THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 55 March 1976 Number 3

Copyright © 1976, American Telephone and Telegraph Company. Printed in U.S.A.

Bell System Centennial: 100 Years of Publishing on Telecommunications

This year the United States is celebrating the bicentennial of its founding—200 years of change, crises, and invention. At the beginning of the nation's second century, in March 1876, the "speaking telephone" was invented. With it began 100 years of research and development in telecommunications—a technology whose influence on the growth of the nation and, indeed, of all western nations, has been profound. The continual improvement and expansion of telecommunications has depended on the depth and diversity of technical research and development. And, if research and development is to encompass all the economic and physical requirements of a growing telecommunications system, it must be founded upon an active exchange of ideas.

Inventiveness in any organization engaged in research and development can be ascribed not only to individual intelligence, but also to the participation of its members in the exchange of ideas—among its own people and throughout the scientific community. Indeed, it is apparent that the *process* of exchange is an active element in the development of ideas. For example, it is recorded that Alexander Graham Bell's invention sprang from a mistranslation of a German text. Bell had the impression from Helmholz's book *Sensations of Tone* that the author had telegraphed vowel sounds over a wire. Although Bell later discovered the mistake, the idea led him to a study of electricity.

In this anniversary article, we note that timely, open publication of advances in telephony has a history as old as the telephone itself.

For example, Bell dated his invention February 14, 1876, found it would carry voice on March 10, and described the instrument in his paper "Researches in Telephony," which he read on May 10 at a meeting of the American Academy of Arts and Sciences. The paper was subsequently printed in Volume XII of the Academy's *Proceedings*. (We found Bell's original paper so refreshing as an expression of his thought and technique that we have reprinted it as the first article in this issue.)

Similar in function to professional journals and magazines, professional societies like the Academy are created to be forums for the exchange of scientific and technical information. As electrical engineering was emerging in the late 19th century as a discipline with an increasing volume of specific knowledge, the American Institute of Electrical Engineers was formed. (Bell was a cofounder and was elected president in 1892.) Discussions at professional society meetings included philosophical and operational topics as well as technical matters. For example, the practical viewpoint of today's systems engineering is clearly identifiable in "Telephone Engineering," J. J. Carty's paper in the 1906 *Transactions* of the AIEE. A year later General Carty became AT&T's Chief Engineer.

In 1876, Sir William Thompson—later Lord Kelvin—observed the operation of the telephone at one of Professor Bell's lectures and reported the discovery to a meeting of the British Association. In 1877, Bell went to England and demonstrated his instruments at a meeting of the same association. The idea caught on so rapidly that only ten years later there were 200,000 telephones in operation in England. Today, even before the new technology of lightwave communications has become commercially feasible for telephone signal transmission, the scientific community throughout the world is keeping pace with the most recent developments in the United States through Bell System publications and Bell System patents.

The founders of the telephone industry early established the policy of open publication that has remained a characteristic of the industry. This policy has been based on the protection of proprietary information afforded by the patent systems of the United States and other countries, and often the publication of new technology is to be found in issued patents. This policy of early publication and patenting is frequently a direct stimulus to invention. As a case in point, the rapid evolution of the telephone transmitter in 1877 and 1878 can be traced in the series of inventions by Edison, Hughes, Blake, Berliner, De Jongh, Mix and Genest, and Hunnings.

In the first thirty years of telecommunications development, Bell System workers depended upon such established professional publications as *Science Magazine, The Proceedings of the American Academy of Arts and Sciences,* the *Philosophical Magazine,* and *Silliman's Journal.* But in 1912, the Bell System started the first of a series of company publications with the first issue of the *Western Electric News,* which combined news of employee activities with articles on new research techniques and technical developments.

A decade later, in 1922, the need for a specialized medium of exchange among scientists and engineers in telecommunications, and specifically between the research and development areas in the Bell System and those in industry, academia, and government, was affirmed in the Foreword to the first issue of *The Bell System Technical Journal*: "A casual examination of recent technical literature dealing with electrical communication would show articles which touch upon almost every branch of human activity, which we designate as science. . . . With this intense and growing interest in the proper application of scientific methods to the solution of the problems of electrical communication, it is natural that a widespread desire should have arisen for a technical journal to collect, print, . . . and make readily available the more important articles relating to the field of the communication engineer. These articles are now appearing in some fifteen or twenty periodicals scattered throughout the world. . . . The need already felt for such a journal will grow keener as new developments extend the scope of the art and the specialization of its engineers of necessity increases."

While *The Bell System Technical Journal* became the primary voice of Bell System research and development, several other technically dedicated publications were started: *The Bell Laboratories Record* was established in 1925 and at the present time provides functionally descriptive articles on the discoveries and developments at Bell Laboratories. *The Bell Telephone Quarterly* (1922–1940), established as a medium of information exchange among the telephone companies, was superseded by *The Bell Telephone Magazine* in 1941; and *The Western Electric Engineer* (begun in 1957) contains articles by Western Electric engineers on all phases of engineering in the manufacture of telecommunications equipment. Parallel with the establishment of these source publications, significant technical papers were published in the *Western Electric Reprint* series, begun in 1919, which evolved into *The Bell Telephone System Technical Publications* (Monographs), published until 1967. On the management side of the business, *The Bell Journal of Economics* was begun in 1970 with the object of encouraging scholarly interest and thought in the application of economics, to the study of regulation, firm and market organization, and the study of interdisciplinary issues in law and economics.

Other telecommunications companies also responded to the need for technical information exchange. Publication of *Electrical Communication* was begun in 1922 by the International Western Electric Company and has been continued since 1925 by the International Telephone and Telegraph Corporation. Now published by General Telephone and Electronics, the *GTE Automatic Electric Technical Journal* was begun in 1913 as *Automatic Telephone. Ericsson Technics* was begun in 1933 by the Swedish firm Telefonaktiebolaget L M Ericsson, and the *Philips Telecommunication Review* was established in 1934 by the Philips' Telecommunicatie Industrie B. V., Netherlands.

One of the oldest continuing journals in the field is *Tele*. Published by the Central Administration of Swedish Telecommunications, *Tele*'s origins can be traced back to 1895. *Telephony*, one of the leading U. S. commercial publications in the field, was started in 1901 and was followed by *Telephone Engineer* in 1909 (now *Telephone Engineer and Management*).

With technical and scientific specialization have come journals to embrace each new field, e.g., optics, acoustics, materials, computers, circuit theory, etc. We find that, in the last decade, almost 19,000 papers by Bell System authors were published in these specialized journals and magazines, nationally and internationally.

This fundamental requisite for free exchange of information, which has been evident since the inception of telecommunications, will be equally necessary in the future if the industry is to maintain its scientific, technical, and, in the final analysis, functional integrity. The Bell System through its own publications and contributions to professional societies and technical journals remains dedicated to this principle.

On October 18, 1892, only sixteen years after the invention of his "speaking telephone," Alexander Graham Bell in New York talks to William H. Hubbard in Chicago at the inauguration of the New York-Chicago telephone line.

REPRINT

PROCEEDINGS

OF THE

AMERICAN ACADEMY

OF

ARTS AND SCIENCES.

VOL. XII.

PAPERS READ BEFORE THE ACADEMY.

I.

RESEARCHES IN TELEPHONY.

By A. Graham Bell.

Presented May 10, 1876, by the Corresponding Secretary.

1. It has long been known that an electro-magnet gives forth a decided sound when it is suddenly magnetized or demagnetized. When the circuit upon which it is placed is rapidly made and broken, a succession of explosive noises proceeds from the magnet. These sounds produce upon the ear the effect of a musical note, when the current is interrupted a sufficient number of times per second. The discovery of "Galvanic Music," by Page,* in 1837, led inquirers in different parts of the world almost simultaneously to enter into the field of telephonic research; and the acoustical effects produced by magnetization were carefully studied by Marrian,† Beatson,‡ Gassiot,§ De la Rive,||

* *C. G. Page.* "The Production of Galvanic Music," Silliman's Journ., 1837, XXXII., p. 396; Silliman's Journ., July, 1837, p. 354; Silliman's Journ., 1838, XXXIII., p. 118; Bibl. Univ. (new series), 1839, II., p. 398.

† *J. P. Marrian.* Phil. Mag., XXV., p. 382; Inst., 1845, p. 20; Arch. de l'Électr., V., p. 195.

‡ *W. Beatson.* Arch. de l'Électr., V., p. 197; Arch. de Sc. Phys. et Nat. (2d series), II., p. 113.

§ *Gassiot.* See "Treatise on Electricity," by De la Rive, I., p. 300.

|| *De la Rive.* Treatise on Electricity, I., p. 300; Phil. Mag., XXXV., p. 422; Arch. de l'Électr., V., p. 200; Inst., 1846, p. 83; Comptes Rendus, XX., p. 1287; Comp. Rend., XXII., p. 432; Pogg. Ann., LXXVI., p. 637; Ann. de Chim. et de Phys., XXVI., p. 158.

Matteucci,[*] Guillemin,[†] Wertheim,[‡] Wartmann,[§] Janniar,[‖] Joule,[¶] Laborde,[**] Legat,[††] Reis,[‡‡] Poggendorff,[§§] Du Moncel,[‖‖] Delezenne,[¶¶] and others.[***]

2. In the autumn of 1874, I discovered that the sounds emitted by an electro-magnet under the influence of a discontinuous current of electricity are not due wholly to sudden changes in the magnetic condition of the iron core (as heretofore supposed), but that a portion of the effect results from vibrations in the insulated copper-wires composing the coils. An electro-magnet was arranged upon circuit with an instrument for interrupting the current,—the rheotome being placed in a distant room, so as to avoid interference with the experiment. Upon applying the ear to the magnet, a musical note was clearly perceived, and the sound persisted after the iron core had been removed. It was then much feebler in intensity, but was otherwise unchanged, —the curious crackling noise accompanying the sound being well marked.

The effect may probably be explained by the attraction of the coils of the wire for one another during the passage of the galvanic current, and the sudden cessation of such attraction when the current is interrupted. When a spiral of fine wire is made to dip into a cup of mercury, so as thereby to close a galvanic circuit, it is well known that the spiral coils up and shortens. Ferguson[†††] constructed a rheotome upon this principle. The shortening of the spiral lifted the end of the

[*] *Matteucci.* Inst., 1845, p. 315; Arch. de l'Électr., V., 389.

[†] *Guillemin.* Comp. Rend., XXII., p. 264; Inst., 1846, p. 30; Arch. d. Sc. Phys. (2d series), I., p. 191.

[‡] *G. Wertheim.* Comp. Rend., XXII., pp. 336, 544; Inst., 1846, pp. 65, 100; Pogg. Ann., LXVIII, p. 140; Comp. Rend., XXVI., p. 505; Inst., 1848, p. 142; Ann. de Chim. et de Phys., XXIII., p. 302; Arch. d. Sc. Phys. et Nat., VIII., p. 206; Pogg. Ann., LXXVII., p. 43; Berl. Ber., IV., p. 121.

[§] *Élie Wartmann.* Comp. Rend., XXII., p. 544; Phil. Mag. (3d series), XXVIII., p. 544; Arch. d. Sc. Phys. et Nat. (2d series), I., p. 419; Inst., 1846, p. 290; Monatscher. d. Berl. Akad., 1846, p. 111.

[‖] *Janniar.* Comp. Rend., XXIII., p. 319; Inst., 1846, p. 269; Arch. d. Sc. Phys. et Nat. (2d series), II., p. 394.

[¶] *J. P. Joule.* Phil. Mag., XXV., pp. 76, 225; Berl. Ber., III., p. 489.

[**] *Laborde.* Comp. Rend., L., p. 692; Cosmos, XVII., p. 514.

[††] *Legat.* Brix. Z. S., IX., p. 125.

[‡‡] *Reis.* "Téléphonie." Polytechnic Journ., CLXVIII., p. 185; Böttger's Notizbl., 1863, No. 6.

[§§] *J. C. Poggendorff.* Pogg. Ann., XCVIII., p. 192; Berliner Monatsber., 1856, p. 133; Cosmos, IX., p. 49; Berl. Ber., XII., p. 241; Pogg. Ann., LXXXVII., p. 139.

[‖‖] *Du Moncel.* Exposé, II., p. 125; also, III., p. 83.

[¶¶] *Delezenne.* "Sound produced by Magnetization," Bibl. Univ. (new series), 1841, XVI., p. 406.

[***] See London Journ., XXXII., p. 402, Polytechnic Journ., CX., p. 16; Cosmos, IV., p. 43; Glösener—Traité général, &c., p. 350; Dove, Repert., VI., p. 58; Pogg. Ann., XLIII., p. 411; Berl. Ber., I., p. 144; Arch. d. Sc. Phys. et Nat., XVI., p. 406; Kuhn's Encyclopædia der Physik, pp. 1014–1021.

[†††] *Ferguson.* Proceedings of Royal Scottish Soc. of Arts, April 9, 1866; Paper on "A New Current Interrupter."

wire out of the mercury, thus opening the circuit, and the weight of the wire sufficed to bring the end down again,—so that the spiral was thrown into continuous vibration. I conceive that a somewhat similar motion is occasioned in a helix of wire by the passage of a discontinuous current, although further research has convinced me that other causes also conspire to produce the effect noted above. The extra currents occasioned by the induction of the voltaic current upon itself in the coils of the helix no doubt play an important part in the production of the sound, as very curious audible effects are produced by electrical impulses of high tension. It is probable, too, that a molecular vibration is occasioned in the conducting wire, as sounds are emitted by many substances when a discontinuous current is passed through them. Very distinct sounds proceed from straight pieces of iron, steel, retort-carbon, and plumbago. I believe that I have also obtained audible effects from thin platinum and German-silver wires, and from mercury contained in a narrow groove about four feet long. In these cases, however, the sounds were so faint and outside noises so loud that the experiments require verification. Well-marked sounds proceed from conductors of all kinds when formed into spirals or helices. I find that De la Rive had noticed the production of sound from iron and steel during the passage of an intermittent current, although he failed to obtain audible results from other substances. In order that such effects should be observed, extreme quietness is necessary. The rheotome itself is a great source of annoyance, as it always produces a sound of similar pitch to the one which it is desired to hear. It is absolutely requisite that it should be placed out of earshot of the observer, and at such a distance as to exclude the possibility of sounds being mechanically conducted along the wire.

3. Very striking audible effects can be produced upon a short circuit by means of two Grove elements. I had a helix of insulated copper-wire (No. 23) constructed, having a resistance of about twelve ohms. It was placed in circuit with a rheotome which interrupted the current one hundred times per second. Upon placing the helix to my ear I could hear the unison of the note produced by the rheotome. The intensity of the sound was much increased by placing a wrought-iron nail inside the helix. In both these cases, a crackling effect accompanied the sound. When the nail was held in the fingers so that no portion of it touched the helix, the crackling effect disappeared, and a pure musical note resulted.

When the nail was placed inside the helix, between two cylindrical pieces of iron, a loud sound resulted that could be heard all over a large room. The nail seemed to vibrate bodily, striking the cylindrical pieces of metal alternately, and the iron cylinders themselves were violently agitated.

4. Loud sounds are emitted by pieces of iron and steel when subjected to the attraction of an electro-magnet which is placed in circuit with a rheotome. Under such circumstances, the armatures of Morse-sounders and Relays produce sonorous effects. I have succeeded in rendering the sounds audible to large audiences by interposing a tense membrane between the electro-magnet and its armature. The armature in this case consisted of a piece of clockspring glued to the membrane. This form of apparatus I have found invaluable in all my experiments. The instrument was connected with a parlor organ, the reeds of which were so arranged as to open and close the circuit during their vibration. When the organ was played the music was loudly reproduced by the telephonic receiver in a distant room. When chords were played upon the organ, the various notes composing the chords were emitted simultaneously by the armature of the receiver.

5. The simultaneous production of musical notes of different pitch by the electric current, was foreseen by me as early as 1870, and demonstrated during the year 1873. Elisha Gray,* of Chicago, and Paul La Cour,† of Copenhagen, lay claim to the same discovery. The fact that sounds of different pitch can be simultaneously produced upon any part of a telegraphic circuit is of great practical importance; for the duration of a musical note can be made to signify the dot or dash of the Morse alphabet, and thus a number of telegraphic messages may be sent simultaneously over the same wire without confusion by making signals of a definite pitch for each message.

6. If the armature of an electro-magnet has a definite rate of oscillation of its own, it is thrown bodily into vibration when the interruptions of the current are timed to its movements. For instance, present an electro-magnet to the strings of a piano. It will be found that the string which is in unison with the rheotome included in the circuit will be thrown into vibration by the attraction of the magnet.

Helmholtz,‡ in his experiments upon the synthesis of vowel sounds caused continuous vibration in tuning-forks which were used as the armatures of electro-magnets. One of the forks was employed as a rheotome. Platinum wires attached to the prongs dipped into mercury.

The intermittent current occasioned by the vibration of the fork traversed a circuit containing a number of electro-magnets between the poles of which were placed tuning-forks whose normal rates of vibration were multiples of that of the transmitting fork. All the

* *Elisha Gray.* Eng. Pat. Spec., No. 974. See "Engineer," March 26, 1875.

† *Paul la Cour.* Telegraphic Journal, Nov. 1, 1875.

‡ *Helmholtz.* Die Lehre von dem Tonempfindungen.

forks were kept in continuous vibration by the passage of the interrupted current. By re-enforcing the tones of the forks in different degrees by means of resonators, Helmholtz succeeded in reproducing artificially certain vowel sounds.

I have caused intense vibration in a steel strip, one extremity of which was firmly clamped to the pole of a U-shaped electro-magnet, the free end overhanging the other pole. The amplitude of the vibration was greatest when the coil was removed from the leg of the magnet to which the armature was attached.

7. All the effects noted above result from rapid interruptions of a voltaic current, but sounds may be produced electrically in many other ways.

The Canon Gottoin de Coma,* in 1785, observed that noises were emitted by iron rods placed in the open air during certain electrical conditions of the atmosphere; Beatson† produced a sound from an iron wire by the discharge of a Leyden jar; Gore‡ obtained loud musical notes from mercury, accompanied by singularly beautiful crispations of the surface during the course of experiments in electrolysis; and Page§ produced musical tones from Trevelyan's bars by the action of the galvanic current.

8. When an intermittent current is passed through the thick wires of a Ruhmkorff's coil, very curious audible effects are produced by the currents induced in the secondary wires. A rheotome was placed in circuit with the thick wires of a Ruhmkorff's coil, and the fine wires were connected with two strips of brass (A and B), insulated from one another by means of a sheet of paper. Upon placing the ear against one of the strips of brass, a sound was perceived like that described above as proceeding from an empty helix of wire during the passage of an intermittent voltaic current. A similar sound, only much more intense, was emitted by a tin-foil condenser when connected with the fine wires of the coil.

One of the strips of brass, A (mentioned above), was held closely against the ear. A loud sound came from A whenever the slip B was touched with the other hand. It is doubtful in all these cases whether the sounds proceeded from the metals or from the imperfect conductors interposed between them. Further experiments seem to favor the latter supposition. The strips of brass A and B were held one in each hand. The induced currents occasioned a muscular tremor in the fingers. Upon placing my forefinger to my ear a loud crackling noise

* See "Treatise on Electricity," by De la Rive, I., p. 800.

† Ibid.

‡ *Gore.* Proceedings of Royal Society, XII., p. 217.

§ *Page.* "Vibration of Trevelyan's bars by the galvanic current." Silliman's Journal, 1850, IX., pp. 105–108.

was audible, seemingly proceeding from the finger itself. A friend who was present placed my finger to his ear, but heard nothing. I requested him to hold the strips A and B himself. He was then distinctly conscious of a noise (which I was unable to perceive) proceeding from his finger. In these cases a portion of the induced currents passed through the head of the observer when he placed his ear against his own finger; and it is possible that the sound was occasioned by a vibration of the surfaces of the ear and finger in contact.

When two persons receive a shock from a Ruhmkorff's coil by clasping hands, each taking hold of one wire of the coil with the free hand, a sound proceeds from the clasped hands. The effect is not produced when the hands are moist. When either of the two touches the body of the other a loud sound comes from the parts in contact. When the arm of one is placed against the arm of the other, the noise produced can be heard at a distance of several feet. In all these cases a slight shock is experienced so long as the contact is preserved. The introduction of a piece of paper between the parts in contact does not materially interfere with the production of the sounds, while the unpleasant effects of the shock are avoided.

When a powerful current is passed through the body, a musical note can be perceived when the ear is closely applied to the arm of the person experimented upon. The sound seems to proceed from the muscles of the fore-arm and from the biceps muscle. The musical note is the unison of the rheotome employed to interrupt the primary circuit. I failed to obtain audible effects in this way when the pitch of the rheotome was high. Elisha Gray* has also produced audible effects by the passage of induced electricity through the human body. A musical note is occasioned by the spark of a Ruhmkorff's coil when the primary circuit is made and broken sufficiently rapidly. When two rheotomes of different pitch are caused simultaneously to open and close the primary circuit, a double tone proceeds from the spark.

9. When a voltaic battery is common to two closed circuits, the current is divided between them. If one of the circuits is rapidly opened and closed, a pulsatory action of the current is occasioned upon the other.

All the audible effects resulting from the passage of an intermittent current can also be produced, though in less degree, by means of a pulsatory current.

10. When a permanent magnet is caused to vibrate in front of the pole of an electro-magnet, an undulatory or oscillatory current of electricity is induced in the coils of the electro-magnet, and sounds

* *Elisha Gray.* Eng. Pat. Spec., No. 2646, see "Engineer," Aug. 14, 1874.

proceed from the armatures of other electro-magnets placed upon the circuit. The telephonic receiver referred to above (par. 4), was connected in circuit with a single-pole electro-magnet, no battery being used. A steel tuning-fork which had been previously magnetized was caused to vibrate in front of the pole of the electro-magnet. A musical note similar in pitch to that produced by the tuning-fork proceeded from the telephonic receiver in a distant room.

11. The effect was much increased when a battery was included in the circuit. In this case, the vibration of the permanent magnet threw the battery-current into waves. A similar effect was produced by the vibration of an unmagnetized tuning-fork in front of the electro-magnet. The vibration of a soft iron armature, or of a small piece of steel spring no larger than the pole of the electro-magnet in front of which it was placed, sufficed to produce audible effects in the distant room.

12. Two single-pole electro-magnets, each having a resistance of ten ohms, were arranged upon a circuit with a battery of five carbon elements. The total resistance of the circuit, exclusive of the battery, was about twenty-five ohms. A drum-head of gold-beater's skin, seven centimetres in diameter, was placed in front of each electro-magnet, and a circular piece of clock-spring, one centimetre in diameter, was glued to the middle of each membrane. The telephones so constructed were placed in different rooms. One was retained in the experimental room, and the other taken to the basement of an adjoining house.

Upon singing into the telephone, the tones of the voice were reproduced by the instrument in the distant room. When two persons sang simultaneously into the instrument, two notes were emitted simultaneously by the telephone in the other house. A friend was sent into the adjoining building to note the effect produced by articulate speech. I placed the membrane of the telephone near my mouth, and uttered the sentence, "Do you understand what I say?" Presently an answer was returned through the instrument in my hand. Articulate words proceeded from the clock-spring attached to the membrane, and I heard the sentence: "Yes; I understand you perfectly."

The articulation was somewhat muffled and indistinct, although in this case it was intelligible. Familiar quotations, such as, "To be, or not to be; that is the question." "A horse, a horse, my kingdom for a horse." "What hath God wrought," &c., were generally understood after a few repetitions. The effects were not sufficiently distinct to admit of sustained conversation through the wire. Indeed, as a general rule, the articulation was unintelligible, excepting when familiar sentences were employed. Occasionally, however, a sentence would come out with such startling distinctness as to render it difficult to believe that the speaker was not close at hand. No sound was audible when the clock-spring was removed from the membrane.

The elementary sounds of the English language were uttered successively into one of the telephones and the effects noted at the other. Consonantal sounds, with the exception of L and M, were unrecognizable. Vowel-sounds in most cases were distinct. Diphthongal vowels, such as *a* (in ale), *o* (in old), *i* (in isle), *ow* (in now), *oy* (in boy), *oor* (in poor), *oor* (in door), *ere* (in here), *ere* (in there), were well marked.

Triphthongal vowels, such as *ire* (in fire), *our* (in flour), *ower* (in mower), *ayer* (in player), were also distinct. Of the elementary vowel-sounds, the most distinct were those which had the largest oral apertures. Such were *a* (in far), *aw* (in law), *a* (in man), and *e* (in men).

13. Electrical undulations can be produced directly in the voltaic current by vibrating the conducting wire in a liquid of high resistance included in the circuit.

The stem of a tuning-fork was connected with a wire leading to one of the telephones described in the preceding paragraph. While the tuning-fork was in vibration, the end of one of the prongs was dipped into water included in the circuit. A sound proceeded from the distant telephone. When two tuning-forks of different pitch were connected together, and simultaneously caused to vibrate in the water, two musical notes (the unisons respectively of those produced by the forks) were emitted simultaneously by the telephone.

A platinum wire attached to a stretched membrane, completed a voltaic circuit by dipping into water. Upon speaking to the membrane, articulate sounds proceeded from the telephone in the distant room. The sounds produced by the telephone became louder when dilute sulphuric acid, or a saturated solution of salt, was substituted for the water. Audible effects were also produced by the vibration of plumbago in mercury, in a solution of bichromate of potash, in salt and water, in dilute sulphuric acid, and in pure water.

14. Sullivan* discovered that a current of electricity is generated by the vibration of a wire composed partly of one metal and partly of another; and it is probable that electrical undulations were caused by the vibration. The current was produced so long as the wire emitted a musical note, but stopped immediately upon the cessation of the sound.

15. Although sounds proceed from the armatures of electro-magnets under the influence of undulatory currents of electricity, I have been

* *Sullivan.* "Currents of Electricity produced by the vibration of Metals." Phil. Mag., 1845, p. 261; Arch. de l'Électr., X., p. 480.

unable to detect any audible effects due to the electro-magnets themselves. An undulatory current was passed through the coils of an electro-magnet which was held closely against the ear. No sound was perceived until a piece of iron or steel was presented to the pole of the magnet. No sounds either were observed when the undulatory current was passed through iron, steel, retort-carbon, or plumbago. In these respects an undulatory current is curiously different from an intermittent one. (See par. 2.)

16. The telephonic effects described above are produced by three distinct varieties of currents, which I term respectively intermittent, pulsatory, and undulatory. *Intermittent currents* are characterized by the alternate presence and absence of electricity upon the circuit; *Pulsatory currents* result from sudden or instantaneous changes in the intensity of a continuous current; and *undulatory currents* are produced by gradual changes in the intensity of a current analogous to the changes in the density of air occasioned by simple pendulous vibrations. The varying intensity of an undulatory current can be represented by a sinusoidal curve, or by the resultant of several sinusoidal curves.

Intermittent, pulsatory, and undulatory currents may be of two kinds,—*voltaic*, or *induced;* and these varieties may be still further discriminated into *direct* and *reversed* currents; or those in which the electrical impulses are all positive or negative, and those in which they are alternately positive and negative.

Telephonic effects can be produced by means of currents of electricity, which are	Intermittent.	Voltaic.	Direct (See par. 1, 2, 3, 4, 5, 6).
			Reversed.
		Induced.	Direct.
			Reversed (See par. 8).
	Pulsatory.	Voltaic.	Direct (See par. 9).
			Reversed.
		Induced.	Direct.
			Reversed.
	Undulatory.	Voltaic.	Direct (See par. 11, 12, 13, 15).
			Reversed.
		Induced.	Direct.
			Reversed (See par. 10).

17. In conclusion, I would say that the different kinds of currents described above may be studied optically by means of König's manometric capsule.* The instrument, as I have employed it, consists

* *König*. "Upon Manometric Flames," Phil. Mag., 1873, XLV., No. 297, 298.

simply of a gas-chamber closed by a membrane to which is attached a piece of clock-spring. When the spring is subjected to the attraction of an electro-magnet, through the coils of which a "telephonic" current of electricity is passed, the flame is thrown into vibration.

I find the instrument invaluable as a rheometer, for an ordinary galvanometer is of little or no use when "telephonic" currents are to be tested. For instance, the galvanometer needle is insensitive to the most powerful undulatory current when the impulses are reversed, and is only slightly deflected when they are direct. The manometric capsule, on the other hand, affords a means of testing the amplitude of the electrical undulations; that is, of deciding the difference between the maximum and minimum intensity of the current.

Copyright © 1976 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 3, March 1976
Printed in U. S. A.

Cross Polarization in Reflector-Type Beam Waveguides and Antennas

By M. J. GANS

(Manuscript received October 29, 1975)

Using the paraxial ray approximation, simple formulas for the cross polarization introduced by curved reflectors are developed. In particular, when the reflectors are quadric surfaces of revolution with the center ray of the beam passing through the foci, the maximum cross-polarized field amplitude throughout a gaussian beam, relative to the on-axis copolarized field, is

$$C = \frac{2\xi\kappa_\perp}{\sqrt{e}} \sin\theta_i,$$

where e is the base of the natural logarithm, ξ is the $1/e$ power radius of the beam, $\kappa_\perp$ is the curvature of the reflector perpendicular to the plane of incidence, and θ_i is the angle of incidence. For such reflectors, the beam fields are accurately represented by a superposition of just two gaussian modes for each plane of polarization: the fundamental mode, which corresponds to the co-polarized gaussian beam, and a higher-order mode, which accounts for the cross-polarized field and the amplitude "space" taper. Transformation of a beam through a general sequence of such reflectors is influenced by three factors: the curved reflectors, longitudinal propagation lengths, and rotations of the plane of incidence. The effect of each factor is described by a 4 × 4 matrix relating the input and output gaussian modes. Several typical beam-reflector systems are analyzed by this method. Theoretical cross-polarization patterns are shown to be in accurate agreement with measurements on a symmetrical dual-reflector system.

