f noise per cycle, N_0, equals the ratio of translated signal energy, A^2T, to translated noise per cycle, $2N_0$. This ratio will be referred to as β.

IDEAL COHERENT RECEPTION

In order to clarify the comparisons which are made later for quantization and quadrature detection, the error probability analysis for ideal coherent reception of orthogonal signals in white Gaussian noise is reviewed here.

---

[8] S. O. Rice, "Mathematical Analysis of Random Noise," Parts I and II, *Bell Sys. Tech. Jour.*, Vol. 23, p. 282, July, 1944; Part III, Vol. 24, p. 46, January, 1945.

If the criterion of maximum a posteriori probability[9] is used, the optimum receiver calculates M "dot-products,"

$$y_j = (X \cdot S_j) = \int\limits_0^T x(t) S_j(t) dt, \qquad (4)$$

and selects the message corresponding to the largest value of y_j. The above dot-product notation will be used to denote the integral of a product of two time functions. The "matched filter" outputs[1,2,10], y_j, can be written

$$y_j = A(S_i \cdot S_j) + (I_c \cdot S_j). \qquad (5)$$

The term matched filter refers to the fact that Equation (4) can be realized using linear filters with unit impulse response $S_j(T-t)$, and that outputs of these filters have maximum peak signal-to-noise mean-square value.[2,10] The term peak signal refers to the filter output at time T, to which the y_j correspond. It follows from Equation (1) that y_j can be written

$$y_j = AT\delta_{ij} + \epsilon_j. \qquad (6)$$

The random variable ϵ_j represents the noise-cross-signal term, $(I_c \cdot S_j)$. ϵ_j has zero mean value, since I_c has zero mean value.

It can easily be shown that the ϵ_j are mutually independent normally distributed random variable with zero mean and variance

$$\sigma^2 = \sigma^2(\epsilon_j) = N_0 T. \qquad (7)$$

Since the signals have been assumed to be equally probable, it follows that the probability of correct reception, q_M, equals the probability that all the $y_j(j \neq i)$ are less than y_i. This probability is [4,5]

$$q_M = \int\limits_{-\infty}^{\infty} \phi\left(t - \frac{m}{\sigma}\right) \left[\Phi(t)\right]^{M-1} dt. \qquad (8)$$

[9] P. M. Woodward, *Probability and Information Theory with Applications to Radar*, Pergamon Press, London, 1953.

[10] H. D. Becker and J. G. Lawton, "Theoretical Comparison of Binary Data Transmission Systems," *Cornell Aeronautical Lab. Inc. Report No. CA-1172-S-1*, May, 1958.

This expression has been referred to as the "M-positions experiment" formula.[3] $\phi(t)$ and $\Phi(t)$ are the standardized normal probability density and distribution functions, respectively· The ratio m/σ is given by

$$\frac{m}{\sigma} = \frac{AT}{\sqrt{N_0 T}} = \sqrt{2\beta.}$$ (9)

As defined earlier, β is the ratio of signal energy to noise spectral density, $A^2 T/2N_0$. A usual notation for β is the ratio E/N_0.

The function in Equation (8) is tabulated in Reference (4) for values of m/σ at intervals of 0.01 from 0 to 0.1, and at intervals of 0.1 from 0.1 to 8.0. The values of M are 2, 3, 4, . . . 20 and 25, 30, 35, . . . 95. When M equals 2, the probability of correct reception as given by Equation (8) becomes

$$q_2 = \Phi \left[\frac{m}{\sqrt{2}\,\sigma} \right] = \Phi \lceil \sqrt{\beta} \rceil.$$ (10)

For large values of β, the probability of error for M-ary reception can be approximated by[11]

$$p_M = 1 - q_M \cong (M-1) \lceil 1 - \Phi(\sqrt{\beta}) \rceil.$$ (11)

This approximation also provides an upper bound for p_M regardless of β. Figure 1 shows the probability of error function p_M as a function of 10 log E/N_0 for various values of M. The data for these curves was obtained from Reference (4).

IDEAL QUADRATURE RECEPTION

Justification for quadrature reception

Suppose the receiver input signal and noise is

$$\begin{aligned}
x_1(t) &= A\, S_i(t) \cos (\omega_0 t + \theta) + n_1(t) \\
&= (A \cos \theta\, S_i(t) + I_c(t)) \cos (\omega_0 t) \\
&\quad - (A \sin \theta\, S_i(t) + I_s(t)) \sin (\omega_0 t).
\end{aligned}$$ (12)

---

[11] C. W. Helstrom, "The Comparison of Digital Communication Systems," *Trans. I.R.E. PGCS*, September, 1960.

This differs from Equation (2) only in that the phase angle θ has been introduced. The angle represents the phase difference between transmitter and receiver local oscillators. Due to this lack of phase synchronization, the translated signal and noise would be

$$x(t) = (A \cos \theta) \, S_i(t) + I_c(t) \tag{13}$$
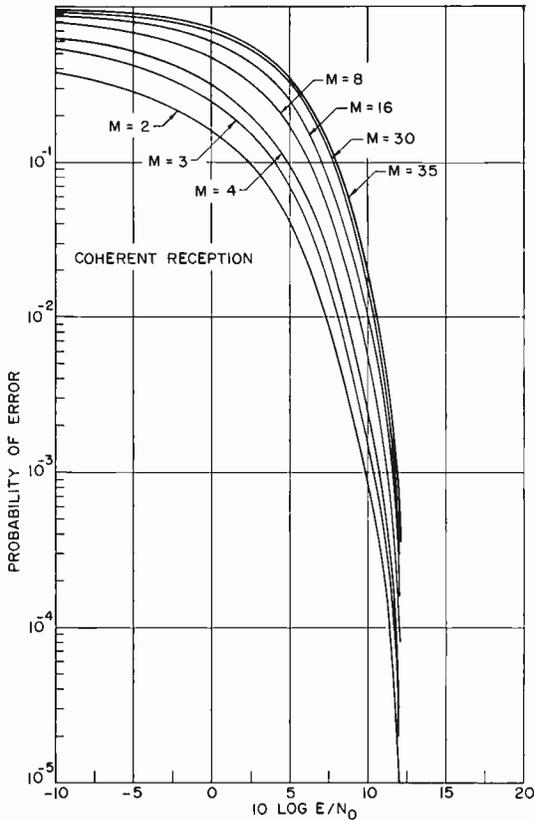
Fig. 1—Error probability curves for coherent M-ary reception.

rather than as shown in Equation (3). As a consequence, if ordinary coherent reception is used, the signal energy will be $E \cos^2 \theta$ rather than E. This will result in a $-20 \log_{10} |\cos \theta|$ decibel loss due to θ. Figure 2 is a plot of the probability distribution function of this loss for a random phase difference which is unformly distributed between 0 and 2π radians. Appendix I contains a derivation of this distribution

function. Note that the 50 per cent point corresponds to a 3-decibel loss, and the 80 per cent point to a 10-decibel loss. For 20 per cent of the time, therefore, we can expect 10 decibels or more loss due to uniformly distributed random phase differences.

Description of Quadrature Reception

In order to eliminate this loss, the signal and noise, $x_1(t)$ is multiplied by a local carrier and by the carrier phase shifted by 90 degrees. Each product wave is then passed through a separate low-pass filter and a set of matched filters. The effect of θ is then reduced by taking the square root of the sum of the squares of pairs of corresponding
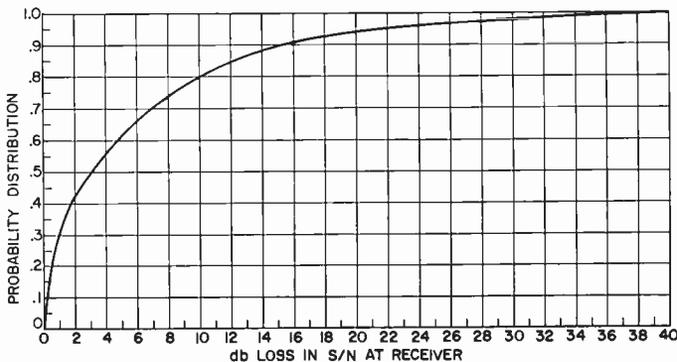
Fig. 2—Distribution function of loss due to random phase asynchronism in coherent reception of orthogonal signals.

matched filter outputs. A block diagram of the quadrature receiver described above is shown in Figure 3. The following is a mathematical description of what occurs.

When the signal plus noise, $x_1(t)$, is multiplied by $\cos(\omega_0 t)$ and $\sin(\omega_0 t)$, respectively, the product functions are (neglecting a scale factor of $\frac{1}{2}$),

$$x^*(t) = \lfloor A \cos\theta \, S_i(t) + I_c(t) \rfloor \tag{14}$$

plus terms with frequency components in the neighborhood of $2\omega_0$, and

$$\hat{x}^*(t) = - [A \sin\theta \, S_i(t) + I_s(t)] \tag{15}$$

plus terms with frequency components in the neighborhood of $2\omega_0$. The effect of the low-pass filtering is to eliminate the high-frequency

terms in the neighborhood of $2\omega_0$. The low-pass filter outputs can therefore be written

$$
\begin{cases}
x^*(t) = AS_i(t) \cos \theta + I_c(t), \\
\hat{x}^*(t) = -AS_i(t) \sin \theta - I_s(t).
\end{cases}
\tag{16}
$$

It follows that $x(t)$ and $\hat{x}(t)$ are statistically independent Gaussian
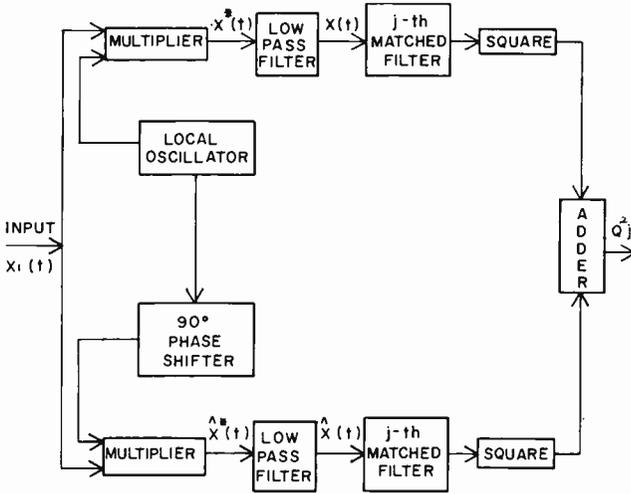
Fig. 3—Block diagram of ideal quadrature detector for j-th signal.

processes. At the output of the j-th matched filters,

$$
\begin{cases}
y_j = (x \cdot S_j) = AT \cos \theta \, \delta_{ij} + \delta_j, \\
\hat{y}_j = (\hat{x} \cdot S_j) = -AT \sin \theta \, \delta_{ij} + \hat{\delta}_j,
\end{cases}
\tag{17}
$$

$$
\begin{cases}
\delta_j = (I_c \cdot S_j), \\
\hat{\delta}_j = -(I_s \cdot S_j),
\end{cases}
\tag{18}
$$

δ_j and $\hat{\delta}_j$ are statistically independent normal random variables with zero mean and variance $\sigma^2 = N_0 T$ (see Equation (7)). The output of the j-th quadrature detector is

$$
Q_j = [y_j^2 + \hat{y}_j^2]^{\frac{1}{2}}.
\tag{19}
$$

The decision process consists of choosing the message corresponding to the largest value of Q_j. The role of quadrature detection in reducing the effect of θ can be seen by applying Equations (17)-(19).

When j is not equal to i, Q_j has a Rayleigh distribution with parameter σ^2. Thus, the probability density of Q_j is

$$f_j(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) . \tag{20}$$

This follows from the fact that Q_j is equal to the square root of the sum of the squares of two independent normal random variables with zero mean and unit variance. When j equals i, the random variables y_i and $\hat{y}_i$ are still independent and normally distributed with variance σ^2, but their mean values are $(AT/2)\cos(\theta)$ and $(-AT/2)\sin\theta$, respectively. The probability density of Q_i is

$$f_i(x) = \frac{x}{\sigma^2} \exp\left[-\frac{1}{2\sigma^2}\left(x^2 + \frac{A^2T^2}{4}\right)\right] I_0\left(\frac{ATx}{2\sigma^2}\right) . \tag{21}$$

This expression is derived[8] by applying a polar coordinate transformation to the joint probability density of y_i and $\hat{y}_i$ and then integrating with respect to the angle variable between 0 and 2π. $I_0(z)$ is the Bessel function of zero order and imaginary argument which is obtained as a consequence of this integration.

If the probability densities in Equations (20) and (21) are normalized by making the transformation $v = x/\sigma$ the probabilities of correct reception remain unchanged. The probability densities become

$$\begin{cases} f_j(v) = v \exp\left(-\frac{v^2}{2}\right) \\ f_i(v) = v \exp\left[-\frac{1}{2}(v^2 + 2\beta)\right] I_0(v\sqrt{2\beta}). \end{cases} \tag{22}$$

The same probability densities arise in the analysis of M-ary non-coherent FSK. The required probabilities of error for M-ary non-coherent FSK reception have been calculated by Reiger in Reference (6). Reiger's formula for the probabilities of error is

$$P_e(M) = \frac{1}{M} \exp(-\beta) \sum_{k=2}^{M} (-1)^k \begin{pmatrix} M \\ k \end{pmatrix} \exp\left(\frac{\beta}{k}\right). \quad (23)$$

The above analysis shows that the use of orthogonal signals in conjunction with ideal quadrature detection leads to the same error probability formulas as for ideal noncoherent FSK.

As in the coherent case (see Equation (11)), it can be shown that for large β,
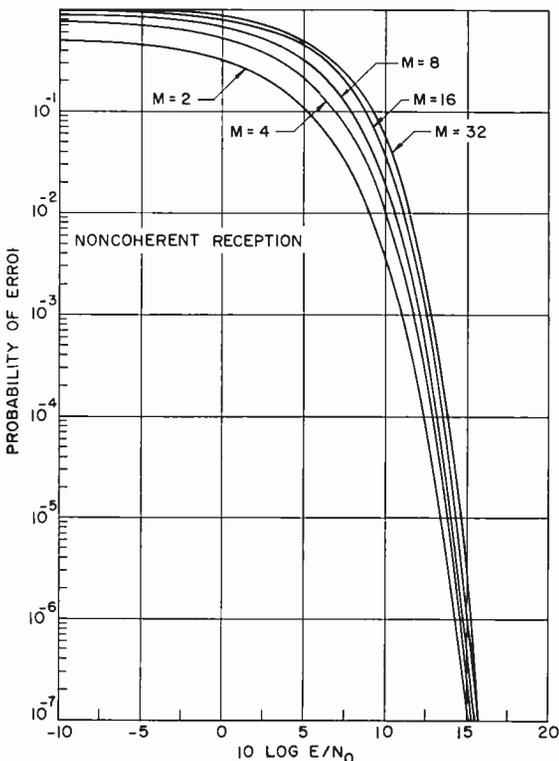
Fig. 4—Error probability curves for noncoherent M-ary reception.

$$P_e(M) \cong (M-1) P_e(2) = \frac{M-1}{2} e^{-\beta/2}, \quad (24)$$

and that this function provides an upper bound for $P_e(M)$ as well. Reiger refers to this in Reference (6). Figure 4 shows curves of $P_e(M)$ as a function of 10 log E/N_0 for various values of M.

Effect of an Arbitrary Number of Quantization Levels on Coherent Reception

In a digital receiver processing system, instead of calculating the exact dot products $(x \cdot S_j)$ as in Equation (4), time quantization can be used, i.e., the band-limited signal plus noise $x(t)$ is sampled periodically at the rate of $2W$ samples per unit time; and, similarly, the band-limited signals $S_j(t)$. The dot products $(x \cdot S_j)$ can then be expressed as[9]

$$(x \cdot S_j) \cong \frac{1}{2W} \sum_{k=1}^{2WT} x\left(\frac{k}{2W}\right) S_j\left(\frac{k}{2W}\right). \qquad (25)$$

The time quantization error will be negligible if the WT product is large. From the definition of $x(t)$ in Equation (3),

$$\begin{aligned}
y_{jk} &= x\left(\frac{k}{2W}\right) S_j\left(\frac{k}{2W}\right) \\
&= AS_i\left(\frac{k}{2W}\right) S_j\left(\frac{k}{2W}\right) + I_c\left(\frac{k}{2W}\right) S_j\left(\frac{k}{2W}\right). \qquad (26)
\end{aligned}$$

A basic assumption in the analysis which follows is that the signals $S_j(t)$ assume only the values ± 1. As a consequence, when $j = i$, the random variables y_{ik} will have mean value A. When $j \neq i$, the y_{jk} will have mean values $\pm A$, depending upon whether there is a polarity agreement or disagreement between the reference sample and the incoming signal sample. From the white Gaussian noise assumption, the y_{jk} are independent Gaussian random variables with variance equal to the average noise power, N.

The process of amplitude quantization consists of quantizing each of the time samples in Equation (25) and then adding them. The "arbitrary" number of quantization levels which are considered here is defined as follows. If the number of quantization intervals r is equal to $2b$, the quantized variables ξ_{jk} can be defined by

$$\xi_{jk} = \begin{cases}
b; & y_{jk} > (b-1)K \\
(b-s); & (b-s-1)K < y_{jk} < (b-s)K \\
-(b-s); & -(b-s)K < y_{jk} < -(b-s-1)K \\
-b; & y_{jk} < -(b-1)K
\end{cases} \qquad (27)$$

$s = 1, 2, 3, \cdots b-1$. K is the quantization boundary constant in units

of the noise r-m-s value. For an even number of intervals, K is the width of the inner intervals; for an odd number of intervals, $2K$ is the width of the inner intervals (see Figure 5).

Similarly, if the number of quantization intervals r is equal to $2b - 1$, the quantized variables ξ_{jk} can be defined by

$$\xi_{jk} = \begin{cases} (b-1) \, ; \, y_{jk} > (2b-3)K \\ (b-s-1) \, ; \, [2(b-s)-3]K < y_{jk} < [2(b-s)-1]K \\ -(b-s-1) \, ; \, -[2(b-s)-1]K < y_{jk} < -[2(b-s)-3]K \\ -(b-1) \, ; \, y_{jk} < -(2b-3)K \end{cases} \tag{28}$$

EVEN NUMBER 2b OF INTERVALS

ODD NUMBER 2b-1 OF INTERVALS
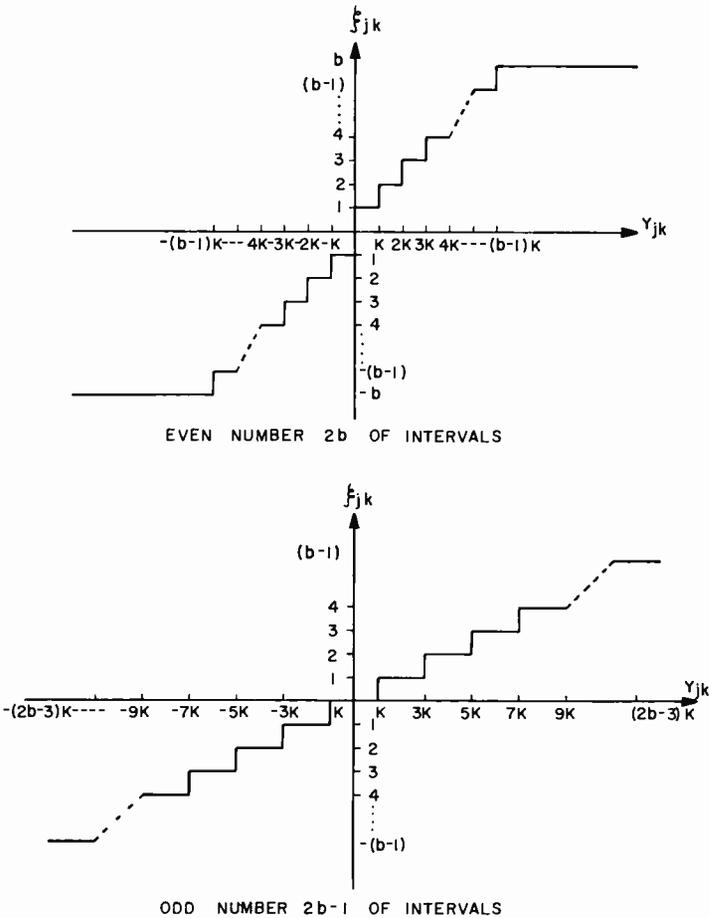
Fig. 5—Quantization transfer characteristics.

$s = 1, 2, 3, \cdots b - 1$. Figure 5 shows the quantization transfer functions defined above.

The j-th quantized matched filter output is given by

$$\xi_j = \frac{1}{2W} \sum_{k=1}^{2WT} \xi_{jk}. \tag{29}$$

While the quantized variables ξ_{jk} have a binomial distribution in the two-level case, they have a multinomial distribution in the general case.[12] The number of categories corresponds to the number of quantization intervals. The associated probabilities for the ξ_{jk} are defined below for an even number of intervals.

$$\begin{cases} p_{s+1} = P[\xi_{jk} = (b - s) \,|\, A] \\ p_{2b-s} = P[\xi_{jk} = -(b - s) \,|\, A] \\ p'_{s+1} = P[\xi_{jk} = -(b - s) \,|-A] \\ p'_{2b-s} = P[\xi_{jk} = (b - s) \,|-A] \end{cases} \tag{30}$$

$s = 0, 1, 2, \cdots, b - 1$. These are the conditional probabilities that the ξ_{jk} assume their respective values under the two possible conditions of positive and negative mean values of y_{jk} equal to $\pm A$.

For an odd number of quantization intervals, the associated probabilities are

$$\begin{cases} q_{s+1} = P[\xi_{jk} = b - s - 1 | A] \\ q_{2b-s-1} = P[\xi_{jk} = -(b - s - 1 | A] \\ q'_{s+1} = P[\xi_{jk} = b - s - 1 |- A] \\ q'_{2b-s-1} = P[\xi_{jk} = -(b - s - 1 |- A] \end{cases} \tag{31}$$

$s = 0, 1, 2, \cdots, b - 1$.

From the symmetry of the normal distribution, it follows that for the even case,

$$p_{s+1} = p'_{2b-s} \; ; \; p_{2b-s} = p'_{s+1} \; ; \tag{32}$$

for the odd case,

$$q_{s+1} = q'_{2b-s-1} \; ; \; q_{2b-s-1} = q'_{s+1} \cdot \tag{33}$$

---

[12] W. Feller, *Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, N. Y., 1950, 1957.

For purposes of comparison with the unquantized case, two basic assumptions will be made; (1) very small receiver input signal-to-noise average power ratio, and (2) a large number $2WT$ of time-quantized samples. As a consequence of the central theorem of probability theory, the ξ_j are approximately normal random variables.[13] Therefore the main problem is to calculate the means and variances of the ξ_j.

From the orthogonality assumption, when j is not equal to i, half of the products

$$S_i \left(\frac{k}{2W} \right) S_j \left(\frac{k}{2W} \right)$$

will be positive, and half will be negative. It follows that the mean values of ξ_j are equal to zero when $j \neq i$ due to the symmetric way the ξ_{jk} have been defined and as a consequence of the symmetry of the normal distribution. That is, for $j \neq i$ in the even case,

$$
\begin{aligned}
E(\xi_j) &= TE(\xi_{jk}) \\
&= \frac{T}{2} \sum_{s=0}^{b-1} (b-s) \left[p_{s+1} - p_{2b-s} + p'_{s+1} - p'_{2b-s} \right] \\
&= 0.
\end{aligned}
\tag{34}
$$

This follows from the identities given in Equation (32). Similarly, in the odd case, for $j \neq i$,

$$
\begin{aligned}
E(\xi_j) &= \frac{T}{2} \sum^{b-1} (b-s-1) \left[q_{s+1} - q_{2b-s-1} + q'_{s+1} - q'_{2b-s-1} \right] \\
&= 0.
\end{aligned}
\tag{35}
$$

This follows from the identities given in Equation (33). In the even case, the mean value of ξ_i is

$$
\begin{aligned}
m' = E(\xi_i) &= TE(\xi_{ik}) \\
&= T \left[b(p_1 - p_{2b}) + \sum_{s=1}^{b-1} (b-s)(p_{s+1} - p_{2b-s}) \right].
\end{aligned}
\tag{36}
$$

In the odd case, the mean value of ξ_i is

---
[13] H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J., 1946.

$$m' = T \left[(b-1)(q_1 - q_{2b-1}) + \sum_{s=1}^{b-2} (b-s-1)(q_{s+1} - q_{2b-s-1}) \right] \cdot$$

$$\text{(37)}$$

In the even case, the variance of the ξ_j for $j \neq i$ is given by

$$(\sigma')^2 = \sigma^2[\xi_j] = \frac{T}{2W} E(\xi_{jk}{}^2)$$

$$= \frac{T}{2W} \sum_{s=0}^{b-1} (b-s)^2 (p_{s+1} + p_{2b-s}). \tag{38}$$

In the odd case,

$$(\sigma')^2 = \sigma^2(\xi_j)$$

$$= \frac{T}{2W} \sum_{s=0}^{b-2} (b-s-1)^2 [q_{s+1} + q_{2b-s}]. \tag{39}$$

The variance of ξ_i is given by

$$\sigma^2[\xi_i] = \frac{T}{2W} \sigma^2(\xi_{ik})$$

$$= \frac{T}{2W} E(\xi_{ik}{}^2) - \frac{T}{2W} [E(\xi_{ik})]^2$$

$$= \sigma^2[\xi_j] - \frac{T}{2W} [E(\xi_{ik})]^2. \tag{40}$$

For very small signal-to-noise ratios, $(T/2W)[E(\xi_{jk})]^2$ becomes very small. Therefore, the variances of the ξ_j are all approximately equal and can be computed from Equations (38) and (39).

As a consequence of the normality of the ξ_j, Equation (8) can be used to calculate probability of correct reception. Instead of m/σ, one needs to use m'/σ'. To calculate quantization loss, the ratio of the mean to standard deviation, m'/σ', needs to be related to the ratio m/σ. This has been done by approximating Equations (36) to (39) for small values of the receiver input signal-to-noise average power ratio. The ratio of m'/σ' to m/σ is a measure of the quantization loss. For any given number of quantization intervals, it is possible to calculate a value for K (see Figure 5) which minimizes the quantization loss. This value of K specifies an optimum set of quantization

thresholds for a given number of quantization intervals. In this
analysis, the receiver input is normalized in terms of the noise r-m-s
value. Thus, the quantization constant K is measured in units of the
noise r-m-s value. Instead of mean values of y_{jk} equal to $\pm A$, the
mean values are $h = \pm A/N^{1/2}$. It follows from Equation (9) that m/σ
is equal to $h\sqrt{2WT}$. Defining the quantization-loss function $g_r(K)$ to
be the ratio of m'/σ' to m/σ, it is shown in Appendix B that for
small h,
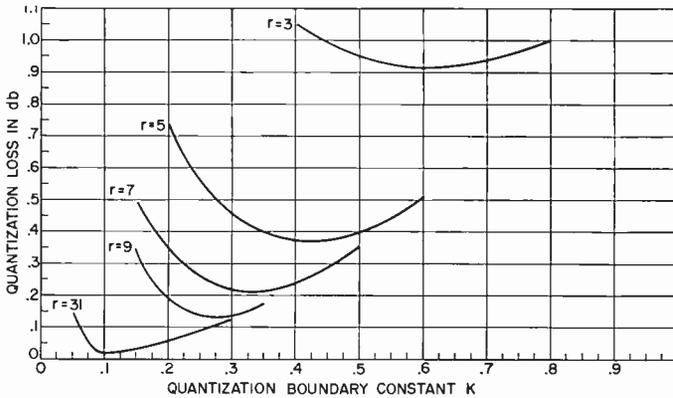
Fig. 6—Quantization loss for an odd number of intervals.

$$\begin{cases}
g_{2b}(K) = \dfrac{\sqrt{2} \displaystyle\sum_{t=0}^{b-1} \phi(Kt)}{\left[\displaystyle\sum_{t=0}^{b-1} (2t+1)[1 - \Phi(Kt)] \right]^{1/2}} & (41) \\[4em]
g_{2b-1}(K) = \dfrac{\sqrt{2} \displaystyle\sum_{t=1}^{b-1} \phi[K(2t-1)]}{\left[\displaystyle\sum_{t=1}^{b-1} (2t-1)[1 - \Phi(K[2t-1])] \right]^{1/2}} & (42)
\end{cases}$$

This function expresses the loss in output signal-to-noise ratio due
to quantization as a function of the number, r, of quantization levels.
The constant K determines the boundaries of the quantization inter-
vals. For a given value of r, a value for K can be found which maxi-
mizes $g_r(K)$, and therefore minimizes the loss in output signal-to-noise
ratio due to quantization. Figure 6 shows the quantization-loss func-
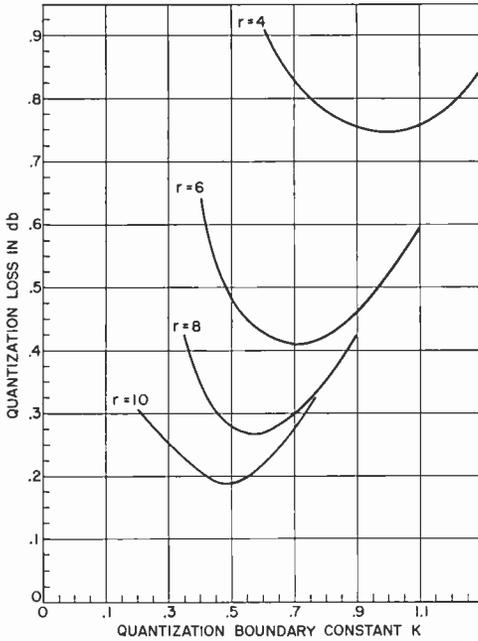
Fig. 7—Quantization loss for an even number of intervals.

tion $g_r(K)$ expressed in decibels as a function of K for various odd values of r. For each value of r, there is a value of K for which the loss is minimized. Figure 7 shows similar plots for even values of r. Figure 8 is a graph of the minimum loss as a function of the number of quantization intervals. Figure 9 shows the optimum values of K as a function of r.
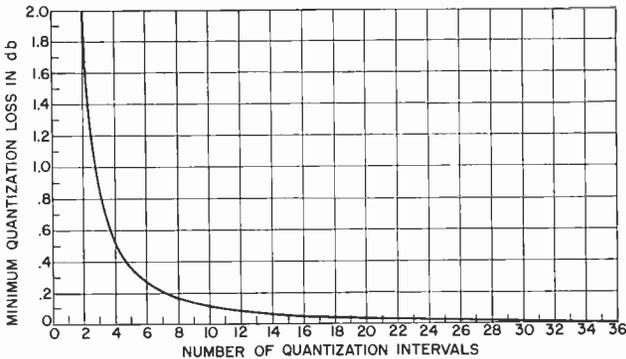
Fig. 8—Minimum quantization loss as a function of the number of quantization intervals.

In the 2-level case, the loss function in Equation (41) reduces to $\sqrt{2/\pi}$, which corresponds to the well-known 2-decibel loss due to two-level quantization.

When b becomes large, the optimum value of K approaches zero in such a way that bK is still large. As a consequence,
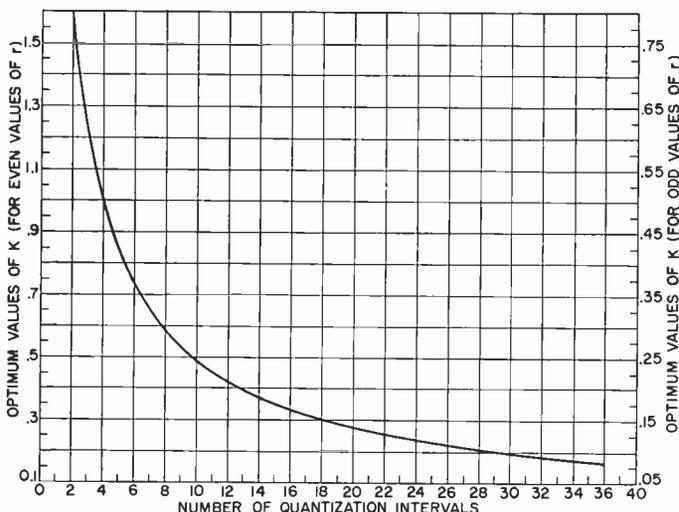
Fig. 9—Optimum values of K as a function of number of quantization intervals. Values of K on left are for even values of r, on the right for odd values of r.

$$g_{2b-1}(K) \sim \frac{\sqrt{2} \displaystyle\int_0^\infty \phi(x)\,dx}{\left[\displaystyle\int_0^\infty x(1 - \Phi|x|)\,dx \right]^{1/2}} = 1. \qquad (43)$$

Similarly,

$$g_{2b}(K) \sim 1. \qquad (44)$$

This merely verifies that as the number of quantization levels increases, the quantization loss goes to zero.