I. INTRODUCTION

At millimeter wavelengths, normal waveguide losses become too large in many applications. For example, long lengths of waveguide are required in satellite earth stations between the transceiver and the reflector antenna focus. To reduce these losses one may use quasi-optical beams[1] which employ reflectors or lenses for refocusing at various intervals, thereby confining the beam within a geometric tube with no (lossy) guiding walls. Long-focal-length, multiple-reflector antennas (e.g., Cassegrainian and Gregorian antennas) may themselves be thought of in the context of beam waveguides.

In another application, periodically refocussed beams of millimeter or submillimeter wavelength electromagnetic waves might be used[2] as a means of distributing large amounts of information in cities. Such a transmission system is referred to as Hertzian cable.

In the above beam waveguide systems, it is desirable to double the system capacity by transmitting separate signals on each of two orthogonal polarizations (e.g., vertical and horizontal linear polarizations). In such dual-polarization systems, cross-polarization coupling introduced by the refocusers can significantly decrease system performance because of crosstalk between the different signals carried on each of the two polarizations.

The purpose of this paper is to describe simple formulas for computing the cross-polarization coupling introduced by sequences of beam refocusers which consist of quadric reflector surfaces arranged with the beam axis passing through their foci.

II. CROSS POLARIZATION OWING TO REFLECTOR CURVATURE

Consider a beam incident on a flat reflector, as in Fig. 1a. We restrict our attention to beams with narrow angular divergence where the paraxial ray approximation applies so that, for example, the beam field may be described in terms of gaussian beam modes.[3] The paraxial ray approximation applies roughly whenever the 3-dB angular divergence of the beam is less than one radian.

The geometrical optics law of reflection from a perfect conductor is*

$$\hat{e}_r = 2\hat{n}(\hat{n}\cdot\hat{e}_i) - \hat{e}_i, \tag{1}$$

where $\hat{e}_i$ and $\hat{e}_r$ are unit vectors in the direction of the incident and reflected field polarizations, respectively, and $\hat{n}$ is the surface unit normal vector. The caret "^" indicates a unit vector. If the polarization of the incident field is a fixed linear polarization throughout the beam and is perpendicular to the surface normal, then

$$\hat{e}_r = -\hat{e}_i; \tag{2}$$

i.e., the reflected field is also a fixed linear polarization throughout the beam. As expected, a flat plate introduces no cross polarization.

In general, a reflector will be curved with two principal radii of curvature,[4] as shown in Fig. 1b. The surface unit normal vector will no longer be perpendicular to $\hat{e}_i$ at all points. In fact, for small displacements Δx and Δy along the directions of maximum and minimum

* See Ref. 9, Sec. 6.11, for example.

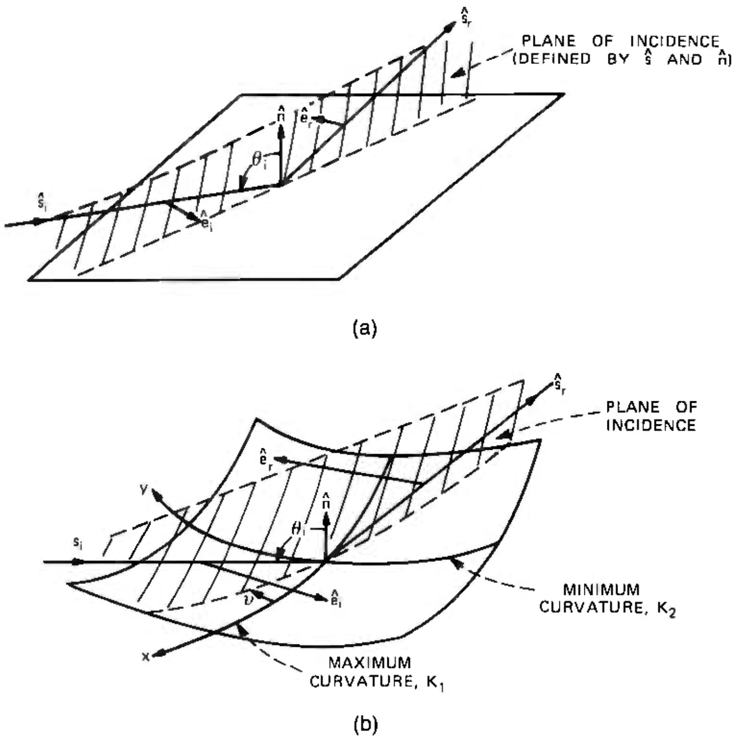


Fig. 1—(a) Reflection of a beam from a flat plate. (b) Reflection of a beam from a curved surface.

curvature, respectively, the unit normal vector will change by

$$\Delta\hat{n} = -\kappa_1(\Delta x)\hat{x} - \kappa_2(\Delta y)\hat{y}, \tag{3}$$

where κ_1 and κ_2 are the maximum and minimum curvatures, respectively, and positive curvature indicates the surface bends toward the incident radiation.

This change in the surface unit normal vector causes the term $(\hat{n}\cdot\hat{e}_i)$ in eq. (1) to change from zero to

$$(\hat{n}\cdot\hat{e}_i) = -\kappa_1(\Delta x)\sin\nu + \kappa_2(\Delta y)\cos\nu, \tag{4}$$

where ν is the angle between the plane of incidence and the direction of maximum curvature as shown in Fig. 1b.

Thus, due to surface curvature, the polarization of the reflected field varies over the surface from that resulting from a flat plate $(-\hat{e}_i)$ by an additional component $2\hat{n}(\hat{n}\cdot\hat{e}_i)$. Part of this component represents the change in the in-line polarization as a consequence of the change in the reflected-ray direction, and part represents cross-polarized signal introduced by the surface curvature. The portion of $\hat{n}$ that is aligned with the cross-polarized field (the field in the plane of incidence and perpendicular to the reflected ray) is of magnitude $\sin \theta_i$. Thus, the ratio of the cross-polarized field to the incident field at a given point is

$$c = 2 \sin \theta_i [-\kappa_1(\Delta x) \sin \nu + \kappa_2(\Delta y) \cos \nu] \tag{5}$$

or

$$c = -2(\Delta\rho) \sin \theta_i \sqrt{(\kappa_1 \sin \nu)^2 + (\kappa_2 \cos \nu)^2} \cos (\phi + \sigma), \tag{6}$$

where

$$\sigma = \arctan \left(\frac{\kappa_2}{\kappa_1} \cot \nu \right).$$

$\Delta\rho$ is the displacement of the reflection point from that of the beam center and ϕ is the angular direction of that displacement relative to the axis of maximum curvature. For a fixed displacement, $\Delta\rho$, the direction, ϕ, that gives maximum cross-polarized signal ratio, c, is $\phi_{\max} = -\sigma$.

If one assumes the incident polarization is in the plane of incidence rather than perpendicular to the plane of incidence, the resulting cross-polarized field is also found to be given by eqs. (5) and (6).

If the incident beam has a gaussian amplitude distribution

$$E_i = E_0 \exp \left\{ -\frac{(\Delta\rho)^2}{2\xi^2} [1 - \sin^2 \theta_i \cos^2 (\phi - \nu)] \right\} \tag{7}$$

(where ξ is the $1/e$ beam intensity radius), one may calculate the ratio of the cross-polarized field relative to the in-line on-axis field (denoted by capital C to differentiate from the lower case c, representing the ratio of in-line and cross-polarized fields at the same point),

$$C = -2(\Delta\rho) \sin \theta_i \sqrt{(\kappa_1 \sin \nu)^2 + (\kappa_2 \cos \nu)^2} \cos (\phi + \sigma) \cdot \exp \left\{ -\frac{(\Delta\rho)^2}{2\xi^2} [1 - \sin^2 \theta_i \cos^2 (\phi - \nu)] \right\}. \tag{8}$$

For a fixed direction ϕ, the radius $\Delta\rho$ at which the relative cross polarization is maximum is

$$\Delta\rho_{C_{\max}} = \frac{\xi}{\sqrt{1 - \sin^2 \theta_i \cos^2 (\phi - \nu)}}, \tag{9}$$

with

$$C_{\max} = -\frac{2\xi \sin \theta_i \sqrt{(\kappa_1 \sin \nu)^2 + (\kappa_2 \cos \nu)^2}}{\sqrt{e}\sqrt{1 - \sin^2 \theta_i \cos^2 (\phi - \nu)}} \cos (\phi + \sigma), \tag{10}$$

and the direction, $\phi_{C_{\max}}$, which provides the greatest cross polarization, is given by

$$\phi_{C_{\max}} = \arctan\left[\frac{\sin^2\theta_i \sin(\nu+\sigma)\cos(\nu+\sigma)}{1-\sin^2\theta_i\sin^2(\nu+\sigma)}\right] - \sigma. \qquad (11)$$

When the plane of incidence coincides with either of the principal curvature planes ($\nu = 0°$ or $90°$), as in the case of quadric surfaces with the beam center ray passing through the surface foci, the expression for the maximum cross polarization of eq. (10) simplifies to

$$\boxed{C_{\max} = \frac{2\xi}{\sqrt{e}}\kappa_\perp \sin\theta_i} \qquad (12)$$

(plane of incidence coincides with either plane of principal curvature), where $\kappa_\perp$ is the curvature in the direction normal to the plane of incidence. Thus, in this case, the maximum cross-polarized field is found in a direction normal to the plane of incidence in the direction of maximum (if $\nu = 0°$) or minimum (if $\nu = 90°$) surface curvature.

In one example, an antenna is formed from two cylindrical mirrors such that $\nu = 0°$ and $\kappa_\perp = 0$ for both mirrors, which by eq. (12) indicates that no cross polarization is generated by the mirrors, in agreement with the results of Ref. 5.

Another example is the offset paraboloid launcher, shown in Fig. 2. The maximum cross-polarization amplitude ratio was derived in Ref. 6 and found to be

$$C_{\max} = \frac{\theta_c \tan(\theta_0/2)}{\sqrt{e}\ln 10}, \qquad (13)$$

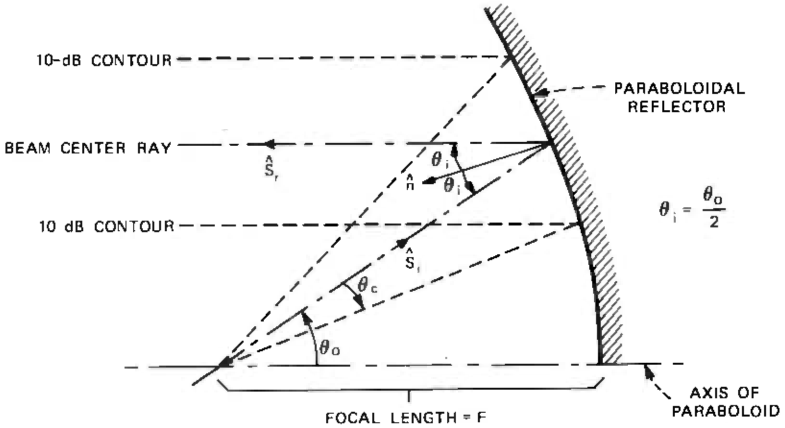


Fig. 2—Offset paraboloid launcher.

where θ_c is the 10-dB half angle of the gaussian beam incident from the focus and θ_0 is the offset angle of the beam center ray.

The $1/e$ beam intensity radius ξ at the paraboloidal reflector is related to the 10-dB half angle, θ_c, by

$$\xi = \frac{\theta_c F \sec^2(\theta_0/2)}{\sqrt{\ln 10}}; \qquad \theta_0 = 2\theta_i, \tag{14}$$

where F is the focal length of the paraboloid. The curvature of the paraboloid in the direction perpendicular to the plane of incidence is

$$\kappa_\perp = \frac{\cos(\theta_0/2)}{2F}. \tag{15}$$

Using eqs. (14) and (15), it is seen that eq. (12) is in agreement with eq. (13).

Another example is the use of cylindrical mirrors in Hertzian cable systems. A typical refocuser mirror arrangement[7] is shown in Fig. 3. The beam remains in a horizontal plane (the plane of incidence) as it is refocused by two cylindrical mirrors both tilted so that their direction of curvature makes an angle $\nu = 50.5$ degrees with the plane of incidence. The output beam has changed direction from the input beam by 225 degrees. The angle of incidence at both mirrors is 33.75 degrees and the curvatures are

$$\kappa_2 = 0,$$

$$\kappa_1 = \frac{1}{66}\text{ meters}^{-1},$$

and the beam radius is

$$\xi = 0.212\text{ meters.}$$

The tilted orientation of the mirrors allows the mirrors to have equal curvature and large aperture efficiency while maintaining sharp focusing and beam symmetry.[7]

The maximum cross polarization for the pair of reflectors is less than twice the maximum cross polarization from either one of the reflectors alone. From eq. (6), σ is zero, and from eq. (11)

$$\phi_{c_{\max}} = 10.515\text{ degrees.}$$

From eq. (10) the maximum cross polarization is

$$20\log_{10}(2C_{\max}) = -48.4\text{ dB.} \tag{16}$$

This is indeed a small value; however, in Hertzian cables with many such refocusers, this cross polarization could accumulate to be a problem.

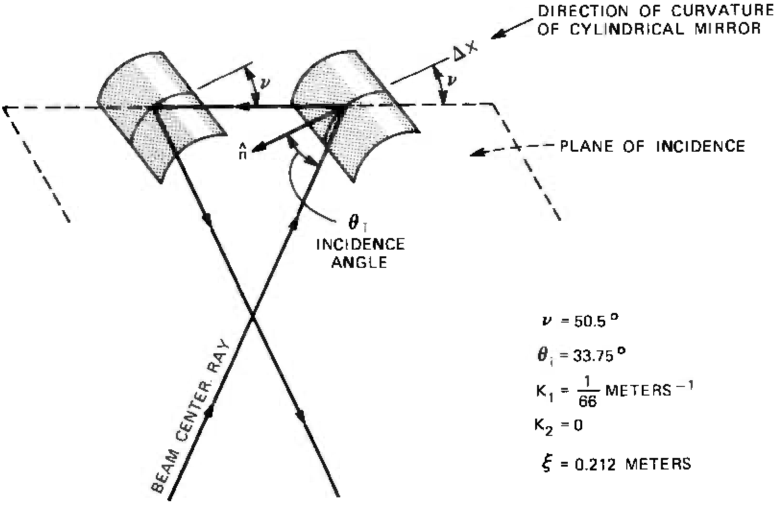


Fig. 3—Typical Hertzian cable refocuser-redirector.

By using geometric optics, eqs. (1) and (7), the cross polarization has been numerically computed for various ellipsoids and paraboloids. The maximum cross polarization was found by a trial-and-error search and compared with that predicted from the paraxial ray result, eq. (12). The comparisons indicate that eq. (12) is accurate to within 0.1 dB for 10-dB half angles of the beam less than 45 degrees.

III. DECOMPOSITION INTO GAUSSIAN BEAM MODES

As described in the previous section, with quadric surface mirrors and the beam center ray passing through the foci of the surfaces, the cross-polarized field resulting from reflection of a perfectly polarized incident gaussian beam is maximum in a direction perpendicular to the plane of incidence and has the maximum value, relative to the in-line polarized field on axis, given by eq. (12), at a distance ξ from the beam center ray.

It is shown in Ref. 6 that this type of reflected field can be represented as the superposition of two gaussian beam modes:[3]

(*i*) Fundamental mode

$$\mathbf{E}_{00} = (H_{00}\hat{x} + V_{00}\hat{y})\,\frac{\sqrt{2\eta}}{\sqrt{\pi}\,\xi_{00}} \cdot \exp\left\{-\,jkz - \frac{\rho^2}{2\xi_{00}^2} + j\left[\arctan\left(\frac{z}{k\xi_{00}^2}\right) - \frac{k\rho^2}{2R_{00}}\right]\right\}, \qquad (17)$$

(*ii*) Higher-order mode

$$\mathbf{E}_{01} = [V_{01}(\hat{x}\cos\alpha - \hat{y}\sin\alpha) - H_{01}(\hat{x}\sin\alpha + \hat{y}\cos\alpha)]\frac{\sqrt{2}\eta\rho}{\sqrt{\pi}\xi_{01}^2}$$

$$\cdot\exp\left\{-jkz - \frac{\rho^2}{2\xi_{01}^2} + j\left[2\arctan\left(\frac{z}{k\bar{\xi}_{01}^2}\right) - \frac{k\rho^2}{2R_{01}}\right]\right\}, \quad (18)$$

where η is the free-space impedance, $\sqrt{\mu_0}/\epsilon_0$, and V_{00}, H_{00}, and V_{01}, H_{01} are the phasor coefficients of the fundamental and higher-order gaussian beam fields for the cases when the incident electric field is in the plane of incidence (V) and perpendicular to the plane of incidence (H), respectively. The subscripts refer to the standard TEM$_{00}$ and TEM$_{01}$ mode notations of Ref. 8, not the "m" and "n" of Ref. 3. $\hat{y}$ and $\hat{x}$ are unit vectors normal to the beam axis in the plane of incidence and normal to the plane of incidence, respectively; ρ, α, and z are cylindrical coordinates, with z denoting distance along the beam axis from the beam waist.

At the beam waist, $z = 0$, the radius of curvature of the phase front of the beam field, R, is infinite, and the field varies with increasing distance, ρ, from the axis at a rate determined by $\bar{\xi}$. For the fundamental mode, the field is maximum on axis and decreases to $1/\sqrt{e}$ of its maximum value at $\rho = \bar{\xi}_{00}$. For the higher-order mode, the field is maximum at $\rho = \bar{\xi}_{01}$ and decreases to $\sqrt{2/e}$ of its maximum value at $\rho = \sqrt{2}\bar{\xi}_{01}$. Away from the beam waist $z \neq 0$, the beam-field amplitude varies with ρ at a rate determined by ξ instead of $\bar{\xi}$, and the phase front has a finite radius of curvature R. ξ and R are determined from $\bar{\xi}$ and z by the following formulas:[8]

$$\xi = \bar{\xi}\sqrt{1 + \left(\frac{z}{k\bar{\xi}^2}\right)^2} \quad (19)$$

and

$$R = z\left[1 + \left(\frac{k\bar{\xi}^2}{z}\right)^2\right]. \quad (20)$$

The choice of eq. (18) as the higher-order mode is based[6] on its ability to approximate simultaneously both the cross-polarization and the "space" taper (amplitude asymmetry from top to bottom of mirror) properties of offset reflectors.

Both modes have a characteristic exponential attenuation with distance from axis, $\exp(-\rho^2/2\xi^2)$, and a spherical wavefront near the axis at constant z, denoted by the term, $\exp(-jk\rho^2/2R)$. As one passes through a beam waist, with increasing z, the on-axis phase advances by π for the fundamental mode and 2π for the higher-order mode (relative to the plane-wave retardation, e^{-jkz}). Thus, if the cross-polarization field (due to the higher-order mode) is in phase with the in-line polarization field at the beam waist, it will be in phase quadrature at large distances, $z \gg k\xi$, from the beam waist.

From the results of Section II and eqs. (17) and (18), we find that, if the higher-order mode is generated by reflection with incidence angle, θ_i, from a quadric surface with curvature, $\kappa_\perp$, perpendicular to the plane of incidence, beam radius, ξ, and reflected phase front radius of curvature R at the reflector, then

$$\xi_{00}(z_r) = \xi_{01}(z_r) = \xi, \tag{21}$$

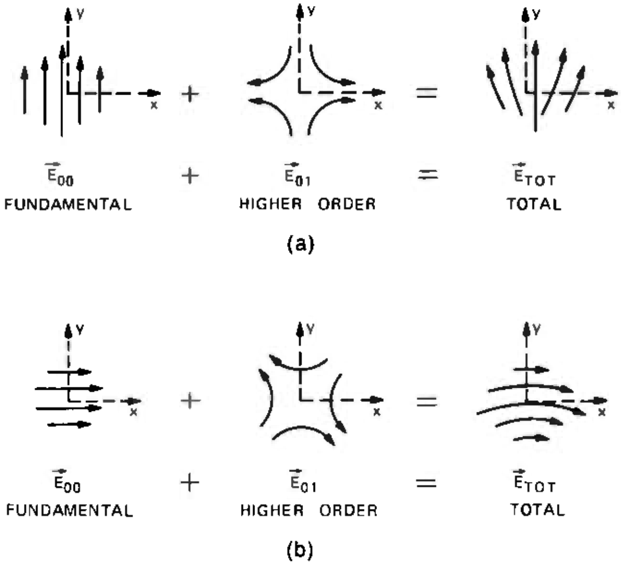
$$R_{00}(z_r) = R_{01}(z_r) = R, \tag{22}$$

and

$$\gamma \triangleq \frac{V_{01}}{V_{00}} = \frac{H_{01}}{H_{00}} = \sqrt{e}C_{\max} = 2\xi\kappa_\perp \sin\theta_i, \tag{23}$$

where the reflector is at $z = z_r$, and the beam waist is at $z = 0$. A picture of a typical aperture-field decomposition into gaussian beam-mode fields is shown in Fig. 4.

Note that, at the reflector z_r, the two modes are in phase with equal beam radii and phase-front curvatures. As one progresses along the beam to an observation point, z_0, the beam radii and phase-front



(a)

(b)

Fig. 4—Two-mode decomposition of aperture field (looking opposite to direction of propagation). (a) Feed horn vertically polarized (parallel to plane of incidence). (b) Feed horn horizontally polarized (perpendicular to plane of incidence).

curvature are still equal

$$\xi_{00}(z_0) = \xi_{00}(z_0) = \bar{\xi}\sqrt{1 + \left(\frac{z_0}{k\bar{\xi}^2}\right)^2} \tag{24}$$

and

$$R_{00}(z_0) = R_{01}(z_0) = z_0\left[1 + \left(\frac{k\bar{\xi}^2}{z_0}\right)^2\right], \tag{25}$$

where $\bar{\xi}$ and z_r are given by[8]

$$\bar{\xi} = \frac{\xi}{\sqrt{1 + (k\xi^2/R)^2}}, \tag{26}$$

and

$$z_r = \frac{R}{1 + (R/k\xi^2)^2}. \tag{27}$$

However, at z_0 there is a relative phase shift between the higher-order mode and the fundamental mode, from eqs. (17) and (18),

$$\Delta\Phi = \Phi_{01}(z_0) - \Phi_{00}(z_0) = \arctan\left(\frac{z_0}{k\bar{\xi}^2}\right) - \arctan\left(\frac{z_r}{k\bar{\xi}^2}\right). \tag{28}$$

This is the relative phase shift near the beam waist mentioned above. When the beam is focusing down towards the beam waist, R and z are negative; when diverging away from the beam waist, R and z are positive.

The power carried by each of the modes in terms of their mode phasors is

$$P = \int_0^\infty \rho d\rho \int_0^{2\pi} d\alpha \frac{|\mathbf{E}|^2}{2\eta} = |A|^2, \tag{29}$$

where A is the phasor of the particular mode in question; i.e., V_{00}, H_{00}, V_{01}, or H_{01}.

IV. MATRIX REPRESENTATION OF BEAM-WAVEGUIDE FACTORS

To keep track of the cross polarization generated by a sequence of factors in a beam-waveguide system, it is useful to represent each factor in terms of its transmission matrix[9] for the fundamental and higher-order modes. We will consider three types of factors that normally affect cross polarization in the reflection process: (*i*) the reflectors per se, (*ii*) the longitudinal propagation length, and (*iii*) the rotation of plane of incidence. See Fig. 5 for an example.

If ξ and R are the same for all modes at the input to a series of reflectors, they remain so throughout the system. Thus, we will assume ξ and R the same for all modes in what follows. If several modes are injected with different pairs of ξ and R, the response to each mode may

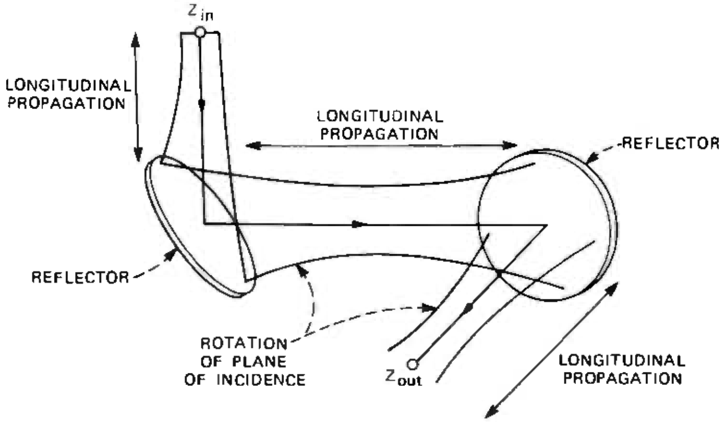
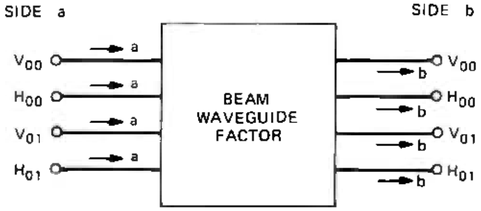


Fig. 5—Factors influencing cross polarization in a reflector-type beam system.

be computed as shown herein and then superposition used to find the total output.

As a dual-mode beam progresses along, undergoing a number of reflections, each factor may be thought of as a reflectionless, passive, eight-port device, as shown in Fig. 6. The coupling between the various modes may be expressed by the matrix equation

$$\mathbf{b} = \boldsymbol{T}\mathbf{a}, \tag{30}$$



$$\vec{b} = T\vec{a}$$

$$\vec{b} = \begin{pmatrix} V_{00b} \\ H_{00b} \\ V_{01b} \\ H_{01b} \end{pmatrix} \qquad \vec{a} = \begin{pmatrix} V_{00a} \\ H_{00a} \\ V_{01a} \\ H_{01a} \end{pmatrix}$$

$$T = \begin{pmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \\ T_{41} & T_{42} & T_{43} & T_{44} \end{pmatrix}$$

Fig. 6—The beam waveguide factor as a reflectionless eight port.

where a is a four vector whose components are the phasors of the input modes,

$$\mathbf{a} = \begin{bmatrix} V_{00a} \\ H_{00a} \\ V_{01a} \\ H_{01a} \end{bmatrix}, \tag{31}$$

and $\mathbf{b}$ is the four vector whose components are the phasors of the output modes

$$\mathbf{b} = \begin{bmatrix} V_{00b} \\ H_{00b} \\ V_{01b} \\ H_{01b} \end{bmatrix}. \tag{32}$$

The properties of the beam factor are described by the four-by-four factor matrix,

$$T = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \\ T_{41} & T_{42} & T_{43} & T_{44} \end{bmatrix}. \tag{33}$$

In general, the matrix T depends on the parameters of the beam propagating through the system. However, it is a simple matter to compute the appropriate matrix for each beam and beam direction one wishes to apply to the system.

4.1 Curved-reflector matrix

To express the beam modes in a form which allows the reflectors to be oriented arbitrarily in space, the beam coordinates at the input and output of a reflector are defined with z in the direction of propagation, y in the plane of incidence perpendicular to z and toward the surface normal, and x normal to z and y (thus normal to the plane of incidence) so that (x, y, z) forms a right-handed cartesian coordinate system, as shown in Fig. 7.

By using the cross-polarization analysis of Section II, the mode definitions of Section III, and conservation of power, the matrix elements applying when a fundamental mode is incident are easily determined:

$$\left.\begin{aligned} &T_{11} = \sqrt{1-\gamma^2}, \quad T_{13} = -\gamma, \quad T_{22} = -\sqrt{1-\gamma^2}, \quad T_{24} = \gamma, \\ &T_{12} = T_{21} = T_{14} = T_{23} = 0, \end{aligned}\right\} \tag{34}$$

where γ is given in eq. (23) as $2\xi\kappa_1 \sin\theta_i$. Note that, for reflectors concave or convex in the direction perpendicular to the plane of incidence, γ is positive or negative, respectively.

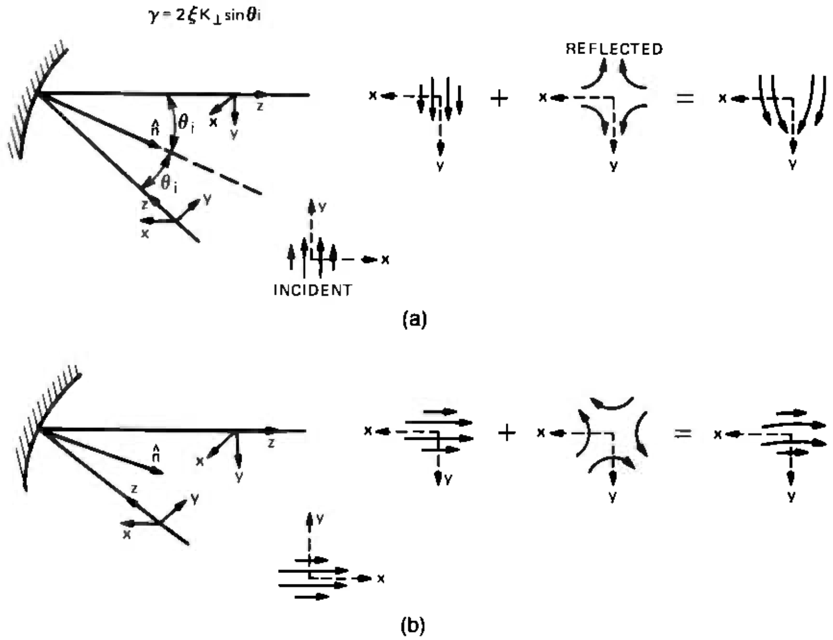


Fig. 7—Reflector matrix components (fields viewed in direction opposite to propagation direction). (a) V_{00} incident. (b) H_{00} incident.

Since the complex conjugate electric field satisfies Maxwell's equations and the boundary conditions on a perfect conductor (time reversal symmetry), the remaining matrix elements follow readily from the above real matrix elements of eq. (34):

$$\left.\begin{array}{l} T_{33} = -\sqrt{1-\gamma^2}, \quad T_{31} = -\gamma, \quad T_{44} = \sqrt{1-\gamma^2}, \quad T_{42} = \gamma, \\ T_{41} = T_{32} = T_{43} = T_{34} = 0. \end{array}\right\} \tag{35}$$

Note that V modes (plane-of-incidence modes) do not couple to H modes (normal-to-plane-of-incidence modes) during reflection from a curved reflector. Thus we have

$$T_{\text{ref}} = \begin{bmatrix} \sqrt{1-\gamma^2} & 0 & -\gamma & 0 \\ 0 & -\sqrt{1-\gamma^2} & 0 & \gamma \\ -\gamma & 0 & -\sqrt{1-\gamma^2} & 0 \\ 0 & \gamma & 0 & \sqrt{1-\gamma^2} \end{bmatrix} \cdot \tag{36}$$

Since the matrix only describes transmission one way, the matrix elements are not necessarily directly related by reciprocity.

4.2 Longitudinal-propagation matrix

As mentioned in Section II, there is a relative phase shift between higher-order modes and their corresponding fundamental modes. In analyzing beam propagation through a system, it is only required to keep track of the relative mode phases to compute the overall cross-polarization coupling. Thus, we will lump all the differential beam-waist phase shifts, $\Delta\Phi$, of eq. (28) with the higher-order modes. As a result, the beam-factor matrix for a longitudinal-propagation length l is

$$T_{lp} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{+j\Delta\Phi} & 0 \\ 0 & 0 & 0 & e^{+j\Delta\Phi} \end{bmatrix}, \tag{37}$$

where

$$\Delta\Phi = \arctan\left(\frac{z_b}{k\xi^2}\right) - \arctan\left(\frac{z_a}{k\xi^2}\right), \tag{38}$$

and z_a and z_b are the positions, relative to the beam waist, of the input and output, respectively.

4.3 Rotation-of-plane-of-incidence matrix

As described in Section 4.1, the beam coordinates are attached to the plane of incidence of each reflector. Thus as one passes from one reflector to another, the plane of incidence may rotate, and what had been a plane-of-incidence mode (V mode) may become a normal-to-plane-of-incidence mode (H mode). From Fig. 4, if one rotates the plane of incidence clockwise by an angle β, the projections of the input modes onto the output modes give the following beam factor matrix for rotation of plane of incidence:

$$T_{\text{rot}} = \begin{bmatrix} \cos\beta & -\sin\beta & 0 & 0 \\ \sin\beta & \cos\beta & 0 & 0 \\ 0 & 0 & \cos 2\beta & -\sin 2\beta \\ 0 & 0 & \sin 2\beta & \cos 2\beta \end{bmatrix}. \tag{39}$$

V. TYPICAL BEAM-WAVEGUIDE APPLICATIONS

In this section, we illustrate the application of the above formulas by considering some typical beam-reflector systems.

5.1 Symmetrical dual reflector

In the symmetrical dual-reflector configuration shown in Fig. 8a, there is no rotation of plane of incidence. The arrangement comprises a curved reflector, followed by a longitudinal propagation length, followed by another reflector. Thus the overall beam system matrix is the product of three beam-factor matrices

$$\boldsymbol{T} = \boldsymbol{T}_{\text{ref }2}\boldsymbol{T}_{lp}\boldsymbol{T}_{\text{ref }1}, \tag{40}$$

where T_{ref} is given by eq. (36) and $\boldsymbol{T}_{lp}$ by eq. (37). Neglecting terms of order γ^2, we have

$$\boldsymbol{T} \doteq \begin{bmatrix} 1 & 0 & -\gamma_1 + e^{j\Delta\Phi}\gamma_2 & 0 \\ 0 & 1 & 0 & -\gamma_1 + e^{j\Delta\Phi}\gamma_2 \\ -\gamma_2 + e^{j\Delta\Phi}\gamma_1 & 0 & e^{j\Delta\Phi} & 0 \\ 0 & -\gamma_2 + e^{j\Delta\Phi}\gamma_1 & 0 & e^{j\Delta\Phi} \end{bmatrix}. \tag{41}$$

From eq. (41), we see that to avoid conversion from a fundamental mode input to a higher-order mode at the output,

$$e^{-j\Delta\Phi} = \frac{\gamma_1}{\gamma_2} = \frac{\xi_1\kappa_{\perp 1}\sin\theta_{i1}}{\xi_2\kappa_{\perp 2}\sin\theta_{i2}}, \tag{42}$$

which implies

$$\xi_1\kappa_{\perp 1}\sin\theta_{i1} = \xi_2\kappa_{\perp 2}\sin\theta_{i2}, \qquad \Delta\Phi = 0 \tag{43}$$

or

$$\xi_1\kappa_{\perp 1}\sin\theta_{i1} = -\xi_2\kappa_{\perp 2}\sin\theta_{i2}, \qquad \Delta\Phi = \pi. \tag{44}$$

Assuming symmetry, $\xi_1 = \xi_2$ and $\theta_{i1} = \theta_{i2}$, and eq. (43) shows that cross polarization is avoided if the two mirrors have equal concave curvature perpendicular to the plane of incidence and are close enough, $\Delta z \ll k\xi^2$, or both far enough to one side or the other of the beam waist so that negligible "beam waist" phase shift takes place. From eq. (44), cross polarization can also be avoided if the reflectors are on opposite sides of the beam waist and in its far field, $\Delta z \gg k\xi^2$, if one reflector is concave and the other convex with equal and opposite curvature normal to the plane of incidence.

Note, from eq. (41), if two identical reflectors are placed symmetrically about the beam waist in the far field, then $\gamma_1 = \gamma_2$ and $\Delta\Phi = \pi$ so the cross-polarization coupling is 6 dB higher than that resulting from just one of the reflectors.

Measurements made by K. C. Kelley[10] on a symmetrical dual-reflector beam-waveguide feed subsystem for a Cassegrainian antenna provide a valuable check on this theory for the combined effect of two of the factors, reflector curvature and longitudinal propagation length. An analysis of his 11-GHz measurements is given in the appendix. The reflectors had equal curvature, the beam size was nearly the same at both curved reflectors, and the relative phase shift was ap-

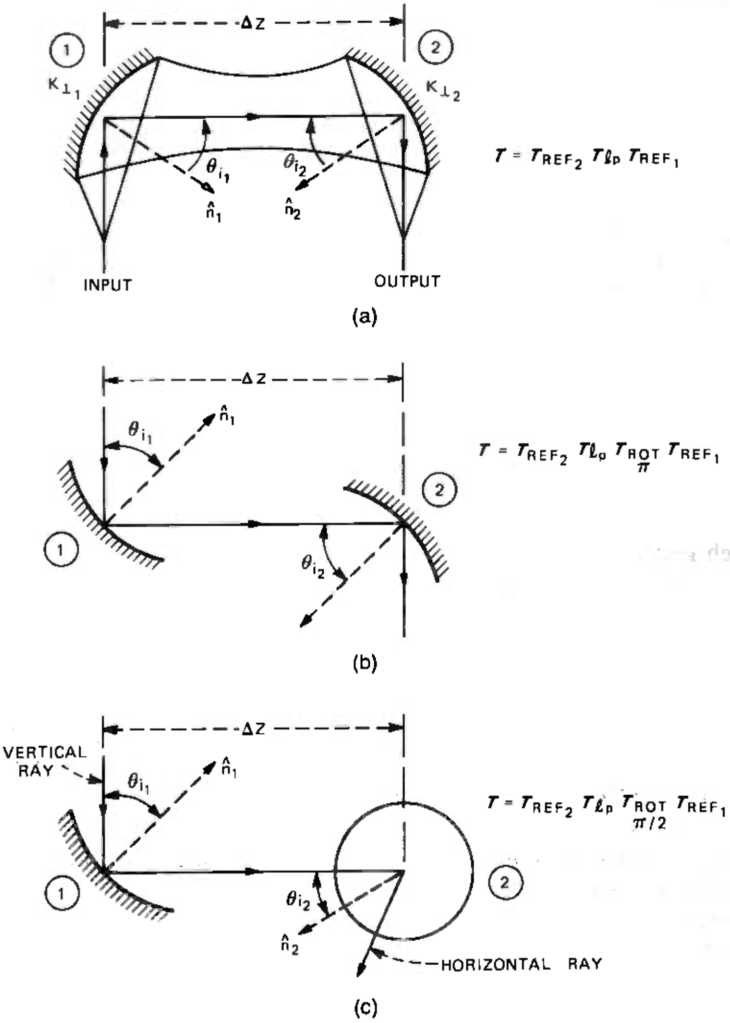


Fig. 8—Some typical beam waveguide applications. (a) Symmetrical dual reflector. (b) Asymmetrical dual reflector. (c) Right-angle dual reflector.

proximately 90 degrees between fundamental and higher-order modes. Thus, from eq. (41), the cross-polarization coupling of the pair at the center frequency is approximately 3 dB higher than that from a single reflector, as confirmed by the measurements. Also, the theoretical frequency dependence of cross-polarization coupling is in approximate agreement with the measurements as shown in the appendix and Fig. 13.

5.2 Asymmetrical dual reflector

In the asymmetrical dual reflector shown in Fig. 8b, the plane of incidence is rotated π radians. The overall matrix is the product of four beam factor matrices,

$$T = T_{\text{ref }2} T_{lp} T_{\text{rot }\pi} T_{\text{ref }1}, \tag{45}$$

where, from eq. (39), since $\beta = \pi$,

$$T_{\text{rot }\pi} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{46}$$

Thus (neglecting terms of order γ^2),

$$T = \begin{bmatrix} -1 & 0 & (\gamma_1 + e^{j\Delta\Phi}\gamma_2) & 0 \\ 0 & -1 & 0 & (\gamma_1 + e^{j\Delta\Phi}\gamma_2) \\ (\gamma_2 + e^{j\Delta\Phi}\gamma_1) & 0 & e^{j\Delta\Phi} & 0 \\ 0 & (\gamma_2 + e^{j\Delta\Phi}\gamma_1) & 0 & e^{j\Delta\Phi} \end{bmatrix}, \tag{47}$$

and the requirement that higher-order modes be avoided is

$$e^{-j\Delta\Phi} = -\frac{\xi_1 \kappa_{11} \sin\theta_{i1}}{\xi_2 \kappa_{12} \sin\theta_{i2}}. \tag{48}$$

Thus the conclusions stated above for the symmetrical dual-reflector configuration with equal (or opposite) curvature on reflectors 1 and 2 apply to the asymmetrical dual-reflector system with opposite (or equal) curvatures on reflectors 1 and 2, respectively.

Note, with closely spaced reflectors in the asymmetrical reflector arrangement ($\Delta z \to 0$, $\theta_{i1} = \theta_{i2}$), a pair of equal curvature mirrors give 6 dB more cross-polarization power coupling than just one of the mirrors, whereas oppositely curved mirrors give cancellation of cross polarization (a well-known property of the Cassegrainian reflector arrangement).

5.3 Right-angle dual reflector

In the right-angle dual reflector shown in Fig. 8c, the plane of incidence is rotated by $\pi/2$ radians. From eq. (39), with $\beta = \pi/2$, the matrix for rotation of the plane of incidence is

$$T_{\text{rot }\pi/2} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \tag{49}$$

and the resulting overall matrix is (neglecting terms of order γ^2)

$$T = \begin{bmatrix} 0 & 1 & -e^{j\Delta\Phi}\gamma_2 & -\gamma_1 \\ -1 & 0 & \gamma_1 & -e^{j\Delta\Phi}\gamma_2 \\ -e^{j\Delta\Phi}\gamma_1 & -\gamma_2 & -e^{j\Delta\Phi} & 0 \\ \gamma_2 & -e^{j\Delta\Phi}\gamma_1 & 0 & -e^{j\Delta\Phi} \end{bmatrix}. \qquad (50)$$

From eq. (50), it is seen that the cross polarization introduced by the first curved reflector cannot be cancelled by the second curved reflector in a right-angle dual-reflector system.

5.4 Confocal beam feed for an offset Cassegrainian antenna

As mentioned in the introduction, an attractive application of beam reflectors is as a feed for a satellite-system ground-station reflector antenna. To show how the above theory may be applied to multiple-reflector antennas, we consider the example of an offset Cassegrainian antenna fed by a beam waveguide. The offset Cassegrainian[11] configuration provides a main reflector aperture with little or no blockage and is shown in Fig. 9 along with a beam reflector feed path from the

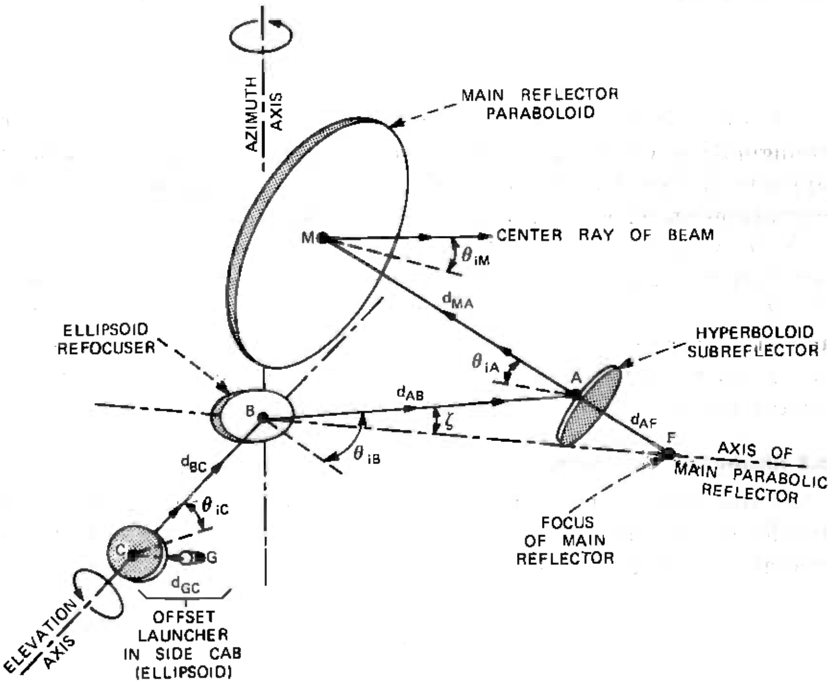


Fig. 9—Beam reflector feed for offset Cassegrainian antenna.

subreflector to a focusing reflector at the elevation axis and on out the elevation axis to an offset launcher[6] in a side cab.

We assume a confocal feed-reflector arrangement as suggested by Arnaud.[12] Subreflector A is an hyperboloid with foci at F and B, refocuser B an ellipsoid with foci at A and C, and the offset launcher reflector is an ellipsoid with foci at G and B. The advantage of this arrangement is that the beam diameters at A and C remain constant with frequency, as do the reflector curvatures, since the beam always seems to originate from the fixed points G, C, B, A, or F. This assumes the feed horn at G has constant beam width and phase center position over the range of frequency variation.

Tracing from G through the beam reflector system, we have the following factors: reflector C, longitudinal propagation length d_{BC}, plane of incidence rotation β_{BC}, reflector B, longitudinal propagation length d_{AB}, plane of incidence rotation $\beta_{AB} = -\pi/2$, reflector A, longitudinal propagation length d_{MA}, plane of incidence rotation $\beta_{MA} = \pi$, and reflector M. Thus the overall matrix

$$\boldsymbol{T} = \boldsymbol{T}_{\mathrm{ref}\,M}\boldsymbol{T}_{\mathrm{rot}\,\pi}\boldsymbol{T}_{lp\,d_{MA}}\boldsymbol{T}_{\mathrm{ref}\,A}\boldsymbol{T}_{\mathrm{rot}\,-\pi/2}\boldsymbol{T}_{lp\,d_{AB}}\boldsymbol{T}_{\mathrm{ref}\,B}\boldsymbol{T}_{\mathrm{rot}\,\beta_{BC}}\boldsymbol{T}_{lp\,d_{BC}}\boldsymbol{T}_{\mathrm{ref}\,C}. \tag{51}$$

Since the cross polarization is small and we may neglect terms of second order ($\gamma^2 \ll 1$), it is easier to add the phasor higher-order mode coupling coefficients as one progresses through the system than to multiply out all the matrices shown in eq. (51).

$$\begin{aligned}\gamma_{VV} = \gamma_{HH} = -\{\gamma_M &+ \exp(j\Delta\Phi_{MA})\gamma_A \\ &+ \sin\beta_{BC}\exp[j(\Delta\Phi_{MA} + \Delta\Phi_{AB} + \Delta\Phi_{BC})]\gamma_C\},\end{aligned} \tag{52}$$

$$\begin{aligned}\gamma_{VH} &= -\gamma_{HV} \\ &= \exp[j(\Delta\Phi_{MA} + \Delta\Phi_{AB})][\gamma_B - \cos\beta_{BC}\exp(j\Delta\Phi_{BC})\gamma_C],\end{aligned} \tag{53}$$

where, for example, γ_{VH} is the "normal to plane of incidence" output higher-order mode, when unit "parallel to plane of incidence" fundamental mode is present at the output, and

$$\Delta\Phi_{BC} = \arctan\left(\frac{z_B}{k\xi_{BC}^2}\right) - \arctan\left(\frac{z_C}{k\xi_{BC}^2}\right), \tag{54}$$

and

$$\gamma_B = 2\xi_B\kappa_{1B}\sin\theta_{iB}, \tag{55}$$

ξ_{BC} being the beam waist radius of the beam traveling from reflector C to reflector B, z_B, and z_C the longitudinal positions of reflectors B and C, respectively, relative to that beam waist, ξ_B the beam radius at reflector B, and κ_{1B} the curvature of reflector B perpendicular to the plane of incidence.

The curvature in the plane perpendicular to the plane of incidence for quadric surfaces of revolution, with beam center rays passing through the foci, may be shown to equal*

$$\kappa_{\perp} = \frac{a \cos \theta_i}{b^2}, \tag{56}$$

where a is the major axis, b the minor axis, and θ_i the angle of incidence of the beam center ray. Using eq. (56) and neglecting diffraction ($\Delta\Phi_{MA} \doteq 0$) between the subreflector A and the main reflector M, one finds that the cross-polarization coupling due to the Cassegrainian combination of M and A is

$$\gamma_{CASS} = \gamma_M + \gamma_A = \gamma_M \frac{\sin^2(\zeta/2)}{\sin^2 \theta_{iM}}, \tag{57}$$

where ζ is the offset angle, relative to the main reflector axis of the beam center ray incident on the subreflector A. Equation (57) is just the result one would obtain from an equivalent parabola[13] with focal length $(e+1)/(e-1)$ times that of reflector M and with beam center ray offset angle ζ.

As frequency decreases from infinity, a beam waist appears on both sides of reflector B; however, $\Delta\Phi_{AB}$ and $\Delta\Phi_{BC}$ remain equal to $\pi/2$.[12] Because of this phase relation, it is not possible to cancel γ_B with γ_C in (53). However, the residual of $\gamma_M + \gamma_A$ in eq. (52) may be cancelled by a special choice of β_{BC}. In fact, if γ_C is adjusted to equal $\gamma_M + \gamma_A$, $\beta_{BC} = \pi/2$ will minimize the cross-polarized modes of both (52) and (53). To maintain $\beta_{BC} = \pi/2$, it is necessary to rotate the offset launcher in Fig. 9 as the antenna is rotated around the elevation axis, just as the fundamental mode polarization at the side cab launcher rotates with antenna elevation angle.

Thus, with $\Delta\Phi_{MA} \doteq 0$, $\gamma_c = \gamma_M + \gamma_A$, and $\beta_{BC} = \pi/2$, we have

$$\gamma_{VV} = \gamma_{HH} \doteq 0 \tag{58}$$

and

$$|\gamma_{VH}| = |\gamma_{HV}| = \gamma_B. \tag{59}$$

From (23) and (25) and $\theta_{iB} = 45°$,

$$\gamma_B = \xi_B \left(\frac{1}{d_{AB}} + \frac{1}{d_{BC}} \right) \tan \theta_{iB} = \frac{1 + d_{AB}/d_{BC}}{k\xi_A}. \tag{60}$$

To satisfy the condition on γ_c, we may choose an incidence angle, θ_{ic}, as follows, from eq. (57),

$$\xi_c \left(\frac{1}{d_{GC}} + \frac{1}{d_{BC}} \right) \tan \theta_{ic} = \gamma_M \frac{\sin^2(\zeta/2)}{\sin^2 \theta_{iM}},$$

* Use the method of Ref. 4, Sec. 19.8, for example.

or

$$\tan \theta_{ic} = \frac{\tan (\zeta/2)}{(d_{BC}/d_{GC}) + 1}. \tag{61}$$

As a specific example to illustrate the cross polarization encountered in practice, consider the following typical antenna dimensions:

$$\frac{d_{AB}}{d_{BC}} = \frac{23}{14}, \quad \frac{d_{BC}}{d_{GC}} = \frac{14}{4}, \quad \zeta = 7.3°, \quad k\xi_A = 4f, \tag{62}$$

where f is the frequency in GHz. Thus, the cross-polarization coupling becomes

$$20 \log_{10} \left(\frac{\gamma_B}{\sqrt{e}} \right) = -8 - 20 \log_{10} f \text{ dB}, \tag{63}$$

e.g., -34 dB at 20 GHz. The incidence angle required on the offset launcher to cancel $\gamma_M + \gamma_A (= -54.3 \text{ dB})$ in (52) becomes 1.6 degrees, which is too small to be practical without blockage, thus other means would be required to reduce γ_C; for example, the launcher could itself be an offset Cassegrainian antenna.

The cross polarization due to the ellipsoid refocuser at B can be reduced by using an additional flat mirror in combination[14] as shown in Fig. 10. With long focal lengths, the beam is essentially of constant width through the combination, and the resultant incidence angle allowing no beam blockage for a beam diameter D depends on the available space h,

$$\theta_{iB} = \frac{1}{2} \arcsin \left(\frac{D}{h} \right). \tag{64}$$

As a specific example, assume there is space available for $D/h = \frac{1}{3}$; whence θ_{iB} is reduced from 45 to 19 degrees and from eq. (60) the cross polarization is reduced 9 dB.

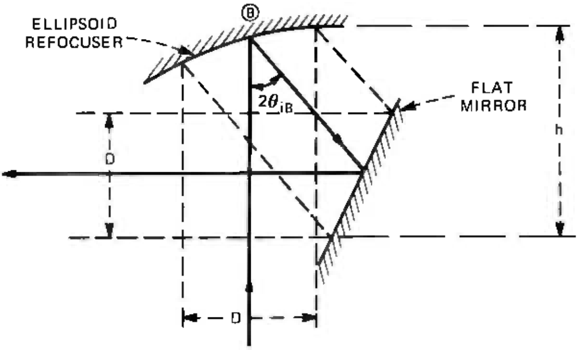


Fig. 10—Combination ellipsoid and flat to reduce θ_{iB}.

VI. CONCLUSIONS

By using the paraxial ray approximation, it has been possible to develop simple formulas for the cross polarization introduced by curved reflectors, e.g., eq. (12). The effect of curved reflectors in a beam-reflector configuration using quadric surfaces of revolution, with the beam center ray passing through the foci, is shown to be accurately characterized by two gaussian modes for each of two planes of polarization. Cross polarization in a general beam-reflector arrangement depends on three factors: reflector curvature, longitudinal propagation length, and rotation of plane of incidence. Using the gaussian modes allows one to represent the effect of the above factors by beam factor matrices which relate the input and output fundamental and higher-order gaussian modes. Some typical beam-reflector configurations were analyzed using these techniques. The theory agrees well with measurements on single reflectors,[6] on a symmetrical dual-reflector system,[10] and with numerical ray tracing computations.

There has been considerable interest in the effect of reflector curvature in beam-reflector configurations. In particular, the work by Mizusawa and Kitsuregawa[15] is worth noting. They show that the symmetric amplitude distribution of an optical beam passing through the foci of a pair of quadric surface-of-revolution reflectors will be preserved if all the foci lie on a straight line and if the eccentricities of the two reflectors are properly related. If both reflectors are ellipsoids or both reflectors are hyperboloids, then the eccentricities must be equal and the exit beam will be parallel to the entrance beam. If one reflector is an ellipsoid and one reflector is an hyperboloid, then one eccentricity must be the inverse of the other eccentricity and the direction of the exit beam is the reflection around the line through the foci of the direction of the entrance beam. By using eq. (12), one can show that only in the case of equal eccentricities does the preservation of amplitude symmetry imply zero cross polarization and then only in the infinite frequency limit where beam waist diffraction is negligible so the relative phase shift between fundamental and higher-order modes is either 0 or 180 degrees.

The frequency dependence of the cross-polarization coupling in a beam-reflector system is an important property not generally indicated in the literature. The paraxial ray approximation for beam diffraction used herein provides a convenient means for computing the frequency dependence of the cross-polarization coupling which, in some cases, can be quite strong; e.g., in eq. (63) for the reflector configuration of Fig. 9 the cross-polarized power varies as the inverse square of the frequency.

VII. ACKNOWLEDGMENTS

It is a pleasure to acknowledge A. A. M. Saleh, D. C. Hogg, and E. A. Ohm for many helpful discussions and J. A. Arnaud for descriptions of Hertzian cable dual-reflector refocusers and confocal triple-reflector arrangements.

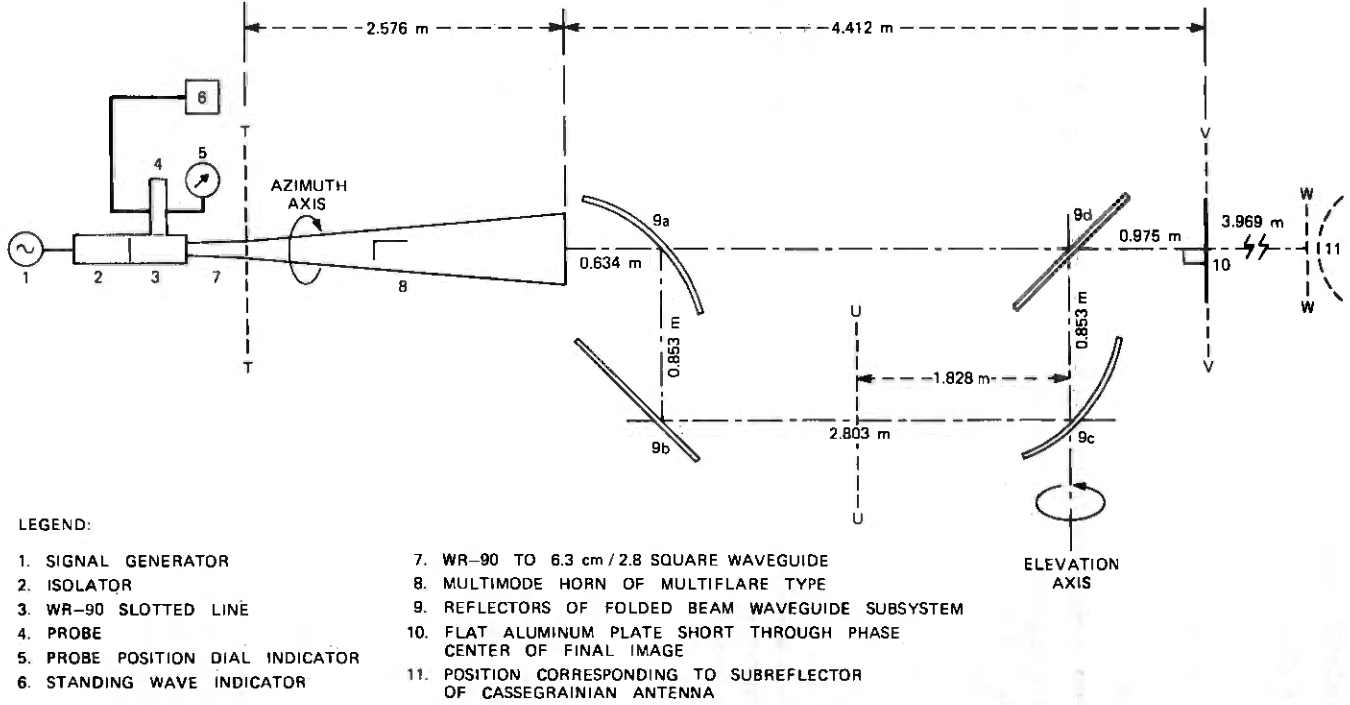
APPENDIX

Comparison of the Matrix Theory With Measurements on a Beam Waveguide

The measurements described in Ref. 10 involved a beam-waveguide feed for a Cassegrainian antenna arranged as shown in Fig. 11. Reflectors 9a and 9c are identical ellipsoids approximately 0.76 m by 1.09 m in size with major and minor axes $a = 3.656$ m and $b = 3.656$ m/$\sqrt{2}$, respectively. Reflectors 9b and 9d are flat mirrors. When one uses the images of the flat mirrors to unfold the beam waveguide, it is seen that a symmetrical dual-reflector type results as shown in Figs. 8a and 12. The feed was designed to produce beam waists approximately at planes *u-u* and *v-v* of Fig. 11 and a beam of the proper diameter and phase curvature at the subreflector position *w-w* to provide a focussed wave reflected from the subreflector toward the main reflector (not shown) of the Cassegrainian antenna. Performance for both vertical and horizontal polarization was measured by rotating the launching horn (No. 8 of Fig. 11) around its axis (azimuth axis). Measurements were made at 10.36, 11.06, and 11.76 GHz for both horn polarizations. The phase and amplitude of the copolarized signal at the subreflector position was measured along the intersection line of plane *w-w* and the beam-bending plane (the plane of the paper in Fig. 11), and the cross-polarized signal (at plane *v-v*) along a line in the beam-bending plane and also along a line perpendicular to the beam-bending plane. The cross-polarized signal along the line in the beam-bending plane always remained below −40 dB relative to the on-axis copolarized signal.

To compare these measurements with theory, the launching horn is assumed to radiate negligible cross-polarized signal and, since measurements of the beam dimensions throughout the reflector system are not available, the beam measurements at the subreflector position (plane *w-w*) will be used to reconstruct the beam dimensions throughout the beam waveguide as shown in the following equations. The theoretical cross polarization will then be computed at plane *v-v* and compared with measurements.