When K is equal to zero, g_r is equal to g_2. That is, too small a value

of K is no better than two-level quantization, in which case the loss is equal to 2.0 decibels.

If b is fixed, and K becomes large,

$$g_{2b-1}(K) \sim [2K\phi(K)]^{1/2} \sim 0, \tag{45}$$

i.e., if K is made too large, the quantization loss approaches a total loss. On the other hand, $g_{2b}(K)$ approaches g_2 as K increases. That is, even though K is taken to be very large, polarity information is still retained. Therefore, the loss cannot exceed the two-level quantization loss.

EFFECT OF AN ARBITRARY NUMBER OF QUANTIZATION LEVELS ON QUADRATURE RECEPTION

Figure 10 is a block diagram of a quadrature receiver employing time sampling and amplitude quantization as defined in the discussion of coherent reception. These processes are performed on the signals $x^*(t)S_j(t)$ and $\hat{x}^*(t)S_j(t)$. The functions $x^*(t)$ and $\hat{x}^*(t)$ are the input signal and noise inphase and quadrature components as defined in Equation (16). Two sets of time samples y_{jk} and $\hat{y}_{jk}$ are obtained. That is,

$$\begin{cases} y_{jk} = AS_i\left(\dfrac{k}{2W}\right) S_j\left(\dfrac{k}{2W}\right)\cos\theta + I_c\left(\dfrac{k}{2W}\right) S_j\left(\dfrac{k}{2W}\right), \\[4mm] \hat{y}_{jk} = -AS_i\left(\dfrac{k}{2W}\right) S_j\left(\dfrac{k}{2W}\right)\sin\theta - I_s\left(\dfrac{k}{2W}\right) S_j\left(\dfrac{k}{2W}\right). \end{cases} \tag{46}$$

Correspondingly, two sets of amplitude quantized time samples ξ_{jk} and $\hat{\xi}_{jk}$ are obtained in accordance with Equation (27) or (28). Instead of a single sum as in Equation (29), there are two sums of quantized products:

$$\begin{cases} \xi_j = \dfrac{1}{2W} \displaystyle\sum_{k=1}^{2WT} \xi_{jk}, \\[4mm] \hat{\xi}_j = \dfrac{1}{2W} \displaystyle\sum_{k=1}^{2WT} \hat{\xi}_{jk}. \end{cases} \tag{47}$$

The j-th quantized quadrature detector output is

$$Q_j^* = [\xi_j^2 + \hat{\xi}_j^2]^{1/2}. \tag{48}$$

Since I_c and I_s are independent normal random variables with zero mean and variance, $N = 2N_0W$; y_{jk} and $\hat{y}_{jk}$ are independent normal variables with mean values

$$\begin{cases} E(y_{jk}) = AS_i\left(\dfrac{k}{2W}\right) S_j\left(\dfrac{k}{2W}\right) \cos\theta, \\[4mm] E(\hat{y}_{jk}) = -AS_i\left(\dfrac{k}{2W}\right) S_j\left(\dfrac{k}{2W}\right) \sin\theta, \end{cases} \tag{49}$$
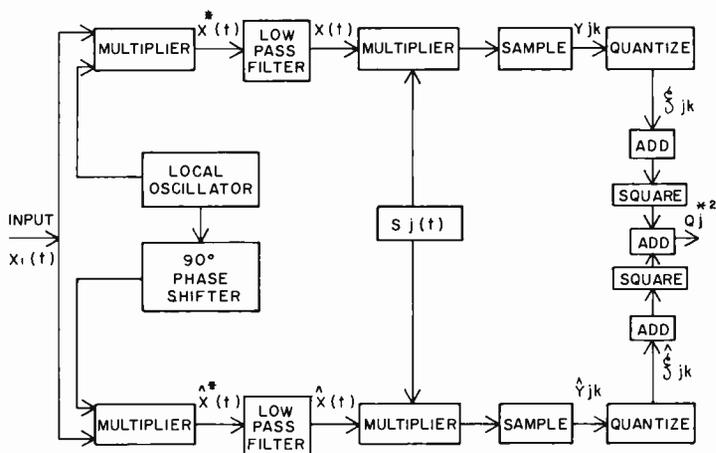
Fig. 10—Block diagram of quadrature detector for j-th signal, using two-level quantization.

and equal variance

$$\sigma^2\lfloor y_{jk}\rfloor = \sigma^2\lfloor \hat{y}_{jk}\rfloor = NS_j^2\left(\frac{k}{2W}\right). \tag{50}$$

If the signals $S_j(t)$ can only assume the values $+1$ and -1,

$$\sigma^2[y_{jk}] = N. \tag{51}$$

The random variables ξ_j and $\hat{\xi}_j$ consist of sums of independent normal random variables. By the central limit theorem for independent ran-

dom variables,[13] when $2WT$ is large, ξ_j and $\hat{\xi}_j$ will be approximately normally distributed. The mean values of ξ_j and $\hat{\xi}_j$, can be written

$$
\left\{
\begin{aligned}
E(\xi_j) &= \frac{1}{2W} \sum_{k=1}^{2WT} E(\xi_{jk}), \\
E(\hat{\xi}_j) &= \frac{1}{2W} \sum_{k=1}^{2WT} E(\hat{\xi}_{jk}).
\end{aligned}
\right.
\tag{52}
$$

If j is not equal to i, these variables have zero mean. For j equal to i, and for small signal-to-noise ratios, the mean values for an even number of intervals are

$$
\left\{
\begin{aligned}
E(\xi_i) &\simeq 2hT\cos\theta \sum_{t=0}^{b-1} \phi(Kt), \\
E(\hat{\xi}_i) &\simeq 2hT\sin\theta \sum_{t=0}^{b-1} \phi(Kt).
\end{aligned}
\right.
\tag{53}
$$

For an odd number of intervals,

$$
\left\{
\begin{aligned}
E(\xi_i) &\simeq 2hT\cos\theta \sum_{t=0}^{b-1} \phi[(2t-1)K], \\
E(\hat{\xi}_i) &\simeq 2hT\sin\theta \sum_{t=0}^{b-1} \phi[(2t-1)K].
\end{aligned}
\right.
\tag{54}
$$

For the even case, there are $2b$ intervals; for the odd case, there are $2b-1$ intervals. The constant K determines the quantization boundaries which provide equal intervals except at the end points. The sign of the mean is dependent upon whether or not the mean values of y_{ik} and $\hat{y}_{ik}$ are positive or negative. This is not relevant here, since only the sum of the squares of the means in Equation (54) is required in the error probability function.

The variances are (for all j including $j = i$)

$$
\sigma^2(\xi_j) = \sigma^2(\hat{\xi}_j) \simeq \frac{T}{W} \sum_{t=0}^{b-1} (2t+1)[1 - \Phi(Kt)]
\tag{55}
$$

for an even number of intervals, and

$$\sigma^2(\xi_j) = \sigma^2(\hat{\xi}_j) \cong \frac{T}{W} \sum_{t=1}^{b-1} (2t-1)\,[\,1 - \Phi([2t-1]K)\,] \qquad (56)$$

for an odd number of intervals.

In summary, when j is not equal to i, ξ_j and $\hat{\xi}_j$ are approximately normally distributed with zero mean and variance given above. ξ_i and $\hat{\xi}_i$ are approximately normally distributed with mean values given in Equations (53) and (54), and variance given above. ξ_j and $\hat{\xi}_j$ are statistically independent since y_{jk} and $\hat{y}_{jk}$ are expressed in terms of the independent noise quadrature components I_c and I_s. It follows that Q_j^* will have a Rayleigh probability density (see Equation (20)) when j is not equal to i. Q_i^* has a probability density of the form given in Equation (21). The analysis in the previous section on coherent reception is applicable here. When the appropriate ratio between sum of squares of the means and the appropriate variance is taken,

$$\frac{E^2(\xi_i) + E^2(\hat{\xi}_i)}{\sigma^2(\xi_i)} = 2\beta g_r^2(K) \ . \qquad (57)$$

The factor $g_r(K)$ represents the quantization loss function derived for coherent reception. Thus, the error-probability function in Equation (23), which has been shown to be applicable to ideal quadrature detection of orthogonal signals, is also applicable when quantization is used provided β in Equation (23) is multiplied by $g_r^2(K)$.

CONCLUSIONS

It has been shown that when the additive noise is stationary, white, and Gaussian, and the ratio of input average signal power to average noise power is small, then the effects of quantization are the same for both coherent and quadrature digital matched filtering for large signal-bandwidth-duration products. Furthermore, the error-probability functions for both ideal and quantized quadrature detection are independent of any slowly varying random phase shift which might be present in the signal carrier.*

ACKNOWLEGMENTS

and B. Glazer in clarifying several points of basic relevance to the analysis. Appreciation is expressed to Miss Barbara Hill, who performed the computations.

APPENDIX I—DISTRIBUTION FUNCTION OF DECIBEL LOSS DUE TO THE RANDOM PHASE ASYNCHRONISM IN COHERENT RECEPTION

The loss is given by

$$X = -10 \log_{10} \cos^2 \theta \text{ decibels} \tag{58}$$

where θ is a random variable assumed to have a uniform distribution in the interval $(0, 2\pi)$. The distribution function x is

$$
\begin{aligned}
F(x) &= P[X \leq x] \\
&= P[-10 \log_{10} \cos^2 \theta \leq x] \\
&= 1 - P[|\cos \theta| \leq 10^{-x/20}].
\end{aligned}
\tag{59}
$$

The frequency function of $\cos \theta$ is

$$p(x) = \frac{1}{\pi \sqrt{1 - x^2}} . \tag{60}$$

It follows that

$$
\begin{aligned}
G(y) &= P[|\cos \theta| \leq y] = \int_{-y}^{y} \frac{dx}{\pi \sqrt{1 - x^2}} \\
&= \frac{2}{\pi} \arcsin (y).
\end{aligned}
\tag{61}
$$

Applying Equation (61) to Equation (59),

$$
\begin{aligned}
F(x) &= 1 - \frac{2}{\pi} \arcsin (10^{-x/20}) \\
&= \frac{2}{\pi} \arccos (10^{-x/20}).
\end{aligned}
\tag{62}
$$

The distribution function $F(x)$ is plotted in Figure 1.

APPENDIX II—DERIVATION OF THE QUANTIZATION LOSS
FUNCTION FOR SMALL h

From the definitions of the p_i in Equation (30),

$$p_{s+1} = P[\xi_{jk} = b - s | h]; \quad (s > 0)$$

$$= \int\limits_{(b-s-1)K}^{(b-s)K} \phi(x - h)\, dx = \Phi[(b-s)K - h] - \Phi[(b-s-1)K - h]$$

$$\cong \Phi[(b-s)K] - \Phi[(b-s-1)K]$$
$$- h[\phi(b-s)K) - \phi(1b-s-1)K)]. \qquad (63)$$

$$p_{2b-s} = P[\xi_{jk} = -(b-s) | h]; \quad (s > 0)$$

$$= \int\limits_{-(b-s)K}^{-(b-s-1)K} \phi(x - h)\, dx = \Phi[(b-s)K + h] - \Phi[(b-s-1)k + h]$$

$$\cong \Phi[(b-s)K] - \Phi[(b-s-1)K]$$
$$+ h[\phi([b-s]K) - \phi([b-s-1]K)]. \qquad (64)$$

$$p_1 = \int\limits_{(b-1)K}^{\infty} \phi(x - h)\, dx$$

$$= 1 - \Phi[(b-1)K - h]$$
$$\cong 1 - \Phi[(b-1)K] + h\phi[(b-1)K]. \qquad (65)$$

$$p_{2b} = \int\limits_{-\infty}^{-(b-1)K} \phi(x - h)\, dx = 1 - \Phi[(b-1)K + h]$$

$$\cong 1 - \Phi[(b-1)K] - h\phi[(1b-1)K] \qquad (66)$$

Applying these approximations to the expressions for m' and σ' in Equations (36) and (38),

$$m' \cong 2hT\left[b\phi([b-1]K) + \sum_{s=1}^{b-1}(b-s)\left(\phi[(b-s-1)K] - \phi[(b-s)K]\right)\right]$$

$$= 2hT\sum_{t=0}^{b-1}\phi(Kt). \qquad (67)$$

$$(\sigma')^2 \cong \frac{T}{W}\left[b^2 \left(1 - \Phi[(b-1)K]\right) + \sum_{s=1}^{b-1} (b-s)^2 \right.$$

$$\left. \left(\Phi[(b-s)K] - \Phi[(b-s-1)K]\right)\right]$$

$$= \frac{T}{W}\left[b^2 - \sum_{t=0}^{b-1} (2t+1)\Phi(Kt)\right]$$

$$= \frac{T}{W} \sum_{t=0}^{b-1} (2t+1)[1 - \Phi(Kt)]. \tag{68}$$

From the definitions of q_i in Equation (31),

$$q_{s+1} = P[\xi_{jk} = b - s - 1 \,|\, h]$$

$$= \int_{[2(b-s)-3]K}^{[2(b-s)-1]K} \phi(x-h)\,dx$$

$$= \Phi\left[(2[b-s]-1)K - h \right] - \Phi\left[(2[b-s]-3)K - h \right]$$

$$\cong \Phi\left[(2[b-s]-1)K \right] - \Phi\left[(2[b-s]-3)K \right]$$

$$- h\left[\phi\left([2(b-s)-1]K\right) - \phi\left([2(b-s)-3]K\right)\right]. \tag{69}$$

$$q_{2b-s-1} = P[\xi_{jk} = -(b-s-1)\,|\,h] = \int_{-[2(b-s)-1]K}^{-[2(b-s)-3]K} \phi(x-h)\,dx$$

$$= \Phi\left[(2[b-s]-1)K + h \right] - \Phi\left[(2[b-s]-3)K + h \right]$$

$$\cong \Phi\left[\left(2(b-s)-1\right)K \right] - \Phi\left(2(b-s)-3\right)K \right]$$

$$+ h\left\{ \phi\left[\left(2(b-s)-1\right)K \right] - \phi\left[\left((2[b-s]-3)K\right)K\right]\right\}. \tag{70}$$

$$q_1 = \int_{(2b-3)K}^{\infty} \phi(x-h)\,dx = 1 - \Phi[(2b-3)K - h]$$

$$\cong 1 - \Phi[(2b-3)K] + h\phi[(2b-3)K]. \tag{71}$$

$$q_{2b-1} = P[\xi_{jk} = -(b-1)]$$

$$= \int_{-\infty}^{-(2b-3)K} \phi(x-h)\,dx$$

$$= 1 - \Phi[(2b-3)K + h]$$

$$\cong 1 - \Phi[(2b-3)K] - h\phi[(2b-3)K]. \tag{72}$$

Applying these approximations to the expressions for m' and σ' in Equations (37) and (39),

$$m' \cong 2hT\left[(b-1)\phi([2b-3]K) + \sum_{s=1}^{b-2}(b-s-1)\right.$$

$$\left.\left(\phi[(2(b-s)-3)K] - \phi[(2(b-2)-1)K]\right)\right]$$

$$= 2hT\sum_{t=1}^{b-1}\phi[(2t-1)K] \tag{73}$$

$$(\sigma')^2 \cong \frac{T}{2W}\left[(b-1)^2\left(1 - \Phi([2b-3]K)\right)\right.$$

$$+ \sum_{s=1}^{b-2}(b-s-1)^2\left[\Phi\left((2(b-s)-1)K\right) - \Phi\left((2[b-s]-3)K\right)\right]\right]$$

$$= \frac{T}{W}\sum_{t=1}^{b-1}(2t-1)\left[1 - \Phi\left((2t-1)K\right)\right] \tag{74}$$

The loss functions given in Equations (41) and (42) can be obtained from Equations (67) and (68) and Equations (73) and (74), respectively, by dividing m'/σ' by

$$\frac{m}{\sigma} = \sqrt{2\beta} = h\sqrt{2WT}.$$

STATISTICAL ANALYSIS OF MULTIPATH JITTER[*]

BY

J. J. BRANDINGER[†] AND H. GOLDMAN[‡]

Summary—The major portion of this paper is devoted to determining the time stability of multipath jitter. After introducing the subject, a summary of first-order statistics is presented, followed by a determination of the autocorrelation and power spectral density functions. The effects of propagation parameters on the quantitative results are examined and, in addition, the relation of multipath jitter to the performance of communication systems is briefly discussed.

INTRODUCTION

MULTIPATH JITTER or delay time variability is a phenomenon which often limits the maximum speed of transmission in a communications channel. Most radio circuits are subject to multipath jitter. That is, the received signal is made up of the resultant of a number of components traversing different paths. A typical multiple-path geometry is illustrated in **Figure 1**. The component of signal traveling via the shortest path arrives first, followed in succession by the components arriving over the longer paths. **Figure 2** presents a plot of transmission time versus great-circle
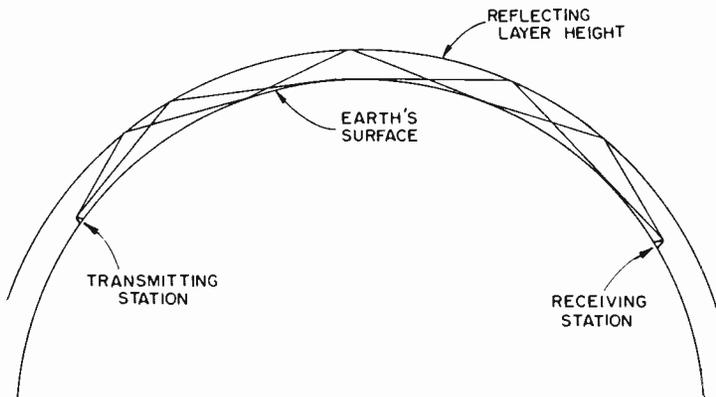
Fig. 1—Multiple-path propagation showing two- and three-hop ray paths.

distance between terminals for a number of typical high-frequency radio-propagation ray paths. It can be seen that path differences may be as great as several hundred miles, resulting in arrival times which differ by amounts ranging from less than 1 to several milliseconds. This type of propagation phenomena is characterized not only by signal amplitude fading, but also by uncertainty in the prediction of the position and length of transmitted keying elements. In the present discussion, the main interest is concentrated on the keying element
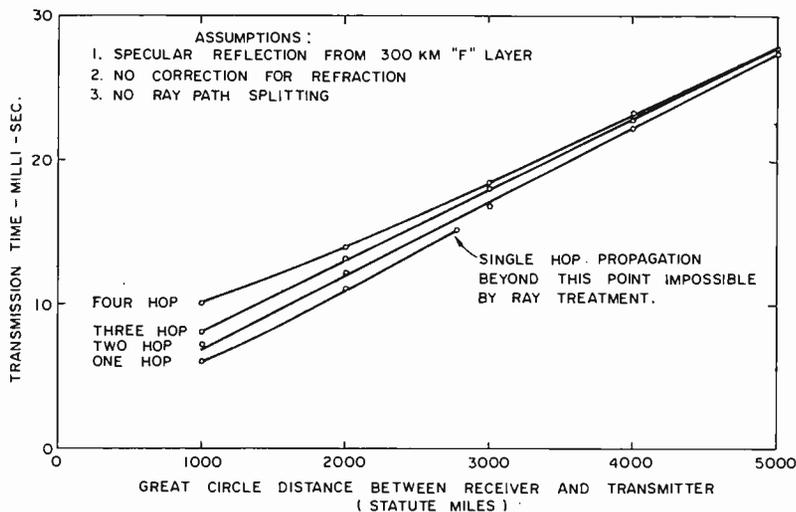
Fig. 2—Transmission time versus great-circle distance between receiver and transmitter for four different hop situations.

timing uncertainties associated with the received signal. For example, in frequency-shift keying (FSK), overlapping at the time of keying-element transition can subject the receiver to two frequencies at the same time. The exact instant of transition at the detector output would depend largely upon which signal happens to be the stronger at the particular moment.

There is a small amount of published data which indicates the relative magnitude of multipath jitter, but very little statistical information has been presented. It is the purpose of this paper to report measurements of multipath jitter recorded on a 5,000-mile, high-frequency (HF) radio circuit. From these measurements the probability distribution, autocorrelation, and power spectral density functions have been obtained.

ANALYSIS OF MULTIPATH JITTER

Probing Techniques

There are several ways to determine experimentally the effects of multipath transmission. Two of the more common techniques involve probing the medium with either short-duration pulses or with frequency shift keying.

The short-pulse technique is used to characterize the multipath jitter by establishing time differences in the arrival of pulses traversing different paths. The transmitted pulse lengths are usually constrained to be less than the time difference between the arrival of successive major paths, but are seldom short enough to resolve the major paths into their respective subpaths. The result is that pulse fading is generally observed. This fading tends to follow a Rayleigh distribution when the transmission path consists of a random combination of subpaths.

In the frequency-shift-keying technique, the time delay between successive signal arrivals is determined by the amount of uncertainty present in the keying-element transition. Unlike the short-pulse technique, it is necessary for FSK elements to be long by comparison with the total amount of uncertainty that might be expected.

The main difference between these two systems lies in their designed response time. The short-pulse system is generally wide-band and is used to examine fine grain structure, while the FSK system provides an integrated picture of the multipath effects in a narrower bandwidth. Since the FSK system is representative of a typical modulation transmission system, results using this technique can be more directly applied in practice.

Extensive measurements using FSK probing techniques have been performed on a long-distance (5,000 miles) radio circuit between Hawaii and Riverhead, New York. A block diagram of the test system is shown in Figure 3. Measurements were performed on both daytime and nighttime frequencies in order to have comparisons with relatively small and large multipath distortions. Generally, the testing schedule and individual recordings were selected so that the contribution of noise to the keying-transition-time distortion was minimized (e.g., signal-to-noise ratio $\geq$ 20 decibels).

Method of Testing

In order to evaluate the received leading and trailing edges of keying elements distorted by multipath jitter, the synchronous keying
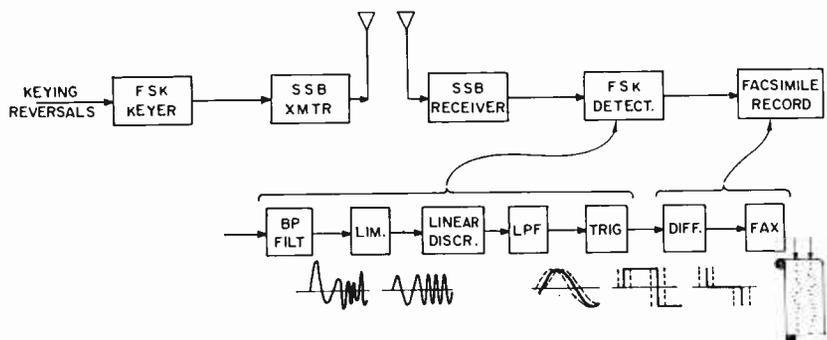
Fig. 3—Block diagram of test circuit.

elements were differentiated and imposed upon facsimile recording equipment. In Figure 3, appropriate waveforms are indicated at several points in the detection and recording locations. The recorder speed was made commensurate with the keying speed, so that in the absence of multipath, a straight line of consecutive dots would be recorded. A typical recording is shown in Figure 4. By determining the distance the dots departed from a straight line, a measure of the peak-to-peak jitter was obtained. For analysis, these facsimile records were divided into about 1-minute intervals and the number of dots that exceeded various threshold values of timing were expressed as a percentage of the total number of dots in the sample. This analysis provided data necessary for determining probability distribution functions.

PROBABILITY DISTRIBUTION FUNCTIONS

Many probability distributions were obtained for the one-minute sample periods. Figure 5 is a typical plot of jitter in milliseconds
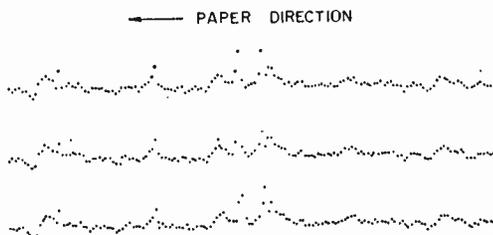
Fig. 4—Jitter recorded at Riverhead (frequency: 7922.5 kilocycles; scale: 1 inch = 13 milliseconds; keying speed: 50 cps reversals; time (EST): 7:05 AM; date: Feb. 15, 1956).

versus the percentage of time the jitter was equal to or greater than the ordinate value. An examination of a large number of these jitter distributions has shown that the distribution can be approximated by drawing a straight line on log-linear paper through a point at (100%, 0) and a measured point at about (1%, X). This is indicated as the dotted line in Figure 5. The expression for this line is $P\ (x \geqq X)$ $= 100\ e^{-NX}$, where $P\ (x \geqq X)$ represents the abscissa scale, X is the ordinate (in milliseconds), and N is a scale factor of 2 for the particular case shown.
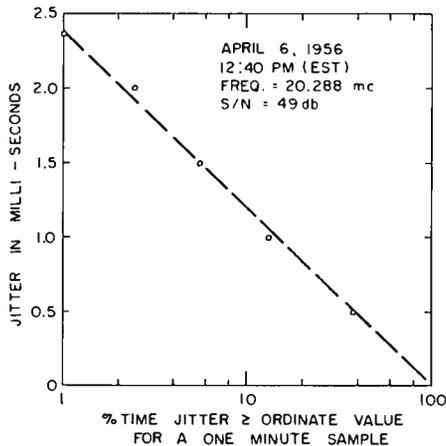
Fig. 5—Distribution of jitter in a one-minute sample (10 millisecond bits).

Approximately 1,000 distribution functions of multipath measured in the one-minute time sample were analyzed in order to determine their 1 per cent jitter values. These distributions represented data from five different frequencies ranging from 7 to 20 megacycles. A summary in the form of a cumulative distribution of these 1 per cent values was then obtained for each frequency. Figure 6 shows the curves of these summarized "1 per cent distributions." It is seen that the multipath jitter has values in the millisecond range, and is generally less than 5 milliseconds. The scale factor N in the distribution function equation lies roughly in the range $1 < N < 5$.

The distribution function can be used to estimate the error rate resulting from multipath jitter.* If it is assumed that the output of

---

* See Appendix.

the FSK system is synchronously sampled in the center of each keying
element, errors result whenever the keying transition is delayed or
advanced by more than one-half of a keying element length. If it is
also assumed that the probability of transition occurrence is 0.5
(remembering that an error due to jitter occurs only when a transi-
tion is present), the probability of error is $P_e = e^{-NK}/2$, where K is
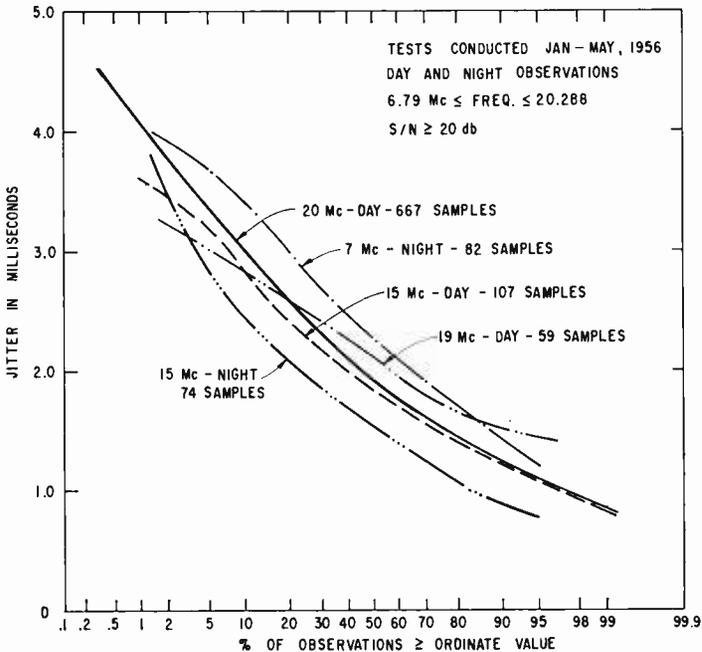
Fig. 6—Cumulative distribution of 1 per cent jitter values.

the keying element length in milliseconds. An estimate of error rate
can also be obtained for the case in which the sampling is not centered
relative to the keying elements. This could result from inaccurate or
delayed correction in the synchronization of the system. Figure 7
shows the statistical model considered, along with plots for several
asymmetric conditions. The parameter ϵ is the displacement in milli-
seconds of the transition relative to its expected position. These curves
are only considered for the cases where $\epsilon \leq K/2$. For example, if
$N = 2$, $K = 5$ milliseconds, and $\epsilon = 0$, $P_e \approx 2.4 \times 10^{-5}$, whereas, for
$\epsilon = 0.75$, $(N\epsilon = 1.5)$, $P_e \approx 2.4 \times 10^{-4}$. This order-of-magnitude in-
crease in probability of error illustrates the type of degradation
resulting from inaccurate pulse-sampling location. Although this

analysis helps to describe the multipath phenomenon and its effects on a first-order statistical basis, it does not convey detailed information on its time variability.
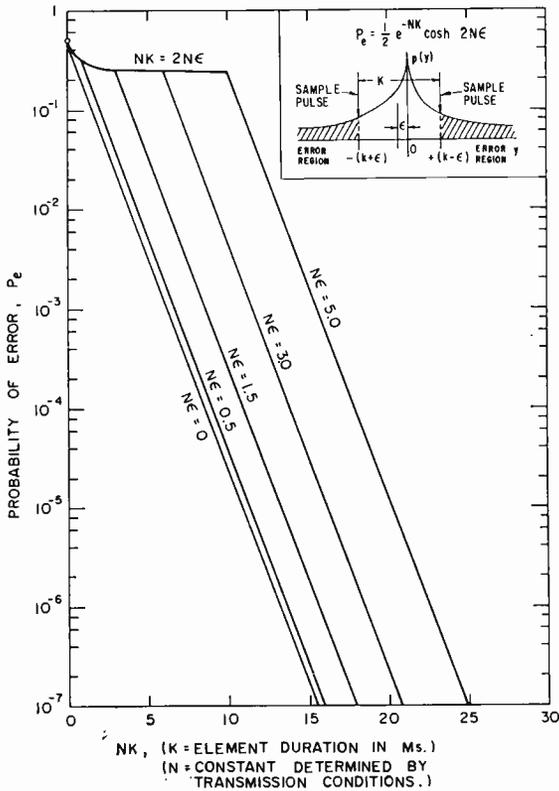
Fig. 7—Multipath-jitter error rate for off-center pulse sampling versus keying-element duration.

AUTOCORRELATION ANALYSIS OF MULTIPATH JITTER

The time variability of multipath jitter is most conveniently studied by means of the autocorrelation function. The autocorrelation function is a measure of the degree of certainty (or lack of it) with which the time of successive received signal transitions can be predicted. Eighteen facsimile records of multipath jitter were selected from a larger collection of data. These samples represent a range of conditions from minimum to maximum for the observed multipath fluctuations. The data were edited to eliminate from consideration obviously noisy portions of the record, and then analyzed by the method of

circular, serial correlation to determine the autocorrelation function,[1] $\phi_{11}(\tau)$. The autocorrelation function is given by the following equation:

$$\phi_{11}(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} f_1(t) \, f_1(t+\tau) \, dt. \tag{1}$$

Three main steps are required to solve this equation: (1) The function is given a time displacement, τ, to generate $f_1(t+\tau)$; (2) $f_1(t)$ and $f_1(t+\tau)$ are multiplied together; and (3) this product is averaged by integration over a complete period. In the above equation, $f_1(t)$ is a function describing the jitter as previously discussed. Its
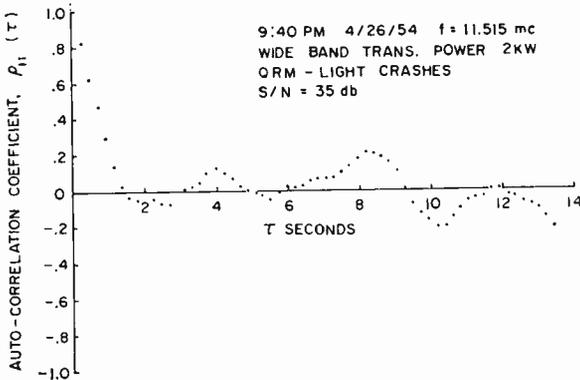
Fig. 8—Autocorrelation plot of multipath jitter.

average value is zero. The detailed calculations were performed on a digital computer. Figure 8 represents one of the 18 correlation functions derived by this method. This curve has been normalized to give the autocorrelation coefficient: $\rho_{11}(\tau) = \phi_{11}(\tau)/\phi_{11}(0)$. The curve consists of 123 sample points, where the spacing between points is 220 milliseconds, and the total sample interval is 27.06 seconds.