Since the horn did not produce a perfectly symmetrical gaussian beam, the average (over both horn polarizations) of the measurements at planes *w-w* and *v-v* are used in the gaussian beam analysis. From the



LEGEND:

1. SIGNAL GENERATOR
2. ISOLATOR
3. WR–90 SLOTTED LINE
4. PROBE
5. PROBE POSITION DIAL INDICATOR
6. STANDING WAVE INDICATOR
7. WR–90 TO 6.3 cm / 2.8 SQUARE WAVEGUIDE
8. MULTIMODE HORN OF MULTIFLARE TYPE
9. REFLECTORS OF FOLDED BEAM WAVEGUIDE SUBSYSTEM
10. FLAT ALUMINUM PLATE SHORT THROUGH PHASE CENTER OF FINAL IMAGE
11. POSITION CORRESPONDING TO SUBREFLECTOR OF CASSEGRAINIAN ANTENNA

Fig. 11—Experimental setup for measurement of subsystem total loss.

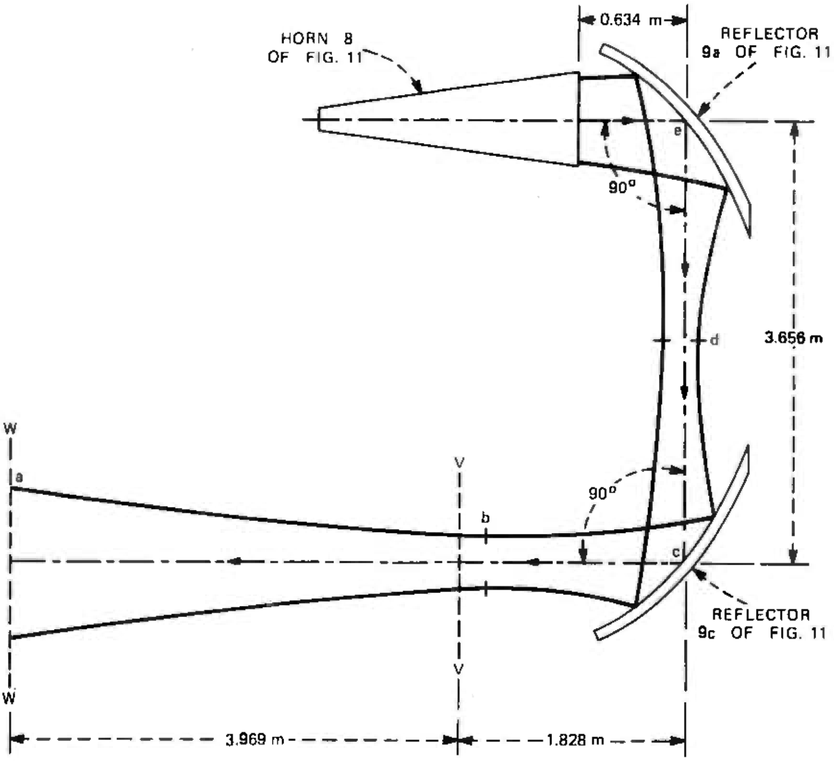


Fig. 12—Unfolded beam waveguide.

measurements at plane *w-w*, the gaussian beam radius ξ_a and phase front radius of curvature R_a at the subreflector position are given in Table I.

Using the beam transformation formulas of Ref. 8, the beam radius, ξ_b, at the beam waist b and the distance from the subreflector to the beam waist z_{ab} are

$$\xi_b = \frac{\xi_a}{\sqrt{1 + (k\xi_a^2/R_a)^2}} \tag{65}$$

Table I — Measured gaussian beam parameters at the subreflector position

Frequency (GHz)	Beam Radius (meters) ξ_a	Phase Front Radius of Curvature (meters) R_a
10.36	0.355	4.177
11.06	0.328	4.623
11.76	0.338	5.027

and

$$z_{ab} = \frac{R_a}{1 + (R_a/k\xi_a^2)^2}, \tag{66}$$

where $k = 2\pi/\lambda$ is the free-space propagation constant.

Going from beam waist b in Fig. 12 to reflector 9b at c, the transformation formulas give the beam parameters on the output side of the reflector,

$$\xi_c = \xi_b \sqrt{1 + \left(\frac{z_{bc}}{k\xi_b^2}\right)^2} \tag{67}$$

and

$$R_{c_{\text{out}}} = z_{bc}\left[1 + \left(\frac{k\xi_b^2}{z_{bc}}\right)^2\right], \tag{68}$$

where z_{bc} is the distance from beam waist b to reflector c in Fig. 12,

$$z_{bc} = 5.797 - z_{ab} \text{ meters.} \tag{69}$$

The radius of curvature of the beam phase front on the input side of reflector c is given by the thin lens formula[9]

$$R_{c_{\text{in}}} = \left[\frac{1}{1.828} - \frac{1}{R_{c_{\text{out}}}}\right]^{-1} \text{ meters,} \tag{70}$$

where 1.828 is the focal length of the ellipsoidal reflectors.

The beam radius ξ_d at the beam waist d and the distance from reflector c to the beam waist z_{cd} are

$$\xi_d = \frac{\xi_c}{\sqrt{1 + (k\xi_c^2/R_{c_{\text{in}}})^2}} \tag{71}$$

and

$$z_{cd} = \frac{R_{c_{\text{in}}}}{[1 + (R_{c_{\text{in}}}/k\xi_c^2)^2]}. \tag{72}$$

The beam radius at reflector e (9a) is

$$\xi_e = \xi_d \sqrt{1 + \left(\frac{z_{de}}{k\xi_d^2}\right)^2}, \tag{73}$$

where the distance from beam waist d to reflector e is

$$z_{de} = 3.656 - z_{cd} \text{ meters.} \tag{74}$$

From Section 5.1, the maximum cross-polarized signal at plane v-v occurs perpendicular to the beam-bending plane at a distance ξ_v from the axis, where ξ_v is the beam radius at plane v-v

$$\xi_v = \xi_b \sqrt{1 + \left(\frac{z_{bv}}{k\xi_b^2}\right)^2}, \tag{75}$$

and the distance from the beam waist b to plane v-v is

$$z_{bv} = z_{ab} - 3.969 \text{ meters.} \tag{76}$$

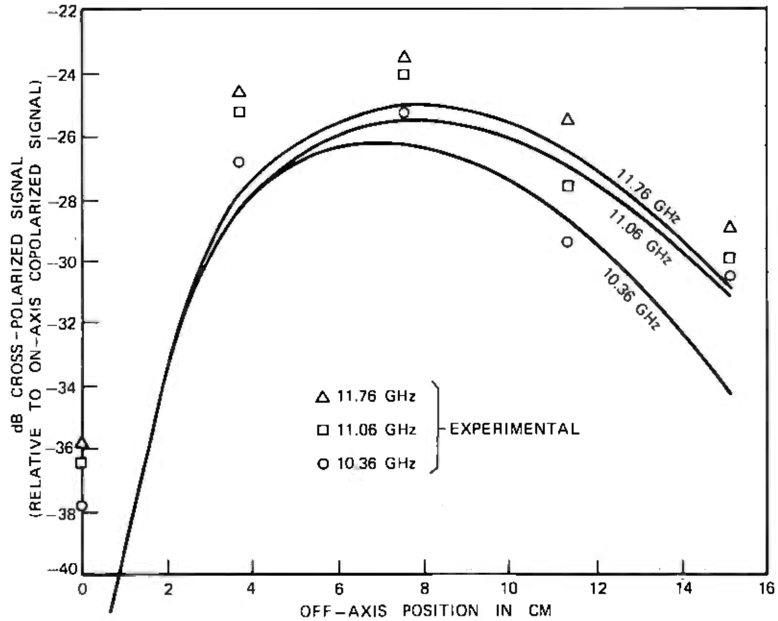


Fig. 13—Comparison of theoretical and measured cross-polarization signal at plane *V-V*.

From eq. (41), the maximum cross-polarized signal amplitude relative to the on-axis copolarized signal is

$$C_{\max} = \frac{-\gamma_c + e^{j\Delta\Phi}\gamma_e}{\sqrt{e}}, \tag{77}$$

where the mode coupling coefficient at reflector c is

$$\gamma_c = 2\xi_c\kappa_{\perp c}\sin\theta_{i_c} = \frac{\xi_c}{1.828}, \tag{78}$$

because the incidence angle $\theta_{i_c} = 45$ degrees and the curvature is $\kappa_{\perp c} = (a/b^2)\cos\theta_{i_c}$ (a and b are the major and minor axes of the ellipsoid, respectively). Similarly,

$$\gamma_e = \frac{\xi_e}{1.828}. \tag{79}$$

$\Delta\Phi$ is the relative phase shift of the higher-order mode relative to the fundamental mode over the longitudinal propagation length between reflectors c and e; from eq. (38),

$$\Delta\Phi = \arctan\left(\frac{|z_{cd}|}{k\xi_d^2}\right) + \arctan\left(\frac{|z_{de}|}{k\xi_d^2}\right). \tag{80}$$

From eq. (18), to find the cross-polarized field at any other radius ρ

instead of $\rho_{c_{\max}} = \xi_v$, one multiplies by the factor:

$$C(\rho) = C_{\max} \left\{ \frac{\rho}{\xi_v} \exp \left[(1 - \rho^2/\xi_v^2)/2 \right] \right\} \cdot \tag{81}$$

Using eqs. (65) through (81) and the values given in Table I, the curves shown in Fig. 13 were computed for the cross-polarized signal power (relative to on-axis copolarized signal) as a function of distance from the axis at plane *v-v* in a direction perpendicular to the beam-bending plane for the three frequencies 10.36, 11.06, and 11.76 GHz. Also shown are the measured values from Ref. 10. The theory is in approximate agreement with the measurements, showing the shape of the curve of cross-polarized signal versus off-axis distance and approximating the absolute level of the maximum cross-polarized signal. The frequency dependence of the theoretical cross-polarized signal is also in the same direction as the measured values.

Theoretically, the cross polarization in the beam-bending plane is negligible, which also is in agreement with the measurements.

REFERENCES

1. D. C. Hogg and R. A. Semplak, "An Experimental Study of Near Field Cassegrainian Antennas," B.S.T.J., *43*, No. 6 (November 1964), pp. 2677–2704.
2. J. A. Arnaud and J. T. Ruscio, "Guidance of 100-GHz Beams by Cylindrical Mirrors," IEEE Trans. on Microwave Theory and Techniques, *MTT-23*, No. 4 (April 1975), pp. 377–379.
3. G. Goubau and F. Schwering, "On the Guided Propagation of Electromagnetic Wave Beams," IRE Trans. on Antennas and Propagation, *AP-9*, No. 3 (May 1961), pp. 248–256.
4. H. S. M. Coxeter, *Introduction to Geometry*, New York: John Wiley, 1969, Secs. 19.4 and 19.5.
5. C. Dragone, "An Improved Antenna for Microwave Radio Systems Consisting of Two Cylindrical Reflectors and a Corrugated Horn," B.S.T.J., *53*, No. 7 (September 1974), pp. 1351–1377.
6. M. J. Gans and R. A. Semplak, "Some Far-Field Studies of an Offset Launcher," B.S.T.J., *54*, No. 7 (September 1975), pp. 1319–1340.
7. J. A. Arnaud and J. T. Ruscio, "Focusing and Deflection of Optical Beams by Cylindrical Mirrors," Appl. Opt., *9*, No. 10 (October 1970), pp. 2377–2380.
8. H. Kogelnik and T. Li, "Laser Beams and Resonators," Appl. Opt., *5*, No. 10 (October 1966), pp. 1550–1567.
9. S. Ramo, J. R. Whinnery, T. VanDuzer, *Fields and Waves in Communication Electronics*, New York: John Wiley, 1965, Sec. 11.09.
10. K. C. Kelley, "Test Data Report for Rantec Model ASF-122 Feed Subsystem," Rantec Proposals No. 62001-TR, April 7, 1972 and No. 6200-TR-1, May 17, 1972.
11. C. Dragone and D. C. Hogg, "The Radiation Pattern and Impedance of Offset and Symmetrical Near-Field Cassegrainian and Gregorian Antennas," IEEE Trans. on Antennas and Propagation, *AP-22*, No. 3 (May 1974), p. 472.
12. J. A. Arnaud, private communication, November 27, 1970.
13. P. W. Hannan, "Antennas Derived from the Cassegrain Telescope," IRE Trans. on Antennas and Propagation, *AP-9*, March 1961, pp. 140–153.
14. D. C. Hogg, private communication, January 30, 1974.
15. M. Mizasawa and T. Kitsuregawa, "A Beam-Waveguide Feed Having A Symmetric Beam for Cassegrain Antennas," IEEE Trans. on Antennas and Propagation, *AP-21*, No. 6 (November 1973), pp. 884–886.

Copyright © 1976 American Telephone and Telegraph Company

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 3, March 1976
Printed in U. S. A.

Jointly Adaptive Equalization and Carrier Recovery in Two-Dimensional Digital Communication Systems

By D. D. FALCONER

(Manuscript received October 22, 1975)

In this paper, we describe a novel receiver structure for two-dimensional-modulated, suppressed-carrier data signals. The receiver consists of a passband equalizer followed by a demodulator which compensates for frequency offset and phase jitter; the demodulator's phase angle is provided by a data-directed, carrier recovery loop, which is shown by analysis and simulation to be capable of tracking relatively high frequency phase jitter. A derivation of the receiver parameters is presented, based on a gradient algorithm for jointly optimizing the equalizer tap coefficients and the carrier phase estimate, to minimize the output mean-squared error. System performance is related to carrier phase-tracking parameters by analysis. Computer simulations confirm the feasibility of the receiver structure.

I. INTRODUCTION

In recent years, a number of double-sideband suppressed-carrier linear-modulation techniques have seen increasing application to the efficient transmission of digital data over band-limited channels. Two-dimensional modulation may be an appropriate designation for these techniques, since they call for coding the transmitted data as two-dimensional data symbols and transmitting the two components by amplitude-modulating two quadrature carrier waves.

Phase-shift keying (PSK) and quadrature amplitude modulation (QAM, sometimes termed QASK), illustrated in Fig. 1, are familar examples. Other two-dimensional modulation examples, characterized by their signal constellations (discrete sets of two-dimensional data symbols), have been extensively studied.[1–3]

This paper presents a unified treatment of adaptive equalization, carrier recovery, and demodulation for two-dimensional-modulated data communication systems. Most previous studies of QAM and PSK systems have treated these receiver functions separately.[4–8] Kobayashi

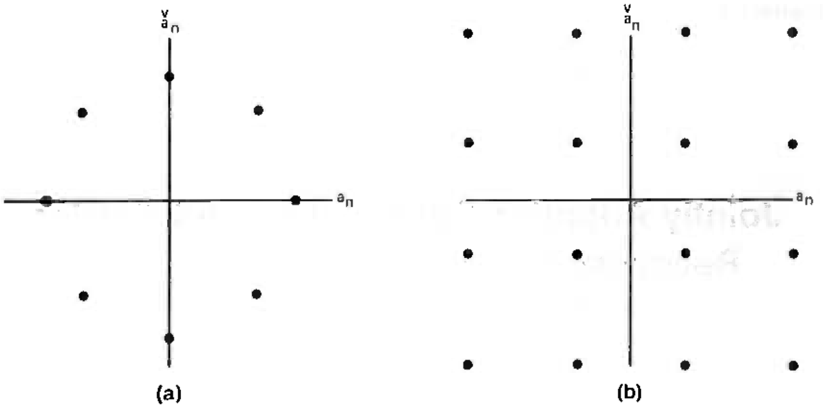


Fig. 1—Examples of two-dimensional signal constellations. (a) 8-phase PSK. (b) 16-point QAM.

presented a unified receiver structure applicable to two-dimensional modulation, based on maximum-likelihood reception.[9] Chang studied a unified linear receiver structure for (one-dimensional) single-sideband modulation systems.[10] A novel feature of the receiver structure presented in this paper is the placement of the carrier phase-tracking and demodulation functions together, after adaptive passband equalization.[†] In a more traditional receiver arrangement,[8–10] baseband equalization follows demodulation and precedes decision-directed phase estimation, thereby introducing a delay of many symbol intervals between these two functions. The decision-directed phase estimate is therefore a *delayed* version of the true channel phase shift affecting the signal that is entering the demodulator. This delay would lead to inaccurate demodulation of a signal perturbed by a time-varying phase shift (phase jitter) introduced by some channels. The receiver structure presented here avoids this source of inaccuracy by placing both the demodulation and phase estimation functions *after* the equalizer.

In Section II we introduce complex notation for two-dimensional bandpass signals and for the effects of linear distortion, phase jitter, and frequency offset. Section III introduces the receiver structure and reviews the function of the passband equalizer. Section IV introduces a mean-squared-error criterion and proposes a gradient algorithm for arriving at a (nonunique) set of equalizer tap coefficients and a carrier phase estimate to minimize it. This ideal gradient algorithm is the motivation for a joint decision-directed equalizer up-

[†] The receiver structure and an equivalent implementation of it are depicted in Figs. 2 and 3, respectively.

dating and demodulation phase-tracking algorithm. It is shown that the phase-tracking algorithm performs essentially the function of a first-order phase-locked loop operating in discrete time. A very simple linear analysis of the loop in Section V illustrates its phase-jitter tracking capability. The receiver's capabilities are further confirmed by the results of simulations, reported in Section VI.

II. BANDPASS SIGNALS AND PHASE JITTER

We consider double-sideband, suppressed-carrier, two-dimensional-modulated data signals specified by

$$s(t) = \sum_n a_n g(t - nT) \cos 2\pi f_c t - \sum_n \check{a}_n g(t - nT) \sin 2\pi f_c t, \qquad (1)$$

where f_c is the carrier frequency, $g(t)$ is a suitably chosen baseband pulse waveform, T is the duration of a symbol interval, and the pair $(a_n, \check{a}_n)$ represents a discrete-valued two-dimensional data symbol. For example, in a 16-point QAM system, each a_n and $\check{a}_n$ is chosen independently from the set $\{\pm 1, \pm 3\}$. In a phase-modulation system (PSK), a_n and $\check{a}_n$ have the form $a_n = \cos \psi_n$ and $\check{a}_n = \sin \psi_n$, the information being coded onto the phase ψ_n. These examples are displayed in Fig. 1.

It is convenient to deal only with the *positive* frequency content of passband spectra. The associated time functions are complex-valued. Thus $s(t) = \mathrm{Re}\,[s(t) + j\check{s}(t)]$, where $\check{s}(t)$ is the Hilbert transform of $s(t)$ and $[s(t) + j\check{s}(t)]$ possesses a Fourier transform consisting of twice the positive frequency part of the spectrum of $s(t)$:

$$s(t) + j\check{s}(t) \equiv \sum_n A_n g(t - nT)\ \exp(j2\pi f_c t), \qquad (2)$$

where $A_n = a_n + j\check{a}_n$. The complex passband waveform $g(t - nT) \times \exp(j2\pi f_c t)$ is said to be *analytic* if its spectrum is nonzero only for positive frequencies. In general, we shall represent real quantities by lower-case letters and complex ones by upper-case letters.

When $s(t)$ is passed through a noisy linear channel, the output is expressed as

$$s'(t) \equiv \mathrm{Re}\,\{\sum_n A_n C(t - nT) \exp[j(2\pi f_c t + \theta)]\} + n(t), \qquad (3a)$$

where $C(t)$ is a complex baseband equivalent impulse response of the combined transmitting filter and channel, θ a phase shift that may be inserted by the channel, and $n(t)$ a realization of additive noise.

Some channels introduce a time-varying phase shift, expressed in general as

$$\theta(t) \equiv \theta + 2\pi\Delta t + \psi(t).$$

Here θ represents a fixed phase shift, Δ a fixed frequency offset, and $\psi(t)$ a random or quasi-periodic waveform that is a manifestation of phase jitter. On voiceband telephone channels, the peak magnitude of the waveform $\psi(t)$ is usually less than about 10 degrees, and its highest frequency spectral component is typically less than 10 percent of the data signal's bandwidth.[11] If the typical rate of variation were comparable to the symbol rate $1/T$, a mathematical model for phase jitter would be critically dependent on the linear filter transfer functions preceding and following the location where the channel phase-modulates the data signal with the phase jitter. However, the assumption of small, relatively slow phase jitter permits us to sidestep this distinction and to model the phase-jitter-perturbed received signal conveniently as

$$s'(t) \equiv \text{Re}\,\{\sum_{n} A_n C(t - nT) \exp[j(2\pi f_c t + \theta_n)]\} + n(t); \qquad (3b)$$

i.e., θ_n is interpreted as the channel phase shift affecting the transmission of the nth data symbol A_n.

III. RECEIVER STRUCTURE

Figure 2 shows the two-dimensional receiver structure. The real-valued received waveform $s'(t)$ first enters a phase splitter, consisting of parallel passband filters with impulse responses $h(t)$ and $\check{h}(t)$, where $\check{h}(t)$ is the Hilbert transform of $h(t)$; thus the complex impulse response defined by $H(t) \equiv h(t) + j\check{h}(t)$ is analytic. An appropriate choice for $H(t)$ is a filter matched to the transmitted pulse, i.e.,

$$H(t) = g(-t) \exp(j2\pi f_c t). \qquad (4)$$

If the channel $C(t)$ were known *a priori*, an optimal choice for $H(t)$ would be a matched filter impulse response

$$C(-t)^* \exp(j2\pi f_c t).$$

The optimality of the complex matched filter and sampler for two-dimensional modulation is brought out in the studies of Kobayashi,[9]

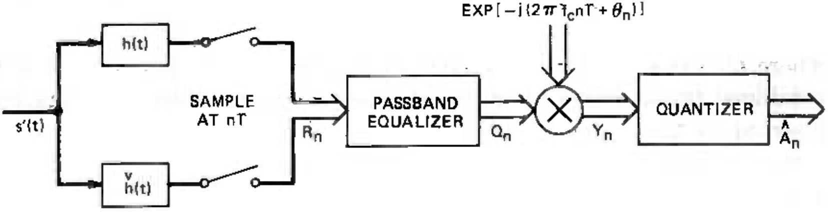


Fig. 2—Two-dimensional receiver.

Ungerboeck,[12] and Ericson and Johansson.[13] Sampling is carried out at the symbol rate $1/T$. We assume a fixed choice of sampling phase and will not be concerned with its optimization. The problem of deriving the optimum sampling phase has been treated previously.[14–16]

The pair of outputs at time nT from the sampler r_n and $\check{r}_n$ can be expressed as a complex sample

$$R_n \equiv r_n + j\check{r}_n,$$

which is of the form

$$R_n = \sum_k A_k X_{n-k} \exp[j(2\pi f_c nT + \theta_k)] + N_n, \tag{5}$$

where $X_n \equiv X(nT)$ is a sample of the overall complex baseband equivalent impulse response and N_n is a complex sample of filtered noise.

The passband linear equalizer[7] with, say, $2M + 1$ complex tap coefficients $\{C_n^*\}_{-M}^{M}$ produces complex passband output samples $\{Q_n\}$ which are a linear combination of sampled inputs; i.e.,

$$Q_n \equiv q_n + j\check{q}_n = \sum_{k=-M}^{M} C_k^* R_{n-k}. \tag{6a}$$

Note that the equalizer's implementation is described either by the above complex expression or by two expressions for the two real outputs, viz.,

$$q_n = \sum_{k=-m}^{M} (c_k r_{n-k} + \check{c}_k \check{r}_{n-k}) \tag{6b}$$

$$\check{q}_n = \sum_{k=-M}^{M} (c_k \check{r}_{n-k} - \check{c}_k r_{n-k}), \tag{6c}$$

where

$$C_k^* \equiv c_k - j\check{c}_k.$$

Expression (6) can also be expressed in vector notation. Define

$$\mathbf{C} \equiv \begin{pmatrix} C_M \\ \vdots \\ C_{-M} \end{pmatrix} \tag{7a}$$

$$\mathbf{R}_n = \begin{pmatrix} R_{n-M} \\ \vdots \\ R_{n+M} \end{pmatrix}. \tag{7b}$$

Then

$$Q_n = \mathbf{C}^*\mathbf{R}_n, \tag{8}$$

where $*$ means complex conjugate transpose.

Ideally, the passband equalizer's sampled impulse response $\{C_k^*\}$ should be such as to yield an overall passband channel with no intersymbol interference; i.e.,

$$\text{ideal } Q_n = A_n \exp[j(2\pi f_c nT + \theta_n)].$$

The information symbol A_n is then recovered by demodulating Q_n to baseband and quantizing the result in accordance with the two-dimensional signal constellation. If the demodulator has a phase estimate $\hat{\theta}_n$, the complex demodulated output is given by

$$Y_n \equiv y_n + j\check{y}_n = Q_n \exp[-j(2\pi f_c nT + \hat{\theta}_n)] \tag{9a}$$

or

$$y_n = q_n \cos(2\pi f_c nT + \hat{\theta}_n) + \check{q}_n \sin(2\pi f_c nT + \hat{\theta}_n) \tag{9b}$$

$$\check{y}_n = -q_n \sin(2\pi f_c nT + \hat{\theta}_n) + \check{q}_n \cos(2\pi f_c nT + \hat{\theta}_n). \tag{9c}$$

The ideal output at time nT is A_n and the receiver *error* is defined by

$$E_n \equiv Y_n - A_n. \tag{10a}$$

For the joint optimization of the equalizer tap coefficients and the demodulator phase, we adopt the following mean-squared-error criterion: minimize ϵ_n, where

$$\epsilon_n \equiv \langle |E_n|^2 \rangle \tag{10b}$$

and the expectation, denoted by $\langle\ \rangle$, is over the data sequence and noise.[†]

The receiver structure shown in Fig. 2 is characterized by the following expression for the complex output sample before quantization. From (6a) and (9a),

$$Y_n = \left[\sum_{k=-M}^{M} C_k^* R_{n-k}\right] \exp[-j(2\pi f_c nT + \hat{\theta}_n)]. \tag{11}$$

An alternative equivalent receiver has a "baseband" structure. Define a new set of tap coefficients by

$$C_k^{*\prime} \equiv C_k^* \exp(-j2\pi f_c kT) \tag{12a}$$

and a set of demodulated received samples by

$$R_n' \equiv R_n \exp(-j2\pi f_c nT). \tag{12b}$$

[†] The "symmetric" mean-squared-error criterion (10b) was proposed by R. D. Gitlin and K. H. Mueller as an improvement to the "single-sided" criterion proposed in Ref. 7.

Then (11) can be re-expressed as

$$Y_n = \left[\sum_{k=-M}^{M} C_k^{*\prime} R'_{n-k} \right] \exp(-j\hat{\theta}_n). \tag{12c}$$

The fully equivalent implementation expressed by (12c) is depicted in Fig. 3. Note that the received samples are demodulated to baseband using a free-running oscillator as in (12b) *before* equalization. However, a second stage of demodulation following baseband equalization remains, whose purpose is to remove the effects of channel phase variation. Again, the delay of the equalizer does not come between this secondary demodulation and the derivation of the phase estimate $\hat{\theta}_n$. The equivalence of the "passband" and "baseband" receiver implementations of Figs. 2 and 3, respectively, gives the system designer some extra flexibility.

IV. OPTIMIZATION OF EQUALIZER TAP COEFFICIENTS AND DEMODULATION PHASE

To bring out the relationships governing the optimal tap vector $\mathbf{C}_n$ and demodulator phase $\hat{\theta}_n$ (both of which may be functions of time), we assume that successive data symbols are uncorrelated; i.e.,

$$\begin{aligned} \langle A_l A_m \rangle &= 0 \qquad \text{all} \quad l, m \\ \langle A_l A_m^* \rangle &= \langle |A|^2 \rangle \delta_{lm}, \end{aligned} \tag{13}$$

where δ_{lm} is the Kronecker delta function. Then for future reference we note, from expressions (5) and (7b), that cross correlation of the data symbols with sampled phase-splitter outputs results in

$$\langle A_n^* \mathbf{R}_n \rangle = \mathbf{X}_n \exp[j(2\pi f_c nT + \theta_n)] \langle |A|^2 \rangle, \tag{14}$$

where

$$\mathbf{X} \equiv \begin{pmatrix} X_{-M} \exp(-j2\pi f_c MT) \\ \vdots \\ X_M \exp(j2\pi f_c MT) \end{pmatrix} \tag{15}$$

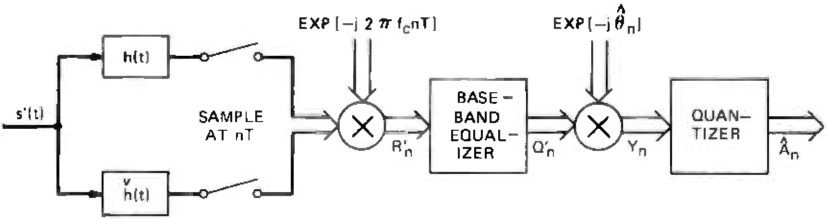


Fig. 3—Equivalent implementation.

is the complex impulse response vector of the combination of the transmitter pulse filter and the channel, truncated to $2M + 1$ samples. The channel correlation matrix or A matrix is defined to be

$$A \equiv \frac{\langle \mathbf{R}_n \mathbf{R}_n^* \rangle}{\langle |A|^2 \rangle}. \tag{16}$$

This is a Hermitian matrix ($A^* = A$) whose l-mth element is

$$A_{lm} = \sum_n X_n X_{n+m-l}^* \exp[j2\pi f_c(l - m)T] + \rho_{l-m}, \tag{17}$$

where $\{\rho_{l-m}\}$ is the noise autocorrelation. Furthermore, it is positive semidefinite. (For any vector $\mathbf{u}$, $\mathbf{u}^* A \mathbf{u} = \langle |\mathbf{u}^* \mathbf{R}_n|^2 \rangle \geqq 0$.)

Using definitions (10), (11), (14), and (16), we can rewrite ϵ_n in terms of A and $\mathbf{X}$, which are fundamental characteristics of the channel.

$$\begin{aligned}\epsilon_n = \{\mathbf{C}_n - A^{-1}\mathbf{X} \exp[-j(\hat{\theta}_n - \theta_n)]\}^* \\ \cdot A\{\mathbf{C}_n - A^{-1}\mathbf{X} \exp[-j(\hat{\theta}_n - \theta_n)]\} + 1 - \mathbf{X}^* A^{-1}\mathbf{X}.\end{aligned} \tag{18}$$

Because the matrix A is positive semidefinite, ϵ_n has the unique minimum

$$\epsilon_{\min} = 1 - \mathbf{X}^* A^{-1}\mathbf{X}, \tag{19}$$

which is achieved when $\mathbf{C}_n$ and $\hat{\theta}_n$ satisfy

$$\mathbf{C}_n = \mathbf{C}_{n\,\mathrm{opt}}(\hat{\theta}_n) \equiv A^{-1}\mathbf{X} \exp[-j(\hat{\theta}_n - \theta_n)]. \tag{20}$$

Observe that the solution (20) is not unique; there is an infinitude of combinations $(\mathbf{C}_n, \hat{\theta}_n - \theta_n)$ that yield the minimum. However, for any specific choice of $\hat{\theta}_n$ (including zero), there is a unique optimum choice of $\mathbf{C}_n$. Indeed, this is a manifestation of the "tap-rotation" property of the passband equalizer which was pointed out by Gitlin, Ho, and Mazo.[7] In particular, when there is no attempt to estimate $\theta_n (\hat{\theta}_n = 0)$, then any amount of frequency offset $\Delta (\theta_n = 2\pi n \Delta T)$ causes $\mathbf{C}_{n\,\mathrm{opt}}$ to "rotate" with frequency Δ. However, a typical adaptive equalizer whose tap coefficients may not be permitted to change by more than about 1 percent from one symbol interval to the next will not be able simultaneously to equalize the channel effectively and to rotate $2\pi\Delta$ radians per symbol interval even for moderate amounts of frequency offset. Similarly, the equalizer taps could not be expected to track typical phase jitter components accurately.

The principal innovation reported in this paper is the joint operation of the adaptive equalizer and a separate phase-tracking loop which removes the major burden of tracking from the slowly adapting equalizer. Assuming this separate phase-angle-tracking algorithm is successful so that the phase error $(\hat{\theta}_n - \theta_n)$ remains constant, we observe, by writing the mean-squared error using definitions (10) and (11) as

$$\epsilon_n \equiv \frac{1}{\langle |A|^2 \rangle} \langle | \mathbf{C}_n^* R_n - A_n \exp[j(2\pi f_c nT + \hat{\theta}_n)] |^2 \rangle, \tag{21}$$

that, if the equalizer's reference signal for the purpose of adapting its tap coefficients is $\{A_n \exp[j(2\pi f_c nT + \hat{\theta}_n)]\}$, the reference signal rotates *in synchronism* with the frequency-offset and phase-jittered carrier of the received signal, and hence the equalizer tap coefficients do not have to rotate if $\theta_n - \hat{\theta}_n$ remains constant.

If the gradients of ϵ_n with respect to the real tap coefficient vectors $\mathbf{c}_n$ and $\check{\mathbf{c}}_n$ are denoted respectively by $\nabla_{\mathbf{c}_n}\epsilon_n$ and $\nabla_{\check{\mathbf{c}}_n}\epsilon_n$ and if we define the gradient with respect to $\mathbf{C}_n$ to be

$$\nabla_{\mathbf{C}_n}\epsilon_n \equiv \nabla_{\mathbf{c}_n}\epsilon_n + j\nabla_{\check{\mathbf{c}}_n}\epsilon_n,$$

then the gradient of the right-hand side of (18) can be written

$$\nabla_{\mathbf{C}_n}\epsilon_n = 2\{A\mathbf{C}_n - \mathbf{X}\exp[-j(\hat{\theta}_n - \theta_n)]\}. \tag{22}$$

Then it follows from expression (18) and from the fact that A is positive semidefinite that $\nabla_{\mathbf{C}_n}\epsilon_n = 0$ is a necessary and sufficient condition for ϵ_n to attain its minimum value.* If the receiver knew $\mathbf{X}\exp(j\theta_n)$, defined by (15), and A, defined by (16), and could calculate this gradient during each symbol interval, then in the nth symbol interval it could use a gradient algorithm to update its estimate of $\mathbf{C}_n$ as follows:

$$\mathbf{C}_{n+1} = \mathbf{C}_n - \frac{\beta}{2}\nabla_{\mathbf{C}_n}\epsilon_n, \tag{23}$$

where the gradient is defined by (22). In this equation, $\mathbf{C}_n$ is the estimate of the correct tap coefficient vector in the nth symbol interval and $\beta/2$ is a positive constant. For the moment, we defer consideration of a more realistic algorithm that does not require prior knowledge of A and $\mathbf{X}$.

Let us now consider the means for providing the estimated sequence $\{\hat{\theta}_n\}$. In general, of course, the true phase jitter angle sequence $\{\theta_n\}$ is a random process. However, the reasonable assumption that it varies slowly with n leads us to treat θ_n as a quasi-static parameter that must be estimated in symbol interval n from present and past received data $\{R_n\}$ and reference information symbols $\{A_n\}$.

Accordingly, the receiver will incorporate an algorithm for updating its estimate $\hat{\theta}_n$, based on a gradient search technique. The derivative

* We assume that matrix A is nonsingular.

of ϵ_n with respect to $\hat{\theta}_n$ is, for a fixed value of $\mathbf{C}_n$,

$$\nabla_{\hat{\theta}_n}\epsilon_n = -2\,\mathrm{Im}\,\{\mathbf{C}_n^*\mathbf{X}\exp[-j(\hat{\theta}_n - \theta_n)]\}. \tag{24}$$

The estimate $\hat{\theta}_n$ is thus updated as

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{\alpha}{2}\nabla_{\hat{\theta}_n}\epsilon_n, \tag{25}$$

where $\alpha/2$ is a constant. In general, α should be large relative to the equalizer's constant β, to ensure that the estimate $\hat{\theta}_n$ can closely track a varying angle θ_n, thereby obviating the need for the passband equalizer taps to follow it closely.

Suppose the angle θ_n is not time-varying ($\theta_n = \theta$). Then the stationary points of the gradient algorithms (23) and (25) are the solutions of the equations

$$\nabla_{\mathbf{C}}\epsilon_n = 0$$

or

$$A\mathbf{C} = \mathbf{X}\exp[-j(\hat{\theta} - \theta)], \tag{26}$$

and

$$\nabla_{\hat{\theta}}\epsilon_n = 0$$

or

$$\mathrm{Im}\,\{\mathbf{C}^*\mathbf{X}\exp[-j(\hat{\theta} - \theta)]\} = 0. \tag{27}$$

It is easy to show from the Hermitian property of A that, if (26) is true, then (27) is true. Furthermore, A is positive semidefinite and thus expression (18) for the mean-squared error shows that the infinite set of stationary points, defined by (26), are the only global minima.

The following question immediately arises: Starting with fixed initial values, $\mathbf{C}_0$ and $\hat{\theta}_0$ and assuming $\theta_n = \theta$ for all n, do the gradient algorithms (23) and (25) jointly converge to a stationary point? Note that by defining

$$\mathbf{Z}_n = \begin{pmatrix} \mathbf{C}_n \\ \hline \hat{\theta}_n \end{pmatrix}, \qquad P = \begin{bmatrix} \frac{\beta}{2} & 0 \\ 0 & \frac{\alpha}{2} \end{bmatrix}$$

and

$$\nabla_{\mathbf{Z}_n}\epsilon_n = \begin{pmatrix} \nabla_{\mathbf{C}_n}\epsilon_n \\ \hline \nabla_{\hat{\theta}_n}\epsilon_n \end{pmatrix},$$

we can combine (23) and (25) by writing

$$\mathbf{Z}_{n+1} = \mathbf{Z}_n - P\nabla_{\mathbf{Z}_n}\epsilon_n. \tag{28}$$

It is shown in Ref. 17 that, if β and α are chosen small enough, the sequence $\{\mathbf{Z}_n\}$ converges in mean-square to a stationary point for which $\nabla_Z \epsilon_n = 0$. As pointed out previously, the stationary points all yield the global minimum of the mean-squared error and thus (28) converges to yield the minimum mean-squared error. The question of rate of convergence will be treated in a later paper.

In a practical situation, the receiver does not know *a priori* the ensemble averages represented by the channel correlation matrix A and the truncated impulse response vector $\mathbf{X}$. In this situation, the receiver can approximate the gradient search algorithm by utilizing the gradients with respect to $\mathbf{C}$ and $\hat{\theta}_n$ of the actual unnormalized squared error

$$|E_n|^2 = |\mathbf{C}_n^*\mathbf{R}_n - A_n \exp[j(2\pi f_c nT + \hat{\theta}_n)]|^2$$

instead of its mean. The A_n used in this calculation is initially an ideal reference known to the receiver, and during normal operation it is the receiver's output decision $\hat{A}_n$ in the nth interval. Thus a decision-directed stochastic approximation algorithm corresponding to (23) and (25) is

$$\begin{aligned}\mathbf{C}_{n+1} &= \mathbf{C}_n - \frac{\beta}{\langle |A|^2\rangle}\{\mathbf{R}_n\mathbf{R}_n^*\mathbf{C}_n - \hat{A}_n^* R_n \exp[-j(2\pi f_c nT + \hat{\theta}_n)]\} \\ &= \mathbf{C}_n - \frac{\beta}{\langle |A|^2\rangle}\mathbf{R}_n(Q_n^* - \hat{Q}_n^*), \qquad (29)\end{aligned}$$

where $\hat{Q}_n = \hat{A}_n \exp[j(2\pi f_c nT + \hat{\theta}_n)]$ is the "rotated" reference for the equalizer in the nth interval, using the receiver's decision $\hat{A}_n$, and

$$\begin{aligned}\hat{\theta}_{n+1} &= \hat{\theta}_n + \frac{\alpha \operatorname{Im}}{|A_n|^2}\{\mathbf{C}_n^*\mathbf{R}_n A_n^* \exp[-j(2\pi f_c nT + \hat{\theta}_n)]\} \\ &= \hat{\theta}_n + \frac{\alpha}{|A_n|^2}\operatorname{Im}(Q_n\hat{Q}_n^*), \qquad (30)\end{aligned}$$

which can also be written as $\hat{\theta}_{n+1} = \hat{\theta}_n + \alpha/|A_n|^2 \operatorname{Im}\{Y_n\hat{A}_n^*\}$.

Expression (30) has a simple heuristic interpretation. Suppose the equalizer has successfully removed all intersymbol interference so that its output, neglecting noise, can be written

$$Q_n \approx A_n \exp[j(2\pi f_c nT + \theta_n)].$$

Then we can write (30) as

$$\hat{\theta}_{n+1} \approx \hat{\theta}_n - \alpha \sin(\hat{\theta}_n - \theta_n). \qquad (31)$$

Equation (31) describes a discrete-time, first-order, phase-locked loop. Because the tracking algorithm makes use of the receiver's decisions, it can be termed a decision-directed tracking loop or a decision feedback loop.[5,6] As expressed in (31), the demodulator phase $\hat{\theta}_n$ is corrected by an amount proportional to the sine of the angular difference between the demodulated output Y_n and the receiver's decision $\hat{A}_n$. The maximum bandwidth of the phase jitter that can be compensated for is a function of the constants α and β. This will be explored in the next section and in a subsequent paper.

The decision-directed phase-tracking principle is well-known for application in systems that do not require adaptive equalization.[5,6] Its appropriateness is further confirmed by studies of maximum-likelihood detection.[9,12]

Note that there is a phase ambiguity in the receiver's decisions inherent in suppressed-carrier systems with symmetric signal constellations, using decision-directed phase tracking. For example, the QAM signal constellation of Fig. 1 is quadrantally symmetric, and therefore constant 90-degree errors in the phase of the receiver's decisions $\{\hat{A}_n\}$ are undetectable. This source of ambiguity is customarily removed by differentially encoding the transmitted data onto the points of the signal constellation, so that phase *differences* between successive decisions $\{\hat{A}_n\}$, rather than absolute phase values, convey information.

V. THE PHASE-TRACKING GAIN CONSTANT α: TRACKING BANDWIDTH CONSIDERATIONS

As pointed out in Section IV, the phase-angle-tracking algorithm is, assuming perfect equalization, basically that of a first-order phase-lock loop with gain constant α. The actual system does not behave quite as simply as this, however, since the passband equalizer, even with a small gain constant β, will also attempt to track the phase to some extent; i.e., the difference equations (29) and (30) are coupled. This coupling and its effect on performance will be explored in a later paper. In this section, we ignore this effect and also the effect of imperfect equalization. Furthermore, in view of the difficulty in analyzing discrete-time phase-locked systems, we make the following linearizing approximation for the steady-state phase error: $|\hat{\theta}_n - \theta_n| \ll \pi$, so that $\sin(\hat{\theta}_n - \theta_n) \approx \hat{\theta}_n - \theta_n$. We can write (31) as the simple linear difference equation

$$\hat{\theta}_{n+1} = (1-\alpha)\hat{\theta}_n + \alpha\theta_n. \tag{32}$$

The case of sinusoidal phase jitter, $\theta_n \equiv \mathrm{Re}\,[J\exp(j\omega nT)]$, is of interest because the phase jitter observed on telephone channels often consists of one or more sinusoids with frequencies $\omega/2\pi$ Hz, which are harmonics of various power line frequencies. The response $\hat{\theta}_n = \mathrm{Re}\,[\hat{J}\exp(j\omega nT)]$ of the linearized phase-locked loop to θ_n is easily found to be given by

$$\frac{\hat{J}}{J} = F(j\omega) = \frac{\alpha}{\exp(j\omega T) - 1 + \alpha}. \tag{33}$$

Now let us consider the effect of additive noise on the linearized phase-tracking algorithm. Assume the complex, equalized, demodulated output can be written

$$Y_n = A_n \exp[-j(\hat{\theta}_n - \theta_n)] + V_n, \tag{34}$$

where $V_n = v_n + j\check{v}_n$ is a complex gaussian random variable with zero mean. Although in general successive noise samples at the equalizer output will be correlated, we assume they are uncorrelated to simplify the results. The effect of this simplification should be minor if the phase-tracking bandwidth is much smaller than the data bandwidth or if the frequency response of the channel and of the equalizer are both nearly flat. Thus if the signal-to-noise ratio is $\langle |A|^2 \rangle / N_0$, $\langle v_n v_m \rangle = \langle \check{v}_n \check{v}_m \rangle = (N_0/2)\delta_{nm}$, and $\langle v_n \check{v}_m \rangle = 0$. Then eq. (31) for updating $\hat{\theta}_n$ can be written, after using the linearizing approximation as in (32),

$$\hat{\theta}_{n+1} = (1 - \alpha)\hat{\theta}_n + \alpha\theta_n + \alpha \operatorname{Im}\left(\frac{V_n}{A_n}\right). \tag{35}$$

The random variable $w_n = \operatorname{Im}(V_n/A_n)$ is not gaussian unless $|A_n|$ is constant (pure phase modulation). However, assuming the information symbols and noise are independent, the $\{w_n\}$ are zero-mean and statistically independent with variance $(N_0/2)\langle 1/|A|^2 \rangle$.

By the superposition principle for linear systems, the error in the output of the phase-locked loop is given by

$$\hat{\theta}_n - \theta_n = \operatorname{Re}\{J[F(j\omega) - 1]\exp(j\omega nT)\} + \nu_n, \tag{36}$$

where the sequence $\{\nu_n\}$ satisfies

$$\nu_{n+1} = (1 - \alpha)\nu_n + \alpha w_n, \tag{37}$$

and therefore has zero mean and steady-state variance

$$\lim_{n \to \infty} \langle \nu_n^2 \rangle = \frac{\alpha N_0}{2(2 - \alpha)} \left\langle \frac{1}{|A|^2} \right\rangle. \tag{38}$$

The mean-squared error in the phase estimate is thus

$$\langle (\hat{\theta}_n - \theta_n)^2 \rangle = \frac{|J|^2}{2} |F(j\omega) - 1|^2 + \langle \nu_n^2 \rangle,$$

which from (33) and (38) is

$$\langle (\hat{\theta}_n - \theta_n)^2 \rangle = \frac{|J|^2}{2} \frac{4 \sin^2(\omega T/2)}{\alpha^2 + 4(1 - \alpha)\sin^2(\omega T/2)} + \frac{\alpha N_0}{2(2 - \alpha)} \left\langle \frac{1}{|A|^2} \right\rangle. \tag{39}$$

The residual RMS phase jitter, given by the square root of the above expression, is plotted as a function of the coefficient α for signal-to-noise ratios $(|A|^2)/N_0$ of 30 dB and 22 dB in Figs. 4a and 4b, respectively. In each case, a 16-point QAM constellation is assumed. The higher the phase jitter frequency ω relative to the symbol rate $1/T$, the greater the residual RMS jitter. The curves in Figs. 4a and 4b show the case of no jitter (in which case, the residual jitter results from noise entering the discrete time-phase-locked loop) and also the cases of 14-degree peak-to-peak jitter with $\omega T/2\pi = 1/48$ and with $\omega T/2\pi = 1/20$. The choice of bandwidth of the decision-directed phase-tracking loop, determined by α, should be governed by the highest expected phase jitter frequency. If the spectrum of the phase jitter is known, a higher-order phase-locked loop may permit more effective phase tracking.

For given values of RMS residual phase jitter, the error probability can be approximated as in Ref. 1. For example, we find from Fig. 4b that the residual RMS phase jitter is about 2.5 degrees in the 16-point QAM systems, for $\alpha = 0.3$, when the channel has a signal-to-noise ratio of 22 dB and 14 degrees peak-to-peak channel phase jitter with frequency 1/48 that of the symbol rate. From Fig. 11 of Ref. 1, we find that the resulting error probability is about 4×10^{-7}. The same system

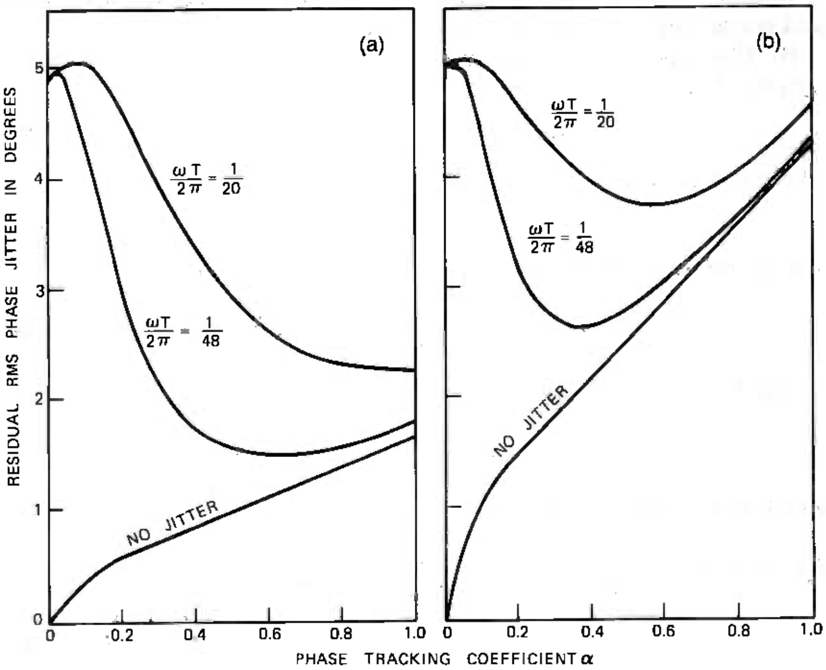


Fig. 4—(a) Residual rms phase jitter for a channel with 30-dB s/n and 14-degree peak-to-peak phase jitter. (b) Residual rms phase jitter for a channel with 22-dB s/n and 14-degree peak-to-peak phase jitter.

with the same value of α in the absence of phase jitter has an error probability of about 5×10^{-8}. The same system with $\alpha = 0$ and no phase jitter has an error probability of about 10^{-8}.

VI. SIMULATION OF THE QAM RECEIVER

The receiver described in this paper with the QAM constellation of Fig. 1 has been simulated on an IBM 370 computer, with 9600-b/s QAM data signals transmitted over real voiceband telephone channels as input. The simulation technique and the evaluation of this and other high-speed modems were reported in Ref. 18. In general, over a variety of different voiceband channels, the QAM system's performance appeared to be superior to that of all other systems tested.

One channel used for transmission of the QAM signals consisted of a Holmdel-to-Murray-Hill voiceband channel plus 50-Hz, 17-degree, peak-to-peak sinusoidal phase jitter which was inserted by a line

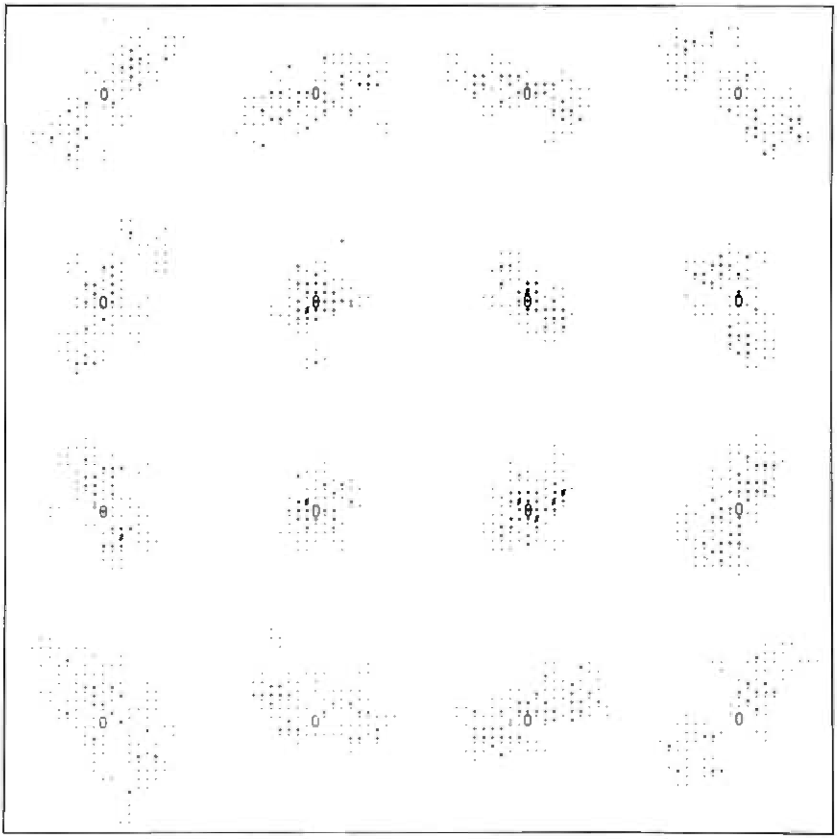


Fig. 5a—Receiver output constellation $\alpha = 0.01$.

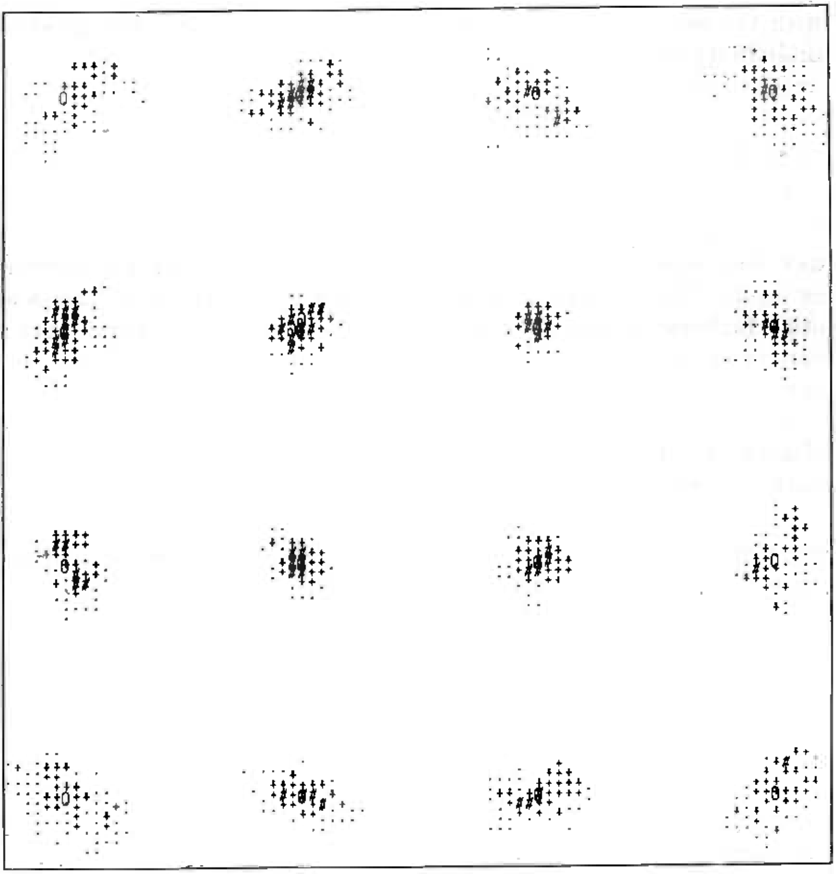


Fig. 5b—Receiver output constellation $\alpha = 0.3$.

simulator. The phase jitter and other impairments contributed by the Holmdel-to-Murray-Hill line alone were not too severe; the worst impairment was second-harmonic distortion, amounting to 32 dB (fundamental to average second-order product).

An illustration of the receiver's effectiveness in tracking and removing sinusoidal jitter from the same recorded data signal is shown in Figs. 5a and 5b, in which the unquantized complex (i.e., two-dimensional) receiver outputs are plotted, $\check{y}_n$ versus y_n. A · indicates that the particular set $(y, \check{y})$ occurred at least once during transmission, a + that it occurred between 4 and 10 times, a # that it occurred between 11 and 20 times, and an @ that it occurred more than 20 times. Thus, these figures are "constellations" or coarsely-quantized two-dimensional histograms of the receiver's demodulated ouput. The coordinates of the possible transmitted information symbols (±1, ±3 for QAM signals) are shown as circles. Figure 5a shows the two-dimensional receiver output constellation for the case when the parameter α is too small to allow the jitter to be tracked; $\alpha = 0.01$. Thus, in this case, the original 17-degree peak-to-peak jitter appeared at the receiver output, resulting in the banana-like shapes lying along the circumferences of circles centered at the origin. Note that, if the only impairment present was additive random noise, we would expect the scatter plots to look like circles centered on the information symbol coordinates and with radii proportional to the rms value of the noise. Figure 5b is a constellation for the case $\alpha = 0.3$, which allows the sinusoidal jitter to be tracked and almost completely removed by the demodulator.

VII. SUMMARY AND CONCLUSIONS

We have proposed a decision-directed demodulator phase-recovery loop coupled with adaptive passband equalization for use in a two-dimensional, suppressed-carrier, data communications system. Accurate compensation of phase jitter and frequency offset is afforded by placing the demodulator and a sufficiently wide bandwidth decision-directed phase-tracking loop together *following* the equalizer.

The derivation of the receiver's adaptive algorithm for jointly setting the equalizer tap coefficients and the carrier phase estimate was based on a gradient search algorithm for minimizing an expression for the receiver's output mean-squared error. This gradient search algorithm was shown to converge in the absence of noise and phase jitter to a nonunique but optimal set of tap coefficients and carrier phase-angle estimate.

Computer simulations using real-channel received waveforms reported here and in Ref. 18 confirm the feasibility of the QAM receiver structure.

Assuming perfect passband equalization and making a simplifying linear approximation, we analyzed the system's residual phase error as a function of carrier tracking loop gain, signal-to-noise ratio, and the amount and frequency of sinusoidal phase jitter. The optimum value of the carrier-tracking-loop-gain parameter α was seen to depend on the noise and phase-jitter parameters, although reasonable design compromises can be made.

A forthcoming paper[19] will explore the adaptation and tracking behavior of the combined equalizer, carrier recovery system, and demodulator in more detail.

The two-dimensional adaptive receiver structure described here can also be extended to systems employing decision feedback equalization. The performance of such a receiver will be reported in a later paper.

VIII. ACKNOWLEDGMENTS

I gratefully acknowledge discussions on the receiver structure with R. D. Gitlin, E. Y. Ho, K. H. Mueller, J. Salz, J. R. Sheehan, and D. A. Spaulding. The computer simulation was carried out with the help of R. R. Anderson and S. B. Weinstein.