Results of Analysis

The results of the 18 correlation curves are summarized in Figure 9, using the mean or average value (μ) of the functions, along with the standard deviation (σ) about the mean. It is seen that the mean value of the correlation coefficient drops rapidly from 1, at $\tau = 0$, to

---

[1] Jules S. Bendat, *Principles and Application of Random Noise Theory*, pp. 25-28, John Wiley & Sons, New York, N.Y., 1958.
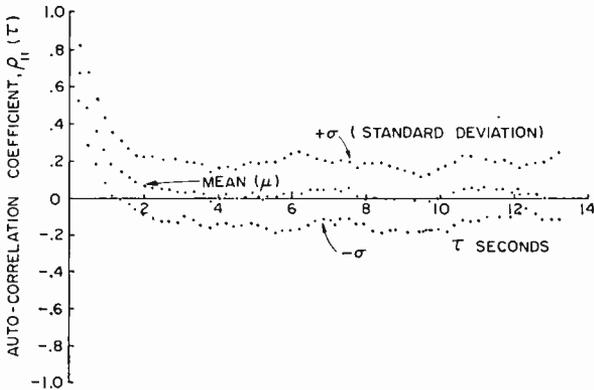
Fig. 9—Mean (μ) and standard-deviation (σ) curves for 18 autocorrelation plots of multipath jitter.

$\rho_{11}(\tau) < 0.1$ in approximately $\tau = 2$ seconds. Beyond 2 seconds, the mean value slowly undulates near $\rho_{11}(\tau) = 0$, closely paralleled by the σ values. An expanded view of Figure 9 covering the first 3.5 seconds is shown in Figure 10.

In evaluating the autocorrelation curve, the point at which the curve crosses the axis, $\rho_{11}(\tau) = 0$, is by definition the point in time (τ) when the function is uncorrelated. Since some functions tend to approach 0 asymptotically, the axis crossing point is not always an adequate reference. A more convenient reference is the time corre-
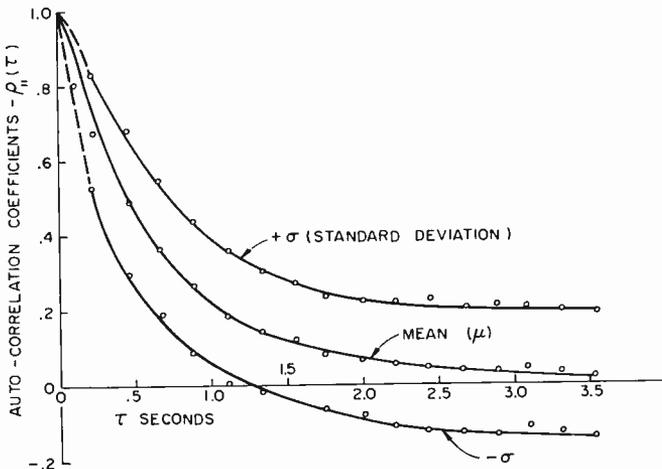
Fig. 10—Expanded view of mean and standard deviations curves for 18 autocorrelation plots of multipath jitter.

sponding to the $1/e$ value of the correlation coefficient, i.e., $\rho_{11}(\tau) = 1/e$, after which time the function is considered "essentially" uncorrelated. In Figure 10, it is seen that the mean value of the auto-correlation coefficient drops to its $1/e$ value in about $\tau = 0.\overline{67}$ second, with the curve asymptotically approaching 0 beyond 3.5 seconds. This means that, on the average, the time of successive keying element transitions would be "essentially" uncorrelated after $0.\overline{67}$ second. Figure 11 shows the extremes of the measured correlation curves, where the minimum and maximum values of the $1/e$ correlation times are 0.28 and 1.9 seconds, respectively.
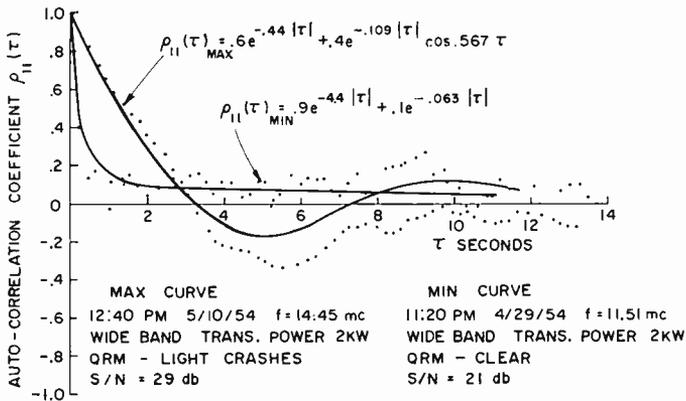
Fig. 11—Equation (2) fitted to extreme autocorrelation curves.

By extrapolation of the 18 autocorrelation curves, an average value of $\tau = 0.\overline{12}$ second was obtained for $\rho_{11}(\tau) = 0.9$. This point has considerable significance, since it shows that for the radio path measured, the successive arrival of keying transitions was highly correlated (i.e., 90 per cent) for time periods up to 120 milliseconds. In a typical telegraph signaling system, using keying elements which are substantially shorter than 120 milliseconds, it might be expected that the transition timing uncertainties would be caused (on a short-term basis) by noise and interference rather than by multipath. The time duration over which 90 per cent correlation is maintained is also useful in estimating the maximum speed of response for automatic pulse sampling and frequency control loops in communication receiver configurations.

Curve Fitting with an Analytical Function

When possible, it is desirable to form an analytical expression to

represent experimental data in order that the performance of other systems can be predicted. Some of the restrictions which should be satisfied in order to generate a usable autocorrelation function are:

(1)   The function should be even with respect to time, with a maximum value at time equal to zero.

(2)   The Fourier transform must be positive (i.e., no power generated by the process).

(3)   The equation should be reasonably simple, with a sufficient number of constants to encompass a wide variety of the data fluctuations.

The equation chosen to approximate the autocorrelation coefficient is given by:

$$\rho_{11}(\tau) = K_1 e^{-a|\tau|} + K_2 e^{-b|\tau|} \cos(c\tau); \quad K_1, K_2, a, b \geqq 0. \qquad (2)$$

The form of this equation, combining exponential decay with periodicity, is similar to the stationary autocorrelation functions encountered in a number of physical processes such as vacuum-tube noise, radar-fading, atmospheric turbulence, etc.[2-4] The constants in Equation (2) must satisfy an additional constraint, namely, $\rho_{11}(0) = 1$. In particular, this means that $K_1 + K_2 = 1$. Within these constraints, the constants of Equation (2) are determined by curve fitting to the actual data. Figure 11 shows the results obtained by a choice of appropriate constants to match the extreme values of the autocorrelation coefficient curves, but does not follow the data in fine detail. It was not con-shortest $1/e$ time. Figure 12 shows a similar fit to the average autocorrelation coefficient curve. It is seen that Equation (2) with adjusted constants gives a good first approximation to the experimental curves, but does not follow the data in fine detail. It was not considered desirable to accomplish a better fit to the data, since a much more complex expression would have been required.

Comparison with Other Published Data at High Frequency

Price has reported a single measurement of the autocorrelation function taken on the received carrier frequency of WWV (5 mega-

[2] S. O. Rice, "Mathematical Analysis of Random Noise," *Bell Syst. Tech. Jour.*, Vol. 23, p. 282, July 1944.

[3] H. M. James, N. B. Nichols, and R. S. Phillips, *Theory of Servomechanism,* MIT Radio Laboratories, Vol. 25, Chapter 6, McGraw-Hill Book Co., New York, N.Y., 1947.

[4] Press, H., Houbolt, J. C., "Some Applications of Generalized Harmonic Analysis to Gust Loads on Airplanes," *J. Aero. Sci.*, Vol. 22, No. 1, January, 1955.

cycles) over a 395-mile path.[5] An approximate scaling of these data are plotted along the dotted line in Figure 12. Price's data show a higher degree of correlation than most of the data measured over the Hawaii–Riverhead path, although some of the extreme values of the autocorrelation function approach these values. The differences between the measurements indicate the amount of decorrelation which tends to occur as the distance between terminals increases and additional ionospheric reflections are included in the transmission path. Price also noted some oscillatory behavior in the correlation-function tail structure.
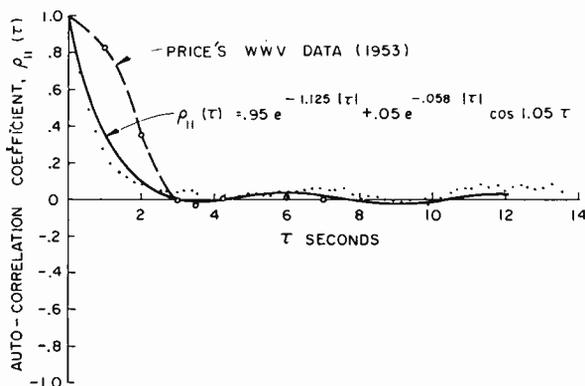
Fig. 12—Plot of Equation (2) to average of 18 autocorrelation functions.

POWER SPECTRAL DENSITY ANALYSIS OF AUTOCORRELATION COEFFICIENT

The power spectral density is defined as the Fourier transform of the autocorrelation function.

$$\psi_{11}(j\omega) = \int_{-\infty}^{\infty} \phi_{11}(\tau) e^{-j\omega\tau} d\tau. \tag{3}$$

The inverse transform is given by:

$$\phi_{11}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_{11}(j\omega) e^{j\omega\tau} d\omega. \tag{4}$$

---

[5] Price, R., "The Autocorrelogram of a Complete Carrier Wave Received over the Ionosphere at Oblique Incidence," *Proc. I.R.E.*, p. 879, Vol. 45, June 1957.

The significance of the power spectral density can be thought of as representing the power dissipation of a voltage $f_1(t)$ across a 1-ohm resistor for all frequencies between ω and $(\omega + d\omega)$. A special case is the total average power dissipated in the resistor which is:

$$P_{\text{TOTAL}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_{11}(j\omega)\, d\omega. \tag{5}$$
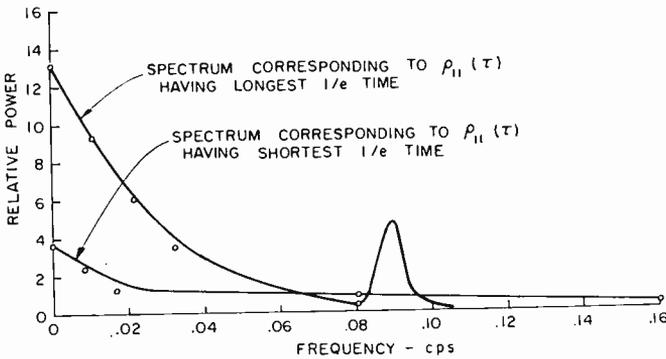
Fig. 13—Power-density spectra for jitter samples corresponding to the extreme autocorrelation coefficient curves.

Substituting $\tau = 0$ in Equation (4) gives

$$\phi_{11}(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_{11}(j\omega)\, e^0 d\omega, \tag{6}$$

which is equivalent to Equation (5). This means that the autocorrelation function for $\tau = 0$ measures the total average power of the time function, $f_1(t)$.

Power spectral density functions computed for the extreme autocorrelation curves are shown in Figure 13, and the spectrum corresponding to the average $\rho_{11}(\tau)$ is shown in Figure 14. Perhaps the most significant point about these results is that the major components of the density function tend to be found between 0 and 0.5 cycle per second. This gives an indication of the degree of periodicity associated with the measured values of multipath jitter.
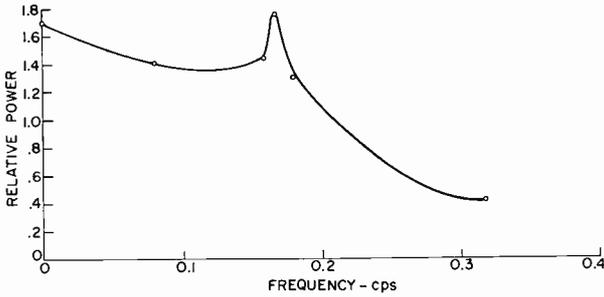
Fig. 14—Power-density spectrum corresponding to average correlation
function shown in Figure 12.

EFFECTS OF EXTERNAL INFLUENCES

Effects of Transmitted Frequency on Multipath Jitter

Past experience has indicated that multipath jitter is influenced
by the frequency of operation. Therefore, a study was made of the
relation between the magnitude of the jitter and the operating fre-
quency. The results of this study are presented in Figure 15. In this
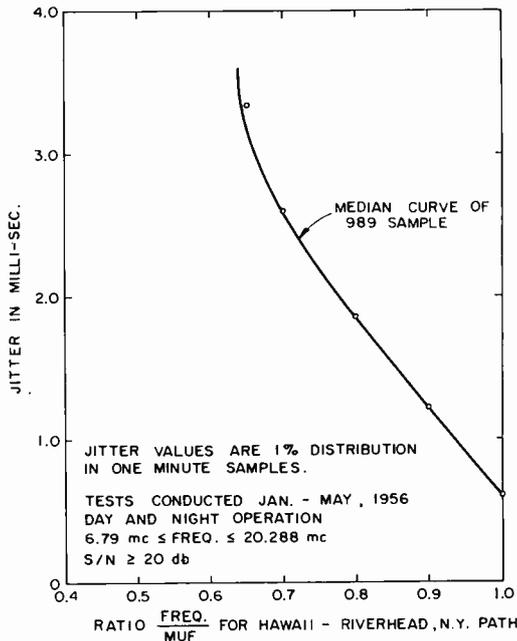
Fig. 15—Multipath jitter versus the ratio of operating frequency to
maximum usable frequency (3000 kilometer path).

plot, F is the frequency used on the radio circuit, and the *MUF* is the maximum usable frequency for a 3,000-kilometer circuit at the time of day a particular measurement of jitter was made. The mass of plotted points were found to be quite scattered, varying as much as ± 2 milliseconds about the median curve. The trend, however, was clear, showing a general reduction in jitter as *F/MUF* approached unity. This was to be expected, since the higher the operating frequency, the fewer the number of possible hops that could be supported by the ionospheric reflecting layers, and consequently the fewer the number of propagation paths that the signal could traverse to the receiver.

Qualitatively, this dependence of multipath jitter with frequency implies that longer $1/e$ correlation times should result as the operating frequency approaches the maximum usable frequency. Table I summarizes data representing measurements taken at about the same time of day (*MUF* approximately constant).

Table I

Frequency *(megacycles)*	*Number of* *Records*	*Average 1/e* *Correlation Time* *(seconds)*
14.5	3	1.32
11.5	3	1.08

These data indicate the expected trend. However, the difference is not great, probably because of the small separation in actual frequencies compared.

Effects of Additive Disturbances

Many factors are involved in determining the effects of additive noise and interference on multipath jitter. The combined effect of these disturbances on S/N is typically influenced by the state of the transmission medium, transmitted power, and receiver bandwidth.

In most of the tests, the transmitted power was either 200 watts or 2 kilowatts, the narrower bandwidth (300 cps) FSK system being associated with the lower power, and the larger bandwidth (1,700 cps) system with the higher power. This leaves the wide-band system with about a 2.5 decibels advantage. Signal-plus-noise-to-noise $((S + N)/N \approx S/N)$ measurements were obtained by examining the quasi-peak values of $S + N$ and N in approximately a 3-kilocycle bandwidth averaged over a 5-minute interval. Data were selected for signal-to-noise ratios equal to, or greater than, 20 decibels in order to consider-

Table II

	Disturbance	Number of Records	Average S/N (decibels)	Average 1/e Correlation Time (seconds)
Case A	Little or Low	4	38.5	0.56
Case B	Moderate to Heavy	4	26.7	0.44

ably reduce the effects of noise. Characteristic noise scattering of the recorded dots was also eliminated from consideration in the scaling process. Some results on the effects of additive disturbances on the correlation coefficient are given in Table II.

These results show that for a decrease of about 12 decibels in S/N, the average correlation coefficient decreases by about 20 per cent. In other words, the average correlation coefficient was not very sensitive to changes in S/N, when this ratio was in the range of 26 to 38 decibels.

Receiver Bandwidth Effects on Correlation Analysis

System bandwidth and keying rise time constants determine the precision with which multipath jitter can be measured. Since a wide-band system has a closer correspondence to the short pulse-sampling technique previously described, it can more accurately define the fine structure of multipath jitter in the presence of noise. In order to obtain some indication of the effects of bandwidth, a careful selection of data was made and the results tabulated as shown in Table III. It is seen that the data processed by the wide-band system is apparently more highly correlated than that in the narrow-band system. Thus, uncertainties in timing associated with random disturbances (i.e., noise) in the narrow-band system are probably due to equipment limitations and are not related to the multipath phenomenon itself.

Table III

	System Bandwidth	Number of Records	Average S/N (decibels)	Average 1/e Correlation Time (seconds)
Case A	Wide	7	28.9	0.91
Case B	Narrow	4	38.5	0.56

Effects of Time of Day on Correlation Analysis

The ionosphere undergoes diurnal variations resulting in changes in the transmission path, and these in turn are reflected in changes in the correlation of multipath jitter. Since daylight transmission tends to be relatively stable, it would be expected, and indeed has been found, that the $1/e$ correlation values tend to be relatively stable during this period. By contrast, the day-to-night transition period is a period of ionospheric change. A tabulation of data for a day-to-night transition is given in Table IV.

Table IV

Time of Day (EST)	Number of Records	Average 1/e Correlation Time (seconds)
5:30 P.M.	1	1.9
9:40 P.M.	2	.67
11:00-11:20 P.M.	2	.47

Although there is only a limited amount of data indicated in this table, it does emphasize the "roughness" of the transmission medium during this period. The ionospheric layer height, which is increasing as the sun leaves the transmission path, evidently results in a strong decorrelation of the multipath jitter. This is consistent with the difficulties that have been experienced with commercial and military communication systems in maintaining reliable information transfer during the transition period.

DISCUSSION OF RESULTS

The results of the preceding study verify quantitatively that signal transmission-time variations over a long distance radio circuit are in the millisecond range and are highly correlated for substantial periods of time. In terms of a practical communications system, this means that the keying element durations should be long enough to absorb up to approximately the 5 milliseconds of jitter occasionally measured. However, even 5 milliseconds may be too short when other limitations of these systems are considered. These limitations can include the timing variations or jitter introduced by noise, transient filter ringing, and sampling pulse position inaccuracies. In addition, the finite length of the sampling pulses also reduces the effective keying element length which can operate in the presence of multipath jitter.

When keying element durations are of the order of 10 milliseconds, the autocorrelation data has indicated that 10 per cent average uncertainty results in the timing between the first and the twelfth elements. Although the correlation data covers a small sampling period (totaling less than 10 minutes), the individual short samples were carefully selected so as to include a wide variety of operating conditions. These data are particularly useful in estimating the length of time over which coherence can be expected in the arrival time of successive keying elements. A number of differentially coherent transmission systems are strongly dependent upon this correlation for their successful operation.

In order to combat the effects of multipath jitter, a number of techniques are available. These include use of special directive antenna arrays (for selection of signals at a single arrival angle), operation at frequencies close to the maximum usable frequency, and use of long keying elements. The use of space diversity reception and increased power have little, if any, effect upon the multipath jitter. When the above techniques have been exhausted, communication theory suggests additional improvements may be obtained by the use of coding to reduce the probability of error. When errors due to the propagation medium occur in a random fashion, the type of coding required is one in which the strategy matches the random process. However, when multipath is present, the errors may not be random, but rather may occur in groups or bursts. In this case, the coding and decoding procedures can take advantage of the non-random error distribution to reduce equipment complexity in the process of reducing error probabilities.

The kind of errors that result from multipath jitter are strongly influenced by the method of signal reception that is used. Many coherent and semicoherent communication systems require signal timing information for their proper operation. If the transmitted signal information is subjected to timing variations of a magnitude comparable to the keying element length, then the coherent systems would be most seriously affected. The correlation analysis which has been presented is very useful in predicting the performance of these systems. In particular, knowledge of the correlation time permits an estimation of the time constant to be used in receiver automatic frequency control (AFC) and pulse-sampling loops. For example, it is not desirable to make the AFC control loop time constant any faster than the 90 per cent correlation time; otherwise, the loop will be subject to fluctuations associated with random noise. Similarly, the sampling pulses which are used to examine the keying elements in

post-detection circuitry should not have their time base changed in large increments, i.e., faster than the 90 per cent correlation time, to again avoid random positioning of the sampling by noise. If shorter time constants are used, it is necessary to anticipate or feed forward the correct information in order to have it associated with the proper keying elements.

CONCLUSIONS

The results of this study have shown that signal transmission time variations over a long distance radio circuit are in the millisecond range, and are highly correlated for substantial time periods. Typically, it was found that for a 5,000-mile HF radio circuit the multipath jitter was less than 5 milliseconds most of the time, and was 90 per cent correlated for time intervals less than 0.12 second. It was also found that most of the major jitter components varied at a rate less than 0.5 cps. Since many communication systems require information about signal timing for their proper operation, the multipath jitter analysis presented should be particularly useful in predicting system performance and in establishing design criteria.

APPENDIX—DERIVATION OF ERROR PROBABILITIES ASSOCIATED WITH THE MULTIPATH JITTER DISTRIBUTION FUNCTION

Symmetrical Pulse Sampling Case

In a binary FSK communications system, the output of the receiver, represented by the FSK detector of Figure 3, would be sampled by decision circuitry to determine the presence of a zero or a one. This sampling is normally done in the center of the keying element, in order to provide maximum protection from transitions displaced by multipath jitter. It was shown that the distribution of transition locations can be approximated by the following equation:

$$P(x \geq X) = e^{-NX}, \qquad\qquad X \geq 0 \qquad\qquad (7)$$

where X represents the peak-to-peak value of the jitter, N is a constant determined by the condition of the transmission medium, and $P(x \geq X)$ is the probability that the jitter (x) is equal to or greater than X. (Multiplying $P(x \geq X)$ by 100 gives the probability in per cent.) This distribution function can also be represented as $P(x \leq X) = 1 - e^{-NX}$, which expresses the probability that x is equal to or less than X. In both cases the representations are for one-sided distributions, i.e., $X \geq 0$. Let a new parameter $y = x/2$ be defined such that

it represents the symmetrical displacement of the transition both plus and minus from a mean value of zero. The distribution function would take the following form:

$$P(y \leqq Y) \begin{cases} = 1 - \dfrac{1}{2} e^{-2NY}, & Y \geqq 0 \\[2mm] = \dfrac{1}{2} e^{2NY}, & Y \leqq 0. \end{cases}$$

(8)

The probability density function $p(y)$ is defined as the derivative of the probability distribution function such that

$$p(Y) = \frac{dP(y \leqq Y)}{dY}, \text{ where } P(y \leqq Y) = \int_{-\infty}^{Y} p(y)\,dy.$$

(9)

The density function $p(y)$ determined from Equation (8) is

$$p(y) = Ne^{-2N|y|}.$$

(10)

A sketch of this statistical model is shown in Figure 16.
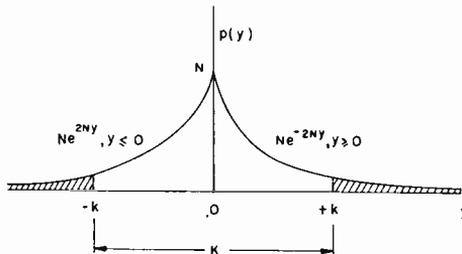
Fig. 16

If locations $-k$ and $+k$ are considered the sampling points, errors occur whenever $|y| \geqq k$. Note that $2k = K$ is assumed to be the time duration of a keying element. The probability of error, P_e, can thus be represented as the summation of the shaded areas under the curves multiplied by the probability that a transition occurs. If 0.5 is assumed for the occurrence of a transition (i.e., equal numbers of zeros and ones), then,

$$P_e = \frac{1}{2} \left[\int_{-\infty}^{y=-k} Ne^{2Ny}\,dy + \int_{y=+k}^{\infty} Ne^{-2Ny}\,dy \right] = \frac{1}{2}\,e^{-2Nk}. \quad (11)$$

This is the expected result, which could have been inferred from the original distribution function, Equation (7). However, the model considered in Figure 16 is a reasonable representation of the transition time displacements actually recorded, and is easily extended to the asymmetric pulse sampling case.

Asymmetric Pulse Sampling Case

In a synchronous communications system there are often transmission delays, which prevent precise location of the receiver sampling pulses at the center of the keying elements. This results in degraded performance (increased error rates), the amount of degradation being dependent upon the amount of asymmetry present. The statistical model considered for this case is illustrated in the insert of Figure 7. The probability of error (P_e) for this configuration is determined by the summation of unequal shaded areas indicated, multiplied by the probability of a transition occurrence. For an assumed 0.5 probability of transition occurrence, P_e is given as

$$P_e = \frac{1}{2} \left[\int_{-\infty}^{y=-(k+\epsilon)} Ne^{2Ny}\,dy + \int_{y=(k-\epsilon)}^{+\infty} Ne^{-2Ny}\,dy \right]$$

$$= \frac{1}{4}\,e^{-2Nk}\,[e^{-2N\epsilon} + e^{+2N\epsilon}]$$

$$P_e = \frac{1}{2}\,e^{-2Nk}\cosh 2N\epsilon = \frac{1}{2}\,e^{-NK}\cosh 2N\epsilon. \quad (12)$$

Equation (12) is the expression plotted in Figure 7. When $\epsilon = 0$, this equation reduces to the symmetric case (Equation (11)).

THE NOISE-POWER PROBABILITY DISTRIBUTION IN A MULTIHOP FM RADIO-RELAY SYSTEM*

BY

JACQUES DUTKA

RCA Surface Communications Division,
and Columbia University
New York, N. Y.

Summary—The noise-power distribution is derived for a multihop FM radio-relay system in which each hop has identical characteristics (e.g., transmitter power, noise figure, and hop length) and is subject to independent Rayleigh fading. The derivations are in the form of integrals involving Bessel functions, and numerical values are computed for particular cases. The asymptotic distribution for a certain normalized sum as the number of hops becomes infinite is also determined.

INTRODUCTION

IN THE design of a radio-relay system, it is of major importance to provide a sufficient but not excessive reserve of effective transmitter power to overcome fading. Reasonably good estimates can be made as to the probability distribution of the depths of fade for a single relay hop. Considerable data has been accumulated, and fading distributions can be related to frequency, time-of-day, and season. For the situation of a radio-relay system consisting of two-or-more hops in tandem, however, the literature dealing with the over-all distribution of noise power due to fading in each relay hop appears to be sparse and inadequate.

In order to design a frequency-modulated radio-relay system made up of n hops, all identical with respect to hop lengths and equipment characteristics such as transmitter power, antenna gains, receiver noise figure, receiver bandwidth, and peak deviation, it is necessary to know the probability distribution of the final signal-to-noise power ratio (or noise power relative to a fixed signal level). This paper presents a solution to this problem based on the assumption that each hop is subjected to the same fading distribution (in this case, the Rayleigh distribution). When diversity is employed on each hop, the analytical formulation of the problem poses no new difficulties. It has been shown experimentally that in the vicinity of 4-to-8 kilomegacycles, a moderately long line-of-sight hop with proper clearance will be

---

* Manuscript received 18 July 1961.

508

subject to a Rayleigh distribution of fades for the worst month of the year; with decreasing frequency, the fading distribution is generally less severe than Rayleigh.

In practice, one does not often encounter, in a system of many tandem hops, the situation where all hops simultaneously undergo Rayleigh fading. Nevertheless, the solution to the problem as stated does set an upper bound to the resultant distribution, and is valuable for that reason.

If a frequency-modulated signal is applied to the input of a non-ideal (noisy) receiver operating above threshold, the instantaneous amplitude and phase of the intermediate-frequency signal would be corrupted by this noise. In a multihop system this i-f signal is further processed in two ways. At some repeaters, the i-f signal is brought up to a new final frequency and amplified prior to retransmission. In this case all peak amplitude fluctuations are wiped clean by a-g-c and/or limiting. In other repeaters, the i-f signal is limited and detected prior to remodulation and retransmission. Here, too, a signal having no peak amplitude fluctuations is retransmitted. In both cases, the noise is retransmitted as phase disturbances; the per-hop contribution has a Rayleigh distribution. The over-all noise power of the system is then the sum of the noise powers for each hop, and the probability distribution of the over-all noise power is the statistical summation of the individual hop distributions.

The mathematical problem of determining the probability distribution of the noise power for an n-hop FM radio relay system was previously stated but left unsolved in a paper by S. Matsuo et al.[1] This paper also treated various kinds of fading problems and obtained approximate results in some cases. More recently, L. F. Fenton,[2] considered a similar problem on the assumption of a log-normal distribution of fades.

In the present paper, there is given a mathematical discussion of the problem in which general expressions for the required probability distribution are obtained in the form of integrals. For a particular case, $n = 2$, the distribution was obtained in explicit form, but in the general case the integrals were evaluated numerically with the aid of a computer for a certain range of numbers of hops and for particular levels. The results thus obtained are reproduced in Table I.

[1] S. Matsuo et al, *Problems Concerning the Reliability of Microwave Relay Links*, published by Nippon Telegraph and Telephone Public Corp., August, 1953.

[2] L. F. Fenton, "The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems," *I.R.E. Trans. Communication Systems*, p. 57, March, 1960.

TABLE I

Noise Power Probability Distribution for an *n*-Hop FM Radio Relay,
Assuming Rayleigh-Distributed Fading, With No Diversity.

(0 db = Median Level for One Hop)

Number of Hops	Noise Power (N) Relative to Median Level for one Hop		Per cent of Time Noise Power is Greater than N
n	*x*	*db*	*%*
2	4	6.0	36.57
	10	10.0	15.12
	100	20.0	1.43
5	5	7.0	81.69
	6	8.0	73.53
	8	9.0	59.43
	16	12.0	30.16
	30	14.8	14.93
	100	20.0	3.90
10	10	10.0	93.09
	15	11.8	76.21
	30	14.8	39.22
	60	17.8	16.91
	100	20.0	9.08
	1000	30.0	0.73
15	15	11.8	97.17
	20	13.0	89.18
	30	14.8	67.69
	50	17.0	38.68
	100	20.0	15.86
	1000	30.0	1.12
30	30	14.8	99.77
	50	17.0	90.46
	80	19.0	61.56
	160	22.0	24.33
	1000	30.0	2.39
	10000	40.0	0.22
60	100	20.0	97.75
	200	23.0	56.78
	400	26.0	19.93
	1000	30.0	5.59
	10000	40.0	0.44
120	360	25.7	77.74
	1000	30.0	15.69
	10000	40.0	0.90
	100000	50.0	0.09

Statement of the Problem

Suppose that the voltage S/N for an individual radio-relay hop has a Rayleigh distribution of the form $P(S/N \leq x) = 1 - \exp(-\lambda x^2)$ where λ is a positive constant. The power S/N is then exponentially distributed with a probability density $\lambda \exp(-\lambda x)$ for $x > 0$. If the signal is kept at a constant unit power level, then the probability density of noise power is given by

$$f(x;\lambda) = \frac{\lambda}{x^2} e^{-\lambda/x} \text{ for } x > 0; \qquad f(x;\lambda) = 0 \text{ for } x \leq 0, \qquad (1)$$

which is a Type V probability density in Pearson's system of curves. This distribution has no finite moments.

The general problem considered here is to find the distribution function of the over-all noise power

$$Y_n = \sum_{m=1}^{n} X_m \qquad (2)$$

where $\{X_m\}$ is a sequence of independent random variables, and X_m, the noise power of the mth hop, has the probability density $f(x;\lambda_m)$.

The case of rth order diversity on each hop is treated briefly later.