REFERENCES

1. G. J. Foschini, R. D. Gitlin, and S. B. Weinstein, "On the Selection of a Two-Dimensional Signal Constellation in the Presence of Phase Jitter and Gaussian Noise," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 927–965.
2. G. J. Foschini, R. D. Gitlin, and S. B. Weinstein, "Optimization of Two Dimensional Signal Constellations in the Presence of Gaussian Noise," IEEE Trans. Commun., *COM-22*, No. 1 (January 1974), pp. 28–37.
3. C. M. Thomas, M. Y. Weidner, and S. H. Durrani, "Digital Amplitude Phase Keying with M-ary Alphabets," IEEE Trans. Commun., *COM-22*, No. 2 (February 1974), pp. 168–180.
4. J. R. O'Neill and B. R. Saltzberg, "An Automatic Equalizer for Coherent Quadrature Carrier Data Transmission Systems," 1966 IEEE International Conference on Communications, Philadelphia, June 15–17, 1966.
5. W. C. Lindsey and M. K. Simon, "Carrier Synchronization and Detection of Polyphase Signals," IEEE Trans. Commun., *COM-20*, No. 6 (June 1972), pp. 441–454.
6. M. K. Simon and J. G. Smith, "Carrier Synchronization and Detection of QASK Signal Sets," IEEE Trans. Commun., *COM-22*, No. 2 (February 1974), pp. 98–106.
7. R. D. Gitlin, E. Y. Ho, and J. E. Mazo, "Passband Equalization of Differentially Phase-Modulated Data Signals," B.S.T.J., *52*, No. 2 (February 1973), pp. 219–238.
8. R. Matyas and P. J. McLane, "Decision-Aided Tracking Loops for Channels with Phase Jitter and Intersymbol Interference," IEEE Trans. Commun., *COM-22*, No. 8 (August 1974), pp. 1014–1023.
9. H. Kobayashi, "Simultaneous Adaptive Estimation and Decision Algorithm for Carrier Modulated Data Transmission Systems," IEEE Trans. Comm. Tech., *COM-19*, No. 3 (June 1971), pp. 268–280.
10. R. W. Chang, "Joint Optimization of Automatic Equalization and Carrier Acquisition for Digital Communication," B.S.T.J., *49*, No. 6 (July–August 1970), pp. 1069–1104.
11. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1311–1347.
12. G. Ungerboeck, "Adaptive Maximum-Likelihood Receiver for Carrier-Modulated Data-Transmission Systems," IEEE Trans. Commun., *COM-22*, No. 5 (May 1974), pp. 624–636.
13. T. Ericson and U. Johansson, "A General Time-Discrete Equivalent to a Time-Continuous Gaussian Channel," IEEE Trans. on Information Theory, *IT-20*, No. 4 (July 1974), pp. 544–549.
14. J. E. Mazo, "Optimum Timing Phase for an Infinite Equalizer," B.S.T.J., *54*, No. 1 (January 1975), pp. 189–201.
15. D. L. Lyon, "Timing Recovery in Synchronous Equalized Data Communication," IEEE Trans. on Commun., *COM-23*, No. 2 (February 1975), pp. 269–274.
16. R. D. Gitlin and J. F. Hayes, "Timing Recovery and Scramblers in Data Transmission," B.S.T.J., *54*, No. 3 (March 1975), pp. 569–593.
17. A. A. Goldstein, *Constructive Real Analysis*, New York: Harper and Rowe, 1967.
18. R. R. Anderson and D. D. Falconer, "Modem Evaluation on Real Channels Using Computer Simulation," National Telecommunications Conference Record, December 1974, San Diego, pp. 877–883.
19. D. D. Falconer, "Analysis of a Gradient Algorithm for Simultaneous Passband Equalization and Carrier Phase Recovery," to be published in B.S.T.J., *55*, No. 4 (April 1976).

Copyright © 1976 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 3, March 1976
Printed in U. S. A.

The Field of a Line Charge Near the Tip of a Dielectric Wedge

By J. A. LEWIS and J. McKENNA

(Manuscript received October 17, 1975)

We calculate the potential of a line charge embedded in a dielectric medium of permittivity ϵ_2 in the presence of a dielectric wedge of permittivity ϵ_1. The potential is calculated with the aid of the Mellin transform, and the answer is given as a definite integral which is then transformed into an infinite series. We show that, for all wedge angles and all ratios ϵ_2/ϵ_1, $\nabla\varphi$ is singular at the tip of the wedge, and we give the strength of the singularity. The results have relevance to the design of contacts on semiconductor devices.

I. INTRODUCTION

Lewis and Wasserstrom[1] have calculated the strength of the field singularity at the tip of a dielectric wedge in the configuration shown in Fig. 1. In particular, with a wedge permittivity ϵ_1 greater than the permittivity ϵ_2 of the surrounding medium and a conductor angle $\beta = \pi$ (the "overhanging electrode"), they found that the tip field was singular for all wedge angles α greater than $\pi/2$. From this analysis, it was concluded that semiconductor devices with undercut edges ($\alpha < \pi/2$) would be advantageous in reducing local field strength and thus preventing breakdown.

Because the analysis of Ref. 1 was strictly local, based on an expansion of the potential in positive powers of the distance from the wedge vertex, multiplied by trigonometric functions of the polar angle, it was felt by some that the results were suspect, since they were not based on the solution of a complete boundary value problem. Here we lay that suspicion to rest by presenting the solution of such a problem, namely the field due to a line charge near a dielectric wedge, as shown in Fig. 2. The solution of this problem, previously treated by Smythe[2] in a somewhat involved fashion, gives Green's function for the composite region. Here we use the Mellin transform, obtaining an expansion of the potential near the wedge tip in terms of the poles of the transform. Based on this analysis, we conclude for the charge-wedge configuration of Fig. 2 that, for *arbitrary* ratios ϵ_2/ϵ_1, the wedge tip field is singular for *all* values of the half-angle α. We show that,

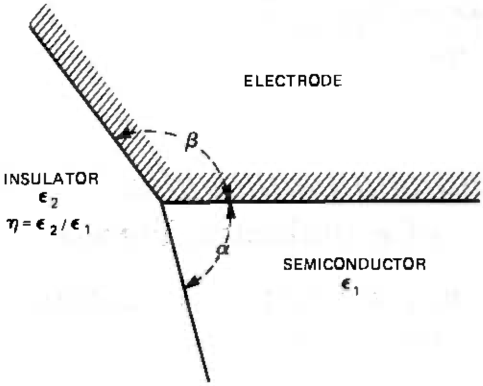


Fig. 1—Electrode, insulator, semiconductor configuration.

when the plane $y = 0$ is replaced by a perfectly conducting sheet, the field singularity due to the line charge is *exactly* as described by Lewis and Wasserstrom.[1] In general, we can conclude that, for any charge distribution for which the resulting potential is neither purely even nor purely odd, the field at the tip of the wedge will be singular for *all* ratios ϵ_2/ϵ_1 and *all* half-angles α.

II. THE PROBLEM

We consider the electrostatic potential due to a line charge of strength q in the presence of a dielectric wedge, as shown in Fig. 2. The charge lies at a distance a from the wedge tip in a dielectric medium with permittivity ϵ_2, while the wedge, with permittivity ϵ_1, occupies the region $-\alpha < \theta < \alpha$. We shall always assume that

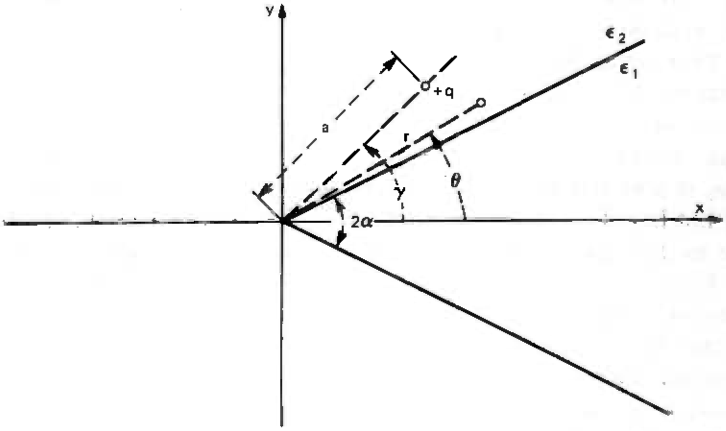


Fig. 2—The dielectric wedge and line charge.

$\gamma > \alpha$, taking into account the case where the charge lies within the dielectric wedge by interchanging ϵ_1 and ϵ_2, replacing α by $\pi - \alpha$ and γ by $\pi - \gamma$ where $\gamma < \alpha$. Finally, instead of working with the dimensional potential $\varphi(x, y)$ and distances (x, y), we introduce the dimensionless potential $u(r, \theta) = (\epsilon_2/q)\varphi(x, y)$, and the dimensionless distance $r = (x^2 + y^2)^{\frac{1}{2}}/a$. Thus, we will calculate the dimensionless potential due to a unit line charge at unit distance from the origin. Although we assume no trapped surface charge on the surface of the wedge, our analysis could be extended to cover this case also. It should be noted that, in these units, a unit line charge located at the origin of a homogeneous medium ($\epsilon_1 = \epsilon_2$) gives the potential

$$u = \frac{1}{2\pi} \ln \frac{1}{r}.$$

In the composite medium, u satisfies Laplace's equation

$$\nabla^2 u = u_{rr} + r^{-1}u_r + r^{-2}u_{\theta\theta} = 0, \tag{1}$$

in the wedge $|\theta| < \alpha$, and the inhomogeneous equation

$$\nabla^2 u = -(2\pi r)^{-1}\delta(r - 1)\delta(\theta - \gamma), \tag{2}$$

where δ is the Dirac delta function, giving the effect of the charge at $(r, \theta) = (1, \gamma)$, for $\alpha < \theta < 2\pi - \alpha$. The problem is completed by the requirement that u and ϵu_θ be continuous across $\theta = \pm\alpha$.

To facilitate further calculations, we split u into the sum of an odd function in y and an even function in y, setting

$$u = \tfrac{1}{2}(v + w),$$

where v and w satisfy eqs. (1) and (2), the continuity conditions, and the boundary conditions

$$v(r, 0) = v(r, \pi) = w_\theta(r, 0) = w_\theta(r, \pi) = 0. \tag{3}$$

Obviously, the pair of problems for v and w are equivalent to the original problem for u. It should be noted, though, that v alone is the potential due to a positive unit line charge at $(1, \gamma)$ and a negative unit line charge at $(1, 2\pi - \gamma)$, in the presence of the dielectric wedge. Alternatively, of course, it can be interpreted as the potential of the unit line charge at $(1, \gamma)$ in the presence of the wedge, when the plane $y = 0$ is replaced by a perfectly conducting sheet. This corresponds to the model of the overhanging electrode used by Lewis and Wasserstrom.[1] Further, w alone is the potential due to positive unit line charges at $(1, \gamma)$ and $(1, 2\pi - \gamma)$ in the presence of the wedge.

We now proceed to calculate v and w, or rather their Mellin transforms, the form of eq. (2) having been chosen to facilitate the application of the transform.

III. THE MELLIN TRANSFORM

The Mellin transform $\bar{v}(\theta, s)$ of $v(r, \theta)$ is given by[3]

$$\bar{v}(\theta, s) = \int_0^\infty r^{s-1} v(r, \theta) dr. \tag{4}$$

If eq. (2) is multiplied by r^{s+1} and integrated from 0 to ∞, after several integrations by parts there results the ordinary differential equation

$$\bar{v}'' + s^2 \bar{v} = -\frac{1}{2\pi} \delta(\theta - \gamma), \tag{5}$$

provided that s lies in the strip $\sigma_1 < \operatorname{Re} s < \sigma_2$, where

$$r^{s+1} v_r \to 0, \qquad r^s v \to 0 \tag{6}$$

for both $r \to 0$ and $r \to \infty$. These terms arise from the integration by parts of $r^{s+1}(v_{rr} + r^{-1} v_r)$. We will determine appropriate values of σ_1 and σ_2 later.

First, let us dispose of the singularity by calculating $\bar{v} = \bar{v}_1$ for a homogeneous medium for which $\eta = \epsilon_2/\epsilon_1 = 1$. Then $\bar{v}_1$ satisfies eq. (5) in $0 < \theta < \pi$ and the boundary conditions

$$\bar{v}_1(0, s) = \bar{v}_1(\pi, s) = 0.$$

The expression

$$\bar{v}_1 = A \sin s\theta - \frac{1}{s} \int_0^\theta \delta(\theta' - \gamma) \sin s(\theta - \theta') d\theta'$$

satisfies the equation and the first boundary condition. A is chosen to satisfy the secondary boundary condition. We finally obtain

$$\bar{v}_1(\theta, s) = \begin{cases} \sin s(\pi - \gamma) \sin s\theta / s \sin s\pi, & 0 < \theta < \gamma, \\ \sin s\gamma \sin s(\pi - \theta)/ s \sin s\pi, & \gamma < \theta < \pi, \\ -\bar{v}_1(2\pi - \theta, s), & \pi < \theta < 2\pi. \end{cases} \tag{7}$$

Now in this case, $v_1(r, \theta)$ is known, and $v_1 \sim r$ for small r and $v_1 \sim 1/r$ for large r, so for (6) to be satisfied for v_1 it is necessary that $-1 < \operatorname{Re} s < 1$.

An analogous calculation yields $\bar{w}$ in the homogeneous medium, viz,

$$\bar{w}_1 = \begin{cases} -\cos s(\pi - \gamma) \sin s\theta / s \sin s\pi, & 0 < \theta < \gamma, \\ -\cos s\gamma \cos s(\pi - \theta)/ s \sin s\pi, & \gamma < \theta < \pi, \\ \bar{w}_1(2\pi - \theta, r), & \pi < \theta < 2\pi. \end{cases} \tag{8}$$

Again in this case, $w_1(r, \theta)$ is known, $w_1 \sim r$ for small r and $w_1 \sim \ln r$ for large r, so for (6) to be satisfied for w_1 it is necessary that $-1 < \operatorname{Re} s < 0$.

We now use these expressions for the potentials due to a line charge in a homogeneous medium to obtain the potentials in the presence of the wedge. Note the way θ and γ are interchanged in eqs. (7) and (8) to make $\bar{v}_1$ and $\bar{w}_1$ continuous. We choose a similar form for $\bar{v}$, setting

$$\bar{v} = \bar{v}_1 + B \begin{cases} \sin s(\pi - \alpha) \sin s\theta, & \text{for } 0 < \theta < \alpha \\ \sin s\alpha \sin s(\pi - \theta), & \text{for } \alpha < \theta < \pi, \end{cases}$$

thus satisfying the differential equations, the boundary conditions at $\theta = 0$, $\theta = \pi$, and the continuity condition

$$\bar{v}(\alpha-, s) - \bar{v}(\alpha+, s) = 0.$$

B is determined from the second continuity condition

$$\bar{v}'(\alpha-, s) - \eta\bar{v}'(\alpha+, s) = 0,$$

where

$$\eta = \epsilon_2/\epsilon_1.$$

We find

$$B = -\frac{(1-\eta)\bar{v}_1'(\alpha, s)}{s[\eta \sin s\alpha \cos s(\pi - \alpha) + \cos s\alpha \sin s(\pi - \alpha)]}.$$

The transform of the odd part of the potential u is then given by

$$\bar{v}(\theta, s) = M(\theta, s)/sP(s, \alpha, \pi), \tag{9}$$

where

$$M(\theta, s) = \begin{cases} P(s, \theta, \theta) \sin s(\pi - \gamma), & 0 < \theta < \alpha \\ P(s, \alpha, \theta) \sin s(\pi - \gamma), & \alpha < \theta < \gamma \\ P(s, \alpha, \gamma) \sin s(\pi - \theta), & \gamma < \theta < \pi \\ -M(2\pi - \theta, s), & \pi < \theta < 2\pi, \end{cases} \tag{10}$$

and

$$P(s, \alpha, \theta) = (1 + \eta) \sin s\theta - (1 - \eta) \sin s(2\alpha - \theta). \tag{11}$$

A similar calculation yields the transform of the even part of the potential, viz,

$$\bar{w}(\theta, s) = N(\theta, s)/sQ(s, \alpha, \pi), \tag{12}$$

where

$$N(\theta, s) = \begin{cases} -R(s, \theta, \theta) \cos s(\pi - \gamma), & 0 < \theta < \alpha \\ -R(s, \alpha, \theta) \cos s(\pi - \gamma), & \alpha < \theta < \gamma \\ -R(s, \alpha, \gamma) \cos s(\pi - \theta), & \gamma < \theta < \pi \\ N(2\pi - \theta, s), & \pi < \theta < 2\pi, \end{cases} \tag{13}$$

and

$$\begin{aligned} Q(s, \alpha, \theta) &= (1 + \eta) \sin s\theta + (1 - \eta) \sin s(2\alpha - \theta), \\ R(s, \alpha, \theta) &= (1 + \eta) \cos s\pi - (1 - \eta) \cos s(2\alpha - \theta). \end{aligned} \tag{14}$$

Next, we must invert $\bar{v}(\theta, s)$, $\bar{w}(\theta, s)$ to obtain $v(r, \theta)$ and $w(r, \theta)$, or rather their forms for small r, since we are primarily interested in the behavior of the potential near the wedge tip.

IV. THE INVERSION INTEGRAL

If the integral (4) defining $\bar{u}(\theta, s)$ converges absolutely for all s in the strip $\sigma_1 < \text{Re} s < \sigma_2$, then $u(r, \theta)$ is given by the inversion integral[3]

$$u(r, \theta) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \bar{u}(\theta, s) r^{-s} ds, \tag{15}$$

where the integration contour in the complex s plane is any vertical straight line $\text{Re}\, s = c$ with $\sigma_1 < c < \sigma_2$. We have already seen from the derivation of $\bar{v}_1$ and $\bar{w}_1$ that $-1 < \sigma_2 < \sigma_1 < 0$. An examination of (9) to (14) shows that, while $\bar{v}(\theta, s)$ is regular at $s = 0$, $\bar{w}(\theta, s)$ has double pole there. Further, as we shortly show, both $\bar{v}$ and $\bar{w}$ have a countably infinite number of poles. They are all real, and the nonzero poles are all simple. The largest of the negative poles, at $s = s_0$, satisfies $-1 < -s_0 < 0$. Since the strip $\sigma_1 < \text{Re}\, s < \sigma_2$ can contain no singularities of $\bar{u}(\theta, s)$, it follows that $\sigma_1 = s_0$, $\sigma_2 = 0$. Assuming s_0 is known, since $\bar{u}(\theta, s) = \frac{1}{2}[\bar{v}(\theta, s) + \bar{w}(\theta, s)]$, eqs. (9) to (15) provide an explicit integral representation of the desired potential $u(r, \theta)$. This expression for u seems much more suitable than the expression given by Smythe[2] for determining the small r behavior of u.

The integral can be evaluated by the residue theorem[4] by closing this contour with large semicircles, to the left for small r and to the right for large r. Examination of the forms for $\bar{v}$ and $\bar{w}$, given by eqs. (9) to (14), reveals that the integrand of eq. (14) vanishes so rapidly on the semicircles that, as the semicircle radii tend to infinity, the semicircles make no contribution to the integral around the contour. The sum of the residues enclosed by the left semicircle thus gives the small r behavior of u; those to the right the large r behavior. It is clear from (11) and (14) that, if $p \neq 0$ is a zero of $P(s, \alpha, \pi)$, then so is $-p$, and, similarly, the nonzero roots of $Q(s, \alpha, \pi)$ come in pairs. Let p_n, q_n, $n = 1, 2, \cdots$ denote the positive roots of P and Q, respectively. Then it follows that, for $r < 1$,

$$u(r, \theta) = -\frac{1}{2} \sum_{n=1}^{\infty} \left\{ \frac{M(\theta, p_n) r^{p_n}}{p_n P'(p_n, \alpha, \pi)} + \frac{N(\theta, q_n) r^{q_n}}{q_n Q'(q_n, \alpha, \pi)} \right\}, \tag{16}$$

while, for $r > 1$,

$$u(r, \theta) = \frac{N(\theta, 0)}{2Q'(0, \alpha, \pi)} \ln r - \frac{1}{2} \sum_{n=1}^{\infty} \left\{ \frac{M(\theta, p_n) r^{-p_n}}{p_n P'(p_n, \alpha, \pi)} + \frac{N(\theta, q_n) r^{-q_n}}{q_n Q'(q_n, \alpha, \pi)} \right\}. \tag{17}$$

The poles of $\bar{v}$ and $\bar{w}$ lie at the zeros of $P(s, \alpha, \pi)$ and $Q(s, \alpha, \pi)$ except, of course, when $M(s, \theta)$ and $N(s, \theta)$ also vanish for the same value of s. For example, $\bar{v}$ has a removable singularity at $s = 0$. Since the zeros also depend on η, we emphasize this by writing $P(s, \alpha, \pi)$ and $Q(s, \alpha, \pi)$ as $P(s, \alpha, \pi; \eta)$ and $Q(s, \alpha, \pi; \eta)$. Then it is simple to show that

$$Q(s, \alpha, \pi; \eta) = \eta P\left(s, \alpha, \pi; \frac{1}{\eta}\right), \tag{18}$$

$$Q(s, \alpha, \pi; \eta) = P(s, \pi - \alpha, \pi; \eta). \tag{19}$$

If we set $s = p$, then $P(s, \alpha, \pi; \eta) = 0$ can be written

$$(1 + \eta) \sin p\pi + (1 - \eta) \sin p(\pi - 2\alpha) = 0,$$

which is identical to eq. (15) of Ref. 1 with $\beta = \pi$, the case of an overhanging electrode. The two smallest values of p for various values of η are then given by Fig. 11 of Ref. 1, here reproduced as Fig. 3. From Fig. 3 and eqs. (18) and (19) we see that, if $0 < \eta < 1$, $0 < \alpha < \pi/2$, or $1 < \eta$, $\pi/2 < \alpha < \pi$, then $p_1 > 1$, $q_1 < 1$, while if $0 < \eta < 1$, $\pi/2 < \alpha < \pi$ or $1 < \eta$, $0 < \alpha < \pi/2$, then $p_1 < 1$, $q_1 > 1$. In all cases, $p_2 > 1$, $q_2 > 1$. If $\rho_1 = \min\,(p_1, q_1)$, we have shown that for $r < 1$, $\nabla u \sim r^{\rho_1 - 1}$ and that for all angles α and ratios η, $\rho_1 < 1$ so the field is always singular at the tip of the wedge. For the case of an overhanging electrode for which the potential is given by v alone, $\nabla v \sim r^{p_1 - 1}$, so we have substantiated the local analysis of Ref. 1 by the solution of a complete boundary value problem.

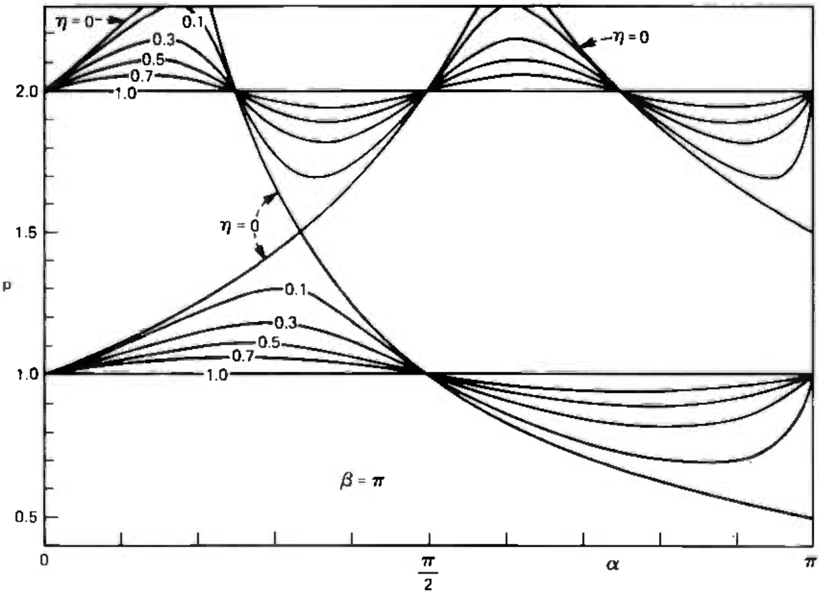


Fig. 3—The zeros of $P(s, \alpha, \pi)$ for various η.

V. REALITY AND SIMPLICITY OF POLES

One minor task remains to complete our analysis. We must show that all the roots of $P(s, \alpha, \pi) = 0$ and $Q(s, \alpha, \pi) = 0$ are real and simple. We write the two equations as

$$\sin \pi s = \pm E \sin (\pi - 2\alpha)s, \tag{20}$$

where $E = (1 - \eta)/(1 + \eta)$. We exclude the case $2\alpha = \pi$, for which the zeros are clearly real and simple. If we set $s = \sigma + i\tau$, the real and imaginary parts of eq. (20) become

$$\begin{aligned} \sin \pi\sigma \cosh \pi\tau &= \pm E \sin (\pi - 2\alpha)\sigma \cosh (\pi - 2\alpha)\tau. \\ \cos \pi\sigma \sinh \pi\tau &= \pm E \cos (\pi - 2\alpha)\sigma \sinh (\pi - 2\alpha)\tau. \end{aligned}$$

Divide the first by $\cosh \pi\tau$, the second by $\sinh \pi\tau$, square, and add to obtain

$$E^2 \left[\sin^2 (\pi - 2\alpha)\sigma \frac{\cosh^2 (\pi - 2\alpha)\tau}{\cosh^2 \pi\tau} + \cos^2 (\pi - 2\alpha)\sigma \cdot \frac{\sinh^2 (\pi - 2\alpha)\tau}{\sinh^2 \pi\tau} \right] = 1. \tag{21}$$

With $2\alpha \neq \pi$, $|\pi - 2\alpha| < \pi$, $\tau \neq 0$, so that $\cosh^2 (\pi - 2\alpha)\tau < \cosh^2 \pi\tau$, $\sinh^2 (\pi - 2\alpha)\tau < \sinh^2 \pi\tau$, eq. (21) implies $E^2 > 1$, which is impossible since $E^2 \leqq 1$ for $0 \leqq \eta < \infty$. By assuming a complex zero, we arrive at a contradiction, so all the zeros of P and Q must be real.

If s is a multiple zero of (20), it must also be a zero of one of the equations obtained by differentiating (20),

$$\cos \pi s = \pm E \left(1 - \frac{2\alpha}{\pi}\right) \cos (2\alpha - \pi)s. \tag{22}$$

If we square and add (21) and (22), we get

$$\frac{1}{E^2} = \sin^2 (\pi - 2\alpha)s + \left(1 - \frac{2\alpha}{\pi}\right)^2 \cos^2 (2\alpha - \pi). \tag{23}$$

Since $(1 - 2\alpha/\pi)^2 < 1$, (23) implies $(1/E^2) < 1$, which is a contradiction. Thus, all the zeros of P and Q must be simple.

REFERENCES

1. J. A. Lewis and E. Wasserstrom, "The Field Singularity at the Edge of an Electrode on a Semiconductor Surface," B.S.T.J., *49*, No. 6 (July–August 1970), pp. 1183–1194.
2. H. R. Smythe, *Static and Dynamic Electricity*, 2nd ed., New York: McGraw-Hill, 1950, pp. 70–72.
3. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, 2nd ed., New York: Oxford, 1948.
4. E. T. Copson, *Theory of Functions of a Complex Variable*, New York: Oxford, 1948.

Copyright © 1976 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 3, March 1976
Printed in U. S. A.

A Note on the Capacity of the Band-Limited Gaussian Channel

By A. D. WYNER

(Manuscript received November 4, 1975)

In this paper we reexamine results of a previous paper[1] *in which the capacity of the continuous-time channel with bandwidth W, average signal power P_0, and additive gaussian noise with flat spectral density N_0 was shown to be approximately W ln $(1 + P_0/N_0W)$ under a number of physically consistent assumptions.*

When one of the models in Ref. 1 is modified by techniques suggested by Slepian in his 1974 Shannon Lecture,[2] *the channel capacity turns out to be exactly W ln $(1 + P_0/N_0W)$.*

I. INTRODUCTION

In his 1974 Shannon Lecture,[2] D. Slepian introduced still another way of resolving the well-known paradoxes that arise when band-limited signals are studied in a physical "real world" context. One such paradox results from the fact that a mathematically band-limited function is determined for all time by its values in an arbitrarily small temporal interval—a highly nonphysical situation. An essential element in Slepian's resolution of these paradoxes is the recognition of the role of measurement accuracy in the determination of signals. To incorporate this into his mathematical model, he introduces the following concept. Two signals $s_1(t)$, $s_2(t)$, $-\infty < t < \infty$ *are really indistinguishable at level* ϵ if

$$\|s_1 - s_2\|^2 \leqq \epsilon, \tag{1}$$

where

$$\|f\|^2 \triangleq \int_{-\infty}^{\infty} f^2(t)dt$$

is the "energy" of the function of $f(t)$. He then says that a signal $g(t)$, $-\infty < t < \infty$, is *bandlimited to* $(-W, W)$ *at level* ϵ if $u_1(t)$ and $u_2(t)$ are really indistinguishable at level ϵ, where

$$U_1(f) = G(f) \tag{2a}$$

and

$$U_2(f) = \begin{cases} G(f), & |f| \leq W, \\ 0, & |f| > W. \end{cases} \tag{2b}$$

Here, U_1, U_2, and G are the Fourier transforms of u_1, u_2, and g respectively, i.e.,

$$U_1(f) = \int_{-\infty}^{\infty} e^{-i2\pi ft} u_1(t) dt,$$

etc. With band-limited functions so defined, paradoxes such as the one mentioned above are resolved, i.e., that $g(t)$ is band-limited to level $\epsilon > 0$ *does not* imply its predictability.

Let us remark that the quantity ϵ in the above definitions represents a limit on the accuracy of the measuring instruments used to determine the frequency spectrum of a signal. Note that $g(t)$ band-limited to a level ϵ *does not* imply that $c \cdot g(t)$ $(c > 1)$ is also so band-limited, even though $g(t)$ and $c \cdot g(t)$ have the same shape. Thus, Slepian's notion of band-limited signals is distinctly different from the usual notion which defines the bandwidth of a signal as a function of its shape.

In this note, we take another look at a related problem—determining the capacity of the band-limited gaussian channel—in the context of Slepian's bandwidth definition. We show that results obtained by the present author[1] have a particularly elegant statement in this new context.

II. STATEMENT OF THE PROBLEM

The definition of the continuous-time, band-limited, additive gaussian noise channel has the following components:

(*i*) Specification of a set $\mathcal{A}(T, W, P_0)$ of allowable channel input signals that are "approximately band-limited" to $(-W, W)$, approximately time-limited to $(-T/2, T/2)$, and with total energy not exceeding P_0T (so that the average power is $\leq P_0$).

(*ii*) Specification of the noise.

The channel output is

$$y(t) = s(t) + z(t),$$

where the channel input $s \epsilon \mathcal{A}(T, W, P_0)$, and the noise $z(t)$ is specified by (*ii*).

We take W and P_0 to be fixed parameters. A *code* with parameters (T, M, P_e) is a set of M functions called code words which belong to $\mathcal{A}(T, W, P_0)$, together with a decoder mapping which associates the received signal $y(t)$, $|t| < T/2$, with one of the M code words. With each of the M code words assumed to be *a priori* equally likely to be transmitted, P_e is the probability that the decoder makes an error.

A number $R \geqq 0$, is said to be *permissible* if for every $\lambda > 0$ there is a $T = T(\lambda)$ sufficiently large that there exists a code with parameters (T, M, P_e), where

$$M \geqq e^{RT}, \qquad P_e \leqq \lambda.$$

The channel capacity C is defined as the supremum of permissible R. Reference 1 has a detailed discussion of this problem and its formulation.

In what follows, we shall specify a set $\mathfrak{A}(T, W, P_0)$ and also specify the noise. The main result is a formula for C. This model is very similar to Model 4 in Ref. 1.

(*i*) Let $\mathfrak{A}(T, W, P_0)$ be the set of functions $s(t)$, $-\infty < t < \infty$, which satisfy

(a) $s(t) = 0, \qquad |t| \geqq T/2,$ (3a)

(b) $\|s\|^2 \leqq P_0 T,$ (3b)

(c) $s(t)$ is band-limited to $(-W, W)$ at level $\epsilon > 0$. (3c)

Thus, $\mathfrak{A}(T, W_0, P_0)$ is a set of strictly time-limited and approximately band-limited signals.

(*ii*) The noise function $z(t)$, is assumed to be a sample from a gaussian noise process with spectral density

$$N(f) = \begin{cases} N_0/2, & |f| < W, \\ 0, & |f| \geqq W. \end{cases} \tag{4}$$

Let us remark at this point that although we assume in our model that the signal is exactly time-limited to $(-T/2, T/2)$ and the noise is exactly band-limited to $(-W, W)$, our results do not exploit these assumptions. In fact, our results will hold if we introduce appropriate approximations here too.

Finally, we must make the assumption that the decoder function is not capable of distinguishing among signals that are arbitrarily close together. Specifically, we assume that if $y_1(t)$, $y_2(t)$, $-T/2 < t < T/2$ are functions that are mapped by the decoder to distinct code words, then

$$\int_{-T/2}^{T/2} [y_1(t) - y_2(t)]^2 dt \geqq \epsilon'. \tag{5}$$

Inequality (5) is equivalent to requiring that the segments of $y_1(t)$ and $y_2(t)$, $|t| < T/2$ (on which the decoding must be done), are really distinguishable at level ϵ'.* Put another way, the receiver must be insensitive to measurement errors of energy $<\epsilon'/4$.

III. THE RESULT

We state our result as a theorem.

Theorem: For the model defined above,

$$C = W \log \left(1 + \frac{P_0}{N_0 W}\right), \tag{6}$$

provided $\epsilon' > 4\epsilon$.

This result is analogous to the "$2WT$" theorem given by Slepian in Ref. 2. Note that (6) holds for every ϵ and ϵ' provided only that $\epsilon' > 4\epsilon$. Thus, the result is independent of the precision with which we can make measurements.

Proof: The theorem follows immediately from the capacity formula (28) given for Model 4 in Ref. 1. Observe that our $\mathcal{A}(T, W, P_0)$ is identical to the set $a_4(T, W, P_0)$, with $\eta = \epsilon/(P_0 T)$. (Note that no changes in the capacity formula will result when we require the channel input signals to have energy *exactly* $P_0 T$.)

Also note that the right member of ineq. (29) of Ref. 1 should be "$4\nu N_0 WT$." Thus, our assumption in (5) is identical to the assumption of (29) in Ref. 1 with $\nu = \epsilon'/(4N_0 WT)$.

It follows that the capacity formula (28) in Ref. 1 holds; that is, for our model

$$C = W \log \left(1 + \frac{P_0}{N_0 W}\right) + \epsilon(\eta, \nu), \tag{7}$$

where $\epsilon(\eta, \nu) \to 0$, as $\eta, \nu \to 0$, provided

$$\frac{\nu}{\eta} > \frac{P_0}{N_0 W}. \tag{8}$$

Since $\eta = \epsilon/(P_0 T)$ and $\nu = \epsilon'/(4N_0 WT)$, both $\eta, \nu \to 0$ as $T \to \infty$. Further, (8) holds if $\epsilon'/\epsilon > 4$, so that (7) becomes (6) as $T \to \infty$.

REFERENCES

1. A. D. Wyner, "The Capacity of the Band-Limited Gaussian Channel," B.S.T.J., *45* (March 1966), pp. 359–395. Also reprinted in *Key Papers in the Development of Information Theory,* ed. D. Slepian, New York: IEEE Press, 1974, pp. 190–193.
2. D. Slepian, "On Bandwidth," 1974 Shannon Lecture, presented at the International Symposium on Information Theory, Notre Dame, Indiana, October 21, 1974, Proceedings of the IEEE, *64*, No. 3 (March 1976), pp. 292–300.

* This assumption requires that the space of received signals contain "null zones" which are not in the domain of the decoder mapping. When the received signal belongs to a null zone, the decoder declares an error.

Copyright © 1976 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 3, March 1976
Printed in U. S. A.

On Optical Data Communication via Direct Detection of Light Pulses

By J. E. MAZO and J. SALZ

(Manuscript received October 16, 1975)

A number of problems are considered relevant to understanding the performance of optical-fiber communication systems that use pulse transmission. The methods used are typically exact solutions or bounds, and we concentrate on simple examples that aid our understanding. Some of our work makes contact with previous studies, particularly by Personick and Hubbard. The major results are:

(i) Presentation of an integral equation for the output density for single-pulse detection with arbitrary avalanche gain

(ii) Exact solution for the probability distribution for gains in physical avalanche diodes

(iii) Bounds on performance when intersymbol interference is present (but no avalanche gain) which suggest that an optimum-bit detector can perform, under practical conditions, only two or three dB better than a simple integrate-and-dump filter, yielding results still many dB from the quantum limit. Thus, in particular, little performance gain is to be expected from equalization techniques.

I. INTRODUCTION AND OVERVIEW

A large part of traditional communication theory has been directed to detecting and processing electrical signals transmitted over wires, cables, or the like. While the physical realization of each of these traditional systems may have led to mathematical treatments designed to handle problems such as linear distortion or fading, which were peculiar to one, or even perhaps several, systems, the principal concern of all mathematical treatments of these time-continuous channels has been the ubiquitous additive gaussian noise. In fact, it would be fair to say that much of the structure of the mathematical treatments used has been dictated by the mathematical properties of this noise. In the absence of noise, many problems would immediately degenerate, at least theoretically, to situations of perfect detection, infinite capacity, etc.

The consideration of some promising optical communication systems seems to alter the above picture. We have in mind the transmission of information by way of light pulses propagating through an optical fiber and subsequently detected by a photodetector that converts electromagnetic energy in the fiber to electrical signals in a circuit. We immediately note certain features which this problem has in common with the traditional problems. For one thing, the fiber can delay, attenuate, or spread the transmitted pulses. For another, the electrical signal after photodetection may be corrupted by additive gaussian noise. Yet there is another fundamental impairment. The electromagnetic signal that propagates in the fiber (which acts as a wave guide) is, under practical considerations, of sufficiently weak intensity that any effective detection mechanism must be based upon the quantum nature of the electromagnetic disturbance. In other words, detection must be based upon photon counting. Here, a new element enters the problem—photon counting is subject to statistical fluctuations. In the quantum case, a signal uncorrupted by any external disturbance still carries with it its own "noise," as it were, which is not additive gaussian. This new noise manifests itself in the following way. The photon-counting process is a time-varying Poisson process whose intensity (or rate) function $\lambda(t)$ varies in direct proportion to the information-bearing pulse train, the latter being thought of in the conventional way (except it must now always be positive). Our purpose here is to explore some of the communication theory of this new situation, paying particular attention to the use of our considerations in proposed fiber-optic communication systems.

The general background of the material that we treat, namely, direct detection of photons in an optical fiber, may be found in works by Personick[1,2] and Foschini et al.[3] Direct detection refers to the processing of the electrical signal at the output of a photodetector as opposed to, say, more esoteric detection schemes based on optimum processing of the existing electromagnetic field, considered as a quantum system. In the case of binary transmission, the choice between a one or a zero is, in the systems considered here, translated into the presence or absence of a short burst of optical power (light) in the fiber. To understand this in more detail, we shall trace the passage of a single pulse through our mathematical model of the system (see Fig. 1). In the case of a one being transmitted, an electrical signal (a square pulse of duration T) turns on our "flashlight," which in this case is a laser or light-emitting diode, and electromagnetic energy is sent into the transmission medium (optical fiber). If a photodetector is placed at the end of the fiber, photons will be detected due to the electromagnetic energy present. Exactly when in time the photons

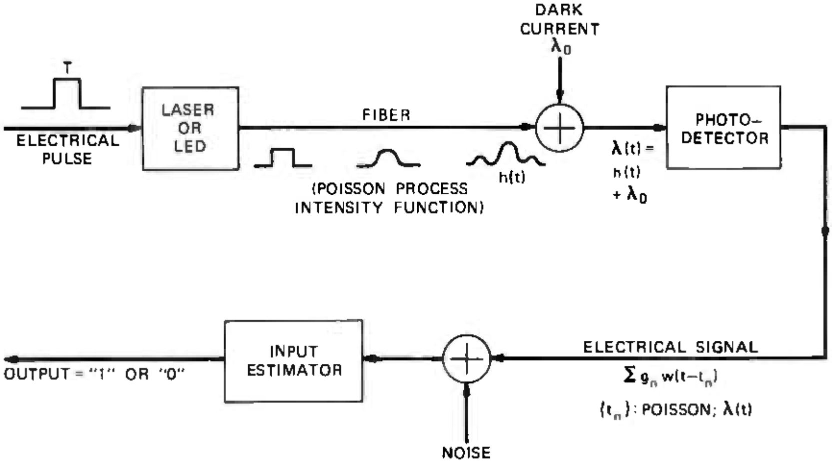


Fig. 1—Passage of a single pulse through the optical system.

register on the detector is random and is the Poisson process spoken of earlier. The probability of receiving a count between time t and $t + dt$ is given by $h(t)dt$, where, owing to effects in the fiber, $h(t)$ is a distorted and attenuated version of the transmitted pulse. The accumulation of distortion as the pulse propagates down the fiber is also sketched in Fig. 1. In practice, a background of counts also exists. This is called the dark current and is modeled by introducing a constant additive intensity function λ_0 before the detector, although some of these counts can originate in the physical detector itself. Typically, transmitted power and transmission loss are adjusted so that on the order of one or two hundred photons per pulse are, on the average, detected. The dark current contributes from about 1 to 5 percent of the counts.

To transmit a zero, we simply do not turn on the transmitting power, and the detector only registers counts resulting from the dark current.

We have been loosely speaking of the output of the photodetector as "counts." The actual electrical current at the output of this device caused by a photon is a wideband pulse $g \cdot w(t)$ (very narrow compared with T, a delta function in the limit), where g=integer-valued random variable or $g \equiv 1$, depending on whether or not an avalanche diode is used. The electrical current at the output of the photodetector is further distorted by gaussian noise whose effect is often lessened in importance when an avalanche diode is used, but not for the $g = 1$ case. In the most literal modeling of the experimental situation, the finite bandwidth of $w(t)$ prevents one from assuming that the Poisson part of the observation is singular, i.e., can be separated out from the background gaussian noise; however, whenever we feel there are insights to be gained from the separation we shall make it.

If we take into account the facts that Personick[4] has shown superposition to hold (approximately) for optical-fiber transmission and that optical power is positive, then we may extend our single-pulse description to a model for transmission of an entire pulse train. If we transmit a sequence of on or off pulses, then the "received signal," defined as that electrical signal on which we may do processing, can be written as

$$\sum_n g_n w(t - t_n) + n(t), \tag{1}$$

where the time points $\{t_n\}$ form a Poisson process having intensity function $\lambda(t)$, with

$$\lambda(t) = \sum_n a_n h(t - nT) + \lambda_0 \tag{2}$$

and

$$\begin{aligned} h(t) \geqq 0 &= \text{distorted pulse} \\ a_n = 0, 1 &= \text{independent, equiprobable data symbols} \\ \lambda_0 \geqq 0 &= \text{dark current} \\ T &= \text{signaling interval} \\ n(t) &= \text{gaussian noise} \\ g_n &= \text{avalanche gain factors} \\ w(t) &= \text{output pulse of photodetector.} \end{aligned} \tag{3}$$

At various stages of our discussion, we may, for interests of simplicity or clarity, idealize or eliminate certain aspects of the full model given by (1), (2), and (3).

The communication theorist is interested in processing the signal (1) to estimate the a_n given in (2). If the distortion is not severe, one may simply process in an intuitive way and (assuming proper synchronization) count the number of photons detected in the appropriate T-second interval. If $g_n = 1$, this is accomplished by integrating the output for T seconds (so-called integrate-and-dump detection). However, the simplicity of this technique demands its investigation even when g_n are random. Neglecting the gaussian noise and assuming g_n are exponential random variables allow one to determine exactly the probability distribution of the output statistic and to determine error rates. This is done in Section II. In Section III we return to the $g = 1$ case to observe the effects of the random gain. In Section IV, Personick's implicit equation for the random gains g_n of actual photodetectors is studied in detail and the exact distribution of these gains is found. Also, the use of Chernoff bounds for bounding the error rate in the general situation is discussed. Section V branches out to include a worst-case analysis of intersymbol interference [the case of appreciable spreading of $h(t)$] using integrate-and-dump detection. A particular example is also computed. Finally, in Section VI, we consider the question of replacing the integrate-and-dump detector with an optimum detector. We know that equalization can achieve considerable improvements for voiceband telephone transmission, but can we expect the same here? Using the lower bound on performance which we derive for the optimum detector and applying this to the example of Section V, we find that performance greatly surpassing that of integrate-and-dump detection cannot be expected.

II. INTEGRATE-AND-DUMP DETECTION—AVALANCHE DETECTORS

As already mentioned in the introduction, a simple way to detect the jth bit in (2) is to integrate the output of the photodetector over the jth T-second interval and compare the random variable thus obtained with a threshold F; if the output is greater than F, a one is declared (pulse present); if it is less than F, a zero is declared (pulse absent). In this section, we discuss the exact error rate for such a situation when pulse overlap in (2) can be neglected, as well as the additive noise. Further, the gains g_n are assumed to be exponentially distributed.

We shall need the moment-generating function (MGF) for the indicated random variable, but we may as well begin by giving the MGF for a general linear filter $P(t)$ rather than simply an integrator. Consider a Poisson point process having an arbitrary intensity function $\lambda(t)$ [not necessarily of the form (2)], and let the nth count be given nonnegative weight g_n, i.e., consider

$$\sum_n g_n \delta(t - t_n), \tag{4}$$

where the sequence of time points $\{t_n\}$ is Poisson with intensity function $\lambda(t)$. If (4) is linearly filtered, with $P(t)$ being the impulse response of the filter, then the output of the filter at time t, $x(t)$ can be shown by elementary calculations to have moment-generating function given by

$$M_X = E \exp [sX] = \exp \left[\int_{-\infty}^{\infty} \lambda(\tau) \{ M_g[sP(t - \tau)] - 1 \} d\tau \right], \tag{5}$$

where $M_g(s)$ is the moment-generating function of the g_n, assumed independent, and we have set $x(t) = X$. In particular, if $P(\tau) = 1$ for $0 < \tau < T$ and zero elsewhere, and if $t = T$, (5) will simplify to

$$M_X = \exp \{ \Lambda [M_g(s) - 1] \}, \tag{6}$$

where

$$\Lambda = \int_0^T \lambda(\tau)d\tau. \tag{7}$$

If the pulse $w(t)$ in (1) (assumed of unit area) is narrow enough so that end effects are negligible when doing the integration and if pulse overlap in (2) is negligible, then (6) and (7) are relevant quantities to consider in determining the error rate for integrate-and-dump detection of (1) and (2). To treat the two separate cases of a one or a zero, we need only replace Λ in (6) by Λ_i, $i = 1$ or 0, where

$$\Lambda_1 = \int_0^T h(t) + T\lambda_0 \tag{8}$$

$$\Lambda_0 = T\lambda_0. \tag{9}$$

While the gaussian noise will be neglected here, let us at least note that to include the effect of the added noise term on the integrated output, we would multiply (6) by the moment-generating function of the noise $M_n(s)$,

$$M_n(s) = \exp\left(\frac{s^2\sigma^2}{2}\right), \tag{10}$$

to obtain the MGF of the new output variable. In (10), the variance of the noise σ^2 is given by

$$\sigma^2 = \frac{N_0}{2}T \tag{11}$$

for the case of the integrator with white noise of two-sided spectral density $N_0/2$, or

$$\sigma^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} N(\omega)|\tilde{P}(\omega)|^2 d\omega, \tag{12}$$

in general, where $N(\omega)$ denotes a general noise spectrum and $\tilde{P}(\omega)$ is the Fourier transform of $P(t)$.

In the special case where the g_n in (4) are continuous variables and are exponentially distributed, i.e.,

$$p(g) = \alpha \exp(-\alpha g), \qquad g > 0, \tag{13}$$

we have

$$M_g(s) = \frac{\alpha}{\alpha - s}, \qquad s < \alpha. \tag{14}$$

At this stage, it is easier to work with the characteristic function version of (6), namely,

$$C_X(\omega) = \exp\{\Lambda[C_g(\omega) - 1]\}, \tag{15}$$

with $C(\omega)$ denoting characteristics functions now, e.g.,

$$C_X(\omega) = E \exp (i\omega X).$$

To obtain our integral equation for $p(x)$, differentiate (15) once with respect to ω, multiply by $\exp(-i\omega x)$, and integrate over x to obtain

$$xp(x) = \Lambda \int_0^x up_g(u)p(x-u)du, \tag{16}$$

where $p(x)$ denotes the density of the random variable X in (14), and $p_g(u)$ denotes the density of the nonnegative gain variable g.

For the exponential gain case (13), an exact solution to (16) can be found. Note that then the variable x has probability $\exp(-\Lambda$ of being zero (no counts) and $p(x)$ will thus contain a δ function at the origin. Introducing this explicitly by writing

$$p(x) = \exp(-\Lambda)\delta(x) + \exp(-\alpha x)f(x), \tag{17}$$

we find

$$xf(x) = (\alpha\Lambda e^{-\Lambda})x + \alpha\Lambda \int_0^x (x-w)f(w)dw, \tag{18}$$

where use has been made of (13). Differentiating (18) twice, we obtain Bessel's equation

$$x^2 f'' + 2xf' - (\alpha\Lambda)xf = 0, \tag{19}$$

where f' stands for differentiation. The appropriate solution of (19) gives, finally, for the density $p(x)$ of the detection statistic

$$p(x) = e^{-\Lambda}\delta(x) + e^{-\Lambda}\sqrt{\frac{\Lambda\alpha}{x}}\, e^{-\alpha x} I_1(2\sqrt{\alpha\Lambda x}), \tag{20}$$

$I_1(\cdot)$ being the modified Bessel function.* The following may be useful in connection with (20):

$$I_1(x) \leqq \frac{\exp(x)}{\sqrt{2\pi x}}, \qquad x \geqq 0 \tag{21}$$

$$I_1(x) \sim \frac{\exp(x)}{\sqrt{2\pi x}}, \qquad x \text{ large} \tag{22}$$

$$I_1(x) \sim \frac{x}{2}, \qquad x \text{ small.} \tag{23}$$

Typically, Λ_1 is in the range of 100 to 200 for a light pulse present and

* This exact result, as well as several useful approximations to it found later in this section, were first derived by Hubbard (Ref. 5) using other techniques.

Λ_0 in the range of 5 to 10 for dark current only. The quantity $1/\alpha$, the average gain, may be 100 or 200. The average number of counts for a pulse is then Λ_1/α so, to within a factor of 2 or so, the decision threshold will be around $\Lambda_1/2\alpha$. Thus, virtually all the area of interest in (20) occurs for $x > 1/\alpha\Lambda_i$ for both $i = 1$ or 2, and (22) may be used and, to excellent accuracy,

$$p(x)dx \approx \frac{(\alpha\Lambda)^{\frac{1}{4}}}{\sqrt{4\pi}} \frac{1}{(x)^{\frac{3}{4}}} \exp\left\{-\frac{(\sqrt{x}-\sqrt{\Lambda}/\alpha)^2}{2(1/2\alpha)}\right\} dx, \qquad x > \frac{1}{\alpha\Lambda}. \tag{24}$$

Equation (24) is slightly more attractive if we write instead the density for $u = \sqrt{x}$, $p_u(u)$,

$$p_u(u)du \approx \left(\frac{\Lambda}{\alpha}\right)^{\frac{1}{4}} \frac{1}{\sqrt{u}} \cdot \frac{\exp\left\{-\dfrac{(u-\sqrt{\Lambda}/\alpha)^2}{2(1/2\alpha)}\right\}}{\sqrt{2\pi}(\sqrt{1}/2\alpha)} du, \qquad u > \frac{1}{\sqrt{\alpha\Lambda}}, \tag{25}$$

showing that $\sqrt{X}$ is, over a rather wide range, gaussian with mean $\sqrt{\Lambda/\alpha}$ and variance $1/2\alpha$. Note $\Lambda/\alpha = EX$, while variance of X is $2\Lambda/\alpha^2$. Also, eq. (25) should not be confused with the central limit theorem version of (24), which is obtained when one writes (for large Λ) $x = (\Lambda/\alpha) + \epsilon$ and ϵ becomes gaussian. Since, from (21), eq. (22) is an upper bound as well as being asymptotic, we have

$$p\left[x > F > \frac{\Lambda}{\alpha}\right] \leq \left(\frac{\Lambda}{F\alpha}\right)^{\frac{1}{4}} Q(\sqrt{2\alpha F} - \sqrt{2\Lambda}), \tag{26}$$

where

$$Q(y) = \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-u^2/2} du \sim \frac{e^{-y^2/2}}{\sqrt{2\pi}y}. \tag{27}$$

Likewise, in the same spirit of approximation that indicates (26) to be an excellent approximation (in addition to being an upper bound), one may write for the lower tail

$$p\left[x < F < \frac{\Lambda}{\alpha}\right] \approx \left(\frac{\Lambda}{F\alpha}\right)^{\frac{1}{4}} Q(\sqrt{2\Lambda} - \sqrt{2\alpha F}). \tag{28}$$

Even for Λ's differing by a factor of 100, the fourth root factor in front of (26) and (28) is weak indeed. Thus, we may, to excellent approximation, find the best threshold by equating the arguments of the Q function for the two cases of error. This results in

$$\sqrt{2\Lambda_1} - \sqrt{2\alpha F} = \sqrt{2\alpha F} - \sqrt{2\Lambda_0}. \tag{29}$$

The left-hand side of (29) comes, of course, from using (28) for a pulse present (the number of counts is then expected to exceed the threshold).

Table I — Tabulation of error rate and threshold for an avalanche detector with exponentially distributed gains

Λ_0	Λ_{1s}	$F_{\text{opt}}(\alpha = 1)$	P_e [eq. (31)]	Quantum Limit
4	100	37.20	2.09×10^{-9}	1.86×10^{-44}
4	200	66.28	8.88×10^{-19}	$\sim 10^{-88}$
4	400	122.1	3.9×10^{-38}	$\sim 10^{-176}$
10	100	46.58	8.03×10^{-8}	1.86×10^{-44}
10	200	77.91	3.61×10^{-16}	$\sim 10^{-88}$
10	400	137.0	3.66×10^{-34}	$\sim 10^{-176}$

Likewise, the right member of (29) comes from using (26) for only dark current where the counts usually fall below the threshold F and an error is made only if they exceed it. We immediately obtain from (29)

$$\sqrt{F_{\text{opt}}} = \sqrt{\frac{\Lambda_1}{4\alpha}} + \sqrt{\frac{\Lambda_0}{4\alpha}}, \tag{30}$$

where, again, Λ_0 is not to be too small, for example, $\Lambda_0 \geqq 2$. In the above, we have in mind, from (8), taking $\Lambda_1 = \Lambda_{1s} + \Lambda_0$ where Λ_{1s} is due to signal alone.

For future comparisons, we should inject at this point the fact that the best detection probability one can obtain with no dark current (or no gaussian noise) is $\frac{1}{2} \exp(-\Lambda_{1s})$, often referred to as the quantum limit.

Table I displays values of the right member of (26), for the optimum F given by (30), i.e., it displays the quantity

$$\left(\frac{\Lambda_0}{F\alpha}\right)^{\frac{1}{2}} Q\left(\sqrt{\frac{\Lambda_1}{2}} - \sqrt{\frac{\Lambda_0}{2}}\right) \tag{31}$$

evaluated for several values of Λ_0 and Λ_{1s}, along with the quantum limit. Note that only αF enters the expressions, and thus the actual value of α plays no role in determining the probabilities for this problem. The fact should also be evident from the scaling properties of the problem. In real applications, $1/\alpha$ would be large so that the electronic circuitry could "see" the pulses above the gaussian noise.

Table I shows (for the parameters shown) about a 7-dB loss relative to the quantum limit, owing to the dark current, and also in part to the random nature of the gain mechanism.*

* To be perfectly clear on this point, it is really the additional (random) gain provided by the avalanche detector that allows one to formulate the physical problem as in (4) without gaussian noise. However, from a mathematical point of view, once (4) is written down, the random gains are hypothesis-insensitive, and thus would be ignored by an optimum detector.

III. INTEGRATE-AND-DUMP DETECTION—PURE POISSON CASE

We now give a brief discussion for the g = 1 case of (4), namely, the random variable X is Poisson,

$$p(X = n) = \frac{e^{-\Lambda}\Lambda^n}{n!} \qquad n = 0, 1, 2, \cdots, \tag{32}$$

$$EX = \Lambda, \qquad \text{var } X = \Lambda^2. \tag{33}$$

The purpose of the remarks will be to shed light on the degradation suffered when the g_n are random, as mentioned at the end of the last section.

If X is Poisson, then the probability that X is larger than or equal to k is

$$\sum_{n=k}^{\infty} \frac{e^{-\Lambda}\Lambda^n}{n!} = \frac{e^{-\Lambda}\Lambda^k}{k!}\left[1 + \frac{\Lambda}{k+1} + \frac{\Lambda^2}{(k+1)(k+2)} + \cdots\right]. \tag{34}$$

If, in addition, we assume $(k + 1) > \Lambda$, then a simple consequence of (34) is that

$$\left(1 + \frac{\Lambda}{k+1}\right)\frac{e^{-\Lambda}\Lambda^k}{k!} < \Pr\left[X \geqq k > \Lambda - 1\right] < \frac{1}{1 - (\Lambda/k + 1)} \cdot \frac{e^{-\Lambda}\Lambda^k}{k!}. \tag{35}$$

Similarly, for the lower tail we have

$$\left(1 + \frac{k}{\Lambda}\right)\frac{e^{-\Lambda}\Lambda^k}{k!} < \Pr\left[x \leqq k < \Lambda\right] < \frac{1}{1 - (k/\Lambda)} \frac{e^{-\Lambda}\Lambda^k}{k!}. \tag{36}$$

Thus, ignoring the weak effects of the coefficient in front, the optimum threshold F for a problem such as the one described in Section II is obtained by equating probabilities such as these in (35) and (36), yielding

$$e^{-\Lambda_0}\Lambda_0^F = e^{-\Lambda_1}\Lambda_1^F \tag{37}$$

or, equivalently, the optimum threshold in this case is

$$F = \frac{\Lambda_1 - \Lambda_0}{\ln(\Lambda_1/\Lambda_0)}. \tag{38}$$

Table II displays the right-hand side of (35) for k given by the rounded-off values of (38). In particular, we see degradation ranging from 3.5 to 4 dB compared to the quantum limits given in Table I. Typically, then, detecting the presence or absence of a single pulse using random amplitudes, as a linear detector might, results in a 3- to 4-dB degradation (for the exponential case), compared with an "ideal"

Table II — Tabulation of error rate and threshold for detection with constant gain

Λ_0	Λ_{1s}	F_{opt}	P_e [eq. (35)]
4	100	30.69	1.17×10^{-17}
4	200	50.87	6.49×10^{-38}
4	400	86.67	2.18×10^{-82}
10	100	41.70	4.21×10^{-14}
10	200	65.69	9.80×10^{-32}
10	400	107.7	3.78×10^{-71}

avalanche detector, which has a large gain but whose distribution is concentrated at a delta function.

The loss due to "gain jitter" suggests a possible remedy. The physical pulse $g_n w(t - t_n)$ in the detection circuits following the avalanche diode should be clearly detectable against the background noise if g_n is sufficiently large; in particular, if it is something like the mean gain G. Suppose this is also true for pulse gains $g_n \geqq fG$, $f < 1$. Now suppose one processed the circuit output of the avalanche diode by first passing it through a pulse detector that detects pulses of height greater than fG and generates a pulse of fixed height if a pulse is detected. The output pulses of this device have fixed gain, which is beneficial, but, on the other hand, we have lost a fraction θ,

$$\theta = \frac{1}{G} \int_0^{fG} \exp(-g/G) dg, \tag{39}$$

of light intensity. Seemingly, by a simple scheme we may have still gained a dB or two in performance. Because of effects such as possible overlap of two close pulses $w(t)$ and even in the pulse shape of $w(t)$ itself, the merits of this proposal are hard to assess without further study. It does appear to be an interesting possibility for a future detailed investigation.

IV. INTEGRATE-AND-DUMP DETECTION—OTHER AVALANCHE GAIN DISTRIBUTIONS

Personick[6] has considered the physics of a class of real avalanche detectors in considerable detail and has derived the following implicit equation for their moment-generating function $M_g(s)$:*

$$s = \ln M - \frac{1}{1-k} \ln [(1-a)M + a], \tag{40}$$

* We shall drop the subscript on the MGF M_g of the gain variable when we refer to the particular M_g given by (40). Also, the k in this section has nothing to do with the k in (35) and (36).

where we have set

$$M \equiv M(s) = \sum_{n=1}^{\infty} e^{sn} p_n. \tag{41}$$

The parameters k and a are related to the physical properties of these photon detectors. Since (40) has never been explicitly solved for $M(s)$, we think it worthwhile to investigate the structure of $M(s)$ implied by (40) in more detail. In addition to yielding structural properties of $M(s)$, we shall find that (40) allows us to determine the p_n of (41) exactly.