Distribution of Y_2

The distribution of Y_2, which can actually be obtained in closed form, will first be derived. It follows from the above that

$$P(X_1 + X_2 \leq y) = \iint\limits_{x_1 + x_2 \leq y} f(x_1;\lambda_1) f(x_2;\lambda_2) \, dx_1 dx_2$$

$$= \lambda_1 \lambda_2 \int_0^y \exp\left\{-\frac{\lambda_1}{x_1}\right\} \frac{dx_1}{x_1^2} \int_0^{y-x_1} \exp\left\{-\frac{\lambda_2}{x_2}\right\} \frac{dx_2}{x_2^2}$$

$$= \lambda_1 \int_0^y \exp\left\{\frac{\lambda_1}{x} + \frac{\lambda_2}{y-x}\right\} \frac{dx}{x^2}$$

$$P(X_1 + X_2 \leq y) = \lambda_1 \int_0^y \exp\left\{-\frac{\lambda_1 y + (\lambda_2 - \lambda_1)x}{x(y-x)}\right\} \frac{dx}{x^2}.$$

To simplify the integral, let $t = \alpha/x$ where $\alpha^2 = \sqrt{\lambda_1/\lambda_2}$. Then

$$P(X_1 + X_2 \leqq y) = \frac{\lambda_1}{\alpha} \int_{\alpha/y}^{\infty} \exp \left\{ -\frac{\lambda_1 y + (\lambda_2 - \lambda_1)\dfrac{\alpha}{t}}{\dfrac{\alpha}{t}\left(y - \dfrac{\alpha}{t}\right)} \right\} dt$$

$$= \frac{\lambda_1}{\alpha} \int_{\alpha/y}^{\infty} \exp \left\{ -\frac{\lambda_1 y t^2 + \alpha(\lambda_2 - \lambda_1)t}{\alpha t y - \alpha^2} \right\} dt.$$

Let $\alpha t y - \alpha^2 = s$. Then, after some simplification,

$$P(X_1 + X_2 \leqq y) = \frac{\sqrt{\lambda_1\lambda_2}}{y} \exp \left\{ -\frac{\lambda_1 + \lambda_2}{y} \right\} \cdot$$

$$\int_0^{\infty} \exp \left\{ -\frac{\sqrt{\lambda_1\lambda_2}}{y}\left(s + \frac{1}{s} \right) \right\} ds,$$

and it remains to evaluate the integral on the right. From a result given by Watson[3]

$$K_1(z) = \frac{z}{4} \int_0^{\infty} \exp \left\{ -\left(\tau + \frac{z^2}{4\tau} \right) \right\} \frac{d\tau}{\tau^2}, \ R(z^2) > 0.$$

Now let $u = \beta/\tau$ and $z = 2\beta$ where $\beta = \sqrt{\lambda_1\lambda_2/y}$. Then

$$2K_1(2\beta) = \int_0^{\infty} \exp \left\{ -\beta \left(u + \frac{1}{u} \right) \right\} du,$$

and on substituting this result in the equation above,

$$P(X_1 + X_2 \leqq y) = \frac{2\sqrt{\lambda_1\lambda_2}}{y} \exp \left\{ -\frac{\lambda_1 + \lambda_2}{y} \right\} K_1 \left(\frac{2\sqrt{\lambda_1\lambda_2}}{y} \right)$$

[3] G. N. Watson, *Bessel Functions*, 2nd edition, Eq. (15), p. 183, Cambridge University Press, London, 1944.

or, if $H_2(y) = P(Y_2 \leqq y)$ is the distribution function of Y_2,

$$H_2(y) = \frac{2\sqrt{\lambda_1\lambda_2}}{y} \exp\left\{-\frac{\lambda_1+\lambda_2}{y}\right\} K_1\left(\frac{2\sqrt{\lambda_1\lambda_2}}{y}\right)$$

DISTRIBUTION OF Y_n

It does not seem possible to extend the convolution method of the preceding section to obtain a closed form for $H_n(y)$, the distribution function of Y_n. Instead, an integral representation of $H_n(y)$ will be obtained using the method of characteristic functions.

The characteristic function $\phi(t;\lambda)$ associated with the probability density $f(x;\lambda)$ *is*

$$\phi(t;\lambda) = \int_{-\infty}^{\infty} e^{itx} f(x;\lambda)\,dx$$

$$= \int_0^{\infty} \frac{\lambda}{x^2} e^{itx} e^{-\lambda/x}\,dx = \int_0^{\infty} \exp\left\{-\left(\frac{\lambda}{x} - itx\right)\right\} \frac{\lambda\,dx}{x^2},$$

and on substituting $x = v\sqrt{\lambda/t}\ e^{i\pi/4}$,

$$\phi(t;\lambda) = \sqrt{\lambda t}\ e^{-i\pi/4} \int_0^{\infty(\exp-i\pi/4)} \exp\left\{-\sqrt{\lambda t}\ e^{-i\pi/4}\left(v + \frac{1}{v}\right)\right\} \frac{dv}{v^2}$$

and it follows from a result given by Watson,[*] that the integral on the right is equal to $2K_1(\sqrt{4\lambda t}\ e^{-i\pi/4})$. Hence

$$\phi(t;\lambda) = \sqrt{4\lambda t}\ e^{-i\pi/4} K_1(\sqrt{4\lambda t}\ e^{-i\pi/4}).$$

Since X_m is a continuous random variable, it follows that

$$Y_n = \sum_{m=1}^{n} X_m$$

is a continuous random variable, and, since $X_1, X_2, \cdots, X_n$ is a set of independent random variables, it follows from the theory of charac-

---

[*] See Reference 3, p. 182, Eq. (8).

teristic functions that the characteristic function of Y_n is

$$\mu_n(t) = \prod_{m=1}^{n} \phi(t;\lambda_m)$$

$$= \prod_{m=1}^{n} \sqrt{4\lambda m t}\, e^{-i\pi/4} K_1(\sqrt{4\lambda m t}\, e^{-i\pi/4})$$

and the probability density of Y_n is

$$h_n(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \mu_n(t)\, dt,\ y > 0$$

while the distribution function of Y_n is

$$H_n(y) - H_n(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1 - e^{-ity}}{it} \mu_n(t)\, dt,$$

where $H_n(0) = 0$.

It does not appear possible to evaluate $h_n(y)$ and $H_n(y)$ in closed form in the general case, so that to obtain numerical results it is necessary to evaluate the above integrals or equivalent forms by numerical integration techniques. Now these integrals are not well suited for numerical evaluation—principally because of the presence of the product

$$\prod_{m=1}^{n} K_1(\sqrt{4\lambda_m t}\, e^{-i\pi/4})$$

in the integrand. So far as is known to the writer, no tables have ever been published from which $K_1(\rho e^{-i\pi/4})$, $\rho > 0$ can be obtained directly, say to seven decimal places, nor is there any convenient algorithm for the calculation of this function (e.g., the calculation of values of $K_1(\rho e^{-i\pi/4})$ from its series representation requires a fair amount of computational effort).

Because of the foregoing, it is convenient to transform the integral representation of $h_n(y)$ by finding an equivalent integral along a suitable contour in the complex plane. This is done in the next section.

The analysis developed above may readily be extended to the case

of rth order diversity on each hop.[†] In this case, the probability density for the linear addition of the power S/N is $(\lambda^{r+1}/\Gamma(r+1))\,x^r e^{-x}$ for $x > 0$, and, if the signal is kept at a constant unit power level, then the probability density of the noise power is given by

$$f_r(x;\lambda) = \frac{\lambda^{r+1}}{\Gamma(r+1)x^{r+2}}\,e^{-\lambda/x} \text{ for } x > 0; \ f_r(x;\lambda) = 0 \text{ for } x \leq 0.$$

The characteristic function of the noise power in this case is

$$\phi_r(t;\lambda) = \frac{2(\sqrt{\lambda t}\,e^{-i\pi/4})^{r+1}}{\Gamma(r+1)}K_{r+1}(\sqrt{4\lambda t}\,e^{-i\pi/4})$$

and the distribution function of the over-all noise power can be obtained in a manner analogous to the derivation of $H_n(y)$ above.

A Contour Integral for $H_n(y)$

An expression for $h_n(y)$ was obtained above in the form

$$h_n(y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ity}\,\mu_n(t)\,dt.$$

Consider the integration of the function $\mu_n(z)\exp(iyz)/(2\pi)$ around a contour consisting of a segment $[\rho, R]$ with $0 < \rho < R$, a quarter circle extending from R to $-iR$, a segment $[-iR, -i\rho]$ and the quarter circle extending from $-i\rho$ to ρ. Now $\mu_n(-t) = \overline{\mu_n(t)}$, and as $R \to \infty$ and $\rho \to 0$ it may be shown that

$$h_n(y) = \frac{1}{4\pi i}\int_0^\infty \exp\left\{-\frac{yt^2}{4}\right\}\left\{\prod_{m=1}^{n}\left[-it\sqrt{\lambda_m}\,K_1(-it\sqrt{\lambda_m})\right]\right.$$
$$\left. -\prod_{m=1}^{n}\left[it\sqrt{\lambda_m}\,K_1(it\sqrt{\lambda_m})\right]\right\}\,dt.$$

If both sides of this equation are integrated with respect to y over

---

[†] An interesting analysis of some types of diversity combining schemes is given by D. G. Brennan, "Linear Diversity Combining Techniques," *Proc. I.R.E.*, Vol. 47, No. 6, p. 1075, June, 1959.

the range (y, ∞) and the order of integration is inverted on the right (a procedure which may be justified by a theorem in T. Bromwich[4]), then the form of the distribution function sought is

$$1 - H_n(y) = \frac{1}{\pi i} \int_0^\infty \left\{ \prod_{m=1}^n \left[-it\sqrt{\lambda_m} K_1(-it\sqrt{\lambda_m}) \right] \right.$$

$$\left. - \prod_{m=1}^n \left[it\sqrt{\lambda_m} K_1(it\sqrt{\lambda_m}) \right] \right\} \exp \left\{ -\frac{yt^2}{4} \right\} \frac{dt}{t} .$$

In particular, for the case $n = 1$,

$$1 - H_1(y) = \frac{1}{\pi i} \int_0^\infty \left[-it\sqrt{\lambda_1} K_1(-it\sqrt{\lambda_1}) \right.$$

$$\left. - it\sqrt{\lambda_1} K_1(it\sqrt{\lambda_1}) \right] \exp \left\{ -\frac{yt^2}{4} \right\} \frac{dt}{t}$$

and since $K_1(-it\sqrt{\lambda}) = (-\pi/2) \left[J_1(t)\sqrt{\lambda}) + iY_1(t\sqrt{\lambda}) \right]$ while $K_1(it\sqrt{\lambda}) = \overline{K_1(-it\sqrt{\lambda})}$,

$$1 - H_1(y) = \frac{1}{\pi i} \int_0^\infty \left[\pi it\sqrt{\lambda_1} J_1(t\sqrt{\lambda_1}) \right] \exp \left\{ -\frac{yt^2}{4} \right\} \frac{dt}{t}$$

$$= \sqrt{\lambda_1} \int_0^\infty J_1(t\sqrt{\lambda_1}) \exp \left\{ -\frac{yt^2}{4} \right\} dt,$$

and by a result in Watson,[*] the integral can be evaluated, giving

$$1 - H_1(y) = \sqrt{\lambda_1} \sqrt{\frac{\pi}{y}} \exp \left\{ -\frac{\lambda_1}{2y} \right\} I_{1/2} \left(\frac{\lambda_1}{2y} \right) ,$$

[4] T. Bromwich, *An Introduction to the Theory of Infinite Series*, 2nd Ed. revised with assistance of T. M. MacRobert, Macmillan and Co., Ltd., London, 1942; p. 504.

[*] See Reference 3, p. 394.

$$= 1 - \exp \left\{ - \frac{\lambda_1}{y} \right\} .$$

The same result is obtained by integrating $f(x, \lambda_1)$ (defined in the first section) with respect to x over the range (y, ∞).

For $n = 2$,

$$1 - H_2(y) = \frac{1}{\pi i} \int_0^\infty \left\{ (-it)^2 \sqrt{\lambda_1 \lambda_2} \, K_1(-it\sqrt{\lambda_1}) \, K_1(-it\sqrt{\lambda_2}) \right.$$

$$- (it)^2 \sqrt{\lambda_1 \lambda_2} \, K_1(it) \sqrt{\lambda_1}) \, K_1(it\sqrt{\lambda_2}) \, \exp \left\{ - \frac{yt^2}{4} \right\} \frac{dt}{t}$$

$$= - \frac{\pi}{2} \int_0^\infty [J_1(t\sqrt{\lambda_1}) \, Y_1(t\sqrt{\lambda_2}) + J_1(t\sqrt{\lambda_2}) \, Y_1(t\sqrt{\lambda_1})] \cdot$$

$$\cdot \exp \left\{ - \frac{yt^2}{4} \right\} \frac{dt}{t} .$$

If this result is compared with the value of $1 - H_2(y)$ which was obtained in the second section, the following relation results:

$$\frac{2\sqrt{\lambda_1 \lambda_2}}{y} \exp \left\{ - \frac{\lambda_1 \lambda_2}{y} \right\} K_1 \left(\frac{2\sqrt{\lambda_1 \lambda_2}}{y} \right) - 1$$

$$= \frac{\pi}{2} \int_0^\infty [J_1(t\sqrt{\lambda_1}) \, Y_1(t\sqrt{\lambda_2}) + J_1(t\sqrt{\lambda_2}) \, Y_1(t\sqrt{\lambda_1})] \exp \left\{ - \frac{yt^2}{4} \right\} \frac{dt}{t} ,$$

which appears to be a new result in the theory of infinite integrals of products of Bessel functions.

NUMERICAL EVALUATION OF $H_n(y)$ IN PARTICULAR CASES

In order to solve the problem which gave rise to this paper, it is necessary to make certain simplifying assumptions. Since there is no specific information available on the distribution of the magnitudes of $\lambda_1, \lambda_2, \cdots, \lambda_n$ for the chain of n radio relay links, it is assumed that $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$, and λ is chosen in such a way as to make

the median of the distribution for one link coincide with the point $x = 1$. That is,

$$\int_0^1 f(x; \lambda) \, dx = e^{-\lambda} = \frac{1}{2},$$

which yields the value $\lambda = \ln 2$. Thus in this case, the form of the distribution function $H_n(y)$ reduces to

$$1 - H_n(y) = \frac{1}{\pi i} \int_0^\infty \left\{ [-it\sqrt{\lambda} \, K_1(-it\sqrt{\lambda})]^n - \lfloor it\sqrt{\lambda} \, K_1(it\sqrt{\lambda}) \rfloor^n \right\} \cdot$$

$$\cdot \exp \left\{ -\frac{yt^2}{4} \right\} \frac{dt}{t}.$$

or

$$H_n(y) = 1 - \int_0^\infty \exp \left\{ -\frac{yt^2}{4\lambda} \right\} \Phi_n(t) \, dt,$$

where

$$\Phi_n(t) = \frac{[-itK_1(-it)]^n - [itK_1(it)]^n}{\pi it}$$

The presence of the strongly damping factor $\exp(-yt^2/(4\lambda))$ in the integrand above suggests that this form is well adapted for numerical integration in contrast to the form for $H_n(y)$ obtained in the third section. But first it is necessary to put $\Phi_n(t)$ in a simpler form. Let

$$-Y_1(t) + iJ_1(t) = re^{i\theta}$$

where

$$r = \sqrt{J_1^2(t) + Y_1^2(t)}$$

$$\theta = \arcsin \left(\frac{J_1(t)}{r} \right).$$

Then

$$\pi t \Phi_n(t) = 2 \, \text{Im} \, \{ [-itK_1(-it) \, |^n \}$$

$$= 2 \left(\frac{\pi t}{2} \right)^n \operatorname{Im} \{ [- Y_1(t) + i J_1(t)]^n \}$$

$$= 2 \left(\frac{\pi t}{2} \right)^n r^n \sin n\theta.$$

Thus the expression for $H_n(y)$ above reduces to

$$H_n(y) = 1 - \left(\frac{\pi}{2} \right)^{n-1} \int_0^\infty t^{n-1} \exp \left\{ - \frac{y t^2}{4\lambda} \right\} r^n \sin n\theta \, dt$$
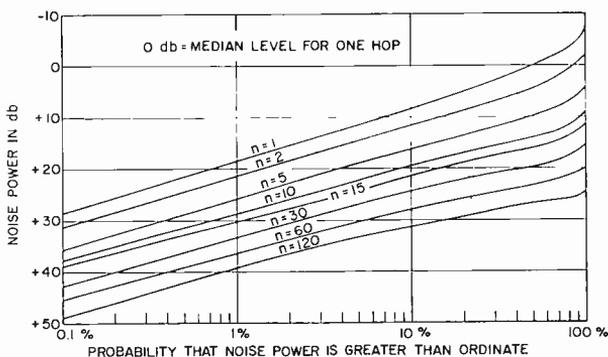
Fig. 1—Noise-power probability distribution for an n-hop FM radio relay (assuming Rayleigh-distributed fading with no diversity).

and this was the form adopted for numerical evaluation with the use of the computer. In the case of rth order diversity, an analogous expression may be obtained for the distribution function.

In the numerical evaluation of the integral for $H_n(y)$ above, use was made of some accurate polynomial approximations to Bessel functions developed by E. E. Allen.[5] The integral was evaluated using Simpson's rule with different choices of intervals of integration, and an analysis of the integration, inherent, and round-off errors showed that the final results could be accepted as having four-digit accuracy. The particular case when $n = 2$ was used as a check on the corresponding values obtained from the computer. The actual values obtained are given in Table I and the curves are shown in Figure 1.

[5] E. E. Allen, *Mathematical Tables and Other Aids to Computation*, Vol. 1, p. 394, Published by National Research Council, Washington, D. C., 1954.

Asymptotic Distribution of Y_n

Let $\{\overline{X}_m\}$ denote a sequence of independent random variables, each of which has a common distribution function $F(x)$, and let Y_n denote the sum $\sum_{m=1}^{n} \overline{X}_m$. It is well known that if certain conditions are satisfied, it is possible to find sequences of real numbers $\{\overline{A}_n\}$ and $\{\overline{B}_n\}$, $\overline{B}_n > 0$, such that the distribution function of the normalized sum $(\overline{Y}_n/\overline{B}_n) - \overline{A}_n$ approaches a limiting distribution as $n \to \infty$.

It frequently happens in many cases which arise in practice that the limiting distribution of the normalized sum is the Gaussian (normal) distribution. That this is not the case here can be seen by applying the following theorem given in Gnedenko and Kolmogoroff.[6]

The distribution function $F(x)$ belongs to the domain of attraction of a (proper) normal law if, and only if, as $T \to \infty$,

$$\frac{T^2 \displaystyle\int_{|x| > T} d\overline{F}(x)}{\displaystyle\int_{|x| > T} x^2 d\overline{F}(x)} \to 0.$$

In the present problem the limit of the ratio is unity.

However, it can be shown in our problem that a limiting distribution, in the sense described above, does exist. For, as is proved in Gnedenko and Kolmogoroff[*], if two sequences of real numbers $\{A_n\}$ and $\{B_n\}$, $B_n > 0$, are defined where

$$A_n \sim -2 \ (ln \ 2 - \gamma - 1/2) + \ln \ (4/n),$$

$$B_n = \lambda \cdot n,$$

then it may be shown that the characteristic function of the normalized sum $(Y_n/B_n) - A_n$, which is $\mu_n(t/B_n) \exp(-iA_nt)$, approaches

$$\mu(t) = \exp \left\{ -\frac{\pi |t|}{2} \left[1 + \frac{2i}{\pi} \operatorname{sgn} (t \ln |t|) \right] \right\}$$

---

[6] B. V. Gnedenko and A. N. Kolmogorov, (translated by K. L. Chung) *Limited Distributions for Sums of Independent Random Variables*, p. 172, Addison Wesley Pub. Co., Cambridge, Mass., 1954.

[*] See Reference 6, p. 175 ff, p. 164 ff, etc.

for fixed t as $n \to \infty$. It follows that the limiting distribution function $G(y)$ of the normalized sum $(Y_n/B_n) - A_n$ is

$$G(y) - G(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1 - e^{-ity}}{it} \mu(t)\,dt$$

$$= \frac{1}{\pi} \int_{0}^{\infty} \exp\left\{-\frac{\pi t}{2}\right\} [\sin(ty + t\ln t) - \sin(t\ln t)]\frac{dt}{t}.$$

For large n, this distribution can be taken as an approximation to the distribution of $(Y_n/B_n) - A_n$, but it is difficult to obtain usable estimates of the magnitude of the resulting error.

AN INVESTIGATION OF SEQUENTIAL DECODING*†

By

WALTER R. WADDEN, DAVID M. JONES, AND JAMES PENNYPACKER

RCA Aerospace Communications and Controls Division
Burlington, Mass.

Summary—The technique of sequential decoding described by Wozencraft makes possible the practical implementation of error-correcting codes having large constraint lengths. This paper describes another sequential decoding scheme, and presents theoretical results on the associated average number of decoding steps and the probability of error. Simulation studies of this decoding scheme are currently being performed on a digital computer. A summary is given of the results obtained to date.

INTRODUCTION

THIS paper is concerned with a means of communicating reliably over a binary symmetric channel (BSC). The BSC is a convenient abstraction which some communication media approximate. Its input consists of sequences of binary signals, which represent zeros and ones. Noise in the channel causes some of these signals
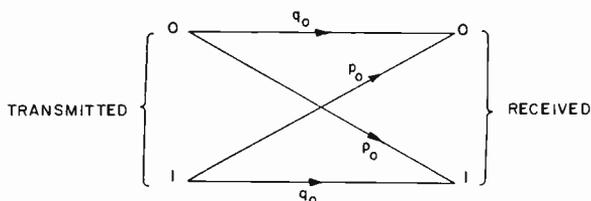
Fig. 1—Binary symmetric channel.

to be delivered incorrectly to the receiver. This noise is statistically defined by the transition probability $p_0 < 1/2$, which is the probability that a transmitted zero will be received as a one, and vice-versa. $q_0 = 1 - p_0$ denotes the probability that a transmitted symbol will be received correctly. The BSC is illustrated in Figure 1.

In the coded-communications situation under consideration, the

channel input is a transformed or "encoded" version of some information sequence of binary digits. This information sequence is denoted by $I(1), I(2), \ldots I(i), \ldots$ and we assume that zeros and ones are equally likely to occur within it.

The transmitted sequence is a function of the I's, and it is generated by the encoder. When it is delivered to the receiver after transmission, some of its digits may have been altered by noise in the channel. The receiver attempts to reconstruct the original information sequence from the received sequence, a process known as decoding.

If the information sequence is transmitted without preliminary encoding, the probability of communicating $I(i)$ incorrectly is p_0. By coding the information sequence into a longer sequence and supplying the receiver with an appropriate decoder, this probability of error can be reduced. If, for instance, the encoder simply triplicates each information bit, the transmitted sequence will be

$$I(1), I(1), I(1), I(2), I(2), I(2), \ldots I(i), I(i), I(i), \ldots$$

and a probability of error smaller than p_0 can be achieved; the decoder makes a maximum likelihood estimate of $I(i)$ by summing the corresponding three digits in the received sequence, printing zero if the sum is zero or one, printing one if the sum is two or three. The probability of error is

$$p_0^3 + 3p_0^2 q_0 < p_0,$$

since $p_0 < 1/2$.

Generalizing, if the encoder repeats each information bit $(2n + 1)$ times and the decoder employs a maximum likelihood estimate, the probability of error is a function of n which approaches zero as $n \to \infty$.

The *rate of transmission*, denoted by R, is the ratio of the number of information digits to the number of corresponding transmitted digits. In the simple coding scheme outlined above, $R = 1/(2n + 1)$. Hence, $R \to 0$ as the probability of error approaches zero. The cost of achieving an arbitrarily low frequency of decoding error in this instance is a correspondingly low rate of transmission.

By adopting a more sophisticated encoding/decoding technique a very low probability of error can be achieved while maintaining a respectable rate of transmission. In fact, Shannon[1] has shown that, for a given channel, an arbitrarily small frequency of error can be

---

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Illinois (1949).

achieved by suitable coding, providing only that R does not exceed a constant known as the channel capacity. For the BSC this constant is $1 + p_0 \log_2 p_0 + q_0 \log_2 q_0$.

This paper describes a coding scheme for which the probability of error can be made arbitrarily small while a fixed positive rate of transmission is maintained (not exceeding channel capacity). The encoding procedure described by Wozencraft[2,3] is a modified version of a convolutional encoding scheme considered by Elias. Wozencraft also specified a procedure for decoding sequentially, i.e., digit by digit. The decoding procedure described in the present paper is also sequential but it differs considerably from Wozencraft's.

The symbol $\oplus$ will be used to denote modulo-2 addition: if A and B are binary digits, then $A \oplus B = 0$, if $A = B$; $A \oplus B = 1$, if $A \neq B$.

The Hamming Distance between two binary sequences of equal length is defined to be the total number of ones obtained by forming the modulo-2 sums of corresponding digits. It is a measure of the extent to which the two sequences differ from each other.

CONVOLUTIONAL ENCODING

Convolutional encoding takes place in two stages. During the first stage, a number of zeros are inserted between the consecutive information bits of the message. $(n_0 - 1)$ zeros are inserted in each case, where n_0 is the reciprocal of R, the required rate of transmission (it is assumed that the reciprocal of R is an integer, but the method and analysis can be extended to cases where this is not so). For example, if $R = 1/2$ the resulting sequence is

$$I(1),\ 0,\ I(2),\ 0,\ \ldots,\ I(i),\ 0,\ \ldots$$

The other stage involves a fixed sequence of $N n_0$ binary digits, called the convolutional generator. It is denoted by

$$G(1),\ G(2),\ G(3),\ \ldots,\ G(N n_0)\ ;$$

$N n_0$ is called the convolutional constraint-length.

The process may best be described with reference to Figure 2, where the case $n_0 = 2$ is illustrated. The sequence of information bits

[2] J. M. Wozencraft, "Sequential Decoding for Reliable Communications," Technical Report No. 325, *Research Laboratory of Electronics, MIT*, August 9 (1957).

[3] J. M. Wozencraft and B. Reiffen, *Sequential Decoding*, Technology Press and Wiley (1961).

and interspersed zeros passes through a shift register Nn_0 bits long. Associated with each bit of the shift register, one-to-one, is a digit of the generator. Before the computation of a transmitted digit, the sequence is shifted one to the right. The transmitted digit is then computed as follows: each generator digit is multiplied by the corresponding digit in the shift register, and the sum mod-2 of these products is formed. Thus, in the case illustrated, the transmitted digit is

$$I(i)\ G(1) \oplus I(i-1)\ G(3) \oplus \ldots \oplus I(i-N+1)\ G(2N-1).$$

SHIFT REGISTER
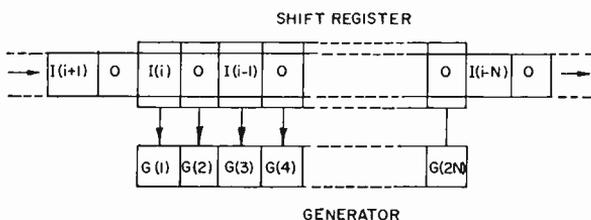
Fig. 2—Encoding for $R = 1/2$.

The sequence is again shifted one to the right, and the next transmitted digit similarly computed. In the case shown it is

$$I(i)\ G(2) \oplus I(i-1)\ G(4) \oplus \ldots \oplus I(i-N+1)\ G(2N).$$

This cycle of shifting and computing continues until all the information bits have passed through the shift register. The resulting sequence of mod-2 sums is the encoded (i.e., transmitted) message.

It is arbitrarily assumed that, before encoding begins, the shift register contains all zeros and the first information bit $I(1)$ is ready to enter. The encoded message will therefore be

$$I(1)\ G(1), I(1)\ G(2), \ldots, I(1)\ G(n_0), I(2)\ G(1) \oplus I(1)\ G(n_0+1),$$

$$I(2)\ G(2) \oplus I(1)\ G(n_0+2), \ldots, I(2)\ G(2n_0) \oplus I(1)\ G(2n_0), \ldots \text{etc.}$$

THE MESSAGE SET

It will be observed that, in general, each transmitted digit is a function of N information bits, and *the influence of each information bit extends over Nn_0 consecutive transmitted digits.* Wozencraft[2,3]

shows that this dependency gives the set of all possible transmitted messages a treelike structure, as illustrated in Figure 3 for the case $n_0 = 2$.

Using Wozencraft's nomenclature, the tree consists of branches, which are connected at nodes. The encoding procedure is equivalent to tracing a path, from left to right, through the tree, and transmitting the digits on the branches lying along this path. Each branch holds n_0 transmitted digits, which are mod-2 sums obtained after a new information bit has entered the shift register. Two branches stem from each node, and the encoder selects the upper, or lower, branch according as the new information bit shifted in is 0 or 1.
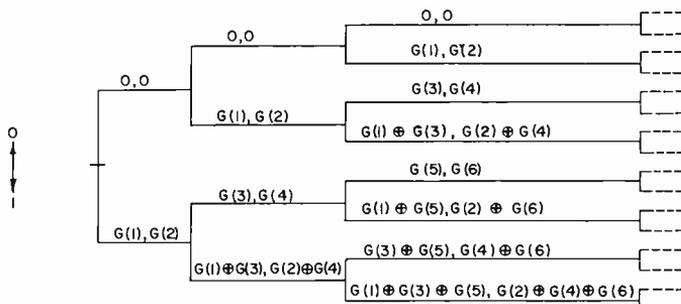
Fig. 3—A convolutional message set.

Thus, the leftmost two branches in the tree contain the two possible configurations for the first n_0 transmitted digits, corresponding to $I(1) = 0$ and $I(1) = 1$, respectively. The next stage in the tree has four branches, containing the second n_0 transmitted digits for the four possible values of the pair $I(1)$, $I(2)$. The number of branches continues to double at each subsequent stage. In Figure 3 the tree is shown only up to its third stage, but of course it actually extends to the full length of the message. The set of all possible transmitted messages, or the convolutional message set, can be represented by a tree of infinite length.

Example

If $I(1) = 1$, $I(2) = 1$, $I(3) = 0$, the first six transmitted digits will be

$$G(1), G(2), G(1) \oplus G(3), G(2) \oplus G(4), G(3) \oplus G(5), G(4) \oplus G(6).$$

SEQUENTIAL DECODING

After the information bits have been processed by the encoder, the coded sequence is transmitted along the binary symmetric channel. Here it is perturbed by noise, and the receiver attempts to extract the original information sequence from the received digits.

The receiver has a mechanism similar to the encoder, so that it can generate the transmitted message corresponding to any information sequence. It is capable, in fact, of generating the treelike convolutional message set. A possible decoding scheme, therefore, would consist in the receiver's generating this set out to the length of the received sequence and picking off that path in the tree which is closest to the received sequence (in the Hamming Distance sense). The information bits corresponding to a path so selected would constitute a maximum-likelihood solution.

Unfortunately, the computational effort involved at the decoder would be excessive under this scheme. This is because the message set, when there are n information bits, contains about $n_0 2^{n+1}$ digits (there are roughly 2^{n+1} branches, each containing n_0 digits). Hence, the amount of computation required in comparing the received sequence with all the digits of this set grows exponentially with n, rendering the scheme impracticable for all save very short messages.

This exponential growth in computational effort is not present in sequential decoding schemes. Here the successive information bits are decoded one at a time, the decoding decision being based in each case upon the Nn_0 *consecutive message digits influenced by the bit in question.* In terms of the message tree, sequential decoders attempt to select a path which closely matches the received sequence, making the selection one branch at a time. In choosing a branch (i.e., in arriving at a decoding decision) the decoder makes distance* measurements between the Nn_0 message digits influenced by the bit it is decoding, and the digits in the corresponding portion of the tree.

The phrase "corresponding portion" requires some amplification. In the case of the first information bit $I(1)$, it refers to the convolutional message set, truncated at the Nth stage. Such a set is represented in Figure 4 with $n_0 = 3$, $N = 4$, and the generator sequence 111101110101.

In the case of $I(2)$, however, it has two possible configurations, one in the upper, one in the lower half of the convolutional tree. Together, these two configurations cover the region in the tree which is influenced by $I(2)$. Each has the form of a tree truncated at the Nth

---

* The word "distance" as used here refers to "Hamming Distance".

stage, and the "corresponding portion" of the convolutional tree is uniquely defined only if a value of $I(1)$ is specified.

Generalizing, the region of influence of information bit $I(i)$ can be divided up into 2^{i-1} nonoverlapping truncated trees, each of which can be specified by assigning appropriate values to $I(i-1)$, $I(i-2)$, ... In decoding $I(i)$ with the sequential schemes, these bits have already been decoded and are assigned values accordingly. Such portions of the message set are called truncated message sets. Each may be considered as having an upper subset S_0 and a lower subset S_1, the paths of S_j corresponding to information sub-sequences of length N and with first information bit equal to j ($j = 0,1$).
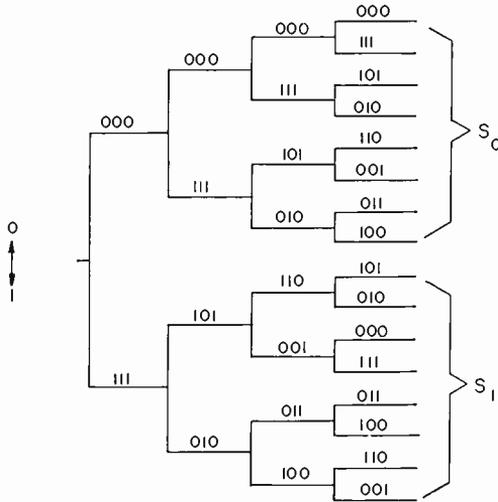
Fig. 4—A truncated message set.

THE PROPOSED SEQUENTIAL DECODING SYSTEM

Each sequential decoding decision is made on the basis of a comparison between Nn_0 digits of the received sequence and the corresponding truncated message set. In identifying this set, which will be denoted by $S (= S_0 \bigcup S_1)$, the decoder must know the true values of preceding information bits. It has already decoded these bits, and it employs the values so obtained.

If it has decoded them correctly, the Nn_0 transmitted digits from which the Nn_0 received digits originated will be situated on some path in S, called the "correct path." In the absence of noise, of course, the Nn_0 received digits will match the digits of this correct path exactly.

In the decoding scheme described in this paper, the decoder searches S for the best match to the received digits it can find, i.e., it picks out the path whose digits are least distant from the Nn_0 received digits. If this path lies in $S_j (j = 0,1)$ the decoder concludes that the correct path also lies in S_j.

Before showing how this search for the best match is accomplished, the error-correcting capability it offers is of interest. It can be shown[3] that, for any truncated message set S in the convolutional message set, the set of distances between a given path in S_0 and the paths of S_1 is identical with the set of distances between a given path in S_1 and the paths of S_0, and this set is invariant with respect to S. The smallest member of this set is denoted by D. D is a function only of the convolutional generator chosen.

It follows that the correct path in S is distant at least D from the paths of the "incorrect subset" (i.e., the subset of S in which it does not lie). Hence, if fewer than $D/2$ noise digits are present in the Nn_0 received digits, the closest path lies somewhere in the "correct subset," and correct decoding ensues.

For example, in a particular convolutional code having $N = 24$, $n_0 = 5$, the minimum distance D was found to be 40. Hence, a sufficient condition for correct decoding was that 19 or fewer noise digits occur in 120 transmitted digits.

If $D/2$ or more noise digits are present in the Nn_0 received digits, decoding may or may not be incorrect, depending on where exactly these noise digits occur.

The above analysis applies to situations where no prior decoding error has been made. Once an error is made by the decoder, subsequent decoding operations are not meaningful, since the decoder does not use the appropriate truncated message sets. It attempts, in fact, to match the received sequence with sets which do not contain the correct path. One would expect its output after this point to bear little resemblance to the correct information sequence. The appropriate truncated sets re-enter the computation only if $N - 1$ subsequent successive information bits get decoded correctly (for the influence of the incorrect bits then vanishes), an event which would seem unlikely to occur, especially if N is large.

This degeneracy in the decoding operation which sets in once an error is made is an importunate feature of all sequential decoders. The use of a feedback channel as a corrective measure has been suggested by Wozencraft and Reiffen.[3]

THE PRINCIPAL MODE OF DECODING (MODE 1)

Each truncated message set S contains roughly 2^{N+1} branches. If the decoder were to compute every digit of every path of S in looking for the best match, it would make about $n_0 2^{N+1}$ computations in all, an overwhelmingly large number for the rather long constraint lengths which it is anticipated that the coding scheme would use.

However, the decoder does not function in this way. In terms of the truncated tree, it computes the digits of each path sequentially, beginning at the left, and, by comparing them to the corresponding digits of the received sequence, it can detect a poor match without following the path out to its full length.

Specifically, the decoder counts the number of disagreements it observes between the received and the computed digits. It rejects the path from further consideration as soon as this number exceeds a given constant T, known as the threshold. The decoder makes a note of all paths which are traced to full length, and of the number of disagreements recorded in each case. The one with the fewest disagreements is the closest path to Nn_0 received digits, and the information bit is decoded accordingly.

If no path can be traced to full length, the threshold is increased by an amount t and the search re-initiated with the new threshold $T + t$. If again none of the paths can be traced to full length, the threshold is once more increased by t, and another search takes place. This cycle of searching and augmenting the threshold continues until the latter becomes large enough to permit the closest path to be traced to full length, at which point decoding occurs. If there are several closest paths, the decoder arbitrarily selects the last one it has found.

The method of search can best be described with reference to Figure 4. Until the threshold is exceeded, the decoder explores the upper branch of each successive node it encounters. Once the threshold is crossed, however, or a path is traced to full length, the decoder backtracks to the node it has just passed, and proceeds out along the lower branch stemming from this node. For example, if the first three received digits are 111 and the threshold is set at 2, the decoder generates the initial 000 in the topmost branch, then backtracks immediately to the first branch of S_1. It then proceeds along the topmost path of S_1.

In this fashion the decoder works its way down through S, tracing each path as far as the threshold will permit and noting any path which goes to full length. In the example given the decoder would examine only one branch of S_0, for the disagreement found is such

that all the paths of S_0 are discarded from further consideration. Only three computations are required to process S_0. This is a somewhat extreme illustration of the way large portions of S can be eliminated from the search when the threshold is exceeded. The smaller the threshold, the sooner it will be exceeded and the quicker will be the search of the tree (and, unfortunately, the less probably will the closest path be traced to full length).

Whatever values are assigned to T and t, the decoder finds the closest path. Hence its performance, as far as the probability of error is concerned, is independent of these parameters. They are therefore chosen in such a fashion as to optimize the computational effort required at the decoder during its search.

The Secondary Mode of Decoding (Mode 2)

In decoding the current information bit $I(i)$ as described, the decoder has also made "preliminary" estimates of the next $(N-1)$ bits; these estimates are the bits which are defined by the closest path, and they will all be correct if this path coincides with the correct path.

The decoder uses these preliminary estimates of $I(i+1)$, $I(i+2)$, ..., $I(i + N - 1)$ to speed up coding. After searching the truncated tree corresponding to $I(i)$, it switches to a secondary mode of decoding (Mode 2). It now attempts to extend the closest path deeper into the convolutional tree, and it does this very simply, computing the contents of the next two branches stemming from this path, and selecting the one which is closer to the corresponding n_0 digits of the received sequence. The result is a preliminary estimate of $I(i + N)$.

The decoder now computes the total distance between the path defined by its preliminary estimates of $I(i + 1), \ldots, I(i + N)$ and the corresponding received digits. If this distance exceeds $D/2$, the decoder reverts to Mode 1, searching the truncated message-set corresponding to $I(i + 1)$. If it does not exceed $D/2$, however, *the decoder concludes that its preliminary estimate of $I(i + 1)$ was correct*, and the decoding of $I(i + 1)$ occurs immediately. It then extends the "closest path" further, selecting the better of the next two branches and thus making a preliminary estimate of $I(i + N + 1)$. If the path defined by its estimates of $I(i + 2), \ldots, I(i + N + 1)$ is more than $D/2$ distant from the received sequence, the decoder reverts to Mode 1. If not, $I(i + 2)$ is decoded immediately, and the "closest path" is again extended. In this way successive information bits can be decoded rapidly.

In this secondary mode of decoding, the computational effort is

very small. Only two branches of the tree are generated in arriving at a decoding decision. Furthermore, the probability of error is no greater than that in Mode 1: an error can only occur if some path in the incorrect subset of S is distant $D/2$ or less from the received sequence, which implies that all the paths in the correct subset are distant $D/2$ or more from the received sequence. It follows that if an error occurs in the secondary decoding mode, so also would it have occurred in Mode 1. The savings in computational effort realized by Mode 2 are often considerable. For example, in a simulation study with $n_0 = 3$, $N = 24$, and $p_0 = .02$, the decoder's Mode 1 required an average of about 300 computations per information bit, yet the over-all average was less than 7. Nearly all bits were decoded in the secondary mode, at the cost of 6 computations each. Whenever a truncated message set was searched, the closet path found coincided in all branches with the correct path, and prolonged operation in Mode 2 subsequently took place. The principal decoding mode was only re-entered when two or three noise digits occurred on the same branch of the tree (an event which leads to making an incorrect preliminary estimate of the corresponding information bit, and a resulting build-up of observed differences until $D/2$ is exceeded).

THE NUMBER OF DECODING COMPUTATIONS

A decoding computation consists of the generation of a digit at the decoder and its comparison with the corresponding received digit. The number of computations taken in decoding governs the speed at which the system can operate, and is therefore of prime importance.

In Mode 2 there are always $2n_0$ computations involved in decoding an information bit. In Mode 1 the number of computations in a function of the noise and is a random variable. A sufficient condition for termination of the search of S is that the threshold exceeds the number of noise digits present. In general, the presence of a large number of noise digits forces the decoder to employ a large threshold. The behavior of the number of computations as a function of the threshold is therefore of interest.

It has been possible to analyze this behavior for the so-called "average generator." The average generator is one whose Nn_0 digits are selected independently and at random. The number of computations for this generator equals the mean number of computations averaged over the ensemble of all possible generators of length Nn_0.

As previously explained, S consists of a correct subset and an incorrect subset. The average number of computations $\overline{M}(K)$ taken in searching the incorrect subset with a threshold $K \geqq 0$ can be shown

(see Appendix) to satisfy the difference relationship:

$$\overline{M}(K)\left[1-(1/2)^{n_0-1}\right] \simeq n_0 + (1/2)^{n_0-1}\left[\binom{n_0}{1}\overline{M}(K-1) + \binom{n_0}{2}\overline{M}(K-2) + \ldots + \overline{M}(K-n_0)\right]$$

provided $n_0 > 1$. Thus, if $n_0 = 2$,

$$\overline{M}(K) \simeq 4 + 2\overline{M}(K-1) + \overline{M}(K-2).$$

The average number of computations $\overline{N}(K)$ taken in searching both subsets is given by

$$\overline{N}(K)\left[1-(1/2)^{n_0-1}\right] \simeq F + (1/2)^{n_0-1}\left[\binom{n_0}{1}\overline{N}(K-1) + \binom{n_0}{2}\overline{N}(K-2) + \ldots + \overline{N}(K-n_0)\right],$$

where F is a well-defined function of the pattern of noise digits.

The method of z-transforms has been used to solve the difference relations for some particular values of n_0. For $n_0 = 2$,

$$\overline{M}(K) \simeq 5.84 \ (2.42)^K.$$

Other rates of transmission also yield exponential expressions for Mode 1.

Tests were run to simulate Mode I operating in the presence of various noise patterns for several convolutional generators. The number of computations was found to depend strongly upon the generator used, but in every case the *rate* at which computations increased with increasing threshold conformed very closely indeed to that predicted for the average generator. And the influence of the actual noise pattern upon this rate was nugatory.

The number of computations taken in searching S in Mode 1, then, increases roughly exponentially with K, the threshold, and the number of computations taken in Mode 2 equals $2n_0$ per bit decoded.

The Average Probability of Error

This is defined to be the probability of decoding error associated with the "average" convolutional generator, and it equals the mean probability of error, averaged over the ensemble of all generators.

Suppose that exactly K noise digits are present in the Nn_0 received digits. A necessary condition for a decoding error is that at least

one path in the incorrect subset is at distance K or less from the received sequence.

The distance of a digit in the incorrect subset from the corresponding received digit is a random variable, and for the average generator it can be shown[2,3] to assume the values zero and one with equal probability. For any path in the incorrect subset, the probability that its distance from the received sequence is K or less is therefore

$$2^{-Nn_0} \sum_{j=0}^{K} \binom{Nn_0}{j}.$$

There are 2^{N-1} paths in the incorrect subset. Hence the probability that the necessary condition for a decoding error will be met is bounded above by

$$2^{N-1}\, 2^{-Nn_0} \sum_{j=0}^{K} \binom{Nn_0}{j}.$$

The total average probability of error is bounded by the expected value of this sum, where K is now treated as the random variable. The significance of this result is that the bound obtained coincides with the upper bound to the average probability of error for block codes of the same constraint-length.[3] It decreases exponentially with increasing N, and can therefore be made arbitrarily small by the choice of a sufficiently large N.

In 1957 J. Wozencraft published a paper containing a description and analysis of convolutional encoding and of a decoding algorithm.[2] The recently published book by J. Wozencraft and B. Reiffen[3] includes a somewhat more refined analysis of the algorithm. It also suggests certain modifications which should improve its performance, while preserving the same basic features.

Perhaps the most salient of these features is the method of searching the truncated message set, S. The threshold, during each search, is allowed to increase in a certain fashion as the decoder penetrates deeper into S. Thus the search is conducted using a sequence of thresholds, each being a function of the depth of penetration. If no path can be traced to full length, the search is resumed with a new sequence of increased thresholds. This process continues until some path is traced to full length, at which point it is concluded that this path lies in the correct subset and decoding occurs forthwith.

There is no guarantee that this path lies in the same subset as the closest path, and the probability of error for this algorithm exceeds

that realized by the scheme described here. In fact, the upper bound obtained[3] for the average probability of error is

$$AN\overline{P}(B),$$

where A is a constant, and $\overline{P}(B)$ is the upper bound for the comparable average block code. As has been stated, the bound for this decoder coincides with $\overline{P}(B)$, which decreases faster with increasing N than does $AN\overline{P}(B)$. However, $\overline{P}(B)$ decays exponentially with increasing N and both bounds therefore share this exponential decay.

The decoder described by Wozencraft was designed to achieve, in addition, a nonexponential increase in decoding effort with increasing N. It has been proved[3] that, under certain restrictions, the average number of computations taken in searching the incorrect subset is bounded above by a quantity which increases more slowly than N. It has not been possible[3] to bound the number of computations taken in searching the correct subset, but heuristic considerations lend support to the conjecture that this number behaves similarly. In other words, Wozencraft's algorithm is conjectured to permit an exponential decay in error probability at the cost of only an algebraic increase in computational effort.

It has been indicated that, in searching S, the number of computations made by the decoder discussed in this paper increases roughly exponentially with the threshold employed. For fixed p_0, the computational effort can be expected to grow exponentially with increasing N. Thus, the effect of making the constraint length larger is to place an exponentially increasing computational burden on the receiver, whereas this increase is conjectured to be only algebraic in the decoder described by Wozencraft. On the other hand, the latter requires a larger value of N to achieve the same probability of error.

SIMULATION

The operation of the decoder and a binary symmetric channel have been simulated on a general-purpose computer. Constraint lengths of 24, 72, and 144 were used in conjunction with a rate of transmission of 1/3, and a constraint length of 120 with a rate of 1/5. The transition probability p_0 was assigned various values between .02 and .12 in each case.

For every set of p_0, R, and N, the number of bits decoded was at least 1000 and often much higher. A total of about 500,000 bits were decoded in all. The average number $\overline{C}$ of computations taken in decoding is given in Table I.

The Generators

A factor which strongly influenced $\overline{C}$ in these experiments was the choice of the convolutional generator. Operation in Mode 2 involves only $2n_0$ computations per decoded bit, and reduces $\overline{C}$ substantially. In the absence of decoding errors, the operation reverts from Mode 2 to Mode 1 when one of two things happens:

(a) More than $\frac{1}{2}n_0$ noise digits are present in a group of n_0 received digits (i.e., a preliminary estimate is incorrect);

(b) More than $D/2$ noise digits are present in a group of Nn_0 consecutive received digits.

Table I—Average Number of Computations

R	Nn_0	p_0	$\overline{C}$	R	Nn_0	p_0	$\overline{C}$
1/3	24	.02	6.1	1/3	144	.04	69.1
		.04	6.5			.06	3200
		.06	8.5	1/5	120	.02	10.0
1/3	72	.02	6.4			.04	10.3
		.04	12.8			.06	13.1
		.06	84.0			.08	41.1
		.08	363			.10	242
1/3	144	.02	8.8			.12	1100
		.03	27.2				

The probability of (b) happening can be reduced by increasing the minimum distance D associated with the generator. A generator was found for which D was 20 in the Code $R = 1/3$, $Nn_0 = 72$. Thus $D/2$ was considerably larger than the expected number of noise digits occurring in 72 transmitted digits for most values of p_0. For the code $R = 1/3$, $Nn_0 = 24$, the minimum distance was 10 (which cannot be exceeded for this code).

For the code $R = 1/3$, $Nn_0 = 144$, the minimum distance was 30. At values of p_0 below .06, Mode 1 was only entered under the circumstances (a) above. At $p_0 = .06$, however, a total of more than 15

noise digits was present in 144 received digits over a span of about 2 per cent of the message, and operation in Mode 1 occurred under circumstances (b), with deleterious consequences to the average number of computations.

Thresholds

The choice of T and t also has an influence on the number of computations. The selection of the thresholds in the decoder described by Wozencraft[3] is based upon the channel transition probability, which is assumed to be known beforehand by the receiver. The decoder described here assumes no prior knowledge of p_0 and the selection of T on entering Mode 1 is made on an empirical basis. If the closest path in the previous Mode 1 was distant d from the received sequence, then T is assigned the value $d + 1$ for the current Mode 1, which is a crude estimate of the number of noise digits present. Obviously, more sophisticated estimates could be made, but this one has the virtue of great simplicity, and it is also quite accurate when the Mode 1's are consecutive. Minimizing computational effort is of particular concern when this is the case, for the incidence of consecutive Mode 1's results in peaks in computation.

In all the simulation studies, t was assigned the value of 3. For the case $R = 1/5$ this choice was arbitrary. An analysis of some additional experimental studies of Mode 1 (already mentioned in connection with the number of computations) has revealed that, for the case $R = 1/3$, the choice seems reasonable. In these experiments T was set initially at 0, and for each of several noise-patterns five cases were run, with $t = 1, 2, 3, 4$, and 5. The fewest computations occurred at $t = 2$ and $t = 3$.

The computer is presently being programmed to simulate decoding schemes employing other threshold settings.

Decoding Errors

No errors in decoding occurred for the code $N = 24$, $R = 1/5$. At the highest noise-level used ($p_0 = 0.12$) 10,000 bits were all decoded correctly. No errors in decoding occurred for the code $N = 48$, $R = 1/3$. At the highest noise level used ($p_0 = .06$), 1,200 bits were all decoded correctly.

Decoding errors for the code $N = 24$, $R = 1/3$, occurred at $p_0 = .08$. Four different runs were made for this noise-level, and a total of 44,000 bits were decoded. In two of these runs the decoder made an error. As had been anticipated, the occurrence of the error was accompanied by a substantial increase in the number of computa-

tions, and led to a degenerate condition in which the decoder continued making errors. This condition did not perpetuate itself indefinitely, however. Eventually the decoder succeeded in decoding 23 consecutive bits correctly, thereby eliminating the effect of the error and restoring meaningful decoding. This phenomenon was totally unexpected.

Unfortunately, the elimination of the degeneracy seems too slow to be useful. The degenerate condition persisted over 30 and 200 decodings in the two cases simulated, and each of these decodings required between 9,000 and 66,000 computations, implying a computational load which no real-time decoder could be expected to handle.

This was not so for the code $N = 8$, $R = 1/3$. Here each error was perpetuated over a span of 5 bits at most. Furthermore, almost all incorrect decoding took place in Mode 2, hence no build-up in computational effort was observed. At $p_0 = .04$ degeneracy occurred on 7 occasions in 80,000 decodings, and 26 bits were decoded erroneously in all. At $p_0 = .06$, 35 errors were made in decoding 13,000 bits. No errors occurred at lower noise levels.

DISCUSSION

In a one-way communications system, the message digits would supposedly arrive at the receiver at a constant rate. The rate at which they are decoded, however, varies considerably in the sequential decoding scheme, because of the computational peaks associated with Mode 1. To compensate for this uneven decoding rate, the receiver may be supplied with a buffer where incoming digits can be stored until the decoder is ready to use them. The question arises: how much buffer storage is needed to handle the worst computational peaks likely to occur?

Clearly, the answer depends partly on the relation between the rate at which message digits arrive, and the computational rate of the decoder. This relation should be such that the statistics of the queue in the buffer are stationary in some sense, i.e., there is no long-term build-up. Even so, the possibility of overflow would not be eliminated; whatever the size of the buffer storage, patterns of noise can occur which produce overflow (unless transmission is very slow indeed compared to the decoder's operating speed). For any positive value of p_0 such patterns have a finite, though perhaps very small, probability of occurring. The question of how much storage is required must, therefore, be answered in probabalistic terms. Because the numbers of computations taken in decoding consecutive informa-

tion bits are not mutually independent, the problem of analyzing the behavior of the queue analytically becomes formidable, and the only feasible approach seems to be experimental.

In particular, the method of setting the threshold, T, and its increment, t, will probably be modified. The code $R = 1/3$, $Nn_0 = 72$ has been re-run with $t = 1$, and with T set equal to the distance d of the closest path from the received sequence in the previous Mode 1. At $p_0 = .02$ these modifications made no appreciable difference. At higher noise levels, a reduction in both the peak and the averages $\bar{C}$ of computation was observed. At $p_0 = .06$ and $.08$ the average $\bar{C}$ was reduced by about 30 per cent. It is anticipated that similar modifications can be profitably made to the other codes.

APPENDIX—AVERAGE NUMBER OF COMPUTATIONS FOR MODE 1

Let $M(K)$ denote the number of computations taken in searching the incorrect sub set S_1 with a threshold K, for a given, fixed noise pattern.

The decoder begins its search of S_1 by comparing the first branch with the first n_0 received digits. If more than K disagreements are noted, the search stops. If K or fewer disagreements are noted, the search is continued beyond the first branch. Now two subsets of S_1 stem from this first branch, an upper subset S_{10} and a lower subset S_{11}, and the threshold with which they are searched is K-D, where D denotes the number of disagreements observed at the first branch. Let $M_0(K)$ and $M_1(K)$ denote the number of computations in these subsets when they are searched with threshold K. It is convenient to define

$$M(K) = M_0(K) = M_1(K) = 0, \qquad \text{for } K < 0.$$

If $K \geqq 0$, the expected number of computations taken to search S_1 is

$$n_0 + [M_0(K) + M_1(K)] \text{ Prob } \{D = 0\}$$
$$+ [M_0(K-1) + M_1(K-1)] \text{ Prob } \{D = 1\}$$
$$+ \cdots$$
$$\vdots$$
$$+ [M_0(K-n_0) + M_1(K-n_0)] \text{ Prob } \{D = n_0\}.$$

Actually, this is the expected number of computations averaged over the first n_0 generator digits. If these are chosen at random, the distance of each from the corresponding received digit is a random variable which assumes the values zero and one with equal probability.[2,3]

It follows that

$$\text{Prob } \{D = 0\} = (\tfrac{1}{2})^{n_0}$$

$$\text{Prob } \{D = 1\} = (\tfrac{1}{2})^{n_0} \begin{pmatrix} n_0 \\ 1 \end{pmatrix}$$

$$\vdots$$

$$\text{Prob } \{D = n_0\} = (\tfrac{1}{2})^{n_0}.$$

Substituting these values into the expression, and averaging over all generator digits, it is found that, for $K \geqq 0$,

$$\overline{M}(K) = n_0 + (\tfrac{1}{2})^{n_0} [\overline{M}_0(K) + \overline{M}_1(K) + \begin{pmatrix} n_0 \\ 1 \end{pmatrix} \overline{M}_0(K-1)$$
$$+ \begin{pmatrix} n_0 \\ 1 \end{pmatrix} \overline{M}_1(K-1) + \cdots + \overline{M}_0(K-n_0) + \overline{M}_1(K-n_0)]. \qquad (1)$$

The next step is to show that

$$\overline{M}_0(K) = \overline{M}_1(K) \simeq \overline{M}(K), \qquad \text{for any } K.$$

To do this, the decoding process must first be considered in some detail. The number of branches in S_1 at stages $1, 2, \ldots, n$ is $1, 2, \ldots, 2^{n-1}$. Each branch, at stage n, belongs to a path of which the ith digit is distant d_i from the corresponding received digit, and the probability that this path survives is

$$\text{Prob} \left\{ \sum_{i=1}^{nn_0} d_i \leqq K \right\}$$
$$= P(nn_0 | K).$$

This expression (for the "average" generator) is independent of the noise pattern and of the particular path under consideration, be-

cause each d_i assumes the values zero and one with equal probability.

It follows that the expected number of paths in S_1 surviving at stage n is

$$2^{n-1} P(nn_0 | K).$$

In proceeding from stage n to stage $n + 1$, the decoder performs $2n_0$ computations for every path which has survived. Therefore, the expected number of computations taken in so doing is

$$2n_0 \, 2^{n-1} P(nn_0 | K).$$

Summing the expression for all stages in S_1,

$$\overline{M}(K) = n_0 + n_0 \sum_{n=1}^{N-1} 2^n P(nn_0 | K) \qquad K \geqq 0. \quad (2)$$

This is an explicit solution to the problem. It follows, for instance, that in the case $n_0 = 3$, $K = 0$,

$$\overline{M}(0) = 3 + 3 \sum_{n=1}^{N-1} 2^{-2n} \simeq 4.$$

However, the evaluation of $\overline{M}$ for higher values of K becomes unattractively complicated. For this reason, the difference-equation approach was favored.

The subsets S_{10} and S_{11} share with S_1 all the statistical properties which were used in the evaluation of $\overline{M}(K)$. The arguments may be repeated with S_{10} (or S_{11}) replacing S_1.

$$\overline{M}_1(K) = \overline{M}_0(K) = n_0 + n_0 \sum_{n=1}^{N-2} 2^n P(nn_0 | K), \qquad K \geqq 0.$$

The sum here has one term fewer than the sum in Equation (2). The term is $2^{N-1} P(Nn_0 - n_0 | K)$ which can be estimated using Chernoff Bounds[3] and, if $n_0 > 1$, is negligibly small for practical codes. For example, if $N = 24$, $n_0 = 3$, and $K = 10$, it is less than 2^{-18}.

Therefore

$$\overline{M}_0(K) = \overline{M}_1(K) \simeq \overline{M}(K).$$

Substituting this result in Equation (1) leads to the final result:

$$\overline{M}(K) \; [1 - (\tfrac{1}{2})^{n_0-1}] \simeq n_0 + (\tfrac{1}{2})^{n_0-1} \left[\binom{n_0}{1} \; \overline{M}(K-1) \right.$$

$$\left. + \binom{n_0}{2} \overline{M}(K-2) + \cdots + \overline{M}(K-n_0) \right], \qquad K \geqq 0, \; n_0 \geqq 2. \; (3)$$

This approach has been adopted in evaluating the total average number of computations $\overline{N}(K)$ taken to search the entire tree (correct and incorrect subsets). The analysis here was complicated by the fact that it was necessary to take into account the noise pattern. A function $J(i)$ was introduced, which represents the number of the stage at which the accumulated noise digits exceed i for the first time. The following was found to be valid:

$$\overline{N}(K) \; [1 - (\tfrac{1}{2})^{n_0-1}] \simeq F + (\tfrac{1}{2})^{n_0-1} \left[\binom{n_0}{1} \overline{N}(K-1) \right.$$

$$\left. + \binom{n_0}{2} \overline{N}(K-2) + \cdots + \overline{N}(K-n_0) \right]$$

where

$$F = 2n_0 J(K) - n_0 (\tfrac{1}{2})^{n_0-1} \left[J(K) + \binom{n_0}{1} J(K-1) + \cdots + J(K-n_0) \right]$$

and $K \geqq 0, \; n_0 \geqq 2$.

A RADIAL-WAVEGUIDE ANTENNA AND MULTIPLE AMPLIFIER SYSTEM FOR ELECTRONIC SCANNING*

BY

C. P. CLASEN, J. B. RANKIN, AND O. M. WOODWARD, JR.

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

Summary—A new antenna system has been developed whose beam can be readily steered electronically through 360° in azimuth. The antenna is a radial waveguide containing radiators arranged in concentric rings. Each ring of radiators is fed by two power amplifiers. One or more independent beams are formed by appropriate adjustment of the amplitudes and phases of the amplifiers. Electronic beam scanning is accomplished by variation of the relative amplifier phases. The phase variation is obtained by a heterodyne delay line network.

A low-power feasibility test model of the antenna, power output amplifiers, and phasing system has been built for UHF; the model gives results which support the theory.

THEORY AND CONSTRUCTION OF RADIAL WAVEGUIDE ANTENNA

THERE has been increasing interest in recent years in ring arrays—that is, a ring or concentric rings of radiators.[1-5] The radial-waveguide antenna belongs to this group. A theoretical concept which is useful in explaining the operation of the antenna is shown in Figure 1. The radial waveguide consists of two parallel circular disks spaced less than a half wave apart so that only the TEM and TE_{no} modes can propagate. The circumferential gap between the disks may be flared as shown into a section of biconical horn to sharpen the elevation pattern and improve the impedance match to free space.

---

[1] H. Chireix, "Antennes a Rayonnement Zenithal Reduit," *L'Onde Electrique*, Vol. 15, p. 440, 1936.

[2] H. L. Knudsen, "The Field Radiated by a Ring Quasi-Array of an Infinite Number of Tangential or Radial Dipoles," *Proc. I.R.E.*, Vol. 41, No. 6, p. 781, June 1953.

[3] H. L. Knudsen, "Radiation from Ring Quasi-Arrays," *Trans. I.R.E. PGAP*, Vol. 4, No. 3, p. 452, July 1956.

[4] A. C. Schell and E. L. Bouché, "A Concentric Ring Array," *I.R.E. WESCON Convention Record*, Part I, p. 212, 1958.

[5] H. P. Neff and J. D. Tillman, "An Electronically Scanned Circular Antenna Array," *I.R.E. International Convention Record*, Part I, p. 41, 1960.

The principle of operation of the antenna may be studied by considering a central vertical radiator surrounded by a number of idealized concentric ring radiators located between the disks. Each of these rings, or mode exciters, is a circular fence of a large number of vertical radiators carrying currents of equal magnitude and progressive phase within each ring. Each ring has an integral number, m, of cycles of phase progression in going around the ring. Successive rings have an additional cycle; thus, the first ring has one cycle, the second ring has two cycles, etc. The sense of the phase progression, clockwise or counterclockwise, is indicated by the sign of m. Each ring of radi-
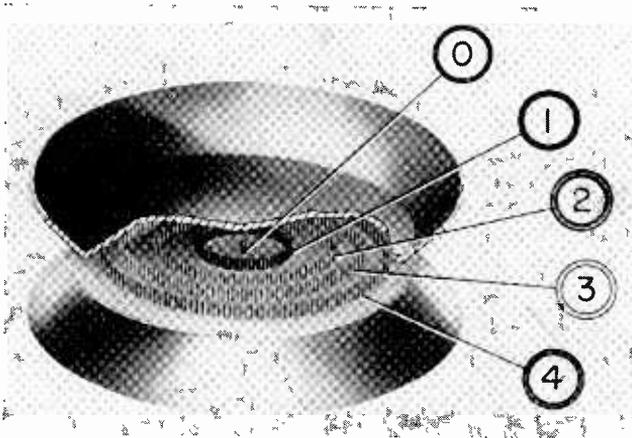
Fig. 1—Idealized radial-waveguide antenna.

ators excites a waveguide mode which propagates radially outward to the circumference and radiates. It will be shown that each ring can be fed so as to excite both the clockwise and counterclockwise modes simultaneously. The identifying symbol for a ring is n, $(n = |m|)$.

From the circular symmetry of the geometry, it is apparent that a one-to-one correspondence exists between the individual mode excitations and their respective far-field patterns in the azimuth plane. Thus, each mode produces a constant-magnitude, progressive phase field pattern. The single central element excites a TEM mode which produces an omnidirectional, constant-phase pattern.

An important feature of this system is that theoretically there is no cross coupling between any of the mode input ports because of the mathematical orthogonality of the mode configurations. Due to this zero mutual coupling and the circular symmetry of the geometry, the

gain and shape of any pattern are independent of the azimuth scan angle. This also means that the loading presented to the amplifier outputs is constant.

Thus far, theoretical rings consisting of an infinite number of elements have been considered. Figure 2, drawn for the three-cycle ring, illustrates how such an idealized ring may be very closely approximated in practice with a finite number of elements. Each ring contains $4n$ equally spaced radiators which are driven in progressive phase quadrature. If the radiators are not too far apart, i.e., not separated by much more than 1/4 wavelength on the arc, the radiation

SIMULATED RING OF 4n
ELEMENTS IN PROGRESSIVE
QUADRATURE PHASING.

TWO SUPERIMPOSED SUB-RINGS
OF 2n ELEMENTS EACH.

RING 3

L   L + λ/4

HYBRID

(+ -)   (+ +)

DECOUPLED INPUT FOR PRODUCING
COUNTER-ROTATING PROGRESSIVELY
PHASED EXCITATIONS.
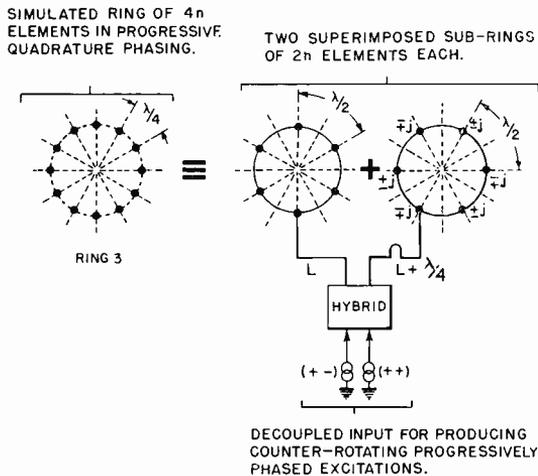
Fig. 2—Separation of ring into subrings and mode-excitation circuit.

from such a ring will closely approximate that from the idealized ring. Appreciably greater spacing will cause higher-order modes to be generated.