To begin with, the gain G, given by $G = Eg$, is

$$G \equiv Eg = \frac{d}{ds} M(s)\bigg|_{s=0}, \tag{42}$$

which, using (40), yields

$$G = \frac{1-k}{a-k}. \tag{43}$$

From (43) we see that the restrictions

$$\begin{aligned} 0 < a \leqq 1 \\ 0 \leqq k < a \end{aligned} \tag{44}$$

are to be imposed on the parameters in (40).

When $a = 1$, (40) gives $M = e^s$, the $g = 1$ case. When $k = 0$, (40) is easily solved to give

$$M(s) = \frac{ae^s}{1-(1-a)e^s}, \qquad k = 0. \tag{45}$$

Equation (45) is the MGF of the discrete geometric distribution having probabilities p_n concentrated on the positive integers, where

$$p_n = \frac{a}{1-a}(1-a)^n, \qquad n = 1, 2, \cdots. \tag{46}$$

It is reasonable to treat the continuous version of this density, and that was done in Section II.

In the general case of (40), the variance may be calculated to give

$$\operatorname{var} g = G^3\left[1 - \frac{(1-a)^2}{1-k}\right] - G^2. \tag{47}$$

If higher moments are desired, they can be obtained recursively from (40). This can be done by expanding $M(s)$ in a power series and equating like powers in s.

In view of the discussion in Section III, one might prefer the detectors represented by (40) that have small variance. A simple investigation of (47) reveals that, for any $a < 1$, $k = 0$ uniquely gives minimum variance. Since even this minimum variance is large (equal to the mean), it may well not be a reliable guide.

Returning to the general case represented in (40), it is evident from the relation

$$M(s) = \sum_{n=1}^{\infty} e^{sn} p_n$$

that the MGF exists for all $s \leqq 0$. However, it does not exist for all positive s, and, in fact, setting $ds/dM = 0$ yields a critical value of M (call it M_c) given by

$$M_c = \frac{a}{1-a} \frac{1-k}{k} \tag{48}$$

and thus a critical value s_c of s given by

$$s_c = s(M_c) = \ln \frac{1-k}{1-a} - \frac{k}{1-k} \ln \frac{a}{k}, \tag{49}$$

beyond which $M(s)$ does not exist. Note that, if $b \neq 0$ (and $a \neq 1$), the value of the MGF at the critical s is finite. This shows that the far-tail behavior of the g variable has an exponential-like tail, with damping factor related to s_c, but in general there is a multiplicative factor, e.g., an inverse power that allows the MGF to be finite at its critical value.

If we let $s_c - s = \delta > 0$, $M_c - M(s) = \Delta > 0$, and write

$$s_c - s \cong s(M) = s(M_c - \Delta) = s(M_c) - \Delta \left. \frac{ds}{dM_g} \right|_{M_c} + \frac{1}{2} \Delta^2 \left. \frac{d^2s}{dM_g^2} \right|_{M_c} + \cdots, \tag{50}$$

we obtain, after evaluating the second derivative in (50), that

$$\Delta \cong \sqrt{\delta} \sqrt{\frac{2}{k}} M_c \tag{51}$$

or, equivalently,

$$M \approx M_c \left[1 - \sqrt{\frac{2}{k}} \sqrt{s_c - s} \right], \tag{52}$$

thus exhibiting a square-root singularity of $M(s)$ in the neighborhood of s_c. This type of behavior is consistent with a far-tail fall-off of the "density" of the g variable being given by

$$\text{const.} \frac{\exp(-s_c g)}{g^{\frac{3}{2}}}. \tag{53}$$

Let us now proceed to the exact solution for the p_n in (41) when $M(s)$ is given by (40). We use instead $z = \exp(s)$ and write, with a slight abuse of notation,

$$M(z) = \sum_{n=1}^{\infty} z^n p_n. \tag{54}$$

Equation (40) becomes, setting $M = M(z)$ when convenient,

$$z = \frac{M}{[M(1-a)+a]^{1/(1-k)}}. \tag{55}$$

In (55) it is useful to make the substitutions

$$M(z) = \frac{a}{1-a} F[(1-a)a^{k/(1-k)}z] \tag{56a}$$

$$u = (1-a)a^{k/(1-k)}z \tag{56b}$$

$$\rho = \frac{1}{1-k} \tag{56c}$$

to obtain

$$u = \frac{F}{[1+F]^{\rho}}, \tag{57}$$

where $F(0) = 0$ and F is regarded as an implicit function of u in the neighborhood of $u = 0$. Equation (57) is a canonical form for the Lagrange inversion formula[7] for obtaining the coefficients c_j in the power series

$$F = \sum_{j=1}^{\infty} c_j u^j. \tag{58}$$

The formula yields, for the present problem,

$$c_j = \frac{1}{j!}\left\{\left(\frac{d}{dF}\right)^{j-1}(1+F)^{\rho j}\right\}_{F=0} \tag{59}$$

or

$$c_1 = 1,$$

$$c_j = \frac{\prod_{s=0}^{j-2}(j\rho - s)}{j!} = \frac{\Gamma[j/(1-k)+1]}{\Gamma(j+1)\Gamma[kj/(1-k)+2]}, \qquad j \geqq 2. \tag{60}$$

From (54) and (56), the probabilities p_j are then given by

$$p_j = \frac{a}{1-a}[(1-a)a^{k/(1-k)}]^j c_j. \tag{61}$$

For (kj) large, we have, from Stirling's asymptotic formula for the gamma function,

$$\Gamma(z + 1) \sim e^{-z}z^{z+\frac{1}{2}}\sqrt{2\pi}, \tag{62}$$

that

$$c_j \sim \frac{1}{\sqrt{2\pi}} \frac{1}{(kj)^{\frac{1}{2}}(1 - k)^{j-1}} \frac{1}{(k^{k/(1-k)})^j}, \qquad \text{as} \quad kj \to \infty. \tag{63}$$

One can show that the behavior given in (63) is, via (61), in complete agreement with (49) and (53).

Remarkably, Personick reports that McIntyre,* from special-case calculations, has conjectured the exact form of (61).

Knowing the p_j does, in principle, allow the exact calculation of the output statistics of the integrate-and-dump filter. The integral equation (16), appropriately interpreted with sums, provides one such way. Instead of discussing this, however, we now turn our attention to bounding techniques. We shall make some remarks directed toward the Chernoff bound, used by Personick[6] for this type of problem.

The Chernoff bound states that, if x has MGF $M_x(s)$, then the probability that x is greater than (less than) F obeys

$$\Pr[x \underset{(<)}{>} F] \leqq \exp(-sF)M_x(s) \qquad \text{for any} \quad s \underset{(<)}{>} 0. \tag{64}$$

One makes the bound as tight as possible by minimizing the right member of (64) over s. This, of course, assumes that $M_x(s)$ is known or can be obtained explicitly as a function of s. For the general class of avalanche diodes for which Personick derives the moment-generating function, we saw that s is given explicitly as a function of M and, in fact, an explicit function of M vs s is difficult to obtain analytically. Personick gets M numerically as a function of s and then proceeds to optimize with respect to s—a rather tedious procedure. We found from our experience that a simpler approach is to eliminate s in (64) by using (40) and then to optimize over M. This optimization still has to be done numerically. Nevertheless, we could generate curves very quickly this way. We do not present these curves here, since they do not reveal more than those which Personick has already published.

For insight concerning the accuracy of the bound for present purposes, we shall apply it below to the problem of exponential gains, for which we have exact solutions available for comparison.

The function appearing in the right member of (64) is, for the exponential gain case,

$$\exp(-sF) \exp\{\Lambda[1/(1 - s) - 1]\}. \tag{65}$$

* In addition to the cited reference of Personick, other experimental properties of avalanche photodiodes may be found in Webb, McIntyre, and Conradi (Ref. 8).

Finding the optimum s is easy in this case, and (64) then yields, for these optimum s,* $s_{\text{opt}} = 1 - \sqrt{\lambda/F}$, and, consequently,

$$\begin{aligned} P[x > F > \Lambda] &\leqq \exp\,[-(\sqrt{\Lambda} - \sqrt{F})^2] \\ P[x < F < \Lambda] &\leqq \exp\,[-(\sqrt{\Lambda} - \sqrt{F})^2]. \end{aligned} \tag{66}$$

From the asymptotic forms of (26) and (28), we see that the Chernoff bound has given us the "right exponent."

From saddle-point considerations, this would be expected to be true in this problem for any $M_g(s)$; however, it by no means has to be true in general, where complex variable (saddle-point) techniques must be resorted to in order to decide the question.

The optimum threshold for single-bit detection that would be obtained by equating the two expressions in (64) (for different Λ's, of course) also results in (30). Table III lists the Chernoff upper bounds to the bit error rate, and these should be compared to the exact answers shown in Table I. Numerically, the Chernoff bound is off by one to two orders of magnitude in error rate due to "coefficient effects." However, even numerically this bound is judged to perform respectably. Also shown in Table III is $s_{\text{opt}} = \alpha[1 - \sqrt{\Lambda/F}]$, where the gain ($\alpha$) effect has been included. For the optimum choice of F, it turns out that the two choices of s_{opt} (due to two possible Λ's) are the negative of each other. Hence, only the positive one is shown in Table III.

If one wishes to include the effects of gaussian noise here, one multiplies the right-hand side of (64) by the appropriate MGF, namely, (10). Instead of finding the optimum s for this problem, one can use the s_{opt} that held for the problem without additive noise (any s of appropriate sign furnishes a bound). The value $\sigma^2 = 10^4$ was used in further Chernoff bound calculations for the $M_g(s)$ given in (41) and may be found in the article by Personick.[6]

V. INTERSYMBOL INTERFERENCE—INTEGRATE-AND-DUMP FILTER

We turn now to the situation where $\lambda(t)$ is given by (2), i.e., a train of interfering pulses instead of just one of them. Personick has claimed that $h(t)$ has a gaussian shape in real fibers and, hence, in practice only a few pulses would be expected to contribute intersymbol interference.

It is evident that, if the filter $P(t)$ that processes the output of the photon detector is always positive, as, for example, for an integrate-and-dump filter, the presence of intersymbol interference increases

* In setting the derivative equal to zero, one must choose the positive s that satisfies $s < 1$, since in the real-variable techniques used here, the MGF of the exponential does not exist for $s \geqq 1$.

Table III — Tabulation of Chernoff bound (CB) for error rate, exponential gain case. Also given are $s_{\text{opt}} = \alpha[1 - \sqrt{\Lambda/F}]$ for a gain $1/\alpha = 100$, and a correction $\exp[s^2_{\text{opt}}\,\sigma^2/2\}$ for $\sigma^2 = 10{,}000$. The latter is a correction for gaussian noise.

Λ_0	Λ_{1s}	CB [eq. (66)]	s_{opt} (gain = 100)	$\exp \frac{s^2_{\text{opt}}\sigma^2}{2}$
4	100	5.05×10^{-8}	6.72×10^{-3}	1.25
4	200	4.16×10^{-17}	7.54×10^{-3}	1.33
4	400	2.70×10^{-36}	8.19×10^{-3}	1.40
10	100	1.49×10^{-6}	5.37×10^{-3}	1.16
10	200	1.16×10^{-14}	6.42×10^{-3}	1.23
10	400	2.01×10^{-32}	7.30×10^{-3}	1.31

the counts observed over any interval. Therefore, if a pulse is present, this intersymbol interference helps detection (helps keep output greater than the threshold) while, if the pulse is absent and no-counts is ideal, it hurts. Hence, the worst-case situation is to evaluate the probability of a one being decoded into a zero when no other pulses are present, while for the reverse error we assume all pulses are on.

Since we are still considering an integrator, i.e., $P(t) = 1$, $|t| < \tau$, we are still to use (6), but now for the two worst cases given we replace Λ in (6) by either

$$\Lambda_1 = 2\tau\lambda_0 + \int_{-\tau}^{\tau} h(t)dt$$

or (67)

$$\Lambda_0 = 2\tau\lambda_0 + \sum_{n\neq 0} \int_{-\tau}^{\tau} h(t - nT)dt.$$

Of course, we assume $\Lambda_0 < \Lambda_1$ for any reasonable operating situation. In addition to the threshold choice, we must also contend with the optimum choice of τ, half the time width of the integration. This latter step is easily handled numerically.

Many calculations may be done and, for the worst-case situation described, nothing new is involved in addition to what has already been discussed. As an illustration, we will deal explicitly with one example. We take $\lambda_0 = 0$, no avalanche gain ($g = 1$), and

$$h(t) = \frac{100}{T}\left[1 - \frac{|\tau|}{T}\right], \tag{68}$$

where T is the pulse repetition rate. Thus, there is considerable overlap from neighboring pulses, but not from others. Also, $\int h(t)dt = 100$,

Table IV — An intersymbol interference example from Section V

$\frac{\tau}{T}$	F	P_e [eq. (35)]
0.1	7	7.3×10^{-5}
0.2	15	1.5×10^{-5}
0.3	25	5.7×10^{-6}
0.4	34	1.5×10^{-5}
0.5	46	5.1×10^{-5}

so the quantum limit for single-pulse detection may be read from Table I.

Table IV gives the worst-case error rate for the above example, using formulas (35) and (38) for the Poisson case. The optimum choice of τ here is 0.3, i.e., 30 percent toward the peak of the neighboring pulse. Also, a 20-percent change in the value of τ does not change the error rate drastically. Note that we are not inferring that one should be careless in the choice of τ, because in calculating Table IV the optimum threshold (F) for each τ is assumed. Also, note the large degradation with respect to the quantum limit caused by the intersymbol interference. For the present example, the error rate averaged over all sequences cannot be much better than shown, because the worst case occurs with probability $\frac{1}{4}$, and hence $(P_e)_{\text{av}}$ cannot be more than a factor of 4 better.

VI. AN INTERSYMBOL INTERFERENCE EXAMPLE AND A LOWER BOUND ON PERFORMANCE

We present now a lower bound on performance which can be readily evaluated for the intersymbol interference problem of the last section [pulses given by (68)]. This lower bound is valid for optimum bit detection and thus sets a limit on how well *any* detector can do in coping with intersymbol interference. In particular, the bound sheds light on the performance in the present situation of suboptimum schemes such as equalization, which have found such wide application in voiceband telephone transmission.

The derivation of the lower bound proceeds along lines used by Mazo[9] to generalize Forney's lower bound for optimum bit-by-bit detection in the gaussian noise. Our approach is to assume that we are optimally detecting the kth bit in a sequence of $(N+1)$ independent bits, i.e., sequences of the form (2) of length $(N+1)$ are being considered. We suppose a_n are binary, equiprobable, and independent. Let $p_1(x|i)$ and $p_0(x|i)$ be the two probability densities of the received signal under the hypotheses $a_n = 1$ or 0, respectively, and i denote conditioning on the ith, $i = 1, \cdots, 2^N$ sequence being transmitted. Then the probability of error for the optimum detector is (in somewhat formal notation)

$$P_e = \frac{1}{2} \int dx \min \left[\frac{1}{2^N} \sum_{i=1}^{2^N} p_1(x|i), \frac{1}{2^N} \sum_{j=1}^{2^N} p_0(x|j) \right], \tag{69}$$

which, as in Ref. 9, can be lower-bounded by

$$P_e \geqq \frac{1}{2^N} P_e \text{ (binary } i, j \text{ problem)}. \tag{70}$$

In (70), P_e (binary i, j problem) is the probability of error which would result for the simple binary problem of distinguishing between sequence i (one having $a_k = +1$) from sequence j (one having $a_k = 0$). The bound (70) holds for all such (i, j) pairs. Finally, (70) holds if the sequences of length $(N + 1)$ are shortened to $N' + 1$, with N being replaced by N' on the right side of (65).

For communication in the Poisson regime, the right member of (70) has no known evaluation as it does for the gaussian case. What is known about the binary problem is the optimum detector, which is linear. The optimum filter $P(t)$ and threshold F are known explicitly if one is deciding between equiprobable intensity functions $\lambda_a(t)$ and $\lambda_b(t)$. In fact, from the work of Bar-David,[10]

$$P(t) = \ln \frac{\lambda_a(t)}{\lambda_b(t)} \tag{71}$$

and

$$F = \int \lambda_a(t) - \int \lambda_b(t). \tag{72}$$

Thus, the set of received impulses is filtered through $P(t)$ and the resulting output variable X at the end of the observation interval is compared to the threshold F, choosing $\lambda_a(t)$ if $X > F$ and $\lambda_b(t)$ otherwise. Assuming $\lambda_a(t)$ is transmitted, the moment-generating function of X is, from (5) and (71) (recall $g = 1$ in this section),

$$\begin{aligned} M_x(s) &= \exp \left[\int \lambda_0(t) [\exp \{ s \ln [\lambda_1(t)/\lambda_0(t)] \} - 1] dt \right] \\ &= \exp \left[\int [\lambda_0^{1-s}(t) \lambda_1^s(t) - \lambda_0(t)] dt \right]. \end{aligned} \tag{73}$$

From this MGF, one can see why the right side of (70) is not known in general.

We now apply (73) to the intersymbol interference of the previous section, where $h(t)$ is given by (68). We choose $N = 2$, $\lambda_1(t)$ to correspond to the pulse sequence (1, 1, 1) and $\lambda_b(t)$ to correspond to the pulse sequence (1, 0, 1). When applied to (70), we interpret the results as applying to the center bit of the sequence. We have, explicitly,*

$$\begin{aligned} \lambda_1(t) &= 1 && \text{for} \quad |t| \leqq 1 \\ \lambda_0(t) &= |t|, && \text{for} \quad |t| \leqq 1 \\ \lambda_1(t) &= \lambda_0(t) && \text{for} \quad |t| > 1. \end{aligned} \tag{74}$$

Since, from (71), $P(t) = 0$ for $|t| > 1$, the detection interval $t \in [-1, 1]$. Using (74) in (73), the decision variable has MGF

$$M_x(s) = \exp\left[\frac{2}{2-s} - 1\right]. \tag{75}$$

Remarkably enough, this is the moment-generating function of the random variable dealt with in Section II; in the notation of that section, it corresponds to $\Lambda = 1$, $\alpha = 2$. The density is given by (20), and the threshold is, from (72) and (74), to be set equal to unity. Putting this all together, (70) becomes

$$P_e \geqq \frac{1}{4} e^{-1} \int_1^\infty \sqrt{\frac{2}{x}}\, e^{-2x} I_1(2\sqrt{2x})\, dx. \tag{76}$$

Or, scaling (76) to reinsert the factor of 100 in front of (68),

$$P_e \geqq \frac{1}{4} e^{-100} \int_{100}^\infty \sqrt{\frac{200}{x}}\, e^{-2x} I_1(2\sqrt{200x})\, dx. \tag{77}$$

So an excellent approximation in the right-hand side of (77) may be evaluated via (26) to give

$$P_e \geqq \tfrac{1}{4}(\tfrac{1}{2})^{\frac{1}{4}} Q(\sqrt{400} - \sqrt{200}) \approx 5.06 \times 10^{-10}. \tag{78}$$

The numerical value of (78) should be compared with Table IV for performance with integrate-and-dump filter and Table I for the quantum limit. Indeed, for this case our bound shows that the optimum detector performance is still far from the quantum limit and, in fact, is roughly only 2.2 dB (comparing powers of 10) better than the integrate-and-dump filter.† The present problem seems to imply that equalization,‡ in particular, cannot be expected to approach the quantum limit bound for the type of distortion found in present optical fibers. In fact, a simple integrate-and-dump receiver with properly

* For the moment, we ignore the factor of 100 in (68) and also set $T = 1$. These are reintroduced only in the final numerical calculations.

† More precisely, the figure is 2.9 dB for strong signals.

‡ Some references on equalization for optical communication systems are Refs. 1 and 11.

chosen threshold compares well with a lower performance bound. The above problem ignored many practical factors, but in fact ignoring them focused even more on the pure intersymbol interference problem in the Poisson regime. It would seem that effects such as dark current and finite width of $w(t)$ would surely make the integrate-and-dump and the optimum detector perform even more equally, and it would seem too much for an equalizer to compensate for gain jitter, which is a rather nonlinear effect.

Another linear filter $P(t)$, which performs better than the integrate-and-dump, may be inferred from (71). This is discussed and evaluated in the appendix for the present problem. This new linear filter has a worst-case exponent approximately 1 dB better than the integrate-and-dump situation.

APPENDIX

A New Filter

We have already noted that (asymptotically) the integrate-and-dump filter performs within 2.9 dB of a lower bound on performance for the optimum processor for our particular example. We now show how a modified $P(t)$ can perform within 2 dB of this bound. We confine ourselves to the worst case again, for which, we recall, the best integrator had $P(t) = 1$ for $|t| \leqq 0.3$ (choosing $T = 1$). The worst case with signal present was $\lambda_1(t) = 1 - |t|$, $|t| < 1$, and $\lambda_0(t) = |t|$, $|t| < 1$, for the worst case with signal absent. Now the optimum filter

$$P(t) = \ln \frac{1 - |t|}{|t|}, \qquad |t| < 1, \tag{79}$$

which distinguishes between these two signals, is not always positive (it is negative for $|t| > \frac{1}{2}$). Therefore, if (79) were used, there could be no claim for a worst-case bound. However, we modify (79) and use

$$P(t) = \ln \frac{1 - |t|}{|t|}, \qquad |t| < \tfrac{1}{2} \tag{80}$$

instead. The filter represented by (80) is always positive, and therefore worst-case claims still obtain. The filter (80) clearly has to outperform our integrate-and-dump one, since the latter integrated only to $|t| = 0.3$, while (80) is optimum for an observation interval $|t| \leqq 0.5$. The optimum threshold for (80) is, from (72),

$$F = \int_{-\frac{1}{2}}^{\frac{1}{2}} \lambda_1(t)dt - \int_{-\frac{1}{2}}^{\frac{1}{2}} \lambda_0(t)dt = \tfrac{1}{2}. \tag{81}$$

Using the Chernoff bound for the case when $\lambda_0(t)$ is sent, we have, from

(64) and (73),

$$P_e \leqq \exp\left\{\int \lambda_0(t)[e^{sP(t)} - 1] - sF\right\}$$

$$= \exp\left\{2\int_0^{\frac{1}{2}} t^{1-s}(1-t)^s dt - \frac{1}{4} - \frac{s}{2}\right\}, \qquad s > 0, \tag{82}$$

where we have used the expression for $\lambda_0(t)$, the filter (80), and threshold (81). If we let $u = (1-t)/t$, then we may write

$$\int_0^{\frac{1}{2}} t^{1-s}(1-t)^s dt = \int_1^\infty \frac{u^s}{(1+u)^3} du. \tag{83}$$

Two integrations by parts give

$$\int_1^\infty \frac{u^s}{(1+u)^3} du = \frac{1}{8} + \frac{s}{4} + \frac{s(s-1)}{2} \int_1^\infty \frac{u^{s-2}}{1+u} du, \tag{84}$$

or, using (84) in (82),

$$P_e \leqq \exp\left[s(s-1)\int_1^\infty \frac{u^{s-2}}{1+u} du\right]. \tag{85}$$

Equation (85) makes it evident that the exponent in (82) will be negative for $0 < s < 1$. If we expand the $1/(1+u)$ part of the integrand in (80) in powers of $(1/u)$ and integrate term by term, the exponent in (85) becomes

$$s(s-1)\sum_{k=0}^\infty \frac{(-1)^k}{k+2-s}$$

$$= s(s-1)\sum_{\substack{k=0 \\ k \text{ even}}}^\infty \frac{1}{(k+2-s)(k+3-s)}. \tag{86}$$

Convergence in (86) can be improved if we write

$$\sum_{k \text{ even}} = \tfrac{1}{2}\sum_{\text{all } k} + \tfrac{1}{2}\sum_{k \text{ even}} - \tfrac{1}{2}\sum_{k \text{ odd}}$$

and use the fact that

$$\sum_{n=1}^\infty \frac{1}{(x+n)(x+n+1)} = \frac{1}{1+x}$$

to obtain

$$s(s-1)\left[\frac{1}{2(2-s)} + \sum_{\substack{k=0 \\ k \text{ even}}}^\infty \frac{1}{(k+2-s)(k+3-s)(k+4-s)}\right]. \tag{87}$$

The optimum s is easily found numerically by plotting (87); we truncated the sum after $k=10$. We find the optimum s is about 0.6, giving a value of (87) of 0.11138. As a check on the possible accuracy of our use of (87), we note that our technique gives 0.10696 when $s = \frac{1}{2}$, for which the exact answer can be shown to be $\pi/8 - \frac{1}{2} \approx$ 0.10730. Thus, the Chernoff bound is

$$\begin{aligned} P_e &\leqq \exp(-0.111\Lambda_0) \\ \Lambda_0 &= \int_{-1}^{1} \lambda_0(t)dt, \end{aligned} \tag{88}$$

while (73) yields as a lower bound something which behaves exponentially as

$$\exp\left(-\Lambda_0\left[\frac{(\sqrt{4}-\sqrt{2})^2}{2}\right]\right) = \exp(-0.172\Lambda_0). \tag{89}$$

The exponent of (88) is 1.9 dB worse than that of (89). Concluding, we note that (80) has a logarithm singularity at $t = 0$. Including dark current in the $\lambda_i(t)$ will remove this, and will also decrease the improvement which this kind of filter provides over the integrate-and-dump filter.

REFERENCES

1. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems, I," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 843–874.
2. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems, II," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 875–886.
3. G. J. Foschini, R. D. Gitlin, and J. Salz, "Optimum Direct Detection for Digital Fiber-Optic Communication Systems," B.S.T.J., *54*, No. 8 (October 1975), pp. 1389–1430.
4. S. D. Personick, "Baseband Linearity and Equalization in Fiber Optic Digital Communication Systems," B.S.T.J., *52*, No. 7 (September 1973), pp. 1175–1194.
5. W. M. Hubbard, "Comparative Performance of Twin-Channel and Single-Channel Optical-Frequency Receivers," IEEE Trans. Commun., *COM-20*, No. 6 (December 1972), pp. 1079–1086.
6. S. D. Personick, "New Results on Avalanche Multiplication Statistics with Applications to Optical Detection," B.S.T.J., *50*, No. 1 (January 1971), pp. 167–190.
7. N. G. DeBruijn, *Asymptotic Methods in Analysis*, Amsterdam: North-Holland, 1961. See p. 22, 2nd edition.
8. P. P. Webb, R. J. McIntyre, and J. Conradi, "Properties of Avalanche Photodiodes," RCA Review, *35*, June 1974, pp. 234–276.
9. J. E. Mazo, "Faster-than-Nyquist Signaling," B.S.T.J., *54*, No. 8 (October 1975), pp. 1451–1462. See Section II.
10. I. Bar-David, "Communication Under the Poisson Regime," IEEE Trans. Inform. Theory, *IT-15*, No. 1 (January 1969), pp. 31–37.
11. David G. Messerschmitt, "Optimum Mean-Square Equalization for Digital Fiber Optic Systems," *International Conference on Communications Conference Record*, Vol. III, paper 43, June 1975.

Contributors to This Issue

David D. Falconer, B.A.Sc., 1962, University of Toronto; S.M., 1963, and Ph.D., 1967, Massachusetts Institute of Technology; post-doctoral research, Royal Institute of Technology, Stockholm, 1966–67; Bell Laboratories, 1967—. Mr. Falconer has worked on problems in coding theory, communication theory, channel characterization, and high-speed data communication. Member, Tau Beta Pi, Sigma Xi, IEEE.

Michael J. Gans, B.S. (E.E.), 1957, Notre Dame University, M.S., 1961, Ph.D. (E.E.), 1965, University of California, Berkeley; Bell Laboratories, 1966—. At Bell Laboratories, Mr. Gans has been engaged in research on antennas for mobile radio and satellite communications.

John A. Lewis, B.S., 1944, Worcester Polytechnic Institute; M.S., 1948, and Ph.D., 1950, Brown University; Bell Laboratories, 1951—. Mr. Lewis has worked on problems in piezoelectricity, heat conduction, and electroplating. He is currently concerned with optical fiber drawing. Member, American Mathematical Society, Society for Industrial and Applied Mathematics, Mathematical Association of America.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. (Physics), 1963, Syracuse University; Research Associate, Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964—. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now working in the Mathematical Research Center. Member, American Physical Society, IEEE.

James McKenna, B.Sc. (Mathematics), 1951, Massachusetts Institute of Technology; Ph.D. (Mathematics), 1961, Princeton University; Bell Laboratories, 1960—. Mr. McKenna has done research in quantum mechanics, electromagnetic theory, and statistical mechanics. He has recently been engaged in the study of nonlinear partial differential equations that arise in solid-state device work, in the theory of stochastic differential equations, and the theory of elastic wave propagation.

Jack Salz, B.S.E.E., 1955, M.S.E., 1956, and Ph.D., 1961, University of Florida; Bell Laboratories, 1961—. Mr. Salz first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems, and is currently a supervisor in the Advanced Data Communications Department. During the academic year 1967–1968, he was on leave as Professor of Electrical Engineering at the University of Florida. Member, Sigma Xi.

Aaron D. Wyner, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, Columbia University; Bell Laboratories, 1963—. Mr. Wyner has been doing research in various aspects of information and communication theory and related mathematical problems. He is presently Head of the Communications Analysis Research Department. He spent the year 1969–1970 visiting the Department of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, and the Faculty of Electrical Engineering, the Technion, Haifa, Israel on a Guggenheim Foundation Fellowship. He has also been a full- and part-time faculty member at Columbia University and the Polytechnic Institute of Brooklyn. He has been chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group, has served as an associate editor of the Group's *Transactions,* and has served as cochairman of two international symposia. He is presently president of the IEEE Information Theory Group. Fellow, IEEE, member, AAAS, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.