This arrangement may be considered as made of two superimposed subrings of $2n$ elements each, as shown in Figure 2. The elements of each subring are energized with alternating polarity at half-wave intervals from a circular transmission line. There will be no cross coupling between the two subrings since the elements of one subring lie in the neutral planes of the other subring. In Figure 2 a "rat-race" type of hybrid is joined to the subrings by two transmission lines differing in length by 1/4 wavelength to provide the required quadrature phasing. A generator or amplifier at the $++$ terminal of the hybrid will excite the progressive phase mode $m = n$ in the radial

waveguide; a generator at the $+-$ terminal will excite mode $m = -n$. Since the modes are decoupled, the $++$ and $+-$ terminals are also decoupled, and the generators can run simultaneously. The ring-feed circuitry was used in the feasibility model of the antenna. However, other types of feed arrangements, such as a corporate network and wide-band hybrids, could be employed to give the required symmetrical element feeding over a greater frequency band.

The details of the constructed ring are pictured in Figure 3. For the sake of clarity, only a single ring (ring 3) is shown. The radiators of one subring are fed from above the upper plate of the radial waveguide and those of the other subring are fed from below the lower
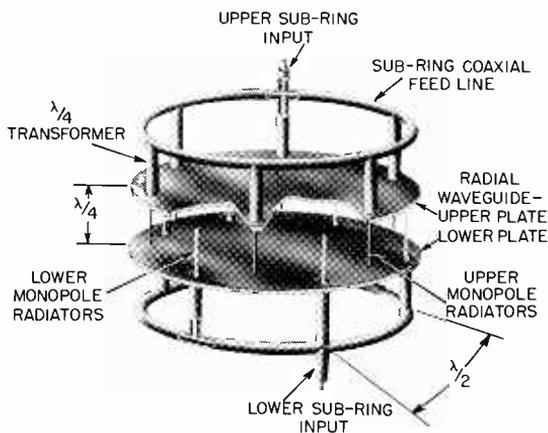
Fig. 3—Construction details for ring 3.

plate. These monopole radiators are joined at half-wave intervals to the subring coaxial feed line by means of quarter-wave transformers, which provide an additional means for matching the monopole impedances to the subring input. A more important function has to do with mutual coupling between monopoles. The modes initiated by any ring propagate through all rings of larger diameter out to the circumference and then radiate. There is a tendency for the monopoles of these outer rings to distort the modes passing by them. The quarter-wave transformers in conjunction with the subring feed lines adequately limit the distortion. The explanation is fairly lengthy and will be omitted.

Figure 4 is a photograph of an antenna, having a central radiator and four concentric rings, which was constructed to operate at a fre-

quency of 765 megacycles. This frequency was arbitrarily chosen in the UHF band to obtain a model with components that could be handled easily. A protective covering around the circumferential opening of the radial waveguide hides the monopole elements. The cables from the upper subring inputs were extended to the circuitry below through a metal tube joining the radial-waveguide plates together at their center. The TEM mode excitation was obtained by feeding the outer

Fig. 4—Radial-waveguide antenna and electronic scanning circuitry.

surface of this tube using an extended 1/4-wavelength coaxial sleeve. Some of the electronic scanning circuitry and measuring equipment is shown in the rack below the antenna. A plan view showing the geometric layout of the monopole elements in the radial waveguide is shown in Figure 5.

Cross-coupling measurements were made between the various inputs to the radial-waveguide antenna. The feed-through from ring one to ring three was down 27 decibels and the feed-through for any other combination of inputs was down between 30 and 40 decibels.

⊗ = UPPER MONOPOLES
⊙ = LOWER MONOPOLES
⊖ & ⊕ = RING FEED TERMINALS
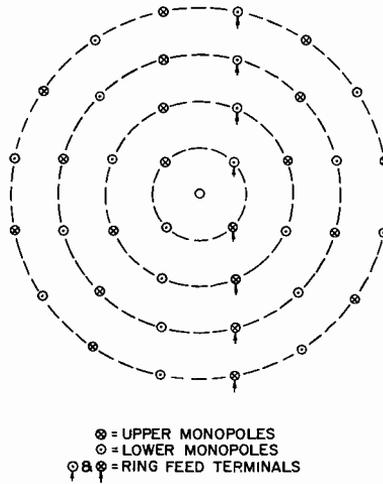
Fig. 5—Plan view of monopole layout.

An alternate method of mode excitation which may have advantages in certain higher-frequency applications is shown in Figure 6. Slots are used, instead of monopoles, to couple to the radial waveguide. The arrangement here is a partially assembled S-band model having four annular, concentric slots. The method of excitation for one subring of ring two is shown here. Coaxial lines from a corporate feed network join the base of the slot at equally spaced intervals. The axial extension of each coaxial inner conductor to the inner or outer lips of the annular slot performs as a balun to produce the appropriate excitation.
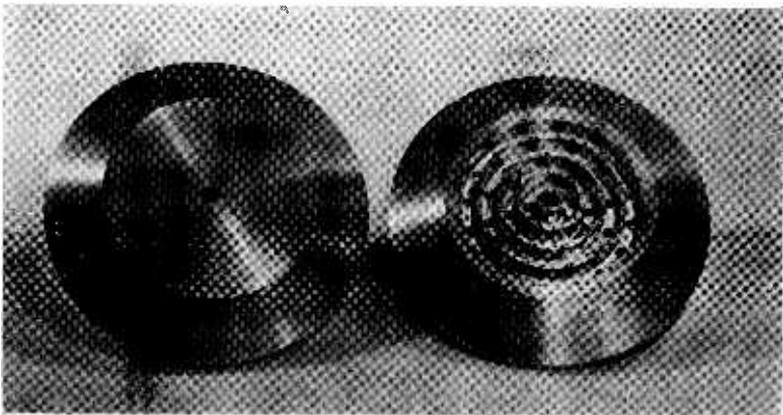
Fig. 6—Annular-slot radial-waveguide antenna for S-band.

Thus far, an antenna has been described which produces radiation components whose azimuth patterns are constant amplitude and progressive phase. Each component is excited by its own generator. There is negligible coupling between generators. This results in a versatile antenna array capable of synthesizing many types of beam shapes, both single and multiple beams, in the azimuth plane. For producing a single maximum-gain beam, the amplitudes and phases of all the mode generators are adjusted to give equal and in-phase field intensities in the desired direction. To change the beam direction, the relative phases of the amplifiers are changed with a linear phase taper.

There is an analogy between pattern synthesis with the radial-waveguide antenna and with a linear array having an odd number of equispaced elements. The radiation from each mode of the radial waveguide corresponds to the radiation from each element of a linear array. The azimuth pattern from each radial-waveguide mode can be written $A_m e^{jm\phi}$, where

$A_m =$ amplitude and omnidirectional part of the phase at a fixed distance from the array center,

$\phi =$ azimuth angle,

$m =$ the integral number of cycles in going around the array. It will be recalled that both clockwise and counterclockwise phase progressions are possible. These are expressed mathematically by positive and negative values of m.

The radiation component from each element of a linear array can be written[6] $A_q e^{jq\psi}$, where

$$\psi = \frac{2\pi d}{\lambda} \cos \phi,$$

$\lambda =$ wavelength,

$d =$ spacing between elements,

$\phi =$ azimuth angle with respect to the line of the array,

$q =$ number of element spacings from the phase reference. q is positive if the element is in the direction $\phi = 0$, negative if in the direction $\phi = 180°$.

The zero-cycle mode corresponds to the radiation from the center element of a linear array, and the two modes produced by the largest ring correspond to the radiation from the two end elements of the linear array. Thus, a large body of linear-array theory may be applied to pattern synthesis for the radial-waveguide antenna, but with the

---

[6] S. A. Schelkunoff, "A Mathematical Theory of Linear Arrays," *Bell Syst. Tech. Jour.*, Vol. 22, No. 1, p. 80, January 1943.

simplification that the variable part of the phase terms is proportional to the azimuth angle rather than a function of both the azimuth angle and the element spacing. Since the radiation phase of each mode is not a function of the element spacing, the beam direction is a function of the excitation phase alone and not of time delay.

The single-beam, maximum-gain azimuth pattern (in the $\phi = 0$ direction) is

$$E_z(\phi) = E_{\max} \frac{\sin \dfrac{M\phi}{2}}{M \sin \dfrac{\phi}{2}} \quad,$$

where M is the number of modes and equals $2N + 1$, and N is the number of cycles in the largest ring. The half-power beam width is approximately $320°/(2N + 1)$ when N is large. The beam width is approximately the same as that of a broadside linear array whose length is the same as the diameter of the largest ring.

A simple means exists for producing monopulse sum and difference patterns. The sum pattern is produced by all the modes adding in phase in the desired azimuthal direction. If the zero mode is not excited, and if the phase of modes with clockwise phase progression (or counterclockwise) is changed by 180°, a difference pattern is obtained. The null of the difference pattern will remain centered on the maximum of the sum pattern as the beams scan. Phasor diagrams and monopulse patterns are shown in Figure 7 for the case where the sum pattern points toward the 0° azimuth direction.

ELECTRONIC SCANNING CIRCUITRY

Transmitting circuitry was designed and built in order to demonstrate the feasibility of the radial-waveguide antenna and electronic scanning circuitry. Monopulse capability was demonstrated by transmitted patterns. A simplified block diagram is shown in Figure 8. The five amplifiers on the right excite the zero mode and the four modes with clockwise phase progression. These are phased to produce a beam in a desired direction. The phasing is done by the power divider and phasing network. The beam direction is controlled by the beam steering control. There is a similar network for the beam formed by the modes with counterclockwise phase progression. Notice that there are two generators for the zero mode. These are both connected to the zero-mode exciter and are run at half power. A composite pattern from the two beams (one with each sense of phase progres-
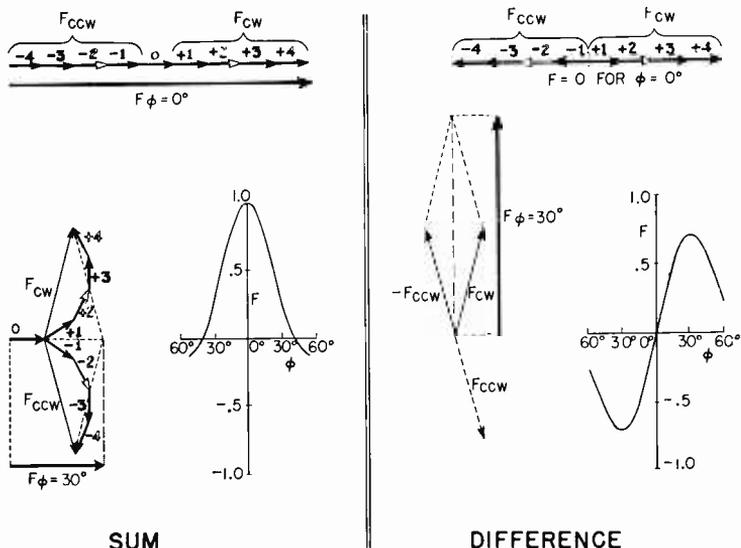
Fig. 7—Phasor diagrams for monopulse sum and difference patterns.

sion) is produced by feeding both beams through one of the hybrid terminals. Feeding the sum terminal of the hybrid feeds the power dividers for the two beams in phase. If the two beams are pointing in the same direction, and if the two zero-mode amplifiers are in phase, a sum pattern will be produced. In other words, a linear phase taper for the nine modes is obtained. The beam is steered by changing the phase taper. If the difference terminal is fed instead, the power divid-
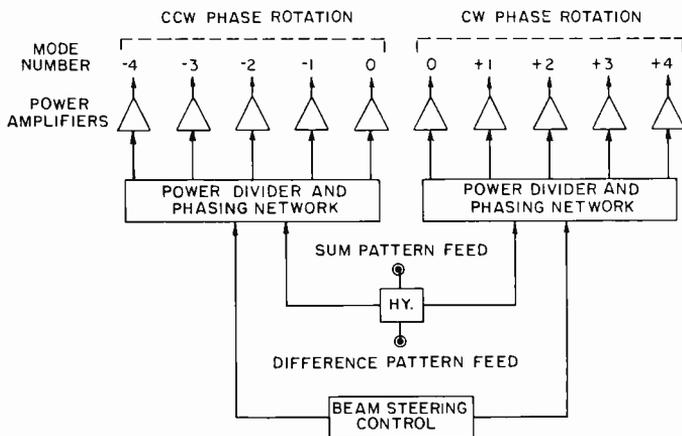
Fig. 8—Block diagram of monopulse circuitry.

ers for the two beams will be fed out of phase, and a monopulse difference pattern will result. In this case, the zero mode is not excited.

Dual beam capability might be provided by feeding the two power dividers with separate beam-steering signals. If the r-f signals are noncoherent, or if the system is used for receiving, the two beams will be completely independent.

Since the natural steering mechanism of the radial-waveguide antenna is phasing rather than delaying, the heterodyne steering system is usable in broad-band applications. Relatively conventional heterodyne phase shifters were constructed for the feasibility test model; these are in the blocks labeled "power divider and phasing network." The networks enable the relative phases of the r-f signals to
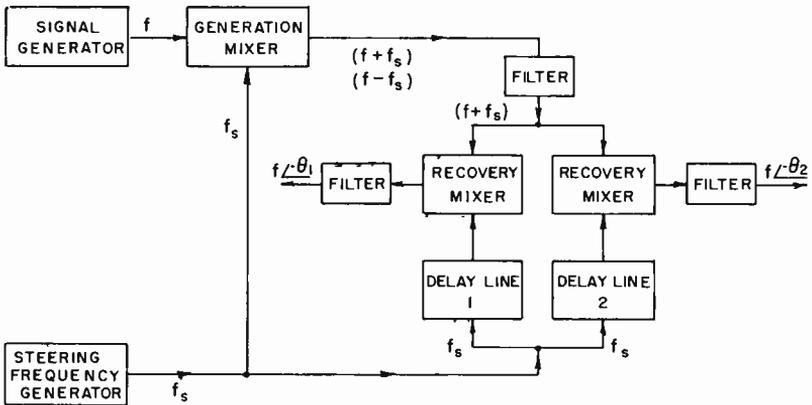
Fig. 9—Block diagram of heterodyne phase shifter.

the power amplifiers to be controlled by varying the frequency of the beam-steering-control generator. A block diagram for controlling the difference in phase of two amplifiers is shown in Figure 9.

Corresponding sum and difference patterns from the antenna, amplifier system, and phasing network are shown in Figures 10-12 for three scan angles: 0°, 90°, and 180°. The beams were steered by varying the frequency of the beam-steering-control generator. The patterns are for equal-amplitude mode excitation. Theoretical patterns are shown in Figure 13. As with a linear array, the side-lobe level can be reduced by grading the amplitudes of the amplifiers.

CONCLUSION

A new antenna system has been developed for electronically steering a directional beam through 360° in the azimuth plane. The feasi-
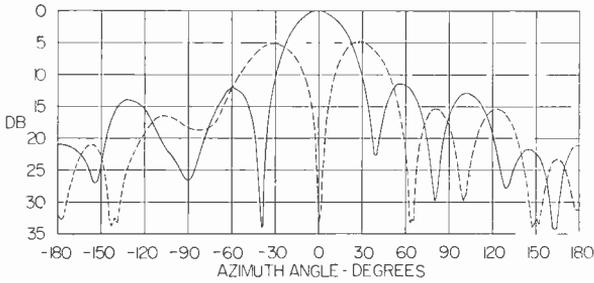
Fig. 10—Measured monopulse patterns for uniform illumination at 0° scan angle.
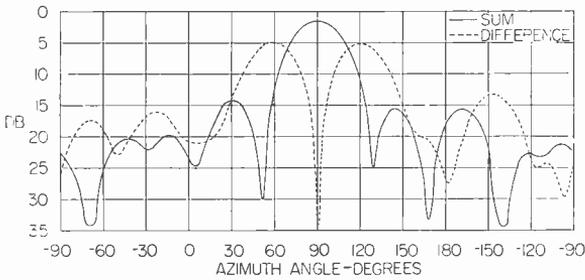
Fig. 11—Measured sum and difference patterns for uniform illumination at 90° scan angle.
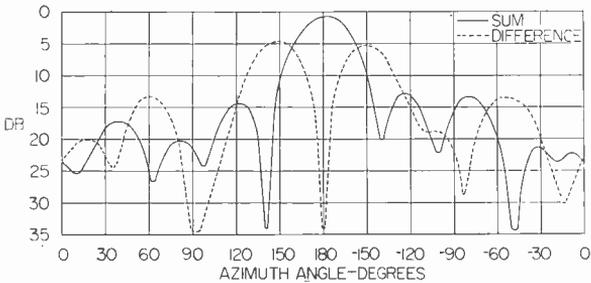
Fig. 12—Measured sum and difference patterns for uniform illumination at 180° scan angle.
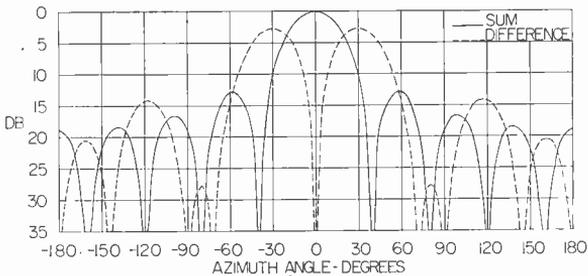
Fig. 13—Theoretical sum and difference patterns for uniform illumination at all scan angles. (Azimuth measures from center of beam.)

bility of the radial-waveguide antenna and electronic scanning circuitry has been confirmed experimentally. Important features of the system are:

1. Beam shape is independent of scan angle.
2. All of the power amplifiers can be used at full power all of the time. In systems employing amplitude scanning,[5] 50 per cent power utilization is the maximum.
3. Monopulse capability in azimuth is achieved with a simple addition to the scanning circuitry.
4. Mutual coupling between inputs to the antenna is very low.

ACKNOWLEDGMENT

CONCEPT FOR AN INTERCONTINENTAL
SATELLITE COMMUNICATION SYSTEM*

BY

EDMUND A. LAPORT† AND S. METZGER‡

Summary—A concept for a high-capacity microwave system for inter-continental telecommunications is discussed. This system provides for simultaneous and continuous access by any desired number of earth stations with full interconnectivity between them. The connectivity and traffic volumes can be varied as desired. The system uses a single repeater in a stabilized stationary satellite, with standard single-sideband suppressed-carrier telephone channels (width and separation) translated to the microwave region. The various station transmitting channels are inter-leaved at the satellite input, and their aggregate frequency-modulates the satellite transmitter. This signal is received by all stations, each of which selects the channels addressed to it by normal carrier techniques. The necessary standard frequency for coordination of transmitting and receiving channels is transmitted through the system from a master station which supervises operational disciplines and capacity assignments through continuous intercommunication channels operating simplex. Telegraphic subchannels are likewise fully versatile in routing to single or multiple destinations. Electronic aspects of the system appear feasible. The necessary rocketry and satellite stabilization technology are being developed.

INTRODUCTION

THE ability to launch artificial earth satellites immediately suggested to many engineers their possible use as vehicles to carry repeaters for global telecommunications. A large area of the world is visible from a satellite, the area increasing with its height. The period of a satellite increases with height; at 22,300 miles altitude, its period is twenty-four hours. At this height, the satellite's period equals the earth's rotational period so that its position can remain fixed relative to the earth if the orbit lies in the earth's equatorial

---

plane. This particular circumstance leads to the term "stationary" or "synchronous" satellite. Line-of-sight paths permit the use of micro-waves for intercontinental circuits.

SYNCHRONOUS SATELLITE

The synchronous satellite has engaged the imagination of com-munication engineers for several years because of its fixed position relative to earth and its very large area of geographical visibility. A primary justification for a satellite communication system is to pro-vide intercontinental telecommunications in quantity and quality com-parable to those realized with overland routes, and preferably at comparable costs per great-circle channel-mile. The present paper describes a concept# for the operational aspects of a global microwave communication system based on the use of stationary satellites.

The size and weight of the satellite will be limited, and therefore its power-generating capability will be limited, which in turn limits the repeater output power. This affects the choice of modulation, and the bandwidth needed to provide signals of desired quality over the distances from a satellite to the visible ground points and the number of repeaters on board for 2-way operation. Furthermore, the cost of placing a satellite repeater in orbit will be such that economic feasi-bility depends upon its having a very high communication capacity and intercontinental traffic sufficient to use the capacity. In the de-scription that follows, the existence of accurately positioned and oriented synchronous satellites, capable of providing a minimum of 10 watts continuous power output, will be assumed.

It will suffice to consider the situation for a single satellite. A global system can be started with an Atlantic satellite and others can be established later.

The earth–space geometry for a synchronous satellite is seen in Figure 1, while Figure 2 illustrates the geographical coverage for two and for three such satellites for earth points from which a satellite has an unrefracted elevation of five degrees or more. This minimum elevation angle is intended to avoid interference with sur-face communications using the same frequencies.

This concept was the basis of the RCA response to the Federal Com-munications Commission's Notice of Inquiry Docket 13522 for space frequency requirements, March 1, 1961, and subsequent testimony before Congressional committees.

Need for Multiple Access to Repeater

Within the communication zone of an Atlantic satellite there will eventually be at least 35 locations that will want direct access to the system. International telecommunications by satellite, therefore, require a sufficient generality of access to the system to permit many routing combinations between many countries, each with varying traffic volumes and time distributions for such traffic. Access to the system by any users should be independent of other traffic that may be circulating through the system between several other earth stations.

The following characteristics and capabilities may be provided with a synchronous-satellite system:
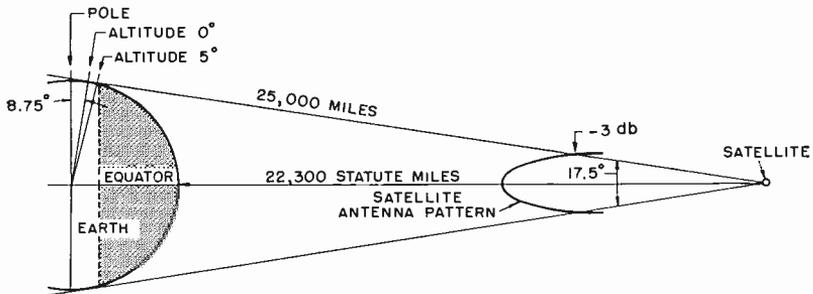
Fig. 1—Earth–space geometry for synchronous satellite.

1. A number of earth stations can make simultaneous use of the single high-capacity repeater in the stationary satellite. Thus, general direct access to the satellite is possible for all users within visibility of the satellite.

2. Every earth station can communicate with any other earth station in the same zone at any time through the same satellite. This means that every user can have his own ground station, located where it is most convenient to him. Transit traffic routing and tolls through other countries are avoided.

3. Each earth station uses only that portion of the system capacity needed for its traffic at any given time.

4. There is continuous, full intercommunication between all earth stations for technical and operational coordination of the system.

5. The satellite repeater uses only two frequency bands — one for communication from ground to satellite, the other for communication from satellite to ground.
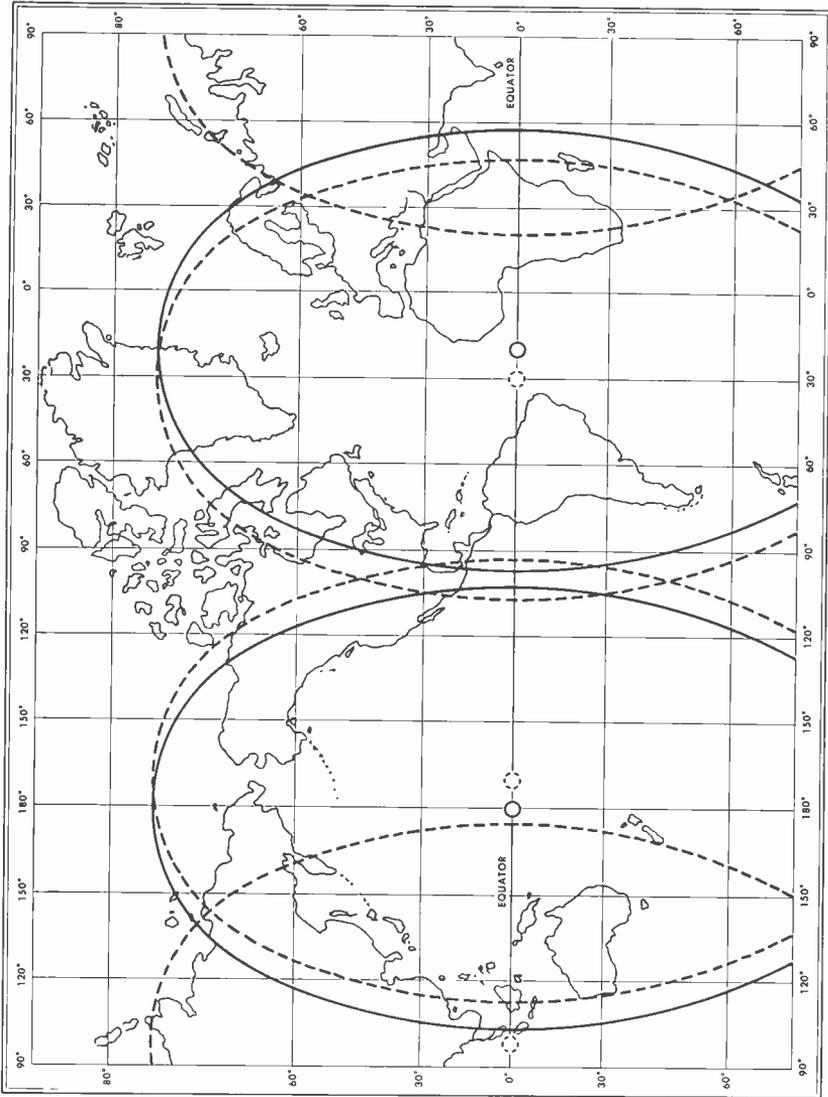
———— APPROXIMATE LOCUS OF POINTS OF 5° ALTITUDE OF TWO SYNCHRONOUS SATELLITES ON THE EQUATOR AT 20°W AND 180°.
– – – – APPROXIMATE LOCUS OF POINTS OF 5° ALTITUDE OF THREE SYNCHRONOUS SATELLITES ON THE EQUATOR AT 30°W, 170°W, AND 96°E.

Fig. 2—Geographical coverage of synchronous satellites.

OPERATIONAL ASPECTS

A suitable *initial* capacity should be at least 1000* duplex telephone channels. It is desirable to be able to shift channels among the various earth stations to accommodate changing demand. Such a capability tends toward uniform loading which favors economy and availability to potentially larger numbers of users. The technical nature of the system will be determined by the degree to which these and many other nontechnical conditions can be satisfied.

From a practical standpoint, it is essential that the satellite communication operation be compatible with existing telecommunication facilities. Thus traffic growth into satellites can supplement other facilities that may exist at the time satellite communication becomes a reality. At that time, satellite communication systems may become the main method for accommodating the expected traffic increases.

For some time into the future, the available power within a satellite will determine the main system parameters. It is preferable to apply the available power to a single repeater carrying 2-way traffic rather than two or more repeaters in a satellite.

The most practicable system providing general access is one using single-sideband suppressed-carrier inputs from the earth.[1] This can be done by translating carrier channels to the frequency bands for which the satellite input is designed. The interleaving of assigned channels at the satellite input from the participating ground stations requires that a standard system frequency be used by all earth stations for their transmission channels and for replacement carriers for detection of their receiving channels. Standard frequencies based on the same primary source could be transmitted from a master control station, to be used by all participating earth stations for system frequency coordination, for transmitting, and for receiving. The single-sideband suppressed-carrier channels from the earth stations are interleaved in their assigned slots at the satellite input to form a substantially continuous band of frequencies. This band may be detected, or it may be translated in frequency and used in turn to frequency-modulate the wave transmitted from satellite to earth after power amplification.

All earth stations, therefore, receive the same signal containing all

---

* 960 channels is a standard grouping for carrier channels.

[1] Metzger, S., "Some Aspects of Synchronous Satellite Communications System," *I.R.E., PGSET*, Fifth National Symposium 1960 on Space Electronics & Telemetry, paper 2.3.

of the channels from all stations. Each station need have only carrier terminal equipment for those particular channels it wants to receive from its correspondents' assigned operating channels designated for its particular circuit(s). The standard reference frequency is also received and used to derive control frequencies for the designated transmission channels and also the replacement carriers for demodulation. The standard pilot frequency passes through the repeater with the other input signals and is available at every earth station for its frequency coordination and for standard time for the system. In this way, an entire system working through one satellite can hold to the frequency disciplines required for SSB channel interleaving.

Full duplex telephony uses two assigned channels for each conversation, one for each direction of transmission. Alternatively, using voice-operated switching techniques, both sides of a conversation can use one channel. A 1-way channel can be used as a simplex party-line channel for supervisory control and load despatching. Another supervisory channel can be used for telecontrol of switching functions by transmitting control signals to all earth stations if automatic or time-assigned despatching functions are to be included in the system operational plan.

As an example, assume that the system is designed for an input capacity of 2000 standard 4-kilocycle single-sideband voice channels, translated into a continuous band of 8 megacycles from, say, 6000 to 6008 megacycles. Let Master Control use Channel 1 for system supervision and let Channel 2 be assigned for general party-line intercommunication. If station A is assigned 60 channels for its use at its discretion to its various correspondents, its ground transmission will be on these assigned carrier-derived channels. It can also use a party-line simplex (press-to-talk) channel for system coordination among its own correspondents and also for participation in the general supervisory communications. Those stations communicating with station A are equipped to receive all, or perhaps only certain designated channels between #1 and #60. In turn, say that station B is assigned twelve transmitting channels (channels 61-72) of which four are designated for permanent working with station A and the remainder available for working with other stations, some on time-shared schedules, and a number reserved for demand use.

Consider further six channels assigned to an international telegraph carrier, two of which are multiplexed with several on-line teleprinter exchange service (TELEX) circuits, another multiplexed for datagram circuits, and the remainder loaded with one-way message traffic. Each of these telegraph circuits may be addressed to a different

country which selects the telegraph channel(s) addressed to it. In reverse, a corresponding country transmits its multiplexed telegraph traffic within one or more of its assigned channels of the system, and each of its addressees receives a designated channel and selects the particular subchannel(s) of interest. In this manner, therefore, there is full versatility of telegraph circuits between any desired combination of points (including multiple-address traffic). These examples apply until the full system capacity is assigned (Figure 3).
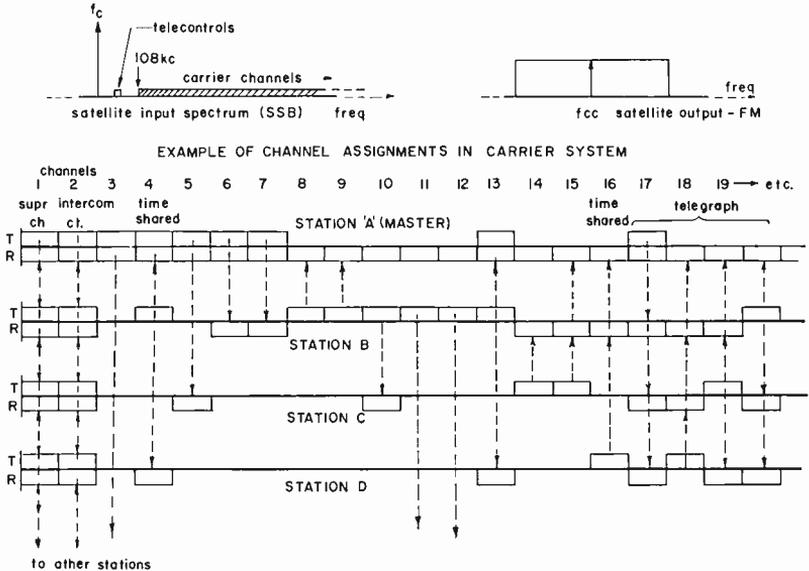
Fig. 3—Example of channel assignments in carrier system.

FLEXIBILITY

It will be recognized that this system plan has the flexibility of operation needed for intercontinental communications of the future, using only one satellite repeater to accommodate a large number of participating users and all classes of normal traffic, including musical programs and television.

A Master Control station, supplying frequency control, load despatching and technical supervision of the satellite and the system, would be equipped to receive and monitor all the channels of the system for proper technical coordination of signal levels, occupancy and even metered time usage if desired. The functional facilities for a master control station are listed in Appendix I.

Voice privacy for telephony could be used in the manner now employed for high-frequency radio telephony. Time assignment speech interpolation may be used between stations employing sufficient numbers of telephone channels.

Figure 1 shows that the distances from the satellite to various visible earth points do not differ greatly, so that the earth–satellite path attenuations will be substantially the same from all stations. This satisfies an essential condition of level control in a single-sideband multiplex system if only the average power per channel is held at a standard level at each earth station, a relatively simple operating requirement. The space-transmission loss is expected to be quite constant so that the integrity of relative signal levels should be preserved at the satellite input. Important inequalities in channel levels will be apparent at the master station monitoring position and corrections can be directed over the supervisory channel to trim antenna alignment or adjust transmitter power, though each station can monitor its own performance. The comparison of channel levels from various stations will also provide indication of the need for correcting satellite antenna attitude by master control. Monitoring standard tone levels from certain earth stations in different geographical locations could provide a sensor system for satellite attitude control, which conceivably could be made automatic.

DISPERSION EFFECTS

The diffuse distributions of free electrons throughout outer space and the relatively dense concentrations in the zones constituting the ionospheres have a potential dispersive effect on electromagnetic wave propagation. The differential delay for a band of signal frequencies depends upon the electron densities along a ray path, the path length, the frequency, and the emission bandwidth.

Computations have been made using available, though uncertain, numbers for electron densities. They indicate that dispersion will be negligible for those frequencies and bandwidths that would be applied for satellite communications.

AIRCRAFT INTRUSION EFFECTS

Aircraft in the beam of an earth station can produce a destructive shadow for the small interval of its transit. The attenuation of the path would depend on the aircraft size and shape, its distance from the station, and the frequency. Diffraction effects may be symmetric

or asymmetric and cause a rise and fall of signal, the minimum being a critical value for producing circuit interruptions. The latter is especially important for a master station that provides standard time for a system of stations using one satellite.

The locations of earth stations should be selected so as to minimize the number of aircraft intrusions. Computations of simplified diffraction from a disc of aircraft dimensions indicates that aircraft must be several miles away to have negligible influence on path loss. A 150-foot disc normal to the path at 24 miles gives a maximum shadow loss of 6 decibels at 6000 megacycles.

DOPPLER EFFECTS

Frequency change due to Doppler shifting results from any component of movement of the satellite in the direction of an earth station. The main cause of Doppler for a stationary satellite is orbital eccentricity, which gives it a radial displacement. Doppler shifts are therefore in the same direction for all earth stations, though the amount varies slightly with the elevation angle of each earth station. Several techniques are available for neutralizing Doppler frequency shift, but not for differential Doppler shift.

Doppler shift being proportional to the frequency, there is also a differential Doppler effect across a band of frequencies. This becomes important for very large bandwidths and for high relative satellite velocities. The synchronous satellite offers minimum values of Doppler and differential Doppler frequency shifts because the relative velocity of movement is intrinsically small.

DELAY

The total 2-way distance of wave travel between two ground stations via the satellite is of the order of 100,000 miles, which introduces a round-trip delay of about 0.55 second. This natural factor is not intolerable in view of the otherwise essentially ideal nature of such a system concept for telephony, as well as for other classes of traffic.

ECHO SUPPRESSION

Echo-suppression requirements are related to the propagation delay and call for new techniques in echo-suppressor design. Such techniques are under development and there is ample evidence at this early date to be optimistic about a satisfactory solution to this problem.

Earth–Station Transmitters and Receivers

The earth-to-satellite transmitters must be capable of amplitude-linear operation. The power output needed will vary with the number of channels to be transmitted. The peak envelope power needed increases rather slowly as the number of active channels (n) increases because of the statistics of loading. The average power varies as $10 \log_{10} n$, but the ratio of peak to average power varies oppositely. Computations indicate that the satellite receiver can have a nominal noise figure and be relatively conventional and therefore simple. Earth station transmitters will not require more than 10 kilowatts peak envelope power, with linearity suitable for CCITT standards for intermodulation. A standardized transmitter design is desirable to assure proper channel interleaving and stability.

The receiver at the earth station should be one having the utmost sensitivity available, regardless of satellite transmitter power output (which tentatively we may estimate to be of the order of 10 watts). The satellite-to-earth transmission by frequency modulation with a deviation ratio of the order of 5 or 6 (e.g., 100 megacycles total receiver noise bandwidth) and received by a paraboloidal antenna of 60-feet diameter, may benefit by using a small amount of FM feedback in the receiver to have a maximum operating margin above its threshold. However, feedback will not be necessary if an 85-foot-diameter antenna is used.

Duplicate Satellites

It will be necessary to provide a reserve satellite in readiness in a displaced position, to be used until an inoperative one can be replaced from the ground. The reserve repeater could be used for television relaying via separate ground transmitting and receiving equipment.

Rocketry

All depends, finally, on the eventual availability of suitable rocketry, stabilization techniques and reliability essential to the realization of a practicable synchronous satellite. Men working in these fields are confident that rockets and guidance systems capable of placing repeaters in synchronous orbit over assigned equatorial locations are attainable.[2]

---

[2] Golden, L. E., "The Synchronous Communication Satellite Mechanical Aspects," *I.R.E.*, *PGSET*, Fifth National Symposium 1960 on Space Electronics & Telemetry, paper 2.2.

Conclusion

The system concept described is one approach to the versatility that will be required for intercontinental telecommunications of the near future in which all nations could participate. The realization of universality of traffic routing and independent access to the system at the satellite repeater requires a stationary satellite, the launching and positioning of which, while beyond our immediate capability, will soon be attained.

Appendix I—Operational Functions

Master-control station transmits:

1. A standard pilot frequency within the carrier band which is received by all stations for deriving their transmitting frequencies, replacement carriers for channel demodulations, and system standard time;

2. telecontrol signals for satellite station holding, attitude adjustments and spare repeater switching;

3. its own traffic channels;

4. supervisory simplex voice intercommunication channel (common to the entire system) for technical supervision and load despatching;

5. system switching signals for selective channel assignments by time allotment or on demand for all ground stations;

6. general simplex party line intercommunication channel for interstation coordination of operations but distinct from master control's supervisory functions.

Master Control station must also keep the second (standby) satellite under control and in readiness in a displaced position. Duplicate radio equipment is required for this purpose.

Master-Control station receives:

1. All traffic channels from all stations for monitoring levels, occupancy, and technical operations, and metering of time usage;

2. standard pilot frequency for monitoring;

3. its own traffic channels;

4. supervisory simplex channel;

5. party-line simplex channel(s);

6. telemetry for monitoring satellite conditions.

A communication ground station transmits:

1. Its own traffic channels;
2. supervisory control simplex channel;
3. party-line simplex channel.

A communication ground station receives:

1. Standard pilot frequency and standard time;
2. its own traffic channels, including reserve channels for demand services;
3. carrier channel switching signals;
4. supervisory simplex channel;
5. party-line simplex channel.

APPENDIX II

The following is an example of how permanent telephone channel assignments could be made in an Atlantic satellite system of 1920 one-way (960 duplex) channels with 108 channels reserved for general demand use. Time sharing of channels can greatly increase the channels available for local busy hours over those shown.

USA Telephony	360
USA Telegraph	48
UK	360
Sweden	12
Norway	12
Finland	12
Germany	120
France	192
Italy	72
Spain	48
Portugal	48
USSR	144
Greece	8
Egypt	8
Central African Countries	8
South Africa	24
Other European Countries	48
Brazil	24
Argentina	24
Venezuela	24
Mexico	24
Canada	96
Colombia	24
Peru	12
Ecuador	12
Central America & Caribbean	12
Uruguay	12
Chile	12
Other South American..	6
System Supervision and intercomm. (simplex) .	6
General on-demand use .	108

THE EFFECT OF LARGE PUMP VOLTAGE ON TUNNEL DIODE DOWN CONVERTER PERFORMANCE*†

BY

H. J. PRAGER AND K. K. N. CHANG

RCA Laboratories
Princeton, N. J.

Summary—In a series of recent measurements on tunnel diode down converters, low noise factors and stable performance under greatly varying input conditions have been observed. In particular, low noise factors have been achieved through the application of relatively large pump voltages, of the order of 250 millivolts.

In this paper, an earlier theoretical analysis, which was based on a small-pump-voltage condition, is extended to include the effects of large pump voltage. The tunnel-diode current–voltage characteristic is represented by a polynomial expansion that includes higher-order terms. This expansion leads to new values of the diode conductance G_0 and the non-linearity term GV_3, and still further to a noise correlation term which is no longer negligibly small. With proper biasing of the diode, this correlation term contributes a partial noise cancellation.

Experimental results are discussed which were obtained with a 440-megacycle tunnel-diode down converter and which were instrumental in establishing and proving the "large pump" theory. A single-sideband system noise factor of 5.3 decibels was measured with a pump voltage of 200 millivolts. The diode was operated under self-bias conditions, which effectively put the bias at about 50 millivolts. In another test, with operation at a fixed bias of 360 millivolts and a pump voltage of 250 millivolts, the system noise factor was 5 decibels. Calculated values based on the "large-pump" theory closely predict these values.

INTRODUCTION

IN AN EARLIER PAPER by the authors on tunnel-diode down converters,[1] the possibilities of low noise factors and high gain were suggested and analyzed. Since then, improvements have been made; these include greater bandwidth and more stable perform-

---

[1] K. K. N. Chang, G. H. Heilmeier, and H. J. Prager, "Low Noise Tunnel Diode Down Converter Having Conversion Gain," *Proc. I.R.E.*, May, 1960.

$$G'_0 = \alpha + \frac{\gamma}{4} V_3{}^2 + \frac{\epsilon}{64} V_3{}^4 \tag{6}$$

and the nonlinearity term GV_3 becomes

$$G'V_3 = -\left(\frac{\beta}{2} V_3 + \frac{\delta}{16} V_3{}^3 + \frac{\zeta}{384} V_3{}^5 \right) \tag{7}$$

The I–V characteristic of a tunnel diode (see Figure 1) may be considered to consist of two regions. Region No. 1 comprises that

Table I

	Region No. 1 at P_1	Region No. 2 at P_3
V_0 (millivolts)	50	360
I_0 (milliamperes)	.96	.15
α	4.60×10^{-3}	1.30×10^{-3}
β	-416×10^{-6}	9.20×10^{-6}
γ	9.12×10^{-6}	$.060 \times 10^{-6}$
δ	-81.6×10^{-9}	37.0×10^{-9}
ϵ	2.98×10^{-9}	1.00×10^{-9}
ζ	-118×10^{-12}	13.5×10^{-12}

part of the curve that turns downward from the peak current I_A, while in region No. 2 the curve turns upward from the valley current I_B. The curves of these two regions can be represented by separate power series. Table I gives a summary of the expansion coefficients in both regions for a typical tunnel diode ($I_A = 1$ milliampere) used in recent experiments. It should be noted that the signs of the coefficients pertaining to the nonlinearity term $G'V_3$ are opposite for the two regions, the significance of which is considered in the discussion of the noise-factor expression.

Figure 2 shows how a power series with the coefficient of Table I representing the current–voltage characteristics fits the measured values. The reverse and forward components of the actually measurable diode current I, and the equivalent shot-noise current, I_e, are also shown. The relation of these currents has been established as follows:

forward current $\qquad\qquad I_F = \dfrac{I}{1 - \exp\,(-qv/kT)}$

reverse current $\qquad\qquad I_R = I_F \exp\,(-qv/kT)$

measurable diode current $\qquad I = |I_F| - |I_R|$

equivalent shot-noise current $\quad I_e = |I_F| + |I_R|$.
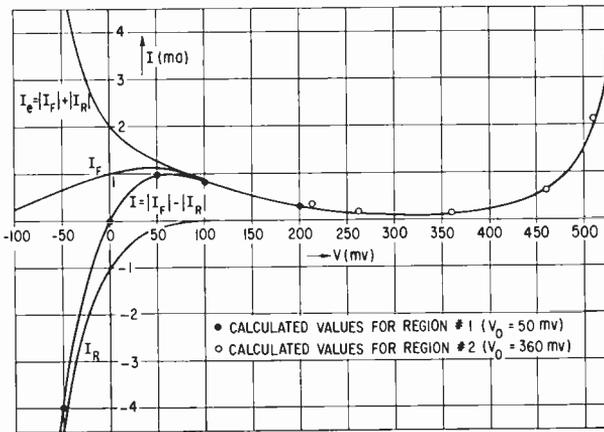
Fig. 2—Comparison of measured and calculated current–voltage characteristic of test diode.

NOISE FACTOR WITH CORRELATION TERM

The derivation of the extended noise-factor expression for the converter circuit (see Figure 3) closely follows the analysis presented in the Reference (1), except that the effect of the correlation term is no longer negligible.

There are two sources of noise present in the converter circuit. The first is due to the *thermal* noise sources in the individual r-f and i-f circuits:

$$|i_{1T}|^2 = 4k\Delta f(G_g T_0 + G_1 T)$$

$$|i_{2T}|^2 = 4k\Delta f(G_L T + G_2 T). \qquad (8)$$

These currents are uncorrelated with one another. The second type of noise is due to the *shot effect* in the diode itself, which produces correlated r-f and i-f currents. The reason for the correlation can be shown briefly as follows.

The shot-noise current of the diode is the sum of the current pulses due to individual electrons traversing the diode. The elementary current pulse can be represented by a delta function:
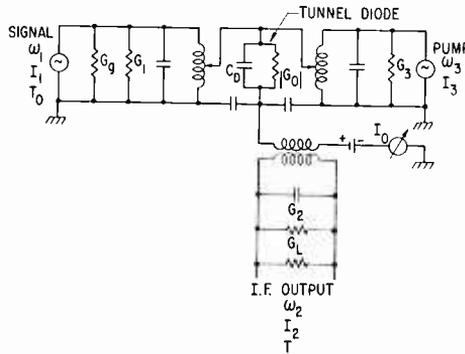
Fig. 3—Converter circuit.

$$i(t) = e \sum_K \delta(t - t_K), \qquad (9)$$

where e is the electron charge and t_K is the time at which the event occurs. The Fourier transform pair of the current are

$$i(t) = e \sum_K \frac{1}{\pi} \int_0^\infty \cos \omega (t - t_K) \, d\omega$$

$$I(\omega) = \int_{-\infty}^\infty i(t) \exp \{-j\omega t\} \, dt. \qquad (10)$$

Thus at time $t = t_K$, a continuous frequency spectrum of sinusoidal waves is generated in the converter circuit. However, because of the circuit resonances, only current waves of a passband $\Delta\omega$ centered at the input frequency ω_1 and at the i-f frequency ω_2 excite the circuit.

The current waves of interest are

$$i_2(t) = e \sum_K \frac{1}{\pi} \int_{\omega_1}^{\omega_1 + \Delta\omega} \cos \omega(t - t_K)\, d\omega$$

$$i_2(t) = e \sum_K \frac{1}{\pi} \int_{\omega_2}^{\omega_2 + \Delta\omega} \cos \omega(t - t_K)\, d\omega. \tag{11}$$

At the particular instant $t = t_K$, the nonlinearity of the diode gives rise to a second i-f current resulting from the interaction of the input circuit voltage and the local-oscillator voltage. Since all of these noise-current components originate from the same delta function at the same instant, the primary i-f noise current is fully correlated with the second i-f noise current which results from conversion of the primary r-f noise current. At the time $t = t_K$, the r-f and i-f shot noise currents are

$$i_{1S} = e \frac{1}{\pi} \int_{\omega_1}^{\omega_1 + \Delta\omega} \cos \omega(t - t_K)\, d\omega$$

$$i_{2S} = e \frac{1}{\pi} \int_{\omega_2}^{\omega_2 + \Delta\omega} \cos \omega(t - t_K)\, d\omega. \tag{12}$$

For the thermal noise currents (i_{1T}, i_{2T}) and the shot noise currents (i_{1S}, i_{2S}), the following equations can be written from the converter network

$$i_{1T} + i_{1S} = V_1 \overline{G_1} - G V_2^* V_3$$

$$i_{2T} + i_{2S} = V_2 \overline{G_2} - G V_1^* V_3 \tag{13}$$

and, solving,

$$V_2 = \frac{(i_{2T} + i_{2S}) + \dfrac{G V_3}{\overline{G_1}} (i_{1T}^* + i_{1S}^*)}{\overline{G_2} - \dfrac{(G V_3)^2}{\overline{G_1}}}. \tag{14}$$

The absolute value squared is

$$|V_2|^2 = |V_2|\ |V_2^*| = \cfrac{|i_{2T}|^2 + |i_{2S}|^2 + \cfrac{(GV_3)^2}{\overline{G_1^2}}\ (|i_{1T}|^2 + |i_{1S}|^2) + \cfrac{GV_3}{\overline{G_1}}\ (i_{1S}i_{2S} + i_{1S}^*i_{2S}^*)}{\left(\overline{G_2} - \cfrac{(GV_3)^2}{\overline{G_1}}\right)^2}, \tag{15}$$

which, averaged over the time distribution, becomes

$$\overline{|V_2|^2} = \cfrac{\overline{|i_{2T}|^2} + \overline{|i_{2S}|^2} + \cfrac{(GV_3)^2}{\overline{G_1^2}}\ (\overline{|i_{1T}|^2} + \overline{|i_{1S}|^2}) + \cfrac{GV_3}{\overline{G_1}}\ \overline{(i_{1S}i_{2S} + i_{1S}^*i_{2S}^*)}}{\left(\overline{G_2} - \cfrac{(GV_3)^2}{\overline{G_1}}\right)^2}. \tag{16}$$

In these equations, i_{1S} and i_{2S} are the original shot-noise currents and $(GV_3/\overline{G_1})\ (i_{1S}i_{2S} + i_{1S}^*i_{2S}^*)$ is the cross-product of these two correlated currents. The averages of the mean square currents over the time distribution are

$$\overline{|i_{1S}|^2} = \overline{|i_{2S}|^2} = 2I(\omega)^2\,df = 2eI_e df = 4kTG_e df, \tag{17}$$

and thus

$$\frac{GV_3}{\overline{G_1}}\ \overline{(i_{1S}i_{2S} + i_{1S}^*i_{2S}^*)} = 2\frac{GV_3}{\overline{G_1}}\ 4kTG_e\Delta f, \tag{18}$$

where G_e is the equivalent noise conductance. The noise correlation term can be either additive or subtractive, depending on the sign of the factor $G'V_3/\overline{G_1}$, which, in turn, depends on the shape of the I–V characteristic of the tunnel diode and the bias point. Thus, the significance of large pump voltages and the sign of the coefficients shown in Table I becomes apparent.

A further comment has to be made about the value of G_e. In region No. 1, where the tunneling noise is predominant, $G_e = eI_e/(2kT)$ where I_e is the equivalent diode shot-noise current under large-pump-voltage conditions, and equals the sum of the averages $|I_F| + |I_R|$. Beyond the valley point in region No. 2, the thermal noise becomes

important and G_e is merely equal to the linear term of the conversion conductance.

Substituting these terms into the expression for the noise factor (see Equation (25) of Reference (1)),

$$F = \frac{V_2{}^2\,G_L}{g_c k T_0 \Delta f} \tag{19}$$

and using Equation (5) of Reference (1),

$$g_c = \frac{4G^2 V_3{}^2 G_g G_L}{[\overline{G_1}\overline{G_2} - G'^2 V_3{}^2]^2}\,,$$

results in

$$F = \frac{G_L}{g_c k T_0 \Delta f}\; \frac{g_c \overline{G_1}{}^2}{4G'^2 V_3{}^2 G_g G_L}\, 4k\Delta f$$

$$\left[\; (G_g T_0 + G_1 T + G_e T)\,\frac{G'^2 V_3{}^2}{G_1{}^2} + 2\,\frac{G' V_3}{\overline{G_1}}\; G_e T\; + G_L T + G_2 T + G_e T\right] \tag{20}$$

which reduces to:

$$F = 1 + \frac{T}{T_0}\left[\;\frac{G_1}{G_g} + \frac{G_e}{G_g} + \frac{\overline{G_1}{}^2}{G_g G'^2 V_3{}^2}\left\{ G_L + G_2 + \left| G_e\left(1 + 2\,\frac{G' V_3}{\overline{G_1}}\right)\right|\right\}\right]\;. \tag{21}$$

This equation is similar to the one derived for the small-pump analysis, except that now the noise correlation $2G' V_3/\overline{G_1}$ is included and the values G_0 and $G V_3$ are replaced with the newly found G'_0 and $G' V_3$.

REDUCTION OF NOISE FACTOR

To minimize the noise factor,

(a)   The shunt admittances G_1 and G_2, i.e., the losses of the r-f and i-f circuit, should be made as low as possible.

(b)   The load admittance G_L should be made small.

(c)   The term G_e/G_g should be made as small as possible. This

deep into the region of negative $I–V$ slope. Because the tunnel diode exhibits this negative $I–V$ slope, it can produce conversion gain. Consequently, the tunnel diode is capable of further reducing the total system noise factor.

To demonstrate this effect, consider a tunnel diode mixer and a crystal-diode mixer with identical 3.2 decibel noise factors. Assume that the noise factor of the following broad-band i-f amplifier, F_{IF}, is 3.0 decibels. With a typical conversion loss for the crystal of $1/g_c = -6$ decibels and a conversion gain for the tunnel diode of $g_c = 5.0$ decibels, the total system noise factor for the crystal system is

$$F_{TOTAL} = 2.09 + \frac{2 - 1}{.25} = 6.09 \text{ or } 7.85 \text{ decibels,}$$

and for the tunnel-diode system is

$$F_{TOTAL} = 2.09 + \frac{2 - 1}{3.16} = 2.40 \text{ or } 3.82 \text{ decibels.}$$

Thus, even with a modest gain, the tunnel-diode converter is effective in reducing the system noise factor.

RESULTS

A few simple calculations will illustrate the reduction of the noise factor through the use of large pump powers when G_0 becomes G_0' (see (d) above). For the specific diode under consideration, the power series coefficients at the bias point P_1 ($V_0 = 50$ millivolts, $I_0 = .96$ milliampere) are given in Table I. These values permit the calculation of the terms G'_0 and $G'V_3$ for various values of pump voltage V_3. Figure 5 shows the results of this calculation for pump voltages up to 250 millivolts. With these terms established, the noise factor, F, and the conversion ratio, g_c, can be calculated for various values of G_g and G_L. Figure 6 shows the results of such calculations. In these curves, G_g ranges from 20 to 180 millimhos, G_L is chosen at a typical value of 3 millimhos, and G_1 and G_2 are assumed to be negligible. G_e is approximated by the shot-noise current at the bias point, $G_e = 20 \times I_e = 20 \times 1.3 \times 10^{-3} = 26$ millimhos. Curves for two values of pump voltage, $V_3 = 200$ millivolts and 225 millivolts, are shown in Figure 6.

Since a low-noise converter can be used to full advantage only if it does not introduce a conversion loss by itself, values for the total
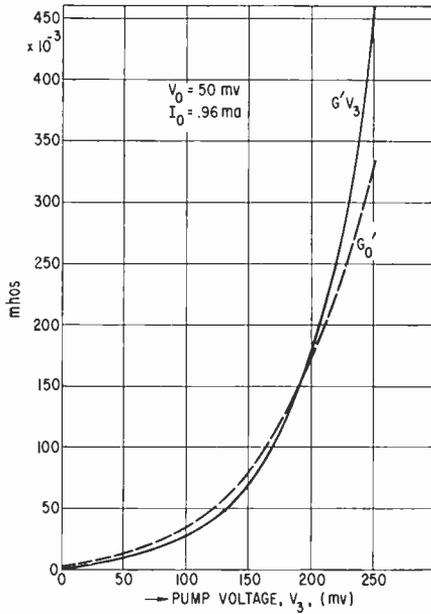
Fig. 5—Calculations of values of G' V_3 and G_0' versus V_3.
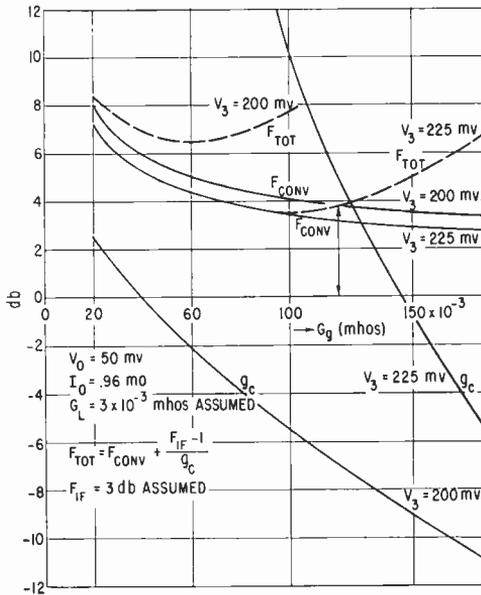
Fig. 6—Calculated values of noise factor and gain for $V_0 = 50$ millivolts and $I_0 = .96$ milliampere.

noise factor of a system have also been calculated. In this calculation, a noise factor of 3 decibels has been assumed for a broad-band i-f amplifier following the converter stage. A reasonable compromise solution for low noise factor and stable gain seems to be reached at a pump voltage of 225 millivolts and a G_g of 120 millimhos. These values yield a calculated converter noise factor of 3.2 decibels, and a total system noise factor of 3.9 decibels.

These calculated results are in good agreement with experimental results. The tunnel diode down converter used for the test set-up was a lumped-parameter circuit operating at a signal frequency of 440 megacycles, a pump frequency of 470 megacycles and an intermediate frequency of 30 megacycles. One of the features of this circuit was a double-tuned input transformer which allowed the r-f impedance of the signal as seen by the diode G_g to be changed and optimized as predicted by the theoretical curves of Figure 6. A total-system, single-sideband noise factor of 5.3 decibels was measured with a pump voltage of 200 millivolts.

Another important feature of this circuit is the use of a self-bias scheme for the tunnel diode. No external d-c bias source is supplied, and the diode establishes its own bias voltage through the rectification of the pump voltage. At these operating conditions, the measured bias voltage across the diode was about 50 millivolts.

A similar calculation can also be performed to explore the possibilities of noise reduction by minimization of the term $[1 + 2 (G'V_3/\overline{G_1})]$. For this case (see (f) above), the diode is biased at $P_2(V_0 = 360$ millivolts, $I_0 = .15$ milliampere), and the corresponding power-series coefficients can again be found in Table I. After calculation of a new series of term for G'_0 and $G'V_3$ for various pump voltages, the noise factor and gain can be determined as a function of G_g and V_3. Figure 7 shows the results of these calculations. It should be noted that the minimum converter noise factor occurs at the point where the correlation term $2(G'V_3/\overline{G_1})$ reaches the value of -1. This is possible because of the negative sign for G' together with a V_3 large enough for their product to cancel out the effective diode noise contribution in the last part of the noise-factor equation.

The minimum total system noise factor in this case occurs at a pump voltage of 225 millivolts and a G_g of about 32 millimhos. The converter noise factor is 4.2 decibels, the conversion gain is about unity, and the total noise factor is 5.55 decibels. As might be expected, the conversion gain is rather low at this bias point because the $I-V$ characteristic near the valley point is very flat.

Experimental results, which were instrumental in establishing and
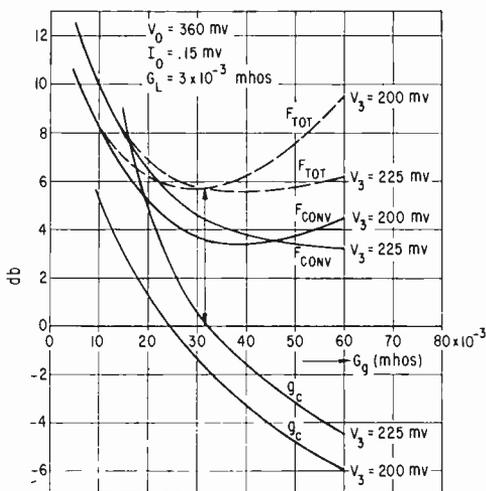
Fig. 7—Calculated values of noise factor and gain for $V_0 = 360$ millivolts and $I_0 = .15$ milliampere.

proving the "large-pump" voltage theory, were obtained with the same tunnel-diode converter described above.* A single-sideband noise factor of 5 decibels was measured for the total system consisting of input filter, converter, and i-f amplifier. The diode bias voltage was 360 millivolts and the pump voltage as measured across the diode was 250 millivolts.

ACKNOWLEDGMENT

The authors wish to thank L. S. Nergaard for his helpful suggestions in the calculation of the shot noise correlation, and also A. Simon and P. Chase for writing the necessary computer programs.

---

* Similar measurements with large pump voltages were carried out independently by F. Sterzer and A. Presser of the RCA Electron Tube Division, and have produced substantiating results.

HIGH-SPEED PRINTING ON ELECTROFAX*†

BY

R. G. OLDEN

RCA Laboratories,
Princeton, N. J.

Summary—The thin-window cathode-ray tube[1] for contact printing has been simplified and improved. An experimental, fully automatic printer for delivering finished Electrofax copy at the rate of 10,000 letters per second has been built. Excellent prints have been obtained both with facsimile and sequential letter recording. The design and construction of the thin-window tube and of the high-speed printer are described. The experimental arrangements for facsimile and sequential letter printing are discussed.

INTRODUCTION

THE efficiency of modern high-speed electronic computers is presently restricted by the relative slowness of the output devices. Electromechanical printers, although surprisingly fast, cannot be expected to match the speed and controllability of electrons. This has been recognized for some time and efforts are underway in many quarters to devise electronic printers.

One of the promising schemes has been a combination of Electrofax and the thin-window cathode-ray tube, models of which were built at the David Sarnoff Research Center in 1957.[1] These demonstrated the practicality of contact printing on Electrofax[2] paper and showed promise of printing speeds of thousands of characters per second. Since that time the technique of making the tube has been fully mastered and its design streamlined. An experimental printer which delivers finished Electrofax prints at the rate of 10,000 characters per second has been built and demonstrated.

THE THIN-WINDOW CATHODE-RAY TUBE

The purpose of the thin-window cathode-ray tube is to permit contact printing on Electrofax paper. The tube consists of a conven-

---

* Registered trade mark.

† Manuscript received July 3, 1961.

[1] R. G. Olden, "A Thin-Window Cathode-Ray Tube for High-Speed Printing with ELECTROFAX," *RCA Review*, Vol. XVIII, No. 3, p. 343, Sept. 1957.

[2] C. J. Young and H. G. Greig, "ELECTROFAX-Direct Electrophotographic Printing on Paper," *RCA Review*, Vol. XV, No. 4, p. 469, Dec. 1954.

tional neck and gun assembly sealed to a wedge-shaped metal body with a metal face-plate (Figure 1). A long narrow slit in the face-plate is covered with a very thin mica window. The phosphor is deposited on the inside of this window. The present tube is designed primarily for computer printout application. The window is made wide enough to accept a line of standard characters. The mica which is sealed over the slit is curved outward to insure good contact with the moving paper web. Its thickness is made just sufficient to safely support atmospheric pressure on the arch, since extra thickness results in loss of resolution
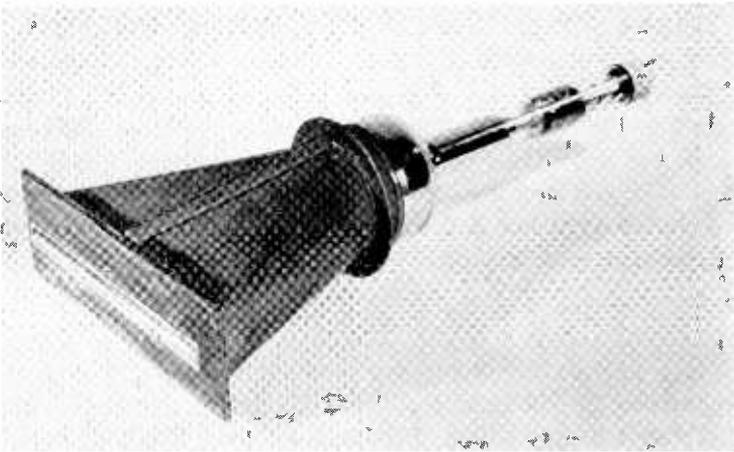
Fig. 1—Thin-window cathode-ray tube.

in the printed image. These considerations have led to the choice of .0025 inch mica for a window .160 inch wide. The length of the window in the present tube is 8½ inches. Natural mica of good quality is available if longer windows are desired.

It is of interest to note that a tube designed for line-by-line facsimile printing would require a slit only .030 to .040 inch wide. In this case it has been found that the mica need be only .001 inch thick and the resolution would be improved by a factor of two.

The main problem in building the thin-window tube was to make a vacuum-tight mica seal. Techniques for making such seals with flat pieces of mica have been known for a long time. For this tube, however, the mica had to be curved outward and the window had to be at least 8 inches long for a full scanning line. Such a seal required a special technique. The materials were selected so that their coefficients of thermal expansion matched closely in the region under consideration

(20° to 600°C.). The face-plate was made of metal, since the forming and machining is easier than with glass when making only a few units. Solder glass was used for the bond between the metal face-plate and the mica window. The adjoining piece also was made of sheet metal by a simple bending technique. A Kovar* ring permitted sealing to the conventional glass neck. All welds were made by heliarc and all seams, except the round flange, were glazed with the solder glass, which insured tight joints and made leak testing unnecessary. The light source for exposing the Electrofax paper is the P-11 phosphor which is settled on the inside of the mica window in the conventional manner. This phosphor closely matches the spectral response of the dye-sensitized Electrofax paper employed in the printing tests. A high-resolution electron gun is used to insure a small spot size.

The vacuum obtained with these tubes ranges from 1.5×10^{-7} mm Hg to 3×10^{-8} mm Hg. The several laboratory-built tubes have held vacuum for more than two years and have undergone well over 1000 hours of operation without showing any sign of deterioration. No thin-window tube has failed in spite of the considerable amount of handling and experimentation.

FACSIMILE PRINTING

Since an existing flying-spot scanner was readily available, the first tests of the tube and associated Electrofax apparatus were made using facsimile recording. Figure 2 shows a block diagram of the arrangement. The master consisted of typewritten material, photographed and made into a lantern slide. The beam of the 5ZP16 flying spot scanning tube was focused on the slide and the transmitted light caught by a 5819 photomultiplier which generated the electrical signal. This signal was then amplified and fed into the thin-window tube where it was reproduced scanning line by scanning line.

The printing speed was arbitrarily set at 10,000 characters per second, which required a paper speed of 19 inches per second.

Several types of Electrofax paper are available; the standard white paper, which is the slowest, and the slightly off-white dye sensitized papers, which are considerably faster. For the tests described here, the off-white paper, dye sensitized with fluorescein sodium, was used because its spectral response curve matches closely the spectral emission of the P-11 phosphor. The average beam current needed to produce the writing speed of 10,000 characters per second varied between 1 and 1.5 microamperes. It can, therefore, be stated that, as far as the

---

* Registered trade mark.
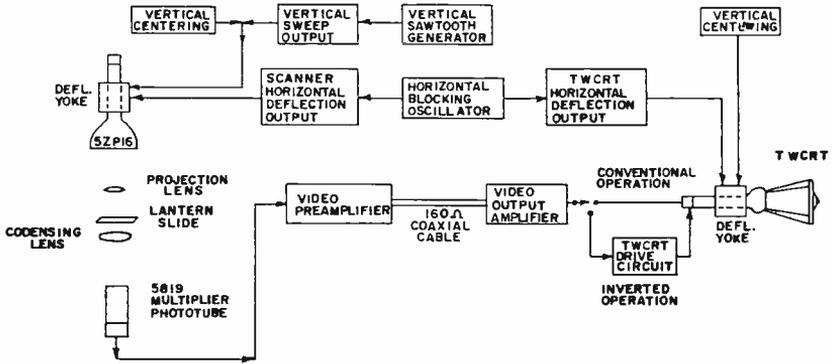
Fig. 2—Block diagram of test arrangement for facsimile recording.

thin-window tube is concerned, much higher printing speeds can be achieved. Test runs have shown that even at speeds corresponding to 150,000 characters per second (a spot speed of 300,000 inches per second) essentially the same definition of 200 lines per inch can be expected. The steep gamma of the Electrofax paper contributes, of course, to better resolution by responding only to the peak of the spot.

A typical facsimile print sample, made at the speed of 10,000 letters per second, is shown in Figure 3. It should be noted that throughout these tests a reverse toner was used for development, i.e., a toner which holds a negative triboelectric charge and therefore adheres to the exposed or relatively positive areas, and is rejected by the negatively charged background.

COMPUTER PRINTOUT

In order to print with binary input, as from a computer, the arrangement shown in the block diagram, Figure 4, was used.[3] Com-

This print was produced on "Electrofax" paper by contact printing from the face of a thin-window cathode ray tube. The thin window is a mica strip, .0025 inch thick, sealed by a special technique to a metal face plate. The height of the window is .130 inch which is sufficient to reproduce a typewritten capital letter. The window is 8 inches in length and was designed to accomodate the writing space on standard 8.5 x 11 paper. For this copy, the image is formed facsimile-wise by line-by-line scanning. The printing speed is 10,000 characters per second with the paper moving at 19 inches per second. At this speed the average beam current is 1.5 microamperes. Since the paper is abrasive, a sheet of "mylar", .0005 inch thick, protects the mica window. The "mylar" shield is renewed from time to time as it becomes worn. With the present geometry of the tube face plate, the diameter of the printed spot is .005 inch. This corresponds to a definition of 200 photographic lines per inch.

Fig. 3—Facsimile print sample (2/3 actual size).

[3] See U. S. Patent No. 2,807,663 to C. J. Young.

puter output was simulated by a six-bit binary punched paper tape. The light from a small incandescent bulb, stretched into a line by a cylindrical lens, is focused on a bank of 1N77B photo diodes which are aligned with the holes in the paper tape. As the light beam is permitted to fall on the photo-diodes by the holes in the rapidly moving tape, an electrical signal is generated which is conducted into the main yoke of the special monoscope tube. The target of this tube comprises a letter chart, Figure 5, containing all the symbols to be reproduced. The electron beam in the monoscope is moved to the
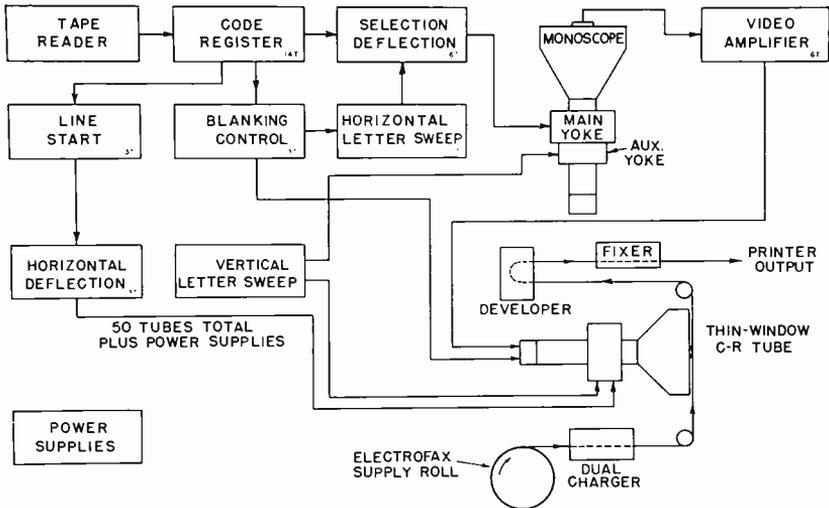
Fig. 4—Block diagram of binary input test arrangement.

selected character and scans it vertically in a small raster at 28 kilocycles. The printout tube has a horizontal sweep of 30 cycles per second, corresponding to the printing speed of 30 type lines per second. This relatively slow speed was chosen as a matter of experimental convenience; as indicated above, much higher speeds are possible. Superimposed on the horizontal sweep is the vertical sweep of 28 kilocycles held in phase with that in the monoscope. In this case the vertical scan is sinusoidal with an amplitude of about twice the letter height. Conventional sawtooth scan has been used in other experiments. After amplification, the signal from the monoscope is impressed on the control grid of the thin-window tube. The necessary circuitry is provided for horizontal deflection and vertical letter sweep in this tube. Figure 6 shows a print sample, made in this manner.

Y Z

Q R S T U V W X

I J K L M N O P

A B C D E F G H

5 6 7 8 9 , #

/ 0 1 2 3 4

% ; & ' — ☼ ●

— + ( ) ❞ : $

Fig. 5—Letter chart in monoscope.

THIS PRINT WAS PRODUCED ON ELECTROFAX PAPER BY CONTACT PRINTING FROM THE
FACE OF A THIN-WINDOW CATHODE-RAY TUBE. THE INPUT WAS PROVIDED BY A
PUNCHED PAPER TAPE. THE ALL-ELECTRONIC PRINTER IS AT PRESENT WRITING AT
A SPEED OF 30 TYPE LINES PER SECOND WITH THE PAPER MOVING AT 5 INCHES
PER SECOND. THE RECORDING METHOD WAS BY RCA ULTRATYPE WITH THE BEAM
CURRENT LESS THAN I MICROAMPERE.

THIS PRINT WAS PRODUCED ON ELECTROFAX PAPER BY CONTACT PRINTING FROM THE
FACE OF A THIN-WINDOW CATHODE-RAY TUBE. THE INPUT WAS PROVIDED BY A
PUNCHED PAPER TAPE. THE ALL-ELECTRONIC PRINTER IS AT PRESENT WRITING AT
A SPEED OF 30 TYPE LINES PER SECOND WITH THE PAPER MOVING AT 5 INCHES
PER SECOND. THE RECORDING METHOD WAS BY RCA ULTRATYPE WITH THE BEAM
CURRENT LESS THAN I MICROAMPERE.

THIS PRINT WAS PRODUCED ON ELECTROFAX PAPER BY CONTACT PRINTING FROM THE
FACE OF A THIN-WINDOW CATHODE-RAY TUBE. THE INPUT WAS PROVIDED BY A
PUNCHED PAPER TAPE. THE ALL-ELECTRONIC PRINTER IS AT PRESENT WRITING AT
A SPEED OF 30 TYPE LINES PER SECOND WITH THE PAPER MOVING AT 5 INCHES
PER SECOND. THE RECORDING METHOD WAS BY RCA ULTRATYPE WITH THE BEAM
CURRENT LESS THAN I MICROAMPERE.

Fig. 6—Print sample, binary input (3/5 actual size).

High-Speed Electrofax Printer

Having demonstrated the practicality and advantages of the thin-window tube it remained to be shown that the Electrofax paper can be processed automatically at web speeds as high as 19 inches per second. A printer designed for this purpose is shown in Figure 7.
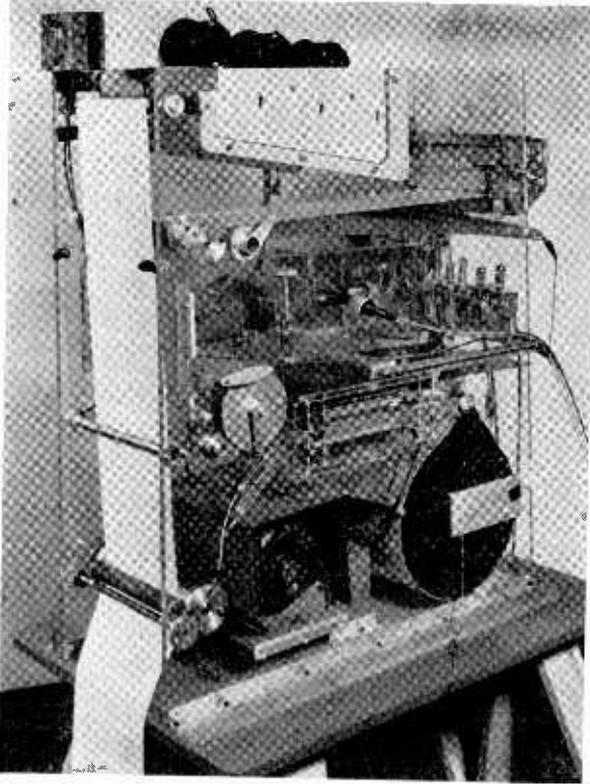
Fig. 7—High-speed printer.

The paper supply roll is housed in a closed compartment which may be humidity controlled when necessary. Upon emerging from this housing the paper is charged by a double corona device. The charged paper passes over the face of the thin-window tube and is fed into the developer. This consists of an applicator brush which brings a mixture of iron filings and toner into contact with the latent image on the paper. A magnetic brush removes iron filings sticking to the charged paper. The developer also contains a powder feed device which meters the toner into the filings. The print then passes through

the fuser which consists of three highly efficient hot air heaters. Here the toner is heated to a temperature above its melting point but below the ignition temperature of the paper. The softened toner fuses into the paper and makes the print permanent. Since the whole metallic front end of the thin-window tube is at the second anode potential of plus 20 kilovolts, the side plates of the machine were made of Lucite* which eliminated the insulation problem, facilitated machining, and is, in general, highly advantageous for an experimental model.

In order to protect the relatively soft mica window from the abrasive actions of the Electrofax paper, a sheet of .0005 inch thick Mylar* is sandwiched between the two members. It has been found that a strip of Mylar remains sufficiently clear through several hundred feet of paper so that the speed of the Mylar web can be very slow compared with that of the paper.

CONCLUSION

High-speed contact printing with a thin-window cathode-ray tube has been proved feasible. Prints of excellent quality have been produced both by facsimile and sequential letter recording. Extremely high printing speeds with good definition can be reached. At the present time the speed limit is imposed, not by the tube, but by problems involved in the high-speed handling and processing of the paper. If a simplification in the processing of the paper can be achieved, then the road to even higher speed printing will be open.

A model of a compact, fully automatic printer has been demonstrated. Permanent prints of high quality have been obtained at the rate of 10,000 characters per second with only 1 microampere of beam current.

Many applications are in sight. Foremost among them is, of course, computer output printing. Besides this there are information retrieval, trace recording, address labeling, and applications in the military and telemetering fields.

ACKNOWLEDGMENTS

---

* Registered trade mark.

RCA TECHNICAL PAPERS†

Any request for copies of papers listed herein should be addressed to the publication to which credited.

"Color Programming on NBC 1960-61," D. Levy, *Broadcast News* (April) 1961

"High-Speed Logic Using Low-Cost Mesa Transistors," D. Gipp, R. D. Lohman, R. R. Painter, and B. Zuk, *Semiconductor Products* (April) 1961

"High Speed Oscilloscope with Electron Optical Magnification Using Four-Pole Lenses," A. B. El-Kareh, *Rev. Sci. Instr.* (April) .. 1961

"Improvements in Color TV Equipment," J. H. Roe, *Broadcast News* (April) 1961

"Injection Electroluminescence," A. G. Fischer, *Solid State Electronics* (April) 1961

"Investigation of the Electrochemical Characteristics of Organic Compounds—VII. Organic Positive Iodine and Aliphatic AzO Compounds," C. K. Morehouse and R. Glicksman, *Jour. Electrochem. Soc.* (April) 1961

"New Low-Light-Level Camera Tube Expands Indoor and Outdoor Color Programming," *Broadcast News* (April) 1961

"A New Parametric Amplifier Using a Semi-Conductor-Diode-Loaded Helix," C. L. Cuccia, *Electronic Products* (April) 1961

"Note on the Mechanism of the Multipactor Effect," F. Paschke, *Jour. Appl. Phys.* (April) (Letters to the Editor) 1961

"Operation of Vidicons in Unusual Environmental Conditions," G. A. Robinson, *Jour. S.M.P.T.E.* (April) 1961

"Some Electron Trajectories in a Nonuniform Magnetic Field," T. N. Chin, *Proc. I.R.E.* (April) (Correspondence) 1961

"The Transidip," L. W. Aurick, *Popular Electronics* (April) 1961

"Transistor Internal Parameters for Small-Signal Representation," W. M. Webster and Coauthors, *Proc. I.R.E.* (April) 1961

"Tunnel-Diode Microwave Oscillators," F. Sterzer and D. E. Nelson, *Proc. I.R.E.* (April) 1961

"The Weighted Resistor Decoder: An Error Analysis," N. Aron and R. Goundry, *Electro-Technology* (April) 1961

"Predicting Spurious Transmitter Signals," J. G. Arnold, *Electronics* (April 21) 1961

"Prepunched Cards Test Data-Link Modules," W. Mergner and J. A. Paschall, *Electronics* (April 21) 1961

"Reduction of Cable Capacitance Loading," H. W. McCord, *Electronic Design* (April 26) (Ideas for Design) 1961

"Computers in Industrial Control," E. W. Engstrom, *Electronics* (April 28) 1961

"Analysis of Lead Telluride with an Accuracy to Better than 0.1%," K. L. Cheng, *Analytical Chemistry* (May) 1961

"Design Considerations for Modulators and Converters Using a New Miniature Beam-Deflection Tube," M. B. Knight and J. T. Maguire, *Electronic Equipment Engineering* (May) 1961

"Determination of Traces of Tellurium Bismuthiol II as a Reagent for Tellurium," K. L. Cheng, *Talanta* (May) 1961

† Report all corrections to *RCA Review*, RCA Laboratories, Princeton, N. J.

"Value Engineering and Product Engineering," C. Fallon, *Trans. I.R.E. PGPEP* (June) 1961

"VHF Television Translators," R. S. Jose, *Broadcast News* (June) 1961

"Effect of Temperature and Doping on the Reflectivity of Germanium in the Fundamental Absorption Region," M. Cardona and H. S. Sommers, Jr., *Phys. Rev.* (June 1) 1961

"Modifying Vidicon Camera Chain for Slow-Scan Television Systems," F. F. Martin and C. T. Shelton, *Electronics* (June 9) 1961

"Computers for Industrial Control," R. W. Sonnenfeldt, *I.R.E. International Convention Record, Part 6, Component Parts; Industrial Electronics; Product Engineering and Production; Reliability and Quality Control; Ultrasonic Engineering* 1961

"Degassing and Permeation of Gases in Tube Materials," R. H. Collins and J. C. Turnbull, *Advances in Electron Tube Techniques*, p. 186, Pergamon Press, New York, N. Y. 1961

"The Effects of Oxygen on Emission of Oxide Cathodes," W. G. Rudy, *Advances in Electron Tube Techniques*, p. 94, Pergamon Press, New York, N. Y. 1961

"Experimental Research on Plasma Thermionic Energy Converters," K. G. Hernqvist, *Progress in Astronautics and Rocketry*, Vol. 3, Academic Press, Inc., New York, N. Y. 1961

"Frequency Division Multiplexing on Transoceanic Cables," W. Lyons, *I.R.E. International Convention Record, Part 8, Communications Systems; Radio Frequency Interference; Vehicular Communications* 1961

"High-Power Load and Attentuator Design," E. M. Leyton and J. T. Coleman, *Conf. Proc. Fifth National Convention of Military Electronics* 1961

"How Vacuum Tube Problems Are Solved by the Use of Radioisotopes," H. A. Stern, *Advances in Electron Tube Techniques*, p. 55, Pergamon Press, New York, N. Y. 1961

"An Improved Loudness Indicator," J. L. Hathaway, *I.R.E. International Convention Record, Part 7, Broadcasting; Broadcast and Television Receivers* 1961

"Materials and Form Factors for Micromodule Inductors," G. G. Hauser, *I.R.E. International Convention Record, Part 6, Component Parts; Industrial Electronics; Product Engineering and Production; Reliability and Quality Control; Ultrasonics Engineering* 1961

"The Measurement of Internal Physiological Phenomena Using Passive-Type Telemetering Capsules," V. K. Zworykin, F. L. Hatke, and Coauthors, *I.R.E. International Convention Record, Part 9, Bio-Medical Electronics, Nuclear Science; Instrumentation* 1961

"The Micro-Module Program," D. Mackey, *Proc. Electronic Components Conf.* 1961

"Micromodules in Avionics—Applicable, Practical." G. R. Sievers and M. DiBartolomeis, *I.R.E. International Convention Record, Part 5, Aerospace and Navigational Electronics; Human Factors in Electronics; Military Electronics; Space Electronics and Telemetry* 1961

"A Miniature-Package 2200-Mc Parametric Amplifier Using a Varactor-Loaded Helix," C. L. Cuccia, *Proc. Electronic Components Conf.* 1961

"New Design Methods for Obtaining 40g Vibrational Capability in Kilowatt-Class Tubes." J. J. Free, *Advances in Electron Tube Techniques*, p. 214, Pergamon Press, New York, N. Y. 1961

"A New Method of Magnetron Tuning and Frequency Stabilization." R. M. Salzer and R. Steinhoff. *I.R.E. International Convention Record, Part 3, Electron Devices; Microwave Theory and Techniques* 1961

AUTHORS

J. J. BRANDINGER received the B.S. degree in Electrical Engineering from Cooper Union, New York City, in 1951. He has done graduate work at Rutgers State University at New Brunswick, New Jersey. In 1951, he became a member of the technical staff at RCA Laboratories at Riverhead, New York, where he worked on the development and testing of diversity receivers. He has since conducted extensive evaluation tests of teletypewriter, facsimile, and voice communications in the high-frequency spectrum. He has also conducted a study of auroral propagation effects. Mr. Brandinger is presently the group leader for the systems Communications Research Group at RCA Laboratories, Princeton, N. J. Mr. Brandinger is a member of the Institute of Radio Engineers, American Geophysical Union, and American Radio Relay League.

K. K. N. CHANG (See *RCA REVIEW*, Vol. XXII, No. 2, June 1961, p. 353.)

HOWARD R. CHARNEY received the B.E.E. degree, Cum Laude, from the Polytechnic Institute of Brooklyn in 1958. Upon graduation, he became a research fellow at the Microwave Research Institute where he worked on the analysis of switching circuit problems. He received the M.E.E. degree in 1960. The same year, he joined RCA Surface Communications Laboratories in New York where he has been engaged in the development of digital communication systems. Mr. Charney is a member of the Institute of Radio Engineers.

CARL P. CLASEN received the A.B. in Physics from Washington University, where he was a member of three honorary societies. He initially joined RCA in June 1953, but was shortly called into the military service. After discharge in 1955, he rejoined RCA and entered the Graduate Study Program in September, 1957. He received the M.S. degree in Physics in June, 1959. Since 1957, he has been intimately associated with RCA's applied research program in the field of electronic scanning. He is presently Senior Scientist of that program at the RCA Antenna Skill Center, Missile & Surface Radar Division, Moorestown, New Jersey.

Jacques Dutka received the B.S. degree from the City College of New York in 1939, the A.M. degree in Mathematics from Columbia University in 1940, and the Ph.D. degree in Mathematics from Columbia in 1943. From 1944 to 1946, he was a member of the Operations Research Group of the U.S. Navy, where he made studies on anti-submarine warfare as well as accuracy studies of aerial attacks on shipping. In 1946, he served as the consultant to the office of Naval Research. From 1947 to 1953, Dr. Dutka was an Assistant Professor of Mathematics at Rutgers University, and from 1953 to 1956, he was employed at the Norton-Ketay Corp., where he specialized in military research and development projects. In 1956, he joined the RCA Surfcom Systems Laboratories as a senior engineer, where he has worked principally on theoretical investigations of communications systems problems and as a consultant on mathematical and systems problems associated with communications projects. Since 1954, he has been teaching at Columbia University and is presently an Adjunct Professor of Electrical Engineering. Dr. Dutka is a member of the American Mathematical Society, the Institute of Mathematical Statistics, and Sigma Xi.

Herbert D. Goldman received the B.E.E. from City College, New York, under a New York State Scholarship. During this period, he was elected Vice President of the Student Metropolitan Council of the AIEE-IRE for the New York City-New Jersey area. He received the M.S.E.E. degree from Syracuse University while working for the General Electric Company as a member of their Honors Program. He is now a member of the Communications Research Group at RCA Laboratories, Princeton, N. J. He is presently the recipient of a Fellowship granted by New York University, where he is working for the Ph.D. Mr. Goldman is member of Tau Beta Pi and Eta Kappa Nu.

David M. Jones received the B.A. degree in Mathematics from the University of Oxford in 1955, and the M.A. in 1960. In 1956, he received the Diplome D'Etudes Superieures from Paris University where he did postgraduate research in Statistics at the Institut Henri Poincare. He was employed by Bristol Aircraft, Ltd., from 1956 to 1957 as a structural engineer. In 1957, he joined Avro Aircraft, Ltd., (Canada) where he worked on problems in dynamic response. From 1958 to 1959 he was employed by Remington Rand, Ltd., where he programmed business and scientific problems for Univac II and S.S. 90. He was in charge of operations research. Since joining RCA in 1959, he has done operations research work on information-transmission networks and warfare models and applied research in communication theory.

EDMUND A. LAPORT attended Northeastern University, Massachusetts University Extension, and McGill University. From 1924 to 1933 he was a radio transmitter engineer with Westinghouse Electric & Manufacturing Company at Springfield and Chicopee Falls, Mass. During 1933-34 he was associated with Paul F. Godley Company, Montclair, N.J. From 1934-36 he was a transmission engineer with Wired Radio, Inc., Ampere, N.J. From 1936 to 1938 he was in charge of high-power transmitter engineering for RCA Victor Division, Camden, N.J.; 1938-1944, Chief Engineer of RCA Victor Company, Limited, Montreal; 1944-1954, Chief Engineer of RCA International Division, New York City; from 1954 to date, Director of Communications Engineering for Radio Corporation of America, New York City, and, since March 1959, in Princeton, N.J. He is a Fellow of the Institute of Radio Engineers, and author of "Radio Antenna Engineering" (McGraw-Hill Book Co., 1952).

GILBERT LIEBERMAN received a B.A. degree in Mathematics at New York University in 1948, and an M.A. degree in Mathematical Statistics at Columbia University in 1949. Since then he has done graduate work in Mathematics and Electrical Engineering at University of Maryland, American University, and the University of Pensylvania. From 1950 to 1955, he was with the U.S. Naval Research Laboratory, Sound Division, Electronics Branch, where he worked on sonar signal processing techniques, analysis of acoustic signal and noise propagation data. From 1955 to 1959, he was with the U.S. Naval Ordnance Laboratory, where he worked on acoustic detection, navigation and guidance systems. During this period, he taught courses in probability and statistics at the University of Maryland. He joined RCA in March of 1959 as a Senior Engineer in the Airborne Systems, Communications Systems Group. His work has involved analysis of digital communication systems, studies of effects of antenna fluctuations and fading on communication system performance, and research on adaptive communication techniques. Mr. Lieberman is a member of the IRE professional groups on Communication Systems, Information Theory, the Institute of Mathematical Statistics, the American Statistical Association, and the Mathematical Association of America.

DAVID G. C. LUCK received his B.S. degree in physics from Massachusetts Institute of Technology in 1927 and his Ph.D. in 1932. In 1927-28 he was a Swope Fellow in Physics, and a Malcolm Cotton Brown Fellow in 1928-29. He was an Assistant in the Department of Physics at Massachusetts Institute of Technology in 1929-32. Dr. Luck joined the Research Division of RCA Victor Company in 1932 and remained with the Victor Division of RCA Manufacturing Company until 1941, working on pulse communication, direction finding, and radio guidance of aircraft. He was transferred to RCA Laboratories Division upon its formation in 1941 and remained a Member of the Technical Staff of the Princeton Laboratories through 1953, working on special electronic systems for military and other aircraft and on color television systems. From 1954 to 1958, he served as staff engineer of the Airborne Systems Department of RCA Defense Electronic Products, acting as a technical adviser on a number of military-system proposals and projects. During 1958 and '59, on leave from RCA, Dr. Luck participated in the work of the Advanced Research Projects Agency of the Department of Defense as a member of their technical advisory staff. He is now a member of the Technical Staff of the Advanced Military Systems operation of RCA Defense Electronic Products.

Dr. Luck is a member of the Institute of Navigation and Sigma Xi, and a Fellow of the Institute of Radio Engineers. He is a recipient of the Ballantine Medal of the Franklin Institute and of the Pioneer Award of the IRE Professional Group on Aeronautical and Navigational Electronics.

C. M. MENGANI received the B.S.E.E. degree from Indiana Technical College in 1949. From 1950 to 1955, he was employed by Sperry Gyroscope Company where he was engaged in the design and development of aircraft engine instruments. In 1955 he joined Control Instrument Company as Project Engineer on the design of character recognition systems. Since August 1959 he has been with the Advanced Digital Techniques group of the RCA Surface Communications Systems Laboratories at New York, N. Y. Mr. Mengani is a member of the Institute of Radio Engineers.

SIDNEY METZGER received the degree of B.S. in E.E. from New York University in 1937, and his M.E.E. degree from Polytechnic Institute of Brooklyn in 1950. From 1939 until 1945, he worked for the U. S. Army Signal Corps Laboratories at Fort Monmouth on radio communications equipment. From 1945 to 1954 he was at the ITT Laboratories as division head in charge of commercial and military multiplex microwave ystems. In 1954, Mr. Metzger joined the RCA Laboratories at Princeton where he worked on communications problems for military systems. He transferred to the Astro-Electronics Division when it was formed in 1958 and is now Manager of Communications Engineering. He was in charge of the group which supplied the satellite and ground based radio equipment for Project SCORE (the "Talking Atlas") in 1958; was responsible for the communication system and equipment for Project TIROS; and is now responsible for the communications system and equipment for Project Relay. He is a Senior Member of the Institute of Radio Engineers, a Senior Member of the American Rocket Society, and a member of Tau Beta Pi and Sigma Xi.

ALAN A. MEYERHOFF received the B.S. degree in E.E. in 1946 and the M.S. degree in 1947 from Rutgers University. From 1947 to 1948 he worked in the Research Division of Philco Corporation. He was then employed by the U.S. Army Signal Research and Development Laboratory at Fort Monmouth, where he did systems engineering and special studies in connection with the development of radio-relay equipment. In 1956 he joined the Surface Communications Systems Laboratory of RCA Defense Electronic Products in New York City where he is now leader of a group dealing with transmission systems engineering. Mr. Meyerhoff is a Senior Member of the Institute of Radio Engineers and a member of Phi Beta Kappa and Tau Beta Pi.

ROGER G. OLDEN received his education in Austria where he graduated as a Mechanical Engineer from the Technical University of Vienna. He then worked for ten years in the mechanical engineering department of the city government of Vienna until the Nazi occupation of Austria forced him to leave. He joined RCA Laboratories in 1942 and has since worked in the fields of facsimile and electrophotography.

JAMES C. PENNYPACKER received the B.S. in Electrical Engineering from M.I.T. in 1960. Upon graduation, he joined RCA as a specialized trainee following which he was assigned to the Aerospace Communications and Controls Division in Burlington, Massachusetts. He has been engaged primarily in work dealing with sequential decoding. He is presently on leave of absence from RCA and is studying for the M.S.S.E. degree at Case Institute of Technology. Mr. Pennypacker is a member of the Institute of Radio Engineers.

H. JOHN PRAGER received the B.S.E.E. degree from the University of Vienna in 1938 and the M.S.E.E. degree from the University of Michigan, Ann Arbor, in 1940. From 1940 to 1942, he was associated with the Research Laboratory of the Detroit Edison Company, Detroit, Michigan. In 1943 he joined the Electron Tube Division of RCA in Harrison, N. J., as a production engineer of small power and receiving tubes. In 1944, he transferred to the Receiving Tube Development Group where he served as Design Engineer, Technical Coordinator, and Unit Leader until 1959. His various assignments there included the design of several entertainment-type receiving tubes, as well as premium industrial tubes, gas-filled tubes, and UHF pencil tubes. From 1954 until 1959, he served as a member of the JETEC-12 Committee, and also was an Evening Instructor at the Newark College of Engineering, Newark, N. J. He is presently a member of the staff of RCA Laboratories, Princeton, N. J., working on semiconductor–microwave devices.

J. B. RANKIN received the B.S. degree from Harvard University in 1947, the B.S.E.E. from the University of Delaware in 1948, and the M.S.E. from Princeton University in 1957. Mr. Rankin joined the technical staff of RCA Laboratories in 1950. In March 1961, he transferred to the RCA Missile and Surface Radar Division at Moorestown. He has been engaged in work on antennas, waveguides, and radio propagation; since 1956, he has specialized in work on electronically steerable antennas. Mr. Rankin is a member of the Institute of Radio Engineers.

WALTER R. WADDEN received the B.S. degree in Electrical Engineering from M.I.T. in 1952. In the same year, he joined RCA in Camden where he worked in the Advanced Development Section on the design of data processing equipment. In 1956 he transferred to the RCA Boston Laboratory where he worked on the design of an airborne digital computer, missile checkout equipment and special purpose digital computers. Since joining the Applied Research Group of the Aerospace Communications and Controls Division, Burlington, Mass., in 1960, he has been engaged in the simulation of sequential decoders using a digital computer.

OAKLEY M. WOODWARD, JR. received the degree of Bachelor of Science in Electrical Engineering from the University of Oklahoma in 1938. He joined the Seismograph Service Corporation of Tulsa, Oklahoma, after graduation. During 1941, he was a research engineer at the RCA Manufacturing Company, Camden, N. J. In January 1942, he became a member of the technical staff at RCA Laboratories, Princeton, N. J. In 1961, he transferred to the RCA Antenna Skill Center, Missile & Surface Radar Division, Moorestown, N. J. Mr. Woodward is a Member of Sigma Xi, and a Senior Member of the Institute of Radio Engineers.

HERMAN ZABRONSKY received the B.S. degree in Mathematics from the City College of New York in 1948 and the M.A. degree in Mathematics from the University of Pennsylvania in 1951. While at the University of Pennsylvania he held a teaching assistantship from 1948 to 1951. From 1951 to 1953 he worked as an associate mathematician for Union Carbide, Oak Ridge, Tenn., working on boundary value problems of heat transfer; and from 1953 to 1958 he worked at the Ford Instrument Co. in New York, on reactor theory, stress analysis, and heat transfer. Mr. Zabronsky joined the Surface Communications Division of RCA in New York in 1958 and worked on problems dealing with the transmission of signals in the presence of noise. He is also a visiting instructor at Stevens Institute, Hoboken, New Jersey where he has lectured in graduate courses on numerical analysis, ordinary and partial differential equations. He is a member of the American Mathematical Society and the Society for Industrial and Applied Mathematics.